Universidad Autónoma de Zacatecas

"Francisco García Salinas"

Unidad Académica de Ingeniería Eléctrica

Asociación de descriptores demográficos y dietéticos con caries dentales para el desarrollo de una herramienta de diagnóstico preventivo

Laura Alejandra Zanella Calzada

Tesis para obtener el grado de

Maestría en Ciencias de la Ingeniería

Con orientación en

Procesamiento de Señales y Mecatrónica

Zacatecas, Zacatecas, 08 de febrero de 2019

Universidad Autónoma de Zacatecas

"Francisco García Salinas"

Unidad Académica de Ingeniería Eléctrica

Asociación de descriptores demográficos y dietéticos con caries dentales para el desarrollo de una herramienta de diagnóstico preventivo

Laura Alejandra Zanella Calzada

Tesis para obtener el grado de

Maestría en Ciencias de la Ingeniería

Con orientación en

Procesamiento de Señales y Mecatrónica

Asesor:

Dr. Carlos E. Galván Tejada

Coasesores:

Dr. Jorge I. Galván Tejada

Mtra. Nubia M. Chávez Lamas

Zacatecas, Zacatecas, 08 de febrero de 2019

Dedicatoria

A mi mamá.

Al igual que cada logro, este es tan tuyo como mío, te amo.

También a Camen y a Meli.

Las amo, siempre están en mi mente y en mi corazón.

Agradecimientos

Agradezco por el inmenso apoyo y compresión de mi mamá, que siempre respeta y respalda mis decisiones, me da sus consejos de mamá y me hace querer ser mejor cada día. Gracias por tu ejemplo de fortaleza, perseverancia, humildad, lucha y crecimiento. Gracias por siempre recordarme tener los pies en la tierra. Gracias por siempre impulsarme a volar alto.

También le agradezco a mis asesores, Eric y Jorge, primero por impulsarme en este posgrado, además de por todo el apoyo, la enseñanza y la paciencia, por dejarme ser parte de sus proyectos y por hacerme crecer no solo como alumna, también como persona.

A la maestra Nubia, por toda su ayuda y apoyo en este proyecto, por guiarnos en una nueva área y trabajar en equipo con nosotros.

Agradezco a los doctores que fueron parte de mi aprendizaje en la maestría, al doctor José María, Villa, Ismael, Ireri, Claudia, Arceo, por mencionar algunos.

A mis compañeros y amigos, porque me ayudaron a sentirme como en casa y a disfrutar mucho más de esta etapa, especialmente gracias a Andrea, David, Brenda, Iván, Ivan, Vero, Luis.

También a mis amigos potosinos que siempre me hacen sentir su apoyo y su cariño, Erick (gracias por tanta ayuda y apoyo en todo, sobretodo gracias por tanta motivación), Saúl, Mario, Chipi, Juanma, Tavo, Elvira.

Gracias a todos por haber estado y seguir estando.

Resumen

La salud oral representa un componente esencial en la calidad de vida de las personas, siendo un factor determinante en la salud general, tomando en cuenta que puede incrementar el riesgo de sufrir otros padecimientos, entre los que se encuentran algunas enfermedades crónicas. En la actualidad, las enfermedades orales son un problema de salud pública, siendo la caries dental la que presenta mayor incidencia, ocurriendo en alrededor del 90 % de la población mundial, presentándose principalmente en las regiones menos desarrolladas. La alta incidencia de este padecimiento se debe a la gran cantidad de factores que la afectan, incluyendo desde procesos biológicos hasta complejas estructuras histórico-culturales y sociales, los cuales interactúan de manera simultánea, provocando que el control de esta enfermedad se vuelva altamente complicado.

En este trabajo se presentan una serie de investigaciones en las que se hizo la búsqueda de la relación entre un conjunto de descriptores demográficos y dietéticos, obtenidos de NHANES 2013 – 2014, y la condición de caries dental. Para el desarrollo de estos trabajos fueron utilizadas diferentes técnicas de inteligencia artificial, tales como algoritmos de aprendizaje automático, algoritmos genéticos, redes neuronales artificiales y aprendizaje profundo.

De acuerdo con los resultados que se presentan, es posible conocer que existe una evidente relación entre los descriptores demográficos y dietéticos con la caries dental, principalmente entre aquellos que contienen información sobre el estado socio-económico de los sujetos, además de la edad, el género, el origen étnico y la ingesta diaria de nutrientes. Además, comparando las diferentes técnicas utilizadas en cada investigación, se puede observar que si los sujetos se separan por grupos de edad es posible seleccionar características específicas que afectan a cada grupo, permitiendo una mejor clasificación de sujetos caso y sujetos control. Por otro lado, si la selección de características se lleva a cabo utilizando una técnica basada en un algoritmo genético, los resultados obtenidos presentan la relación de sensibilidad - especificidad más significativa estadísticamente.

En conclusión, las herramientas computacionales utilizadas en este trabajo permiten analizar información con el fin de encontrar complejas relaciones entre los datos. De esta manera se presenta una serie de trabajos preliminares para el posible desarrollo de una herramienta de diagnóstico asistido por computadora, basada en determinantes demográficos y dietéticos, que podría apoyar a los especialistas en la reducción de la alta prevalencia que presenta la condición de caries dentales, especialmente en las regiones menos desarrolladas.

Palabras clave: Salud oral; Caries dental; Descriptores demográficos; Descriptores dietéticos; NHANES; Inteligencia artificial; Diagnóstico asistido por computadora.

Lista de Figuras

Lista de Tablas

Contenido

1. Introducción

Las enfermedades crónicas son los principales problemas de salud pública en la mayor parte del mundo. El patrón de enfermar se ha transformado y las enfermedades bucodentales son consideradas como uno de los principales problemas de salud pública debido a su alta incidencia y prevalencia en todas las regiones del mundo, y como en todas las enfermedades, la mayor carga es en las poblaciones desfavorecidas y marginadas socialmente, representando un problema, ya que el tratamiento de estas enfermedades es extremadamente caro y no es fiable para la mayoría de los países de bajos y medianos ingresos [1].

La salud oral es de gran influencia en la calidad de vida de las personas, siendo un factor determinante en la salud general de los individuos, convirtiéndose en un punto importante en el cuidado de la salud. Por lo tanto, las serias repercusiones que podrían implicar las enfermedades orales, tales como dolor, sufrimiento, deterioro del funcionamiento y efectos en la calidad de vida, deben de ser consideradas.

Un aspecto importante a tomar en cuenta es que estas enfermedades incrementan el riesgo de sufrir enfermedades crónicas, tales como enfermedades cardiovasculares, cerebrovasculares, diabetes mellitus y enfermedades respiratorias [2].

Dentro de las enfermedades orales, la caries dental es el padecimiento más frecuente y se define como una enfermedad crónica evitable de origen multifactorial que afecta los tejidos duros de los dientes, presentándose con una prevalencia de alrededor del 60 al 90 % en todo el mundo [3, 4].

Hay una cantidad significativa de factores de riesgo y determinantes relacionados con la presencia de caries dental, y entre mayor sea el grado de exposición a estos factores,

mayor será la probabilidad de contraer o desarrollar esta condición.

Los determinantes y las condiciones de la caries dental son de categorías multicausales, multisectoriales e interdisciplinarias que engloban una serie de situaciones relacionadas con los procesos históricos y políticos que cada país experimenta. Además, se ha descrito cómo el desarrollo de esta condición depende de la frecuencia del consumo de ciertos alimentos, las características de la comida, el tiempo de exposición, entre otros, incluyendo la dificultad del acceso a servicios médicos dentales especializados. Estos factores regularmente interactúan de manera simultánea, encontrándose entre ellos variables de diferentes órdenes, desde procesos biológicos hasta estructuras histórico-culturales complejas y relaciones sociales, convirtiendo a la salud oral en un fenómeno complejo [5, 6].

Por lo tanto, debido a la dificultad que representa el control de la incidencia de esta enfermedad, se han desarrollado una serie de estudios en donde se implementan algoritmos y se desarrollan diferentes análisis a través de diagnóstico asistido por computadora (CADx, por sus siglas en inglés), donde se utilizan modelos de predicción para conocer el comportamiento futuro de los complejos datos relacionados con la caries dental, con fines clínicos y de investigación [7].

Estos estudios se realizan con el fin de priorizar, identificar y resolver este problema de salud oral, además de que con los modelos de predicción que se generan es posible crear herramientas que favorezca el decremento de la incidencia y la prevalencia de la caries dental. Entre estas herramientas se encuentran una serie de estudios que permiten evaluar la calidad de vida y la relación con la salud oral, tales como, Dental Impact [8], Dental Impact on Daily Living [9], Oral Health Impact Profile [10] y Oral Impacts on Daily Performances [11], entre otros [12].

La *National Health and Nutrition Examination Survey* (NHANES) es un programa diseñado para evaluar la salud de adultos y niños en Estados Unidos de América (E.U.A.), en donde se realizan una serie de encuestas enfocadas en diferentes aspectos de la vida cotidiana, con el fin de recolectar información que pueda ser significativa para conocer y determinar los factores de riesgo de diferentes padecimientos, entre los que se encuentran la caries dental. NHANES tiene como objetivo el brindar información

que permita desarrollar herramientas para mejorar la calidad de vida de las personas, previniendo y reduciendo la prevalencia de estas condiciones.

Entre las herramientas de CADx que se han utilizado para el análisis de los factores que influyen en las caries dental, se encuentran las redes neuronales artificiales (RNA). Las RNAs son modelos no lineales semiparamétricos, que permiten la integración de variables y manejan con facilidad grandes cantidades de datos comparados con análisis lineales. Además, son un método que está implicado en procesos de construcción de sistemas que permiten clasificar, modelar y predecir información, siendo modelos matemáticos basados en un principio de aprendizaje que tiene su fundamento en los conceptos de inteligencia artificial (IA) y la respuesta biológica del cerebro humano. Tienen elementos de procesamiento, llamados neuronas, que son las unidades del sistema que pueden ser ajustadas o entrenadas a través de un proceso de aprendizaje y generalización. La información de cada de las neuronas es agrupada y procesada [13].

Por otro lado, un método que también ha sido utilizado en el desarrollo de estos modelos es el uso de algoritmos genéticos (AG), los cuales son procedimientos de búsqueda que se basan en el principio de evolución de selección natural, emulando este proceso usando mecanismos como la sobrevivencia del subconjunto de variables más efectivas, mutación y cruce para mejorar las combinaciones. De esta manera, los AG computacionales generan una serie de cromosomas, que se forman con variables medidas que sobreviven a procesos de evolución de generaciones que se van creando artificialmente. Los cromosomas que sobreviven finalmente, son modelos formados por una o múltiples variables que sobrevivieron durante las generaciones pasadas, basádonse en una función de costo que determina su aporte al cromosoma final [14].

Estos son ejemplos de herramientas de IA que se utilizan ampliamente para el estudio de diferentes enfermedades, entre las que se encuentran los padecimientos de salud oral.

Con base en esta descripción general del problema de salud oral, en el presente estudio se pretenden analizar los determinantes que afectan el problema que representa la caries dental, basado en factores dietéticos y demográficos, presentando como principal contribución el desarrollo de modelos predictivos a través de diferentes herramientas de

CADx, con el fin de desarrollar un modelo multivariado que permita conocer el futuro comportamiento de esta condición y así tener una herramienta de diagnóstico preventivo de bajo costo que apoye el decremento en la prevalencia e incidencia de esta condición.

1.1. Planteamiento del problema

La Organización Mundial de la Salud (OMS) definió a la salud como un estado físico, mental y social saludable, y no solo como la ausencia de enfermedades. Esta definición ha evolucionado desde ese concepto a una serie de escalas cuantitativas que permiten medir el estado general de salud. Por lo tanto, en 1994, la OMS propuso el concepto de calidad de vida como una percepción individual de la posición de vida, dentro del contexto de un ambiente cultural y la relación con los objetivos personales, expectativas, estándares y preocupaciones.

En la salud oral, los padecimientos bucodentales se han tornado en uno de los principales problemas de salud pública, en donde la caries dental está situada en el primer lugar de prevalencia.

Además, de acuerdo con la Organización Panamericana de la Salud (OPS), el desarrollo de este padecimiento depende de la frecuencia del consumo de ciertos elementos en la dieta, las características de estos elementos, el tiempo de exposición, escasas medidas preventivas en salud oral, entre otros, los cuales interactúan de manera simultánea.

Otro problema que representan las enfermedades orales, principalmente para las regiones desfavorecidas y socialmente marginadas, es el alto costo de los tratamientos, los cuales no son factibles en la mayoría de los países con bajos y medianos ingresos, ya que de acuerdo con la OMS, estos padecimientos son la cuarta causa más costosa de tratar. Por lo que la alta incidencia de este padecimiento se ve principalmente favorecida debido al restringido acceso que tienen estas regiones al servicio de salud y tratamiento de los padecimientos orales [15].

1.2. Justificación

Debido a la importancia que la salud oral representa en la calidad de vida de las personas, una gran cantidad de análisis sobre determinantes dietéticos, demográficos y sociales, que presentan influencia en la salud oral, se han convertido en un interesante nicho de estudio. Sin embargo, estos estudios se pueden ver afectados por diferentes condiciones que dependen de la región de los individuos, la herencia genética, e incluso de los servicios de salud y políticos de cada región.

El desarrollo de estos estudios se hace con el fin priorizar, identificar y resolver los problemas de salud oral, además de generar modelos de predicción que ayuden a reducir la incidencia y prevalencia de la enfermedades orales.

La caries dental es un desafío, ya que es una enfermedad evitable pero que se puede volver crónica, la cual ha permitido el desarrollo de instrumentos para el reconocimiento de los diferentes determinantes que contribuyen en la modificación negativa de la salud oral provocando esta condición, ya que de acuerdo a la literatura, su alta incidencia es consecuencia de la gran diversidad de factores de riesgo que la afectan. Entre estos determinantes se encuentran la conjugación de factores biológicos (anatomía dental y la dieta), con evidencia histórica, observacional, estudios clínicos y experimentales, nivel socio-económico, área de residencia, nivel educativo, ocupación, características del hogar, ingresos, oportunidades educativas, cuidado dental, edad, sexo, entre otros.

Además, de acuerdo con recientes estudios, la prevalencia de caries dental es mayor en áreas rurales que en áreas urbanas, volviéndose importante el abastecimiento de herramientas de servicios de salud de bajo costo en áreas rurales con el fin de que tengan un acceso justo y equitativo en la atención de la salud bucodental.

Por lo tanto, debido a la dificultad que representa el control de la incidencia de las caries dental, se han presentado investigaciones que implementan algoritmos y realizan análisis basados en CADx donde se usan modelos de predicción para saber el comportamiento futuro de estos determinantes complejos de manera simultánea, con el fin de poder realizar un diagnóstico preventivo que permita evitar el desarrollo de esta enfermedad y de esta manera reducir su alta incidencia.

En el presente trabajo se propone el análisis de una serie de determinantes basados en factores demográficos y dietéticos a través de diferentes técnicas de IA, con el fin de determinar cuáles son los determinantes que presentan información significativa en el desarrollo de caries dental, obteniendo un modelo multivariado que permite clasificar entre sujetos con presencia actual o pasada de esta condición y sujetos con ausencia de la misma, y de esta manera generar una herramienta de diagnóstico preventivo de bajo costo que apoye en la disminución de su incidencia.

A través de este análisis se desarrolla la vinculación del área odontológica y el área de ingeniería, en donde se logra la formación de recurso humano capacitado para el abordaje en el área biomédica, presentando una herramienta preliminar para los especialistas con aplicación en la mejora de la salud oral, pronosticando y diagnosticando la caries dental a través de diferentes modelos desarrollados con base en técnicas de IA.

1.3. Preguntas de investigación

- ¿Cuáles son los determinantes demográficos obtenidos de NHANES 2013 – 2014 que presentan información relevante para la descripción de sujetos con presencia de caries dental?

- ¿Cuáles son los determinantes dietéticos obtenidos de NHANES 2013 – 2014 que presentan información relevante para la descripción de sujetos con presencia de caries dental?

- ¿Cuáles son los determinantes demográficos y dietéticos obtenidos de NHANES 2013 – 2014 que, en conjunto, permiten modelar con base en IA la clasificación de sujetos con presencia de caries dental de sujetos control?

- ¿Cómo se demuestra que el modelo desarrollado contiene los determinantes demográficos y dietéticos que aportan la información más significativa para la clasificación de sujetos con presencia de caries y sujetos control?

- ¿Cuál es la herramienta de IA que permite el mejor modelado para la clasificación de sujetos con presencia de caries dental de sujetos control?

1.4. Objetivo general

Desarrollar un modelo multivariado como herramienta de diagnóstico preventivo con base en técnicas de IA, con el fin de clasificar sujetos con presencia de caries dental, de acuerdo con la información obtenida de determinantes demográficos y dietéticos recopilada en las bases de datos de NHANES 2013 – 2014.

1.4.1. Objetivos específicos

- Determinar dentro de la información contenida en la base de datos demográfica de NHANES 2013 – 2014 cuales son los determinantes que permiten la descripción de sujetos con presencia de caries dental.

- Determinar dentro de la información contenida en la base de datos dietética de NHANES 2013 – 2014 cuales son los determinantes que permiten la descripción de sujetos con presencia de caries dental.

- Definir los determinantes demográficos y dietéticos que en conjunto permiten modelar la clasificación de sujetos con presencia de caries dental y sujetos control.

- Validar que el modelo desarrollado contiene los determinantes demográficos y dietéticos que aportan la información más significativa en la clasificación de sujetos con presencia de caries y sujetos control.

- Comparar las diferentes herramientas de IA utilizadas para el modelado de la clasificación de sujetos con presencia de caries de sujetos control, con el fin de identificar la que presenta el mejor comportamiento en la precisión.

1.5. Hipótesis

Un modelo multivariado desarrollado a través de diferentes técnicas de IA contenido por descriptores demográficos y dietéticos, obtenidos de las bases de datos de NHANES 2013 – 2014, permite la clasificación de sujetos con presencia de caries dental, con el fin de obtener una herramienta de diagnóstico preventivo para la disminución de la alta incidencia de esta condición.

1.6. Estructura de la tesis

La estructura con la que se organiza el presente trabajo es la siguiente:

Capítulo 1 Se presenta una breve introducción del tema que se aborda, mostrando el contexto del problema, los motivos que justifican la investigación del mismo y la búsqueda de una solución. También se incluyen las preguntas de investigación, objetivos, general y específicos, e hipótesis.

Capítulo 2 Se presenta una amplia descripción de los temas teóricos y de referencia requeridos para la realización de este trabajo, analizando de manera detallada el marco teórico de los conceptos, técnicas y métodos que permitieron adquirir la información utilizada.

Capítulo 3 Se muestran los artículos que han sido desarrollados a lo largo de este periodo de investigación, los cuales han permitido discutir y llegar a conclusiones objetivas a través de las comparación de sus diferentes metodologías y resultados obtenidos.

Capítulo 4 Se presenta la discusión de los diferentes resultados logrados en cada uno de los artículos de investigación realizados.

Capítulo 5 Se presenta un resumen de las conclusiones alcanzadas a través de los diversos resultados obtenidos, al igual que la conclusión global y puntual a la que fue posible llegar.

2. Marco teórico y de referencia

Con el fin de comprender el estudio de la caries dental, sus factores de riesgo y sus herramientas de diagnóstico, tratamiento y pronóstico, en este capítulo se profundiza en las definiciones de estos puntos, explicando las técnicas actuales del área clínica que permiten la evaluación de este padecimiento, además de mostrar la persistente necesidad de mejorar los métodos que combaten su prevalencia e incidencia, presentando una serie de algoritmos y técnicas computacionales que podrían brindar una solución al convertir este problema en un modelo computacional desarrollado a través del análisis de datos.

2.1. Trabajo relacionado

Debido al gran impacto que tiene la condición de caries dental alrededor de todo el mundo, tomando en cuenta que es una enfermedad de salud pública, una serie de trabajos han sido desarrollados con el fin de definir los factores de riesgo que apoyan su incidencia y prevalencia, además de buscar herramientas que permitan evitar la aparición y el desarrollo de las mismas.

En el trabajo de Pereira et al. [16], se propone investigar como afectan los factores del estado de salud oral y la vulnerabilidad social en la calidad de vida de los adolescentes que participaron en la encuesta de "SB Sao Paulo 2015". La metodología se basó en buscar la relación entre diferentes variables (Paulista Social Vulnerability Index score, género, color de piel, ingreso familiar, edad, caries no tratadas, pérdida de dientes, condición periodental, entre otros) y si afectan en la calidad de vida, a través de un análisis de regresión logística. Se pudo concluir que la vulnerabilidad social no está asociada con el impacto oral; sin embargo, las condiciones de salud oral y las variables socio-demográficas

sí se encuentran relacionadas, afectando en las actividades diarias de los adolescentes.

En el trabajo de Jieyi et al. [17], se estudió el estado de las caries y su asociación con factores demográficos, ingesta de azúcar y comportamientos relacionados con la salud oral, en niños de 5 años en Hong Kong. La experiencia de la caries fue medida con el índice Decayed, Missing, and Filled Teeth (dmft), la cual se analizó a través una regresión de binomio negativo cero inflado (ZINB, por sus siglas en inglés) para buscar su relación con los diferentes factores. Las conclusiones de este trabajo pudieron sugerir que la prevalencia de caries en los niños está relacionada con la frecuencia con que ingieren azúcar, su atención dental y su nivel socio-económico.

En la investigación de Vega et al. [18] proponen determinar la asociación entre las salidas dentales clave y la ingesta de azúcar en inmigrantes y adultos nacidos en E.U.A. de origen mexicano. Las variables que fueron utilizadas para este trabajo incluyen la práctica de cuidado dental, información socio-demográfica y de aculturación, y el estimado de la ingesta de azúcar en comidas y bebidas. Para evaluar las asociaciones se utilizó el método de regresión lineal, prueba t de 2 muestras y ANOVA. Los resultados permitieron concluir que sí existe una asociación entre la ingesta de azúcar y el estado de salud dental en adultos de origen mexicano.

En el trabajo de Henry et al. [19], se desarrolla un estudio para describir la relación entre los tratamientos de caries dental y diversos factores demográficos entre adultos jóvenes de Israel. La examinación de los datos se hizo de manera cruzada con un análisis multivariado (modelo lineal generalizado), obteniendo entre los resultados que las mayores necesidades de tratamiento dental fueron estadísticamente significativas entre los participantes pertenecientes a familias de inmigrantes, pudiendo concluir que las variables socio-demográficas tienen una influencia significativa en las necesidades de tratamientos dentales.

Por otro lado, Wu et al. [20], propuso un estudio para identificar los indicadores útiles de la salud oral que permiten conocer los riesgos de la malnutrición en adultos mayores, basándose en parámetros clínicos e información socio-demográfica. Los resultados mostraron un impacto negativo de la salud oral en relación con la calidad de vida, mientras

que ninguno de los indicadores clínicos (caries dental, número de dientes, etc.) mostró asociaciones con el riesgo de tener malnutrición. Por lo que fue posible concluir que la pérdida de dientes de los adultos mayores y la falta de tratamiento para la caries dental están asociadas con la calidad de vida, lo que podría indicar mayor probabilidad de malnutrición.

Oyedele et al. [21] proponen un trabajo para determinar los factores que influyen en la aparición de la caries dental en una población de estudiantes de primaria y secundaria de los suburbios de Nigeria. Las variables que fueron analizadas incluyeron la edad, género, estado socio-económico, estado de salud oral, tipo de crianza, rango de nacimiento, tamaño de la familia y la presencia de caries dental. Los resultados mostraron que los niños de entre 11 y 16 años de edad tenían menor probabilidad de desarrollar caries dental, mientras que el último hijo de la familia presentó mayor probabilidad de desarrollar esta condición y los niños de familias grandes presentó menos. Además, se encontró que los primeros molares permanentes y los segundos molares primarios se veían más afectados por caries. Con base en esto se pudo concluir que la edad, la higiene oral, el rango de nacimiento y el tamaño de la familia son determinantes significantes de la caries dental.

Igualmente, en la investigación de Ghazal et al. [22] se conduce la evaluación de covariables dependientes del tiempo en relación con la caries dental de los dientes permanentes entre niños afroamericanos con el estado socio-económico bajo. Ninguno de los niños presentó caries dental en el comienzo del estudio y las revisiones se realizaron anualmente. El análisis consistió en un modelado de riesgos de Cox extendido multivariado y bivariado para buscar la relación entre las covariables dependientes y no dependientes del tiempo, y el tiempo del evento. Los resultados multivariados mostraron que un mayor consumo de agua fue asociado con un menor riesgo de padecer caries dental, mientras que el consumo de leche, bebidas azucaradas y jugos presentaron un mayor riesgo. Con base en estos resultados fue posible concluir que el consumo de agua en la niñez tiene un efecto protector significativo en la prevención de caries dental.

Finalmente, Brito et al. [23] presentaron un estudio con el fin de determinar la prevalencia de la caries dental y factores asociados en niños de Recife, Brasil. Se realizó un estudio

transversal, analítico, descriptivo, en donde se analizaron variables socio-demográficas y de comportamiento. Con el fin de validar los resultados se construyó un árbol de decisión inductivo. De esta manera, los resultados mostraron que los factores asociados con la caries dental fueron la manera en la que los niños se cepillan los dientes, la frecuencia del cepillado, un ingreso menor a la media en el hogar y la educación primaria incompleta de los padres.

2.2. Salud

De acuerdo con la OMS, salud se define como "un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades". Esta cita entró en vigor desde el 7 de abril de 1948 y no ha sido modificada desde esta fecha [24]. Por otro lado, investigaciones definen salud como "la capacidad de un cuerpo para adaptarse a las nuevas amenazas y enfermedades", basando este concepto en que la ciencia moderna ha aumentado la conciencia humana con respecto a las enfermedades y su funcionamiento.

Dentro de los factores que forman parte de una buena salud se encuentra el lugar en el que vive la persona, estado del ambiente que la rodea, genética, ingreso económico, acceso a servicios de salud, nivel de educación y relaciones con amigos y familia. Estos factores se pueden dividir en tres aspectos principales:

Ambiente social y económico: considera el nivel de riqueza de la familia o comunidad.

Ambiente físico: conformado por parásitos que existe en el área o niveles de contaminación.

Características y comportamiento de la persona: formados por los genes con los que nace la persona y las decisiones del estilo de vida.

Según la OMS, entre mayor sea el estado socio-económico de una de una persona, más probable es que goce de buena salud, buena educación, trabajo bien remunerado y acceso a atención médica cuando se vea amenazada la salud. Mientras que la gente con menor estado socio-económico es más propensa a padecer tensiones con respecto a la vida diaria,

dificultades financieras, desempleo, marginación y discriminación, aumentando el riesgo de mala salud.

Además, un bajo nivel socio-económico suele significar un menor acceso a la atención médica, teniendo menor esperanza de vida que quien tiene acceso a ella.

Para preservar la salud, uno de los puntos más importantes es tener un estilo de vida saludable, incluyendo una dieta equilibrada y nutritiva, ejercicio regular, pruebas de enfermedades que pueden implicar un riesgo, aprender a manejar el estrés, realizar actividades que brindan propósitos y conexión con otras personas y mantener un estado mental positivo.

Sin embargo, la promoción y preservación de la salud continúa siendo un problema alrededor del mundo, no solo en países en desarrollo, sino también en países desarrollados, tomando en cuenta que las enfermedades crónicas son uno de los principales problemas de salud pública en la mayor parte del mundo.

En la actualidad, el patrón de enfermedad ha sido transformado, considerando a las enfermedades de salud oral como uno de los problemas más importantes de salud pública debido a su alta incidencia y prevalencia en todas las regiones del mundo [24].

2.3. Salud oral

La salud oral se puede definir como "la ausencia de dolor orofacial, cáncer de boca o de garganta, infecciones y llagas bucales, enfermedades periodentales, caries, pérdida de dientes y otras enfermedades y trastornos que limitan en la persona afectada la capacidad de morder, masticar, sonreír y hablar, al tiempo que repercuten en el bienestar psicosocial", de acuerdo con la OMS [15].

La salud oral es considerada una parte integral y esencial de la salud en general, ya que puede comprometer la calidad de vida de todos los grupos sociales, incluyendo niños, adolescentes, adultos y adultos mayores sin distinguir entre edad, sexo, nivel social, cultural, económico, educacional, estado civil, historial previo de enfermedades orales, niveles de factores microbiológicos, herencia y el ambiente en el que se encuentran [25].

Debido a esto, la salud oral presenta un componente esencial en la calidad de vida debido a su influencia como factor determinante en la salud general de los individuos y las comunidades, convirtiéndose en un punto relevante en el cuidado de la salud. Por lo tanto, las serias repercusiones en términos de dolor y sufrimiento, deterioro de ciertas funciones y el efecto en la calidad de vida deben de ser consideradas también.

Por otro lado, la vigilancia epidemiológica de las enfermedades orales o bucodentales se vuelve importante en la medida que brinda elementos útiles para la planificación, programación, organización, integración, control y dirección de los programas de salud oral, al tiempo que orienta la atención a la población [2].

La prevalencia de estas enfermedades, además de depender fuertemente de la edad y la región geográfica (ya que existen en todas las poblaciones, aunque varían en su prevalencia y severidad), también se ve favorecida por la disponibilidad y accesibilidad de los servicios de salud bucodental, al igual que por diversos determinantes sociales.

A pesar de que las enfermedades de salud oral se presentan alrededor de todo el mundo, su mayor carga se encuentra especialmente en las poblaciones desfavorecidas y marginadas socialmente, en donde la cobertura de la atención de estos padecimientos es reducida, además de ser pocos los programas públicos de salud oral [1]. Cabe mencionar que la mayoría de las enfermedades y afecciones orales necesitan tratamientos con atención odontológica; sin embargo, la disponibilidad de estos servicios es limitada o inaccesible, haciendo que sus tasas de utilización sean especialmente bajas entre los adultos mayores, los habitantes de zonas rurales y las personas con bajos niveles de ingresos y de estudios, ya que, de acuerdo con la OMS, el tratamiento de varias condiciones orales es extremadamente caro y no es factible en la mayoría de los países de medianos y bajos ingresos, siendo la cuarta causa más cara para tratar en los países más industrializados, representando entre el 5 y 10 % del gasto sanitario [2], por lo que una medida que pudiera evitar el elevado costo de los tratamientos odontológicos aplicando medidas eficaces de prevención y promoción de la salud presentaría un beneficio de alto impacto.

Las dos principales enfermedades orales infecciosas son la caries dental y la enfermedad periodontal, y a pesar de que se crearon diversos programas de salud enfocados en la hi-

giene dental para evitar desarrollar estas condiciones, combatiendo el problema de caries que existía por falta de higiene bucal, el rol de ésta se volvió relativo en la prevención del desarrollo de caries ya que la prevalencia continuó [26].

2.4. Caries dental

En la salud oral, la caries dental ha sido un reto debido a que es una enfermedad crónica, pero evitable [27]. La presencia de esta condición desde la niñez hasta la edad adulta genera dolor, incapacidad de sonreír y problemas para probar, masticar y tragar, lo cual puede afectar en el autoestima, expresión, comunicación, estética facial, absentismo escolar en la niñez, afectando en el desempeño académico, mientras que en la edad adulta puede tener repercusiones en el trabajo, afectando en la producción económica [28].

La raíz de la caries dental hace referencia a una disolución química localizada en una superficie dental producida por la actividad metabólica en un depósito microbiano, llamado biofilm o placa dental, cubriendo la superficie del diente en todo momento.

El biofilm dental es una biomasa microbiana compuesta por bacteria residente de la saliva, el cual se daña cuando se hace el cepillado de los dientes. Estos microorganismos metabolizan el azúcar de la dieta y, como producto de desecho, se produce ácido. Este ácido puede desmineralizar diferentes componentes de los dientes, tales como el esmalte, la dentina y el cemento, mostrados en la imagen **2-1** [29], provocando lesiones que se pueden manifestar clínicamente en una variedad de formas, entre ellas la caries dental.

La dentición humana varía en su anatomía, presentando dos tipos de superficies, suaves y de mordedura, que se caracterizan por tener una forma más lisa o más tortuosa, respectivamente. Las superficies de mordedura presentan áreas incompletas de esmalte unidas, vistas clínicamente como fosas, fisuras y surcos. Ambas superficies pueden desarrollar lesiones por caries; sin embargo, las superficies de mordedura son las que tienen el mayor riesgo de desarrollarlas, debido a que las áreas de fosas y fisuras representan sitios de estancamiento para los sustratos del biofilm, beneficiando de manera desproporcionada la experiencia de esta condición.

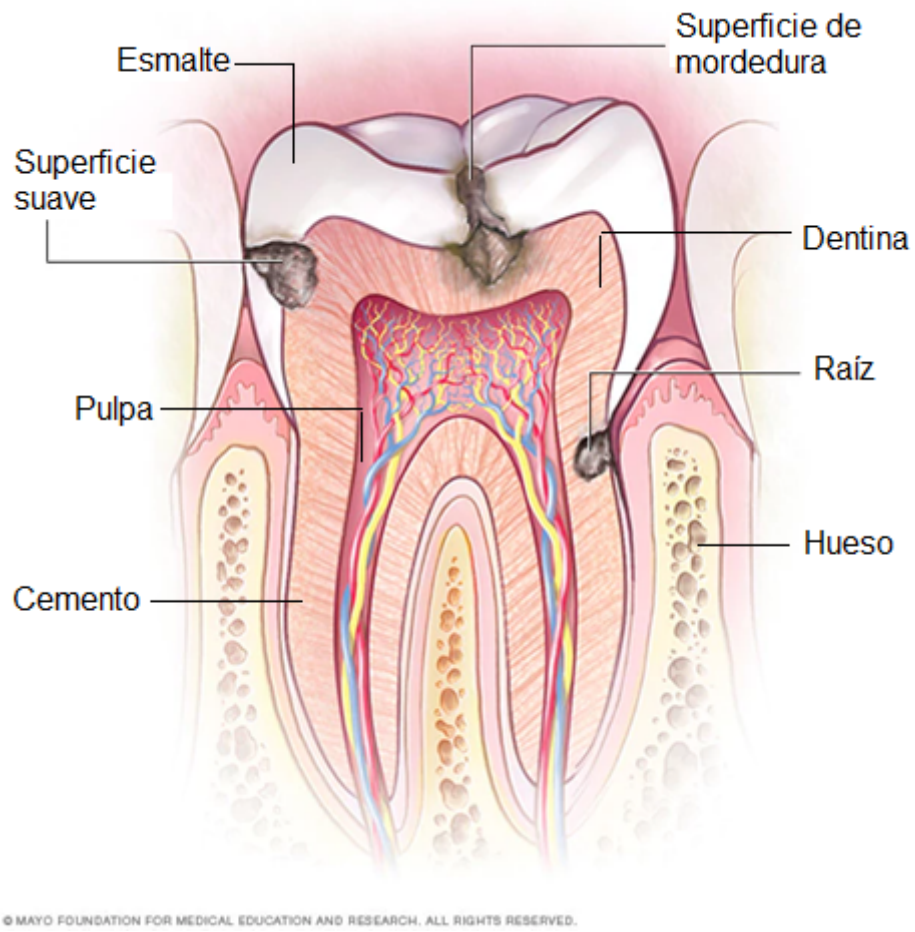

Figura 2-1.: Anatomía del diente y superficies donde se presenta la caries dental.

Principalmente, los primeros molares permanentes son los dientes permanentes más propensos a la caries dental, seguidos por los segundos permanentes molares, y aunque los premolares también presentan superficies con fosas y fisuras, tienen mucha menos prevalencia de caries en comparación con los molares.

Hasta la actualidad, el tratamiento más común para combatir la caries dental, son las restauraciones realizadas con amalgamas.

En la Figura **2-2** [30] se muestra un ejemplo de los primeros y segundos molares permanentes, presentando en (A) las fosas y las fisuras que están presentes en estos dientes, en donde, como se mencionó anteriormente, se suelen desarrollar la caries dental con mayor frecuencia, mientras que en (B) los molares se muestran con presencia de restauraciones realizadas con amalgama por caries previas.

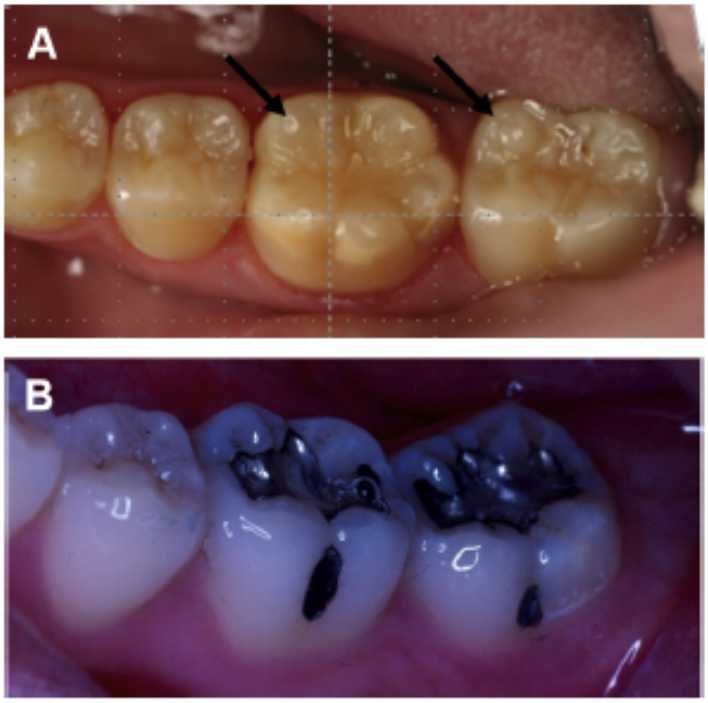

Figura 2-2.: Primero y segundo molares permanentes, (A) presenta las fosas y fisuras donde regularmente se desarrollan las caries, (B) presenta restauraciones con amalgama.

Por otro lado, ha sido históricamente conocido que la prevalencia de la caries dental era mayor en las zonas urbanas que en las zonas rurales en algunas partes del mundo; sin embargo, de acuerdo con estudios recientes, fue encontrado que a través del tiempo, la prevalencia y cantidad de caries dental han sido significativamente reducidas en las zonas urbanas, mientras que en las zonas rurales se han visto incrementadas, convirtiéndose en un aspecto importante de suministros en estas zonas, con el fin de tener acceso justo e igualitario a los servicios de cuidado de salud oral [31].

También, de acuerdo con estudios epidemiológicos, se ha revelado que la caries dental es la enfermedad crónica con mayor prevalencia en la comunidad pediátrica [32], presentándose en alrededor del 60 y 90 % de los niños en edad escolar, donde la dentición es temporal; mientras que en los adultos se presenta en cerca del 100 %, donde la dentición es permanente [5,6]. Además, es importante mencionar que los niños que presentan caries en la dentición primaria tienen mayor probabilidad de desarrollar caries en la dentición permanente.

Una medida clave que ha sido establecida en la epidemiología dental para estandarizar

el diagnóstico de la caries dental es el índice de dientes permanentes cariados, perdidos y obturados (CPO) o más conocido por sus siglas en inglés *Decayed*, *Missing*, *Filled* (DMF). Este índice es aplicado en los dientes permanentes y es expresado como el número total de dientes o superficies que se encuentran deteriorados, faltantes u obstruídos en un sujeto. Un ejemplo de la dentición permanente se observa en la Figura **2-3** (A).

Cuando el índice es aplicado a los dientes completos, es llamado índice CPOD o $DMFT$ y las puntuaciones tienen un rango de 0 a 28 o 32, dependiendo de si se incluyen los terceros molares en la puntuación. Si el índice es aplicado solo a las superficies de los dientes (cinco en la parte posterior del diente y cuatro en la parte anterior), es llamado CPOS o $DMFS$, y las puntuaciones tienen un rango de 0 a 128 o 148, dependiendo también de si se incluyen los terceros molares en la puntuación.

Si el índice es aplicado a dientes temporales se define como índice de dientes primarios cariados, con indicación de extracción y obturados (CEO) o es escrito en letras minúsculas en sus siglas en inglés, dmf, y se refiere a la experiencia de caries en niños de acuerdo con el número total de dientes o superficies deteriorados, faltantes u obstruídos. Un ejemplo de esta dentición se observa en la Figura **2-3** (B). El índice CEOD o $dmft$ tiene un rango de 0 a 20, mientras que el índice CEOS $dmfs$ tiene un rango de 0 a 88.

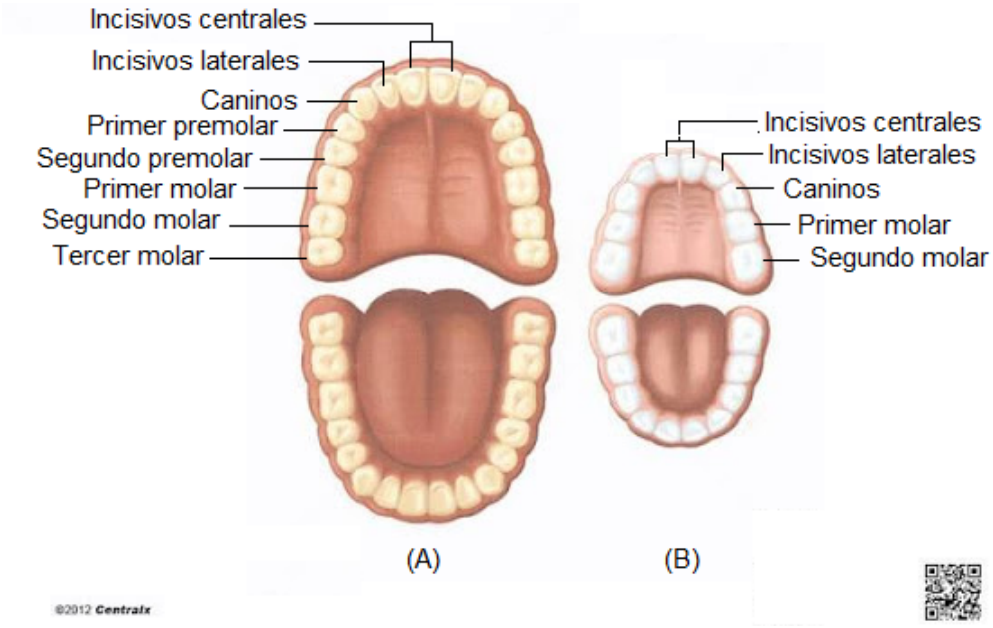

Figura 2-3.: Ejemplo de (A) dentición permanente y (B) dentición temporal.

Sin embargo, a pesar de que este índice puede proporcionar datos importantes sobre la caries dental, presenta algunas limitaciones. Una de estas limitaciones es que los investigadores han notado una cantidad significativa de *bias* o tendencia y variabilidad en los datos, la cual es necesario tomar en cuenta para el desarrollo de trabajos [33].

2.4.1. Factores de riesgo

Algunos factores de riesgo en el padecimiento de enfermedades bucodentales, entre ellas caries, son la mala alimentación, el tabaquismo, el consumo nocivo de alcohol y la falta de higiene bucodental, además de diversos determinantes sociales [15].

También, existen algunos factores de riesgo específicos para la caries en las fosas y en las fisuras dentales. Entre estos riesgos se encuentra el biofilm, la morfología de las fisuras, la mineralización del esmalte y la etapa de la erupción de los dientes. La presencia visual del biofilm en la superficie de una fisura que entró en erupción parcialmente también ha sido asociada con un mayor riesgo de caries dental [30].

Por otro lado, de acuerdo con el reporte de la OPS, se describe que el desarrollo de la caries dental depende de la frecuencia en el consumo de ciertos componentes, como carbohidratos, las características de los alimentos, el tiempo de exposición, eliminación del biofilm y la susceptibilidad del huésped, sumándose las escasas medidas preventivas en salud oral, entre otros. Un importante aspecto a considerar son las deficiencias crónicas nutricionales (nutrición, sobrepeso, obesidad), ya que de acuerdo con estudios, el índice de masa corporal (IMC) es un indicador de la relación de peso-talla que describe la caries dental, entre otros padecimientos, encontrando que los niños con estas deficiencias presentan un riesgo mayor de desarrollar biofilm, que, como se mencionó anteriormente, generalmente continúa en caries [34, 35].

En la Figura **2-4** [36] se presentan algunos de los factores más importantes, sociales y dentro de la cavidad oral, que determinan la tasa relativa del desarrollo de la caries dental. Los factores que se presentan dentro de la cavidad oral indican como actúan los determinantes biológicos de la caries en el nivel de la superficie del diente; mientras que los factores al nivel social presentan la fuerte influencia que pueden tener la educación,

el conocimiento, el comportamiento y las actitudes en algunos determinantes biológicos, como la calidad de la higiene oral, la selección de la comida, el uso de fluoruros y el flujo de saliva, entre otros [5, 6].

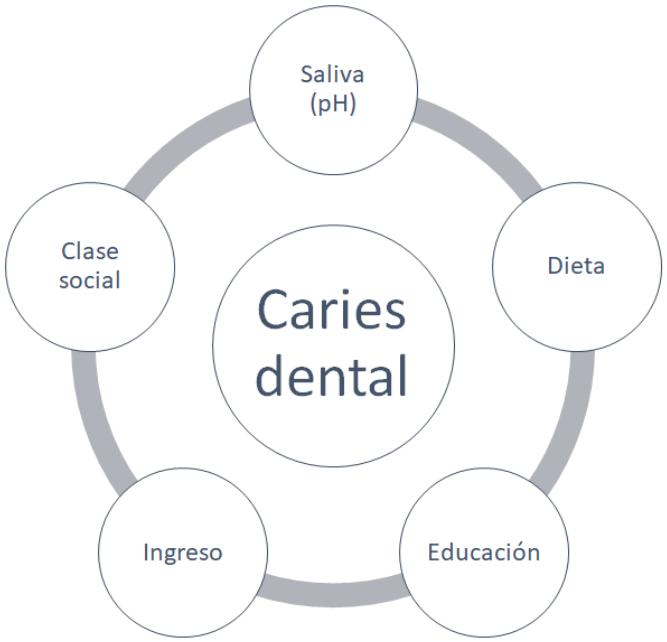

Figura 2-4.: Algunas de las principales variables que determinan el desarrollo de la caries dental.

Además, es importante tomar en cuenta que estos puntos no solo generan un impacto negativo en la salud, sino que también presentan un incremento en el riesgo o predisposición de sufrir descompensaciones, provocando diabetes, artritis, eventos trombóticos y nacimientos prematuros [28].

2.5. NHANES

NHANES es un programa de estudios diseñado para evaluar el estado de salud y nutricional de niños y adultos en E.U.A., siendo el único que recopila en una encuesta la combinación de entrevistas y examinaciones físicas. Además, NHANES es un programa mayor del *National Center for Health Statistics* (NCHS), el cual es parte de los *Centers for Disease Control and Prevention* (CDC) y tiene la responsabilidad de producir estadísticas vitales de salud para la nación [37].

Este programa inició en los comienzos de los años 60s y ha sido manejado como una serie de encuestas enfocado en diferentes grupos de población o temas de salud. Después, en 1999, la evaluación se convirtió en un programa continuo que cambió su enfoque a una variedad de medidas de salud y de nutrición para conocer necesidades emergentes. La evaluación examina una muestra nacionalmente representativa de alrededor de 5,000 personas cada año (utilizando la evaluación 2013–2014 para el desarrollo de esta investigación). Estas personas se encuentran en localidades situadas en diferentes regiones del país, de las cuales 15 son visitadas cada año.

Los resultados de esta evaluación son usados para determinar la prevalencia de las principales enfermedades y sus factores de riesgo, además de evaluar el estado nutricional y su asociación con la promoción de la salud y la prevención de enfermedades.

De manera que los datos contenidos en estas encuestas son usados para estudios epidemiológicos e investigación en ciencias de salud, apoyando al desarrollo de políticas sanas de salud pública, a la dirección y diseño de programas y servicios de salud, y a expandir el conocimiento de la salud globalmente.

2.5.1. Contenido de las encuestas

Las encuestas de NHANES contienen datos sobre la prevalencia de diferentes enfermedades crónicas en la población, produciendo estimaciones de las condiciones no diagnosticadas previamente, así como de las conocidas e informadas por los encuestados.

El contenido también se compone por factores de riesgos, que se refieren a aquellos aspectos del estilo de vida, la herencia o el entorno de una persona que pueden aumentar las posibilidades de desarrollar cierta enfermedad o condición. Entre los factores de riesgo que son estudiados, se encuentran el tabaquismo, el consumo de alcohol, la actividad sexual, el uso de drogas, la aptitud y actividad física, el peso y la ingesta dietética. También se incluyen datos de ciertos aspectos de salud reproductiva como el uso de anticonceptivos orales y la práctica de amamantamiento.

Entre las enfermedades, las condiciones médicas y los indicadores de salud estudiados se incluyen anemia, enfermedad cardiovascular, diabetes, exposiciones ambientales, en-

fermedades de los ojos, pérdida de la audición, enfermedades infecciosas, enfermedad del riñón, nutrición, obesidad, salud oral, osteoporosis, funcionamiento físico y actividad física, historial reproductivo y comportamiento sexual, enfermedades respiratorias y enfermedades de transmisión sexual.

La muestra de las encuestas es seleccionada para representar a población de U.S.A de todas las edades. Para reproducir estadísticas confiables, NHANES sobremuestrea personas de 60 años o mayores, americanos, asiáticos, africanos e hispánicos.

Todos los participantes visitan al médico y tienen incluidas entrevistas dietéticas y medidas del cuerpo. También, a todos, excepto a los más jóvenes, se les toma una muestra de sangre y se les aplica una prueba dental.

2.5.2. Participantes

La población objetivo principal para NHANES son los residentes civiles no institucionalizados de E.U.A. El diseño de la selección de la población está basado en el muestreo de un gran número de subgrupos específicos que presentan características de salud pública particulares con el fin de aumentar la confiabilidad y precisión de los indicadores del estado de salud.

La selección de los participantes se lleva a cabo en cinco etapas, representadas en la Figura **2-5**. En la primera etapa (A), todas las localidades de E.U.A. se dividen en 15 grupos de acuerdo a sus características. Después, de cada grupo se selecciona una localidad, obteniendo 15 localidades en total para las encuestas de NHANES de cada año. En la segunda etapa (B), dentro de cada localidad se forman pequeños grupos contenidos con un gran número de hogares, y de ellos se seleccionan de 20 a 24 grupos. Posteriormente, todas las casas o departamentos de cada grupo son identificados en la tercera etapa (C), y una muestra de alrededor de 30 hogares es seleccionada dentro de cada grupo. En la cuarta etapa (D), todos los entrevistadores de NHANES van a cada hogar seleccionado y preguntan por información específica (edad, raza y género) a cada miembro de la familia. Finalmente, en la quinta etapa (E), se seleccionan aleatoriamente a través de un algoritmo computacional a todos, algunos o ningún miembro del hogar

para ser participante de las encuestas.

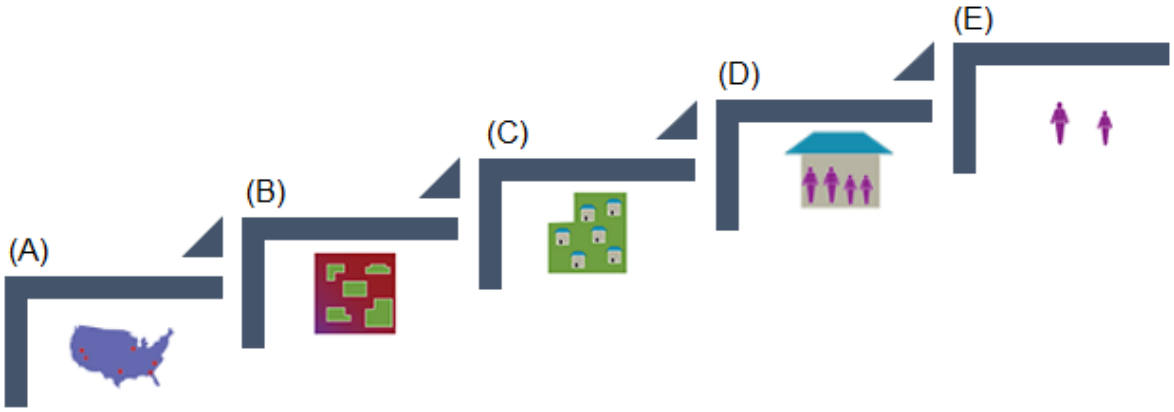

Figura 2-5.: Diagrama de la selección de sujetos de NHANES.

2.5.3. Encuestas

Las encuestas incluyen preguntas demográficas, socio-económicas, dietéticas y relacionadas a la salud. Además, se presenta un componente de examinación que consiste en medidas médicas, dentales y fisiológicas, al igual que pruebas de laboratorio administrados por personal médico altamente calificado.
La preguntas que forman parte de cada encuesta contienen información que permite conocer características de la calidad de vida de los sujetos:

- Demográfica: Brinda información a nivel individual, familiar y del hogar sobre diferentes temas de la demografía del lugar que se habita (nivel socio-económico, tamaño de la familia, situación política, entre otros).

- Dietética: Brinda información detallada sobre la ingesta dietética diaria consumida en diferentes periodos de tiempo por los participantes (periodos de 24 horas y de 30 días).

- Examinación: Brinda información sobre diferentes medidas y evaluaciones de la salud general de los participantes (presión sanguínea, medidas del cuerpo, radiografías de diferentes partes del cuerpo, salud oral, gusto y olfato, entre otros).

- Laboratorio: Brinda información sobre los resultados obtenidos de diferentes pruebas de laboratorio a las que se sometieron los participantes (albumina, creatinina, mercurio en la sangre, colesterol, hepatitis, virus de inmunodeficiencia humana (VIH), virus del papiloma humano (VPH), entre otros).

- Cuestionario: Brinda información sobre preguntas sobre diferentes temas de la salud general de los participantes (uso de alcohol, dermatología, diabetes, uso de drogas, seguro de salud, condiciones médicas, entre otros.)

2.5.3.1. Determinantes demográficos

Los determinantes demográficos de NHANES proporcionan información sobre diferentes tópicos, como de la entrevista y examinación de muestras de pesos corporales, estado de embarazo, ingreso del hogar y de la familia, tamaño del hogar y de la familia, cuestionario de lenguaje, composición del hogar (número de niños y de adultos mayores), demografía sobre el hogar de la persona de referencia, género, edad, raza, educación, estado civil, estado del servicio militar, país de nacimiento, ciudadanía y años de residencia en E.U.A.

En la Tabla **2-1** se muestra el número y el porcentaje de los sujetos examinados que participaron en las encuestas de NHANES 2011–2014, subdivididos de acuerdo con su raza.

Tabla 2-1.: Número y porcentaje de participantes examinados mostrados por raza de NHANES 2011–2014.

	Hispanos		No hispanos				Total
	Mexicanos americanos	Otros hispanos	Blancos, raza única	Negros, raza única	Asiáticos, raza única	Otros, incluidos multirraciales	
2011–2014	3,001	1,941	6,379	4,780	2,234	816	19,151
n (%)	15.7	10.1	33.3	25.0	11.7	4.3	100

En el Apéndice A se muestran los determinantes obtenidos con la encuesta de datos demográficos.

2.5.3.2. Determinantes dietéticos

Los determinantes dietéticos de NHANES tienen el objetivo de presentar información sobre la ingesta dietética con el fin de estimar los tipos y cantidades de alimentos y bebidas consumidos durante diferentes periodos de tiempo, y así estimar la ingesta de energía, nutrientes, y otros componentes de la comida.

Estos determinantes proporcionan información sobre diferentes intereses dietéticos. Entre ellos se encuentra la ingesta total de sal, antiácidos, vitaminas, minerales, hierbas, bocadillos, harinas, agua embotellada, pescados y mariscos, si el participante se encuentra en algún tipo de dieta para bajar peso y los componentes agregados en la preparación de la comida.

En el Apéndice B se muestran los determinantes obtenidos con la encuesta de datos dietéticos.

2.6. Diagnóstico asistido por computadora

El análisis computacional de datos médicos puede contribuir a comprender mejor algunas enfermedades, además de diagnosticar problemas asociados con etapas tempranas en sujetos con riesgo de padecer alguna enfermedad, contribuyendo así en la mejora de los sistemas de salud [38]. Las estadísticas computacionales y las herramientas bioinformáticas permiten encontrar relaciones inesperadas entre las características que se están analizando y los síntomas de enfermedades [39, 40].

El control de la alta incidencia y prevalencia de la caries dental se ha vuelto un reto debido a que una gran cantidad de factores riesgo interactúan de manera simultánea en su favor. Con base en esto, se han desarrollado una serie de investigaciones en donde se implementan algoritmos y se realiza el análisis de datos, buscando relación entre ellos, utilizando herramientas de CADx, con el fin de encontrar modelos de clasificación y de predicción que permiten conocer el diagnóstico automático y el posible comportamiento futuro de esta condición, brindando un apoyo en la odontología clínica y de investigación [7, 41].

El CADx permite realizar el análisis de diferentes tipos de contenidos digitales, entre ellos, las bases de datos masivas o minería de datos, al igual que las imágenes médicas [42].

El principal aporte del CADx en el área clínica se encuentra en la identificación de datos de interés y la búsqueda de relación entre ellos y los diferentes padecimientos. Haciendo uso de diferentes tareas de aprendizaje, el sistema comienza a identificar patrones y a "aprenderlos"haciendo uso de un primer conjunto de datos finito, conocido como el "conjunto de entrenamiento", una vez que el sistema se adaptó, se usa un conjunto de datos nuevo, conocido como "conjunto de prueba", con el que se auto-evalúa para validar el resulado obtenido [43].

Dentro de las ventajas que los sistemas CADx presentan, se encuentra el apoyo a los especialistas en salud, que aunque no permiten dar un diagnóstico completo a partir de la fuente, sí ayuda con la interpretación de datos presentados en grandes cantidades e imágenes médicas, mejorando el rendimiento y la consistencia del diagnóstico, ya que, estos sistemas no se ven afectados por las limitaciones de distracción y de la fatiga, entre otros factores que afectan a los seres humanos [44].

2.7. Inteligencia artificial

La IA es uno de los campos más novedosos en la ciencia y la ingeniería, y no solo sigue el principio de entender, sino que también de construir entidades inteligentes, teniendo como objetivo el desarrollar técnicas que permitan a los algoritmos aprender.

En la actualidad, la IA es capaz de realizar una serie diversa de actividades en una gran cantidad de subcampos. Algunas aplicaciones son el reconocimiento de voz, planeamiento de logística, robótica, traducción automática, vehículos autónomos y el CADx, entre otras.

Historicamente, la IA tiene cuatro enfoques principales:

Pensamiento humano Se refiere a la automatización de actividades que se asocian con el pensamiento humano, como la toma de decisiones, resolución de problemas, aprendizaje, entre otras (Bellman, 1978);

Pensamiento racional Se refiere al estudio de las facultades mentales a través del uso de modelos computacionales (Charniak y McDermott, 1985);

Acto humano Se refiere al estudio de como las computadoras hacen cosas que, hasta el momento, las personas hacen mejor (Rich y Knight, 1991);

Acto racional se refiere a la inteligencia computacional como el estudio del diseño de los agentes inteligentes (Pool et al., 1998).

La IA fue fundada en parte como una rebelión contra las limitaciones que existían en los campos de la teoría del control y las estadísticas. En términos de metodología, la IA adoptó el método científico desde 1987, y para ser aceptada, las hipótesis deben de ser sometidas a rigurosos experimentos empíricos, y los resultados deben de ser estadísticamente analizados para conocer su importancia [45, 46].

La IA se puede dividir en tres subcampos, la inteligencia artificial angosta (IAA), la inteligencia artificial general (IAG) y la inteligencia artificial superior (IAS). La IAA logra realizar una sola tarea y es la etapa en la que se encuentra la IA actualmente. Por otro lado, la IAG busca un algoritmo universal para aprender y actuar en cualquier entorno que sea capaz de dar soluciones y razonar, presentando habilidades cognitivas humanas. Finalmente, la IAS es una inteligencia que presenta capacidades superiores a las de un ser humano [46, 47].

Finalmente, en resumen, las técnicas de IA consisten en crear algoritmos que a partir de información particular con la que se induce al sistema, se puedan reconocer patrones y generalizar comportamientos. Esta inducción se considera completa cuando el sistema ha observado todos los casos particulares, por lo que la generalización obtenida se considera válida; sin embargo, en la mayoría de los casos es imposible obtener una inducción completa, por lo que el enunciado general queda sometido a un cierto grado de incertidumbre y en consecuencia no se puede considerar como un proceso de inferencia formalmente válido; no obstante, los algoritmos de IA pueden superar la incertidumbre haciendo uso de la minería de datos, ya que al unir la IA con la minería de datos el sistema puede aprender en forma automática, a través del aprendizaje automático [48].

2.7.1. Minería de datos

La minería de datos se define como el proceso de descubrir patrones en las grandes cantidades de datos, definidas como "big data".

Big data comienza con un gran volumen de fuentes heterogéneas y autónomas, con un control distribuido y descentralizado, buscando explorar relaciones complejas entre los datos a través de la minería de datos. Entre mayor sea el volumen de la big data, mayor será la complejidad y las relaciones de los datos.

Estas características hacen que el descubrir conocimiento útil de la big data sea un reto. Las fuentes son heterogéneas porque son recolectadas con diferentes esquemas o protocolos, además de provenir de diferentes aplicaciones, provocando que los datos tengan diversas representaciones.

Por otro lado, las fuentes al ser autónomas son capaces de generar y recolectar información sin implicar un control centralizado [49].

En la minería se tiene el conjunto de datos sin procesar que es almacenado electrónicamente y la búsqueda es automática o semi-automática, facilitada por una computadora. Esta búsqueda tiene el fin de resolver problemas a través del análisis de los datos que se encuentran recopilados en las bases de datos, distinguiendo patrones que permitan identificar comportamientos característicos.

Los patrones descubiertos deben de ser significativos, permitiendo tener alguna ventaja y hacer la predicción no trivial en datos nuevos, con el fin de conocer comportamientos futuros y poder tomar ventaja sobre diferentes problemas, por ejemplo, el diagnóstico preventivo de diferentes enfermedades.

Existen dos extremos para la expresión de patrón, caja negra y caja transparente. En la caja negra la estructura interna es efectivamente incomprensible, mientras que la caja transparente revela la estructura del patrón; sin embargo, con las dos es posible tener predicciones exitosas.

La diferencia en los patrones, es que cuando son minados están representados en términos de una estructura que puede ser examinada, razonada y usada para informar de futuras predicciones. Estos patrones se conocen como "estructurales"porque capturan la

estructura de las decisiones de forma explícita, es decir, que ayudan a explicar algo sobre los datos [50].

Estos descriptores se pueden volver muy complejos y típicamente son expresados como conjuntos de reglas, los cuales sirven para explicar lo que se ha aprendido, o en otras palabras, explicar las bases de las nuevas predicciones. La búsqueda de descriptores estructurales se hace a través de diversas técnicas de aprendizaje automático, con el fin de conocer lo que se quiere aprender, ya que, de acuerdo con la experiencia, se ha demostrado que el conocimiento explícito que es adquirido de las estructuras es tan necesario como la habilidad de predecir correctamente nuevos ejemplos [51].

2.7.2. Aprendizaje automático

El aprendizaje automático o *machine learning* es un tipo de IA que proporciona a los sistemas la capacidad de aprender, sin ser programados explícitamente. Este aprendizaje se centra en el desarrollo de sistemas informáticos que pueden cambiar cuando se exponen a nuevos datos y estos sistemas se clasifican de acuerdo con los objetos ingresados al sistema de la forma siguiente:

Regresión: intentan predecir un valor real a partir de valores almacenados.

Clasificación (binaria/multiclase): intentan predecir la clasificación de objetos sobre un conjunto de clases prefijadas.

Ranking: intentan predecir el orden óptimo de un conjunto de objetos según un orden de relevancia predefinido.

Por lo tanto, en la solución de un nuevo problema que se desea resolver por medio de aprendizaje automático se busca dentro de las clases anteriores la más semejante a la solución deseada, ya que depende de esta clasificación la forma en la que se puede medir el error cometido entre la predicción y la realidad.

Por otro lado, el aprendizaje automático también se clasifica de acuerdo al tipo de salida que produce el problema y de cómo se aborde el tratamiento de los ejemplos en [52]:

Aprendizaje supervisado: genera una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema, donde la base de conocimientos del

sistema está formada por ejemplos etiquetados de los que se conoce su clasificación correcta.

Aprendizaje no supervisado: el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formados únicamente por entradas al sistema, sin conocer su clasificación correcta, por lo que se busca que el sistema sea capaz de reconocer patrones para poder etiquetar las nuevas entradas.

Aprendizaje semi-supervisado: es una combinación de los algoritmos supervisados y los no supervisados, teniendo ejemplos clasificados y no clasificados.

Transducción: es similar al aprendizaje supervisado, pero su objetivo no es construir de forma explícita una función, sino únicamente tratar de predecir las categorías en las que caen los siguientes ejemplos basándose en los ejemplos de entrada, sus respectivas categorías y los ejemplos nuevos al sistema.

Aprendizaje multi-tarea: engloba todos los métodos de aprendizaje que usan conocimiento previamente aprendido por el sistema para tratar problemas parecidos a los aprendidos.

El trabajo computacional de los métodos de aprendizaje automático se divide regularmente en dos partes: entrenamiento y prueba. El proceso de entrenamiento consiste en evaluar el comportamiento de los datos y hacer un análisis de los mismos, en donde se establecen parámetros específicos que ayudan a relacionar la salida del sistema con su entrada; mientras que el proceso de prueba consiste en aplicar el algoritmo entrenado [53].

Este diseño experimental de los algoritmos de aprendizaje automático es llamado validación cruzada, y consiste en dividir de manera aleatoria y proporcional los datos disponibles en un conjunto de datos de entrenamiento y un conjunto de datos de prueba. El conjunto de datos de entrenamiento debe contener una mayor cantidad de datos que el conjunto de datos de prueba, regularmente de una base de datos se usan 70 % de los datos para entrenamiento y 30 % de los datos para prueba. El motivo de esta división es validar al algoritmo con un conjunto de datos diferente al conjunto de datos de donde el algoritmo aprendió. Después, regularmente se suele hacer un análisis estadístico que permite conocer la precisión del modelo desarrollado, estimaciones del error, aptitud,

exactitud, valor de sensibilidad vs especificidad, falsos negativos y falsos positivos [54]. El algoritmo de aprendizaje automático al inicio debe contar con las características o variables independientes de los datos de entrenamiento y sus respectivas salidas, como se muestra en la Figura **2-6** (A) y (B). Si la cantidad de características se quiere reducir, es necesario contemplar una etapa de extracción de características, mostrado en la Figura **2-6** (C), en donde a través de diferentes técnicas se eliminan aquellas características que no representan información significativa para el objetivo del proceso. Después, el algoritmo de aprendizaje automático comienza su etapa de entrenamiento, en donde aprende la relación que existe entre los datos de entrada y los datos de salida, como se observa en la Figura **2-6** (D). Este procedimiento se hace con el objetivo de buscar una generalización en esta relación y encontrar un modelo o patrón específico que permita que se cumpla (Figura **2-6** (E)).

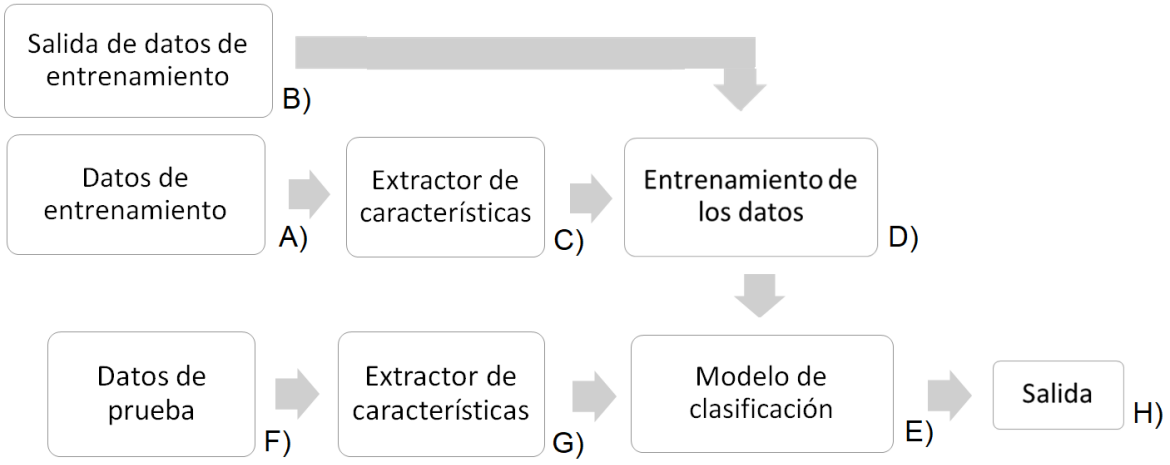

Figura 2-6.: Diagrama del proceso simple de los algoritmos de aprendizaje automático.

Finalmente, cuando el algoritmo ya aprendió la forma de relacionar las características con su salida correspondiente a través de un modelo, se evalúa si éste cumple con su tarea aplicando una prueba ciega, en donde el modelo se ve sometido a un conjunto de datos nuevo (datos de prueba, Figura **2-6** (F),(G)) sin tener conocimiento de sus salidas, con el fin de medir la capacidad que tiene para calcular sus salidas, Figura **2-6** (H), con base en el conocimiento aprendido en la etapa anterior y de esta manera saber si el modelo es apto para cumplir el objetivo inicial, clasificando los datos de salida de acuerdo a los

datos de entrada de manera correcta [55].

2.7.3. Algoritmos genéticos

Por otro lado, los algoritmos genéticos son una técnica que imita a la evolución biológica para solucionar problemas. Un AG se compone por procedimientos que simulan mecanismos de supervivencia de conjuntos de variables que demuestran ser las más efectivas para lograr un objetivo específico. Estos procedimientos se basan en hacer una mutación y el cruce de variables tantas veces como sea necesario para ir mejorando continuamente las combinaciones de las mismas y finalmente encontrar un conjunto de variables o modelo multivariado que sea el que mejor cumpla con ese objetivo; entonces, los algoritmos genéticos se componen por tres procedimientos básicos que son: selección, cruzamiento y mutación [56]. Estos algoritmos pueden utilizar diferentes técnicas para seleccionar a los individuos que deben ser copiados para la siguiente generación, algunas de estas técnicas son:

Selección elitista: se garantiza la selección de los individuos más aptos de cada generación; en donde, si el individuo no demuestra ser el mejor, es eliminado y ya no es copiado para la siguiente generación.

Selección proporcional a la aptitud: los individuos más aptos tienen más probabilidad de ser seleccionados; sin embargo, no es seguro ni es claro que se pueda hacer.

Selección por rueda de ruleta: es una forma de selección proporcional a que un individuo sea seleccionado de acuerdo a la diferencia entre su aptitud y la de sus competidores.

Selección escalada: en este método, al incrementarse la aptitud media de la población, la fuerza de la presión selectiva también aumenta y la función de aptitud se hace más discriminatoria. Este método puede ser útil cuando el sistema tenga tantos individuos que todos tengan una aptitud relativamente alta y sólo les distingan pequeñas diferencias.

Selección por torneo: en este método se eligen subgrupos de individuos de la población para que estos individuos compitan entre ellos, eligiendo únicamente al individuo que gane de cada subgrupo para la reproducción.

Selección por rango: en este método a cada individuo de la población se le asigna un rango numérico basado en su aptitud y la selección se basa en este ranking. La ventaja de este método es que puede evitar que los individuos muy aptos ganen al inicio a expensas de los menos aptos, lo que reduciría la diversidad genética de la población y podría obstaculizar la búsqueda de una solución aceptable.

Selección generacional: en este método la descendencia de los individuos seleccionados en cada generación se convierte en toda la siguiente generación.

Selección jerárquica: en este método los individuos atraviesan múltiples rondas de selección en cada generación. Las evaluaciones de los primeros niveles son más rápidas y menos discriminatorias, mientras que los que sobreviven hasta niveles más altos son evaluados más rigurosamente. La ventaja de este método es que reduce el tiempo total de cálculo al utilizar una evaluación más rápida y menos selectiva para eliminar a la mayoría de los individuos que se muestran poco o nada prometedores, y sometiendo a una evaluación de aptitud más rigurosa y computacionalmente más costosa sólo a los que sobreviven a esta prueba inicial.

Una vez que la selección ha elegido a los individuos más aptos, éstos son alterados de manera aleatoria con el fin de mejorar su aptitud para la siguiente generación. Las dos estrategias básicas de los métodos de cambio, como se mencionó anteriormente, son: mutación y cruzamiento. El método de cambio más sencillo es la mutación, Figura **2-7**, en donde, al igual que una mutación en los seres vivos cambia un gen por otro. Una mutación en un AG también causa pequeñas alteraciones en puntos concretos de la codificación del individuo [57].

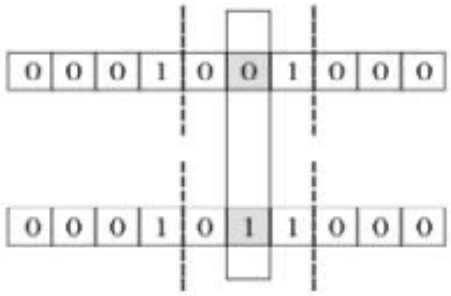

Figura 2-7.: Método de cambio de mutación.

El segundo método de cambio es el cruzamiento, y consiste en seleccionar a dos individuos para que intercambien segmentos de su código genético, produciendo una descendencia artificial cuyos individuos son combinaciones de sus antecesores. Este proceso pretende simular el proceso análogo de la recombinación que se da en los cromosomas durante la reproducción sexual. Las dos técnicas más comunes de cruzamiento son: cruzamiento de un punto y cruzamiento uniforme. En el cruzamiento de un punto, Figura **2-8** (A), se establece un punto de intercambio en un lugar aleatorio del genoma de los dos individuos, en donde uno de los individuos contribuye todo su código anterior a ese punto y el otro individuo contribuye todo su código a partir de ese punto para producir una descendencia [57]. En el cruzamiento uniforme, Figura **2-8** (B), se establece que el valor de una posición en el genoma de la descendencia corresponde al valor en esa posición del genoma de uno de los antecesores o del otro; la probabilidad de seleccionar cada uno de los antecesores es de 50 % [57].

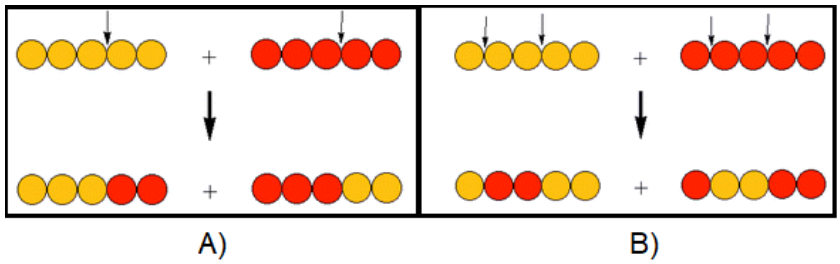

Figura 2-8.: Método de cambio de A) cruzamiento de un punto y B) cruzamiento uniforme.

Resumiendo, el proceso que lleva a cabo un AG, se comienza con un problema específico a resolver, en donde la entrada del AG es un conjunto de soluciones a ese problema y una métrica llamada "función de aptitud", que permite evaluar cuantitativamente a cada solución candidata del problema. Estas "soluciones candidatas" pueden ser soluciones que ya se sabe que funcionan, con el objetivo de que el AG las mejore, pero se suelen seleccionar aleatoriamente (Figura **2-9** (A)). Una vez que las soluciones candidatas son seleccionadas, el AG evalúa a cada una de las soluciones candidatas con la función de aptitud (Figura **2-9** (B)). Las primeras soluciones candidatas seleccionadas tendrán una eficiencia mínima con respecto a la solución del problema y la mayoría no mostrará un buen desempeño; ya que fueron elegidas aleatoriamente, sin conocimiento previo de una

solución al problema. Sin embargo, por simple azar, algunas pocas soluciones candidatas pueden ser prometedoras, pudiendo mostrar características que muestren, aunque no sea de forma fuerte y totalmente perfecta, cierta capacidad de solución al problema (Figura **2-9** (C)).

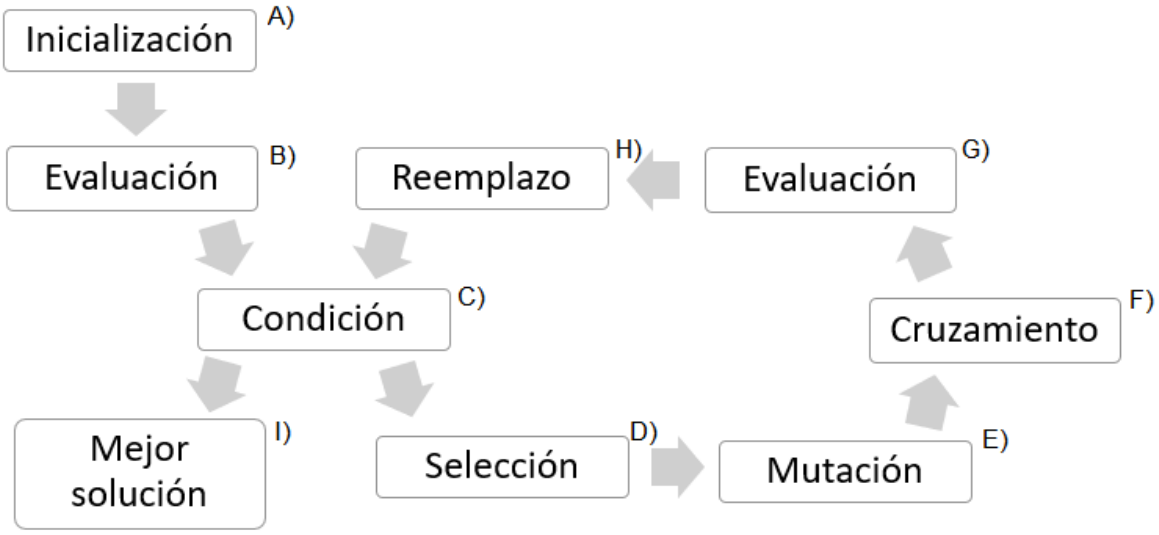

Figura 2-9.: Diagrama del proceso simple de los algoritmos genéticos.

Una vez que las soluciones candidatas son seleccionadas, el AG evalúa a cada una de las soluciones candidatas con la función de aptitud (Figura **2-9** (B)). Las primeras soluciones candidatas seleccionadas tendrán una eficiencia mínima con respecto a la solución del problema y la mayoría no mostrará un buen desempeño; ya que fueron elegidas aleatoriamente, sin conocimiento previo de una solución al problema. Sin embargo, por simple azar, algunas pocas soluciones candidatas pueden ser prometedoras, pudiendo mostrar características que muestren, aunque no sea de forma fuerte y totalmente perfecta, cierta capacidad de solución al problema (Figura **2-9** (C)). Las soluciones candidatas prometedoras se conservan (Figura **2-9** (D)) y se les permite reproducirse, haciendo múltiples copias de ellas; en esas copias son introducidos algunos cambios aleatorios durante el proceso de copiado, representando las mutaciones genéticas que pueden sufrir los descendientes en una población (Figura **2-9** (E)) y con el cruzamiento de los genes se intercambia información genética de las soluciones candidatas, representando la reproducción sexual en una población (Figura **2-9** (F)). Después, la descendencia continúa

con la siguiente generación, forma un nuevo conjunto de soluciones candidatas que van a ser sometidas nuevamente a una evaluación de aptitud (Figura **2-9** (G)). Las soluciones candidatas que empeoren o que no mejoren con los cambios en su código son eliminadas, y de nuevo; por simple azar, las variaciones introducidas en la población pudieron haber mejorado a algunos individuos, presentándolos como las mejores soluciones del problema, más completos o más eficientes (Figura **2-9** (H)). Este proceso es repetido durante varias iteraciones hasta obtener una solución lo suficientemente buena para presentar la mejor solución al problema que el sistema sea capaz de entregar (Figura **2-9** (I)) [58].

2.7.4. Redes neuronales artificiales

La neurociencia es el estudio del sistema nervioso, particularmente del cerebro. Aunque la manera exacta en la que el cerebro permite el pensamiento continúa siendo un misterio para la ciencia, el hecho de que permita el pensamiento ha sido apreciado por miles de años por la evidencia de que tiene la fragilidad de poder quedar incapacitado.

Es sabido que el cerebro consiste de células nerviosas o neuronas, representadas gráficamente en la Figura **2-10** [46]. Cada neurona consiste de un cuerpo celular o soma, contenido por un núcleo. En la parte externa del cuerpo celular se encuentra una serie de fibras llamadas dendritas y una fibra especialmente larga llamada axón, que puede llegar a medir desde un centímetro hasta un metro. Una neurona hace conexión con 10 a 10,000 neuronas a través de uniones conocidas como sinapsis. Las señales son propagadas de neurona a neurona por complejas reacciones electroquímicas, y son las encargadas de controlar la actividad cerebral de término corto, además de hacer cambios de término largo en la conectividad de las neuronas. Estos mecanismos forman la base del aprendizaje en el cerebro [46].

En la actualidad se tienen datos del mapeo que existe entre las áreas del cerebro y las partes del cuerpo que controlan o de donde son recibidos los sensores de entrada; sin embargo, no se tiene una completa compresión y casi no hay teoría de cómo es almacenada la memoria individual.

La IA se inspiró en la hipótesis del funcionamiento del cerebro para crear las RNAs. En

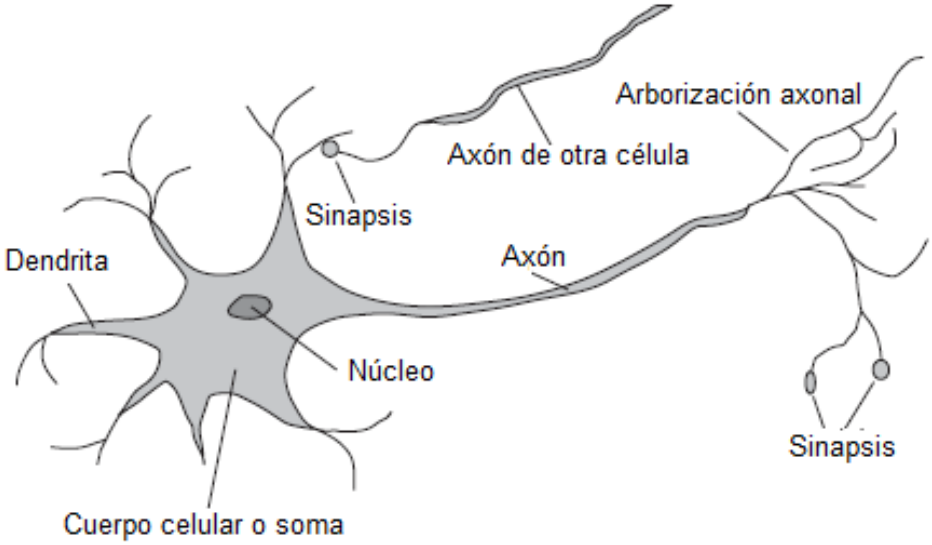

Figura 2-10.: Partes de una célula nerviosa o neurona.

la Figura **2-11** [46] se muestra un modelo matemático simple de una unidad neuronal, creado por McCulloch y Pitts (1943), y una RNA es una colección de estas unidades conectadas.

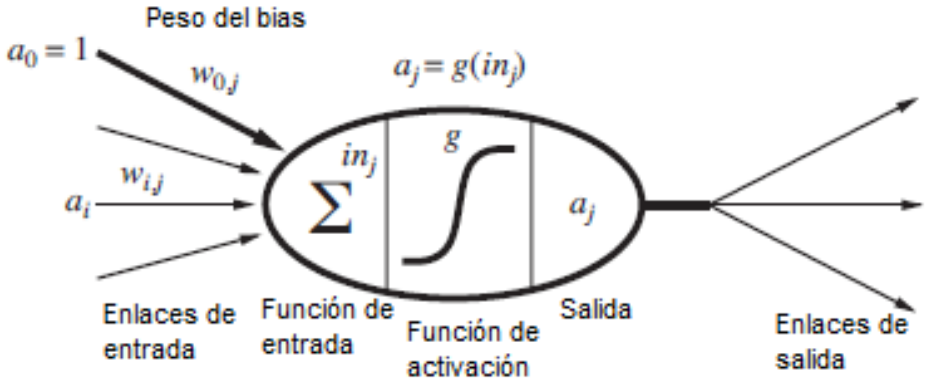

Figura 2-11.: Modelo matemático simple de una neurona.

Las RNAs conectan las unidades o nodos a través de enlaces direccionados. Un enlace que va de la unidad i a la unidad j, sirve para propagar la activación a_i de i a j. Cada enlace cuenta con un peso numérico $w_{i,j}$ asociado con él, el cual determina la fuerza y el signo de la conexión. Como en los modelos de regresión lineal, cada unidad tiene un entrada ficticia $a_0 = 1$ con un peso asociado, $w_{0,j}$. Cada unidad j primero calcula la

suma de sus entradas a través de la Ecuación 2-1.

$$in_j = \sum_{i=0}^{n} w_{i,j} a_i \tag{2-1}$$

Después, la función de activación g es aplicada a esta suma para obtener la salida, como se muestra en la Ecuación 2-2.

$$a_j = g(in_j) = g(\sum_{i=0}^{n} w_{i,j} a_i) \tag{2-2}$$

Regularmente, la función de activación g es un umbral rígido (salida de 0 o 1), en donde la unidad es llamada "perceptrón", o una función logística o sigmoidea ($\frac{1}{1+e^{-z}}$), en donde la unidad es llamada "percetrón sigmoideo". Ambas funciones de activación no lineales aseguran la importante propiedad de que la RNA completa pueda representar una función no lineal.

Una vez que se decidió el modelo matemático para las neuronas individuales, se conectan entre ellas para formar una red.

Existen dos métodos fundamentales para realizar esta tarea. Una RNA $feed-forward$ o que se alimenta hacia adelante, tiene conexiones en una sola dirección, es decir, forma un gráfico direccionado acíclico, donde cada nodo recibe la entrada de los nodos anteriores y entrega la salida hacia los nodos posteriores, sin que existan ciclos. Una red que se alimenta hacia adelante representa una función de su entrada actual, por lo que no tiene otro estado interno mas que los mismos pesos [59, 60].

Por otro lado, una red recurrente retroalimenta sus entradas con sus propias salidas. Esto quiere decir que los niveles de activación de la red forman un sistema dinámico que puede lograr un estado estable o exhibir un comportamiento oscilatorio o incluso caótico. Además, la respuesta de la red a una entrada dada depende de su estado inicial, el cual depende de sus entradas previas.

Por lo tanto, las redes recurrentes pueden soportar memoria a corto plazo, lo que genera un comportamiento interesante como modelos cerebrales, aunque también más difícil de entender.

Las RNAs generalmente se arreglan en capas, de las cuales cada unidad puede recibir entradas únicamente de las unidades que se encuentran en la capa anterior o pueden estar retroalimentadas, dependiendo del tipo de RNA. Las capas que se encuentran intermedias en la RNA se denominan como "capas ocultas" y están contenidas por "unidades ocultas". Entre más capas ocultas haya, más compleja se vuelve la RNA.

En una RNA multicapa con múltilples salidas m, el vector de error en la capa de salida se puede describir como $y - h_w(x)$, donde y es la salida esperada y $h_w(x)$ es la salida calculada; sin embargo, el error en las capas ocultas se vuelve complejo de calcular porque los datos de entrenamiento no permiten saber que valor deberían de tener los nodos ocultos. Este problema se resuelve con $back - propagration$ o propagación hacia atrás, en donde el error se puede propagar de la capa de salida hacia las capas ocultas. Este proceso emerge directamente de la derivación del gradiente de error general.

En la capa de salida, la regla de la actualización de los pesos para minimizar el error se calcula con la Ecuación 2-3.

$$w_i = w_i + \alpha(y - h_w(x)) \times h_w(x)(1 - h_w(x)) \times x_i \tag{2-3}$$

Al tener múltiples unidades de salida, Err_k es el kth componente del vector de error $y - h_w$, y el vector de error modificado es $\Delta_k = Err_k \times g'(in_k)$, por lo que la regla de la actualización de los pesos se convierte en la Ecuación 2-4.

$$w_{j,k} = w_{j,k} + \alpha \times a_j \times \Delta_k \tag{2-4}$$

Para actualizar las conexiones entre las unidades de entrada y las unidades ocultas, es necesario definir una cantidad análoga al término del error para los nodos de salida. Esta cantidad se define con el error propagado hacia atrás, en donde el nodo oculto j es "responsable"de una fracción del error Δ_k de cada nodo de salida al que está conectado. Entonces, los valores Δ_k son divididos de acuerdo con la fuerza de la conexión entre el nodo oculto y el nodo de salida y son propagados hacia atrás para brindar los valores Δ_j de la capa oculta.

La regla de propagación para los valores Δ se calcula con la Ecuación 2-5.

$$\Delta_j = g'(in_j) \sum_k w_{j,k} \Delta_k \tag{2-5}$$

Ahora la regla para la actualización de los pesos entre las entradas y la capa oculta se convierte en la misma que la regla de actualización de la capa de salida, y se calcula con la Ecuación 2-6 [46].

$$w_{i,j} = w_{i,j} + \alpha \times a_i \times \Delta_j \tag{2-6}$$

Por lo tanto, el proceso de la propagación hacia atrás se puede resumir en dos pasos principales:

- Calcular los valores Δ para la unidades de salida, usando el error observado.
- Comenzando con la capa de salida, se repite lo siguiente para cada capa de RNA, hasta llegar a la primera capa oculta:
  - Propagar los valores Δ hacia la capa previa.
  - Actualizar los pesos entre las dos capas.

2.7.5. Aprendizaje profundo

El aprendizaje profundo o *deep learning* es una forma de aprendizaje automático que permite que las computadoras aprendan de la experiencia y el entendimiento del mundo en términos de una jerarquía de conceptos. Debido a que la computadora reúne aprendizaje de la experiencia, no es necesaria la operación humana de manera formal para especificar todo el conocimiento requerido por la computadora. La jerarquía de conceptos hace que la computadora aprenda conceptos complicados a través de su propia construcción [61, 62].

El aprendizaje profundo se puede definir como una técnica compuesta por múltiples capas que procesan información de manera no lineal y la transforman en una representación de un nivel de mayor complejidad y abstracción, para la extracción y transformación

supervisada o no supervisada de características, y para el análisis de patrones y clasificación [63].

La clave del aprendizaje profundo es que las capas de características no son diseñadas por humanos, sino que son aprendidas de los datos usando un procedimiento de aprendizaje de propósito general.

La forma más común del aprendizaje profundo y del aprendizaje automático en general, es el aprendizaje supervisado. Con este enfoque se calcula una función objetivo que mide el error o la distancia entre las puntuaciones de salida y los patrones deseados de las puntuaciones. Después, el sistema modifica sus parámetros ajustables internos para reducir este error. Estos parámetros son los pesos, y son números que definen la función entrada-salida. En un sistema típico de aprendizaje profundo hay miles de millones de estos pesos ajustables, al igual que hay miles de millones de ejemplos etiquetados con los que se hace el entrenamiento.

Para ajustar de manera apropiada el vector de los pesos, el algoritmo de aprendizaje calcula un vector de gradiente que, por cada peso, indica por cuanto debería incrementar o reducir el error si los pesos fueran incrementados por una pequeña cantidad. Por lo tanto, el vector de pesos es ajustados en la dirección opuesta al vector de gradiente.

La función objetivo, promediada a través de todos los ejemplos de entrenamiento, puede ser vista como un tipo de "campo con relieve" en un espacio multidimensional de valores de pesos. El vector de gradiente negativo indica la dirección del descenso en el campo, acercándose hacia un mínimo, donde el error de salida es menor en promedio.

En la práctica, el procedimiento que más se utiliza es el descenso de gradiente estocástico (DGE), el cual consiste en mostrar una serie de ejemplos al vector de entrada, calculando las salidas y los errores a través del cálculo del gradiente promedio para esos ejemplos y ajustando de acuerdo a eso los pesos. Este proceso se repite para múltiples conjuntos pequeños de ejemplos del conjunto de datos de entrenamiento hasta que el promedio de la función objetivo deje de decrementar. Se le llama estocástico porque cada conjunto pequeño de ejemplos da un estimado ruidoso del gradiente promedio de todos los ejemplos.

Después del entrenamiento, el rendimiento del sistema es medido en un conjunto de ejemplos diferente, llamado conjunto de prueba. Este proceso sirve para probar la habilidad de generalización del sistema, la cual produce respuestas sensibles a nuevas entradas que no fueron vistas durante el entrenamiento.

Por lo tanto, una arquitectura de aprendizaje profundo es un pilar de múltiples capas de módulos simples, las cuales son sometidas a aprendizaje y a mapeos de cálculos no lineales de entrada-salida. Cada módulo en el pilar transforma su entrada para incrementar la selectividad y la invarianza de la representación. Con múltiples capas no lineales, con una profundidad mínima de 5 a 20, un sistema puede implementar funciones extremadamente intrincadas de sus entradas que son simultáneamente sensibles a detalles minuciosos e insensibles a variaciones irrelevantes de los datos [64].

Una manera de calcular el gradiente de una función objetivo es a través del procedimiento de propagación hacia atrás, el cual se refiere a una aplicación de la regla de la cadena para derivaciones. La regla de la cadena permite saber como son compuestos dos efectos, el pequeño cambio de x en y y el de y en z. El cambio de Δx en x se transforma primero en un cambio Δy en y, siendo multiplicado por $\frac{\delta y}{\delta x}$, como se muestra en la Ecuación 2-7.

$$\Delta y = \frac{\delta y}{\delta x}\Delta x \tag{2-7}$$

De manera similar, el cambio de Δy crea un cambio Δz en z, como se muestra en la Ecuación 2-8.

$$\Delta z = \frac{\delta z}{\delta y}\Delta y \tag{2-8}$$

Después, se vuelve Δx en Δz a través del producto de las derivadas parciales, como se muestra en la Ecuación 2-9.

$$\Delta z = \frac{\delta z}{\delta y}\frac{\delta y}{\delta x}\Delta x \tag{2-9}$$

Finalmente, se sustituyen las ecuaciones para crear la regla de la cadena, como se muestra en la Ecuación 2-10.

$$\frac{\delta z}{\delta x} = \frac{\delta z}{\delta y}\frac{\delta y}{\delta x} \tag{2-10}$$

En la Figura **2-12** [64] se presenta una RNA multicapa en donde primero se hace el cálculo de alimentación hacia adelante, yendo desde la capa de entrada hasta la capa de salida a través de las capas ocultas.

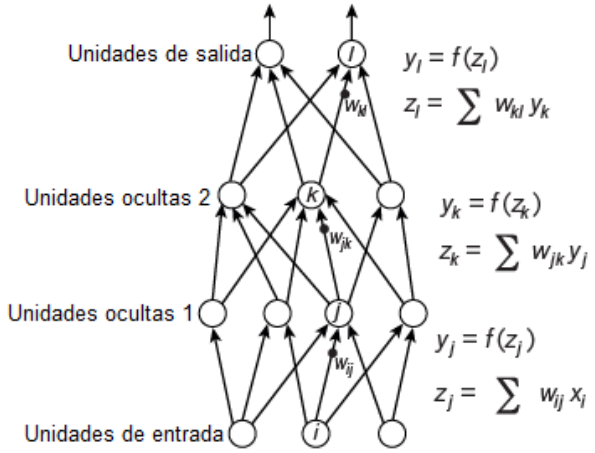

Figura 2-12.: RNA multicapa.

En cada capa primero se calcula la entrada total z de cada unidad, la cual es la suma ponderada de las salidas de las unidades de la capa anterior. Después, se aplica una función no lineal $f(\cdot)$ a z para obtener la salida de la unidad.

Dentro de las funciones no lineales usadas se encuentran la unidad lineal rectificada (ReLU, por sus siglas en inglés), $f(z) = max(0, z)$, y las funciones sigmoideas más convencionales, como la tangente hiperbólica, $f(z) = (exp(z) - exp(-z))/(exp(z) - exp(-z))$, y la función logística, $f(z) = 1/(1 + exp(-z))$.

En la Figura **2-13** [64] se muestran las ecuaciones y los pasos para el cálculo de la propagación hacia atrás. En cada capa oculta se calcula la derivada del error con respecto a la salida de cada unidad, siendo la suma ponderada de la derivada de los errores con respecto al total de entradas de las unidades de la capa anterior. Después, la derivada del error se convierte con respecto a la salida en la derivada del error con respecto a la entrada a través de multiplicarlo por el gradiente de $f(z)$.

En la capa de salida, la derivada del error con respecto a la salida de una unidad es

calculada diferenciando la función de costo, obteniendo $y_l - t_l$ si la función de costo por unidad l es $0.5(y_l - t_l)^2$, donde t_l es el valor objetivo. Una vez que es calculado el valor de $\delta E/\delta z_k$, la derivada del error para el peso w_{jk} en la conexión de la unidad j en la capa posterior es $y_j \delta E/\delta z_k$ [50, 60].

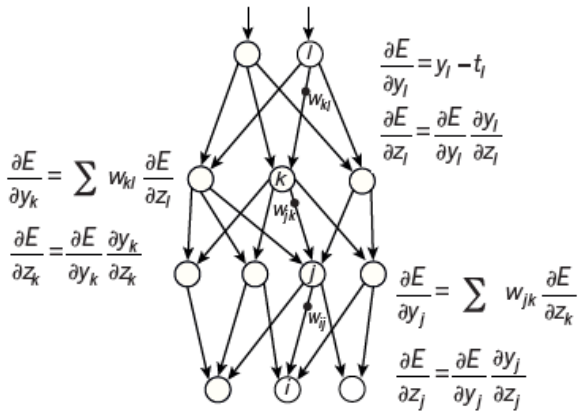

Figura 2-13.: Cálculo de la propagación hacia atrás en la RNA multicapa.

2.7.6. Evaluación

Evaluar es la clave para hacer un progreso real a través de la IA. Existe una serie de métodos de evaluación que se ajustan a diferentes necesidades que pueden estar presentes en los problemas.

2.7.6.1. Entrenamiento y prueba

Dentro de estos métodos se encuentra el de entrenamiento y prueba. Para problemas de clasificación es común medir el rendimiento de clasificación en términos de la tasa de error. El clasificador predice la clase de cada ejemplo, si es correcto, se cuenta como un éxito, sino como un error. La tasa de error se refiere a la proporción de errores obtenida del conjunto total de ejemplos y mide el rendimiento general del clasificador.

Para predecir el rendimiento de un clasificador en datos nuevos, es necesario evaluar su tasa de error en un conjunto de datos que no fue parte de la información que se le presentó al clasificador, que son los datos de entrenamiento. Este conjunto de datos independiente se refiere a los datos de prueba y se asume que ambos conjuntos, el de

entrenamiento y el de prueba, son pruebas representativas del problema subyacente.

También se suelen usar tres conjuntos de datos para calcular la tasa de error: los datos de entrenamiento, los datos de validación y los datos de prueba. Los datos de entrenamiento son sometidos a uno o más métodos de aprendizaje para crear el o los clasificadores. Los datos de validación son usados para optimizar los parámetros de los clasificadores o para seleccionar alguno en particular. Finalmente, los datos de prueba son usados para calcular la tasa de error del método final optimizado.

Generalmente, entre mayor sea la muestra de entrenamiento, mejor será el clasificador, aunque el resultado comienza a disminuir una vez que se excede cierto volumen de datos de entrenamiento, y entre mayor sea la muestra de prueba, más precisa será la estimación del error, la cual puede ser estimada estadísticamente [50, 65].

2.7.6.2. Predicción del rendimiento

En estadística, una sucesión de eventos independientes que triunfan o fracasan se conoce como "proceso de Bernoulli". En un proceso de Bernoulli donde su número de pruebas es N, S son las pruebas exitosas y la tasa de éxito observada es $f = S/N$, se puede calcular la tasa verdadera de éxito, p, la cual hace referencia a un intervalo de confianza.

La media y la varianza de una prueba de Bernoulli, con una tasa de éxito p, son p y $p(1-p)$, respectivamente. Si se toman N pruebas de un proceso de Bernoulli, la tasa de éxito esperada $f = S/N$ es una variable aleatoria con la misma media, p, y la varianza es reducida por un factor de N a $p(1-p)/N$. Para un valor de N grande, la distribución de esta variable aleatoria se aproxima a la distribución normal. La probabilidad de que una variable aleatoria X, con media de cero, tenga un cierto rango de confianza de ancho de $2z$ es $Pr[-z \leq X \leq z] = c$.

Para una distribución normal, los valores de c y los valores correspondientes de z son dados en tablas establecidas de estadística [50].

2.7.6.3. Validación cruzada

En la validación cruzada se decide un número específico de "pliegues" o de particiones de los datos. La manera estándar de predecir la tasa de error en una técnica de aprendizaje, es usar validación cruzada de 10 pliegues o $10-foldcross-validation$, en donde los datos son divididos de manera aleatoria en 10 partes, donde las clases son representadas en aproximadamente las mismas proporciones que en el conjunto de datos completo. Cada parte se presenta en un diferente turno y el esquema de aprendizaje es entrenado con las nueve décimas partes restantes del conjunto de datos. Este procedimiento se repite 10 veces, logrando que al final cada ejemplo haya sido usado exactamente una vez para prueba. Finalmente, el estimado de los 10 errores es promediado para calcular estimado general del error.

Específicamente, el número 10 en las particiones ha sido probado por numerosos conjuntos de datos, con diferentes técnicas de aprendizaje, mostrando que es el número indicado de pliegues para tener el mejor estimado del error, teniendo evidencia teórica que lo respalda en la literatura [66].

2.7.6.4. Leave-one-out

La validación de $Leave-one-out$ o "dejar uno fuera", es simplemente una validación cruzada con n pliegues, donde n es el número de ejemplos en el conjunto de datos. En cada turno un ejemplo es dejado fuera, y el método de aprendizaje es entrenado con todos los ejemplos restantes. Se juzga por su corrección en la instancia restante, uno o cero para el éxito o el fracaso, respectivamente. Los resultados de los n juicios, uno por cada miembro del conjunto de datos, son promediados, representando el error final estimado.

Este procedimiento tiene dos ventajas principales. Una de ellas es que, entre mayor cantidad de datos sea usada para el entrenamiento en cada caso, la posibilidad de que el clasificador sea más preciso incrementa. La otra ventaja es que el procedimiento es determinístico, es decir, no usa muestreo aleatorio. Sin embargo, se presenta la desventaja de que tiene un alto costo computacional, ya que el procedimiento de aprendizaje debe

de ser ejecutado n veces, por lo que no es factible para bases de datos muy grandes, pero en bases de datos pequeñas se obtiene el estimado más preciso posible [65].

2.7.6.5. Contando el costo

Además de medir la precisión de las evaluaciones, es necesario medir el costo de hacer decisiones o clasificaciones erróneas. Optimizar la tasa de clasificación sin considerar el costo de los errores regularmente conduce a resultados extraños.

En un caso de dos clases, con clases *si* y *no*, una predicción tiene las cuatro posibles salidas mostradas en la Tabla **2-2**. Los verdaderos positivos (VP) y verdaderos negativos (VN) son las clasificaciones correctas. Un falso positivo (FP) ocurre cuando la salida es predicha incorrectamente como *si* (positiva) cuando en realidad es *no* (negativa). Un falso negativo (FN) ocurre cuando la salida es predicha incorrectamente como *no* (negativa) cuando en realidad es *si* (positiva).

Tabla 2-2.: Posibles salidas de una predicción de dos clases.

		Clase predicha	
		Si	No
Clase actual	Si	Verdaderos positivos	Verdaderos negativos
	No	Falsos positivos	Verdaderos negativos

La tasa de VP es su división entre el número total de positivos, $TP + FN$; mientras que la tasa de FP es su división entre en el número total de negativos, $FP + TN$. La tasa de éxito general es el número de la clasificaciones correctas dividido entre el número total de clasificaciones, como se muestra en la Ecuación 2-11. Finalmente, la tasa de error es 1 - Ecuación 2-11 [50].

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{2-11}$$

2.7.6.6. Curvas ROC

Las curvas ROC (*receiver operating characteristic*) son una técnica gráfica para evaluar los diferentes esquemas en el aprendizaje automático, en donde el aprendiz está

tratando de seleccionar muestras de ejemplos de prueba que tienen una alta proporción de positivos. Estas curvas describen el rendimiento de un clasificador sin considerar la distribución de la clases o el costo del error.

En el eje vertical se grafica el número de positivos incluidos en la muestra, expresado como el porcentaje del número total de positivos, contra el número de negativos incluidos en la muestra, expresado como el porcentaje del número total de negativos, en el eje horizontal. Cada punto de la gráfica corresponde a dibujar una línea de cierta posición de la lista de clasificaciones, contando los *si*'s y *no*'s que hay sobre ella y graficándolos de manera vertical y horizontal, respectivamente [67]. En la Figura **2-14** [68] se muestra el ejemplo de una curva ROC que presenta una proporción de sensibilidad - especificidad significativa.

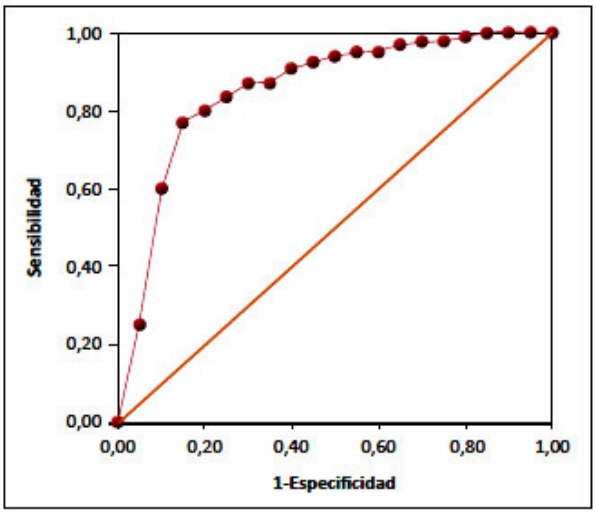

Figura 2-14.: Ejemplo de una curva ROC.

3. Artículos publicados

ISSN 1870-4069

An Analysis of Dietary and Demographic Data in Oral Health, Data from the National Health and Nutrition Examination Survey: A Preliminary Study

Nubia M. Chávez-Lamas[1], Laura A. Zanella-Calzada[2], Carlos E. Galván-Tejada[2]

[1] Universidad Autónoma de Zacatecas, Unidad Académica de Odontología, Clínica Comunitaria de Tacoaleche, Zacatecas, Zacatecas, Mexico

[2] Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, Zacatecas, Zacatecas, Mexico

{lzanellac, ericgalvan}@uaz.edu.mx

Abstract. Dietary and demographics features has an influence in the general health status of the population in the world. Therefore in this paper is proposed an univariate analysis of 7 dietary and demographic features to study the impact on the oral status in order recognize the different determinants that contribute to modify in a negative way the oral health status. Univariate analysis is carried on applying an multi objective linear regression and evaluated in terms of area under the curve (AUC), p-value, sensitivity and specificity. Additionally, a multivariate model is done to evaluate confounder to increase the AUC. Preliminary results shows that individually water source is an important feature that affects oral health status.

Keywords: oral health, health status, linear regression, statistical analysis, univariate model, multivariate model.

1 Introduction

The World Health Organization (WHO), defined health as a physical, mental and social healthy status, not only the absence of diseases. This definition as evolved from that conceptual concept to a serie of quantitative scales that allows be measured of the general health status, therefor in 1994, the WHO propose the concept of quality of life as an individual perception of his life position, inside the context of cultural environment and the relationship with their objectives, expectations, standards and concerns. This new aspects that comprise the definition of health implies that behavior patterns, expectations

and cultural are independent for each group, consequently, to evaluate this, three dimensions are proposed: General (lifestyle), particular (life conditions) and singular (type of life) [18,13].

Oral health is included in the general dimension, being an essential component to life quality of individuals, hence, nowadays oral health is a determining factor in individuals general health and of communities. For that reason, several studies about which characteristics of life, diet, demographic, social and others influence in the oral health are becoming an interesting niche study [18,13,7,15]. However, these studies can be influenced by several characteristics from the particular individuals, as was mentioned before, therefore, this features to evaluate oral health implies that varies depending on the condition of the country, genetic heritage, even political and public health services for each country [18].

These studies are done using a methodological scientific (survey), which is useful to prioritize, identify and solve oral health, moreover allows to generate prediction models that helps to decrease the incidence and prevalence of oral diseases. Is well-known that one of the main illness in oral health that affects 90 percent of the world population is dental caries and periodontal illness in a 80 percent [19], besides those are concentrated mainly in communities less favored by what are considered as oral health problems [15]. In last three decades, researchers develop different purpose surveys to evaluate life quality and the relationship with oral health, for instance, Social Impacts Of Dentals Disease, Geriatric Oral Health Assessment index, Dental Impact, Dental Impact on Daily Living, Oral Health Impact Profile, Oral Impacts on Daily Performances [9].

Through these instruments it is possible to recognize the different determinants that contribute to modify in a negative way the oral health status since its appearance depends on the conjugation of biological factors such as dental anatomy, diet where it is widely demonstrated the relationship between consumption and the appearance of oral diseases, both by historical evidence, observational, clinical studies and experimentation; socio-economic level, area of residence, educational level, occupation, housing characteristics, income, opportunities for general education, health, dental care as well as age and sex [17,4,14,8,19,6]

There are many risk factors and determinants of importance related to oral diseases and it is evident that the greater the degree of exposure to risk, the greater the probability of contracting or developing a condition [8], hence the importance of considering the present analysis that aims to determine some of the socioeconomic and dietary characteristics that directly influence oral health status. These may be considered as parameters that provide information on the normal or pathological state of an individual at a given time, a risk group and a specific place, in addition to helping to understand the oral health-disease process, in addition to establishing specific primary health care measures such as promotion, prevention, diagnosis, treatment and rehabilitation in a timely manner [2].

Therefore, the main contribution of this paper is to analyze as univariate models, individual demographic and dietary characteristics and the relationship

of these with a general oral health status, as well as how an interaction of these characteristics (as a multivariate model) with the general status.

The analysis is carried on using a multi objective logistic regression and evaluated using a Receiver Operating Characteristic (ROC) curve, which allows us to study the sensitivity and specificity of each characteristic, just as the model comprised by all the selected characteristics in order to explain the relationship between demographic and dietary features with the oral health status.

This paper is organized as follows, in section 2 is presented a description of the data set used for this research and methods to carry on the study. In section 3 the experimentation conditions are presented. Results from univariate and multivariate analysis are shown in section 4. Finally conclusions and future work is described in section 5.

2 Materials and Methods

In this section is described the data set of National Health and Nutrition Examination Survey (NHANES), patients selection and the methods used to carry on the univariate analysis.

2.1 Data set Description

The NHANES is a national program that design studies to perform a survey to assess the health and nutritional status of adults and children in the United States, including all ethnic groups. The survey combines interviews and physical examinations, allowing to develop studies using clinical, para-clinical and demographic characteristics (features) of individuals. NHANES is a program founded by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.

Content Description NHANES survey include several types of interviews, to cover a wide range of features, including demographic, socioeconomic, dietary, and health-related questions, described in detail in 1. One examination component that is critical to this study is dental care examination, which is carried on by trained medical personnel.

Table 1. NHANES data description.

Questionnaire Type	Description
Demographics	The demographics file provides individual, family, and household level information.
Examinations	Public health significance in areas of surveillance, prevention, treatment, dental care utilization, health policy, evaluation of Federal health programs.
Dietary	Total nutrient intake.
Laboratory	Laboratory tests, which includes Cholesterol, Fasting Questionnaire, Hepatitis Tests, HIV, Urinary tests.
Questionnaire	information on: Acculturation, Alcohol Use, Health Insurance, Income.
Medication	Past 30 days, used or taken medication.

Meta Data Health and Nutrition Examination Surveys are comprised by interviews applied to 27,631 persons. In Table 2 are described in detail the number of persons and the features/parameters of the groups included in the sample.

Table 2. Patient demographic information.

Characteristic	NHANES
Age of civilian	All ages from birth
Geographic areas	Unaited States
Average number of sample persons per household	2
Number of study locations	60
Domains for oversampling	Predesignated: 87 subdomains of sex-age groups for non-Hispanic black persons, non-Hispanic non-black Asian persons, and Hispanic persons. Oversampled: Hispanic persons, non-Hispanic black persons, non-Hispanic non-black.
Number of selected persons	27,631
Number of interviewed persons	20,491
Number of examined persons	19,644

2.2 Data Analysis

In this work, as mentioned before, were conducted an univariate and multivariate searches. The univariate search was conducted using the demographics and dietary features being subjected to a statistical analysis; while the multivariate search was conducted using the complete feature set, including the examinations features.

The statistical analysis consisted of submitting each of the features to a linear regression to obtain the univariate models; then, all features together developed a multivariate model trough a linear regression too.

Linear regression is an analysis which consists on a statistical technique for modeling the relationship between features. The simplest representation of a model obtained by this method is represented in Equation 1, where y is the dependent variable or the outcome feature, β_0 is the intercept, β_1 is the slope and x is the independent variable or the analyzed feature. Models can be composed by the number of terms needed. Finally, the difference between the real outcome and the outcome proposed by the model is the error of the model, ϵ; this variable

is known as a statistical error due to it's a random variable that measures the model failure for fitting the data exactly [16]:

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{1}$$

The statistical validation consisted on obtaining the P-value, the odds ratio and the area under the receiver operating characteristic (ROC) curve, known as the AUC.

The P-value or the observance significance level is the smallest value obtained where the null hypothesis can be rejected [13]; while the AUC is a standard method to evaluate the accuracy of the classification model [12], finally, the odds ratio represents the ratio of the probability that an event of interest occurs against the probability that the same event doesn't occur [3].

All the data analysis were realized with the free software *R* (version 3.3.1) [20] and its packages, *pROC* (Version 1.8, 2015-06-10) [21], *epitools* (Version 0.5-7, 2012-09-30) [1], *ResourceSelection* (Version 0.2-6, 2016-02-15) [10] and *randomForest* (version 4.6-12, 2011-10-18) [11].

3 Experiments

The process realized in the univariate and multivariate models experimentation and their statistical analysis is described in this section. In Figure 1 is presented a flowchart of the followed methodology.

Firstly, the NHANES data analysis and the features selection for this research were realized by an expert dentist, according to the information obtained from the literature. Then, a new dataset with the extracted features was obtained (Figure 1 A)). Followed by a data preprocessing, where is initially described the imputation of missing data (Figure 1 B)) and then, for the univariate analysis, it was necessary removing the features that are unable to contribute meaningful information due to their contents. The univariate models development and models validation were realized in base of an statistical process (Figure 1 C)).

Finally, for the multivariate model development all features were taking into account and the model validation is also carried out (Figure 1 D)). The validation process is performed with the intention of evaluating the contributions of the results.

3.1 Dataset Preprocessing

The dataset used for this research was mostly contained by demographic and examination features, also a dietary feature was present [5] .

- Demographic features:
  - DMDMARTL: Describes the marital status,
  - INDFMIN2: Describes the annual family income,

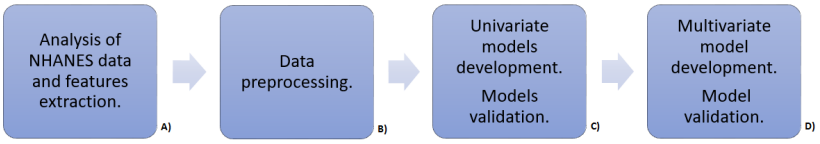

Fig. 1. Flowchart of the methodology followed. A) Dataset analysis and features extraction, B) Univariate models development and their statistical validation, C) Multivariate model development and its statistical validation.

  - RIAGENDR: Describes the gender of the participant,
  - DMDHHSIZ: Describes the total number of people in the household,
  - DMDEDUC3: Describes the education level on youths from 6 to 19 years,
  - DMDEDUC2: Describes the education level on adults from 20 years old.
- Dietary feature:
  - DR1TWS: Describes the tap water source.
- Outcome feature:
  - OHDEXSTS: Describes the overall oral health exam status (from 1 being complete to 3 being not done).

Examination features were initially removed for the univariate analysis, since they contain information about the health center where the patient was treated and each health center is represented by a different feature, therefore, each feature presents a large amount of missing data, becoming impossible to perform an univariate analysis for them.

The remaining features contained some missing values represented as Not a Number (NaN). These missing values were imputed through the $rfimpute$ function from $randomForest$ package, consisting on the substitution of NaN's for the value of the median of the column where the features are found.

3.2 Univariate Models Development and Models Validation

The development of the univariate models was realized by a linear regression process between each of the features and the outcome feature, which was related to the general status of oral health.

After the univariate models were obtained, they were subjected to an univariate statistical analysis, obtaining their P-values, odds ratios, AUC and ROC curves.

3.3 Multivariate Model Development and Model Validation

For the multivariate model development, which was realized by a linear regression between the feature set and the outcome feature, was also included the examination features that weren't used for the univariate analysis. The

information contained in the examination features was all joined in only one feature, avoiding problems in the model development because of the missing data, taking into consideration that for each patient these features only contained information in one of them, because they were related to the health center where the patients were attended and most of them were attended only in one.

Finally, for the multivariate model validation were obtained the P-value, AUC and ROC curve of the model. Odds ratio wasn't calculated for this model because this parameter is obtained for specific features and not for a set of them.

4 Results

Results obtained from the univariate statistical analysis are presented in Table 3, which is contained by every feature of the dataset and its respective P-value, odds ratio and AUC, in ascendant order according to the AUC values. For this analysis, features related with the health center where the patients were treated (examination features), weren't taken into account.

Table 3. Univariate statistical analysis.

Feature	P-value	AUC	Odds ratio	2.5%	97.5%
DMDMARTL	0.683	0.496	0.999	0.994	1.003
INDFMIN2	0.326	0.501	0.999	0.998	1.000
RIAGENDR	0.299	0.502	0.999	0.999	1.000
DMDHHSIZ	0.585	0.510	1.001	0.996	1.006
DMDEDUC3	0.744	0.521	1.005	0.997	1.003
DMDEDUC2	0.472	0.521	1.003	0.993	1.013
DR1TWS	0.101	0.540	1.000	0.999	0.101

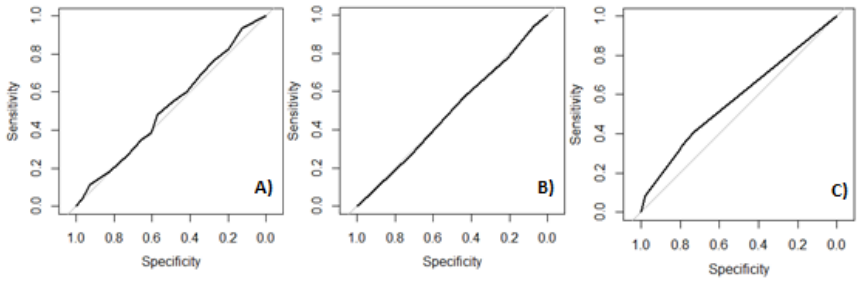

Fig. 2. ROC curves obtained from the most significant features in the univariate analysis, A) DMDEDUC3, B) DMDEDUC2, C) DR1TWS.

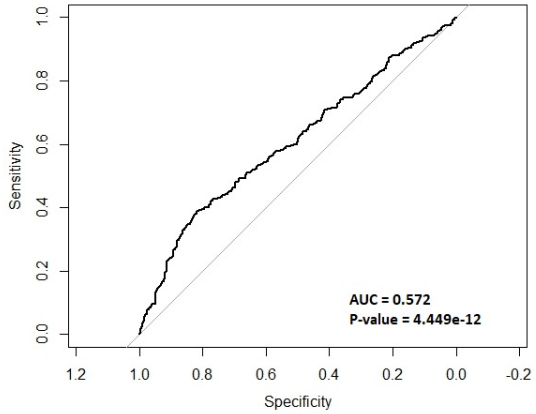

Fig. 3. ROC curve obtained from the multivariate analysis (AUC = 0.572, P-value = 4.449e-12).

From the univariate analysis results, the three most representative univariate models, according to their AUC values, present their ROC curves in Figure 2. In Figure 2 A) is shown the ROC curve of DMDEDUC3 feature, in Figure 2 B) is shown the ROC curve of DMDEDUC2 feature and in Figure 2, C) is shown the ROC curve of DR1TWS feature.

In multivariate statistical analysis, the ROC curve generated is shown in Figure 3, obtaining a P-value of 4.449e-12 and a value of AUC of 0.572. For this analysis, features related to the health center where the patients were treated (examination features) were taken into account.

5 Conclusion and Future Work

It is well known that the health problems that arouse the greatest interest are those that represent a risk of death or permanent disability, and carry with them the doubt as to the possibility of attacking a certain person.

Commonly, oral health problems do not arouse the spontaneous interest of the community, which is why identifying the determinants through such analysis in the population allows informing, educating and motivating appropriate oral health care because it depends on whether the population shows interest in receiving different treatments and thus improve oral health.

This analysis has shown that around a disease there are many factors that cause or aggravate it. When referring to the issue of socio-cultural health determinants (SHD) has become a latent concern for health professionals and oral health professionals since most of the problems derive from inequality or inequity

in health and are linked to other determinants which produce different effects in each risk group with respect to their conditions and lifestyle, generated in the short, medium or long term.

Demographic features that were used for this research showed that none of them can significantly predict the health status of the population in an univariate approach, this can be observed at the statistical analysis where all of the P-values were > 0.06, which is taken as the standard value to consider a feature to be significant; also, the AUC values and the ROC curves showed a similar result, since the higher AUC is 0.540, which means that the true positives / true negatives proportion is 54% for this specific feature, DR1TWS (dietary feature). Odds ratios didn't show a higher probability than 1 for any feature in relationship with a health condition. Although the predictive capacity of this feature is better than a blind test, its contribution isn't so statically significant.

Multivariate model showed better statistical results than univariate models. In this approach, where the examination or health centers features were also used, the P-value obtained, 4.449e-12, is statistically highly significant, which represents that the alternative hypothesis has a high probability of being true, it means that the model has an influence on the health condition. The AUC and the ROC curve presented a true positives / true negatives proportion of 57.2%, which is higher than the AUC values of the univariate models. The improvement of statistical values in the multivariate model allows to conclude that the demographic and dietary features can help to evaluate the health status in combination with the examination features. By this, it's possible to consider that social security, economic income, type of health service and other similar factors may help to determine the patient's health status, specifically the oral health, according to the data used for this work.

As future work is proposed add more dietary and medical condition features, but including a clever feature selection than can lead to a high AUC model cleaning features that can decrease specificity and sensitivity of a oral health prediction model, in addition, complex techniques of machine learning, as neural networks, elastic networks or similar can be used to tackle this problem.

References

1. Aragon, T.: EpiTools: Epidemiology Tools. R package version 0.5-7 (2016)
2. Arango, V., Sandra, S.: Biomarcadores para la evaluación de riesgo en la salud humana. Revista Facultad Nacional de Salud Pública 30(1), 75–82 (2012)
3. Bland, J.M., Altman, D.G.: The odds ratio. Bmj 320(7247), 1468 (2000)
4. Castañeda Abascal, I.E., Lok Castañeda, A., Molina, L., Manuel, J.: Prevalencia y factores pronósticos de caries dental en la población de 15 a 19 años. Revista Cubana de Estomatología 52, 21–29 (2015)
5. for Disease Control, C., Prevention, et al.: National health and nutrition examination survey, 2011-2012 (2013)
6. Escalona, T.P., Ortiz, H.R.C., Palomino, Y.P., Tamayo, M.I., Rodríguez, M.I.R.: 08-relación entre factores de riesgos y caries dental relationship between risk factors and dental caries. MULTIMED Revista Médica Granma 19(4) (2017)

7. Espinosa González, L.: Cambios del modo y estilo de vida; su influencia en el proceso salud-enfermedad. Revista Cubana de Estomatología 41(3), 0–0 (2004)
8. Gispert Abreu, E.d.l.Á., Castell-Florit Serrate, P., Herrera Nordet, M.: Salud bucal poblacional y su producción intersectorial. Revista Cubana de Estomatología 52, 62–67 (2015)
9. González, C.F., Franz, L.N., Sanzana, N.D.: Determinantes de salud oral en población de 12 años. Revista clínica de periodoncia, implantología y rehabilitación oral 4(3), 117–121 (2011)
10. Lele, S.R., Keim, J.L., Solymos, P., Solymos, M.P.: Package ResourceSelection (2017)
11. Liaw, A., Wiener, M.: The randomforest package. R News 2(3), 18–22 (2002)
12. Lobo, J.M., Jiménez-Valverde, A., Real, R.: Auc: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography 17(2), 145–151 (2008)
13. Martínez Abreu, J., Capote Femenias, J., Bermúdez Ferrer, G., Martínez García, Y.: Determinantes sociales del estado de salud oral en el contexto actual. MediSur 12(4), 562–569 (2014)
14. Martínez Abreu, J., Castell-Florit Serrate, P., Llanes Llanes, E., Morales Aguiar, D.R., Sánchez Barrera, O., et al.: Componente bucal y determinantes sociales en el análisis de la situación de salud. Revista Cubana de Estomatología 52, 53–61 (2015)
15. Medina Solís, C.E.: Políticas de salud bucal en México: disminuir las principales enfermedades. Una descripción (2006)
16. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis. John Wiley & Sons (2015)
17. Narváez Chávez, A.M.: Asociación entre el conocimiento de los padres sobre salud bucal y uso de técnicas educativas con relación a la presencia de biofilm y caries en infantes. Master's thesis, Quito: UCE (2017)
18. Ojeda-Garcés, J.C., Oviedo-García, E., Salas, C.E.S.: Streptococcus mutans y caries dental. Odontologica 26(1) (2013)
19. Ospina, D., Herrera, Y., Betancur, J., Agudelo, H.B., López, A.P.: Higiene bucal en la población de san francisco antioquia y sus factores relacionados. Revista Nacional de Odontología 12(22), 23–30 (2016)
20. Ripley, B.D.: The R project in statistical computing. MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network 1(1), 23–25 (2001)
21. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., Müller, M.: pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics 12(1), 77 (2011)

Multivariate features selection from demographic and dietary descriptors as caries risk determinants in oral health diagnosis: data from NHANES 2013-2014

Laura A. Zanella-Calzada[1]*, Carlos E. Galván-Tejada[1]*, Nubia M. Chávez-Lamas[2], Jorge I. Galván-Tejada[1], José M. Celaya-Padilla[1]
[1]Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, 98000 Zacatecas, Zac, México
[2]Clínica Comunitaria de Tacoaleche, Unidad Académica de Odontología, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, 98000 Zacatecas, Zac, México
Email: {lzanellac, ericgalvan}@uaz.edu.mx

***Abstract*—The main pathology that make up the epidemiological profile of oral health status is dental caries. This pathology is considered a challenge because its high prevalence, besides being a chronic but preventable disease, which can be caused by a series of different demographic, dietary and laboratory data, among others. Therefore, in this research two multivariate models were proposed for the classification between control patients ('0'), presence of caries ('1') and presence of restorations ('2'). Analyzing 189 demographic and dietary features from NHANES 2013-2014, by FBS method based on the P-value they were reduced to 24 features, which were separated in two sets according to their demographic or dietary information. Both datasets were subjected to a statistical analysis based on linear regression, obtaining their residuals, AUC, ROC and OR values, in order to validate their classification accuracy. For each model were obtained significant statistical results and none of them presented problems related to outliers, according to their residuals. Both models shown P-values > 0.05 for most features and small OR confidence intervals. Finally, demographic data model obtained an AUC = 0.664 and dietary data model obtained an AUC = 0.633, averaging the AUC values for each outcome, proving to be statistically significant. By these results it's concluded that these specific demographic and dietary features are significant determinants for estimating the oral health status patients, based on their likelihood of developing caries.**

***Index Terms*—NHANES, Oral health, Fast backward variable selection, Statistical analysis.**

I. Introduction

Epidemiological studies are the starting point for planning, implementing and evaluating health programs. For the oral component of the health-disease process, the World Health Organization (WHO), as a technical advisory body for public health, has promoted epidemiological studies in different regions of the world, with the purpose of knowing how health and disease are distributed and, based on this information, to advance in the development of policies and programs for the care and transformation of this object of study, according to the political, economic and social reality of each country.

Dental caries (90%), periodontal disease (80%) and malocclusions (75%) have been recognized for several decades as the main pathologies that make up the epidemiological profile of oral health status, which are associated with different determinants and risk factors, in these studies the most commonly used epidemiological indexes in oral health are Decayed, Missing, Filled (DMF) and Decayed, Missing, Filled in Temporal (DMFT) for dental caries in permanent and temporal dentition, respectively; the Community Periodontal Index (CPI) for periodontal disease and the Angle classification (modification Dewey - Anderson) for malocclusions [1].

Oral health is considered to be an integral and essential part of health in general and this can compromise the quality of life considering all groups of people such as infants, adolescents, young people, adults and older adults without distinguishing between age, sex, social level, cultural, low economic, educational, marital status, previous caries history, current caries index, levels of microbiological factors, drawing of family caries, the environment in which they are found and in particular the type of drinking water [2].

It was historically known that the prevalence of dental caries was higher in urban areas than rural in some parts of the world, however, according to recent studies, it was found that over time, the prevalence and quantity of dental caries in urban areas have been significant reduced while in rural areas have been increased, becoming important to supply rural areas in order to have fair and equal access of all citizens to oral health care [3]. On the other hand, according to Petersen, besides the economic position, studies present that some factors in the daily diet may affect the prevalence of caries, as the sugar intake [4].

In oral public health, dental caries is a challenge because it is a chronic and preventable disease. In the early stages of human life, teeth are generally free of the disease, but throughout their life they can acquire it by different factors such as those mentioned above and this is the main reason for

978-1-5386-2363-3/18/$31.00 ©2018 IEEE

attention of such conditions. [5].

The presence or absence of some afflictions due to caries from childhood to adulthood generate pain, inability to smile, swallow, chew and taste mainly, in such a way that these can compromise the psychosocial well-being and influence self-esteem, expression, communication, facial aesthetics, school absenteeism in children, affecting their academic performance, in the working life of the adult population and therefore low economic production; generating a negative impact on health, as well as an increased risk or predisposition to suffer from decompensated diabetes, arthritis, thrombotic events and premature birth [6].

A very important aspect to consider is the chronic nutritional deficiencies (nutrition, overweight or obesity) as these cause delays in the growth of girls which directly affects their physical and mental development, restricting learning opportunities in the different stages of their lives.In addition, they are also at greater risk of developing oral cavity diseases because studies indicate that the Body Mass Index (BMI) is an indicator of the weight-to-size ratio used to describe the nutritional status of the population [7], [8], dental caries and platelet aggregation are directly related, finding that children with these characteristics have a higher risk of developing plaque, which results in caries. There is also the possibility that it may jeopardize the development of craniofacial growth affecting the formation of tissues such as: bone, periodontal ligament and teeth, finding defects in dental enamel related to hypocalcemia or vitamin and mineral deficiency [9].

Dental caries has been described in the scientific literature under different names, it is a multifactorial disease that is characterized by localized and progressive demineralization of the inorganic portions of the tooth and the subsequent deterioration of its organic part, has a high degree of morbidity and high prevalence [10]–[12]. Thus, there are several approaches to identify caries risk determinants, Lam C. et al. [13] found that dental visits, brushing frequency, lower parental perceived importance of baby teeth, and weaning onto solids are determinants associated with plaque accumulation, however they only validate these factors in toddlers. I. B. Fernandes et al. [14] conducted a cross-sectional study of 274-children and their mothers addressing a demographic/socio-economic status, they propose a Poisson regression to carry on the analyses, founding that dental caries can be found in mothers of children aged 1, which demonstrate a relationship between demographic/socio-economic status and caries with children. Other type of risk factor are associated with genetic and dietary data, A. Lips et al. [15], demonstrated an association between genetic polymorphisms and risk of dental caries for most of the salivary proteins. Diabetic and hypertensive patients has dietary risk factors that could lead to oral health problems, Asif Ahmed et al. [16] conducted an statistical analysis using clinical data and clearly identifies that patients with oro-dental problems were hemodynamically stressed.

As mentioned before, according to literature, the study of oral health based on demographic features in Latin America, specifically in Mexico, isn't as large as in the eastern side, where it has been found that this type of features are important for the prevention and prognosis of dental caries; based on the relationship that socioeconomic problems are very similar in both regions, it would be feasible to develop a search in some demographic features in order to find a tool that may facilitate the diagnosis of caries, specially in less developed regions, providing better opportunities in the prevention of this condition and the improvement of oral health.

Therefore, in this work, is analyzed a 189 features set from NHANES 2013-2014 contained by demographic and dietary data, which was reduced to 105 features by data preprocessing. These 105 features were subjected to a fast backward selection (FBS) based on the P-value; by this process were obtained a demographic dataset and a dietary dataset. Then, an evaluation process was carried out for both datasets, obtaining by statistical analysis the residuals and area under the receiver operating characteristic curve (AUC, ROC) values for each multivariate model and, for each feature contained in models were calculated the P-value and odds ratio (OR) confidence intervals with the purpose of validate their classification accuracy.

This paper is organized as follows, after this introduction, materials and methods used in this study are described in section II. Section III presents experimentation carried on, results and discussions of risk determinants founded are described in section IV. Finally, conclusions and future work is presented in section V.

II. Materials and methods

The dataset that was used for this work, from the National Health and Nutrition Examination Survey (NHANES, 2013-20114) is described in this section, as well as the patient's selection and methods that were used for the development of models.

A. Dataset description

NHANES is a federal agency that produces data and materials for public domain from the health and nutritional status of adults and children in the United States, including ethnic groups. The survey combines interviews and physical examinations, allowing to develop studies using clinical, paraclinical and demographic information of subjects. NHANES is an initiative that was founded by the Centers for Disease Control and Prevention (CDC) and the National Center for Health Statistics (NCHS).

1) Content Description: NHANES data include a series of different type of characteristics / features, among them are included demographic data (individual, family, and household level information), dietary data (total nutrient intake), and health-related information (public health significance in areas of surveillance, prevention, treatment, dental care utilization, health policy, evaluation of Federal health programs). A health-related component that is critical to this study is oral health examination, which is carried on by trained medical personnel.

Data that were analyzed for this work were the demographic and dietary, with a total of 189 features and 9812 subjects, being the oral health dentition the feature used for the patient's classification. Oral health dentition is a feature that describes the status of patient's dentition in terms of their condition of caries or restorations.

B. Data analysis

In this work was conducted a multivariate approach with demographic and dietary data, subjecting all features to a preprocessing step to be followed by a FBS method based on the P-value for features selection and finally, a statistical analysis was carried out to evaluate the behavior of the developed models.

The preprocessing step consisted on manually eliminate all features that presented a high percentage of missing data, which were represented by $Not\ a\ Number$ (NaN), then, all the remaining missing values were imputed by the $rfimput$ function from the $randomForest$ package (Version 4.6-12, 2015-10-06), which consists on supplying all NaN with the median of the values from the column where the missing value is located [17].

Finally, columns that were contained by singular values were eliminated; two values are singular if they are multiples of each other.

FBS process consisted on initially subjecting all features to a multinomial logistic regression to obtain a multivariate model.

Logistic regression (LR) is an analysis which consists on a statistical technique for modeling the relationship between the independent features and the dependent feature. This method belongs to the statistical methods where the contribution of different factors in the occurrence of a simple event is measured. The main objective of LR is to model the influence of the probability of an event. The simplest representation of a model obtained by this method is presented in Equation 1, where y from $logit(y = 1)$ is the dependent variable or the outcome feature that it's necessary to be subjected to a logarithmic transformation ($logit$) because the initial equation of the model is of exponential type and by this transformation it's possible to use it as a lineal function, w is an offset term that can be included, β_1 is the slope and x is the independent variable or the analyzed feature. Models can be composed by the number of terms needed [18].

$$logit(y = 1) = w + \beta_1 x \tag{1}$$

From the developed multivariate model, it was realized a FBS from fbw function, which performs a numerically stable version of FBS, using a method based on Lawless and Singhal [19]. The mechanism of the FBS procedure consists on using multiple regressions in that the variable least significant by some criterion (P-value for this work) es removed at each step, until a stopping rule is reached [20]. The P-value or the observance significance level is the smallest value obtained where the null hypothesis can be rejected [21]. For this work, when all remaining features have a significant P-value defined by a threshold, the iterative process will be finished. In FBS it's necessary to be in a situation where the number of observations is greater than the number of features with the purpose of fitting the model.

From the features set obtained by the FBS algorithm, were selected two multivariate models. One of them contained all demographic features while the other contained all dietary features.

Both models were subjected to a linear regression for a later statistical analysis.

Linear regression is an analysis which consists on a statistical technique for modeling the relationship between features, differing from LR in the dependent feature, where for linear regression it is represented by continuous values while in LR it is represented by discrete values. A model obtained by this method is represented in Equation 2, where y is the dependent variable or the outcome feature, β_0 is the intercept, β_1 is the slope and x is the independent variable or the analyzed feature. Models can be composed by the number of terms needed. Finally, the difference between the real outcome and the outcome proposed by the model is the error of the model, ϵ; this variable is known as a statistical error due to it's a random variable that measures the model failure for fitting the data exactly [22].

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

The statistical validation consisted on obtaining the residuals, OR with their confidence intervals, ROC curves and AUC values.

Residuals are obtained from the linear regression, which represents the difference between the observed values of the dependent variable and the predicted values that were obtained by the developed model.

OR is the value of the relation that exists between a feature and an outcome. It represents the probability that an outcome will occur by a particular feature, compared to the probability that the outcome will occur with this feature absent. OR confidence intervals are used to estimate the precision of the OR value; a large confidence interval indicates a poor accuracy of the value, whereas a small confidence interval indicates a high accuracy [23].

AUC value is a standard method to evaluate the accuracy of the classification model based on the relationship between the specificity and the sensitivity [24].

Sensitivity is defined as the proportion of subjects with one condition that were classified as positive; this value is calculated by Equation 3, where TP represents the quantity of true positives and FP represents the quantity of false positives.

$$PPV = \frac{TP}{TP + FP} \tag{3}$$

Specificity is defined as the proportion of subjects without a condition that were classified as negative; this value is

calculated by Equation 4, where TN represents the quantity of true negatives and FN represents the quantity of false negatives.

$$NPV = \frac{TN}{TN + FN} \tag{4}$$

All data analysis was realized with the free software R (version 3.3.1) [25] and its packages, $pROC$ (Version 1.8, 2015-06-10) [26], rms (Version 5.1-1, 2017-05-01) [27], $MASS$ (Version 7.3-45, 2015-11-10) [28], $DAAG$ (Version 1.22, 2015-08-22) [29], and $rminer$ (Version 1.4.2, 2016-08-22) [30].

III. Experiments

The followed experimentation process for multivariate models development and their statistical analysis and validation is presented in this section.

Firstly, NHANES data analysis and features selection were realized by an expert dentist, according to the information obtained from literature.Two databases were used for this research, a demography dataset and a dietary dataset.

For both databases was carried out a data preprocessing procedure, beginning with a manual extraction of features and the imputation of missing values (Figure 1 A). Then, all features were subjected to a feature selection process for the development of two multivariate models using a FBS based on the P-value (Figure 1 B).

Finally, it was realized a statistical analysis and validation process, subjecting both databases to a linear regression to obtain the residuals, OR and AUC values, with their correspondent ROC curves (Figure 1 C).

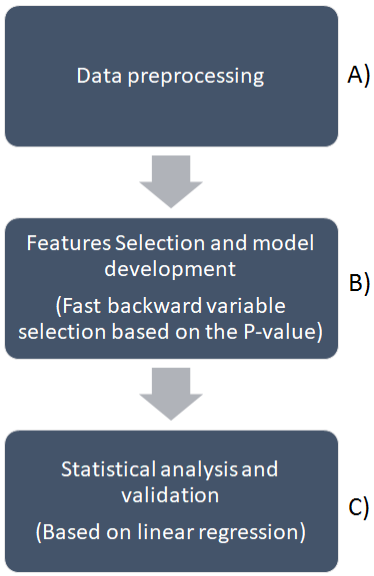

Fig. 1. Flowchart of the methodology followed. A) Database preprocessing, B) Multivariate models development, C) Statistical analysis.

A. Dataset preprocessing

The two datasets used for this research were obtained from NHANES 2013-2014 and they were contained with demographic and dietary information, obtaining a total of 189 features from 9812 patients, with an outcome feature of the oral health dentition based on the caries condition of patients from the examination database, where control patients were represented as '0', patients with presence of caries were represented as '1' and patients with restorations were represented as '2'. All patients (men and women) were in an age range from 1 to 84 years of age [31].

From both databases were manually extracted all features with at least 30% of missing data and with presence of singular values, retaining only 105 features.

The 105 remaining features contained some missing values ($< 30\%$) that were imputed by the $rfimput$ function from $randomForest$ package for R, where all missing data were supplied by the median value of the column where the data were located.

B. Models development and validation

For the development of models, all features were initially subjected to a logistic regression and, later, to a FBS method based on the P-value. For this step was used the $fastbw$ function from rms package. The selected threshold was a P-value < 0.005.

From FBS were selected 24 features from the 105, which according with their P-values, they were the most significant. Then, this features set was divided in two datasets; the first dataset contained all demographic features, while the second contained all dietary features.

Finally, both datasets were subjected to a validation process based on statistical analysis. It was realized a linear regression for each dataset and after obtaining their correspondent general models, were calculated a series ofstatistical parameters.

IV. Results and Discussions

Results obtained from data preprocessing shown a total of 105 remaining features from the original database, with demographic and dietary information. Then, after a FBS, 24 features were selected from the total features set. Features were separated according to their information content.

Features set correspondent to demographic information was contained by 10 features:

- Demographic features

  RIDRETH1:Recode of reported race and Hispanic origin information.

  RIDAGEYR: Age in years of the participant at the time of screening.

  RIAGENDR: Describes the gender of the participant.

  RIDEXMON: Six month time period when the examination was performed.

RIDEXAGM: Age in months of the participant at the time of examination. Reported for persons aged 19 years or younger at the time of examination.

DMDEDUC3: Highest grade or level of school completed or highest degree received.

DMDHHSZA: Number of children aged 5 years or younger in the household (HH).

DMDHRMAR: HH reference person's marital status.

INDFMIN2: Total family income (reported as a range value in dollars).

RISK.GROUP: From 1 to 4, depending on the age rank (0 to 9 years old = group 1, 10 to 19 = group 2, 20 to 59 = group 3, 60 or more = group 4).

As it is possible to observe, four features from the demographic data model are related to age, indicating that according to the patients age, they can be more / less likely to have the dentition problem of caries, or at least have suffered it before and treated by restorations. Also, if a patient is man or woman can represent a difference in the probability of suffering caries, according to the RIAGENDR feature. An interest feature from this model is INDFMIN2, which can represent the patients economic status and it could indicate that the better this status, the better the dentition health, because they've more possibilities to access to corrective or preventive oral health treatments, therefore, patients with lower economic status may be more prone to suffer caries.

Features set correspondent to dietary information was contained by 14 features:

- Dietary features

WTDR2D: Dietary two-day sample weight.

DR1TNUMF: Total number of foods/beverages reported in the individual foods file.

DR1TKCAL: Energy (kcal).

DR1TPROT: Protein (gm).

DR1TSUGR: Total sugars (gm).

DR1TFIBE: Dietary fiber (gm).

DR1TFOLA: Total folate (mcg).

DR1TVC: Vitamin C (mg).

DR1TMAGN: Magnesium (mg).

DR1TTHEO: Theobromine (mg).

DR1TS060: SFA 6:0 (Hexanoic) (gm).

DR1TS080: SFA 8:0 (Octanoic) (gm).

DR1TP205: PFA 20:5 (Eicosapentaenoic) (gm).

DR1TP226: PFA 22:6 (Docosahexaenoic) (gm).

From dietary data model is possible to observe that all features are part of the food balance of patients. The four features, DR1TS060, DR1TS080, DR1TP205, DR1TP226, are part of the fatty acids group, where the first two correspond to saturated fats and the last two correspond to polyunsaturated fats, which are very common elements in the daily diet.

From statistical analysis were obtained the residuals values for each model, contained in Table I, which essentially are the difference between the observed response values and the response values predicted by models, where min represents the minimum value obtained by the model, $1Q$ represents the value of quarter 1 (25%), $Median$ represents the value of quarter 2 (50%), $3Q$ represents the value of quarter 3 (75%) and max represents the maximum value.

TABLE I
RESIDUALS VALUES OF BOTH OF THE DEMOGRAPHIC AND DIETARY DATA MODELS.

Dataset	min	1Q	Median	3Q	Max
Demographic	-1.839	-0.584	-0.097	0.717	1.807
Dietary	-2.273	-0.867	0.024	0.871	1.562

From results in Table I, according to $Q1$ and $Q2$ values from demographic data model, which represent an internal limit of (-3.140, 2.018); and $Q1$ and $Q2$ from dietary data model, which represent an internal limit of (-2.605, 2.609), and according to their respective min and max values, it is possible to observe that they don't exist significant outliers that could be affecting the datasets behavior in the development of the model. Also, according to these five parameters, it is presented a data distribution that looks very similar to normal distribution.

In Figure 2 is presented a graph using 1% of the total demographic data for the obtained residuals between the observed and the predicted values, while in Figure 3 is presented a graph using 1% of the total dietary data for the obtained residuals between the observed and the predicted values. Predicted values are represented as "○" and observed values are represented as "•" in different sizes and color range (between black and red), where the smallest size and black color mean a small residual value between observed and predicted, and the biggest size and red color mean a high residual value.

From both graphs (Figures 2 and 3) is evident that there isn't pattern of behavior in residuals with respect to the dependent variable, which means that these random features are uncorrelated between them.

OR and P-values obtained from demographic data are presented in Table II; while OR and P-values obtained from dietary data are presented in Table III.

TABLE II
OR VALUES FROM DEMOGRAPHIC DATA.

Feature	P-value	OR	2.5%	97.5%
RIAGENDR	<2e-16	1.051	1.017	1.085
RIDAGEYR	0.002	1.010	1.009	1.011
RIDRETH1	<2e-16	0.970	0.958	0.983
RIDEXMON	5.38e-06	1.014	0.981	1.047
RIDEXAGM	<2e-16	1.002	1.002	1.002
DMDEDUC3	0.003	0.991	0.986	0.997
DMDHHSZA	6.36e-10	0.928	0.907	0.907
DMDHRMAR	0.073	0.997	0.994	1.000
INDFMIN2	0.001	1.001	1.000	1.003
RISK.GROUP	9.16e-08	0.878	0.838	0.921

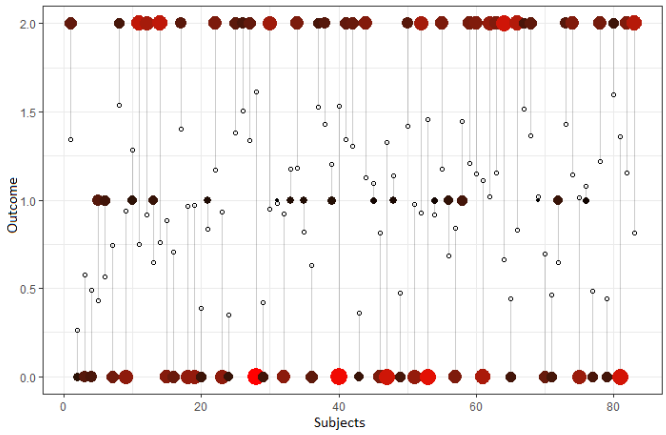

Fig. 2. Representation of residuals from the demographic data model using 1% of the total demographic data.

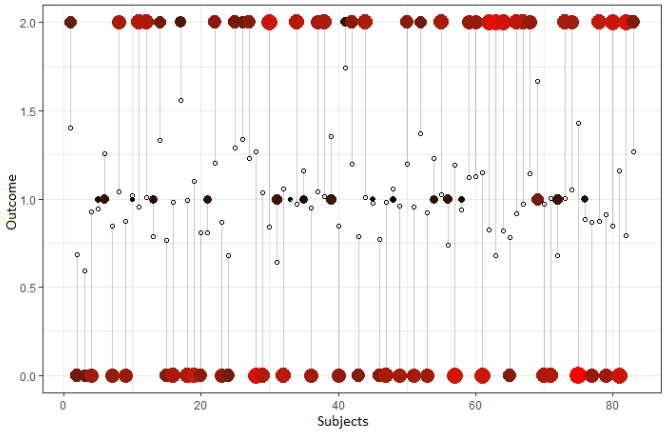

Fig. 3. Representation of residuals from the diet data model using 1% of the total dietary data.

TABLE III
OR VALUES FROM DIETARY DATA.

Feature	P-value	OR	2.5%	97.5%
WTDR2D	<2e-16	1.000	1.000	1.000
DR1TNUMF	8.83e-05	1.006	1.003	1.010
DR1TKCAL	7.88e-08	1.000	1.000	1.000
DR1TPROT	0.101	0.999	0.998	1.000
DR1TSUGR	3.29e-12	0.998	0.998	0.998
DR1TFIBE	3.98e-4	1.005	1.002	1.008
DR1TFOLA	0.444	0.999	0.999	1.000
DR1TVC	0.020	0.999	0.999	0.999
DR1TMAGN	7.75e-08	1.000	1.000	1.000
DR1TTHEO	0.646	1.000	0.999	1.000
DR1TS060	0.004	0.899	0.837	0.966
DR1TS080	<2e-16	0.865	0.841	0.890
DR1TP205	0.386	0.820	0.525	1.283
DR1TP226	0.092	1.240	0.965	1.593

As it is shown in Tables II and III, all features present OR values with reduced confidence intervals, and most of them present a P-value < 0.05, meaning that both models are contained by statistical significant features.

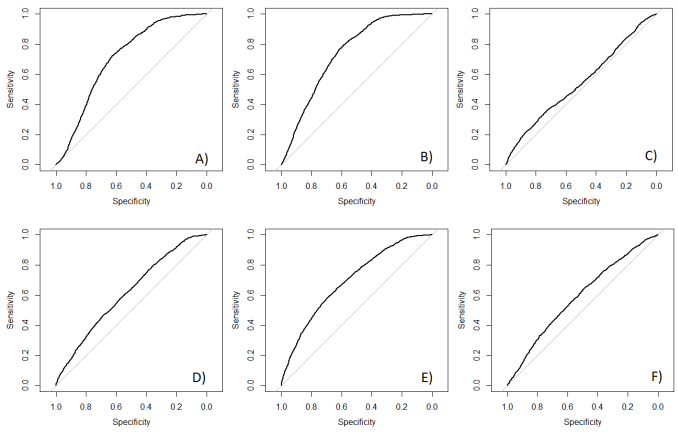

Fig. 4. ROC curves obtained from demographic dataset (A) = 0, B) = 1, C) = 2) and dietary dataset (D) = 0, E) = 1, F) = 2) for each of the outcome data.

Finally, ROC curves obtained by the sensitivity / specificity proportion from demographic and dietary data are presented in Figure 4, where ROC curves from section A), B) and C) correspond to the demographic dataset for the three outcome values 0, 1 and 2, respectively. The three ROC curves together obtained an AUC mean of 0.664.

ROC curves from section D), E) and F) correspond to the dietary dataset for the three outcome values 0, 1, and 2, respectively. The three ROC curves together obtained an AUC mean of 0.633.

ROC curves developed by the demographic data model (Figure 4 A), B) and C)) present a significant sensitivity / specificity relationship, which means that most of the patients classified for the three outcomes, by this model, were true positives and true negatives.

On the other hand, ROC curves developed by the dietary data model (Figure 4 D), E) and F)), also present a significant sensitivity / specificity relationship, which means that most of the patients were correctly classified.

V. CONCLUSIONS AND FUTURE WORK

In most of the worldwide population, health problems that are mainly taken into account are those that represent a risk of death or permanent disability. Oral health problems do not arouse the spontaneous interest of the community, being one of the main reasons why it is difficult to prevent them, due to this, it is an important task to identify by different analysis the determinants of oral health problems, informing population in a correct and accurate way the care they must have to make easier the improvement of their oral health.

For this research were used demographic and dietary features that presented significant properties for the patients classification according to their dentition health.

From the 24 features model obtained by a FBS research, were developed two multivariate models; one model was contained by the 10 features that belonged to demographic

data, being most of them related to the age of patients. The second model was contained by the 14 features that belonged to dietary data, which most of them were related to the food balance of patients.

Results obtained from statistical analysis applied to each model in Table I and Figures 2, 3, allow to conclude that the predicted data by both models has an appropriate behavior, very similar to a normal distribution, besides, minimum and maximum values are within the internal limits, which means that isn't present any problem related to outliers that could be affecting the distribution of their data.

P-values and OR values with their respective confidence intervals shown in Tables II and III, for both models, that all features were statistical significant, according to their P-values lower than 0.05 in most cases and their small OR confidence intervals. Finally, in Figure 4 is demonstrated that both models are capable to patients classification for each condition, because, according to the sensitivity / specificity ratio of all curves together, were obtained AUC values $>$ 0.6, which means they're statistical significant and making possible to develop two emergent models based on these 24 specific demographic and dietary features to diagnose the oral health status of patients based on their probability to develop caries, reducing the diagnostic survey formed by the initial 189 features.

For future work it is proposed a deeper analysis based on the risk group (age) where the subjects belong, looking for features that may depend on it to become subjects more vulnerable to the caries condition, developing a specific model for each risk group in order to improve the accuracy in the classification of caries / restoration / control patients.

REFERENCES

[1] Y. Romero, D. Carrillo, N. Espinoza, and N. A. Díaz, "Perfil epidemiológico en salud bucal de la población escolarizada del municipio campo elías del estado mérida," *Acta Bioclínica*, vol. 6, no. 11, pp. 3–24, 2016.

[2] M. A. T. Tamez, L. T. de Mendoza, E. G. R. Peña, and P. C. C. Martínez, "Salud bucodental en escolares de estrato social bajo," *RESPYN*, vol. 6, no. 2, 2017.

[3] P. Srisilapanan, N. Korwanich, S. Jienmaneechotchai, S. Dalodom, N. Veerachai, W. Vejvitee, and J. Roseman, "Estimate of impact on the oral health-related quality of life of older thai people by the provision of dentures through the royal project," *International journal of dentistry*, vol. 2016, 2016.

[4] P. E. Petersen and H. Ogawa, "Prevention of dental caries through the use of fluoride–the who approach," *Community dental health*, vol. 33, no. 2, pp. 66–8, 2016.

[5] T. P. Escalona, H. R. C. Ortiz, Y. P. Palomino, M. I. Tamayo, and M. I. R. Rodríguez, "08-relación entre factores de riesgos y caries dental relationship between risk factors and dental caries," *MULTIMED Revista Médica Granma*, vol. 19, no. 4, 2017.

[6] S. Díaz-Cárdenas, M. A. Meisser-Vidal, L. R. Tirado-Amador, N. Fortich-Mesa, L. Tapias-Torrado, and F. D. González-Martínez, "Impacto de salud oral sobre calidad de vida en adultos jóvenes de clínicas odontológicas universitarias," *International journal of odontostomatology*, vol. 11, no. 1, pp. 5–11, 2017.

[7] C. Espinoza, D. Fiorella, P. M. Ramos Mejía *et al.*, "Asociación del índice de masa corporal con la presencia de caries dental en escolares de 6 a 12 años," 2017.

[8] S. FALEIROS, A. ORMEÑO, M. PINTO, R. TAPIA, C. DÍAZ, H. GARCÍA *et al.*, "Prevalencia de caries en alumnos de educación básica y su asociación con el estado nutricional," *Revista chilena de pediatría*, vol. 81, no. 1, pp. 28–36, 2010.

[9] A. B. Hernández Cadena, "Relación de la caries dental con índice de masa corporal de niños 5-12 años de edad de las comunidades rurales de la parroquia cangahua, ecuador," B.S. thesis, Quito: Universidad de las Américas, 2017., 2017.

[10] R. V. Sarmiento, F. P. Barrionuevo, Y. S. Huamán, and M. C. Loyola, "Prevalencia de caries de infancia temprana en niños menores de 6 años de edad, residentes en poblados urbano marginales de lima norte," *Revista Estomatológica Herediana*, vol. 21, no. 2, pp. 79–86, 2011.

[11] C. Oropeza-Oropeza, N. Molina-Frechero, E. Castañeda-Castaneira, C. Zaragoza-Rosado, and C. Cruz Leyva, "Caries dental en primeros molares permanentes de escolares de la delegación tláhuac." *Revista ADM*, vol. 69, no. 2, 2012.

[12] B. J. Cardozo, M. M. Gonzalez, S. R. Pérez, P. A. Vaculik, and E. G. Sanz, "Epidemiología de la caries dental en niños del jardín de infantes pinocho de la ciudad de corrientes," *Revista de la Facultad de Odontología*, vol. 9, no. 1, pp. 35–41, 2017.

[13] C. U. Lam, L. Khin, A. Kalhan, R. Yee, Y. Lee, M.-F. Chong, K. Kwek, S. Saw, K. Godfrey, Y. Chong *et al.*, "Identification of caries risk determinants in toddlers: Results of the gusto birth cohort study," *Caries Research*, vol. 51, no. 4, pp. 271–282, 2017.

[14] I. Fernandes, A. Sá-Pinto, L. S. Marques, J. Ramos-Jorge, and M. Ramos-Jorge, "Maternal identification of dental caries lesions in their children aged 1–3 years," *European Archives of Paediatric Dentistry*, pp. 1–6, 2017.

[15] A. LIPS, L. S. ANTUNES, L. A. ANTUNES, A. V. B. PINTOR, D. A. B. d. SANTOS, R. BACHINSKI, E. C. KÜCHLER, and G. G. ALVES, "Salivary protein polymorphisms and risk of dental caries: a systematic review," *Brazilian Oral Research*, vol. 31, 2017.

[16] A. Ahmed, K. Ikram, H. Masood, and M. Urooj, "Identification of relationship between oral disorders & hemodynamic parameters." *Pakistan Oral & Dental Journal*, vol. 37, no. 2, 2017.

[17] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[18] V. Berlanga Silvente and R. Vilà Baños, "Cómo obtener un modelo de regresión logística binaria con spss," *REIRE. Revista d'Innovació i Recerca en Educació, 2014, vol. 7, num. 2*, 2014.

[19] J. Lawless and K. Singhal, "Efficient screening of nonnormal regression models," *Biometrics*, pp. 318–327, 1978.

[20] Y. Nguyen, D. Nettleton, H. Liu, and C. K. Tuggle, "Detecting differentially expressed genes with rna-seq data using backward selection to account for the effects of relevant covariates," *Journal of agricultural, biological, and environmental statistics*, vol. 20, no. 4, pp. 577–597, 2015.

[21] J. Martínez Abreu, J. Capote Femenias, G. Bermúdez Ferrer, and Y. Martínez García, "Determinantes sociales del estado de salud oral en el contexto actual," *MediSur*, vol. 12, no. 4, pp. 562–569, 2014.

[22] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2015.

[23] M. Szumilas, "Explaining odds ratios," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 19, no. 3, p. 227, 2010.

[24] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "Auc: a misleading measure of the performance of predictive distribution models," *Global ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.

[25] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[26] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "proc: an open-source package for r and s+ to analyze and compare roc curves," *BMC bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[27] F. E. Harrell Jr, *rms: Regression Modeling Strategies*, 2017, r package version 5.1-1. [Online]. Available: https://CRAN.R-project.org/package=rms

[28] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[29] J. Maindonald and W. J. Braun, "Daag: Data analysis and graphics data and functions. r package version 1.20," 2014.

[30] P. Cortez, "Data mining with neural networks and support vector machines using the r/rminer tool," *Advances in data mining. Applications and theoretical aspects*, pp. 572–583, 2010.
[31] C. for Disease Control, Prevention *et al.*, "National health and nutrition examination survey, 2011-2012," 2013.

An analysis of demographic and dietary data with an oral health approach, a preliminary study using genetic algorithms

Laura A. Zanella-Calzada[1][0000−0002−8049−8077], Carlos E. Galvn-Tejada[1][0000−0002−7635−4687], Nubia M. Chvez-Lamas[2], Mara Del Carmen Gracia-Cortez[2], and Jorge I. Galvn-Tejada[1]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, México, {lzanellac, ericgalvan, gatejo}@uaz.edu.mx

[2] Unidad Académica de Odontología, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, México, {nubiachavez, gacc005340}@uaz.edu.mx

Abstract. Oral health is one of the main components in the quality of life of people, since it is a determining factor in general health that affects the risk of suffering from other conditions, such as chronic diseases. Dental caries is the condition that most affects oral health worldwide, and occurs in about 90 % of people. The high prevalence in dental caries is caused by the diverse elements that interact simultaneously in their favor, such as the nutritional and socioeconomic elements. Based on this problem, this study proposes the analysis of a series of dietetic and demographic features compiled by the National Health and Nutrition Examination Survey 2013 – 2014, in order to obtain a model that allows the automatic classification of subjects according to their oral health status, presence or absence / restorations of caries. The methodology was carried out in three steps, starting with a data preprocessing, followed by a feature selection using a genetic algorithm and a final validation through a statistical analysis. The developed model is made up of five features (of the initial 188), and reached an area under the curve of 0.748, which is a statistically significant value. According to the results obtained, it was possible to conclude that the proposed model presents a preliminary result for the development of a low-cost support tool that helps the diagnosis of dental caries and the reduction of this condition.

Keywords: Oral health · Dental caries · Classification · Feature selection · Genetic algorithm · Statistical analysis.

1 Introduction

Chronic diseases are one of the main problems in public health and, nowadays, the pattern of diseases has changed, turning oral diseases into an important public health problem throughout the world. Oral diseases have a high incidence

and prevalence in all regions of the world, especially in disadvantaged and socially marginalized populations. According to the World Health Organization (WHO), the treatment of various oral conditions is extremely expensive and is not feasible in most low and middle income countries, being the fourth most expensive cause to treat [1]. Oral health presents an essential component in the quality of life due to its great influence on general health, since this condition can increase the risk of chronic diseases, such as: cardiovascular and cerebrovascular, diabetes mellitus and respiratory [2].

The most common condition in oral health is dental caries, which according to the WHO, affects between 60 and 90 % of children between five and 17 years. There is a series of determinants that favor this condition, such as carbohydrate consumption, food characteristics, plaque removal, among others, that interact simultaneously with variables that correspond to different orders of biological processes, complex historical-cultural structures, social relations, socioeconomic level, educational level, among others [3, 4].

Due to the difficulty of controlling the incidence of caries, which is caused by the large number of factors that influence, recent studies have implemented algorithms and performed analyzes based on computer-aided diagnosis (CADx) to develop prediction and classification models for preventive diagnosis and reduction of dental caries prevalence, looking for the main factors that affect this condition [5].

In this study is presented an analysis of two types of determinants that affect the important condition in oral health, dental caries. These determinants are demographic and dietary features that have been collected by the National Health and Nutrition Examination Survey (NHANES), 2013 – 2014. With a feature selection through a genetic algorithm, a multivariate model es developed that presents significant results in the classification of subjects with presence of this condition of subjects with absence, validated under a statistical analysis.

This work is organized as follows; in section 2 is presented the methodology followed in three main steps, data preprocessing, feature selection and validation. In section 3, the results obtained are shown and in section 4 these results are discussed. Finally, conclusions are briefly described in section 5.

2 Methodology

The methodology of this work is presented in Figure 1, which is a flowchart of the stages that were followed for the development of this study.

A data preprocessing step (A) was performed to avoid any problem related to missing data or outliers that could affect the later stages. Then, in (B) a feature selection step is presented, which was carried out using the genetic algorithm "Galgo" to find the features that present the most significant behavior for the correct and automated classification of the subjects. Finally, (C) presents a validation step, where the features that were selected were subjected to a logistic regression, obtaining a general model for the classification of subjects that allows to know the true positive and true negative rates through the calculation of

the receiver operating characteristic (ROC) curve and the area under the curve (AUC).

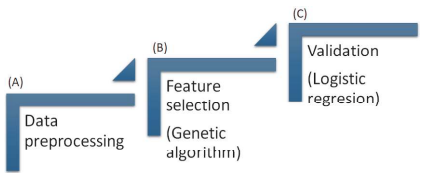

Fig. 1. Flowchart of the methodology followed.

All the methodology was performed using "R" (version 3.4.4) [6], which is a free software environment for statistical computing and graphics.

2.1 Data description

NHANES is a program of studies designed to assess the health and nutritional status of children and adults in the United States of America (USA), was founded by the Centers of Disease Control and Prevention (CDC) and the National Center for Health Statistics (NCHS). The surveys conducted by this program are unique, since it combines interviews and physical examinations [7]. NHANES collects information from different type of data and, in turn, this information is included in six main contexts; demographic, dietary, examination, laboratory, questionnaire and limited access. For this work, the demographic and dietary datasets were used;

- Demographic: it provides individual, family and household level information in different topics (income of households and families, size of households and families, pregnancy status, among others).
- Dietary: it provides detailed information on dietary intake, in order to estimate the types and amount of food and beverages consumed, in addition to estimating the intake of energy, nutrients, and other food components.

These datasets were contained by 188 features, and they were used as input features; while the condition of dental caries (absence or presence / restorations) was used as output feature. The subjects that were contained in those datasets belong to different counties in the USA and they were randomly selected with a computer algorithm by NHANES. The total number of subjects was 9812 (3690 controls / 6122 cases), 4982 females and 4830 males, and they were in a range age between zero and 89 years old.

2.2 Data preprocessing

The two main purposes of the data preprocessing stage were to avoid problems of missing data and outliers.

Initially, all incomplete cases were manually removed, eliminating those subjects that presented ≥ 70 % of missing data. Of the rest of the subjects, all the missing values were imputed through the "nearestneighborimpute" function of the package "FRESA.CAD" (version 3.0.1) [8]. This function searches for any "NA" (not available) that is present in the dataset and looks for the row of complete observations that have the closest interquartile range normalized Manhattan distance to the rows that present missing values. In case that more than one row have similar minimum distances, the median value is used.

Then, the data was normalized using Z normalization, forcing the data to have a standard normal distribution. It was calculated with Equation 1, where x_i refers to the data, μ to the mean value and σ to the standard deviation.

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Finally, the singular values were removed, eliminating those columns that presented multiple values among themselves or the same value along the entire rows. This step was performed to avoid redundant or non-significant information.

2.3 Feature selection

The feature selection was carried out through "Galgo" (version 1.1) [9], an R package for the selection of multivariate variable using genetic algorithms. In addition, Galgo presents a series of functions for the analysis of the population contained in the models and for the reconstruction and characterization of the models.

The Galgo procedure begins with a set of subsets of features or genes (named chromosomes) contained by data that were randomly selected. Then, each chromosome is evaluated based on its ability to predict a dependent variable, measuring its level of accuracy. The general idea is to replace the initial set of data with new data that conforms variants of chromosomes with a higher classification accuracy and repeat this process until achieving a desired level of accuracy. The progressive improvement of the chromosomes is carried out through a series of steps that resemble the process of natural selection, which consists of three steps; selection, mutation and crossover.

Looking for the increment of the solution space that is explored, independent chromosomes can evolve in isolated environments (named niches). Chromosomes can occasionally migrate between niches ensuring that that good solutions can recombine.

The classification methods that are included in Galgo are "k-nearest-neighbors", "nearest centroid", "support vector machines", "neural networks", "classification trees" and "discriminant functions" [10].

The main four steps that Galgo follows are briefly described.

1. **Setting-up the analysis.** In this step is necessary to specify the input and the output features, the statistical model, the desired accuracy, the error estimation scheme and the parameters that decide the search environment of the genetic algorithm. It is important to mention that the estimation of the error can be defined in two levels; a validation strategy for training / testing, and within the training process using k-fold cross-validation, random splits or re-substitution.
2. **Searching for relevant multivariate models.** Starting with a random set of chromosomes, the genetic algorithm procedure will look for a diverse collection of statistically significant local solutions. A sufficiently large number of chromosomes must be selected to have a good representation of the solution space.
3. **Refinement and analysis of the set of selected chromosomes.** The chromosomes that were selected have a length that was previously defined. Although the models have reached the desired classification accuracy, a refinement step is carried out, since there is a possibility that not all the model genes have a significant contribution to accuracy. Through a backward selection strategy, only those genes that effectively contribute to the classification task are selected.
4. **Development of a representative statistical model.** Finally, in this part of the analysis is obtained a single representative model of the selected chromosomes. For this step, a forward selection strategy was implemented based on the step-wise inclusion of the most frequent genes in the chromosome.

For this work, the settings used in this experimentation were the following: chromosomes composed of five genes, with 200 evolutionary processes in 200 big bangs, these settings were selected given the number of features and patients to avoid overfitting.

On the other hand, the classification model used was the nearest centroid, which is a supervised method that assigns to the observations the class label of the training samples whose mean or centroid is the closest to the observations. For each set of data points belonging to the same class, the centroid vectors are calculated, if there are k classes in the training set, there are k centroid vectors. The test samples are classified in the class with the nearest centroid.

For the training procedure, given the labeled training samples $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$, with class labels $y_i \in Y$, the class centroids $\boldsymbol{\mu}_k$ are calculated with Equation 2, where C_k is the set of indices of samples belonging to class $k \in Y$.

$$\boldsymbol{\mu_k} = \frac{1}{|C_k|} \sum_{l \in C_k \boldsymbol{x}_i} \tag{2}$$

The prediction function $\hat{y}$, which is the class assigned to an observation $\boldsymbol{x}$ is calculated with Equation 3

$$\hat{y} = min_{k \in Y} ||\boldsymbol{\mu}_k - \boldsymbol{x}|| \tag{3}$$

2.4 Validation

The validation stage was carried out in order to evaluate the multivariate model resulting from the feature selection. Initially, the model was subjected to a logistic regression (LR) to obtain a general model for the classification of subjects using the set of selected features. LR is an analysis that consists of a statistical technique to model the relationship between the features. This method belongs to the statistical methods where the contribution of different factors in the occurrence of a simple event is measured. The main objective of LR is to model the influence of the probability of an event. The simplest representation of a model obtained by this method is presented in Equation 4, where y from $logit(y = 1)$ is the dependent variable or the outcome feature that it's necessary to be subjected to a logarithmic transformation ($logit$) because the initial equation of the model is of exponential type and by this transformation it's possible to use it as a lineal function, w is an offset term that can be included, β_1 is the slope and x is the independent variable or the analyzed feature. Models can be composed by the number of independent variables needed [12].

$$logit(y = 1) = w + \beta_1 x \tag{4}$$

On the other hand, the AUC value is a standard method used for the evaluation of the accuracy obtained by the model, and is calculated with the relationship between the specificity and the sensitivity [13].

The sensitivity parameter is referred to the proportion of data that belongs to a condition and it's classified as positive. This value is obtained with Equation 5, where TP represents the quantity of true positives and FP represents the quantity of false positives.

$$PPV = \frac{TP}{TP + FP} \tag{5}$$

The specificity parameter is referred to the proportion of subjects without a condition that are classified as negative. This value is obtained with Equation 6, where TN represents the quantity of true negatives and FN represents the quantity of false negatives.

$$NPV = \frac{TN}{TN + FN} \tag{6}$$

3 Results

The results obtained from the methodology performed are presented in this section.

From the preprocessing step, a series of features were eliminated according to the missing data and singular values that were present in the dataset, maintaining a total of 136 dietary and demographic features. Then, those features were submitted to the Galgo algorithm for a feature selection.

Figure 2 presents a graph where all the features are ordered according to their degree of appearance in the chromosomes throughout the iterations. The features shown in black have the highest frequency, while the features shown in gray have the lowest.

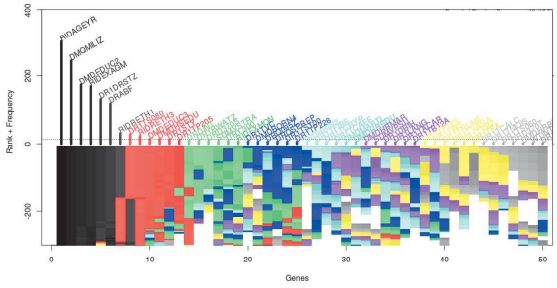

Fig. 2. Graph of the frequency with which the features appear on chromosomes, ordered descending. Vertical axis presents the frequency number and horizontal axis presents the chromosomes.

Figure 3 presents the performance graph of forward selection, showing the classification accuracy achieved for each feature included in the model. The features are arranged according to their frequency of appearance, from the highest to the lowest.

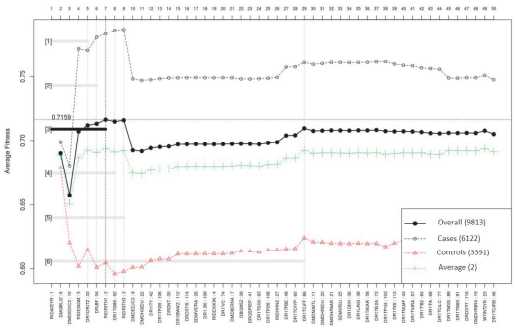

Fig. 3. Graph of the forward selection step. Vertical axis presents the accuracy value and horizontal axis presents the chromosomes ordered according with their frequency appearance.

After the forward selection step, the backward elimination process was performed, obtaining a final model that presented the best accuracy without any redundant information. This model was contained by the five features that are described in Table 1.

Table 1. Features obtained from the backward elimination step.

Feature	Description
RIDAGEYR	Age in years, at the time of the screening interview, is reported for survey participants between the ages of 1 and 79 years of age.
RIDEXAGM	Age in months of the participant at the time of examination.
DMQMILIZ	Have you ever served on active duty in the U.S. Armed Forces, Military reserves, or National Guard?
DMDEDUC2	What is the highest grade or level of school you have completed or the highest degree you have received?
DR1DRSTZ	Dietary recall status.

Finally, after the feature selection, it was carried out the validation step, where the ROC curve presented in Figure 4 was obtained, having an AUC of 0.748.

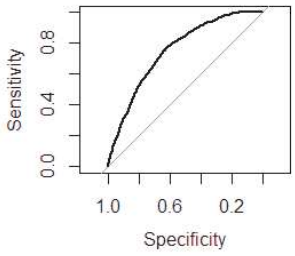

Fig. 4. ROC curve obtained from the classification of subjects.

4 Discussion

The results obtained show a multivariate model, contained by demographic and dietary features, which was obtained through the Galgo genetic algorithm, which presents a statistically significant performance in the classification of subjects who show absence of dental caries from those who have presence or restorations of they.

Figure 2 presents the graph of the frequency that each feature obtained according to the number of times they appeared in the different chromosomes that developed throughout the Galgo process, where it is possible to observe that from the total features, there where seven with a significant number of frequency, which are those presented in black; RIDAGEYR, DMQMILIZ, DMDEDUC2, RIDEXAGM, DR1DRSTZ, DRABF and RIDRETH1. Then, in Figure 3 the graph of the forward selection step is observed, where the accuracy obtained for each feature that is included in the model is presented; when the model is contained by the seven most frequent features, the best accuracy is achieved. In addition, it is possible to observe that the classification of the cases (-○-) is much more accurate than the classification of the controls (-△-), making that the overall (-●-) classification reduces its accuracy. This may occur due to the low significant information that the features may be presenting for the control subjects or to the number of controls presented in the data set, which can be resolved by increasing the number of subjects and / or features. Then, Table 1 includes the five features that were selected through the backward elimination step, eliminating two of the features that were selected in the previous step; DRABF and RIDRETH1. These features were removed because they didn't present any significant information for the classification of subjects or their contribution was redundant.

The information presented in the multivariate model is related to the age of the subjects, the educational level, the state of the diet and the duty of activity in USA services. As mentioned above, one of the main factors that affect the state of oral health is age, with young people being the most affected by dental caries. On the other hand, the state of dietary recall is used to know the individual foods and the total intake of nutrients, which indicates the quality and integrity of the subjects, being one of the determinants of dental caries that was also mentioned above.

The educational level and the participation in the services of the USA are features related to the socioeconomic level. The educational level is significant because the low income regions have less educational opportunities; while participation in the services of the USA is an important feature taking into account that the main objective population of NHANES is the non-institutionalized civilian resident population of the USA, which may be people who decided to emigrate in search of better economic opportunities and one of the requirements for the Armed Forces of the USA, the Military Reserves and the National Guard is being a citizen or a permanent resident. Therefore, it is possible to justify the selection of those features for the developed model.

Finally, Figure 4 presents the ROC curve obtained according to the true positives and true negatives calculated with the developed model; its AUC value was 0.748, which is statistically significant since it means that from the total subjects, 74.8 % were correctly classified.

5 Conclusions

This paper presents the analysis of a series of demographic and dietary features, in order to develop a model that can present a tool for specialists in the preventive diagnosis of dental caries and the reduction of the incidence of this condition. This analysis was performed in three main stages, data preprocessing, feature selection and validation. In the stage of feature selection, a multivariate model was developed that contained the features that provided the most significant information in the classification of control and case subjects, while the validation stage allowed the evaluation of this multivariate model, obtaining statistically significant results.

Therefore, the model proposed in this work may represent a low-cost preliminary tool that helps in the diagnosis of dental caries and in the possible prediction of them, both for high and low income regions, to reduce their high incidence and avoid the high cost treatments that this condition represents.

References

1. World Health Organization (NCHS): Oral Health, http://www.who.int/oral-health/disease-burden/global/en/. Last accessed 05 June 2018
2. Ridao Marín, D.: Desarrollo de un Sistema de Ayuda a la Decisión para Tratamientos Odontológicos con Imágenes Digitales. Universidad de Málaga: Málaga, Spain, 10–12 (2017)
3. Espinoza Solano, M.; León-Manco, R.A.: Prevalencia y experiencia de caries dental en estudiantes según facultades de una universidad particular peruana. Rev. Estomatol. Hered. **25**(3), 187–193 (2015)
4. Acuña Aguilar, L.D.; Porras Cerón, D.; Ríos Rueda, L.D.: Prevalencia de Lesiones Cariosas y Factores Asociados Presentes en Pacientes con SNdrome de Down en las Fundaciones Fundown y san Luis Guanella de Bucaramanga. Universidad Santo Toms: Bucaramanga, Colombia, 12–16 (2017)
5. Gispert Abreu, E.D.L.Á.; Castell-Florit Serrate, P.; Herrera Nordet, M.: Salud bucal poblacional y su produccin intersectorial. Rev. Cubana Estomatol.**52**, 62–67 (2015)
6. R: A Language and Environment for Statistical Computing, https://www.R-project.org/. Last accessed 17 June 2018
7. National Health and Nutrition Examination Survey Data. 2013–2014, http://www.cdc. gov/nchs/nhanes.htm. Last accessed 17 June 2018
8. FRESA.CAD: Feature Selection Algorithms for Computer Aided Diagnosis, https://CRAN.R-project.org/package=FRESA.CAD. Last accessed 17 June 2018
9. GALGO: an R package for multivariate variable selection using genetic algorithms, http://bioinformatica.mty.itesm.mx/galgo2. Last accessed 17 June 2018
10. Trevino, V., and Falciani, F.: GALGO: an R package for multivariate variable selection using genetic algorithms. Bioinformatics **22**(9), 1154–1156 (2006)
11. Das, T.: Machine Learning algorithms for Image Classification of hand digits and face recognition dataset. Machine Learning **4**(12), 640–649 (2017)
12. Montgomery, Douglas C and Peck, Elizabeth A and Vining, G Geoffrey: Introduction to linear regression analysis. John Wiley & Sons, Location (2015)
13. Berlanga Silvente, Vanesa and Vilà Baños, AUC: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography **17**(2), 145–151 (2008)

International Journal of
Environmental Research and Public Health

Article

A Case–Control Study of Socio-Economic and Nutritional Characteristics as Determinants of Dental Caries in Different Age Groups, Considered as Public Health Problem: Data from NHANES 2013–2014

Laura A. Zanella-Calzada [1,†], Carlos E. Galván-Tejada [1,*,†], Nubia M. Chávez-Lamas [2], Ma. del Carmen Gracia-Cortés [2], Arturo Moreno-Báez [1], Jose G. Arceo-Olague [1], Jose M. Celaya-Padilla [3], Jorge I. Galván-Tejada [1] and Hamurabi Gamboa-Rosales [1]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, Mexico; lzanellac@uaz.edu.mx (L.A.Z.-C.); morenob20@uaz.edu.mx (A.M.-B.); arceojg@uaz.edu.mx (J.G.A.-O.); gatejo@uaz.edu.mx (J.I.G.-T.); hamurabigr@uaz.edu.mx (H.G.-R.)

[2] Clínica Comunitaria de Tacoaleche, Unidad Académica de Odontología, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, Mexico; nubiachavez@uaz.edu.mx (N.M.C.-L.); gacc005340@uaz.edu.mx (M.d.C.G.-C.)

[3] CONACYT, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, Mexico; jose.celaya@uaz.edu.mx

* Correspondence: ericgalvan@uaz.edu.mx; Tel.: +52-492-544-0968

† These authors contributed equally to this work.

Received: 14 April 2018; Accepted: 26 April 2018; Published: 10 May 2018

Abstract: One of the principal conditions that affects oral health worldwide is dental caries, occurring in about 90% of the global population. This pathology has been considered a challenge because of its high prevalence, besides being a chronic but preventable disease which can be caused by a series of different demographic, dietary, among others. Based on this problem, in this research a demographic and dietary features analysis is performed for the classification of subjects according to their oral health status based on caries, according to the age group where the population belongs, using as feature selector a technique based on fast backward selection (FBS) approach for the development of three predictive models, one for each age range (group 1: 10–19; group 2: 20–59; group 3: 60 or more years old). As validation, a net reclassification improvement (NRI), AUC, ROC, and OR values are used to evaluate their classification accuracy. We analyzed 189 demographic and dietary features from National Health and Nutrition Examination Survey (NHANES) 2013–2014. Each model obtained statistically significant results for most features and narrow OR confidence intervals. Age group 2 obtained a mean NRI = −0.080 and AUC = 0.933; age group 3 obtained a mean NRI = −0.024 and AUC = 0.787; and age group 4 obtained a mean NRI = −0.129 and AUC = 0.735. Based on these results, it is concluded that these specific demographic and dietary features are significant determinants for estimating the oral health status in patients based on their likelihood of developing caries, and the age group could imply different risk factors for subjects.

Keywords: NHANES; dental caries; fast backward variable selection; net reclassification improvement; computer-aided diagnosis; statistical analysis; predictive multivariate models

1. Introduction

Health is a condition that presents difficulties in its description due to its different definitions. According to the World Health Organization (WHO), health can be defined as a physical, mental, and social healthy status and not only as the absence of diseases. Quality of life is included, defined as

an individual perception of life position, in the context of the cultural environment and the individual's relationships with the objectives, expectations, standards, and concerns in life. Due to this concept, there are some factors that affect the health of subjects. These factors can be evaluated through three different dimensions: general (lifestyle), particular (life conditions), and singular (type of life) [1,2].

Oral health is considered to be an integral and essential part of health, and can compromise the quality of life of all groups of people, including infants, adolescents, young people, adults, and older adults without distinguishing between age; sex; social level; culture; economic, educational, and marital status; previous caries history; current caries index; levels of microbiological factors; drawing of family caries; and the environment in which they are found [1–3].

These characteristics have supported and facilitated the recognition of different determinants that contribute to negatively modifying oral health status, since its appearance depends on the conjugation of biological factors such as dental anatomy, diet consumption, and the appearance of oral diseases, where the relationship between dietary and demographic features and the appearance of dental caries is widely demonstrated [4–9]. In oral public health, dental caries are a challenge because it is a condition that represents a chronic and preventable disease, affecting around 90% of the global population [8], becoming an oral health problem. In the early stages of life, teeth usually do not present problems related to this disease, but throughout life, dental caries can be developed by different factors such as those mentioned above, being the main reason for the attention of such conditions [9].

Due to the difficulty of controlling the incidence of caries, some studies have been developed that have implemented algorithms and performed analysis based on computer-aided diagnosis (CADx) in order to provide an automated tool for the detection of caries based on different types of features, with the aim of obtaining a better prognosis for patients and identifying the risk factors that can contribute to their prevention. One of the main CADx tools that has been used for the detection of diseases is the net reclassification index (NRI), which is a measure for evaluating the improvement in the prediction performance of the disease gained by adding a marker to baseline predictors. Other statistical parameters that are commonly used in clinical research include the receiver operating characteristic (ROC) curve, the area under the curve (AUC), the benefit function, and the Brier score, among others [10].

Based on this general oral health problem description, the main contribution of this paper is to analyze the relationship between demographic and dietary features and the dental caries status, in order to develop a multivariate model (based on CADx) for three different age groups for the classification of subjects according to their dental caries status (presence/absence), in addition to looking for the different features that affect the age groups. These multivariate models were developed through a statistical approach using fast backward selection (FBS) for the feature selection, NRI for the validation, in order to know the reclassification proportion of subjects, and the ROC curve, which allows the sensitivity (true positives) and specificity (true negatives) of the model's classification to be obtained.

According to the contributions mentioned above, the hypothesis of this work makes reference to the possibility of developing a multivariate model through statistical analysis—based on demographic and dietary features that were provided from National Health and Nutrition Examination Survey (NHANES) 2013–2014—that is able to automatically classify between patients with the presence and absence of caries, in order to find a tool that provides information about the risk factors that make subjects vulnerable to dental caries, according to their age range.

Related Work

There are several approaches to identifying caries risk determinants. In the work of Lam C. et al. [11], it was found that dental visits, brushing frequency, lower parental perceived importance of baby teeth, and weaning onto solids are determinants associated with plaque accumulation; however, they only validated these factors in children. On the other hand, I. B. Fernandes et al. [12] presented a cross-sectional study of 274 children and their mothers based on demographic/socio-economic status; they used a Poisson

regression to perform the analysis, finding that dental caries can be mainly found in mothers of children aged 1, which demonstrates a relationship between demographic/socio-economic status and caries with children. Other types of risk factors are associated with genetic and dietary data, like in the work of A. Lips et al. [13], where the association between genetic polymorphisms and risk of dental caries was demonstrated for most of the salivary proteins. Diabetic and hypertensive patients have dietary risk factors that can lead to oral health problems, demonstrated by Asif Ahmed et al. [14] through a statistical analysis using clinical data, aiming to identify that patients with oro-dental problems were hemodynamically stressed.

This paper is organized as follows. Section 2 provides a description of the data set used for this research, the statistical methods performed, as well as the experimentation conditions. The results obtained from the statistical analysis are presented in Section 3. Discussion and conclusions are described in Section 4, and finally, future work is briefly mentioned in Section 5.

2. Materials and Methods

The development of this work was performed using the data from the National Health and Nutrition Examination Survey (NHANES, 2013–2014). These data are described in this section, as well as the patient selection, data preprocessing, and methods for the acquisition of models.

2.1. Study Design

The study design of this work is presented in Figure 1, beginning with the data acquisition from NHANES 2013–2014 in (A), selecting three different types of features from the public datasets: dietary, demographic, and examination. A brief data preprocessing is carried out in (B), applying different techniques to solve the missing data problem and to remove any singular value presented in features, and to separate the data in three different datasets, according to the age range of the participants. Then, the feature selection is presented in (C), which is performed using the statistical technique FBS, obtaining three multivariate models (one for each dataset) containing the most statistically significant features. Finally, a validation process for each multivariate model was performed using the NRI technique in addition to the statistical parameters AUC, ROC curve, and OR, in order to evaluate the model's accuracy.

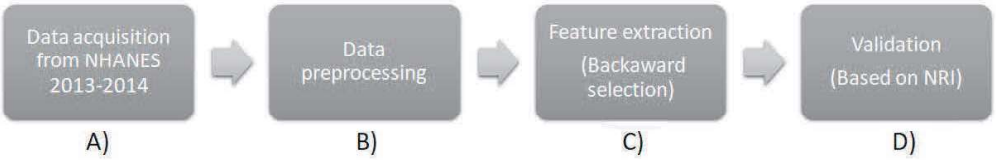

Figure 1. Flowchart of the steps followed in methodology. NHANES: National Health and Nutrition Examination Survey; NRI: net reclassification index.

2.2. Setting

For this work, the public data of the NHANES program were used. NHANES collects a repository of information from US participants based on a series of different questionnaires in order to obtain data for health status knowledge. The data used were from the 2013–2014 period, and from three different types of questionnaires: dietary, demographic, and examination.

2.3. Dataset Description

NHANES is a federal agency that produces data and materials for the public domain from the health and nutritional status of adults and children in the United States, including ethnic groups. The survey combines interviews and physical examinations, allowing the development of studies using the clinical, para-clinical, and demographic information of subjects. NHANES is an initiative

that was founded by the Centers for Disease Control and Prevention (CDC) and the National Center for Health Statistics (NCHS) [15].

The main objectives of NHANES are to estimate the number and percentage of persons in the US population with selected diseases and risk factors; to monitor trends in the prevalence, awareness, treatment, and control of selected diseases; to monitor trends in risk behaviors and environmental exposures; to study the relationship between diet, nutrition, and health; to explore emerging public health issues and new technologies; and to provide baseline health characteristics that can be linked to mortality data from the National Death Index or other administrative records.

NHANES data include a series of different types of characteristics, which are called features in this work; among them are included demographic data (individual, family, and household level information), dietary data (total nutrient intake), and health-related information (public health significance in areas of surveillance, prevention, treatment, dental care utilization, health policy, and evaluation of Federal health programs). A health-related component that is critical to this study is the oral health examination, which is carried out by trained medical personnel and is related to the presence or absence of dental caries.

2.3.1. Participants

The content of these features was obtained from subjects that were submitted to a series of different questionnaires, related to the different features. These subjects were women and men belonging to different counties in the USA, divided into 15 groups based on their characteristics, and were randomly selected with a computer algorithm by NHANES. The algorithm consists of a complex multistage probability design to choose the participants, with a series of stages. The first selection is related to primary sampling units (PSUs), which are counties or small groups of them; the next selection is about segments within PSUs that constitute a group of households; then, a selection of specific households is performed; and finally, a selection of individuals within a household.

2.3.2. Variables

The data used are a collection of 190 features (described in Appendix A). From these features, 189 correspond to demographic and dietary data, which were used as descriptors or inputs. Demographic data refers to individual-, family-, and household-level information, while dietary data refers to the total nutrient intake. The remaining feature was contained by the dental caries status, and it was set as output. This output feature describes a patient's dentition; if the subject has suffered from caries or restorations they are assigned the value "1", while if the subject has not, they are assigned the value "0".

2.4. Statistical Methods

All data preprocessing, analysis, and validation were carried out using the free software *R* (version 3.4), which is an environment for statistical computing and graphics that provides tools for different data mining techniques [16]. Packages for *R* that were used are *rms* (version 5.1.2) [17], *MASS* (version 7.3.49) [18], *rminer* (version 1.4.2) [19], *pROC* (version 1.10.0) [20], *randomForest* (version 4.6.12) [21], and *nricens* (version 1.5) [22].

2.4.1. Data Preprocessing

The first preprocessing step consisted of manually removing those input features that presented a high percentage of missing data ($\geq$70%) or singular values, which made their use impossible for the statistical analysis of this work. Missing data were represented as Not a Number (NaN), while singular features presented the same value in all rows of a column or presented multiple values with another feature between each other.

The remaining input features presented <20% of missing values, and were imputed using the *rfimput* function from the *R* package *randomForest*. The imputation based on this function consists of replacing all *NaN* with the median value of the column where the missing data are located [21].

The next part in the preprocessing step consisted of dividing the database in three different datasets, where subjects were classified into three different age groups, according to the age range where they belong. The age ranges were defined by the contribution of Dr. Nubia M. Chávez-Lamas, who is an expert in oral health and an author of this work.

Finally, after the classification of subjects in the age groups, it was necessary to repeat the process of removing all features that presented singular values for group 2 (10–19 years), due to the poor quantity of data that were contained within it. Those features were removed because they presented the same value in all their rows, which means that all subjects presented the same information for those features, being useless for this analysis, where we are looking for the most relevant differences that subjects present. For age groups 3 and 4 it was not necessary to remove features due to singular values.

2.4.2. Feature Selection

After all data were subjected to a preprocessing step, three different datasets were obtained containing subjects belonging to the three different age ranges. Then, each dataset was subjected separately to a feature selection process using an FBS approach in order to select the features that presented the best performance for the classification of patients.

The FBS process consisted of initially subjecting all the features to a multivariate logistic regression (LR), in order to obtain a general model based on the relationship between input features and the output feature.

LR is an analysis that consists of a statistical technique to model the relationship between the input/independent features and the output/dependent feature, using binary data. It measures the contribution of different factors to the occurrence of a simple event, and its main objective is to model the influence of the probability of this event. In Equation (1), the simplest representation of a model obtained by this method is presented, where y is the output feature. y must be subjected to a logarithmic transformation, represented as $logit$, because the model is initially exponential. Through this transformation, it is possible to use it as a lineal function; w is an offset term that can be included, and is also known as bias; $\beta_1 \ldots \beta_n$ are the slopes or coefficients calculated to obtain the real solution of the equation, and $x_1 \ldots x_n$ are the input features to be analyzed. Models can be composed by the number of input features needed [23].

$$logit(y = 1) = w + \beta_1 x_1 + +\beta_2 x_2 + \ldots + \beta_n x_n \quad (1)$$

Then, after applying LR, a feature selection process was carried out using the fbw function from the *rms* package, which performs a numerically stable version of FBS, using a method based on Lawless and Singhal [24].

FBS belongs to the stepwise methods that are used for feature selection, which are techniques that add and/or delete features through iterations, keeping only those features that present the best aptitude in the required task. This method starts subjecting the model, contained by all n features, to multiple regressions, and eliminating features one at a time. At each step, the feature that was selected to be deleted presented the least inflation in the residual sum of squares. The iterative technique continues until only one feature is left in the model or until a stopping rule or threshold is satisfied. For this work, the stopping rule selected was the p-value because it is one of the main parameters used for this technique as a threshold, according to different works [25,26].

The p-value or the observance significance level is the smallest value obtained where the null hypothesis can be rejected and is calculated using the sampling distribution of the test statistic under the null hypothesis [2].

In FBS, it is necessary to be in a situation where the number of observations is greater than the number of features, in order to avoid problems of overfitting. It is important to mention that

the threshold that was selected for the feature selection in age group 2 (p-value < 0.5) was different from the selected for the age groups 3 (20–59 years) and 4 (60 or more years) (p-value < 0.005). This difference between age groups was due to the significant difference in the number of subjects and features that were contained in the age groups, being specially remarkable in age group 2, causing difficulties in finding a general model for this group.

Finally, a multivariate model was developed for each age group. These multivariate models were contained by a series of demographic and dietary features that were selected due to their significant performance in the classification of subjects with the presence or absence of caries.

After this step, all models were subjected to a validation process based on an NRI approach and an analysis of sensitivity (true positives) and specificity (true negatives).

2.4.3. Validation

NRI is a very popular measure that was introduced in 2008 as a new statistic for evaluating the performance of prediction models based on a cross-validation process. This evaluation is carried by adding a marker to a set of reference predictors to predict a binary outcome, and it is based on the principle on comparing if a model B can predict an outcome better than a model A. One of the main purposes of the use of prediction models is to assess whether the addition of a new prediction feature improves the discrimination of who will experience an event from those who will not. A very simple measure that is used to validate the discrimination capacity of the models is to calculate the difference between the average of the predicted risk obtained in those that have a value of the binary result and those who have the other value; the greater this difference, the better the discrimination.

NRI is a technique that has demonstrated clinical utility based on the improvement of clinical decision-making that a specific model can achieve. Through a statistical validation (comparing a first model vs. a second model), NRI measures the proportion of subjects with the condition that is correctly reclassified up to high risk and incorrectly reclassified down to low risk; and on the other hand, the proportion of subjects without the condition that is correctly reclassified down to low risk and incorrectly reclassified up to high risk. This validation process is a confounder-adjusted estimate that obtains as result the reclassification value and the proportion rates, which represent a confidence interval of the precision [10,22].

For this work, NRI was performed using the *nricens* package, which provides the functions to estimate the NRI parameters for risk prediction models with two main approaches: time to event, and binary response data. The binary response data approach was selected, which can be calculated with the *nribin* function. The risk category can be calculated using LR. Additionally, confidence intervals were calculated using the percentile bootstrap method. For the risk category calculation, it was necessary to specify a *cut* parameter, which represents the cutoff values of the risk category. Then, from this function a parameter was returned with its confidence intervals (i.e., the point and the intervals estimated by the NRI process and its components). The *up* and *down* values were also returned, which are logical values that allow the determination of the number of subjects that belong to UP (high risk) and DOWN (low risk) parameters, with their respective reclassification tables of objects that correspond to all, case, and control subjects [22].

NRI has gained great popularity in a series of biomedical applications; nevertheless, it has been demonstrated that some problems related to the fitting of the risk models may be present. Based on this caution, the results obtained under NRI are backed up with the statistical parameters OR, AUC, and ROC curve [27].

The OR parameter is defined as the value of the relationship between an input feature and an outcome, and it is represented as the probability that this specific outcome will occur under a particular feature, against the probability that the outcome will occur with the absence of that feature. Additionally, the 95% confidence intervals (from 2.5% to 97.5%) are calculated [28]. Equation (2) represents the calculation of OR, where P_a is the probability of a first event occurring, while P_b is the probability of a second event occurring, taking into account that the first one has already occurred.

$$ODDSratio = \frac{\frac{P_a}{1-P_a}}{\frac{P_b}{1-P_b}} \quad (2)$$

The AUC value is a standard method that evaluates the accuracy of the classification model based on the relationship between the specificity and the sensitivity, obtained by calculating the ROC curve [29].

Sensitivity is defined as the proportion of subjects with one condition that were classified as positive; this value is calculated by Equation (3), where TP represents the number of true positives and FP represents the number of false positives.

$$PPV = \frac{TP}{TP + FP} \quad (3)$$

Specificity is defined as the proportion of subjects without a condition that were classified as negative; this value is calculated by Equation (4), where TN represents the number of true negatives and FN represents the number of false negatives.

$$NPV = \frac{TN}{TN + FN} \quad (4)$$

3. Results

Results obtained from the feature extraction using a FBS approach for each dataset (corresponding to the three different age groups) and the validation of each multivariate model using NRI, OR, ROC, and AUC are presented in this section.

3.1. Participants

The subject selection process is shown in Figure 2, where from the total 14,332 persons that were selected from the different survey locations, 10,175 completed the interview and 9812 were examined. From the 9812 total subjects, 6122 belonged to cases and 3690 belonged to controls. Case subjects were those that presented dental caries before or during the interviews, while control subjects were those that did not present dental caries before or during the interviews. The number of women were 4831 and the number of were men 4982.

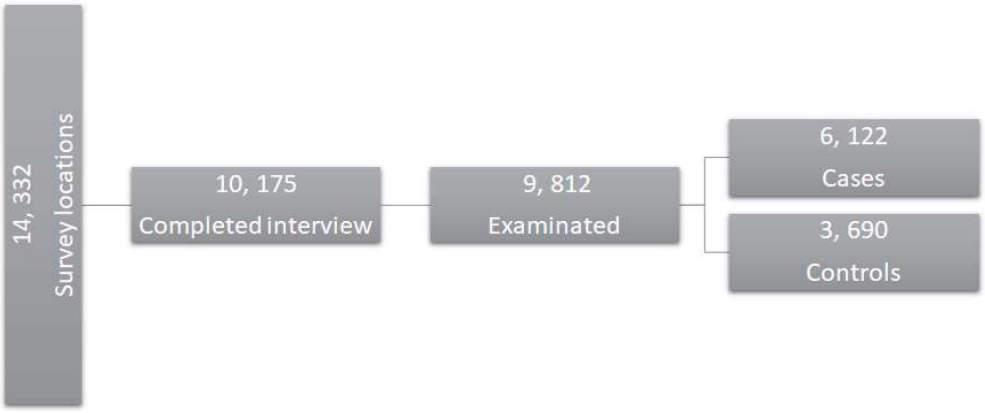

Figure 2. Subject selection process.

Descriptive Data

The main target population for NHANES is the non-institutionalized civilian resident population of the USA. There have been a series of changes in the design of the population selection in order

to sample larger numbers of specific subgroups that present particular public health characteristics, thus to increase the reliability and precision of health status indicators. NHANES started these design changes in 2011, including in its population the oversampled subgroups survey cycle:

- Hispanic persons;
- Non-Hispanic black persons;
- Non-Hispanic Asian persons;
- Non-Hispanic white and other* persons at or below 130% of the poverty level; and
- Non-Hispanic white and other* persons aged 80 years and older.

3.2. Outcome Data and Main Results

Here are reported the preprocessing, statistical analysis, and validation steps. For the preprocessing are presented the number of features that were removed due to missing data and singular values, in addition to the imputation technique of the remaining missing data. Then, for the statistical analysis are presented the multivariate models that were developed using the FBS method, including all features for each model and their description. Finally, the validation results are shown, presenting the values obtained for the NRI technique and the values obtained for the AUC and OR, besides the ROC curve.

3.2.1. Preprocessing

The first step of the statistical analysis that was performed was a preprocessing. The first part of the preprocessing consisted of manually removing 85 features from the 189 total input features because they presented a high percentage of missing data or singular values. Then, from the 104 remaining input features, there were some data that presented <20% of missing values, which were imputed using the $rfimput$ function from the *R* package *randomForest*.

Once all data were complete, they were separated in three different datasets according to the age group where the population belonged. The age groups and the number of subjects that were contained in each group are presented in Table 1. Age group 1 was discarded from the analysis of this work because it was not contained by any subject. Therefore, only age groups 2, 3, and 4 were retained. Finally, the process of removing all features that presented singular values for age group 2 was repeated, eliminating 48 features and keeping 56. For age groups 3 and 4 it was not necessary to remove features due to singular values.

Table 1. Developed age groups where subjects were classified.

Age Group	Age	Subjects	Case/Control
1	0–9	0	0
2	10–19	112	50/62
3	20–59	7603	4551/3052
4	60≤	2097	1521/576

3.2.2. Feature Selection and Validation

After the preprocessing step, a feature selection was performed in order to obtain a multivariate model for each age group containing the most statistically significant features from the total features set.

The multivariate model obtained for age group 2 contained 33 features, that obtained for age group 3 contained 18 features, and that obtained for age group 4 contained 10 features, out of a total of 104 features.

Table 2 presents the 33 features contained in the model obtained for age group 2, with their respective description, OR values, and confidence intervals.

Table 2. Odds ratio (OR) values and confidence intervals (2.5%/97.5%) obtained by the statistical analysis of the multivariate model corresponding to age group 2, containing 15 demographic features and 18 dietary features. PSU: primary sampling unit.

Age Group 2: 112 Subjects (50 Cases/62 Control), 1.14% of the Total Subjects.				
Feature	**Description**	**OR**	**2.5%**	**97.5%**
	Demographic features			
RIAGENDR	Gender of the participant.	6.269	1.0817	50.844
RIDRETH1	Recode of reported race and Hispanic origin information.	4.17×10^{-2}	2.66×10^{-3}	0.425
RIDRETH3	Recode of reported race and Hispanic origin information, with Non-Hispanic Asian category.	8.391	1.458	63.056
RIDEXMON	Six-month time period when the examination was performed.	1.00×10^{2}	7.763	2967.384
RIDEXAGM	Age in months of the participant at the time of examination.	1.045	1.007	1.093
DMDEDUC3	Highest grade or level of school completed or the highest degree received.	5.07×10^{-1}	2.49×10^{-1}	0.888
DMDMARTL	Marital status.	2.243	9.76×10^{-1}	5.793
DMDFMSIZ	Total number of people in the family.	2.03×10^{-1}	5.38×10^{-2}	0.562
DMDHHSZA	Number of children aged 5 years or younger in the household (HH).	3.640	6.62×10^{-1}	24.505
DMDHRGND	HH reference person's gender.	2.56×10^{-1}	2.63×10^{-2}	1.842
DMDHREDU	HH reference person's education level.	8.235	1.962	47.976
SDMVPSU	Masked variance unit pseudo-PSU variable for variance estimation.	6.12×10^{-2}	2.49×10^{-3}	0.864
INDHHIN2	Total household income (reported as a range value in dollars).	9.44×10^{-1}	8.91×10^{-1}	0.990
INDFMIN2	Total family income (reported as a range value in dollars).	1.151	1.023	NA
INDFMPIR	A ratio of family income to poverty guidelines.	4.259	4.60×10^{-1}	40.677
	Dietary features			
WTDRD1	Dietary day one sample weight.	1.000	1.000	1.000
WTDR2D	Dietary two-day sample weight.	9.99×10^{-1}	9.99×10^{-1}	1.000
DR1DRSTZ	Dietary recall status.	4.972	1.721	17.311
DBQ095Z	Type of salt usually added to food at the table.	1.102	1.006	NA
DBD100	Frequency with which ordinary salt is added to the food on the table.	2.87×10^{-1}	6.86×10^{-2}	1.003
DRQSPREP	Frequency with which ordinary salt or seasoned salt is added in cooking or preparing foods in the household.	1.612	6.37×10^{-1}	4.444
DR1TKCAL	Energy (kcal).	9.81×10^{-1}	9.54×10^{-1}	1.005
DR1TPROT	Protein (g).	1.138	1.010	1.311
DR1TCARB	Carbohydrate (g).	1.129	1.005	1.298
DR1TSUGR	Total sugars (g).	9.39×10^{-1}	8.82×10^{-1}	0.986
DR1TFIBE	Dietary fiber (g).	6.95×10^{-1}	5.01×10^{-1}	0.903
DR1TTFAT	Total fat (g).	4.78×10^{-1}	2.28×10^{-1}	0.859
DR1TSFAT	Total saturated fatty acids (g).	3.168	1.548	7.820
DR1TMFAT	Total monounsaturated fatty acids (g).	2.409	1.228	5.476
DR1TPFAT	Total polyunsaturated fatty acids (g).	2.349	1.235	5.126
DR1TLYCO	Lycopene (mcg).	1.000	9.99×10^{-1}	1.000
DR1TFA	Folic acid (mcg).	9.74×10^{-1}	9.54×10^{-1}	0.991
DR1TB12A	Added vitamin B12 (mcg).	2.469	1.244	5.555

After the validation process using NRI for age group 2, the graph from Figure 3 was obtained, where the risk category under the LR technique is presented, defining as cutoff the range from 0.5

to 0.7. Subjects obtaining a classification probability >0.5 belong to controls and subjects obtaining a classification probability ≥0.7 belong to cases. The NRI values obtained were $NRI = -0.080$, $+NRI = -0.080$, $-NIR = 0$, with the proportion rates $Pr(Up|Case) = 0.040$, $Pr(Down|Case) = 0.120$, $Pr(Down|Control) = 0.048$, $Pr(Up|Control) = 0.048$.

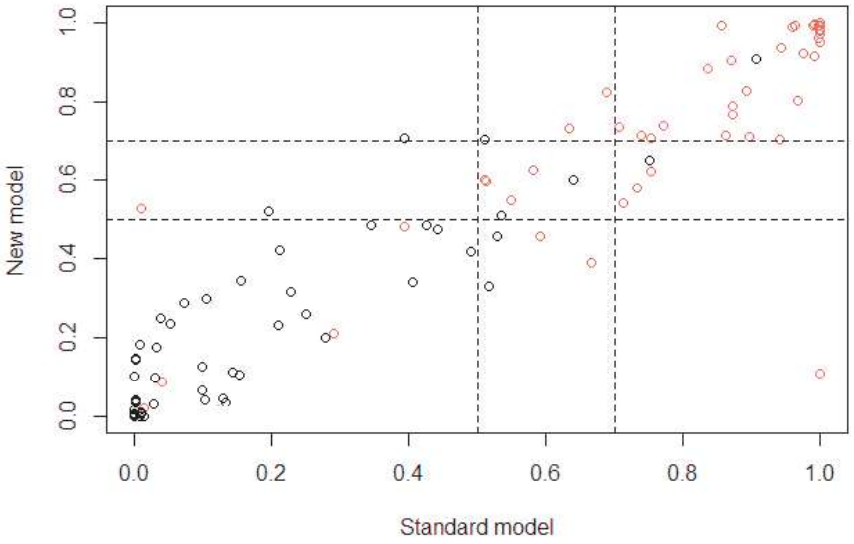

Figure 3. Plot of the risk category based on the NRI calculus for age group 2; horizontal axis represents the standard model or the multivariate model before the feature selection and vertical axis represents the new model or the multivariate model after the feature selection (control ○, case ○).

Table 3 presents three reclassification tables, measuring the true and false positives and the true and false negatives by comparing the results obtained in the classification of subjects using the standard model or the model before the feature selection and the new model generated through the feature selection for age group 2.

Table 3. Reclassification table obtained for model 2 using NRI, 0.5 and 0.7 values represent the cutoff range used to calculate the risk category. Values in red represent true negatives, values in green represent true positives, and values in the cutoff range are presented in gray. False positives and false negatives are presented in black.

Age Group 2: 112 Subjects (50 Cases/62 Control), 1.14% of the Total Subjects.											
All Subjects				**Case**				**Control**			
New				New				New			
Standard	<0.5	<0.7	≥0.7	Standard	<0.5	<0.7	≥0.7	Standard	<0.5	<0.7	≥0.7
<0.5	57	1	1	<0.5	4	0	0	<0.5	53	1	1
<0.7	4	6	3	<0.7	2	4	2	<0.7	2	2	1
≥0.7	1	4	35	≥0.7	1	3	34	≥0.7	0	1	1

Table 4 presents the 18 features that were contained in the model obtained for age group 3, with their respective description, OR values, and confidence intervals.

Table 4. OR values and confidence intervals (2.5%/97.5%) obtained by the statistical analysis of the multivariate model corresponding to age group 3, containing 6 demographic features and 12 dietary features.

Age Group 3: 7603 Subjects (4551 Cases/3052 Control), 77.48% of the Total Subjects.				
Feature	**Description**	**OR**	**2.5%**	**97.5%**
	Demographic features			
RIDAGEYR	Age in years of the participant at the time of screening.	1.029	1.025	1.033
RIDRETH1	Recode of reported race and Hispanic origin information.	0.915	0.876	0.956
RIDEXAGM	Age in months of the participant at the time of examination.	1.006	1.005	1.006
DMDEDUC2	Highest grade or level of school completed or the highest degree received.	1.340	1.270	1.414
DMDHHSZB	Number of children aged 6–17 years old in the household (HH).	1.128	1.076	1.183
DMDHSEDU	HH reference person's spouse's education level.	0.837	0.789	0.888
	Dietary features			
WTDR2D	Dietary two-day sample weight.	1.000	1.000	1.000
DR1DRSTZ	Dietary recall status.	0.923	0.886	0.962
DR1TKCAL	Energy (kcal).	0.992	0.989	0.996
DR1TPROT	Protein (g).	1.031	1.017	1.046
DR1TCARB	Carbohydrate (g).	1.029	1.016	1.042
DR1TTFAT	Total fat (g).	1.071	1.040	1.103
DR1TCHOL	Cholesterol (mg).	1.001	1.000	1.001
DR1TFA	Folic acid (mcg).	1.001	1.000	1.001
DR1TCHL	Total choline (mg).	0.998	0.997	0.999
DR1TIRON	Iron (mg).	0.969	0.957	0.982
DR1TALCO	Alcohol (g).	1.062	1.038	1.088
DR1TS080	SFA 8:0 (Octanoic) (g).	0.564	0.445	0.708

Then, after the validation process using NRI for age group 3, the graph in Figure 4 was obtained, where the risk category is presented, defining the range from 0.5 to 0.7 as cutoff, where subjects that obtain a classification probability >0.5 belong to controls and subjects that obtain a classification probability ≥0.7 belong to cases. The NRI values obtained were $NRI = -0.024$, $+NRI = -0.019$, $-NIR = -0.005$; with the proportion rates $Pr(Up|Case) = 0.055$, $Pr(Down|Case) = 0.074$, $Pr(Down|Control) = 0.055$, $Pr(Up|Control) = 0.060$.

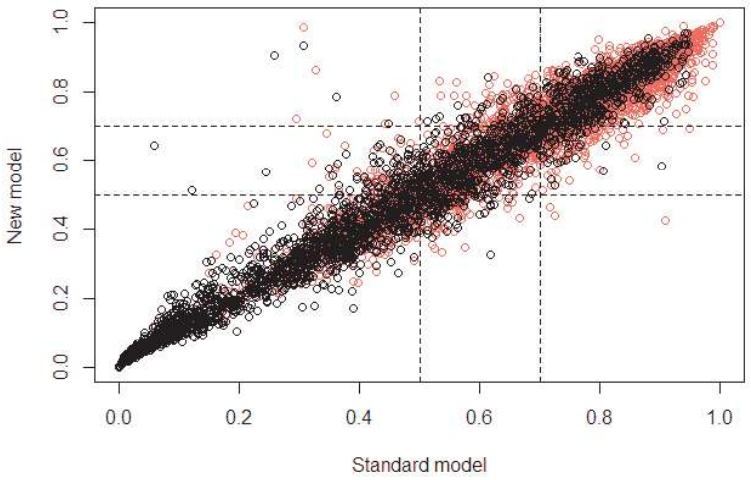

Figure 4. Plot of the risk category based on the NRI calculus for age group 3. The horizontal axis represents the standard model or the multivariate model before the feature selection and the vertical axis represents the new model or the multivariate model after the feature selection (control ○, case ○).

Table 5 presents three reclassification tables, measuring the true and false positives as well as the true and false negatives by comparing the results obtained in the classification of subjects using the standard model or the model before the feature selection and the new model generated through the feature selection for age group 3.

Table 5. Reclassification table obtained for model 3 using NRI; 0.5 and 0.7 values represent the cutoff range used to calculate the risk category. Values in red represent true negatives, values in green represent true positives, and values in the cutoff range are presented in gray. False positives and false negatives are presented in black.

Age Group 3: 7603 Subjects (4551 Cases/3052 Control), 77.48% of the Total Subjects.											
All Subjects				**Case**				**Control**			
New				New				New			
Standard	<0.5	<0.7	≥0.7	Standard	<0.5	<0.7	≥0.7	Standard	<0.5	<0.7	≥0.7
<0.5	2244	195	8	<0.5	610	83	5	<0.5	1634	112	3
<0.7	223	1252	236	<0.7	124	745	166	<0.7	99	507	70
≥0.7	3	284	3158	≥0.7	3	214	2601	≥0.7	0	70	557

Table 6 presents the 10 features that were contained in the model obtained for age group 4, with their respective description, OR values, and confidence intervals.

Table 6. OR values and confidence intervals (2.5%/97.5%) obtained by the statistical analysis of the multivariate model corresponding to age group 4, containing four demographic features and six dietary features.

Age Group 4: 2097 Subjects (1521 Cases/576 Control), 21.37% of the Total Subjects.				
Feature	**Description**	**OR**	**2.5%**	**97.5%**
	Demographic features			
RIDRETH1	Recode of reported race and Hispanic origin information.	0.777	0.704	0.856
RIDEXAGM	Age in months of the participant at the time of examination.	1.004	1.003	1.006
DMDEDUC2	Highest grade or level of school completed or the highest degree received.	1.471	1.339	1.618
INDFMPIR	Ratio of family income to poverty guidelines.	1.205	1.112	1.308
	Dietary features			
DR1DRSTZ	Dietary recall status.	0.770	0.712	0.833
DR1TATOC	Vitamin E as alpha-tocopherol (mg).	1.087	1.049	1.128
DR1TATOA	Added alpha-tocopherol (Vitamin E) (mg).	0.909	0.870	0.951
DR1TVD	Vitamin D (D2 + D3) (mcg).	0.958	0.938	0.980
DR1TPHOS	Phosphorus (mg).	1.000	1.000	1.000
DR1TS160	SFA 16:0 (Hexadecanoic) (g).	0.943	0.921	0.964

Then, after the validation process using NRI for age group 4, the graph from Figure 5 was obtained, presenting the risk category, defining as cutoff the range from 0.5 to 0.7, where subjects that obtain a classification probability >0.5 belong to controls and subjects that obtain a classification probability ≥0.7 belong to cases. The NRI values obtained were $NRI = -0.129$, $+NRI = -0.013$, $-NIR = -0.116$. The proportion rates were $Pr(Up|Case) = 0.083$, $Pr(Down|Case) = 0.096$, $Pr(Down|Control) = 0.081$, $Pr(Up|Control) = 0.197$.

Table 7 presents three reclassification tables, measuring the true and false positives as well as the true and false negatives by comparing the results obtained in the classification of subjects using the standard model or the model before the feature selection and the new model generated by the feature selection for age group 4.

After the validation step based on the NRI parameters, the ROC and AUC parameters were calculated for a second validation. Figure 6 presents in (A) the ROC curve for age group 2, obtaining an AUC value of 0.933. (B) presents the ROC curve for age group 3, obtaining an AUC value of 0.787. (C) presents the result for age group 4, obtaining an AUC value of 0.735.

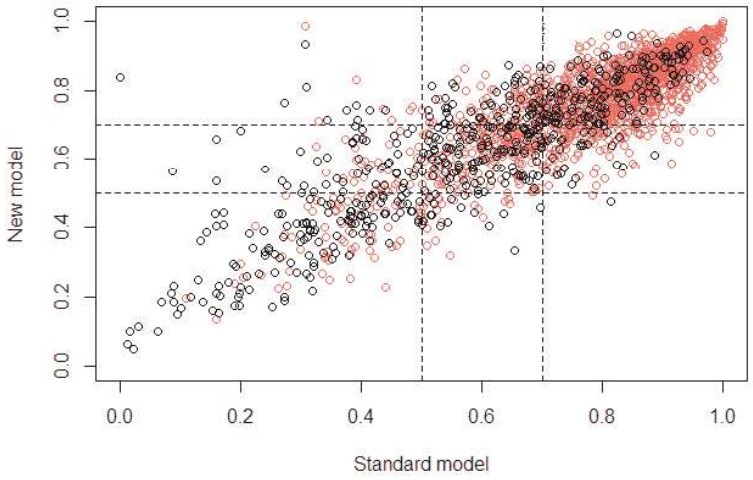

Figure 5. Plot of the risk category based on the NRI calculus for age group 4. The horizontal axis represents the standard model or the multivariate model before the feature selection, and the vertical axis represents the new model or the multivariate model after the feature selection (control ○, case ○).

Table 7. Reclassification table obtained for model 4 using NRI, 0.5 and 0.7 values represent the cutoff range used to calculate the risk category. Values in red represent true negatives, values in green represent true positives, and values in the cutoff range are presented in gray. False positives and false negatives are presented in black.

All Subjects				**Case**				**Control**			
New				New				New			
Standard	<0.5	<0.7	≥0.7	Standard	<0.5	<0.7	≥0.7	Standard	<0.5	<0.7	≥0.7
<0.5	193	81	16	<0.5	65	33	7	<0.5	28	48	9
<0.7	42	299	144	<0.7	24	177	87	<0.7	18	122	57
≥0.7	2	150	1170	≥0.7	1	122	1005	≥0.7	1	28	165

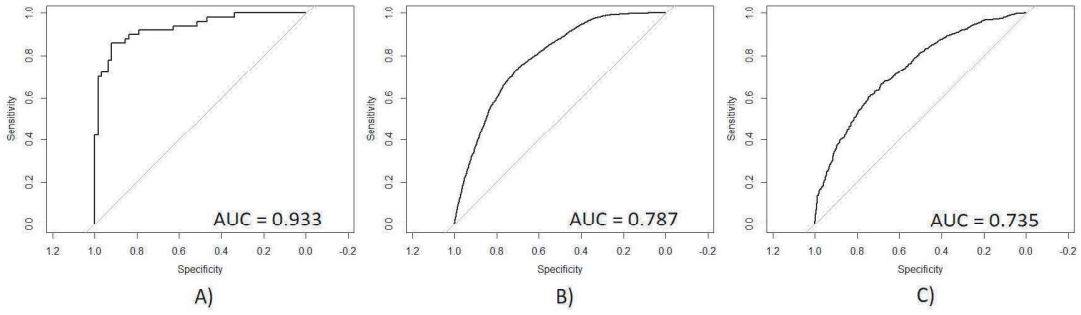

Figure 6. Receiver operating characteristic (ROC) curves of the generated multivariate models for the detection of caries. The horizontal axis represents the 1-specificity measure and the vertical axis represents the sensitivity measure. (**A**) Age group 2, (**B**) Age group 3, (**C**) Age group 4. AUC: area under the curve.

4. Discussion and Conclusions

This section presents the discussion and conclusions for the reported results obtained for this case–control study, where the main objective was to find a multivariate model that allows for classification of subjects with the presence of caries from subjects with their absence, according to their age, looking for the demographic and dietary features that bring the most descriptive information for cases and controls. This helps to indicate when a subject is at risk of suffering from dental caries, making it possible to take preventive measures and thus decreasing this public health problem.

The three multivariate models that were obtained for the age groups present different characteristics in the classification of subjects. Some of the features are present in all three models; nevertheless, there are remarkable differences in specific features that were selected for each group, besides the quantity of features that are contained in each. It is important to remark that due to the difference in the number of subjects belonging to each age group, it was necessary to use two different cutoff values in the selection process; nevertheless, the impact of this difference was not significant according to the results obtained, since it did not cause problems of overfitting and the validation remained consistent between the age groups. Therefore, it is possible to compare them.

In Table 2 it is possible to observe that according to the OR values obtained, most features in the model were statistically significant and the proportion of demographic features was very similar to the proportion of dietary features (15 features and 18 features, respectively), which means that for age group 2, both types of features provided important information in very similar proportion for the classification of subjects. On the other hand, the values obtained from the NRI analysis showed that the reclassification of subjects through the developed multivariate model had a mean proportion of -0.080, which means that the classification was 0.080 better if the standard model (all features) was used to classify cases instead of the new model obtained trough the feature selection. The correct reclassification of subjects with presence of caries had a proportion of 0.040, while the incorrect had a proportion of 0.120. The correct reclassification of control subjects had a proportion of 0.048, which is the same as that obtained for the incorrect reclassification. These values are very significant considering that the reclassification was performed using 33 features instead of 56. In Table 3 it is easier to observe that the reclassification presented a confusion problem to classify subjects in the upper cutoff (i.e., the cases) subjects, showing a higher error in that proportion in comparison with the reclassification proportion of the lower cutoff (i.e., the controls) subjects, being evident in the reclassification of cases subjects exclusively, which means that the features that were selected for this age group may be presenting similar values for some subjects in both type of outcomes, case and control, inducing a classification error in the new model. Figure 3 is a graphical represenation indicating that most subjects were classified in a similar way for both models (standard and new), according to the linear behavior that is shown. Nevertheless, the classification of case and control subjects in the cutoff region can be wrong because it is the range of values where the threshold between both outcomes is present, which means that the subjects that are part of this range of values may belong to controls with a similar probability of belonging to cases.

For age group 3, in Table 4 it is possible to observe that according to the OR values obtained, all features presented a probability of subjects being case very similar to them being control, which means that any feature provided more information to classify cases than the others. The proportion of demographic features was reduced in comparison to the proportion of dietary features (6 demographic features and 12 dietary features), which means that for age group 3, there were more dietary features that provided information for the classification of subjects than demographic fetures. On the other hand, the values obtained from the NRI analysis show that the reclassification of subjects through the developed multivariate model had a mean proportion of -0.024, which means that the classification was 0.024 better if the standard model (all features) was used instead of the new model. From this mean value, 0.019 proportion belongs to the classification of cases and 0.005 proportion belongs to controls; this value is not relatively significant in comparison with the NRI values obtained for age group 2. The correct reclassification of subjects with presence of caries had

a proportion of 0.055, while the incorrect had a proportion of 0.074. The correct reclassification of control subjects had a proportion of 0.055, while the incorrect had a proportion of 0.060. Based on these values, it is possible to say that for this new model is was easier to classify control subjects than case. However, the increase of the error in the classification of case subjects is not very significant, considering that the reclassification was performed using 18 features instead of 104. In Table 5 it is easier to observe that the reclassification presented a small proportion of false positives and false negatives; however, the cutoff range presented a significant value of uncertainty due to the high number of subjects that were between the threshold values, which means that the features that were extracted for this age group may be presenting similar values for some subjects in both types of outcome, inducing an error in the classification through the new model. Figure 4 allows graphical observation that most subjects were classified similarly for both models (standard and new), according to the linear behavior that is shown. Nevertheless, the classification of case and control subjects in the cutoff region can be wrong because it is the range of values where subjects may belong to controls with a similar probability of belonging to cases.

Finally, for age group 4, in Table 6 it is possible to observe that according to the OR values obtained, all features presented a probability of subjects being case very similar to the probability of them being control, as in age group 3. This multivariate model presented the smallest number of features in comparison with age groups 2 and 3, and the proportion of demographic features was very similar to the proportion of dietary features (four demographic features and six dietary features), which means that for this age group, both types of features provided important information and in very similar proportion for the classification of subjects. The statistical values obtained from the NRI analysis show that the reclassification of subjects through the developed multivariate model had a mean proportion of −0.129, which means that the classification was 0.129 better if the standard model (all features) was used instead of the new model, being the highest value in the three age groups, where the standard model classified with a better behavior than the new one. This may occur due to the small number of features that are part of this age group. The correct reclassification of subjects with the presence of caries had a proportion of 0.083, while the incorrect had a proportion of 0.096. The correct reclassification of control subjects had a proportion of 0.081, while the incorrect had a proportion of 0.197. Based on these values it is possible to say that for this new model it is easier to classify control subjects than case; however, the increase of the error in the classification of case subjects was not very significant considering that the reclassification was performed using 10 features instead of 104. In Table 7 it is easier to observe that the reclassification presented a very similar proportion of false positives and false negatives. The lower cutoff (i.e., the controls) presented the highest error in the reclassification, and the subjects that were located in the cutoff range also presented a significant value, which means that the features that were extracted for this age group may be presenting similar values for some subjects in both type of outcomes, inducing an error in the classification through the new model. Figure 5 demonstrates that most subjects were classified similarly for both models (standard and new), according to the linear behavior that is shown. Nevertheless, the classification of case and control subjects in the cutoff region can be wrong because it is the range of values where subjects may belong to controls with a similar probability of belonging to cases, besides being remarkable that a significant number of control subjects appeared in those values that would be assigned to case subjects, causing the uncertainty value due to the false positives.

According to the NRI values obtained using the new models, all of them presented statistically significant reclassification values, obtaining parallel proportions of true positives/true negatives, considering that the number of subjects was very different for each age group, as the number of features for each model.

It is important to remark that age group 2 contained the smallest number of subjects but the largest number of features, which means that the smaller the quantity of data, the more difficult it is to generate a general model that correctly classify them, making it necessary to use more features.

On the other hand, Figure 6 presents the ROC obtained for each age group with their respective AUC values. (A) shows a very significant ROC with a sensitivity/specificity proportion of 0.933, which means that 93.3% of subjects from age group 2 were correctly classified using the new model. (B) shows the ROC curve for age group 3 with a sensitivity/specificity proportion of 0.787, which means that 78.7% of the total subjects were correctly classified using the new model. (C) shows the ROC curve for age group 4 with a sensitivity/specificity proportion of 0.735, which means that 73.5% of the subjects were correctly classified. According to these results, it is possible to observe that age group 2 had a remarkably better performance than the other two age groups, which presented very similar AUC values. The accuracy of age group 2 may be reached because this group had a smaller number of subjects and a higher number of features on its multivariate model. However, this relationship of subjects and features may cause an overfit problem, making difficult to generalize this multivariate model. On the other hand, the AUC values obtained for age groups 3 and 4 were also statistically significant, which means that these models presented a good performance in calculating the correct outcome for each subject using the smallest number of data possible, extracted from a very large quantity of data, which means that age groups 3 and 4 did not present the problem of overfitting like age group 2.

According to these results, it is possible to conclude that the use of FBS for feature selection presents a good performance, based on the statistical validation using NRI, ROC, AUC, and OR. The feature selection was used to develop three different multivariate models for the classification of control and case subjects, taking into account the age groups to which the patients belonged in order to know if the age of subjects changed the risk of suffering caries, based on their demographic and dietary information.

All multivariate models that were developed were significant, and each presented different characteristics from the others. The model of age group 2 presented the highest accuracy in the classification of subjects and the proportion of demographic and dietary features was very similar, which means that for subjects that were between 10 and 19 years old (age group 2), both types of features influenced the development of caries. This may occur because at this age range, aside from the importance of the good feeding (information that is present in dietary features), it is important to teach or educate children ranging from 10 years old to teenagers to have good oral health, taking into account the environment in which they are developed (information that is present in demographic features).

For age group 3 a model was developed that presented double the number of dietary features than demographic, which means that subjects between 20 and 59 years old (age group 3) need the information of dietary features more than demographic features to know their dental status and their risk of developing dental caries. Finally, for age group 4, it was more important for subjects that are 60 years old or more (age group 4) to pay more attention in their feeding than in their demographic status in order to avoid caries problems.

Through this statistical analysis it was possible to find which demographic and dietary features were the most significant to bring information about the development of caries in three different groups of people, being a case/control study. These groups of people correspond to three different age ranges. Nevertheless, it is important to consider that one of the main limitations of these results is the type of subjects that were used for the statistical analysis and the number of subjects in the datasets. These databases collected a series of data from a great diversity of subjects, which may represent a general view of the public health problem, but for a centralization of the problem, it would be important to use a more specific dataset with the demographic and dietary features of the population under study (i.e., Mexican subjects). On the other hand, the number of subjects corresponding to each age group is unbalanced, which may induce a bias in the feature selection and the validation process. Another limitation is presented for age groups 3 and 4, which obtained an uncertainty value of around 30%, representing a relative significant error for the prediction of dental caries. Finally, even when the most accurate model was obtained using the data from age group 2, it is important to note that in this age range (10–19 years), those subjects that are younger are dependent on their relatives for their

demographic and dietary circumstances, making dental caries reduction not an exclusive problem of the subjects of study, but the environment that surrounds them.

5. Future Work

As future work we propose a different feature selection approach for the development of models based on the genetic algorithm Galgo, validated through a forward selection and a backward elimination process, in order to find the similarities and the differences between those results and the obtained in this work, looking for the improvement of the accuracy in the classification of subjects. Additionally, for the validation stage, a random forest method may be used in order to test the accuracy of the models based on a decision tree approach, ensuring the certainty of the model behavior.

On the other hand, we propose the implementation of an app with the purpose of providing an automated calculation of the probability that a subject presents the risk of developing dental caries, based on a questionnaire with the required information, according to the multivariate model developed. This approach may represent a tool for a preliminary diagnosis that may be available independently of the demographic situation, helping to reduce the high incidence of dental caries.

Finally, it has been proven that the demographic situation can strongly influence the prevalence of many conditions or diseases. Based on this, we propose the analysis of oral health data from exclusively Mexican subjects (Zacatecas state) and the comparison of those results with those obtained in this work, in order to prove the differences and the similitudes between the oral health of both populations, and also to develop a preliminary model for the diagnosis and prognosis of caries for each age group in this specific city.

Author Contributions: L.A.Z.-C. and C.E.G.-T. performed the study and performed the study design and data analysis. N.M.C.-L., M.d.C.G.-C., and A.M.-B. contributed to materials and methods (data base and patients selection criteria) used in this study. C.E.G.-T., J.I.G.-T., J.M.C.-P. performed statistical analysis and statistical validation with critical feedback to authors. J.M.C.-P., J.G.A.-O., and L.A.Z.-C. provided data base explanation. H.G.-R. provided feedback from results. All authors interpreted findings from the analysis and drafted the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1

Table A1. Description of demographic features.

Feature	Description
AIALANGA	Language of the MEC ACASI Interview Instrument.
DMDBORN4	In what country was Sample Person (SP) born?
DMDCITZN	Is SP a citizen of the United States?
DMDHHSIZ	Total number of people in the household.
DMDHHSZE	Number of adults aged 60 years or older in the household.
DMDEDUC2	Highest grade or level of school completed or the highest degree received.
DMDHHSZB	Number of children aged 6–17 years old in the household (HH).
DMDHSEDU	HH reference person's spouse's education level.
DMDMARTL	Marital status.
DMDFMSIZ	Total number of people in the Family.
DMDHHSZA	Number of children aged 5 years or younger in the household.
DMDHRGND	HH reference person's gender.
DMDHREDU	HH reference person's education level.
DMDEDUC3	Highest grade or level of school completed or the highest degree received.

Table A1. *Cont.*

Feature	Description
DMDHRAGE	HH reference person's age in years.
DMDHRBR4	HH reference person's country of birth.
DMDHREDU	HH reference person's education level.
DMDHRMAR	HH reference person's marital status.
DMDYRSUS	Length of time the participant has been in the US.
DMQADFC	Did SP ever serve in a foreign country during a time of armed conflict or on a humanitarian or peace-keeping mission?
DMQMILIZ	Has SP ever served on active duty in the U.S. Armed Forces, military Reserves, or National Guard?
FIAINTRP	Was an interpreter used to conduct the Family interview?
FIALANG	Language of the Family Interview Instrument.
FIAPROXY	Was a Proxy respondent used in conducting the Family Interview?
INDFMPIR	A ratio of family income to poverty guidelines.
INDHHIN2	Total household income (reported as a range value in dollars).
INDFMIN2	Total family income (reported as a range value in dollars).
LATVPSU	Variance unit: PSU variable for variance estimation.
LATVSTRA	Variance unit: stratum variable for variance estimation.
MIAINTRP	Was an interpreter used to conduct the MEC CAPI interview?
MIALANG	Language of the MEC CAPI Interview Instrument.
MIAPROXY	Was a Proxy respondent used in conducting the MEC CAPI Interview?
RIAGENDR	Gender of the participant.
RIDAGEEX	Age in years of the participant at the time of examination. Individuals aged 959 months and older are topcoded at 959 months.
RIDAGEMN	ge in months of the participant at the time of screening. Reported for persons aged 24 months or younger at the time of exam.
RIDRETH3	Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category.
RIDEXMON	Six month time period when the examination was performed.
RIDEXPRG	Pregnancy status for females between 20 and 44 years of age at the time of MEC exam.
RIDAGEYR	Age in years of the participant at the time of screening.
RIDRETH1	Recode of reported race and Hispanic origin information.
RIDEXAGM	Age in months of the participant at the time of examination.
RIDSTATR	Interview and examination status of the participant.
SDMVPSU	Masked variance unit pseudo-PSU variable for variance estimation.
SDDSRVYR	Data release cycle.

Table A2. Description of demographic features.

Feature	Description
SDMVSTRA	Masked variance unit pseudo-stratum variable for variance estimation.
WTLAF8YR	Subsample 8-year fasting weight for participants aged 12 years and older who were examined in the morning sessions in Los Angeles, California.
WTLAI8YR	Full sample 8-year interview weight for participants in Los Angeles, California.
WTLAM8YR	Full sample 8-year MEC exam weight for participants in Los Angeles, California.

Appendix A.2

Table A3. Description of dietary features.

Feature	Description
WTDR2D	Dietary two-day sample weight.
DR1DRSTZ	Dietary recall status.
DR1TPROT	Protein (g).
DR1TCARB	Carbohydrate (g).
DR1TTFAT	Total fat (g).
DR1TCHOL	Cholesterol (mg).
DR1TFA	Folic acid (mcg).
DR1TCHL	Total choline (mg).
DR1TIRON	Iron (mg).
DR1TALCO	Alcohol (g).
DR1TS080	SFA 8:0 (Octanoic) (g).
DR1TATOC	Vitamin E as alpha-tocopherol (mg).
DR1TATOA	Added alpha-tocopherol (Vitamin E) (mg).
DR1TVD	Vitamin D (D2 + D3) (mcg).
DR1TPHOS	Phosphorus (mg).
DR1TS160	SFA 16:0 (Hexadecanoic) (g).
WTDRD1	Dietary day one sample weight.
DBQ095Z	Type of salt usually add to food at the table.
DBD100	Frequency with which ordinary salt is added to the food on the table.
DRQSPREP	Frequency with which ordinary salt or seasoned salt is added in cooking or preparing foods in the household.
DR1TKCAL	Energy (kcal).
DR1TSUGR	Total sugars (g).
DR1TFIBE	Dietary fiber (g).
DR1TSFAT	Total saturated fatty acids (g).
DR1TMFAT	Total monounsaturated fatty acids (g).
DR1TPFAT	Total polyunsaturated fatty acids (g).
DR1TLYCO	Lycopene (mcg).
DR1TFA	Folic acid (mcg).
DR1TB12A	Added vitamin B12 (mcg).
DR1_300	Was the amount of food that you ate yesterday much more than usual, usual, or much less than usual?
DR1_320Z	Total plain water drank yesterday—including plain tap water, water from a drinking fountain, among others.
DR1_330Z	Total tap water drank yesterday—including filtered tap water and water from a drinking fountain.
DR1BWATZ	Total bottled water drank yesterday (g).
DR1DAY	Intake day of the week.
DR1DBIH	Number of days between intake day and the day of family questionnaire administered in the household.
DR1SKY	What type of salt was it? (Was it ordinary or seasoned salt, lite salt, or a salt substitute?).
DR1STY	Did SP add any salt to her/his food at the table yesterday?
DR1TACAR	Alpha-carotene (mcg).
DR1TBCAR	Beta-carotene (mcg).
DR1TCAFF	Caffeine (mg).
DR1TCALC	Calcium (mg).
DR1TCOPP	Copper (mg).
DR1TCRYP	Beta-cryptoxanthin (mcg).
DR1TFDFE	Folate as dietary folate equivalents (mcg).
DR1TFF	Food folate (mcg).
DR1TFOLA	Total folate (mcg).
DR1TLZ	Lutein + zeaxanthin (mcg).
DR1TM161	MFA 16:1 (Hexadecenoic) (g).
DR1TM181	MFA 18:1 (Octadecenoic) (g).
DR1TM201	MFA 20:1 (Eicosenoic) (g).

Table A4. Description of dietary features.

Feature	Description
DR1TM221	MFA 22:1 (Docosenoic) (g).
DR1TMAGN	Magnesium (mg).
DR1TMOIS	Moisture (g).
DR1TNIAC	Niacin (mg).
DR1TNUMF	Total number of foods/beverages reported in the individual foods file.
DR1TP182	PFA 18:2 (Octadecadienoic) (g).
DR1TP183	PFA 18:3 (Octadecatrienoic) (g).
DR1TP184	PFA 18:4 (Octadecatetraenoic) (g).
DR1TP204	PFA 20:4 (Eicosatetraenoic) (g).
DR1TP205	PFA 20:5 (Eicosapentaenoic) (g).
DR1TP225	PFA 22:5 (Docosapentaenoic) (g).
DR1TP226	PFA 22:6 (Docosahexaenoic) (g).
DR1TPOTA	Potassium (mg).
DR1TRET	Retinol (mcg).
DR1TS040	SFA 4:0 (Butanoic) (g).
DR1TS060	SFA 6:0 (Hexanoic) (g).
DR1TS100	SFA 10:0 (Decanoic) (g).
DR1TS120	SFA 12:0 (Dodecanoic) (g).
DR1TS140	SFA 14:0 (Tetradecanoic) (g).
DR1TS180	SFA 18:0 (Octadecanoic) (g).
DR1TSELE	Selenium (mcg).
DR1TSODI	Sodium (mg).
DR1TTHEO	Theobromine (mg).
DR1TVARA	Vitamin A as retinol activity equivalents (mcg).
DR1TVB1	Thiamin (Vitamin B1) (mg).
DR1TVB12	Vitamin B12 (mcg).
DR1TVB2	Riboflavin (Vitamin B2) (mg).
DR1TVB6	Vitamin B6 (mg).
DR1TVC	Vitamin C (mg).
DR1TVK	Vitamin K (mcg).
DR1TWS	When you drink tap water, what is the main source of the tap water? Is the city water supply; a well or rain cistern; a spring; or something else?
DR1TZINC	Zinc (mg).
DRABF	Indicates whether the sample person was an infant who was breast-fed on either of the two recall days.
DRD340	During the past 30 days did you eat any types of shellfish listed on this card?
DRD350A	Clams eaten during the past 30 days.
DRD350AQ	Number of times clams were eaten in the past 30 days.
DRD350B	Crabs eaten during the past 30 days.
DRD350BQ	Number of times crab was eaten in the past 30 days.
DRD350C	Crayfish eaten during the past 30 days.
DRD350CQ	Number of times crayfish was eaten in the past 30 days.
DRD350D	Lobsters eaten during the past 30 days.
DRD350DQ	Number of times lobster was eaten in the past 30 days.
DRD350E	Mussels eaten during the past 30 days.
DRD350EQ	Number of times mussels were eaten in the past 30 days.
DRD350F	Oysters eaten during the past 30 days.
DRD350FQ	Number of times oysters were eaten in the past 30 days.
DRD350G	Scallops eaten during the past 30 days.
DRD350GQ	Number of times scallops were eaten in the past 30 days.
DRD350H	Shrimp eaten during the past 30 days.
DRD350HQ	Number of times shrimp was eaten in the last 30 days.
DRD350I	Other shellfish (ex. octopus, squid) eaten during the past 30 days.
DRD350IQ	Number of times other shellfish (ex. octopus, squid) was eaten in the past 30 days.
DRD350J	Other unknown shellfish eaten during the past 30 days.
DRD350JQ	Number of times other unknown shellfish was eaten in the past 30 days.
DRD350K	Refused to give detailed information on shellfish eaten during the past 30 days.

Table A5. Description of dietary features.

Feature	Description
DRD360	During the past 30 days did you eat any types of fish listed on this card?
DRD370A	Breaded fish products eaten during the past 30 days.
DRD370AQ	Number of times breaded fish products were eaten in the past 30 days.
DRD370B	Tuna eaten during the past 30 days.
DRD370BQ	Number of times tuna was eaten in the past 30 days.
DRD370C	Bass eaten during the past 30 days.
DRD370CQ	Number of times bass was eaten in the past 30 days.
DRD370D	Catfish eaten during the past 30 days.
DRD370DQ	Number of times catfish was eaten in the past 30 days.
DRD370E	Cod eaten during the past 30 days.
DRD370EQ	Number of times cod was eaten in the past 30 days.
DRD370F	Flatfish eaten during the past 30 days.
DRD370FQ	Number of times flatfish was eaten in the past 30 days.
DRD370G	Haddock eaten during the past 30 days.
DRD370GQ	Number of times haddock was eaten in the past 30 days.
DRD370H	Mackerel eaten during the past 30 days.
DRD370HQ	Number of times mackerel was eaten in the past 30 days.
DRD370I	Perch eaten during the past 30 days.
DRD370IQ	Number of times perch was eaten in the past 30 days.
DRD370J	Pike eaten during the past 30 days.
DRD370JQ	Number of times pike was eaten in the past 30 days.
DRD370K	Pollock eaten during the past 30 days.
DRD370KQ	Number of times pollock was eaten in the past 30 days.
DRD370TQ	Number of times other type of fish was eaten in the past 30 days.
DRD370V	Refused to give detailed information on fish eaten during the past 30 days.
DRQSDIET	Are you currently on any kind of diet, either to lose weight or for some other health-related reason?
DRQSDT1	What kind of diet are you on? (Is it a weight loss or low calorie diet: low fat or cholesterol diet; low salt or sodium diet; sugar free or low sugar diet; low fiber diet; high fiber diet; diabetic diet; or another type of diet?).
DRD370L	Porgy eaten during the past 30 days.
DRD370LQ	Number of times porgy was eaten in the past 30 days.
DRD370M	Salmon eaten during the past 30 days.
DRD370MQ	Number of times salmon was eaten in the past 30 days.
DRD370N	Sardines eaten during the past 30 days.
DRD370NQ	Number of times sardines were eaten in the past 30 days.
DRD370O	Sea bass eaten during the past 30 days.
DRD370OQ	Number of times sea bass was eaten in the past 30 days.
DRD370U	Other unknown type eaten during the past 30 days.

References

1. Ojeda-Garcés, J.C.; Oviedo-García, E.; Salas, C. Odontologica ISSN 0120-971X, volumen 26 N. 1 primer semestre de 2013, Streptococcus mutans y caries dental, recibido: febrero de 2013. *Universidad CES* **2013**, *26*, 44–56.
2. Martínez Abreu, J.; Capote Femenias, J.; Bermúdez Ferrer, G.; Martínez García, Y. Determinantes sociales del estado de salud oral en el contexto actual. *MediSur* **2014**, *12*, 562–569.
3. Tamez, M.A.T.; de Mendoza, L.T.; Peña, E.G.R.; Martínez, P.C.C. Salud bucodental en escolares de estrato social bajo. *RESPYN* **2017**, *6*, 20–26.
4. Narváez Chávez, A.M. Asociación Entre el Conocimiento de los Padres Sobre Salud Bucal y uso de Técnicas Educativas con Relación a la Presencia de Biofilm y Caries en Infantes. Master's Thesis, Universidad Central del Ecuador, Quito, Ecuador, 2017.
5. Castañeda Abascal, I.E.; Lok Castañeda, A.; Molina, L.; Manuel, J. Prevalencia y factores pronósticos de caries dental en la población de 15 a 19 años. *Rev. Cuba. Estomatol.* **2015**, *52*, 21–29.
6. Martínez Abreu, J.; Castell-Florit Serrate, P.; Llanes Llanes, E.; Morales Aguiar, D.R.; Sánchez Barrera, O. Componente bucal y determinantes sociales en el análisis de la situación de salud. *Rev. Cuba. Estomatol.* **2015**, *52*, 53–61.

7. Gispert Abreu, E.d.l.Á.; Castell-Florit Serrate, P.; Herrera Nordet, M. Salud bucal poblacional y su producción intersectorial. *Rev. Cuba. Estomatol.* **2015**, *52*, 62–67.
8. Ospina, D.; Herrera, Y.; Betancur, J.; Agudelo, H.B.; López, A.P. Higiene bucal en la población de San Francisco Antioquia y sus factores relacionados. *Rev. Nac. Odontol.* **2016**, *12*, 23–30. [CrossRef]
9. Escalona, T.P.; Ortiz, H.R.C.; Palomino, Y.P.; Tamayo, M.I.; Rodríguez, M.I.R. 08-Relación entre factores de riesgos y caries dental. *MULTIMED Rev. Med. Granma* **2017**, *19*, 1–13.
10. Pepe, M.S.; Fan, J.; Feng, Z.; Gerds, T.; Hilden, J. The net reclassification index (NRI): A misleading measure of prediction improvement even with independent test data sets. *Stat. Biosci.* **2015**, *7*, 282–295. [CrossRef] [PubMed]
11. Lam, C.U.; Khin, L.; Kalhan, A.; Yee, R.; Lee, Y.; Chong, M.F.; Kwek, K.; Saw, S.; Godfrey, K.; Chong, Y.; et al. Identification of Caries Risk Determinants in Toddlers: Results of the GUSTO Birth Cohort Study. *Caries Res.* **2017**, *51*, 271–282. [CrossRef] [PubMed]
12. Fernandes, I.; Sá-Pinto, A.; Marques, L.S.; Ramos-Jorge, J.; Ramos-Jorge, M. Maternal identification of dental caries lesions in their children aged 1–3 years. *Eur. Arch. Paediatr. Dent.* **2017**, *18*, 197–202. [CrossRef] [PubMed]
13. Lips, A.; Antunes, L.S.; Antunes, L.A.; Pintor, A.V.B.; Santos, D.A.B.d.; Bachinski, R.; Küchler, E.C.; Alves, G.G. Salivary protein polymorphisms and risk of dental caries: A systematic review. *Braz. Oral Res.* **2017**, *31*. [CrossRef] [PubMed]
14. Ahmed, A.; Ikram, K.; Masood, H.; Urooj, M. Identification of relationship between oral disorders & hemodynamic parameters. *Pak. Oral Dent. J.* **2017**, *37*, 202–204.
15. National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. 2013–2014. Available online: http://www.cdc.gov/nchs/nhanes.htm (accessed on 17 November 2017).
16. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
17. Harrell, F.E., Jr. rms: Regression Modeling Strategies, R Package Version 5.1-1. 2017. Available online: https://CRAN.R-project.org/package=rms (accessed on 17 November 2017).
18. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002; ISBN 0-387-95457-0.
19. Cortez, P. rminer: Data Mining Classification and Regression Methods, R Package Version 1.4.2. 2016. Available online: https://CRAN.R-project.org/package=rminer (accessed on 17 November 2017).
20. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef] [PubMed]
21. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
22. Inoue, E. nricens: NRI for Risk Prediction Models with Time to Event and Binary Response Data, R Package Version 1.5. 2017. Available online: https://CRAN.R-project.org/package=nricens (accessed on 17 November 2017).
23. Berlanga Silvente, V.; Vilà Baños, R. Cómo obtener un modelo de regresión logística binaria con SPSS. *REIRE Rev. d'Innov. Recer. Educ.* **2014**, *7*, 105–118.
24. Lawless, J.; Singhal, K. Efficient screening of nonnormal regression models. *Biometrics* **1978**, *34*, 318–327. [CrossRef]
25. Nguyen, Y.; Nettleton, D.; Liu, H.; Tuggle, C.K. Detecting Differentially Expressed Genes with RNA-seq Data Using Backward Selection to Account for the Effects of Relevant Covariates. *J. Agric. Biol. Environ. Stat.* **2015**, *20*, 577–597. [CrossRef] [PubMed]
26. Derksen, S.; Keselman, H.J. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *Br. J. Math. Stat. Psychol.* **1992**, *45*, 265–282. [CrossRef]
27. Schousboe, J.T.; Langsetmo, L.; Taylor, B.C.; Ensrud, K.E. Fracture Risk Prediction Modeling and Statistics: What Should Clinical Researchers, Journal Reviewers, and Clinicians Know? *J. Clin. Densitom.* **2017**, *20*, 280–290. [CrossRef] [PubMed]
28. Szumilas, M. Explaining odds ratios. *J. Can. Acad. Child Adolesc. Psychiatry* **2010**, *19*, 227. [PubMed]
29. Lobo, J.M.; Jiménez-Valverde, A.; Real, R. AUC: A misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **2008**, *17*, 145–151. [CrossRef]

Redes neuronales profundas para el diagnóstico de caries utilizando características socioeconómicas y nutricionales como determinantes

Laura A. Zanella-Calzada[1], Carlos E. Galván-Tejada[1], Nubia M. Chávez-Lamas[2], María Del Carmen Gracia-Cortez[2], and Jorge I. Galván-Tejada[1]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, México,
{lzanellac, ericgalvan, gatejo}@uaz.edu.mx

[2] Unidad Académica de Odontología, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, México,
{nubiachavez, gacc005340}@uaz.edu.mx

Resumen La salud oral representa un componente esencial en la calidad de vida de las personas. Los padecimientos bucodentales se han tornado en uno de los principales problemas de salud pública, en donde la caries ocupa el primer lugar en incidencia. Este padecimiento ocurre dependiendo del consumo de ciertos elementos nutricionales interactuando de manera simultánea con diferentes factores, como los socioeconómicos. En este trabajo se realiza el análisis de determinantes dietéticos y demográficos, en relación con la situación oral de un conjunto de sujetos, específicamente con la presencia y ausencia / restauración de caries. La metodología se lleva cabo utilizando una red neuronal artificial (RNA) densa como herramienta de diagnóstico asistido por computadora, buscando un modelo generalizado que permita clasificar sujetos, y posteriormente ser evaluado a través de un análisis de validación cruzada. Los resultados obtenidos fueron estadísticamente significativos, obteniendo una precisión $\simeq 0.69$ y un AUC $\simeq 0.70$. Con base en los resultados es posible concluir que el modelo desarrollado es capaz de clasificar sujetos con ausencia de caries, de sujetos con presencia o restauración de las mismas, con un bajo nivel de error.

Abstract. Oral health represents an essential component in the quality of life of people. Oral diseases have become one of the main public health problems, where caries occupies the first place in incidence. This condition occurs depending on the consumption of certain nutritional elements interacting simultaneously with different factors, such as socioeconomic factors. In this work is performed the analysis of dietary and demographic determinants, in relation to the oral situation of a set of subjects, specifically with the presence and absence / restorations of

caries. The methodology is carried out using a dense artificial neural network (ANN) as a computer-aided diagnosis tool, looking for a generalized model that allows to classify subjects. This classification was later evaluated through an analysis based on cross-validation. The results obtained were statistically significant, obtaining a precision $\simeq 0.69$ and an AUC $\simeq 0.70$. Based on the results, it is possible to conclude that the model developed is able to classify subjects with absence of caries from subjects with presence or restorations, with high accuracy.

1. Introducción

Las enfermedades crónicas son los principales problemas de salud pública en la mayor parte del mundo. El patrón de enfermar se ha transformado y las enfermedades bucodentales son consideradas como uno de los principales problemas de salud pública debido a su alta incidencia y prevalencia en todas las regiones del mundo, y como en todas las enfermedades, la mayor carga es en las poblaciones desfavorecidas y marginadas socialmente. Esta característica representa un problema importante, ya que de acuerdo a la Organización Mundial de Salud (OMS), las enfermedades bucales son la cuarta causa más costosa de tratar [1].

La caries es el padecimiento más frecuente y según la Organización Mundial de la Salud (OMS) afecta entre un 60 % a 90 % de los niños en edad escolar entre 5 a 17 años. Además, el reporte de la Organización Panamericana de la Salud (OPS), describe que el desarrollo de dicho padecimiento depende de la frecuencia en el consumo de carbohidratos, las características de los alimentos, el tiempo de exposición, eliminación de la placa y la susceptibilidad del huésped, sumándose las escasas medidas preventivas en salud oral y la dificultad de hacer uso de servicios médico odontológicos especializados. Estos factores interactúan de manera simultánea, y sus variables corresponden a distintos órdenes desde procesos biológicos, hasta complejas estructuras histórico-culturales y las relaciones sociales, el nivel socioeconómico, el nivel educacional, entre otros, volviendo complejos los fenómenos de salud [2, 3].

En el presente estudio se pretende analizar los determinantes que afectan esta situación de salud oral con base en factores dietéticos y demográficos.

Debido a la dificultad que representa el control de la incidencia de la condición de caries por su alta prevalencia y la gran cantidad de factores que pueden influir en este estado, en la actualidad se han desarrollado algunos estudios que han implementado algoritmos y realizado análisis basados en diagnóstico asistido por computadora donde los modelos de predicción se utilizan cuando se requiere conocer a futuro el comportamiento de datos complejos altamente relacionados; éstos tienen utilidades en la odontología clínica e investigativa [4]

Tal es el caso de las redes neuronales artificiales (denominadas habitualmente como RNA o en inglés ANN) método utilizado para la predicción de los padecimientos bucales además de ser modelos matemáticos basados en un principio de aprendizaje que tiene su fundamento en los conceptos de inteligencia artificial

y la respuesta biológica del cerebro humano. Así mismo, están implicadas en procesos de construcción de sistemas que permiten clasificar, modelar y predecir información. Las RNA son modelos no lineales semiparamétricos, que permiten la integración de variables y manejan con facilidad grandes cantidades de datos comparados a los análisis lineales. Tienen unos elementos de procesamiento o neuronas, que son las unidades del sistema que pueden ser ajustadas o entrenadas a través de un proceso de aprendizaje y generalización. La información de cada una de estas neuronas es agrupada y procesada [5].

Con base en esta descripción general del problema de salud oral, específicamente de la caries, la principal contribución de este trabajo es determinar que dadas las características socioeconómicas y de dieta, el utilizar una red neuronal podrá determinar si el paciente presenta un estado de salud óptimo o con presencia de reparaciones o caries, de tal manera que se pueda tener una clasificación que permita saber las diferencias de las características que afectan a la población, además de buscar la predicción futura de la salud oral.

2. Materiales y Métodos

En esta sección se describen las bases de datos que fueron utilizadas para este trabajo, obtenidas de la National Health and Nutrition Examination Survey (NHANES 2013-2014), al igual que la descripción de los sujetos y los métodos utilizados para la clasificación automática de los mismos.

2.1. Descripción de los datos

NHANES es un programa de estudios diseñado para evaluar el estado de salud y nutricional de niños y adultos en EE. UU., fue fundada por Centers for Disease Control and Prevention (CDC) y National Center for Health Statistics (NCHS). Las encuestas realizadas por este programa son únicas, ya que combina entrevistas y exámenes físicos [6].

NHANES recolecta información de diferentes tipos de datos, y a su vez, estos datos se engloban en seis principales contextos; Demográfico, Dietético, Examinación, Laboratorio, Cuestionario y Acceso limitado.

En este trabajo fueron utilizadas las bases de datos de tres de los contextos mencionados:

- Demográfico: proporciona información individual, familiar y de nivel de hogar en diferentes temas (ingreso de hogares y familias, tamaños de hogares y familias, estado de embarazo, entre otros).
- Dietético: proporciona información detallada de la ingesta dietética, con el fin de estimar los tipos y las cantidades de alimentos y bebidas consumidos, además de estimar la ingesta de energía, nutrientes, y otros componentes alimenticios.
- Examinación: proporciona información del estado de salud, indicadores de riesgo de enfermedades, y acceso a servicios preventivos y de tratamiento, desde diferentes aspectos, entre ellos salud oral.

Las características analizadas en este trabajo fueron 189, de las cuales, 188 pertenecían a datos demográficos y dietéticos. Estas características fueron utilizadas como variables de entrada para la clasificación de los sujetos, mientras que la característica restante pertenecía a los datos de examinación, describiendo el estado de salud oral de los sujetos con base en la presencia / restauración o ausencia de caries, la cual fue utilizada como variable de salida.

En la Figura 1 se presenta el total de sujetos contenidos en las bases de datos utilizadas. De 9812 sujetos, 3690 eran sujetos control y 6122 eran sujetos caso. Los sujetos se encontraban en un rango de edad de 0 a 80 años y, 4830 pertenecían al género masculino y 4982 al género femenino.

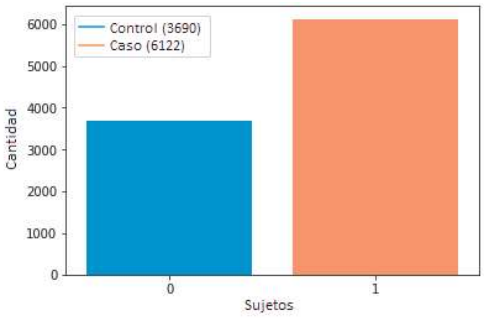

Figura 1. Gráfica de la cantidad de sujetos utilizados, clasificados de acuerdo a su estado de salud oral. Los sujetos contenidos en la barra con '0' presentaron ausencia de caries, mientras que los sujetos contenidos en la barra con '1' presentaron presencia / restauración de caries.

2.2. Análisis de los datos

El análisis de los datos de este trabajo se realizó a través de un enfoque multivariado, sometiendo a las características demográficas y dietéticas a una RNA profunda para la clasificación de sujetos de acuerdo a su salud oral y posteriormente, se llevó a cabo una validación estadística con el fin de evaluar los resultados obtenidos de la RNA desarrollada para estos datos específicos.

Preprocesamiento de los datos El preprocesamiento de los datos consistió inicialmente en eliminar de manera manual todas las características que presentaron un alto porcentaje de datos faltantes (mayores al 30 %). Después, de las características restantes se imputaron los datos faltantes que se presentaron en bajo porcentaje. Las columnas que contenían valores singulares también fueron eliminadas debido a que no proporcionaban información significativa.

Finalmente, la base de datos fue separada en dos conjuntos, seleccionando un conjunto destinado a entrenamiento contenido con el 70 % de los datos y un conjunto destinado a prueba contenido con el 30 % restante de los datos.

Clasificación de los datos Para la clasificación de los datos se utilizó una RNA recurrente densa, construida con el paquete *keras*.

Las redes neuronales artificiales buscan una solución a una tarea determinada, con base en la correlación de variables o características, a través de un aprendizaje o entrenamiento que asemeja el comportamiento de las redes neuronales biológicas. Por medio de diferentes capas compuestas por nodos o neuronas, las redes neuronales buscan un modelo de relación entre las variables de entrada y la variable de salida. La cantidad de nodos que existe por cada capa es configurable, al igual que la cantidad de capas y, dependiendo de los datos, regularmente existe mayor capacidad de la red entre mayor sea la cantidad de capas y de neuronas; sin embargo, si no se tiene la cantidad adecuada de estos elementos, se pueden provocar problemas de sobre ajuste.

Las redes presentan tres elementos principales, 1) un conjunto de sinapsis o conexiones, caracterizadas por un 'peso', en donde la señal de entrada se conecta a una neurona a través de su producto con el peso de esa conexión; 2) un sumador, que agrega las contribuciones de una señal ponderadas por todos los pesos; y 3) una función de activación, que equivale a una función de transferencia, la cual afecta a las neuronas, permitiendo limitar la amplitud de la salida de la red, brindando un rango permisible para la señal de salida en términos de valores finitos. Entre las funciones de activación más comunes se encuentran la función lineal, cuadrática, geométrica, logística, función de unidad lineal rectificada, entre otras.

Por otro lado, se dice que una RNA es densa cuando cada nodo de una capa tiene conexión con todos los nodos de la siguiente capa; mientras que el término de recurrencia hace referencia a que la red presenta al menos un lazo de retroalimentación, que brindará un impacto más profundo en la capacidad de aprendizaje de la red, al igual que en su desempeño, ya que permite optimizar su comportamiento, a través de un parámetro que da conocimiento del comportamiento de los datos y permite orientar el ajuste de la configuración de la red para mejorar su precisión.

Cabe mencionar que el número de iteraciones o épocas que la RNA tendrá para la optimización de su comportamiento con base en la recurrencia también es configurable y dependerá del tipo de datos. Para saber cuál es el número apto de épocas, se tienen algunos parámetros que permiten evaluar el comportamiento a través de cada iteración, como la función de pérdida y la precisión [7].

Con base en esto, *keras* presenta la ventaja de que permite diseñar la arquitectura de la RNA de acuerdo con el tipo de datos, configurando el tipo de RNA, cantidad de nodos, validación, función de pérdida, entre otros [8].

Evaluación Finalmente, para la evaluación de la clasificación que se obtuvo con la red neuronal, se midieron tres parámetros de validación diferentes; la función de pérdida, la precisión y la curva ROC.

Las redes neuronales son entrenadas en su mayoría usando métodos de gradiente a través de un proceso iterativo del descenso de la función de pérdida. Una pérdida es diseñada para tener la propiedad principal de que entre menor sea su valor, mejor será el modelo que se ajuste a sus datos y además, será diferenciable, lo que permitirá optimizar la red de manera directa, dando información de la capacidad del sistema. Por lo tanto, la función de pérdida se basa en la búsqueda del mínimo global, que corresponde al mínimo error, basado en un factor de aprendizaje. Algunas de las técnicas más comunes para calcular la función de pérdida son el error cuadrático medio, error absoluto medio, entropía cruzada binaria, Poisson, entre otras [9].

Por otro lado, la precisión permite medir el comportamiento de la RNA a través de una función no diferenciable, es decir, esta métrica no permite optimizar la red; sin embargo, si permite seleccionar el modelo que muestre el comportamiento más apto en el entrenamiento de la red, y se basa en el cálculo del promedio de las diferencias que existen entre la clasificación calculada por el modelo de la RNA y la clasificación real de los datos, como se muestra en la Ecuación 1, en donde la precisión está representada como 1 - $error$. V_{pred} hace referencia al valor de clasificación que se calculó por la red neuronal, mientras que V_{real} hace referencia al valor de clasificación verdadero [10].

$$error = V_{pred} - V_{real} \tag{1}$$

Finalmente, la curva ROC es un método estándar que permite evaluar la precisión con la que el modelo clasifica, basado en la relación entre la sensibilidad y la especificidad. La sensibilidad es definida como la proporción de sujetos con una condición que fueron clasificados como positivos, es decir, los valores predictivos positivos (VPP); este valor es calculado con la Ecuación 2, donde VP representa la cantidad de verdaderos positivos y FP representa la cantidad de falsos positivos.

$$VPP = \frac{VP}{VP + FP} \tag{2}$$

La especificidad es definida como la proporción de sujetos sin una condición que fueron clasificados como negativos, es decir, los valores predictivos negativos (VPN); este valor es calculado con la Ecuación 3, donde VN representa la cantidad de verdaderos negativos y FN representa la cantidad de falsos negativos [11].

$$VPN = \frac{VN}{VN + FN} \tag{3}$$

El análisis de este trabajo fue realizado en *Python* (versión 3.6), utilizando los paquetes *keras* (versión 2.1.5) [12], *scipy* (versión 1.0.0) [13], *pandas* (versión 0.22.0) y *sklearn* (versión 0.19.1) [14].

3. Experimentación

En esta sección se presenta el proceso que se llevó a cabo para la clasificación de sujetos con presencia / ausencia de caries, con base en la información obtenida de una serie de características socioeconómicas y nutricionales, obtenidas de las bases de datos públicas de NHANES 2013-2014.

En la Figura 2 se presenta el diagrama de flujo de los pasos seguidos para la experimentación de este trabajo. La sección A) representa el preprocesamiento de los datos, en donde se hace una reducción de características debido a la alta cantidad de datos faltantes y valores singulares que se presentan, además de separar los datos en dos conjuntos, uno para entrenamiento y uno para prueba. La sección B) hace referencia a la clasificación de los datos por medio de una RNA que se diseñó con el fin de modelar la clasificación de sujetos de acuerdo con su estado de salud oral. Finalmente, en C) se presenta la evaluación del comportamiento de la RNA, con el fin de conocer la precisión con la que el modelado clasifica a los sujetos.

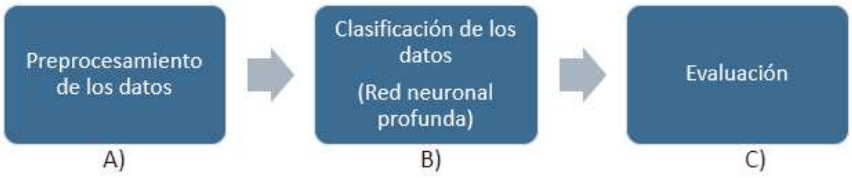

Figura 2. Diagrama de flujo de la metodología seguida.

3.1. Preprocesamiento de los datos

El preprocesamiento de los datos consistió inicialmente en eliminar de manera manual todas las características que presentaron un alto porcentaje de datos faltantes ($> 30\,\%$), representados como *Not a Number* (NaN). Después, de las características restantes, las que presentaron un bajo porcentaje de datos faltantes ($\leq 30\,\%$) fueron imputadas con la función $rfimput$ del paquete para R, $randomForest$ (versión 4.6-12, 2015-10-06) [15], que consiste en reemplazar todos los NaN con la media de los valores que presenta la columna donde se localiza el dato faltante.

Las columnas que contenían valores singulares también fueron eliminadas debido a que no proporcionaban información significativa. Dos valores son singulares si tienen valores múltiples entre ellas o si presentan el mismo valor en toda la columna.

Finalmente, la base de datos fue separada en dos conjuntos, seleccionando de manera aleatoria balanceada el 70 % de los datos para un conjunto destinado a entrenamiento y el 30 % restante destinado a un conjunto de prueba.

3.2. Clasificación de los datos

Se hizo la clasificación de los sujetos de acuerdo con la presencia / ausencia de caries a través de una RNA que se diseñó con base en los datos de los sujetos, utilizando el paquete *keras*.

keras es una interfaz de programación de aplicaciones de redes neuronales de alto nivel, escrita en *Python*. Fue desarrollada con el enfoque de permitir una experimentación rápida, permitiendo la creación de prototipos de manera sencilla [12].

La RNA que se diseñó estuvo compuesta por una serie de capas ocultas densas y de pérdida, como se muestra en el diagrama de la Figura 3. A la capa de entrada (A) se le asignaron 104 neuronas, haciendo referencia a las 104 características del conjunto de datos. La primera capa oculta densa (B) estuvo formada por 100 neuronas, al igual que la primera capa oculta de pérdida (C), además de tener un porcentaje de "pérdida"de los datos del 50 %. A la segunda capa oculta densa (D) se le asignaron 1000 neuronas, al igual que a la segunda capa oculta de pérdida (E), con un porcentaje de "pérdida"del 25 %. La tercera capa oculta densa (F) estuvo compuesta por 100 neuronas y la tercera oculta de pérdida (G) también, con un porcentaje de "pérdida"del 50 %. Finalmente, la cuarta capa oculta densa (H) o capa de salida se caracterizó por dos neuronas, haciendo referencia a las dos salidas o clasificaciones posibles.

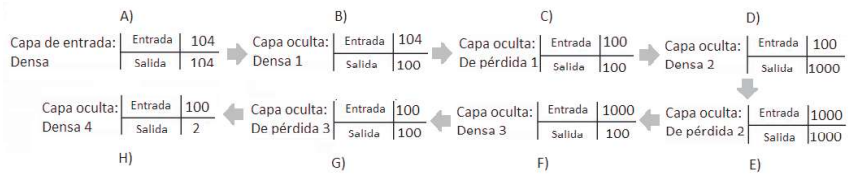

Figura 3. Diagrama del diseño de la RNA construída.

La capa de entrada y las capas ocultas densas, utilizaron como función de activación la función de unidad lineal rectificada ($RELU$, por sus siglas en inglés), que consiste en asignar 0 a los valores de las neuronas < 0, y respetar el valor de las neuronas cuando ≥ 0, como se muestra en la Ecuación 4 [16].

$$RELU(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{4}$$

La última capa oculta densa utilizó como función de activación la función exponencial normalizada o $softmax$, que representa una forma general de la función logística y se emplea para comprimir un vector de valores arbitrarios, en un vector de valores reales en el rango [0, 1]. En la Ecuación 5 se muestra esta función, en donde $\sigma(z)$ representa el vector K-dimensional, z, de valores binarios [17].

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, j = 1, ..., K \tag{5}$$

Finalmente, las capas de pérdida fueron incluidas con el fin de evitar problemas de sobre ajuste, como se mencionó anteriormente, y su funcionamiento se basa en asignar a un porcentaje de los datos el valor de cero y así no puedan ser tomados en cuenta para la clasificación de los sujetos en esa capa. El subconjunto de datos seleccionado por estas capas va cambiando con cada iteración, de manera que en cada época la clasificación se hará omitiendo un porcentaje de datos diferente.

Por otro lado, debido a que la RNA era recurrente, fue posible optimizar su comportamiento. Se eligió el algoritmo de optimización *Adam*, que basa su funcionamiento en los algoritmos de descenso de gradiente estocástico, haciendo uso del promedio del primer y segundo momento de los gradientes para adaptar la tasa del parámetro de aprendizaje. Específicamente, *Adam* calcula el promedio móvil exponencial del gradiente y del gradiente cuadrado, y controla las tasas de desintegración de ese promedio móvil [18].

Cabe mencionar que la RNA se entrenó con un número de épocas de 100.

3.3. Evaluación

Con el fin de validar los resultados de clasificación obtenidos por la RNA se evaluaron tres parámetros; la función de pérdida, la precisión y la curva ROC. La función de pérdida y la precisión se calcularon en cada una de las épocas, permitiendo saber si el comportamiento de la RNA iba mejorando; mientras que la curva ROC se calculó con base en el promedio del comportamiento general de la RNA.

La función de pérdida que se utilizó está preestablecida en *keras* como $binary - crossentropy$ y calcula el valor de la entropía cruzada para problemas de clasificación binario. Este método utiliza la distancia Kullback-Leibler, que es una medida entre dos funciones de densidad g y h, conocida como la entropía cruzada entre g y h, como se muestra en la Ecuación 6. Su funcionamiento se basa en iteraciones, generando un conjunto aleatorio de valores estimando el valor que se quiere obtener y posteriormente, actualizando los parámetros para generar valores $mejores$ o más aproximados en términos de la distancia $Kullback - Leiber$, en la siguiente iteración.

$$D(g,h) = \int_g (x) ln \frac{g(x)}{h(x)} \mu(dx) = \int_g (x) lng(x) \mu(dx) - \int_g (x) lnh(x) \mu(dx) \tag{6}$$

La precisión se calculó con la función $binary - accuracy$, que calcula la tasa de precisión media a través de todas las predicciones para problemas de clasificación binaria.

4. Resultados

Las dos bases de datos utilizadas para este trabajo tenían un contenido de 188 características demográficas y dietéticas en total, además de utilizarse una característica que contenía el estado de salud oral como variable de salida para la clasificación de los datos.

Después del preprocesamiento de los datos, se conservaron únicamente 105 características en total, de las cuales 25 pertenecían a características demográficas y 79 a características dietéticas, la característica restante hace referencia a la variable de salida.

Finalmente, del conjunto total de 9812 sujetos, 3690 pertenecían a sujetos control y 6122 pertenecían a sujetos caso. Este conjunto se dividió en dos subconjuntos, uno para entrenamiento con el 70 % de los datos, contenido por 2596 sujetos control y 4272 sujetos caso, y uno para prueba con el 30 % restante de los datos, contenido por 1094 sujetos control y 1850 sujetos caso. Estos datos se pueden observar gráficamente en la Figura 4.

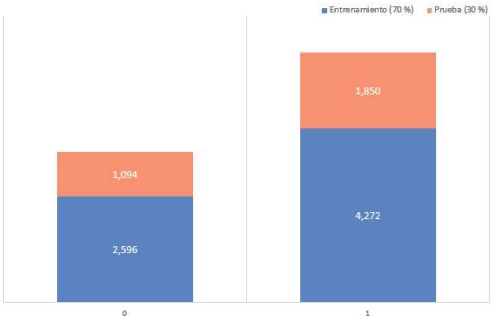

Figura 4. Gráfica de la cantidad de sujetos pertenecientes a cada conjunto de datos, entrenamiento y prueba. Los contenidos en la barra con '0' presentaron ausencia de caries, mientras que los sujetos contenidos en la barra con '1' presentaron presencia / restauración de caries.

El conjunto de datos que fue designado para entrenamiento se usó para el aprendizaje de la red, el cual fue evaluado en cada época a través de la precisión y la función de pérdida. La gráfica del comportamiento de la precisión se presenta en la Figura 5 A), donde la línea azul representa el comportamiento de los datos de entrenamiento, obteniendo una precisión final de 0.69; mientras que la línea naranja representa el comportamiento de los datos de prueba, obteniendo una precisión final de 0.68.

La gráfica del comportamiento de la función de pérdida se presenta en la Figura 5 B), donde la línea azul representa el comportamiento de los datos de entrenamiento, obteniendo un valor final de 0.58; mientras que la línea naranja

representa el comportamiento de los datos de prueba, obteniendo un valor final de 0.60.

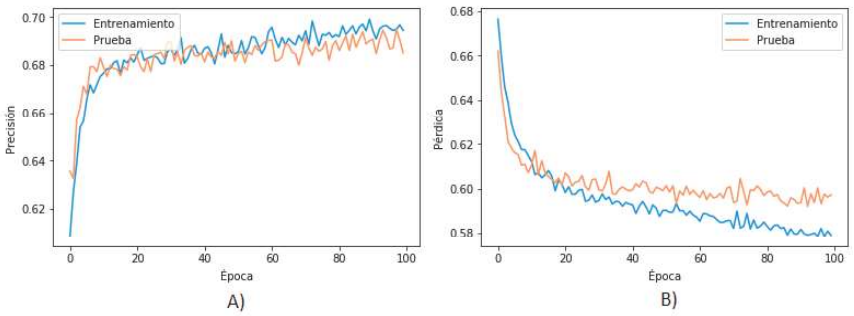

Figura 5. Gráficas del comportamiento de la precisión (A) y de la función de pérdida (B) de la RNA en cada época.

En la gráfica de la Figura 6 se presentan las curvas del comportamiento medio de la RNA, en donde la línea rosa hace referencia a la proporción de sensibilidad y especificidad para la clase '0' o los sujetos con ausencia de caries, obteniendo un AUC de 0.69. La línea azul claro hace referencia a la proporción de sensibilidad y especificidad para la clase '1' o los sujetos con presencia / restauración de caries, obteniendo un AUC de 0.69. La línea punteada naranja hace referencia a la curva calculada del micro-promedio de la proporción de sensibilidad y especificidad para la clasificación de los sujetos, obteniendo un AUC de 0.75; y la línea punteada azul oscuro hace referencia a la curva calculada del macro-promedio, obteniendo un AUC de 0.69.

En el micro-promedio se hace la suma de los verdaderos positivos, falsos positivos y falsos negativos del sistema para diferentes conjuntos, como se muestra en la Ecuación 7, en donde VP_1 se refiere a los verdaderos positivos del conjunto uno, VP_2 se refiere a los verdaderos positivos del conjunto dos, FP_1 se refiere a los falsos positivos del conjunto uno y FP_2 se refiere a los falsos positivos del conjunto dos.

$$micro - promedio = \frac{VP_1 + VP_2}{VP_1 + VP_2 + FP_1 + FP_2} \tag{7}$$

Por otro lado, el macro-promedio es un método directo; toma la media de la precisión del sistema en diferentes conjuntos. Se calcula con la Ecuación 8, en donde P_1 se refiere al promedio del conjunto uno y P_2 se refiere al promedio del conjunto dos.

$$macro - promedio = \frac{P_1 + P_2}{2} \tag{8}$$

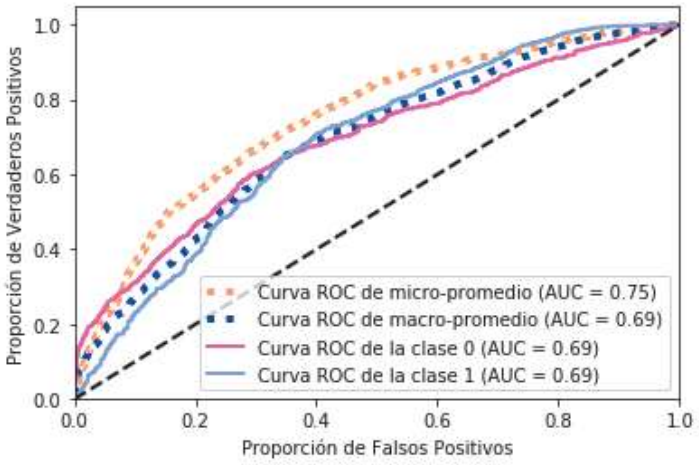

Figura 6. Gráfica de las curvas ROC generadas con base en el comportamiento promedio de la RNA en la clasificaión de sujetos.

5. Discusión y Conclusiones

De los dos conjuntos de datos que se obtuvieron a través del paso de preprocesamiento, es posible observar que el conjunto de sujetos caso contenía mayor cantidad de datos que el conjunto de sujetos control, como se observó en la Figura 4; esto se debió a que los sujetos caso englobaron a los sujetos con presencia de caries y con restauraciones.

La RNA densa, recurrente, fue entrenada durante 100 épocas con el conjunto de datos de entrenamiento y evaluada en cada una de las épocas con el conjunto de datos de prueba. Se utilizó el algoritmo de optimización *Adam* para la mejora del comportamiento de la RNA en la retroalimentación y en la validación se usaron la precisión y la función de pérdida, calculada con entropía cruzada, para saber cómo se va modificando el comportamiento de la RNA. La validación se realizó en ambos conjuntos de datos, de entrenamiento y de prueba, con el fin de asegurar que no hubiera sobre ajuste en la clasificación de los sujetos; es decir, si el comportamiento de estos parámetros es similar para ambos conjuntos de datos, significa que la RNA está clasificando con base en un modelo generalizado, ya que logra clasificar con una precisión y una función de pérdida parecida en el conjunto de datos con la que se entrenó y en un conjunto de datos "desconocido". En la gráfica de la Figura 5 A) se observa que el comportamiento de la precisión para ambos conjuntos es similar, siendo ligeramente superior para los datos de entrenamiento, comportamiento que es normal ya que son los datos en los que se está basando el modelado de la RNA. Al ser este comportamiento similar, es posible saber que el modelo de la RNA se generalizó a través del

aprendizaje al poder clasificar datos con una precisión similar a la que obtuvo con los datos que fue entrenada. Por otro lado, en la gráfica de la Figura 5 B) se observa que el comportamiento de la función de pérdida también es similar para ambos conjuntos de datos. Es notable que para el conjunto de datos de entrenamiento este parámetro es menor en comparación que para el conjunto de datos de prueba; sin embargo, al igual que en el caso de la precisión, resulta normal debido a que fueron los datos usados para el aprendizaje. También, al tener un comportamiento similar permite saber que el modelo no tiene problemas de sobre ajuste y que logra clasificar datos "desconocidos"de manera parecida a como clasifica datos çonocidos".

Finalmente, en la Figura 6 se puede observar que todas las curvas presentaron valores estadísticamente significativos, obteniendo valores de AUC ≥ 0.69. Estas curvas son generadas con base en la proporción de verdaderos positivos y verdaderos negativos, en donde para la clase 0, la clase 1 y el macro-promedio, se calculó un AUC de 0.69, lo que implica que el 69 % de los sujetos fueron clasificados de manera adecuada para cada una de las clases y para el promedio general de las mismas; mientras que para la curva generada para el micro-promedio, se calculó un AUC de 0.75, haciendo referencia a que el 75 % de los sujetos fueron clasificados de manera adecuada cuando la proporción fue medida por subconjuntos de datos.

Por lo tanto, de acuerdo con los resultados obtenidos, es posible concluir que la cantidad de datos que se utilizaron para modelar la clasificación de los sujetos fue apta para que la RNA tuviera un aprendizaje generalizado, basando esto en que la clasificación de los datos de prueba tuvo un comportamiento similar al obtenido en el aprendizaje. Este comportamiento se observó en las gráficas de precisión y de función de pérdida, permitiendo ver que la cantidad de épocas con la que se entrenó la RNA fueron suficientes para que estos parámetros tuvieran un comportamiento estable en ambos conjuntos de datos. La precisión alcanzada con este modelado fue de alrededor de 0.70, valor que es estadísticamente significativo ya que implica que el 70 % de las veces se clasificará de manera correcta a los sujetos con presencia o con ausencia / restauración de caries. La función de pérdida disminuyó en aproximadamente un 10 % desde que inició el entrenamiento y comenzó a tener cambios menores a lo largo de las épocas, implicando una aproximación al mínimo buscado. Como evaluación final, todas las curvas ROC tuvieron un AUC estadísticamente significativo, implicando que alrededor del 70 % de las veces, los sujetos fueron clasificados correctamente en proporción de verdaderos positivos y verdaderos negativos, valor que corrobora el resultado obtenido con la precisión. Con base en esto, es posible clasificar sujetos con ausencia de caries de sujetos con presencia / restauración, a través de datos demográficos y dietéticos, con una precisión estadísticamente significativa.

Referencias

1. Daniel Ridao Marín. Desarrollo de un sistema de ayuda a la decisión para tratamientos odontológicos con imágenes digitales. 2017.

2. Miguel Espinoza Solano and Roberto Antonio León-Manco. Prevalencia y experiencia de caries dental en estudiantes según facultades de una universidad particular peruana. *Revista Estomatológica Herediana*, 25(3):187–193, 2015.
3. Laura Daniela Acuña Aguilar, Daniela Porras Cerón, Laura Daniela Ríos Rueda, et al. Prevalencia de lesiones cariosas y factores asociados presentes en pacientes con síndrome de down en las fundaciones fundown y san luis guanella de bucaramanga. 2018.
4. Estela de los Ángeles Gispert Abreu, Pastor Castell-Florit Serrate, and Mirtha Herrera Nordet. Salud bucal poblacional y su producción intersectorial. *Revista Cubana de Estomatología*, 52:62–67, 2015.
5. Tania Camila Niño Sandoval, Sonia Victoria Guevara Pérez, Fabio Augusto González, Robinson Andrés Jaque, and Clementina Infante Contreras. Uso de redes neuronales artificiales en predicción de morfología mandibular a través de variables craneomaxilares en una vista posteroanterior. *Universitas Odontológica*, 35(74), 2016.
6. Centers for Disease Control, Prevention, et al. National health and nutrition examination survey, 2011-2012, 2013.
7. Simon S Haykin, Simon S Haykin, Simon S Haykin, and Simon S Haykin. *Neural networks and learning machines*, volume 3. Pearson Upper Saddle River, NJ, USA:, 2009.
8. Francois Chollet. *Deep learning with Python*. Manning Publications Co., 2017.
9. Carlos Antona Cortés. Herramientas modernas en redes neuronales: la librería keras. B.S. thesis, 2017.
10. Maxwell Nye and Andrew Saxe. Are efficient deep representations learnable? 2018.
11. Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.
12. François Chollet et al. Keras: Deep learning library for theano and tensorflow. *URL: https://keras. io/k*, 7:8, 2015.
13. Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed ¡today¿].
14. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
15. Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
16. Alessio Lomuscio and Lalit Maganti. An approach to reachability analysis for feed-forward relu neural networks. *arXiv preprint arXiv:1706.07351*, 2017.
17. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.
18. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

bioengineering

Article

Deep Artificial Neural Networks for the Diagnostic of Caries Using Socioeconomic and Nutritional Features as Determinants: Data from NHANES 2013–2014

Laura A. Zanella-Calzada [1,†], Carlos E. Galván-Tejada [1,*,†], Nubia M. Chávez-Lamas [2], Jesús Rivas-Gutierrez [2], Rafael Magallanes-Quintanar [1], Jose M. Celaya-Padilla [3], Jorge I. Galván-Tejada [1] and Hamurabi Gamboa-Rosales [1]

1 Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, México; lzanellac@uaz.edu.mx (L.A.Z.-C.); tiquis@uaz.edu.mx (R.M.-Q.); gatejo@uaz.edu.mx (J.I.G.-T.); hamurabigr@uaz.edu.mx (H.G.-R.)
2 Unidad Académica de Odontología, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Zac, México; nubiachavez@uaz.edu.mx (N.M.C.-L.); rigj002959@uaz.edu.mx (J.R.-G.)
3 CONACYT—Universidad Autónoma de Zacatecas—Jardín Juarez 147, Centro, Zacatecas 98000, Zac, Mexico; jose.celaya@uaz.edu.mx
* Correspondence: ericgalvan@uaz.edu.mx; Tel.: +52-492-544-0968
† These authors contributed equally to this work.

Received: 31 May 2018; Accepted: 15 June 2018; Published: 18 June 2018

Abstract: Oral health represents an essential component in the quality of life of people, being a determinant factor in general health since it may affect the risk of suffering other conditions, such as chronic diseases. Oral diseases have become one of the main public health problems, where dental caries is the condition that most affects oral health worldwide, occurring in about 90% of the global population. This condition has been considered a challenge because of its high prevalence, besides being a chronic but preventable disease which can be caused depending on the consumption of certain nutritional elements interacting simultaneously with different factors, such as socioeconomic factors. Based on this problem, an analysis of a set of 189 dietary and demographic determinants is performed in this work, in order to find the relationship between these factors and the oral situation of a set of subjects. The oral situation refers to the presence and absence/restorations of caries. The methodology is performed constructing a dense artificial neural network (ANN), as a computer-aided diagnosis tool, looking for a generalized model that allows for classifying subjects. As validation, the classification model was evaluated through a statistical analysis based on a cross validation, calculating the accuracy, loss function, receiving operating characteristic (ROC) curve and area under the curve (AUC) parameters. The results obtained were statistically significant, obtaining an accuracy $\simeq 0.69$ and AUC values of 0.69 and 0.75. Based on these results, it is possible to conclude that the classification model developed through the deep ANN is able to classify subjects with absence of caries from subjects with presence or restorations with high accuracy, according to their demographic and dietary factors.

Keywords: NHANES; oral health; dental caries; classification multivariate models; computer-aided diagnosis; artificial neural networks; deep learning; statistical analysis

1. Introduction

Chronic diseases are the main problems in public health worldwide. The pattern of disease has been transformed and oral diseases are considered one of the main public health problems due to its high incidence and prevalence in all regions of the world, and, as in all diseases, the greatest burden

is on populations disadvantaged and socially marginalized; treatment of the conditions is extremely expensive and is not feasible in most low and middle income countries. This characteristic represents an important problem, since, according to the World Health Organization (WHO), oral diseases are the fourth most expensive cause to treat in the most industrialized countries [1].

Oral health is an essential component for quality of life of people due to its influence as a determinant factor in general health of individuals and communities becoming a relevant point in health care. Therefore, the serious repercussions in terms of pain and suffering, impairment of function and effect on quality of life should also be considered. An important aspect to consider is that oral diseases increase the risk of chronic diseases, such as: cardiovascular and cerebrovascular, diabetes mellitus and respiratory. On the other hand, epidemiological surveillance of oral diseases becomes important insofar as it provides useful elements for the planning, programming, organization, integration, control and direction of the oral health program, at the same time that it guides the attention to the population [2].

Caries is the most frequent condition and according to the WHO; it affects between 60% and 90% of children of school age between 5 to 17 years old. The determinants and conditions of oral health status are multicausal, multisectoral and interdisciplinary categories that encompass a series of situations related to the historical and political process that each country experiences. In addition, the report of the Organización Panamericana de la Salud (OPS) describes how the development of this condition depends on the frequency of carbohydrate consumption, the characteristics of the food, the time exposure, plaque removal and susceptibility of the guest, adding the few preventive measures in oral health and the difficulty of making use of specialized dental medical services. These factors interact simultaneously, and their variables correspond to different orders from biological processes, to complex historical-cultural structures and social relationships, socioeconomic level, educational level, among others, making health phenomena complex [3,4].

There are too many risk factors and important determinants related to the presence of caries, and it is clear that the greater the degree of risk exposure, the greater the probability of contracting or developing it. Due to the difficulty of representing the control of the incidence of caries due to its high prevalence and the large number of factors that can influence this state, at present, some studies have implemented algorithms and performed analyses based on computer-aided diagnosis (CAD) where prediction models are used when it is necessary to know in the future the behavior of highly related complex data; these have utilities in clinical and research dentistry [5].

An example of these algorithms are the artificial neural networks (ANNs), a method used for the prediction of diseases, such is the case of oral diseases, as well as being mathematic models based on a principle of learning that is based on the concepts of artificial intelligence and the biological response of the human brain. Likewise, they are involved in systems' construction processes that allow classifying, modeling and predicting information. ANNs are semiparametric nonlinear models, which allow the integration of variables and easily handle large amounts of data compared to linear analyses. They have processing elements or neurons, which are the units of the system that can be adjusted or trained through a process of learning and generalization. The information of each of these neurons is grouped and processed [6].

According to the literature, dental caries incidence has been a problem that has been studied by a series of researchers, trying to identify caries risk determinants. Lam et al. [7] found that dental visits, brushing frequency, lower parental perceived importance of baby teeth, and weaning onto solids are determinants associated with plaque accumulation, validating these results specifically in children. In the work of Fernandes et al. [8], a cross-sectional study of 274 children and their mothers based on demographic/socio-economic status is presented; through a Poisson regression approach, an analysis was performed, obtaining that dental caries can be mainly found in mothers of children aged 1, demonstrating the relationship between demographic/socio-economic status and caries.

As mentioned in the text, there are different types of risk factors that are associated with genetic and nutritional data. Lips et al. [9] presented a work where the association between genetic polymorphisms and risk of dental caries is demonstrated for most of the salivary proteins.

In addition, diabetic and hypertensive patients have dietary risk factors that can be caused by oral health problems. This affirmation was demonstrated in the work of Asif Ahmed et al. [10] through a statistical analysis using clinical data, in order to identify that patients with oro-dental problems were hemodynamically stressed.

In the present study, it is intended to analyze the determinants that affect this oral health situation based on dietary and demographic factors. Based on this general description of the oral health problem, specifically caries, the main contribution of this work is to determine that, given the socioeconomic and dietary features, using an ANN can determine if the patient presents an optimal state of health or with the presence of repairs or caries, obtaining a classification system that allows for knowing the differences in the features that affect the population, as well as looking for the future prediction of oral health.

This method is an important tool for human resources and dental services because it reflects information collected in the population groups studied, which could unveil that oral health problems are not simple conditions, but they are obtained from different processes that have their trigger, which can enhance or mitigate them, impacting positively or negatively on the state of oral and general health.

Related Work

The condition of dental caries has been described in the scientific literature under different terms, as a multifactorial disease that is characterized by localized and progressive demineralization of the inorganic portions of the tooth and the subsequent deterioration of its organic part, with a high degree of morbidity and high prevalence [11–13].

A related work that suggests CADx to improve oral health prediction is developed by Zhang et al. [14], where an autoregressive integrated moving average model is performed and a grey predictive model for the estimation of the national prevalence of early childhood caries from 2014 to 2018, finding that the highest annual prevalence would be of 55.8%, being attributed to the socioeconomic developments and the public health service. In the work of Asif et al. [15], a caries preventive tool based on the influence of genetic pattern using the frequency of occurrence of fingerprint patterns among children, using as a validation measure the p-value, finding that dermatoglyphics could be an appropriate method to explore the possibility of a noninvasive and early predictor for dental caries. Chapple et al. [16] performed a systematic appraisal to identify potential risk factors for caries and periodontal diseases using genetic, role of diet and nutrition risk factors, obtained as results that the genetic contribution to these conditions present an attributable risk up to 50%, while controlled diabetes and obesity are common acquired factors. Hayes et al. [17] had as an objective to determine the risk indicators associated with root caries conducting a prospective longitudinal study using a regression analysis with data related to hygiene habits, diet, smoking habits and education level and as the outcome the root caries experience, suggesting a correlation between root caries and the variables poor plaque control, xerostomia, coronal decay and exposed root surfaces, obtaining a prevention tool for the root caries condition. In the work of Sudhir et al. [18], evaluating Caries Management by Risk Assessment (CAMBRA) as a tool for caries risk prediction was proposed, using as validation parameter OR value and ROC curve, obtaining that CAMBRA was found to be 47.62% with a specificity of 80% and AUC was found to be 0.638, which means that it is valid and highly predictive in determining the caries risk. Finally, Twetman [19] proposed to summarise the findings of recent systematic reviews covering caries risk assessment, finding that caries risk assessment should be carried out at the subject's first dental visit and reassessments should be done during childhood; multivariate models display a better accuracy than the use of single predictors; there is no clearly superior method to predict future caries and no evidence to support the use of one model, program, or technology before the other; and the risk category should be linked to appropriate preventive care with recall intervals based on the individual need.

This work is organized as follows: Section 2 briefly describes the materials and methods used for the development of the data analysis, including data preprocessing, data classification and data evaluation. Section 3 exposes the results obtained. Finally, results are discussed in Section 4 and concluded in Section 5.

2. Materials and Methods

In this section, the process that was carried on for the classification between subjects with presence of caries and subjects with absence was presented, based on the information obtained from a series of socioeconomic and nutritional features. The databases that were used in this work were obtained from the National Health and Nutrition Examination Survey (NHANES, 2013–2014). In addition, the description of the subjects and the methods used for their classification are presented.

In Figure 1, the flowchart of the steps followed for the experimentation of this work is presented. Section (A) presents the data acquisition of the public databases "Demographic" and "Dietary" from NHANES 2013–2014. Section (B) represents the data preprocessing, where a feature reduction is performed due to the high amount of missing data and singular values that are presented, in addition to separating the data in two sets, one for training and one for testing. Section (C) refers to the data classification of subjects according to their oral health status. Finally, in section (D), the evaluation of the ANN performance is presented, in order to know the accuracy with which the model classifies the subjects.

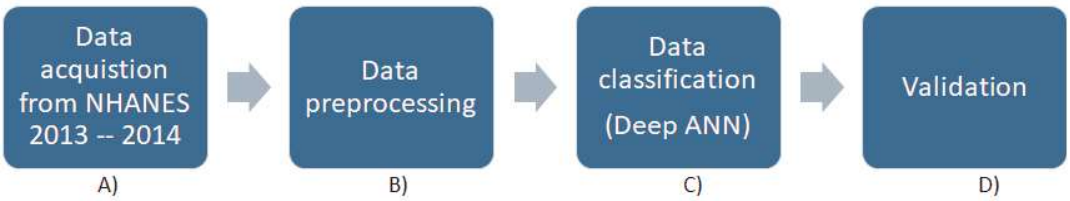

Figure 1. Flowchart of the methodology followed.

2.1. Features Description

NHANES is a program of studies designed to evaluate the health and nutritional status of children and adults in United States, and it was founded by Centers of Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS). The surveys conducted by this program are unique, since it combines interviews and physical exams [20].

NHANES collects information from different types of data, and, in turn, this information is included in six main contexts; demographic, dietetic, examination, laboratory, questionnaire and limited access. These contexts are contained by the information described as follows:

- Demographic: it provides individual, family and household level information in different topics (income of households and families, size of households and families, pregnancy status, among others).
- Dietetic: it provides detailed information on dietary intake, in order to estimate the types and amount of food and beverages consumed, in addition to estimating the intake of energy, nutrients, and other food components.
- Examination: it provides information of the health status, indicators of disease risk, and access to preventive and treatment services, from different aspects, including oral health.
- Laboratory: it provides information of the results obtained from laboratory analysis of different components (components of urine, proteins, triglycerides, plasma, among others).
- Questionnaire: it provides information of the data obtained from the interviews conducted through a system of computer-aided personal interview, of different topics (alcohol use, cardiovascular health, dermatology, among others).

- Limited access: it provides similar information to that found in the questionnaire; however, it isn't publicly available.

There were 189 features analyzed for this work, of which 188 belonged to demographic data, described in Appendix A, and dietary data, described in Appendix B. These features were used as input variables for the classification of subjects, while the remaining feature belonged to examination data, describing the oral health status of the subjects based on the presence/restoration or absence of caries, for which it was used as an output feature.

2.2. Subjects Description

The subjects of these databases were submitted to a series of different questionnaires, related to the different features. These subjects belong to different counties in the USA and they were randomly selected with a computer algorithm by NHANES.

Figure 2 presents a flowchart of the randomly selection process followed by NHANES. (A) presents the first stage that corresponds to the sampling of counties. In this step, all the counties are divided into 15 groups according to their characteristics. Then, from each group, one county is selected, forming the 15 counties in the NHANES surveys for each year. (B) presents the second stage, corresponding to the sampling of segments, where from each county smaller groups are formed (with a large number of households in each group), selecting between 20 and 24 of these groups. (C) presented the third stage, which is referred to the sampling of households, where from all of the houses that correspond to the groups that were selected in the past stage, a sample of about 30 households are selected within each group. Finally, (D) presents the fourth stage, corresponding to the sampling of persons, where the NHANES interviewers go to the selected households and ask for the information of the surveys (age, race, and gender). The members of the households that responded to the surveys were randomly selected through a computer algorithm; they can be selected some, all, or none of the household members.

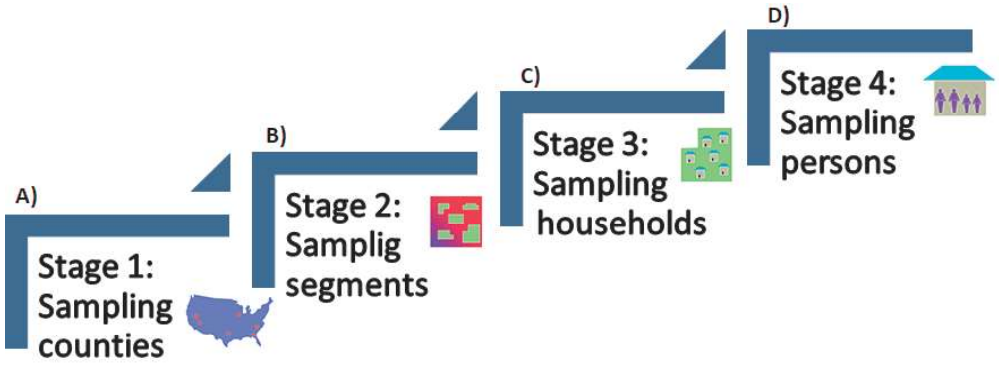

Figure 2. Flowchart of the stages followed by NHANES 2013–2014 for the sampling of participants.

The main target population for NHANES is the non-institutionalized civilian resident population of the USA. The design of the population selection is based on the sampling of a larger number of specific subgroups that present particular public health characteristics, in order to increase the reliability and precision of health status indicators. NHANES started these design changes in 2011, including in its population the oversampled subgroups survey cycle:

- Hispanic persons;
- Non-Hispanic black persons;
- Non-Hispanic Asian persons;
- Non-Hispanic white and other* persons at or below 130% of the poverty level; and
- Non-Hispanic white and other* persons aged 80 years and older.

This random selection of subjects avoid presenting any bias problem, favoring the important variety of demography characteristics of the subjects reducing the probability that the methodology followed could be influenced by the data used.

Figure 3 presents the total of subjects contained in the database used. From 9812 subjects, 3690 were control subjects and 6122 were case subjects. The subjects were in an age range of 0 to 80 years old; 4830 belonged to the masculine gender and 4982 to the feminine gender.

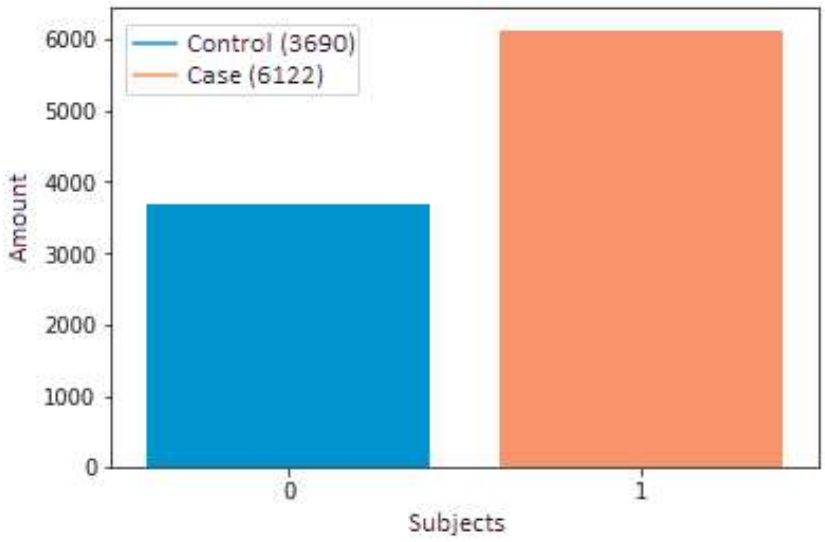

Figure 3. Graph of the total number of subjects used, classified according to their oral health status. The subjects contained in the bar with '0' presented absence of caries, while the subjects contained in the bar with '1' presented presence/restoration of caries.

It is important to mention that this specific database was chosen in order to show the consistency of the data and the results obtained in the relationship to its background work. In this work [21] three different models obtained through a fast backward selection (FBS) of features are presented, each model corresponding to a different age group, then a posterior evaluation using a net reclassification improvement (NRI) technique is performed, besides the AUC parameter, obtaining a maximum true positive-true negative rate of 0.787. On the other hand, this work [22] presents a multivariate model obtained through a FBS method based on the *p*-value, in order to classify between three different classes, "caries", "restorations" and "control"; this model was evaluated using a statistical analysis, obtaining a maximum AUC value of 0.664. Finally, this work [23] presents a univariate analysis using a linear regression approach in order to classify between subjects with the presence of caries and with absence; then, from the most significant univariate models, a multivariate model contained by three features was developed, which obtained an AUC value of 0.572.

2.3. Data Analysis

The data analysis of this work was performed through a multivariate approach, subjecting the demographic and dietary features to a deep ANN for the classification of subjects according to their health and then a statistical validation was carried out in order to evaluate the results obtained from the ANN developed for these specific data.

2.3.1. Data Preprocessing

The data preprocessing consisted initially in the manual elimination of the features that presented a high percentage of missing values (>30%), represented as "Not a Number" (NaN). Then, of the remaining features, those that presented a low percentage of missing data (≤30%) were imputed with

the "rfimput" function, of the "randomForest" package (version 4.6-12, 6 October 2015) [24], for R, which consists of replacing all NaN with the average of the values that present the column where the missing data is located.

The columns containing singular values were also removed because they didn't provide significant information. Two values are singular if they have multiple values between them or if they present the same value in the whole column.

Finally, the database was separated into two sets, randomly and balanced selecting a set for the training stage, contained with 70% of the data and a set for the testing stage, contained with the remaining 30% of data.

2.3.2. Data Classification

The classification of subjects was performed according to the presence or absence of caries through a dense ANN that was designed based on the data of the subjects, using the package "Keras". "Keras" is a high-level ANN application programming interface written in Python. It was developed with the approach of allowing fast experimentation, facilitating the creation of prototypes in a simple way [25].

ANNs seek a solution to a specific task, based on the correlation between features, through learning or training that resembles the behavior of biological neural networks. Using different layers composed of nodes or neurons, the ANNs look for a model of the relationship between the input features and the output feature. The number of nodes that exist for each layer is configurable, as is the number of layers and, depending on the data, there is usually a greater capacity of the network, the greater the number of layers and neurons; however, if there isn't the right amount of these elements, problems of overfitting can be caused .

ANNs present three main elements: (1) a set of synapses or connections, characterized by a "weight", where the input signal is connected to a neuron through its product with the weight of that connection; (2) an adder, which adds the contributions of a signal weighted by all the weights; and (3) an activation function, which is equivalent to a transfer function, affecting the neurons, allowing for limiting the amplitude of the network output, providing a permissible range for the output signal in terms of finite values. Among the most common activation functions are the lineal, quadratic, geometric, logistic, and rectified linear unit function, among others.

On the other hand, an ANN is dense when each node of a layer is connected with every node of the next layer; while the recurrence term refers to the network presenting at least one feedback loop, which will provide a deeper impact on the learning capacity of the network, as well as on its performance, since it allows for optimizing its behavior through a parameter that gives knowledge of the behavior of the data and allows for guiding the adjustment of the configuration of the network to improve its accuracy.

The ANN designed was composed by a series of dense and dropout hidden layers, as shown in the diagram of Figure 4. The input layer (A) was assigned 104 neurons, making reference to the 104 features of the dataset. The first dense hidden layer (B) was composed of 100 neurons, as was the first dropout hidden layer (C), which had a loss percentage of 50%. The second dense hidden layer (D) was assigned 1000 neurons, as was the second dropout hidden layer (E), which had a loss percentage of 25%. The third dense hidden layer (F) was composed of 100 neurons, as was the third dropout hidden layer (G), which had a loss of 50%. Finally, the fourth dense hidden layer was the output layer (H), which was characterized by two neurons, making reference to the two outputs or possible classifications.

The input layer and the dense hidden layers used as activation function, the Rectified Linear Unit (ReLU) function, which consists of assigning 0 to the values of the neurons that are <0, and to respect the value of the neurons when their values are ≥ 0, as shown in Equation (1) [26]:

$$\mathrm{ReLU}(x) = \begin{cases} x, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (1)$$

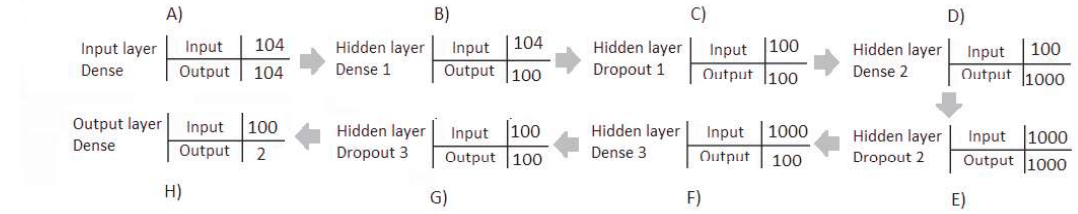

Figure 4. Diagram of the ANN designed.

The output layer used the Normalized Exponential function as an activation function, also known as "Softmax", which represents a general form of the logistic function and it is used to compress a vector of arbitrary values into a vector of real values in a range of $[0, 1]$. This function is shown in Equation (2), where $\sigma(z)$ represents the K-dimensional vector, z, of binary values [27]:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \; j = 1, ..., K. \tag{2}$$

Finally, the dropout hidden layers were included in order to avoid overfitting problems, as mentioned before, and its performance is based on assigning the value of 0 to a percentage of the data and thus they are not taken into account for the classification of subjects in that layer. The data subset selected by these layers is changing with each iteration, causing that in each epoch the classification will be done omitting a different percentage of data. On the other hand, due to the recurrence of the ANN, it was possible to optimize its performance. The optimization algorithm, "Adam", was selected, which bases its operation on stochastic gradient descent algorithms, making use of the average of the first and second moments of the gradients to adapt the rate of the learning parameter. Specifically, "Adam" calculates the exponential moving average of the gradient and the square gradient, and controls the decay rates of that moving average [28]. Some of the benefits that present "Adam" are that it is a method that is straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and one of the most important points is that is suitable for problems that are large in terms of data or features. In addition, the parameters have intuitive interpretations and typically require little tuning [28].

It is important to mention that the number of iterations or epochs that the ANN will have for the optimization of its behavior, based on the recurrence, is also configurable and will depend on the type of data. To know what is the appropriate number of epochs, there are some parameters that allow for evaluating the behavior through each iteration, such as the loss function and the accuracy [29]. The number of the epochs that were selected through a series of tests with different numbers were 100 epochs, since they demonstrated having the best performance in terms of the accuracy of the ANN.

Based on the above, "keras" has the advantage that it allows for designing the architecture of the ANN according to the type of data, configuring the type of ANN, number of nodes, validation, loss function, among others [30,31].

2.3.3. Evaluation

Finally, in order to validate the results of classification obtained by the ANN, three parameters were evaluated; loss function, accuracy and ROC curve. The loss function and accuracy were calculated on each epoch, allowing to know if the performance of the ANN was improving, while the ROC curve was calculated based on the average of the general performance of the ANN.

ANNs are mostly trained using gradient methods through an iterative process of decreasing the loss function. A loss is designed to have the main property that the lower its value, the better the model that fits its data and, in addition, it will be differentiable, which will optimize the network directly, giving information of the capacity of the system. Therefore, the loss function is based on the search

of the global minimum, which corresponds to the minimum error, based on a learning factor. Some of the most common techniques to calculate the loss function are the mean square error, mean absolute error, binary cross-entropy, and Poisson, among others [31].

The loss function that was used is preset in "keras" as "binary-crossentropy" and calculates the cross-entropy value for binary classification problems. This method uses the Kullback–Leibler distance, which is a measure between two density functions g and h, known as the cross-entropy between g and h, as shown in Equation (3). Its operation is based on iterations, generating a random set of values estimating the value that wants to be obtained and then actualizing the parameters in the next iteration to generate "better" values or more approximate, in terms of the Kullback–Leibler distance [32]:

$$D(g,h) = \int g(x) ln \frac{g(x)}{h(x)} \mu(dx) = \int g(x) lng(x) \mu(dx) - \int g(x) lnh(x) \mu(dx). \quad (3)$$

On the other side, the accuracy allows for measuring the performance of the ANN through a non-differentiable function. This metric doesn't allow to optimize the network, but it allows to select the model that shows the most suitable performance in the training of the network, and it is based on the calculation of the average of the differences that exist between the classification calculated by the ANN model and the true classification of the data, as shown in Equation (4), where the accuracy is reported as 1-*error*. V_{pred} refers to the classification value that was calculated by the ANN, while V_{actual} refers to true classification value [33]:

$$error = V_{pred} - V_{true}. \quad (4)$$

In this work, the accuracy was calculated with the "binary-accuracy" function from "keras", which calculates the average accuracy rate across all predictions for binary classification problems.

The ROC curve is a standard method that allows for evaluating the precision with which the model classifies, based on the relationship between sensitivity and specificity. Sensitivity is defined as the proportion of subjects with a condition that were classified as positive, which means the positive predictive values (PPV); this value is calculated with Equation (5), where TP represents the number of true positives and FP represents the number of false positives [34,35]:

$$PPV = \frac{TP}{TP + FP}. \quad (5)$$

Specificity is defined as the proportion of subjects without a condition that was classified as negative; this means the negative predictive values (NPV); this value is calculated with Equation (6), where TN represents the number of true negatives and FN represents the number of false negatives [34,35]:

$$NPV = \frac{TN}{TN + FN}. \quad (6)$$

The ROC curves were calculated to obtain the true positive and true negative rate for each class, for the macro-average precision and for the micro-average precision.

The macro-average precision is obtained through the sum of the true positives, false positives and false negatives of the system, for different sets, as shown in Equation (7), where TP_1 refers to the true positives of the set one, TP_2 refers to the true positives of the set two, FP_1 refers to the false positives of the set one and FP_2 refers to the false positives of the set two [36,37]:

$$\text{Micro-average} = \frac{TP_1 + TP_2}{TP_1 + TP_2 + FP_1 + FP_2}. \quad (7)$$

On the other hand, the macro-average precision is a direct method; it takes the average of the accuracy of the system in different sets. It is calculated with Equation (8), where A_1 refers to the average of the set one and A_2 refers to the average of the set two [36,37]:

$$\text{Micro-average} = \frac{A_1 + A_2}{2}. \quad (8)$$

The analysis of this work was performed in Python (version 3.6), using the packages, "Keras" (version 2.1.5) [25], "Scipy" (version 1.0.0) [38], "Pandas" (version 0.22.0) [39] and "Sklearn" (version 0.19.1) [36].

3. Results

The two databases that were used in this work were contained by a total of 188 demographic and dietary features, besides a feature being used that was contained by the oral health status as an output feature for the classification of subjects.

After the data preprocessing, only 105 features were conserved, of which 25 belonged to demographic features and 79 to dietary features, the remaining feature refers to the output feature.

Of the total set of 9812 subjects, 3690 belonged to controls and 6122 belonged to cases. This dataset was divided into two subsets, one for training which was contained by 70% of the data (2596 controls/4272 cases), and one for testing which was contained with the remaining 30% of the data (1094 controls/1850 cases). These data are graphically shown in Figure 5.

Figure 5. Graph of the number of subjects that belongs to each dataset, training and testing. The content in the bar with "0" presented a status of absence of caries, while the contest in the bar with "1" presented a status of presence/restoration of caries.

The dataset that was designed for the training stage was evaluated at each of the 100 epochs through the accuracy and loss function values. This number of epochs was selected through a comparison between the values obtained using different number of epochs; Table 1 presents the values obtained for ten different epochs.

On the other hand, in Table 2, a comparison of results using different number of layers and neurons is shown. The number of epochs selected was 100, establishing this number according to the result obtained from the previous table. It is possible to observe that the most statically significant values of accuracy and loss function were obtained using an ANN designed with seven layers, four dense layers and three dropout layers. The dense layers were contained by 104, 1000, 100 and two neurons,

in descendant order, while the dropout layers represented 0.50, 0.25 and 0.50 percentages, in descendant order too.

Table 1. Comparison of the accuracy and loss function values obtained using different number of epochs.

Epochs	Accuracy	Loss Function	Processing Time (s)
10	0.67	0.60	10.44
30	0.68	0.60	32.02
50	0.68	0.59	55.46
80	0.69	0.59	93.57
100	0.69	0.58	124.55
150	0.68	0.60	188.21
200	0.68	0.61	246.74
300	0.69	0.62	369.06
500	0.70	0.62	646.86
1000	0.70	0.66	4152.96

Table 2. Comparison of the accuracy and loss function values obtained using different number of layers and neurons.

Layers Dense/Dropout	Neurons	Accuracy	Loss Function	Processing Time (s)
2/1	104>0.50>2	0.68	0.60	36.84
3/1	104>0.50>1000>2	0.68	0.59	71.10
3/2	104>0.25>1000>0.50>2	0.69	0.61	93.86
4/1	104>1000>0.50>100>2	0.68	0.73	117.63
4/2	104>0.50>1000>0.50>100>2	0.68	0.59	119.77
4/3	104>0.50>1000>0.25>100>0.50/2	0.69	0.58	124.55
5/1	104>100>1000>0.50>100>2	0.68	0.91	129.53
5/2	104>100>0.50>1000>0.25>100>2	0.68	0.65	131.66
5/3	104>100>0.50>1000>0.25>100>0.50>2	0.69	0.64	139.37
5/4	104>0.25>100>0.50>1000>0.25>100>0.50>2	0.68	0.59	138.38

The graph of the performance of the accuracy is shown in Figure 6A, where the blue line represents the behavior of the training data, obtaining a final accuracy of 0.69, while the orange line represents the behavior of the testing data, obtaining a final accuracy of 0.68.

The graph of the performance of the loss function is shown in Figure 6B, where the blue line represents the behavior of the training data, obtaining a final value of 0.58, while the orange line represents the behavior of the testing data, obtaining a final value of 0.60.

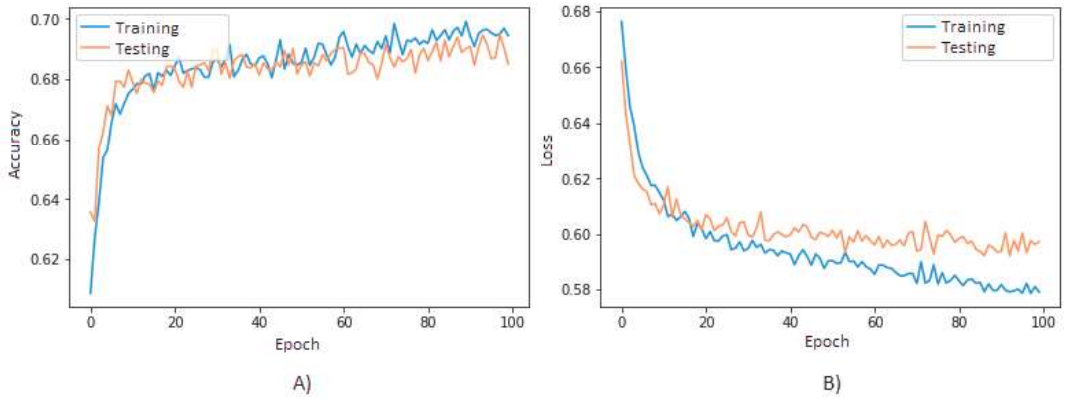

Figure 6. Graphs of the performance of the accuracy (**A**) and the loss function (**B**) of the ANN in each epoch.

Both accuracy and loss function values were obtained in order to know the performance of the classification of subjects on each iteration; nevertheless, the accuracy is a parameter that may

show an optimistic response if the data presents any bias. Therefore, there were also measured the specificity and sensitivity parameters in order to ensure that the classification of subjects was statistically significant.

The graph of the Figure 7 shows the ROC curves of the mean performance of the ANN, where the pink line refers to the proportion of sensitivity and specificity for the class "0", which are the subjects that presented absence of caries, obtaining an AUC value of 0.69. The light blue line refers to the proportion of sensitivity and specificity for the class "1", which are the subjects that presented presence/restoration of caries, obtaining an AUC value of 0.69. The dotted orange line refers to the curve calculated with the micro-average of the proportion of sensitivity and specificity for the classification of subjects, obtaining an AUC value of 0.75; the dotted dark blue line refers to the curve calculated with the macro-average, obtaining an AUC value of 0.69.

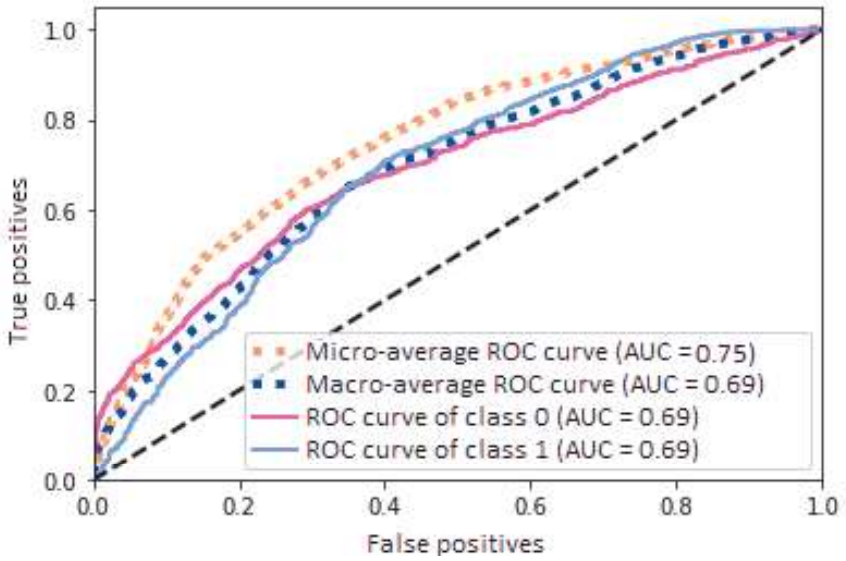

Figure 7. Graph of the ROC curves generated based on the average performance of the ANN in the classification of subjects.

4. Discussion

Of the two datasets that were obtained through the preprocessing step, it is possible to observe that the set of cases contains more data than the set of controls as it is observed in Figure 5; this was due to the fact that the cases included the subjects with presence of caries and with restorations. Nevertheless, the number of subjects that were contained on each dataset was enough to train the ANN, obtaining statistically significant results.

The data were separated in two sets based on an aleatory and balanced selection; one of the sets was contained with 70% of the data and the other dataset was contained with 30%, as shown in Figure 5. The dataset that was contained with the largest amount of data had the purpose of training the ANN. The dense ANN was trained during 100 epochs and was optimized through the optimization algorithm, "Adam", which had the purpose of improving the performance of the ANN with feedback.

Finally, for the validation of the ANN modeling, the accuracy and the loss function in each of the epochs with the dataset that was contained with the least amount of data were obtained. The loss function was calculated using cross-entropy in order to know how the performance of the ANN was modified. The validation was carried out in both datasets, training and testing, in order to ensure that there was no adjustment in the classification of the subjects; that is, if the performance of these parameters is similar for both datasets, it means that the ANN is classifying based on a generalized

model, since it manages to classify with an accuracy and a loss function similar in the training dataset and in the testing dataset.

In the graph of Figure 6A, it is observed that the performance of the accuracy is similar for both datasets, being slightly higher for the training data, which is normal since they are the data on which the ANN modeling is being based. As the performance is similar for both datasets, it is possible to know that the ANN model became generalized through the learning by being able to classify unknown data with a similar accuracy to that obtained with the training data. The final value obtained for the accuracy was 0.69, which is statistically significant since it means that at least 69% of the data was correctly classified.

On the other hand, in the graph of Figure 6B, it is observed that the performance of the loss function is also similar for both datasets. It is notable that, for the training dataset, this parameter is smaller compared to that obtained with the testing dataset; however, as in the case of the accuracy parameter, this performance is normal because the testing dataset was unknown for the model. In addition, having a similar performance allows for knowing that the model doesn't have overfitting problems and that it manages to classify unknown data in a similar way as it classifies known data. The final value obtained for the accuracy had an average of 0.59. This value isn't close to the ideal, which is zero; however, it is possible to observe that the data remains tending toward zero for both datasets, showing that the testing dataset decreases slower than the training dataset. This problem can be solved increasing the amount of data, removing features that are not significant or that are redundant for the model and increasing the size if the ANN or the number of epochs.

Although the graph of the loss function doesn't present very favorable results, it allows for knowing that the modeling of the ANN is robust among the two datasets, in addition to continuing to decrease the loss function and reducing the error through the epochs, which is one of the main purposes of the training stage.

Finally, Figure 7 shows that all curves presented statistically significant values, obtaining AUC values $\geq$0.69. These curves are generated based on the proportion of true positives and true negatives, where for class "0", class "1" and the macro-average, an AUC value of 0.69 was obtained, which implies that 69% of the subjects were classified appropriately for each of the classes and for their general average, while. for the curve generated for the micro-average, an AUC value of 0.75 was obtained, which implies that 75% of the subjects were classified appropriately when the measure was based in subsets of data. The ROC curves of class "0" and class "1" presented a very similar performance, besides obtaining the same AUC value, which means that the classification performance is equitable and the generalization of the model allows for classifying subjects of both classes.

On the other hand, the AUC value is greater for the micro-average than for the macro-average; this may be because the calculation of the macro-average is based in obtaining the true positives/true negatives proportion through the whole amount of data, while, for the micro-average, the calculation of the ROC curve is based in subsets of data, where some subsets can present better AUC values than obtained with the whole set, thus achieving a better average of classification.

Is it important to remark that this generalized model presents an impartial performance, since it was trained with data that cover a reasonable spectrum of the possible outcomes; in addition, the moment of time and geographical location are parameters that were taken into account in the initial data acquisition; regardless of whether the data was acquired in the USA, most of the subjects were from other regions, of which a percentage belongs to Mexican subjects, providing robustness in the results and avoiding any bias problem related to the structure, the moment of time and the geographical location of the data.

Finally, the results obtained allow for supporting the main motivation of this work, which was to develop a tool for human resources and dental services, since this model is able to give a statistically significant response of the possible development of caries based on information that doesn't require any extra equipment for its acquisition, only the willingness of people to answer the demographic and

diet surveys, providing support to the dental specialists in the work of reducing the high incidence of dental caries

Based on these discussions, it can be proposed as future work to add a stage in the methodology that consists of a step for the feature selection using a genetic algorithm approach. This stage will help to remove redundant information and correlated features, keeping only those features that provide the most significant information for the classification of subjects, besides reducing the computational cost. Additionally, for the validation of the feature selection, a forward selection and backward elimination may be used in order to test the accuracy of the set of features obtained through the selection process, ensuring the certainty of the model behavior.

5. Conclusions

According to the results obtained, it is possible to conclude that the amount of data that was used to model the classification of the subjects was adequate for the preliminary generalized learning of the ANN, basing this on the performance that the classification of cases had, being similar to the classification of controls. This performance was observed in the accuracy and loss function graphs, which allows for recognizing that the number of epochs that were assigned to the ANN was sufficient to maintain stable performance of the parameters in both datasets.

The accuracy achieved with this modeling was around 0.70, a value that is statistically significant since it implies that 70% of the time will be correctly classified to the subjects with presence or absence/restoration of caries.

The loss function decreased by approximately 10% since the beginning of the training, showing minor changes throughout the epochs, implying an approximation to the global minimum sought.

As a final evaluation, all the ROC curves obtained an AUC value statistically significant, implying that, around 70% of the time, subjects were correctly classified, according to the true positives and true negatives proportion, a value that corroborates the accuracy obtained. Nevertheless, a methodology was proposed based in convolutional neural networks in order to identify features in a spectral or spatial domain to obtain time-independent abstract features, looking for the improvement in the modeling for the classification of subjects.

Based on this, it is possible to classify subjects with the absence of caries from subjects with presence/restorations, through demographic and dietary data, with a statistically significant accuracy, demonstrating that the socioeconomic and nutritional status are important determinants in the development of caries.

Then, according to the results obtained, it was demonstrated that the demographic situation can significantly affect the prevalence of dental caries. Based on this, the analysis of oral health data from exclusively Mexican subjects is proposed, comparing those results with the ones obtained in this work, with the purpose of proving how demography can influence the oral health status.

On the other hand, even though the results do not show an ideal behavior, they give preliminary knowledge of the benefit that its implementation would have in a real environment, since the requirements for its use are minimal, no in-depth knowledge of the techniques used is required, results are presented quickly and the computational cost is very low, besides the accuracy obtained being statistically significant. In addition, an important point to take into account is that the performed experimental tests were based in a real production environment, since all the subjects that were part for the development of this work were real control and cases, and the demographic and dietary databases were obtained from real information.

The hardware tool that is required for the implementation of this work is a computer, while the free software tool that is required is Python, which is a programming language that allows for working quickly and integrating systems more effectively. As extra information, it is mentioned that the computer that was used for the development of this work was a laptop Acer Aspire F5-573-70LX 15.6" (Acer America Corporation, 333 West San Carlos Street, Suite 1500, San Jose, CA 95110), Intel Core i7-7500U 2.70 GHz (Plot 6, Bayan Lepas Technoplex Medan Bayan Lepas 11900 Bayan Lepas Penang,

Malaysia, Georgetown, Pulau Pinang, Malasia), 16 GB, 1 TB + 128 GB Solid State Drive, Windows 10 Home (15010 NE 36th Street, Microsoft Campus, Building 92, Redmond, WA 98052), 64-bit; and the version of Python that was used is 3.6 [40].

Finally, based on the last point, it is evident that the implementation of the model proposed in this work could provide an easy, free and fast tool that helps the specialists in the preventive diagnosis of dental caries, besides offering an option that collaborates in the decrement of the incidence of this public health problem.

Author Contributions: C.E.G.-T. and L.A.Z.-C., performed the study. C.E.G.-T., L.A.Z.-C., and J.I.G.-T. performed the study design and data analysis. N.M.C.-L. and J.R.-G. contributed to materials and methods (selection of patients from database) used in this study. C.E.G.-T., J.I.G.-T., and J.M.C.-P. performed statistical analysis and statistical validation with critical feedback to authors. J.M.C.-P., R.M.-Q. and L.A.Z.-C. contributed with the neural network implementation used in this study. L.A.Z.-C. and H.G.-R. provide feedback from results. All authors interpreted findings from the analysis and drafted the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Description of demographic features.

Feature	Description
AIALANGA	Language of the MEC ACASI Interview Instrument.
DMDBORN4	In what country was Sample Person (SP) born?
DMDCITZN	Is SP a citizen of the United States?
DMDHHSIZ	Total number of people in the Household.
DMDHHSZE	Number of adults aged 60 years or older in the household.
DMDEDUC2	Highest grade or level of school completed or the highest degree received.
DMDHHSZB	Number of children aged 6–17 years old in the household (HH).
DMDHSEDU	HH reference person's spouse's education level.
DMDMARTL	Marital status.
DMDFMSIZ	Total number of people in the Family.
DMDHHSZA	Number of children aged 5 years or younger in the household.
DMDHRGND	HH reference person's gender.
DMDHREDU	HH reference person's education level.
DMDEDUC3	Highest grade or level of school completed or the highest degree received.
DMDHRAGE	HH reference person's age in years.
DMDHRBR4	HH reference person's country of birth.
DMDHREDU	HH reference person's education level.
DMDHRMAR	HH reference person's marital status.
DMQADFC	Did SP ever serve in a foreign country during a time of armed conflict or on a humanitarian or peace-keeping mission?
DMDYRSUS	Length of time the participant has been in the US.
DMQMILIZ	Has SP ever served on active duty in the U.S. Armed Forces, military Reserves, or National Guard?
FIAINTRP	Was an interpreter used to conduct the Family interview?
FIALANG	Language of the Family Interview Instrument.
FIAPROXY	Was a Proxy respondent used in conducting the Family Interview?
INDFMPIR	A ratio of family income to poverty guidelines.
INDHHIN2	Total household income (reported as a range value in dollars).
INDFMIN2	Total family income (reported as a range value in dollars).
LATVPSU	Variance unit: PSU variable for variance estimation.
LATVSTRA	Variance unit: stratum variable for variance estimation.
MIAINTRP	Was an interpreter used to conduct the MEC CAPI interview?
MIALANG	Language of the MEC CAPI Interview Instrument.
MIAPROXY	Was a Proxy respondent used in conducting the MEC CAPI Interview?
RIAGENDR	Gender of the participant.
RIDAGEEX	Age in years of the participant at the time of examination. Individuals aged 959 months and older are topcoded at 959 months.

Table A1. *Cont.*

Feature	Description
RIDAGEMN	ge in months of the participant at the time of screening. Reported for persons aged 24 months or younger at the time of exam.
RIDRETH3	Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category.
RIDEXMON	Six month time period when the examination was performed.
RIDEXPRG	Pregnancy status for females between 20 and 44 years of age at the time of MEC exam.
RIDAGEYR	Age in years of the participant at the time of screening.
RIDRETH1	Recode of reported race and Hispanic origin information.
RIDEXAGM	Age in months of the participant at the time of examination.
RIDSTATR	Interview and examination status of the participant.
SDMVPSU	Masked variance unit pseudo-PSU variable for variance estimation.
SDDSRVYR	Data release cycle.
SDMVSTRA	Masked variance unit pseudo-stratum variable for variance estimation.
WTLAF8YR	Subsample 8-year fasting weight for participants aged 12 years and older who were examined in the morning sessions in Los Angeles, CA, USA.
WTLAI8YR	Full sample 8-year interview weight for participants in Los Angeles, CA, USA.
WTLAM8YR	Full sample 8-year MEC exam weight for participants in Los Angeles, CA, USA.

Appendix B

Table A2. Description of dietary features.

Feature	Description
WTDRD1	Dietary day one sample weight.
DBQ095Z	Type of salt usually add to food at the table.
DR1TKCAL	Energy (kcal).
DR1TSUGR	Total sugars (g).
DR1TFIBE	Dietary fiber (g).
DR1TSFAT	Total saturated fatty acids (g).
DR1TMFAT	Total monounsaturated fatty acids (g).
DR1TPFAT	Total polyunsaturated fatty acids (g).
DR1TLYCO	Lycopene (µg).
DR1TFA	Folic acid (µg).
DR1TB12A	Added vitamin B12 (µg).
DR1_300	Was the amount of food that you ate yesterday much more than usual, usual, or much less than usual?
DR1_320Z	Total plain water drank yesterday—including plain tap water, water from a drinking fountain, water from a water cooler, bottled water, and spring water.
DR1_330Z	Total tap water drank yesterday—including filtered tap water and water from a drinking fountain.
DR1BWATZ	Total bottled water drank yesterday (g).
DR1DAY	Intake day of the week.
DR1DBIH	Number of days between intake day and the day of family questionnaire administered in the household.
DR1SKY	What type of salt was it? (Was it ordinary or seasoned salt, lite salt, or a salt substitute?).
DR1STY	Did SP add any salt to her/his food at the table yesterday?
DR1TACAR	Alpha-carotene (µg).
DR1TBCAR	Beta-carotene (µg).
DR1TCAFF	Caffeine (mg).
DR1TCALC	Calcium (mg).
DR1TCOPP	Copper (mg).
DR1TCRYP	Beta-cryptoxanthin (µg).
DR1TFDFE	Folate as dietary folate equivalents (µg).
DR1TFF	Food folate (µg).
DR1TFOLA	Total folate (µg).
DR1TLZ	Lutein + zeaxanthin (µg).
DR1TM161	MFA 16:1 (Hexadecenoic) (g).
DR1TM181	MFA 18:1 (Octadecenoic) (g).
DR1TM201	MFA 20:1 (Eicosenoic) (g).
DR1TM221	MFA 22:1 (Docosenoic) (g).
DR1TMAGN	Magnesium (mg).
DR1TMOIS	Moisture (g).
DR1TNIAC	Niacin (mg).

Table A2. *Cont.*

Feature	Description
DR1TNUMF	Total number of foods/beverages reported in the individual foods file.
DR1TP182	PFA 18:2 (Octadecadienoic) (g).
DR1TP183	PFA 18:3 (Octadecatrienoic) (g).
DR1TP184	PFA 18:4 (Octadecatetraenoic) (g).
DR1TP204	PFA 20:4 (Eicosatetraenoic) (g).
DR1TP205	PFA 20:5 (Eicosapentaenoic) (g).
DR1TP225	PFA 22:5 (Docosapentaenoic) (g).
DR1TP226	PFA 22:6 (Docosahexaenoic) (g).
DR1TPOTA	Potassium (mg).
DR1TRET	Retinol (µg).
DR1TS040	SFA 4:0 (Butanoic) (g).
DR1TS060	SFA 6:0 (Hexanoic) (g).
DR1TS100	SFA 10:0 (Decanoic) (g).
DR1TS120	SFA 12:0 (Dodecanoic) (g).
DR1TS140	SFA 14:0 (Tetradecanoic) (g).
DR1TS180	SFA 18:0 (Octadecanoic) (g).
DR1TSELE	Selenium (µg).
DR1TSODI	Sodium (mg).
DR1TTHEO	Theobromine (mg).
DR1TVARA	Vitamin A as retinol activity equivalents (µg).
DR1TVB1	Thiamin (Vitamin B1) (mg).
DR1TVB12	Vitamin B12 (µg).
DR1TVB2	Riboflavin (Vitamin B2) (mg).
DR1TVB6	Vitamin B6 (mg).
DR1TVC	Vitamin C (mg).
DR1TVK	Vitamin K (µg).

Table A3. Description of dietary features.

Feature	Description
DR1TWS	When you drink tap water, what is the main source of the tap water? Is the city water supply; a well or rain cistern; a spring; or something else?
DR1TZINC	Zinc (mg).
DRABF	Indicates whether the sample person was an infant who was breast-fed on either of the two recall days.
DRD340	During the past 30 days did you eat any types of shellfish listed on this card?
DRD350A	Clams eaten during the past 30 days.
DRD350AQ	Number of times clams were eaten in the past 30 days.
DRD350B	Crabs eaten during the past 30 days.
DRD350BQ	Number of times crab was eaten in the past 30 days.
DRD350C	Crayfish eaten during the past 30 days.
DRD350CQ	Number of times crayfish was eaten in the past 30 days.
DRD350D	Lobsters eaten during the past 30 days.
DRD350DQ	Number of times lobster was eaten in the past 30 days.
DRD350E	Mussels eaten during the past 30 days.
DRD350EQ	Number of times mussels were eaten in the past 30 days.
DRD350F	Oysters eaten during the past 30 days.
DRD350FQ	Number of times oysters were eaten in the past 30 days.
DRD350G	Scallops eaten during the past 30 days.
DRD350GQ	Number of times scallops were eaten in the past 30 days.
DRD350H	Shrimp eaten during the past 30 days.
DRD350HQ	Number of times shrimp was eaten in the last 30 days.
DRD350I	Other shellfish (ex. octopus, squid) eaten during the past 30 days.
DRD350IQ	Number of times other shellfish (ex. octopus, squid) was eaten in the past 30 days.
DRD350J	Other unknown shellfish eaten during the past 30 days.
DRD350JQ	Number of times other unknown shellfish was eaten in the past 30 days.
DRD350K	Refused to give detailed information on shellfish eaten during the past 30 days.
DRD360	During the past 30 days did you eat any types of fish listed on this card?
DRD370A	Breaded fish products eaten during the past 30 days.
DRD370AQ	Number of times breaded fish products were eaten in the past 30 days.
DRD370B	Tuna eaten during the past 30 days.

Table A3. *Cont.*

Feature	Description
DRD370BQ	Number of times tuna was eaten in the past 30 days.
DRD370C	Bass eaten during the past 30 days.
DRD370CQ	Number of times bass was eaten in the past 30 days.
DRD370D	Catfish eaten during the past 30 days.
DRD370DQ	Number of times catfish was eaten in the past 30 days.
DRD370E	Cod eaten during the past 30 days.
DRD370EQ	Number of times cod was eaten in the past 30 days.
DRD370F	Flatfish eaten during the past 30 days.
DRD370FQ	Number of times flatfish was eaten in the past 30 days.
DRD370G	Haddock eaten during the past 30 days.
DRD370GQ	Number of times haddock was eaten in the past 30 days.
DRD370H	Mackerel eaten during the past 30 days.
DRD370HQ	Number of times mackerel was eaten in the past 30 days.
DRD370I	Perch eaten during the past 30 days.
DRD370IQ	Number of times perch was eaten in the past 30 days.
DRD370J	Pike eaten during the past 30 days.
DRD370JQ	Number of times pike was eaten in the past 30 days.
DRD370K	Pollock eaten during the past 30 days.
DRD370KQ	Number of times pollock was eaten in the past 30 days.
DRD370TQ	Number of times other type of fish was eaten in the past 30 days.
DRD370V	Refused to give detailed information on fish eaten during the past 30 days.
DRQSDIET	Are you currently on any kind of diet, either to lose weight or for some other health-related reason?
DRQSDT1	What kind of diet are you on? (Is it a weight loss or a low calorie diet: low fat or cholesterol diet; low salt or sodium diet; sugar free or low sugar diet; low fiber diet; high fiber diet; diabetic diet; or another type of diet?).
DRD370L	Porgy eaten during the past 30 days.
DRD370LQ	Number of times porgy was eaten in the past 30 days.
DRD370M	Salmon eaten during the past 30 days.
DRD370MQ	Number of times salmon was eaten in the past 30 days.
DRD370N	Sardines eaten during the past 30 days.
DRD370NQ	Number of times sardines were eaten in the past 30 days.
DRD370O	Sea bass eaten during the past 30 days.
DRD370OQ	Number of times sea bass was eaten in the past 30 days.
DRD370U	Other unknown type eaten during the past 30 days.

Table A4. Description of dietary features.

Feature	Description
WTDR2D	Dietary two-day sample weight.
DR1DRSTZ	Dietary recall status.
DR1TPROT	Protein (g).
DR1TCARB	Carbohydrate (g).
DR1TTFAT	Total fat (g).
DR1TCHOL	Cholesterol (mg).
DR1TFA	Folic acid (µg).
DR1TCHL	Total choline (mg).
DR1TIRON	Iron (mg).
DBD100	Frequency with which ordinary salt is added to the food on the table.
DRQSPREP	Frequency with which ordinary salt or seasoned salt is added in cooking or preparing foods in the household.
DR1TALCO	Alcohol (g).
DR1TS080	SFA 8:0 (Octanoic) (g).
DR1TATOC	Vitamin E as alpha-tocopherol (mg).
DR1TATOA	Added alpha-tocopherol (Vitamin E) (mg).
DR1TVD	Vitamin D (D2 + D3) (µg).
DR1TPHOS	Phosphorus (mg).
DR1TS160	SFA 16:0 (Hexadecanoic) (g).
DRD370UQ	Number of times other unknown type of fish was eaten in the past 30 days.
DRD370V	Refused to give detailed information on fish eaten during past 30 days.
DRDINT	Indicates whether the sample person has intake data for one or two days.
DRQSDIET	Are you currently on any kind of diet, either to lose weight or for some other health-related reason?

References

1. Oral Health. Available online: http://www.who.int/oral_health/disease_burden/global/en/ (accessed on 5 June 2017).
2. Ridao Marín, D. *Desarrollo de un Sistema de Ayuda a la Decisión para Tratamientos Odontológicos con Imágenes Digitales*; Universidad de Málaga: Málaga, Spain, 2017; pp. 10–12.
3. Espinoza Solano, M.; León-Manco, R.A. Prevalencia y experiencia de caries dental en estudiantes según facultades de una universidad particular peruana. *Rev. Estomatol. Hered.* **2015**, *25*, *3*, 187–193. [CrossRef]
4. Acuña Aguilar, L.D.; Porras Cerón, D.; Ríos Rueda, L.D. *Prevalencia de Lesiones Cariosas y Factores Asociados Presentes en Pacientes con SíNdrome de Down en las Fundaciones Fundown y san Luis Guanella de Bucaramanga*; Universidad Santo Tomás: Bucaramanga, Colombia, 2017; pp. 12–16.
5. Gispert Abreu, E.D.L.Á.; Castell-Florit Serrate, P.; Herrera Nordet, M. Salud bucal poblacional y su producción intersectorial. *Rev. Cubana Estomatol.* **2015**, *52*, 62–67.
6. Niño, T.C.; Guevara, S.V.; González, F.A.; Jaque, R.A.; Infante, C. Uso de redes neuronales articiales en predicción de morfología mandibular a través de variables craneomaxilares en una vista posteroanterior. *Univ. Odontol.* **2016**, *35*, 1–28.
7. Lam, C.U.; Khin, L.; Kalhan, A.; Yee, R.; Lee, Y.; Chong, M.F.; Kwek, K.; Saw, S.; Godfrey, K.; Chong, Y.; et al. Identification of Caries Risk Determinants in Toddlers: Results of the GUSTO Birth Cohort Study. *Caries Res.* **2017**, *51*, 271–282. [CrossRef] [PubMed]
8. Fernandes, I.; Sá-Pinto, A.; Marques, L.S.; Ramos-Jorge, J.; Ramos-Jorge, M. Maternal identification of dental caries lesions in their children aged 1–3 years. *Eur. Arch. Paediatr. Dent.* **2017**, *18*, 197–202. [CrossRef] [PubMed]
9. Lips, A.; Antunes, L.S.; Antunes, L.A.; Pintor, A.V.B.; Santos, D.A.B.D.; Bachinski, R.; Küchler, E.C.; Alves, G.G. Salivary protein polymorphisms and risk of dental caries: A systematic review. *Braz. Oral Res.* **2017**, *31*. [CrossRef] [PubMed]
10. Ahmed, A.; Ikram, K.; Masood, H.; Urooj, M. Identification of relationship between oral disorders & hemodynamic parameters. *Pak. Oral Dent. J.* **2017**, *37*, 202–204.
11. Sarmiento, R.V.; Barrionuevo, F.P.; Huamán, Y.S.; Loyola, M.C. Prevalencia de caries de infancia temprana en niños menores de 6 años de edad, residentes en poblados urbano marginales de Lima Norte. *Rev. Estomatol. Herediana* **2011**, *21*, 79–86. [CrossRef]
12. Oropeza-Oropeza, C.D.; Molina-Frechero, N.; Castañeda-Castaneira, E.; Zaragoza-Rosado, C.D.; Cruz Leyva, C.D. Caries dental en primeros molares permanentes de escolares de la delegación Tláhuac. *Rev. ADM* **2012**, *69*, 63–68.
13. Cardozo, B.J.; Gonzalez, M.M.; Pérez, S.R.; Vaculik, P.A.; Sanz, E.G. Epidemiología de la caries dental en niños del Jardín de Infantes "Pinocho" de la ciudad de Corrientes. *Rev. Fac. Odontol.* **2017**, *9*, *1*, 35–41.
14. Zhang, X.; Zhang, L.; Zhang, Y.; Liao, Z.; Song, J. Predicting trend of early childhood caries in mainland China: A combined meta-analytic and mathematical modelling approach based on epidemiological survey. *Sci. Rep.* **2017**, *7*, 6507. [CrossRef] [PubMed]
15. Asif, S.M.; Babu, D.B.; Naheeda, S. Utility of Dermatoglyphic Pattern in Prediction of Caries in Children of Telangana Region, India. *J. Contemp. Dent. Pract.* **2017**, *18*, 490–496. [CrossRef] [PubMed]
16. Chapple, I.L.; Bouchard, P.; Cagetti, M.G.; Campus, G.; Carra, M.C.; Cocco, F.; Nibali, L.; Hujoel, P.; Laine, M.L.; Lingstrom, P.; et al. Interaction of lifestyle, behaviour or systemic diseases with dental caries and periodontal diseases: Consensus report of group 2 of the joint EFP/ORCA workshop on the boundaries between caries and periodontal diseases. *J. Clin. Periodontol.* **2017**, *44*, s39–s51. [CrossRef] [PubMed]
17. Hayes, M.; Da Mata, C.; Cole, M.; McKenna, G.; Burke, F.; Allen, P.F. Risk indicators associated with root caries in independently living older adults. *J. Dent.* **2016**, *51*, 8–14 [CrossRef] [PubMed]
18. Sudhir, K.M.; Kanupuru, K.K.; Fareed, N.; Mahesh, P.; Vandana, K.; Chaitra, N.T. CAMBRA as a Tool for Caries Risk Prediction Among 12-to 13-year-old Institutionalised Children-A Longitudinal Follow-up Study. *Oral Health Prev. Dent.* **2016**, *14*, 355–362. [PubMed]
19. Twetman, S. Caries risk assessment in children: How accurate are we? *Eur. Arch. Paediatr. Dent.* **2016**, *17*, 27–32. [CrossRef] [PubMed]
20. National Health and Nutrition Examination Survey Data. 2013–2014. Available online: http://www.cdc.gov/nchs/nhanes.htm (accessed on 5 November 2017).

21. Zanella-Calzada, L.A.; Galván-Tejada, C.E.; Chávez-Lamas, N.M.; Gracia-Cortés, M.D.C.; Moreno-Báez, A.; Arceo-Olague, J.G.; Celaya-Padilla, J.M.; Galván-Tejada, J.I.; Gamboa-Rosales, H. A Case–Control Study of Socio-Economic and Nutritional Characteristics as Determinants of Dental Caries in Different Age Groups, Considered as Public Health Problem: Data from NHANES 2013–2014. *Int. J. Environ. Res. Public Health* **2018**, *15*, 957. [CrossRef] [PubMed]
22. Zanella-Calzada, L.A.; Galván-Tejada, C.E.; Chávez-Lamas, N.M.; Galván-Tejada, J.I.; Celaya-Padilla, J.M. Multivariate features selection from demographic and dietary descriptors as caries risk determinants in oral health diagnosis: Data from NHANES 2013–2014. In Proceedings of the International Conference on Electronics, Communications and Computers (CONIELECOMP), Cholula, Mexico, 21–23 Feburary 2018; pp. 217–224.
23. Chávez-Lamas, N.M.; Zanella-Calzada, L.A.; Galván-Tejada, C.E. An Analysis of Dietary and Demographic Data in Oral Health, Data from the National Health and Nutrition Examination Survey: A Preliminary Study. *Adv. Comput. Netw. Appl.* **2017**, *142*, 79–88.
24. Liaw, A.; Wiener, M. Classication and regression by randomforest. *R News* **2002**, *2*, 18–22.
25. Francois Chollet. Keras: Deep Learning Library for Theano and Tensorflow. pp. 145–151. Available online: https://keras.io/k (accessed on 5 May 2018).
26. Lomuscio, A.; Maganti, L. An approach to reachability analysis for feed-forward relu neural networks. *arXiv* **2017**. [CrossRef]
27. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
28. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**. [CrossRef]
29. Haykin, S.S. *Neural Networks and Learning Machines*; Pearson: Upper Saddle River, NJ, USA, 2009; Volume 3, pp. 1–46.
30. Chollet, F. *Deep Learning with Python*; Manning Publications Co.: Greenwich, CT, USA, 2017; Volume 1, pp. 1–93.
31. Antona Cortés, C. Herramientas Modernas en Redes Neuronales: La LibreríA Keras. Bachelor's Thesis, UAM, Departamento de Ingeniería Informática, Madrid, Spain, 2017; pp. 21–38.
32. Helene Bischel, S. El método de la EntropíA Cruzada. Algunas Aplicaciones. Master's Thesis, Universidad de Almería, Almería, Spain, 2015; pp. 4–14.
33. Nye, M.; Saxe, A. *Are Efficient Deep Representations Learnable?* UAM, Departamento de Ingeniería Informática: Madrid, Spain, 2017; pp. 1–4.
34. Lobo, J.M.; Jiménez-Valverde, A.; Real, R. AUC: A misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* **2008**, *17*, 145–151. [CrossRef]
35. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef] [PubMed]
36. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
37. Rosa, K.D.; Shah, R.; Lin, B.; Gershman, A.; Frederking, R. Topical clustering of tweets. In Proceedings of the ACM SIGIR: SWSM, Beijing, China, 28 July 2011.
38. Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python. Available online: https://docs.scipy.org/doc/scipy (accessed on 5 May 2018).
39. Wes McKinney. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 51–56.
40. Python Software Foundation. Python Language Reference, Version 3.6. Available online: http://www.python.org (accessed on 5 June 2018).

4. Discusión

Esta sección muestra un resumen de las discusiones presentadas en los artículos publicados mencionados anteriormente con relación a este trabajo de tesis, así como una discusión general de acuerdo con los resultados obtenidos.

En el artículo "An Analysis of Dietary and Demographic Data in Oral Health Data from the National Health and Nutrition Examination Survey: A Preliminary Study" [41] se presenta un estudio en donde se lleva a cabo un análisis univariado de un reducido conjunto de determinantes demográficos y dietéticos seleccionados por un especialista, con el fin de conocer su impacto en el estado de salud oral. El análisis univariado se basó en el método de regresión lineal y los resultados se evaluaron posteriormente en términos del AUC y el p-valor.

De acuerdo con los resultados, los determinantes demográficos no fueron estadísticamente significativas de manera univariada; sin embargo, cuando se analizaron de manera multivariada, incluyendo los determinantes dietéticos y los determinantes demográficos en un mismo modelo, la evaluación presentó mejores resultados.

Es posible identificar que algunos factores demográficos, tales como la seguridad social, el ingreso económico y el tipo de servicio de salud, pueden ser determinantes en el estado de salud oral de los pacientes.

Por otro lado, en el artículo "Multivariate features selection from demographic and dietary descriptors as caries risk determinants in oral health diagnosis: data from NHANES 2013-2014" [3] se proponen dos modelos multivariados, uno con datos demográficos y uno con datos dietéticos, para la clasificación entre sujetos control, sujetos con presencia de caries y sujetos con restauraciones de caries. Estos modelos fueron desarrollados utilizan-

do un método de selección rápida hacia adelante basado en el p-valor y posteriormente fueron sometidos a una etapa de validación, a través de un análisis estadístico.

De acuerdo con los resultados que se obtuvieron, el modelo contenido con determinantes demográficos muestra que la edad, el sexo y el estado socio-económico son determinantes importantes en la presencia, ausencia o restauración de caries, presentando una probabilidad favorecedora para alguno de los tres estados.

Por otro lado, el modelo contenido con determinantes dietéticos presentó información relacionada con el balance de alimentos, encontrando entre ella determinantes del grupo de ácidos grasos, los cuales son elementos comunes en la dieta diaria.

La etapa de validación permitió evaluar estos modelos, presentando resultados estadísticamente significativos, por lo que la clasificación de sujetos se puede realizar con base en la información de los modelos, demográfica y dietética, de manera independiente, con un rendimiento significativo.

En el trabajo titulado "An analsysis of demographic and dietary data with an oral health approach, a preliminary study usign genetic algorithms" [69], se analizaron nuevamente los determinantes demográficos y dietéticas, llevando a cabo una selección de características a través de un AG, con el fin de desarrollar un modelo multivariado basado en las características que presentaran el mejor rendimiento para clasificar sujetos con presencia o restauración de caries de sujetos control. Las características seleccionadas fueron sometidas a una etapa de validación, en donde se utilizó el método de regresión logística y un análisis estadístico, utilizando la curva ROC y el AUC para conocer la proporción de sensibilidad y especificidad.

Después de una etapa de selección hacia adelante y una eliminación hacia atrás, fueron seleccionadas cinco características, de las cuales cuatro pertenecían a información demográfica y una pertenecía a información dietética. El resultado de la evaluación del modelo multivariado presentó un valor estadísticamente significativo.

La información demográfica contenida en las características seleccionadas se relaciona con la edad de los sujetos, el nivel académico y la actividad en los servicios de E.U.A. Estos resultados están acordes con la literatura, ya que, de acuerdo con diferentes estudios,

la edad es un factor importante en la salud oral. El nivel académico y la actividad en los servicios de la nación son características relacionadas con el nivel socio-económico, el cual es un factor de riesgo en el desarrollo de la caries dental, en donde las regiones de menores ingresos son las más afectadas por esta condición.

Por otro lado, el determinante relacionado con la información dietética hace referencia a la ingesta total de nutrientes, indicando la calidad y la integridad de la vida de los sujetos, los cuales también son determinantes en esta enfermedad.

Cabe resaltar que en este trabajo los determinantes demográficos presentaron mayor peso que las características dietéticas, mostrando la gran relevancia que presentan los factores del ambiente en el que se desenvuelven los sujetos, además de la edad.

En el artículo "A Case-Control Study of Socio-Economic and Nutritional Characteristics as Determinants of Dental Caries in Different Age Groups, Considered as Public Health Problem: Dara from NHANES 2013-104" [67], se presentó un estudio de caso - control donde se realiza el análisis de una serie de determinantes demográficos y dietéticos, buscando clasificar sujetos con presencia de caries dental de sujetos control, utilizando un selector de características basado en una técnica de selección rápida hacia adelante. Los sujetos fueron divididos en cuatro diferentes grupos de edad (grupo 1: 0-9 años, grupo 2: 10-19 años, grupo 3: 20-59 años, grupo 4: 60 años o más), con el fin de comprobar si la edad representa un factor de riesgo para el desarrollo de caries. Finalmente, una etapa de validación, basada en la técnica de "Net Reclassification Improvement", fue llevada a cabo para la evaluación de los resultados.

Por cada grupo de edad se obtuvo un modelo multivariado, con excepción del grupo de edad 1, que no estaba contenido por ningún sujeto. Estos modelos concuerdan en algunas características por las que están compuestos; sin embargo, hay algunas diferencias importantes en las características que fueron seleccionadas para cada grupo de edad.

En el grupo de edad 2, el modelo está contenido por una cantidad balanceada de determinantes demográficos y dietéticos (15 y 18, respectivamente), implicando que ambas presentan información relevante para clasificar la presencia y la ausencia de caries; es decir, que el tener una buena alimentación tiene una importancia similar a tener una ade-

cuada educación en el cuidado de la salud oral, tomando en cuenta también el ambiente en el que se desarrollan los sujetos.

Por otro lado, el grupo de edad 3 presenta un modelo con el doble de características dietéticas que demográficas (12 y 6), lo que quiere decir que en este rango de edad hay más información dietética que permite diferenciar a los sujetos.

Finalmente, el grupo de edad 4 también presenta una cantidad balanceada de determinantes demográficos y dietéticos (4 y 6, respectivamente), dando una relevancia semejante a la información de ambas para clasificar sujetos, al igual que en el grupo de edad 2.

Es importante señalar que la cantidad de características contenidas en cada modelo multivariado disminuyó en relación con el incremento del rango de edad, por lo que es posible que en el rango de edad del grupo 2 exista una mayor cantidad de factores de riesgo para el desarrollo de caries dental que en los otros grupos de edad, es decir, los factores de riesgo de este padecimiento tienden a disminuir con la edad.

En el artículo "Redes neuronales profundas para el diagnóstico de caries utilizando características socioeconómicas y nutricionales como determinantes" [70] se presenta el análisis del conjunto de descriptores demográficos y dietéticos a través de una RNA densa, con el fin de encontrar un modelo generalizado que permitiera encontrar la relación entre estos descriptores y dos posibles estados, ausencia o, presencia o restauración de caries dental. El modelo es posteriormente evaluado con un enfoque de validación cruzada y con un análisis estadístico.

Los resultados dan a conocer que el modelo obtenido tiene un comportamiento generalizado, ya que el rendimiento de la clasificación de los sujetos es similar en el conjunto de datos de entrenamiento y en el conjunto de datos de prueba. Además, la evaluación llevada a cabo a través del análisis estadístico valida el comportamiento del modelo, presentando valores estadísticamente significativos, siendo posible clasificar sujetos con ausencia o presencia de esta condición utilizando esta arquitectura de RNA, basándose en datos demográficos y dietéticos.

Finalmente, en el artículo "Deep Artificial Neural Networks for the Diagnostic of Caries Using Socioeconomic and Nutritional Features as Determinants: Data from NHANES

2013–2014" [60] se presenta el análisis del conjunto de determinantes demográficos y dietéticos utilizando una RNA densa, con el fin de encontrar un modelo generalizado que permitiera la relación de los mismos con sujetos que presentaban ausencia de caries o, presencia o restauración. Como evaluación se utilizó el método de validación cruzada utilizando una serie de parámetros que permitiera medir el rendimiento del modelo.

Los resultados obtenidos permitieron conocer que la cantidad de datos, al igual que las características de la RNA fueron adecuados para el análisis llevado a cabo, tomando en cuenta que las RNAs son específicas para cada caso en particular.

De acuerdo con la precisión y a la función de pérdida que presentó la RNA a lo largo de las épocas, es posible saber que la clasificación de los sujetos mejoró a través del entrenamiento y que el modelo alcanzó un nivel de generalización adecuado para poder diferenciar entre controles y casos.

El rendimiento que presentó la RNA fue confirmado a través de los datos calculados con la validación, siendo todos estadísticamente significativos.

Finalmente, el resultado de este estudio confirma que los datos demográficos y dietéticos sometidos a la RNA, permiten distinguir a sujetos con presencia o restauración de caries de aquellos que no presentan estas condiciones de manera generalizada.

De acuerdo con las diferentes metodologías que se muestran en los artículos de investigación, es posible identificar resultados estadísticamente significativos, aprobando de esta manera la hipótesis propuesta inicialmente. Sin embargo, algunas metodologías presentan ventajas sobre otras.

De acuerdo con el trabajo presentado en [41], se observa que las características tienen un mejor comportamiento de manera multivariada que de manera univariada al desarrollar una serie de modelos; sin embargo, a pesar de presentar un valor de AUC y un p-valor significativos, los resultados se ven afectados por una posible confusión entre dos condiciones, restauración y presencia de caries. Por otro lado, en [3], donde se hace una selección de características a través de la técnica de selección rápida hacia adelante, buscando el desarrollo de dos modelos, uno contenido por determinantes demográficos y el otro contenido por determinantes dietéticos, con el fin de clasificar a los sujetos en dos

condiciones, uniendo a los sujetos con restauración y presencia en una misma salida y a los sujetos con ausencia en otra, se logran mejorar los valores de AUC en comparación con el estudio anterior. En [69] la selección de características se lleva a cabo a través de un AG, buscando clasificar a los sujetos en dos salidas, en donde el valor de AUC se vio incrementado, mientras que el número de características necesarias para la clasificación se redujo. Después, en [67] se lleva a cabo un preprocesamiento inicial en donde los sujetos fueron separados en cuatro grupos de edad, obteniendo así un modelo específico para cada grupo, en donde, de acuerdo con los valores de AUC obtenidos, la relación de sensibilidad - especificidad mejoró de manera significativa, presentando la precisión más alta dentro de todos los trabajos; sin embargo, es importante mencionar que también se presenta la mayor cantidad de características dentro de cada uno de los modelos, por lo que puede implicar un mayor costo computacional.

Finalmente, en [70] y [60] se proponen dos estructuras de RNA a las que se somete el conjunto completo de determinantes demográficos y dietéticos. Ambas presentan valores de precisión y de función de pérdida semejantes; sin embargo, en [60] se presenta un análisis en donde se hace una serie de pruebas para la selección de la cantidad de capas, neuronas y épocas, eligiendo los determinantes que permitieron hacer más eficiente a la RNA, con el menor costo computacional, logrando así reducir estas variables, manteniendo una precisión estadísticamente significativa. No obstante, en comparación con los trabajos anteriores, la RNA no presenta un aporte significativo en su precisión para el enfoque de este estudio.

Es conveniente señalar que dentro de los determinantes seleccionados hay similitudes en todos los trabajos mencionados anteriormente. El conjunto de descriptores demográficos presenta énfasis en aquellos relacionados con el nivel socio-económico, principalmente, lo cual se explica debido a que, de acuerdo con la literatura, las regiones menos desarrolladas y por ende, con menor ingreso, son aquellas que presentan mayores problemas de salud oral ya que no se cuenta con el recurso económico para llevar a cabo los tratamientos, además de tener un acceso restringido a los sistemas de salud. Por otro lado, el conjunto de descriptores dietéticos presenta relevancia en la ingesta diaria de nutrientes totales y

de comida, característica que está directamente relacionada con el nivel socio-económico, ya que la ingesta de una dieta completa se puede ver afectada por el ingreso económico, y ésta a su vez afecta el nivel nutricional, debilitando los sistemas del cuerpo y haciéndolos vulnerables a diferentes condiciones, entre las que se encuentra la caries dental.

5. Conclusiones

En este trabajo se presenta la recopilación de una serie de investigaciones que se realizaron para llevar a cabo diferentes metodologías con el fin de desarrollar un modelo que permitiera relacionar un conjunto de descriptores demográficos y dietéticos, obtenidos de NHANES 2013 – 2014, con la restauración o actual padecimiento de caries dental, clasificándolos de sujetos control.

De acuerdo con los diferentes resultados obtenidos, es posible concluir que existe una evidente relación entre la demografía y la dieta de los sujetos con el estado del padecimiento de caries dental; sin embargo, hay un conjunto de estos descriptores que presentan resultados más significativos, es decir, que brindan el mayor aporte para lograr clasificar sujetos caso y control con el menor error logrado a través de las diferentes técnicas de IA implementadas.

Las técnicas utilizadas para la selección de descriptores, clasificación de sujetos y validación de resultados permitieron desarrollar una serie de modelos multivariados que clasifican la condición de presencia o ausencia de caries dental con una precisión estadísticamente significativa, concluyendo a través de éstos que, si los sujetos son clasificados inicialmente por grupos de edad, es posible desarrollar un modelo específico contenido por las características que afectan en mayor proporción a los sujetos de acuerdo con la etapa de edad en la que se encuentran.

Por otro lado, si la selección de los descriptores utilizados en estos trabajos se realiza por medio de un AG, el conjunto de características obtenido permite desarrollar un modelo con una precisión para clasificar sujetos similar a la precisión promedio obtenida cuando se hace un preprocesamiento inicial clasificando a los sujetos por grupos de edad,

generando así un modelo general que puede identificar sujetos caso de sujetos control independientemente de su etapa de edad, además de presentar una cantidad menor de características que las contenidas en los modelos desarrollados para cada grupo de edad, reduciendo así el costo computacional.

Además, los resultados enfocados en los descriptores demográficos permiten conocer que las características relacionadas con el nivel socio-económico, tales como el nivel académico, la cantidad de personas que habitan el hogar y el ingreso económico, al igual que la edad, el género, el origen étnico y el estado civil, son las que brindan la información más relevante al diferenciar a sujetos con presencia pasada o actual de caries de sujetos que no han estado relacionados con esta condición.

Los resultados relacionados con los descriptores dietéticos se enfocan en las características relacionadas con la dieta general diaria, la ingesta de sal, calorías, proteínas, carbohidratos, azúcares, ácidos grasos, ácido fólico, vitaminas y minerales en el agua, ya que son las que presentan la mejor aptitud para la clasificación de sujetos.

Por lo tanto, el desarrollo de una serie de metodologías que de manera sistemática analizan una serie de descriptores demográficos y dietéticos a través de métricas computacionales y estadísticas, tiene la capacidad de encontrar las asociaciones con la condición de caries dental que de manera individual podrían resultar nulas.

El uso de herramientas de IA, enfocadas en bioinformática, presentan un potencial significativo en el estudio de diversas condiciones y enfermedades, ya que permiten el manejo de una gran cantidad de información y encontrar patrones que pueden no resultar evidentes.

De esta manera, a través de este trabajo se presenta el desarrollo preliminar de una herramienta que puede fungir como apoyo para los médicos especialistas en el diagnóstico de caries dental, tomando en cuenta que, efectivamente, tanto los datos demográficos como los dietéticos, son de alto impacto en la salud oral, también presentando un posible aporte en el diagnóstico preventivo y pronóstico de este padecimiento.

Además, con el seguimiento de este trabajo podría presentarse una posible opción de bajo costo para las regiones menos desarrolladas y con menor acceso a servicios de salud, ya

que el fin de esta investigación preliminar esta enfocada a no requerir de software o hardware especializado para su implementación.

A. Descripción de descriptores demográficos

Tabla A-1.: Descripción de descriptores demográficos.

Descriptor	Descripción
AIALANGA	Idioma del instrumento usado en la entrevista MEC ACASI.
DMDBORN4	¿En qué país nació la persona de muestra?
DMDCITZN	¿La persona de muestra es ciudadana de E.U.A.?
DMDHHSIZ	Número total de gente en el hogar.
DMDHHSZE	Número de adultos de 60 años o mayores en el hogar.
DMDEDUC2	Mayor grado o nivel escolar completado o el mayor grado recibido por participantes de 20 años o mayores.
DMDHHSZB	Número de niños de 6 a 17 años en el hogar.
DMDHSEDU	Nivel escolar de la pareja de la persona de referencia del hogar.
DMDMARTL	Estado civil.
DMDFMSIZ	Número total de gente en la familia.
DMDHHSZA	Número de niños de cinco años o menores en el hogar.
DMDHRGND	Género de la persona de referencia del hogar.
DMDEDUC3	Mayor grado o nivel escolar completado o el mayor grado recibido por participantes de 6 a 19 años.
DMDHRAGE	Edad en años de persona de referencia del hogar.
DMDHRBR4	País de nacimiento de la persona de referencia del hogar.
DMDHREDU	Nivel escolar de la persona de referencia del hogar.
DMDHRMAR	Estado civil de la persona de referencia del hogar.
DMQADFC	¿La persona de muestra sirvió alguna vez en un país extranjero durante una época de conflicto armado o en una misión humanitaria o de mantenimiento de la paz?
DMDYRSUS	Periodo de tiempo que el participante ha estado en E.U.A.
WTLAI8YR	Peso total de la entrevista de 8 años para los participantes en Los Ángeles, California.
WTLAM8YR	Peso total del examen MEC de 8 años para participantes en Los Ángeles, California.

Tabla A-2.: Descripción de descriptores demográficos.

Descriptor	Descripción
LATVPSU	Unidad de varianza: Unidades de pseudomuestreo primario para la estimación de la varianza.
LATVSTRA	Unidad de varianza: variable de estrato para la estimación de la varianza
MIAINTRP	¿Se utilizó un intérprete en la entrevista MEC CAPI?
MIALANG	Idioma del instrumento usado para la entrevista MEC CAPI.
MIAPROXY	¿Se utilizó un encuestado Proxy en la entrevista MEC CAPI?
RIAGENDR	Género del participante.
RISK.GROUP[a]	Grupo de riesgo al que pertenece la persona de referencia del hogar.
SIALANG	Idioma (inglés o español) utilizado durante la entrevista.
SIAPROXY	Indica si se utilizó un informante proxy durante la entrevista.
SIAINTRP	Indica si se utilizó un intérprete durante la entrevista.
WTINT2YR	Peso de muestra de 2 años que deben de ser usados para todos los análisis de NHANES 2013–2014.
WTMEC2YR	Peso de muestra de 2 años que deben de ser usados para todos los análisis de NHANES 2013–2014.
RIDAGEEX	Edad en años del participante en el momento del examen. Las personas de 959 meses y mayores tienen un código superior a los 959 meses.
RIDAGEMN	Edad en meses del participante en el momento de la selección. Informado para personas de 24 meses o menos en el momento del examen.
RIDRETH3	Recolección de información de raza e información de origen hispano, con categoría asiática no hispana.
RIDEXMON	Periodo de seis meses de cuando se realizó el examen.
RIDEXPRG	Estado de embarazo para mujeres entre 20 y 44 años de edad al momento del examen MEC.
RIDAGEYR	Edad en años del participante al momento del estudio.
RIDEXAGM	Edad en meses del participante al momento de la examinación.
RIDSTATR	Estado de la entrevista y examinación del participante.
SDMVPSU	Variable pseudo-PSU de la unidad de varianza enmascarada para la estimación de la varianza.
SDDSRVYR	Ciclo de lanzamiento de datos.
SDMVSTRA	Variable de pseudoestrato de la unidad de varianza enmascarada para la estimación de la varianza.
WTLAF8YR	Submuestra 8 años de peso en ayunas para participantes de 12 años en adelante que fueron examinados en las sesiones de la mañana en Los Ángeles, California.

[a]Los grupos de riesgo se dividen por rangos de edad, grupo 1: 0-9 años, grupo 2: 10-19 años, grupo 3: 20-59 años, grupo 4: 60 años o más.

Tabla A-3.: Descripción de descriptores demográficos.

Descriptor	**Descripción**
DMQMILIZ	¿La persona de muestra ha servido en servicio activo a las fuerzas armadas, reservas militares, o a la guardia nacional de E.U.A.?
FIAINTRP	¿Se utilizó un intérprete en la entrevista familiar?
FIALANG	Idioma del instrumento usado en la entrevista familiar.
FIAPROXY	¿Se utilizó un encuestado Proxy en la entrevista familiar?
INDFMPIR	Una relación entre el ingreso familiar y las pautas de pobreza.
INDHHIN2	Ingreso anual del hogar (reportado en un rango de valores en dólares).
INDFMIN2	Ingreso total del hogar cuando hay más de una familia o no familiares (reportado en un rango de valores en dólares).
RIDRETH1	Recolección de información de raza e información de origen hispano.

B. Descripción de descriptores dietéticos

Tabla B-1.: Descripción de descriptores dietéticos.

Descriptor	Descripción
WTDRD1	Peso de la muestra dietética de un día.
DBQ095Z	Tipo de sal que se agrega usualmente a la comida en la mesa.
DR1TKCAL	Energía (kcal).
DR1TSUGR	Azúcares totales (g).
DR1TFIBE	Fibra dietética (g).
DR1TSFAT	Total de ácidos grasos saturados (g).
DR1TMFAT	Total de ácidos grasos monosaturados (g).
DR1TPFAT	Total de ácidos grasos polisaturadoss (g).
DR1TLYCO	Licopeno (μg).
DR1TFA	Ácido fólico (μg).
DR1TB12A	Vitamina B12 agregada (μg).
DR1_300	¿La cantidad de comida ingerida ayer fue mucho más de lo normal, usual o mucho menos de lo normal?
DR1_320Z	Total de agua potable bebida ayer, incluyendo el agua del grifo, el agua de una fuente para beber, el agua de un enfriador de agua, el agua embotellada y el agua de manantial.
DR1_330Z	Total de agua potable bebida ayer, incluyendo el agua filtrada del grifo y agua de una fuente para beber.
DR1BWATZ	Total agua filtrada del grifo bebida ayer (g).
DR1DAY	Día de admisión de la semana.
DR1LANG	¿En qué idioma habló principalmente el encuestado?
DR1DBIH	Número de días entre el día de admisión y el día del cuestionario familiar administrado en el hogar.
DR1MNRSP	¿Quién fue el principal encuestado en la entrevista?
DR1HELPD	¿Quién ayudó a responder en la entrevista?
DR1SKY	¿Qué tipo de sal era? (¿Era sal común o condimentada, sal ligera o un sustituto de la sal?)?.
DR1STY	¿La persona de muestra agregó sal a su comida ayer?
DR1TLZ	Luteína + zeaxantina (μg).
DR1TM161	Ácidos grasos monosaturados 16:1 (Hexadecenoico) (g).
DR1TM181	Ácidos grasos monosaturados 18:1 (Octadecenoico) (g).

Tabla B-2.: Descripción de descriptores dietéticos.

Descriptor	Descripción
DR1TM201	Ácidos grasos monosaturados 20:1 (Eicosenoico) (g).
DR1TM221	Ácidos grasos monosaturados 22:1 (Docosenoico) (g).
DR1TMAGN	Magnesio (mg).
DR1TMOIS	Humedad (g).
DR1TNIAC	Niacina (mg).
DR1TNUMF	Total de alimentos/bebidas reportados en el archivos de las comidas individuales.
DR1TP182	Ácidos grasos polisaturados 18:2 (Octadecadienoico) (g).
DR1TP183	Ácidos grasos polisaturados 18:3 (Octadecatrienoico) (g).
DR1TP184	Ácidos grasos polisaturados 18:4 (Octadecatetraenoico) (g).
DR1TP204	Ácidos grasos polisaturados 20:4 (Eicosatetraenoico) (g).
DR1TP205	Ácidos grasos polisaturados 20:5 (Eicosapentaenoico) (g).
DR1TP225	Ácidos grasos polisaturados 22:5 (Docosapentaenoico) (g).
DR1TP226	Ácidos grasos polisaturados 22:6 (Docosahexaenoico) (g).
DR1TPOTA	Potasio (mg).
DR1TRET	Retinol (μg).
DR1TS040	Ácidos grasos saturados 4:0 (Butanoico) (g).
DR1TS060	Ácidos grasos saturados 6:0 (Hexanoico) (g).
DR1TS100	Ácidos grasos saturados 10:0 (Decanoico) (g).
DR1TS120	Ácidos grasos saturados 12:0 (Dodecanoico) (g).
DR1TS140	Ácidos grasos saturados 14:0 (Tetradecanoico) (g).
DR1TS180	Ácidos grasos saturados 18:0 (Octadecanoico) (g).
DR1TSELE	Selenio (μg).
DR1TSODI	Sodio (mg).
DR1TTHEO	Teobromina (mg).
DR1TVARA	Vitamina A como equivalentes de actividad de retinol (μg).
DR1TVB1	Tiamina (Vitamina B1) (mg).
DR1TVB12	Vitamina B12 (μg).
DR1TVB2	Riboflavina (Vitamina B2) (mg).
DR1TVB6	Vitamina B6 (mg).
DR1TVC	Vitamina C (mg).
DR1TVK	Vitamina K (μg).
DR1EXMER	Código del entrevistador.
DR1TWS	Cuando se bebe agua del grifo, ¿cuál es la fuente principal del agua del grifo? ¿Es el suministro de agua de la ciudad; un pozo o cisterna de lluvia; un resorte, o algo mas?
DRD370S	Walleye comido durante los últimos 30 días.
DRD370SQ	Número de veces que se comió walleye en los últimos 30 días.
DR1TZINC	Zinc (mg).
DRABF	Indique si la persona de muestra era un bebé que fue amamantado en cualquiera de los dos días de recuperación.
DRD340	Durante los últimos 30 días, ¿se ingirió algún tipo de mariscos enumerados en esta tarjeta?
DRD350A	Almejas comidas durante los últimos 30 días.
DRD350AQ	Número de veces que se comieron almejas en los últimos 30 días.
DRD350B	Cangrejos comidos durante los últimos 30 días.
DRD350BQ	Número de veces que se comió cangrejo en los últimos 30 días.
DRD350C	Cangrejo de río comido durante los últimos 30 días.
DRD350CQ	Número de veces que se comió cangrejos en los últimos 30 días.

Tabla B-3.: Descripción de descriptores dietéticos.

Descriptor	Descripción
DRD350D	Langostas comidas durante los últimos 30 días.
DRD350DQ	Número de veces que se comió langosta en los últimos 30 días.
DRD350E	Mejillones comidos durante los últimos 30 días.
DRD350EQ	Número de veces que se comieron mejillones en los últimos 30 días.
DRD350F	Ostras comidas durante los últimos 30 días.
DRD350FQ	Número de veces que se comieron ostras en los últimos 30 días.
DRD350G	Vieiras comidas durante los últimos 30 días.
DRD350GQ	Número de veces que se comieron vieiras en los últimos 30 días.
DRD350H	Camarones comidos durante los últimos 30 días.
DRD350HQ	Cantidad de veces que se comió camarones en los últimos 30 días.
DRD350I	Otros mariscos (por ejemplo, pulpo, calamar) comidos durante los últimos 30 días.
DRD350IQ	Número de veces que otros mariscos (por ejemplo, pulpo, calamar) se comieron en los últimos 30 días.
DRD350J	Otros mariscos desconocidos comidos durante los últimos 30 días.
DRD350JQ	Número de veces que otros mariscos desconocidos se comieron en los últimos 30 días.
DRD350K	Se negó a dar información detallada sobre los mariscos consumidos durante los últimos 30 días.
DRD360	Durante los últimos 30 días, ¿comió algún tipo de pescado en esta tarjeta?
DRD370A	Productos de pescado empanados durante los últimos 30 días.
DRD370AQ	Número de veces que se comieron productos de pescado empanados en los últimos 30 días.
DRD370B	Atún comido durante los últimos 30 días.
DRD370BQ	Número de veces que se comió atún en los últimos 30 días.
DRD370C	Bajo comido durante los últimos 30 días.
DRD370CQ	Número de veces que se comió bajo en los últimos 30 días.
DRD370D	Bagre comido durante los últimos 30 días.
DRD370DQ	Número de veces que se comió bagre en los últimos 30 días.
DRD370I	Perca comido durante los últimos 30 días.
DRD370IQ	Número de veces que se comió perca en los últimos 30 días.
DRD370K	Pollock comido durante los últimos 30 días.
DRD370KQ	Número de veces que se comió colín en los últimos 30 días.
DRD370T	Otro tipo de pescado comido durante los últimos 30 días.
DRD370TQ	Número de veces que se consumió otro tipo de pescado en los últimos 30 días.
DRQSDIET	¿Actualmente se tiene algún tipo de dieta, ya sea para perder peso o por alguna otra razón relacionada con la salud?
DRQSDT1	¿En qué tipo de dieta estás? (¿Es una dieta para bajar de peso o baja en calorías?: dieta baja en grasas o colesterol; dieta baja en sal o sodio; dieta libre de azúcar o baja en azúcar; dieta baja en fibra; dieta alta en fibra; dieta para la diabetes; u otro tipo) de la dieta?).
DRD370L	Porgy comido durante los últimos 30 días.
DRD370LQ	Número de veces que se consumió porgy en los últimos 30 días.
DRD370M	Salmón comido durante los últimos 30 días.
DRD370MQ	Número de veces que se comió salmón en los últimos 30 días.
DRD370N	Sardinas comidas durante los últimos 30 días.
DRD370NQ	Número de veces que se comieron sardinas en los últimos 30 días.

Tabla B-4.: Descripción de descriptores dietéticos.

Descriptor	**Descripción**
DRD370U	Otro tipo de pescado desconocido comido durante los últimos 30 días.
WTDR2D	Peso de la muestra dietética de dos días.
DR1DRSTZ	Estado de recuperación dietética.
DR1TPROT	Proteínas (g).
DR1TCARB	Carbohidratos (g).
DR1TTFAT	Grasa total (g).
DR1TCHOL	Colesterol (mg).
DR1TFA	Ácido fólico (μg).
DR1TCHL	Colina total (mg).
DR1TIRON	Hierro (mg).
DBD100	Frecuencia con la que se agrega sal ordinaria a la comida en la mesa.
DRQSPREP	Frecuencia con la que se agrega sal ordinario o sazonada al cocinar o preparar comida en el hogar.
DR1TALCO	Alcohol (g).
DR1TATOC	Vitamina E como alfa-tocoferol (mg).
DR1TATOA	Alfa-tocoferol agregado (vitamina E) (mg).
DR1TVD	Vitamina D (D2 + D3) (μg).
DR1TPHOS	Fósforo (mg).
DRD370UQ	Número de veces que se consumió otro tipo de pescado desconocido en los últimos 30 días.
DRD370V	Se negó a dar información detallada sobre los pescados consumidos durante los últimos 30 días.
DRDINT	Indica si la persona de la muestra tiene datos de ingesta durante uno o dos días.
DR1TACAR	Alfa-caroteno (μg).
DR1TBCAR	Beta-caroteno (μg).
DR1TCAFF	Cafeína (mg).
DR1TCALC	Calcio (mg).
DR1TCOPP	Cobre (mg).
DR1TCRYP	Beta-criptoxantina (μg).
DR1TFDFE	Folato como equivalente de folato en la dieta (μg).
DR1TFF	Folato para alimentos (μg).
DR1TFOLA	Folato total (μg).

Bibliografía

[1] World Health Organization (NCHS). Oral health. accessed on 17 September 2018.

[2] D. Ridao Marín. Desarrollo de un sistema de ayuda a la decisión para tratamientos odontológicos con imágenes digitalesantes sociales del estado de salud oral en el contexto actual. *Universidad de Málaga: Málaga, Spain*, pages 10–12, 2017.

[3] Laura A Zanella-Calzada, Carlos E Galván-Tejada, Nubia M Chávez-Lamas, Jorge I Galván-Tejada, and José M Celaya-Padilla. Multivariate features selection from demographic and dietary descriptors as caries risk determinants in oral health diagnosis: Data from nhanes 2013–2014. In *Electronics, Communications and Computers (CONIELECOMP), 2018 International Conference on*, pages 217–224. IEEE, 2018.

[4] Jeannette Ramírez Mendoza, Marco A Rueda Ventura, Manuel Higinio Morales García, and Alicia Gallegos Ramírez. Prevalencia de caries dental y maloclusiones en escolares de tabasco, méxico. *Horizonte Sanitario*, 11(1):13–23, 2014.

[5] R.A. Espinoza Solano, M.; León-Manco. Prevalencia y experiencia de caries dental en estudiantes según facultades de una universidad particular peruana. volume 25, pages 217–224, 2015.

[6] D.; Ríos Rueda L.D. Acuña Aguilar, L.D.; Porras Cerón. Prevalencia de lesiones cariosas y factores asociados presentes en pacientes con síndrome de down en las fundaciones fundown y san luis guanella de bucaramanga. pages 12–16, 2017.

[7] Estela de los Ángeles Gispert Abreu, Pastor Castell-Florit Serrate, and Mirtha He-

rrera Nordet. Salud bucal poblacional y su producción intersectorial. *Revista Cubana de Estomatología*, 52:62–67, 2015.

[8] Ronald P Strauss. The dental impact profile. *Measuring oral health and quality of life*, 13:81, 1996.

[9] Anna T Leao and Aubrey Sheiham. The dental impact on daily living. *Measuring Oral Quality of Life*, 1997.

[10] Gary D Slade and A John Spencer. Development and evaluation of the oral health impact profile. *Community dental health*, 11(1):3–11, 1994.

[11] Supreda Adulyanon and Aubrey Sheiham. Oral impacts on daily performances. *Measuring oral health and quality of life*, 151:160, 1997.

[12] C Fernández González, L Núñez Franz, and N Díaz Sanzana. Determinantes de salud oral en población de 12 años. *Revista clínica de periodoncia, implantología y rehabilitación oral*, 4(3):117–121, 2011.

[13] Tania Camila Niño Sandoval, Sonia Victoria Guevara Pérez, Fabio Augusto González, Robinson Andrés Jaque, and Clementina Infante Contreras. Uso de redes neuronales artificiales en predicción de morfología mandibular a través de variables craneomaxilares en una vista posteroanterior. *Universitas Odontológica*, 35(74), 2016.

[14] Leping Li, Clarice R Weinberg, Thomas A Darden, and Lee G Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method. *Bioinformatics*, 17(12):1131–1142, 2001.

[15] World Health Organization (NCHS). Oral health. accessed on 24 September 2018.

[16] Inara Pereira da Cunha, Antônio Carlos Pereira, Antônio Carlos Frias, Vladen Vieira, Marcelo de Castro Meneghim, Marília Jesus Batista, Karine Laura Cortellazzi,

and Jaqueline Vilela Bulgareli. Social vulnerability and factors associated with oral impact on daily performance among adolescents. *Health and quality of life outcomes*, 15(1):173, 2017.

[17] Kitty Jieyi Chen, Sherry Shiqian Gao, Duangporn Duangthip, Samantha Kar Yan Li, Edward Chin Man Lo, and Chun Hung Chu. Dental caries status and its associated factors among 5-year-old hong kong children: a cross-sectional study. *BMC oral health*, 17(1):121, 2017.

[18] S Vega-López, NM Lindberg, GJ Eckert, EL Nicholson, and G Maupomé. Association of added sugar intake and caries-related experiences among individuals of mexican origin. *Community dentistry and oral epidemiology*, 2018.

[19] Dan Henry Levy, Alon Livny, Harold Sgan-Cohen, and Nirit Yavnai. The association between caries related treatment needs and socio-demographic variables among young israeli adults: a record based cross sectional study. *Israel journal of health policy research*, 7(1):24, 2018.

[20] LL Wu, KY Cheung, PYP Lam, and Xiaoli Gao. Oral health indicators for risk of malnutrition in elders. *The journal of nutrition, health & aging*, 22(2):254–261, 2018.

[21] TA Oyedele, AD Fadeju, YI Adeyemo, CL Nzomiwu, and AM Ladeji. Impact of oral hygiene and socio-demographic factors on dental caries in a suburban population in nigeria. *European Archives of Paediatric Dentistry*, 19(3):155–161, 2018.

[22] Tariq S Ghazal, Steven M Levy, Noel K Childers, Knute D Carter, Daniel J Caplan, John J Warren, and Justine L Kolker. Survival analysis of caries incidence in african-american school-aged children. *Journal of public health dentistry*, 2018.

[23] Gomes MC Barbosa Neves ÉT Barbosa AS de Medeiros CA de Aquino MM Granville-Garcia AF de Menezes VA Brito ÁS, Clementino MA. Sociodemographic

and behavioral factors associated with dental caries in preschool children: Analysis using a decision tree. *Journal of Indian Society of Pedodontics and Preventive Dentistry*, 36(3):244–249, 2018.

[24] World Health Organization (NCHS). Health. accessed on 24 September 2018.

[25] Miriam Alveza Treviño Tamez, Liliana Tijerina de Mendoza, Esteban Gilberto Ramos Peña, and Pedro César Cantú Martínez. Salud bucodental en escolares de estrato social bajo. *RESPYN*, 6(2), 2017.

[26] World Health Organization. *Oral health surveys: basic methods*. World Health Organization, 2013.

[27] Tahiris Paneque Escalona, Héctor Rafael Castillo Ortiz, Yoannis Piquera Palomino, Mauren Infante Tamayo, and María Isabel Ramírez Rodríguez. 08-relación entre factores de riesgos y caries dental relationship between risk factors and dental caries. *MULTIMED Revista Médica Granma*, 19(4), 2017.

[28] Shyrley Díaz-Cárdenas, Madera Anaya Meisser-Vidal, Lesbia Rosa Tirado-Amador, Natalia Fortich-Mesa, Liliana Tapias-Torrado, and Farith Damián González-Martínez. Impacto de salud oral sobre calidad de vida en adultos jóvenes de clínicas odontológicas universitarias. *International journal of odontostomatology*, 11(1):5–11, 2017.

[29] Sanitas. Anatomía del diente. accessed on 04 December 2018.

[30] John Timothy Wright. The burden and management of dental caries in older children. *Pediatric Clinics*, 65(5):955–963, 2018.

[31] Patcharawan Srisilapanan, Narumanas Korwanich, Sutha Jienmaneechotchai, Supranee Dalodom, Nontalee Veerachai, Warangkana Vejvitee, and Jeffrey Roseman. Estimate of impact on the oral health-related quality of life of older thai people by the provision of dentures through the royal project. *International journal of dentistry*, 2016, 2016.

[32] Amal Elamin, Malin Garemo, and Andrew Gardner. Dental caries and their association with socioeconomic characteristics, oral hygiene practices and eating habits among preschool children in abu dhabi, united arab emirates—the noplas project. *BMC oral health*, 18(1):104, 2018.

[33] ECM Lo. Caries process and prevention strategies: epidemiology. *USA: Crest®+ Oral-B® at dentalcare. com Continuing Education Course, Recuperado de www. dentalcare. com/en-US/dental-education/continuingeducation/ce368/ce368. aspx*, 2014.

[34] Castellares Espinoza, Dacy Fiorella, Pedro Martín Ramos Mejía, et al. Asociación del índice de masa corporal con la presencia de caries dental en escolares de 6 a 12 años. 2017.

[35] SIMONE FALEIROS, ANDREA ORMEÑO, MAYERLING PINTO, REBECA TAPIA, CARLOS DÍAZ, HERNÁN GARCÍA, et al. Prevalencia de caries en alumnos de educación básica y su asociación con el estado nutricional. *Revista chilena de pediatría*, 81(1):28–36, 2010.

[36] Edwina AM Kidd and Ole Fejerskov. *Essentials of dental caries*. Oxford University Press, 2016.

[37] Centers for Disease Control and National Center for Health Statistics (NCHS) Prevention (CDC). National health and nutrition examination survey data. accessed on 20 September 2018.

[38] Gemma L Nedjati-Gilani, Torben Schneider, Matt G Hall, Claudia AM Wheeler-Kingshott, and Daniel C Alexander. Machine learning based compartment models with permeability for white matter microstructure imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 257–264. Springer, 2014.

[39] JI Galván-Tejada, A Martinez-Torteya, S Totterman, J Farber, V Treviño, and

JG Tamez-Pena. A wide association study of predictors of future knee pain: data from the osteoarthritis initiative. *Osteoarthritis and Cartilage*, 20:S85, 2012.

[40] JI Galván-Tejada, JG Arceo-Olague, H Luna-García, H Gamboa-Rosales, JM Celaya-Padilla, LA Zanella-Calzada, R Magallanes-Quintanar, and CE Galván-Tejada. General linear models for pain prediction in knee osteoarthritis: Data from the osteoarthritis initiative. *REVISTA MEXICANA DE INGENIERÍA BIOMÉDICA*, 39(1):30, 2018.

[41] Nubia M Chávez-Lamas, Laura A Zanella-Calzada, and Carlos Eric Galván-Tejada. An analysis of dietary and demographic data in oral health, data from the national health and nutrition examination survey: A preliminary study. *Research in Computing Science*, 142:79–88, 2017.

[42] Carlos D Luna-Gómez, Laura A Zanella-Calzada, Miguel A Acosta-García, Jorge I Galván-Tejada, Carlos E Galván-Tejada, and José M Celaya-Padilla. Can multivariate models based on moaks predict oa knee pain? data from the osteoarthritis initiative. In *Medical Imaging 2017: Computer-Aided Diagnosis*, volume 10134, page 1013445. International Society for Optics and Photonics, 2017.

[43] JJ Suárez. *Desarrollo de un sistema de diagnóstico asistido por computador para detección de nódulos pulmonares en tomografía computarizada multicorte*. PhD thesis, Tesis PhD, 2009.

[44] Salvatore Gitto, Giorgia Grassi, Chiara De Angelis, Cristian Giuseppe Monaco, Silvana Sdao, Francesco Sardanelli, Luca Maria Sconfienza, and Giovanni Mauri. A computer-aided diagnosis system for the assessment and characterization of low-to-high suspicion thyroid nodules on ultrasound. *La radiologia medica*, pages 1–8, 2018.

[45] Nils J Nilsson. *Principles of artificial intelligence*. Morgan Kaufmann, 2014.

[46] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[47] Paul R Cohen and Edward A Feigenbaum. *The handbook of artificial intelligence*, volume 3. Butterworth-Heinemann, 2014.

[48] Miriam Frey and Mario Larch. Statistical learning in the age of "big data" and machine learning. 2017.

[49] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.

[50] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[51] Richard J Roiger. *Data mining: a tutorial-based primer*. Chapman and Hall/CRC, 2017.

[52] Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

[53] Stephen Marsland. *Machine learning: an algorithmic perspective*. CRC press, 2015.

[54] Fernando Sancho Caparrini. Introducción al aprendizaje automático. 2015.

[55] Angel Garcia, Israel Gonzalez, Ricardo Colomo-Palacios, Jose Luis Lopez, and Belen Ruiz. Methodology for software development estimation optimization based on neural networks. *IEEE Latin America Transactions*, 9(3):384–398, 2011.

[56] Salah Bouktif, Ali Fiaz, Ali Ouni, and Mohamed Serhani. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, 11(7):1636, 2018.

[57] Adam Marczyk. Algoritmos genéticos y computación evolutiva. *Departamento de Informática, universidad de Colorado*, 2004.

[58] Sankar K Pal and Paul P Wang. *Genetic algorithms for pattern recognition*. CRC press, 2017.

[59] Yuichiro Anzai. *Pattern recognition and machine learning*. Elsevier, 2012.

[60] Laura Zanella-Calzada, Carlos Galván-Tejada, Nubia Chávez-Lamas, Jesús Rivas-Gutierrez, Rafael Magallanes-Quintanar, Jose Celaya-Padilla, Jorge Galván-Tejada, and Hamurabi Gamboa-Rosales. Deep artificial neural networks for the diagnostic of caries using socioeconomic and nutritional features as determinants: Data from nhanes 2013–2014. *Bioengineering*, 5(2):47, 2018.

[61] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[62] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

[63] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[64] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[65] Carlos E Galván-Tejada, Laura A Zanella-Calzada, Jorge I Galván-Tejada, José M Celaya-Padilla, Hamurabi Gamboa-Rosales, Idalia Garza-Veloz, and Margarita L Martinez-Fierro. Multivariate feature selection of image descriptors data for breast cancer with computer-assisted diagnosis. *Diagnostics*, 7(1):9, 2017.

[66] Yan Zhang, Danjv Lv, Ran Guo, TG Dietterich, ZH Zhou, C Zhang, Y Ma, LI Kuncheva, L Rokach, ZH Zhou, et al. Data mining: Practical machine learning tools and techniques. *Journal of Software Engineering*, 11(1):97–136, 1997.

[67] Laura A Zanella-Calzada, Carlos E Galván-Tejada, Gracia-Cortés Ma del Carmen Moreno-Báez Arturo Chávez-Lamas, Nubia M, Celaya-Padilla Jose M Arceo-Olague, Jose G, Jorge I Galván-Tejada, and Hamurabi Gamboa-Rosales. A case–control study of socio-economic and nutritional characteristics as determinants of dental caries in different age groups, considered as public health problem: Data

from nhanes 2013–2014. *International journal of environmental research and public health*, 15(5):957, 2018.

[68] Jaime Cerda and Lorena Cifuentes. Uso de curvas roc en investigación clínica: Aspectos teórico-prácticos. *Revista chilena de infectología*, 29(2):138–141, 2012.

[69] Galván-Tejada Carlos E Chávez-Lamas-Nubia M Zanella-Calzada, Laura A, Ma del Carmen Gracia-Cortés, and Jorge I Galván-Tejada. An analsysis of demographic and dietary data with an oral health approach, a preliminary study usign genetic algorithms.

[70] Galván-Tejada Carlos E Chávez-Lamas-Nubia M Zanella-Calzada, Laura A, Ma del Carmen Gracia-Cortés, and Jorge I Galván-Tejada. Redes neuronales profundas para el diagnóstico de caries utilizando características socioeconómicas y nutricionales como determinantes.