

UNIVERSIDAD AUTÓNOMA DE ZACATECAS
“Francisco García Salinas”



**“Biomarcador para la Detección de
Enfermedades Cardiacas”**

Tesis para obtener el grado de:
Maestro en Ciencias del Procesamiento de la Información

Presenta:

ITIC. Isamar Aparicio Montelongo

Director:

Dr. José María Celaya Padilla

Co-Directores:

Dr. Huizilopoztli Luna García

Dr. Hamurabi Gamboa Rosales

Asesores:

Dr. Juan Rubén Delgado Contreras

Dr. José Guadalupe Arceo Olague

Zacatecas, Zac., noviembre de 2023



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL



Zacatecas, Zac., 02 de Octubre de 2023.


C. Isamar Aparicio Montelongo
Estudiante de la MCPI
PRESENTE

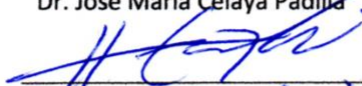
At'n: Dr. Huizilopoztli Luna García
Responsable de la MCPI

Nos es grato comunicarle que después de haber sometido a revisión académica la propuesta de Tesis titulada **"Biomarcador para la Detección de Enfermedades Cardiacas"**, presentada por el estudiante Ing. Isamar Aparicio Montelongo y habiendo efectuado todas las correcciones indicadas por este Comité Tutorial, se **AUTORIZA** el documento de tesis para su impresión.

Sin más por el momento reciban un cordial saludo.

COMITÉ TUTORIAL
PROCESAMIENTO Y ANÁLISIS DE DATOS





Dr. José María Celaya Padilla


Dr. Hamurabi Gamboa
Rosales



Dr. Juan Rubén Delgado
Contreras



Dr. Huizilopoztli Luna García


Dr. José Guadalupe Arceo
Olague

c.c.p. *Interesado.*

c.c.p. *Responsable de la Maestría en Ciencias del Procesamiento de la Información.*



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL

**COORDINACIÓN DE
INVESTIGACIÓN Y POSGRADO**

Carta de similitud núm.512/ IyP
Zacatecas, Zacatecas 22/noviembre/2023

Dr. Huizilopoztli Luna García
Responsable de la MCPI – UAZ
Presente

Estimado Dr. Huizilopoztli,

Después de saludarlo, sirva el presente oficio para notificar que el documento

*"Biomarcador para la detección de enfermedades
cardiacas"
de Isamar Aparicio Montelongo*

Fue analizado con el software Copyleaks, con la intención de detectar similitudes; el resultado en cuestión fue

22.5 % de similitud

De acuerdo a lo anterior, el porcentaje se considera **ACEPTABLE** de acuerdo a los estándares internacionales.

Atentamente
"Somos Arte, Ciencia y Desarrollo Cultural"

Dr. Carlos Francisco Bautista Capetillo
Coordinador de Investigación y Posgrado
Universidad Autónoma de Zacatecas



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL



Zacatecas, Zac., 02 de Octubre de 2023.
Carta de cesión de derechos

A QUIEN CORRESPONDA

El suscrito, C. **Isamar Aparicio Montelongo**, alumna del Programa de **Maestría en Ciencias del Procesamiento de la Información**, con número de matrícula **20204469** y adscrito a la Unidad Académica de Ingeniería de la Universidad Autónoma de Zacatecas, desea informar que soy la autora intelectual del trabajo de Tesis titulado "**Biomarcador para la Detección de Enfermedades Cardiacas**". Bajo la dirección del **Dr. José María Celaya Padilla**, cedo los derechos de este trabajo a la **Universidad Autónoma de Zacatecas** con el propósito de su difusión para fines académicos e investigativos.

Se solicita a los usuarios de esta información que se abstengan de reproducir el contenido textual, gráficos o datos del trabajo sin la autorización expresa del autor y/o los directores de la investigación. Para obtener dicho permiso, pueden contactarme a través del correo electrónico isa.aparicio186@gmail.com o comunicarse con el responsable del Programa de Maestría, quien derivará la solicitud a los directores del trabajo de investigación. Si se concede el permiso, se espera que el usuario muestre la debida gratitud y cite la fuente apropiadamente.

Agradezco de antemano su atención a esta solicitud y quedo a su disposición para cualquier consulta adicional. Reciba un cordial saludo.

ATENTAMENTE

Isamar Aparicio Montelongo



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL



CONAHCYT
CONSEJO NACIONAL DE HUMANIDADES
CIENCIAS Y TECNOLOGÍAS

AGRADECIMIENTO ESPECIAL.

Agradezco al Programa Nacional de Posgrados de Calidad del Consejo Nacional de Humanidades, Ciencias y Tecnología (CONAHCYT) por el invaluable respaldo financiero durante mi Maestría en Ciencias del Procesamiento de la Información (2021-2023). Este apoyo ha sido fundamental para mi formación académica y contribuirá significativamente a mi desarrollo profesional. ¡Gracias por hacer posible este logro!

Agradecimientos

Quiero expresar mi más sincero agradecimiento a mis padres, pilares fundamentales en mi trayectoria, cuyo apoyo incondicional ha sido la base de todos mis logros. Sin ustedes, este camino no habría sido posible, jamás me alcanzaré la vida para pagarles todo lo que han hecho por mí. mis hermanos, agradezco su constante motivación que ha sido mi impulso para seguir adelante. A mis sobrinos, quienes con su alegría y energía han llenado mi vida de momentos especiales, gracias por ser la luz que ilumina mi camino.

A mis asesores y maestros, Dr. José María Celaya Padilla, Dr. Huizilopoztli Luna García, Dr. José Guadalupe Arceo Olague y Dr. Juan Rubén Delgado Contreras, les agradezco por sus valiosas enseñanzas a lo largo de la maestría, que han enriquecido mi conocimiento y perspectiva profesional.

A mis amigas Flor Carolina, Cenyoalli, Elizabeth Aranda, Milagros Sandoval, amigos Man Kit y Eduardo González quienes, con sus largas conversaciones y salidas para desestresarnos, han sido parte de mi apoyo emocional, ayudándome a enfrentar este nuevo reto en mi carrera.

Y a ti, Mafe, mi más profundo agradecimiento. Gracias por escucharme y estar siempre que lo necesito, por ser mi brazo derecho en todo. Tus palabras de aliento, consejos, tiempo dedicado y tus chistes han sido mi refugio en momentos de estrés. Gracias por tanto amor.

Gracias infinitas a cada uno de ustedes por ser parte importante de mi éxito y felicidad.

Dedicatoria

A mis padres, quienes han sido mi fuente inagotable de apoyo, inspiración y sacrificio, siendo su amor incondicional y sabios consejos mi faro en este viaje académico. Expreso mi gratitud a mis amigos y seres queridos, cuya paciencia y aliento han sido fundamentales en los momentos desafiantes. Además, dedico este trabajo a mi amada familia, en especial a mi esposa, mi roca y motivación incondicional. Su paciencia y comprensión han sido mi faro en días oscuros, y su amor impulsa cada logro. También agradezco a mis leales compañeros perrunos, cuya presencia alegre ilumina mis jornadas de estudio. A mi familia, gracias por creer en mí y por compartir este viaje. Cada página lleva la huella de su amor y aliento. Este logro es tan suyo como mío. A través de cada desafío, su apoyo ha sido mi mayor fortaleza. Finalmente, dedico esta tesis a mi propia determinación y perseverancia, recordándome que cada desafío superado es un paso más hacia el crecimiento personal y profesional.

Resumen

Las enfermedades cardíacas (EC) son trastornos que afectan el funcionamiento del corazón, siendo la principal causa de morbilidad y muerte ya que anualmente se cobran más de 17,3 millones de vidas y se estima que para 2030 esta cantidad ascienda a 23,6 millones de fallecimientos a nivel mundial [2]. Durante el primer semestre de 2021 en México hubo 113,899 decesos por Enfermedades del corazón [4]. Si no se detectan a tiempo, pueden provocar complicaciones graves como insuficiencia cardíaca, infarto de miocardio, accidente cerebrovascular, etcétera. La detección temprana es esencial para iniciar tratamientos adecuados y prevenir consecuencias devastadoras. La comunidad médica utiliza características clínicas como resultados de electrocardiograma, análisis de sangre, ecocardiograma y pruebas de esfuerzo para detectar EC. La inteligencia artificial en el campo de la medicina ha tenido un impacto significativo en diversas áreas como predicción, prevención, personalización de tratamiento, descubrimiento de medicamentos, entre otras, por lo tanto, el objetivo de esta investigación es desarrollar un biomarcador mediante técnicas de aprendizaje automático para predecir estas afecciones. Analizando características antropométricas, hábitos alimenticios, aspectos de salud y otras enfermedades secundarias, como la diabetes, se exploró un conjunto de 253,680 observaciones con 19 características. Se utilizaron algoritmos genéticos para identificar factores de riesgo, así como regresión logística, random forest y KNN para clasificar a los pacientes en ocho grupos distintos, considerando la presencia de diabetes, rangos de edad y género. Con una precisión superior al 80% en diversos grupos, se concluye que la implementación de herramientas de inteligencia artificial y la consideración de información general de pacientes, tanto sanos como enfermos, pueden contribuir a prever enfermedades cardíacas. Esto permite un diagnóstico oportuno adaptado a las necesidades individuales de cada paciente y contribuye a la prevención de futuras complicaciones cardiovasculares.

Abstract

Cardiac diseases (CD) are disorders that affect the functioning of the heart, being the leading cause of morbidity and death, claiming over 17.3 million lives annually. It is estimated that by 2030, this number will rise to 23.6 million deaths worldwide [2]. During the first half of 2021, there were 113,899 deaths due to heart diseases in Mexico [4]. If not detected in time, they can lead to serious complications such as heart failure, myocardial infarction, stroke, etc. Early detection is essential to initiate appropriate treatments and prevent devastating consequences. The medical community uses clinical features such as electrocardiogram results, blood analysis, echocardiograms, and stress tests to detect CD. Artificial intelligence in the field of medicine has had a significant impact in various areas such as prediction, prevention, treatment personalization, drug discovery, among others. Therefore, the objective of this research is to develop a biomarker using machine learning techniques to predict these conditions. Analyzing anthropometric characteristics, dietary habits, health aspects, and other secondary diseases, such as diabetes, a dataset of 253,680 observations with 19 features was explored. Genetic algorithms were used to identify risk factors, and logistic regression, random forest, and KNN were employed to classify patients into eight different groups, considering the presence of diabetes, age ranges, and gender. With an accuracy exceeding 80% in various groups, it is concluded that the implementation of artificial intelligence tools and the consideration of general patient information, both healthy and sick, can contribute to predicting cardiac diseases. This allows for a timely diagnosis tailored to the individual needs of each patient and contributes to the prevention of future cardiovascular complications.

Contenido General

Agradecimientos	ii
Dedicatoria	iii
Resumen	iv
Abstract	v
Índice de Figuras	viii
Índice de Tablas	9
Capítulo 1. Introducción	1
1.1. Antecedentes	2
1.2. Planteamiento del problema de investigación	2
1.3. Justificación del problema de investigación	4
1.4. Preguntas de investigación	4
1.5. Objetivo general	5
1.5.1. Objetivos específicos	5
1.6. Hipótesis	5
1.7. Trabajo a realizar	5
1.8. Estructura de la tesis	6
Capítulo 2. Marco Teórico	7
2.1. Enfermedades Cardiacas	7
2.1.1. Factores de riesgo cardiovascular	7
2.2. Diabetes	8
2.2.1. Complicaciones comunes de diabetes	9
2.3. Prevención	10
2.4. Biomarcador	11
2.5. Inteligencia Artificial (IA)	11
2.5.1. Técnicas de IA	11
2.5.2. Machine Learning (ML)	12
2.6. Algoritmos	14
2.6.1. Regresión logística	15

2.6.2. Random forest	16
2.6.3. K vecinos más cercanos (KNN).....	17
2.6.4. Métricas de desempeño y validación cruzada (CV).....	18
2.6.5. Selección de características.....	20
2.7. Principales estudios relacionados	25
2.8. Contribuciones y limitaciones de estudios previos.....	26
2.9. Comparación entre los trabajos relacionados y la propuesta de investigación.....	29
2.10. Propuesta de investigación	30
2.11. Modelo o esquema general de investigación	30
Capítulo 3. Método y propuesta de investigación.....	31
3.1. Modelo de investigación.....	31
3.2. Descripción de la propuesta.....	31
3.2.1. Adquisición de datos.....	31
3.2.2. Preprocesamiento de datos.....	32
3.2.3. Selección de características.....	34
3.2.4. Obtención del modelo de predicción	34
3.2.5. Evaluación del modelo.....	35
Capítulo 4. Resultados	36
4.1. Caso de estudio 1: Diabetes.....	36
4.1.1. Adquisición de datos.....	36
4.1.2. Preprocesamiento de datos.....	37
4.1.3. Selección de características.....	37
4.1.4. Obtención del modelo de predicción	39
4.1.5. Evaluación del modelo.....	39
4.2. Caso de estudio 2: Grupos de edad.....	40
4.2.1. Adquisición de datos.....	40
4.2.2. Preprocesamiento de datos.....	40
4.2.3. Selección de características.....	40
4.2.4. Obtención del modelo de predicción	42
4.2.5. Evaluación del modelo.....	42
4.3. Caso de estudio 3: Sexo biológico.....	43

4.3.1. Adquisición de datos.....	43
4.3.2. Preprocesamiento de datos.....	44
4.3.3. Selección de características.....	44
4.3.4. Obtención del modelo de predicción	45
4.3.5. Evaluación del modelo.....	45
Capítulo 5. Conclusiones.....	47
5.1. Objetivos alcanzados	48
5.2. Hipótesis/proposiciones demostradas.....	48
5.3. Contribuciones de la investigación.....	49
5.4. Trabajos publicados.....	49
Referencias.....	52
Anexos.....	57

Índice de Figuras

Figura 1.1 Defunciones registradas en Zacatecas de enero-junio 2021 [4].	1
Figura 1.2 Número de fallecimientos registrados en Zacatecas en 2020, según las principales causas de mortalidad [13]......	3
Figura 2.1 Etapas genéricas para llevar a cabo un proyecto de Machine Learning [52]......	12
Figura 2.2 Ejemplo de clasificación de aprendizaje supervisado [54].	13
Figura 2.3 Ejemplo de clasificación de aprendizaje no supervisado [54].	14
Figura 2.4 Función sigmoide para regresión logística [57]......	16
Figura 2.5 Algoritmo de Random Forest [58]......	17
Figura 2.6 Representación de posibles curvas ROC [64]......	19
Figura 2.7 Ejemplo de área bajo la curva [60].	19
Figura 2.8 Validación cruzada [65]......	20
Figura 2.9 Intercambio de información genética entre dos individuos[66].	21
Figura 2.10 Una mutación modifica al azar parte del cromosoma de los individuos [66]......	22
Figura 2.11 Ciclo del algoritmo genético: Selección (Se) → Cruzamiento (Cr) → Mutación (Mu) → Evaluación (f(x)) → Reemplazo (Re). El símbolo ? es la condición de término y x* es la mejor solución [66]......	22
Figura 2.12 Representación esquemática del procedimiento de AG [67]......	24
Figura 2.13 Proceso General de Investigación (elaboración propia) [68]......	31
Figura 3.1 Modelo de investigación implementado (elaboración propia)......	31
Figura 3.2 Estadística de pacientes Diabéticos (elaboración propia).	33
Figura 4.1 Frecuencia de aparición de las características más importantes obtenidas por Galgo en el grupo de Diabéticos (elaboración propia).	38
Figura 4.2 Selección hacia adelante en el grupo de Diabéticos (elaboración propia).	38
Figura 4.3 Comparativa de modelos de predicción con RLO, RF y KNN en el grupo de Diabéticos (elaboración propia).	39
Figura 4.4 Promedios de métricas de desempeño, obtenidas en grupos de diabetes del modelo de	

RLO (elaboración propia).	40
Figura 4.5 Frecuencia de aparición de las características más importantes obtenidas por Galgo en el grupo de Adultos (elaboración propia).	41
Figura 4.6 Selección hacia adelante en el grupo de Adultos (elaboración propia).	42
Figura 4.7 Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de Adultos (elaboración propia).	43
Figura 4.8 Promedios de métricas de desempeño, obtenidas en grupos de Edad del modelo de random forest (elaboración propia).	43
Figura 4.9 Frecuencia de aparición de las características más importantes obtenidas por Galgo en el grupo de Hombres (elaboración propia).	44
Figura 4.10 Proceso de selección hacia adelante en el grupo de Hombres (elaboración propia).	45
Figura 4.11 Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de Hombres (elaboración propia).	46
Figura 4.12 Promedios de métricas de desempeño, obtenidas en grupos de Sexo del modelo de random forest (elaboración propia).	46
Figura 5.1 Artículo publicado en CIICTA 2021 [72].	50
Figura 5.2 Artículo publicado en COMIA, 2022 [73].	51

Índice de Tablas

Tabla 2.1 Matriz de confusión [61].	18
Tabla 2.2 Principales estudios relacionados (elaboración propia).	26
Tabla 2.3 Comparación de trabajos relacionados y la propuesta de investigación (elaboración propia).	29
Tabla 3.1 Características del conjunto de datos de indicadores de salud de Enfermedades Cardíacas (elaboración propia).	31
Tabla 4.1 Características utilizadas del conjunto de datos original (elaboración propia).	36
Tabla 4.2 Características más significativas y porcentaje de desempeño en grupos de diabetes (elaboración propia).	37
Tabla 4.3 Características más significativas en grupos de Edad (elaboración propia).	41
Tabla 4.4 Características más significativas en grupos de Sexo (elaboración propia).	44

Capítulo 1. Introducción

De acuerdo con la Organización Mundial de la Salud (OMS) las Enfermedades Cardiacas (EC) son la principal causa de muerte a nivel mundial, constituyen un conjunto de trastornos del corazón y los vasos sanguíneos que incluyen cardiopatías coronarias y enfermedades cerebrovasculares [1]. Se estima que anualmente a nivel global se cobran 17.9 millones de vidas [2].

Este problema es mucho mayor en países en vía de desarrollo que en desarrollados y se considera que millones de personas padecen de factores de riesgo que no son comúnmente diagnosticados, tales como hipertensión arterial, tabaquismo, diabetes, hipercolesterolemias y dieta inadecuada [3]. Según datos del INEGI, en México durante el primer semestre de 2021 hubo 113,899 fallecimientos por EC y en Zacatecas las enfermedades del corazón ocuparon el primer lugar como causa de muerte en el estado con 1,511 decesos (806 hombres y 705 mujeres) como se muestra en la Figura 1.1 [4].

Rango	Total	Hombre	Mujer
1	Enfermedades del corazón 1,511	Enfermedades del corazón 806	Enfermedades del corazón 705
2	COVID-19 1,361	COVID-19 783	COVID-19 578
3	Diabetes mellitus 797	Agresiones (homicidios) 609	Diabetes mellitus 385
4	Agresiones (homicidios) 676	Diabetes mellitus 412	Tumores malignos 326
5	Tumores malignos 646	Tumores malignos 320	Influenza y neumonía 129

Figura 1.1 Defunciones registradas en Zacatecas de enero-junio 2021 [4].

Con base en la información anterior, es una necesidad atender oportunamente este tipo de enfermedades dando pauta a la integración de Inteligencia Artificial (IA) para generar una herramienta que ayude al personal de salud y a las personas portadoras de alguna patología, a obtener un diagnóstico certero y oportuno acorde a sus necesidades [5]. La tecnología y la medicina siguen un camino paralelo durante las últimas décadas. Los avances tecnológicos van modificando el concepto de salud y las necesidades sanitarias están influyendo en el desarrollo de la tecnología. La IA está formada por una serie de algoritmos lógicos suficientemente entrenados desde de los cuales las máquinas son capaces de tomar decisiones para casos concretos a partir de normas generales. Esta tecnología tiene aplicaciones en el diagnóstico, detección y seguimiento de pacientes con una evaluación pronóstica e individualizada. Por ejemplo, los modelos de machine learning (ML) podrían usarse para rastrear los signos vitales (en los monitores cardíacos) de los enfermos que reciben cuidados intensivos, alertando a los médicos si aumentan ciertos factores de riesgo, así, la inteligencia artificial puede recopilar los datos de esos dispositivos y buscar afecciones más complejas [6].

Sin embargo, estas técnicas no han sido aprovechadas en su totalidad debido a que aún se siguen dando sugerencias de manera general y no se han explorado las opciones de dividir por sexo biológico y edad para dar recomendaciones personalizadas, es por ello que en la presente investigación se dará enfoque en desarrollar un biomarcador que permita clasificar personas propensas a presentar una Enfermedad Cardíaca mediante aprendizaje supervisado.

1.1. Antecedentes

Las Enfermedades Cardiovasculares (ECV) son la principal causa de muerte en el mundo y en la mayoría de los países desarrollados, en donde se estima que causan 1.9 millones de muertes al año. Sin embargo, una alta proporción de estas enfermedades se pueden prevenir si se sigue una dieta saludable, se hace ejercicio físico y evitar el consumo de tabaco, entre otras medidas [1].

Las ECV como el infarto de miocardio y el accidente cerebrovascular, forman parte de las denominadas Enfermedades no transmisibles (ENT). Tres de cada cuatro personas padecen una ENT en países en vías de desarrollo, 4.45 millones de personas mueren al año por causa de alguna de ellas y, de esa cifra, 1.5 millones fallecen antes de los 70 años. Las Enfermedades Cardiovasculares provocan 1.9 millones de muertes al año, el cáncer, 1.1 millones; la diabetes, 260,000; y las enfermedades respiratorias crónicas, 240,000 [7].

En el mundo, las EVC tienen mayor índice de mortalidad, ya que anualmente se cobran más de 17.3 millones de vidas, y se estima que para 2030 esta cantidad ascienda a 23.6 millones de muertes por alguna de estas afecciones [2].

En México, las causas de fallecimiento varían según la edad y el sexo biológico de las personas, sin embargo, de acuerdo con cifras del INEGI, en 2020 se registraron 1,086,743 decesos. Las Enfermedades del Corazón ocupan el primer lugar en la tasa de mortalidad con el 38.7%, seguidas por complicaciones derivadas de covid-19 y diabetes mellitus [8].

Dentro de las complicaciones de las Enfermedades Cardíacas (EC), la función cognitiva disminuye rápidamente en los pacientes que presentan ataques cardíacos, lo que sugiere que la prevención de dichos ataques podría ayudar a preservar la salud cerebral. Una forma de enlazar la salud cerebral y cardíaca es a través de factores de riesgo compartidos, como el estilo de vida de las personas y el control de la presión arterial [9].

La detección y control de los factores de riesgos cardiovasculares (FRCV) tales como: hipertensión arterial (HTA), diabetes mellitus (DM), obesidad, dislipidemia y tabaquismo, entre otros, sigue siendo un método fundamental para prevenir estas patologías. La identificación de estos factores, permite establecer estrategias y medidas de control en los sujetos que aún no padecen la enfermedad, o si la han presentado, llevar un chequeo pertinente [10].

1.2. Planteamiento del problema de investigación

Los datos sobre mortalidad del INEGI muestran que en el 2020 fallecieron 218,704 personas por distintas enfermedades del corazón. Cabe señalar que para el 2019 la cifra de defunciones era de 156,041, es decir, hubo un incremento de 62,663 personas fallecidas entre ambos años. Esa cantidad supera el total de las generadas por la cuarta causa de muerte en el país (influenza y neumonía), cuyo número se ubicó ligeramente por debajo de los 60 mil decesos en el 2020 [8].

De acuerdo con la Organización para la Cooperación y el Desarrollo Económicos (OCDE), lamentablemente México es de los países que más avanzan en el número de fallecimientos por Enfermedades Cardiovasculares, mientras que las naciones en desarrollo las mitigan gracias a los programas de detección, prevención y tratamiento oportuno. Los padecimientos afectan en mayor medida a las naciones de ingresos bajos y medios, más del 80% de las defunciones son por esta causa e impactan casi por igual a hombres y mujeres, señala la Organización Mundial de la Salud (OMS), la cual también prevé que para 2030 más de 23 millones de personas morirán por alguna

ECV, principalmente por cardiopatías y accidentes cerebrovasculares. Muchos de esos decesos podrían evitarse con una alimentación saludable que reduzca el consumo de sal, ejercicio físico y sin consumo de tabaco, así como un diagnóstico temprano. En México hay una acumulación de los factores de riesgo como diabetes, obesidad, hipertensión arterial, sedentarismo y colesterol elevado [11].

Durante el 2020, en Zacatecas las Enfermedades del corazón fueron la primera causa de muerte en el estado con 2,565 casos, seguido de covid-19 con 2,432 defunciones y diabetes mellitus con 1,588 fallecimientos [12], como se observa en la Figura 1.2.

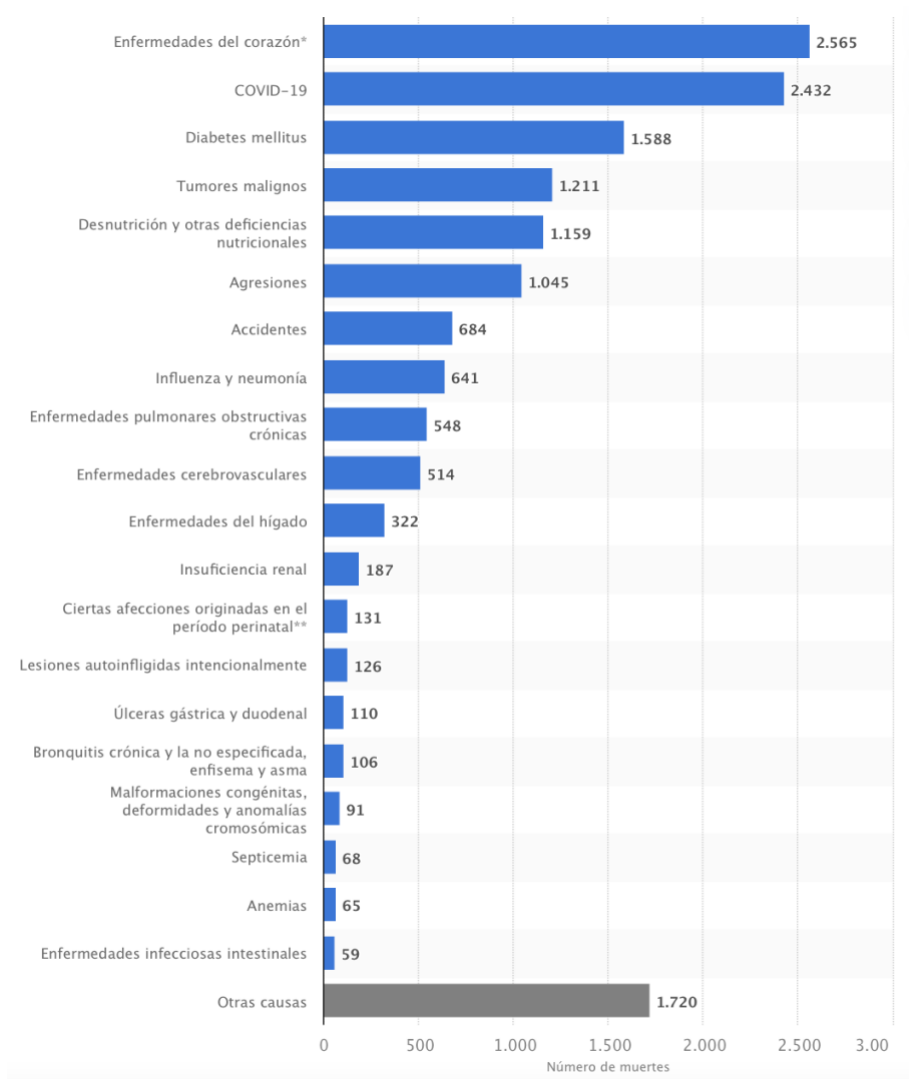


Figura 1.2 Número de fallecimientos registrados en Zacatecas en 2020, según las principales causas de mortalidad [13].

Tomando en cuenta lo anterior, la comunidad científica ha intentado mitigar los efectos de estas enfermedades mediante nuevos tratamientos, así como el diagnóstico oportuno que ayude a evitar secuelas a largo plazo, por medio del uso de Inteligencia Artificial, técnicas de machine learning y análisis de datos, como Daniel González Cedillo en 2019 [14], Chicco & Jurman en 2020 [15], Ali et al. en 2020 [16], Gallego Valcárcel David & Lucas Monsalve Delly Fabián en 2021 [17] y Faiyaz Waris & Koteeswaran en 2021 [18], quienes han creado modelos de predicción de pacientes propensos a presentar una Enfermedad Cardíaca, desarrollando una comparativa de distintos algoritmos de aprendizaje supervisado (naive bayes, KNN, random forest, regresión

lineal, logística, SVM, redes neuronales, árboles de decisión y refuerzo del gradiente) obteniendo resultados satisfactorios con un alto índice de precisión. Sin embargo, en cada trabajo hace falta especificar las características utilizadas en la implementación de los modelos de clasificación y con ello aportar información a futuras investigaciones.

En la presente investigación se pretende identificar las características o factores de riesgo más importantes para la detección temprana de Enfermedades Cardíacas con ayuda de métodos de selección de características como algoritmos genéticos, y con ello, crear un modelo de clasificación por medio de técnicas de aprendizaje automático (regresión logística, random forest, KNN). Comparando los resultados obtenidos de cada método se pretende contribuir en la predicción de pacientes potenciales a desarrollar una complicación cardíaca, extrayendo estos datos para después invitar a la sociedad en general a tomar medidas apropiadas para favorecer su salud.

1.3. Justificación del problema de investigación

De acuerdo con la Organización Mundial de la Salud, las enfermedades del corazón son la principal causa de muerte en todo el mundo [1]. Según datos del INEGI en 2020, México presentó 218,704 decesos de su población a consecuencia de estas patologías [8].

México se encuentra ante un grave problema de salud pública, la obesidad y las enfermedades no transmisibles relacionadas son ya una pandemia que afecta tanto a la salud individual y poblacional, afectando directamente la productividad, economía y bienestar del país. Las Enfermedades Crónicas No Transmisibles (ECNT), en particular el cáncer y las cardiometabólicas como Enfermedad Cardiovascular, hipertensión arterial y diabetes tipo 2, son sin duda el principal reto para el sistema de salud de nuestro país, tanto por su magnitud como el gran impacto en mortalidad prematura, deterioro de la calidad de vida y los costos de atención de sus complicaciones [19].

Los investigadores han tratado de contribuir mediante la incorporación de nuevas tecnologías que permiten la detección de enfermedades, como la inteligencia artificial (IA) que es el campo científico de la informática que se centra en crear programas y mecanismos que pueden mostrar comportamientos inteligentes. En otras palabras, la IA es el concepto según el cual las máquinas piensan como seres humanos, de modo que, es capaz de analizar grandes cantidades de datos (Big Data), identificar patrones y tendencias, formulando predicciones de manera automática, con rapidez y precisión [20].

Debido a la necesidad de fortalecer la atención primaria de salud para la detección oportuna de Enfermedades Cardíacas, se crea un biomarcador (modelo), a través de la implementación de Inteligencia Artificial y técnicas de Machine Learning (Regresión logística, Random forest, KNN), identificando los principales factores de riesgo de este padecimiento (mediante algoritmos genéticos), así mismo, este permitirá una clasificación de pacientes que muestran mayor incidencia a la predisposición de alguna complicación debido a la patología, ayudando a presentar sugerencias de hábitos saludables a la sociedad que muestre alta exposición de padecer una enfermedad del corazón, lo cual permita evitar estas afecciones.

1.4. Preguntas de investigación

- ¿Por qué ciertos factores de riesgo presentes en EC tienen mayor relevancia para la predicción

de esta patología?

- ¿Es posible crear un modelo de IA que tenga potencial clínico para ser usado en una aplicación con un desempeño por arriba del 0.70 en la predicción de EC?
- ¿Cuáles son los mejores algoritmos para predecir las Enfermedades Cardiacas?
- ¿Qué factores presentan los diabéticos para llegar a contraer mayor riesgo de cardiopatías?
- ¿Qué factores de riesgo se mantienen constantes a lo largo de la vida de las personas para desarrollar una EC?
- ¿Qué factores de riesgo intervienen para determinar el riesgo de contraer alguna Enfermedad Cardiacas dependiendo del sexo biológico?

1.5. Objetivo general

Generar un biomarcador que permita detectar Enfermedades Cardiacas por medio de características antropométricas, mediante la implementación de técnicas de aprendizaje automático, como algoritmos genéticos, Regresión logística, Random forest y KNN.

1.5.1. Objetivos específicos

- Identificar los estudios relacionados respecto al tema de investigación.
- Buscar conjuntos de datos sobre EC.
- Determinar las características más significativas que causan EC mediante algoritmos genéticos.
- Generar modelos de predicción por medio de técnicas de machine learning.
- Comparar los riesgos presentes de EC en pacientes no diabéticos, prediabéticos y diabéticos.
- Comparar los riesgos presentes de EC en los distintos rangos de edad.
- Evaluar los riesgos presentes en EC y segmentarlo por sexo biológico.

1.6. Hipótesis

A partir de datos y el uso de algoritmos genéticos, en combinación con Regresión logística, Random forest y KNN es posible crear un biomarcador que permita detectar Enfermedades Cardiacas.

1.7. Trabajo a realizar

Con el fin de generar un biomarcador que permita predecir EC en personas diabéticas, este trabajo pretende analizar un conjunto de datos denominado: “Indicadores de salud de Enfermedades Cardiacas” [21], el cual consta de 22 características binarias u ordinales, con un total de 253,680 pacientes entre sanos y enfermos, de los cuales, 141,974 son hombres (56%), 111,706 mujeres (44%) con un rango de edad entre 18 y 80 años en adelante.

Se llevará a cabo un preprocesamiento, consistirá en dividir en ocho subconjuntos diferentes (No diabéticos, Prediabéticos, Diabéticos, Jóvenes, Adultos, Adultos Mayores, Hombres y Mujeres) la información y con ello realizar un procesamiento de datos, una vez teniendo los grupos separados, se realizará una selección de características a través de un algoritmo genético, este permitirá seleccionar el mejor subconjunto de variables más significativas que favorezcan en la generación de modelos de clasificación para cada grupo, y por último, evaluar dichos modelos mediante validación cruzada bajo las métricas de desempeño como área bajo la curva (AUC), sensibilidad, especificidad y exactitud para evitar el sobreajuste de datos, logrando un resultado más óptimo y preciso.

1.8. Estructura de la tesis

El resto del documento se distribuye de la siguiente manera:

- Capítulo 2. Describe el marco teórico de la presente investigación, trabajos relacionados con temas de Enfermedades Cardiacas, modelos de clasificación, así como la aplicación de aprendizaje automático.
- Capítulo 3. Describe la metodología utilizada en la investigación, además del análisis e implementación de algoritmos genéticos para la obtención de características más significativas, y generación de modelos de clasificación en la predicción de Enfermedades Cardiacas, segmentados en diversos grupos.
- Capítulo 4. Describe los resultados obtenidos en la implementación de análisis de datos y técnicas de machine learning.
- Capítulo 5. Conclusiones, objetivos alcanzados y trabajos publicados.

Capítulo 2. Marco Teórico

2.1. Enfermedades Cardiacas

La enfermedad cardiaca se refiere a cualquier afección del corazón incluyendo una gran variedad de enfermedades, por ejemplo:

- Enfermedad de las arterias coronarias. Es el tipo más común de afección cardíaca, se desarrolla cuando las arterias que suministran sangre al corazón se obstruyen con placa (contiene colesterol y otras sustancias), esto hace que se endurezcan y estrechen [22].
- Problemas en el ritmo cardíaco (arritmias). Arritmia se refiere a un ritmo cardíaco irregular. Ocurre cuando los impulsos eléctricos que coordinan los latidos del corazón no funcionan adecuadamente y como resultado, el corazón puede latir demasiado rápido, demasiado lento o de forma errática [22].
- Defectos cardíacos de nacimiento (defectos cardíacos congénitos). Son el tipo más común de defecto al nacimiento siendo el resultado de un desarrollo cardiaco anormal [23].
- Enfermedad de las válvulas cardíacas (estenosis aórtica). Es la enfermedad más prevalente en todo el mundo, caracterizada por fibrocalcificación y engrosamiento de las válvulas lo cual trae como consecuencia que el corazón no pueda bombear sangre del ventrículo derecho a la arteria pulmonar y también se disminuye el flujo sanguíneo del ventrículo izquierdo a la aorta [24].
- Infección del corazón. Se produce cuando los gérmenes (virus, bacterias, parásitos) llegan hasta el corazón o las válvulas cardíacas [25].

La comunidad científica ha estudiado y encontrado algunas características que típicamente se les conoce como factores de riesgo, los cuales nos ayudan a llevar un control de las Enfermedades Cardiacas. Por lo tanto, la mayoría de las patologías anteriormente mencionadas comparten un conjunto de factores de riesgo que intervienen en su desarrollo, los cuales se pueden observar en la sección 2.1.1.

2.1.1. Factores de riesgo cardiovascular

Se entiende como factor de riesgo cardiovascular (FRCV) aquella característica biológica, condición y/o modificación del estilo de vida, que aumenta la probabilidad de padecer o de fallecer por cualquier causa de una enfermedad cardiovascular (ECV) en aquellos individuos que lo presentan a medio y largo plazo [26]. En otras palabras, son diversas características de una persona que aumentan la probabilidad de desarrollar una enfermedad. Entre los factores de riesgo de EC se incluyen los siguientes:

- Edad. El envejecimiento aumenta el riesgo de que las arterias se dañen y se estrechen, y de que el músculo cardíaco se debilite o engrose [27].
- Sexo biológico. En general, los hombres corren mayor riesgo de padecer una enfermedad cardíaca. El riesgo en las mujeres aumenta después de la menopausia [27].
- Antecedentes familiares. Los antecedentes familiares de enfermedades cardiacas

aumentan tu riesgo de padecer enfermedad de las arterias coronarias, especialmente si uno de tus padres la desarrolló a temprana edad (antes de los 55 años en caso de un familiar hombre, como tu hermano o tu padre, y antes de los 65 años en caso de una familiar mujer, como tu madre o hermana) [27].

- Tabaquismo. La nicotina contrae los vasos sanguíneos y el monóxido de carbono puede dañar su revestimiento interno, lo que los vuelve más propensos a la aterosclerosis. Los ataques cardíacos son más comunes en fumadores que en no fumadores [28].
- Mala alimentación. Una dieta con alto contenido de grasas, sal, azúcar y colesterol puede contribuir al desarrollo de una enfermedad cardíaca [28].
- Presión arterial alta. La presión arterial alta no controlada puede producir el endurecimiento y engrosamiento de las arterias, lo que estrecha los vasos por los que circula la sangre [14].
- Niveles altos de colesterol en la sangre. Los niveles altos de colesterol en la sangre pueden aumentar el riesgo de que se formen placas y de padecer aterosclerosis [14].
- Diabetes. La diabetes aumenta la exposición de sufrir enfermedades cardíacas. Ambas afecciones comparten factores de riesgo similares, como la obesidad y la presión arterial alta [27].
- Obesidad. El exceso de peso normalmente empeora otros factores de riesgo de enfermedades cardíacas [28].
- Inactividad física. La falta de ejercicio también está relacionada con muchas formas de enfermedad cardíaca y con algunos de sus otros factores de riesgo [25].
- Estrés. El estrés no tratado puede dañar las arterias y empeorar otros factores de riesgo de enfermedades cardíacas [29].
- Mala higiene dental. Es importante cepillarse los dientes y las encías, y usar hilo dental con frecuencia, y hacerse chequeos dentales periódicos. Si los dientes y las encías no están sanos, los gérmenes pueden entrar al torrente sanguíneo y llegar al corazón, lo que produce endocarditis [25].

Como se puede observar, si no existe una buena salud en las personas, el conjunto de estos factores de riesgo puede derivar en el desarrollo de EC y diabetes (descrita a continuación), sin embargo, pueden prevenirse o aminorarse mediante un estilo de vida saludable (descrito en la sección 2.3).

2.2. Diabetes

La diabetes mellitus (DM) es una enfermedad crónica caracterizada por la presencia de glucosa alta en el torrente sanguíneo debido a una deficiencia en la producción de la insulina o un mal uso de la misma. La insulina es una hormona producida en el páncreas que permite la absorción de la glucosa proveniente de los alimentos que ingerimos en las células del cuerpo, para aprovecharla en forma de energía. Al no tener una producción o función adecuada de esta hormona los niveles de glucosa son altos (hiperglucemia) y con el tiempo esto se va asociando con daños importantes y fallas en varios órganos y tejidos [30].

Entre los diferentes tipos de diabetes más comunes encontramos a la diabetes mellitus tipo 1

(DM1) y diabetes mellitus tipo 2 (DM2), en el primer caso es causado por una reacción autoinmune donde las defensas atacan a las células del páncreas que producen insulina, como resultado el cuerpo produce menos o nada de esta hormona provocando de esta manera un bajo control de la glucosa en la sangre. Las causas de este deterioro aún son desconocidas, pero están ligadas en una combinación de las condiciones genéticas y medio ambiente. Este tipo de diabetes afecta frecuentemente más a niños y adultos jóvenes, los síntomas más comunes que presentan los pacientes son: sed anormal y bocas seca, pérdida repentina de peso, orinar más de lo común, falta de energía o cansancio, hambre constante, visión borrosa e incontinencia nocturna. En tanto la DM2 representa alrededor del 90% de todos los casos con este padecimiento, generalmente se caracteriza por la resistencia a la insulina y los niveles de glucosa alcanzan niveles muy altos, lo que hace que sistema siga liberando más de esta hormona agotando eventualmente al páncreas, lo que resulta que con el tiempo el órgano produzca cada vez menos insulina causando niveles aún más altos de azúcar en la sangre. Este tipo de diabetes se presenta más común en adultos mayores, pero se observa cada vez con más frecuencia en niños, adolescentes y adultos jóvenes debido al aumento de casos de obesidad, inactividad física y la mala alimentación, los síntomas de DM2 son similares a los del tipo 1 pero suman otros como: Cicatrización lenta de heridas, infecciones recurrentes de la piel, hormigueo o entumecimiento en manos y pies. Estos síntomas suelen ser leves o ausentes, por lo que las personas con este tipo de diabetes pueden vivir varios años con la afección antes de ser diagnosticados [31].

Si no existe un buen control o cuidado de diabetes, esta puede dar paso a una o más complicaciones de la enfermedad como las que se describen en la sección 2.2.1.

2.2.1. Complicaciones comunes de diabetes

Las personas con diabetes tienen un nivel de riesgo alto de desarrollar un número severo de problemas de salud. Constantemente los niveles de glucosa en sangre afectan el corazón, vasos sanguíneos, ojos, riñones, nervios y dientes. Por esto es importante mantener los niveles, no solo de glucosa sino también la presión y colesterol lo más cercano a niveles óptimos para prevenir complicaciones derivadas de la diabetes. Entre las complicaciones más comunes por este padecimiento son las siguientes [32]:

- **Enfermedades cardiovasculares (ECV):** La diabetes afecta frecuentemente el corazón y vasos sanguíneos causando complicaciones fatales, como la enfermedad de arterias coronarias y accidentes cerebrovasculares. Además, la presión arterial alta, el colesterol alto y la glucosa alta contribuyen a aumentar el riesgo de desarrollar complicaciones cardiovasculares, siendo estas la principal causa de muertes en pacientes diabéticos [27].
- **Insuficiencia renal o nefropatía diabética:** Esta enfermedad es más común en personas diabéticas que en aquellas sin diabetes, y es causada por el daño a los pequeños vasos sanguíneos en los riñones, lo que lleva que este órgano se vuelva menos eficiente o fallen completamente [33].
- **Neuropatía diabética:** Cuando los niveles de glucosa y presión son demasiado altos conducen a dañar los nervios de todo el cuerpo lo que puede derivar en problemas de digestión, disfunción eréctil y otras funciones. Afecta comúnmente a las extremidades en especial los pies, el daño a los nervios en estas partes del cuerpo puede provocar dolor, hormigueo y pérdida de sensibilidad. Este último particularmente puede derivar en lesiones desapercibidas lo que conlleva a infecciones graves y posibles amputaciones, por esto los diabéticos son 25 veces más propensos a sufrir este tipo de amputaciones respecto

a los no diabéticos [34].

- **Retinopatía diabética:** A parte de los niveles de glucosa, la presión arterial y el colesterol alto son las principales causas de una visión reducida o ceguera, y la gran mayoría de las personas diabéticas desarrollarán alguna forma esta enfermedad y es la principal causa de ceguera en la población económicamente activa. Por ello es importante mantener control ocular regular y monitoreo constante de niveles de glucosa y lípidos si se es diabético [35].
- **Complicaciones en el embarazo:** Con cualquier tipo de diabetes que tenga la madre, la glucosa alta en el torrente sanguíneo durante el embarazo puede derivar en problemas en el parto, traumatismos tanto para el niño como la madre, en casos especiales el niño puede sufrir caída repentina de glucosa después del nacimiento, y los niños expuestos durante mucho tiempo a niveles de glucosa en el útero tienen un mayor riesgo de desarrollar diabetes en el futuro [36].
- **Complicaciones bucales:** La periodontitis o inflamación de las encías es una causa frecuente de pérdida de piezas dentales asociada también con un mayor riesgo de enfermedad cardiovascular, los diabéticos son más propensos a desarrollar este padecimiento por el descontrol en los niveles de glucosa [32].

De todas las complicaciones anteriormente mencionadas, esta investigación estará centrada en las Enfermedades Cardiovasculares y/o Cardíacas, ya que los pacientes con diabetes tienen de dos a tres veces más probabilidades de desarrollar una ECV que las personas sin diabetes, debido a que los altos niveles de glucosa en sangre pueden hacer que el sistema de coagulación sanguínea sea más activo, lo que aumenta el riesgo de coágulos de sangre. Por lo tanto, estas enfermedades en conjunto aumentan los casos de mortalidad en la población.

A continuación, se presentan algunas recomendaciones preventivas que contribuyan en aminorar el desarrollo de una afección cardíaca.

2.3. Prevención

Algunas medidas de estilo de vida pueden ayudar a reducir el riesgo de Enfermedad Cardíaca. Estas incluyen:

- Seguir una dieta balanceada: optar por una dieta saludable para el corazón que sea rica en fibra y contenga cereales integrales, frutas y vegetales frescos. Además, puede ayudar comer menos alimentos procesados y grasas, sal y azúcar agregadas [37].
- Hacer ejercicio regularmente: esto puede ayudar a fortalecer el corazón y el sistema circulatorio, reducirá el colesterol y mantendrá la presión arterial normal [37].
- Mantener un peso corporal moderado: verificando el índice de masa corporal (IMC) [38].
- Dejar o evitar fumar: fumar es un importante factor de riesgo para enfermedades cardíacas y cardiovasculares [38].
- Limitar el consumo de alcohol: las mujeres no deben consumir más de una bebida estándar por día y los hombres no deben consumir más de dos bebidas estándar por día [38].
- Controlar las afecciones subyacentes: buscar tratamiento para los padecimientos que afectan la salud del corazón, como presión arterial alta, obesidad y diabetes [37].

Cada año mueren más personas por EC que por cualquier otra causa. Más de tres cuartas partes

de los fallecimientos relacionados con cardiopatías ocurren en países en proceso de desarrollo, por lo cual, la Ciencia y la Tecnología han concentrado sus esfuerzos para tratar de mitigar el aumento de decesos por estas patologías, a través del diagnóstico oportuno, nuevos tratamientos, medicina personalizada, entre otros, que ayude a evitar secuelas a largo plazo mediante técnicas de inteligencia artificial que se enfocan en el desarrollo de biomarcadores que permiten estimar el riesgo de contraer una enfermedad en una persona sana.

2.4. Biomarcador

El concepto de "biomarcador", un acrónimo de "marcador biológico", se refiere a una amplia subcategoría de datos médicos del paciente, es decir, indicadores que se pueden medir de manera precisa y reproducible. Los datos médicos contrastan con síntomas que se limitan a aquellos indicios de salud o enfermedad percibidos por los propios pacientes [39].

Por ello, es posible crear un biomarcador en conjunto con herramientas tecnológicas como la inteligencia artificial (descrito en la sección 2.5) y técnicas derivadas para contribuir en el diagnóstico oportuno de diferentes enfermedades.

2.5. Inteligencia Artificial (IA)

La inteligencia artificial es la serie de tecnologías que sirven para emular características o capacidades exclusivas del intelecto humano. Si bien no hay una definición exacta sobre lo que significa, el término se aplica cuando una máquina imita las funciones cognitivas que los humanos asocian con otras mentes humanas, como aprender o resolver problemas, entre otros [40].

En informática, la inteligencia artificial se ha centrado tradicionalmente en problemas de alto nivel; en impartir habilidades de alto nivel para usar el lenguaje, abstracciones y conceptos de la forma y resolver tipos de problemas ahora reservados para los humanos (McCarthy et al., 1955) [41].

En 1968 Marvin Minsky definió a la IA como la ciencia de hacer que la máquinas hagan cosas que requerirían inteligencia si las hubiera hecho un humano (Marvin Minsky, 1968). En 2018 la comisión europea se refirió a la inteligencia artificial como sistemas que muestran un comportamiento inteligente: analizando su entorno pudiendo realizar diversas tareas con cierto grado de autonomía para alcanzar objetivos específicos (European Commission, 2018) [41].

2.5.1. Técnicas de IA

Como se mencionó en la sección anterior, la IA es una serie de tecnologías y metodologías que ayudan a resolver un problema. A continuación, se mencionan las técnicas más utilizadas:

- Aprendizaje automático (Machine Learning, Deep learning) [42].
- Sistemas basados en conocimiento [43].
- Visión computacional [44].
- Procesamiento de voz y lenguaje natural [45].
- Lógica difusa [46].

- Redes neuronales [47].
- Computación evolutiva [48].
- Sistemas multiagente [49].
- Robótica [50].
- Técnicas heurísticas [51].

Todas las técnicas mencionadas anteriormente hacen referencia a distintos campos de investigación y desarrollo de Inteligencia Artificial. Para este trabajo se utilizará Machine Learning o aprendizaje automático (descrito en la sección 2.5.2) aplicado a análisis de datos en el ámbito clínico.

2.5.2. Machine Learning (ML)

Machine Learning (por sus siglas en inglés) se fundamenta en las matemáticas, estadística y computación. Los desarrollos en computación llegaron hasta la década de 1940, pero antes de ese momento hubo varios métodos y técnicas matemáticas y estadísticas que sirvieron de base para lo que hoy se conoce como machine learning: el teorema de Bayes (1763), el ajuste por mínimos cuadrados (1805) y las cadenas de Márkov (1913) [52].

ML o también llamado “aprendizaje automático” es una técnica de inteligencia artificial que permite que un sistema aprenda de datos a través de programación de cadenas de código, es decir, que los ordenadores puedan aprender por sí mismos, sin embargo, no es un proceso simple (Hurwitz y Kirsch, 2018) [41].

El aprendizaje en un sistema de ML consiste en ajustar los parámetros de un modelo en función de los datos recibidos. Este conjunto de datos (data set), contiene variables tanto independientes como dependientes. Las variables independientes son aquellas columnas del data set usadas por el algoritmo para generar un modelo que prediga lo mejor posible las variables dependientes (son resultado de una correlación entre variables independientes). Sin embargo, la aplicación de esta técnica no se centra únicamente en implementar un modelo y entrenarlo, sino que, cuenta con una serie de pasos a seguir para aumentar la posibilidad de éxito [52] como se muestra en la Figura 2.1, cabe señalar que estas etapas son genéricas y pueden cambiar acorde a los objetivos de cada proyecto de ML.

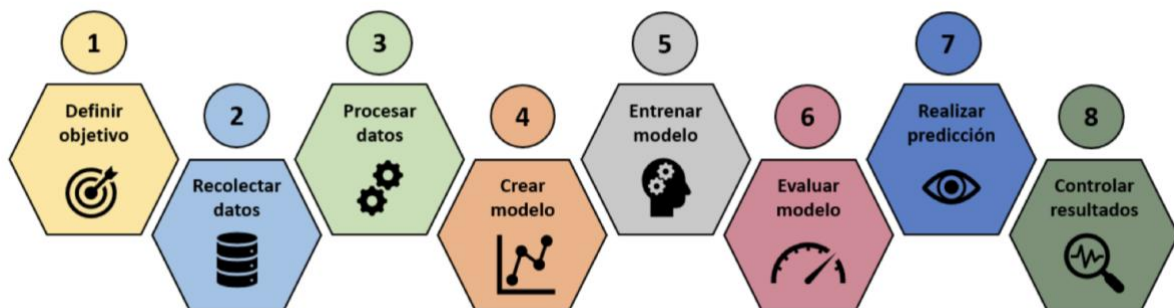


Figura 2.1 Etapas genéricas para llevar a cabo un proyecto de Machine Learning [52].

Existen varios tipos de técnicas de ML (se describen a continuación), que utilizan análisis de datos y algoritmos que aprenden iterativamente de la información, permitiendo a los ordenadores encontrar referencias escondidas sin tener que indicar de manera explícita dónde buscar. Según

el problema dado y de los datos disponibles es la que se debe seleccionar.

2.5.2.1 Aprendizaje supervisado

Comienza típicamente con un conjunto establecido de datos y una cierta comprensión de cómo se clasifican estos datos. El aprendizaje supervisado tiene la intención de encontrar patrones en datos que se pueden aplicar a un proceso de analítica. Estos datos tienen características etiquetadas que definen el significado de los datos [53].

El aprendizaje supervisado requiere un conjunto de datos de entrenamiento que incluya ejemplos para la entrada, así como respuestas etiquetadas o valores objetivo para la salida. Un ejemplo podría ser la predicción de los usuarios activos suscritos a una plataforma de mercado en el plazo de un mes como salida (considerada como la variable objetivo o variable y) basada en diferentes características de entrada, como el número de productos vendidos o las opiniones positivas de los usuarios (a menudo denominadas características de entrada o variables x). Los pares de datos de entrada y salida del conjunto de entrenamiento se utilizan entonces para calibrar los parámetros abiertos del modelo ML. Una vez que el modelo ha sido entrenado con éxito, puede utilizarse para predecir la variable objetivo y a partir de puntos de datos nuevos o no vistos de las características de entrada x . En cuanto al tipo de aprendizaje supervisado, podemos distinguir además entre los problemas de regresión, en los que se predice un valor numérico (por ejemplo, el número de usuarios), y los problemas de clasificación, en los que el resultado de la predicción es una afiliación de clase categórica, como "mirones" o "compradores" [42].

En los problemas de aprendizaje supervisado, el análisis comienza con un conjunto de datos que contiene ejemplos de entrenamiento con etiquetas correctamente asociadas. El algoritmo aprenderá la relación entre los datos y sus etiquetas, aplicará esa relación aprendida para clasificar datos completamente nuevos que el modelo no haya visto antes. En la Figura 2.2 se puede observar un ejemplo, cuando se aprende a clasificar imágenes de gatos, el modelo toma miles de imágenes de gatos junto con la etiqueta "gato". El algoritmo aprenderá esta relación y cuando se le muestre una nueva imagen, esta vez sin ninguna etiqueta, podrá aplicar esa relación aprendida y determinar si es un gato o no [54].



Figura 2.2 Ejemplo de clasificación de aprendizaje supervisado [54].

2.5.2.2 Aprendizaje no supervisado

Se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar. La comprensión del significado detrás de estos datos requiere algoritmos que clasifican los datos con

base en los patrones o clústeres que encuentra. El aprendizaje no supervisado lleva a cabo un proceso iterativo, analizando los datos sin intervención humana [53].

El aprendizaje no supervisado tiene lugar cuando el sistema de aprendizaje debe detectar patrones sin etiquetas o especificaciones preexistentes. Así, los datos de entrenamiento sólo consisten en variables x con el objetivo de encontrar información estructural de interés, como grupos de elementos que comparten propiedades comunes (lo que se conoce como clustering), o representaciones de datos que se proyectan desde un espacio de alta dimensión a otro de menor (lo que se conoce como reducción de la dimensionalidad). Un ejemplo destacado de aprendizaje no supervisado en los mercados electrónicos es la aplicación de técnicas de clustering para agrupar clientes o mercados en segmentos, con el fin de realizar una comunicación más específica para el grupo objetivo [42].

En la Figura 2.3 se muestra un ejemplo de esto, en donde existe un conjunto de imágenes de distintos animales, el algoritmo no supervisado realizará una agrupación para cada uno de los tipos de animales de acuerdo a las características y similitudes que poseen (este agrupamiento sería el resultado final del modelo). El algoritmo aprenderá a agrupar los tipos de animales, por lo tanto, cuando se le introduzca un nuevo animal, podrá aplicar esa relación aprendida y determinar a qué grupo pertenece [54].

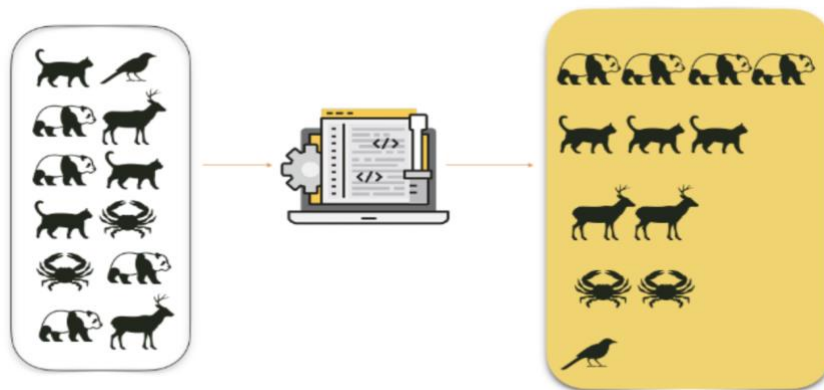


Figura 2.3 Ejemplo de clasificación de aprendizaje no supervisado [54].

De estos dos tipos de aprendizaje automático, la que más se utiliza para desarrollar biomarcadores en el ámbito clínico es el aprendizaje supervisado [55], mismo que se implementará para realizar la presente investigación. Típicamente esta técnica consta de una metodología, la cual se describe a continuación.

2.6. Algoritmos

Un algoritmo es una serie de pasos que describen el proceso que se debe seguir para solucionar un problema específico.

La metodología general que se utiliza para crear un modelo de machine learning consiste en una serie de etapas como obtención de datos, preprocesamiento, extracción y selección de características, entrenamiento, pruebas del modelo y por último análisis de resultados [56].

Dependiendo de la tarea de aprendizaje, el campo ofrece varias clases de métodos de ML, cada uno de ellos con múltiples especificaciones y variantes, incluyendo modelos de regresión, algoritmos basados en instancias, árboles de decisión, métodos bayesianos, redes neuronales

artificiales, entre otros.

A continuación, se explican brevemente algunos de los algoritmos de clasificación empelados en la experimentación de la presente investigación: regresión logística, random forest y K vecinos más cercanos (KNN por sus siglas en inglés).

2.6.1. Regresión logística

La regresión logística es una técnica de aprendizaje automático que proviene del campo de la estadística. Es un algoritmo que en lugar de resolver problemas de valor continuo atiende problemas de clasificación, en los que se obtienen valores binarios entre 0 y 1.

Mide la relación entre la variable dependiente, la afirmación que se desea predecir, con una o más variables independientes, el conjunto de características disponibles para el modelo. Para ello utiliza una función logística que determina la probabilidad de la variable dependiente. Como se ha comentado anteriormente, lo que se busca en estos problemas es una clasificación, por lo que la probabilidad se ha de traducir en valores binarios. Se utiliza un valor umbral (generalmente es 0.5, aunque se puede aumentar o reducir para gestionar el número de falsos positivos o falsos negativos) para establecer los valores de probabilidad para que la clasificación se considere cierta o falsa.

La función que relaciona la variable dependiente con las independientes también se le llama función sigmoide. La función sigmoide es una curva en forma de S que puede tomar cualquier valor entre 0 y 1, pero nunca valores fuera de estos límites. La ecuación que define la función sigmoide se muestra en la ecuación 1:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

donde x es un número real. En la ecuación 1 se puede ver que cuando x tiene a menos infinito el cociente tiende a cero. Por otro lado, cuando x tiende a infinito el cociente tiende a la unidad. En la Figura 2.4 se muestra una representación gráfica de la función logística (función sigmoide) [57].

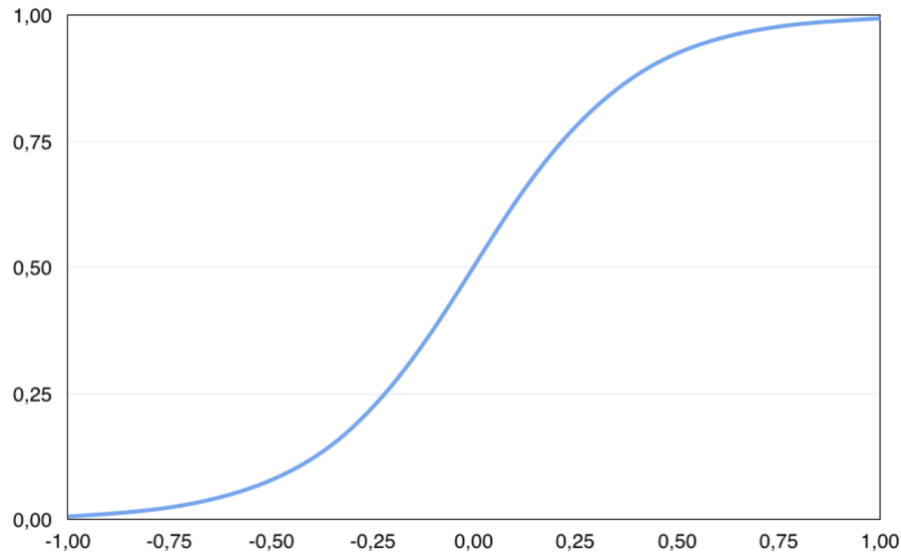


Figura 2.4 Función sigmoide para regresión logística [57].

2.6.2. Random forest

El algoritmo random forest es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento: los resultados obtenidos se combinan a fin de obtener un modelo único más robusto, en comparación con los resultados de cada árbol por separado. Cada árbol se obtiene mediante un proceso de dos etapas:

1. Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables m (predictores) de forma que $m < M$ (donde M = total de predictores).
2. Cada árbol crece hasta su máxima extensión.

Cada árbol generado por el algoritmo random forest contiene un grupo de observaciones aleatorias (elegidas mediante bootstrap, que es una técnica estadística para obtener muestras de una población donde una observación se puede considerar en más de una muestra). Las observaciones no estimadas en los árboles (también conocidas como “out of the bag”) se utilizan para validar el modelo. Las salidas de todos los árboles se combinan en una salida final Y (conocida como ensamblado) que se obtiene mediante alguna regla (generalmente el promedio, cuando las salidas de los árboles del ensamblado son numéricas y, conteo de votos, cuando las salidas de los árboles del ensamblado son categóricas), esto se puede observar en la Figura 2.5 [58].

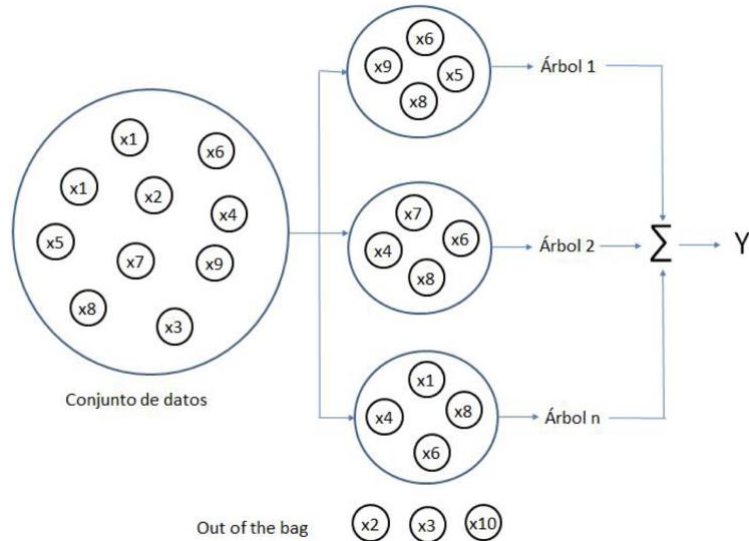


Figura 2.5 Algoritmo de random forest [58].

2.6.3. K vecinos más cercanos (KNN)

El método KNN (K-Nearest Neighbors) o k-vecinos más cercanos es un algoritmo de aproximación y clasificación no paramétrico, el cual realiza una clasificación de una base de datos basado en el cálculo de distancias entre una muestra a clasificar y un conjunto de referencia. Es usado como método de clasificación de objetos que consiste en hacer pruebas a través de ejemplos cercanos en un lugar determinado de los elementos KNN donde se hace una aproximación de los elementos más cercanos. El algoritmo comprende dos etapas:

- 1) Etapa de entrenamiento en donde el algoritmo busca una combinación de parámetros óptimos.
- 2) Etapa de prueba o aplicación que, ante nuevos datos, el algoritmo clasifica y asigna una clase de acuerdo con parámetros anteriormente obtenidos.

Luego de identificar las clases, calcula la distancia (comúnmente es euclidiana dada por la ecuación 2) con los nuevos datos a clasificar, seleccionando las k instancias más cercanas para clasificar. Finalmente agrega los resultados obtenidos con las k seleccionadas para así obtener los datos deseados [59].

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (2)$$

Donde x_i es el punto uno, y_i es el punto dos y (x_i, y_i) dentro de la raíz cuadrada son las coordenadas de los puntos.

Como se pudo observar, los algoritmos anteriormente especificados son algunos de los más utilizados en trabajos realizados del ámbito de la salud, los cuales serán implementados en la presente investigación; sirven para crear modelos de predicción, para después evaluar su desempeño obtenido mediante diversas métricas como las que se mencionan en la siguiente sección.

2.6.4. Métricas de desempeño y validación cruzada (CV)

Un algoritmo se evalúa para determinar si realizará una buena predicción, basándose en cómo funciona el modelo con nuevos y futuros datos, para ello se utilizan varias métricas de evaluación que se emplearon en la investigación para validar de manera empírica el desempeño del modelo [60].

La matriz de confusión es una herramienta que permite obtener el desempeño de un algoritmo, se aplica en problemas de clasificación binaria (2 clases). En la Tabla 2.1 se muestra que está compuesta por Verdaderos Positivos (VP), Falsos Negativos (FN), Falsos Positivos (FP) y Verdaderos Negativos (VN). Cada columna de la matriz representa el número de predicciones de cada clase y las filas interpretan los valores reales [61].

Tabla 2.1 Matriz de confusión [61].

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Mediante la matriz de confusión se pueden obtener algunas métricas de evaluación como:

- Exactitud (accuracy)
- Sensibilidad
- Especificidad

Exactitud (ecuación 3) es una métrica que indica el porcentaje de predicciones clasificadas correctamente, tanto para la clase positiva y negativa para la clasificación binaria como para cualquier clase en caso de clasificación múltiple [62].

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN} \quad (3)$$

La sensibilidad (ecuación 4) indica la tasa de clasificación positiva, es decir, la proporción de casos positivos que el modelo predijo correctamente (verdaderos positivos) [63].

$$Sensibilidad = \frac{VP}{VP + FN} \quad (4)$$

Especificidad (ecuación 5) es la capacidad de un algoritmo para predecir un falso positivo, es decir, que tan bien predice los verdaderos negativos [63].

$$Especificidad = \frac{VN}{VN + FP} \quad (5)$$

La curva ROC (Receiver Operating Characteristic) por sus siglas en inglés, es una técnica muy utilizada para evaluar el desempeño de los clasificadores binarios. Son gráficas bidimensionales en las cuales la tasa de VP (sensibilidad) se grafica sobre el eje Y, y la tasa de FP (1-especificidad) sobre el eje X. Este par de valores produce un punto en el espacio ROC, el cual es delimitado por las coordenadas (0,0), (0,1), (1,1) y (1,0). Existen clasificadores que producen salidas continuas que pueden ser consideradas una estimación de la probabilidad de que una instancia sea miembro de una clase (positiva o negativa). Por lo tanto, si variamos el umbral para el cual una instancia pertenece a una clase, podremos producir diferentes puntos ROC, entonces al conectar todos los puntos incluyendo (0,0) y (1,0) obtendremos la curva ROC empírica para el clasificador. En el caso de los clasificadores discretos que sólo producen una etiqueta de la clase, podemos calcular las tasas de VP y FP a través de cortes progresivos de los datos [64]. En la Figura 2.6 se muestra una representación de las curvas ROC que un modelo pudiera alcanzar, siendo el valor de 1 el valor máximo, lo que quiere decir que tiene un excelente desempeño.

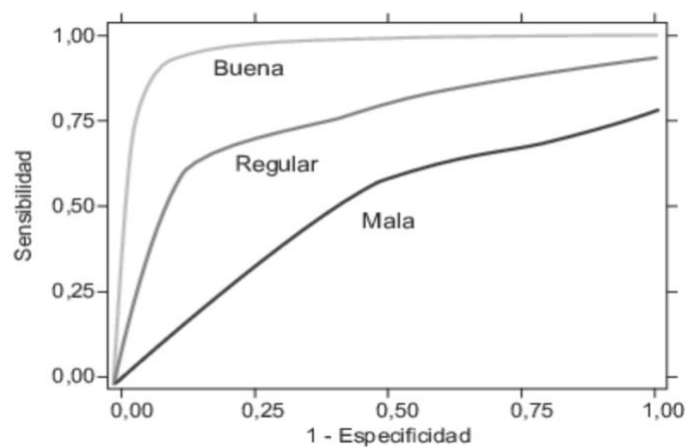


Figura 2.6 Representación de posibles curvas ROC [64].

El área bajo la curva (AUC) es una métrica de precisión estándar de Machine Learning (ML) para modelos de clasificación binaria, se obtiene a partir de la curva ROC. Mide su capacidad de predecir una mayor puntuación para ejemplos positivos en comparación con negativos, devuelve un valor decimal comprendido entre 0 y 1; los valores de AUC cercanos a 1 indican un modelo de aprendizaje automático muy preciso [60]. En la Figura 2.7 se muestra un ejemplo de esta métrica.

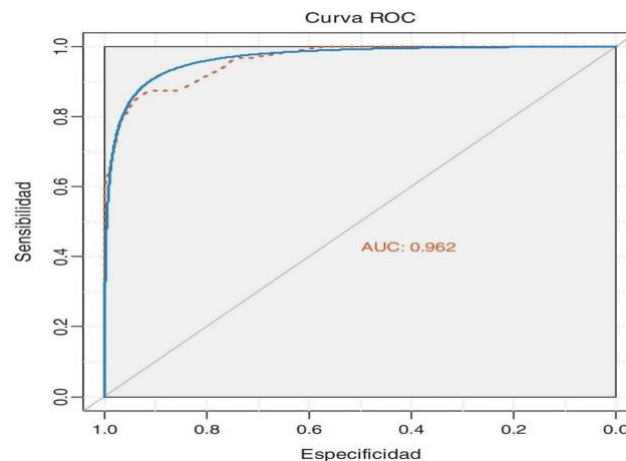


Figura 2.7 Ejemplo de área bajo la curva [60].

Por medio de validación cruzada K-Fold (K-Fold Cross-Validation) se evalúa el rendimiento de los modelos de predicción generados. Es una estrategia que permite estimar la capacidad predictiva de estos, cuando se aplican a nuevas observaciones, haciendo uso únicamente de los datos de entrenamiento. Consiste en dividir los datos de forma aleatoria en k particiones de aproximadamente el mismo tamaño, k-1 folds se usan para entrenar el modelo y una partición se utiliza como prueba. Este proceso se repite k veces utilizando una partición distinta como validación en cada iteración. Genera k estimaciones del error cuyo promedio se emplea como estimación final [65], como se muestra en la Figura 2.8.

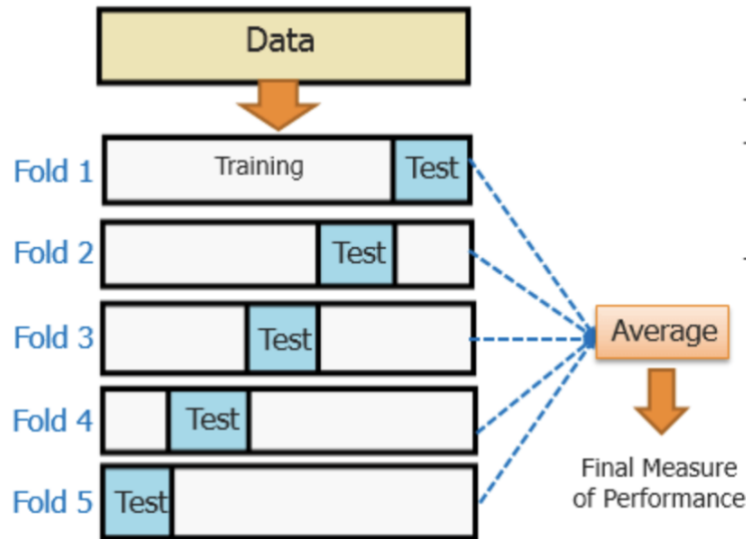


Figura 2.8 Validación cruzada [65].

2.6.5. Selección de características

Uno de los principales problemas al construir un modelo de los algoritmos vistos anteriormente, es seleccionar cuáles características de todas las posibles se van a utilizar, a esto se le conoce como reducción de dimensionalidad. Dentro de los métodos de selección de variables, están los siguientes:

- Algoritmos genéticos (inspirados en la selección natural)
- Estocásticos (Lasso, Boruta)
- Estadísticos (PCA)
- Inteligencia colectiva (colonia de hormigas)

El algoritmo genético es una técnica de programación basado en la reproducción de los seres vivos e imita la evolución biológica como estrategia para resolver problemas de optimización. En general, los Algoritmos Genéticos (AG) son parte de la inteligencia artificial; es decir, la resolución de problemas mediante el uso de programas de computación que imitan el funcionamiento de la inteligencia natural.

Los AG fueron introducidos por un par de científicos norte-americanos, John Holland [1929–] en los años 1970, y presentados en 1989 por David Goldberg [1953–] como un método de

optimización de búsqueda global, debido a que este tipo de métodos explora todo el espacio de soluciones del problema permitiendo salir de posibles óptimos locales e ir en busca de óptimos globales. Para entender lo que es optimización hay que considerar que la programación matemática intenta resolver procesos que tienen diferentes posibles soluciones, pero sólo una de ellas corresponde al óptimo global, es decir la solución que se ajusta mejor a las condiciones del mismo problema. Cualquier otra solución que se parezca al óptimo global es un óptimo local.

El procedimiento funciona haciendo evolucionar una población de individuos, o conjunto de posibles soluciones del problema, sometiéndola a acciones aleatorias semejantes a las que actúan en la evolución biológica, tales como mutaciones y recombinaciones genéticas ; así como también a una selección de acuerdo con algún criterio, en función del cual se decide cuáles son los individuos más adaptados, que sobreviven, y cuáles los menos aptos, que son descartados.

En resumen, un AG consiste de los siguientes pasos:

- Inicialización: se genera aleatoriamente una población inicial constituida por posibles soluciones del problema, también llamados individuos (cromosomas).
- Evaluación: aplicación de la función de evaluación a cada uno de los individuos.
- Evolución: uso de operadores genéticos (como son selección, reproducción y mutación), esto tiene como fin mejorar la población de soluciones mediante la aplicación repetitiva de las operaciones de cruzamiento, mutación y selección.
- Término: el AG deberá detenerse cuando se alcance la solución óptima, pero ésta generalmente se desconoce, por lo que se utilizan varios criterios de detención.

El cruzamiento, o reproducción, es sinónimo de apareamiento entre dos individuos de diferente sexo. Así, en un AG se debe escoger a una pareja de individuos que cruzarán sus cromosomas para generar dos descendientes donde se combinan las características de ambos cromosomas padres. La selección de los individuos que funcionarán como padres puede ser aleatoria o permitiendo que al menos uno de los padres pertenezca al grupo de mejores individuos, tal como se lleva a cabo en la naturaleza donde los cromosomas del individuo más fuerte o mejor adaptado son transferidos a su descendencia. El intercambio de la información genética del par de individuos también se puede llevar a cabo de diferentes formas. Una de ellas se ilustra en la Figura 2.9, donde aleatoriamente se ha seleccionado un punto de corte común a ambos padres, y que sirve como referencia para intercambiar su información genética para producir dos hijos con características diferentes a los padres, aunque éstos hereden parte de su información genética.

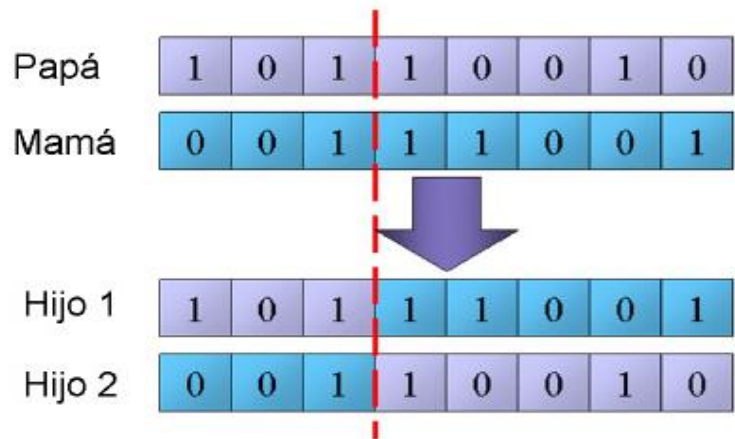


Figura 2.9 Intercambio de información genética entre dos individuos[66].

En la evolución una mutación es un suceso bastante poco común, sucede aproximadamente una en cada mil apareamientos. En la mayoría de los casos las mutaciones son letales o no tienen sentido, pero en promedio, contribuyen a la diversidad genética. En un AG tendrán el mismo papel y la misma frecuencia. Una vez establecida la frecuencia de mutación (muy baja), se genera un número entre 0 y 1 de manera aleatoria y si ese número es menor que la frecuencia de mutación se permite que un gen del cromosoma cambie su información; si no, se dejará como está. La mutación modifica al azar parte del cromosoma de los individuos (ver Figura 2.10), y permite alcanzar zonas del espacio de búsqueda que no estaban cubiertas por los individuos de la población actual.

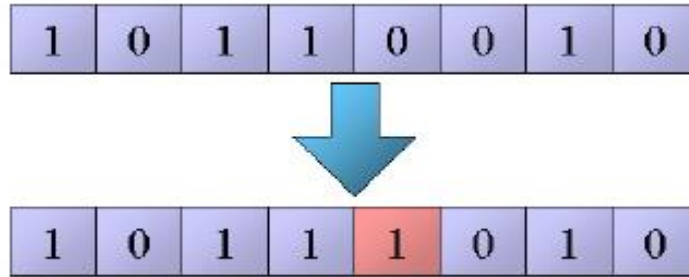


Figura 2.10 Una mutación modifica al azar parte del cromosoma de los individuos [66].

Finalmente, una vez aplicados los operadores genéticos, se seleccionan los mejores individuos para conformar la población de la generación siguiente. Este proceso se realiza por medio de la evaluación de cada individuo con la función de aptitud y se reemplaza la población original. El AG se deberá detener cuando se alcance la solución óptima, por lo general ésta se desconoce, así que se deben utilizar otros criterios de detención. Normalmente se usan dos criterios: 1) correr el AG un número máximo de iteraciones (generaciones), y 2) detenerlo cuando no haya cambios en la población. Mientras no se cumpla la condición de término se repite el ciclo: Selección (Se) → Cruzamiento (Cr) → Mutación (Mu) → Evaluación ($f(x)$) → Reemplazo (Re). Ver Figura 2.11, donde (?) es la condición de término y x^* es la mejor solución [66].

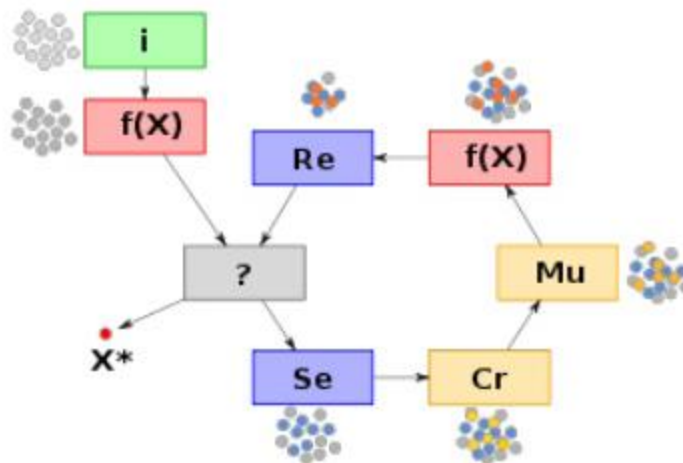


Figura 2.11 Ciclo del algoritmo genético: Selección (Se) → Cruzamiento (Cr) → Mutación (Mu) → Evaluación ($f(x)$) → Reemplazo (Re). El símbolo ? es la condición de término y x^* es la mejor solución [66].

En resumen, el procedimiento funciona haciendo evolucionar conjuntos de variables (cromosomas) que se ajustan a determinados criterios a partir de una población aleatoria inicial

mediante ciclos de replicación diferencial, recombinando y mutando los cromosomas más aptos. Aunque su aplicación ha estado razonablemente extendida, sólo se hizo muy popular cuando se dispuso de ordenadores suficientemente potentes. A continuación, se describe paso a paso del procedimiento en el contexto de un problema de clasificación:

- Etapa 1: inicialmente se crea una serie de conjuntos de variables aleatorias (cromosomas) formando una población de cromosomas (nicho).
- Etapa 2: se evalúa la capacidad de cada cromosoma de la población para predecir la pertenencia a un grupo de cada muestra del conjunto de datos, a esto se le llama función de aptitud. Para ello se entrena un modelo estadístico donde el AG comprueba la precisión de la predicción y asigna una puntuación a cada cromosoma que es igual a la exactitud resultante de la función de aptitud.
- Etapa 3: cuando un cromosoma tiene una puntuación superior a un valor predefinido (fitness), este se selecciona y el procedimiento se detiene; en caso contrario, el procedimiento continúa en la etapa 4.
- Etapa 4: la población de cromosomas se replica. Los cromosomas con una mayor puntuación de fitness generarán una descendencia más numerosa.
- Etapa 5: la información genética contenida en los cromosomas parentales replicados se combinan mediante un cruce genético. Dos cromosomas parentales seleccionados al azar se utilizan para crear dos nuevos cromosomas. Este mecanismo de cruce permite explorar mejor las posibles soluciones recombinando los cromosomas buenos.
- Etapa 6: a continuación, se introducen mutaciones en el cromosoma de forma aleatoria. Estas mutaciones producen nuevos genes que serán tomados en cuenta en los cromosomas.

El proceso se repite desde la etapa 2 hasta obtener un cromosoma preciso. El ciclo de replicación (etapa 4), cruce genético (etapa 5) y mutaciones (etapa 6) se denomina generación [67]. En la Figura 2.12 se muestra una representación esquemática del procedimiento de algoritmos genéticos.

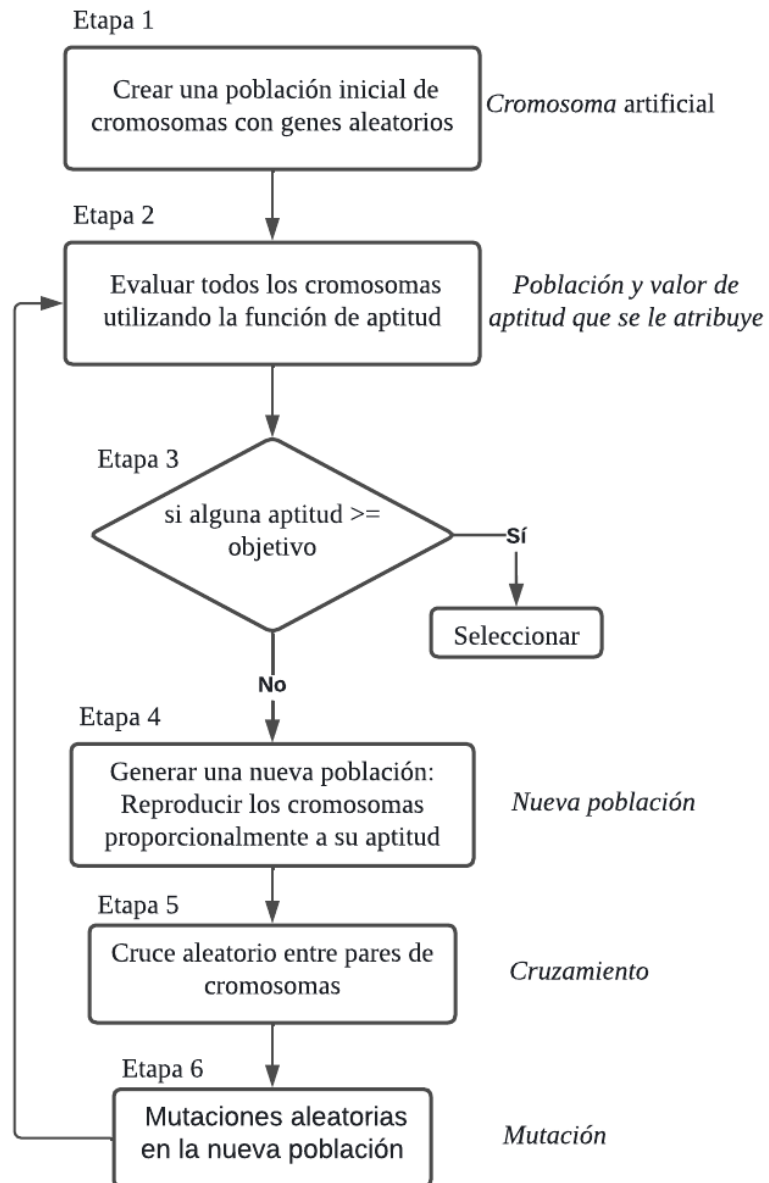


Figura 2.12 Representación esquemática del procedimiento de AG [67].

2.7. Principales estudios relacionados

La metodología propuesta incluye varios pasos, desde la selección de características hasta la construcción del modelo con algunas técnicas vistas anteriormente. A continuación, se presentan principales estudios relacionados con esta investigación, los cuales aplican la metodología de Machine Learning (sección 2.5.2).

1. Daniel González Cedillo, (2019). Diagnóstico de enfermedades cardíacas con los algoritmos supervisados naive bayes.
2. Chicco, D., & Jurman, G. (2020). El aprendizaje automático puede predecir la supervivencia de los pacientes con insuficiencia cardíaca a partir de la creatinina sérica y la fracción de eyección.
3. Ali, F., El-Sappagh, S., Islam, S. M. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020). Un sistema de monitorización sanitaria inteligente para la predicción de Enfermedades Cardíacas basado en el aprendizaje profundo conjunto y la fusión de características.
4. Gallego Valcárcel David, & Lucas Monsalve Delly Fabián. (2021). Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por Insuficiencia Cardíaca con datos clínicos.
5. Faiyaz Waris, S., & Koteeswaran, S. (2021). Predicción precoz de enfermedades cardíacas mediante un novedoso método de aprendizaje automático denominado clasificador de vecinos KNN mejorado en python.

En cada uno de los trabajos mencionados anteriormente se toman en cuenta temas como técnicas de inteligencia artificial, predicción de Enfermedades Cardíacas, selección de características y algoritmos de machine learning, los cuales son palabras clave para el desarrollo de la presente investigación.

El primero de Daniel González Cedillo en 2019 [14] desarrolló un sistema predictivo para determinar pacientes que puedan ser propensos a sufrir alguna EC, a través del uso de naive bayes.

El trabajo de Chicco, D., & Jurman en 2020, en cuya investigación realizan una comparativa de diez clasificadores de aprendizaje automático (random forest, gradient boosting, decision tree, regresión lineal, naive bayes, SVM, redes neuronales y KNN, por mencionar algunos) para predecir la supervivencia de pacientes con Insuficiencia Cardíaca y categorizar las características clínicas correspondientes a los factores de riesgo más importantes.

En la investigación de Ali, F., El-Sappagh, S., Islam, S. M. R., Kwak, D., Ali, A., Imran, M., & Kwak, K. S. (2020), proponen un sistema sanitario inteligente para la predicción de Enfermedades Cardíacas utilizando enfoques de fusión de características y aprendizaje profundo. Llevaron a cabo una comparativa del modelo propuesto, contra seis clasificadores existentes (naive bayes, redes neuronales, decision tree, random forest, regresión logística y SVM).

En el trabajo de Gallego Valcárcel David, & Lucas Monsalve Delly Fabián, 2021, implementaron modelos de clasificación utilizando técnicas de aprendizaje automático (redes neuronales, máquina de soporte de vectores y random forest), para predecir el riesgo de fallecer por Insuficiencia Cardíaca a partir de datos clínicos de pacientes recopilados. Apoyándose de

técnicas de reducción de variables (análisis de componentes principales y eliminación hacia atrás), para la selección de características.

El trabajo de Faiyaz Waris, S., & Koteeswaran, S. (2021), proponen mejorar el algoritmo de K vecinos cercanos (KNN, por sus siglas en inglés) y, con ello, corroborar que es más preciso que el KNN normal, en predicción temprana de Enfermedades Cardíacas.

En la siguiente sección se muestra el desempeño de los algoritmos utilizados en cada uno de los trabajos antes mencionados, así como sus contribuciones y limitaciones.

2.8. Contribuciones y limitaciones de estudios previos

En la Tabla 2.2, se muestra el desempeño obtenido en los algoritmos implementados, así mismo contribuciones y limitaciones de los estudios relacionados.

Tabla 2.2 Principales estudios relacionados (elaboración propia).

Trabajo	Contribución	Técnica utilizada	Exactitud	Limitación
Diagnóstico de enfermedades cardíacas con los algoritmos supervisados Naives Bayesian, [14]	Modelo predictivo para determinar pacientes que puedan ser propensos a sufrir enfermedades cardíacas, a través del uso de Naive Bayes, utilizando 11 características: Edad, sexo, tipo de dolor de pecho, presión arterial en reposo, colesterol sérico, glucemia en ayunas, resultados de electrocardiograma en reposo, máximo ritmo cardíaco alcanzado, angina inducida por ejercicio, depresión inducida por el ejercicio relativo al descanso y tipo de pendiente del pico de ejercicio.	Naive bayes	86.81%	Pequeño conjunto de datos. No hay selección de variables, debido a que analizan las mismas que en otras investigaciones.
Machine learning can predict survival of patients with heart failure	Comparación de diferentes clasificadores para predecir pacientes propensos a sufrir alguna EC. Dando 11 características más significativas:	Random forest	74%	Analiza un pequeño conjunto de datos (299 pacientes). Falta añadir información
		Gradient boosting	73.8%	
		Decision tree	73.7%	

from serum creatinine and ejection fraction alone, [15]	Creatinina sérica, fracción de eyección, edad, sodio sérico, hipertensión arterial, anemia, plaquetas, creatinina fosfocinasa, tabaquismo, sexo y diabetes.	Regresión lineal	73%	adicional sobre las características físicas de los pacientes (altura, peso, índice de masa corporal, entre otros) y sus antecedentes laborales para detectar otros factores de riesgo de enfermedades cardiovasculares
		Naive bayes	69.6%	
		SVM	68.4%	
		Redes neuronales	68%	
		KNN	62.4%	
A Smart Healthcare Monitoring System for Heart Disease Prediction Based On Ensemble Deep Learning and Feature Fusion, [16]	Sistema inteligente para la predicción de Enfermedades Cardíacas, comparando del modelo propuesto, contra clasificadores existentes. Utilizando los 15 atributos: Edad, sexo, tipo de dolor torácico, presión arterial, colesterol, fumador, nivel de azúcar en sangre, antecedentes familiares, lectura de electrocardiograma, frecuencia cardíaca máxima, angina inducida por ejercicio, pico de depresión inducido por ejercicio, tipo de pendiente del pico de ejercicio, número de vasos mayores (0-3) coloreados por fluoroscopia y prueba de esfuerzo del corazón en minutos.	Modelo propuesto	83.5%	Cantidad de variables analizadas
		Naive bayes	80.4%	
		Redes neuronales	77.6%	
		Decision tree	74.8%	
		Random forest	73.7%	
		Regresión logística	73.7%	
		SVM	71.8%	
Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por insuficiencia	Implementación de tres modelos de aprendizaje automático para predecir el riesgo de fallecimiento por insuficiencia cardíaca, a partir de 12 características: Sexo, fumador, diabetes, presión sanguínea, anemia,	SVM	82.29%	Análisis de un conjunto de datos muy pequeño (299 registros) y cantidad de variables utilizadas
		Redes neuronales	82.28%	
		Random forest	81.28%	

cardíaca con datos clínicos, [17]	edad, tiempo (días en los que el paciente fue diagnosticado), fracción de eyección, sodio, creatinina sérica, plaquetas y creatinfosfoquinasa (CPK).			
Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python, [18]	Mejorar el algoritmo de KNN, y corroborar que es más preciso que el KNN normal, en predicción temprana de Enfermedades Cardíacas. Utilizando 13 características:	KNN modificado	93%	Cantidad de variables analizadas
	Edad, sexo, tipo de dolor torácico, presión arterial, colesterol, glucemia en ayunas, resultados de electrocardiograma, frecuencia cardíaca, angina inducida por ejercicio, pico de depresión inducido por ejercicio, tipo de pendiente del pico de ejercicio, resultado de fluoroscopia, resultado de prueba estrés de thalium.	KNN normal	88%	

2.9. Comparación entre los trabajos relacionados y la propuesta de investigación

En la Tabla 2.3 se presenta la comparación entre los trabajos relacionados y la propuesta de investigación.

Tabla 2.3 Comparación de trabajos relacionados y la propuesta de investigación (elaboración propia).

TRABAJO	MODELOS DE PREDICCIÓN DE EC	CONJUNTO DE DATOS MAYOR A 1000 INSTANCIAS	MÉTODO DE SELECCIÓN DE CARACTERÍSTICAS INTELIGENTE	VARIABLES ANALIZADAS	SEGMENTACIÓN POR EDAD, SEXO, DIABETES	VALIDACIÓN CRUZADA
Biomarcador para la predicción de Enfermedades Cardíacas	✓	✓	✓	11	✓	✓
Diagnóstico de enfermedades cardíacas con los algoritmos supervisados Naives Bayesian	✓	✗	✗	11	✗	✗
Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone	✓	✗	✓	11	✗	✗
A Smart Healthcare Monitoring System for Heart Disease Prediction Based On Ensemble Deep Learning and Feature Fusion	✓	✗	✗	15	✗	✗
Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por insuficiencia cardíaca con datos clínicos	✓	✗	✗	12	✗	✓
Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python	✓	✗	✗	13	✗	✓

Como se puede observar, la mayoría de los trabajos relacionados utilizan una cantidad menor a 1000 instancias, no todos implementan un proceso de selección de características inteligente, ni validación cruzada, por lo que se requiere un análisis más profundo. Es por ello, que en esta investigación se plantea trabajar con un conjunto de datos de mayor dimensión, considerando información de personas que radican en Estados Unidos, dado que la mayoría de sus pacientes tiene etnicidad similar a la de los mexicanos por lo que nos permitiría obtener conocimiento más amplio en la detección de enfermedades, además de que la infraestructura y políticas de aquel país permiten tener acceso a este tipo de datos de manera abierta para la comunidad científica y académica.

De acuerdo a lo anterior, se optó por realizar un primer acercamiento a la predicción de Enfermedades Cardíacas con datos ya existentes en la comunidad médica de ese país, lo suficientemente amplio para generar un modelo preciso y eficaz y poder ser implementado en un contexto como el de México. Explorando la selección de las variables más significativas, por medio de algoritmos genéticos, para después generar modelos predictivos con regresión logística, random forest y K vecinos más cercanos (KNN), realizar una evaluación mediante validación cruzada evitando un sesgo en la información, además, efectuar una comparativa de dichos modelos y determinar cuál es el mejor modelo de predicción.

2.10. Propuesta de investigación

Debido a la necesidad de la comunidad médica y científica de detectar de manera temprana las Enfermedades Cardíacas, se propone realizar biomarcador con ayuda de técnicas de inteligencia artificial para determinar las variables más significativas en la predicción de esta patología, estableciendo un modelo de clasificación a partir de un conjunto de datos antropométricos en lugar de datos clínicos (comúnmente considerados cuando el paciente ya presenta síntomas de alguna enfermedad), para lograr un biomarcador preventivo que contribuya en el retardo de aparición de algún padecimiento cardíaco.

2.11. Modelo o esquema general de investigación

La investigación científica sigue un proceso sistemático para recolectar evidencias que después se analizan para responder las preguntas del estudio y generar nuevos conocimientos científicos [68]. El proceso general de investigación que se llevará a cabo en este trabajo está basado en el método científico de Hernández Sampieri (2017), consta de 11 fases como se muestra en la Figura 2.13.



Figura 2.13 Proceso General de Investigación (elaboración propia) [68].

Capítulo 3. Método y propuesta de investigación

3.1. Modelo de investigación

Derivado del procedimiento para la creación de un modelo de Machine Learning, se generó un modelo de investigación que busca determinar cuáles son los factores de riesgo para clasificar pacientes propensos a sufrir una Enfermedad Cardíaca. El proceso general de investigación propuesto se muestra en la Figura 3.1. En primera instancia, se obtiene el conjunto de datos que será analizado, posteriormente se realiza un procesamiento de dicha información para después dar paso a la extracción de características mediante métodos de selección, en cuarta instancia se procede a la construcción del modelo de clasificación utilizando algoritmos de aprendizaje supervisado, con ello llevar a cabo las respectivas predicciones y conocer su desempeño. Finalmente, por medio de validación cruzada se efectuará la evaluación del modelo evitando un sobreajuste de los datos.

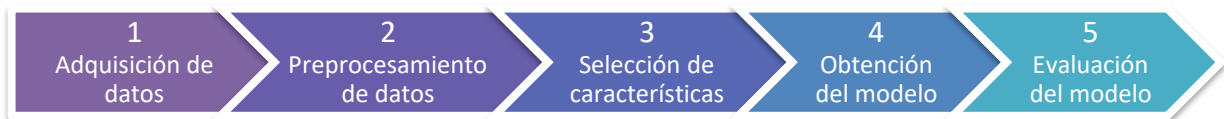


Figura 3.1 Modelo de investigación implementado (elaboración propia).

3.2. Descripción de la propuesta

A continuación, se describe cada etapa del modelo de investigación propuesto en la sección anterior (3.1).

3.2.1. Adquisición de datos

Los datos que se utilizaron para este estudio, fueron obtenidos de la plataforma Kaggle [21] en su repositorio de data sets, es un conjunto de datos abiertos que contiene 253,680 observaciones y 22 características entre clínicas y generales (como se puede observar en la Tabla 3.1), la mayoría son de tipo categórica y permitirán generar modelos predictivos para Enfermedades Cardíacas.

Tabla 3.1 Características del conjunto de datos de indicadores de salud de Enfermedades Cardíacas (elaboración propia).

HeartDiseaseorAttack	HvyAlcoholConsump
HighBP	AnyHealthcare

HighChol	NoDocbcCost
CholCheck	GenHlth
BMI	MentHlth
Smoker	PhysHlth
Stroke	DiffWalk
Diabetes	Sex
PhysActivity	Age
Fruits	Education
Veggies	Income

3.2.2. Preprocesamiento de datos

El conjunto de datos cuenta con un total de 253,680 pacientes, de los cuales 141,974 son hombres (56%) y 111,706 son mujeres (44%) en un rango de edad de 18 a 80 años en adelante; siendo más sobresaliente el grupo de 60 a 64 años en ambos sexos con el 13% para cada uno. Se encuentra el 91% de personas sanas (229,787) y 9% de pacientes con alguna EC (23,893), siendo el grupo de mujeres enfermas con mayor prevalencia con 11% (13,688) de los casos y 9% (10,205) en hombres, aproximadamente.

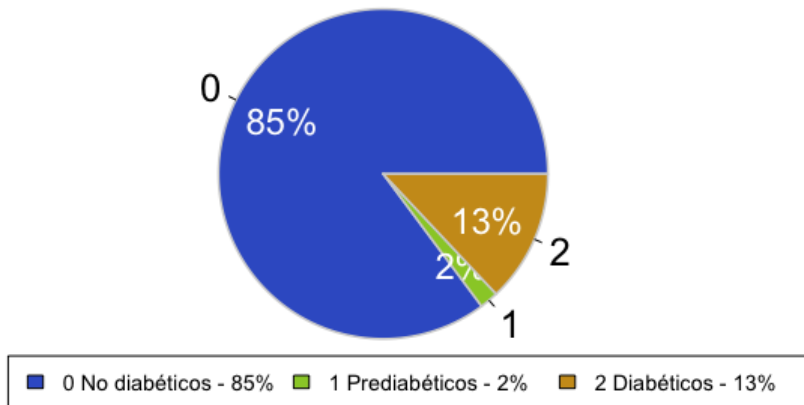
En un primer análisis demográfico, es más notable la cantidad de pacientes del sexo femenino que presentan un diagnóstico positivo de Enfermedad Cardíaca, encima de hombres (ver Tabla 3.2), en un rango de edad entre 60 a 64 años. Cuentan con un nivel de educación de grado universitario y estadísticamente, la moda como ingreso anual promedio es 75,000 dólares (aproximadamente) por grupo (hombres y mujeres).

Tabla 3.2 Datos demográficos del conjunto de datos (elaboración propia).

	Sector poblacional		Enfermos	Sanos
Hombres	Jóvenes	6,736	10,205	131,769
	Adultos	66,125		
	Adultos mayores	69,113		
Mujeres	Jóvenes	6,562	13,688	98,018
	Adultos	51,943		
	Adultos mayores	53,201		

Además, se muestra un alto índice de pacientes no diabéticos (120,959 hombres y 92,744 mujeres), seguido de diabéticos con un promedio entre el 13 y 15% de la población total y por último las personas prediabéticas con el 2% (ver Figura 3.2).

Hombres



Mujeres

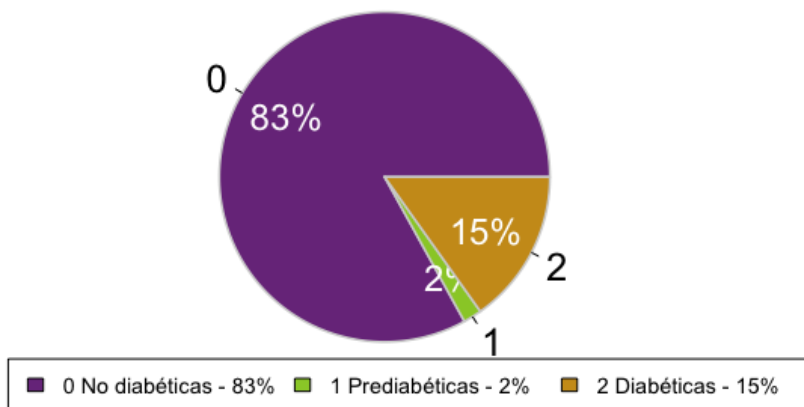


Figura 3.2 Estadística de pacientes Diabéticos (elaboración propia).

En esta investigación se explorarán varias opciones: la primera es diabetes en la cual se quiere conocer las características que permiten clasificar a pacientes no diabéticos, prediabéticos y diabéticos debido a las condiciones multifactoriales que presenta esta enfermedad, ya que actualmente tiene un alto índice de prevalencia en México. La segunda opción es descubrir la diferencia de factores de riesgo presentes entre grupos de edad (jóvenes entre 18 a 29 años, adultos de 30 a 59 años y adultos mayores de 60), ya que no es lo mismo que un joven haga ejercicio determinado a que un adulto mayor repita la misma rutina, debido a que este último tiene un rango de movilidad reducido en comparación con una persona de 20 años. Como tercera opción es por sexo biológico, teniendo en cuenta que cada uno tiene diferentes hábitos tanto saludables como de ejercicio, añadiendo principalmente las diferencias biológicas, es por ello que se desea conocer las características que están presentes en cada sexo. Derivado de lo anterior se pretende contribuir en el diagnóstico oportuno de las Enfermedades Cardíacas, así como en el tratamiento personalizado acorde a la fisiología que presenta cada paciente.

Por lo tanto, la muestra será probabilística estratificada ya que se pretende comparar los resultados entre segmentos de la población. Los grupos se dividirán de la siguiente manera como se muestra en la Tabla 3.3:

Tabla 3.3 Experimentos aplicados en la investigación (elaboración propia).

	Subconjunto	Descripción
Experimento 1	Grupo 1: no diabéticos (213,703)	En este experimento se pretende conocer si hay diferencia entre las Enfermedades Cardiacas respecto a las características que presentan en personas sanas, prediabéticas y diabéticas.
	Grupo 2: prediabéticos (4,631)	
	Grupo 3: diabéticos (35,346)	
Experimento 2	Grupo 4: pacientes jóvenes entre 18 y 29 años (13,298)	En este experimento se quiere conocer cuáles características para predecir Enfermedades Cardiacas se mantienen a lo largo de la vida de los pacientes sin tomar en cuenta otros factores.
	Grupo 5: pacientes adultos entre 30 y 59 años (118,068)	
	Grupo 6: pacientes adultos mayores de 60 años (122,314)	
Experimento 3	Grupo 7: hombres (229,787)	Para este experimento es de gran interés si las características son iguales para hombres y mujeres.
	Grupo 8: mujeres (23,893)	

3.2.3. Selección de características

Con el objetivo de seleccionar las características más significativas para la predicción de Enfermedades Cardiacas, como se mostró en la sección 3.2.1 el conjunto de datos cuenta con 22 características, implica una posible combinación de 22 factorial, es decir, habría que generar millones de modelos diferentes para poder encontrar cuáles son los mejores, con el fin de evitar esto, se sugiere el uso de algoritmos genéticos ya que son una técnica de programación inspirada en la reproducción de los seres vivos imitando la evolución biológica. Son parte de la inteligencia artificial, siendo un método de optimización de búsqueda global.

Para esta investigación se propone la implementación del algoritmo genético Galgo [67], el cual funciona haciendo evolucionar una población de individuos, o conjunto de soluciones posibles del problema, sometiéndola a acciones aleatorias semejantes a las que actúan en la evolución biológica, tales como mutaciones y recombinaciones genéticas; así como también a una selección de acuerdo con algún criterio, en función del cual se decide cuáles son los mejores individuos que sobreviven, y los menos aptos son descartados [66].

3.2.4. Obtención del modelo de predicción

Después de obtener las características más significativas, se procede a la creación de modelos de predicción, implementando los algoritmos de aprendizaje supervisado vistos en la sección 2.6 para la creación de modelos de predicción y con ayuda de algunas métricas de desempeño como área bajo la curva y matriz de confusión (vistas en la sección 2.6.4) se obtendrá el nivel de rendimiento de cada modelo.

3.2.5. Evaluación del modelo

Por medio de validación cruzada (vista en la sección 2.6.4), la cual se efectúa particionando el conjunto de datos original en subconjuntos aleatorios conformado por el 70% para entrenamiento y 30% para pruebas, realizando este proceso 3 veces ($k=3$), es decir 3 particiones diferentes, con la intención de evitar un sesgo en las predicciones.

Capítulo 4. Resultados

En este capítulo se presentan los resultados ordenados por los casos de estudios que se establecieron en los objetivos de la sección 1.5.1, implementando la metodología vista en el capítulo anterior en cada uno de los tres casos de estudio que se presentan a continuación.

En primer caso se desea clasificar pacientes no diabéticos, prediabéticos y diabéticos debido a la disposición multifactorial que tiene esta enfermedad y el alza de mortalidad en nuestro país.

En el segundo, se pretende conocer las causas de riesgo presentes en cada sector poblacional (jóvenes, adultos y adultos mayores), en vista de que cada uno tiene diversas características metabólicas.

En el tercer caso, se busca saber cuáles características clínicas y no clínicas derivan al desarrollo de padecer alguna complicación cardíaca a causa de que existen diferencias biológicas en cada sexo.

El propósito principal de los casos de estudio es conocer diferencias (en condiciones multifactoriales de diabetes, factores de riesgo presentes en grupos de edad y biológicas por sexo) que hay entre cada uno, con ello propiciar al desarrollo de un biomarcador para la predicción de Enfermedades Cardíacas y así contribuir en el diagnóstico oportuno de estas patologías.

A continuación, se presentarán los resultados obtenidos en los distintos casos de estudio efectuados en esta investigación.

4.1. Caso de estudio 1: Diabetes

En el capítulo 3 se propuso una metodología, la cual, consta de cinco etapas que se aplicaron a lo largo del primer caso de estudio de diabetes (4.1 a 4.1.5).

4.1.1. Adquisición de datos

Para este caso en particular, se utilizó el conjunto de datos visto en la sección 3.2.1 de 253,680 observaciones y 22 características, de las cuales tres de ellas se excluyeron de los experimentos (AnyHealthcare, NoDocbcCost, Stroke) debido a que eran variables poco significativas y provocaría un sesgo en las predicciones, por lo tanto, se tomaron en cuenta 19 (ver Tabla 4.1), incluyendo la característica clasificatoria de Enfermedades Cardíacas. Cabe mencionar, que para cada caso de estudio varía el número de observaciones analizadas por subconjunto.

Tabla 4.1 Características utilizadas del conjunto de datos original (elaboración propia).

HeartDiseaseorAttack	HvyAlcoholConsump	PhysActivity
HighBP	GenHlth	Fruits
HighChol	MentHlth	Veggies
CholCheck	PhysHlth	Education
BMI	DiffWalk	Income

Smoker	Sex	
Diabetes	Age	

4.1.2. Preprocesamiento de datos

Se emplearon tres subconjuntos del experimento 1 (visto en la Tabla 3.3): No diabéticos (213,703), Prediabéticos (4,631) y Diabéticos (35,346).

4.1.3. Selección de características

El proceso de selección de características se realizó mediante la librería Galgo, para el cuál se utilizaron 500 soluciones máximas, también conocidas como cromosomas, conteniendo 5 genes (chromosomeSize=5) que corresponden a modelos desarrollados, utilizando un clasificador del centroide más cercano (classification.method="nearcent") con una precisión de clasificación del 100% (goalFitness=1), la cual también se le conoce como función objetivo, ya que representa el valor de precisión que desea obtener el cromosoma, siendo este el criterio de detención. Para la selección de características se empleó el método selección hacia adelante, el cual es iterativo, comenzando con un modelo en la que no se tiene ninguna variable y en cada iteración se va añadiendo una variable hasta que la adición de nuevas características no mejore el rendimiento del modelo.

Los algoritmos genéticos obtuvieron entre cuatro y once características más significativas para cada uno de los grupos (ver Tabla 4.2).

Tabla 4.2 Características más significativas y porcentaje de desempeño en grupos de diabetes (elaboración propia).

SUBCONJUNTO	CARACTERÍSTICAS MÁS SIGNIFICATIVAS
No Diabéticos	Dificultad para caminar , control de colesterol, consumo excesivo de alcohol y sexo
Prediabéticos	Salud física , edad , frutas y salud mental
Diabéticos	Dificultad para caminar , colesterol alto, sexo , salud física , salud en general, consumo excesivo de alcohol , edad , presión arterial alta, IMC, salud mental y fumador

El eje horizontal en la Figura 4.1 muestra los genes ordenados por rango, y el eje vertical muestra frecuencia de aparición de las características y el rango esta codificado por colores en cada una de las evoluciones anteriores, donde "DiffWalk", "HighChol" y "Sex" (aparecen en color negro) para el subconjunto de Diabéticos representan la frecuencia más alta y las variables en gris

presentan la más baja. En la sección de Anexos (A y B) se pueden consultar las frecuencias de los subconjuntos faltantes analizados en este caso de estudio (no diabéticos y prediabéticos).

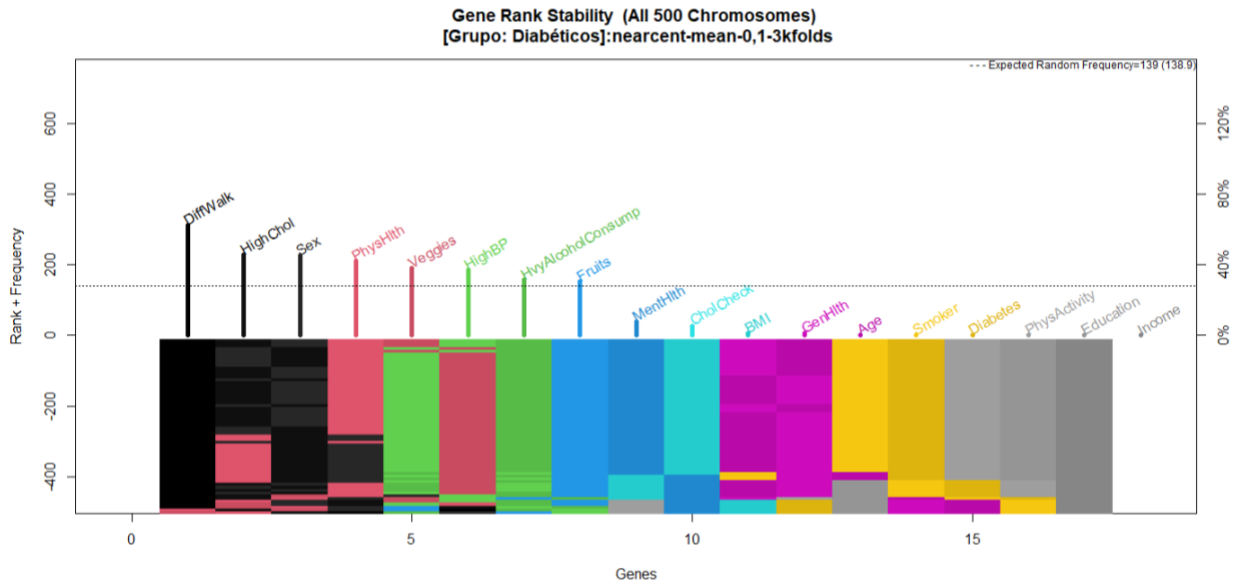


Figura 4.1 Frecuencia de aparición de las características más importantes obtenidas por Galgo en el grupo de Diabéticos (elaboración propia).

El eje horizontal de la Figura 4.2 representa las variables ordenadas por su rango de significancia, mientras que el eje vertical muestra la precisión de clasificación. La línea con puntos negros simboliza el promedio general de precisión (0.6683) para el caso de pacientes Diabéticos cuyo valor más alto es 1. Las líneas discontinuas en color negro y rojo representan la precisión por clase (personas sanas y enfermas respectivamente). La línea verde especifica el promedio de ambas clases. Del método de selección hacia adelante se obtuvieron 3 modelos de los cuales el último fue el mejor (línea negra continua) contiene catorce características más significativas: “DiffWalk”, “HighChol”, “Sex”, “PhysHlth”, “Veggies”, “HighBP”, “HvyAlcoholConsump”, “Fruits”, “MentHlth”, “CholCheck”, “BMI”, “GenHlth”, “Age” y “Smoker”. En los Anexos F y G se muestran los resultados de este proceso en pacientes no diabéticos y prediabéticos.

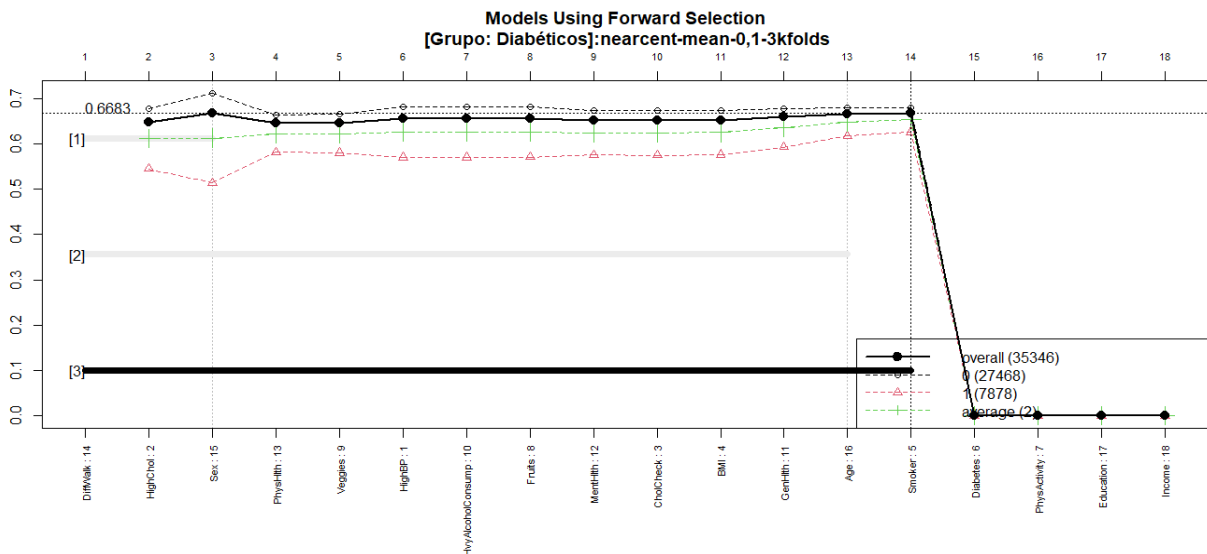


Figura 4.2 Selección hacia adelante en el grupo de Diabéticos (elaboración propia).

4.1.4. Obtención del modelo de predicción

Después de obtener las características más significativas, se procede a generar los modelos de predicción utilizando regresión logística (RL), ya que en primera instancia es un algoritmo sencillo, fácil de entrenar sobre gran cantidad de datos y rara vez existe un sobreajuste en comparación con otros algoritmos. Random forest (RF), se considera un método muy preciso, flexible y fácil de usar. La razón principal es que toma el promedio de todas las predicciones, lo que anula los sesgos [65]. K vecinos más cercanos (KNN) es un algoritmo sencillo y se utiliza en problemas de regresión o clasificación. Los modelos fueron creados en el entorno de software libre R y Rstudio, el cual sirve para realizar cálculos numéricos, estadísticos, entre otros, así como para crear gráficas y figuras de calidad mediante diversas librerías.

A su vez, se emplearon algunas métricas de evaluación como la curva ROC, siendo una herramienta estadística utilizada para clasificar a los individuos de una población en dos grupos (uno que represente un evento de interés y otro que no) de aquí se deriva el área bajo la curva (AUC), la cual es una métrica de precisión estándar para modelos de clasificación binaria, mide la capacidad de predecir eventos positivos en comparación con negativos. Además de los elementos que componen la matriz de confusión como exactitud, sensibilidad y especificidad (vistas en la sección 2.6.4).

4.1.5. Evaluación del modelo

Para determinar la precisión de los modelos de predicción se utilizó validación cruzada (vista en la sección 3.2.5) para evitar algún tipo de sesgo. En la siguiente Figura 4.3 se muestra el promedio de AUC obtenido de la comparativa entre los algoritmos estudiados en el presente trabajo de investigación (vistos en la sección 2.6). Siendo regresión logística el mejor clasificador para los tres grupos de diabetes, con el 73.5% para pacientes Diabéticos. En la sección de Anexos (K y L) se pueden consultar los resultados de los subconjuntos restantes analizados en este caso de estudio.

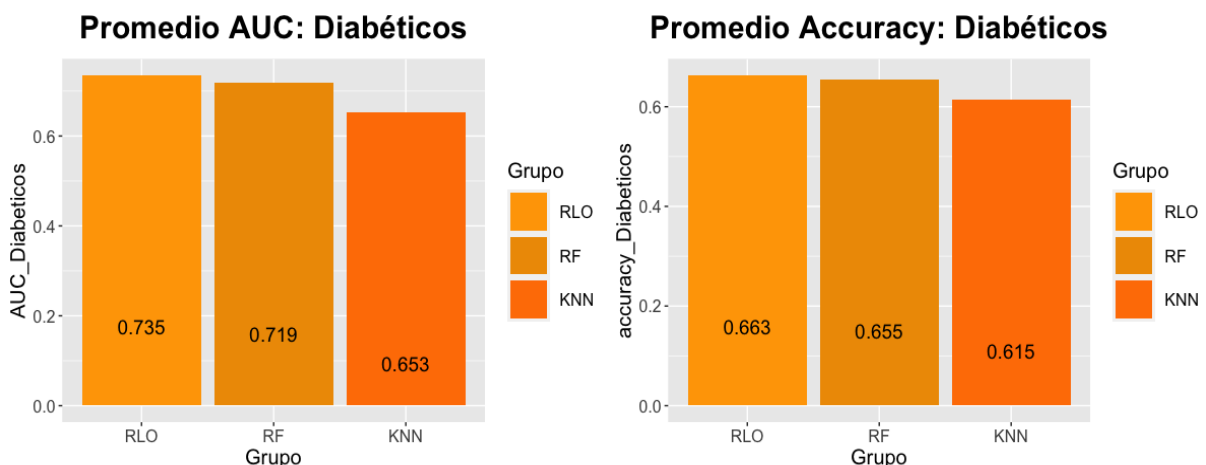


Figura 4.3 Comparativa de modelos de predicción con RLO, RF y KNN en el grupo de Diabéticos (elaboración propia).

Se obtuvo el promedio de las métricas de desempeño de los modelos de predicción, como se muestra en la Figura 4.4. Consiguiendo una media entre el 66 y 61% de accuracy entre los diversos grupos del primer caso de estudio. Así mismo, consiguiendo un AUC del 68.3% para personas

No diabéticas, 69.2% en Prediabéticos y 73.5% para pacientes Diabéticos. Cabe señalar que para este caso se utilizó regresión logística, siendo el algoritmo que arrojó predicciones más óptimas.

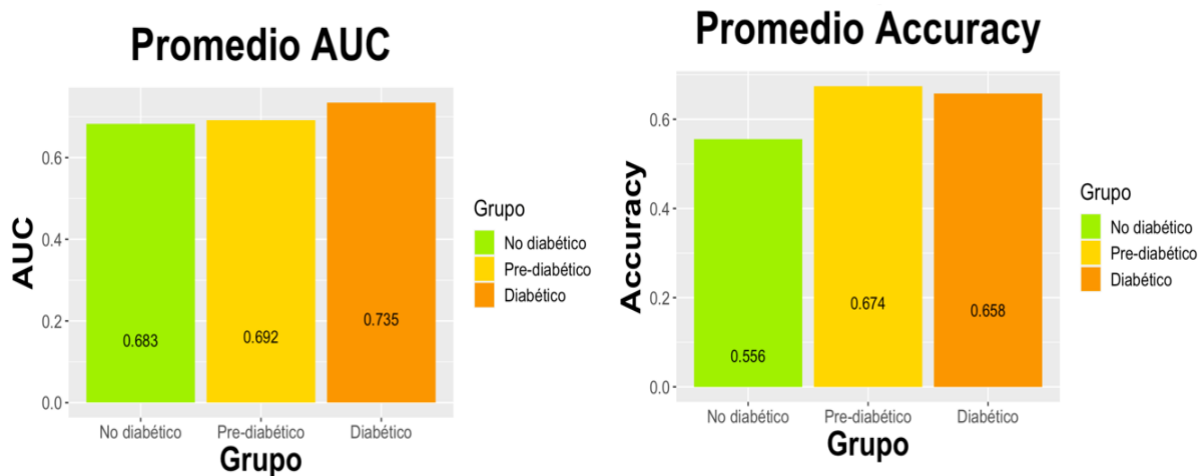


Figura 4.4 Promedios de métricas de desempeño, obtenidas en grupos de diabetes del modelo de RLO (elaboración propia).

4.2. Caso de estudio 2: Grupos de edad

Al igual que en el caso de diabetes, se aplicó la misma metodología de cinco etapas que se efectuaron en el transcurso del segundo caso de estudio (4.2 a 4.2.5).

4.2.1. Adquisición de datos

El siguiente caso de estudio, al igual que el anterior, se utilizó el conjunto original con 253,680 datos y 19 características.

4.2.2. Preprocesamiento de datos

Se realizó una división de tres subconjuntos, pero enfocados en sectores poblacionales como jóvenes (13,298), adultos (118,068) y adultos mayores (122,314), como se observó anteriormente en la Tabla 3.3.

4.2.3. Selección de características

Para el segundo caso de estudio, las variables más significativas obtenidas por el algoritmo genético se muestran a continuación en la Tabla 4.3. Cabe señalar que en los tres casos de estudio se utilizaron los mismos parámetros para la selección de características (visto en la sección 3.2.3).

Tabla 4.3 Características más significativas en grupos de Edad (elaboración propia).

SUBCONJUNTO	CARACTERÍSTICAS MÁS SIGNIFICATIVAS
Jóvenes	Diabetes, control de colesterol, frutas
Adultos	Edad, diabetes
Adultos mayores	Salud física, fumador

En la Figura 4.5 se observa una gráfica donde el eje horizontal representa las variables ordenadas por su nivel de significancia, y el eje vertical muestra la frecuencia de aparición de dichas características siendo “Diabetes”, “Age” y “CholCheck” las más frecuentes (mostrándose en color negro) para el subconjunto de Adultos. La parte inferior donde aparecen los colores representan las evoluciones previas, en donde menos cambio de color quiere decir que la característica es estable y, por consiguiente, muy significativa. En los Anexos C y D, se puede consultar las frecuencias de los subconjuntos faltantes analizados en este caso de estudio (jóvenes y adultos mayores).

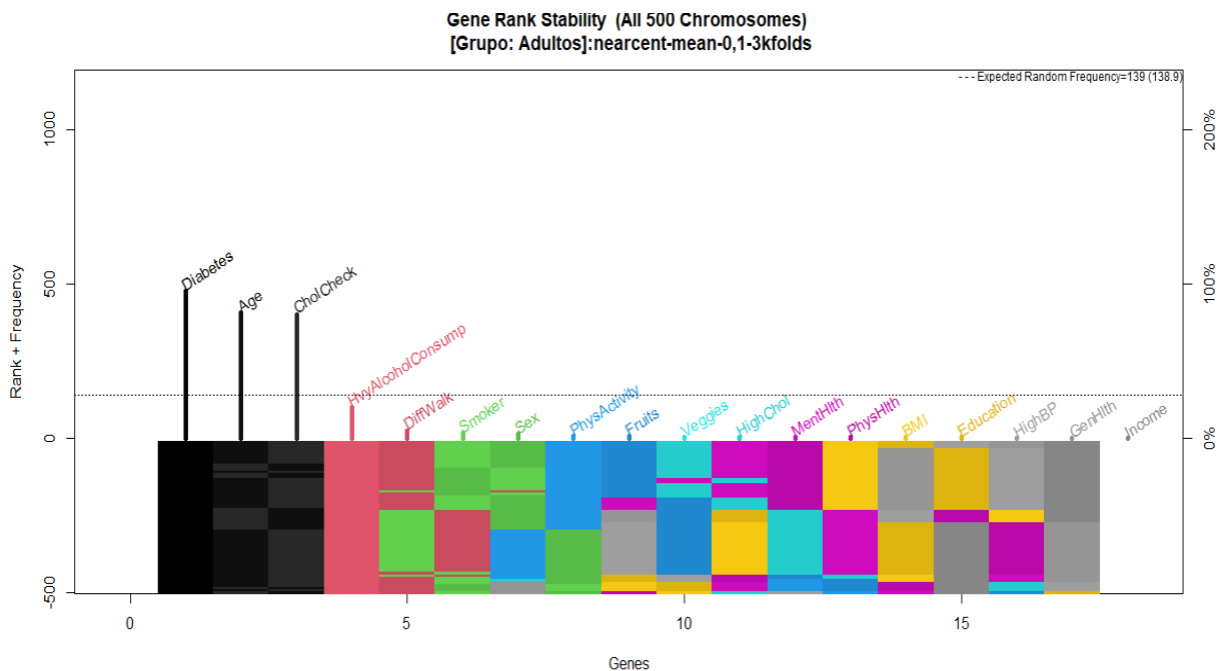


Figura 4.5 Frecuencia de aparición de las características más importantes obtenidas por Galgo en el grupo de Adultos (elaboración propia).

El proceso de selección hacia adelante va añadiendo características al modelo para mejorar su precisión hasta que este ya no mejore. En la Figura 4.6 el eje horizontal simboliza los genes ordenados por su rango de aparición. El eje vertical muestra la exactitud de la clasificación. La línea con puntos negros representa la precisión general (muestras mal clasificadas divididas por

el número total de muestras). Las líneas discontinuas de colores (negro y rojo) interpretan la precisión por clase (personas sanas y enfermas). Para este subconjunto en particular (Adultos) el procedimiento de selección obtuvo tres modelos, el mejor de estos fue el que está etiquetado con el número 1, incluyendo solamente dos características más significativas (“Diabetes”, “Age”), obteniendo así una precisión del 0.8804. Los Anexos H e I muestran los resultados de este proceso en pacientes jóvenes y adultos mayores.

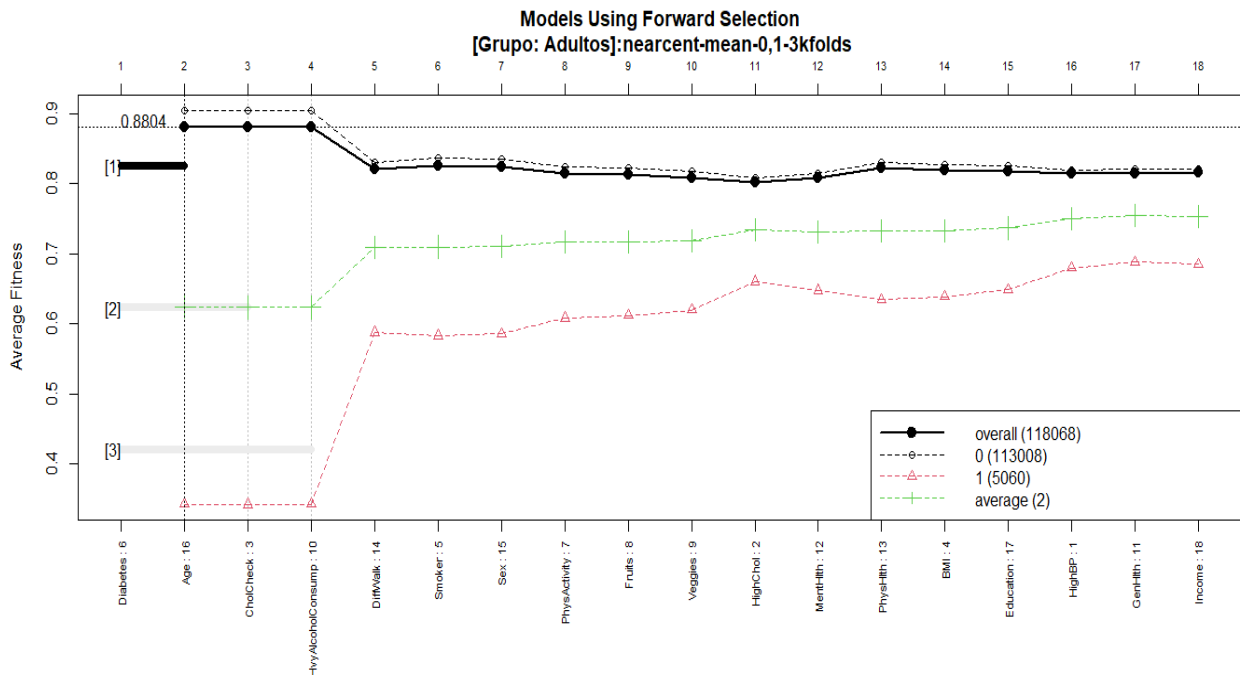


Figura 4.6 Selección hacia adelante en el grupo de Adultos (elaboración propia).

4.2.4. Obtención del modelo de predicción

Al igual que el caso de diabetes, se crearon modelos de predicción mediante los algoritmos de regresión logística, random forest y KNN. Además del uso de las métricas de evaluación aplicadas a cada modelo para después pasar a la etapa de evaluación, donde se obtienen los promedios de cada métrica mediante validación cruzada.

4.2.5. Evaluación del modelo

Un buen modelo de clasificación debe proporcionar predicciones precisas, por lo cual, es necesario aplicar una validación cruzada sobre los datos y de esta manera evitar un sobreajuste. En la Figura 4.7 se muestra una comparativa de los diversos algoritmos utilizados en esta investigación (vistos en el capítulo 2), obteniendo resultados iguales en el caso de AUC en regresión logística y random forest para el grupo de Adultos con casi el 72% de precisión, 58% para pacientes Jóvenes y un 63.9% para Adultos mayores (ver Anexos). KNN no es un método con alto grado de precisión, pero las predicciones son estables. Los porcentajes de accuracy alcanzados van entre 98% y 78% siendo cifras favorables para la precisión de los modelos en el presente caso de estudio. Se pueden consultar los Anexos M y N para consultar los resultados de los grupos restantes, analizados en este caso de estudio.

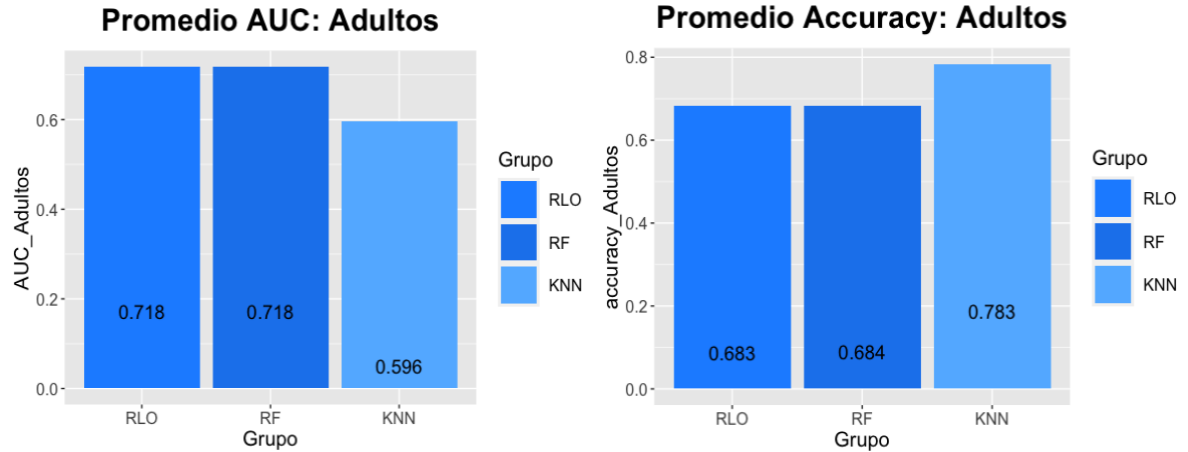


Figura 4.7 Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de Adultos (elaboración propia).

El promedio general de las métricas de desempeño utilizadas se muestra en la Figura 4.8, abordando el algoritmo de random forest, el área bajo la curva que se obtuvo fue 57.4% para Jóvenes, 71.8% en Adultos y 63.9% en Adultos Mayores. El porcentaje de exactitud (accuracy) ronda entre el 98 y 68% entre los diversos grupos, siendo cifras favorables para los modelos de predicción generados en el presente caso de estudio.

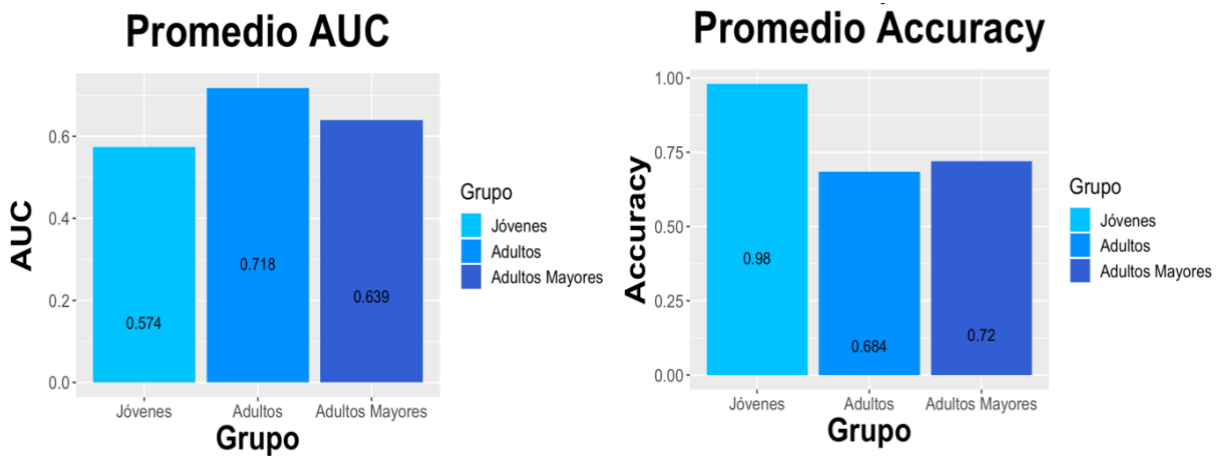


Figura 4.8 Promedios de métricas de desempeño, obtenidas en grupos de Edad del modelo de random forest (elaboración propia).

4.3. Caso de estudio 3: Sexo biológico

Para el último caso de estudio (4.3 a 4.3.5) se empleó la metodología vista en el capítulo 3, a continuación, se describe detalladamente cada una de las etapas.

4.3.1. Adquisición de datos

Para el tercer caso de estudio se utilizó el conjunto original con 253,680 datos y 19 características para su análisis.

4.3.2. Preprocesamiento de datos

Se realizó una división de dos subconjuntos, conformados por 229,787 pacientes hombres y 23,893 mujeres, como se observó en la Tabla 3.3.

4.3.3. Selección de características

Utilizando la misma metodología (vista en el capítulo 3) de los casos anteriores, las variables más significativas obtenidas por el algoritmo genético para el tercer caso de estudio se muestran a continuación en la Tabla 4.4.

Tabla 4.4 Características más significativas en grupos de Sexo (elaboración propia).

SUBCONJUNTO	CARACTERÍSTICAS MÁS SIGNIFICATIVAS
Hombres	Diabetes, consumo excesivo de alcohol , control de colesterol
Mujeres	Dificultad para caminar, control de colesterol , consumo excesivo de alcohol , verduras

En la Figura 4.9, se muestra la frecuencia de aparición de las características más importantes (eje vertical) en para el grupo de Hombres. El eje horizontal muestra las variables ordenadas por rango, siendo las de color negro más significativas: "Diabetes", "HvyAlcoholConsump" y "CholCheck". Al igual que en sección 4.2.3, los colores representan las evoluciones que se van creando a lo largo del ciclo de los algoritmos genéticos (visto en el capítulo 2), el color gris representa la frecuencia más baja. Se puede consultar el Anexo E para observar la frecuencia de aparición de variables del subconjunto de Mujeres analizado en este caso de estudio.

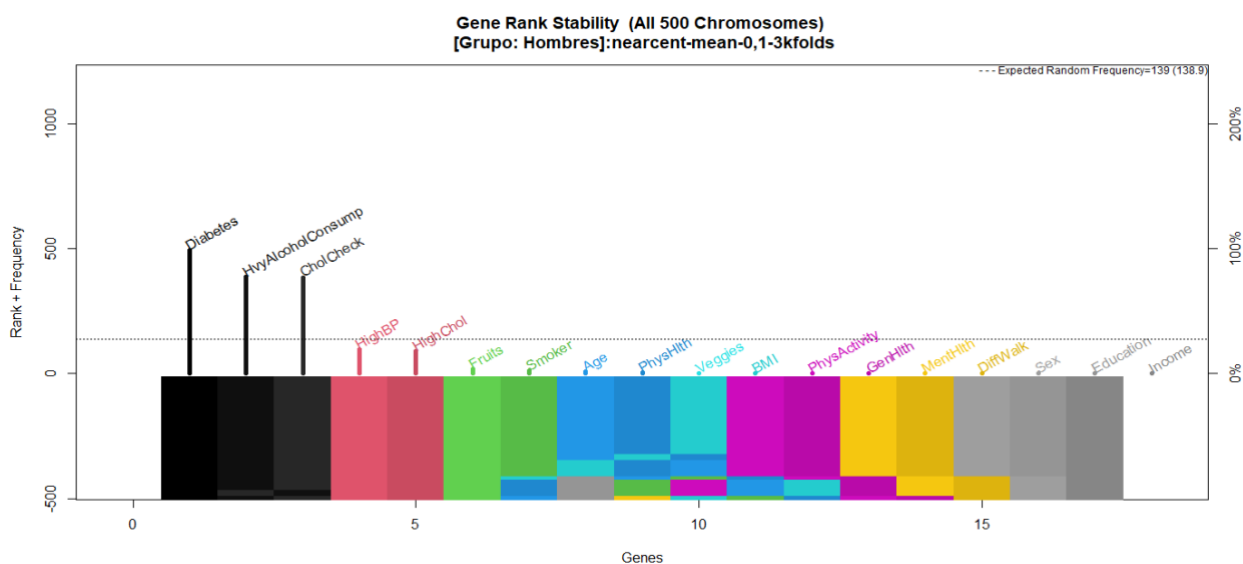


Figura 4.9 Frecuencia de aparición de las características más importantes obtenidas por Galgo en el grupo de Hombres (elaboración propia).

En la Figura 4.10 se observa el proceso de selección hacia adelante, el segundo modelo (línea sólida en negro) alcanza mejor desempeño y estabilidad con tres características: “Diabetes”, “HvyAlcoholConsump” y “CholCheck” (eje horizontal), el eje vertical representa el 0.8322 de precisión en la clasificación de personas sanas y enfermas (línea punteada color negro y rojo). El Anexo J muestra el resultado obtenido del mismo proceso en el grupo de Mujeres.

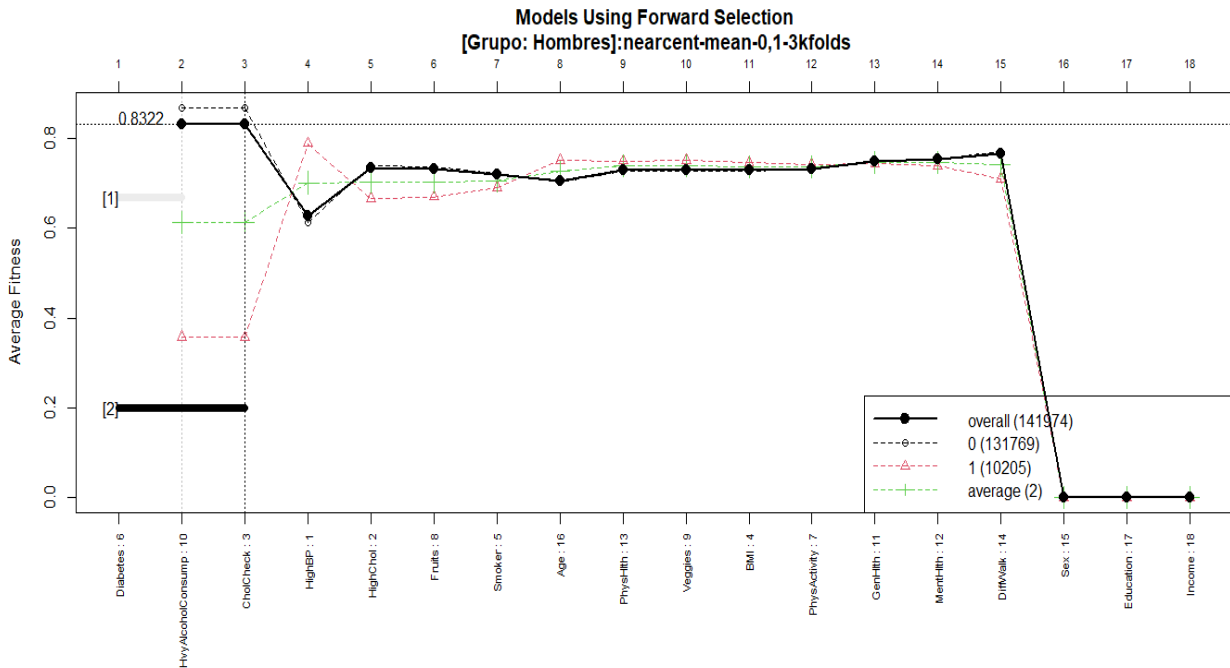


Figura 4.10 Proceso de selección hacia adelante en el grupo de Hombres (elaboración propia).

4.3.4. Obtención del modelo de predicción

De la misma manera que en anteriores casos de estudio, luego de obtener las características más significativas, se elaboró una comparativa de algunos algoritmos de aprendizaje supervisado mencionados anteriormente (capítulo 2), para así determinar cuál método es más preciso para la predicción de Enfermedades Cardiacas tanto en hombres como en mujeres.

4.3.5. Evaluación del modelo

En la Figura 4.11 se observan los resultados de la comparativa de los algoritmos vistos anteriormente. Logrando mejor precisión random forest para Hombres con el 68.8% y en el caso de las Mujeres el valor es el mismo para RF y regresión logística con 63.4% (ver Anexo O). KNN presenta un valor ligeramente bajo a comparación de los otros métodos de aprendizaje supervisado.

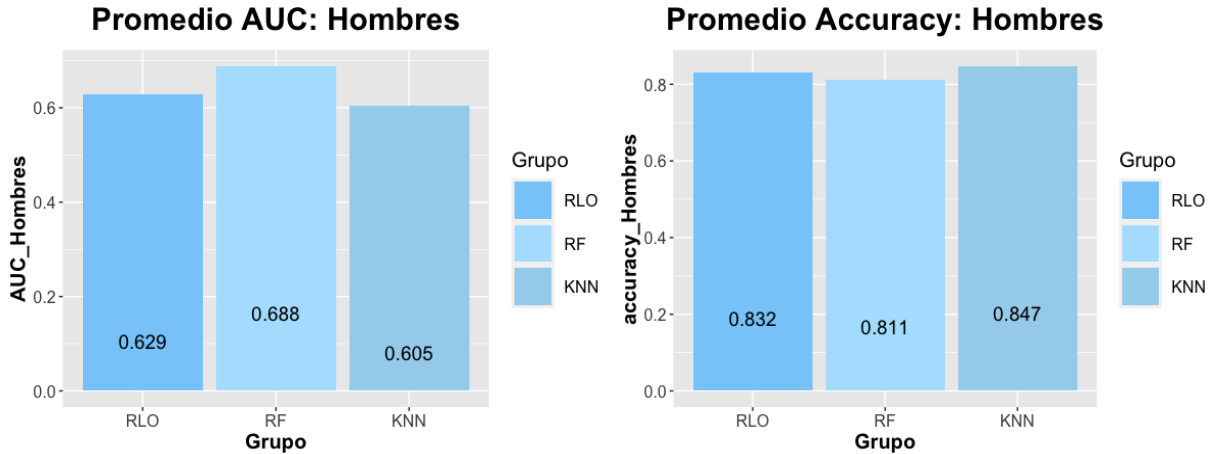


Figura 4.11 Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de Hombres (elaboración propia).

Se calculan las métricas de evaluación como se vio en la sección 2.6.4, obteniendo el promedio general de cada una, como se muestra en la Figura 4.12. Consiguiendo un área bajo la curva (AUC) del 62.9% para Hombres y 63.4% en Mujeres mediante el algoritmo de random forest y un accuracy de 83 y 82% respectivamente, siendo buenos valores para la predicción de dichos modelos.

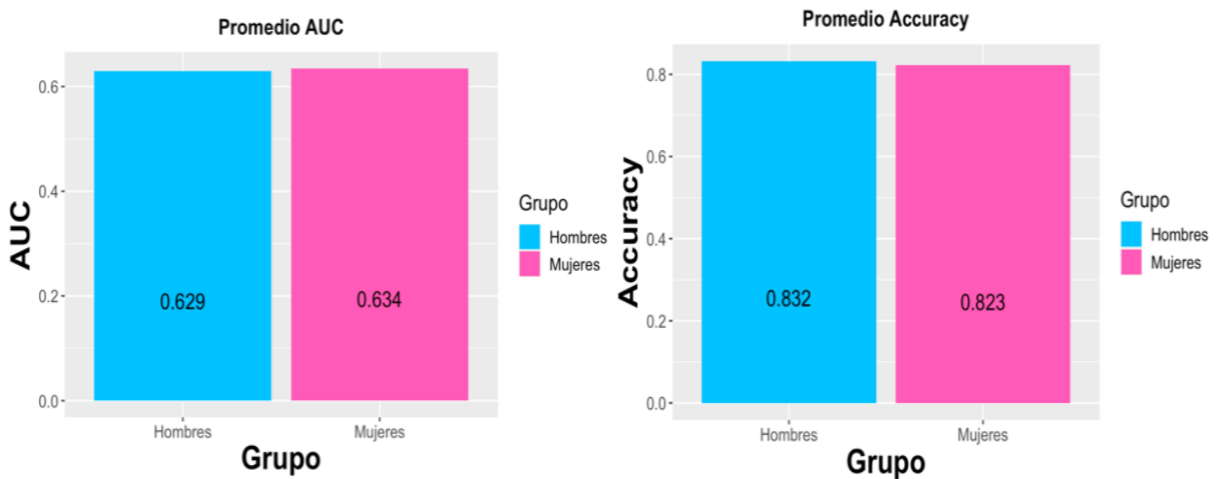


Figura 4.12 Promedios de métricas de desempeño, obtenidas en grupos de Sexo del modelo de random forest (elaboración propia).

Capítulo 5. Conclusiones

Con esta investigación se puede afirmar que la inteligencia artificial y sus técnicas son herramientas tecnológicas muy importantes en la actualidad, sus múltiples aplicaciones ayudan en determinados problemas que existen hoy en día a nivel mundial. En particular, para el diagnóstico oportuno, el aprendizaje automático (machine learning) juega un papel muy importante para predecir enfermedades, como es el caso de regresión logística y random forest, que son algoritmos utilizados en la presente investigación, arrojando buenos resultados en la predicción de Enfermedades Cardíacas, utilizando datos generales de pacientes tanto sanos como enfermos.

Mediante datos antropométricos de una persona y en conjunto con algoritmos de machine learning, es posible crear un biomarcador (modelo) que permita predecir (con más del 60% de precisión) Enfermedades Cardíacas en pacientes no diabéticos, prediabéticos, diabéticos, jóvenes (de 18 a 29 años), adultos (de 30 a 59 años), adultos mayores de 60 años, hombres y mujeres. Es muy importante resaltar que, en la actualidad, la mayoría de las investigaciones en el área médica en conjunto con la ciencia y la tecnología (como se observó en la sección 2.7), se centran más en un historial clínico y dejan de lado hábitos cotidianos que pueden llegar a ser un detonante en estos pacientes, por lo tanto, tomar en cuenta estos últimos datos es fundamental para prevenir enfermedades y/o complicaciones a futuro.

En el primer caso de estudio, se identificaron las características más importantes para clasificar a pacientes en diversos casos de diabetes debido a que esta enfermedad presenta condiciones multifactoriales, por lo cual, la dificultad para caminar, un mal control de colesterol, sexo y un alto consumo de alcohol, son factores de riesgo que nos pueden ayudar a predecir Enfermedades Cardíacas en personas no diabéticas, mientras que un bajo consumo de frutas, edad, mal estado de salud física y mental son características importantes para detectar una EC en pacientes prediabéticos y diabéticos, para este último grupo, es importante mencionar que el riesgo aumenta considerando otras variables como la dificultad para caminar, colesterol alto, género, mal estado de salud en general, alto consumo de alcohol, presión alta, índice de masa corporal y haber consumido más de 100 cigarrillos a lo largo de la vida. Según María Victoria Ramos [69] de la Revista Uruguaya de Cardiología el descenso de peso y la actividad física regular puede retrasar el avance de prediabetes a diabetes mellitus tipo 2, además la ingesta disminuida de calorías mejora la calidad de vida. Es recomendable suspender la ingesta de alcohol y tabaco para reducir incremento del riesgo a desarrollar alguna Enfermedad Cardiovascular (EVC).

En cuanto al segundo caso de estudio, se logró obtener las características que se mantienen en los diversos grupos de edad, en específico para Jóvenes y Adultos, es fundamental saber si el paciente es diabético, ya que es una enfermedad clave para que un individuo sea propenso a sufrir un episodio cardíaco, además de tener un mal control de colesterol y malos hábitos alimenticios, por lo tanto, muestra un alto índice de padecer alguna complicación cardíaca. La edad, también es un atributo importante que hay que tomar en cuenta en los Adultos, mientras que el mal estado de salud física y ser fumador son indicios que se deben considerar en Adultos Mayores. Durante el paso de los años la probabilidad de padecer una EVC se hace mayor, el corazón y el resto del organismo experimentan cambios en el transcurso del envejecimiento, por lo que es importante realizar un seguimiento en el control de la mayor cantidad de factores de riesgo desde temprana edad mediante cambios en el estilo de vida y evitar futuras complicaciones [70].

En el tercer y último caso de estudio, se determinaron los factores de riesgo que existen en

hombres y mujeres por las diferencias sobre hábitos saludables, de ejercicio, añadiendo principalmente las diferencias biológicas, por lo que, el consumo excesivo de alcohol y no tener un buen control de colesterol, son características que se comparten en ambos sexos. Añadiendo la diabetes como parte fundamental para detectar una Enfermedad Cardíaca en hombres debido a las condiciones multifactoriales y las complicaciones que esta conlleva. Sin embargo, el bajo consumo de verduras y la dificultad para caminar son factores que ayudan en la predicción de estas enfermedades en mujeres. Los varones presentan mayor riesgo de sufrir EC a una edad más temprana que las féminas que tienen el efecto protector del estrógeno. En ellas, el riesgo se iguala cuando llegan a la menopausia. Una dieta equilibrada, ejercicio físico regular y abstinencia del consumo de tabaco son medidas que ayudan a aminorar la probabilidad de padecer un evento cardíaco [70].

En general, la modificación en los hábitos de vida, en su mayoría lo referente a la alimentación, constituye en un aspecto fundamental en la prevención de Enfermedades Cardíacas, es decir, se recomienda consumir diariamente frutas y verduras, realizar ejercicio físico, y abandonar hábitos tóxicos ayuda a disminuir el riesgo de presentar eventos cardiovasculares [71]. Sin embargo, es esencial destacar la importancia del tratamiento personalizado acorde a las necesidades de cada persona en los diferentes casos de estudio analizados en la presente investigación. Siendo los grupos más vulnerables personas diabéticas debido a condiciones múltiples que esta patología presenta, al igual que los hombres por su factor genético y adultos mayores de 60 años derivado de su condición clínica.

5.1. Objetivos alcanzados

El primer objetivo planteaba identificar los estudios relacionados respecto al tema de investigación para ello, se logró distinguir cinco de ellos, así como sus diferentes contribuciones y limitaciones de cada uno. Posteriormente se realizó una búsqueda de conjuntos de datos sobre Enfermedades Cardíacas, para después determinar las características más significativas que causan EC mediante algoritmos genéticos, dando paso a la generación de modelos de machine learning que para esta investigación se implementó regresión logística, random forest y K-vecinos más cercanos (KNN). Después se realizó una comparativa de los riesgos presentes a desarrollar una EC en pacientes no diabéticos, prediabéticos, diabéticos, también en los distintos rangos de edad (jóvenes, adultos y adultos mayores), así mismo realizando una evaluación de dichos riesgos en hombres y mujeres.

5.2. Hipótesis/proposiciones demostradas

La hipótesis planteada inicialmente afirmaba que “a partir de datos y el uso de algoritmos genéticos, en combinación con Regresión logística, Random Forest y KNN es posible crear un biomarcador que permita detectar Enfermedades Cardíacas”. La hipótesis se demuestra de manera positiva, ya que se logró desarrollar un biomarcador para predecir EC. La precisión alcanzada en este estudio, situada por encima del 80% es similar a trabajos relacionados (mencionados en las secciones 2.6 y 2.7 del Capítulo 2) que se centraron en características clínicas obteniendo un promedio del 83%, esta investigación introduce un enfoque integral al incorporar características antropométricas, hábitos alimenticios, estado de salud física y mental, así como la consideración de hábitos tóxicos como el consumo de alcohol y tabaco, y enfermedades secundarias.

El hallazgo sugiere que la inclusión de información más completa, que va más allá de las características clínicas, puede servir para crear una herramienta potencialmente valiosa para la comunidad médica, ofreciendo nuevas perspectivas y estrategias en la detección temprana y gestión de enfermedades cardíacas acorde a las necesidades de cada persona.

5.3. Contribuciones de la investigación

Hasta el momento se han obtenido dos publicaciones que se muestran en la siguiente sección. También hubo participación y colaboración en un proyecto internacional con la Universidad Internacional de La Rioja en Colombia (UNIR), el objetivo principal fue desarrollar un prototipo de plataforma web utilizando herramientas de software libre y en conjunto con técnicas de inteligencia artificial, modelos de machine learning obteniendo el riesgo de padecer “Diabetes”, “Nefropatía diabética” y “Enfermedades Cardíacas”, solo después de que el usuario se haya registrado previamente, aceptado los términos y condiciones correspondientes, para posteriormente contestar una serie de preguntas médicas y generales sobre su estado de salud. Finalmente analizar los datos registrados por el usuario, arrojar un resultado final por cada patología y realizar recomendaciones acordes a las necesidades de cada paciente. En la sección de Anexos (P al X) se muestran las diferentes pantallas del prototipo de baja calidad.

5.4. Trabajos publicados

A lo largo del desarrollo de la presente investigación se ha logrado la publicación de dos artículos, el primero “Implementación de Algoritmos de Aprendizaje Automático para la Predicción de pacientes Diabéticos” [72] (ver Figura 5.1), tiene un enfoque hacia la detección de pacientes con Diabetes, publicado en el Congreso Internacional de Ingenierías, Ciencias y Tecnología Aplicada, CIICTA 2021. Se puede encontrar en la revista No. 12 “Investigación Aplicada, un enfoque en la Tecnología” a partir de la página 609 – 621.

Implementación de Algoritmos de Aprendizaje Automático para la Predicción de pacientes con Diabetes

Man Kit Liao-Li¹, Isamar Aparicio-Montelongo², José M. Celaya-Padilla³, Carlos E. Galván-Tejada⁴, Miguel Cruz⁵

¹ Universidad Autónoma de Zacatecas /Alumno, e-mail: mankit9lio@gmail.com

² Universidad Autónoma de Zacatecas /Alumno, e-mail: 20204469@uaz.edu.mx ³ Universidad Autónoma de Zacatecas /Docente, e-mail: jose.celaya@uaz.edu.mx ⁴ Universidad Autónoma de Zacatecas /Docente, e-mail: ericgalvan@uaz.edu.mx

⁵ Centro Médico Nacional Siglo XXI del Instituto Mexicano del Seguro Social (IMSS)/Médico Colaborador

Línea de investigación: Sistemas computacionales

Resumen

Según datos del INEGI, la diabetes ocupa el tercer lugar de las principales causas de muerte en México [1], con el paso del tiempo, el índice de muertes al año ha ido en aumento, además, esta enfermedad se empieza a presentar en edades cada vez más tempranas. Hoy en día y gracias al avance tecnológico, la implementación de modelos predictivos mediante técnicas de aprendizaje automático (ML) y minería de datos hacen posible diagnosticar enfermedades en etapas tempranas. El propósito de este artículo es presentar un análisis de un conjunto de datos que contiene información de pacientes diabéticos y personas sanas. Con ello, se pretende ayudar en la prevención y diagnóstico de esta patología, y predecir futuras complicaciones durante el control de este padecimiento. Se utilizan métricas como el área bajo la curva (AUC) para medir el rendimiento de las predicciones.

Palabras clave— Diabetes, Algoritmos de Aprendizaje Automático, Complicaciones patológicas, Predicción, AUC.

Abstract

According to data from INEGI, diabetes is the third leading cause of death in Mexico [1], over time, the rate of deaths per year has been increasing, in addition, this disease begins to present in age earlier and earlier. Nowadays and thanks to technological advances, the implementation of predictive models using machine learning (ML) techniques and data mining make it possible to diagnose diseases in early stages. The purpose of this article is to present an analysis of a data set that contains information on diabetic patients and healthy people. With this, it is intended to help in the prevention and diagnosis of this pathology, and to predict future complications during the control of this condition. Metrics such as area under the curve (AUC) are used to measure the performance of predictions.

Keywords— Diabetes, Machine Learning Algorithms, Pathological Complications, Prediction, AUC.

Figura 5.1 Artículo publicado en CHIITA 2021 [72].

El segundo artículo fue “Predicción de Enfermedades Cardíacas Derivadas de Diabetes, mediante Algoritmos Genéticos: Caso de Estudio” (ver Figura 5.2) [73], el cual tiene la capacidad de predecir Enfermedades Cardíacas en pacientes sanos, prediabéticos, diabéticos, jóvenes, adultos y adultos mayores mediante regresión logística y algoritmos genéticos. Se publicó en el Congreso Mexicano de Inteligencia Artificial (COMIA), 2022.

Predicción de enfermedades cardíacas derivadas de diabetes, mediante algoritmos genéticos: caso de estudio

Isamar Aparicio-Montelongo¹, José M. Celaya-Padilla², Huizilopoztli Luna-García²,
Carlos E. Galván-Tejada², Jorge I. Galván-Tejada², Hamurabi Gamboa-Rosales²

¹ Consejo Nacional de Ciencia y Tecnología,
Unidad Académica de Ingeniería Eléctrica,
México

² Universidad Autónoma de Zacatecas,
México

{20204469, jose.celaya, hlugar, ericgalvan,
gatejo, hamurabigr}@uaz.edu.mx

Resumen. En México, según datos del INEGI las principales causas de muerte son Enfermedades Cardíacas (EC) y Diabetes Mellitus (DM). En el primer semestre de 2021 se registraron 579,586 decesos, ocupando los primeros lugares solo después de COVID-19. Correspondiendo a las EC, como la segunda más importante con un 19.7%. Gracias al avance tecnológico, en la actualidad, es posible crear modelos para el diagnóstico oportuno de patologías, mediante técnicas de Inteligencia Artificial (IA). El objetivo de esta investigación es implementar algoritmos genéticos (AG) como método de selección de características posteriormente, generar un modelo multivariado con Regresión Logística para conocer si es posible considerarlo como una herramienta eficaz en la detección de pacientes propensos a sufrir un episodio cardíaco derivado de DM.

Palabras clave: Enfermedades cardíacas, inteligencia artificial, predicción, diabetes, algoritmos genéticos, selección de características.

Prediction of Diabetes-Related Heart Disease Using Genetic Algorithms: A Case Study

Abstract. In Mexico, according to INEGI data, the main causes of death are Heart Disease (CD) and Diabetes Mellitus (DM). In the first semester of 2021, 579,586 deaths were registered, occupying the first places only after COVID-19. Corresponding to CD, as the second most important with 19.7%. Thanks to

Referencias

- [1] Organización Mundial de la Salud, «Enfermedades cardiovasculares». mayo de 2020. [En línea]. Disponible en: https://www.who.int/es/health-topics/cardiovascular-diseases#tab=tab_1
- [2] Organización Panamericana de la Salud, «Día Mundial del Corazón: Enfermedades Cardiovasculares causan 1,9 millones de muertes al año en las Américas». 2020. Accedido: 28 de marzo de 2022. [En línea]. Disponible en: https://www3.paho.org/hq/index.php?option=com_content&view=article&id=7257:2012-dia-mundial-corazon-enfermedades-cardiovasculares-causan-1-9-millones-muertes-ano-americas&Itemid=4327&lang=fr
- [3] Chevez Elizondo Daniel, A. A. Kristel, Salas Ureña Fabio, Robledo Guzmán Alison, L. C. Ernesto, y A. V. María, «Factores de riesgo cardiovascular», *Med. - Programa Form. Mdica Contin. Acreditado*, pp. 6-9, 2020, doi: 10.1016/s0211-3449(05)73753-x.
- [4] Instituto Nacional de Estadística Geografía e Informática, «Estadísticas de defunciones registradas, enero-junio 2021», vol. 2021, pp. 1-40, 2022.
- [5] J. F. Avila-Tomás, M. A. Mayer-Pujadas, y V. J. Quesada-Varela, «La inteligencia artificial y sus aplicaciones en medicina I: introducción antecedentes a la IA y robótica», *Aten. Primaria*, vol. 52, n.º 10, pp. 778-784, dic. 2020, doi: 10.1016/j.aprim.2020.04.013.
- [6] IBM, «¿Qué es la inteligencia artificial en la medicina?» 2021. Accedido: 9 de junio de 2022. [En línea]. Disponible en: <https://www.ibm.com/mx-es/topics/artificial-intelligence-medicine>
- [7] Organización Mundial de la Salud, «Enfermedades no transmisibles». 2021. Accedido: 27 de abril de 2022. [En línea]. Disponible en: <https://www.who.int/es/news-room/fact-sheets/detail/noncommunicable-diseases>
- [8] Instituto Nacional de Estadística Geografía e Informática, «Comunicado de prensa núm. 592/21 28 de octubre de 2021 Pág. 2/4», 2021. [En línea]. Disponible en: <https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2021/EstSociodemo/DefuncionesRegistradas2020preliminar.pdf>
- [9] Asociación Americana del Corazón, «Los supervivientes de ataques cardíacos podrían experimentar una disminución más rápida de la función cerebral.» febrero de 2022. Accedido: 29 de marzo de 2022. [En línea]. Disponible en: <https://www.heart.org/en/news/2022/02/03/heart-attack-survivors-could-experience-more-rapid-brain-function-declines>
- [10] N. Hierrezuelo Rojas, J. T. Álvarez Cortés, J. Cruz Llaugert, A. J. Limia Dominguez, y P. J. Carlos Finlay Servicio Asistencia Médica Santiago de Cuba, «Factores de riesgo asociados a enfermedades cardiovasculares. Policlínico Ramón López Peña», vol. 27, n.º 4, 2021, [En línea]. Disponible en: <http://www.revcardiologia.sld.cu/>
- [11] Gómez Álvarez Enrique, «Enfermedades del corazón, pandemia permanente». septiembre de 2020. Accedido: 28 de marzo de 2022. [En línea]. Disponible en: https://www.dgcs.unam.mx/boletin/bdboletin/2020_811.html
- [12] Instituto Nacional de Estadística Geografía e Informática, «Principales causas de mortalidad por residencia habitual, grupos de edad y sexo del fallecido». noviembre de 2020. Accedido: 4 de mayo de 2022. [En línea]. Disponible en: <https://www.inegi.org.mx/sistemas/olap/registros/vitales/mortalidad/tabulados/ConsultaMortalidad.asp>
- [13] Instituto Nacional de Estadística Geografía e Informática, «*statistic_id650286_mexico_-principales-causas-de-mortalidad-en-el-estado-de-zacatecas-2020*», 2020.
- [14] C. Daniel González Cedillo, «Diagnóstico de enfermedades cardíacas con los algoritmos

supervisados Naives Bayesian», 2019.

[15] D. Chicco y G. Jurman, «Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone», *BMC Med. Inform. Decis. Mak.*, vol. 20, n.º 1, pp. 1-16, 2020, doi: 10.1186/s12911-020-1023-5.

[16] F. Ali *et al.*, «A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion», *Inf. Fusion*, vol. 63, pp. 208-222, 2020, doi: 10.1016/j.inffus.2020.06.008.

[17] Gallego Valcárcel David y Lucas Monsalve Delly Fabián, «Modelos De Aprendizaje Automático Para La Predicción Del Riesgo De Fatalidad Por Insuficiencia Cardíaca Con Datos Clínicos», 2021.

[18] S. Faiyaz Waris y S. Koteeswaran, «Heart disease early prediction using a novel machine learning method called improved K-means neighbor classifier in python», *Mater. Today Proc.*, n.º xxxx, pp. 1-7, 2021, doi: 10.1016/j.matpr.2021.01.570.

[19] Secretaría de Gobernación, «Diario Oficial de la Federación», pp. 14-27, 2020.

[20] Salesforce, «Inteligencia Artificial: ¿Qué es?» junio de 2017.

[21] Kaggle y Teboul Alex, «Heart Disease Health Indicators Dataset». diciembre de 2021. Accedido: 6 de marzo de 2022. [En línea]. Disponible en: <https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset>

[22] Medical New Today, «Enfermedad de las arterias coronarias», 2019. [En línea]. Disponible en: <https://www.medicalnewstoday.com/articles/es/327293#pronosticoHaydiferentesenfermedadescardiacasafectanalcorazondeidiferentesmaneras.Enlasseccionessiguienteseseanalizarnalgunstiposdiferentesdeenfermedadcardiacaconmaldetalle>.

[23] Barrera Domínguez Erick Armando, «Factores de riesgo asociados a defectos al nacimiento en mujeres atendidas en el Hospital de la Mujer Puebla», 2021, [En línea]. Disponible en: <https://repositorioinstitucional.buap.mx/handle/20.500.12371/13821>

[24] B. M. Valbuena Oscar, Vera Miguel, «A current review of computational techniques for diseases characterizing associated with the aortic valve», 2020. [En línea]. Disponible en: http://bonga.unisimon.edu.co/bitstream/handle/20.500.12442/6848/Revisi%C3%B3n_actual_tecnicas_computacionales.pdf?sequence=1&isAllowed=y

[25] Mayo Clinic, «Enfermedad cardíaca». diciembre de 2021. Accedido: 1 de marzo de 2022. [En línea]. Disponible en: <https://www.mayoclinic.org/es-es/diseases-conditions/heart-disease/symptoms-causes/syc-20353118?p=1>

[26] F. Arrieta *et al.*, «Diabetes mellitus and cardiovascular risk: Working group recommendations of Diabetes and Cardiovascular Disease of the Spanish Society of Diabetes (SED, 2015)», en *Atencion Primaria*, Elsevier Doyma, may 2016, pp. 325-336. doi: 10.1016/j.aprim.2015.05.002.

[27] Federación Internacional de Diabetes, «Diabetes y enfermedades cardiovasculares». 2021. [En línea]. Disponible en: <https://www.idf.org/our-activities/care-prevention/cardiovascular-disease.html>

[28] B. S. Martínez, V. V. Falcón, N. G. Martínez, y G. E. V. Vizúete, «Case-control study on risk factors for type 2 diabetes mellitus in older adults», *Univ. Soc.*, vol. 12, n.º 4, pp. 156-164, 2020.

[29] G. Blanc-Pihuave, L. Cevallos-Torres, y J. Arteaga-Vera, «Modelo computacional de clasificación de aprendizaje de máquina supervisado, para el análisis de datos cardiovasculares y pronóstico médico», *Ecuadorian Sci. J.*, vol. 4, n.º 2, pp. 71-79, sep. 2020, doi: 10.46480/esj.4.2.83.

[30] Federación Internacional de Diabetes, «¿Qué es la Diabetes?» 2020. [En línea].

Disponible en: <https://idf.org/aboutdiabetes/what-is-diabetes.html>

[31] Organización Mundial de la Salud, «Diabetes». 2021. Accedido: 26 de abril de 2022. [En línea]. Disponible en: <https://www.who.int/es/news-room/fact-sheets/detail/diabetes>

[32] Federación Internacional de Diabetes, «Complicaciones de la diabetes». 2020. [En línea]. Disponible en: <https://www.idf.org/aboutdiabetes/complications.html>

[33] Federación Internacional de Diabetes, «La diabetes y los riñones», 2021, [En línea]. Disponible en: <https://www.idf.org/our-activities/care-prevention/diabetes-and-the-kidney.html>

[34] Federación Internacional de Diabetes, «El pie diabético». 2020. Accedido: 9 de mayo de 2022. [En línea]. Disponible en: <https://www.idf.org/our-activities/care-prevention/diabetic-foot.html>

[35] Federación Internacional de Diabetes, «La diabetes y el ojo». 2020. Accedido: 9 de mayo de 2022. [En línea]. Disponible en: <https://www.idf.org/our-activities/care-prevention/eye-health.html>

[36] Federación Internacional de Diabetes, «Diabetes gestacional». 2020. Accedido: 9 de mayo de 2022. [En línea]. Disponible en: <https://www.idf.org/our-activities/care-prevention/gdm.html>

[37] Humphreys J, «ACC19 – Nuevas Guías de Prevención Primaria ACC/AHA». 2019. [En línea]. Disponible en: <https://www.siacardio.com/academia/guias/acc19-nuevas-guias-de-prevencion-primaria-accaha-2019/>

[38] Villar Moya Raul, «Manejo del riesgo cardiovascular en adultos con diabetes tipo 2». 2022. [En línea]. Disponible en: <https://www.siacardio.com/editoriales/prevencion-cardiovascular/siacprevent/diabetest2/>

[39] K. Strimbu y J. A. Tavel, «What are biomarkers?», *Curr. Opin. HIV AIDS*, vol. 5, n.º 6, pp. 463-466, nov. 2010, doi: 10.1097/COH.0b013e32833ed177.

[40] N. Gestión, «¿Qué es la inteligencia artificial y para qué sirve? | TECNOLOGIA», Gestión. Accedido: 28 de junio de 2023. [En línea]. Disponible en: <https://gestion.pe/tecnologia/inteligencia-artificial-historia-origen-funciona-aplicaciones-categorias-tipos-riesgos-nnda-nnlt-249002-noticia/>

[41] García Peñalvo Francisco José, «Una introducción a la inteligencia artificial», 2019. [En línea]. Disponible en: <https://repositorio.grial.eu/bitstream/grial/1639/1/IA.pdf>

[42] C. Janiesch, P. Zschech, y K. Heinrich, «Machine learning and deep learning», 2021, doi: 10.1007/s12525-021-00475-2/Published.

[43] Guzmán Bazán Yuniel, *VI Congreso Internacional de Investigación Cualitativa en Ciencia y Tecnología*. Editorial Científica 3Ciencias, 2018. doi: 10.17993/IngyTec.2018.31.

[44] Izetta Javier, «Técnicas de Inteligencia Artificial aplicadas a Problemas de Visión por Computadora», may 2020, [En línea]. Disponible en: http://sedici.unlp.edu.ar/bitstream/handle/10915/103379/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

[45] Sancho Escrivá José Vicente, «Utilidad de las nuevas tecnologías en la mejora de la comunicación médico-paciente en el área de salud mental: aportaciones de la inteligencia artificial y el procesamiento del lenguaje natural», PhD Thesis, Universitat Jaume I, Castelló de la Plana, 2021. doi: 10.6035/14111.2021.301397.

[46] O. D. Castrillón, W. Sarache, y S. Ruiz-Herrera, «Predicción del rendimiento académico por medio de técnicas de inteligencia artificial», *Form. Univ.*, vol. 13, n.º 1, pp. 93-102, feb. 2020, doi: 10.4067/S0718-50062020000100093.

[47] E. Serna M, *Desarrollo e Innovación en Ingeniería*. 2017. [En línea]. Disponible en: https://www.researchgate.net/publication/331498946_Principios_y_caracteristicas_de_las_redes_neuronales_artificiales

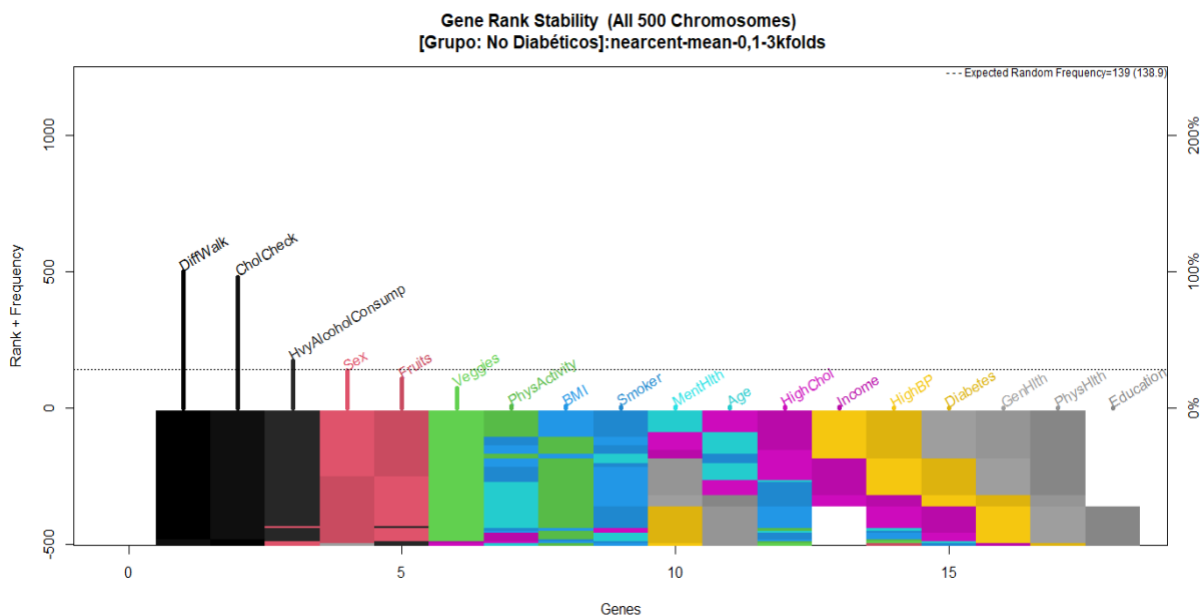
[48] Bravo Andrés y Restrepo Diego, «Modelo para la detección de riesgo de crédito en

- entidades financieras utilizando técnicas de Inteligencia Artificial», PhD Thesis, 2020. [En línea]. Disponible en: <https://repositorio.utp.edu.co/server/api/core/bitstreams/d815b335-981a-4176-a9d0-490a020cd396/content>
- [49] Montoya Hernández Angélica Yohana, «Inteligencia artificial al servicio de la auditoría: Una revisión sistemática de literatura», mar. 2020, [En línea]. Disponible en: https://media.proquest.com/media/hms/PFT/1/ezJ5G?_s=mkgoMzeMb2Vp13aRyCc%2BsdInvA%3D
- [50] Cotino Lorenzo, «Riesgos e Impactos del Big Data, la Inteligencia Artificial y la Robótica. Enfoques, Modelos y Principios de la Respuesta del Derecho», 2019, [En línea]. Disponible en: https://www.researchgate.net/profile/Lorenzo-Hueso/publication/349494641_Riesgos_e_impactos_del_big_data_la_inteligencia_artificial_y_la_robotica_y_enfoques_modelos_y_principios_de_la_respuesta_del_Derecho/links/6038e67ba6fdcc37a85250cf/Riesgos-e-impactos
- [51] Coto Jiménez Marvin, «Consideraciones para la incorporación de la Inteligencia Artificial en un programa de pregrado de Ingeniería Eléctrica», ago. 2021, [En línea]. Disponible en: <https://www.scielo.sa.cr/pdf/aie/v21n2/1409-4703-aie-21-02-00529.pdf>
- [52] Maisueche Cuadrado D. Alberto, «Utilización del Machine Learning en la Industria 4.0», PhD Thesis, 2019. [En línea]. Disponible en: <https://uvadoc.uva.es/bitstream/handle/10324/37908/TFM-I-1372.pdf?sequence=1&isAllowed=y>
- [53] IBM, «¿Qué es Machine Learning?» 2021. [En línea]. Disponible en: <https://www.ibm.com/mx-es/analytics/machine-learning>
- [54] Ligdi Gonzalez, «Clasificación de Machine Learning». 2018. Accedido: 14 de diciembre de 2022. [En línea]. Disponible en: <https://aprendeia.com/clasificacion-de-machine-learning/>
- [55] C.-P. E. Pedrero Víctor Reynaldos-Grandón Katuska, Ureta-Achurra Joaquín, «Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia», 2021.
- [56] CEC, «Machine Learning, un gran aliado para la ciberseguridad». 2021. Accedido: 8 de junio de 2022. [En línea]. Disponible en: <https://noticias.cec.es/index.php/2021/12/13/machine-learning-un-gran-aliado-para-la-ciberseguridad/>
- [57] Rodríguez Daniel, «La regresión logística». julio de 2018. [En línea]. Disponible en: <https://www.analyticslane.com/2018/07/23/la-regresion-logistica/>
- [58] J. J. Espinosa Zúñiga, «Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito», *Ing. Investig. Tecnol.*, vol. 21, n.º 3, pp. 1-16, jul. 2020, doi: 10.22201/fi.25940732e.2020.21.3.022.
- [59] L. A. Bermeo Varon, J. E. Vargas Vásquez, y N. Erazo Benavidez, «Aplicación del algoritmo de K-NN en la asignación de órdenes de trabajo de mantenimiento correctivo para equipos biomédicos», *Comput. Electron. Sci. Theory Appl.*, vol. 3, n.º 1, pp. 39-47, mar. 2022, doi: 10.17981/cesta.03.01.2022.05.
- [60] L. González, «Amazon Machine Learning Guía para desarrolladores», pp. 94-96, 2016.
- [61] Barrios Arce Juan Ignacio, «La matriz de confusión y sus métricas». 2019. Accedido: 18 de mayo de 2022. [En línea]. Disponible en: <https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/>
- [62] N. H. Quiroz, M. L. Posadas Martínez, E. Rossi, D. Giunta, y M. Risk, «Aprendizaje automático aplicado en área de la salud. Parte 2», *Rev. Hosp. Ital. B. Aires*, vol. 42, n.º 1, pp. 56-58, mar. 2022, doi: 10.51987/revhospitalbaires.v42i1.152.
- [63] R. A. I. Irizarry, «Introducción a la Ciencia de Datos - Análisis de datos y algoritmos de predicción con R», *CRC Press*, pp. 538-539, 2021.

- [64] Mena Camaré Luis Javier, «Aprendizaje Automático a partir de Conjuntos de Datos No Balanceados y su Aplicación en el INSTITUTO NACIONAL DE ASTROFÍSICA, OPTICA Y ELECTRÓNICA», 2008. [En línea]. Disponible en: <https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/533/1/MenaCaLJ.pdf>
- [65] R. Amat, «Validación de modelos predictivos: Cross-validation, OneLeaveOut, Bootstrapping.» 2020. [En línea]. Disponible en: https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap
- [66] Ramón Garduño Juárez, «Algoritmos genéticos», 2018, [En línea]. Disponible en: <https://conogasi.org/articulos/algoritmos-geneticos/>
- [67] V. Trevino y F. Falciani, «GALGO An R package for Genetic Algorithm Searches (Customized for Variable Selection in Functional Genomics)», 2006. [En línea]. Disponible en: <http://www.wag-inc.org>
- [68] Hernández Sampieri Roberto, Méndez Valencia Sergio, Mendoza Torres Christian Paulina, y Cuevas Romo Ana, *Fundamentos de Investigación*. 2017.
- [69] Ramos María Victoria, «Novedades de la Guía Europea 2019 sobre diabetes, prediabetes y enfermedades cardiovasculares», *Rev. Urug. Cardiol.*, vol. 3, n.º 1, mar. 2020, doi: 10.29277/cardio.35.1.10.
- [70] Valle Muñoz Alfonso, «Fundación Española del Corazón». 2017. Accedido: 28 de febrero de 2023. [En línea]. Disponible en: <https://fundaciondelcorazon.com/prevencion/marcadores-de-riesgo.html>
- [71] A. E. Pinilla-Roa y M. D. P. Barrera-Perdomo, «Prevention of diabetes mellitus and cardiovascular risk: Medical and nutritional approach», *Rev. Fac. Med.*, vol. 66, n.º 3, pp. 459-468, 2018, doi: 10.15446/revfacmed.v66n3.60060.
- [72] Liao-Li Man Kit, Aparicio-Montelongo Isamar, Celaya-Padilla José M., y Galván-Tejada Carlos E., «Implementación de Algoritmos de Aprendizaje Automático para la Predicción de pacientes con Diabetes .», *Rev. Investig. Apl. Un Enfoque En Tecnol.*, pp. 609-621, 2021.
- [73] Aparicio-Montelongo Isamar, Celaya-Padilla José M., Luna-García Huizilopoztli, Galván-Tejada Carlos E., y Gamboa-Rosales Hamurabi, «Predicción De Enfermedades Cardíacas Derivadas De Diabetes , mediante Algoritmos Genéticos : Caso de Prediction of Diabetes-Related Heart Disease Using Genetic Algorithms : A Case Study .», pp. 159-172, 2022.

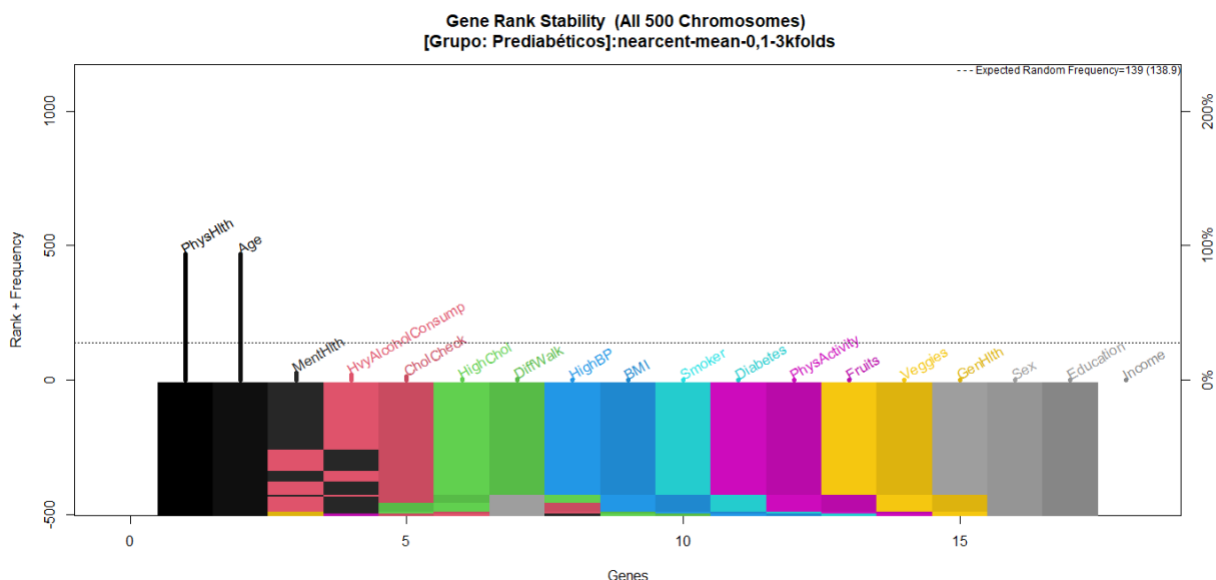
Anexos

El eje horizontal de los Anexos A-E muestra los genes ordenados por rango, y el eje vertical muestra frecuencia de aparición de las características y el rango esta codificado por colores en cada una de las evoluciones anteriores, donde "DiffWalk", "CholCheck" y "HvyAlcoholConsump" (aparecen en color negro) para el subconjunto de No diabéticos, representando la frecuencia más alta y las variables en gris presentan la más baja.



Anexo A. Frecuencia de aparición de las características más importantes obtenidas por galgo en el grupo de No diabéticos (elaboración propia).

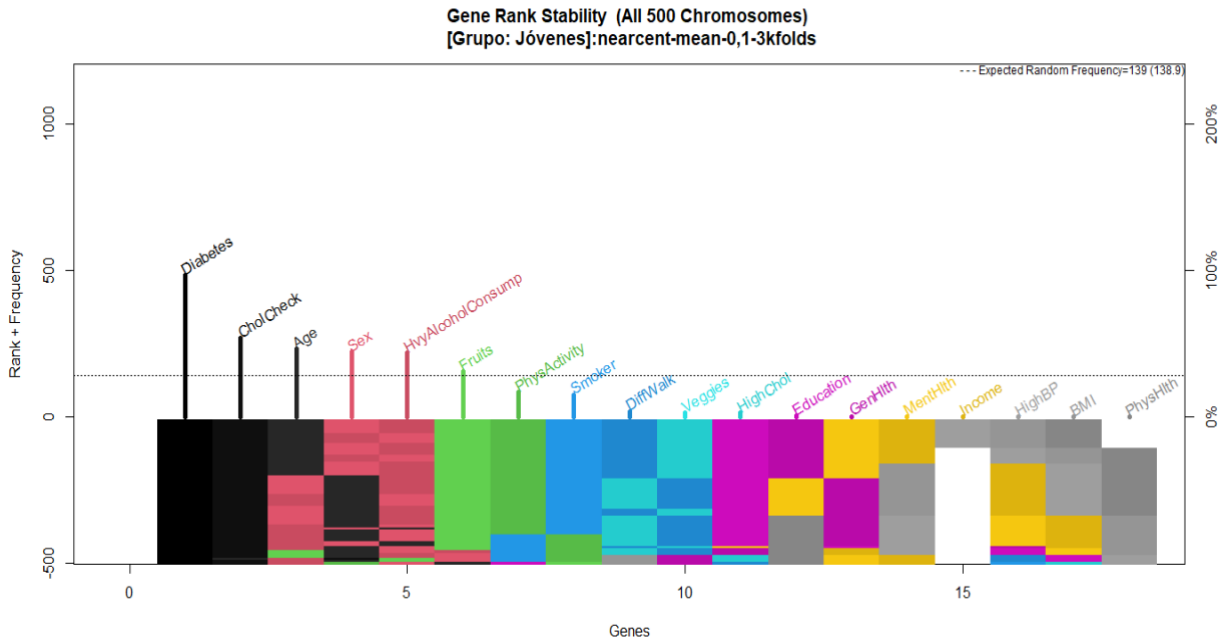
“PhysHlth”, “Age” y “MentHlth” representan las variables más frecuentes para los pacientes prediabéticos.



Anexo B. Frecuencia de aparición de las características más importantes obtenidas por galgo en el grupo de Prediabéticos

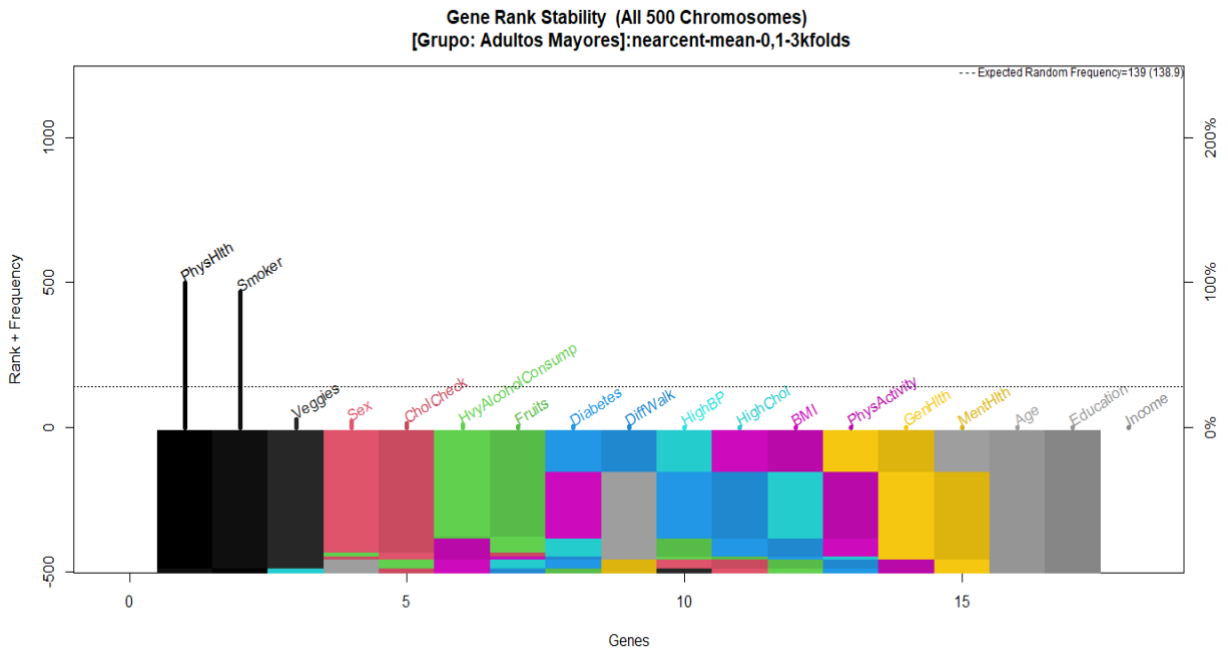
(elaboración propia).

Las características más frecuentes para los Jóvenes son: “Diabetes”, “CholCheck” y “Age”.



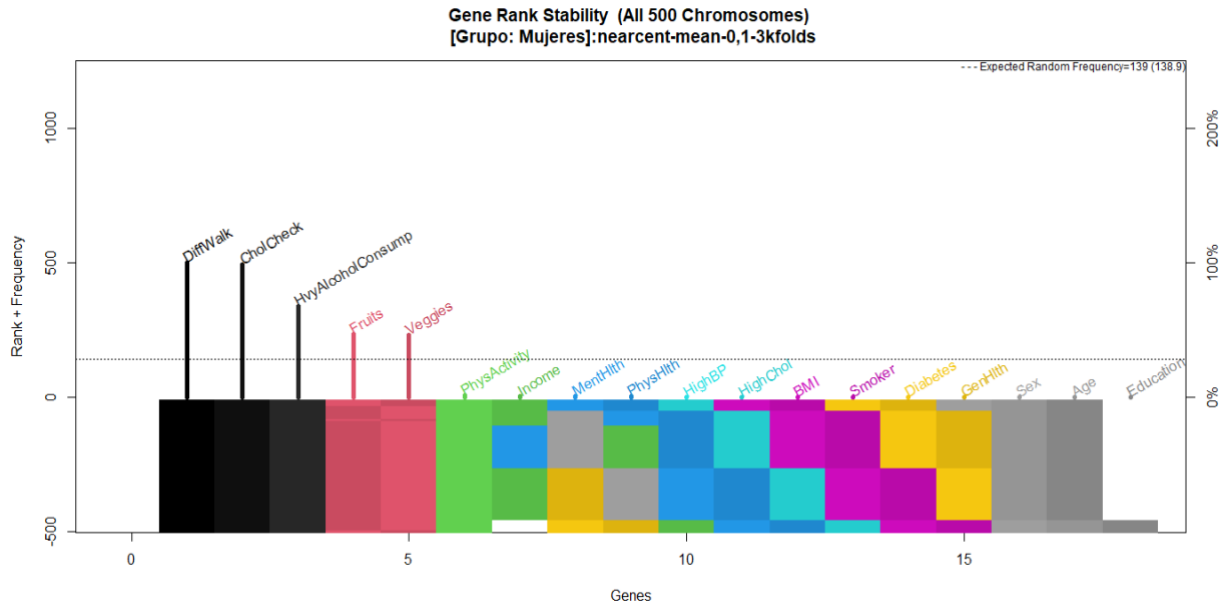
Anexo C. Frecuencia de aparición de las características más importantes obtenidas por galgo en el grupo de Jóvenes (elaboración propia).

“PhysHlth”, “Smoker” y “Veggies” representan las características más frecuentes para el grupo de Adultos Mayores.



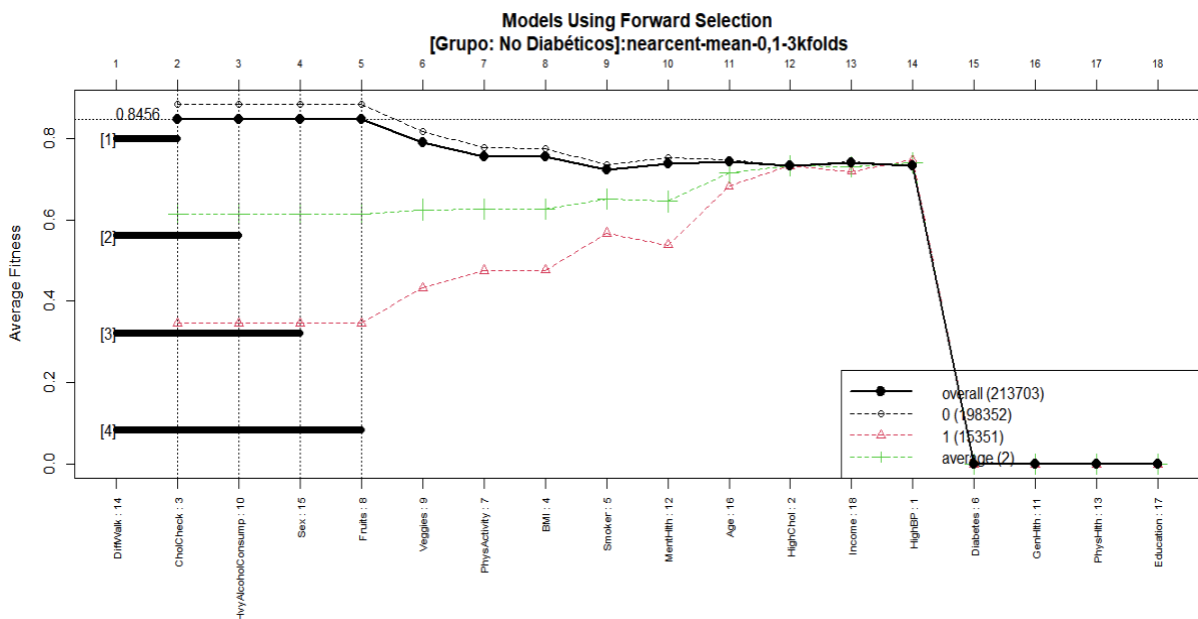
Anexo D. Frecuencia de aparición de las características más importantes obtenidas por galgo en el grupo de Adultos mayores (elaboración propia).

“DiffWalk”, “CholCheck” y “HvyAlcoholConsump” son las características más significativas en el grupo de mujeres.



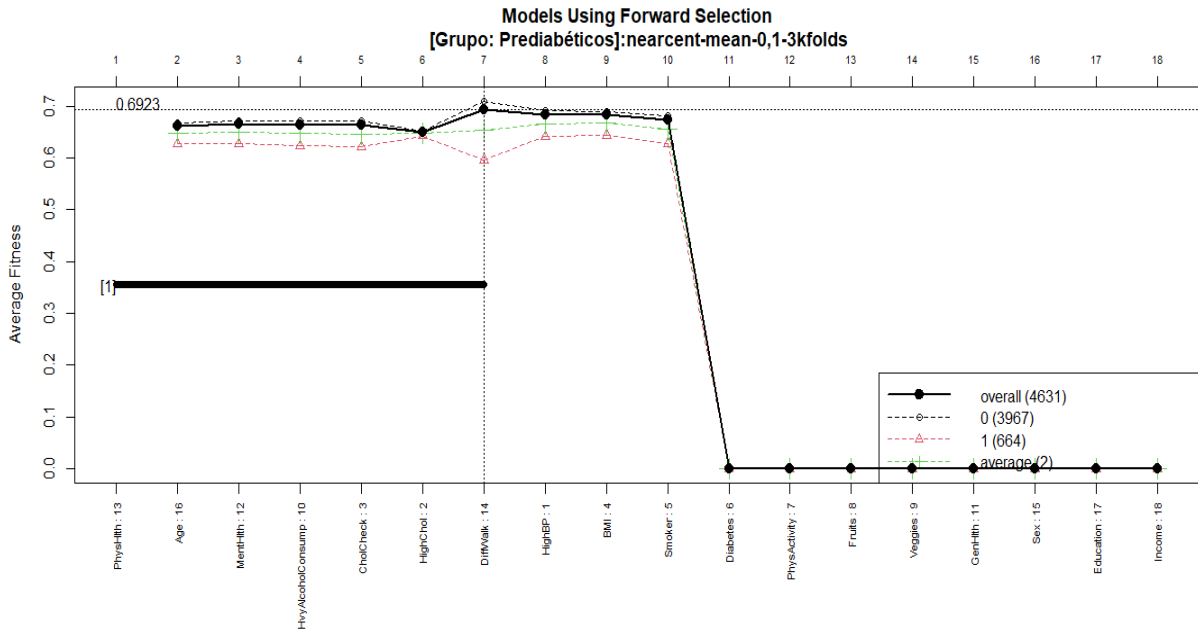
Anexo E. Frecuencia de aparición de las características más importantes obtenidas por galgo en el grupo de Mujeres (elaboración propia).

En los Anexos F-J se muestra el proceso de selección hacia adelante, donde el eje horizontal simboliza los genes ordenados por su rango de aparición. El eje vertical muestra la exactitud de la clasificación. La línea con puntos negros representa la precisión general. Las líneas discontinuas de colores (negro y rojo) interpretan la precisión por clase (personas sanas y enfermas). Para el grupo de No diabéticos el procedimiento de selección obtuvo cuatro modelos, contienen desde 2 hasta 5 variables (“DiffWalk”, “CholCheck”, “HvyAlcoholConsump”, “Sex” y “Fruits”), obteniendo así una precisión del 0.8456.



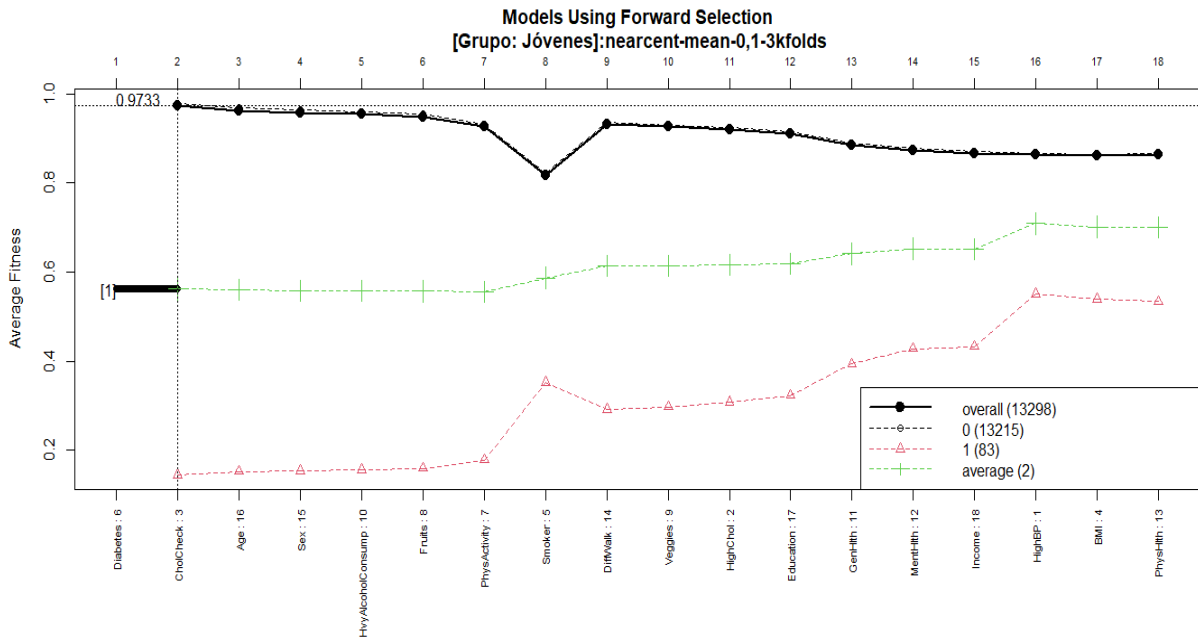
Anexo F. Proceso de selección hacia adelante en el grupo de No diabéticos (elaboración propia).

Forward selection arrojó únicamente un modelo con 7 características para el subconjunto de Prediabéticos, las cuales son: “PhysHlth”, “Age”, “MentHlth”, “CholCheck”, “HighChol” y “DiffWalk”, obteniendo 0.6923 de precisión.



Anexo G. Proceso de selección hacia adelante en el grupo de Prediabéticos (elaboración propia).

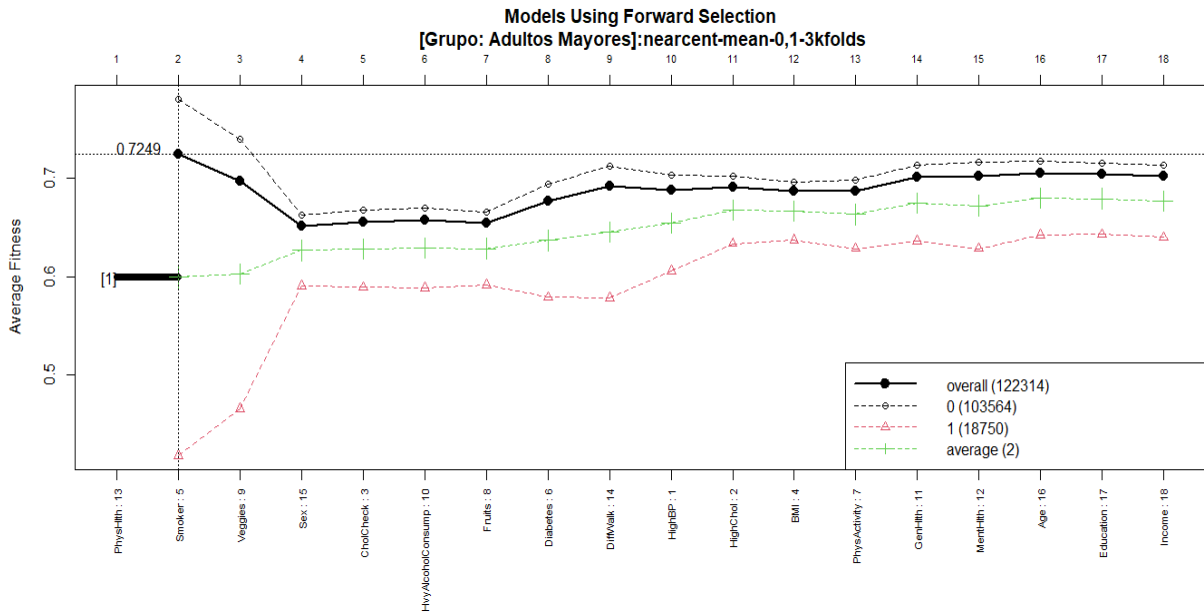
Igual que en el conjunto anterior, para el caso de Jóvenes, solamente se obtuvo un modelo con dos características (“Diabetes” y “CholCheck”), representando 0.9733 de precisión, siendo un valor alto en clasificación.



Anexo H. Proceso de selección hacia adelante en el grupo de Jóvenes (elaboración propia).

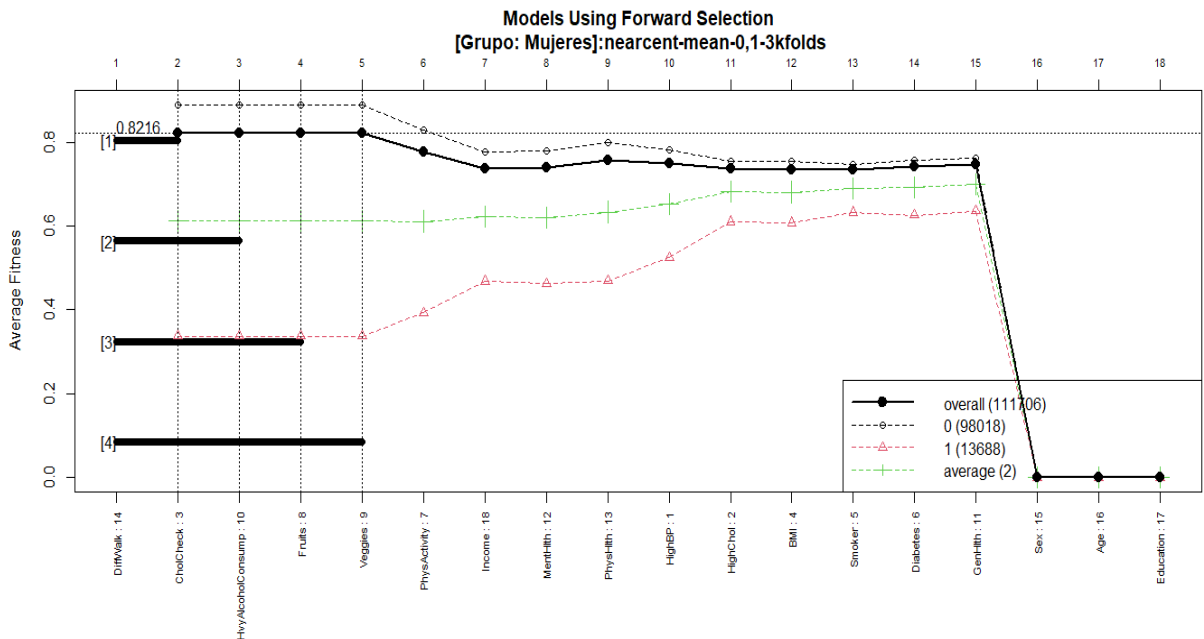
“PhysHlth” y “Smoker” son las características más significativas del modelo generado por la

selección hacia adelante en el grupo de Adultos mayores, obteniendo 0.7249 de precisión.



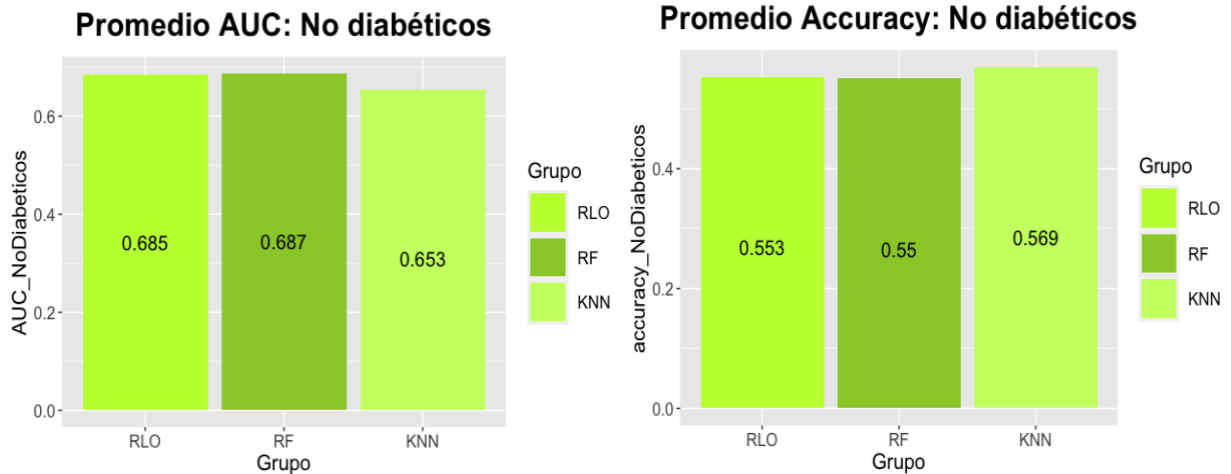
Anexo I. Proceso de selección hacia adelante en el grupo de Adultos mayores (elaboración propia).

En el Anexo I se puede observar que el método de selección hacia adelante generó cuatro modelos que contienen entre dos y cinco características más importantes (“DiffWalk”, “CholCheck”, “HvyAlcoholConsump”, “Fruits” y “Veggies”) en el grupo de Mujeres, obteniendo 0.8216 de precisión.



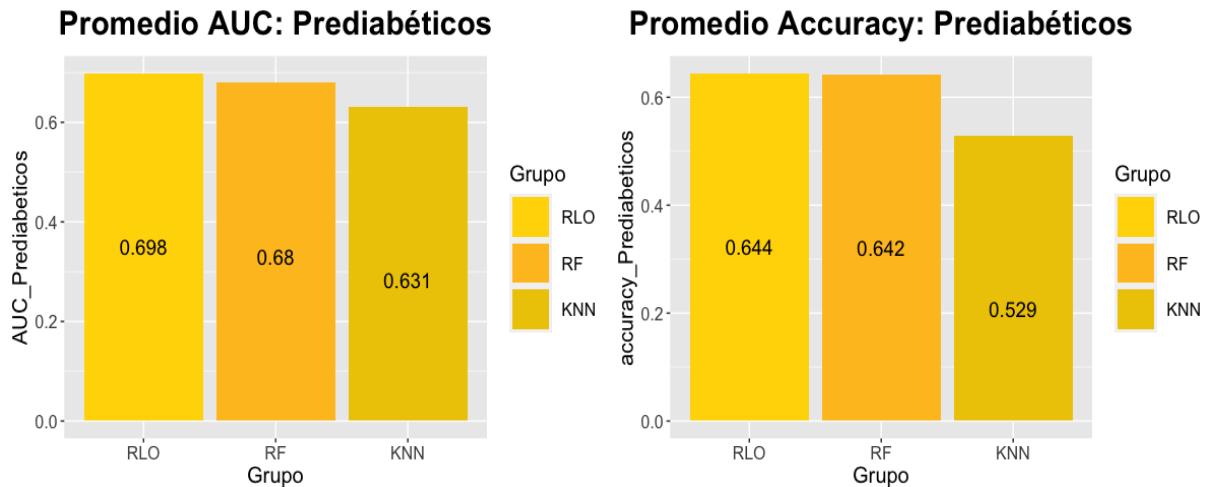
Anexo J. Proceso de selección hacia adelante en el grupo de Mujeres (elaboración propia).

El promedio de la validación cruzada para los conjuntos faltantes del caso de estudio 1 se pueden observar en el Anexo K y L, la comparativa muestra que en el grupo de No diabéticos el mejor clasificador fue random forest (RF) con 68.7% de área bajo la curva solo dos décimas encima de regresión logística (RLO), mientras que KNN obtuvo mejor desempeño en exactitud con 56.9%, esto significa que es más estable en las predicciones.



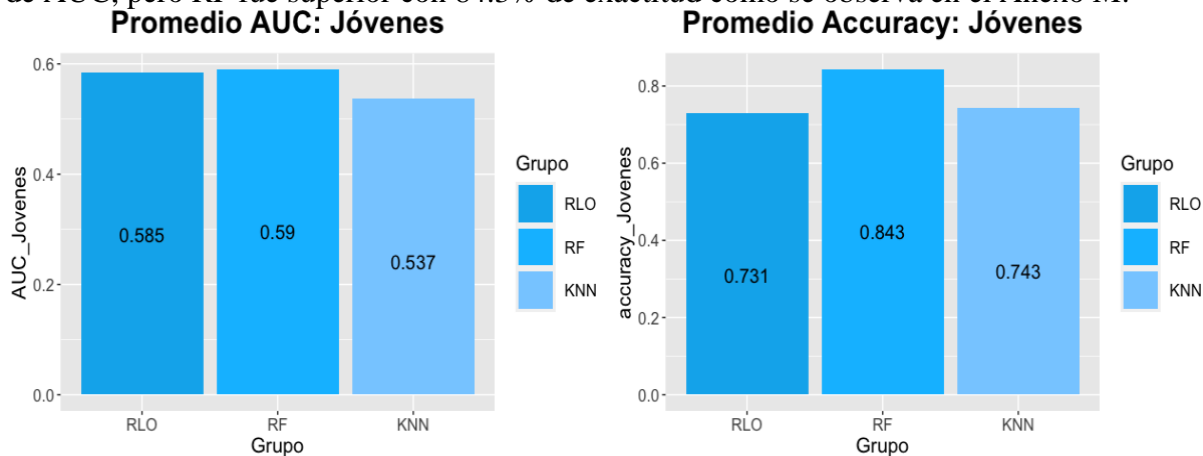
Anexo K. Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de No diabéticos (elaboración propia).

El algoritmo que mostro mejores predicciones para el caso de Prediabéticos fue regresión logística en ambas métricas (AUC y Accuracy) con 69.8% y 64.4% respectivamente.



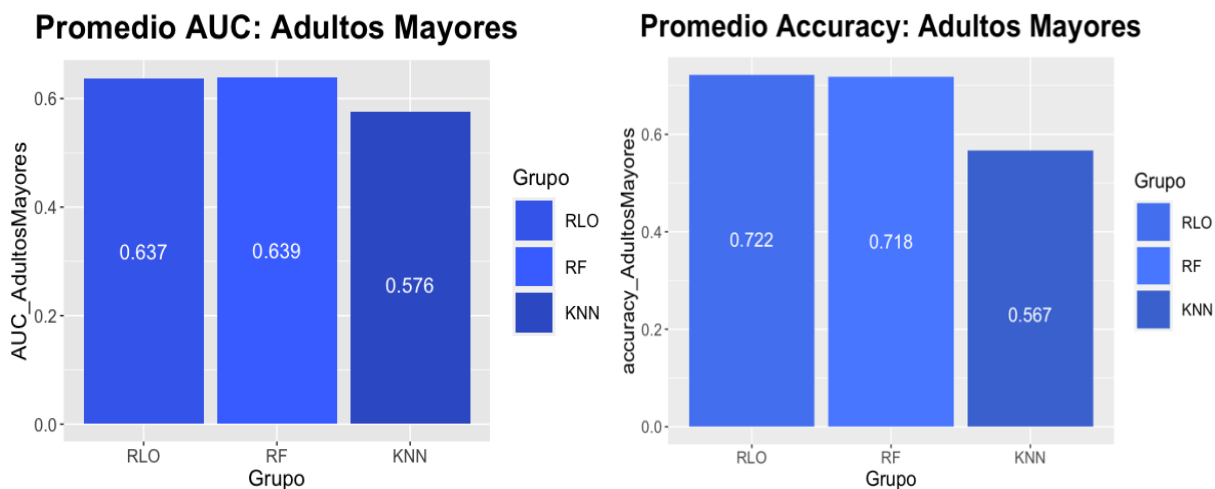
Anexo L. Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de Prediabéticos (elaboración propia)..

Nuevamente el modelo de RLO mostró mejor desempeño para el caso de Jóvenes con el 58.5% de AUC, pero RF fue superior con 84.3% de exactitud como se observa en el Anexo M.



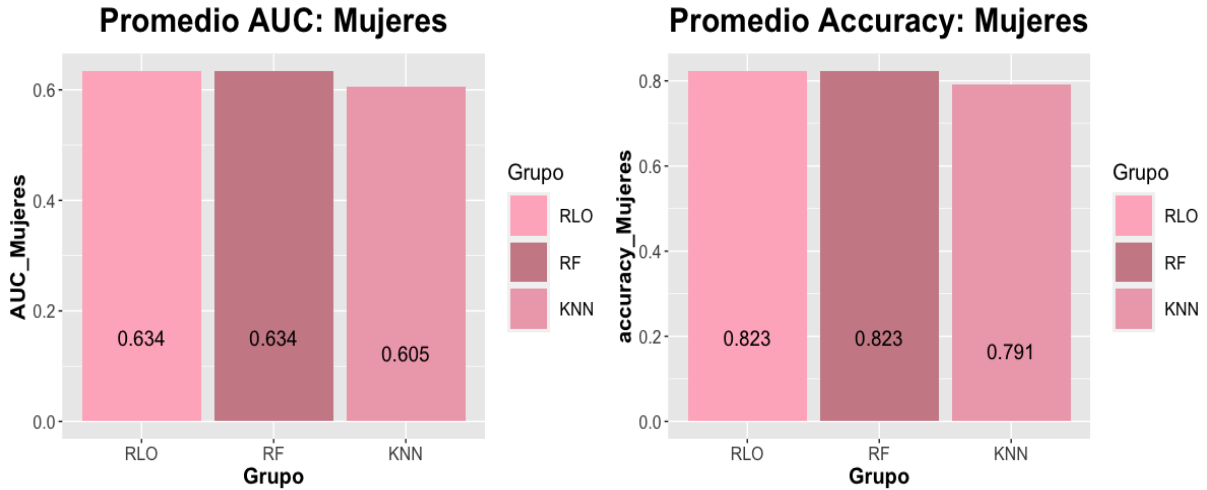
Anexo M. Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de Jóvenes (elaboración propia).

En el caso de Adultos mayores, random forest obtuvo mayor área bajo la curva con el 63.9% y regresión logística presentó 72.2 % de accuracy (ver Anexo N).



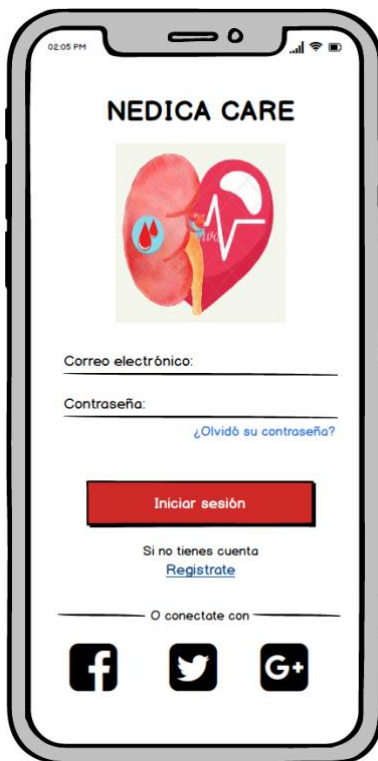
Anexo N. Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de Adultos mayores (elaboración propia).

Para este caso en particular (Mujeres), el desempeño de las predicciones generadas por los algoritmos, tanto para RLO como RF obtienen el 63.4% y 82.3% de AUC y accuracy respectivamente (ver Anexo O). Mientras que para todos los casos anteriores KNN muestra que sus clasificaciones se mantienen más estables y el desempeño es un poco menor comparado con los demás modelos.

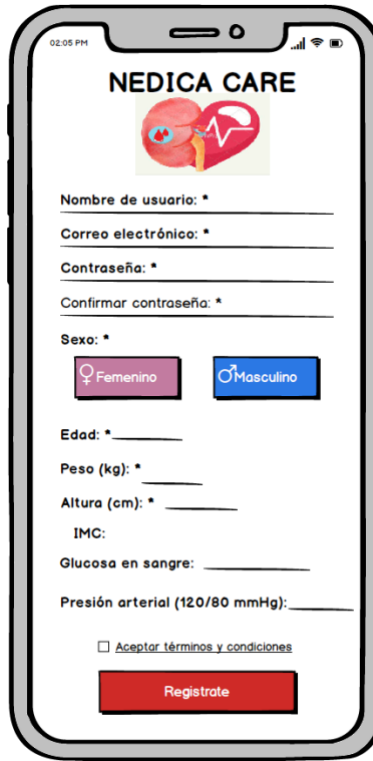


Anexo O. Comparativa de modelos de predicción bajo la métrica de AUC y Accuracy con RLO, RF y KNN en el grupo de Mujeres (elaboración propia).

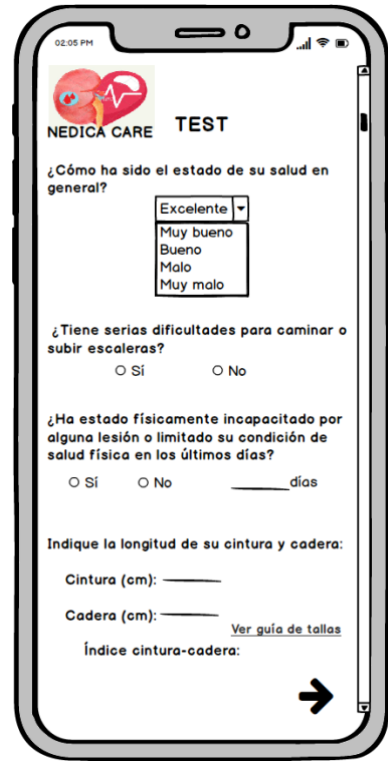
A continuación, en los Anexos P hasta X se muestra el maquetado del prototipo de baja calidad (elaboración propia) trabajado en el proyecto con la Universidad Internacional de La Rioja en Colombia.



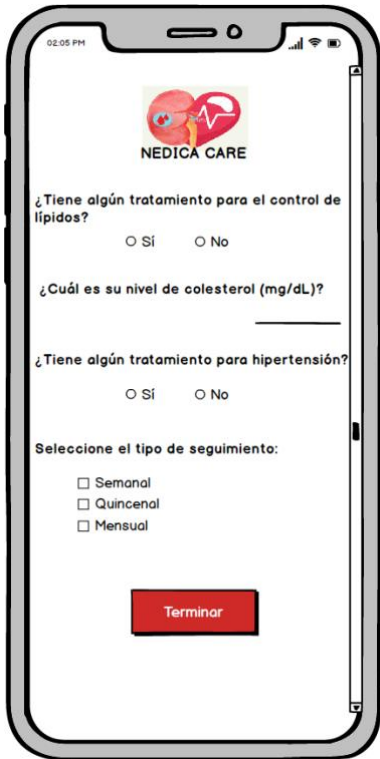
Anexo P. Iniciar sesión.



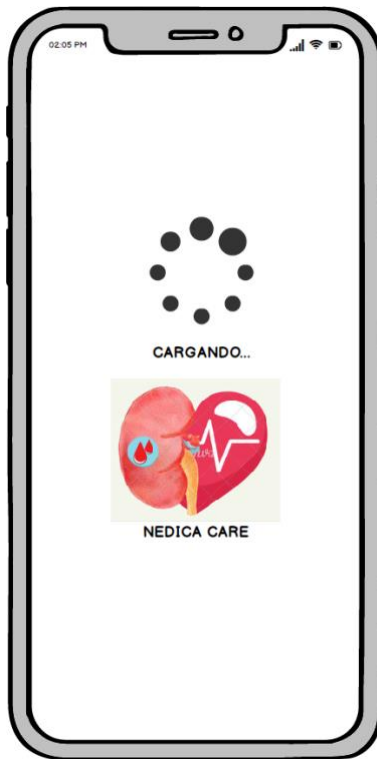
Anexo Q. Registro.



Anexo R. Test parte 1.



Anexo S. Test parte 2.



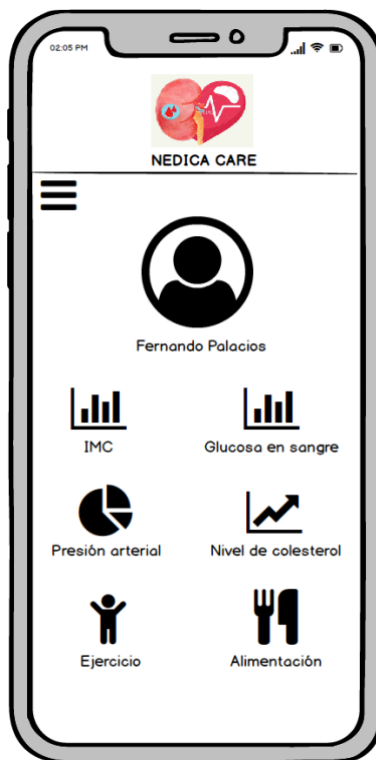
Anexo T. Procesando los datos.



Anexo U. Datos incompletos.



Anexo V. Resultados finales.



Anexo W. Pantalla principal.



Anexo X. Menú de opciones.

Se obtuvo una constancia en el Congreso Internacional de Ingenierías, Ciencias y Tecnología Aplicada, CIICTA 2021 (ver Anexo Y), por la participación de coautoría en el artículo “Implementación de Algoritmos de Aprendizaje Automático para la Predicción de pacientes Diabéticos” [72].



Anexo Y. Constancia de participación en el Congreso Internacional de Ingenierías, Ciencias y Tecnología Aplicada (CIICTA) en 2021.

De igual manera, se obtuvo un certificado de participación (ver Anexo Z) en el Congreso Mexicano de Inteligencia Artificial (COMIA) por la presentación del artículo “Predicción de Enfermedades Cardíacas Derivadas de Diabetes, mediante Algoritmos Genéticos: Caso de Estudio” [73].



Anexo Z. Certificado de participación en el Congreso Mexicano de Inteligencia Artificial (COMIA) en 2022.