# UNIVERSIDAD AUTÓNOMA DE ZACATECAS


## DESARROLLO DE BIOMARCADORES BASADOS EN DETERMINANTES SOCIOECONÓMICOS Y DEMOGRÁFICOS PARA EL DIAGNÓSTICO DE ENFERMEDADES MULTIFACTORIALES NO TRANSMISIBLES POR MEDIO DE REDES NEURONALES PROFUNDAS.

**Vanessa del Rosario Alcalá Ramírez**

Tesis de Maestría

presentada a la Unidad Académica de Ingeniería Eléctrica

de acuerdo a los requerimientos de la Universidad para obtener el Grado de

**MAESTRO EN CIENCIAS DE LA INGENIERÍA**

Directores de tesis: Dr. Carlos Eric Galván Tejada. y Dra. Nubia M. Chávez Lamas

UNIDAD ACADÉMICA DE INGENIERÍA ELÉCTRICA

20 de Septiembre del 2019

# DESARROLLO DE BIOMARCADORES BASADOS EN DETERMINANTES SOCIOECONÓMICOS Y DEMOGRÁFICOS PARA EL DIAGNÓSTICO DE ENFERMEDADES MULTIFACTORIALES NO TRANSMISIBLES POR MEDIO DE REDES NEURONALES PROFUNDAS.

Vanessa del Rosario Alcalá Ramírez

Directores de tesis: Dr. Carlos Eric Galván Tejada. y Dra. Nubia M. Chávez Lamas

RESUMEN

La incidencia y prevalencia de dos enfermedades no transmisibles las cuales son diabetes mellitus y caries denta, está en aumento, motivo principal por el que en los últimos años el estudio de la relación que hay entre estos dos padecimientos ha incrementado. Tanto la caries como la diabetes, son enfermedades multifactoriales, que tienen diversos factores de riesgo que contribuyen su inicio y progresión. Los factores de riesgo pueden clasificarse en biológicos, ambientales o socioconductuales, por mencionar algunos. La caries es problema de salud pública a nivel mundial y es conocida como la enfermedad no transmisible más extendida. La diabetes actualmente está afectando a la población mexicana en niveles preocupantes, ocupando el primer lugar en prevalencia de este padecimiento.

Debido a que ambas enfermedades son prevenibles, en este trabajo se propone ell uso de una Red Nuronal Artificial que sea capaz de clasificar a los sujetos con presencia o ausencia de estas afecciones, utilizando 31 características que describen el estado del paciente. El modelo es evaluado mediante análisis estadístico tomando en cuenta la precisión, función de pérdida, el área bajo la curva (AUC por sus siglas en inglés) y la curva de características operativas receptoras (ROC por sus siglas en inglés). Se obtuvieron resultados estadísticamente significativos tieniendo una precisión de 0.99, AUC de 0.99 para las curvas ROC.

**Palabras clave:** Diabetes mellitus, NHANES 2013-2014, Redes neuronales artificiales, Diagnóstico asistido por computadora, Caries.

# BIOMARKERS DEVELOPMET BASED ON SOCIOECONOMIC AND DEMOGRAPHIC DETERMINANTS FOR THE DIAGNOSTIC OF MULTIFACTORIAL NON-TRANSMISSIBLE DISEASES THROUGH DEEP NEURONAL NETWORKS.

Vanessa del Rosario Alcalá Ramírez

Thesis supervisors: Dr. Carlos Eric Galván Tejada. y Dra. Nubia M. Chávez Lamas

ABSTRACT

The incidence and prevalence of two non-transmissible diseases, there are diabetes mellitus and dental caries, is increasing, for this reason, in the recent years the relationship between dental caries and diabetes mellitus studies are rising. Dental caries and diabetes are multifactorial diseases, they have many risks factors that contribute to the onset and progression. Risk factors can be classified as biological, environmental or socio-behavioral. Caries is a public health problem worldwide and is known as the most widespread noncommunicable disease. Diabetes is currently affecting the Mexican population at worrying levels, occupying the first place in prevalence of this condition.

Due to both diseases are preventable, this work proposes the use of an Artificial Neural Network that is able to classifying subjects with the presence or absence of caries and diabetes, using 31 characteristics that describe the patient's condition.

The model is evaluated by statistical analysis taking into account the accuracy, loss function, the area under the curve (AUC) and the receiver operating characteristics curve (ROC). Statistically significant results were obtained with an accuracy of 0.99, AUC of 0.99 for ROC curves.

**Keywords:** Diabetes mellitus, NHANES 2013-2014, Artificial Neural Network, Computer-aided diagnosis, Caries.

# Dedicatoria.

A mi padre José Luis Alcalá Castro, quien siempre ha confiado en mi.

A mi madre María Luisa Ramírez Barrera, que me ha dado la fortaleza necesaria para cumplir con cada meta.

A mis hermanos; Ana Luisa Alcalá Ramírez, Luis Francisco Alcalá Ramírez y Alejandra Alcalá Ramírez quienes son mi motivación en cada meta de mi vida.

A mi sobrino Luis Zabdiel Alcalá Esquivel, quien me da fuerza y me motiva con cada sonrisa.

A mi novio Misael del Río Torres que siempre me ha apoyado incondicionalmente.

# Agradecimientos

Agradezco por darme cada día motivación a mi padre, así mismo agradezco a mi madre que aunque no esté físicamente conmigo, siempre ha sido fuente de fortaleza en cada paso de mi vida.

De igual manera agradezco a cada uno de mis hermanos, por apoyarme y confíar en mi, al igual a mi sobrino, quien ilumina mi vida día a día.

Agradezco a mi novio por siempre apoyar mis decisiones y estar a mi lado en cada momento que lo necesito.

Por otro lado, agradezco al Dr. Carlos Eric Galván Tejada por todo el apoyo brindado durante mi estancia en la maestría, así como a todos los doctores que fueron parte de mi aprendizaje y de mi crecimiento académico.

Además, agradezco al Centro Médico Siglo XXI por permitirme realizar una estancia académica, en la cual adquirí diferentes conocimientos.

También quiero agradecer a mis amigos y compañeros, por sus consejos, por compartir sus experiencias y conocimientos conmigo y por apoyarme en mi crecimiento personal y profesional.

# Contenido General

Pag.

# Lista de figuras

# Lista de tablas

# Capítulo 1

# Introducción

En este capítulo se explica el concepto general referente a las enfermedades no transmisibles, principalmente diabetes y caries. Dado el alto número de incidencias de ambas enfermedades, es importante encontrar la o las relaciones presentes entre caries y diabetes, para que una vez identificadas sea posible el desarrollo de biomarcadores basados en determinantes socioeconómicos y demográficos que ayuden a la detección temprana de dichos padecimientos, debido a esto, se plantean un conjunto de hipótesis, las cuales se analizan bajo el desarrollo del objetivo general y de los objetivos específicos correspondientes a esta investigación.

Las Enfermedades No Transmisibles (ENT), también conocidas como Enfermedades Crónicas No Transmisibles (ECNT), tienden a ser de larga duración y solo se controlan, son la principal causa de muerte y discapacidad en el mundo, según la Organización Panaméricana de la Salud (OPS) el término ENT hace referencia a un grupo de enfermedades que no son causadas principalmente por una infección aguda, además tienen consecuencias para la salud a largo plazo y con frecuencia crean necesidad de tratamientos y cuidados de por vida [1]. Las ENT son el resultado de la combinación de factores genéticos, fisiológicos, ambientales y conductuales, y afectan a todos los grupos de edad, a todas las regiones y países, centrandose en los países de ingresos bajos y medios, así como en personas de edad avanzada [2].

Según la Dirección Nacional de Promoción de la Salud y Control de Enfermedades No Transmisibles, entre las ENT más comunes se encuentran las enfermedades cardiovasculares,

como pueden ser infartos de miocardio o accidentes cerebrovasculares; el cáncer; las enfermedades respiratorias crónicas tales como neumopatía obstructiva crónica o asma; la enfermedad renal; la diabetes y algunas enfermedades bucodentales [3]. Estas enfermedades representan un 63% de muertes que se producen a nivel mundial, aproximadamente el 80% se encuentran en los países de bajos y medios ingresos afectando por igual a hombres y mujeres [4].

La Organización Mundial de la Salud (OMS) define diabetes como una enfermedad crónica, incurable, dinámica, de agravamiento progresivo y de larga duración, la cual se presenta cuando el páncreas no produce suficiente insulina o cuando el cuerpo no puede usar eficientemente la insulina producida, debido a esto, los niveles de glucosa en la sangre incrementan por lo que el cuerpo libera más cantidad de insulina teniendo como consecuencia que la producción de insulina disminuya constantemente, provocando altos niveles de azúcar en la sangre, lo cual se conoce como hiperglucemia [5, 6].

La diabetes puede clasificarse dentro de las siguientes categorias:

- **Diabetes tipo 1:** Consiste en la destrucción autoinmune de las células $\beta$, lo que generalmente conduce a una deficiencia absoluta de insulina.

- **Diabetes tipo 2:** Ocurre debido a la pérdida progresiva de la secreción de insulina de células $\beta$ con frecuencia provocando resistencia a la insulina.

- **Diabetes Gestacional:** Este tipo de diabetes se genera a partir de la gestación, siendo diagnosticada en el segundo o tercer trimestre del embarazo.

- **Tipos de diabetes específicos debidos a otras causas:** Dentro de estos tipos de diabetes se encuentra la diabetes neonatal, o la diabetes que se presenta al inicio de la madurez en jóvenes, otras causas son fibrosis quística y la pancreatitis, además de la diabetes inducida por drogas o productos químicos, como puede ser el uso de glucocorticoides en el tratamiento del VIH/SIDA o después de un transplante de órganos [7].

.

Dentro de la clasificación de los tipos de diabetes mellitus resaltan dos, las cuales son aquellas que se presentan en la población con mayor frecuencia, estas son la diabetes mellitus tipo 1 y tipo 2, esta última se caracteriza por hiperglucemia con grados variables de resistencia insulínica, disminución de la secreción de insulina e incremento en la producción de glucosa hepática [8, 9].

La diabetes en etapas tempranas no suele producir síntomas, cuando se detecta tardíamente y no se trata adecuadamente ocasiona complicaciones de salud, como pueden ser infartos del corazón; ceguera; falla renal; amputación de extremidades inferiores y muerte prematura [10], lo cual lleva a estimadar que la esperanza de vida de individuos con diabetes se reduce entre 5 y 10 años [11].

Por otro lado, dentro de las enfermedades no transmisibles y especifícamente tratandose de salud oral o enfermedades bucodentales, destaca la caries, se define como un padecimiento de origen multifactorial, con diferentes riegos que contribuyen a su aparición y su progreso, estos factores de riesgo se pueden categorizar en biológicos, ambientales o socio-culturales. Es importante mencionar que la caries es considerada como una enfermedad prevenible no transmisible que se presenta en la mayoría de la población a lo largo de su vida, con una prevalencia que ronda entre el 60 y 90% mundialmente, afectando los tejidos duros de los dientes, la caries comparte una serie de factores conductuales, socioeconómicos y de estilo de vida con otras enfermedades no transmisibles como son la diabetes y el sobrepeso, por lo que de igual manera debería estar sujeto a un modelo similiar de control como las enfermedades crónicas [12, 13, 14].

De acuerdo con la OPS el desarrollo de caries con frecuencia está relacionada con el consumo de carbohidratos, placa cariogénica, saliva, características de la comida, tiempo de exposición, la eliminación de la placa y la susceptibilidad del paciente, así como, las pocas medidas preventivas en la salud oral y el acceso limitado a los servicios dentales especializados [15].

## 1.1 Planteamiento del problema.

Existen grupos de personas de todas las edades, regiones y países que son afectados por enfermedades no transmisibles ya que son altamente vulnerabes a los factores de riesgo que provocan la aparición de estas enfermedades tales como son las dietas no saludables, falta de actividad física, exposición al tabaco y el uso constante del alcohol, según algunos estudios, hay 15 millones de muertes causadas por estas enfermedades en el grupo de edad que abarca desde los 30 años hasta los 69 años [16]. Un factor relevante para que incremente el riesgo de desarrollar ENT son las afecciones orales, las cuales son parte de las enfermedades que afectan la calidad de vida de las personas, según el *Global Burden of Disease Study* realizado en 2016, se estimó que las enfermedades orales afectan al menos a 3.58 billones de personas en el mundo, siendo la caries,a cual se define como una padecimiento de origen multifactorial, el padecimiento con mayor prevalencia de la cual se estima que a nivel global 2.4 billones de personas sufren de caries en los dientes permanentes[17, 18]. El desarrollo de esta enfermedad se concentra principalmente en poblaciones socialmente marginadas, representa un problema debido a que los tratamientos para esta condición son altamente costosos, además, de acuerdo con la OMS, los padecimientos orales ocupan el cuarto lugar dentro de las causas más costosas por tratar [19]. Una vez que se tiene un padecimiento oral, incrementa el riesgo de desarrollar otras ENT tales como problemas cardiovasculares, cerebrovasculares y diabetes mellitus.

La diabetes mellitus también es una ENT y según los datos de la OMS y de la Federación Internacional de Diabetes (FID), el número de personas con este padecimiento está incrementando de manera exponencial en el mundo, siendo hoy en día una de las principales causas de muerte y dicapacidad en el mundo, presentando uno de los más grandes retos del siglo XXI para la salud [19, 20]. La Colaboración de Factor de Riesgo de las ENT estima que el número de personas con diabetes se ha cuadriplicado entre 1980 y 2014, la prevalencia estandarizada por edad entre hombres adultos se duplico durante este tiempo y la prevalencia estandarizada por edad entre mujeres adultas incrementó en un 60% [20]. Además, la diabetes tiene una connotación importante debido a que en el 2017 aproximadamente 451 millones de personas con edades entre los 18 y 99 años padecían diabetes, se predice que esta cifra incremente hasta

los 693 millones de personas para el 2045 [21]. Es importante mencionar que México ocupa el sexto lugar con mayor cantidad de diabéticos, lo que corresponde a 6.4 millones de personas en nuestro país [22].

El problema se centra en que debido a que la caries y la diabetes son ENT multifactoriales, se dificulta el control de incidencia y prevalencia, además de que los diagnósticos no se suelen obtener en etapas tempranas, lo que provoca el desarrollo de otros problemas en el cuerpo humano, en el caso de la caries, una atención tardía puede tener una realción con la presencia de diabetes mellitus, además de que dificulta el diagnóstico y tratamiento de dichos padecimientos, por su lado, la diabetes mellitus tiene consecuencias que van desde ceguera hasta muerte prematura. En ambas afecciones, se requiere de tratamientos continúos administrados tanto por profesionales de la salud como por el mismo paciente o familiares, teniendo como consecuencia la generación de gastos extraordinarios [23].

## 1.2 Justificación.

El control de las enfermedades no transmisibles es actualmente una de las prioridades en la Salud Pública dada la evolución ascendente de la mortalidad que producen y el costo económico, sanitario y social que provoca. Tanto la diabetes como la caries, tienen un origen multifactorial, por lo cual es difícil decir las causas precisas que provocan este padecimento, sin embargo, la implementación de algoritmos y diferentes análisis a través del diagnóstico asistido por computadora (CADx, por sus siglas en inglés) se proponen para dar un soporte al diagnóstico preventivo y la reducción de los altos índices de prevalencia, buscando el desarrollo de modelos de clasificación y predicción que contengan los principales factores que potencializan el desarrollo de estos padecimientos e identificar a la población que se encuentra con mayor riesgo [24, 25].

Por otro lado, según algunos estudios, estos padecimientos suelen presentarse con mayor recurrencia en áreas rurales, por lo que por medio de la generación de modelos que puedan proporcionar información acerca un diágnostico en etapas tempranas de ambas enfemedades se puede generar una herramienta basada en CADx de bajo costo que ayude al sector salud a

proporcionar información relevante del paciente con la finalidad de que se tenga la atención adecuada para el control de incidencias y prevalencias de estas afecciones.

Por medio de algunas técnicas de Inteligencia Artificial (IA), en la presente investigación se propone el ánalisis de factores demográficos y socioeconómicos, con lo cual se pretende encontrar los determinates que proporcionen información relevante sobre ambos padecimientos, y así permita clasificar cuando un paciente padece ambas afecciones o en si en caso contrario, no ha desarrollado ni caries ni diabetes.

## 1.3 Objetivo General

Desarrollar biomarcadores basados en determinantes socioeconómicos, dietéticos y demográficos, obtenidos de las bases de datos de NHANES 2013-2014 que ayuden a la detección oportuna de enfermedades multifactoriales no transmisibles, con el fin de clasificar sujetos presencia de caries y diabetes.

## 1.4 Objetivos Específicos

- Identificar las características que provocan el desarrollo de enfermedades no transmisibles como son caries y diabetes mellitus en el paciente con base en la información de las bases de datos NHANES 2013-2014.
- Desarrollar Redes Neuronales Artificiales Profundas (RNA) de diferentes arquitecturas para la comparación de la precisión en la clasificiación de pacientes con enfermedades no transmisibles.
- Evaluar los modelos, obtenidos de las diferentes arquitecturas de las redes neuronales implementadas, a través de diferentes métricas.
- Analizar y evaluar el modelo con mejor desempeño , buscanado obtener un porcentaje mayor al 95% de precisión en la clasificación de pacientes con enfermedades no transimisibles correspondientes a caries y diabetes mellitus.

## 1.5 Hipótesis

Es posible clasificar con una precisión mayor al 95% aquellos pacientes con enfermedades no transmisibles, tal como son caries y diabetes mellitus, a través de la implementación de Redes Neuronales Artificiales Profundas entrenadas con descriptores demográficos, dietéticos y socioeconómicos, obtenidos de las bases de datos de NHANES 2013-2014.

## 1.6 Estado del Arte

De acuerdo con la literatura, se han realizado diferentes implementaciones tanto para el estudio de caries y diabetes de manera individual como para el estudio de ambas enfermedades simultáneamente por medio de sistemas CADx.

Para el caso específico de diabetes, Carnimeo et al. [26] propone la detección automática de síntomas de diabetes en imagenes de retina haciendo uso de una red neuronal con perceptron multicapa. Este trabajo consiste en entrenar la red neuronal utilizando algoritmos para evaluar el umbral global óptimo que pueda minimizar los errores de clasificación de pixeles. El rendimiento del sistema es evaluado por un índice adecuado para proporcionar una medida porcentual en la detección de regiones sospechosas del ojo basado en el subsistema neuro-difuso. Además, Chen et al. [27] propone un sistema 5G-Smart Diabetes basado en CADx, en el cual se utilian diferentes técnicas de aprendizaje automático y big data para realizar un análisis de pacientes que sufren diabetes. Por otro lado, Cappon et al. [28] desarrolla una herramienta basada en RNA, optimizando y personalizando el cálculo del bolus a través del monitoreo de los niveles de glucosa, obteniendo información útili y accesible sobre los pacientes, y esto permite saber si un paciente desarrolló diabtes.

Dentro de los estudios enfocados a caries, Ghazal et al. [29] se realiza un análisis en un modelado de riesgos multivariado y bivariado para buscar la relación entre las covariables dependientes y no dependientes del tiempo en ralación con la caries de los diantes permanentes entre niños afroamericanos con un estao socio-económico bajo. Lee et al. [30] proponen la detección y diagnóstico de caries haciendo uso del aprendizaje profundo, implementando RNA convolucionales, donde se evaluó la eficiencia de la RNA convolucional en la detección

y dignóstico de caries en radiografías periapicales, se utilizaron 3000 imagenes de las cuales el 80% fueron para entrenar y validar la red neuronal y el 20% para realizar pruebas, de igual manera, se implementó la red pre-entrenada GoogleNet Inception v2. Por otro lado, en la investigación de Brito et al. [31] se propone un estudio que tiene la finalidad de determinar la prevalencia de caries y los factores asociados a niños en Brasil. En dicho estudio se analizaron variables socio-demográficas y de comportamiento por medio de un árbol de decisión inductivo.

Finalmente, hay algunos estudios que involucran el análisis de ambas enfermedades, tal como la investigación de Lai et al. [32] donde se desarrolla una evaluación de la diferencia en la experiencia de caries en niños diabéticos y no diabéticos, obteniendo que se encontró un mayor número de sujetos sin caries en sujetos diabéticos con buen control metabólico ($p < 0.1$ ), lo cual indica que los niños diabéticos con un mal metabolismo tienen riesgo alto de padecer caries. Existe otro estudio presentado en el trabajo de Barylo et al. [33] que se enfoca en los efectos que tiene la diabetes mellitus en pacientes con salud oral a través de un análisis estadístico utilizando pruebas de estudiantes, según los resultados obtenidos, se concluye que la diabetes tiene un efecto directo en la salud oral, mostrando que el nivel de atención dental médica y preventiva debe aumentarse para pacientes con diabetes. Por otro lado en el trabajo presentado por Latti et al. [34] se presenta una evaluación de los efectos de la diabetes mellitus sobre los microorganismos de la caries dental responsables de la caries, a través de una técnica de índice de búsqueda en profundidad (DFS, por sus siglas en inglés). Los resultados mostraron que la caries dental, entre otros factores, aumenta en los diabéticos a comparación de los sujetos control, existiendo una relación entre diabetes mellitus, microbiota oral y caries. En el caso de Ferizi et al. [35] se propone un trabajo basado en CADx, en el cual se busca la influencia de Diabetes Mellitus Tipo 1 (DMT1) sobre la caries. Por medio de una comparación estadística, los datos recopilados de un conjunto de controles y casos (presencia de DMT1 y caries) se analizaron mediante la prueba de *chi-squared test* y la prueba *Mann-Whitney U-test*, obteniendo una relación significativa entre los sujetos con DMT1 y la presencia de caries ($p < 0.001$), siendo posible concluir que la DMT1 tiene una parte importante en la salud oral, ya que parece que las personas con DMT1 presentan un mayor riesgo de caries.

Pinho et al.[36] proponen la evaluación del impacto de problemas orales dependiendo de la calidad de vida de pacientes que presentan DMT2 por medio del análisis de datos basados en estadística descriptica, análisis bivariante y regresión logistica. Los resultados exponen que la prevalencia del impacto en la calidad de vida relacionada con la salud bucal fue del 47%. En el análisis multivariado, las variables que permanecieron significativamente asociadas con un impacto negativo en la calidad de vida fueron la xerostomía, necesidad de dentadura postiza y periodontitis. Por último la investigación de Song et al. [37] explica la asociación entre DMT2 y la caries no tratada, con la finalidad de identificar la diabetes como un factor de riesgo de la caries, los resultados muestran que la prevalencia de caries y diabetes no controlada fue un 26% más alta a comparación de aquellos que tienen una tolerancia normal a la glucosa.

## 1.7 Estructura de la tesis.

El presente trabajo de investigación, está compuesto por 5 capítulos distribuidos como a continuación de explica.

- El capítulo 1 contiene una introducción al trabajo desarrollado, incluyendo la problemática, justificación, hipótesis, objetivos, etc.
- El capítulo 2 contiene un panorama general sobre las enfermedades no transmisibles, a su vez se explican las herramientas y conceptos utilizados para el desarrollo de esta investigación.
- En el capítulo 3 se explica la metodología y experimentación que se llevo a cabo a lo largo del desarrollo del presente trabajo.
- En el capítulo 4 se abarcan los resultados obtenidos y un apartado de discusión sobre los resultados generales del trabjo.
- Finalmente, en el capítulo 6 se encuentran las conclusiones y el trabajo futuro propuesto.

# Capítulo 2

# Redes Neuronales Artificiales enfocadas a enfermedades no transmisibles.

En este capítulo se describirán las circunstancias por las cuales se desarrollan las enfermedades no transmisibles, de esta manera se proporciona un panorama sobre el enfoque en el que se está desarrollando la presente investigación, dando a conocer los datos, herramientas y algoritmos con los cuales fue posible la extracción y el análisis de la información. Con base en los datos experimentales y realizando un análisis de correlación se lograr primeramente detectar las caracteristicas y la relación entre ellas que provoca la aparición de enfermedades no transmisibles, una vez identificadas se muestra el desarrollo de los biomarcadores pertinentes para la detección de diabetes a través de RNA profundas describiendo la arquitectura que logró el mayor porcentaje de precisión en la clasificación de pacientes con y sin las enfermedades no transmisibles.

## 2.1 Enfermedades no transmisbles.

Existen cuatro enfermedades no transmisibles principales las cuales son: enfermedades cardiovasculares, cáncer, diabetes y enfermedades respiratorias crónicas. Aproximadamente 40 millones de personas mueren al año. Incluyendo 15 millones de personas que murieron en un rango de edad que va de los 30 a los 69 años. Arriba del 80% de las estas muertes prematuras se centran en países con ingresos medios o bajos.[38] Las factores de riesgo más comunes que

provocan el desarrollo de enfermedades no transmisibles se encuentran englobados en cuatro comportamientos particulares los cuales son:

- Uso del tabaco.
- Falta de actividad física.
- Alimentación no adecuada.
- Uso perjudicial del alcohol.

Los comportamientos anteriores son la llave de los cambios psicológicos y metabólicos, tales son: incremento de la presión arterial, sobrepeso/ obesidad, incremento de glucosa en la sangre y el incremento del colesterol [4].

En este estudio se realiza un enfoque a la investigación de dos enfermedades multifactoriales no transmisibles, las cuales son diabetes y caries. La diabetes tipo 2 es un desorden metabólico multifactorial y poligénico, esta patogenesia está influenciada por diversos factores de riesgo ambientales y genéticos[39]. La diabetes tipo 2 está caracterizada por hiperglucemia, con variaciones en los grados de insulina, con un daño con respecto a la secreción de insulina y un incremento en la producción de la glucosa hepática[40].

La prevalencia en el mundo de diabetes tipo dos está incrementando de manera rápida y se predice que incremente a 225 millones para el final de la decada y 300 millones para el año 2025 [41]. En México, la prevalencia de diabetes ha incrementado dramáticamente, y se ha estimado que el 10% de los adultos tienen este padecimiento, además de que la diabetes y sus complicaciones son la primer causa de muertes en mujeres mexicanas y la segunda causa en hombres [42].

La caries es una condición frecuente en salud oral, la cual está definida como una enfermedad multifactorial y es considerada el principal problema de salud pública del mundo, siendo la enfermedad no transmisible más común.
De acuerdo con la Organización Mundial de la Salud, la caries afecta entre un 60% y un 90%

de la población. La mayor desventaja de este padecimiento es la marginación social, lo cual representa un problema tomando en cuenta que los tratamientos para esta condición pueden ser muy costosos económicamente, las enfermedades bucales ocupan el cuarto lugar de los padecimientos más caros de tratar de acuerdo con la Organización Mundial de la Salud, lo cual limita el acceso a esta atención en algunos paises. Esta condición puede provocar dolor e infecciones en una región especifica que va progresando a través de la pulpa dental, donde si no se realiza un tratamiento oportuno, puede provocar absesos dentales [43]. Esto representa muchos factores de riesgo que contribuyen al progreso de la condición [44]. De acuerdo con la OPS, el desarrollo de caries depende de la frecuencia del consumo de carbohidratos, placas cariógenicas, saliva, las caracteristicas de la comida, el tiempo de exposición, eliminación de placas y susceptibilidad del paciente, así como las escasas medidas de prevención correspondientes a salud oral y el acceso limitado a los servicios dentales especializados.

## 2.2 Diabetes Mellitus.

La Diabetes Mellitus (DM) se define como una enfermedad sistémica, crónicodegenerativa, de carácter heterogéneo, con grados de predisposición hereitaria y con participación de diversos factores ambientales [45]. Perner Serena afirma que la DMT2 no se distribuye igual en los diferentes grupos de la sociedad, esto debido a que se presenta mayor incidencia en las personas con un nivel socioeconómico bajo, a su vez, tienen mayor número de complicaciones y una tasa más alta de mortalidad, en este estudio, se le atribuyen las consecuencias mencionadas a la diferencia que existe para el acceso de alimentos, la disponibilidad de los mismos, la falta de espacios para realizar actividad física. [46].

La DM pertenece un grupo de enfermedades metabólicas y es consecuencia de la deficiencia en el efecto de la insulina causada por una alteración en la función endocrina del páncreas o por la alteración de los tejidos efectores, que pierden su sensibilidad a la insulina. [47].

En la figura 2.1 se muestra la fisiopatología celular correspondiente a la diabetes, en la primer fila, se presenta el proceso que se lleva a cabo cuando el sujeto no presenta ningún tipo de diabetes, en la segunda columna se observa la manera en que funciona el páncreas si el sujeto

presenta diabetes tipo 1, la cual ocurre cuando existe una falla en la producción pancreática de insulina, y finalmente en la tercer columna se muestran la actividad célular cuando el sujeto tiene diabetes tipo 2, la cual se presenta cuano las células no responden apropiadamente a la insulina. La DMT2 se asocia con la falta de adaptación al incremento en la demanda de insulina, además de pérdida de la masa celular por la glucotoxicidad. En este tipo de diabetes, el receptor de la insulina presenta alteraciones en su función.

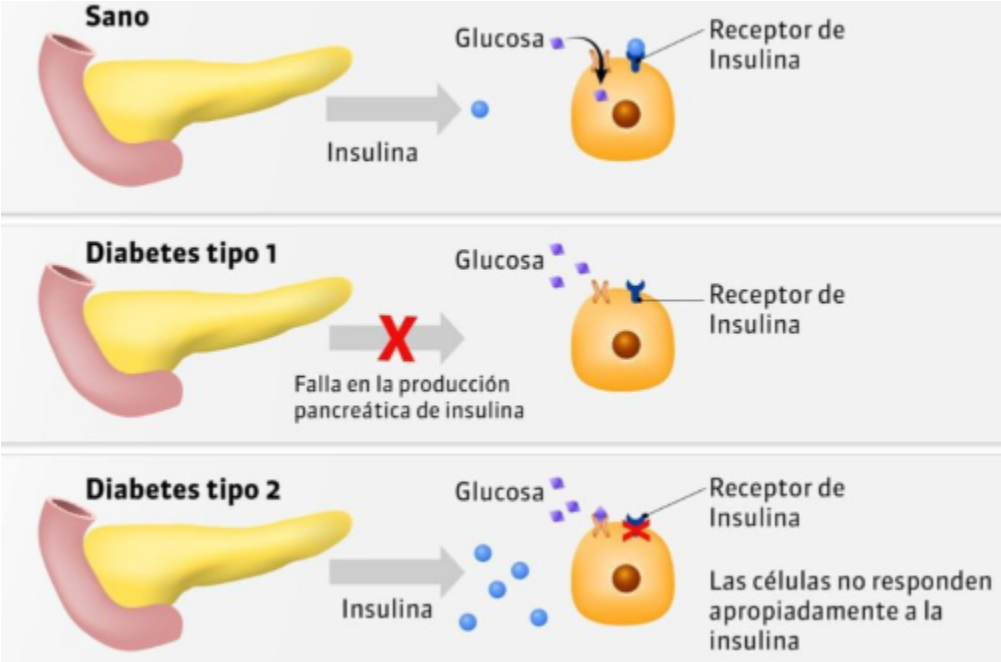


Figura 2.1 Fisiopatología celular de la diabetes tipo 1 y 2.

### 2.2.1 Factores de Riesgo de la Diabetes.

De acuerdo con la Norma Oficial Mexicana 015 (NOM-015) y la OMS, los factores de riesgo para la DM, son aquellos que incrementan la posibilidad para e desarrollo del padecimiento, principalmente [45]:

- Antecedentes hereitarios.
- Edad mayor a 45 años.
- Falta de actividad física (caminar menos de media hora todos los días).

- Sobrepeso y obesidad, definidos como un Índice de Masa Corporal (IMC) mayor a 25 y 30 respectivamente y una circunferencia de cintura mayor de 80 cm para las mujeres y 90 cm en los hombres.
- Estrés prolongado, dado el vínculo que tienen ciertas sustancias o moléculas elementales para el metabolismo de la glucosa.
- Consumo de alcohol, más de dos copas al día en hombres y más de una en mujeres.
- Consumo de tabaco.
- Consumo de medicamentos como las tiazidas, glucocorticoides, difenilhidantoina y bloqueadores beta-adrenérgico .

Es de suma importancia hacer un cambio en los hábitos diarios de los sujetos con diabetes, ya que de no existir alguna modificación se puede desarrollar complicaciones tales como complicaciones en la piel, complicaciones en los ojos, neuropatía, complicaciones en los pies, cetoacidosis, enfermedad renal, derrame cerebral, síndrome hiperglucémico hiperosmolar no ctósico, gastroparesia, enfermedades del corazón, salud mental, problemas en el embarazo, entre otras, todas estas complicaciones disminuyen la esperanza de vida del sujeto además de que incrementa los gastos médicos, por lo que se la situación del sujeto se vuelve aún más delicada.

## 2.3 Caries Dental.

Se considera caries dental como un proceso patológico complejo que afecta a las estructuras dentarias y se caracteriza por un desequilibrio bioquímico; de no ser revertido a favor de los factores de resistencia, conduce a cavitación y alteraciones del complejo dentino-pulpar[48]. Además, la caries es un proceso complejo multifactorial que afecta a los tejidos dentales. Este padecimiento causa la destrucción del estmalte dental cuando el proceso dinámico de desmineralización y remineralización constante se altera por el exceso de producción de los ácidos. La caries se puede observar en primer instancia como una opacidad o decoloración del esmalte dental que, si no recibe las medidas de control, avanza hasta llegar a generar cavidades y efectos en otros tejidos dentales. Cuando la caries no es tratada adecuadamente, puede producir

problemas al comer y/o dormir, a su vez, la caries es la causa principal de ausentismo escolar y laboral [49].

En la figura 2.2 se muestra la composición dental, la cual se divide en tres partes principales:

- Corona, la cual se compone del esmalte y la dentina.
- Encía, incluye la gengiva y la cavidad pulpar.
- Raíz, involucra el conducto, cemento, periodonto, hueso maxilar, los nervios y vasos sanguíneos.

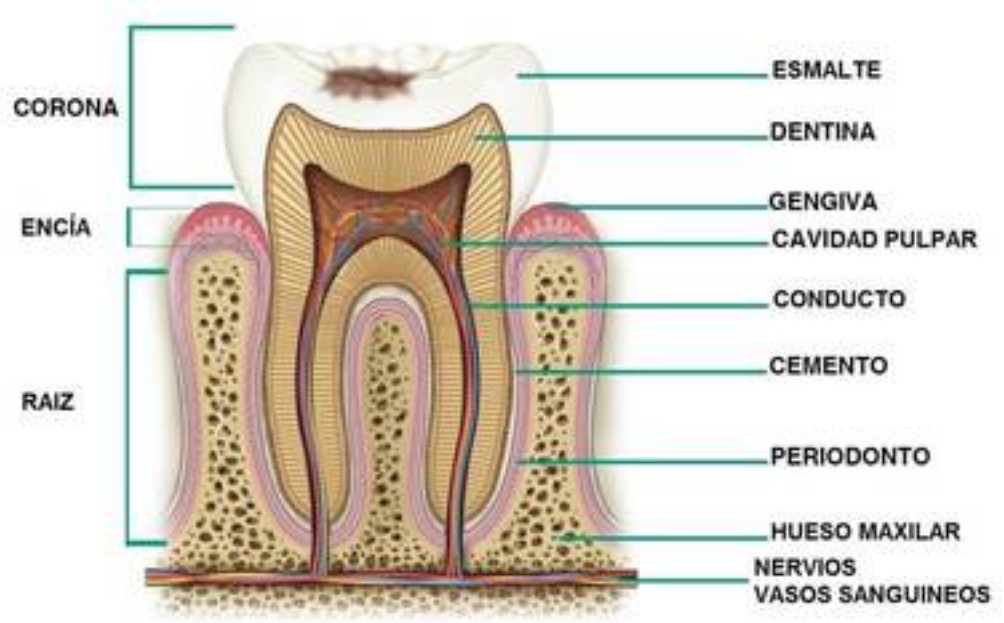


Figura 2.2 Composición dental.

En la figura 2.3 se muestra con mayor detalle la manera en que afecta al diente la caries no tratada adecuadamente, en primer lugar se encuentra la caries superficial, la cual aparece en el esmalte del diente, de ahí pasa a la caries avanzada, ocurre cuando la caries profundiza hasta llegar a la dentina, posteriormente pasa a ser pulpitis, locual significa que la caries avanzó hasta llegar a la pulpa dental y finalmente es posible desarrollar periodontitis, la cual es una infección grave de las encías que daña el tejido blando y destruye el hueso que sostiene los dientes, las consecuencias pueden llegar a tener pédida del diente afectado.

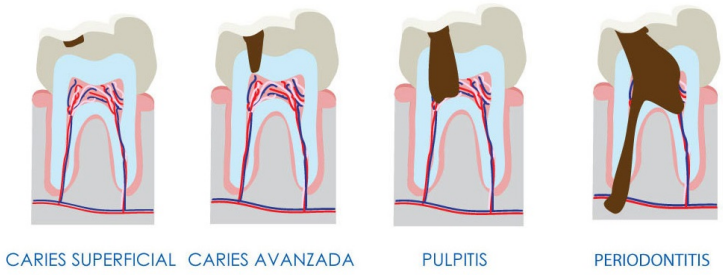


Figura 2.3 Caries dental y caries dental avanzada.

### 2.3.1 Factores de Riesgo de la Caries.

Existen diferentes factores que se ven involucrados en el desarrollo de caries. Según Newbrun estos se pueden clasificar en cuatro conjuntos principales, los cuales son el factor huésped, bacterias, tiempo y sustrato.

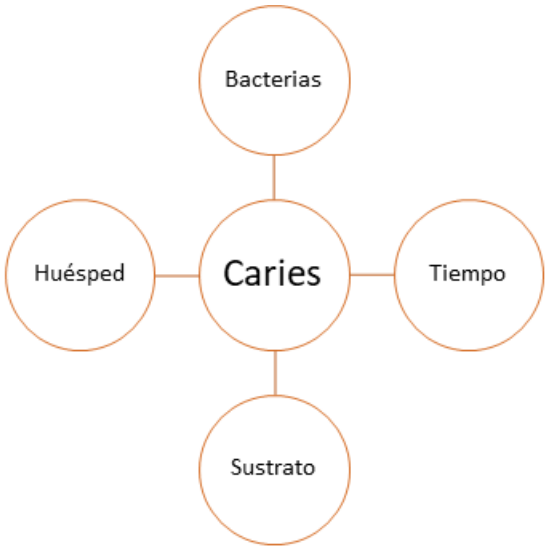


Figura 2.4 Factores de riesgo involucrados en el desarrollo de caries.

### 2.3.2 Huésped.

El factor huésped se contiene a dos grupos principales, los cuales son dientes y saliva.

El diente tiene áreas que son susceptibles al desarrollo de caries como las fosas y fisuras, a su vez, la posición de los dientes tiene relaicón directa con el acúmulo de placa, una vez que se presenta este padecimeinto, comienza a existir dolor en la zona calcificada, esto se debe a que el esmalte se vuelve fácil de destruir. A pesar de que la saliva suele ser un factor protector, también es un factor de riesgo, debido a que la presencia de carbohidratos retenidos y microorganismos bucales se encuentran en medios constantes y expuestos, como lo es la saliva.

### 2.3.3 Bacterias.

Este factor se presenta debido a que la cavidad bucal contiene placa bacteriana, *Streptococcus mutans y Lactobacillus spp*.

La placa bacteriana hace referencia a un ecosistema micribiano compuesto de diferentes estructuras microbianas que suelen agruparse sobre la superficie de la estructura dentaria. Debido a que las bacterias requieren del huésped para su sobrevivencia, se considera parasitaria. Para el caso de *Streptococcus mutans*, se habla de una bacteria Gram positiva, la cual tiene la capacidad de adherirse a la superficie del diente. Cabe destacar que entre esta bacteria y la caries no existe una relación absoluta, ya que pueden existir grandes cantidaes de esta bacteria y aun así no tener un progreso de caries. Los *Lactobacillus spp*, son recurrentes debido a la ingesta frecuente de carbohidratos, una vez que se presentan estas bacterias se acelera el progreso de caries en el sujeto.

### 2.3.4 Tiempo.

Newbrun agrega el factor tiempo con la finalidad de hacer más precisos los modelos que le anteceden, una vez que se añade este nuevo factor, se determina que existen los llamados factores etiológicos moduladores que al igual que los factores etióligicos primarios son causantes de caries; entre ellos se encuentra el tiempo, la edad, salud general, flúor, nivel de instrucción, nivel socio-económico, experiencias anteriores de caries, grupo epidemiológico al igual que variables de comportamiento.

### 2.3.5 Sustrato.

Con el factor sustrato, se hace referencia principalmente a los factores de dieta tomando en cuenta el tipo y proporción del alimento que el sujeto consume diariamente. Por ejemplo, en el estudio de Bart et al. se encontró que existe una relación intima entre la caries y el consumo de azúcares, aunque también concluyen que es importante realizar más estudios con la intención de determinar los factores que aporten información sobre porque no todas las personas con un consumo elevando de azúcar desarrollan caries.

## 2.4 Aprendizaje automático.

El aprendizaje automático, también conocido como *machine learning*, es una disciplina en ciencias de la computación, donde las computadoras son programadas para aprender patrones en conjuntos de datos. El objetivo principal de esta disciplina es obtener un modelo predictivo basado en reglas matemáticas y estadísticas tomando en cuenta las características que se encuentan en una base de datos.

En la Figura 2.5 se muestra el proceso que se sigue al implementar el aprendizaje automático. Como entrada a los algoritmos de machine learnig, se tienen características y etiquetas sobre un conjunto de muestras. Las características son todas aquellas medidas o datos que se encuentran en todas las muestras, estas se pueden encontrar en crudo o con alguna transformación matemática, por otro lado, las etiquetas hacen referencia a lo que el modelo pretende predecir (la salida del modelo). El siguiente paso es hacer que el modelo aprenda ya sea a través del aprendizaje no supervisado o aprendizaje superisado, dependiendo del problema que se presente, una vez entrenado el algoritmo se hacen predicciones y validaciones con datos nuevos.

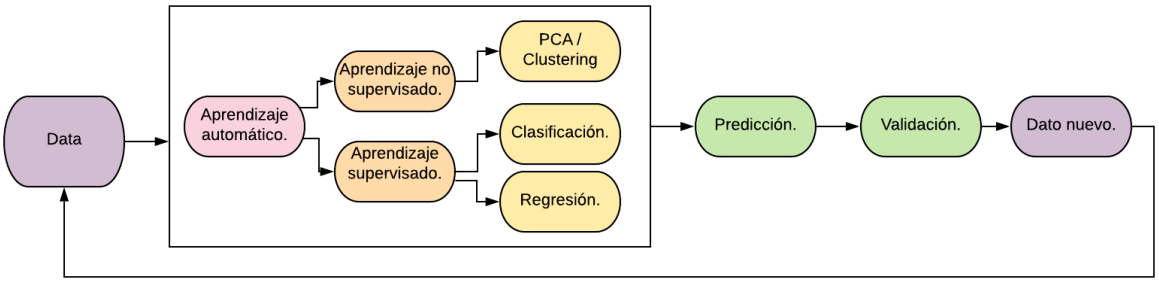


Figura 2.5 Proceso del aprendizaje automático.

**Aprendizaje no supervisado:** Existen ocasiones en las se desconocen las llamadas etiquetas para los datos de entrada o en algunos casos las etiquetas son erróneas, para este tipo de problemas se hace uso de técnicas de aprendizaje no supervisado. Estos métodos agrupan subconjuntos de las muestras, las cuales tienen caracteríticas similares, una vez que se tiene un nuevo dato a clasificar, se tienen dos opciones, la primera consiste en asignar el dato nuevo a un cluster existente, la segunda opción es ejecutar nuevamente el agrupamiento incluyendo el o los datos nuevos. También tienen la ventaja de que permiten visualizar dimensiones grandes de entrada de datos.

**Aprendizaje supervisado:** Para poder hacer uso de métodos supervisados, es fundamental contar con las etiquetas correspondientes a los datos de entrada, debido a que dichas etiquetas se utilizan para entrenar el modelo y así reconocer patrones predictivos y lograr clasificar datos nuevos.

## 2.5 Redes Neuronales Artificiales Profundas.

Desde hace algunos años, muchos procesos computacionales han sido investigados en conjunto con las nuevas técnicas de inteligencia artificial, minería de datos y el uso de diferentes algoritmos, dentro de este conjunto de herramientas se encuentran las redes neuronales artificiales. Una red neuronal es un sistema adaptativo que puede cambiar los parámetros de su estructura para clasificar un problema basado en la información interna o externa que fluye a

través de la red. Una red neuronal también puede clasificarse como un instrumento de modelación no lineal y puede usarse para modelar sistemas con entradas y salidas complejas, por lo que existen numerosas conexiones entre los nodos de los datos[50, 51].

Las RNAs tienen su base en las Redes Neuronales Biológicas (RNB), en la figura 2.6 [52] se muestran los componentes de las neuronas. Se puede considerar una neurona como una unidad estructural y funcional del tejido nervioso, su principal objetivo es desarrollar operaciones de síntesis y procesamiento de información. Se estima que el cerebro humano cuenta con más de cien millones de neuronas y sinapsis (conexiones entre neuronas) en el sistema nervioso, lo que hace posible el procesamiento de información. Las dendritas reciben señales de entrada procedentes de otras neuronas por medio de las sinapsis. La neurona se compone de tres partes principales, comenzando por el cuerpo de la neurona; después las dentrinas, que reciben las entradas; finalmente el axón, que es el encargado de llevar la salida de la neurona a las dentrinas de otras neuronas por medio de diferencias de potencial eléctrico [52, 53].

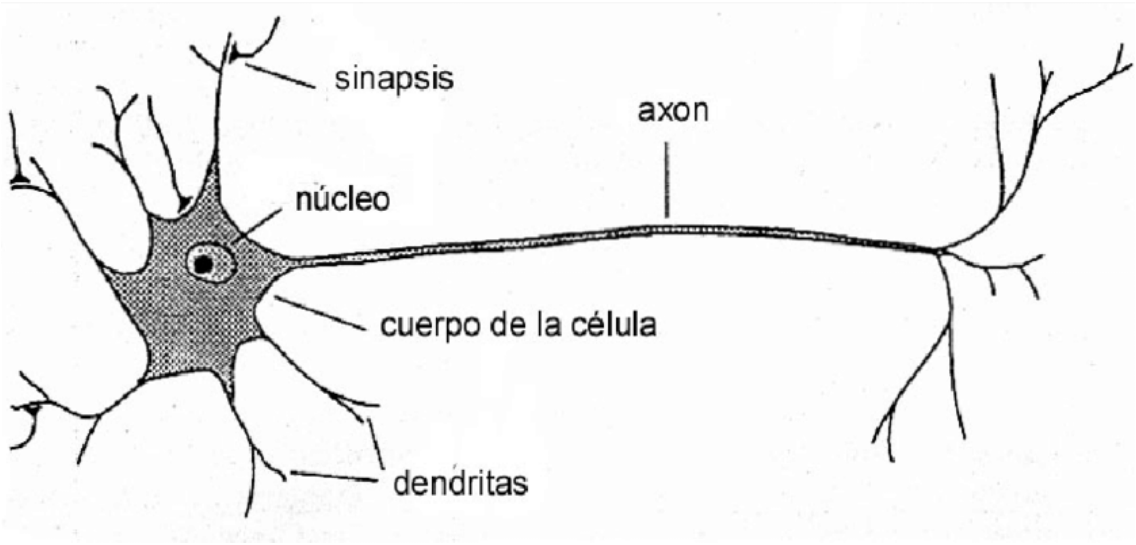


Figura 2.6 Componentes principales de la neurona.

En la figura 2.7 [52] se muestra el modelo matemático de una neurona con sus elementos principales, los cuales son un conjunto de sinapsis o conexiones caracterizadas por su peso; un

sumador $\Sigma$, el cual produce la suma ponderada de las entradas según los pesos de las conexiones; y finalmente una función de activación, tambien conocida como función de transferencia, la cual tiene como objetivo limitar la amplitud de la salida generada por la neurona. En algunas ocasiones se incluye el llamado umbral, cuya tarea es controlar el nivel a partir del cual la neurona produce su salida [52].

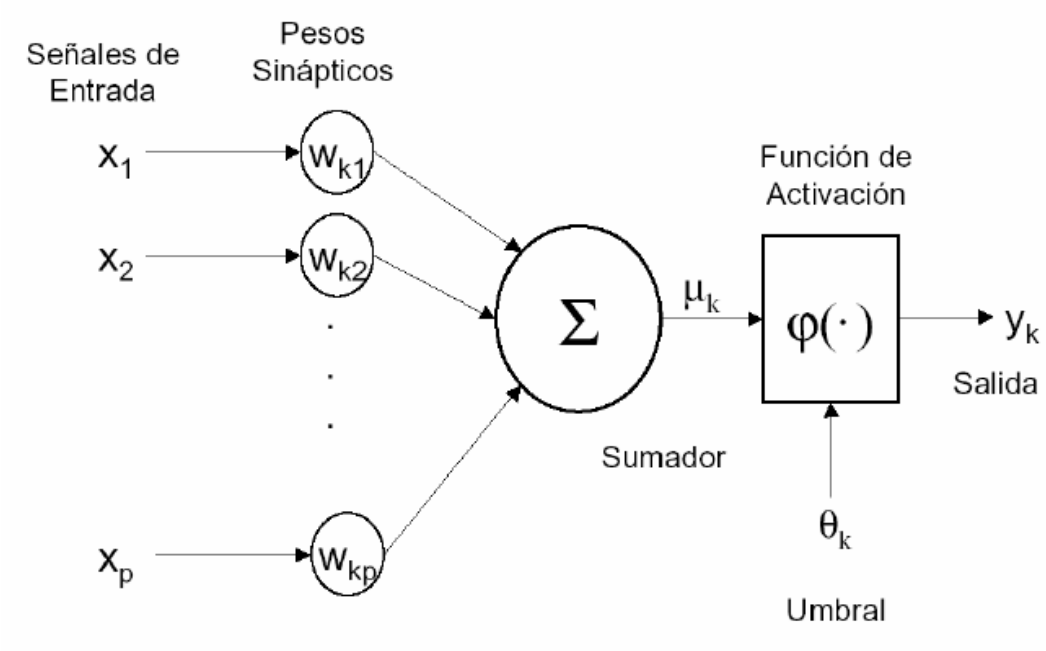


Figura 2.7 Modelo matemático de una neurona.

Las RNA pueden implementarse dentro del llamado aprendizaje profundo, también conocido como *deep learning*, el cual es un subconjunto de aprendizaje automático. Se habla de aprendizaje profundo cuando existen modelos contruidos por múltiples capas de procesamiento que permiten el aprendizaje de representaciones jerarquicas de diferentes niveles de abtracción [54]. En el aprendizaje profundo se hace principalmente uuso del aprendizaje supervisado, ya que así es posible medir el error o la distancia entre las puntuaciones de salida y los patrones deseados de dichas puntuaciones, una vez con este cálculo, se modifican los pesos para hacer que el error se reduzca. Estos pesos se pueden ajustar de manera adecuada calculando un vector de gradiente, el cual, por cada peso incrementará o en su defecto reducirá el error.

En términos generales, una RNA profunda intenta modelar la relación entre las entidades de entrada y la característica de salida, tomando en cuenta tres elementos principales:

- **Pesos,** se refiere a un conjunto de conexiones que son los elementos que conectan la señal de entrada con una neurona a través del cálculo de su producto,
- **Función de activación,** que afecta a las neuronas que limitan la amplitud de la salida con un valor finito,
- Un elemento que resume las contribuciones de una señal ponderada.

Además las RNAs profundas acumulan sus conocimientos detectando los patrones y las relaciones en los datos y aprenden a través de la experiencia, no de la programación. Una RNA profunda se forma a partir de cientos de unidades individuales, llamadas neuronas, conectadas con coeficientes, llamadas ponderaciones, que constituyen la estructura neuronal y están organizadas en capas. El número de capas es modificable como el número de neuronas en cada capa. El poder de los cálculos neuronales proviene de la conexión de las neuronas en una red. Cada elemento de procesamiento tiene entradas ponderadas, una función de transferencia y una salida. El comportamiento de una RNA profunda está determinado por las funciones de transferencia de sus neuronas, la regla de aprendizaje y la arquitectura en sí. Los pesos son los parámetros ajustables y, en ese sentido, una RNA profunda es un sistema parametrizado [55].

### 2.5.1 Perceptron multicapa.

El perceptrón multicapa se muestra en la figura 2.8, consiste en una RNA profunda formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón. El perceptrón multicapa puede ser totalmente o localmente conectado. En el primer caso cada salida de una neurona de la capa “i” es entrada de todas las neuronas de la capa “i+1”, mientras que en el segundo cada neurona de la capa “i” es entrada de una serie de neuronas (región) de la capa “i+1”.

Las capas pueden clasificarse en tres tipos:

- **Capa de entrada:** Se constituye por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce procesamiento.
- **Capas ocultas:** Formada por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de coapas posteriores.
- **Capa de salida:** Neuronas cuyos valores de salida se corresponden con las salidas de toda la Red Neuronal Artificial.

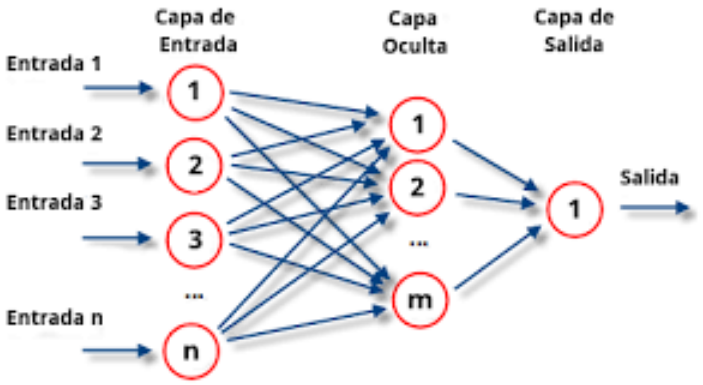


Figura 2.8 Ejemplo de perceptron multicapa.

#### 2.5.1.1 Propagación hacía atrás.

En una red de propagación hacía atrás existe una capa de entrada con n neuronas y una capa de salida con m neuronas y al menos una capa oculta de neuronas internas. Cada neurona de una capa (excepto las de entrada) recibe entradas de todas las neuronas de la capa anterior y envía su salida a todas las neuronas de la capa posterior (excepto las de salida). No hay conexiones hacia atrás feedback ni laterales entre las neuronas de la misma capa.

La aplicación del algoritmo tiene dos fases, una hacia adelante y otra hacia atrás. Durante la primera fase el patrón de entrada es presentado a la red y propagado a través de las capas hasta llegar a la capa de salida. Obtenidos los valores de salida de la red, se inicia la segunda fase, comparándose éstos valores con la salida esperada para así obtener el error. Se ajustan los pesos de la última capa proporcionalmente al error. Se pasa a la capa anterior con una

retropopagación del error, ajustando los pesos y continuando con este proceso hasta llegar a la primera capa. De esta manera se han modificado los pesos de las conexiones de la red para cada patrón de aprendizaje del problema, del que conocíamos su valor de entrada y la salida deseada que debería generar la red ante dicho patrón.

La técnica de propagación hacía atrás requiere el uso de neuronas cuya función de activación sea continua, y por lo tanto, diferenciable. Generalmente, la función utilizada será del tipo sigmoidal.

**Algoritmo de entrenamiento.**

- **Paso 1.** Inicializar los pesos de la red con valores pequeños aleatorios.

- **Paso 2.** Presentar un patrón de entrada y especificar la salida deseada que debe generar la red.

- **Paso 3.** Calcular la salida actual de la red. Para ello se presentan las entradas a la red y se calcula la salida de cada capa hasta llegar a la capa de salida, ésta será la salida de la red.

  Los pasos son los siguientes:

  1. Se calculan las entradas netas para las neuronas ocultas procedentes de las neuronas de entrada.

     – Para una neurona j oculta:

     $$net_{pj}^{h} = \sum_{i=1}^{N} w_{ji}^{h} x_{pi} + \theta_{j}^{h} \tag{2.1}$$

     en donde el índice h se refiere a magnitudes de la capa oculta; el subíndice p, al pésimo vector de entrenamiento, y j a la jésima neurona oculta. El término $\Theta$ puede ser opcional, pues actúa como una entrada más.

  2. Se calculan las salidas de las neuronas ocultas:

     $$y_{pj} = f_{j}^{h}(net_{pj}^{h}). \tag{2.2}$$

3. Se realizan los mismos cálculos para ibtener las salidas de las neuronas de salida:

$$net^o_{pk} = \sum_{j=1}^{L} w^o_{kj} y_{pj} + \theta^o_k \tag{2.3}$$

$$y_{pk} = f^o_k(net^o_{pk}). \tag{2.4}$$

- **Paso 4.** Calcular los términos de error para todas las neuronas.

  Si la neurona *k* es una neurona de la capa de salida, el valor de delta es:

$$\delta^0_{pk} = (d_{pk} - y_{pk} f^o_k(net^o_{pk})) \tag{2.5}$$

  La función $f$ debe de ser derivable. En general se dispone de dos formas de función de salida:

  – La función lineal:

$$f_k(net_{jk}) = net_{jk} \tag{2.6}$$

  – La función sigmoidal:

$$f_k(net_{jk}) = \frac{1}{1 + e^{-net_{jk}}} \tag{2.7}$$

- **Paso 5.** Actualización de los pesos:

  Se hace uso de un algoritmo recursivo, comenzando por las neuronas de salida y trabajando hacía atrás hasta llegar a la capa de entrada, ajustando los pesos de la siguiente forma:

  – Para los pesos de las neuronas de la capa de salida:

$$w^0_{kj}(t+1) = w^o_{kj}(t) + \Delta w^o_{kj}(t+1) \tag{2.8}$$

$$\Delta w^o_{kj}(t+1) = \alpha \delta^o_{pj} y_{pj} \tag{2.9}$$

– Para los pesos de las neuronas de la capa oculta:

$$w_{ji}^{h}(t+1) = w_{ji}^{h}(t) + \Delta w_{ji}^{h}(t+1) \tag{2.10}$$

$$\Delta w_{ji}^{h}(t+1) = \alpha \delta_{pj}^{h} y_{pj} x_{pi} \tag{2.11}$$

- **Paso 6.** El proceso se repite hasta que el término de error de la ecuación 2.12 resulta aceptablemente pequeño para cada uno de los patrones aprendidos.[56][57]

$$E_p = \frac{1}{2} \sum_{k=1}^{M} \delta_{pk}^{2} \tag{2.12}$$

Las aplicaciones del perceptron multicapa son variadas, frecuentemente se utilizan para pronósticos y en su mayoría para la clasificación de diferentes eventos. En este proyecto se hizo uso del perceptron multicapa con un entrenamiendo basado en la propagación hacia atrás para la clasificación de pacientes con y sin ENT como diabetes y caries.

### 2.5.2 Funciones de activación.

La función de activación es requerida para que una vez que la entrada neta ha sido calculada, se transforme en el valor de activación, o activación simplemente y una vez hecho esto se puede aplicar una función de salida que es la encargada de transformar el valor de la entrada neta en el valor de salida del nodo [58]. De manera simple podemos decir que las funciones de activación son una función que limitan la salida de la señal a un valor finito.
Las funciones de activación comumente deben de ser no lineales, a continuación se describe la función de activación ReLu y Softmax, las cuales fueron implementadas en la presente investigación.

#### 2.5.2.1 Unidad Lineal Rectificada (ReLu).

Actualmente es común aplicar la función de activación ReLu en la implementación de Redes Neuronales Artificiales de aprendizaje profundo, especificamente para las capas ocutas. Como se muestra en la ecuación 2.13 la función ReLu indica que la entrada va a ser igual a la

misma entrada si dicha entrada es mayor o igual que 0, en caso contrario, la salida corresponderá a 0. La función ReLu es una función de activación no diferenciable en z=0.

$$ReLu(z) = \begin{cases} \text{si } z < 0 & 0 \\ \text{si } z \geq 0 & z \end{cases} \tag{2.13}$$

Las funciones de activación ReLu son las funciones de activación más simples, además esta función es más rápida al entrenar una RNA de gran tamaño. Por otro lado, tiene menos problema con el desvanecimiento del gradientes en modelos profundos, pero una desventaja que conlleva es que la RNA puede detener el proceso que esté realizando si es que se hace uno de una tasa de aprendizaje muy alta.

En la figura 2.9 se muestra la gráfica relacionada con el comportamiento de la función de activación ReLu, la cual indica que todos aquellos elementos positivos se mantienen sin cambios, mientras que por otro lado los valores correspondientes a números negativos se convierten en ceros.

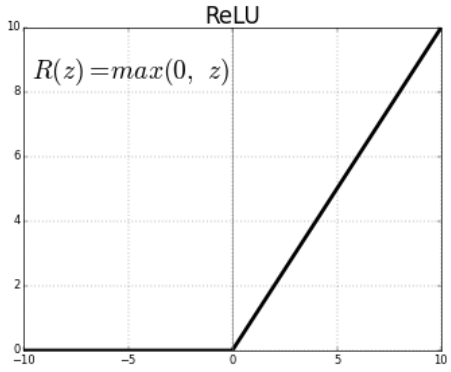


Figura 2.9 Gráfica de la función de activación ReLu.

#### 2.5.2.2 Softmax.

La función Softmax es utilizada para forzar la salida de tal modo que la suma total de los valores sea igual a uno. Esto se refiere a que la salida de la función softmax es equivalente a una distribución probabilistica categorica, la cual indica la probabilidad de que alguna de las clases que se tengan sea verdadera. La principal ventaja de usar softmax es el rango de la probabilidad de la salida. El rango se encontrará entre 0 y 1, y la suma de todas las probabilidades será igual

a uno. Si la función softmax es utilizada para un modelo de multiclasificación, esta regresará la probabilidad de que la salida corresponda a cada una de las clases y la clase ideal tendrá la probabilidad más alta.

Matemáticamente la función softmax se muestra en la ecuación 2.14, donde z es un vector que contiene las entradas que corresponden a la capa de salida, j se encarga de indexar las unidades de salida por lo que j=1,2, . . . , K. Esta ecuación, calcula la exponencial de los valores de entrada proporcionados y la suma del exponencial de todos los valores correspondientes al vector de entrada. Entonces la proporción del exponencial de los valores de entrada y la suma exponencial de los valores da como resultado la salida de la función softmax.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{K=1}^{K} e^{z_k}} \tag{2.14}$$

En la figura 2.10 se muestra la gráfica relacionada con el comportamiento de la función de activación Softmax, la cual se puede visualizar con valores del 0-9, por lo que tenemos 10 clases y se muestra la probabilidad de que un dígito pre-establecido corresponda a una de las clases especificadas.

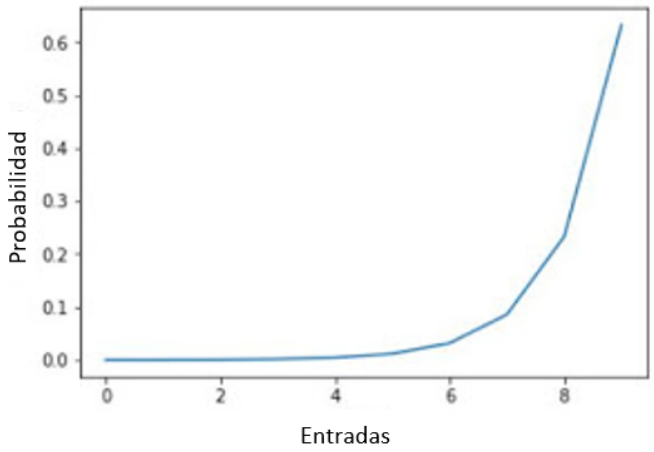


Figura 2.10 Implemantación de la función Softmax.

### 2.5.3 Tipos de capas.

Para el diseño de la arquitectura de una RNA, es importante conocer los tipos de capas con las que se puede trabajar, las cuales tienen objetivos específicos por cumplir. A continuación se describen de manera breve algunas de ellas.

#### 2.5.3.1 Densa.

La capa densa consiste en una matriz de pesos creada por la capa, a su vez está compuesta por un vector creado por la misma capa llamado bias. Para poder implementar este tipo de capa, es importante indicar las unidades o neuronas con las que se pretende trabajar y del mismo modo se tiene que indicar la función de activación con la que se va a estar trabajando en los cálculos correspondientes a dicha capa.
Existen difententes argumentos que pueden incluirse en esta capa, pero los anteriormente mencionados son aquellos indispensables para un correcto funcionamiento.

#### 2.5.3.2 Dropout.

Esta capa consiste en establecer de manera aleatoria una tasa fraccionaria de unidades de entrada a 0 para cada actualización durante el tiempo de entrenamiento, este tipo de capa ayuda principalmente a evitar un sobreajuste. Aquellas unidades que se mantienen son escaladas según la operación 2.15, donde rate se refiere a una tasa establecia entre 0 y 1. Haciendo uso de esta ecuación, aseguramos que la suma no cambie en tiempo de entrenamiento ni en tiempo de inferencia. Además de indicar la tasa, cabe resaltar que existen diferentes argumentos que pueden incluirse en este tipo de capa, aunque para el presente trabajo solo se indicó la tasa, el cual es el argumento base para el correcto funcionamiento de la RNA.

$$\frac{1}{1 - rate} \tag{2.15}$$

### 2.5.4 Parámetros para la validación de la Red Neuronal Artificial.

Los parámetros que se utilizan para la validación de la red neuronal artificial es la función de pérdida, la precisión y la curva ROC (acrónimo de Receiver Operating Characteristic), la cual se basa en el promedio general de la Red Neuronal Artificial Profunda.

Es posible determinar cuando el modelo se ajusta mejor a los datos porque el valor de la función de pérdida baja debido a que se está acercando al mínimo global, lo cual representa el error mínimo. A su vez, la función de pérdida es capaz de optimizar la retroalimentación hacía atrás de la red con información de la capacidad del sistema[59].

El parámetro correspondiente a la Entropía Binaria Cruzada fue elegido como un método para calcular la función de pérdida, la cual está incluída en el paquete de Keras. Este método hace uso del principio de la distancia Kullback-Leiber, la cual es calculada con la ecuación 2.16 y consiste en la medida entre dos funciones de densidad *g* and *h*. La entropía cruzada es un método iterativo que genera un conjunto de valores aleatorios que son actualizados con el objetivo de generar valores aproximados [60].

$$D(g,h) = \int g(x) ln \frac{g(x)}{h(x)} \mu(dx) = \int g(x) lng(x) \mu(dx) - \int g(x) lnh(x) \mu(dx) \qquad (2.16)$$

Por otro lado, la función de precisión seleccionada obtiene el promedio de la precisión basada en el total de predicciones, esto es utilizado para problemas de clasificación, esta función es llamada precisión-binaria, la cual está dentro del paquete Keras. La precisión es calculada con la ecuación 2.17, la cual se basa en la diferencia entre la clasificación calculada y la clasificación real y es representada como 1 - error, $V\_pred$ es el valor calculado de la clasificación y $V\_true$ es el valor real calculado. Este valor es obtenido para cada modelo, dando la opción para seleccionar el modelo que presenta un mejor rendimiento [61].

$$error = V_{pred} - V_{true} \qquad (2.17)$$

También se obtuvo la curva ROC, la cual es utilizada para medir la precisión del modelo clasificatorio y se basa en la especificidad y sensitividad. La especificidad representa el número de sujetos correspondientes a los controles que fueron correctamente clasificados y es calculado como en la ecuación 2.18 donde $NPV$ representa el valor negativo predicho, $TN$ representa los verdaderos negativos y $FN$ los falsos negativos [62].

$$NPV = \frac{TN}{TN + FN} \tag{2.18}$$

La sensitividad representa el número de sujetos que corresponden a casos y que fueron correctamente clasificados, esto se obtiene con la ecuación 2.19 donde $PPV$ representa los valores positivos predichos, $TP$ los verdaderos positivos y $FP$ representa los falsos positivos[62].

$$PPV = \frac{TP}{TP + FP} \tag{2.19}$$

Para cada clase se calcula a curva ROC, así como el macro-promedio y micro-promedio.

El macro-promedio es la precisión promedio en diferentes conjuntos arbitarios, y es calculado con la ecuación 2.20 donde $A_1$ representa el promedio del conjunto 1 y $A_2$ representa el promedio del conjunto 2.

$$Macro - promedio = \frac{A_1 + A_2}{2} \tag{2.20}$$

El micro-promedio es la suma total de los verdaderos positivos y los falsos negativos para diferentes conjuntos aleatorios y se calcula con la ecuación 2.21, donde $TP_1$ representa los verdaderos positivos del conjunto 1, $TP_2$ representa los verdaderos positivos del conjunto dos, $FP_1$ representa los falsos positivos del conjunto y y $FP_2$ representa los falsos positivos del conjunto dos [63].

$$Micro - promedio = \frac{TP_1 + TP_2}{TP_1 + TP_2 + FP_1 + FP_2} \tag{2.21}$$

## 2.6 Herramientas de implementación para la Red Neuronal Artificial.

Para la implementación de las Redes neuronales artificiales se hizo uso de las siguientes herramientas:

- Python es un lenguaje de programación interpretado, orientado a objetos y de alto nivel con semántica dinámica. Sus estructuras de datos integradas de alto nivel, combinadas con la tipificación dinámica y el enlace dinámico, lo hacen muy atractivo para el rápido desarrollo de aplicaciones, así como para un lenguaje de scripting o pegamento para conectar componentes existentes entre sí. Python es una herramienta simple y fácil de aprender donde la sintaxis enfatiza la legibilidad, reduciendo el costo del mantenimiento del programa. Python admite módulos y paquetes, lo que fomenta la modularidad del programa y la reutilización del código. El intérprete de Python y la extensa biblioteca estándar están disponibles en formato fuente o binario sin cargo para todas las plataformas principales, y se pueden distribuir libremente[64]

- Keras es una API de Redes Neuronales Artificiales profundas de alto nivel, diseñada para realizar una experimentación rápida usando RNA profunda , enfocada en ser fácil de usar, modular y extensible. Fue desarrollado como parte del esfuerzo de investigación del proyecto ONEIROS (Sistema operativo de robot inteligente neuroelectrónico de composición abierta) [65].

- Tensorflow es una biblioteca de software de código abierto para la programación del flujo de datos en una variedad de tareas. Es una biblioteca de matemáticas simbólicas, utilizada para aplicaciones de aprendizaje automático como Redes Neuronales Artificiales Profundas [66].

# Capítulo 3

# Implementación de las Redes Neuronales Profundas para la identificación de biomarcadores en enfermedades no transmisibles.

Este capítulo contiene información correspondiente a los pasos que se siguieron para realizar la experimentación, asi como la interpretación de los resultados obtenidos.

Los pasos seguidos para llevar a cabo este trabajo se muestran en la figura 3.1. Donde A) corresponde a la adquisición de la base de datos NHANES 2013-2014, B) se enfoca en el pre-procesamiento de los datos necesario para posteriormente pasar a C) donde se realiza la implementación de la RNA, así como la validación y obtención de resultados de la misma.

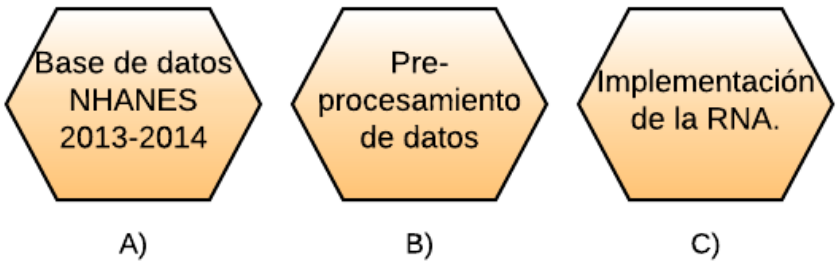


Figura 3.1 Metodología de trabajo del proyecto.

## 3.1 Base de Datos NHANES 2013-2014.

National Health and Nutrition Examination Survey (NHANES) es un programa que diseña una serie de estudios para llevar a cabo una entrevista con la intención de medir el estatus nutricional y de salud en adultos y niños en los Estados Unidos, la cual fue fundada por el Centro Nacional de Estadisticas de Salud(NCHS por sus siglas en inglés).

En el año 1956, la Ley Nacional de Encuestas de Salud, otorgó autorización legistaltiva para que po medio d euna encuesta se proporcionara información estadística sobre la cantidad, distribución, efectos de enfermedades y discapacidad en los Estados Unidos. La información es obtenida de tres fuentes principales

1. Datos de gente recolectados por entrevistas directas.

2. Pruebas clínicas, medidas, y examinaciones físicas de personas seleccionadas para la muestra.

3. Lugares donde las personas reciben atencion médica, ya sean hospitales, clínicas y/o oficinas de doctores.

Las primeras tres encuestas fueron realizadas en los años 60, y consistian en algunas enfermedades crónicas que se presentaban en adultos desde lo 18 hasta los 79 años de edad, en la siguente encuesta se incluyeron a niños de 6 a 11 años de edad, la tercer encuesta se enfocó únicamente en jóvenes de 12 a 17 años de edad. El conjunto de las 3 encuestas tenían un tamaño de muestra aproximado de 7,500 individuos.

La colección de datos que se realiza actualmente, comenzó a principios del año 1999, y anualmente se realiza un examen al sujeto. Todos los datos recolectados son útiles para determinar la prevalencia de algunas enfermedades, así como sus factores de riesgo y lograr la prevención de ellas. Toda la información es utilizada para la realización de estudios epidemiológicos y diferentes investigaciones en el área de ciencias de la salud.

**Selección de los participantes.**

El proceso de selección de los participantes para las encuestas NHANES consiste en cinco pasos:

1. Todos los condados de Estados Unidos se divide en 15 grupos dependiendo de las características, posteriormente se selecciona un condado de cada grupo y de ahí se obtienen los 15 condados en donde se realizarán las encuestas de cada año.

2. Una vez seleccionados los condados, estos se dividen en grupos más pequeños, de los cuales se seleccionan entre 20 y 24 de los nuevos grupos generados.

3. Posteriormente se identifican todas las casas que se encuentran dentro de los grupos seleccionados y se elige una muestra de aproximadamente 30 viviendas por cada grupo.

4. Los entrevistadores de NHANES van a cada una de las viviendas que fueron seleccionadas en el paso anterior, y piden información correspondiente a edad, raza y sexo de todas las personas que residen en dichas viviendas.

5. Finalmente, se hace uso de un algoritmo que aleatoriamente selecciona a algunos, a todos o a ninguno de los miembros del hogar.

La encuesta combina exámenes físicos y entrevistas, lo que permite desarrollar estudios a través de diferentes características de los individuos.

La estructura general de la encuesta de NHANES incluye varios tipos de entrevistas, incluidas preguntas demográficas, dietéticas y relacionadas con la salud. En la tabla 3.1 se presentan las descripciones de los conjuntos de datos utilizados en este trabajo.

En la encuesta participaron un total de 27,631 sujetos. Los sujetos que participaron en las entrevistas se seleccionaron al azar a través de un algoritmo que consiste en un diseño complejo de múltiples etapas con una serie de etapas y los sujetos.

La muestra de personas incluidas en el conjunto de datos de NHANES tiene como principal población objetivo a la población residente no nacionalizada de los EE. UU., Incluidas las personas hispanas, negras no hispanas, asiáticas no hispanas, blancas no hispanas y otras

Tabla 3.1 Descripción de los tipos de cuestionarios realizados para generar el dataset.

| Tipo de Cuestionario | Descripción |
|---|---|
| Demográfico | El archivo demográfico proporciona información individual, familiar y nivel de hogar. |
| Evaluación | Importancia para la salud pública en área de vigilancia, prevención, tratamiento, cuidado dental, políticas de salud, evaluación de programas de salud federales, entre otros. |
| Cuestionario | Información sobre: aculturación, consumo de alcohol, estatus de diabetes, seguro de salud, entre otras |

personas con 130% del nivel de pobreza, además de las personas blancas no hispanas y otras personas de 80 años o más.

El conjunto de datos demográficos contiene 9,801 sujetos de los cuales 4,826 son hombres y 4,975 son mujeres, ambos pertenecen a un rango de edad que va desde 0 hasta 80 años, conteniendo un total de 39 características.

## 3.2 Preprocesamiento de los datos.

El primer paso para el preprocesamiento de los datos consiste en mantener solo a aquellos sujetos que presentaron información correspondiente a caries y diabetes, una vez filtrados estos datos, se mezclaron para crear solo un conjuntos de datos. Como segundo paso, se detectaron aquellos sujetos que presentaban información incompleta, o que no presentaban un estado positivo o negativo para ambas enfermedades (caries y diabetes), dichos sujetos fueron eliminados del conjunto de datos.

En el tercer paso se eliminaron las características que presentaban $\geq 70$ % de datos faltantes o valores singulares. La característica utilizada como resultado presentó dos estados: presencia de caries y diabetes, etiquetada como ” 1” , y la ausencia de caries y diabetes, etiquetada como ”0”.

Posteriormente, el método Z-score se usó para normalizar las 31 características, en el conjunto de datos, este método se seleccionó porque transforma los datos en una distribución con media 0 y desviación estándar 1 y debido a que generalmente los datos no están definidos en la misma escala numérica. Una vez aplicado el método Z-score, los datos son adecuados para la clasificación.

Finalmente, tomando en cuenta la validación de los resultados a obtener, se dividió el conjunto de datos en dos sub conjuntos, uno correspondiente a entrenamiento, el cual contiene el 70% de los datos, y el 30% restante se utiliza para realizar las pruebas de validación.

## 3.3 Implementación de una RNA profunda para la clasificación de los datos.

Con base en la estructura de la base de datos propuesta, se propuso la implementación de perceptron multicapa con el algoritmo de propagación hacía atrás (backpropagation), los cuales se describen a continuación.

Es importante saber que las RNA artificiales profundas se caracterizan principalmente por [67]:

- Tener una inclinación natural a adquirir el conocimiento a través de la experiencia, el cual es almacenado, al igual que en el cerebro, en el peso relativo de las conexiones interneuronales.
- Tienen alta plasticidad y gran adaptabilidad, son capaces de cambiar dinámicante junto con el medio.
- Poseen un alto nivel de tolerancia a fallas, es decir, pueden sufrir un daño considerable y continuar teniendo un buen comportamiento, al igual como ocurre en los sistemas biológicos.
- Tener un comportamiento altamente no-lineal, lo que les permite procesar información procedente de otros fenómenos no-lineales.

La implementación de la RNA profunda específicamente diseñada para este conjunto de datos se llevó a cabo utilizando la paquetería ”Keras” y ”Tensorflow” en Python. La clasificación de los sujetos puede ser; paciente de control, lo que significa que no padece ni caries ni diabetes y se encuentra etiquetado como ”0”; o en caso contrario están los pacientes que tienen presencia de ambas afecciones, los cuales se etiquetan como ”1”.

Una vez identificadas las variables y las muestras a analizar con sus respectivas etiquetas, se procedió a diseñar la RNA profunda, la cual se muestra en la figura 3.2, el objetivo de esta RNA profunda es clasificar como ”0” la ausencia de caries y diabetes, y como ”1” la presencia de ambas ENT. La estructura propuesta de la RNA profunda se compone de la siguiente manera:

- A la capa de entrada se le asignaron 31 neuronas que representan las 31 características en el conjunto de datos.
- La primera capa oculta dropout tiene un porcentaje de pérdida del 25%.
- La primera capa oculta densa fue compuesta por 100 neuronas.
- La segunda capa oculta dropout tiene un un porcentaje de pérdida del 50%.
- La segunda capa oculta densa fue asignada con 500 neuronas.
- La tercera capa oculta dropout tiene un una pérdida de 25%
- La tercera capa oculta densa estaba compuesta por 100 neuronas.
- La cuarta capa oculta dropout tiene un una pérdida del 50%.
- Finalmente, la cuarta capa oculta densa corresponde a la capa de salida y se encuentra caracterizada por dos neuronas que se refieren a las dos posibles clasificaciones.

El algoritmo de optimización implementado fue ”Adam”, el cual calcula la media móvil exponencial del gradiente y el gradiente cuadrado. Este proceso se basa en el algoritmo de descenso del gradiente estocástico, utilizando el promedio del primer y segundo momento de los gradientes, con el propósito de controlar el deterioro de esa media móvil [68].

Se hizo uso de dos funciones de activación, la primera conocida como ReLu, la cual se utilizó en las capas densas, excepto en la capa de salida, para esta última se hizo uso de la función softmax.

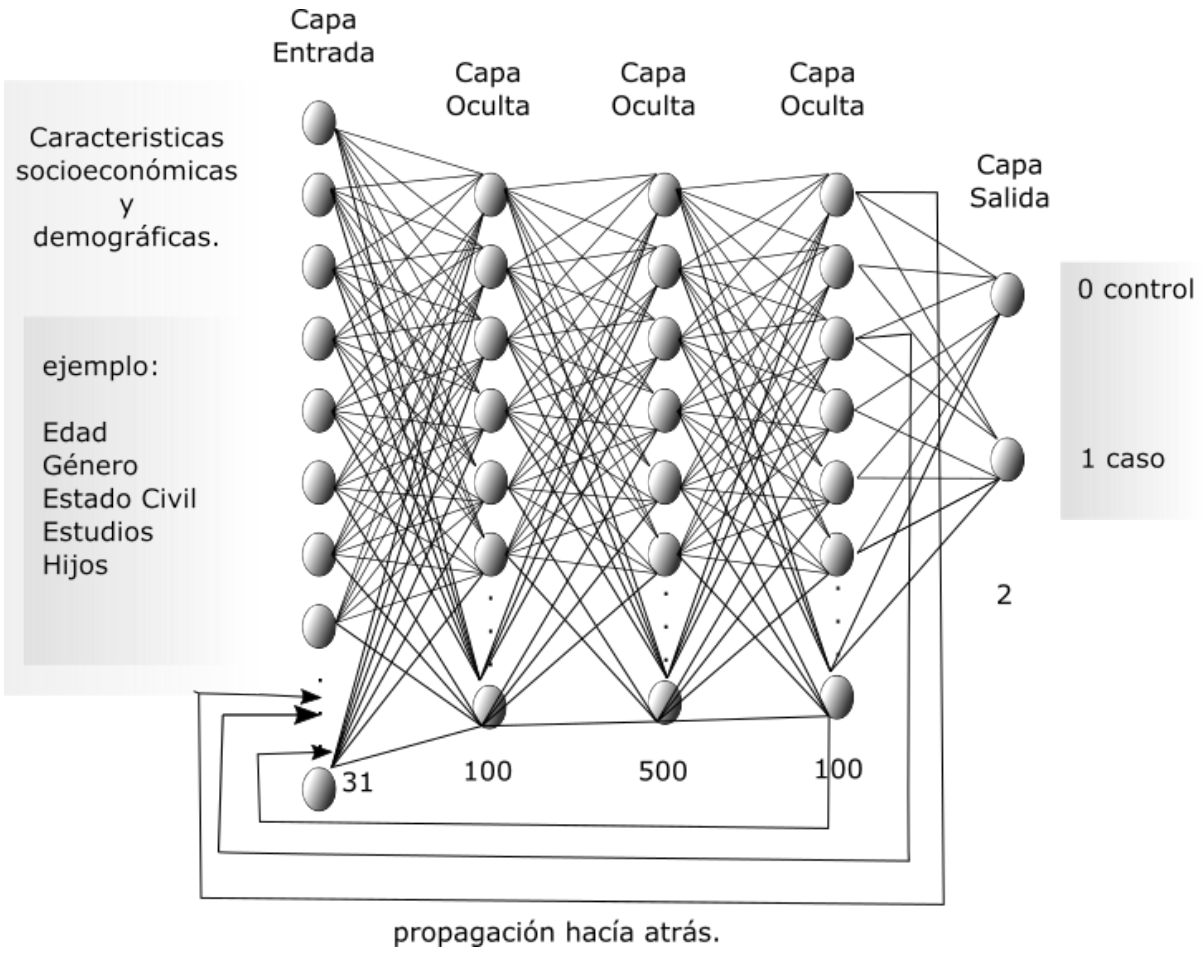


Figura 3.2 Red Neuronal Profunda utilizada como clasificador de enfermedades no transmisibles.

Por otro lado, existe un parámetro llamado *épocas*, el cual es un parámetro de la RNA configurable, la épocas hacen referencia al número de iteraciones que ejecutará la RNA. Antes de realizar la clasificación de los sujetos, se realizó una prueba donde el objetivo es determinar el número de épocas óptimo para la RNA, para ello se hizo una variación en dicho número, en la tabla 3.2 se muestran los resultados de la prueba realizada. Una vez vistos los resultados correspondientes a la precisión, función de pérdida y tiempo de procesamiento en segundos, variando las épocas de la siguiente manera: 10, 50, 100, 150, 200 y 300, se determinó que las

épocas óptimas para este trabajo son 100.

Tabla 3.2 Resultado de los valores de precisión y función de pérdida utilizando diferente número de épocas.

| Épocas | Precisión | Función de pérdida | Tiempo de Procesamiento (segundos) |
|---|---|---|---|
| 10 | 0.9759 | 0.0824 | 3.6252 |
| 50 | 0.9916 | 0.0216 | 15.4416 |
| 100 | 0.9988 | 0.0048 | 29.2827 |
| 150 | 0.9992 | 0.0088 | 45.2142 |
| 200 | 0.9984 | 0.0047 | 58.9482 |
| 300 | 0.9980 | 0.0160 | 89.2401 |

Después de determinar el número de épocas, se hicieron pruebas con diferentes arquitecturas propuestas, como se observa en la tabla 3.3, en las cuales se hicieron variaciones en el número y tipo de capas, cantidad de neuronas por capas, se verificó la precisión, función de pérdida y tiempo de procesamiento en segundos. Con base en los resultados obtenidos del procesamiento, se determinó hacer uso de la arquitectura que consta de cinco capas densas y cuatro de tipo dropout, dichas capas se componen de la siguiente manera:

- La capa de entrada consta de 31 neuronas correspondientes a las características utilizadas para el análisis.
- La primer capa oculta de tipo dropout tiene un porcentaje de pérdida del 25%.
- La primer capa tipo densa se compone de 100 neuronas.
- La segunda capa tipo dropout corresponde a un porcentaje de pérdida del 50%
- La segunda capa oculta de tipo densa con 500 neuronas.
- La tercer capa oculta de tipo dropout tiene una pérdida del 25 %.

- La tercer capa de tipo densa se compone de 100 neuronas.
- La cuarta capa de tipo dropout tiene una pérdida del 50%.
- Finalmente, la cuarta capa de tipo densa corresponde a la capa de salida, contien dos neuronas, las cuales hacen referencia a las dos posibles clasificaciones.

Tabla 3.3 Valores de precisión, función de pérdida y tiempo de procesamiento con diferentes números de capas y neuronas.

| Capas Densas/Dropout | Neuronas | Precisión | Función de pérdida | Tiempo de procesamiento (segundos) |
|---|---|---|---|---|
| 2/0 | 31 >2 | 1 | 0.0003 | 12.4599 |
| 3/1 | 31 >100 >0.5 >2 | 1 | 0.0002 | 15.0060 |
| 3/2 | 31 >0.25 >100 >0.5 >2 | 0.9984 | 0.0048 | 15.9131 |
| 4/1 | 31 >100 >0.5 >500 >2 | 0.9988 | 0.0028 | 20.7666 |
| 4/2 | 31 >0.25 >100 >0.5 >500 >2 | 0.9988 | 0.0052 | 21.7124 |
| 4/3 | 31 >0.25 >100 >0.5 >500 >0.25 >2 | 0.9988 | 0.0044 | 24.0035 |
| 5/1 | 31 >100 >0.5 >500 >100 >2 | 0.9992 | 0.0048 | 25.5583 |
| 5/2 | 31 >0.25 >100 >0.5 >500 >100 >2 | 0.9992 | 0.0028 | 26.7241 |
| 5/3 | 31 >0.25 >100 >0.5 >500 >0.25 >100 >2 | 0.9996 | 0.0018 | 28.8960 |
| 5/4 | 31 >0.25 >100 >0.5 >500 >0.25 >100 >0.5 >2 | 0.9964 | 0.0099 | 29.8431 |

# Capítulo 4

# Resultados y Discusión

En el paso del preprocesamiento correspondiente al capítulo 3, una vez realizados los filtros descritos, los 9,801 sujetos se redujeron a 3,552 de los cuales 1,812 son de género masculino y 1,740 de género femenino. En la Figura 4.1 se muestra el número de sujetos correspondientes a controles y casos. Además las 39 características fueron reducidas a 31 más la característica que proporciona información sobre el diagnóstico.

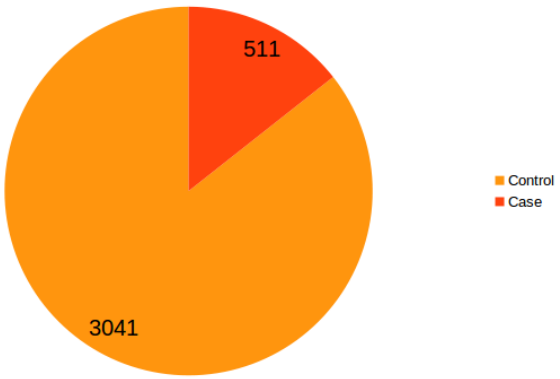


Figura 4.1 Gráfico que indica el número de casos y controles que se encuentran en la base de datos.

Después de la reducción en el número de sujetos, el conjunto de datos quedó dividido con 2,125 controles y 361 casos para el entrenamiento de la RNA, y 916 controles y 150 casos para realizar las pruebas a la RNA entrenada.

Una vez que se tienen todos parámetros a utilizar, se procede a ejecutar la RNA, obteniendo los resultados que a continuación se presentan.

En la figura 4.2 se muestra el comportamiento de la precisión, donde la línea azul se refiere a los datos de entrenamiento, alcanzando un valor de 0.9964, la línea naranja corresponde a los datos de prueba, alcanzando un valor de 0.9906.

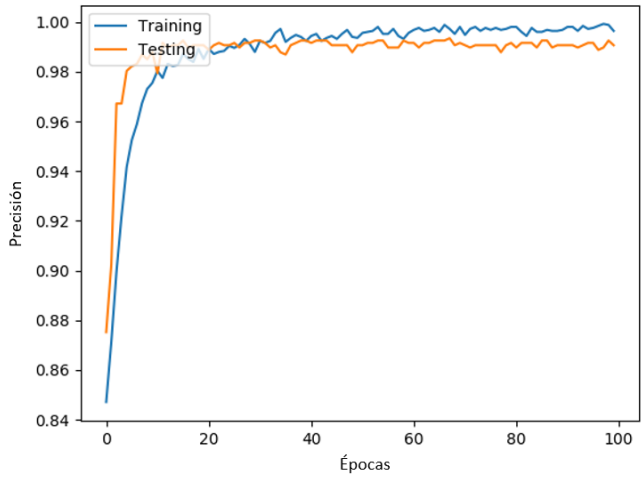


Figura 4.2 Gráfico de el comportamiento de la precisión

La figura 4.3 muestra el comportamiento de la función de pérdida, donde la línea azul hace referencia a los datos de entrenamiento, alcanzando un valor de 0.0099 y la línea naranja hace referencia a los datos de prueba, alcanzando un valor de 0.0945.

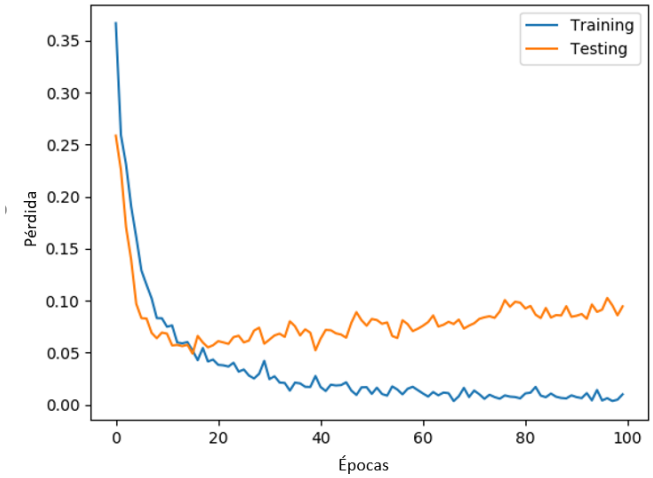


Figura 4.3 Gráfico del comportamiento de la función de pérdida.

Finalmente, en la Figura 4.4 se muestran las curvas ROC que representan el rendimiento de la red neuronal. La línea rosa corresponde a la curva ROc de la clase ”0” o de los sujetos de control, la cual obtuvo un AUC = 0.99. La línea color azul claro correspoende a la curva ROC de la clase ”1” o los sujetos caso, donde se obtuvo un AUC = 0.99. La línea que se observa con puntos color naranja corresponde a la curva ROC del micro-promedio obteniendo un AUC = 1 y por último, la línea de puntos color azul oscuro corresponde al macro-promedio, en donde se obtuvo un AUC= 0.99.

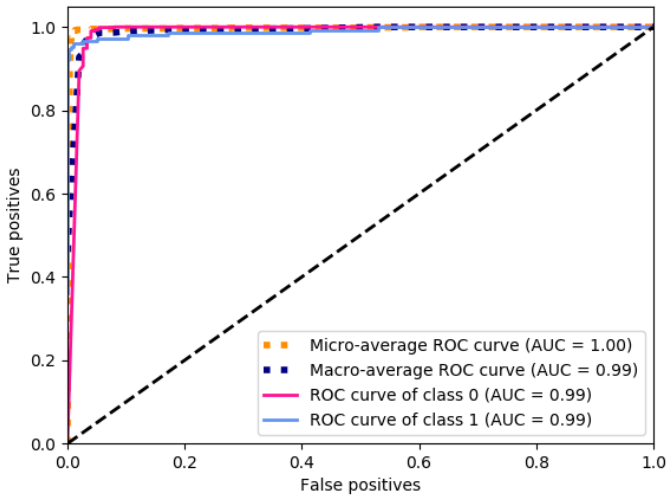


Figura 4.4 Curvas ROC obtenidas con el promedio del rendimiento de la RNA.

Los datos utilizados para el desarrollo de este trabajo, se dividieron en dos conjuntos, de los cuales uno corresponde al 70% de los datos, los cuales se emplearon para el entrenamiento de la RNA, el segundo conjunto contiene el 30% del resto de los datos, los cuales se realizó la prueba de la RNA. El entrenamiento de la RNA corrsponde en un total de 100 épocas o iteraciones, las cuales fueron seleccionadas a partir de una comparación de difentes valores como se muestra en la tabla 3.2. La RNA implementada hace uso del algoritmo Adam, con la finalidad de mejorar el comportamiento RNA con cada retroalimentación. La cantidad de casos y controles aparentemente no se encuentra equilibrada, pero fue suficiente para obtener una ejecución significativa con respecto a la clasificación de los sujetos. Con los datos correspondientes a las pruebas, se hizo la validación de los resultados obtenidos en la etapa de

entrenamiento, tomando en cuenta las métricas de precisión y función de pérdida. En la figura 4.2 y 4.3 es posible observar que el comportamiento del entrenamiento y de las pruebas siguen un patron, lo cual indica que se logró una generalización en el proceso de aprendizaje. La figura 4.2 indica que el 99% de los sujetos se lograron clasificar correctamente entre casos y controles. Por otro lado la figura 4.3 indica que tanto para el conjunto de entrenamiento como para el de pruebas, la función de pérdida decrese, lo cual indica que la RNA se acerca a un mínimo global. Con los resultados obtenidos en la figura 4.4, se observa que en todas las curvas ROC el valor correspondiente al AUC es estadísticamente significativo $\geq 0.99$, este dato indica que la RNA es capaz de clasificar el 99% de los sujetos de manera correcta.

# Capítulo 5

# Conclusiones y trabajo futuro.

De acuerdo con los resultados obtenidos, es posible concluir que la base de datos es adecuada para este trabajo, demostrando que es posible clasificar sujetos que tienen presencia de caries y diabetes y a su vez pacientes con ausencia de ambas enfermedades, haciendo uso de las 31 características demográficas.

Las curvas ROC obtenidas muestran un comportamiento similar, lo cual indica que la generalización del modelo permite clasificar sujetos tanto con presencia de ambas ENT como con ausencia de ambas.

Además con el presente trabajo se demuestra que la situación demográfica y socio-económica es un factor importante para el desarrollo de caries y diabetes.

Con las características seleccionadas para el desarrollo de esta investigación, fue posible la creación de un modelo que logra clasificar si un paciente padece caries y diabetes o en caso contrario, la ausencia de las mismas, con una precisión estadísticamente significativa.

Estos resultados basados en descriptores demográficos muestran la relación que existe con respecto a características correspondientes al nivel socio-económico, como son la cantidad de personas que viven en el mismo hogar; estado civil; origen étinico; edad; nivel académico e ingreso económico. Dicha relación permite diferenciar a sujetos que padecen las ENT de pacientes que no tienen ningún padecimiento de este tipo.

Es importante recalcar que este trabajo proporciona un conocimiento preliminar de las ventajas que tiene el hacer uso de este tipo de herramientas en el área de la salud, este modelo, puede ser utilizado como una herramienta de bajo costo y que dé soporte a los especialistas de

la salud con el objetivo de proporcionar diagnósticos preventivos de padecer diabetes y caries, buscando así disminuir la incidencia tan alta existente para estas dos enfermedades, principalmente en regiones rurales o con menor desarrollo económico, donde se complica el acceso a los diferentes servicios de salud necesarios para diagnósticar por los métodos tradicionales estas ENT.

**Trabajo Futuro.**

Como trabajo futuro se propone en primer lugar la reducción de características por medio de algoritmos genéticos para ver si es posible mantener un porcentaje de clasificación estadíticamente significativo pero con una cantidad menor de caracteríticas analizadas.

Además, se puede hacer uso de la misma base de datos pero con otras técnicas de aprendizaje automático y realizar una comparativa de los resultados obtenidos.

También se propone realizar el presente trabajo pero aplicado a una base de datos que contenga datos específicamente de la población mexicana, esto con el objetivo de prevenir o en su caso diagnósticar en etapas tempranas la caries y diabetes.

Finalmente, se puede hacer uso del modelo presentado en esta investigación para crear una herramienta de bajo costo que sirva de auxiliar a los médicos para la detección de estas enfermedades no transmisibles.

# Apéndice A: Descripción de descriptores demográficos.

| Caracteristica | Descripción |
|---|---|
| RIAGENDR | Género del participante. |
| RIDAGEYR | Edad en años de los participantes al momento de la toma de muestra. |
| RIDRETH1 | Información sobre raza y origen hispano. |
| RIDRETH3 | Información sobre raza y origen hispano con categoría de Asiaticos no hispano |
| RIDEXMON | Seis meses del periodo cuando fue realizada la evaluación. |
| DMQMILIZ | Fue servidor activo de las fuerzas armadas, reserva militar o guardía nacional en Estado Unidos |
| DMDBORN4 | País de nacimiento |
| DMDCITZN | Es ciudadano estadounidense. |
| DMDEDUC3 | Ultimo grado de estudios recibido |
| DMDEDUC2 | Ultimo grado de estudio completado. |
| DMDMARTL | Estatus marital |
| FIALANG | Idioma del instrumento de la entrevista. |
| DMDHHSIZ | Número de personas en el hogar. |
| DMDFMSIZ | Número total de personas en la familia. |
| DMDHHSZA | Número de niños con 5 años o menos. |
| DMDHHSZB | Número de niños de 6 a 17 años. |
| DMDHHSZE | Número de adultos con 60 años o más. |
| DMDHRGND | Género de referencia de las personas en el hogar. |
| DMDHRAGE | Edad de referencia de las personas en el hogar. |
| DMDHRBR4 | Lugar de referencia de las personas en el hogar. |
| DMDHREDU | Nivel de referencia de las personas en el hogar. |
| DMDHRMAR | Estatus marital de referencia de las personas en el hogar. |
| DMDHSEDU | Nivel de educación de referencia de las personas en el hogar. |
| WTINT2YR | Muestra completa del peso en dos años. |
| WTMEC2YR | Muestra completa de examen MEC por dos años. |
| SDMVPSU | Unidad de varianza enmascarada variable pseudo-PSU para la estimación de la varianza |
| SDMVSTRA | Unidad de varianza enmascarada variable peudo–stratum para la estimación de la varianza |
| INDHHIN2 | Ingreso total del hogar (reportado en un rango del valor de dolares) |
| INDFMIN2 | Ingreso total de la familia (reportado en un rango de valor en dolares.) |
| INDFMPIR | Una proporción de ingresos familiares a lineamientos de pobreza. |
| DX | Salida del dataset 0 para ausencia de caries y diabetes y 1 para presencia de ambas. |

Figura A.1 Descripción de las caracteristicas correspondientes al dataset utilizado.

# Apéndice B: Contribuciones.

## Deep Artificial Neural Networks for the Diagnosis of Dental Caries and Diabetes: Data from NHANES 2013–2014

**Vanessa Alcalá-Rmz[1], Carlos E. Galván-Tejada[1], Laura A. Zanella-Calzada[1], Nubia M. Chávez-Lamas[2], Jorge I. Galván-Tejada[1], Manuel Haro-Márquez[3].**

{vdrar.06, ericgalvan, lzanellac, nubiachavez, gatejo}@uaz.edu.mx, mharo@cozcyt.gob.mx

[1]Unidad Académica de Ingenieréa Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, 98000, Zacatecas, Zac, Mexico;
[2]Clínica Comunitaria de Tacoaleche, Unidad Académica de Odontología, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, 98000, Zacatecas, Zac, Mexico;
[3]Laboratorio de Software Libre, Consejo Zacatecano de Ciencia y Tecnología, Av. de la Juventud 504, Zona A, Javier Barros Sierra, 98090, Zacatecas, Zac, Mexico.

**Resumo**: *En los últimos años, el estudio de la relación entre dos enfermedades no transmisibles, caries dental y diabetes mellitus, está aumentando, dada la incidencia y prevalencia de éstas. La caries dental es una enfermedad multifactorial, con diversos factores de riesgo que contribuyen a su inicio y progresión. Los factores de riesgo pueden clasificarse como biológicos, ambientales o socioconductuales. También, la caries dental es un importante problema de salud pública a nivel mundial y es la enfermedad no transmisible más extendida. La diabetes también es una enfermedad no transmisible que está afectando a la población mexicana en niveles preocupantes. México ocupa el primer lugar en prevalencia de diabetes con un 15,9 %. La caries dental y la diabetes son enfermedades crónicas pero prevenibles, por esta razón en este trabajo propusimos una red neuronal artificial (RNA) para clasificar a los sujetos con presencia / ausencia de caries dental y diabetes mellitus. El análisis se basó en 31 características que determinan el estado del paciente. El modelo propuesto se evalúa mediante un análisis estadístico basado en el cálculo de la función de pérdida, la precisión, el área bajo la curva (AUC, por sus siglas en inglés) y la curva de características operativas receptoras (ROC, por sus siglas en inglés). Los resultados obtenidos son significativos, con una precisión de 0,99, valores de AUC de 0,99 y curvas ROC de clase 1 para micro-promedio y 0,99 para macro-promedio.*

**Palabras claves**: Diabetes mellitus, Caries dental, NHANES 2013-2014, Red neuronal artificial, Diagnóstico asistido por computadora, Análisis estadístico.

**Abstract**: *In recent years, the study of the relationship between two noncommunicable diseases, dental caries and diabetes mellitus, is increasing, given the incidence and prevalence of these. Dental caries is a multifactorial disease, with many risk factors contributing to their initiation and progression. The risk factors can be categorized as biological, environmental or socio-behavioral, also dental caries is a major public health problem globally and is the most widespread noncommunicable disease. Diabetes is also a noncommunicable disease that is affecting the Mexican population in worrying levels. Mexico has the first place in prevalence diabetes with a 15.9 %. Dental caries and diabetes are chronic but preventable diseases, for this reason in this work we proposed an artificial neural network to classify subjects with presence/absence of dental caries and diabetes mellitus, analysis was based on 31 features that determines the patient status. Proposed model is evaluated through a statistical analysis based on the calculus of loss function, accuracy, area under the curve (AUC) and Receiving Operating Characteristics (ROC) curve. The results obtained are significant, with an accuracy of 0.99, AUC values of 0.99 and a ROC curves of class of 1 for micro-average and 0.99 for macro-average.*

**Keywords**: Diabetes mellitus, Dental caries, NHANES 2013-2014, Artificial Neural Network, Computer-aided diagnosis, Statistical analysis.

### 1 Introduction

Noncommunicable diseases (NCDs), also known as chronic diseases, tend to be of long duration and are the result of a combination of genetic, physiological, environmental and behaviors factors [1]. People of all age groups, regions and countries are affected by NCDs. These conditions are often associated with older age groups, but evidence shows that 15 million of all deaths attributed to NCDs occur between the ages of 30 and 69 years. Of the "premature" deaths, over 85% are estimated to occur in low- and middle-income countries. Children, adults and the elderly are all vulnerable to the risk factors contributing to NCDs, whether from unhealthy diets, physical inactivity, exposure to tobacco smoke or the harmful use of alcohol [1].

A determinant factor that increases the risk of developing NCDs are oral diseases, which are part of the components that most affect the quality of life of people, being an important point in health care. The most frequent condition in oral health is dental caries, defined as a multifactorial disease and considered a major public health problem worldwide, being the most widespread NCD. According to the Global Burden of Disease Study, in 2015, ranking first for decay of permanent teeth (2.3 billion people) and 12th for deciduous teeth (560 million

children). Besides, according to the World Health Organization (WHO), dental caries affects between 60 % and 90 % of the population.

The greatest burden of this disease is in disadvantaged and socially marginalized populations, which represents a problem taking into account that the treatment of these conditions is extremely expensive, with oral diseases being the fourth most expensive cause to treat, according to the WHO, limiting their access in countries [2].

Dental caries causes pain and a local systemic infection in a specific region that progresses towards the dental pulp, which if left untreated can develop a dental abscess [3]. It presents many risk factors contributing to their initiation and progression, which can be categorized as biological, environmental and socio-behavioral [4], [5]. According to the Organización Panamericana de la Salud, the development of dental caries depends on the frequency of carbohydrate consumption, cariogenic plaque, saliva, the characteristics of the food, the time exposure, plaque removal and susceptibility of the guest, as well as the few preventive measures in oral health and the limited access to specialized dental medical services.

The factors mentioned above interact in a simultaneous way, and they correspond to different orders from biological processes, to complex historical-cultural structures and social relationships, socioeconomic level, education level, among others, making oral health a complex phenomenon.

It is important to mention that among the NCDs that increase their risk of occurrence due to oral diseases are cardiovascular and cerebrovascular diseases, as well as diabetes mellitus. Diabetes mellitus is a NCD that according to the WHO and the International Diabetes Federation (IDF), the number of people with this disease is increasing very fast all over the world, being nowadays one of the leading causes of death and disability worldwide and represents one of the greatest challenges of the 21 century for the health and countries development [6], [7].

The NCD Risk Factor Collaboration (NCD-RisC) estimates that the number of people with diabetes quadrupled between 1980 and 2014. Age-standardized prevalence among adult men doubled during that time (from 43% to 90% ), and age-standardized prevalence among adult women increased by 60% ( from 50% to 79%) [7], coupled of this, diabetes has an important global connotation considering that in 1994 there were 100 million people with this disease, 165 million in the year 2000 and forecast 300 million in 2025. In the United States of America (USA), there is a significant percentage of diabetes mellitus cases with a forecast of 65 million in the year 2025 [8], [9].

Diabetes is defined as a multifactorial and polygenic metabolic disorder, and its pathogenesis is influenced by diverse environmental and genetic risk factors [10]. This disease is characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both [11], consisting in a complex disorder involving profound alterations in the metabolism of carbohydrates, fats and proteins [12]. Among its main effects are long-term damage, dysfunction and failure of various organs, as well as the vascular complications that shorten the life expectancy of those suffering from this disease, with cardiovascular manifestations being one of the recent diagnoses at the time of detection of diabetes in approximately 25 % of patients. [13].

Also, diabetic patients are associated with intensive loss of fluid due to polyuria, reduced response to infections, impaired connective tissue metabolism, and various microvascular changes. These factors are responsible for various oral diseases in diabetic patients including xerostomia, salivary gland dysfunction, increased susceptibility to bacterial, viral and fungal infection, periapical abscesses, loss of teeth, taste impairment and dental caries, among others [14].

Type 2 diabetes mellitus (T2DM) is an expanding global health problem, which 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. Besides, T2DM has been closely linked to the epidemic of dental caries, since the prevalence of this condition is higher in patients with T2DM compared to non-diabetic patients [15], [16], [17], [18], [19].

One of the main reasons that makes it difficult to control the incidence of diabetes and dental caries is that both are multifactorial diseases; nevertheless, the implementation of algorithms and the development of analyzes with different approaches has been proposed, based on computer-aided diagnosis (CADx), to support the preventive diagnosis and the reduction of the high prevalence, looking for the development of prediction and classification models contained by the main factors that affect these conditions [20], [21].

One of the algorithms that have been implemented based on CADx tools are Artificial Neural Networks (ANN), which are defined as mathematical models based on a principle of learning that follow the concept of artificial intelligence and the biological response of the human brain. ANN is a nonlinear method that allows the integration of variables and easily handle large amounts of data compared to linear analyzes, being one of the reasons why this procedure is useful for complex pattern recognition problems. The analysis with this algorithm may be better able to cope with the variability because this system uses mathematical "weights" to decide the probability that the input data belongs to a particular output. These weights are adjusted by training the network with labeled data or with known outputs. Subsequently, the network can be tested with the unknown data, providing a probability of such data belonging to a particular output [22].

According to the literature, the implementation of CADx systems has been widely used for the study of the relationship between dental caries and diabetes, looking for the development of supportive tools that may help to the preventive and automated diagnosis of these conditions, in order to reduce their high incidence. In the

study of Lai et al. [23] is developed an evaluation of the difference in caries experience in diabetic and no diabetic children, obtaining that a higher number of caries free subjects was found in diabetic subjects in good metabolic control ($p < 0.1$), indicating that diabetic children with bad metabolic control are prone to a high caries risk.

Ferizi et al. [24] propose a work based in a CADx analysis looking for the influence of type 1 diabetes mellitus (T1DM) on dental caries. Through a statistical comparison, the data collected from a set of controls and cases (presence of T1DM and dental caries) was analyzed using the chi-square test and Mann-Whitney U-test, obtaining a significant relationship between subjects with T1DM and the presence of dental caries ($p < 0.001$), being possible to conclude that T1DM has an important part in oral health, since it appears that those with T1DM present a higher risk for caries. In the work of Montalvan et al. [25] is presented an expert system that proposes the nutritional diagnosis of dental caries and diabetes, among other diseases. The system is able to give an automated diagnosis based on the symptoms of the patient, subsequently generating a suitable diet according to their detected nutritional status. This system uses the multilayer perceptron (MLP) algorithm, which is a multilayer ANN that assigns values to each of the nodes, looking to minimize the error function. Also, Barylo et al. [26] developed a study of the effects that diabetes mellitus presents on patients' oral health through a statistical analysis using a student's test. According to the obtained results, diabetes has a direct effect on oral health, showing that the level of medical and preventive dental care must be increased for diabetes patients. Song et al. [27] present a study of the association between T2DM and untreated dental caries, in order to identify diabetes as a risk factor of caries, showing through their results that the prevalence of untreated caries uncontrolled T2DM subjects was about 26 % higher than those with normal glucose tolerance levels. Finally, in the work of Latti et al. [23] is performed an evaluation of the effects of diabetes mellitus on dental caries micro-organisms responsible for caries, through a depth-first search (DFS) index technique. Results shown dental caries, among other factors, increases in diabetics than in control subjects, existing a relationship between diabetes mellitus, oral microbiota and dental caries.

In the present study, the relationship that exists between several features that involve the appearance of diabetes and dental caries in the same patient is analyzed, looking for the information that possible links these two NCDs. The contribution of this study is presented in two parts, first, to demonstrate that the same features can contribute to describe the presence of dental caries and diabetes, and second, the use of an ANN to determine if the patient has developed diabetes and dental caries simultaneously, obtaining a classification system that allows to look for a possible future prediction of diabetes and dental caries, presenting preliminary results.

## 2 Materials and Methods

In this work is presented the development of a technique based on ANNs to classify the presence of dental caries and diabetes or the absence of dental caries and diabetes in a subject.

In this section is described the data from the National Health and Nutrition Examination Survey (NHANES) 2013-2014 that were used for this work, as well as the data preprocessing, data classification and validation of the results.

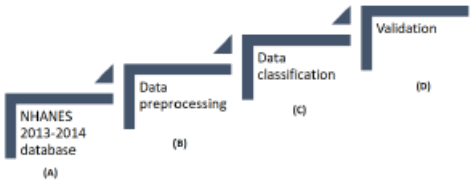


Figure 1: Flowchart of the methodology followed. (A) Dataset used from the NHANES 2013-2014, (B) Data preprocessing, (C) Data classification and (D) Validation of the ANN performance.

Figure 1 presents the flowchart of the methodology followed. Section (A) is shown the database used, which specifically corresponds to demographic data. In section (B) is presented a preprocessing step where it was necessary to prepare the data for a correct analysis of the features, then, in section (C) is shown the data classification of subjects according to their status of diabetes and dental caries. Section (D) makes reference to a validation step based on the receiver operating characteristic (ROC) curve and the area under the curve (AUC) in order to measure the accuracy and specificity rate of the ANN performance in the classification of the patients.

### 2.1 Data Description

NHANES is a program that designs a series of studies to perform a survey in order to measure the health and nutritional status of adults and children in the United States (U.S.), which was founded by the National Center for Health Statistics (NCHS). The survey combines physical examinations and interviews, allowing to develop studies through different features of individuals.

The general structure of the NHANES survey includes several types of interviews, including demographic, dietary and health-related questions. In Table 1 are presented the descriptions of the datasets used in this work.

### 2.2 Meta Data

The NHANES program interviewed 27,631 subjects. The subjects that participated in the interviewees were randomly selected through an algorithm which consists of a complex multistage probability design with a series of stages, and the subjects.

The selection was performed with women and men that belong to different counties from the U.S. These counties were divided into 15 groups, then, from each group one county was randomly selected obtaining 15 counties. After that, a sampling of segments was performed to each

| Questionnaire | Description |
|---|---|
| Demographic | The demographics file provides individual, family, and household level information. |
| Examination | Public health significance in areas of surveillance, prevention, treatment, dental care utilization, health policy, evaluation of Federal health programs, among others. |
| Questionnaire | Information on: acculturation, alcohol use, diabetes status, health insurance, income, among others. |

county, selecting between 20 and 24 segments, and for each segment there was selected a sample of about 30 households. Finally, the sampling of people was interviewed with the information of the survey.

The sampling of people contained in the NHANES data set has as main target population the non-institutionalized civilian resident population of the U.S, including Hispanic, Non-Hispanic black, Non-Hispanic Asian, Non-Hispanic white and other persons at or below 130 % of the poverty level, besides Non-Hispanic white and other persons aged 80 years and older.

The demographic data set contains 9,801 subjects (male = 4,826 / female = 4,975) that belong to an age range from 0 to 80 years old, and a total of 39 features.

### 2.3 Data Preprocessing

Initially, only those subjects that presented the information referenced to the dental caries and diabetes status in all the data sets were kept, being subsequently mixed in one data set. Then, the subjects that presented incomplete information were eliminated, as well as those that did not present a positive or a negative status in both conditions, dental caries and diabetes. On the other hand, the features that presented > 70 % of missing data or singular values were removed.

The feature used as output presented two states: presence of dental caries and diabetes, labeled as '1' and absence of dental caries and diabetes, labeled as '0'.

The Z-score method was used to normalize the 31 features in the dataset, this method was selected because transforms the data to a distribution with mean 0 and standard deviation 1 and due to usually the data are not defined in the same numeric scale. Once applied Z-score method, the data are adequate for the classification.

Finally, for the purpose of validating the proposed analysis, the data set was divided in two subsets, one for training, containing 70 % of the data and one for testing, containing the 30 % remaining data.

### 2.4 Data Classification

There were two possible classifications for the patients, control patients, '0', which are those with absence of dental caries and diabetes, and case patients, '1', which are those with presence of dental caries and diabetes. This step was performed using a dense ANN that was specifically designed for these data set, using the packages "Keras" and "TensorFlow", for Python.

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, makes it very attractive for rapid application development, as well as for a scripting or glue language to connect existing components together. Python is a simple and easy tool to learn where syntax emphasizes readability, reducing the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed [28].

Keras is a high-level ANN API, designed to perform fast experimentation using deep ANN, focusing on being user-friendly, modular, and extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System) [29].

Tensorflow is an open-source software library for data flow programming across a range of tasks. It is a symbolic math library, used for machine learning applications such as ANN [30].

On the other hand, ANNs search a solution to a task through the correlation between features, based in the learning or training process that imitates the behavior of biological neural networks.

ANNs gather their knowledge by detecting the patterns and relationships in data and learn through experience, not from programming. A deep ANN is formed from hundreds of single units, called neurons, connected with coefficients, called weights, which constitute the neural structure and are organized in layers. The number of layers is modifiable as the number of neurons in each layer. The power of neural computations comes from connecting neurons in a network. Each processing element has weighted inputs, a transfer function and one output. The behavior of an ANN is determined by the transfer functions of its neurons, the learning rule, and the architecture itself. The weights are the adjustable parameters, and, in that sense, an ANN is a parameterized system [31].

The deep ANN that was designed for this work presented the structure mentioned below:

The input layer was assigned with 31 neurons that represents the 31 features in the data set.

- The first dropout hidden layer had a loss percentage of 25%.
- The first dense hidden layer was composed by 100 neurons.

- The second dropout hidden layer had a loss percentage of 50%.
- The second dense hidden layer was assigned with 500 neurons.
- The third dropout hidden layer had a loss of 25%.
- The third dense hidden layer was composed by 100 neurons.
- The fourth dropout hidden layer had a loss of 50%.
- Finally, the fourth dense hidden layer was the output layer and it was characterized by two neurons that refers to the two possible classifications.

The optimization algorithm implemented was ``Adam", which calculates the exponential moving average of the gradient and the square gradient. This process is based in the stochastic gradient descent algorithm, using the average of the first and second moments of the gradients, with the purpose of controlling the decay of that moving average [32].

Besides, there were implemented two activation functions, the first one is called rectified linear unit (ReLU), which was used in the dense layers (except in the output layer). This function assigns '0' to the neurons that present a value lower than '0' and assigns the original value when this is upper or equal to '0'. ReLU is shown in Equation 1 [33].

$$ReLU(z) = \begin{cases} if\ z < 0 & 0 \\ if\ z \geq 0 & z \end{cases} \quad (1)$$

The second activation function is called normalized exponential or Softmax. This function is obtained from a general logistic function, compressing a vector of arbitrary values into a vector of values located in [0,1]. Equation 2 represents this function, where $\sigma(z)$ is a $K$-dimensional vector of $z$ [34].

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{K=1}^{K} e^{z_K}}, j = 1, \ldots, K \quad (2)$$

Also, it is important to consider the number of epochs, which is a configurable parameter that refers to the number of iterations of ANN. The main objective is to find the optimal number of epochs through some test varying the epoch number. For this work, there were established 100 epochs according to the results obtained through many tests with different number of epochs, obtaining the best accuracy and loss function with 100 epochs.

## 2.2 Validation

The parameters calculated in this stage of the methodology were the loss function, accuracy and ROC curve. These values were calculated on each epoch and the ROC curve was obtained with the average of the general behavior of the deep ANN. It is possible to determine when the model is fitting better to the data because the value of the loss function goes down because the global minimum was found or is approaching it, which represents the minimum error. Also, the loss function is able to optimize the network feeding back with information of the capacity of the system [35].

"Binary Cross-Entropy" was chosen as method to calculate the loss function, which is included in the Keras package. This method uses the Kullback-Leibler distance principle, which consists in a measure between two density functions $g$ and $h$. Cross-entropy is an iterative method that generates a set of random values that are updated in order to generate approximate values [36]. For this work the accuracy function selected obtains the average accuracy based in the total predictions and it is used in binary classification problems, this function is called "binary-accuracy", which is contained in the keras package. The accuracy parameter is calculated with the Equation 3, which is based in the difference between the calculated classification and the real classification and is represented as 1-error, $Vpred$ is the calculated classification value and $Vtrue$ is the true classification value. This value is obtained for each model, giving the option to select the model that present the better performance [37].

$$error = V_{pred} - V_{true} \quad (3)$$

There was also obtained the ROC curve, which is used to measure the precision of the classification model and it is based on the specificity and sensitivity rate. The specificity represents the number of control subjects that were correctly classified, and it is calculated with Equation 4, where $NPV$ represents the negative predictive value, $TN$ represents the true negatives and $FN$ represents the false negatives [38].

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

The sensitivity represents the number of case subjects that were correctly classified, and it is calculated with Equation 5 where $PPV$ represents the positive predictive value, $TP$ represents the true positives and $FP$ represents the false positives [39].

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

The ROC curves were calculated for each class, as well as for the macro-average and micro-average precision. The macro-average precision is the average accuracy in different arbitrary sets, calculated with Equation 6, where $A_1$ represents the average of the set one and $A_2$ represents the average of the set two.

$$Macro - average = \frac{A_1 + A_2}{2} \quad (6)$$

The micro-average precision is the sum of the total true positives, false positives and false negatives for different aleatory sets, calculated with Equation 7, where $TP_1$ represents the true positives of the set one, $TP_2$ represents the true positives of the set two, $FP_1$ represents the false positives of the set one and $FP_2$ represents the false positives of the set two [40].

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

As extra information it is mentioned that the computer that was used for the development of this work was a laptop Toshiba Satellite S55T-B5233", Intel Core i7-7500U 2.70GHz, 16GB, 500GB SSD, Ubuntu 16.4, 64-bit; and the version of Python used is 2.7.

All the analysis was performed using the packages, Keras 2.1.5, Scipy 1.0.1, Pandas 0.22.0, Sklearn 0.191 and Tensorflow 1.7.0.

## 3 Results

In the preprocessing step, the 9,801 subjects were reduced to 3,552 (male = 1,812 / female = 1,740). In Figure 2 is shown the number of control and case subjects. Besides, the 39 demographic futures were reduced to 32, shown in Table 2.

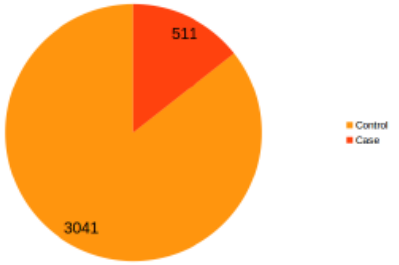


Figure 2: Graph of the number of patients contained in the data set.

Then, the data set was divided in a training set (2125 controls / 361 cases), and a testing set (916 controls / 150 cases), as shown in Figure 3.

Then, the 32 features were subjected to the deep ANN constructed. The training data set was evaluated at each epoch through the accuracy and the loss function. The number of epochs was tested with different values as shown in Table 2, in order to find the number of epochs that present the best result.

Also, a comparison of the results using different type of layers, number of layers and number of neurons, was performed, as shown in Table 3. The number of epochs selected was 100, in a deep ANN with nine layers (five dense layers and four dropout layers). The dense layers contained: 31, 100, 500, 100 and 2 neurons, and the dropout layers presented: 0.50, 0.25 and 0.50 percentages.


Figure 3: Graph of the number of patients that belongs to the training and testing data sets.

Table 2: Results of the accuracy and loss function values using different number of epochs.

| Epochs | Accuracy | Loss function | Processing time (s) |
|---|---|---|---|
| 10 | 0.9759 | 0.0824 | 3.6282 |
| 50 | 0.9916 | 0.0216 | 15.4416 |
| 100 | 0.9988 | 0.0048 | 29.2827 |
| 150 | 0.9992 | 0.0088 | 45.2142 |
| 200 | 0.9984 | 0.0047 | 58.9482 |
| 300 | 0.9980 | 0.0160 | 89.2401 |

In Figure 4 is shown the behavior of the accuracy, where the blue line refers to the training data, reaching an accuracy value of 0.9964, and the orange line refers to the testing data, reaching an accuracy value of 0.9906.

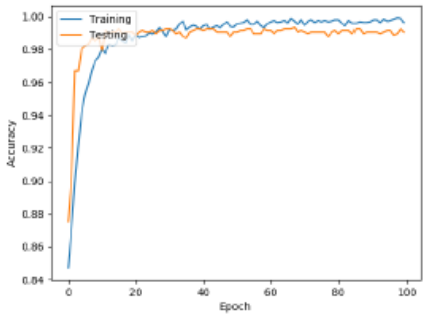


Figure 4: Graph of the accuracy behavior.

In Figure 5 is shown the behavior of the loss function, where the blue line refers to the training data, reaching a loss function value of 0.0099, and the orange line refers to the testing data, reaching a loss function value of 0.0945.

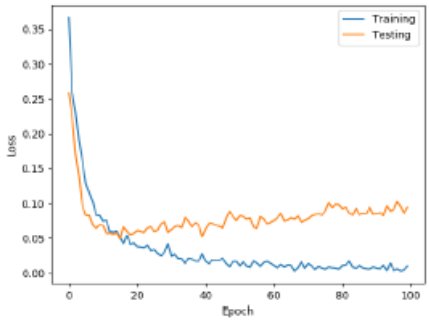


Figure 5: Graph of the loss function behavior.

Finally, in Figure 6 are presented the ROC curves that represent the performance of the deep ANN. The pink line refers to the ROC curve of the class '0' or the control subjects, which obtained an AUC = 0.99. The light blue line refers to the ROC curve of the class '1' or the case subjects, which obtained an AUC = 0.99. The dotted orange line refers to the ROC curve of the micro-average, which obtained an AUC = 1, and the dotted dark blue line refers to the ROC curve of the macro-average, which obtained an AUC = 0.99.

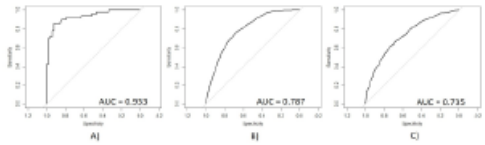


Figure 6: ROC curves obtained with the average performance of the ANN.

## 4 Discussion

The data set used in this work was separated in two through an aleatory and balanced selection. One set was used for the training of the deep ANN, which consisted in 100 epochs. This number of epochs was selected based in a comparison of different values, as shown in Table 2. Then, the ANN was optimized with the Adam algorithm, in order to improve the behavior of the ANN through a feedback.

On the other hand, the test data set was used to evaluate the ANN behavior, validating the results of the training stage through the accuracy and lost function values.

The relationship between controls and cases is apparently disproportionate, however, the number of data for both conditions was enough to reach a significant performance in the classification of subjects.

According to the graph of the accuracy, 99 % of the subjects were correctly classified between cases and controls, and according to the graph of the loss function, it is possible to observe a decreasing behavior for both data sets, which implies that the ANN is approaching to the global minimum.

Besides, Figure 4 and Figure 5 show that the behavior of the training and the testing data follow a pattern, indicating that the model developed through the constructed ANN reached a generalization in the learning process.

On the other hand, even when the accuracy and loss function values were obtained to measure the behavior of the classification, the accuracy may provide a better result than the real if the data has an intrinsic bias. Due to this, the AUC value was also obtained to check if the results were statistically significant.

In Figure 6 is shown that all the ROC curves obtained an AUC value statistically significant $\geq$ 0.99, which means that the ANN model was able to classify 99 % of the data correctly.

All the ROC curves presented a similar performance, which implies that the classification performance is equitable, and the generalization of the model allows to classify subjects of both classes.

The difference of the AUC value that is presented between the macro-average and micro-average, where the value of the micro-average is greater than the value of the macro-average, may be due to the calculation of these parameters, since the macro-average is obtained from true positives and true negatives rate using the whole amount of data, and the micro-average is obtained from subsets of data, where some subsets could present better AUC value than the obtained with the whole set.

## 4 Conclusions

According to the results obtained from the evaluation step, it was possible to conclude that the database was adequate for this work, demonstrating that is possible to classify subjects that have presence of dental caries and diabetes from subjects with absence of dental caries and diabetes using the information of the 32 demographic features.

Therefore, through this work it is demonstrated that the demographic situation is an important factor for the development of both, dental caries and diabetes.

Finally, it is important to remark that this work gives a preliminary knowledge of the advantages that this kind of implementations could have in health, since it is conclude that the developed model can be used as a tool to support specialists for the simultaneous preventive diagnosis and prediction of diabetes and dental caries, looking for decreasing the high incidence of these diseases.

## 4 Future Work

Based on this work, it can be proposed as a future work the analysis of a data set with exclusively Mexican subjects information and a comparison with the results obtained here, with the propose of proving how demography can influence in the absence or presence of dental caries and diabetes.

Table 3: Values of accuracy, loss function and processing time with different number of layers and neurons.

| Layers Dense / Dropout | Neurons | Accuracy | Loss function | Processing time (s) |
|---|---|---|---|---|
| 2/0 | 31 >2 | 1 | 0.0003 | 12.4599 |
| 3/1 | 31 >100 >0.5 >2 | 1 | 0.0002 | 15.0060 |
| 3/2 | 31 >0.25 >100 >0.5 >2 | 0.9984 | 0.0048 | 15.9131 |
| 4/1 | 31 >100 >0.5 >500 >2 | 0.9988 | 0.0028 | 20.7666 |
| 4/2 | 31 >0.25 >100 >0.5 >500 >2 | 0.9988 | 0.0052 | 21.7124 |
| 4/3 | 31 >0.25 >100 >0.5 >500 >0.25 >2 | 0.9988 | 0.0044 | 24.0035 |
| 5/1 | 31 >100 >0.5 >500 >100 >2 | 0.9992 | 0.0048 | 25.5583 |
| 5/2 | 31 >0.25 >100 >0.5 >500 >100 >2 | 0.9992 | 0.0028 | 26.7241 |
| 5/3 | 31 >0.25 >100 >0.5 >500 >0.25 >100 >2 | 0.9996 | 0.0018 | 28.8960 |
| 5/4 | 31 >0.25 >100 >0.5 >500 >0.25 >100 >0.5 >2 | 0.9964 | 0.0099 | 29.8431 |

## Referencias bibliográficas


[1] Noncommunicable diseases. Avaiable online: http://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. Accessed: September 20, 2018.

[2] World Health Organization (NCHS). Oral health. urlhttp://www.who.int/oralhealth=diseaseburden=global=en=:accessedonJune052017.

[3] 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-yearold hong kong children: a cross-sectional study. BMC oral health, 17(1):121, 2017.

[4] RH Selwitz, AI Ismail, and NB Pitts. Dental caries. 2007.

[5] WHO. Sugars and dental caries. 2017.

[6] WHO. Global status report on noncommunicable diseases 2010.

[7] G. Etienne. Trends in diabetes: sounding the alarm. 387:1485–1486, 2016.

[8] Carty D Mc and P. Zimmet. Diabetes1994 to 2010. global estimates and projection. Kobe,Japan:International Diabetes Federation Congress, 1994.

[9] colectivo de autores. Guas ALAD de Diagnstico, Control y Tratamiento de la Diabetes Mellitus Tipo 2., 2006.

[10] M Cruz, A Valladares-Salgado, J Garcia-Mena, K Ross, M Edwards, J Angeles-Martinez, C Ortega-Camarillo, J Escobedo de la Pea, A. I. Burguete-Garcia, N. Wacher-Rodarte, R Ambriz, R Rivera, A L D'artote, J Peralta, Esteban J Parra, and J Kumate. Candidate gene association study conditioning on individual ancestry in patients with type 2 diabetes and metabolic syndrome from mexico city. 26:261–270, 2010.

[11] American diabetes association. Diagnosis and classification of diabetes mellitus. 2010.

[12] John R. Turtle. What is diabetes mellitus? 1969.

[13] MM Bolet Socarras, J M Blanco, A Vazquez, D Gonzles, and ME. Licea. Factores de riesgo de ateroesclerosis en la diabetes mellitus tipo 2. Revista Cubana Med, 42:17–25, 2003.

[14] IDF WHO. Diabetes action now. 58, 2008.

[15] I Singh, P Singh, A Singh, T Singh, and R Kour. Diabetes an inducing factor for dental caries: A case control analysis in jammu. Journal of International Society of Preventive Community Dentistry, 2016.

[16] Ralph A. DeFronzo, Ele Ferrannini, Leif Groop, Robert R Henry, William H Herman, Jens Juul Holst, Frank B Hu, C. Ronald Kanh, Itamar Raz, Gerald I Shulman, Donal C Simonson, Marcia A Testa, and Ram Weiss. Type 2 diabetes mellitus. 2015.

[17] R.V. Sarmiento, F.P. Barrionuevo, Y.S. Huamn, and M.C. Loyola. Prevalencia de caries de infancia temprana en nios menores de 6 aos de edad, residentes en poblados urbano marginales de lima norte. Rev. Estomatol., 21:79–86, 2011.

[18] C.D. Oropeza-Oropeza, N. Molina-Frechero, C.D. Castaeda-Castaneira, E.and Zaragoza-Rosado, and C.D. Cruz Leyva. Caries dental en primeros molares permanentes de escolares de la delegación tlahuac. Rev. ADM., 69:63–68, 2012.

[19] B.J. Cardozo, M.M. Gonzalez, S.R. Prez, P.A. Vaculik, and E.G. Sanz. Epidemiologa de la caries dental en nios del jardn de infantes pinocho de la ciudad de corrientes. 2017.

[20] E.d.l.A. Gispert Abreu, P. Castell-Florit Serrate, and M. Herrera Nordet. Salud bucal poblacional y su produccin intersectorial. 52:62–67, 2015.

[21] P. Singh, A.B. Pal, M. Anburajan, and J. Kumar. Computer-aided diagnosis of diabetes mellitus using thermogram of open mouth. Computing and Intelligent Engineering, 52:62–67, 2018.

[22] GG Gardner, D Keating, TH Williamson, and AT Elliott. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. British Journak of Ophtalmology 1996, 80:940–944, 1996.

[23] Bhagyashri Ramachandra Latti, Jitendra V Kalburge, Sanjeev Bhimashankar Birajdar, and Ramachandra Girimallappa Latti. Evaluation of relationship between dental caries, diabetes mellitus and oral microbiota in diabetics. Journal of oral and maxillofacial pathology: JOMFP, 22(2):282, 2018.

[24] Lul¨ejeta Ferizi, Fatmir Dragidella, Lidvana Spahiu, Agim Begzati, and Vjosa Kotori. The influence of type 1 diabetes mellitus on dental caries and salivary composition. International Journal of Dentistry, 2018, 2018.

[25] P. Montalvn, K. Michay, S. Snchez, and P. Snchez. Desarrollo e implementacin de un sistema experto nutricional que permita diagnosticar enfermedades generales nutricionales de acuerdo a los sntomas y emitir su tratamiento correspondiente en el rea de enfermera del departamento de orientacin y bienestar estudiantil (dobe) de la unidad educativa san jos de calasanz de la ciudad de loja. 2012.

[26] S Barylo, Kanishyna, and LI Shkilniak. The effects of diabetes mellitus on patients' oral health. Wiadomosci lekarskie (Warsaw, Poland: 1960), 71(5):1026–1031, 2018.

[27] Stefano Lai, Maria Grazia Cagetti, Fabio Cocco, Dina Cossellu, Gianfranco Meloni, Guglielmo Campus, and Peter Lingstr¨om. Evaluation of the difference in caries experience in diabetic and nondiabetic childrena case control study. PloS one, 12(11):e0188451, 2017.

[28] Python Community. What is python?. available on: https://www.python.org/doc/essays/blurb/ (accessed on september 2018).

[29] Francois Chollet. Keras: Deep learning library for theano and tensorflow. available on: https://keras. io/k (accessed on June 2018).

[30] Google. Tensorflow. available on: https://www.tensorflow.org/ (accessed on september 2018).

[31] S Agatonovic-Kustrin and R Beresford. Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research. Journal of Pharmaceutical and Biomedical Analysis, 22:717–727, 2000.

[32] D P Kingma and J Ba. Adam: A method for stochastic optimization. 2014.

[33] Alessio Lomuscui and Lalit Maganti. An approach to reachability analysis for feed-forward relu neural networks. 2017.

[34] Nicholas Carlini and David Wanger. Towards evaluating the robustness of neural network. 2017.

[35] C Antona Corts. Herramientas modernas en redes neuronales: la librera keras. 2017.

[36] S Kullback and R Leibler. On information and sufficiency, anals of mathematical statistics. pages 76–86, 1951.

[37] Maxwell Nye and Andrew Saxe. Are efficient deep representations learnable? 2017.

[38] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology, 1982.

[39] Francois Chollet. Keras: Deep learning libary for theano and tensofirflow. avaiable on: https://keras. io/k. 2011.

International Journal of
*Environmental Research and Public Health*

MDPI

*Article*

# Identification of Diabetic Patients through Clinical and Para-Clinical Features in Mexico: An Approach Using Deep Neural Networks

**Vanessa Alcalá-Rmz [1,†], Laura A. Zanella-Calzada [1,†], Carlos E. Galván-Tejada [1,*,†], Alejandra García-Hernández [1], Miguel Cruz [2], Adan Valladares-Salgado [2], 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; vdrar.06@uaz.edu.mx (V.A.-R.); lzanellac@uaz.edu.mx (L.A.Z.-C.); alegarcia@uaz.edu.mx (A.G.-H.); gatejo@uaz.edu.mx (J.I.G.-T.); hamurabigr@uaz.edu.mx(H.G.-R.)

[2] Unidad de Investigación Médica en Bioquímica, Hospital de Especialidades, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Av. Cuauhtémoc 330, Col. Doctores, Del. Cuauhtémoc, Ciudad de México CP 06720, Mexico; miguel.cruzlo@imss.gob.mx (M.C.); adan.valladares@imss.gob.mx (A.V.-S.)

* Correspondence: ericgalvan@uaz.edu.mx; Tel.: +52-49-2544-0968

† These authors contributed equally to this work.

Received: 29 December 2018; Accepted: 21 January 2019; Published: 29 January 2019

check for updates

**Abstract:** Diabetes is a chronic and noncommunicable but preventable disease that is affecting the Mexican population at worrying levels, being the first place in prevalence worldwide. Early diabetes detection has become important to prevent other health conditions that involve low organ yield until the patient death. Based on this problem, this work proposes the architecture of an Artificial Neural Network (ANN) for the automated classification of healthy patients from diabetics patients. The analysis was performed used a set of 19 para-clinical features to determine the health status of the patients. The developed model was evaluated through a statistical analysis based on the calculation of the loss function, accuracy, area under the curve (AUC) and receiving operating characteristics (ROC) curve. The results obtained present statistically significant values, with accuracy of 0.94 and AUC values of 0.98. Based on these results, it is possible to conclude that the ANN implemented in this work can classify patients with presence of diabetes from controls with significant accuracy, presenting preliminary results for the development of a diagnostic tool that can be supportive for health specialists.

**Keywords:** type 2 diabetes; Artificial Neural Network; net reclassification improvement; computer-aided diagnosis; statistical analysis

## 1. Introduction

According to the World Health Organization (WHO) and the International Diabetes Federation (IDF), the number of people with diabetes is increasing very quickly all over the world. Diabetes is one of the Noncommunicable Diseases (NCD), also known as chronic diseases, which are characterized by being of long duration and the result of a combination of different factors: genetic, environmental and behavior factors [1].

Diabetes is a major cause of death and disability worldwide and represents one of the greatest challenges of the 21st century for health and development. The NCD Risk Factor Collaboration (NCD-RisC) estimated that the number of people with diabetes quadrupled between 1980 and 2014. Age-standardized prevalence among adult men doubled during that time (from 4.3% to 9.0%), and

age-standardized prevalence among adult women increased by 60% (from 5.0% to 7.9%) [2]. Moreover, diabetes has an important global connotation considering that, in 2017, 451 million people aged 18–99 years lived with diabetes and the number of people are predicted to rise to 693 million by 2045 [3,4].

Diabetes is defined as a group of metabolic diseases characterized by hyperglycemia resulting from defects in insulin secretion, insulin action, or both [5]. It consists of a complex disorder involving profound alterations in the metabolism of carbohydrates, fats and proteins [6].

In the scientific literature, this condition has been described under different terms, as a multifactorial and polygenic metabolic disorder, and its pathogenesis is influenced by diverse environmental and genetic risk factors [7]. Specifically, type 2 diabetes (T2D) is an expanding global health problem, closely linked to the epidemic of obesity [8].

Among the main effects of diabetes are long-term damage, dysfunction and failure of various organs, and vascular complications that shorten the life expectancy of those who suffer from this disease. For example, about 25% of persons with recent diagnosis have cardiovascular manifestations at the detection moment [9].

Therefore, there is an urgent need to implement population-based interventions that prevent diabetes and enhance its early detection [10].

There are several approaches to identify diabetes. For decades, the diagnosis of diabetes has been based on glucose criteria, either the FPG (Fasting Plasma Glucose) or the 2-h plasma glucose (2-h PG) value after a 75-g OGTT (Oral Glucose Tolerance Test) or A1C criteria [5,11].

Currently, algorithms based on computer-aided diagnosis (CADx) have been implemented for the diagnosis of diabetes through prediction models to know the future behavior of some data related to this disease.

Within these algorithms are found Artificial Neural Networks (ANN), which are based on mathematical models following the learning principle of artificial intelligence, as well as the natural response of the human neurons. ANNs are a nonlinear technique that integrates a set of variables through many data; for that reason, this procedure is useful for complex pattern recognition problems [12].

The implementation of ANN in this area is a tool that could be used by health services [13,14] because it gives information collected from diabetes cases and control cases (non-diabetic patients). Once the dataset has been analyzed, it is possible to improve the diabetes diagnosis, impacting in a positive way the health quality of the persons.

Carnimeo et al. [15] proposed the automatic detection of diabetic symptoms in retinal images by using a multilevel perceptron neural network. The network is trained using algorithms for evaluating the optimal global threshold that can minimize pixel classification errors. System performance is evaluated by an adequate index to provide a percentage measure in the detection of eye suspect regions based on neuro-fuzzy subsystem.

In addition, Cappon et al. [16] proposed the development of a tool based on ANN, optimizing and personalizing the calculation of the bolus calculation through continuous monitoring of glucose levels, obtaining useful and accessible information about patients that allows knowing if they have diabetes.

On the other hand, Chen et al. [17] proposed the 5G-Smart Diabetes system based on CADx, using different techniques of machine learning and big data to perform the analysis of patients suffering diabetes. In addition, the data sharing mechanism and personalized data analysis model for 5G-Smart Diabetes are presented in this work.

A continuing problem regarding diabetes despite its extensive study by different researchers is the difficulty of its early diagnosis and the identification of risks factors that may help to reduce the high incidence of this disease.

According to the above, in the present study, the objective was to analyze the relationship between anthropometric and biochemical features involved in the condition of diabetes. The main contribution of this work was to determine if a subject is presenting diabetes based in a set of specific features,

which were analyzed through an ANN, obtaining a classification system that allows the identification of diabetic patients from controls.

Therefore, the novel aspect of this work was the analysis of this type of data through a tool of artificial intelligence looking for the relationship between the features used that allows automatically classifying subjects with presence of diabetes from controls to support the diagnosis of these patients through an automatic tool.

The rest of the paper is structured as follows. Section 2 describes the materials required for the development of this work as well as the methodology proposed, which consists in three main steps: data acquisition, data classification and validation. Section 3 presents the results obtained for each of the stages proposed. In Section 4, the results are discussed. In Section 5, the final conclusions are drawn.

## 2. Materials and Methods

The process performed for the classification between patients with presence of diabetes and patients with absence of this disease is presented in this section, as well as the description of the data and the validation of the results.

The methodology followed in this work consists in three main steps: (A) data acquisition; (B) data classification; and (C) validation.

Data were acquired from the general hospital "Centro Médico Siglo XXI" with information from Mexican patients. All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of "Instituto Mexicano del Seguro Social" and "Comisión Nacional de Investigación Científica" (R-2011-785-018). The patient data were then classified according to diabetes status. Finally, The ANN's performance was evaluated, based on the objective of accurately classifying patients.

### 2.1. Subjects Description

The total number of patients contained in the dataset used for this work, is 1019, all Mexican, of which 499 correspond to non-diabetic patients (controls) and 520 to diabetic patients (cases). The age of the patients are between 35 and 65 years old, and 502 are female patients while 517 patients are male.

The criteria for the subjects to be part of the study were that the cases had less than five years of evolution and without other diseases. For the controls, the criterion was that they did not present any disease or metabolic syndrome. Pregnant women were excluded.

The diagnosis of the controls was made at the same time as the cases. The anthropometric and biochemical information was obtained at the same time the sample was taken and the sample was processed on the same day.

The glucose quantification was carried out for all participants in the overnight fasting of 12 h. The quantification was performed by the glucose oxidase method with the ILAB300 plus Instrument Laboratory, Bedford, MA, USA. The reference value was 70–100 mg/dL.

Finally, to ensure statistical significance, the sample was calculated with a 95% confidence level, with a 5% confidence interval over a estimated population size of 17,000,000 (estimated diabetic people in Mexico).

### 2.2. Features Description

In this work, 19 para-clinical features were analyzed, which are described in Table 1. They were used as input features to an ANN, while as output feature the health status based on the condition of diabetes of the patients was used.

**Table 1.** Features description.

| Feature | Description |
|---|---|
| Age | Patient age at the time of analysis. |
| Gender | Patient gender (0—male/1—female). |
| Education | Studies concluded by the patient, (1—elementary school/2—secondary school, 3—high school/4—bachelor's degree). |
| Weight | Patient weight in kilograms. |
| Height | Patient height in centimeters. |
| Waist | Patient waist perimeter in centimeters. |
| Hip Perimeter | Patient hip perimeter in centimeters. |
| BMI | Body Mass Index based on weight and height of a patient. |
| WHR | Waist Hip-Ratio based on the circumference of the waist to that of the hips. |
| SBP | Systolic Blood Pressure based on the pressure in the blood vessels when the heart beats. |
| DBP | Diastolic Blood Pressure based on the pressure in the blood vessels when the heart rests between beats. |
| Glucose | Blood glucose levels in terms of milligrams. |
| MMO Glucose | Blood glucose levels in terms of a molar concentration. |
| Insulin | Patient insulin in the blood. |
| HOMA | Homeostatic Model Assessment based on insulin resistance and beta-cell function. |
| Cholesterol | Fat-like substance that is found in all cells in the patient body. |
| LDL | Stands of low-density lipoprotein in the patient body. |
| HDL | Stands for high-density lipoprotein in the patient body. |
| TR | Triglycerides based on a type of fat (lipids) found in the patient body. |
| Output | Diabetes status (0—control/1—case). |

### 2.3. Data Analysis

In this work, a data analysis based on a multivariate approach was performed, looking for the classification of patients according with their diabetes status. The input features represented the input layer of a deep ANN. Afterwards, a statistical validation was carried out to evaluate the results.

### 2.4. Data Preprocessing

The dataset was composed of 19 features, which were normalized through the Z-core method. This method transforms the data to a normal distribution with mean 0 and standard deviation 1. Z-score was used with the purpose of adequate data for the classification, because usually the data are not defined in the same numeric scale.

Once the dataset was normalized, it was randomly divided and balanced into two sets:

- The first one was the training set, which corresponded to the training stage, involving 70% of all data.
- The second was the test set, which corresponded to the test stage, involving the remaining 30%.

#### 2.4.1. Data Classification

The patients were classified based in an ANN approach according to their condition; the control patients were labeled with "0", which are those who have not developed diabetes; and the case patients were labeled with "1", which are those who have developed diabetes. This step was performed using a dense ANN that was specifically designed for this dataset, using the packages Tensorflow and Keras, for Python.

Tensorflow is an open-source software, symbolic math library focused in the dataflow programming based on the ranging of tasks, which is widely applied in different machine learning applications, such as ANN [18]. Keras is a high-level ANN API written in Python. It was created to perform fast experimentation using deep ANN, and it presents three maim characteristics: user-friendly, modular, and extensible [19].

ANNs can have different layers, which are composed by nodes or neurons. In general terms, an ANN tries to find a model that best describes the relationship between the input features and the output feature. The number of layers is modifiable as the number of neurons in each layer [20].

ANNs have three main elements:

- As mentioned above, an ANN has a set of connections that are called weights. The weights are the elements that connect the input signal with a neuron through the calculation of their product.
- The activation function affects the neurons, limiting the amplitude of the output with a finite value.
- An element summarizes the contributions of a weighted signal.

There are many types of layers but in this work two different layers were applied: a dense layer, which consists of a matrix of weights created by the layer and a vector of values called bias; and a dropout layer, which consists of randomly establishing a fractional rate, the main aim of this type of layer being to avoid an overfitting problem.

The deep ANN designed for this work is shown in Figure 1, and the details of each layer are described below:

1. Input layer: 19 neurons (the data set has 19 features).
2. Dropout hidden layer: loss percentage of 25%.
3. Dense hidden layer: 100 neurons.
4. Dropout hidden layer: loss percentage of 50%.
5. Dense hidden layer: 500 neurons.
6. Dropout hidden layer: loss percentage of 25%.
7. Dense hidden layer: 100 neurons.
8. Dropout hidden layer: loss percentage of 50%.
9. Output layer: 2 neurons (control/diabetic).

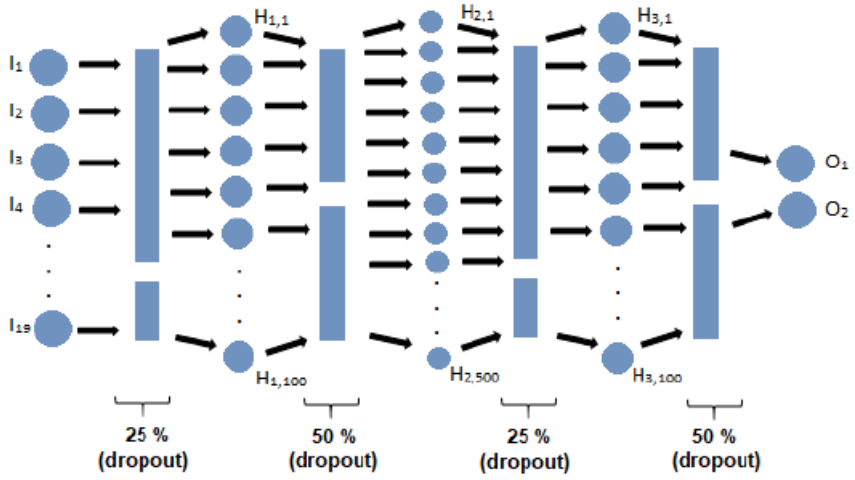


**Figure 1.** Graphic representation of the Artificial Neural Network (ANN) implemented.

For each dense layer, except the output dense layer, the activation function Rectified Linear Unit (ReLU) was added, which assigns 0 to the neurons that present a value lower than 0 and assigns the same value when this is equal or above 0. This function is calculated with Equation (1) [21].

$$ReLu(z) = \begin{cases} \text{if } z < 0 & 0 \\ \text{if } z \geq 0 & z \end{cases} \quad (1)$$

For the output layer, the activation function Softmax, also known as Normalized Exponential function, was added. Softmax is based on a general logistic function, compressing a vector of arbitrary values into a vector of values ranging [0, 1]. Equation (2) represents this function, where $\sigma(z)$ refers to the $K$-dimensional vector, $z$ [22].

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{K=1}^{K} e^{z_k}}, j = 1, \ldots, K \quad (2)$$

The optimization algorithm implemented was "Adam", which is based in the stochastic gradient descent algorithm, using the average of the first and second moments of the gradients, adapting the learning parameter. In other words, this algorithm calculates the exponential moving average of the gradient and the square gradient, controlling the decay of that moving average [23].

The epoch number is a configurable value based on different parameters. For this work, there were established 100 epochs according to the results obtained through many tests with different number of epochs, where the best accuracy was reached using 100 epochs.

#### 2.4.2. Evaluation

For the evaluation stage, the loss function and accuracy were calculated on each epoch, and the receiving operating characteristics (ROC) curve was obtained with the average of the general behavior of the deep ANN.

When the value of the loss function goes down, it indicates that the model is fitting better to the data it is trying to model, looking for the global minimum, which represents the minimum error. In addition, this function optimizes the network feeding back with information of the system capacity [24]. The algorithm chosen to calculate the loss function in this work was "binary cross-entropy", which is included in the Keras package for Python, and is able to calculate the cross-entropy parameter specifically in binary classification problems. This method is based on the Kullback–Leibler distance, which is calculated with Equation (3). This distance is a measure between two density functions $g$ and $h$. Cross-entropy is an iterative method that generates a set of random values that are updated looking to generate more approximate values [25].

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

The accuracy is the parameter that calculates the average performance of the ANN based on the difference between the classification calculated and the real classification, as shown in Equation (4), calculating the accuracy as 1-error, where $V_{pred}$ is the classification value calculated and $V_{true}$ is the true classification value. It does not optimize the network but it obtains this value for each of the models, giving the option to select the model that presents the better performance [26].

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

The accuracy function selected for this work was "binary-accuracy" function from the Keras package. This function obtains the average accuracy based in the total predictions and it is used in binary classification problems specifically.

The ROC curve is a parameter used to measure the classification precision of the model, through the sensitivity and specificity. Sensitivity refers to the proportion of subjects with a positive condition that were correctly classified and it is calculated with Equation (5), where $PPV$ is the positive predictive value, $TP$ are the true positives and $FP$ are the false positives [27].

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

Specificity refers to the proportion of subjects with a negative condition that were correctly classified and it is calculated with Equation (6), where $NPV$ is the negative predictive value, $TN$ are the true negatives and $FN$ are the false negatives [27].

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

For each class, the ROC curves were obtained, as well as the ROC curve of the macro-average and the micro-average precision. The micro-average precision refers to the sum of the total true positives, false positives and false negatives for different sets, and it is calculated with Equation (7), where $TP_1$ are the true positives of one set, $TP_2$ are the true positives of a second set, $FP_1$ are the false positives of the first set and $FP_2$ are the false positives of the second [19].

$$Micro - average = \frac{TP_1 + TP_2}{TP_1 + TP_2 + FP_1 + FP_2} \tag{7}$$

Finally, the macro-average precision is a value that calculates the average accuracy in different arbitrary sets and it is obtained with Equation (8), where $A_1$ is the average of one set and $A_2$ is the average of a second set.

$$Macro - average = \frac{A_1 + A_2}{2} \tag{8}$$

The implementation of this work was done with a laptop Toshiba Satellite S55T-B5233", Intel Core i7-7500 U 2.70 GHz, 16 GB, 500 GB SSD, Ubuntu 16.4, 64-bit; and with Python version 2.7 (Toshiba, Tokio, Japan) [28].

All analyses were performed in Python 2.7.12, using the packages, Keras 2.1.5, Scipy 1.0.1, Pandas 0.22.0, Sklearn 0.191 and Tensorflow 1.7.0.

## 3. Results

The dataset was initially divided in two subsets, one for training containing 70% of the data (349 controls/364 cases), and one for testing containing 30% of the data (150 controls/156 cases).

The dataset used for the training of the deep ANN was evaluated at each epoch with the accuracy and loss function parameters. The number of epochs was tested with different values, as shown in Table 2, looking for the number of epochs that present the best result.

**Table 2.** Accuracy and loss function values using different number of epochs.

| Epochs | Accuracy | Loss Function | Processing Time (s) |
|---|---|---|---|
| 10 | 0.93 | 0.21 | 1.59 |
| 50 | 0.94 | 0.20 | 5.12 |
| 100 | 0.96 | 0.21 | 9.35 |
| 150 | 0.94 | 0.23 | 13.96 |
| 200 | 0.93 | 0.25 | 18.23 |
| 300 | 0.93 | 0.30 | 27.18 |
| 500 | 0.94 | 0.29 | 44.12 |
| 1000 | 0.93 | 0.39 | 86.99 |

In addition, a comparison of the results using different type of layers, number of layers and number of neurons was made, which is shown in Table 3. The structure selected is presented in bold, which is characterized by a number of epochs of 100, in a deep ANN with nine layers, five dense layers and four dropout layers. The dense layers contained 19, 100, 500, 100 and 2 neurons, and the dropout layers presented 0.50, 0.25 and 0.50 percentages.

**Table 3.** Accuracy, loss function and processing time with different number of layers and neurons.

| Layers Dense/Dropout | Neurons | Accuracy | Loss Function | Processing Time (s) |
|---|---|---|---|---|
| 2/0 | 19 > 2 | 0.94 | 0.20 | 3.83 |
| 2/1 | 19 >0.5 >2 | 0.94 | 0.20 | 4.37 |
| 3/1 | 19 > 100 > 0.5 > 2 | 0.96 | 0.20 | 4.83 |
| 3/2 | 19 > 0.25 > 100 > 0.5 > 2 | 0.95 | 0.19 | 5.20 |
| 4/1 | 19 > 100 > 0.5 > 500 > 2 | 0.97 | 0.25 | 6.59 |
| 4/2 | 19 > 0.25 > 100 > 0.5 > 500 > 2 | 0.96 | 0.22 | 6.97 |
| 4/3 | 19 > 0.25 > 100 > 0.5 > 500 > 0.25 > 2 | 0.96 | 0.23 | 7.63 |
| 5/1 | 19 > 100 > 0.5 > 500 > 100 > 2 | 0.98 | 0.31 | 8.17 |
| 5/2 | 19 > 0.25 > 100 > 0.5 > 500 > 100 > 2 | 0.96 | 0.21 | 8.62 |
| 5/3 | 19 > 0.25 > 100 > 0.5 > 500 > 0.25 > 100 > 2 | 0.96 | 0.22 | 9.21 |
| **5/4** | **19 > 0.25 > 100 > 0.5 > 500 > 0.25 > 100 > 0.5 > 2** | **0.96** | **0.21** | **9.50** [1] |

[1] Structure selected for the Artificial Neural Network (ANN).

Figure 2 shows the behavior of the accuracy, where the blue line represents the training data, with an accuracy value of 0.96, and the orange line represents the testing data, with an accuracy value of 0.94.

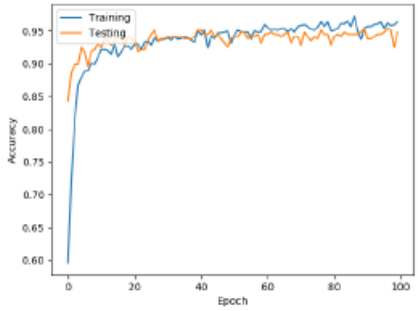


**Figure 2.** Accuracy behavior.

Figure 3 shows the behavior of the loss function, where the blue line represents the training data, with a loss function value of 0.11, and the orange line represents the testing data, with a loss function value of 0.23.

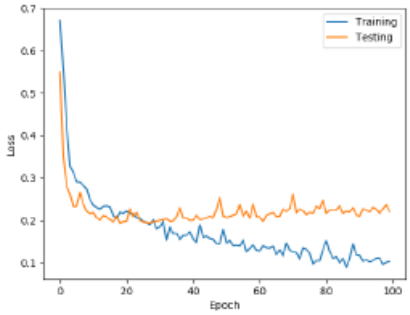


**Figure 3.** Loss function behavior.

Even when the accuracy and loss function values were obtained to measure the behavior of the classification of subjects, the accuracy is a parameter that may provide a better result than the real

accuracy bias is present in the data. Due to this situation, the area under the curve (AUC) value was calculated to check that the results obtained were statistically significant.

Figure 4 presents the ROC curves obtained based in the performance of the deep ANN. The ROC curve for the class 0 (control patients) is presented in pink, with an AUC value of 0.98. The ROC curve for the class 1 (case patients) is presented in light blue, with an AUC value of 0.98. The ROC curve calculated with the micro-average is presented in dotted orange, obtaining an AUC value of 0.98. Finally, the ROC curve calculated with the macro-average is presented in dotted dark blue, with an AUC value of 0.98.

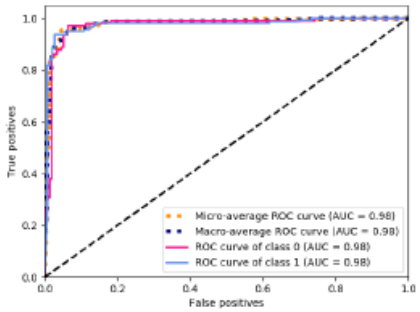


**Figure 4.** Receiving operating characteristics (ROC) curves obtained with the average performance of the ANN. AUC: area under the curve.

## 4. Discussion

The dataset used in this work was classified into two subsets through an aleatory and balanced selection. One subset was used for the training of the deep ANN, which consisted of 100 epochs. This number of epochs was selected based in a comparison of different values, as shown in Table 2. Then, the ANN was optimized with the Adam algorithm to improve the behavior of the ANN through feedback.

The other subset was used for testing the ANN behavior, and it validated the results of the training stage through the accuracy and loss function values.

Figures 2 and 3 show that the behaviors of the training data and testing data follow the same pattern, which indicates that the model obtained through the ANN became generalized with the learning process by being able to classify unknown data with the same performance as the known data.

It is important to remark that Figure 3 shows that the loss function is decreasing while the deep ANN is doing the training, as well as in the testing stage, which implies an approximation to the global minimum, while, in Figure 2, the behavior of the accuracy improves with the increase of the epochs for both datasets, indicating that both are improving their classification capacity in a similar proportion based on the learning.

The validation step allowed knowing that all ROC curves obtained, as presented in Figure 4, achieved statistically significant AUC values of around 0.98, which means that the ANN model presented a sensitivity—specificity rate that is able to classify the data with only 2% error.

Then, according to the results of the validation stage, the generalization of the model allowed significant values to be obtained for both classes, "0" and "1", and for both measures, micro-average and macro-average.

It is important to mention that obtaining similar results in the micro-average and the macro-average represents great stability and robustness in this analysis, since the micro-average is calculated through the average of subsets of data and the macro-average is calculated using

the complete dataset, implying that the data can be classified correctly regardless of how many are presented.

The positive results obtained are due to the features subjected to the ANN designed, as presented in Table 1, and the relationship between them. According to the literature, some of features are strongly related to diabetes disease and they can provide information to determine if a patient is prone to this condition. For example, the Body Mass Index (BMI), which is based on the weight and height, indicates the patient is prone to diabetes when it is $\geq$2.

In addition, compared to other works, this research presents higher values in the results of the obtained accuracy, as presented in Table 4.

**Table 4.** Accuracy presented in related works.

| Work | Description | Accuracy |
|---|---|---|
| Ndaba et al. [29] | Diabetes classification based on a regression ANN | 86.00% |
| Soltani et al. [30] | Diabetes diagnosis based on a probabilistic ANN | 89.56% |
| Sejdinović et al. [31] | Diabetes classification on an ANN | 93.90% |
| Chen et al. [17] | Diabetes classification model based on boosting algorithms | 95.30% |

On the other hand, one of the limitations of this work was related to the number of features used. It could be interesting to propose a dataset containing a greater number of features, looking for the improvement of the current results, to determine the main risk factors in the Mexican population.

In addition, it could be beneficial to create a device that implements the model developed to automate the diagnosis of diabetes and yield faster and more reliable results, supporting the initial diagnosis of the specialist.

It is important to mention that the implementation of the developed model does not need high computational cost, which is an advantage because it is not necessary to acquire any special hardware. Besides, for the development of this work, only free software was used, thus it does not imply any cost in licenses.

Finally, one of the points that most support the results is that the dataset contained the information of Mexican patients, which can help to improve the Mexican health through data-based tools developed with their own demography.

## 5. Conclusions

The results were validated through an evaluation step, being possible to conclude that the database was adequate for this work, demonstrating that is possible to classify persons with diabetes and persons who have not developed it, using the information of the 19 features previously mentioned.

The implementation of an ANN that was trained and tested with para-clinical data was to show the importance of having these types of data in the development of diabetes. Therefore, obtaining an accuracy of 0.96 allowed checking this hypothesis since the classification of cases from controls was correct 96% of the time.

The strong relationship between some features caused the good classification of persons according to the implemented model.The features used for this work represent the important risk factors to develop diabetes.

Because the dataset is based on Mexican people with presence and absence of diabetes, this work gives the advantage of knowing the important features to determinate this condition, and it is possible to create an auxiliary system that helps specialists to support their first diagnostic.

Based on these results, it is possible to conclude that these features present significant information for the classification of Mexican people with presence of diabetes from those with absence, which means that this information could be also used as a tool to support specialists for the preventive diagnosis and prediction of diabetes, helping to decrease the high incidence rate of this disease in Mexican population.

**Author Contributions:** Conceptualization, C.E.G.-T.; Data curation, V.A.-R., L.A.Z.-C. and C.E.G.-T.; Formal analysis, V.A.-R., L.A.Z.-C. and C.E.G.-T.; Investigation, A.V.-S.; Methodology, L.A.Z.-C., C.E.G.-T., J.I.G.-T. and H.G.-R.; Project administration, J.I.G.-T. and H.G.-R.; Resources, J.I.G.-T. and H.G.-R.; Software, V.A.-R.; Supervision, C.E.G.-T., A.G.-H. and M.C.; Validation, A.G.-H., M.C. and A.V.-S.; Visualization, L.A.Z.-C. and A.G.-H.; Writing—original draft, V.A.-R., L.A.Z.-C., C.E.G.-T., A.G.-H., M.C. and A.V.-S.; and Writing—review and editing, V.A.-R., C.E.G.-T., A.G.-H., M.C. and A.V.-S. All authors interpreted findings from the analysis and drafted the manuscript.

**Funding:** This research received no external funding.

**Acknowledgments:** This work was supported by the Fondo Sectorial de Investigación en Salud y Seguridad Social (SSA/IMSS/ISSSTECONACYT) project 150352

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

## References

1. WHO. *Global Status Report on Noncommunicable Diseases 2010*; World Health Organization: Geneva, Switzerland, 2011.
2. Etienne, G.K. Trends in diabetes: Sounding the alarm. *Lancet* **2016**, *387*, 1485–1486.
3. McCarty, D.J.; Zimmet, P. Diabetes 1994 to 2010: Global estimates and projection. In Proceedings of the Internacional Diabetes Institute, Kobe, Japan, 6–11 November 1994.
4. Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; da Rocha Fernandes, J.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* **2018**, *138*, 271–281. [CrossRef] [PubMed]
5. American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care* **2010**, *37* (Suppl. 1), S81–S90.
6. Turtle, J.R. What is diabetes mellitus? *Australas. Ann. Med.* **1969**, *18*, 59–73. [CrossRef] [PubMed]
7. Cruz, M.; Valladares-Salgado, A.; Garcia-Mena, J.; Ross, K.; Edwards, M.; Angeles-Martinez, J.; Ortega-Camarillo, C.; Escobedo de la Peña, J.; Burguete-Garcia, A.I.; Wacher-Rodarte, N.; et al. Candidate gene association study conditioning on individual ancestry in patients with type 2 diabetes and metabolic syndrome from Mexico City. *Diabetes/Metab. Res. Rev.* **2010**, *26*, 261–270. [CrossRef] [PubMed]
8. DeFronzo, R.A.; Ferrannini, E.; Groop, L.; Henry, R.R.; Herman, W.H.; Juul Holst, J.; Hu, F.B.; Kanh, C.R.; Raz, I.; Shulman, G.I.; et al. Type 2 diabetes mellitus. *Nat. Rev. Dis. Prim.* **2015**, *1*, 15019. [CrossRef] [PubMed]
9. Socarras, M.B.; M Blanco, J.; Vazquez, A.; Gonzáles, D.; Licea, M. Factores de riesgo de ateroesclerosis en la diabetes mellitus tipo 2. *Rev. Cub. Med.* **2003**, *42*, 17–25.
10. Majid, E. Worldwide trends in diabetes since 1980: A pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet* **2016**, *387*, 1513–1530.
11. International Expert Committee. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes Care* **2009**, *32*, 1327–1334. [CrossRef] [PubMed]
12. Gardner, G.; Keating, D.; Williamson, T.; Elliott, A. Automatic detection of diabetic retinopathy using an artificial neural network: A screening tool. *Br. J. Ophtalmol.* **1996**, *80*, 940–944. [CrossRef]
13. Chae, S.; Kwon, S.; Lee, D. Predicting infectious disease using deep learning and big data. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1596.10.3390/ijerph15081596. [CrossRef] [PubMed]
14. Irles, C.; González-Pérez, G.; Carrera Muiños, S.; Michel Macias, C.; Sánchez Gómez, C.; Martínez-Zepeda, A.; Cordero González, G.; Laresgoiti Servitje, E. Estimation of neonatal intestinal perforation associated with necrotizing enterocolitis by machine learning reveals new key factors. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2509.10.3390/ijerph15112509. [CrossRef] [PubMed]
15. Carnimeo, L.; Giaquinto, A. An intelligent system for Improving Detection of Diabetic Symptoms in Retinal Images. In Proceedings of the IEEE International Conference on Information Technology in Biomedicine, Larnaca, Cyprus, 5–7 November 2006.
16. Cappon, G.; Vettoretti, M.; Marturano, F.; Facchinetti, A.; Sparacino, G. A Neural-Network-Based approach to personalize insuline bolus calculating using continuous glucose monitoring. *SAGE J.* **2018**, *12*, 265–272.
17. Chen, M.; Yang, J.; Zhou, J.; Hao, Y.; Zhang, J.; Youn, C.H. 5G-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds. *IEEE Commun. Mag.* **2018**, *56*, 16–23. [CrossRef]
18. Google. Tensorflow. Available online: https://www.tensorflow.org/ (accessed on 15 June 2018).

19. Chollet, F. Keras: Deep Learning Library for Theano and Tensorflow. Available online : https://keras.io (accessed on 15 June 2018).
20. Agatonovic-Kustrin, S.; Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed Anal.* **2000**, *22*, 717–727. [CrossRef]
21. Lomuscui, A.; Maganti, L. An approach to reachability analysis for feed-forward relu neural networks. *arXiv* **2017**, arXiv:1706.07351.
22. Carlini, N.; Wanger, D. Towards evaluating the robustness of neural network. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017.
23. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980. Available online: https://arxiv.org/pdf/1412.6980.pdf (accessed on 29 December 2018).
24. Antona Cortés, C.Herramientas Modernas en Redes Neuronales: La Librería Keras. Bachelor's Thesis, Universidad Autónoma de Madrid, Madrid, Spain, 2017.
25. Kullback, S.; Leibler, R. On information and sufficiency. *Anals Math. Stat.* **1951**, *22*, 76–86. [CrossRef]
26. Nye, M.; Saxe, A. Are efficient deep representations learnable? In Proceedings of the International Conference on Learning Representations ICLR 2018 Workshop, Vancouver, BC, Canada, 30 April–3 May 2018.
27. 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]
28. Community, P. What Is Python? Available online: https://www.python.org/doc/essays/blurb/ (accessed on 1 September 2018).
29. Ndaba, M.; Pillay, A.W.; Ezugwu, A.E. An improved generalized regression neural network for type II diabetes classification. In Proceedings of the International Conference on Computational Science and Its Applications, Melbourne, VIC, Australia, 2–5 May 2018; Springer: Cham, Switzerland, 2018; pp. 659–671.
30. Soltani, Z.; Jafarian, A. A new artificial neural networks approach for diagnosing diabetes disease type II. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 89–94. [CrossRef]
31. Sejdinović, D.; Gurbeta, L.; Badnjević, A.; Malenica, M.; Dujić, T.; Čaušević, A.; Bego, T.; Mehmedović, L.D. Classification of prediabetes and type 2 Diabetes using Artificial Neural Network. In *CMBEBIH 2017*; Springer: Singapore, 2017; pp. 685–689.


# Referencias


[1] OPS, "Enfermedades no transmisibles." Avaiable online: https://www.paho.org. accessed on: April 21, 2019.

[2] WHO, "Noncommunicable diseases." Avaiable online: http://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. accessed on: September 20, 2018.

[3] dirección nacional de promoción de la salud y control de enfermedades no transmisibles, "¿qué son las enfermedades no transmisibles?." Avaiable online: http://www.msal.gob.ar/ent/index.php/informacion-para-ciudadanos/ique-son-icuales-son. accessed on: April 16, 2019.

[4] WHO, "Datos y cifras de enfermedades no transmisibles." Avaiable online: https://www.who.int/features/factfiles/noncommunicable_diseases/facts/es/index9.html. accessed on: April 21, 2019.

[5] WHO, "Oral health," accessed on May 9, 2019.

[6] IDF, "Type 2 diabetes." Avaiable online: https://idf.org/52-about-diabetes.html. accessed on: May 6, 2019.

[7] A. D. Association *et al.*, "2. classification and diagnosis of diabetes: standards of medical care in diabetes—2018," *Diabetes care*, vol. 41, no. Supplement 1, pp. S13–S27, 2018.

[8] T. L and M. P. S., "Diagnóstico clínico y tratamiento," p. 324, 2000.

[9] R. Calderón, "Observaciones sobre diabetes mellitus al final del milenio.," pp. 6–9, 2000.

[10] R. DeFronzo, "Pathogenesis of type 2 diabetes mellitus," pp. 787–835, 2004.

[11] E. DonnellyR, A. Garder, and I. Morris A., "Abc of vascular disease: Vascular complications of diabetes," pp. 1062–1066, 2000.

[12] S. Twetman, "Prevention of dental caries as a non-communicable disease," *European journal of oral sciences*, vol. 126, pp. 19–25, 2018.

[13] L. A. Zanella-Calzada, C. E. Galván-Tejada, N. M. Chávez-Lamas, J. I. Galván-Tejada, and J. M. Celaya-Padilla, "Multivariate features selection from demographic and dietary descrptors as caries risk determinants in oral health diagnosis: Data from nhanes 2013-2014.," *Electronics, Communications and Computers (CONIELECOMP)*, pp. 217–222, 2018.

[14] J. Ramírez Mendoza, M. A. Rueda Ventura, M. H. Morales García, and A. Gallegos Ramírez, "Prevalencia de caries dental y maloclusiones en escolares de tabasco, méxico," vol. 11, no. 1, pp. 13–23, 2012.

[15] WHO, "Sugars and dental caries," 2017.

[16] WHO, "Noncommunicable diseases." Avaiable online: http://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. accessed on: November 15, 2018.

[17] G. B. D. disease, injury incidence, and prevalence collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the global burden of disease study 2016," *Lancet*, vol. 390, pp. 1211–1259, 2017.

[18] N. Kassebaum, E. Bernabé, M. Dahiya, B. Bhandari, C. Murray, and W. Marcenes, "Global burden of untreated caries: a systematic review and metaregression," *Journal of dental research*, vol. 94, no. 5, pp. 650–658, 2015.

[19] WHO, "Global status report on noncommunicable diseases 2010," *Geneva*, 2010.

[20] G. Etienne, "Trends in diabetes: sounding the alarm.," vol. 387, pp. 1485–1486, 2016.

[21] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes, A. Ohlrogge, and B. Malanda, "Idf diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045," *Diabetes research and clinical practice*, vol. 138, pp. 271–281, 2018.

[22] M. Cruz, A. Valladares-Salgado, J. Garcia-Mena, K. Ross, M. Edwards, J. Angeles-Martinez, C. Ortega-Camarillo, J. Escobedo de la Peña, A. I. Burguete-Garcia, N. Wacher-Rodarte, R. Ambriz, R. Rivera, A. L. D'artote, J. Peralta, E. J. Parra, and J. Kumate, "Candidate gene association study conditioning on individual ancestry in patients with type 2 diabetes and metabolic syndrome from mexico city," vol. 26, pp. 261–270, 2010.

[23] C. Grosso, "Síndrome metabólico y riesgo vascular," 2013.

[24] E. Gispert Abreu, P. Castell-Florit Serrate, and M. Herrera Nordet, "Salud bucal poblacional y su producción intersectorial.," *Revista Cubana de Estomatología*, vol. 52, pp. 62–67, 2015.

[25] P. Singh, A. Pal, M. Anburajan, and J. Kumar, "Computer-aided diagnosis of diabetes mellitus using thermogram of open mouth," *Computing and Intelligent Engineering*, vol. 52, pp. 62–67, 2018.

[26] L. Carnimeo and A. Giaquinto, "An intelligent system for improving detection of diabetic symptoms in retinal images.," *IEEE International Conference on Information Technology in Biomedicine.*, 2006.

[27] M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang, and C. H. Youn, "5g-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds.," *IEEE Communications Magazine*, vol. 56, pp. 16–23, 2018.

[28] G. Cappon, M. Vettoretti, F. Marturano, A. Facchinetti, and G. Sparacino, "A neural-network-based approach to personalize insuline bolus calculating using continuous glucose monitoring.," *SAGE journals*, vol. 12, pp. 265–272, 2018.

[29] T. S. Ghazal, S. M. Levy, N. K. Childrens, K. D. Carter, J. J. W. Caplan, and J. L. Kilker, "Survival analysis of caries incidence in africanamerican school-aged children.," *Journal of public health dentistry*, 2018.

[30] J.-H. Lee, D.-H. Kim, S.-N. Jeong, and S.-H. Choi, "Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm," *Journal of dentistry*, vol. 77, pp. 106–111, 2018.

[31] Á. S. Brito, M. A. Clementino, M. C. Gomes, É. T. B. Neves, A. de Sousa Barbosa, C. A. de Medeiros, M. M. de Aquino, A. F. Granville-Garcia, V. A. de Menezes, *et al.*, "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*, vol. 36, no. 3, p. 244, 2018.

[32] S. Lai, M. G. Cagetti, F. Cocco, D. Cossellu, G. Meloni, G. Campus, and P. Lingström, "Evaluation of the difference in caries experience in diabetic and non-diabetic children—a case control study," *PloS one*, vol. 12, no. 11, p. e0188451, 2017.

[33] Barylo, Kanishyna, and L. Shkilniak, "The effects of diabetes mellitus on patients' oral health.," *Wiadomosci lekarskie (Warsaw, Poland: 1960)*, vol. 71, no. 5, pp. 1026–1031, 2018.

[34] B. R. Latti, J. V. Kalburge, S. B. Birajdar, and R. G. Latti, "Evaluation of relationship between dental caries, diabetes mellitus and oral microbiota in diabetics," *Journal of oral and maxillofacial pathology: JOMFP*, vol. 22, no. 2, p. 282, 2018.

[35] L. Ferizi, F. Dragidella, L. Spahiu, A. Begzati, and V. Kotori, "The influence of type 1 diabetes mellitus on dental caries and salivary composition," *International Journal of Dentistry*, vol. 2018, 2018.

[36] R. C. M. Pinho, R. V. de Sousa, B. d. C. F. Vajgel, S. M. Paiva, and R. Cimões, "Evaluation of oral health-related quality of life in individuals with type 2 diabetes mellitus," *Brazilian Journal of Oral Sciences*, vol. 18, 2019.

[37] W. Fu, C. Lv, L. Zou, F. Song, X. Zeng, C. Wang, S. Yan, Y. Gan, F. Chen, Z. Lu, *et al.*, "Meta-analysis on the association between the frequency of tooth brushing and diabetes mellitus risk," *Diabetes/metabolism research and reviews*, p. e3141, 2019.

[38] world health organization and united nations development Programme., "Noncommunicable diseases: what ministries of agriculture need to know.," 2018. Accessed: Mayo, 2019.

[39] expert commitee on the diagnosis and classification of diabetes mellitus., "Report of the expert commitee on the diagnosis and classification of diabetes mellitus.," 2003.

[40] L. K. Billings and J. C. Florez, "The genetics of type 2 diabetes: what have we learned from gwas?," *Annals of the New York Academy of Sciences*, vol. 1212, no. 1, pp. 59–77, 2010.

[41] P. Zimmet, "The burden of type 2 diabetes: are we doing enough?," *Diabetes and Metabolism*, vol. 29, no. 4, Part 2, pp. 6S9 – 6S18, 2003. Cardiovascular benefits of metformin.

[42] J. A. Rull, C. A. Aguilar-Salinas, R. Rojas, J. M. Rios-Torres, F. J. Gómez-Pérez, and G. Olaiz, "Epidemiology of type 2 diabetes in mexico," *Archives of Medical Research*, vol. 36, no. 3, pp. 188–196, 2005.

[43] K. J. Chen, S. S. Gao, D. Duangthip, S. K. Y. Li, E. C. M. Lo, and C. H. Chu, "Dental caries status and its associated factors among 5-year-old hong kong children: a cross-sectional study," *BMC oral health*, vol. 17, no. 1, p. 121, 2017.

[44] R. H. Selwitz, A. I. Ismail, and N. B. Pitts, "Dental caries," *The Lancet*, vol. 369, no. 9555, pp. 51–59, 2007.

[45] S. de salud, "Norma oficial mexicana nom-015-ssa2-2010, para la prevención, tratamiento y control de la diabetes mellitus." Avaiable online: http://www.salud.gob.mx/unidades/cdi/nom/m015ssa24.html. accessed on: April 21, 2019.

[46] S. M. Perner, "Respuesta a la carta "la evidencia publicada y las transformaciones en el abordaje de la diabetes".." Avaiable online: http://www.scielo.org.ar/scielo.php?script=sci_arttextpid=S1851-82652014000200012. accessed on: February 9, 2019.

[47] K. Plonosky, "The past 200 years in diabetes.," *N Engl J Med*, vol. 367, pp. 1332–1340, 2012.

[48] J. D. de Estrada Riverón, J. A. P. Quiñonez, and I. H.-G. Fuentes, "Caries dental y ecología bucal, aspectos importantes a considerar," *Rev Cubana Estomatol*, vol. 43, no. 1, pp. 47–55, 2019.

[49] H. Heymann, J. Swift, and A. Ritter, ". sturdevanant's art and science of operative dentistry.," *Mosby ELsevier*, 2006.

[50] M. Najib, N. M. Ali, M. M. Arip, M. A. Jalil, and M. Taib, "Classification of agarwood using ann," *International Journal of Electrical and Electronic Systems Research*, vol. 5, pp. 20–34, 2012.

[51] M. R. Kumar, B. Sethi, and S. Bhutia, "Review on classification of digital images using artificial neural network," *International Journal of Engineering Science Invention*, 2017.

[52] F. Farfán, *Control Cerebral de Interfases: Análisis Exploratorio de Técnicas Paramétricas Digitales para la Detección y Cuantificación de Estados Mentales*. PhD thesis, 01 2005.

[53] M. Quiñones, L. Feller, C. Tarazona, and G. Teodosio, "Aplicación de redes neuronales artificiales sobre la violencia de la mujer por su pareja según la encuesta demográfica y de salud familiar, endes 2016," 2018.

[54] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[55] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ann) modeling and its application in pharmaceutical research," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 22, pp. 717–727, 2000.

[56] J. F. y D.M. Skapura, "Redes neuronales: Algoritmos, aplicaciones y técnicas de programación.," 1993.

[57] B. M. y Alfredo Sanz, "Redes neuronales y sistemas difusos," 2002.

[58] F. A. y Skapura M., "Neural networks: Algorithms, applications, and programming techniques.," 1991.

[59] C. Antona Cortés, "Herramientas modernas en redes neuronales: la librería keras.," 2017.

[60] S. Kullback and R. Leibler, "On information and sufficiency, anals of mathematical statistics," pp. 76–86, 1951.

[61] M. Nye and A. Saxe, "Are efficient deep representations learnable?," pp. 1–4, 2017.

[62] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve," *Radiology*, 1982.

[63] F. Chollet, “Keras: Deep learning libary for theano and tensofrflow. avaiable on: https://keras. io/k.,” 2011.

[64] P. Community, “What is python?. available on: https://www.python.org/doc/essays/blurb/ (accessed on september 2018),”

[65] F. Chollet, “Keras: Deep learning library for theano and tensorflow. available on: https://keras. io/k (accessed on june 2018),” (accessed on September 2018).

[66] Google, “Tensorflow. available on: https://www.tensorflow.org/ (accessed on september 2018),”

[67] S. C. Izaurieta Fernando, “Redes neuronales artificiales,”

[68] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014.