Universidad Autónoma de Zacatecas
"Francisco García Salinas"
Unidad Academica de Ingenieria Electrica
Doctorado en Ciencias de la Ingeniería
(DOCII)

Desarrollo de Biomarcadores para la Detección de Diabetes Mellitus Tipo 2 y sus Comorbilidades por Medio de Modelos de Machine Learning

Que como parte de los requisitos
para obtener el grado de:

Doctor en Ciencias de la Ingeniería

presenta:
Jorge Alejandro Morgan Benita

Zacatecas, Zac. Agosto de 2024

Universidad Autónoma de Zacatecas

"Francisco García Salinas"

Unidad Academica de Ingenieria Electrica

Doctorado en Ciencias de la Ingeniería

(DOCII)

Desarrollo de Biomarcadores para la Detección de Diabetes Mellitus Tipo 2 y sus Comorbilidades por Medio de Modelos de Machine Learning

Que como parte de los requisitos
para obtener el grado de:

Doctor en Ciencias de la Ingeniería

presenta:

Jorge Alejandro Morgan Benita

dirigido por:

Dr. José María Celaya Padilla

Sinodales:

Dr. José María Celaya Padilla **Director de tesis**	Firma
Dr. Huizilopoztli Luna García **Co-director de tesis**	Firma
Dr. Carlos Eric Galván Tejada **Co-director de tesis**	Firma
Dra. Erika Yanneth Acosta Cruz **Vocal**	Firma
Dr. Klinge Orlando Villalba Condori **Vocal**	Firma

Dr. Jorge Isaac Galván Tejada
Unidad Academica de Ingeniría
Director

Dr. José Ismael de la Rosa Vargas
Doctorado en Ciencias de la Ingeniería
Responsable del programa

Zacatecas, Zac. Agosto de 2024 México

DOCTORADO EN CIENCIAS DE LA INGENIERÍA

CIRCULAR

Se suplica de la manera más atenta a los señores Doctores que se citan al calce, se sirvan revisar y en su caso aprobar la impresión del **TRABAJO DE TESIS** que se anexa a la presente, del pasante de Doctorado en Ciencias de la Ingeniería C. **JORGE ALEJANDRO MORGAN BENITA** con matrícula 42105044.

A T E N T A M E N T E

Zacatecas, Zac. 25 de noviembre 2024

DR. JOSÉ ISMAEL DE LA ROSA VARGAS
RESPONSABLE DEL PROGRAMA
DOCTORADO EN CIENCIAS DE LA INGENIERÍA

REVISÓ Y APROBÓ

DR. JOSÉ MARÍA CELAYA PADILLA

DR. HUIZILOPOZTLI LUNA GARCIA

DR. CARLOS ERIC GALVÁN TEJADA

DRA. ERIKA YANNETH ACOSTA CRUZ.

DR. KLINGE ORLANDO VILLALBA CONDORI

Carta de similitud núm. 722/IyP
Zacatecas, Zacatecas 11/octubre/2024

Dr. José Ismael de la Rosa Vargas
Responsable del DCI – UAZ
Presente

Estimado Dr. Ismael,

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

"Desarrollo de Biomarcadores para la detección de Diabetes Mellitus Tipo 2 y sus Comorbilidades por Medio de Modelos de Machine Learning"
de Jorge Alejandro Morgan Benita

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

15 % 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"

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

C. JORGE ALEJANDRO MORGAN BENITA.
P R E S E N T E:

De acuerdo al oficio de fecha 27 de mayo del 2024, en el cual solicita se le señale el tema a desarrollar para su trabajo de tesis del Programa de Doctorado en Ciencias de la Ingeniería, le manifiesto a Usted lo siguiente:

Se aprueba su solicitud, designando como directores de tesis a los profesores **DR. DR. JOSÉ MARÍA CELAYA PADILLA Y HUIZILOPOZTLI LUNA GARCÍA** mismo que acordaron acordes en fijar a usted el tema titulado:

"DESARROLLO DE BIOMARCADORES PARA LA DETECCIÓN DE DIABETES MELLITUS TIPO 2 Y SUS COMORBILIDADES POR MEDIO DE MODELOS DE MACHINE LEARNING",

Le comunico a Usted que dispone de un plazo máximo de **SEIS MESES,** a partir de la presente fecha para la conclusión de su trabajo de tesis. Asimismo, le indico que, una vez concluido su documento de tesis deberá remitir cinco copias del mismo para su revisión y aprobación por parte de la Comisión Revisora, que se nombrará en su oportunidad para su aprobación, o en su caso indicarle las correcciones que fueran pertinentes, antes de la impresión de la versión final.

A T E N T A M E N T E

Zacatecas, Zac. 25 de Noviembre del 2024.

UNIDAD ACADÉMICA DE INGENIERÍA ELÉCTRICA U.A.Z.

DR. JORGE ISSAC GALVÁN TEJADA
DIRECTOR DE LA UNIDAD ACADÉMICA DE INGENIERÍA ELÉCTRICA.

C. JORGE ALEJANDRO MORGAN BENITA.
P R E S E N T E.

La Dirección de la Unidad Académica de Ingeniería Eléctrica le notifica a Usted que la Comisión Revisora de su documento de Tesis de Doctorado, está integrada por los profesores Dr. José María Celaya Padilla, Dr. Huizilipoztli Luna García, Dr. Carlos Eric Galván Tejada, Dra. Erika Yanneth Acosta Cruz y Dr. Klinge Orlando Villalba Condori, ha concluido la revisión del mismo y ha dado la aprobación para su respectiva presentación.

Por lo anterior, se le autoriza la impresión definitiva de su documento de Tesis de Doctorado a fin de dar trámite a la sustentación de su Examen de Grado, a presentarse el 28 de noviembre del 2024.

A T E N T A M E N T E .
Zacatecas, Zac., 25 de Noviembre 2024.

UNIDAD ACADÉMICA DE INGENIERÍA ELÉCTRICA U.A.Z.

DR. JORGE ISSAC GALVÁN TEJADA.
DIRECTOR DE LA UNIDAD ACADÉMICA DE INGENIERÍA ELÉCTRICA

LIC. SAMANTA DECIRÉ BERNAL AYALA.
JEFA DEL DPTO. DE SERVICIOS ESCOLARES
P R E S EN T E:

Por este medio hacemos de su conocimiento, que la Tesis denominada " **DESARROLLO DE BIOMARCADORES PARA LA DETECCIÓN DE DIABETES MELLITUS TIPO 2 Y SUS COMORBILIDADES POR MEDIO DE MODELOS DE MACHINE LEARNING",** está **revisada y aprobada** para su impresión, misma que desarrolló el alumno del Programa de Doctorado en Ciencias de la Ingeniería que se menciona a continuación**: JORGE ALEJANDRO MORGAN BENITA.**

Sin otro particular de momento nos despedimos de Usted con un saludo cordial.

A T E N T A M E N T E
Zacatecas, Zac., 25 Noviembre del 2024.

REVISÓ Y APROBÓ

DR. JOSÉ MARIA CELAYA PADILLA.
ASESOR DE TESIS.

Resumen

La diabetes mellitus tipo 2 (DMT2) es una de las enfermedades más comunes que existen en la actualidad, su detección y tratamiento temprano son cruciales para reducir el riesgo a padecer complicaciones o comorbilidades asociadas. En esta tesis se presentan tres casos de estudio en los que se analizan diversos conjuntos de datos para la detección de diabetes, su progresión, sus complicaciones y comorbilidades por medio del desarrollo de biomarcadores potenciales, incluyendo datos antropométricos, clínicos, metabolómica y la diferenciación de estos por sexo incluyendo rangos específicos para cada característica.

II

Abstract

Type 2 diabetes mellitus (T2DM) is one of the most common diseases that exist today, its early detection and treatment are crucial to reduce the risk of suffering from complications or associated comorbidities. This thesis presents three case studies in which various data sets are analyzed for the detection of diabetes, its progression, its complications and comorbidities through the development of potential biomarkers, including anthropometric data, clinical data, metabolomics and the differentiation of these by sex including specific ranges for each feature.

Dedicatoria

A todos los que han estado a mi lado desde el comienzo de cada primer paso y me han regalado su valioso tiempo para acompañarme en este camino... Gracias por todo.

VI

Agradecimientos

Agradezco a mi director el Dr. José María Celaya Padilla por sus innumerables aportes a mi trabajo. Su inteligencia, carácter y carisma me proporcionaron todo un modelo a seguir. Doctor, sus consejos y apoyo los llevaré y recordaré siempre, ¡Muchas gracias!

Agradezco a mi codirector el Dr. Huizilopoztli Luna García por todo el apoyo brindado a lo largo de estos 3 años, gracias a usted conocimos lugares y personas increíbles, valoramos el tiempo dedicado a todas esas actividades que nos han puesto en el mapa en más de una ocasión, también, el incorporarnos a un grupo de personas en el que construimos una serie de relaciones de colaboración bastante productivas. El apoyo brindado en todos los ámbitos lo tendré siempre en cuenta doctor, ¡Muchas gracias!

Al Dr. Carlos Eric Galván Tejada por el apoyo brindado desde el principio, sus observaciones y conocimiento, me han servido en numerosas ocasiones haciendo mi trabajo más profesional y me ha proporcionado una serie de objetivos específicos para trazar mi ruta en el contexto académico. Estoy muy agradecido sobre todo, de que me haya dado la oportunidad de ser parte del programa que dirige. ¡Mil gracias!

A la Dra. Erika Yanneth Acosta Cruz a quien a pesar de que conocí hace poco, me ha proporcionado una serie de observaciones bastante valiosas para hacer de este trabajo algo de mucha más calidad. ¡Gracias!

Al Dr. Klinge Orlando Villalba-Condori a quien a pesar de que no conozco en persona, nos ha brindado el apoyo en más de una ocasión, haciendo de nuestra última publicación una realidad. ¡Gracias!

Agradezco ampliamente el apoyo brindado por el Conahcyt, gracias al cual pude dedicarme por completo a algo que actualmente me apasiona y me dio la oportunidad de conocer personas grandiosas.

Agradezco al Dr. Hamurabi Gamboa Rosales y al Consejo Zacatecano de Ciencia Tecnología e innovación por todo el apoyo brindado en diversas áreas, presentando una perspectiva de los retos que existen tanto dentro como fuera del contexto académico. ¡Gracias!

Por último pero no menos importante, agradezco a mis paperamigos los doctores Gaby y Carlitos, que decirles chicos, sin ustedes la verdad que ahorita todavía estaría batallando con las publicaciones, sus aportes a mi trabajo siempre los tendré presentes, pero más aún un placer llamarlos amigos y colegas. ¡Gracias!

Índice general

Índice de figuras

Índice de tablas

Capítulo 1
Introducción

La diabetes mellitus tipo 2 (DMT2) es una de las enfermedades más comunes que aquejan a la población mundial. Según el informe más reciente de la Federación Internacional de Diabetes (IDF por sus siglas en inglés), la incidencia global de DMT2 entre adultos fue de 536.6 millones de personas en el año 2021, lo que constituye aproximadamente el 10.5 % de la población mundial. Se estima que para el año 2045, habrá 783.2 millones de personas, o el 12.2 % de la población mundial, viviendo con diabetes (IDF, 2022b). En el año 2021 se reportaron más de 6.7 millones de muertes, lo que equivale a aproximadamente una muerte cada 5 segundos (IDF, 2022a).

México ocupa el puesto 25 a nivel mundial con 13.5 millones de adultos con diabetes (World Population Review, 2021). Según los datos de mortalidad del 2021 proporcionados por el Instituto Nacional de Estadística y Geografía (INEGI), se registraron 142,546 muertes atribuidas a la diabetes mellitus, de las cuales 72,324 corresponden a hombres y 70,219 mujeres, más 3 no especificadas, las cuales se situaron en el rango de 0 a 85 años. Dentro de este amplio rango de edades, las cifras en el grupo de 55 años en adelante superan las 10,000 muertes (INEGI, 2022). Hubo una tendencia al alza en la tasa de mortalidad en México durante el periodo comprendido entre 2011 y 2016, seguido de un descenso de 2016 a 2019 y, de nuevo, un incremento en 2020-2021, lo que demuestra el creciente impacto de la diabetes en la sociedad (INEGI, 2021).

La obesidad es uno de los factores de riesgo más comunes de la DMT2 según el observatorio global de la obesidad (RisC, 2022), sin embargo, no es determinante para un diagnóstico, por lo que es necesario identificar y desarrollar biomarcadores que proporcionen un camino inicial y una serie de recomendaciones sustentadas y avaladas por la evidencia analizada, tanto para pacientes como para la población mexicana en general. Tomando en cuenta que la población total Mexicana (alrededor de 130 millones en 2021 (World Population Review, 2021)) y los 2 reportes trimestrales de este año 2021 por parte del Sistema de Vigilancia Epidemiológica Hospitalaria de DMT2 se tiene que la cantidad de casos confir-

mados de DMT2 es la mínima parte de esta, 3,831 casos el primer trimestre 2021 y 9,636 casos el segundo trimestre de 2021 (DVEENT, 2021), dando menos del 1 %. Aún con el aparentemente bajo número de descensos, México es uno de los países con mayor índice de obesidad en el mundo presentándose en adultos hombres una tasa de obesidad del 32.22 %, en mujeres de un 41.28 %, en niños un 20.29 % y en niñas un 14.97 %, posicionándose mundialmente entre los primeros 50 países.

La enfermedad de la diabetes puede ser detectada a través de una serie de biomarcadores biológicos, extraídos de diversas pruebas de laboratorio o toma de medidas antropométricas en consultorio y que, dependiendo del tipo de diabetes que se trate, presentan indicadores o puntos de partida para proporcionar un diagnóstico y un tratamiento adecuado. La clasificación establecida por la OMS identifica tres tipos principales de diabetes: la tipo 1, la tipo 2 y la gestacional (WHO, 2022).

Además de estos biomarcadores biológicos, la comunidad científica y tecnológica ha explorado nuevas herramientas para mejorar la detección de la DMT2, entre ellas el uso de técnicas de aprendizaje automático o Machine Learning (ML) y con ello el desarrollo de biomarcadores bioinformáticos o digitales, los cuales a través del análisis y tratamiento de datos específicos, se identifican patrones y relevancia estadística proporcionando una herramienta valiosa en el apoyo diagnóstico. Estas herramientas con aplicaciones bioinformáticas permiten analizar datos clínicos y biológicos de pacientes para desarrollar predicciones y apoyo diagnóstico de manera temprana a enfermedades como la DMT2 (Iglay *et al.*, 2016). La aplicación de algoritmos de ML en el análisis de historiales clínicos permite extraer relaciones, predicciones y patrones de comportamiento que pueden ser utilizados para clasificar o proporcionar valores estadísticos relevantes, dando lugar a indicadores o biomarcadores que pueden sugerir la posibilidad de padecer la DMT2 con un nivel de precisión variable según el modelo o conjunto de modelos empleados (Ngiam and Khor, 2019).

En contraste con el desarrollo de biomarcadores químicos que requieren procedimientos de laboratorio, el desarrollo de biomarcadores mediante ML se beneficia de la disponibilidad de grandes conjuntos de datos clínicos, lo que hace que el proceso sea más económico, rápido y escalable a nivel poblacional. En esta investigación, se excluyen las características ya establecidas como biomarcadores efectivos: los niveles de glucosa en sangre o los resultados de las pruebas FPG, OGTT o HbA1c (Kavakiotis *et al.*, 2017). Sin embargo, estas características se mantienen como puntos de referencia para evaluar la efectividad de los nuevos biomarcadores potenciales propuestos, esto incluye todas aquellas características que podrían no tener una relación directa con la diabetes por sí mismas, pero que podrían mejorar la efectividad de los modelos en conjuntos y que estas, a su

vez, estén disponibles en las bases de datos analizadas.

En el ámbito médico, la aceptación y aplicación de modelos de ML requieren alta precisión, generalmente establecida en un mínimo del 90 % de área bajo la curva (AUC por sus siglas en inglés) (García-Carretero *et al.*, 2021). Este nivel de precisión determina si un conjunto de características analizadas puede considerarse un posible biomarcador, ya sea de manera individual o en combinación con otros biomarcadores, lo que resulta en un aporte de apoyo diagnóstico seguro y confiable. Para alcanzar este nivel de precisión, es necesario utilizar una variedad de modelos y realizar pruebas comparativas de métricas como la sensibilidad y especificidad. Esto implica la combinación, integración, normalización y estandarización de pruebas, así como el cálculo y la comparación del AUC, la construcción de curvas ROC (por sus siglas en inglés de Receiver Operating Characteristic, que se traduce como característica operativa del receptor en español), matrices de confusión y otras herramientas estadísticas integradas en la investigación. Además, se requiere una base de datos correctamente balanceada con datos relevantes y características cuidadosamente seleccionadas por personal experto en el área. La calidad de la base de datos influye significativamente en la capacidad de estimar el riesgo de desarrollar la enfermedad. En el contexto de la detección de la DMT2, la mejora continua de la precisión de los modelos de inteligencia artificial (IA) es esencial (Hamdi *et al.*, 2018). Esta mejora se logra mediante la identificación o selección de características y la integración y/o combinación de modelos existentes o generados a la medida de las necesidades del experimento, con el propósito de identificar con mayor exactitud a las personas susceptibles de desarrollar la enfermedad y evaluar el riesgo de complicaciones a corto o mediano plazo. En el proceso de análisis de la información de las bases de datos proporcionadas, se emplean parámetros, indicadores, tendencias y patrones para desarrollar un biomarcador predictivo de la enfermedad de la DMT2 y su progresión hacia otras enfermedades ya sean complicaciones o comorbilidades asociadas.

En este trabajo de investigación se planteó identificar qué biomarcadores son los más relevantes para el desarrollo de la enfermedad y analizar información actualizada, organizada y categorizada adecuadamente para implementar un modelo de ML comparable, replicable y estructurado, además se identifican los rangos correspondientes a cada característica de manera individual, realizando comparativas por sexo. Esto se lleva a cabo utilizando una muestra representativa de la población mexicana contenida en las bases de datos del Hospital Siglo XXI y el laboratorio de Metabolómica y Proteómica de la Universidad Autónoma de Zacatecas. Con estas bases de datos, se identificaron los factores de mayor relevancia para determinar el nivel de riesgo de padecer la enfermedad, su progresión, sus complicaciones y comorbilidades asociadas. Se seleccionaron e implementaron

algoritmos de ML apropiados, los cuales se integraron y combinaron para obtener biomarcadores potenciales que mantengan un alto grado de precisión o que se acerquen a él. Este enfoque permitirá avanzar en la detección temprana y precisa de la DMT2, contribuyendo así a una mejor atención médica y prevención de complicaciones asociadas a esta enfermedad. La implementación de nuevos parámetros de control y una adecuada categorización y registro en México es esencial para identificar tanto a los individuos aparentemente saludables como a aquellos con un alto riesgo de padecer DMT2. Esto permitiría identificar a personas que, con los cuidados adecuados, pueden evitar o minimizar el riesgo de desarrollar la enfermedad. Este enfoque además, presenta numerosos beneficios para las instituciones de salud, como la reducción del uso de medicamentos, hospitalizaciones y exámenes asociados al tratamiento de la enfermedad. Además, representa un importante ahorro económico y una mejora significativa en la calidad de vida del adulto mexicano promedio. La formalización de pre-diabéticos y personas con alto riesgo de DMT2 permite una intervención temprana y personalizada, enfocada en la prevención y el control de los factores de riesgo. Esto incluye cambios en el estilo de vida, como la dieta y el ejercicio, así como el seguimiento regular y la educación sobre la enfermedad. Un registro adecuado también facilita la investigación epidemiológica y el desarrollo de políticas de salud pública dirigidas a la prevención y el manejo de la DMT2 en México. En resumen, establecer nuevos parámetros de control y una categorización y registro adecuados en México sería de suma importancia para abordar de manera efectiva el creciente problema de la diabetes en el país y mejorar la salud y el bienestar de la población.

1.1. Planteamiento del problema

La efectividad de los modelos de ML depende de las técnicas de selección de características implementadas, los criterios de inclusión y exclusión presentes en el experimento, el conjunto de modelos utilizados, la base de datos que requiere validación por parte de un comité médico y de sujetarse a protocolos internacionales como el de Helsinki, antes de su aplicación clínica directa. Uno de los desafíos de los biomarcadores informáticos para la detección de la diabetes reside en lograr modelos de alta precisión similares a la que proporciona el biomarcador de la glucosa. Este desafío puede ser difícil de alcanzar debido a diversas razones:

- Limitaciones en el presupuesto de investigación para la extracción de otro tipo de muestras (Sadoughi *et al.*, 2016).
- Acceso limitado a equipos sofisticados para el entrenamiento de modelos

robustos (Boutilier *et al.*, 2021).

- Falta de conocimiento en el desarrollo de modelos avanzados de ML (Tuppad and Patil, 2022).
- La complejidad de los datos médicos y los protocolos de recolección (Lipscombe *et al.*, 2018).
- Protección de información privada (Klonoff and Price, 2016).

Cabe destacar que la identificación de biomarcadores potenciales para la DMT2 va más allá de simplemente comparar estos indicadores con los niveles de glucosa. Hay muchas otras características y biomarcadores que pueden estar relacionados con la enfermedad, y explorar estas diferentes características puede requerir más recursos y tiempo de investigación.

El tiempo que los médicos dedican a analizar a cada paciente puede variar según diversos factores, y como seres humanos, están sujetos a cometer errores y tienen limitaciones en términos de energía de actividad o fisiológica. Además, el sistema actual de pruebas para la detección de la DMT2, que depende principalmente del biomarcador de glucosa, puede resultar en diagnósticos lentos y poco optimizados debido a su dependencia de pruebas de laboratorio o pruebas desechables. ML tiene el potencial de reducir significativamente los tiempos de diagnóstico y ayudar a los médicos a realizar esta función de manera más rápida y precisa, no solo para un paciente, sino para un conjunto de pacientes. Actualmente el uso de ML en la práctica médica aún no está ampliamente adoptado o normalizado, y existen varias razones para ello, como la falta de educación, o actualización sobre las capacidades y beneficios del ML, la resistencia al cambio en el sistema de atención médica, la falta de acceso a datos de calidad o la necesidad de validación clínica de los resultados presentados por los modelos de ML.

La DMT2 presenta desafíos importantes desde múltiples perspectivas:

- En el contexto social, los costos asociados a la atención médica de calidad son evidentes. Si bien las pruebas y los análisis de laboratorio para la detección de la DMT2 pueden no ser excesivamente costosos para la media poblacional, las demoras en la implementación y el acceso limitado a la atención o el seguimiento adecuado, pueden dar lugar a complicaciones que si requieren de pruebas y atención más especializada, siendo por ende, más costosa. Otro obstáculo importante para el manejo eficaz de la DMT2 es la ausencia de diagnósticos y enfoques terapéuticos personalizados, las técnicas de diagnóstico temprano más avanzadas tienden a ser muy costosas, pues requieren equipo especializado, esto sin contar con el seguimiento

por parte de médicos especialistas en endocrinología o ya para la atención de comorbilidades en el área de la cardiología, la oftalmología, nutrición o incluso psicología reduciendo las opciones para la detección temprana y las intervenciones oportunas, en particular en áreas con recursos limitados (Pérez-Lozano *et al.*, 2023).

- Por el lado de la bioinformática, tenemos que la integración de la glucosa como característica principal en los modelos de aprendizaje automático para detectar DMT2 a menudo conduce a un sobreajuste, donde los modelos se vuelven demasiado dependientes de esta única variable. Esto no sólo reduce la generalización de los modelos, sino que también descuida otras características predictivas potenciales, como los marcadores metabólicos o antropométricos, que podrían mejorar la precisión y amplitud de las predicciones de diabetes, complicación y comorbilidad, tendientes a la medicina personalizada (Mejía *et al.*, 2023).

- Por último en el eje de sistemas inteligentes identificamos que el uso generalizado de modelos de ML para el análisis de datos en el sector sanitario subraya la necesidad de experimentos sólidos y replicables. La falta de estándares de diseño consistentes en las aplicaciones de IA conduce a distintos grados de éxito en la detección y predicción de DMT2. Para abordar estos desafíos es esencial crear experimentos de IA centrados en el análisis de datos inteligentes, empleando técnicas integrales de selección de características e incorporando diversos modelos para mejorar el poder predictivo y garantizar que las herramientas de diagnóstico puedan ser precisas y replicables en diferentes poblaciones y conjuntos de datos (Pedrero *et al.*, 2021).

1.2. Antecedentes

La antropometría humana posee una historia extensa y compleja, con sus raíces conceptuales remontándose a trabajos como la obra anatomofisiológica de Claudio Galeno de Pérgamo en la antigua Grecia (De Cervantes, 2024). Galeno realizó contribuciones fundamentales al conocimiento del cuerpo humano, sentando las bases para la observación sistemática y el estudio de las proporciones anatómicas. Sin embargo, las aplicaciones clínicas de la antropometría, tal como las conocemos hoy, encuentran sus fundamentos en trabajos mucho más recientes. Uno de los pioneros en establecer una base científica para la antropometría moderna fue el naturalista alemán Johann Sigismund Elsholtz, quien en el siglo XVII presentó su tesis titulada “Anthropometria”, la cual se convirtió en uno de

los primeros textos de referencia en la materia (Albrizio, 2007). Elsholtz es reconocido por ser uno de los primeros en estudiar y cuantificar las dimensiones corporales en un contexto clínico y científico, un esfuerzo que posteriormente influiría en el desarrollo de métodos para evaluar la salud y el estado físico a través de medidas antropométricas. Su obra sentó precedentes en la relación entre las características físicas y el diagnóstico médico, avanzando hacia una comprensión más profunda de cómo la estructura y las proporciones del cuerpo pueden reflejar condiciones de salud específicas. A lo largo de los siglos, derivado de múltiples estudios en biometría y la identificación de una variedad de proteínas, moléculas y perfiles bioquímicos, se han ajustado y refinado las métricas antropométricas. Estas métricas determinantes han evolucionado con el propósito de optimizar el diagnóstico y la prevención de enfermedades, permitiendo que la antropometría no solo sea una medida de las proporciones físicas, sino también una herramienta esencial para la identificación temprana de factores de riesgo en enfermedades metabólicas, cardiovasculares y otras afecciones.

Uno de los datos que se ha tomado como biomarcador clave para la diabetes durante mas de 50 años es la HbA1c descubierta en los 60s. El precursor de lo que hoy conocemos como HbA1c se llama Rahbar (2005) que identificó la subfracción de la hemoglobina glicada A1c en pacientes diabéticos. Esta, apareció como una ”banda de hemoglobina anormal de movimiento rápido”menor en pacientes diabéticos durante los exámenes de rutina para detectar variantes de hemoglobina. Posteriormente, este descubrimiento resultó ser un marcador biomolecular importante con implicaciones clínicas y patológicas. Desde entonces, la medición de la HbA1c en pacientes diabéticos se ha convertido en un método establecido para evaluar el control de la diabetes a largo plazo, lo que marca un avance importante en la calidad de la atención al paciente diabético en este siglo. En particular, la HbA1c fue el primer ejemplo observado de glicación de proteínas no enzimática in vivo, presentando una propuesta sólida que luego derivaría en una investigación exhaustiva sobre las reacciones de Maillard dentro de los sistemas biológicos. Esto incluye la exploración de productos finales de glicación y lipoxidación avanzada (AGE/ALE), que contribuyen a las complicaciones de la diabetes y diversas enfermedades relacionadas con la edad. Este descubrimiento permanece vigente como uno de los más importantes en torno a la diabetes en la actualidad.

Desde la década de 1960, con el surgimiento del ML a finales de la década de 1950, se han logrado utilizar con éxito los primeros modelos de clasificación. Desde entonces, estos modelos han sido continuamente utilizados y mejorados a lo largo de los años con el propósito de apoyar al desarrollo de herramientas y técnicas para el diagnóstico médico (Kononenko, 2001). Principalmente y antes del ejercicio de recolección de muestras o de información debemos atender

primeramente qué datos se requiere obtener para para poderlos procesar con un sentido específico. Además debemos comprender que los procesos de recolección de datos para su posterior análisis con ML han sido variados, desde muy simples como una entrevista breve hasta altamente técnicos como el análisis de metabolitos. En el caso específico de la DMT2, los modelos de ML pueden utilizar datos clínicos como la HbA1c, datos biométricos como la presión arterial o datos antropométricos como el índice de masa corporal (BMI por sus siglas en inglés), antecedentes familiares y otros datos de interés para predecir la probabilidad de desarrollar la enfermedad. Estos modelos han surgido a lo largo de los años integrando algoritmos capaces de aprender a distinguir entre diferentes perfiles y hacer predicciones precisas (Cahn *et al.*, 2020). Todas las características y rangos resultantes analizados por estos modelos se conocen como biomarcadores bioinformáticos. Los primeros modelos de ML pudieron ofrecer predicciones preliminares que luego fueron validadas por profesionales expertos en el área médica antes de tomar decisiones clínicas importantes. Las métricas de evaluación del modelo, como la *Sensitivity*, *Specificity* y el AUC, surgen poco después para proporcionar una medida de la precisión y confiabilidad de las predicciones, a partir de la validación observada por los expertos, proporcionando la retroalimentación necesaria para validar los resultados de estas métricas (Rainio *et al.*, 2024).

El campo de la metabolómica surgió a principios de la década de 2000, aprovechando tecnologías desarrolladas en genómica y proteómica. Se basa en técnicas analíticas sofisticadas como la espectroscopia de resonancia magnética nuclear, la espectrometría de masas y la cromatografía. Es un enfoque novedoso para el descubrimiento de biomarcadores. Junto con la genómica, la metabolómica puede proporcionar una comprensión sistémica de las causas subyacentes de la patología, lo que resalta su importancia en las ciencias clínicas y su capacidad para guiar intervenciones clínicas. La metabolómica es un enfoque potente y potencialmente de alto rendimiento para el descubrimiento de biomarcadores a nivel molecular, aunque su proceso es lento. Para una adaptación más rápida del descubrimiento de biomarcadores, se han empleado tecnologías portátiles y dispositivos móviles, junto con extracción inteligente de datos, aprendizaje profundo en IA (Drupad *et al.*, 2017). Actualmente existen más de 200,000 metabolitos diferentes identificados en el cuerpo humano, según la base de datos del metaboloma humano 5.0 (Wishart *et al.*, 2021).

1.3. Justificación

La implementación de estrategias de detección temprana y tratamiento oportuno son primordiales para abordar este importante problema de salud pública

y mejorar la calidad de vida de la población afectada por la DMT2 en México. La detección temprana de la DMT2 puede ayudar a prevenir o retrasar la aparición de complicaciones (William H. *et al.*, 2015). La eficacia de la detección de la DMT2 puede conducir a una identificación más temprana de la enfermedad, un mejor control glucémico y una reducción de las tasas de complicaciones (Alieva *et al.*, 2022). Los objetivos actuales de la medicina moderna en torno a la diabetes buscan profundizar en la relación la información de perfiles clínicos con diagnóstico concluyente de DMT2 con los factores observables para ser comparados con otros perfiles clínicos de personas sin diagnóstico de DMT2 pero con factores de riesgo, como sobrepeso, colesterol LDL elevado o presión arterial alta, entre otros (Liao *et al.*, 2022). Los enfoques impulsados por ML para el descubrimiento de biomarcadores pueden acelerar la traducción de los hallazgos de la investigación a la práctica clínica al simplificar el proceso de validación y aprobación regulatoria. Al priorizar biomarcadores candidatos con alta precisión predictiva y relevancia clínica, los modelos de ML permiten una asignación eficiente de recursos y facilitan la toma de decisiones basada en evidencia en entornos de atención médica. Esto puede tener un impacto significativo en la mejora de la atención médica y la gestión de la salud de los pacientes con DMT2 y otras enfermedades crónicas.

La mayoría de la literatura revisada se centra en la implementación de modelos para la predicción de la DMT2 utilizando biomarcadores ampliamente conocidos en su análisis, sin embargo, el enfoque orientado al desarrollo de nuevos biomarcadores tendientes a la medicina personalizada, a través de la selección de características y el reconocimiento de rangos y diferencias por sexo dentro de esos biomarcadores desarrollados, ha sido poco explorado. Se sabe que ML ofrece una solución para manejar volúmenes masivos de datos al detectar patrones y realizar predicciones, sin embargo, para garantizar resultados novedosos o predicciones de alta calidad y usabilidad clínica, los datos deben ser respaldados por expertos en el campo y extraídos mediante técnicas clínicas estandarizadas. Además, estos conjuntos de datos se deben someter a escrutinio de protocolos adecuados avalados por una junta médica para su liberación y posterior uso en los experimentos o análisis para los cuales fueron creados. La cantidad de herramientas para aprovechar estos datos, aunque bastas, tienden a dejar sesgos o nichos de oportunidad que han llevado a la necesidad de incorporar nuevas metodologías y herramientas automáticas que permitan predecir y determinar si una persona desarrollará la enfermedad (predicción para la prevención y no detección para el tratamiento).

Los enfoques para justificar esta tesis basado en las limitantes encontradas en todos los trabajos relacionados presentes son:

1. Limitantes en la revisión de literatura: Una revisión exhaustiva de la literatura sobre la predicción de la DMT2 utilizando modelos de ML puede ser una

tarea ardua debido a la gran cantidad de trabajos relacionados con relevancia significativa en los últimos años sobre diabetes. Identificar los modelos y conjuntos de características que han demostrado ser efectivos en estudios previos dará una base sólida y confiable sobre este trabajo y ayudará a identificar posibles lagunas en la investigación existente.

2. La implementación de técnicas de selección de características diversas independientemente de la cantidad o calidad de los datos en el conjunto analizado: Se requiere profundizar en los conocimientos obtenidos de la revisión de literatura realizando distintas pruebas de selección de un conjunto de características relevantes para su estudio. Además de la prueba de FPG, OGTT o HbA1c para detección de la DMT2, hay que considerar la importancia de características antropométricas, y su intervención o interacción también otros biomarcadores y factores de riesgo relacionados con la DMT2, como la presión arterial y el perfil lipídico. Si bien es cierto que en la literatura se encontró relación entre características antropométricas y la enfermedad así como el papel del perfil lipídico en el desarrollo de la DMT2, no se encontró un artículo que combine o valide estas características en conjunto, así como también no se encontró una tabla o resultados que proporcionen un rango ni biomarcadores diversos a la glucosa que en conjunto predicen la DMT2.

3. La existencia y desarrollo de distintos modelos de ML basados en limitados conjuntos de datos de dominio público: La base de datos más utilizada en la literatura es la denominada conjunto de datos PIMA, un dataset de 768 observaciones con 8 características, y una salida diagnóstico muestra un panorama bastante limitado tanto en antropometría como características clínicas, contando además con una característica en particular llamada "*Pregnancies*", integrando el número de embarazos del paciente. Esta característica puede producir una limitante, error o sesgo en los modelos debido a que en el dataset se encuentran tanto hombres como mujeres. Esta problemática presenta una oportunidad para realizar implementaciones en conjuntos de datos más robustos.

4. Exclusión de biomarcadores relacionados con la glucosa: Esta exclusión no está presente en la literatura revisada, por lo que se requiere un análisis sobre su exclusión para validar otras características, ya sean, antropométricas o clínicas, y que en conjunto puedan dar la misma eficacia o similar a la de los niveles de glucosa. Cabe señalar que el biomarcador de la hemoglobina glicada podría en ocasiones no ser adecuado para diagnosticar diabetes debido a su amplio rango normal, la falta de estandarización entre laboratorios y la susceptibilidad a afecciones que afectan la vida útil de los glóbulos

rojos. También existen personas que no pueden tomar la prueba como pacientes con ciertas condiciones que afectan los glóbulos rojos, como anemia, trastornos sanguíneos, insuficiencia renal o enfermedad hepática, pacientes que toman medicamentos que pueden provocar un aumento rápido de la glucosa, como esteroides o antipsicóticos, pacientes con síntomas recientes (menos de 2 meses al momento de la prueba), niños y jóvenes, pacientes con daño pancreático agudo, pacientes embarazadas, entre otros.

1.4. Hipótesis

La integración de modelos de ML, técnicas estadísticas y técnicas de selección de características utilizando datos antropométricos, biométricos y clínicos permite desarrollar un conjunto de rangos de valores en biomarcadores potenciales relacionados con la DMT2, sus complicaciones y sus comorbilidades. Estos biomarcadores posibilitan el desarrollo de herramientas de predicción y apoyo diagnóstico con una alta capacidad discriminatoria, representada por una tasa entre las predicciones correctas y las predicciones totales de 90 % o superior.

1.5. Objetivos de las preguntas de investigación

En esta sección se presentan los objetivos producto de la pregunta de investigación:

1. ¿Qué características clínicas, biométricas y antropométricas están más fuertemente asociadas con la aparición y progresión de la DMT2, sus complicaciones y sus comorbilidades relacionadas?

2. ¿Cuál es el impacto de las técnicas de preprocesamiento en la calidad, cantidad y usabilidad de los conjuntos de datos disponibles públicamente o de manera privada, para el análisis de la DMT2?

3. ¿Qué técnicas de selección de características son más efectivas para identificar las características clave que influyen en la progresión de la DMT2 y sus comorbilidades asociadas?

4. ¿Cómo se comparan los diferentes algoritmos de aprendizaje automático en términos de su capacidad para identificar y desarrollar biomarcadores confiables para predecir la DMT2, sus complicaciones y sus comorbilidades?

5. ¿Cuáles son las fortalezas y limitaciones de los modelos de aprendizaje automático para el desarrollo de biomarcadores de DMT2 cuando se evalúan utilizando métricas como precisión, sensibilidad, especificidad y puntuación F1?

6. ¿Cómo se comparan los biomarcadores derivados del aprendizaje automático para la DMT2 en cuanto a precisión y eficacia con las herramientas y pautas

clínicas establecidas para el diagnóstico y el tratamiento de la enfermedad, sus complicaciones y sus comorbilidades?

7. ¿Cuáles son las diferencias específicas de género en los valores de los biomarcadores para la detección temprana de la DMT2 y cómo influyen estas diferencias en el diagnóstico y la progresión de la enfermedad en pacientes masculinos y femeninos?

1.5.1. Objetivo general

Desarrollar un biomarcador para la detección de la DMT2, sus complicaciones y sus comorbilidades, a partir de datos clínicos, biométricos y antropométricos integrados a modelos de ML, que al ser evaluados presenten una tasa entre las predicciones correctas y las predicciones totales de 90 % o superior.

Objetivos específicos

1. Identificar y compilar una lista de por lo menos 10 características clínicas y antropométricas relevantes que puedan estar asociadas con el desarrollo y la progresión de la DMT2, sus complicaciones y sus comorbilidades.

2. Preprocesar al menos 3 conjuntos de datos proporcionados u obtenidos públicamente, utilizando imputación de datos, escalado/normalización y codificación de datos categóricos.

3. Desarrollar, implementar y evaluar al menos 3 técnicas de selección de características para determinar las características más informativas para su uso en modelos de ML.

4. Implementar y comparar el rendimiento de al menos 5 modelos de ML para el desarrollo de biomarcadores para la DMT2 y sus comorbilidades.

5. Evaluar el rendimiento de los modelos desarrollados utilizando métricas de evaluación estándar: precisión, *Sensitivity*, *Specificity* y *F1-Score*.

6. Comparar el rendimiento de los biomarcadores desarrollados con las herramientas y directrices clínicas existentes para el diagnóstico y manejo de la DMT2, sus complicaciones y sus comorbilidades.

7. Identificar los rangos en los valores y las diferencias existentes entre pacientes masculinos y femeninos de los biomarcadores potenciales identificados para la detección temprana de DMT2, sus complicaciones y sus comorbilidades.

Capítulo 2
Marco teórico y de referencia

2.1. Diabetes

La diabetes mellitus es un trastorno metabólico crónico caracterizado por niveles elevados de glucosa en sangre, como resultado de una producción insuficiente de insulina, resistencia a la insulina o una combinación de ambas (Collyns *et al.*, 2021). De acuerdo con la definición de la Organización Mundial de la Salud (OMS), la diabetes mellitus (DM) se caracteriza por manifestar insuficiencia en la producción de insulina en el páncreas o cuando el organismo no utiliza eficazmente la insulina producida (WHO, 2022). A medida que avanza esta enfermedad surgen comorbilidades asociadas a la hiperglucemia, que pueden inducir daño renal directamente o a través de modificaciones hemodinámicas, dando lugar a la nefropatía diabética (ND), una de las complicaciones más comunes que afecta de un 30 % a un 40 % de los pacientes (Schena and Gesualdo, 2005). A su vez, está asociada al desarrollo de enfermedades cardiovasculares, ceguera y amputaciones de miembros inferiores (Julia and Carol, 2016), representando un desafío significativo para la salud pública (Iglay *et al.*, 2016). Existen varios tipos de diabetes, cada uno con causas, factores de riesgo y tratamientos distintos, siendo el tipo 2 el más común y desarrollándose principalmente debido a un estilo de vida inactivo, falta de ejercicio y sobrepeso (World Health Organization, 2022).

Los tipos de diabetes son:

La diabetes tipo 1 (DMT1), de naturaleza hereditaria o congénita, se caracteriza por una deficiencia en la producción de insulina por parte del páncreas, lo que resulta en una afección crónica (WHO, 2022). Es un trastorno autoinmune en el que el sistema inmunológico ataca y destruye por error las células β productoras de insulina en el páncreas. Como resultado, se produce poca o ninguna insulina, que es esencial para regular los niveles de azúcar en sangre. La DMT1 generalmente se manifiesta en niños, adolescentes o adultos jóvenes, aunque puede desarrollarse a cualquier edad. El cuerpo no puede producir suficiente insulina,

las personas con DMT1 requieren terapia con insulina de por vida para mantener niveles normales de glucosa en sangre y prevenir complicaciones asociadas con la enfermedad (Collyns *et al.*, 2021).

La DMT2 es una afección progresiva que se caracteriza por una deficiencia relativa de insulina, causada por la disfunción de las células β pancreáticas (las cuales sintetizan y secretan insulina y amilina) y la resistencia a la insulina (Chatterjee *et al.*, 2017). Los principales factores de riesgo de la DMT2 incluyen la obesidad, la inactividad física, los malos hábitos alimentarios, la predisposición genética y la edad avanzada. Aunque tradicionalmente se diagnostica en adultos mayores de 45 años, las crecientes tasas de obesidad han llevado a diagnósticos más frecuentes en personas más jóvenes. El tratamiento para la DMT2 implica una combinación de modificaciones en el estilo de vida, como dieta y ejercicio, junto con medicamentos orales y, en algunos casos, terapia con insulina para controlar los niveles de glucosa en sangre (DeFronzo *et al.*, 2015).

La diabetes gestacional ocurre durante el embarazo cuando los cambios hormonales provocan resistencia a la insulina, lo que afecta la capacidad del cuerpo para regular eficazmente el azúcar en sangre. Las mujeres obesas o que tienen antecedentes familiares de DMT2 tienen un mayor riesgo de desarrollar esta afección. El control de la diabetes gestacional generalmente implica ajustes en la dieta, aumento de la actividad física y, a veces, terapia con insulina para mantener niveles saludables de glucosa en sangre durante el embarazo. Las mujeres que experimentan diabetes gestacional tienen un mayor riesgo de desarrollar DMT2 en el futuro (McIntyre *et al.*, 2019).

2.1.1. Glucagón y su inferencia en la diabetes

El glucagón conocida también como hormona hiperglucemiante, es una hormona peptídica sintetizada por las células α de los islotes pancreáticos o islotes de Langerhands, desempeña un papel fundamental en la homeostasis de la glucosa y tiene implicaciones sustanciales en la fisiopatología de la DMT2 (vea figura 2.1(Molina, 2024)). En condiciones fisiológicas normales, el glucagón contrarresta los efectos de la insulina al promover la producción de glucosa hepática mediante la glucogenólisis y la gluconeogénesis, particularmente durante el ayuno o los estados de hipoglucemia. Su secreción es estimulada principalmente por los niveles bajos de glucosa en sangre, el sistema nervioso autónomo y los aminoácidos. El glucagón ejerce sus efectos uniéndose a receptores acoplados a la proteína G en los hepatocitos, activando la adenilato-ciclasa y aumentando posteriormente los niveles de AMP cíclico intracelular (AMPc). Esta activación estimula la proteína quinasa A (PKA) y cataliza aún más la activación de la glucógeno fosforilasa, lo que culmina en la degradación del glucógeno y la acti-

vación de la vía gluconeogénica, que en última instancia aumenta los niveles de glucosa en sangre (Jiang and Zhang, 2003).

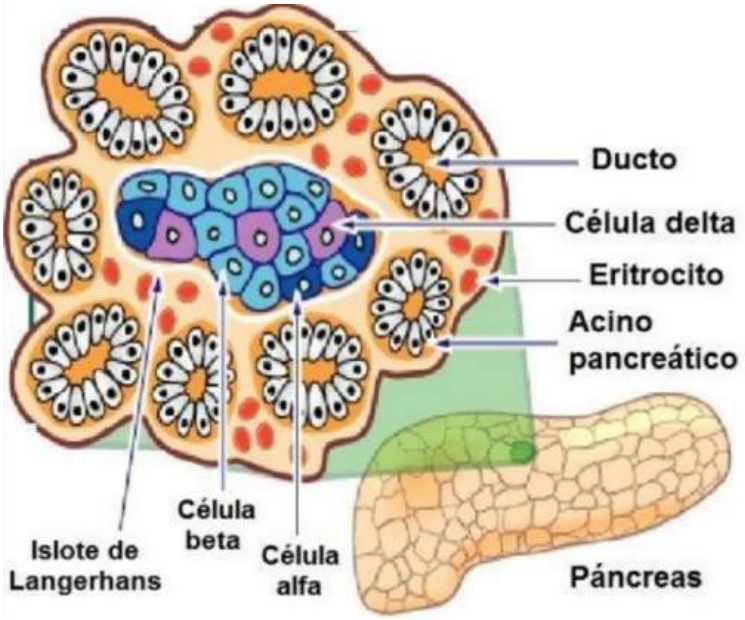


Figura 2.1: Glucagón - Una hormona peptídica sintetizada por las células α de los islotes pancreáticos o islotes de Langerhands y desempeña un papel fundamental en la homeostasis de la glucosa. La insulina, por otro lado, es una hormona peptídica secretada por las células β y facilita la captación de glucosa uniéndose al receptor de insulina, un receptor de tirosina quinasa transmembrana.

En la DMT2, se observa un patrón atípico de secreción de glucagón, a menudo denominado hiperglucagonemia, donde los niveles de glucagón están elevados incluso en estados hiperglucémicos. Esta secreción paradójica de glucagón contribuye a la hiperglucemia característica de la DMT2. Los mecanismos subyacentes a la hiperglucagonemia en la DMT2 son multifactoriales e implican alteración de la señalización en las células α, resistencia a la insulina y desregulación de las interacciones paracrinas dentro de los islotes pancreáticos. En individuos sanos, la insulina inhibe directamente la secreción de glucagón; sin embargo, en la DMT2, la resistencia a la insulina altera este efecto inhibidor. Además, la propia hiperglucemia debería suprimir la liberación de glucagón a través de la secreción de somatostatina por las células δ en los islotes pancreáticos. Sin embargo, en la DM2, esta inhibición paracrina a menudo se altera, lo que da como resultado una secreción lasciva de glucagón. Un efecto importante de la hiperglucagonemia crónica en la DMT2 es el aumento de la producción de glucosa hepática, que contribuye significativamente a la hiperglucemia en ayunas. Con el tiempo, este

estado exacerba la resistencia a la insulina, ya que el exceso de glucosa en el torrente sanguíneo provoca una demanda continua de insulina, lo que lleva a la disfunción de las células β y, en última instancia, empeora la condición diabética. Además, la desregulación del glucagón en la DMT2 también se ha relacionado con la lipólisis y la cetogénesis, aunque estos procesos son menos pronunciados en la DMT2 en comparación con la DMT1 debido a la actividad parcial de la insulina que aún inhibe una producción significativa de cetonas (Lund *et al.*, 2014).

2.1.2. Resistencia a la insulina

La resistencia a la insulina altera la respuesta celular a esta hormona, particularmente en el músculo esquelético, el tejido adiposo y el hígado. Tras la unión de la insulina, se desarrollan una serie de eventos intracelulares que se inician con la autofosforilación del receptor y activan las proteínas del sustrato del receptor de insulina. Esto, a su vez, desencadena la vía fosfoinositol 3-quinasa (PI3K)/Akt, que desempeña un papel crucial en la absorción y el metabolismo de la glucosa. Sin embargo, en estados de resistencia a la insulina, esta vía de señalización está significativamente atenuada (Cefalu, 2001).

En el músculo esquelético, que es responsable de una parte sustancial de la eliminación de glucosa posprandial, la resistencia a la insulina da como resultado una reducción de la captación de glucosa y la síntesis de glucógeno. Esto se debe a una translocación deficiente del transportador de glucosa tipo 4 (GLUT4) a la superficie celular. Como consecuencia, la glucosa permanece en el torrente sanguíneo, lo que contribuye a la hiperglucemia. En el tejido adiposo, la resistencia a la insulina disminuye la acción antilipolítica de la insulina, lo que resulta en una mayor liberación de ácidos grasos libres a la circulación. Los niveles elevados de ácidos grasos libres alteran aún más la señalización de la insulina y promueven la gluconeogénesis en el hígado, lo que exacerba la hiperglucemia. Además, los niveles elevados de FFA contribuyen a un estado inflamatorio crónico de bajo grado, con citocinas proinflamatorias como TNF-α, IL-6 y resistina que alteran las vías de señalización de la insulina a través de mecanismos que involucran la fosforilación de serina de las proteínas del IRS, lo que inhibe la señalización posterior (Zierath *et al.*, 2000).

En el tejido hepático, la resistencia a la insulina altera la capacidad de la insulina para suprimir la gluconeogénesis y la glucogenólisis hepáticas, que contribuyen de manera importante a la hiperglucemia en ayunas. Normalmente, la insulina actúa para regular negativamente la expresión de enzimas gluconeogénicas, como la fosfoenolpiruvato carboxiquinasa (PEPCK) y la glucosa-6-fosfatasa (G6Pasa), y para promover la síntesis de glucógeno. Sin embargo, en el contexto de la resistencia a la insulina, estas acciones regulatorias se ven comprometi-

das. El aumento resultante en la producción de glucosa hepática contribuye de manera crítica a la hiperglucemia observada en la DMT2 (Kotronen *et al.*, 2008).

La obesidad, particularmente la adiposidad visceral, es un factor de riesgo importante para la resistencia a la insulina. El tejido adiposo en la obesidad se vuelve disfuncional, liberando niveles elevados de ácidos grasos libres y adipocinas (como la leptina y la resistina) al tiempo que disminuye la adiponectina, una hormona sensibilizadora de la insulina. Además, el depósito ectópico de grasa en tejidos no adiposos como el hígado y los músculos también contribuye a la resistencia a la insulina a través de mecanismos que alteran la función mitocondrial y promueven el estrés oxidativo (Ye, 2013).

Además de la hiperglucemia, la resistencia a la insulina también se asocia con hiperinsulinemia debido al aumento compensatorio de la secreción de insulina por parte de las células β pancreáticas para mantener la euglucemia en las primeras etapas de la resistencia a la insulina. Sin embargo, este mecanismo compensatorio no es sostenible y, a medida que la función de las células β disminuye con el tiempo, la secreción de insulina se vuelve insuficiente, lo que precipita la transición de la resistencia a la insulina a la DM2 manifiesta. Se cree que la pérdida progresiva de la función de las células β implica glucotoxicidad y lipotoxicidad, por lo que la exposición crónica a niveles elevados de glucosa y ácidos grasos libres altera aún más la secreción e insulina y conduce a la apoptosis de las células β (Shanik *et al.*, 2008).

2.1.3. Glucosa

La glucosa, es un monosacárido de seis carbonos y la principal fuente de energía para las células del cuerpo, siendo esencial para diversas funciones metabólicas. En un individuo sano, los niveles de glucosa en sangre están estrechamente regulados por una interacción compleja entre la secreción de insulina, la producción de glucosa hepática y la captación celular de glucosa. Esta regulación mantiene los niveles de glucosa dentro de un rango fisiológico de aproximadamente 70-100 mg/dL en ayunas. Sin embargo, en personas con DMT2, la homeostasis de la glucosa se ve significativamente alterada debido a una combinación de resistencia a la insulina y disfunción progresiva de las células β, lo que conduce a hiperglucemia crónica, que es la base de la fisiopatología y las complicaciones asociadas con la DMT2 (Bergman, 1989).

2.1.4. Biomarcadores en la diabetes

En el ámbito clínico, los biomarcadores se definen como medidas cuantificables de procesos biológicos o patológicos, que pueden incluir desde moléculas

biológicas como proteínas o metabolitos que se encuentran en la sangre, la orina o los tejidos, hasta características físicas como imágenes o mediciones antropométricas (Lock and Bonventre, 2008). Los biomarcadores pueden usarse para detectar enfermedades o controlar su progresión indicando un estado biológico, en el caso de la diabetes, los biomarcadores ayudan a diagnosticar la afección de manera temprana, evaluar su gravedad y predecir complicaciones o utilizarse para el desarrollo de nuevos medicamentos y enfoques terapéuticos.

Los tres biomarcadores más utilizados actualmente para la detección de la diabetes mellitus son:

- Glucosa plasmática en ayunas (Fasting Plasma Glucose o FPG por sus siglas en inglés)
- Test de tolerancia oral a la glucosa (Oral Glucose Tolerance Test u OGTT por sus siglas en inglés)
- Hemoglobina glicada (HbA1c o prueba de hemoglobina A1c)

La glicohemoglobina se produce mediante una reacción de cetoamina entre la glucosa y el aminoácido N-terminal de la cadena β de la hemoglobina. La cantidad de glucohemoglobina formada refleja los niveles promedio de glucosa en sangre durante las 8 a 10 semanas anteriores, lo que la convierte en un indicador valioso para monitorear el control de la glucosa en sangre a largo plazo.

Estos biomarcadores químicos y clínicos se utilizan como métricas para clasificar el estado de salud de un individuo, ya sea como sano, prediabético o diabético.

La tabla 2.1 presenta los rangos de referencia para cada prueba y su clasificación asociada, según la Asociación Americana de Diabetes (ADA por sus siglas de American Diabetes Association). Por ejemplo, los valores de FPG inferiores a 100 mg/dL se consideran normales, mientras que valores entre 100 mg/dL y 125 mg/dL indican prediabetes, y 126 mg/dL o más señalan diabetes (William H. *et al.*, 2015).

Existen tambien otros biomarcadores clínicos como:

El péptido C que es una sustancia producida por el páncreas en cantidades iguales a la insulina cuando la proinsulina, la molécula precursora, se divide en insulina y péptido C. Medir los niveles de péptido C es una herramienta de diagnóstico crucial para evaluar cuánta insulina produce el cuerpo. Este biomarcador es especialmente útil para distinguir entre DMT1 y DMT2. En la DMT1, donde el sistema inmunológico destruye las células β productoras de insulina, los niveles de péptido C suelen ser muy bajos o ausentes, lo que refleja la incapacidad del cuerpo para producir insulina. Por el contrario, las personas con

Tabla 2.1: Biomarcadores clínicos relacionados con la glucosa.

Prueba	**Normal**	**Prediabetes**	**Diabetes**
FPG	< 100 mg/dL	100 mg/dL – 125 mg/dL	≥ 126 mg/dL
OGTT	< 140 mg/dL	140 mg/dL – 199 mg/dL	≥ 200 mg/dL
HbA1c	$< 5.7\%$	$5.7\% - 6.4\%$	$\geq 6.5\%$

La clasificación de la enfermedad de DMT2 se da estableciendo un rango de valores en los que se proporciona la métrica o biomarcador para definir si una persona tiene DMT2 y si la tiene, en qué grado se manifiesta, estos datos fueron establecidos por la ADA.

DMT2 suelen tener niveles de péptido C normales o incluso elevados, lo que indica resistencia a la insulina en lugar de una falta total de producción de insulina. Por tanto, la prueba del péptido C ayuda a guiar el diagnóstico y el tratamiento al diferenciar entre estos dos tipos de diabetes. También se utiliza para controlar la función pancreática en personas con diabetes a lo largo del tiempo (Jones and Hattersley, 2013).

Los triglicéridos y el colesterol desempeñan un papel importante en la salud de las personas con DMT2. Los niveles elevados de triglicéridos y los niveles bajos de HDL (lipoproteína de alta densidad o colesterol "bueno") se observan comúnmente en pacientes con DMT2. Estos desequilibrios de lípidos contribuyen al desarrollo de aterosclerosis, donde se acumulan depósitos de grasa en las arterias, lo que aumenta el riesgo de enfermedades cardiovasculares. Las complicaciones cardiovasculares, como enfermedades cardíacas y accidentes cerebrovasculares, son comorbilidades frecuentes en la diabetes, controlar los niveles de triglicéridos y colesterol es crucial para reducir el riesgo general. El manejo eficaz mediante cambios en el estilo de vida, medicación y seguimiento regular es esencial para minimizar el impacto de estos factores de riesgo en los pacientes con DMT2 (Adiwinoto *et al.*, 2024).

Los biomarcadores inflamatorios, como la interleucina-6 (IL-6) y la proteína C reactiva (PCR), están estrechamente relacionados con la resistencia a la insulina y el desarrollo de DMT2. Estas moléculas sirven como indicadores de inflamación sistémica, que es una afección subyacente común en personas con DMT2. Los niveles elevados de PCR e IL-6 sugieren que el cuerpo está experimentando una inflamación crónica de bajo grado, un estado que contribuye a la progresión de la resistencia a la insulina y empeora el metabolismo de la glucosa. La inflamación sistémica también se asocia con un mayor riesgo de complicaciones cardiovasculares, el seguimiento de estos biomarcadores inflamatorios puede ser útil tanto para diagnosticar la DMT2 como para controlar sus comorbilidades. Abordar la

inflamación mediante terapias dirigidas o intervenciones en el estilo de vida puede ayudar a mejorar la sensibilidad a la insulina y reducir el riesgo general de complicaciones (Mohammad *et al.*, 2023).

Las enzimas hepáticas, específicamente la alanina aminotransferasa (ALT) y la aspartato aminotransferasa (AST), con frecuencia están elevadas en personas con DMT2. Estos niveles elevados de enzimas pueden indicar la presencia de enfermedad del hígado graso no alcohólico (NAFLD), una comorbilidad común en pacientes diabéticos. La NAFLD ocurre cuando se acumula un exceso de grasa en el hígado, lo que con el tiempo provoca inflamación y daño hepático. La DMT2 y la NAFLD comparten factores de riesgo similares, como la obesidad, la resistencia a la insulina y el síndrome metabólico, la presencia de niveles elevados de ALT y AST puede servir como biomarcadores importantes para diagnosticar y controlar las complicaciones hepáticas en pacientes con DMT2. La detección y el tratamiento tempranos de la NAFLD son cruciales para prevenir la progresión a afecciones hepáticas más graves, como la cirrosis (Bi *et al.*, 2024).

2.1.5. biomarcadores bioinformáticos

Los biomarcadores bioinformáticos sirven como variables cuantificables que pueden incorporarse en modelos de ML y análisis estadísticos para predecir la aparición, la progresión y la respuesta al tratamiento de una enfermedad. Utilizando conjuntos de datos a gran escala y herramientas computacionales, los ingenieros en ciencia de datos pueden emplear métodos de selección de características y modelos predictivos con ML determinando con alta precisión un posible diagnóstico, a partir de patrones y correlaciones extraídas de los datos analizados (Alur *et al.*, 2023).

2.1.6. Comorbilidades y complicaciones

Las comorbilidades y complicaciones son conceptos distintos pero interrelacionados, ya que ambos describen afecciones asociadas con una enfermedad primaria. Las comorbilidades se refieren a una o más afecciones o enfermedades médicas coexistentes que están presentes junto con una enfermedad primaria pero que no son necesariamente causadas por ella. Estas afecciones pueden haberse desarrollado de forma independiente y compartir potencialmente factores de riesgo comunes, como la edad, el estilo de vida o las predisposiciones genéticas, con la enfermedad primaria. Las comorbilidades pueden influir en el curso y el pronóstico de la enfermedad primaria, añadiendo a menudo complejidad a los enfoques terapéuticos. Por ejemplo, en la DMT2, la hipertensión y la dislipidemia son comorbilidades comunes. Si bien estas afecciones no son causadas

directamente por la diabetes, con frecuencia coexisten debido a factores fisiopatológicos compartidos como la obesidad, la inflamación crónica y la resistencia a la insulina. Los médicos que atienden a pacientes con comorbilidades a menudo necesitan coordinar tratamientos que aborden múltiples afecciones, con el objetivo de mitigar posibles interacciones farmacológicas y optimizar los resultados de los pacientes. En contraste, las complicaciones son condiciones o patologías secundarias que surgen como consecuencia directa de una enfermedad primaria. Las complicaciones generalmente se desarrollan cuando la enfermedad primaria progresa o se maneja mal, lo que resulta en un daño mayor a sistemas de órganos o funciones fisiológicas específicas. En la DMT2, las complicaciones comunes incluyen retinopatía, neuropatía, nefropatía y cardiopatía. Estas complicaciones ocurren debido a una hiperglucemia prolongada, que daña los vasos sanguíneos y los tejidos con el tiempo. A diferencia de las comorbilidades, que pueden existir de forma independiente, las complicaciones están causalmente relacionadas con la enfermedad primaria y a menudo indican un empeoramiento de la gravedad de la enfermedad (Iglay *et al.*, 2016).

2.1.7. Grupos de edad con mayor riesgo de diabetes

La DMT1 se diagnostica comúnmente en niños, adolescentes y adultos jóvenes, generalmente menores de 30 años. Los períodos pico de diagnóstico ocurren durante la primera infancia y la adolescencia, con las tasas de incidencia más altas observadas en personas de entre 5 años. y 14 (Ouyang *et al.*, 2024).

La DMT2 se diagnostica con mayor frecuencia en adultos de 45 años o más, pero debido a las crecientes tasas de obesidad y factores relacionados con el estilo de vida, las poblaciones más jóvenes, incluidos adolescentes y adultos jóvenes, se ven cada vez más afectadas. Este cambio en la demografía ha generado preocupación sobre la prevalencia de la DMT2 entre personas más jóvenes que tradicionalmente no están asociadas con la afección. Las personas mayores de 65 años tienen un mayor riesgo de desarrollar la enfermedad, especialmente si tienen otros factores de riesgo como obesidad, hipertensión o un estilo de vida sedentario (Li *et al.*, 2024).

La diabetes gestacional ocurre durante el embarazo y afecta principalmente a mujeres de 25 años o más. El riesgo es notablemente mayor para aquellos con antecedentes familiares de DMT2 u obesidad. A medida que aumenta la edad materna, especialmente en mujeres mayores de 30 años, también aumenta la probabilidad de desarrollar diabetes gestacional. Esta afección es una preocupación importante porque no solo plantea riesgos para la salud de la madre y el bebé durante el embarazo, sino que también aumenta las posibilidades de que la madre desarrolle DMT2 más adelante en la vida (Kumar *et al.*, 2022).

2.1.8. Tratamientos para la diabetes

Las modificaciones en el estilo de vida desempeñan un papel importante en el control de la diabetes y abarcan diversas estrategias destinadas a mejorar la salud general y el control de la glucosa en sangre. Los cambios en la dieta pueden incluir: una alimentación equilibrada rica en cereales integrales, frutas, verduras, proteínas magras y grasas saludables, esencial para mantener estables los niveles de azúcar en sangre. Estos grupos de alimentos proporcionan los nutrientes necesarios al tiempo que ayudan a regular la ingesta de energía. El ejercicio es otro componente clave, ya que la actividad física regular mejora la sensibilidad a la insulina, lo que permite que el cuerpo utilice la glucosa de manera más efectiva. Además, el control del peso es particularmente importante para las personas con DMT2, donde incluso una pérdida de peso modesta puede conducir a mejoras sustanciales en la sensibilidad a la insulina y la salud metabólica general (Zhu *et al.*, 2024).

Las estrategias terapéuticas para la resistencia a la insulina incluyen: intervenciones en el estilo de vida, como la reducción de peso y el aumento de la actividad fisica, que mejoran la sensibilidad a la insulina en los tejidos periféricos. Los agentes farmacológicos, como la metformina que es uno de los medicamentos recetados con más frecuencia, actúan para mejorar la sensibilidad a la insulina, principalmente en el hígado, mientras que las tiazolidinedionas actúan sobre el receptor γ activado por el proliferador de peroxisomas en el tejido adiposo para reducir la resistencia a la insulina. Otro tratamiento serían las sulfonilureas, que funcionan estimulando el páncreas para que produzca más insulina, aumentando así la cantidad de insulina disponible para reducir los niveles de glucosa en sangre. Además, existen los inhibidores de SGLT2 que son una clase más nueva de medicamentos que ayudan a los riñones a eliminar el exceso de glucosa del torrente sanguíneo, lo que no solo ayuda a controlar el azúcar en la sangre sino que también favorece la pérdida de peso y puede ofrecer beneficios cardiovasculares (Bailey, 2024).

La terapia con insulina es un tratamiento vital para controlar la diabetes, particularmente en la DMT1 y en algunos casos de DMT2. Para las personas con DMT1, las inyecciones diarias de insulina son esenciales para la supervivencia, ya que sus cuerpos no pueden producir insulina debido a un ataque autoinmune a las células β productoras de insulina en el páncreas. Estas inyecciones proporcionan un suministro constante de insulina para regular eficazmente los niveles de glucosa en sangre (Schaffner *et al.*, 2024).

En el caso de la DMT2, la terapia con insulina puede resultar necesaria cuando las modificaciones en el estilo de vida y los medicamentos orales no logran mantener un control adecuado del azúcar en sangre. Esto puede ocurrir en eta-

pas avanzadas de la enfermedad o en individuos con resistencia significativa a la insulina. La terapia con insulina para la DMT2 puede ayudar a mejorar el control glucémico y reducir el riesgo de complicaciones asociadas con niveles elevados de azúcar en sangre prolongados. La flexibilidad en la dosis y la capacidad de adaptar el tratamiento a las necesidades individuales hacen de la terapia con insulina una opción crucial en el tratamiento integral de ambos tipos de diabetes (Schaffner *et al.*, 2024).

Los antagonistas del receptor de glucagón y los agonistas del receptor de GLP-1 (que aumentan la secreción de insulina e inhiben la liberación de glucagón) se han mostrado prometedores para reducir los niveles de glucosa en sangre al normalizar la secreción y la acción del glucagón. Estos enfoques tienen como objetivo restablecer el equilibrio fisiológico entre la insulina y el glucagón, mejorando así el control glucémico y potencialmente ralentizando la progresión de la DM2. Estos medicamentos inyectables mejoran la secreción de insulina en respuesta a las comidas y al mismo tiempo suprimen la liberación de glucagón, lo que conduce a un mejor control de la glucosa en sangre. Estos medicamentos no solo ayudan a reducir los niveles de azúcar en sangre, sino que también pueden promover la pérdida de peso, lo que los hace beneficiosos para muchos pacientes con DMT2 (Kukova *et al.*, 2024). Uno de estos medicamentos es la Semaglutida.

Para los pacientes con obesidad mórbida y DMT2, la cirugía bariátrica representa una opción de tratamiento potencialmente transformadora. Esta cirugía para bajar de peso puede dar lugar a reducciones significativas del peso corporal y se ha asociado con mejoras sustanciales en el control de la diabetes. En algunos casos, los pacientes pueden lograr la remisión completa de su diabetes después del procedimiento. Este enfoque aborda uno de los factores de riesgo más importantes y puede conducir a una mayor sensibilidad a la insulina y mejores resultados metabólicos (Courcoulas *et al.*, 2024).

2.1.9. Comorbilidades y complicaciones de la diabetes

La diabetes a menudo conduce a otros problemas de salud, conocidos como comorbilidades, que pueden afectar significativamente la calidad de vida y aumentar el riesgo de mortalidad.

Entre las comorbilidades más comúnes se encuentran:

- Enfermedades cardiovasculares: los diabéticos tienen un riesgo mucho mayor de sufrir enfermedades cardíacas, como enfermedad de las arterias coronarias, accidentes cerebrovasculares e hipertensión. Los niveles elevados de glucosa en sangre pueden dañar los vasos sanguíneos y provocar aterosclerosis (endurecimiento de las arterias).

- Enfermedad renal (nefropatía diabética): la diabetes a largo plazo puede dañar el sistema de filtrado de los riñones y provocar enfermedad renal crónica o insuficiencia renal.

- Daño a los nervios (neuropatía diabética): los niveles altos de azúcar en sangre pueden dañar los nervios de todo el cuerpo, provocando dolor, hormigueo o pérdida de sensibilidad, especialmente en las manos y los pies.

- Daño ocular (retinopatía diabética): la diabetes puede dañar los vasos sanguíneos de la retina, provocando pérdida de la visión o ceguera si no se trata.

2.1.10. Enfermedades cardiovasculares

Las enfermedades cardiovasculares son una preocupación importante para las personas con diabetes, ya que tienen un riesgo notablemente mayor de desarrollar diversas afecciones relacionadas con el corazón. Este mayor riesgo se puede atribuir a múltiples factores, incluidos los efectos directos de los niveles altos de glucosa en sangre en el sistema cardiovascular y la presencia de factores de riesgo comunes asociados con la diabetes (Sasako, 2023).

Uno de los principales mecanismos por los cuales la diabetes contribuye a las enfermedades cardiovasculares es a través del daño causado a los vasos sanguíneos por los niveles crónicamente elevados de glucosa en sangre. Con el tiempo, un nivel alto de azúcar en sangre puede provocar una afección conocida como aterosclerosis, caracterizada por el endurecimiento y estrechamiento de las arterias. La aterosclerosis ocurre cuando la glucosa reacciona con proteínas y lípidos en la sangre, formando productos finales de glicación avanzada. Estos compuestos nocivos pueden promover la inflamación y el estrés oxidativo, dañando las células endoteliales que recubren los vasos sanguíneos (Sasako, 2023).

Como resultado de este daño, las arterias se vuelven menos flexibles y acumulan depósitos de grasa, colesterol y otras sustancias. Esta acumulación estrecha las arterias y restringe el flujo sanguíneo, lo que puede provocar eventos cardiovasculares graves, como enfermedad de las arterias coronarias y ataques cardíacos. En la enfermedad coronaria, las arterias coronarias que suministran sangre al músculo cardíaco se estrechan o bloquean, lo que reduce el suministro de oxígeno al corazón y puede provocar dolor en el pecho (angina) o ataques cardíacos (Sasako, 2023).

Además de la enfermedad coronaria, la diabetes aumenta significativamente el riesgo de sufrir un accidente cerebrovascular. Los mismos procesos ateroscleróticos que afectan las arterias coronarias también pueden afectar las arterias que

suministran sangre al cerebro. Si estas arterias se bloquean o se rompen, puede provocar un derrame cerebral, que puede tener consecuencias devastadoras, incluida la discapacidad a largo plazo o la muerte (Sasako, 2023).

La hipertensión, o presión arterial alta, es otra comorbilidad común entre los diabéticos que exacerba el riesgo de enfermedades cardiovasculares. El daño a los vasos sanguíneos causado por un nivel alto de azúcar en sangre puede provocar un aumento de la rigidez y la resistencia arterial, lo que contribuye a una presión arterial elevada. Además, la hipertensión en sí misma puede dañar aún más los vasos sanguíneos y aumentar la carga de trabajo del corazón, agravando los riesgos asociados con la diabetes (Sasako, 2023).

2.1.11. Nefropatía diabética

La nefropatía diabética, o enfermedad renal (ND), es una complicación importante y común de la diabetes a largo plazo, que afecta a un número sustancial de personas con DMT1 y DMT2. Esta afección resulta de la exposición prolongada a niveles altos de glucosa en sangre, que con el tiempo pueden dañar el intrincado sistema de filtrado de los riñones (Gao *et al.*, 2023).

Fisiopatología

Los riñones desempeñan un papel crucial en la filtración de productos de desecho y el exceso de líquido de la sangre, la regulación de los niveles de electrolitos y el mantenimiento del equilibrio general de líquidos en el cuerpo. Dentro de los riñones hay estructuras diminutas llamadas nefronas, cada una de las cuales consta de un glomérulo (una red de pequeños vasos sanguíneos) y un túbulo renal. Los glomérulos filtran la sangre, mientras que los túbulos reabsorben los nutrientes esenciales y excretan los productos de desecho en la orina (Gao *et al.*, 2023).

En la ND, los niveles elevados de glucosa en sangre provocan una serie de cambios patológicos en los riñones:

- Hiperglucemia y cambios metabólicos: los niveles elevados de glucosa en el torrente sanguíneo hacen que los riñones trabajen más para filtrar la sangre. Con el tiempo, este aumento de la carga de trabajo puede provocar cambios estructurales en los riñones, incluido el engrosamiento de la membrana basal glomerular.

- Aumento de la presión glomerular: los niveles altos de glucosa también provocan cambios en el flujo sanguíneo renal, lo que provoca un aumento de la presión dentro de los glomérulos. Esta presión elevada puede dañar el

delicado aparato de filtrado, provocando una fuga de proteínas en la orina (proteinuria).

- Inflamación y fibrosis: la hiperglucemia promueve la inflamación dentro de los riñones y la producción de diversos factores de crecimiento, lo que lleva a la fibrosis (cicatrización) de los tejidos renales. Esta cicatrización perjudica aún más la función renal y contribuye a la progresión de la nefropatía.

Manifestaciones clínicas

La ND generalmente se desarrolla durante varios años y puede progresar a través de varias etapas:

- Microalbuminuria: esta etapa temprana se caracteriza por la presencia de pequeñas cantidades de proteína (albúmina) en la orina. A menudo es asintomático, pero la detección periódica de los niveles de albúmina en orina es esencial para una detección temprana.

- Proteinuria: a medida que la afección progresa, mayores cantidades de proteína comienzan a filtrarse a la orina. Esta etapa puede ir acompañada de síntomas como hinchazón (edema) en piernas, tobillos y pies debido a la retención de líquidos.

- Enfermedad Renal Crónica (ERC): con el tiempo, la función renal disminuye aún más, lo que lleva a la ERC. Los pacientes pueden experimentar fatiga, cambios en la producción de orina, presión arterial alta y desequilibrios electrolíticos.

- Enfermedad renal en etapa terminal (ESRD): en casos graves, la ND puede progresar a ESRD, donde los riñones ya no pueden filtrar eficazmente los productos de desecho de la sangre. Esta etapa a menudo requiere diálisis o trasplante de riñón para sobrevivir.

Factores de riesgo y gestión

Varios factores aumentan significativamente el riesgo de desarrollar ND, incluida la duración de la diabetes, el control glucémico deficiente, la hipertensión y la predisposición genética. Cuanto más tiempo una persona tiene diabetes, mayor es la probabilidad de daño renal, los niveles persistentemente altos de azúcar en sangre son un factor de riesgo importante para la nefropatía. Además, la hipertensión sirve como causa y consecuencia del daño renal, lo que exacerba aún más la progresión de la nefropatía, mientras que los antecedentes familiares

de enfermedad renal pueden aumentar la susceptibilidad de un individuo. Para gestionar y prevenir eficazmente el avance de la ND. Mantener niveles óptimos de glucosa en sangre mediante modificaciones en el estilo de vida, como dieta y ejercicio, junto con medicamentos, es crucial para frenar la progresión del daño renal. Además, controlar la hipertensión con medicamentos como inhibidores de la enzima convertidora de angiotensina (ECA) o bloqueadores de los receptores de angiotensina II (BRA) puede ayudar a salvaguardar la función renal. La monitorización periódica, incluida la detección sistemática de microalbuminuria y evaluaciones de la función renal (como la creatinina sérica y la tasa de filtración glomerular estimada), es vital para la detección e intervención tempranas. Finalmente, las modificaciones en el estilo de vida que promuevan una dieta equilibrada baja en sodio y proteínas, actividad física regular y evitar fumar pueden contribuir aún más al control eficaz de la diabetes y la preservación de la salud renal (Gao *et al.*, 2023).

2.1.12. Neuropatía diabética

La neuropatía diabética es una complicación común y debilitante de la DMT2, resultante de una hiperglucemia prolongada que daña los nervios periféricos. Esta afección abarca diversas formas de daño a los nervios que afectan distintas regiones del cuerpo y presentan una variedad de síntomas clínicos. Ocurre principalmente debido a cambios metabólicos, vasculares y neurotróficos causados por niveles altos de glucosa en sangre, que comprometen la función nerviosa con el tiempo. La fisiopatología de la neuropatía diabética implica múltiples mecanismos, incluido el estrés oxidativo por niveles elevados de glucosa, productos finales de glicación avanzada (AGE) e insuficiencia vascular. La hiperglucemia conduce a la activación de la vía de los polioles, que convierte la glucosa en sorbitol en las células nerviosas. Este proceso agota cofactores esenciales como el NADPH, aumenta el estrés oxidativo y provoca desequilibrios osmóticos y metabólicos que dañan las fibras nerviosas. Además, la enfermedad vascular asociada a la diabetes restringe el flujo sanguíneo a los nervios, privándolos de oxígeno y nutrientes, lo que agrava la condición de disfunción nerviosa. El tipo más común es la neuropatía periférica, que a menudo afecta los pies y las piernas, seguidos de las manos y los brazos en una distribución en “guante y media”. Los pacientes con neuropatía periférica pueden experimentar síntomas como entumecimiento, hormigueo, ardor o dolores agudos, que tienden a empeorar por la noche. Con el tiempo, la pérdida de sensibilidad puede provocar lesiones no reconocidas, lo que aumenta el riesgo de úlceras e infecciones que, en última instancia, pueden provocar amputaciones si no se tratan adecuadamente (Braffett *et al.*, 2024).

La neuropatía autónoma afecta el sistema nervioso autónomo y altera funciones corporales involuntarias como la frecuencia cardíaca, la digestión y el control de la vejiga. Por ejemplo, la neuropatía autonómica puede causar problemas gastrointestinales como la gastroparesia, donde el retraso en el vaciado del estómago provoca náuseas, hinchazón y niveles erráticos de glucosa en sangre. La neuropatía autonómica cardiovascular, otra manifestación, puede provocar una pérdida de la variabilidad de la frecuencia cardíaca, hipotensión ortostática (una caída repentina de la presión arterial al estar de pie) y un mayor riesgo de infarto de miocardio silencioso, ya que es posible que los pacientes no experimenten el dolor torácico típico debido a la reducción. sensación (Verrotti *et al.*, 2014).

La neuropatía proximal (también conocida como amiotrofia diabética) es una forma menos común que afecta los muslos, las caderas, las nalgas o las piernas, causando dolor intenso, debilidad muscular y, en algunos casos, atrofia muscular. Esta forma suele ocurrir en adultos mayores con diabetes tipo 2 y puede provocar una discapacidad significativa y una movilidad reducida (Pascoe *et al.*, 1997).

Por último, la neuropatía focal o mononeuropatía afecta nervios específicos, a menudo en la cabeza, el torso o la pierna, y causa un dolor intenso y repentino. Por ejemplo, la mononeuropatía craneal puede afectar los nervios que controlan los músculos oculares, lo que provoca visión doble o párpados caídos. En otros casos, la neuropatía focal puede afectar el nervio femoral, provocando dolor y debilidad en el muslo (Vinik *et al.*, 2004).

2.1.13. Retinopatía diabética

La retinopatía diabética es una complicación grave y progresiva de la DMT2 que afecta la retina, la capa de tejido sensible a la luz en la parte posterior del ojo. Surge debido a niveles elevados prolongados de glucosa en sangre, que provocan daños en la microvasculatura de la retina. Con el tiempo, este daño vascular puede comprometer el flujo sanguíneo de la retina, lo que provoca isquemia (disminución del suministro de sangre) y posterior lesión retiniana. La patogénesis de la retinopatía diabética implica varias etapas. Inicialmente, en la fase de retinopatía diabética no proliferativa, los vasos sanguíneos de la retina desarrollan microaneurismas, pequeñas hinchazones en forma de globos en las paredes de los vasos sanguíneos. Estos microaneurismas pueden romperse, provocar pequeñas hemorragias y provocar un aumento de la permeabilidad vascular. Los líquidos y las proteínas pueden filtrarse hacia la retina, lo que provoca edema (hinchazón) de la retina, lo cual es especialmente preocupante si afecta a la mácula, el área responsable de la visión central nítida. El edema dentro de la mácula, conocido

como edema macular diabético, es una de las principales causas de discapacidad visual en la retinopatía diabética (Ivanescu *et al.*, 2024).

A medida que la enfermedad progresa, puede avanzar a la etapa de retinopatía diabética proliferativa, caracterizada por el crecimiento de vasos sanguíneos nuevos y frágiles en la superficie de la retina. Este proceso, llamado neovascularización, ocurre en respuesta a la isquemia, cuando la retina intenta compensar la reducción del flujo sanguíneo. Sin embargo, estos nuevos vasos son propensos a romperse, lo que provoca una hemorragia vítrea (sangrado en la sustancia gelatinosa que llena el ojo) y también pueden formar tejido cicatricial fibroso que tira de la retina, aumentando el riesgo de desprendimiento de retina. condición que puede resultar en ceguera permanente si no se trata. Clínicamente, la retinopatía diabética puede ser asintomática en sus primeras etapas, lo que subraya la importancia de exámenes oculares periódicos para los pacientes con diabetes. Los síntomas, cuando ocurren, pueden incluir visión borrosa o fluctuante, manchas oscuras o "moscas volantes.[en] el campo de visión y problemas de visión de los colores. Los casos avanzados pueden provocar una pérdida grave de la visión o ceguera (Ivanescu *et al.*, 2024).

El tratamiento de la retinopatía diabética implica un enfoque multidisciplinario. El control estricto de los niveles de glucosa en sangre, la presión arterial y los niveles de lípidos es crucial para frenar la progresión de la enfermedad. En los casos en los que la retinopatía ha avanzado se requieren tratamientos específicos. Estos incluyen inyecciones intravítreas de agentes anti-VEGF (factor de crecimiento endotelial vascular), que pueden reducir la formación anormal de vasos sanguíneos y el edema retiniano. La terapia de fotocoagulación con láser también se utiliza para sellar los vasos con fugas y reducir la neovascularización. En casos de desprendimiento de retina o hemorragia vítrea persistente, puede ser necesaria una cirugía de vitrectomía para restaurar o preservar la visión (Ivanescu *et al.*, 2024).

2.1.14. Metabolómica

La metabolómica es un campo integral y altamente especializado dentro de las ciencias biológicas y la investigación médica que implica el estudio del conjunto completo de metabolitos (pequeñas moléculas involucradas en procesos metabólicos) dentro de un sistema biológico, como una célula, tejido u organismo. La metabolómica, que surge de los avances en genómica, proteómica y química analítica, permite a los investigadores examinar el estado fisiológico en tiempo real de un sistema mediante el análisis de los cambios dinámicos en estos metabolitos. Como los metabolitos son productos posteriores de los procesos celulares, proporcionan una lectura directa de la actividad bioquímica y el estado ge-

neral de una célula u organismo, lo que refleja cambios debidos a enfermedades, factores ambientales o intervenciones farmacológicas. El análisis metabolómico suele seguir uno de dos enfoques: dirigido o no dirigido. En la metabolómica dirigida, se cuantifican metabolitos conocidos específicos, mientras que la metabolómica no dirigida busca perfilar tantos metabolitos como sea posible, lo que puede revelar cambios metabólicos inesperados o nuevos biomarcadore (Klein and Shearer, 2015).

En la investigación clínica y biomédica, la metabolómica es fundamental en el descubrimiento de biomarcadores para diversas enfermedades, como la diabetes, el cáncer y las enfermedades cardiovasculares. Al estudiar los cambios metabólicos asociados con la progresión de la enfermedad o la respuesta al tratamiento, la metabolómica puede identificar biomarcadores que ayudan en el diagnóstico temprano, el pronóstico y la medicina personalizada. Además, la metabolómica tiene amplias aplicaciones en farmacología para estudiar el metabolismo de los fármacos, detectar efectos tóxicos y comprender la variabilidad interindividual en las respuestas a los fármacos. La metabolómica a menudo se lleva a cabo junto con otras ciencias "ómicas", como la genómica (estudio de genes y ADN), la transcriptómica (estudio de la expresión del ARN) y la proteómica (estudio de proteínas). La integración de estas disciplinas, conocida como multiómica, proporciona una comprensión más completa de los mecanismos celulares, permitiendo observar cómo convergen factores genéticos y ambientales a nivel metabólico. Esta visión holística es especialmente valiosa en el estudio de enfermedades complejas como la DMT2, donde interactúan factores genéticos, ambientales y metabólicos (Van Der Greef *et al.*, 2006).

Además de las aplicaciones clínicas, la metabolómica está estrechamente relacionada con la nutrigenómica, que examina la relación entre la nutrición y la expresión genética. La metabolómica también tiene implicaciones en la investigación del microbioma, ya que los perfiles metabólicos de la microbiota intestinal pueden afectar la salud del huésped e influir en enfermedades como la obesidad y la diabetes. La metabolómica ambiental, otro subconjunto, estudia los efectos de los contaminantes y factores estresantes ambientales en el metabolismo del organismo, proporcionando información sobre la toxicidad y la salud ecológica (Trujillo *et al.*, 2006).

2.2. Modelos de aprendizaje automático

El aprendizaje supervisado (*Supervised learning* en inglés) implica entrenar un modelo en un conjunto de datos etiquetados. Su objetivo es aprender a partir de un mapeo desde las características de entrada hasta las etiquetas de sali-

da, lo que habilita al modelo para realizar predicciones sobre datos no observados. Entre los algoritmos más comunes utilizados en el aprendizaje supervisado se encuentran la regresión lineal, la regresión logística, los *decision trees*, las máquinas de soporte vectorial y las redes neuronales.

En términos generales, el aprendizaje supervisado se define como un paradigma de ML donde el algoritmo aprende a partir de datos etiquetados, compuestos por pares de entrada y salida. El objetivo principal radica en aprender una función de mapeo: $f : X \rightarrow Y$, donde X representa el espacio de entrada y Y el espacio de salida correspondiente.

En este proceso, el aprendizaje supervisado busca aprender una regla general que asigne las entradas a las salidas. Los conjuntos de datos utilizados en este tipo de aprendizaje están previamente etiquetados (Gonzalez, 2020a).

2.2.1. Regresión Logística

La regresión logística (LR por las siglas de *Logistic Regression* en inglés) es un modelo logístico binario que se utiliza para estimar la probabilidad de una respuesta dicotómica (binaria) basada en una o más variables predictivas o independientes, es decir, que la presencia de un factor de riesgo aumenta la probabilidad de un resultado dado un porcentaje específico (Gonzalez, 2020b). Es además, un método estadístico utilizado para problemas de clasificación binaria, donde el resultado o la variable dependiente es categórico y típicamente dicotómico. A diferencia de la regresión lineal, que predice resultados continuos, la regresión logística predice la probabilidad de que ocurra el resultado de manera discreta.

La regresión logística se encarga de modelar la probabilidad de que una entrada dada x pertenezca a una clase particular. El modelo de regresión logística estima la probabilidad $P(y = 1|x)$ en función de las variables independientes $x = (x_1, x_2, \ldots, x_p)$.

La función logística, también conocida como función sigmoidea, se utiliza para asignar los valores predichos a probabilidades. Se define como:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.1}$$

donde z es una combinación lineal de las variables de entrada:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{2.2}$$

La salida de la función logística siempre está entre 0 y 1, lo que la hace adecuada para la estimación de probabilidad.

El modelo de regresión logística se puede expresar como:

$$P(y = 1|x) = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p) \tag{2.3}$$

De manera equivalente,

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}} \tag{2.4}$$

La probabilidad de la clase alternativa (es decir, $y = 0$) es:

$$P(y = 0|x) = 1 - P(y = 1|x) \tag{2.5}$$

La regresión logística a menudo se interpreta en términos de probabilidades y log-odds (logit). Las probabilidades de que ocurra un evento se definen como la relación entre la probabilidad de que ocurra el evento y la probabilidad de que no ocurra:

$$Probabilidad = \frac{P(y = 1|x)}{P(y = 0|x)} \tag{2.6}$$

Tomando el logaritmo natural de las probabilidades se obtiene el log-odds o logit:

$$Logit(P) = \log\left(\frac{P(y = 1|x)}{1 - P(y = 1|x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{2.7}$$

Los parámetros $\beta_0, \beta_1, \ldots, \beta_p$ normalmente se estiman utilizando el método de máxima verosimilitud. La función de verosimilitud $L(\beta)$ representa la probabilidad de los datos observados en función de los parámetros del modelo:

$$L(\beta) = \prod_{i=1}^{n} P(y_i|x_i)^{y_i}(1 - P(y_i|x_i))^{1-y_i} \tag{2.8}$$

Tomando el logaritmo natural de la función de verosimilitud se obtiene la función logarítmica de verosimilitud:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(P(y_i|x_i)) + (1 - y_i) \log(1 - P(y_i|x_i))] \tag{2.9}$$

Los parámetros contenidos en esta ecuación se estiman maximizando la función de probabilidad logarítmica, a menudo utilizando técnicas de optimización numérica como el ascenso de gradiente o el método de Newton-Raphson.

Los coeficientes β_j en la regresión logística representan el cambio en las probabilidades logarítmicas del resultado para un cambio de una unidad en el predictor x_j, manteniendo constantes todos los demás predictores. Específicamente,

para un aumento de unidad en x_j, las probabilidades del resultado se multiplican por e^{β_j}:

$$RP = e^{\beta_j} \tag{2.10}$$

Una razón de probabilidades mayor que 1 indica una asociación positiva entre el predictor y el resultado, mientras que una razón menor que 1 indica una asociación negativa.

La mejora de ajuste de un modelo de regresión logística se puede evaluar mediante varios métodos:

- **Desviación:** La desviación es una medida del ajuste del modelo a los datos, similar a la suma residual de cuadrados en la regresión lineal. Una desviación más baja indica un mejor ajuste.
- **AUC:** El AUC evalúa la capacidad del modelo para discriminar entre las dos clases de resultados. Un AUC más alto indica un mejor rendimiento del modelo.

2.2.2. Máquinas de Soporte Vectorial

Las máquinas de soporte vectorial (SVM por sus siglas, referente a, *Support Vector Machines* en inglés) tienen como objetivo encontrar un hiperplano que separe de la mejor forma posible dos clases diferentes de puntos de datos. "De la mejor forma posible" implica el hiperplano con el margen más amplio entre las dos clases (Burges, 1998).

Las SVM hacen referencia a un subconjunto de las observaciones de entrenamiento que identifican la ubicación del hiperplano de separación. El algoritmo SVM estándar está formulado para problemas de clasificación binaria, los problemas multiclase normalmente se reducen a una serie de problemas binarios.

Dado un conjunto de entrenamiento que consta de datos etiquetados (es decir, entradas de ejemplo y sus salidas deseadas) $(X_{train}, Y_{train}) = (x_1, y_1), ..., (x_l, y_l)$, el aprendizaje supervisado aprende una regla general que asigna entradas a las salidas. Esto es como un maestro (experto en etiquetado de datos) que le da a un estudiante un problema (encontrar la relación de mapeo entre entradas y salidas) y sus soluciones (datos de salida etiquetados) y le dice al estudiante que averigüe cómo resolver otros problemas similares: encontrar el mapeo de las características de muestras invisibles a sus etiquetas correctas o valores objetivo en el futuro (Zhang, 2020).

Las SVM son un método versátil y potente para tareas de clasificación y regresión. Su capacidad para crear límites de decisión complejos utilizando un

Tabla 2.2: Tipos de SVM y sus kernels de Mercer.

Tipo de SVM	**Kernel de Mercer**	**Descripción**
Función de base radial (RBF) o gaussiana	$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$	Aprendizaje de una clase. σ representa la anchura del kernel.
Lineal	$K(x_1, x_2) = x_1^T x_2$	Aprendizaje de dos clases.
Polinómica	$K(x_1, x_2) = (x_1^T x_2 + 1)^\rho$	ρ representa el orden del polinomio.
Sigmoide	$K(x_1, x_2) = \tanh(\beta_0 x_1^T x_2 + \beta_1)$	Representa un kernel de Mercer solo para determinados valores β_0 y β_1.

kernel 2.2, junto con una sólida base teórica, los convierte en una herramienta valiosa en el conjunto de herramientas del profesional del ML. Las fortalezas clave de SVM incluyen su robustez al sobreajuste, especialmente en espacios de alta dimensión, y su efectividad en problemas de clasificación tanto lineales como no lineales (Mathworks, 2024).

Para un conjunto de datos linealmente separable, SVM encuentra un hiperplano definido por la ecuación:

$$w \cdot x - b = 0 \tag{2.11}$$

donde:

- w es el vector de peso perpendicular al hiperplano.
- x es el vector de características.
- b es el término de sesgo.

El objetivo de SVM es maximizar el margen $\frac{2}{\|w\|}$ sujeto a la restricción de que todos los puntos de datos están clasificados correctamente. Esto lleva al siguiente problema de optimización:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{2.12}$$

Sujeto a las restricciones:

$$y_i(w \cdot x_i - b) \geq 1, \quad \forall i \tag{2.13}$$

donde y_i es la etiqueta de clase para el i-ésimo punto de datos.

Cuando los datos no son separables linealmente, SVM se puede ampliar para manejar límites de decisión no lineales utilizando el truco del kernel. Una función central $K(x_i, x_j)$ calcula el producto escalar de los puntos de datos en un espacio de características de dimensiones superiores, asignando implícitamente los datos originales a este espacio.

Usando kernel, el problema de optimización se convierte en:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{n} \alpha_i \tag{2.14}$$

Sujeto a:

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{2.15}$$

donde α_i son los multiplicadores de Lagrange y C es el parámetro de regularización que controla el equilibrio entre maximizar el margen y minimizar el error de clasificación.

Los vectores de soporte son los puntos de datos que se encuentran más cerca del límite de decisión. Estos puntos son críticos ya que definen la posición y orientación del hiperplano. Los puntos de datos que no son vectores de soporte no afectan al modelo final.

En la práctica, los conjuntos de datos a menudo no son perfectamente separables. Para manejar estos casos, SVM introduce un margen suave que permite algunas clasificaciones erróneas. El problema de optimización se modifica para incluir un término de penalización por puntos mal clasificados:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \tag{2.16}$$

Sujeto a:

$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i \tag{2.17}$$

donde ξ_i son las variables de holgura que miden el grado de clasificación errónea.

2.2.3. K-vecinos más cercanos

El algoritmo K-vecinos más cercanos (KNN por sus siglas de *K-nearest neighbors* en inglés), está basado en una instancia de tipo supervisado de ML. Puede

usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos) este un método que busca en las observaciones más cercanas a la que se está tratando de predecir y clasificar el punto de interés basado en la mayoría de datos que le rodean (Na8, 2020).

El algoritmo clasificador KNN funciona según el principio de encontrar K un número predefinido de muestras de entrenamiento que estén más cercanas a un nuevo punto y predecir una etiqueta para nuestro punto único utilizando estas muestras. La distancia euclidiana es la medida de distancia más utilizada, cuando se trata de datos densos o continuos, se recomienda encarecidamente hacer uso de esta. El clasificador KNN también se clasifica como un algoritmo de aprendizaje no generalizado basado en instancias. Almacena registros de datos de entrenamiento en un espacio multidimensional. Vuelve a calcular distancias euclidianas y predice la clase objetivo para cada nueva muestra y el valor de K. Como resultado, no genera un modelo interno generalizado (Mucherino *et al.*, 2009).

La distancia euclidiana se define por:

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2}, \tag{2.18}$$

donde x_i es una muestra de entrada con p características ($x_{i1}, x_{i2}, \ldots, x_{ip}$), n es el número total de muestras de entrada ($i = 1, 2, \ldots, n$) y p el número total de características ($j = 1, 2, \ldots, p$).

$d(x_i, x_l)$ es la distancia euclidiana entre los puntos. Luego, en orden no descendente, calculamos el n distancias euclidianas. Tomamos el primero k distancias de esta lista ordenada porque k es un número entero. Luego encontramos el $k - puntos$ que corresponden a estos $k - distancias$. Finalmente, definimos k_i como el número de puntos en la $iesima$ clase entre los puntos k.

Su efectividad depende de la elección de la métrica de distancia y del valor de K. Si bien es computacionalmente intensivo y sensible a la maldición de la dimensionalidad, sigue siendo una opción popular para muchas aplicaciones del mundo real debido a su naturaleza intuitiva y flexibilidad. KNN almacena todos los casos disponibles y clasifica los casos nuevos basándose en una medida de similitud. Se denomina algoritmo de aprendizaje diferido porque no aprende explícitamente un modelo durante la fase de entrenamiento. En cambio, retrasa el cálculo hasta la fase de predicción.

La elección de la métrica de distancia es crucial para el desempeño de KNN. Las métricas de distancia comúnmente utilizadas incluyen:

- **Distancia euclidiana:** La métrica de distancia más utilizada, que se define como:

$$d(x, x') = \sqrt{\sum_{i=1}^{n}(x_i - x'_i)^2} \tag{2.19}$$

- **Distancia de Manhattan:** También conocida como distancia L1 o distancia de cuadra de ciudad, definida como:

$$d(x, x') = \sum_{i=1}^{n} |x_i - x'_i|$$

- **Distancia de Minkowski:** Una generalización de las distancias euclidiana y de Manhattan, definida como:

$$d(x, x') = \left(\sum_{i=1}^{n} |x_i - x'_i|^p \right)^{1/p}$$

  donde p es un parámetro que determina el tipo de distancia. Para $p = 2$, se convierte en distancia euclidiana, y para $p = 1$, se convierte en distancia de Manhattan.

- **Distancia de Hamming:** Se utiliza para variables categóricas, definidas como el número de posiciones en las que los elementos correspondientes son diferentes.

El algoritmo KNN se puede resumir en los siguientes pasos:

1. **Elija el número de vecinos K:** Esta es una constante definida por el usuario.
2. **Calcule la distancia:** Calcule la distancia entre la instancia de consulta y todas las muestras de entrenamiento utilizando una métrica de distancia adecuada.
3. **Ordenar las distancias:** Ordenar las distancias calculadas en orden ascendente.
4. **Seleccione los vecinos más cercanos:** Elija las primeras K entradas de la lista ordenada.
5. **Asigne la etiqueta:** Para la clasificación, asigne la etiqueta de clase más frecuente entre los K vecinos más cercanos. Para la regresión, calcule el promedio de las respuestas de los K vecinos más cercanos.

El rendimiento del algoritmo KNN depende en gran medida de la elección de K. Un valor pequeño de K hace que el algoritmo sea sensible al ruido, mientras que un valor grande de K lo hace computacionalmente intensivo y puede hacer que el algoritmo no se ajuste adecuadamente. Un enfoque común es utilizar validación cruzada para encontrar el valor óptimo de K.

2.2.4. Redes neuronales artificiales

Las redes neuronales artificiales (ANN por sus siglas de *Artificial neural networks* en inglés), están basadas en el funcionamiento de las redes de neuronas biológicas. En el caso de las neuronas artificiales, la suma de las entradas multiplicadas por sus pesos asociados determina el "impulso nervioso" que recibe la neurona. Este valor, se procesa en el interior de la célula mediante una función de activación que devuelve un valor que se envía como salida de la neurona (XERIDIA, 2021).

Una red neuronal artificial está formada por neuronas artificiales conectadas entre sí y agrupadas en diferentes niveles que denominamos capas. Una capa es un conjunto de neuronas cuyas entradas provienen de una capa anterior (o de los datos de entrada en el caso de la primera capa) y cuyas salidas son la entrada de una capa posterior.

Las ANN están inspiradas en el cerebro y están compuestas por neuronas artificiales interconectadas capaces de realizar ciertos cálculos sobre sus entradas (Hornik *et al.*, 1989). Los datos de entrada activan las neuronas en la primera capa de la red cuya salida es la entrada a la segunda capa de neuronas en la red. De manera similar, cada capa pasa su salida a la siguiente capa y la última capa genera el resultado. Las capas entre las capas de entrada y salida se denominan capas ocultas. Cuando se utiliza una ANN como clasificador, la capa de salida genera la categoría de clasificación final (Zhang, 2020).

$$y = \sum_{i=1}^{n} (w_i * x_i) + bias \tag{2.20}$$

La unidad básica de una ANN es la **neurona**, también conocida como **nodo** o **perceptrón**. Una neurona recibe una o más entradas, las procesa y produce una salida. La estructura fundamental de una ANN incluye:

- **Capa de entrada:** Esta capa recibe las señales de entrada.
- **Capas ocultas:** Estas capas realizan cálculos intermedios y extracción de características. Puede haber una o más capas ocultas en una ANN.
- **Capa de salida:** Esta capa produce la salida final de la red.

Cada neurona de una capa está conectada a cada neurona de la capa siguiente. Las conexiones tienen pesos asociados w que se ajustan durante el entrenamiento para minimizar el error de la red. La salida de una neurona se calcula de la siguiente manera:

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{2.21}$$

donde x_i son las entradas, w_i son los pesos y b es el término de sesgo. La activación de la neurona a se calcula usando una función de activación ϕ:

$$a = \phi(z) \tag{2.22}$$

Las funciones de activación comunes incluyen:

- **Sigmoidea:** $\phi(z) = \frac{1}{1+e^{-z}}$
- **Tangente hiperbólica (tanh):** $\phi(z) = \tanh(z)$
- **Unidad Lineal Rectificada(ReLU):** $\phi(z) = \max(0, z)$
- **Softmax:** Se utiliza en la capa de salida para tareas de clasificación, definida como $\phi(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$

Entrenar una ANN implica ajustar los pesos y sesgos para minimizar el error entre la salida prevista y la salida real. Este proceso normalmente se realiza mediante retropropagación y descenso de gradiente.

La **Retropropagación (Backpropagation en inglés)** es un algoritmo utilizado para calcular el gradiente de la función de pérdida con respecto a cada peso mediante la regla de la cadena. Los pasos involucrados en la retropropagación son:

1. **Paso hacia adelante:** Calcula la salida de la red para una entrada determinada.
2. **Pérdida de cálculo:** Calcule el error entre la salida prevista y la salida real usando una función de pérdida, como el error cuadrático medio (MSE) para regresión o pérdida de entropía cruzada para clasificación.
3. **Pase hacia atrás:** Calcule el gradiente de la función de pérdida con respecto a cada peso usando la regla de la cadena.

4. **Actualizar pesos:** Ajuste los pesos usando el descenso de gradiente:

$$w_{ij} = w_{ij} - \eta \frac{\partial L}{\partial w_{ij}}$$

donde η es la tasa de aprendizaje, L es la función de pérdida y $\frac{\partial L}{\partial w_{ij}}$ es el gradiente de la función de pérdida con respecto al peso w_{ij}.

2.2.5. Centroide más cercano

El centroide más cercano (Nearcent abreviado de *Nearest Centroid* en inglés) es uno de los clasificadores más simples, sin embargo, es capaz de clasificar datos sin ninguna selección de características, por ejemplo, espectros de masas sin procesar (I., 2005). Además, es extremadamente rápido y requiere baja potencia computacional, proporciona una base para la evaluación de algoritmos de selección de características y permite probar una serie de algoritmos que antes no eran aplicables. Nearcent y KNN proporcionan enfoques similares cuando hay conocimiento limitado en la distribución, brindando validación mutua de los resultados de las clasificaciones.

El clasificador de centroide más cercano es un algoritmo simple e intuitivo que se utiliza en el reconocimiento de patrones y el ML para tareas de clasificación. Asigna una etiqueta de clase a una muestra en función de la proximidad de la muestra al centroide (media) de las clases en el espacio de características.

$$\vec{\mu_\ell} = \frac{1}{|C_\ell|} \sum_{i \in C_\ell} x_i \tag{2.23}$$

Se le dieron muestras de entrenamiento etiquetadas. $(\vec{x_1}, y_1)$, ...,$(\vec{x_n}, y_n)$ con etiquetas de clase $y_i \in Y$, calcular los centroides por clase $\vec{\mu}_\ell = \frac{1}{|C_\ell|} \sum_{i \in C_\ell} \vec{x}_i$ donde C_ℓ es el conjunto de índices de muestras pertenecientes a la clase $\ell \in Y$.

El clasificador de centroide más cercano funciona de la siguiente manera:

1. **Entrenamiento:** Calcule el centroide para cada clase. El centroide es el vector medio de todas las muestras que pertenecen a esa clase.

2. **Predicción:** Asigne cada muestra de prueba a la clase cuyo centroide esté más cerca de la muestra en términos de una métrica de distancia elegida.

Sea $X = x_1, x_2, \ldots, x_n$ el conjunto de datos de entrenamiento con n muestras, donde cada muestra $x_i \in R^d$ es un vector d-dimensional. Sean $Y = y_1, y_2, \ldots, y_n$ las etiquetas correspondientes, donde $y_i \in 1, 2, \ldots, K$ y K es el número de clases.

Para cada clase $k \in 1, 2, \ldots, K$, calcula el centroide c_k. El centroide c_k se define como la media de todas las muestras en la clase k:

$$c_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i \tag{2.24}$$

donde C_k es el conjunto de todas las muestras que pertenecen a la clase k, y $|C_k|$ es el número de muestras en la clase k.

Para una muestra de prueba x, calcule la distancia desde x a cada centroide c_k. La muestra se asigna a la clase k^* cuyo centroide es el más cercano a x:

$$k^* = \arg\min_k d(x, c_k) \tag{2.25}$$

donde $d(x, c_k)$ es la distancia entre x y c_k. Las métricas de distancia comunes incluyen:

- **Distancia euclidiana:**

$$d(x, c_k) = \sqrt{\sum_{j=1}^{d} (x_j - c_{kj})^2}$$

- **Distancia de Manhattan:**

$$d(x, c_k) = \sum_{j=1}^{d} |x_j - c_{kj}|$$

- **Similitud del coseno:**

$$d(x, c_k) = 1 - \frac{x \cdot c_k}{\|x\| \|c_k\|}$$

2.2.6. Bosques aleatorios

Los bosques aleatorios (*random forest* o RF por sus siglas en inglés) es un algoritmo de ML utilizado tanto para la clasificación como para la regresión. Se basa en la creación de múltiples árboles de decisión durante el entrenamiento y devuelve la clase que es el modo de las clases (clasificación) o la media de las predicciones (regresión) de cada árbol. Este enfoque mejora la precisión y controla el sobreajuste, haciendo de RF una técnica robusta en el ámbito de la minería de datos y la inteligencia artificial.

Un árbol de decisión es un modelo predictivo que utiliza un conjunto de reglas de decisión basadas en las características del conjunto de datos. Cada nodo interno del árbol representa una prueba sobre una característica, cada rama representa el resultado de la prueba y cada hoja representa una clase o un valor de predicción. La forma matemática de un árbol de decisión puede describirse como:

$$f(X) = \sum_{i=1}^{n} I(X \in R_i) \cdot C_i \tag{2.26}$$

donde X es el conjunto de características, R_i es la región en la que X cae, C_i es la clase o el valor de predicción asociado, e I es la función indicadora que es 1 si X está en R_i y 0 en caso contrario.

RF utiliza una técnica llamada *bagging* para construir sus árboles. El proceso consiste en tomar múltiples muestras de los datos de entrenamiento con reemplazo, lo que se conoce como muestreo bootstrap. Esto significa que cada árbol en el bosque se entrena con un subconjunto diferente de datos, lo que introduce diversidad en el modelo y reduce la varianza. Formalmente, si D es el conjunto de datos original, el conjunto D_i para el árbol i se define como:

$$D_i = \{x_1, x_2, \ldots, x_n\} \quad \text{con} \quad x_j \sim D \text{ para } j = 1, 2, \ldots, n \tag{2.27}$$

Al construir cada árbol, RF selecciona un subconjunto aleatorio de características en cada nodo para determinar la mejor división. Este proceso ayuda a hacer que el modelo sea más robusto y evita que se base demasiado en una sola característica. La selección aleatoria de características se puede expresar como:

$$S = \{f_1, f_2, \ldots, f_m\} \quad \text{donde } f_i \sim F \text{ (con } m \ll p\text{)} \tag{2.28}$$

donde F es el conjunto completo de características y p es el número total de características en el conjunto de datos.

Una vez que todos los árboles han sido construidos, las predicciones se realizan mediante la agregación de las predicciones de todos los árboles individuales. Para la clasificación, se utiliza la mayoría de votos:

$$\hat{y} = \text{mode}(y_1, y_2, \ldots, y_n) \tag{2.29}$$

donde y_i son las predicciones de los árboles individuales. Para la regresión, se calcula el promedio de las predicciones:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{2.30}$$

2.3. Clasificación y regresión

ML implica aprender de conjuntos de datos de entrenamiento para resolver dos problemas básicos: regresión, que se utiliza para predecir salidas continuas, y clasificación, que se emplea para predecir salidas discretas. La clasificación está estrechamente relacionada con el reconocimiento de patrones, ya que su objetivo es entrenar un modelo para reconocer o clasificar muestras desconocidas basándose en un conjunto de datos de entrenamiento. Por otro lado, la regresión implica diseñar un modelo predictivo basado en datos de entrenamiento para predecir valores continuos para muestras desconocidas.

2.3.1. Clasificación

La clasificación constituye una actividad recurrente en la vida humana. Se plantea un problema de clasificación cuando se requiere asignar un objeto a un grupo o categoría predefinida, en base a un conjunto de atributos observados luego de clasificar varios grupos o categorías. En esencia, se busca determinar a qué clase pertenece una muestra de prueba específica entre varias clases disponibles. Esta fase, conocida como la etapa de prueba, contrasta con el período de entrenamiento previo (Zhang, 2020). Los posibles resultados de una tarea de clasificación se denominan clases o etiquetas, estos resultados provienen del conjunto de datos utilizado para probar un modelo ya entrenado que consta de ejemplos con etiquetas de clase ya conocidas, estos datos ayudan al modelo a conocer las relaciones entre las entidades de entrada y las etiquetas de salida. Para probar un modelo ML entrenado se requiere de un conjunto de datos independiente que es utilizado para evaluar su rendimiento, este conjunto de datos consta de valores que no se observaron durante la fase de entrenamiento, lo que permite una evaluación imparcial del poder predictivo del modelo, la función o algoritmo matemático que asigna características de entrada a etiquetas predichas (Liaw *et al.*, 2002).

Los principales desafíos en la clasificación:

- Datos desbalanceados: Cuando una clase supera significativamente a otras, puede dar lugar a modelos sesgados. Técnicas como el remuestreo, la ponderación de clases y la generación de datos sintéticos, pueden ayudar a abordar este problema.

- Ingeniería de características: La calidad de las características afecta significativamente el rendimiento del modelo. La selección e ingeniería de características efectivas son cruciales para construir clasificadores robustos.

- Sobreajuste: Cuando un modelo aprende ruido en los datos de entrenamiento, es posible que tenga un rendimiento deficiente con datos nuevos. Las técnicas de regularización, la validación cruzada y los métodos de conjunto ayudan a mitigar el sobreajuste.

2.3.2. Regresión

En el ámbito del modelado estadístico y el ML, la regresión se presenta como un proceso estadístico diseñado para estimar las relaciones entre variables. Su enfoque se dirige hacia la relación entre una variable dependiente x y una o más variables independientes (llamadas "Predictores") β. Específicamente, la variable dependiente es el resultado o variable objetivo que el modelo pretende predecir, mientras que las variables independientes son las características de entrada o predictores utilizados. La regresión consiste en analizar cómo varía el valor típico de la variable dependiente cuando se modifica cualquiera de las variables independientes, manteniendo fijas las demás variables independientes. Es una tarea centrada en predecir resultados continuos basados en datos de entrada, que a diferencia de la clasificación, que categoriza de manera discreta, los modelos de regresión estiman productos de valor real, esto hace que la regresión sea esencial para numerosas aplicaciones donde la comprensión y predicción de relaciones numéricas son cruciales (Zhang, 2020).

Regresión lineal: La regresión lineal modela la relación entre la variable dependiente y una o más variables independientes ajustando una ecuación lineal a los datos observados.

Para una regresión lineal simple con una variable independiente, el modelo es:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.31}$$

donde y es la variable dependiente, x es la variable independiente, β_0 es la intercepción, β_1 es la pendiente, y ϵ es el término de error.

Regresión lineal múltiple: Extiende este concepto a múltiples variables independientes, con el modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon \tag{2.32}$$

Regresión polinomial: La regresión polinómica o polinomial modela la relación entre la variable dependiente y la variable independiente como un polinomio de $n$ésimo grado. Capta relaciones no lineales mediante la introducción de términos polinomiales de las variables independientes.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n + \epsilon \tag{2.33}$$

Regresión Ridge y LASSO Se trata de versiones regularizadas de regresión lineal que introducen penalizaciones para evitar el sobreajuste.

Regresión Ridge: Agrega una penalización igual al cuadrado de la magnitud de los coeficientes a la función de pérdida.

$$P = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.34}$$

P: Pérdida.

Regresión LASSO: Añade una penalización igual al valor absoluto de la magnitud de los coeficientes.

$$P = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.35}$$

P: Pérdida. Otros modelos que utilizan regresión son: Los Decision Trees, estos modelos de regresión predicen el valor de la variable dependiente aprendiendo reglas de decisión inferidas de las características. Los datos se dividen en subconjuntos según los valores de las características, formando una estructura de árbol. Otro modelo que utiliza regresión consta de un método de conjunto que crea múltiples Decision Trees y combina sus predicciones para mejorar la precisión y controlar el sobreajuste llamado RF. Cada árbol se basa en un subconjunto aleatorio de datos y características, y la predicción final es el promedio de las predicciones de todos los árboles individuales. Para tareas de regresión, las redes neuronales constan de capas de nodos o neuronas que transforman los datos de entrada mediante pesos aprendidos para predecir valores continuos.

2.3.3. Reconocimiento y clasificación de patrones

El reconocimiento de patrones es un campo dedicado al estudio de métodos diseñados para categorizar datos en clases distintas. De acuerdo con el autor Watanabe (Watanabe, 1985), un patrón se define como “opuesto al caos”, es una entidad, vagamente definida, a la que se le podría dar un nombre”. Este concepto se aplica ampliamente en el reconocimiento biométrico humano (como el reconocimiento facial, de huellas dactilares y del iris) y en la identificación de diversos objetivos (como aviones y barcos). En estas aplicaciones, la extracción de características de los objetos es un aspecto crucial. Por ejemplo, cuando un objetivo se considera un sistema lineal, el parámetro objetivo se convierte en una característica de la señal del objetivo. Dado un patrón, su reconocimiento y clasificación pueden implicar una de las dos tareas siguientes (Watanabe, 1985):

- clasificación supervisada en la que el patrón de entrada se identifica como miembro de una clase predefinida,
- clasificación no supervisada en la que el patrón se asigna a una clase desconocida hasta ahora. Las tareas de reconocimiento y clasificación de patrones se dividen habitualmente en cuatro bloques distintos:
    * Representación de datos (adquisición y preprocesamiento),
    * Selección o extracción de características,
    * Agrupación,
    * clasificación.

Los enfoques para el reconocimiento de patrones propuestos por Jain *et al.* (2000) son:

- Coincidencia de plantillas,
- clasificación estadística,
- Coincidencia sintáctica o estructural,
- Redes neuronales.

2.4. Imputación de datos

Con el fin de tratar o lidiar con los datos faltantes que forman parte de un conjunto de observaciones con características especiales que incluyen: datos agrupados, agregados, redondeados, censurados o truncados, se utilizan una serie de reglas de inserción de caracteres o valores (Medina and Galván, 2007).

La imputación de datos consiste en el tratamiento de los valores faltantes:

2.4.1. Análisis con datos completos

El análisis con datos completos (Listwise case deletion en inglés), Consiste en trabajar únicamente con las observaciones que disponen de información completa para todas las variables y se desecha el resto de las demás variables (Medina and Galván, 2007).

2.4.2. Análisis con los datos disponibles

En contraste con el caso anterior, el análisis con los datos disponibles (pairwise deletion en inglés) utiliza distintos tamaños de muestra, por lo que también se le conoce como *pairwise deletion* o *pairwise inclusion* (Medina and Galván, 2007).

2.5. Selección de características

En la revisión de la literatura realizada, se presentan los distintos alcances y aportaciones sobre la detección de la DMT2, en esta investigación se promueve que es más relevante la selección de características que los modelos utilizados para la evaluación de estos conjuntos. La relevancia en este trabajo radica en la obtención de biomarcadores potenciales por medio de diversas técnicas de selección y la implementación de modelos ML de ensamble, siendo las características verdaderamente relevantes las que aparecen repetidamente, independientemente de la técnica implementada. Esto implica que no importa cuantos modelos se implementen o cuantas técnicas de tratamiento se apliquen al conjunto de datos, la selección de las características da la pauta para la identificación de los biomarcadores potenciales. Cada técnica de selección de características tiene su aproximación en cuanto a su evaluación interna, siendo las de reducción de dimensionalidad o el uso de modelos de ML como evaluadores del rendimiento de los conjuntos de características seleccionados o correspondientes a una iteración en particular dentro de la implementación de la selección de características. En esta investigación a su vez se tomó en cuenta los rangos de valores de cada característica obtenida, resarciendo así el objetivo de brindar biomarcadores potenciales por medio de conjuntos personalizados de características y valores dependientes al paciente que se está evaluando, esta herramienta desarrollada, presenta así, una premisa de medicina personalizada para ese paciente.

La selección de características es el método mediante el cual elegimos meticulosamente los atributos más pertinentes de la base de datos. En el campo de la investigación biomédica, el objetivo principal de la selección de características es identificar variables clínicamente significativas y estadísticamente sólidas, excluyendo las que no están relacionadas o son ruidosas. Hay numerosos métodos disponibles para la selección de características, cada uno con su propio conjunto de ventajas y limitaciones, también a su vez casa técnica de selección depende del dataset y de su preprocesamiento (Hutter *et al.*, 2013).

Las modelos más utilizados para la selección de características son:

2.5.1. Selección hacia adelante

La selección hacia adelante (forward selection en inglés), Es una herramienta genérica independiente del modelo que se puede utilizar para resolver este problema. Específicamente, este método identifica conjuntos de modelos con entradas que son suficientes en conjunto para lograr una buena precisión predictiva (Hutter *et al.*, 2013).

La selección directa ofrece un enfoque eficiente para reducir la dimensionali-

dad de datos de alta dimensión. La razón fundamental para emplear la selección directa secuencial (SFS) radica en su eficacia para revelar las intrincadas interacciones entre características. SFS es un método contenedor de selección de características, este algoritmo se inicia con un conjunto de características vacío y comienza agregando la característica con la puntuación de importancia de característica más alta a este conjunto vacío. Luego, estas características se clasifican según su importancia, lo que permite extraer las de mejor rendimiento utilizando un clasificador de árbol adicional. Una propuesta hecha por Selim Buyrukoglu y Ayhan Akbasis fue implementar este conjunto de técnicas de selección de características y modelos de ML para identificar biomarcadores para la detección temprana de DMT2, con resultados prometedores (Buyrukoğlu and Akbaş, 2022).

2.5.2. LASSO

La técnica LASSO (*Least Absolute Shrinkage and Selection Operator* en inglés) impone una restricción a la suma de los valores absolutos de los parámetros del modelo, la suma debe ser menor que un valor fijo (Fonti and Belitser, 2017).

LASSO se utiliza en este trabajo ya que tiene el valor P menor de los selectores de características propuestos por Liu *et al.* (2019), los otros selectores son: desviación absoluta recortada suavizada (SCAD), probabilidad penalizada cóncava minimax (MCP), regresión logística por pasos y detección iterativa de independencia segura (ISIS). El enfoque gradual agrega de forma iterativa la mejor variable en cada ciclo, lo que normalmente produce un conjunto aceptable de variables. Sin embargo, puede verse limitado por una tendencia a quedarse estancado en óptimos locales. Por el contrario, el mejor enfoque de subconjunto explora sistemáticamente todo el espacio de patrones de covariables, pero puede resultar difícil de manejar cuando se trata de conjuntos de datos que contienen decenas o cientos de variables, un escenario común en los datos clínicos actuales.

Se implementa LASSO, como una solución rápida y sólida como se presenta en el estudio realizado por Kocbek *et al.* (2022), fueron conjuntos ajustados a modelos lineales generalizados y similares a través de la máxima verosimilitud penalizada. La implementación LASSO desarrollada resuelve el problema:

$$\min_{(\beta_0,\,\beta)\,\in\,R^{p+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda P_a(\beta)\right], \tag{2.36}$$

donde:

$$P_a(\beta)=(1-\alpha)\frac{1}{2}\|\beta\|_{\ell_2}^2+\alpha\|\beta\|_{\ell_1} \tag{2.37}$$

$$= \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right] \tag{2.38}$$

que es la penalización de red elástica (Zou and Hastie, 2005). P_α es un compromiso entre la penalización de la *ridge-regression* o regresión Ridge $(\alpha = 0)$ y la penalización de LASSO $(\alpha = 1)$. β_0 es el coeficiente constante y β un vector de coeficientes. La penalización de la elastic-net o red elástica está controlada por α y cierra la brecha entre la regresión LASSO (α=1, el valor predeterminado) y la regresión Ridge (α =0). El parámetro de sintonización λ controla la fuerza general de la pena. Esta implementación utilizó un enfoque de validación cruzada, que realizó una validación cruzada k veces para glmnet y devuelve un valor para λ (Friedman *et al.*, 2010).

Se sabe que la penalización ridge reduce los coeficientes de los predictores correlacionados a medida que se acercan entre sí, mientras que LASSO tiende a elegir un coeficiente correlacionado y descartar los demás. La penalización de red elástica realizada por el paquete glmnet (Friedman *et al.*, 2010) Combina estos dos: si los predictores están correlacionados en grupos, α=0.5 tiende a seleccionar u omitir todo el grupo de características, lo que da como resultado un conjunto de hiperparámetros utilizables para el modelo.

2.5.3. Algoritmos genéticos

Está inspirado en el proceso genético de la biología de los organismos. Los algoritmos genéticos (AG) constan de varias soluciones llamadas cromosomas o individuos. Cada cromosoma en un AG binario incluye varios genes con valores binarios 0 y 1, que determinan los atributos de cada individuo. Un conjunto de los cromosomas se componen para formar una población. El mérito de cada cromosoma se evalúa utilizando una función de ajuste. Son heurísticas de búsqueda inspiradas en los principios de la selección natural y la genética. Se utilizan para encontrar soluciones aproximadas a problemas de optimización y búsqueda. Los AG pertenecen a la clase de algoritmos evolutivos, que utilizan mecanismos inspirados en la evolución biológica, como la reproducción, la mutación, la recombinación y la selección. Los AG ofrecen una solución como métodos de optimización heurística para la selección de variables en modelos de regresión multivariable. Este documento presenta una implementación paso a paso para utilizar AG para la selección de características. Proporciona no sólo conocimientos teóricos sino también una implementación práctica en forma de código R que se puede adaptar a diversos requisitos de análisis de datos. La implementación de la selección de características mediante algoritmos genéticos, específicamente con GALGO, es

de suma importancia. GALGO ofrece un enfoque sistemático y eficiente para elegir las características más relevantes de conjuntos de datos médicos complejos. Este método, como se demuestra en este artículo, es fundamental para mejorar el rendimiento de los modelos predictivos, asegurando que solo se incluyan las variables clínicamente más significativas y que se excluyan las variables de ruido irrelevantes. Esto tiene un impacto directo en la calidad y confiabilidad de los procesos de toma de decisiones médicas, lo que lo convierte en una contribución crucial al campo de la atención médica y la investigación médica (Zhang *et al.*, 2018).

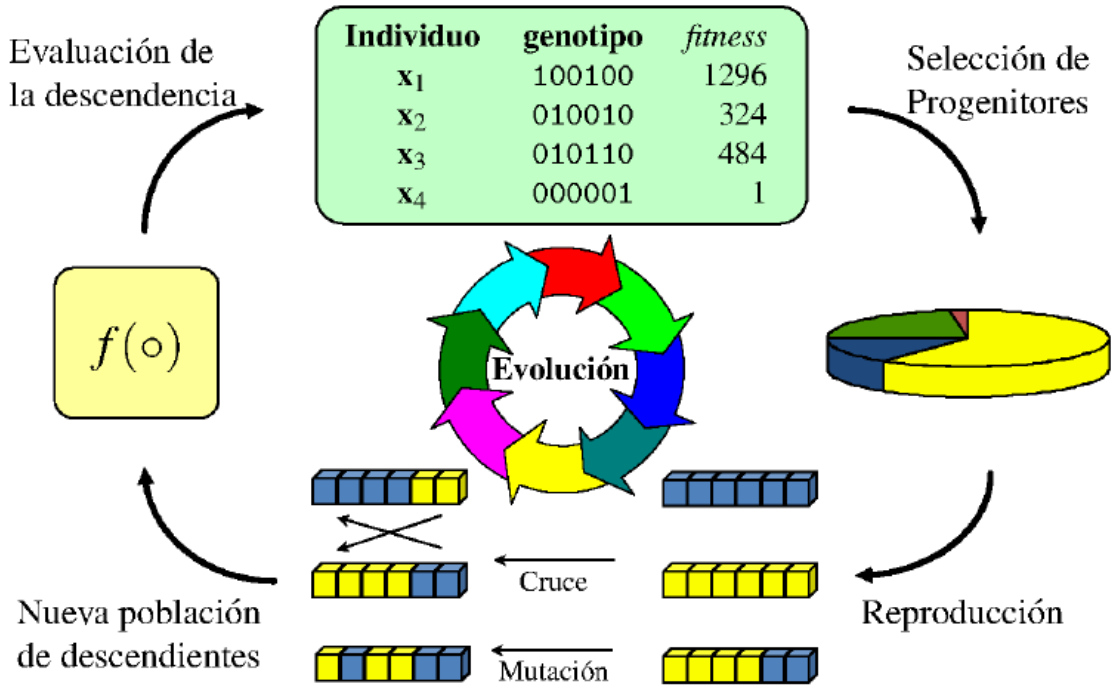


Figura 2.2: Mapa explicativo de un algorítmo genético

Los cromosomas aptos se seleccionan para la generación de nuevos cromosomas como se muestra en la figura 2.2 basada en la propuesta por Holland (1992). En ese paso, se seleccionan dos cromosomas aptos y combinados a través de un paso cruzado para producir una nueva descendencia (o solución). Luego, la mutación se aplica a la población para aumentar la aleatoriedad de los individuos para disminuir la posibilidad de quedarse atascado en el óptimo local (Ghamisi and Benediktsson, 2015).

Los algoritmos genéticos constan de:

1. **Población**: La población es un conjunto de posibles soluciones al problema

dado. Cada individuo de la población representa una posible solución y, a menudo, está codificado como una cadena de bits, caracteres o números.

2. **Cromosomas y genes**: En el contexto de los AG, un cromosoma representa una solución individual y está compuesto de genes. Cada gen codifica una parte de la solución y el conjunto de genes (cromosomas) determina la calidad de la solución.

3. **Función de *fitness***: La función fitness evalúa qué tan buena es una solución. Asigna una puntuación de ajuste a cada individuo de la población en función de qué tan bien resuelve el problema. El objetivo del AG es maximizar (o minimizar) la función fitness.

Los algoritmos genéticos constan de 3 operadores característicos:

1. **Selección**: La selección es el proceso de elegir individuos de la población para crear descendencia para la próxima generación. Los individuos con mayor ajuste (*fitness* en inglés) tienen una mayor probabilidad de ser seleccionados. Los métodos de selección comunes incluyen:

   - **Selección de la ruleta**: La probabilidad de seleccionar un individuo es proporcional a su ajuste.
   - **Selección de torneo**: se elige al azar un subconjunto de individuos y se selecciona el que tiene mejor ajuste del subconjunto.

2. ***Crossover* (recombinación)**: *Crossover* es el proceso de combinar dos cromosomas parentales para producir descendencia. Su objetivo es crear nuevos individuos que hereden los mejores rasgos de sus padres. Las técnicas de cruce comunes incluyen:

   - **Cruce de punto único**: se selecciona un punto de cruce y los segmentos antes y después de este punto se intercambian entre los padres.
   - **Cruce de puntos múltiples**: se seleccionan múltiples puntos de cruce y los segmentos se intercambian alternativamente entre los padres.
   - **Cruce uniforme**: Cada gen se elige independientemente de uno de los padres con una cierta probabilidad.

3. **Mutación**: La mutación introduce cambios aleatorios en genes individuales para mantener la diversidad genética dentro de la población. Evita que el algoritmo quede atrapado en óptimos locales. Un método de mutación común es:

   Mutación de inversión de bits: se invierte un bit seleccionado aleatoriamente en el cromosoma (0 se convierte en 1 y viceversa).

Su proceso es el siguiente:

1. **Inicialización**: comience con una población inicial de individuos generados aleatoriamente.
2. **Evaluación**: Evalúa el ajuste de cada individuo de la población utilizando la función de ajuste.
3. **Selección**: seleccione individuos según su aptitud para formar un grupo de cruce.
4. **Cruce y mutación**: aplique operadores de cruce y mutación al grupo de cruce para producir una nueva generación de individuos.
5. **Reemplazo**: Reemplazar la población actual por la nueva generación.
6. **Terminación**: Repita los pasos de evaluación, selección, cruce y mutación hasta que se cumpla una condición de terminación.

Sea $P(t)$ la población en la generación t, y $f(i)$ el ajuste del individuo i en la población. Los pasos clave del AG se pueden formular matemáticamente de la siguiente manera:

1. **Selección**: La probabilidad de seleccionar el individuo i puede venir dada por:
$$P(i) = \frac{f(i)}{\sum_{j=1}^{N} f(j)} \tag{2.39}$$
donde N es el tamaño de la población.
2. ***Crossover***: Para cruce de un solo punto, si $x_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ y $x_2 = (x_{21}, x_{22}, \ldots, x_{2n})$ son dos cromosomas padres, y c es el punto de cruce, la descendencia y_1 y y_2 son:
$$y_1 = (x_{11}, x_{12}, \ldots, x_{1c}, x_{2(c+1)}, \ldots, x_{2n}) \tag{2.40}$$
$$y_2 = (x_{21}, x_{22}, \ldots, x_{2c}, x_{1(c+1)}, \ldots, x_{1n}) \tag{2.41}$$
3. **Mutación**: Para la mutación de inversión de bits, si $x = (x_1, x_2, \ldots, x_n)$ es un cromosoma y m es el punto de mutación, el cromosoma mutado y es:
$$y = (x_1, x_2, \ldots, -x_m, \ldots, x_n) \tag{2.42}$$
donde $-x_m$ denota el bit invertido de x_m.

Los algoritmos genéticos se utilizan en esta investigación, para:

- **Optimización**: Encontrar soluciones óptimas a problemas complejos en la identificación de biomarcadores para la detección de DMT2 tomando como base un conjunto de datos previamente etiquetado.
- **ML**: Selección de características, ajuste de hiperparámetros y arquitecturas de distintos modelos para evaluación de conjuntos.

2.5.4. Eliminación recursiva de características

La eliminación recursiva de características (*recursive feature elimination* o RFE por sus siglas en inglés) es una técnica de selección de características comúnmente utilizada en ML y modelado estadístico para mejorar el rendimiento y la interpretabilidad del modelo mediante la identificación y selección de las características más relevantes, esto lo hace por medio de un proceso iterativo que elimina sistemáticamente características menos importantes para mejorar la precisión predictiva del modelo y reducir el sobreajuste. RFE es particularmente útil cuando se trata de conjuntos de datos de alta dimensión, ya que puede reducir de manera eficiente el espacio de características mientras mantiene la precisión del modelo.

La idea principal detrás de RFE es construir un modelo de forma iterativa, evaluar la importancia de las características y eliminar las características menos significativas en función del rendimiento del modelo. RFE tiene como objetivo encontrar un subconjunto de características que contribuyan más a predecir la variable objetivo. Por lo general, esto se hace utilizando un estimador predefinido (por ejemplo, regresión lineal, máquinas de vectores de soporte o bosque aleatorio) para evaluar la importancia de las características en cada paso. El proceso RFE se puede resumir de la siguiente manera:

1. **Inicializar:** Inicia con todas las características del conjunto de datos.
2. **Ajuste del modelo:** Entrena un modelo de ML con todo el conjunto de características y evalúa la importancia o el peso de cada característica. Para los modelos lineales, los coeficientes (β_i) pueden representar la importancia de las características, mientras que para otros modelos (por ejemplo, modelos basados en árboles), se pueden utilizar puntuaciones de importancia de las características.
3. **Clasificación y eliminación de características:** Identifica las características menos importantes según un criterio, como el coeficiente absoluto más pequeño o la puntuación de importancia más baja, y las elimina del conjunto de características.

4. **Repetir:** Vuelve a colocar el modelo con el conjunto de características reducido y repite el proceso hasta alcanzar un número específico de características o hasta que se cumpla otro criterio de parada.

5. **Seleccione características óptimas:** Las características restantes al final de este proceso se consideran las más relevantes para el modelo de predicción.

Sea $X = \{x_1, x_2, \ldots, x_n\}$ que represente el conjunto de características y Y la variable de destino. En cada iteración, nuestro objetivo es encontrar y eliminar características que menos contribuyen al rendimiento del modelo.

Considere un modelo de regresión lineal como ejemplo, donde el modelo predictivo se define como:

$$\hat{Y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon$$

donde β_0 es la intersección, β_i son los coeficientes de característica y ϵ es el término de error. La magnitud absoluta de β_i se puede utilizar para medir la importancia de la característica x_i.

Para modelos como SVM, la importancia de las características se puede derivar de los pesos en la función de decisión. Para los modelos basados en árboles (por ejemplo, bosques aleatorios), la importancia a menudo se calcula basándose en la reducción de la impureza o el índice de Gini a partir de divisiones que involucran cada característica:

$$\text{Importancia}(x_i) = \sum_{t \in T} I(t) \cdot p(t|x_i)$$

donde $I(t)$ es la impureza en el nodo t, $p(t|x_i)$ es la probabilidad de seleccionar x_i en esa división, y T es el conjunto de todos los nodos.

Los criterios de parada para RFE pueden definirse por varios factores:

1. **Número de funciones:** Deténgase cuando quede un número predefinido de funciones.

2. **Rendimiento del modelo:** Deténgase si eliminar características no mejora aún más la precisión del modelo.

3. **Puntuaciones de validación cruzada:** Utilice la validación cruzada para monitorear el rendimiento y seleccionar funciones con un error de validación mínimo.

Normalmente, las puntuaciones de validación cruzada se utilizan para seleccionar el subconjunto óptimo mediante la evaluación de la precisión predictiva en diferentes iteraciones, lo que garantiza que el conjunto de características elegido se generalice bien a datos invisibles.

RFE se utiliza ampliamente en aplicaciones con datos de alta dimensión, como bioinformática, procesamiento del lenguaje natural y reconocimiento de imágenes. Sin embargo, RFE puede resultar costoso desde el punto de vista computacional para conjuntos de datos muy grandes, ya que requiere ajustes repetidos del modelo. La elección del estimador también es crucial, ya que RFE se basa en el método de clasificación de características del modelo, lo que lo hace sensible a diferentes tipos de modelos y potencialmente introduce sesgos si el modelo no se adapta bien a los datos.

2.5.5. Criterio de información de Akaike

El criterio de información de Akaike (Akaike information criterion o AIC por sus siglas en inglés) es una herramienta valiosa que se utiliza para evaluar la calidad relativa de los modelos estadísticos. Además de su aplicación en modelado estadístico, el AIC ha demostrado ser muy eficaz en procesos de selección de características y aplicaciones de ML, arrojando resultados favorables.

En el contexto de la selección de características, el AIC se emplea para generar modelos considerando inicialmente todas las características disponibles en el conjunto de datos. Emplea una técnica de predicción de ajuste llamada regresión por pasos, que permite agregar o eliminar selectivamente características del conjunto completo de características (Cavanaugh and Neath, 2019).

El uso del AIC sobre el AICc se debe a que ha demostrado ser una mejor herramienta en modelos con más observaciones que características. Se representa de la siguiente manera:

$$AIC = 2k - 2\ln(\hat{L}). \tag{2.43}$$

Los criterios de información son herramientas esenciales en la selección de modelos, ya que proporcionan un medio para evaluar y comparar modelos en función de su bondad de ajuste y complejidad. Existe una variedad de herramientas de criterios de información, incluido el Criterio de información bayesiano (BIC) y el Criterio de información de Schwarz, los cuales incorporan una penalización más fuerte por la complejidad del modelo en comparación con el AIC, pero no pueden gestionar colecciones complejas de modelos en problemas de selección de variables de alta dimensión. El BIC se deriva de un marco bayesiano y tiende a favorecer modelos más simples en comparación con el AIC. Otra alternativa es el Criterio de información de desviación (DIC), que se utiliza en la comparación de

modelos bayesianos, especialmente para modelos jerárquicos, este criterio equilibra la bondad del ajuste con la complejidad del modelo, similar a AIC y BIC. Otro, llamado Criterio de Información Corregida (CIC), tiene como objetivo mejorar las limitaciones de otros criterios, particularmente en términos de corrección de sesgos. El CIC proporciona un ajuste más matizado para tamaños de muestra pequeños en comparación con el AIC y el AICc. Otra comparación realizada en este estudio es con el Criterio de Información Enfocada (FIC), que se utiliza para seleccionar modelos en función de su desempeño con respecto a un parámetro de enfoque específico o funcional del modelo. El FIC evalúa modelos basándose en una combinación de sesgo y varianza respecto del parámetro de interés, proporcionando un criterio adaptado a objetivos específicos. Cada criterio tiene sus propios puntos fuertes y es adecuado para diferentes escenarios. El AIC y el AICc se utilizan generalmente para enfoques frecuentistas, mientras que el BIC y el DIC son más comunes en contextos bayesianos. El CIC ofrece ajustes mejorados para muestras pequeñas, mientras que el FIC proporciona un enfoque de selección de modelo específico (Vrieze, 2012; Cavanaugh, 1997; Claeskens and Hjort, 2003; Emiliano *et al.*, 2014).

2.6. Métricas de validación de modelos

Son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, haciendo uso únicamente de los datos de entrenamiento (Rodrigo, 2020).

2.6.1. Área bajo la curva

En clasificación binaria, el área bajo la curva (AUC por las siglas de *area under the curve* en inglés) una métrica de evaluación que es el valor del AUC que traza la tasa de verdaderos positivos (en el eje y) en relación con la tasa de falsos positivos (en el eje x), va de 0.5 (el peor) a 1 (el mejor) (Microsoft, 2021).

2.6.2. Sensibilidad

La sensibilidad (*Sensitivity* en inglés), también conocida como tasa de verdaderos positivos (*True Positive Rate* en inglés o TPR), es una métrica que se utiliza para evaluar el rendimiento de un modelo de clasificación binaria, particularmente en el diagnóstico médico, donde identificar correctamente los casos positivos es crucial.

La *Sensitivity* mide la proporción de casos positivos reales que el modelo identifica correctamente como positivos. En otras palabras, cuantifica la capacidad

del modelo para detectar casos positivos verdaderos entre todos los casos positivos reales en el conjunto de datos.

Matemáticamente, la TPR se calcula mediante la siguiente fórmula:

$$TPR = \frac{TP}{TP + FN} \tag{2.44}$$

TP Verdaderos positivos (*True Positives* en inglés) se refiere al número de instancias que el modelo predice correctamente como positivas.

FN Falsos negativos (*False Negatives* en inglés) se refiere al número de instancias que el modelo predice incorrectamente como negativas pero que en realidad son positivas.

En el diagnóstico médico, la *Sensitivity* es crucial ya que determina la capacidad del modelo para identificar correctamente a los individuos con una condición o enfermedad particular. Un valor de *Sensitivity* alto indica que el modelo tiene una tasa baja de casos positivos perdidos, lo cual es deseable, especialmente cuando los falsos negativos pueden tener consecuencias graves, como no diagnosticar una enfermedad o afección (Morgan-Benita *et al.*, 2022b)

2.6.3. Especificidad

La especificidad (*Specificity* en inglés), también conocida como tasa de verdaderos negativos (*True Negative Rate* en inglés o TNR), es una métrica que se utiliza para evaluar el rendimiento de un modelo de clasificación binaria, particularmente en aplicaciones de seguridad y diagnóstico médico.

La *Specificity* mide la proporción de casos negativos reales que el modelo identifica correctamente como negativos. En otras palabras, cuantifica la capacidad del modelo para evitar clasificar erróneamente los casos negativos como positivos.

Matemáticamente, TNR se calcula mediante la siguiente fórmula:

$$TNR = \frac{TN}{FP + TN} \tag{2.45}$$

TN Verdaderos negativos (*True Negatives* en inglés) se refiere al número de instancias que el modelo predice correctamente como negativas.

FP Falsos Positivos (*False Positives* en inglés) se refiere al número de instancias que el modelo predice incorrectamente como positivas pero que en realidad son negativas.

La *Specificity* complementa la *Sensitivity* al evaluar el rendimiento general de un modelo de clasificación binaria. Mientras que la *Sensitivity* se centra en identificar correctamente los casos positivos, la *Specificity* se centra en identificar correctamente los casos negativos.

En el diagnóstico médico, la *Specificity* es crucial ya que determina la capacidad del modelo para descartar correctamente a individuos sin una condición o enfermedad particular. Un valor alto de *Specificity* indica que el modelo tiene una tasa baja de falsos positivos, lo cual es deseable, más aún, cuando los falsos positivos pueden conducir a tratamientos o intervenciones innecesarias, empeorando la condición de la persona (Morgan-Benita *et al.*, 2022b).

2.6.4. Precisión

La precisión (*Precision* en inglés), también conocida como valor predictivo positivo (Positive Predictive Value en inglés o PPV), es una métrica crucial que se utiliza para evaluar el rendimiento de un modelo de clasificación binaria, particularmente en situaciones donde el costo de los falsos positivos es alto.

La precisión mide la proporción de predicciones positivas realizadas por el modelo que realmente son correctas. En otras palabras, cuantifica la capacidad del modelo para evitar predecir falsamente resultados positivos.

Matemáticamente, PPV se calcula mediante la siguiente fórmula:

$$PPV = \frac{TP}{TP + FP} \tag{2.46}$$

La precisión proporciona información sobre la confiabilidad de las predicciones positivas realizadas por el modelo. Un valor de precisión alto indica que el modelo tiene una tasa baja de falsos positivos, lo cual es deseable, especialmente cuando los falsos positivos pueden tener consecuencias importantes, como tratamientos o intervenciones innecesarias.

La precisión es crucial a la hora de interpretar los resultados de una prueba diagnóstica. Un valor alto de precisión implica que la prueba puede identificar con precisión a personas con una condición o enfermedad particular, minimizando la probabilidad de falsas alarmas. Sin embargo, la precisión debe interpretarse junto con la *Sensitivity* (tasa de verdaderos positivos), ya que ambas métricas proporcionan información complementaria sobre el rendimiento del modelo.

2.6.5. Valor predictivo negativo

El valor predictivo negativo (*Negative Predictive Value* en inglés o NPV), es una métrica que se utiliza para evaluar el rendimiento de un modelo de clasificación binaria, particularmente en escenarios donde identificar correctamente los casos negativos es crucial.

El NPV mide la proporción de predicciones negativas realizadas por el modelo que realmente son correctas. En otras palabras, cuantifica la capacidad del modelo para identificar correctamente resultados negativos.

Matemáticamente, el NPV se calcula mediante la siguiente fórmula:

$$NPV = \frac{TN}{TN + FN} \tag{2.47}$$

FN (*False Negatives* en inglés) se refiere al número de instancias que el modelo predice incorrectamente como negativas pero que en realidad son positivas.

El NPV proporciona información sobre la confiabilidad de las predicciones negativas realizadas por el modelo. Un valor alto de NPV indica que el modelo tiene una tasa baja de falsos negativos, lo cual es deseable, especialmente cuando los falsos negativos pueden tener consecuencias importantes, como no detectar una enfermedad o afección.

El NPV es crucial a la hora de interpretar los resultados de una prueba diagnóstica para descartar una enfermedad. Un valor NPV alto implica que la prueba puede identificar con precisión a las personas sin la enfermedad, minimizando la probabilidad de una falsa tranquilidad.

De manera similar a PPV, el NPV debe interpretarse junto con la *Sensitivity* (tasa de verdaderos positivos), ya que ambas métricas proporcionan información complementaria sobre el rendimiento del modelo. Mientras que el NPV se centra en la precisión de las predicciones negativas, la *Sensitivity* se centra en la capacidad del modelo para detectar casos positivos verdaderos.

2.6.6. Tasa de falsos positivos

La tasa de falsos positivos (*False Positive Rate* en inglés o FPR), es una métrica que se utiliza para evaluar el rendimiento de un modelo de clasificación binaria, particularmente en escenarios donde identificar correctamente los casos negativos es crucial.

FPR mide la proporción de casos negativos reales que el modelo identifica incorrectamente como positivos. En otras palabras, cuantifica la tendencia del modelo a clasificar falsamente instancias negativas como positivas.

Matemáticamente, FPR se calcula utilizando la siguiente fórmula:

$$FPR = \frac{FP}{FP + TN} \tag{2.48}$$

FP Falsos positivos (*False Positives* en inglés) se refiere al número de instancias que el modelo predice incorrectamente como positivas pero que en realidad son negativas.

TN Verdaderos Negativos (*True Negatives* en inglés) se refiere al número de instancias que el modelo predice correctamente como negativas.

FPR proporciona información sobre la *Specificity* del modelo, es decir, su capacidad para evitar predicciones falsas positivas. Un valor de FPR bajo indica

que el modelo tiene una tasa baja de falsos positivos, lo cual es deseable, especialmente cuando los falsos positivos pueden tener consecuencias importantes, como tratamientos o intervenciones innecesarias.

Un valor de FPR bajo implica que la prueba de diagnóstico tiene una baja probabilidad de indicar falsamente la presencia de una enfermedad en individuos que en realidad están sanos. Esto ayuda a reducir la cantidad de falsas alarmas y garantizar que únicamente las personas con la enfermedad sean señaladas para una evaluación o tratamiento adicional.

FPR se utiliza a menudo junto con la *Sensitivity* (tasa de verdaderos positivos) para evaluar el rendimiento general de un modelo de clasificación binaria. Mientras que FPR se centra en los falsos positivos, la *Sensitivity* se centra en los verdaderos positivos, proporcionando información complementaria sobre la capacidad del modelo para clasificar correctamente los casos positivos.

2.6.7. Tasa de descubrimiento falso

La tasa de falso descubrimiento (*False Discovery Rate* en inglés o FDR), es una métrica utilizada para evaluar el rendimiento de un modelo de clasificación binaria, particularmente en escenarios donde controlar la tasa de predicciones falsas positivas es crucial.

FDR mide la proporción de predicciones positivas realizadas por el modelo que en realidad son incorrectas. En otras palabras, cuantifica la velocidad a la que las predicciones positivas resultan falsas.

Matemáticamente, FDR se calcula utilizando la siguiente fórmula:

$$FDR = \frac{FP}{FP + TP} \tag{2.49}$$

FP Falsos positivos (*False Positives* en inglés) se refiere al número de instancias que el modelo predice incorrectamente como positivas pero que en realidad son negativas. TP Verdaderos positivos (*True Positives* en inglés) se refiere al número de instancias que el modelo predice correctamente como positivas.

FDR proporciona información sobre la confiabilidad de las predicciones positivas realizadas por el modelo. Un valor FDR bajo indica que el modelo tiene una tasa baja de predicciones falsas positivas, lo cual es deseable, especialmente cuando los falsos positivos pueden tener consecuencias significativas, como tratamientos o intervenciones innecesarias.

Un valor FDR bajo implica que la prueba de diagnóstico tiene una baja probabilidad de indicar falsamente la presencia de una enfermedad en individuos que en realidad están sanos. Esto ayuda a reducir la cantidad de falsas alarmas y garantizar que solo las personas con la enfermedad sean señaladas para una

evaluación o tratamiento adicional. FDR se utiliza a menudo junto con *Precision* (valor predictivo positivo) para evaluar el rendimiento general de un modelo de clasificación binaria. Mientras que *Precision* se centra en la precisión de las predicciones positivas, FDR se centra en la tasa de predicciones falsas positivas, proporcionando información complementaria sobre el rendimiento del modelo.

2.6.8. Tasa de falsos negativos

La tasa de falsos negativos (*False Negative Rate* en inglés o FNR), es una métrica utilizada para evaluar el rendimiento de un modelo de clasificación binaria, particularmente en escenarios donde identificar correctamente los casos positivos es crucial.

FNR mide la proporción de casos positivos reales que el modelo identifica incorrectamente como negativos. En otras palabras, cuantifica la tendencia del modelo a pasar por alto casos positivos.

Matemáticamente, FNR se calcula mediante la siguiente fórmula:

$$FNR = \frac{FN}{FN + TP} \tag{2.50}$$

FN (*False Negatives* en inglés) se refiere al número de instancias que el modelo predice incorrectamente como negativas pero que en realidad son positivas. TP (*True Positives* en inglés) se refiere al número de instancias que el modelo predice correctamente como positivas.

FNR proporciona información sobre la *Sensitivity* del modelo, es decir, su capacidad para identificar correctamente los casos positivos. Un valor FNR bajo indica que el modelo tiene una tasa baja de falsos negativos, lo cual es deseable, especialmente cuando los falsos negativos pueden tener consecuencias importantes, como no detectar una enfermedad o afección.

Un valor FNR bajo implica que la prueba de diagnóstico tiene una baja probabilidad de pasar por alto individuos con una enfermedad o condición particular. Esto ayuda a garantizar que el modelo identifique correctamente a las personas que requieren una evaluación o tratamiento adicional. FNR se utiliza a menudo junto con la *Sensitivity* (tasa de verdaderos positivos) para evaluar el rendimiento general de un modelo de clasificación binaria. Mientras que *Sensitivity* se centra en identificar correctamente los casos positivos, FNR se centra en la tasa de predicciones falsas negativas, proporcionando información complementaria sobre el rendimiento del modelo.

2.6.9. Exactitud

La exactitud (*Accuracy* en inglés o ACC), es una utilizada para evaluar el rendimiento de un modelo de clasificación binaria. Mide la exactitud general de las predicciones del modelo considerando tanto casos verdaderos positivos como verdaderos negativos.

La ACC representa la proporción de predicciones correctas realizadas por el modelo entre todas las instancias del conjunto de datos.

Matemáticamente, ACC se calcula mediante la siguiente fórmula:

$$ACC = \frac{TP + TN}{P + N} \tag{2.51}$$

TP Verdaderos positivos (*True Positives* en inglés) se refiere al número de instancias que el modelo predice correctamente como positivas. TN Verdaderos negativos (*True Negatives* en inglés) se refiere al número de instancias que el modelo predice correctamente como negativas. P se refiere al número total de casos positivos en el conjunto de datos. N se refiere al número total de casos negativos en el conjunto de datos.

La ACC proporciona una evaluación integral del rendimiento del modelo en clases tanto positivas como negativas. Un valor de ACC alto indica que el modelo tiene una alta proporción de predicciones correctas, lo que sugiere que funciona bien en todo el conjunto de datos. La ACC refleja la exactitud general de la prueba de diagnóstico al identificar correctamente tanto a los individuos con una enfermedad (verdaderos positivos) como a los individuos sin la enfermedad (verdaderos negativos). Si bien la ACC es una métrica comúnmente utilizada para evaluar modelos de clasificación binaria, puede no ser adecuada para conjuntos de datos desequilibrados donde el número de instancias en cada clase es significativamente diferente. En tales casos, la ACC puede ser engañosa, ya que una alta proporción de predicciones puede deberse a la clase mayoritaria.

2.6.10. Puntuación F1

La Puntuación F1 (*F1-Score* en inglés), es una métrica comúnmente utilizada para evaluar el rendimiento de un modelo de clasificación binaria. Combina *Precision* y recuperación (*Sensitivity*) en una sola métrica, proporcionando una evaluación equilibrada del rendimiento del modelo.

La *F1-Score* es particularmente útil en situaciones en las que existe un desequilibrio entre las clases o cuando tanto los falsos positivos como los falsos negativos son igualmente importantes.

Matemáticamente, *F1-Score* se calcula mediante la siguiente fórmula:

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.52}$$

TP Verdaderos positivos (*True Positives* en inglés) se refiere al número de instancias que el modelo predice correctamente como positivas. FP Falsos positivos (*False Positives* en inglés) se refiere al número de instancias que el modelo predice incorrectamente como positivas pero que en realidad son negativas. FN falsos negativos (*False Negatives* en inglés) se refiere al número de instancias que el modelo predice incorrectamente como negativas pero que en realidad son positivas.

Una *F1-Score* alta indica que el modelo tiene una *Precision* y una *Sensitivity* altas, lo que sugiere que funciona bien en términos de minimizar los falsos positivos y los falsos negativos. La *F1-Score* proporciona una evaluación integral del desempeño de la prueba de diagnóstico para identificar correctamente a las personas con una enfermedad y minimizar la cantidad de casos perdidos.

La *F1-Score* es más útil cuando el coste de los falsos positivos y los falsos negativos es similar. En situaciones en las que un tipo de error es más costoso que el otro, métricas alternativas como *Precision* o *Sensitivity* pueden ser más apropiadas.

2.7. Diabetes y Machine Learning en este estudio

Hay innumerables aplicaciones sólidas de ML que han dado buenos resultados en la detección de la DMT2, pero estas están principalmente delimitadas por el entorno, los conjuntos de datos utilizados, la elección de técnicas para el tratamiento de los datos, la elección de las técnicas óptimas para una selección óptima de características, la implementación de diferentes modelos de ML, ya sea haciendo comparaciones entre ellos o integrándose para que funcionen juntos, y finalmente, distintas métricas de evaluación para validar estos conjuntos de características dentro de los modelos de ML implementados. Debido a la variedad de datos con los que trabajamos y los diferentes objetivos entre cada experimento, los resultados no siempre son apropiados o aplicables en otros contextos, dificultando hacer comparaciones entre modelos o técnicas y la discusión entre ellos (Collins *et al.*, 2011).

El contexto trabajado en esta tesis se centra en la investigación y análisis de las mejores técnicas y modelos aplicados a un segmento de la población mexicana/latinoamericana para el desarrollo de biomarcadores para la detección de la DMT2, en este contexto existen menos trabajos relacionados a comparación de experimentos e investigación en mayores sectores de población de diabéticos per cápita, destacando los estudios realizados en los continentes europeo y asiático.

En latinoamérica, la accesibilidad a los conjuntos de datos médicos es restringida y limitada, esto debido a la política interna y al poco uso de información médica clasificada como confidencial con fines de investigación, lo que genera a su vez, pocas bases de datos que cumplan con los protocolos establecidos por el gobierno, instituciones privadas u organizaciones internacionales, funcionando solo para objetivos específicos, sesgando otros estudios relacionados y limitando la posibilidad de mejorar la instrumentación o los tratamientos. Los conjuntos de datos que incluyen población hispanoamericana/latinoamericana, lo hacen en porcentajes por debajo de 5 %, siendo en la mayoría de los casos la población más baja a la cual se someten a este tipo de análisis o experimentación (Misra *et al.*, 2023).

Los casos de estudio realizados en esta tesis se integraron con dos conjuntos de datos proporcionados, uno proporcionado por la Unidad de Investigación Médica en Bioquímica, Centro Médico Nacional Siglo XXI ubicado en la Ciudad de México y otro por la Unidad de Investigación Biomédica ubicada en Zacatecas, México. Todos los pacientes mexicanos incluidos en el primer conjunto de datos firmaron una carta de consentimiento informado y el protocolo cumple con los criterios de Helsinki (hel, 2017), que fueron aprobados por el Comité de Ética del Instituto Mexicano del Seguro Social bajo el número R-2011-785-018. En el segundo conjunto de datos, todos los pacientes mexicanos firmaron una carta de consentimiento informado y los datos incluidos en el conjunto de datos del IMSS cumplen con la opinión de aprobación R-2017-785-131 de acuerdo con el protocolo “Análisis de metabolómica y transcriptómica diferencial en orina y suero de pacientes con prediabetes, diabetes y ND para identificar posibles biomarcadores pronósticos de daño renal”, que cumple con los criterios aprobados por el Comité Nacional de Investigación Científica y Ética y sigue los estándares éticos internacionales de la convención de Helsinki para estudios de investigación en seres humanos. Estos conjuntos de datos tienen información sobre metabolómica, antropométrica, clínica y pruebas de laboratorio. Estas evaluaciones se pueden combinar para medir la progresión desde un sujeto aparentemente sano o sin diabetes (control) hasta la ND.

La gran mayoría de los estudios presentados como trabajos relacionados, han desarrollado modelos con una precisión basada principalmente en los niveles de Glucosa (Obtenidos mediante prueba de FPG o OGTT o el HbA1c y seleccionando características como complemento haciendo una comparación entre varios otros que usan apilamiento, conjunto o reflejan los resultados de cada modelo individualmente, se puede afirmar que: La dependencia de los niveles de glucosa o Hb1Ac para diagnosticar DMT2 es una práctica común pero no es concluyente y fortalecer los resultados de los modelos validados individualmente con otras características es obligatorio, para ello, el conjunto de modelos integra los re-

sultados de cada modelo utilizado en un único conjunto, identificando el mejor escenario posible con una selección de características definidas por una técnica soportada, y con la omisión de los niveles de glucosa y HbA1c como parte de los modelos implementados. , existe la posibilidad de identificar otras características que puedan ser consideradas individualmente o agrupadas, que tengan el potencial de convertirse en biomarcadores de DMT2 para que puedan ser probados, replicados y utilizados en el diagnóstico por el área médica.

La propuesta en los tres casos de estudio presentes en esta tesis es implementar modelos de ML en ensamble para clasificar pacientes con y sin DMT2, basado en en datos clínicos. Se podría utilizar estos datos como un enfoque no invasivo, práctico, extremadamente rápido y preciso similar a la FPG, OGTT o HbA1c. Este modelo busca tener exclusivamente características no relacionadas con la glucosa. Con el uso de modelos multivariados, es decir, el uso de múltiples características (características de glucosa no invasivas y no relacionadas en este trabajo) aplicadas en un modelo para que se puedan encontrar correlaciones o patrones, se utiliza para la detección de DMT2. El modelo de conjunto de ML propuesto se centra en clasificar a los 1787 pacientes con una entrada de los 48 disponibles en la base de datos procesada. El conjunto de datos para este estudio se adquirió de la "Unidad de Investigación Médica en Bioquímica, Centro Médico Nacional Siglo XXI, IMSS", con la información de pacientes mexicanos. La base de datos contiene datos antropométricos, tratamiento médico, complicaciones, perfil lipídico y presión arterial. Cabe mencionar que no existen conjuntos de datos públicos con este tipo de pacientes.

2.7.1. Rendimiento esperado en algoritmos de ML

Se espera que un algoritmo de ML tenga el siguiente rendimiento esperado (Klaine *et al.*, 2017):

- **Escalabilidad**: este parámetro se puede definir como la capacidad de un algoritmo para poder manejar un aumento en su escala, como alimentar más datos al sistema, agregar más características a los datos de entrada o agregar más capas en una red neuronal, sin aumentando ilimitadamente su complejidad.

- **Tiempo de entrenamiento**: esta es la cantidad de tiempo que tarda un algoritmo de ML en estar completamente entrenado y formar la capacidad de hacer sus predicciones.

- **Tiempo de respuesta**: este parámetro está relacionado con la agilidad de un sistema de ML y representa el tiempo que tarda un algoritmo, después

de haber sido entrenado, para hacer una predicción para la función deseada de redes autoorganizadas.

- **Datos de entrenamiento**: esta métrica del algoritmo de ML es la cantidad y el tipo de datos de entrenamiento que necesita un algoritmo. Los algoritmos respaldados por más datos de entrenamiento suelen tener una mayor precisión, pero también requieren más tiempo para entrenarse.
- **Complejidad**: la complejidad de un sistema se puede definir como la cantidad de operaciones matemáticas que realiza para lograr la solución deseada. Este parámetro puede determinar si ciertos algoritmos son más adecuados para ser implementados en el lado del usuario.
- **Precisión**:Se espera que las redes del futuro sean mucho más inteligentes y rápidas, permitiendo tipos de aplicaciones y requisitos de usuario muy diferentes. La implementación de algoritmos que tengan una alta precisión es necesaria para garantizar una buena operatividad de ciertas funciones de red autoorganizadas.
- **Tiempo de convergencia**: esta métrica de un algoritmo, diferente del tiempo de respuesta, se relaciona con la rapidez con la que acepta que la solución encontrada para ese problema en particular es la solución óptima en ese momento.
- **Fiabilidad de la convergencia**: este parámetro representa la susceptibilidad de algún algoritmo a quedarse atascado en los mínimos locales y cómo las condiciones iniciales pueden afectar su rendimiento.

El descubrimiento de biomarcadores se ha basado en ensayos bioquímicos, técnicas de imagen y análisis de datos clínicos (Xia *et al.*, 2012). Pero, con el advenimiento del ML, ha surgido un creciente interés en aprovechar métodos computacionales para identificar y validar nuevos biomarcadores para la DMT2. Estos métodos permiten analizar grandes conjuntos de datos de manera eficiente y rápida, lo que acelera el desarrollo de herramientas de apoyo diagnóstico más efectivas y precisas para la DMT2, especialmente cuando se aplican en poblaciones de gran escala (Huynh-Thu *et al.*, 2012).

Mediante técnicas de selección de características, la reducción de dimensionalidad y la optimización de modelos, los algoritmos de ML pueden priorizar biomarcadores informativos y al mismo tiempo minimizar el ruido y el sobreajuste. En el contexto de la DMT2, los modelos de ML pueden facilitar la detección temprana, la estratificación del riesgo y las recomendaciones de tratamiento personalizadas basadas en perfiles de pacientes individuales (Leclercq *et al.*, 2019). Al analizar datos longitudinales de pacientes, incluidos registros médicos electrónicos,

imágenes médicas y datos de sensores portátiles, los algoritmos de ML pueden proporcionar información oportuna sobre la progresión de la enfermedad y la respuesta a la terapia.

La aplicación de modelos de ML para el descubrimiento de biomarcadores para la DMT2 va más allá del diagnóstico y el tratamiento, abarcando estrategias de tratamiento personalizadas y medidas preventivas. Al aprovechar datos específicos de los pacientes, incluida la predisposición genética, los perfiles metabólicos y los factores del estilo de vida, los algoritmos de ML pueden adaptar las intervenciones para abordar los factores de riesgo individuales y optimizar los resultados de salud. Un área de investigación destacada es la identificación de biomarcadores predictivos capaces de anticipar la aparición de la DMT2 antes de que los síntomas clínicos se manifiesten. La detección temprana de la prediabetes o de personas con alto riesgo de desarrollar DMT2 permite intervenciones específicas, como modificaciones en el estilo de vida, intervenciones dietéticas y terapias farmacológicas, dirigidas a prevenir o retrasar la progresión de la enfermedad. La DMT2 es prácticamente asintomática en sus etapas más tempranas, resulta difícil que las personas busquen atención médica o realicen cambios en su estilo de vida, incluso cuando tienen un riesgo elevado de desarrollar la enfermedad (Nedyalkova *et al.*, 2020).

Los modelos de ML proporcionan un marco integral que contiene diversas modalidades para el análisis de datos, incluyendo identificación de marcadores genéticos, parámetros bioquímicos, estudios de imágenes y variables clínicas, con el fin de desarrollar predicciones sólidas para la estratificación del riesgo de DMT2. Al analizar conjuntos de datos a gran escala de estudios en cohortes longitudinales y registros médicos electrónicos con los que se cuenta, los algoritmos de ML pueden identificar patrones e interacciones sutiles que contribuyen a la susceptibilidad a la DMT2. En los últimos años, los avances en ML, especialmente en el aprendizaje profundo, han impulsado avances significativos en diversas áreas, incluyendo el procesamiento del lenguaje natural, la visión por computadora y la atención médica. Los modelos de aprendizaje profundo, como las redes neuronales convolucionales (CNN por sus siglas en inglés) y recurrentes (RNN por sus siglas en inglés), han demostrado un rendimiento excepcional en tareas como el reconocimiento de imágenes, el procesamiento de voz y el diagnóstico médico. Sin embargo, debido a la complejidad en el desarrollo de bases de datos que incluyan video, imágenes o voz, y la necesidad existente de ajustarse a la nuevas técnicas y modelos que van surgiendo producto de la investigación continúa en esta área, se busca que la obtención de datos antropométricos y clínicos sea con presupuestos más reducidos y tiempos más cortos para la recolección (Vamathevan *et al.*, 2019).

La integración de modelos de ML con tecnologías de salud digitales, como

aplicaciones de salud móviles, sensores portátiles y dispositivos de monitoreo remoto, permite el seguimiento en tiempo real de parámetros fisiológicos y patrones de comportamiento. Estos flujos de datos proporcionan información valiosa sobre las trayectorias de salud individuales, lo que facilita la intervención temprana y la evaluación personalizada del riesgo de DMT2 y sus complicaciones. Los modelos de ML pueden ayudar en la selección y optimización de regímenes de tratamiento para personas con DMT2. Al analizar datos de ensayos clínicos, registros médicos electrónicos y evidencia del mundo real, los algoritmos de ML pueden identificar subgrupos de pacientes que probablemente respondan favorablemente a terapias específicas y guiar las decisiones de tratamiento en consecuencia (Makroum *et al.*, 2022). El desarrollo y la identificación de biomarcadores para la DMT2 se fundamentan en diversas teorías biológicas y clínicas, con el propósito de comprender los mecanismos subyacentes de la enfermedad e identificar indicadores confiables para su detección, diagnóstico y tratamiento. En esta investigación, se exploran distintos enfoques para identificar biomarcadores más robustos que respalden un diagnóstico multifactorial.

A pesar de los avances en metabolómica y ML, se debe tener en cuenta que algunos procedimientos, como el panel de glicerolípidos, son costosos y complejos. Esto resalta la necesidad de desarrollar procedimientos menos invasivos y de bajo costo para la detección y prevención de la DMT2, lo que podría hacer que estos métodos sean más accesibles para una mayor parte de la población. Las diferencias entre hombres y mujeres en cuanto a la epidemiología, las causas y los mecanismos biológicos del síndrome metabólico son cada vez más reconocidas. Estas diferencias se reflejan en aspectos como las tasas de disglucemia, la distribución de la grasa corporal, el tamaño y la función de los adipocitos, así como en la regulación hormonal del peso corporal y la acumulación de grasa. Además, la disminución de los niveles de estrógeno en las mujeres puede tener un impacto en los factores de riesgo metabólicos. Estas diferencias metabólicas pueden tener implicaciones significativas en la relación entre el perfil metabólico, el sexo y la DMT2, así como en sus comorbilidades (Huang *et al.*, 2020). Por ejemplo, los pacientes con síndrome metabólico pueden presentar variaciones en los niveles de azúcar en sangre entre los sexos, lo que aumenta el riesgo de desarrollar DMT2. Asimismo, las diferencias en la función y el tamaño de los adipocitos, junto con los mecanismos hormonales relacionados, pueden influir en el peso corporal y el almacenamiento de grasa de manera diferente en hombres y mujeres. Estos hallazgos resaltan la importancia de considerar el sexo como un factor importante en la evaluación y el tratamiento del síndrome metabólico y sus complicaciones, como la DMT2.

Una de las líneas de esta investigación se centra en la resistencia a la insulina, una afección central de la DMT2 que se caracteriza por una alteración en

la acción de la insulina en tejidos clave como los músculos, el hígado y el tejido adiposo. Para el desarrollo de modelos, nos centramos en el uso del aprendizaje supervisado y no supervisado. Cada uno de los algoritmos y técnicas fueron diseñados para abordar tareas específicas y desafíos particulares dentro de cada experimento propuesto. Los algoritmos supervisados utilizados están alineados a la dicotomía que se buscó implementar que, aunque variados, gracias a las revisiones sistemáticas de literatura hechas por otros autores más la realizada por el autor de esta tesis para documentar y respaldar el estado del arte de los productos científicos realizados, se encontraron los modelos y técnicas más relevantes para los conjuntos de datos utilizados y los objetivos propuestos para la experimentación. Como aportación más importante en esta tesis se resalta que se hizo hincapié en las técnicas de selección de características implementadas por encima de los modelos de ML utilizados para validarlas. Estas técnicas y modelos fueron probados y respaldados en productos científicos publicados en revistas internacionales JCR dando aval a su importancia.

Aunque los conjuntos de datos utilizados no incluyen directamente adiponectina, leptina u otras características específicamente relacionadas con el tejido adiposo, sí contienen un perfil lipídico, complementado con el peso y el BMI, lo cual permite abordar esta falta de información. Este perfil lipídico incluye lipoproteínas de baja y alta densidad, colesterol y triglicéridos. Además, características como el peso y el BMI son esenciales, ya que las personas con cierto grado de obesidad tienden a tener un riesgo mayor de desarrollar DMT2, además de otras comorbilidades como el daño cardiaco o la ND.

Otra línea observada es la disfunción de las células β, esta disfunción progresiva y la pérdida de células β pancreáticas contribuyen a la patogénesis de la DMT2. Los biomarcadores que reflejan la función de las células β, como los niveles FPG y OGTT, la relación proinsulina/insulina y los niveles de péptido C, pueden proporcionar información sobre el estado de las células β y ayudar en la detección temprana y el seguimiento de la DMT2. Esta característica incluida en el conjunto de datos corresponde a la lectura de niveles de FPG, corroborada con la prueba de HbA1c para personas diagnosticadas con DMT2, nos proporcionó un fundamento sólido en los datos ligados a cada paciente que poseía este dato. La HbA1c solo se realiza cuando existe una posibilidad fehaciente de que ese paciente podría tener DMT2 no confirmada, por ello, este dato sólo se incluye en personas con niveles de FPG por encima de 125 mg/dl.

Gran parte de nuestra investigación se basa en la desregulación del metabolismo de la glucosa y los lípidos. Los biomarcadores relacionados con el metabolismo de la glucosa, como los resultados de la HbA1c, la FPG y la OGTT, se utilizan comúnmente para diagnosticar y monitorear la DMT2, pero además, los biomarcadores lipídicos como los niveles de triglicéridos, el colesterol LDL, el

colesterol HDL y colesterol no HDL se asocian con el riesgo de DMT2 y complicaciones cardiovasculares. Cabe aclarar que los factores genéticos y epigenéticos contribuyen a la susceptibilidad individual a la DMT2. Existen diversos estudios que corroboran la genética de la población latinoamericana como propensa al desarrollo de la enfermedad de la DMT2. Las modificaciones epigenéticas, incluida la metilación del ADN, las modificaciones de histonas y los ARN no codificantes, también influyen en la patogénesis de la DMT2 y pueden servir como biomarcadores potenciales para el diagnóstico y pronóstico de enfermedades, sin embargo son soluciones sumamente costosas y poco viables en estudios a gran escala. En los conjuntos de datos valorados y analizados en esta investigación, se tiene exclusivamente población de origen latinoamericano (Prandi *et al.*, 2022).

Por último se toma en cuenta la microbiota y salud intestinal, como la evidencia emergente que sugiere un vínculo entre la composición de la microbiota intestinal, la permeabilidad intestinal y la salud metabólica, incluida la DMT2. Los biomarcadores que reflejan la diversidad de la microbiota intestinal, los metabolitos microbianos y los marcadores de la integridad de la barrera intestinal, pueden proporcionar información valiosa sobre el eje intestino-páncreas y su relevancia para el desarrollo de DMT2 (Han and Lin, 2014).

Para la inclusión de modelos de ML y técnicas de selección de características adecuadas se analizan distintas teorías. Una de ellas es por medio de la comprensión de las vías y los mecanismos biológicos subyacentes implicados en la patogénesis de la DMT2, siendo esencial para identificar biomarcadores relevantes. Las teorías relacionadas con la resistencia a la insulina, la disfunción de las células β, la inflamación, el estrés oxidativo y la desregulación metabólica proporcionan un marco para seleccionar biomarcadores candidatos. Los algoritmos de ML pueden analizar datos ómicos de alta dimensión, como en nuestro caso, la metabolómica, para descubrir nuevos biomarcadores asociados con estos procesos biológicos. Como en el caso de los conjuntos de datos utilizados que nos proporcionan datos relevantes en estas vías (Reel *et al.*, 2021).

Las técnicas de ML, como los algoritmos de integración de datos multiómicos, pueden combinar información de diferentes capas ómicas para identificar firmas de biomarcadores sólidas que capturen la complejidad de la fisiopatología de la DMT2. Los métodos de selección de características diseñados para datos multiómicos, como las técnicas de reducción de dimensionalidad (como LASSO que es utilizado en esta tesis), ayudan a priorizar características relevantes y mejorar la interpretabilidad del modelo (Wu *et al.*, 2019). Estas técnicas son descritas a detalle más adelante en este capítulo.

En cuanto a la incorporación de datos clínicos y fenotípicos, en esta investigación es incluida información demográfica, historial médico, mediciones antropométricas y factores de estilo de vida, mejora el poder predictivo de los modelos

de biomarcadores para la evaluación y el pronóstico del riesgo de DMT2. Los algoritmos de selección de características compatibles con tipos de datos heterogéneos, como la eliminación recursiva de características con validación cruzada (RFECV), algoritmos genéticos o LASSO, pueden identificar características clínicas informativas y sus interacciones con biomarcadores moleculares. La interpretabilidad de los modelos de ML es crucial para comprender la relevancia biológica de los biomarcadores identificados y traducir las predicciones del modelo en conocimientos prácticos para la toma de decisiones clínicas.

La validación y reproducibilidad rigurosas de los modelos de biomarcadores son esenciales para garantizar su generalización y utilidad clínica. Las estrategias de validación cruzada, los cohortes de validación externa y los estudios de replicación independientes validan la solidez y confiabilidad de los biomarcadores identificados. Además, la presentación de informes transparentes sobre las metodologías de desarrollo de modelos, el cumplimiento de las directrices estándar y el intercambio de datos de acceso abierto facilitan la reproducibilidad y fomentan la colaboración en la investigación de biomarcadores para la DMT2.

2.8. Metodología de la investigación

Es esencial validar los aportes de esta investigación tanto en la comunidad médica como en la sociedad, asegurándose de que estén respaldados por fundamentos sólidos y fundamentados en el método científico. Para lograrlo, se empleó el marco de investigación establecido por Roberto Hernández Sampieri, específicamente su metodología de investigación cuantitativa. Esta metodología se divide en ocho etapas esenciales que guían el proceso de investigación sistemática y mejoran la confiabilidad de los hallazgos. Cada una de estas etapas se detalla detalladamente en la figura 2.3, sirviendo como hoja de ruta para el diseño de la investigación (Hernández-Sampieri and Mendoza, 2019).

El enfoque cuantitativo de Hernández Sampieri es particularmente útil en la investigación científica y médica, ya que se centra en la recopilación y análisis de datos numéricos para probar hipótesis, medir variables y determinar relaciones entre ellas. Este método sigue una secuencia estructurada, que comienza con la definición del problema, la revisión de la literatura y la formulación de hipótesis. A continuación viene el diseño del estudio, donde los investigadores deciden los métodos para recopilar datos cuantificables, seguido de la recopilación de datos a través de instrumentos como encuestas, experimentos o ensayos clínicos (Hernández-Sampieri and Mendoza, 2019).

1. El primer paso del modelo de Sampieri es definir y formular el problema de investigación. Esta etapa implica identificar claramente el problema, es-

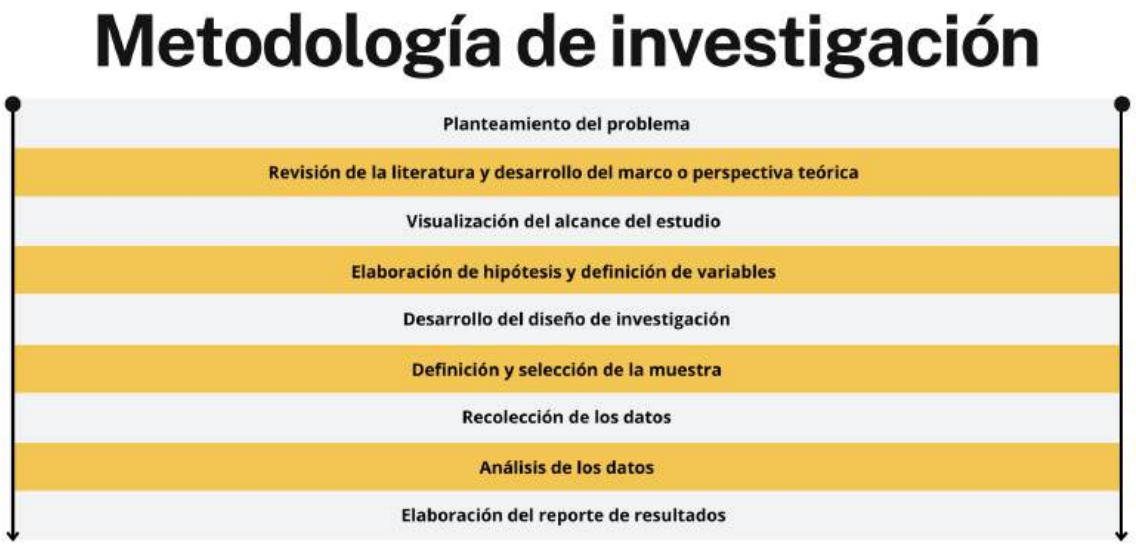


Figura 2.3: Fases del proceso de investigación cuantitativa. Imagen de elaboración propia

tablecer los objetivos de la investigación y refinar las hipótesis. Una comprensión profunda del problema garantiza que la investigación esté bien enfocada, lo que hace que los pasos posteriores sean más eficientes.

2. A continuación se realiza la revisión bibliográfica, que implica recopilar y analizar estudios previos y perspectivas teóricas relacionadas con el tema.

3. Gracias al contexto proporcionado por la revisión bibliográfica se identifican lagunas en el conocimiento existente, lo que garantiza que la investigación actual se base en una base sólida y aporte ideas novedosas, de esta manera se visualiza el alcance del estudio.

4. El cuarto paso implica la formulación de la hipótesis. En la investigación cuantitativa, este paso es donde los investigadores establecen una declaración clara y comprobable basada en el problema identificado y la revisión bibliográfica. La hipótesis establece el escenario para toda la investigación al predecir relaciones o resultados entre variables.

5. Después de la formulación de la hipótesis viene el desarrollo del diseño y la planificación de la investigación. Esta fase incluye la selección de un tipo de estudio, la determinación del tamaño de la muestra y la elección de las técnicas adecuadas de recopilación de datos. Los investigadores pueden decidir entre diseños transversales o longitudinales, según los objetivos de

la investigación, mientras que la planificación garantiza que la metodología se ajuste a la hipótesis y pueda responder eficazmente a las preguntas de investigación.

6. La etapa de definición y selección de la muestra radica en que datos se requieren recopilar para el contexto de esta investigación, en este caso se requieren datos de pacientes relacionados con la diabetes y otras comorbilidades, lo que proporcionaría la materia prima para el análisis posterior.

7. Una vez seleccionada y definida la muestra se procede a la recopilación de los datos. En esta fase se adquirieron los datasets directamente y no se realizaron tomas de muestra.

8. Una vez recopilados los datos, la siguiente etapa es el análisis de datos. Los investigadores utilizan herramientas y métodos estadísticos para procesar los datos, evaluar los resultados y probar las hipótesis. Al aplicar técnicas de selección de características y modelos ML, se puede confirmar o refutar la hipótesis, lo que genera información valiosa sobre el problema de investigación.

9. El noveno paso es la interpretación de los resultados y la elaboración del reporte de resultados, en la que los resultados del análisis de datos se traducen en conclusiones significativas. Esta es la etapa en la que los investigadores evalúan las implicaciones más amplias de sus hallazgos y determinan si coinciden con el conocimiento existente en el campo o lo desafían. La presentación y difusión de los hallazgos implica recopilar los resultados en informes, publicaciones o presentaciones. Otro elemento a tomar en consideración sería, brindar recomendaciones basadas en los resultados del estudio, asegurando que la investigación no solo contribuya al conocimiento académico, sino que también, que tenga aplicaciones prácticas en la práctica médica, la política de atención de la salud o la investigación científica en el futuro.

Una de las fortalezas de este enfoque es su énfasis en el análisis de datos, empleando típicamente herramientas estadísticas para validar hipótesis, medir la efectividad de las intervenciones y extraer conclusiones. Este proceso asegura que los resultados no solo sean sólidos sino también replicables. Las etapas finales del modelo de Hernández Sampieri incluyen la interpretación de los datos, la presentación de los hallazgos y la formulación de recomendaciones basadas en los resultados. Cada uno de estos pasos garantiza un alto nivel de precisión y generalización, lo que hace que las contribuciones de la investigación sean más confiables y aplicables en contextos más amplios.

Capítulo 3
Estado del arte

3.1. Diabetes

La DMT2 ha sido objeto de diversas aproximaciones y tratamientos a lo largo de los años. Recientemente, estudios como el de Hallberg *et al.* (2019), titulado "Reversing Type 2 Diabetes: A Narrative Review of the Evidence", han arrojado datos impactantes que contrastan con las creencias convencionales. Este estudio sugiere la posibilidad de revertir la enfermedad, lo que ofrece esperanza a los pacientes diabéticos al considerar la posibilidad de alcanzar un estado prediabético o incluso de normalidad. Este enfoque contrasta con la visión más tradicional de la DMT2 como una enfermedad crónica e incurable que sólo puede ser controlada con tratamientos farmacológicos de por vida.

Otro estudio proporciona evidencia interesante sobre las diferencias en las vías metabólicas entre hombres y mujeres con COVID-19 grave. Se observaron variaciones significativas en el metabolismo de los lípidos, la vía de las pentosas fosfato, el metabolismo de los ácidos biliares y el procesamiento de aminoácidos aromáticos, como el triptófano y la tirosina, entre otros. Estas diferencias metabólicas entre los sexos podrían contribuir a la comprensión de las disparidades observadas en la gravedad y el pronóstico de la enfermedad entre hombres y mujeres con COVID-19. Además, el estudio señala que los métodos estadísticos no supervisados revelaron un dimorfismo sexual significativo en las relaciones entre los parámetros clínicos específicos de los pacientes y sus perfiles metabólicos generales. Esto sugiere que las características clínicas y metabólicas pueden interactuar de manera diferente en hombres y mujeres con COVID-19, lo que resalta la importancia de considerar el sexo como un factor relevante en la evaluación y el tratamiento de la enfermedad. Estos hallazgos subrayan la necesidad de investigaciones adicionales para comprender mejor las diferencias metabólicas entre los sexos y su impacto en la fisiopatología y el tratamiento de COVID-19 (Rocio Diaz *et al.*, 2022).

3.2. Diabetes y machine learning

El uso de encuestas para recopilar datos clínicos de la población puede ser ineficiente, costoso y poco confiable en muchos casos. En cambio, los hospitales y centros de salud pueden obtener datos directamente de personas con diagnósticos conocidos o comorbilidades, utilizando técnicas adecuadas de obtención de datos y registrando los resultados de manera sistemática. Es importante comparar los perfiles clínicos de personas con diagnóstico concluyente de DMT2 con aquellos que no tienen un diagnóstico, pero presentan factores de riesgo conocidos, como sobrepeso, niveles elevados de lipoproteínas de baja densidad (LDL), colesterol alto o presión arterial sistólica (SBP por sus siglas en inglés) irregular. El uso de técnicas de ML en el ámbito médico ha demostrado ser una herramienta poderosa para resolver problemas de diagnóstico y predicción de diversas enfermedades, incluida la DMT2. Los modelos de ML pueden analizar datos clínicos y de salud para identificar patrones, relaciones y características relevantes que pueden ser utilizadas para predecir la presencia o riesgo de enfermedades (Kavakiotis *et al.*, 2017).

Thyde *et al.* (2021) propusieron un enfoque para detectar la adherencia a las inyecciones de insulina basal administradas una vez al día, utilizando técnicas de ML. Este estudio empleó datos simulados generados por el paciente virtual de Medtronic, con el objetivo de evaluar la viabilidad de la detección de la adherencia basada en datos de monitoreo continuo de glucosa (CGM). Se exploraron varios modelos, incluyendo regresión logística, perceptrones multicapa, redes neuronales convolucionales y técnicas de extracción automática y dependiente de expertos de características. La validación de los modelos se realizó mediante la curva ROC (Característica Operativa del Receptor).

Por otro lado, Fujihara *et al.* (2021) propusieron el desarrollo de modelos de ML para apoyar la toma de decisiones sobre el inicio del tratamiento con insulina en pacientes japoneses con DMT2. Este estudio analizó los datos de 4860 sujetos, de los cuales 293 recibieron tratamiento con insulina y 4567 no. Su objetivo fue evaluar la capacidad de los modelos de ML para predecir el inicio del tratamiento con insulina según lo determinado por especialistas, así como examinar si estos modelos podrían respaldar la toma de decisiones de los médicos generales en cuanto al inicio de la insulina en pacientes con DMT2. Se emplearon modelos de regresión logística y redes neuronales, y la validación se llevó a cabo mediante la curva ROC.

Otro estudio relevante fue llevado a cabo por Lele Yang*et al.*, quienes propusieron la integración de datos metabolómicos con un enfoque en el ML para descubrir marcadores cuantitativos (*Q-markers*) en la preparación de Jinqi Jiangtang contra la DMT2. Utilizando muestras químicas de seis lotes de Coptidis Rizoma,

cinco lotes de Astragali Radix y seis lotes de Lonicerae Japonicae Flos, su objetivo fue desarrollar un enfoque de IA que permitiera la identificación rápida y fácil de los *Q-markers* con bioactividad en la preparación Jinqi Jiangtang. Para este propósito, emplearon la ANN de retropropagación y validaron el modelo utilizando medidas de *Precision* y error cuadrático medio (Yang *et al.*, 2021).

Por otra parte, Sevil *et al.* (2021) realizaron un estudio que involucró pruebas físicas y de estrés psicológico agudo utilizando un dispositivo portátil para medir la concentración de glucosa. Se reclutaron 12 sujetos con diabetes tipo 1 para las pruebas físicas y 8 sujetos adicionales para experimentos que incluyeron pruebas físicas y estrés psicológico agudo. El objetivo fue predecir la concentración de glucosa utilizando este dispositivo portátil. Los modelos utilizados en este estudio incluyeron k-nearest neighbors, análisis discriminante lineal, Decision Trees, aprendizaje por conjunto, máquinas de vectores de soporte, regresión de proceso gaussiano y redes neuronales profundas con memoria a largo plazo y corto plazo (LSTM). La validación de los modelos se realizó mediante el valor p.

En otro estudio llevado a cabo por Carrillo-Larco *et al.* (2021), se propuso identificar grupos de personas con DMT2 y evaluar la frecuencia de estos grupos en países de América Latina y el Caribe. Para ello, se seleccionaron 13 conjuntos de datos de diferentes países de la región. El objetivo fue agrupar a personas con DMT2 dentro de la población general de América Latina y el Caribe. Los modelos utilizados incluyeron el Análisis de Componentes Principales (PCA) y el algoritmo de agrupamiento *K-Means*. La validación se llevó a cabo utilizando el coeficiente de Jaccard.

Wang (2021) propuso mejorar la eficiencia discriminativa de los modelos predictivos para la DMT2 mediante el uso de puntuaciones de riesgo genético (GRS) y el ML. Utilizando una muestra poblacional de 5,712 participantes, el objetivo fue desarrollar un modelo de predicción combinando GRS con factores de riesgo convencionales para DMT2. Se emplearon modelos como ANN, RF y máquinas de aumento de gradiente (GBM), validados mediante el AUC y el índice de reclasificación.

Nath (2021) investigó la capacidad de la grasa corporal para predecir la capacidad de ejercicio en personas con DMT2, utilizando un enfoque de ML. Se seleccionaron 1348 pacientes después de un proceso de preprocesamiento, y se exploraron características como el porcentaje de grasa subtotal, la edad, los niveles séricos de triglicéridos y la presión arterial. Se emplearon varios modelos de ML, como Bosques Aleatorios, Incremento de Gradiente, Máquinas de Vectores de Soporte, Redes Neuronales Multicapa y Regresor de Apilamiento, validados mediante el MSE.

En un enfoque innovador, Nedyalkova (2021) presentó un procedimiento original basado en el clustering de *K-Means* para identificar variables clínicas perti-

nentes capaces de separar eficazmente pacientes con DMT2 en grupos similares. El estudio se realizó con 52 pacientes, de los cuales 51 padecían hipertensión arterial, 48 tenían enfermedad coronaria isquémica, 51 presentaban polineuropatía diabética y 47 mostraban microangiopatía diabética. El objetivo fue proporcionar resultados hipotéticos basados en una compleja relación con el estado de salud del grupo y visualizar la capacidad predictiva del estado de salud en curso. Se empleó un modelo denominado *K-Means* Combinatorial, el cual se validó utilizando la probabilidad de discriminación.

Syed and Khan (2020) llevaron a cabo un estudio transversal basado en cuestionarios para investigar la prevalencia y la asociación entre los factores de riesgo convencionales de la diabetes. La muestra poblacional consistió en 4896 participantes, de los cuales 990 fueron casos de diabetes y 3906 no presentaban la enfermedad. El objetivo del estudio fue estimar la prevalencia de la diabetes y evaluar la asociación entre la exposición a los factores de riesgo y la presencia de la enfermedad. Se emplearon diversos modelos, incluyendo regresión logística, perceptrón promedio, Naïve Bayes, redes neuronales, máquinas de vectores de soporte, máquinas de soporte vectorial profundas locales, árbol de decisiones, *K-Means* y árbol de decisiones mejorado. La validación de estos modelos se realizó utilizando métricas como *Precision*, *F1-Score* y AUC.

You *et al.* (2019), propusieron aplicar una metodología de ML para evaluar el desempeño del programa DIABETIMSS en pacientes con DMT2 atendidos en clínicas de medicina familiar en México. Utilizaron datos de 78,894 pacientes con DMT2, de los cuales 37,767 recibieron atención a través del programa DIABETIMSS. El objetivo fue investigar la significancia del programa DIABETIMSS en el control glucémico de los pacientes. Se emplearon modelos de PCA y árboles de regresión, y se validaron utilizando valores de P-Value.

Fitriyani *et al.* (2019), presentaron un modelo de predicción de enfermedades llamado DPM (Disease Prediction Model por sus siglas en inglés), que utiliza un método de detección de valores atípicos basado en bosques de aislamiento (iForest), una técnica de sobremuestreo de minorías sintéticas (SMOTE Tomek) para equilibrar los datos y un enfoque de conjunto para predecir enfermedades. Utilizaron tres conjuntos de datos que incluían individuos con y sin DMT2 e hipertensión. El objetivo fue proporcionar una predicción temprana de la DMT2 y la hipertensión basada en los factores de riesgo de cada individuo. Utilizaron modelos como isolation forest, SMOTETomek y ensemble learning, y se validaron mediante métricas como *Precision*, *F1-Score*, ACC y AUC.

Existen contribuciones significativas en las mejoras al rendimiento de los modelos al agruparlos y mejorarlos (Cohen *et al.*, 2021; Du *et al.*, 2019), y en el tratamiento de sesgos (Barda *et al.*, 2020), lo que proporciona herramientas para generar resultados más sólidos. El uso de modelos de ML para apoyar el trata-

miento o seguimiento del estado físico de pacientes con DMT2 presenta importantes ventajas, como la reducción del seguimiento invasivo continuo de glucosa en personas insulinodependientes.

Hasan *et al.* (2020) proponen un conjunto ponderado para la predicción de la diabetes, utilizando diversos modelos de ML como k-vecinos más cercanos, Decision Trees, bosques aleatorios, AdaBoost, Naïve Bayes y XGBoost. Este enfoque incluye la estandarización de datos con Z-Score, la reducción de dimensionalidad mediante análisis de componentes principales, validación cruzada de 5 veces y perceptrón multicapa (MLP).

Por otro lado, Deberneh and Kim (2021) propusieron el uso de un conjunto por apilamiento y votación suave, que incluyó modelos de RF, SVM y XGBoost para predecir el estado de un paciente como no diabético, prediabético o diabético. Los modelos se construyeron utilizando 12 características similares a las analizadas en el experimento presentado, como FPG, HbA1c, triglicéridos, BMI, edad, ácido úrico y sexo. El conjunto logró una *Precision* del 78 %.

El marco propuesto por El-Sappagh *et al.* (2019) combina varios algoritmos de ML, como KNN, *Naïve Bayes*, árbol de decisión, SVM, árbol de decisión difuso, ANN y LR, para clasificar la diabetes. Este marco se evaluó utilizando datos reales recopilados de registros médicos electrónicos de hospitales universitarios en Mansura, Egipto. Los resultados obtenidos fueron una *Precision* del 90 %, una *Sensitivity* del 90.2 %, y una ACC del 94.9 %.

Por otro lado, Kumari *et al.* (2021) propusieron un clasificador de votación suave por conjunto que utiliza tres algoritmos de ML: RF, LR y Naïve Bayes, para realizar una clasificación binaria. La metodología propuesta fue evaluada empíricamente utilizando diferentes clasificadores base como *AdaBoost*, LR, SVM, RF, *Naïve Bayes*, *Bagging*, *GradientBoost*, *XGBoost* y *CatBoost*. Los criterios de evaluación utilizados fueron *Precision*, *Sensitivity*, ACC y *F1-Score*.

El enfoque de conjunto propuesto muestra resultados prometedores en la clasificación de diabetes en el conjunto de datos PIMA Indians, con una ACC del 79.04 %, una *Precision* del 73.48 %, una *Sensitivity* del 71.45 % y un *F1-Score* del 80.6 %. Estos resultados sugieren que el clasificador de votación suave por conjunto puede ser efectivo para la detección de la diabetes en este conjunto de datos específico. Además, la eficacia de esta metodología se evaluó en un conjunto de datos de cáncer de mama, donde logró una ACC del 97.02 %.

Singh and Singh (2020) desarrollaron un sistema de aprendizaje conjunto evolutivo basado en apilamiento, denominado “NSGA-II-Stacking”, con el objetivo de predecir la aparición de DMT2 en un período de cinco años. Utilizaron el conjunto de datos de diabetes PIMA Indians, que es de acceso público. En el proceso de preprocesamiento de datos, se identificaron y se imputaron los valores faltantes y los valores atípicos utilizando los valores medianos. Para la selección de mo-

delos base, se empleó un algoritmo de optimización multiobjetivo que maximiza simultáneamente la precisión de la clasificación y minimiza la complejidad del conjunto de modelos. Este enfoque podría ser útil para predecir la incidencia de DMT2 en poblaciones con características similares a las del conjunto de datos PID.

Un estudio proporcionado por Frimpong *et al.* (2021) presenta un modelo de red neuronal artificial feedforward (FFANN), diseñado para conjuntos de datos numéricos y textuales, para abordar las limitaciones de los modelos ANN existentes en el diagnóstico médico. Con una arquitectura optimizada, FFANN maximiza las capas y los nodos para un aprendizaje eficaz de las características del conjunto de datos y al mismo tiempo mitiga los problemas de sobreajuste y desajuste. El modelo propuesto muestra una precisión mejorada y proporciona una herramienta valiosa para la detección de enfermedades en pacientes con diabetes dentro del contexto más amplio de la medicina molecular.

3.3. Metabolómica y machine learning

La implementación de modelos de ML que aprovechan conjuntos de datos completos del metaboloma, combinados con biomarcadores conocidos como glucosa, manosa y $\alpha-$hidroxibutirato (comúnmente utilizados como factores de riesgo clínicos), ha demostrado ser prometedora en la predicción de la progresión a la DMT2. Algunos de los biomarcadores predictivos identificados incluyen α-tocoferol, bradicinina, hidroxiprolina, y otros metabolitos como X-12063 y X-13435, que han mostrado un potencial significativo para realizar predicciones precisas sobre la progresión de esta enfermedad (Peddinti *et al.*, 2017). Estos avances representan pasos importantes hacia una detección y tratamiento más eficaces de la DMT2 en sus etapas tempranas.

Con el fin de diseñar planes de prevención y no esquemas de solución, se presentan soluciones basadas en el uso de la metabolómica proporcionando lecturas de los estados de la DMT2 antes de que aparezcan los síntomas. La identificación de nuevos biomarcadores plasmáticos para la pérdida de masa celular β funcional en la etapa de prediabetes asintomática puede ser una solución a este problema, con metabolómica dirigida y no dirigida. Este estudio se realizó en ratones e identificó el 1.5-anhydroglucitol como asociado con la pérdida de masa celular β funcional y descubrió similitudes metabólicas entre el hígado y el plasma, lo que proporciona información sobre los efectos sistémicos causados por la disminución temprana de β-células, esta desoxihexosa refleja una disminución progresiva de la masa funcional de células β en la etapa prediabética asintomática. Estos hallazgos establecieron una base a aplicar en cohortes humanas para

poder validarlos (Li *et al.*, 2019). Una forma de identificar comorbilidades es a través de posibles biomarcadores de metabolitos o un cohorte de población mediante metabolómica dirigida y enfoques de ML. Un caso de biomarcadores potenciales en la predicción de la ND son la esfingomielina C18:1 y la fosfatidilcolina diacil C38:0 identificadas específicamente en individuos hiperglucémicos (Huang *et al.*, 2020).

En el estudio realizado por Zeng *et al.* (2022) se examinó el papel preciso de la microbiota intestinal en la patogénesis de la DMT2 y se identificó, entre otras cosas, que la colina plasmática (por desviación estandar del cambio transformado en logaritmo: razón de probabilidades 1.36 (intervalo de confianza del 95% 1.16, 1.58) fue positiva. La colina en plasma entonces, demuestra ser un clasificador potencial de la diabetes y proporciona controles que se distinguen con precisión, integrando datos sobre la colina y ciertas especies de microbiota, así como los factores de riesgo tradicionales. Esto proporciona una visión novedosa del metabolismo de la colina que relaciona el metabolismo alterado de la glucosa y la diabetes con la microbiota.

Otra forma de comprender la función metabólica en los órganos y el desarrollo y progresión de la DMT2 es comparar el cribado metabólico bidimensional en muestras de tejido de tejidos metabólicos clave, como el suero, el tejido adiposo visceral, el hígado, los islotes pancreáticos o el músculo esquelético, de individuos en diferentes estados de DMT2. De esta manera, las carnitinas son significativamente más altas en el hígado, mientras que las lisofosfatidilcolinas fueron significativamente más bajas en los músculos y el suero de los sujetos con diabetes. Otros hallazgos mostraron que las lisofosfatidilcolinas son significativamente más bajas en el músculo y el suero de sujetos con prediabetes y el ácido glicodesoxicólico fue significativamente mayor en el hígado (Diamanti *et al.*, 2019). Los biomarcadores ampliamente establecidos, como la FPG o la resistencia a la insulina (HOMA), en combinación con la metabolómica, pueden tener áreas bajo la curva más grandes (Savolainen *et al.*, 2017; Liu *et al.*, 2017), la inclusión de estos metabolitos como parte de los modelos ML brinda nuevas posibilidades, haciendo que la detección sea más robusta y precisa. Otro biomarcador como único candidato en la metabolómica urinaria no dirigida presenta el papel del metaboloma urinario 3-hidroxidecanoil-carnitina para la identificación de individuos con riesgo de DMT2 (Salihovic *et al.*, 2020).

Otros enfoques apoyan las relaciones entre la diabetes T2DM y los marcadores de metabolitos, como la glutamina, la glicina y los aminoácidos aromáticos. Metabolitos como la glucosa, la fructosa, los aminoácidos y los lípidos suelen estar alterados en pacientes con DMT2 (Arneth *et al.*, 2019). La metabolómica y la proteómica proporcionan biomarcadores circulantes como herramientas eficaces para la detección, el diagnóstico y el pronóstico de la diabetes mellitus (Chen and

Gerszten, 2020). Una asociación directa de los aminoácidos de cadena ramificada y una asociación inversa de la glicina con la DMT2, se ha encontrado en varios estudios realizados con la metabolómica como base del experimento en población caucásica (Satheesh *et al.*, 2020). Algunos módulos de metabolitos, incluidos los metabolitos relacionados con la diabetes en alimentos vegetales y esteroides androgénicos previamente informados, se asociaron con una mayor calidad de la dieta, un menor riesgo de diabetes y cambios longitudinales favorables en HOMA para la resistencia a la insulina (Chai *et al.*, 2022).

La patogénesis de la DN, puede diagnosticarse tempranamente con biomarcadores no invasivos tomando como base el ciclo de la urea, el ciclo de los TCA, la glucólisis y el metabolismo de los aminoácidos, siendo el ácido láctico, el ácido hipúrico, la alantoína (en orina) y la glutamina (en sangre), los en lo más alto, como sugiere un metanálisis (Roointan *et al.*, 2021). La valina (o betaína) y el 3-(4-metil-3-pentenil)tiofeno se asociaron con un mayor riesgo de enfermedad renal terminal (Zhang *et al.*, 2022). Los biomarcadores de pronóstico proporcionados por la metabolómica tienen potencial para descubrir mecanismos en la progresión de la ND. Estudios recientes presentan antígenos diana potenciales en la nefropatía membranosa, con la firma de péptidos urinarios, esto agrega información de pronóstico a la albúmina urinaria e implica a las proteínas inflamatorias circulantes como mediadores potenciales de la ND, lo que demuestra la importancia de la bioenergética renal como factor modificable en la lesión renal aguda (Dubin and Rhee, 2019).

Las características antropométricas como el BMI, la relación cintura-cadera (WHR) y la relación cintura-altura (WHtR) pueden ser útiles como factores de predicción, debido a la naturaleza de cada característica, uno puede ser más útil que el otro (Moosaie *et al.*, 2021). Otras características antropométricas basadas en las mediciones de BP y HbA1c proporcionan varias diferencias estadísticamente significativas que pueden identificarse en relación con el origen étnico (Spurr *et al.*, 2020). La metabolómica es un enfoque prometedor para identificar biomarcadores que podrían ayudar en la detección temprana y la prevención de la DMT2 y la comorbilidad de la nefropatía. La metabolómica se ha utilizado para identificar posibles biomarcadores de DMT2.

Un estudio desarrolló un protocolo basado en farmacometabolómica, resonancia magnética nuclear y espectrometría de masas para determinar el mejor enfoque de tratamiento basado en biomarcadores metabolómicos identificados. Esto también se ha utilizado para predecir la progresión de la DMT2 mediante imágenes de EM, sin duda, una tecnología prometedora que utiliza procesamiento de imágenes y videos para identificar y proporcionar información precisa sobre moléculas en las superficies de los tejidos (Saigusa *et al.*, 2021). Sin embargo, la naturaleza compleja y de alta dimensión de los datos metabolómicos plantea un

desafío a la hora de identificar biomarcadores relevantes. La detección temprana de DMT2 a través de modelos de ML es posible, con un conjunto de datos que proporciona valores de características recopilados a través de un período de tiempo específico, este conjunto de datos también se puede mejorar con técnicas de refuerzo como XGBoost o LightGBM para hacer predicciones más precisas (Kopitar *et al.*, 2020a). Como estos procedimientos como el panel de glicerolípidos son complejos y tienen un costo diez veces mayor que un panel de lípidos o un hemograma completo (los precios varían en cada país y dependen de otros factores como pólizas de seguro, laboratorios o instituciones de salud), surge la necesidad de procedimientos menos invasivos y surge el análisis de bajo costo.

Las manifestaciones clínicas del síndrome metabólico son procesos fisiológicos subyacentes que a menudo se pasan por alto. Este descuido es cada vez más significativo, especialmente dado el preocupante aumento de su prevalencia entre las mujeres jóvenes. En este aspecto conciso, existen distinciones entre hombres y mujeres en cuanto a la epidemiología, las causas, los mecanismos biológicos y la presentación clínica del síndrome metabólico. Específicamente, las diferencias sexuales notables abarcan cosas como: diferentes tasas de disglucemia, disparidades en la distribución de la grasa, variaciones en el tamaño y función de los adipocitos, regulación hormonal del peso corporal y la acumulación de grasa, impacto de la disminución de los estrógenos en los factores de riesgo, entre otros (Aruna D., 2014). Estas alteraciones metabólicas pueden enfatizar una relación directa entre el perfil metabólico, el sexo, la DMT2 y sus comorbilidades. Los pacientes con síndrome metabólico presentan variaciones en la aparición de niveles anormales de azúcar en sangre entre los sexos, estas alteraciones aumentan el riesgo de padecer DMT2, además, la función y tamaño de las células grasas o adipocitos y los mecanismos hormonales que rigen el peso corporal y el almacenamiento de grasa exhiben sexo.

3.4. Comorbilidades asociadas a la diabetes y machine learning

Allen *et al.* (2022) presenta que la ND tiende a manifestarse en aproximadamente la mitad de los pacientes con DMT2, el enfoque en la investigación de la diabetes debe estar dirigido a superar los desafíos asociados con los diagnósticos retrasados, que a menudo resultan de prácticas de detección inconsistentes. Este estudio implementó Decision Trees y dos variaciones de este modelo, RF y árboles potenciados por gradiente (XGB) capaces de predecir diferentes etapas de DKD dentro de un período de 5 años después del diagnóstico inicial de DMT2. Los resultados indicaron que los modelos superaron el rendimiento de la pun-

tuación de riesgo de los Centros para el Control y la Prevención de Enfermedades tanto en la prueba de reserva como en los conjuntos de datos externos.

Otra investigación realizada por Chan *et al.* (2021) mejorar la predicción de la progresión en DKD, formulando y validando una puntuación de riesgo de pronóstico aprendida por máquina llamada KidneyIntelX™, este modelo innovador integra registros médicos electrónicos (EHR) y biomarcadores que involucran el entrenamiento de un modelo de RF. El modelo fue evaluado por AUC, los Valores Predictivos Positivos y Negativos (PPV/NPV) y el Índice de Reclasificación Neta (NRI). Se realizó un análisis comparativo con un modelo clínico convencional y las categorías Enfermedad renal: mejora de los resultados globales (KDIGO). Esta investigación tenía como objetivo predecir un resultado compuesto, específicamente la disminución en la tasa de filtración glomerular estimada (eGFR) de ≥ 5 ml/min por año, $\geq 40\,\%$ de disminución sostenida o insuficiencia renal en un lapso de 5 años. La correlación entre la derivación y la validación del modelo de ML fue fundamental para evaluar su eficacia y confiabilidad en la predicción de resultados cruciales en la progresión de la ERC.

Nagaraj and Kieneker (2021) investigaron el potencial de utilizar el índice de edad renal (KAI), una métrica de la desviación entre la edad biológica (BA) y la edad cronológica (CA), para evaluar la función renal en pacientes con DKD. El KAI se desarrolló entrenando algoritmos de ML LR, RF, SVM y red neuronal de retroalimentación (FNN), en tres conjuntos de datos: PREVEND, RENAAL e IDNT de individuos sanos y pacientes con DKD con el algoritmo FNN que tiene el mejor rendimiento para predecir CA en función de varios marcadores clínicos y luego se utilizan para predecir BA para cada paciente. El KAI se calculó como la diferencia entre BA y CA. Los investigadores encontraron que el KAI era significativamente mayor en pacientes con DKD que en individuos sanos. Esto sugiere que la DKD acelera el proceso de envejecimiento, lo que lleva a una edad biológica más alta de lo que se esperaría según la edad cronológica.

La predicción de la nefropatía según el estudio realizado por Ou *et al.* (2023b) incluyen LR, clasificador de árbol adicional, RF, GBM, aumento de gradiente extremo (XGBoost) y máquina de aumento de gradiente ligero para predecir el riesgo de desarrollar enfermedades renales en etapa terminal en DMT2 recién diagnosticada. Los resultados apuntados en esta investigación acortan la brecha entre la predicción de cada enfermedad y agilizan el diagnóstico precoz previniendo comorbilidades más graves.

Otro estudio de cohorte retrospectivo realizado por Lin *et al.* (2017) cuyo objetivo fue desarrollar un modelo de predicción de ND e insuficiencia cardíaca (IC) para pacientes con DMT2. Siguiendo los procedimientos del Framingham Heart Study, un modelo de regresión de riesgos proporcionales de Cox identificó predictores clave. La puntuación de riesgo incorporó factores como la edad, el sexo,

la aparición de diabetes, la presión arterial, los medicamentos, la creatinina, la HbA1c, la presión arterial sistólica, la retinopatía, la albuminuria, los medicamentos para la diabetes y la hiperlipidemia. El modelo demostró alta precisión y discriminación para predecir, ofreciendo un valioso potencial de detección para la prevención temprana en pacientes con DMT2. Kanda *et al.* (2022) Al desarrollar un modelo de predicción de DKD para pacientes con DMT2 siguiendo los procedimientos del Framingham Heart Study, un modelo de regresión de riesgos proporcionales de Cox identificó predictores clave, ya que la puntuación de riesgo incorporó factores como la edad, el sexo, la aparición de diabetes, la presión arterial, los medicamentos, la creatinina, la HbA1c, la presión arterial sistólica, presión arterial, retinopatía, albuminuria, medicamentos para la diabetes e hiperlipidemia.

Rodríguez-Romero *et al.* (2019) llevaron a cabo un análisis retrospectivo del ensayo "*Action to Control Cardiovascular Risk in Diabetes*"(ACCORD), se centra en diferenciar los predictores tempranos y tardíos estratificando datos longitudinales según el tiempo posterior a la inscripción de los pacientes. Entre los algoritmos evaluados, los métodos RF y Simple LR demostraron un rendimiento superior, como factores basales clave como la tasa de filtración glomerular (TFG), la creatinina urinaria, la albúmina urinaria, el potasio, el colesterol, las lipoproteínas de baja densidad y la relación albúmina-creatinina urinaria, se identificaron como predictores de nefropatía. Los primeros predictores incluyeron valores iniciales de TFG, presión arterial sistólica, FPG y potasio en el mes 4.

Agliata *et al.* (2023) presentan un enfoque innovador que utiliza ANN para predecir la DMT2, lo que ofrece una promesa significativa para el manejo y la prevención de enfermedades. Este estudio emplea un clasificador binario entrenado desde cero, explorando asociaciones previamente desconocidas entre parámetros de salud y la aparición de diabetes. Aprovechando tres conjuntos de datos, incluida la encuesta bienal de NHANES, MIMIC-III y MIMIC-IV, la investigación muestra la capacidad predictiva superior a largo plazo de la ANN. Este trabajo sienta las bases para una evaluación precisa de los riesgos y la detección temprana, contribuyendo a los avances en las estrategias de prevención de la diabetes. Los estudios dicotómicos han mostrado resultados prometedores y un conjunto diferente de herramientas a lo largo de los años para el diagnóstico, pronóstico y detección temprana de enfermedades. Glmnet LASSO proporciona un modelo de regresión versátil para predecir DMT2 no diagnosticada en un estudio que evalúa modelos basados en ML entre RF, XGBoost y LightGBM frente a modelos de regresión tradicionales, utilizando lotes de 6 meses de datos entrantes simulados. Glmnet muestra una mejora con datos adicionales, mientras que LightGBM demuestra la mayor estabilidad de selección de variables a lo largo del tiempo, lo que demuestra sus sólidas capacidades predictivas en el diagnóstico de

DMT2 (Kopitar *et al.*, 2020b).

El uso de XGBoost surgió como el modelo más eficaz, demostrando una alta *Sensitivity* y *Specificity*. Para las mujeres, la diabetes y las artralgias fueron marcadores clave, mientras que la ND y el dolor en el pecho fueron cruciales para los hombres. La disnea, la hipertensión y la polipnea fueron factores de riesgo universales. La edad resultó ser el factor demográfico más influyente en la letalidad. Estos marcadores ofrecen información valiosa para la clasificación inicial, especialmente en regiones con recursos limitados (Rojas-García *et al.*, 2023).

El papel de los lípidos en la interacción entre infecciones virales, respuestas metabólicas del huésped y reacciones inmunes. Centrándose en el COVID-19, una investigación de Castañé *et al.* (2022) utilizaron las firmas lipidómicas de pacientes con COVID-19 positivo, comparándolas con pacientes con enfermedades inflamatorias y voluntarios sanos como control. Se empleó cromatografía líquida combinada con espectrometría de masas y herramientas de ML para interpretar datos matizados, desentrañando distintos perfiles lipídicos asociados con COVID-19 y otras afecciones inflamatorias.

Al proponer un marco híbrido que mitiga los problemas mal planteados, de sobreajuste y de desequilibrio de clases, empleando técnicas de minería de datos, específicamente una combinación del algoritmo de regresión Lasso para la selección y regularización de variables, y el clasificador de ANN, un estudio de Singh Y. y Tiwari M. logra una notable *Precision* de clasificación del 93 % utilizando el conjunto de datos de mujeres de los indios Pima como sujetos. Al reducir eficazmente el tiempo de computación mediante la regresión Lasso, este enfoque híbrido ofrece previsibilidad e interoperabilidad mejoradas. Los hallazgos sugieren que estas metodologías de extracción de datos son muy prometedoras para ayudar a los médicos a realizar diagnósticos precisos de diabetes mellitus, contribuyendo a mejorar la toma de decisiones médicas (Singh and Tiwari, 2022).

El osteopontin (OPN), es un un biomarcador presentado por Moszczuk *et al.* (2022) que emerge como un candidato prometedor para la detección de nefropatía, incluidos pacientes con diversas glomerulopatías (GN) comprobadas por biopsia, como nefropatía por inmunoglobulina A (NIgA), nefropatía membranosa (NM) y nefritis lúpica (NL), junto con un grupo control. Utilizando Boruta y RF, se reveló que la OPN mostró un aumento significativo en la NIgA en comparación con otras GN al inicio del estudio, lo que permitió el desarrollo de un algoritmo de ML con una *Precision* del 87 % para diferenciar la IgAN de otras GN basándose únicamente en los niveles de OPN en orina.

3.5. Selección de características y biomarcadores

Las características identificadas como factores de riesgo de la ND abarcan varios elementos. Estos comprenden la excreción urinaria de albúmina, los niveles de glucosa, la presión arterial, la dislipidemia, la obesidad, el tabaquismo, la duración de la diabetes, la edad, el sexo y la retinopatía. Además, los factores de riesgo reconocidos incluyen estrés oxidativo, inflamación, antecedentes genéticos, origen étnico e hiperfiltración glomerular (Slieker *et al.*, 2021). Estos diversos factores contribuyen colectivamente al intrincado panorama de la susceptibilidad a la ND. Comprender y abordar estos determinantes multifactoriales es crucial para estrategias integrales de gestión y prevención. Un nomograma presentado en un estudio chino contribuye a esta evaluación y presenta una lista de características como predictores que incluyen SBP, DBP, FPG, HbA1c, triglicéridos totales, creatinina sérica, nitrógeno ureico en sangre y BMI. Todas estas características se han presentado anteriormente como biomarcadores no solo para la ND sino también como biomarcadores para predecir la DMT2 en las primeras etapas, lo que generó el foco de diversos estudios para establecer el control de estas características como tratamiento para estas enfermedades.

Un estudio realizado por Sabitha and Durgadevi (2022) tuvo como objetivo subrayar la importancia del preprocesamiento de datos, la selección de características y el aumento de datos en la predicción de enfermedades, utilizando técnicas en estos dominios para mejorar la eficacia de los algoritmos de clasificación en el diagnóstico y predicción de la diabetes, empleando una propuesta método en el conjunto de datos PIMA Indians, una comparación sistemática de tres categorías reveló que el preprocesamiento de datos, RFE con selección de características de regresión de bosque aleatoria y el aumento de sobremuestreo SMOTE lograron puntuaciones de *Precision* notables superiores al (80 %) entre seis clasificadores (LR, RF, DT, SVC, GNB y KNN). La eliminación recursiva de características con RF se emplea para la selección de características significativas y la estimación del sistema revela una asociación sólida de la diabetes con el BMI y el nivel de glucosa, extraída mediante el enfoque a priori. Estos enfoques se pueden comparar con el enfoque ANN, lo que ofrece potencial apoyo a los profesionales médicos en las decisiones de tratamiento (Tiwari and Singh, 2021).

La aplicación del modelo de aprendizaje disperso LASSO demostró ser eficaz para descubrir patrones epidemiológicos asociados con la retinopatía diabética (RD) y otras comorbilidades, hay estudios que utilizaron el ML para predecir el riesgo de RD con base en registros médicos, destacando a LASSO como una excelente opción para pacientes con alto riesgo de RD. Análisis dimensional de registros médicos electrónicos, lo que demuestra su eficacia no sólo al proporcionar poder discriminativo sino también al sobresalir en la selección de variables, enfa-

tizando su potencial para avanzar en nuestra comprensión y predicción de la RD dentro del ámbito del análisis de atención médica (Oh *et al.*, 2013). Aprovechando la técnica LASSO para identificar factores de riesgo cruciales, se realizó una comparación exhaustiva con otros algoritmos (LR, RF, SVM, LDA, NB y Treebag). Posteriormente, se puede emplear el análisis LR multivariado para construir nomogramas para la predicción individualizada, con evaluación a través de curvas características operativas del receptor y calibración, la superioridad de LASSO está probada para la predicción del riesgo de diabetes. En particular, los nomogramas incluyeron factores específicos para la prediabetes y su progresión a diabetes, lo que demuestra una discriminación sólida y se confirma mediante curvas de calibración bien ajustadas (Ou *et al.*, 2023a).

Kocbek *et al.* (2022) emplearon LASSO para la selección de características, utilizando un modelo de regresión logística para predecir la DMT2 no diagnosticada. Este modelo incorporó características como FPG, edad, sexo, BMI y circunferencia de la cintura, logrando un AUC de 0.818 (81.8%) en el conjunto de datos de prueba. Este enfoque demostró ser una herramienta de predicción efectiva e interpretable, lo que puede contribuir a una mayor confianza en los modelos de predicción por parte de los expertos en atención médica.

En cuanto a la combinación de modelos, KNN se emplea como un metaclasificador que combina las predicciones de los alumnos base. Los resultados comparativos demuestran que el método NSGA-II-Stacking propuesto supera significativamente a varios enfoques de ML individuales y a los enfoques de conjunto convencionales. En términos de métricas de rendimiento, el sistema propuesto alcanza la mayor *Precision* del 83.8%, *Sensitivity* del 96.1%, *Specificity* del 79.9%, métrica f del 88.5% y AUC del 85.9%. Liu *et al.* (2019) propusieron el uso de algoritmos de ML para mejorar la precisión de las predicciones de DMT2 utilizando sistemas de puntuación de riesgo no invasivos. LASSO, la desviación absoluta recortada suavizada (SCAD) y la probabilidad penalizada cóncava minimax (MCP), que se usan comúnmente en la selección de variables para modelos de alta dimensión, se utilizaron para seleccionar automáticamente valores no invasivos significativos. Factores de riesgo para la DMT2. También se aplicaron a este conjunto de datos un método de selección de modelo más conservador para el modelo de dimensiones ultra altas, el procedimiento ISIS para la selección de variables en regresión logística y la regresión logística por pasos tradicionales. La SVM, los métodos basados en árboles y la red neuronal fueron tres técnicas de ML comúnmente utilizadas para la predicción del riesgo de diabetes, los resultados del modelo conjunto con un selector de características son AUC de 0.802 (0.780, 0.825), *Sensitivity* de 0.662 (0.614, 0.709) y *Specificity* de 0.702 (0.676, 0.728).

Otros estudios han utilizado algoritmos genéticos como herramienta de selec-

ción de características, estudios recientes han demostrado que la combinación de investigaciones se centra en optimizar los modelos de detección de diabetes mediante la integración de dos técnicas de selección de características como el criterio de información de Akaike y los algoritmos genéticos se pueden combinar con algoritmos clasificadores destacados, para demostrar desempeño superior (García-Domínguez *et al.*, 2023). Abordar la tarea crucial de seleccionar metabolitos diferenciales es vital para aplicaciones tanto biológicas como clínicas, sin embargo, hay miles de metabolitos y se necesita un enfoque de selección de características para desarrollar análisis complejos de conjuntos de datos metabolómicos, aprovechando la efectividad de modelos como las SVM como clasificador básico.

En la comunidad de ML, existe una tendencia creciente hacia el empleo de modelos complejos, particularmente en campos como la visión por computadora y el procesamiento del lenguaje natural (NLP). Estos modelos complejos a menudo exigen importantes recursos computacionales. Sin embargo, los modelos más simples pueden seguir siendo muy eficaces, especialmente cuando los datos abundan. La selección de características se vuelve crucial para mejorar la precisión del modelo. El enfoque de eliminación recursiva de características con validación cruzada (RFECV) para predecir la diabetes tipo II tiene como objetivo aumentar la precisión de la clasificación. Al superar desafíos como el sobreajuste, el enfoque de esta técnica incorpora métodos de preprocesamiento adicionales, luego los resultados se pueden evaluar con algoritmos de ML clásicos, incluidos LR, ANN, *Naïve Bayes*, SVM y *Decision Trees* (Misra and Yadav, 2020).

Un método diseñado por Lin *et al.* (2011) emplea la eliminación de características recursivas de la SVM (SVM-RFE) para una identificación significativa de características, además, se utilizan GA y RF para capturar las interacciones de los metabolitos y el rendimiento individual, respectivamente, validando en un conjunto de datos de metabolómica plasmática de ratas. En enfermedades hepáticas, el enfoque identifica 31 metabolitos importantes, lo que muestra un efecto sinérgico de los tres métodos de selección para obtener información metabólica integral sobre diferentes enfermedades hepáticas. Esta técnica también se ha aplicado en conjuntos de datos humanos. El algoritmo genético puede identificar un conjunto óptimo para detectar DMT2, utilizando un modelo como GA basado en RF (miRDM-rfGA) como algoritmo de selección de características, puede generar subconjuntos y compararlos con configuraciones que utilizan características tradicionales los métodos de selección (prueba F y LASSO) (Park and Nam, 2023).

La selección de características recursivas (RFE) se ha comparado con otros selectores de características que utilizan el paquete Boruta y el algoritmo genético (GA) aplicado al conocido conjunto de datos de diabetes PIMA Indians y se ha

descubierto que los boruta los superan utilizando el algoritmo ID3 del árbol de decisión. Estos resultados presentados por Sadhasivam J.*et al.*, indican que la aplicación de algoritmos de selección de características en un conjunto de datos estándar, como el conjunto de datos PIMA Indians, no produce cambios en la precisión, sugiere que la información en un conjunto de datos estándar ya está preprocesada y no requiere procesamiento adicional. Sin embargo, cuando se aplican a un conjunto de datos recopilados de un hospital local, estos métodos de selección de características aumentan significativamente la precisión del modelo (Sadhasivam *et al.*, 2021). En términos de precisión, el paquete Boruta es el algoritmo de selección de características más eficaz para el conjunto de datos local de infección por diabetes, sin embargo, estos resultados pueden ser refutados, ya que el modelo utilizado para la validación fue solo Decision Trees.

Capítulo 4
Materiales y métodos

La metodología empleada para probar nuestra hipótesis se basó en procesos establecidos de ML. Cada método para el desarrollo de modelos de ML se adapta a las necesidades específicas de la organización o el problema particular que se está abordando. Nuestra investigación sigue de cerca el marco de trabajo esbozado por Alberto Maisueche Cuadrado en su tesis de maestría en la Universidad de Valladolid (Maisueche Cuadrado, 2019), ya que se alinea bien con la metodología propuesta por Hernández Sampieri. Este enfoque enfatiza una estrategia integral de resolución de problemas para ML, no solo centrada en la selección y entrenamiento de modelos, sino que también incluye una serie de pasos críticos para maximizar el éxito. Las etapas de este proceso se detallan en la figura 4.1.

El proceso de experimentación fue dividido en 3 casos de estudio, En el primer caso se presenta una primer solución al problema de clasificación en el dataset SIGLO XXI realizando una implementación de LASSO como selector de características y un modelo ensamble validado por AUC, como punto a resaltar en este primer caso se realizó la exclusión de los biomarcadores conocidos de la glucosa FPG y HbA1c. Para el segundo caso se realizó una exahustiva selección de características con algorítmos genéticos validada por distintos modelos, en este

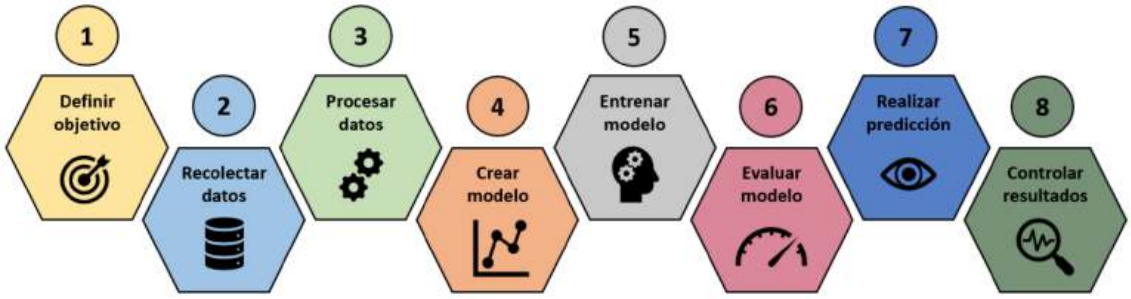


Figura 4.1: Diagrama de flujo de la metodología ML

.

segundo caso solo se realizó un análisis de la metabolómica involucrada en la detección de la DMT2 y su comorbilidad más común la ND. Por último, en el tercer caso se realiza un estudio completo en el que se incluyen ahora el criterio de información de Akaike (AIC), tres técnicas de selección de características (LASSO, AG y RFE) y un modelo ensamble. En este último estudio el enfoque principal fue la separación de los dataset por sexo y la mejora en el algorítmo de selección de características RFE combinando el AIC con RFE con SVM y un kernel lineal y una comparativa iterativa con la métrica ACC. En cada caso la metodología presenta ligeros cambios en la primera fase (preprocesamiento de los datos), así como también en las fases de selección de características (distintas técnicas en cada caso), implementación de modelos (distintos modelos) y su evaluación (Con las métricas establecidas en el marco teórico).

4.1. CASO 1: Selección LASSO / Ensamble / dataset SIGLO XXI

Para el primer caso de estudio se presenta la siguiente metodología en la figura 4.2, una metodología basada en la propuesta por Akhtar *et al.* (2021) que se explica de la siguiente manera: el primer paso es realizar un análisis preliminar de los datos de la muestra, seguido del procesamiento e imputación de los datos, preparación de los pliegues y separación de los datos en prueba y entrenamiento, realizando la selección de características, desarrollando los modelos e integrándose en el conjunto para finalmente validar los modelos con los datos de prueba y extraer el AUC, *Sensitivity* y *Specificity* (Morgan-Benita *et al.*, 2022b).

4.1.1. Muestra

La base de datos fue proporcionada por el Centro Médico Nacional Siglo XXI ubicado en la Ciudad de México. Todos los pacientes mexicanos firmaron una carta de consentimiento informado y el protocolo cumple con los criterios de Helsinki los cuales fueron aprobados por el Comité de Ética del Instituto Mexicano del Seguro Social bajo el número R-2011-785-018. Incluye 1787 pacientes, 898 casos positivos de DMT2 y 889 pacientes controles, según género, esta base de datos incluye: 892 hombres y 895 mujeres. En la figura 4.3 se muestra la correlación inicial entre las características (Morgan-Benita *et al.*, 2022b).

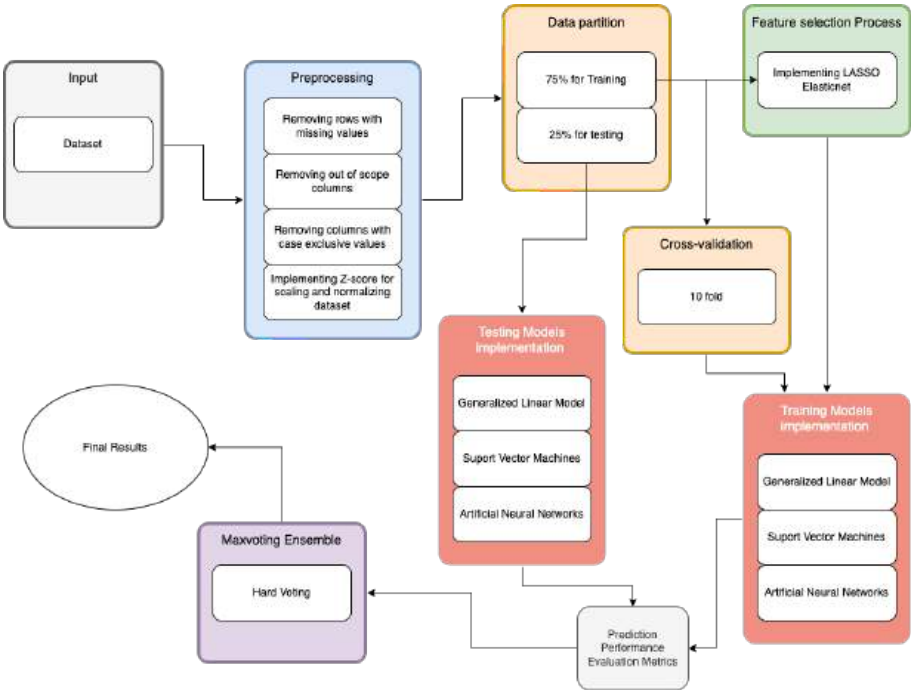


Figura 4.2: Diagrama de flujo de la metodología propuesta en el caso de estudio 1.

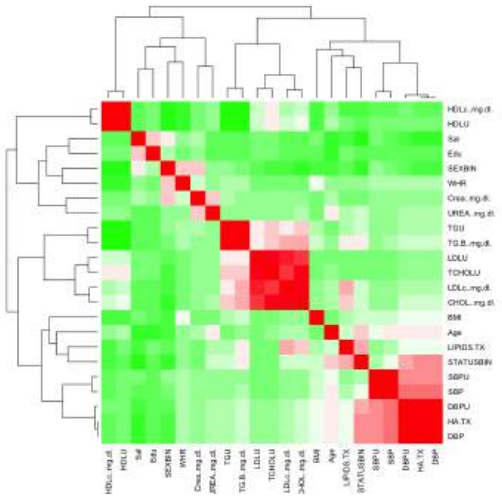


Figura 4.3: Mapa de calor de correlación de características

4.1.2. Tratamiento de datos

Este paso del proceso incluye el tratamiento de datos, comenzando con la imputación de datos, enfocándose en valores faltantes y características fuera del alcance de este estudio, el siguiente paso son los valores de imputación de datos incluidos en la base de datos y el paso final es la normalización de los datos , que corresponden a todas las características incluidas en LASSO para el proceso de selección de características que se muestra en la tabla 4.2.

4.1.3. Imputación de datos

Este trabajo utiliza el criterio de exclusión de trabajar solo con las observaciones que tienen datos completos para todas las variables y descartar todas las demás, eliminando todas las características fuera de alcance o inutilizables (Scheffer, 2013), se eliminaron las características que identifican a los pacientes y los ID, ya que solo eran datos de identificación interna del paciente y un número consecutivo, respectivamente. Como en el conjunto de datos existe una serie de características y En los pacientes con datos faltantes se eliminaron todos los datos nulos o no disponibles (NA), ya que las características identificadas con NA eran exclusivas de los casos (pacientes positivos para DMT2) y muestran datos inexistentes en los pacientes control (pacientes negativos para DMT2) o estos. Los datos no fueron recolectados ni registrados, por lo que no fueron tomados en consideración. Con base en este criterio se eliminaron todos los medicamentos y su cantidad de ingesta diaria: Glibenclamida, Dosis de Glibenclamida, metformina, Dosis de metformina, Pioglitazona, Dosis de Pioglitazona, Rosiglitazona, Dosis de Rosiglitazona, Acarbosa, Dosis de Acarbosa, Insulina y Dosis de Insulina. Datos de 3 pacientes, estos pacientes (en los 3 casos) no tenían datos sobre la característica Hipertensión bajo tratamiento, también se eliminaron las características Filtrado glomerular, Edad del diagnóstico (edad en la que se presenta el diagnóstico positivo de DMT2, datos exclusivamente en casos), HbA1c y complicaciones de DMT2, ya que solo hay datos actuales de pacientes con DMT2. La característica de complicaciones de la DMT2 contiene comorbilidades asociadas con la DMT2 y también están fuera del alcance de este experimento, ya que contiene NA en algunos de los casos y NA en todos los pacientes de control.

La función de glucosa es un biomarcador bien conocido, se elimina del alcance de este experimento para revisar el rendimiento de las otras características. Esta característica en la prueba de modelos univariados tenía más del 90 % de AUC, lo que la convierte en la característica más importante del conjunto de datos analizado.

Todas las características eliminadas se enumeran en la tabla 4.1 y las características analizadas se enumeran en la tabla 4.2.

4.1.4. Normalización de datos

La mayoría de las características en el conjunto de datos analizado tienen un rango diferente de valores, se implementa Z-Score para que las características puedan compararse entre sí o tomarse como parte del modelo final configurándose con mayor precisión en el mismo rango. El Z-Score consiste en: calcular la media de los valores de una característica, calcular la desviación estándar de los mismos valores y finalmente calcular el Z-Score con la siguiente fórmula (Wiesen, 2018):

$$Z = \frac{x - \bar{x}}{\sigma_x} \tag{4.1}$$

donde x es la puntuación bruta, $\bar{x}$ la media de todas las x y σ la desviación estándar de x. Este proceso da como resultado un valor en un rango estándar, que se aplica a diferentes características. Se ha implementado una función genérica cuyo método predeterminado centra y/o escala las características de una matriz numérica.

4.2. CASO 2: Algoritmos Genéticos / Metabolómica

La metodología de este estudio consta de seis etapas, como se muestra en la figura 1 y se explica de la siguiente manera: la primera etapa describe el conjunto de datos utilizado (figura 1A). En la segunda etapa, se realizó un análisis preliminar de los datos seleccionando sujetos de acuerdo con los datos aplicando los criterios de inclusión (figura 1B). Posteriormente, los datos del conjunto de datos se separaron en tres grupos: Control - Prediabetes, Control - Diabetes y Control - Nefropatía Diabética (figura 1C). En la cuarta etapa, se implementa la selección de características a través de un algoritmo genético. Mencionado (figura 1D). En la quinta etapa, los modelos de ML (SVM, KNN y Nearcent) se desarrollaron utilizando las características principales de la etapa anterior (figura 1E). Finalmente, los modelos fueron validados tomando en consideración diferentes métricas (*Precision*, *Sensitivity*, *Specificity* y AUC) para determinar el rendimiento de nuestros modelos (figura 1F) (Morgan-Benita *et al.*, 2022a).

Los cuadrados azules se refieren a la metodología de análisis de datos, mientras que los cuadrados blancos detallan la tarea involucrada en cada paso. **(A)** El conjunto de datos de metabolómica se obtiene de la Unidad de Investigación Médica Biomédica. **(B)** El conjunto de datos se analiza y se crean nuevos conjuntos de datos seleccionando sujetos de acuerdo con los criterios descritos en la tabla 4.5. **(C)** Los datos se normalizan mediante la normalización del cociente probabilístico (PQN) y las observaciones del conjunto de datos se analizan y separan

Tabla 4.1: Características descartadas.

Características	**Descripción**	**Valores posibles**
Age DX	Edad de diagnóstico de DM	Numérico entero
Glucose	Niveles de glucosa en la sangre	Numérico Flotante
HbA1c	Hemoglobina Glicosilada	Numérico Flotante
GFR	Tasa de filtración glomerular (Prueba de sangre que comprueba que tan buen funcionamiento tienen los riñones)	Numérico entero
Glibenclamide	Tratamiento farmacológico	0 - No 1 - Si
Metformin	Tratamiento farmacológico	0 - No 1 - Si
Pioglitazone	Tratamiento farmacológico	0 - No 1 - Si
Rosiglitazone	Tratamiento farmacológico	0 - No 1 - Si
Acarbose	Tratamiento farmacológico	0 - No 1 - Si
Insulin	Tratamiento farmacológico	0 - No 1 - Si
Complications T2DM	Complicaciones asociadas con DMT2	NEUROPATHY - Tiene neuropatía RETINOPATHY - Tiene retinopatía

Todas las características de esta tabla fueron excluidas del análisis mediante imputación de datos.

en tres grupos: Control - Prediabetes, Control - Diabetes y Control - Nefropatía diabética. **(D)** Se implementa el uso de algoritmos genéticos para extraer las principales características de los datos. **(E)** Utilizando las características principales para la detección de Prediabetes, Diabetes y Nefropatía Diabética en pacientes, se generan varios modelos usando las máquinas de vectores de soporte, k-vecino más cercano y centroide más cercano. **(F)** La validación de nuestros resultados se realiza utilizando diferentes métricas (validación cruzada por GALGO y ACC

Tabla 4.2: Descripción de características y posibles valores.

Característica	**Descripción**	**Valores posibles**	**P-Value univariado por regresión logística**
Educación	Estudios concluidos por el paciente	1 - Primaria 2 - Secundaria 3 - Técnica 4 - Preparatoria 5 - Universidad 6 - Posgrado	0.00118
Salario	Ingreso mensual	1 - Menos de 2000.00 2 - Entre 2000.00 y 5000.00 3 - Más de 5000.00	2^{-16}
Sexo	Sexo del paciente	0 - Masculino 1 - Femenino	6.7^{-16}
Edad	Edad del paciente en años	Entero numérico	2^{-16}
WHR	Índice cintura-cadera	Numérico	2.89^{-5}
BMI	Índice de masa corporal	Numérico	9.35^{-16}
Urea	Producto de desecho resultante de la descomposición de proteína en el cuerpo del paciente	Entero numérico	2^{-16}
Creatinina	Producto de desecho producido por los músculos como parte de actividad diaria regular	Numérico	0.000456
Tratamiento de lípidos	Niveles de lípidos con tratamiento	1 - Tratamiento de lípidos 2 - Sin tratamiento de lípidos.	0.956

Todas las características de esta tabla se incluyeron en el análisis mediante imputación de datos.

Tabla 4.3: Descripción de características y posibles valores (Continuación).

Característica	**Descripción**	**Valores posibles**	**P-Value univariado por regresión logística**
Cholesterol	Sustancia parecida a la grasa que se encuentra en todas las células del cuerpo del paciente.	Numérico	2^{-16}
HDL	Lipoproteína de alta densidad (corregida por medicación).	Numérico	3.77^{-12}
LDL	Lipoproteína de baja densidad (corregida por medicación).	Numérico	2^{-16}
Triglycerides	Tipo de grasa que se encuentra en el cuerpo del paciente.	Numérico	2^{-16}
TCHOLU	Colesterol total (sin corregir).	Entero numérico	0.258
HDLU	Lipoproteína de alta densidad (sin corregir).	Entero numérico	2.12^{-5}
LDLU	Lipoproteína de baja densidad (sin corregir).	Entero numérico	0.240
TGU	Triglicéridos (sin corregir).	Entero numérico	2^{-16}

Todas las características de esta tabla se incluyeron en el análisis mediante imputación de datos.

Tabla 4.4: Descripción de características y posibles valores (Continuación).

Característica	**Descripción**	**Valores posibles**	**P-Value univariado por regresión logística**
SBP	Presión arterial sistólica (corregida por medicamento)	Entero numérico	2^{-16}
DBP	Presión arterial diastólica (corregida con medicamentos)	Entero numérico	2^{-16}
SBPU	Presión arterial sistólica (sin corregir)	Entero numérico	1.28^{-10}
DBPU	Presión arterial diastólica (sin corregir)	Entero numérico	2^{-16}
HA-TX	Tratamiento de la hipertensión	0 - No en tratamiento de hipertensión. 1 - En el tratamiento de la hipertensión	0.959
Output	Clasificador de pacientes	0 - Paciente negativo para DMT2 1 - Paciente positivo para DMT2	-

Todas las características de esta tabla se incluyeron en el análisis mediante imputación de datos.

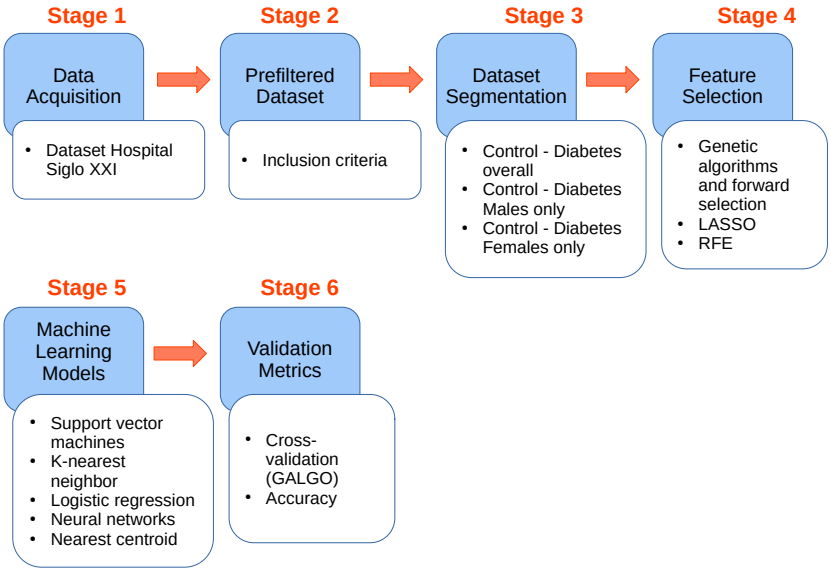


Figura 4.4: Diagrama de flujo de la metodología propuesta en el caso de estudio 2.

promedio) para determinar cuál de los modelos tiene mejor rendimiento.

4.2.1. Muestra

La metodología implementada para obtener los datos metabolómicos de cada uno de los 80 pacientes, se presenta en subsecciones: preparación de muestras, Controles de Calidad (QC) y Garantía de Calidad (QA), Cromatografía Líquida de Ultra Rendimiento (UPLC)-Método de Espectrometría de Masas para Análisis Lipidómico , análisis de datos y análisis estadístico.

4.2.2. Preparación de la muestra

Se extrajeron 100 uL de plasma descongelado en hielo con 300 uL de isopropanol preenfriado (grado LCMS, Honeywell, Charlotte, NC, EE. UU.), se agitaron durante 1 min y se incubaron a -20 °C durante 1 h para la precipitación de proteínas. Posteriormente, la mezcla de extracción se centrifugó a 15800 g durante 15 min y se recogieron los sobrenadantes. Para el análisis, cada alícuota se transfirió a viales de LC y se diluyó a 1:20 con una mezcla de isopropanol/acetonitrilo/agua (2:1:1, v: v: v). El orden de preparación de la muestra se

obtiene de un proceso aleatorio a partir de la selección de la muestra y se volvió a aleatorizar a partir del orden del análisis de la muestra para garantizar que no hubiera sesgos sistemáticos.

4.2.3. Controles de Calidad (QC) y Garantía de Calidad (QA)

Estos procesos se relacionan con los procedimientos aplicados en preparación para la adquisición de datos (QA) y durante/después de la adquisición de datos (QC). Como parte de los procedimientos de control de calidad, el equipo fue sometido a un mantenimiento completo antes de cada bloque analítico. Este mantenimiento incluyó tanto el sistema de cromatografía, el analizador de masas y la fuente de nitrógeno. La limpieza del cono de muestra y de la fuente de iones se realizó entre lotes analíticos. La calibración y el ajuste manual también se realizaron inmediatamente antes de analizar las muestras. Un conjunto de plasma humano de todos los participantes en el estudio sirvió como réplica técnica en todo el conjunto de datos (QC). La variabilidad general del proceso se determinó calculando la desviación estándar relativa mediana para todos los metabolitos endógenos presentes en el 100 % de las muestras de matriz (QC) agrupadas. Las muestras experimentales se aleatorizaron en toda la plataforma con diez muestras de control de calidad al principio (para el equilibrio del instrumento y la columna) y una muestra de control de calidad adquirida cada inyección de quince muestras.

4.2.4. Cromatografía líquida de ultra rendimiento (UPLC): método de espectrometría de masas para análisis lipidómico

El análisis se realizó utilizando un ACQUITY UPLC I-Class (Waters Corp., Milford, MA, EE. UU.) acoplado a un espectrómetro de masas de tiempo de vuelo (ToF) cuadrupolo XEVO-G2 XS (Waters, Manchester, NH, EE. UU.) con una fuente de ionización por electropulverización. Las muestras se analizaron tanto en modo positivo (ESI+) como negativo (ESI-). Para la separación de lípidos se utilizó una columna UPLC CSH C18 (2.1 x 100 mm, 1.7 µm) con una elución de disolventes en gradiente binario. La fase móvil A: 10 mmol de acetato de amonio con 0.1 % de ácido fórmico en acetonitrilo/agua (60:40, v:v) y B: 10 mmol de acetato de amonio con 0.1 % de ácido fórmico en isopropanol/acetonitrilo (90:10, v:v). Las fases móviles se administraron a un caudal de 0.4 ml/min, inicialmente con 60 % A, seguido de un gradiente lineal hasta 57 % A durante 2 min, y luego el porcentaje de A se redujo al 50 % en 0.1 min. Durante los siguientes 9.9 minutos, el gradiente aumentó hasta 46 % de A, y luego la cantidad de A se redujo hasta 30 % en 0.1 min. Durante 5.9 min, la cantidad de A disminuyó al 1 % y

volvió a las condiciones iniciales (60 %) al final de los 20 min. La temperatura de la columna se ajustó a 55°C y el volumen de inyección fue 5 uL. Los datos se adquirieron utilizando el modo de ionización por electropulverización positiva con el voltaje capilar establecido en 3.2 kV, el voltaje del cono en 40 eV y la temperatura de la fuente en 120 °C. El gas de desolvatación fue nitrógeno, con un caudal de 900 L/h y una temperatura de 550 °C. Los datos se adquirieron de m/z 50-1200 en modo MSE en el que la energía de colisión se alternaba entre baja energía (6 eV) y alta energía (aumentada de 15 a 40 eV).

4.2.5. Análisis de datos

Los datos sin procesar se procesaron bajo parámetros predeterminados como un archivo UNIFI, que se utilizó para exportarlo a Progenesis IQ (versión. Waters, Milford). Para la alineación, se excluyeron los tiempos de retención inferiores a 1 min y después de 14 min, debido a interferencias con los picos de extracción del blanco. Se definió un pico de 0.06 s. La deconvolución se realizó automáticamente, considerando como aductos M+H, M+Na y M+NH4. Sin embargo, se realizó una inspección manual, eliminando aquellas características con alineación incorrecta en cromatogramas y masa neutra. Se exportó un archivo Excel y se calculó una señal de ruido para cada muestra en función del blanco de extracción. Se eliminaron todas las características con una relación señal/ruido $(S/N) < 3$ en el 80 % de las muestras. Además, la RSD se calculó tomando como referencia los QC. También se eliminaron las características con $RSD > 20\,\%$. Finalmente, se realizó una inspección manual sobre la supuesta identificación mediante la búsqueda de masa precisa en HMDB, LipidBlast y METLIN. Se eliminaron aquellas características sin identificación putativa. La identificación putativa se asignó según el tiempo de retención, la masa exacta y el patrón de fragmentación.

4.2.6. Análisis estadístico

El análisis estadístico se realizó con MetaboAnalyst 5.0. A continuación se considera cada caso.

4.2.7. Conjunto de datos IMSS

El conjunto de datos utilizado en este estudio fue proporcionado por la Unidad de Investigación Biomédica ubicada en Zacatecas, México, que está incorporada al IMSS. Todos los pacientes mexicanos firmaron una carta de consentimiento informado y los datos incluidos en el conjunto de datos del IMSS cumplen con

Tabla 4.5: Criterios de inclusión.

Criterios de inclusión
1. La edad de los pacientes debe ser mayor de 18 años.
2. No habrá distinción de género, educación, etnia, raza, estado civil.
3. Los conjuntos de datos deben contener únicamente la metabolómica de cada tema.
4. El conjunto de datos debe distinguir los controles de: prediabetes, DMT2 y DN.
5. Los datos de cada característica en cada materia deberán estar completos.

la aprobación del dictamen R-2017-785-131, según el protocolo "Análisis metabolómico y transcriptómico diferencial en orina y suero de pacientes prediabéticos, diabéticos y con nefropatía diabética para identificar potenciales biomarcadores pronósticos de daño renal", que cumple con los criterios aprobados por el Comité Nacional de Investigación Científica y Ética y sigue los estándares éticos internacionales de la convención de Helsinki para estudios de investigación en humanos. El conjunto de datos del IMSS cuenta con información de pruebas metabolómicas, antropométricas, clínicas y de laboratorio. Estas evaluaciones se pueden combinar para medir la progresión de la prediabetes, la diabetes y la nefropatía diabética.

4.2.8. Inclusión de datos

El conjunto de datos del IMSS incluye 375 pacientes y 842 características. En este conjunto de datos, se aplicó un filtro para seleccionar solo los pacientes y las características que cumplían con los criterios de inclusión indicados en la tabla 4.5. El conjunto de datos filtrado resultante, después de aplicar los criterios de inclusión enumerados en Table 4.5, contiene información correspondiente a 79 pacientes (41 mujeres/ 38 hombres), edad (52.34 ± 10.45), metabolómica y diagnóstico (19 pacientes positivos para prediabetes, 20 pacientes positivos para DMT2, 20 pacientes positivos para DN y 20 pacientes de control).

4.2.9. Normalización de los datos

La normalización implementada en este estudio es la PQN y es la siguiente: para cada función, la media de salida se calcula sobre todas las muestras. Luego se genera un vector de referencia. Se calcula la mediana entre el vector de referencia resultante y cada muestra, obteniendo un vector de coeficientes

relacionados. Luego cada muestra se divide por el valor medio del vector de coeficientes, este valor medio es diferente para cada muestra. El propósito de PQN es tener en cuenta los cambios de concentración de algunas características de los metabolitos que afectan regiones limitadas de los datos. El enfoque PQN supone que los cambios en las concentraciones de analitos individuales influyen sólo en partes del espectro, mientras que los cambios en la concentración general de una muestra influyen en todo el espectro. A diferencia de la normalización integral, que supone que la integral total, que cubre todas las señales, es función únicamente de la dilución, PQN supone que la intensidad de la mayoría de las señales es función únicamente de la dilución. Por lo tanto, se calcula como factor de normalización un cociente más probable entre las señales del espectro correspondiente y un espectro de referencia, que reemplaza a la integral total como marcador de la concentración de la muestra. Este cociente más probable para un espectro específico se puede derivar de la distribución de señales de un espectro dividida por la señal correspondiente de un espectro de referencia (Dieterle *et al.*, 2006).

$$I(i) = \frac{I^{old}(i)}{\sum_k (\int_{j_k^l}^{j_k^u} (I(x))^n \mathrm{d}x)^{\frac{1}{n}}} \tag{4.2}$$

donde $I^{old}(i)$ y $I(i)$ son las intensidades de la variable i que es característica espectral, longitud de onda, bin y desplazamiento químico, antes y después de la normalización, respectivamente, k es un índice de las regiones espectrales utilizadas para la normalización, j_k^l y j_k^u son los límites inferior y superior, respectivamente, de la normalización región k, para la cual se integran las potencias n de las intensidades $I(x)$ (Dieterle *et al.*, 2006).

4.2.10. Selección de características

El conjunto de datos utilizado para realizar este estudio tiene más de 700 metabolitos diferentes, cada uno de los cuales tiene un significado potencial para convertirse en un biomarcador o parte de él para resolver un problema de clasificación, sin embargo, esta tarea podría convertirse en un proceso complejo y costoso desde el punto de vista computacional. Con algoritmos genéticos esta compleja tarea se puede realizar y resolver. GALGO (V. and F., 2006) es un GA implementado en este estudio, como paquete de R el software utilizado para realizar las selecciones de características en los 5 conjuntos de este artículo (control-prediabetes, control-T2DM, prediabetes-T2DM, control-DN y T2DM-DN). Para este estudio, GA crea una población inicial de cromosomas constituida por conjuntos aleatorios de metabolitos. El ajuste de los cromosomas se evalúa comparando su capacidad para detectar correctamente cada etapa en la progresión

de la DMT2 ($control \rightarrow prediabetes \rightarrow DMT2 \rightarrow DN$). Dependiendo de la puntuación de ajuste obtenida, la población de cromosomas continúa replicándose y los cromosomas se cruzan y mutan, ya que los cromosomas más aptos producirán descendencia de la próxima generación. El proceso solo se detiene cuando cumple con el criterio objetivo (en este estudio se establece en 1) o cuando los bigbangs (iteraciones) han alcanzado el límite propuesto (3 veces el número de metabolitos en este caso redondeado a 2300 en todas las implementaciones de GALGO). Luego, el resultado de GA Blast (la implementación del modelo GALGO) se envía a un proceso de selección directa para obtener el modelo con mejor rendimiento (podría ser uno o más), luego este modelo o conjunto de características está listo para usarse en un modelo ML. La selección directa se utiliza ampliamente en algoritmos genéticos como complemento para presentar el mejor resultado posible del modelo de las implementaciones de GALGO como ejemplo: en Alzheimer (Sánchez-Reyna *et al.*, 2021), en COVID-19 (Celaya-Padilla *et al.*, 2021) o Retinopatía diabética (Shen *et al.*, 2021).

GALGO permite el uso de diferentes criterios o parámetros de modelo, en este estudio, KNN, Nearcent y SVM se configuraron como se muestra en la tabla 4.6.

4.2.11. Desarrollo de los modelos

Una vez seleccionadas las características, se inicia el desarrollo de los modelos con partición de datos, este proceso es para asegurar que los resultados del entrenamiento y las pruebas sean lo más precisos posible, evitando sobreajustes o desajustes, los modelos se emparejaron con los utilizados en las implementaciones de GALGO, de esta manera la selección será consistente con los resultados de las implementaciones de ML como se presenta en la tabla 4.7.

4.3. CASO 3: Configuración de rangos de valores por sexo

La metodología para este estudio comprende seis etapas, representadas visualmente en la figura 4.5, y se explica a continuación. En la etapa inicial, se expone la utilización del conjunto de datos del IMSS Siglo XXI (representado en la figura 4.5, Etapa 1). Pasando a la segunda etapa, profundizamos en un análisis de datos preliminar que implica la selección de sujetos en función de criterios de inclusión específicos, indicado por la figura 4.5, Etapa 2. Posteriormente, los datos dentro del conjunto de datos se clasifican en tres grupos distintos, a saber, Control-Diabetes en general, Control-Diabetes Hombres y Control-Diabetes Mujeres, como se ilustra en la figura 4.5, Etapa 3. La cuarta etapa involucra la

Tabla 4.6: Parámetros de GALGO.

Modelo	**Parámetro**	**Valor**
KNN	*classification.method*	knn
	chromosomeSize	5
	maxSolutions	2300
	maxGenerations	200
	goalFitness	1
Nearest Centroid	*classification.method*	nearcent
	chromosomeSize	5
	maxSolutions	2300
	maxGenerations	200
	goalFitness	1
SVM	*classification.method*	svm
	svm.kernel	radial
	chromosomeSize	5
	maxSolutions	2300
	maxGenerations	200
	goalFitness	1

implementación de selección de características a través de RFE, LASSO y GA con GALGO y FS, como se detalla en la figura 4.5, Etapa 4. La quinta etapa explica el desarrollo de los modelos ML, incluidos SVM, KNN y Nearcent, LR. , y ANN, utilizando las características principales de las etapas anteriores (ver figura 4.5, Etapa 5). Finalmente, para evaluar el desempeño de nuestros modelos, se elabora la validación de los modelos, teniendo en cuenta varias métricas como *Precision*, *Sensitivity*, *Specificity*, *F1-Score* y AUC, como se representa en la figura 4.5, etapa 6 (Morgan-Benita *et al.*, 2024).

El diagrama de flujo que ilustra nuestra metodología propuesta se presenta a continuación. Los cuadrados azules representan el proceso de análisis de datos, mientras que los cuadrados blancos describen las tareas específicas dentro de cada paso. Etapa 1: Para comenzar se adquiere el conjunto de datos antropométricos y clínicos del Hospital Siglo XXI. Etapa 2: El conjunto de datos se

Tabla 4.7: Desarrollo de los modelos.

Sub-conjunto de datos	**Modelo GALGO**	**Modelo ML**
Control-prediabetes	knn	K-Nearest Neighbours
	nearcent	Nearest Centroid
	svm	Support Vector Machines
Control-T2DM	knn	K-Nearest Neighbours
	nearcent	Nearest Centroid
	svm	Support Vector Machines
Prediabetes-T2DM	knn	K-Nearest Neighbours
	nearcent	Nearest Centroid
	svm	Support Vector Machines
Control-DN	knn	K-Nearest Neighbours
	nearcent	Nearest Centroid
	svm	Support Vector Machines
T2DM-DN	knn	K-Nearest Neighbours
	nearcent	Nearest Centroid
	svm	Support Vector Machines

somete a análisis, lo que da como resultado la creación de nuevos conjuntos de datos, se eliminaron un conjunto de características presentadas en la tabla 4.1 logradas mediante la selección de sujetos con base en los criterios descritos en la tabla 4.5 y las características finales incluidas en el estudio que se muestran en la tabla 4.2. Etapa 3: Posteriormente, las observaciones del conjunto de datos se segmentan en tres grupos distintos, a saber, Control-Diabetes en general, Control-Diabetes masculinos y Control-Diabetes femeninas. Etapa 4: GA y FS se combinan para extraer características de datos esenciales, luego LASSO y RFE se comparan con la implementación de AIC para proporcionar el mejor resultado de modelo posible para determinar el modelo con el mayor rendimiento. Etapa 5: Para aprovechar las características clave para la detección de diabetes y sus diferencias entre pacientes masculinos y femeninos, se generaron múltiples modelos implementando las técnicas SVM, KNN, LR, ANN y Nearcent incluidas en el modelo de conjunto. Etapa 6: El proceso de validación implicó el uso de varias métricas, incluida la validación cruzada a través de GALGO y la ACC promedio

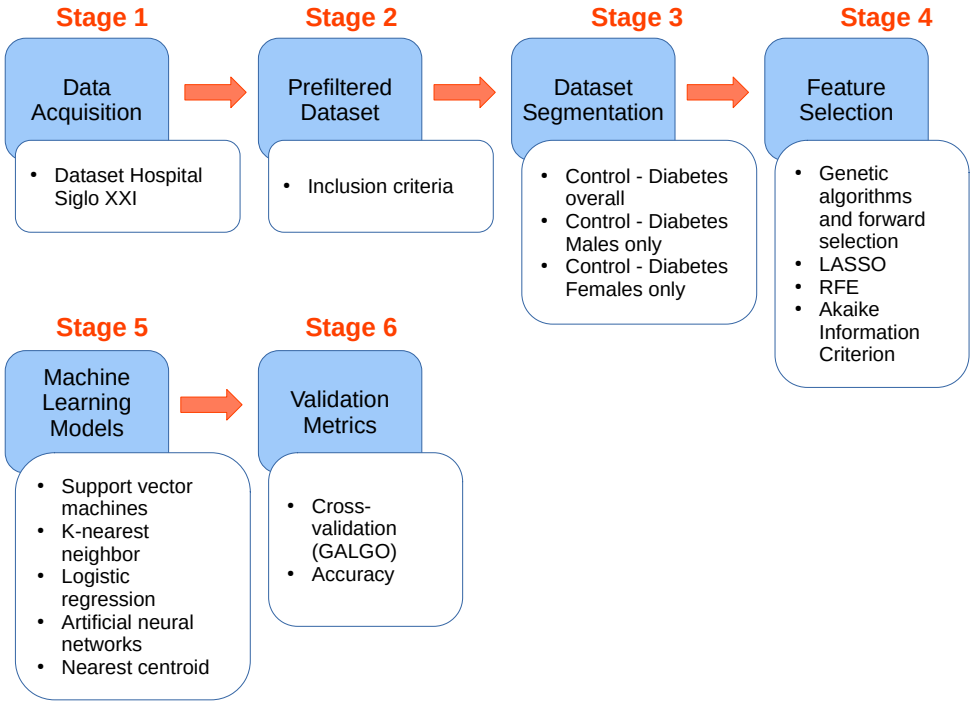


Figura 4.5: Metodología propuesta en el caso de estudio 3.

y la validación cruzada de 10 veces para todos los modelos implementados, con una separación del 70 % para entrenamiento y del 30 % para pruebas. El diagrama de flujo completo se presenta en la figura 4.5.

4.3.1. Muestra

El conjunto de datos se originó en el Centro Médico Nacional Siglo XXI en la Ciudad de México. Todos los participantes mexicanos dieron su consentimiento mediante la firma de una carta de consentimiento informado. El protocolo del estudio cumplió con los criterios de Helsinki y recibió la aprobación del Comité de Ética del Instituto Mexicano del Seguro Social bajo el número R-2011-785-018. El conjunto de datos incluye un total de 1726 personas, con 887 casos de DMT2 y 839 sujetos de control. En términos de distribución por sexo, el conjunto de datos abarca 855 hombres y 871 mujeres. El conjunto de datos contiene 41 características, incluidos datos antropométricos, clínicos y de identificación.

4.3.2. Tratamiento de datos

En esta fase del proceso, los datos se sometieron a tratamiento, comenzando con la imputación de datos para abordar los valores faltantes y las características más allá del alcance del estudio. Posteriormente, los valores imputados fueron incorporados a la base de datos. El último paso implicó normalizar los datos.

4.3.3. Imputación de datos

Este estudio implementó un criterio de exclusión, trabajando exclusivamente con observaciones que presentaban datos completos para todas las variables y descartando otras con características faltantes o irrelevantes. Se excluyeron las características que podrían identificar a los pacientes, como las identificaciones, ya que solo representaban datos internos de identificación del paciente y números consecutivos. Se eliminaron los datos nulos o no disponibles (NA), particularmente para las características asociadas exclusivamente con casos positivos de DMT2, que muestran datos inexistentes en pacientes de control o casos en los que no se registraron datos.

En consecuencia, se eliminaron todos los medicamentos y sus cantidades de ingesta diaria (glibenclamida, dosis de glibenclamida, metformina, dosis de metformina, pioglitazona, dosis de pioglitazona, rosiglitazona, dosis de rosiglitazona, acarbosa, dosis de acarbosa, insulina y dosis de insulina). Además, se excluyeron los datos de tres pacientes con valores faltantes para la característica "hipertensión bajo tratamiento". También se eliminaron características como la TFG, la edad de diagnóstico (específicamente para los casos de DMT2 positivo), la HbA1c y las complicaciones de la DMT2, dado que estos datos estaban presentes exclusivamente para los pacientes con DMT2 positivo.

Los pacientes que tuvieron valores negativos en "WHR", "BMI", "Creatinineƨ "HDLc"fueron eliminados para preservar la integridad de los valores.

La característica de complicaciones de la DMT2, que contiene comorbilidades asociadas con la DMT2, se excluyó por estar fuera del alcance del experimento. Contenía valores de NA en algunos casos y era completamente NA para todos los pacientes de control.

La característica de la glucosa, un biomarcador muy conocido, también fue eliminada del alcance del experimento. El motivo de su exclusión fue evaluar mejor el rendimiento de otras características, teniendo en cuenta que su alto AUC, superior al 90 % en pruebas de modelos univariados, ya significa su prominencia en el conjunto de datos analizado.

Las características de salario y educación también fueron eliminadas del alcance de este estudio.

4.3.4. Selección de características

El conjunto de datos empleado en esta investigación comprendió una gama diversa de 1726 pacientes distintos y se dividió nuevamente en conjuntos de datos masculinos y femeninos, proporcionando diferentes características y creando subconjuntos que potencialmente sirvieron como biomarcadores o ayudaron a abordar los desafíos de clasificación. Sin embargo, gestionar esta tarea resultó computacionalmente exigente y complejo. Para abordar esta complejidad, empleamos un GA utilizando el paquete GALGO R 1.4 (V. and F., 2006). GALGO facilita la selección de características para categorizar individuos como casos positivos de T2DM o DN.

En nuestra metodología, se utilizó GALGO para inicializar una población de cromosomas, cada uno de los cuales contenía conjuntos de características seleccionados al azar, el ajuste de los cromosomas se evaluó en función de su eficacia para categorizar con precisión a los individuos en sus respectivas etapas de DMT2 o DN según sus puntuaciones de ajuste, cruce, mutación o contribución a la generación de la siguiente descendencia. Este proceso iterativo continuó hasta que se cumplieron condiciones específicas, como lograr el objetivo predefinido (establecido en 1 en este estudio) o alcanzar el límite de iteración predeterminado, en este caso dos veces el número de observaciones (redondeado a 2000) para todas las implementaciones de GALGO en el conjunto de datos completo (1726 pacientes) y en los subconjuntos de datos separados por hombres y mujeres.

Después de la ejecución de GALGO, el resultado se sometió a un procedimiento FS para identificar el modelo o conjunto de características de mejor rendimiento. Luego, este modelo o conjunto de características seleccionado se preparó para su integración en un modelo ML. FS, una técnica empleada a menudo en los AG, se empleó para mejorar el resultado de GALGO con el fin de presentar el modelo más óptimo.

GALGO proporcionó flexibilidad en cuanto a los criterios o parámetros del modelo. En este estudio, lo configuramos para acomodar KNN, Nearcent, ANN, LR y SVM, como se especifica en la tabla correspondiente en el conjunto de datos completo de Siglo XXI (utilizando 1726 pacientes en total) (Table 4.11).

Con los mismos modelos de validación y utilizando los mismos criterios, este estudio también utilizó GALGO configurado para acomodar KNN, Nearcent, ANN, LR y SVM, como se especifica en las tablas respectivas en los conjuntos de datos Siglo XXI Masculino (855 pacientes masculinos) y Femenino (871 pacientes femeninos) (tabla 4.12).

Otra técnica implementada después de las ejecuciones de GALGO fue la técnica FS. El objetivo principal es identificar las características más pertinentes e informativas que mejoran el rendimiento del modelo y al mismo tiempo mitigan

la complejidad y el riesgo de sobreajuste.

El proceso comienza con la inicialización de un conjunto de características vacío o un conjunto mínimo de características consideradas relevantes según el conocimiento previo o la experiencia en el dominio. Posteriormente, se realiza una evaluación de características que implica entrenar un modelo utilizando las características seleccionadas y medir su rendimiento utilizando métricas de evaluación elegidas, como *Precision*, *F1-Score*, AUC u otras en un conjunto de datos de validación o validación cruzada.

El proceso de selección de características contempla entonces la adición de una característica del conjunto restante de características al conjunto actual. La selección puede basarse en varios criterios, incluida la correlación con la variable objetivo, valores p de pruebas estadísticas u otros factores relevantes. Luego se reevalúa el modelo, se entrena un nuevo modelo utilizando el conjunto ampliado de características y se evalúa su rendimiento en el mismo conjunto de datos de validación o validación cruzada.

Para determinar si se mantiene o excluye la característica agregada, se realiza una comparación entre el rendimiento del nuevo modelo con la característica agregada y el modelo anterior sin la característica de acuerdo con la métrica de evaluación elegida. Si el rendimiento muestra una mejora, entonces la característica se conserva en el conjunto, en caso contrario, se elimina. Este proceso se repite iterativamente a lo largo de los pasos 3 a 5 hasta que se cumple un criterio de parada predefinido. Este criterio de parada puede basarse en factores tales como un número predeterminado de características a seleccionar o un nivel específico de rendimiento del modelo deseado. Cuando se cumple el criterio de parada, el conjunto final de características seleccionadas se emplea en la construcción del modelo (García-Domínguez *et al.*, 2023).

LASSO fue implementado con el objetivo resolver un problema de optimización minimizando una función de pérdida que cuantifica el error entre los resultados previstos y los resultados reales en un conjunto de datos. La función objetivo se define de la siguiente manera:

$$L(\beta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}|\ beta_j| \tag{4.3}$$

donde N representa el número de muestras, y_i es la variable objetivo (estado T2DM) para la i-ésima muestra, X_i representa el vector de características para i-ésima muestra, que incluye p características, β es un vector de coeficientes asociados con cada característica y λ es el parámetro de regularización, que controla la intensidad de la penalización.

La característica clave de LASSO es la adición del término de regularización L1 $\lambda\sum_{j=1}^{p}|\beta_j|$ a la función de pérdida. Este término fomenta que los coeficientes

(β) sean escasos, lo que significa que obliga a que muchos coeficientes sean exactamente cero. Como resultado, LASSO promueve la selección de características eliminando eficazmente características irrelevantes o menos importantes. El problema de optimización se resuelve para encontrar los valores de los coeficientes (β) que minimizan la función objetivo. Este proceso implica ajustar los valores de β para minimizar el término de error y al mismo tiempo aplicar el término de penalización. Como resultado de la regularización L1, algunos coeficientes se reducen a cero durante la optimización, seleccionando efectivamente un subconjunto de las características más relevantes. Estas características seleccionadas se consideran las más influyentes para la detección de DMT2 (Ou *et al.*, 2023a).

por último se realizó una implementación de RFE incluyendo el AIC y una incorporación de la métrica de *accuracy* como parámetro primario de evaluación de cada conjunto de características proporcionado por RFE. La idea de esta implmenetación es identificar el mejor conjunto de características tomando en cuenta su tamaño con AIC y evaluando el rendimiento de cada modelo a través de la métrica, comparando cada resultado y seleccionando solo el mejor conjunto resultante en cada iteración.

El proceso ejecutado fue el siguiente:

- Paso 1: Cargar Bibliotecas y Conjunto de Datos
  - Importar las bibliotecas necesarias para la manipulación de datos, selección de características y modelado.
  - Cargar el conjunto de datos y almacenarlo en `df`.
  - Separar `df` en la matriz de características X (todas las columnas excepto `DX.bl`) y la variable objetivo y (la columna `DX.bl`).
- Paso 2: Dividir y Escalar los Datos
  - Dividir X y y en conjuntos de entrenamiento (70 %) y prueba (30 %).
  - Estandarizar las características utilizando `StandardScaler` para que tengan media cero y varianza unitaria.
- Paso 3: Inicializar Seguimiento de Mejores Modelos
  - Crear variables de seguimiento (`best_aic`, `best_selected_features`, `best_accuracy`, `best_model`) para cada modelo (Regresión Logística, SVM, Bosque Aleatorio) para almacenar valores óptimos.
- Paso 4: Selección de Características con RFE-CV

- Definir un bucle para reducir el número de características de manera iterativa, comenzando con todas las características y reduciendo hasta un mínimo de 4 características.
- Para cada modelo:
  - Regresión Logística con RFE-CV
    - ◇ Inicializar `LogisticRegression` con un límite de máximo de iteraciones.
    - ◇ Realizar RFE-CV usando `RFECV` en X y y.
    - ◇ Recuperar las características seleccionadas del arreglo de soporte.
    - ◇ Calcular el Criterio de Información de Akaike (AIC):

$$\text{AIC} = 2k - 2 \times \text{Verosimilitud} \tag{4.4}$$

      donde k es el número de características $+1$ (por el término de intercepción).
    - ◇ Evaluar la precisión en el conjunto de prueba.
    - ◇ Si el AIC y la precisión son mejores que los registros anteriores, almacenar los valores actualizados.
  - Máquina de Vectores de Soporte (SVM) con RFE-CV
    - ◇ Inicializar `SVR` con un kernel lineal y realizar RFE-CV.
    - ◇ Recuperar las características seleccionadas.
    - ◇ Calcular el AIC usando la puntuación del modelo en el entrenamiento.
    - ◇ Evaluar la precisión en el conjunto de prueba.
    - ◇ Si los nuevos valores de AIC y precisión son mejores, almacenarlos.
  - Bosque Aleatorio (RF) con RFE-CV
    - ◇ Inicializar `RandomForestClassifier` y realizar RFE-CV.
    - ◇ Recuperar las características seleccionadas.
    - ◇ Calcular el AIC usando su puntuación en el entrenamiento.
    - ◇ Evaluar la precisión en el conjunto de prueba.
    - ◇ Si los nuevos valores de AIC y precisión son mejores, almacenarlos.

- Paso 5: Imprimir Resultados Finales
  - Imprimir las mejores características seleccionadas, AIC y precisión para cada modelo.

Tabla 4.8: Características eliminadas.

Característica	**Descripción**
plate_info	ID del paciente del hospital
ID	Número de identificación consecutivo en el conjunto de datos
Educativo	Nivel educativo
Sal	Salario
Edad DX (casos)	Años con enfermedad diabética diagnosticada
GLU (mg/dL)	Niveles de glucosa
HbA1c	Hemoglobina glicosilada
GLIBENCLAMIDA	Si el paciente tiene tratamiento con glibenclamida
GLIBEN_MG_DIA	Glibenclamida recetada en miligramos
METFORMINA	Si el paciente recibe tratamiento con metformina
METFOR_MG_DIA	Metformina recetada en miligramos
PIOGLITAZONA	Si el paciente tiene tratamiento con pioglitazona
PIOGLI_MG_DIA	Pioglitazona prescritas en miligramos
ROSIGLITAZONA	Si el paciente tiene tratamiento con rosiglitazona
ROSIGLI_MG_DIA	Rosiglitazona prescrita en miligramos
ACARBOSA	Si el paciente tiene tratamiento con Acarbosa
ACARBO_MG_DIA	Acarbosa recetada en miligramos
INSULINA	Si el paciente tiene tratamiento con insulina
INSUL_UI_DIA	Insulina recetada en miligramos
TIPO_COMPLICACION DE DT2	T2DM complicaciones

Tabla 4.9: Criterios de inclusión.

1. Los pacientes deben tener al menos 18 años.
2. No deberá haber diferenciación en los datos obtenidos en función del sexo, educación, origen étnico, raza o estado civil.
3. Los conjuntos de datos deben comprender exclusivamente datos antropométricos y clínicos de cada individuo.
4. El conjunto de datos debe ser capaz de distinguir sujetos control de aquellos con T2DM.
5. El conjunto de datos de cada sujeto debe incluir información completa para todas las características.
6. Los datos no contienen valores negativos.
7. No contiene características relacionadas con la glucosa.

Tabla 4.10: Características utilizadas en la experimientación.

Característica	**Descripción**
Sex	Sexo de los pacientes
Age	Edad del paciente en años
WHR	Relación cintura-cadera
BMI	Índice de masa corporal
Urea	Producto de desecho resultante de la degradación de proteínas en el cuerpo del paciente
Creatinine	Producto de desecho producido por los músculos como parte de actividad diaria regular
Lipid treatment	Niveles de lípidos en tratamiento
Cholesterol	Sustancia parecida a la grasa que se encuentra en todas las células del cuerpo del paciente
HDL	Lipoproteínas de alta densidad (corregido por medicación)
LDL	Lipoproteínas de baja densidad (corregido por medicación)
Triglycerides	Tipo de grasa que se encuentra en el cuerpo del paciente
TCHOLU	Colesterol total (no corregido con medicación)
HDLU	Lipoproteínas de alta densidad (no corregidas con medicamentos)
LDLU	Lipoproteínas de baja densidad (no corregidas con medicamentos)
TGU	Triglicéridos (no corregidos con medicación)
SBP	Presión arterial sistólica (corregida por medicación)
DBP	Presión arterial diastólica (corregida por medicación)
SBPU	Presión arterial sistólica (no corregida con medicamentos)
DBPU	Presión arterial diastólica (no corregida con medicación)
HA-TX	Sujeto bajo tratamiento de hipertensión
LIPIDS-TX	Sujeto bajo tratamiento con lípidos

Tabla 4.11: Parámetros de GALGO en el conjunto de datos Siglo XXI overall.

Modelo	**Parámetro**	**Valor**
KNN	*classification.method*	'knn'
	chromosomeSize	5
	maxSolutions	2000
	maxGenerations	60
	goalFitness	0.9
Nearcent	*classification.method*	'nearcent'
	chromosomeSize	5
	maxSolutions	2000
	maxGenerations	60
	goalFitness	0.9
Artificial Neural Network	*classification.method*	'nnet'
	chromosomeSize	5
	maxSolutions	2000
	maxGenerations	60
	goalFitness	0.9
Logistic Regression	*classification.method*	'user'
	classification.userFitnessFunc	logreg.R.predict
	chromosomeSize	5
	maxSolutions	2000
	maxGenerations	60
	goalFitness	0.9
Support Vector Machines	*classification.method*	'svm'
	svm.kernel	'radial'
	chromosomeSize	5
	maxSolutions	2000
	maxGenerations	60
	goalFitness	0.9

Tabla 4.12: Parámetros de GALGO en el conjunto de datos Siglo XXI Masculino/Femenino.

Model	**Parameter**	**Value**
KNN	*classification.method*	'knn'
	chromosomeSize	5
	maxSolutions	1600
	maxGenerations	60
	goalFitness	0.9
Nearcent	*classification.method*	'nearcent'
	chromosomeSize	5
	maxSolutions	1600
	maxGenerations	60
	goalFitness	0.9
Artificial Neural Network	*classification.method*	'nnet'
	chromosomeSize	5
	maxSolutions	1600
	maxGenerations	60
	goalFitness	0.9
Logistic Regression	*classification.method*	'user'
	classification.userFitnessFunc	logreg.R.predict
	chromosomeSize	5
	maxSolutions	1600
	maxGenerations	60
	goalFitness	0.9
Support Vector Machines	*classification.method*	'svm'
	svm.kernel	'radial'
	chromosomeSize	5
	maxSolutions	1600
	maxGenerations	60
	goalFitness	0.9

Capítulo 5
Resultados

Se realizaron 3 casos de estudio con diferentes aproximaciones en 2 bases de datos distintas:

Caso de Estudio 1:

Ensamble por votación y selección de características con LASSO.

Los datos analizados consisten en 1787 pacientes, 898 casos positivos de DMT2 y 889 pacientes control, todos ellos con las características descritas en la tabla 4.1 y la tabla 4.2. El tratamiento de datos consistió en imputación de datos (Imputación de datos) y normalización de datos (Normalización de datos), dando como resultado datos balanceados y normalizados y 10 características descartadas (tabla 4.1). Las 23 características procesadas en LASSO (Selección de características) con penalización de red elástica obtuvieron un conjunto de 12 características para los modelos integrados en el modelo de conjunto, estas características se muestran en la tabla 5.27.

Todos los resultados en las tablas 5.2, 5.4, 5.6 y 5.8 fueron calculados por:

$$\textit{Sensitivity} = \frac{T_p}{(T_p + F_n)}, \tag{5.1}$$

$$\textit{Specificity} = \frac{T_n}{(F_p + T_n)}, \tag{5.2}$$

$$\textit{Precision} = \frac{T_p}{(T_p + F_p)}, \tag{5.3}$$

$$\textit{Negative Predictive Value} = \frac{T_n}{(T_n + F_n)}, \tag{5.4}$$

$$\textit{False Positive Rate} = \frac{F_p}{(F_p + T_n)}, \tag{5.5}$$

$$\textit{False Negative Rate} = \frac{F_n}{(F_n + T_p)}, \tag{5.6}$$

$$Accuracy = \frac{(T_p + T_n)}{(T_p + T_n + F_p + F_n)}, \tag{5.7}$$

$$F1\text{-}Score = \frac{2T_p}{(2T_p + F_p + F_n)} \tag{5.8}$$

donde:

T_p =Verdadero positivo, número de sujetos con DMT2 correctamente clasificados.
F_p =Falso positivo, número de sujetos sanos clasificados incorrectamente.
T_n =Verdadero negativo, número de sujetos sanos clasificados correctamente.
F_n =Falso negativo, número de sujetos con DMT2 clasificados como sanos.

Estas métricas proporcionan cuál de los modelos implementados es mejor para identificar a los pacientes con DMT2.

Tabla 5.1: Estructura de la matriz de confusión.

Valores verdaderos	**Predicho (verdadero)**	**Predicho (falso)**
Verdadero	T_p	T_n
Falso	F_p	F_n

Los modelos desarrollados mostraron un buen desempeño en la implementación del conjunto, el modelo SVM con kernel radial tuvo un AUC de 92.8 % ± 3 % con el conjunto de prueba de 25 %, en la matriz de confusión del En el modelo SVM, se observa que la *Sensitivity* es de 0.8750 (87.5 %) y es menor que la *Specificity* de 0.9238 (92.38 %) como se presenta en en la tabla 5.3.

El modelo de ANN de una sola capa tuvo un AUC de 90.5 % ± 3 % con el conjunto de prueba de 25 %, y en la matriz de confusión del modelo de ANN, se observa que la *Sensitivity* de 0.8559 (85.59 %) es inferior a la *Specificity* de 0.9175 (91.75 %), este segundo modelo muestra la menor *Sensitivity* y *Specificity* en comparación con el modelo SVM como se muestra en la tabla 5.5.

El modelo GLM tuvo un AUC de 90.5 % ± 3 % con el conjunto de pruebas del 25 %, en la matriz de confusión del modelo GLM se observa que la *Sensitivity* de 0.8487 (84.87 %) es menor que la *Specificity* de 0.9167 (91.67 %) como es mostrado en la tabla 5.7. Este modelo preestablece la *Sensitivity* y *Specificity* más bajas de todos los modelos, incluido el conjunto.

El modelo de conjunto con votación máxima tuvo un AUC de 90.5 % ± 3 %, este AUC muestra coherencia con los 3 modelos que integran una solución robusta al problema de clasificación. En la matriz de confusión del modelo de conjunto,

Tabla 5.2: Valores de métrica de la matriz de confusión SVM.

Métrica	**Valor**
Sensitivity	0.8750
Specificity	0.9238
Precision	0.9269
Negative Predictive Value	0.8700
False Positive Rate	0.0762
False Negative Rate	0.1250
Accuracy	0.8982
F1-Score	0.9002

Tabla 5.3: Matriz de confusión de SVM.

Valores verdaderos	**Predicho (Verdadero)**	**Predicho (Falso)**
Verdadero	203	16
Falso	29	194

Tabla 5.4: Resultado en las métricas de la matriz de confusión de ANN.

Measure	**Value**
Sensitivity	0.8559
Specificity	0.9175
Precision	0.9224
Negative Predictive Value	0.8475
False Positive Rate	0.0825
False Negative Rate	0.1441
Accuracy	0.8846
F1-Score	0.8879

Tabla 5.5: Matriz de confusión ANN.

Valores verdaderos	**Predicho (Verdadero)**	**Predicho (Falso)**
Verdadero	202	17
Falso	34	189

Tabla 5.6: Valores de métrica de la matriz de confusión GLM.

Measure	**Value**
Sensitivity	0.8487
Specificity	0.9167
Precision	0.9224
Negative Predictive Value	0.8386
False Positive Rate	0.0833
False Negative Rate	0.1513
Accuracy	0.8801
F1-Score	0.8840

Tabla 5.7: Matriz de confusión GLM.

Valores verdaderos	**Predicho (Verdadero)**	**Predicho (Falso)**
True	202	17
False	36	187

Tabla 5.8: Valores de métrica de la matriz de confusión de Maxvoting Ensemble.

Measure	**Value**
Sensitivity	0.8788
Specificity	0.9242
Precision	0.9269
Negative Predictive Value	0.8744
False Positive Rate	0.0758
False Negative Rate	0.1212
Accuracy	0.9005
F1-Score	0.9022

Tabla 5.9: Matriz de confusión del modelo Ensamble.

Valores verdaderos	**Predicho (Verdadero)**	**Predicho (Falso)**
Verdadero	203	16
Falso	28	195

se verifica que la *Sensitivity* de 0.8788 (87.88 %) es menor que la *Specificity* de 0.9242 (92.42 %) que se muestra en la tabla 5.9, valores por encima del promedio de la *Sensitivity* y *Specificity* de los modelos implementados.

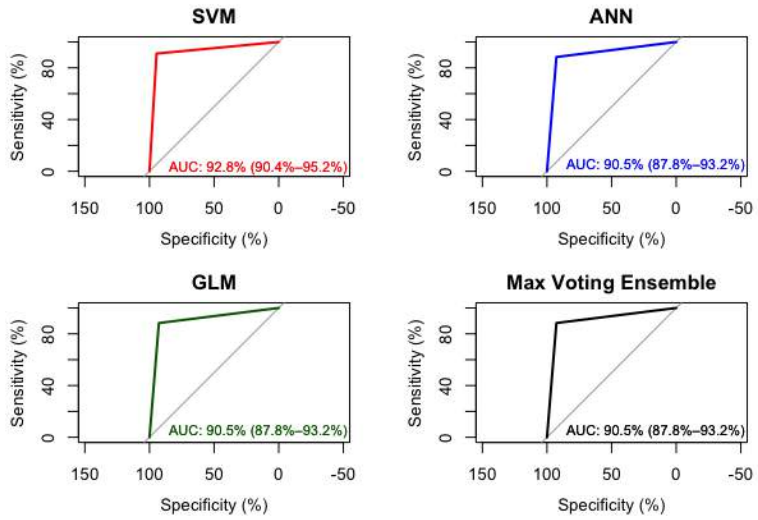


Figura 5.1: AUC in SVM, ANN, GLM y Modelo Ensamble.

El modelo de conjunto mostró un AUC similar a los modelos ANN y GLM con un 90.5% como promedio, siendo el modelo SVM un 2.3% mejor que los otros modelos en este caso, como se muestra en la tabla 5.1.

Caso de estudio 2:

Propuesta de selección de características e identificación metabolómica.

La metodología propuesta (figura 4.4), presenta en 6 pasos el proceso seguido en este estudio: Adquisición de datos, Selección de sujetos, Preprocesamiento de datos, Selección de características, clasificación y validación. Una descripción de la metodología para obtener la muestra, que incluye: preparación de la muestra, controles de calidad (QC) y aseguramiento de la calidad (QA), Cromatografía líquida de ultra rendimiento (UPLC) -Espectrometría de masas 158 Método para análisis lipidómico, análisis de datos y análisis estadístico. La normalización de datos se presenta en la subsección. El conjunto de datos obtenido se describe en la sección de conjunto de datos con los criterios de inclusión presentados en la tabla 4.5, la inclusión de datos se proporciona como en la sección con el mismo nombre. Después de la inclusión de datos, el proceso de selección de características comienza con la implementación de los algoritmos genéticos, se requirieron 15 ejecuciones para obtener cada combinación de las etapas como se muestra en la tabla 4.11. Cada conjunto de características obtenidas en los algoritmos genéticos integran un modelo como se presenta en la tabla 4.7, los modelos son: KNN, Nearcent y SVM. Por último, la implementación en R se presenta en la sección de implementación.

Tabla 5.10: Modelos GALGO - ACC promedio.

Subconjunto de datos	**Modelo GALGO**	**Modelo ML**	**ACC promedio**
Control-prediabetes	knn	K-Nearest Neighbours	0.923
	nearcent	Nearest Centroid	0.925
	svm	Support Vector Machines	0.911
Control-T2DM	knn	K-Nearest Neighbours	0.889
	nearcent	Nearest Centroid	0.932
	svm	Support Vector Machines	0.962
Prediabetes-T2DM	knn	K-Nearest Neighbours	0.743
	nearcent	Nearest Centroid	0.934
	svm	Support Vector Machines	0.873
Control-ND	knn	K-Nearest Neighbours	0.958
	nearcent	Nearest Centroid	0.962
	svm	Support Vector Machines	0.932
T2DM-ND	knn	K-Nearest Neighbours	0.8679
	nearcent	Nearest Centroid	0.926
	svm	Support Vector Machines	0.882

5.0.1. Resultados de GALGO

Las 15 ejecuciones de GALGO con diferentes conjuntos de datos que combinaron las muestras y las compararon proporcionaron una ACC promedio presentada en la tabla 5.10 y un grupo de características presentadas en las siguientes secciones.

Características obtenidas por el método GALGO KNN en el conjunto de datos Control-Prediabetes

Las características obtenidas por GALGO y el mejor modelo de selección directa, en este caso el modelo 2, como se presenta en la figura 5.0.1, con una ACC promedio de 0.923 5.10 se presentan en la tabla 5.11. Derivado de la ba-

ja cantidad de datos frente a la gran cantidad de características incluidas en el modelo, este resultado prueba que los metabolitos resultantes incluidos son extremadamente significativos, como se presenta en la figura de estabilidad de rango genético 5.3 y el ajuste figura 5.4.

FSM representa el rendimiento de los modelos más compactos y precisos después de emplear la metodología de selección directa. El eje vertical muestra la precisión de la clasificación. El eje horizontal representa las características ordenadas, cada una de las cuales corresponde a un número. La línea negra sólida representa los modelos más compactos y precisos. La selección directa muestra que la precisión de la clasificación se muestra en el eje vertical. El eje horizontal representa las características ordenadas, que están representadas por números. El modelo más compacto y preciso está representado por la línea negra continua.

La frecuencia de los genes y la estabilidad del rango de los genes en los modelos se determinan aplicando GA con KNN, para la selección de las características principales en el conjunto de datos. (A) La frecuencia genética muestra el número de veces que una característica ha estado presente en los modelos. (B) El rango genético muestra la estabilidad y la frecuencia de cada característica en los modelos, ordenadas por rango.

Evolución de la puntuación máxima de ajuste a través de generaciones. El eje horizontal representa una generación determinada, mientras que el eje vertical representa la puntuación de ajuste. El ajuste promedio, trazada con una línea continua azul, considera todos los modelos. La ajuste inacabado promedio, trazada con una línea continua de color cian, considera todas las búsquedas que fallaron para una generación determinada y representa la expectativa promedio en el peor de los casos. El ajuste objetivo de GA establecida se traza con la línea de puntos roja.

Características obtenidas por el método GALGO KNN en el conjunto de datos Control-T2DM

Las características obtenidas por GALGO y el mejor modelo de selección directa, en este caso el modelo 3 como se presenta en la figura 5.5, con una ACC promedio de 0.8893 presente en la tabla 5.10 y en la tabla 5.12. Derivado de la baja cantidad de datos frente a la gran cantidad de características incluidas en el modelo, este resultado prueba que los metabolitos resultantes incluidos son extremadamente significativos, como se presenta en la figura de estabilidad de rango genético 5.6 y el ajuste figura 5.7.

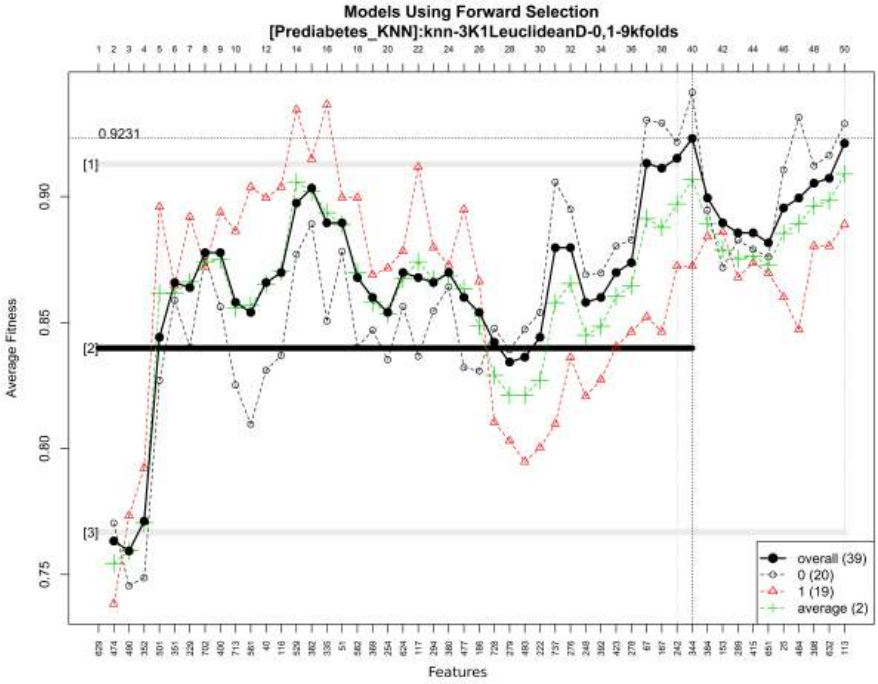


Figura 5.2: Prediabetes FSM-KNN

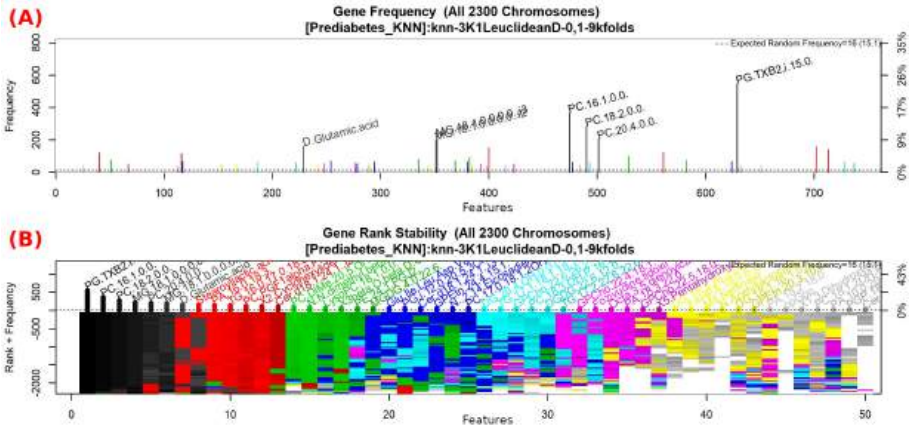


Figura 5.3: Prediabetes Frecuencia-KNN.

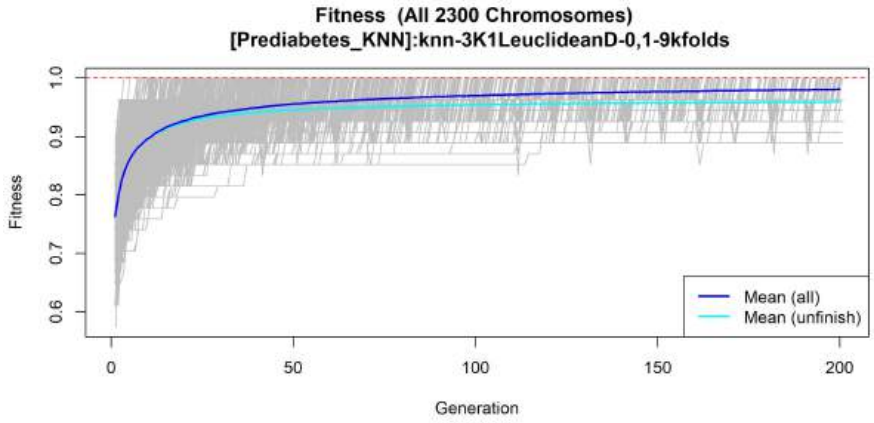


Figura 5.4: Prediabetes Evolución-KNN.

Tabla 5.11: Características obtenidas por el método GALGO KNN en el conjunto de datos Control - Prediabetes.

Características resultantes
PG(TXB2/i-15:0), PC(16:1/0:0), PC(18:2/0:0), MG(18:1/0:0/0:0).i3, PC(20:4/0:0), MG(18:1/0:0/0:0) i2, D-Glutamic acid, Stearoyllactic acid, PA(16:0/18:3), TG(16:0/17:0/18:3), PC(PGF1alpha/P-18:0)., 2-Cyclohexylidenecyclohexanone, Cer(d18:1/24:1) i2, PC(DiMe/20:5-OH), N,N-dimethyl-beta-alanine, LysoPC(18:1/0:0), 3beta-O-beta-D-Glucopyranosiduronic acid (1_2)-beta-D-glucopyranosyloxy]-machaerinic acid_-lactone, PE(24.1/PGJ2), MGDG(4:0/22:6), Glu-Ile-Leu-Asp-Val, PG(i-12:0/20:5-3OH), Cer(d18:1/24:1) i3, GPEtn(24:4/15:1), N-(2,14-Eicosadienoyl)piperidine, PC(17:0/18:1-2OH), DG(20:4-OH/0:0/i-18:0), TG(17:1/18:1/18:1), GPCho.22.6.26.2, PC(18:3-OH/P-16:0) i2, DGDG.22.6.12.0, TG.17.2.22.6.22.6, GPCho.22.4.18.1, Fozivudine.tidoxil, Octa.3.5.dienoylcarnitine, PA.18.3.22.0, GPCho(22:5/18:0), X5.Pentahydroxy.5.cucurbiten.11.one.3..glucosyl..1..6..glucoside., DG.18.3.26.2.0.0, FAHFA.18.0.6.O.16.0, M8.Nelfinavir

Todas las características de esta tabla se obtuvieron con GALGO con una implementación del modelo ML KNN, 2300 Big bangs y 200 generaciones.

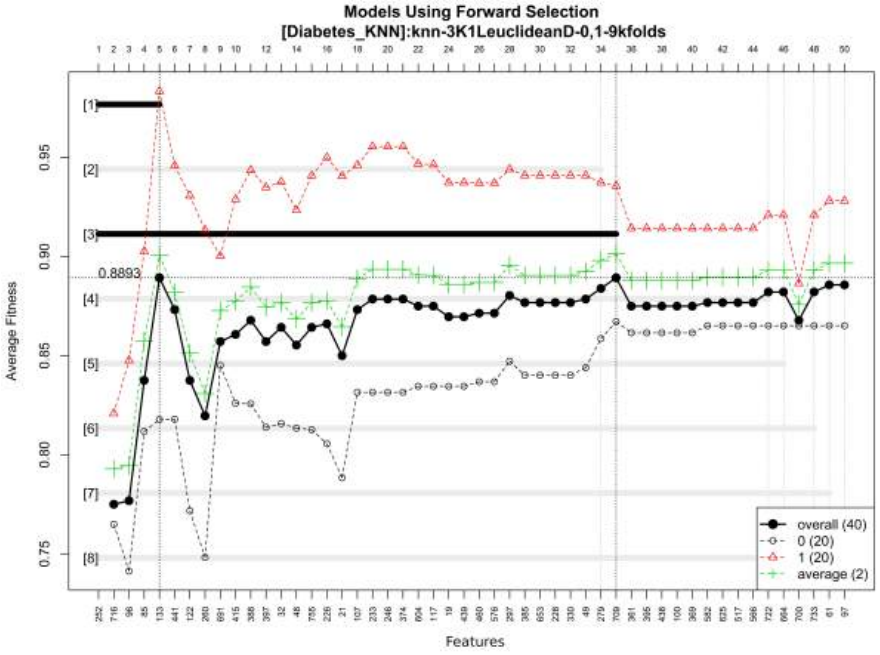


Figura 5.5: Diabetes FSM-KNN

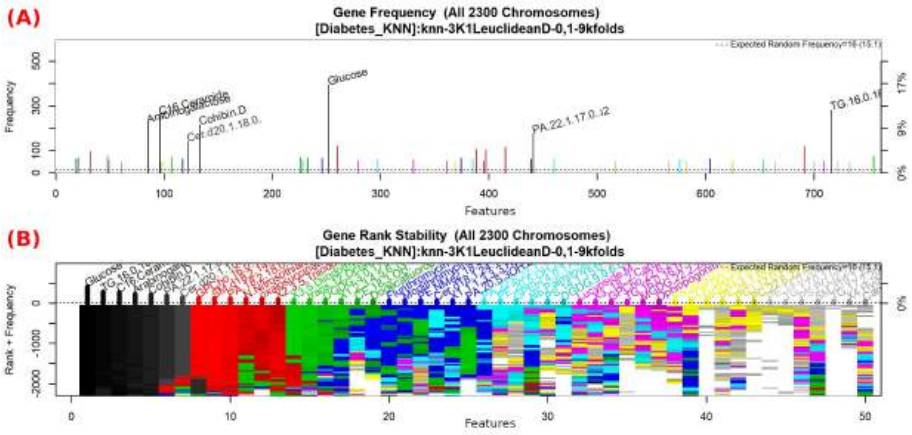


Figura 5.6: Diabetes Frecuencia-KNN.

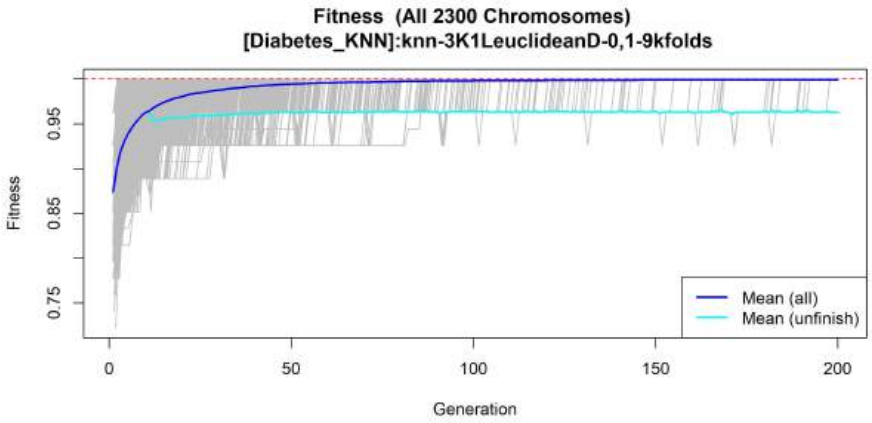


Figura 5.7: Diabetes Evolución-KNN.

Tabla 5.12: Características obtenidas por el método GALGO KNN en el conjunto de datos Control - T2DM.

Características de resultados
Ganoderic acid-C2, TG(16:0/17:1/18:1), butyl methacrylate, cholestan-3-one

Todas las características de esta tabla se obtuvieron con GALGO con una implementación del modelo ML KNN, 2300 Big bangs y 200 generaciones.

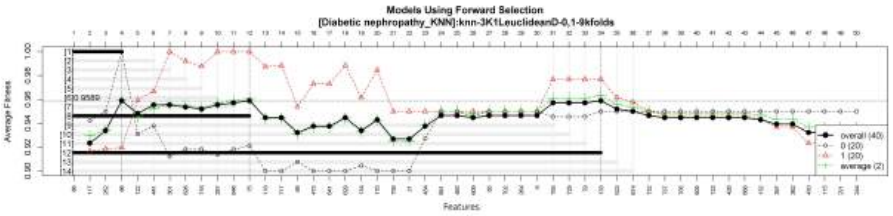


Figura 5.8: Nefropatía Diabética FSM-KNN

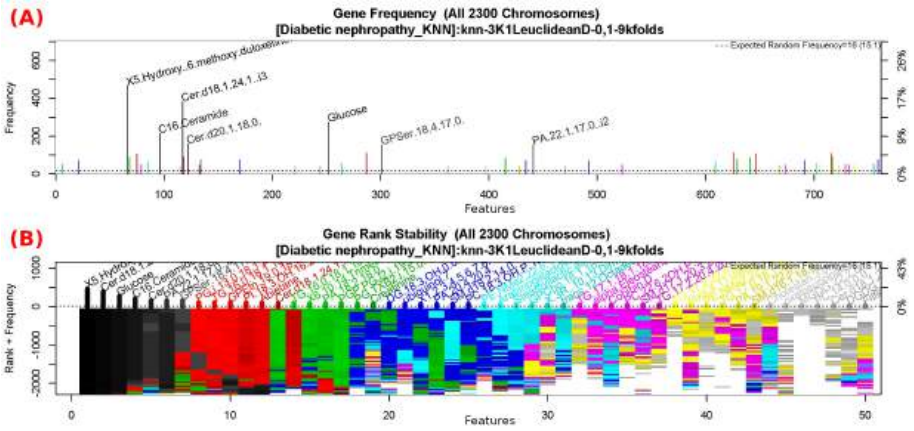


Figura 5.9: Nefropatía Diabética Frecuencia-KNN.

Características obtenidas por el método GALGO KNN en el conjunto de datos de Control-Nefropatía diabética

Las características obtenidas por GALGO y el mejor modelo de selección directa, en este caso el modelo 2 como se presenta en la figura 5.8, con una ACC promedio de 0.958 presente en la tabla 5.10 y en la tabla 5.13. Derivado de la baja cantidad de datos frente a la gran cantidad de características incluidas en el modelo, este resultado prueba que los metabolitos resultantes incluidos son extremadamente significativos, como se presenta en la figura de estabilidad de rango genético 5.9 y el ajuste figura 5.10.

Características obtenidas por el método GALGO KNN en el conjunto de datos Prediabetes-T2DM

Las características obtenidas por GALGO y el mejor modelo de selección directa, en este caso el modelo 2 como se presenta en la figura 5.11, con una ACC promedio de 0.743 presente en la tabla 5.10 y en la tabla 5.14. Derivado de la baja cantidad de datos frente a la gran cantidad de características incluidas en

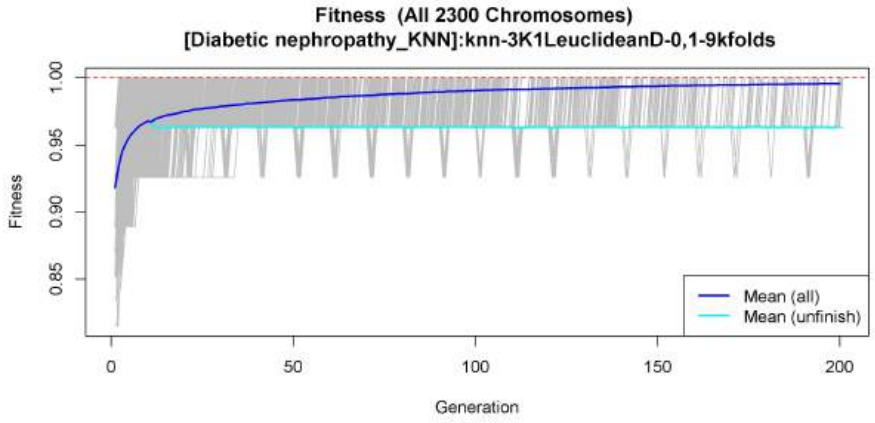


Figura 5.10: Nefropatía Diabética Evolución-KNN.

Tabla 5.13: Características obtenidas por el método GALGO KNN en el conjunto de datos Control - Nefropatía diabética.

Características resultantes
5beta-Cholestanone, Cer(d18:1/24:1) i2, Ganoderic acid-C2, Butyl methacrylate

Todas las características de esta tabla se obtuvieron con GALGO con una implementación del modelo ML KNN, 2300 Big bangs y 200 generaciones.

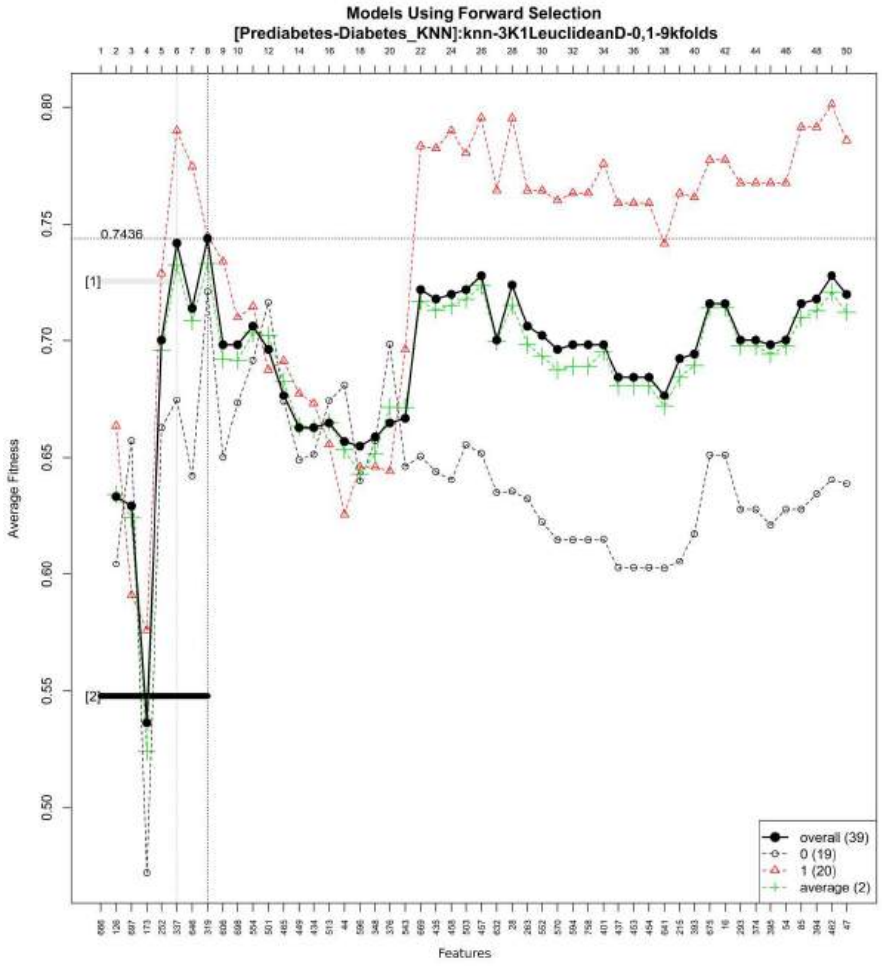


Figura 5.11: Prediabetes-Diabetes FSM-KNN

el modelo, este resultado prueba que los metabolitos resultantes incluidos son extremadamente significativos, como se presenta en la figura de estabilidad de rango genético 5.12 y la figura *fitness* 5.13.

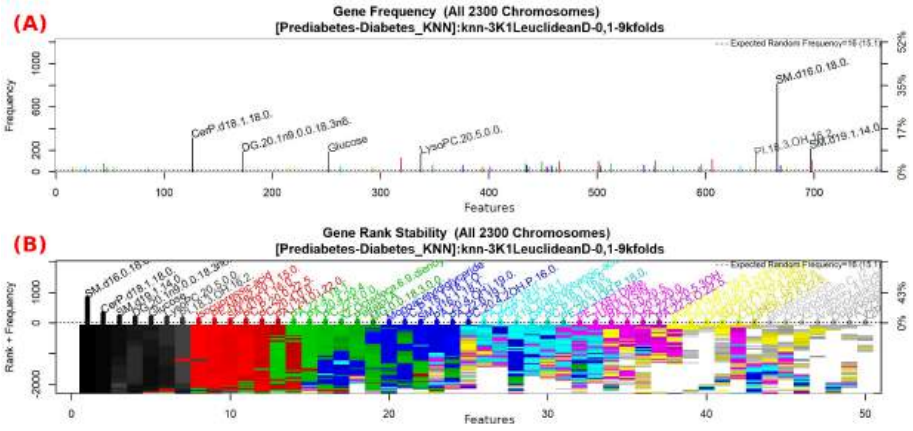


Figura 5.12: Prediabetes-Diabetes Frecuencia-KNN.

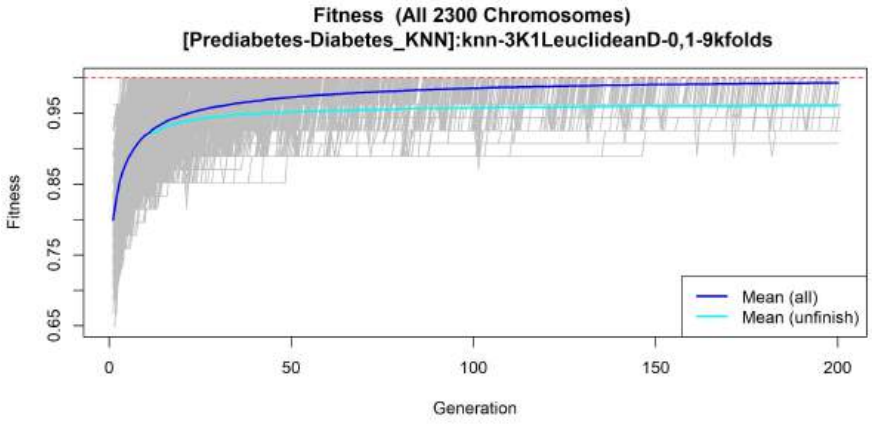


Figura 5.13: Prediabetes-Diabetes Evolución-KNN.

Tabla 5.14: Características obtenidas por el método GALGO KNN en el conjunto de datos Prediabetes - T2DM.

Características resultantes
"Rimiterol", "CerP.d15.0.2.0.", "SM.d19.0.20.3.OH.", "DG.20.0.0.0.18.3n6.", "Ganoderic.acid.C2", "LysoPC.18.3.0.0.", "PI.18.3.OH.16.0.", "Isobehenic.acid"

Todas las características de esta tabla se obtuvieron con GALGO con una implementación del modelo ML KNN, 2300 Big bangs y 200 generaciones.

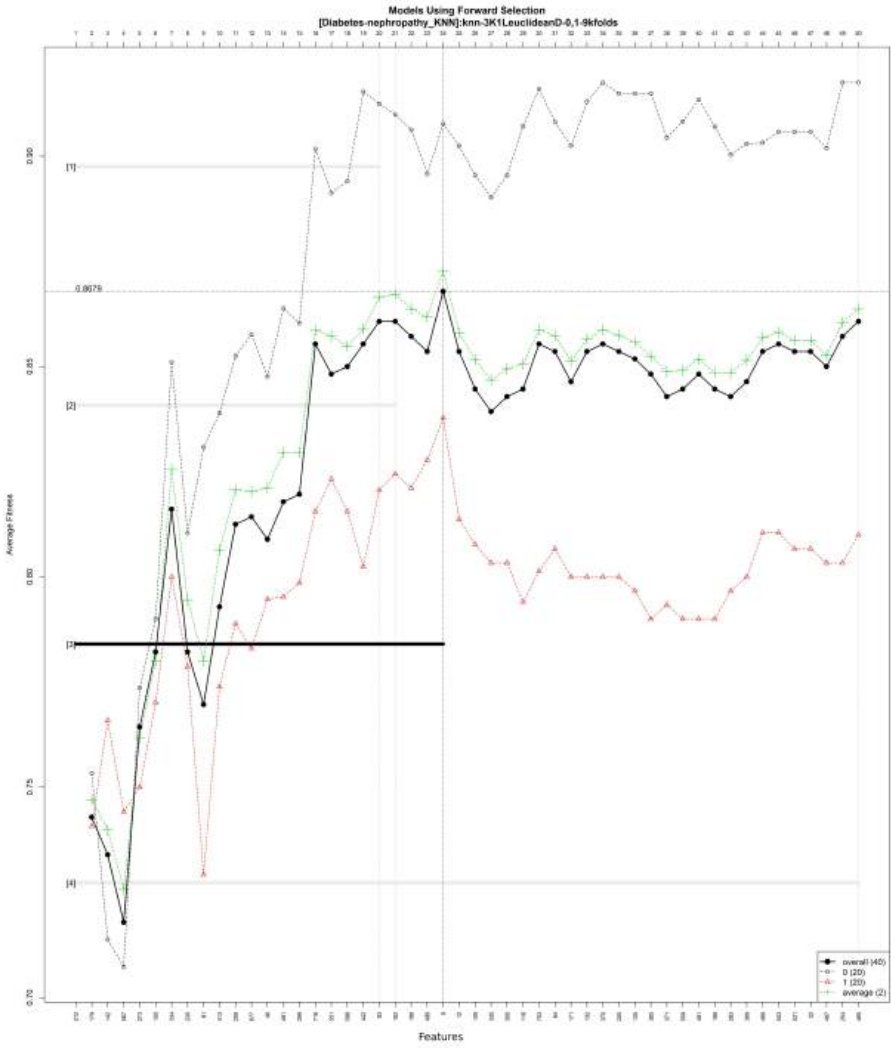


Figura 5.14: Diabetes-Nefropatía Diabética FSM-KNN

Resultados SVM DMT2-Nefropatía diabética

Las características obtenidas por GALGO y el mejor modelo de selección directa, en este caso el modelo 2 como se presenta en la figura 5.14, con una ACC promedio de 0.958 5.10 se presentan en la tabla 5.15. Derivado de la baja cantidad de datos frente a la gran cantidad de características incluidas en el modelo, este resultado prueba que los metabolitos resultantes incluidos son extremadamente significativos, como se presenta en la figura de estabilidad de rango genético 5.15 y la figura *fitness* 5.16.

Caso de Estudio 3: Comparativa en técnicas de selección implementadas, mo-

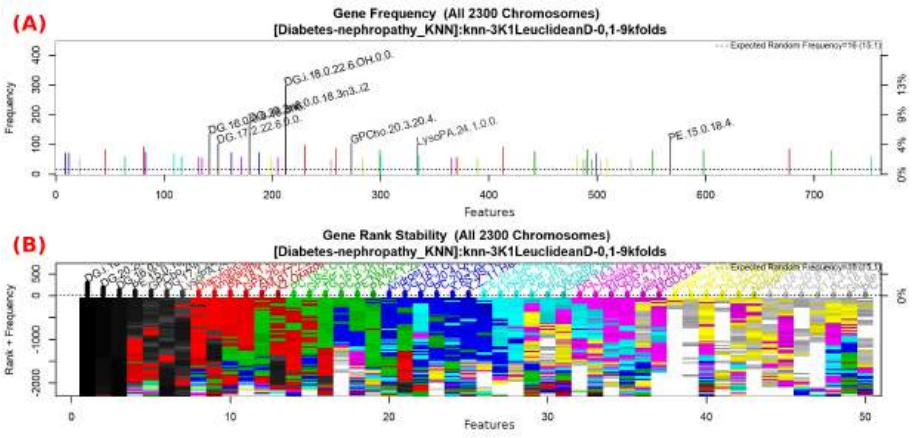


Figura 5.15: Diabetes-Nefropatía Diabética Frecuencia-KNN.

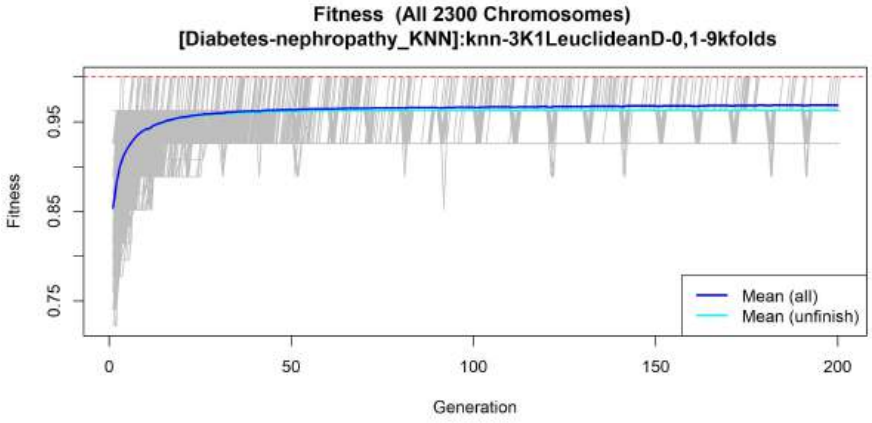


Figura 5.16: Diabetes-Nefropatía Diabética Evolución-KNN.

Tabla 5.15: Características obtenidas por el método GALGO KNN en el conjunto de datos T2DM - Nefropatía diabética.

Características resultantes
"DG.i.18.0.20.4.OH.0.0.", "DG.20.2n6.0.0.18.3n3.", "DG.16.0.0.0.18.1n7.", "PE.14.0.20.5.3OH.", "GPCho.20.2.20.2.", "DG.16.1n7.0.0.20.3n6..i2", "LysoPA.22.2.0.0.", "D.Glutamic.acid", "Alpha.Ionone", "PA.18.1.18.1.", "GPA.20.0.25.0.", "SM.d17.1.LTE4.", "X2.Octenoylcarnitine", "PC.18.2.0.0.", "GPEtn.4.0.18.4.", "TG.16.0.17.1.18.1.", "PC.P.20.0.14.0.", "PE.NMe.20.1.18.0.", "PA.22.1.17.0..i2", "Americine", "DG.18.1n9.0.0.20.4n6.", "DG.20.4.OH.i.14.0.0.0.", "PC.20.3.0.0.", "X.4E.15E..Bilirubin"

Todas las características de esta tabla se obtuvieron con GALGO con una implementación del modelo ML KNN, 2300 Big bangs y 200 generaciones.

delo ensamble y establecimiento de rangos en biomarcadores.

Los datos analizados consisten en 3 conjuntos de datos (Muestra 1.1), uno llamado Siglo XXI en general que incluye 1726 pacientes, incluye 887 casos de DMT2 y 839 sujetos de control, los otros 2 conjuntos de datos son una separación de pacientes masculinos y femeninos. , cada conjunto de datos abarca 855 hombres y 871 mujeres respectivamente. El tratamiento de los datos consistió en la imputación de datos y los criterios de inclusión (Tratamiento de datos 1.2), lo que dio como resultado datos equilibrados de los 3 conjuntos de datos y 12 características descartadas. Las 21 características procesadas en el conjunto de datos general de Siglo XXI y 20 en los conjuntos de datos Masculino y Femenino con las implementaciones de selección de características (Selección de características1.4) obtienen diferentes conjuntos de características para los modelos integrados en el modelo de conjunto.

Resultados de la selección de características

Los datos analizados consistentemente proporcionan un modelo con más de 0.8 de ACC que puede ser probado por el conjunto, en esta prueba ciega realizada con bajas pérdidas en comparación con el modelo entrenado, el modelo para las métricas de la prueba ciega se muestra en la siguiente subsección.

Características obtenidas por el método GALGO KNN

Las características obtenidas por GALGO y el mejor modelo de selección directa, como se muestra en la tabla 5.16, 9 de 21 características se obtuvieron con

una ACC promedio de 0.8312, como se muestra en la figura 5.17. Debido al bajo número de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.18 y la figura de ajuste 5.19.

La frecuencia genética y la estabilidad del rango genético de los modelos se determinan utilizando GA con KNN para seleccionar las características principales en el conjunto de datos. El gráfico superior 5.30, se muestra la frecuencia genética de veces que aparece una característica en los modelos. El gráfico del medio, ordenado por rango, muestra la estabilidad del rango genético y la frecuencia de cada característica en los modelos. El gráfico inferior muestra el número de generaciones empleadas.

La evolución del ajuste más alto a lo largo de generaciones. El eje horizontal representa una generación particular y el eje vertical representa la puntuación de ajuste. El ajuste promedio, representado por una línea continua azul, tiene en cuenta todos los modelos. El ajuste inacabado promedio, representada por una línea continua de color cian, tiene en cuenta todas las búsquedas fallidas para una generación determinada y representa la expectativa promedio en el peor de los casos. La línea de puntos roja representa el objetivo de *fitness* de GA establecido.

Las características obtenidas por GALGO para el conjunto de datos masculino y el mejor modelo de selección directa, como se muestra en la tabla 5.17, 11 de 20 características se obtuvieron con una ACC promedio de 0.8714, como se muestra en la figura 5.20. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.21 y la figura de ajuste 5.22.

Las características obtenidas por GALGO para el conjunto de datos femenino y el mejor modelo de selección directa, como se muestra en la tabla 5.17, 14 de 20 características se obtuvieron con una ACC promedio de 0.8312, como se muestra en la figura 5.23. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.24 y la figura de ajuste 5.25.

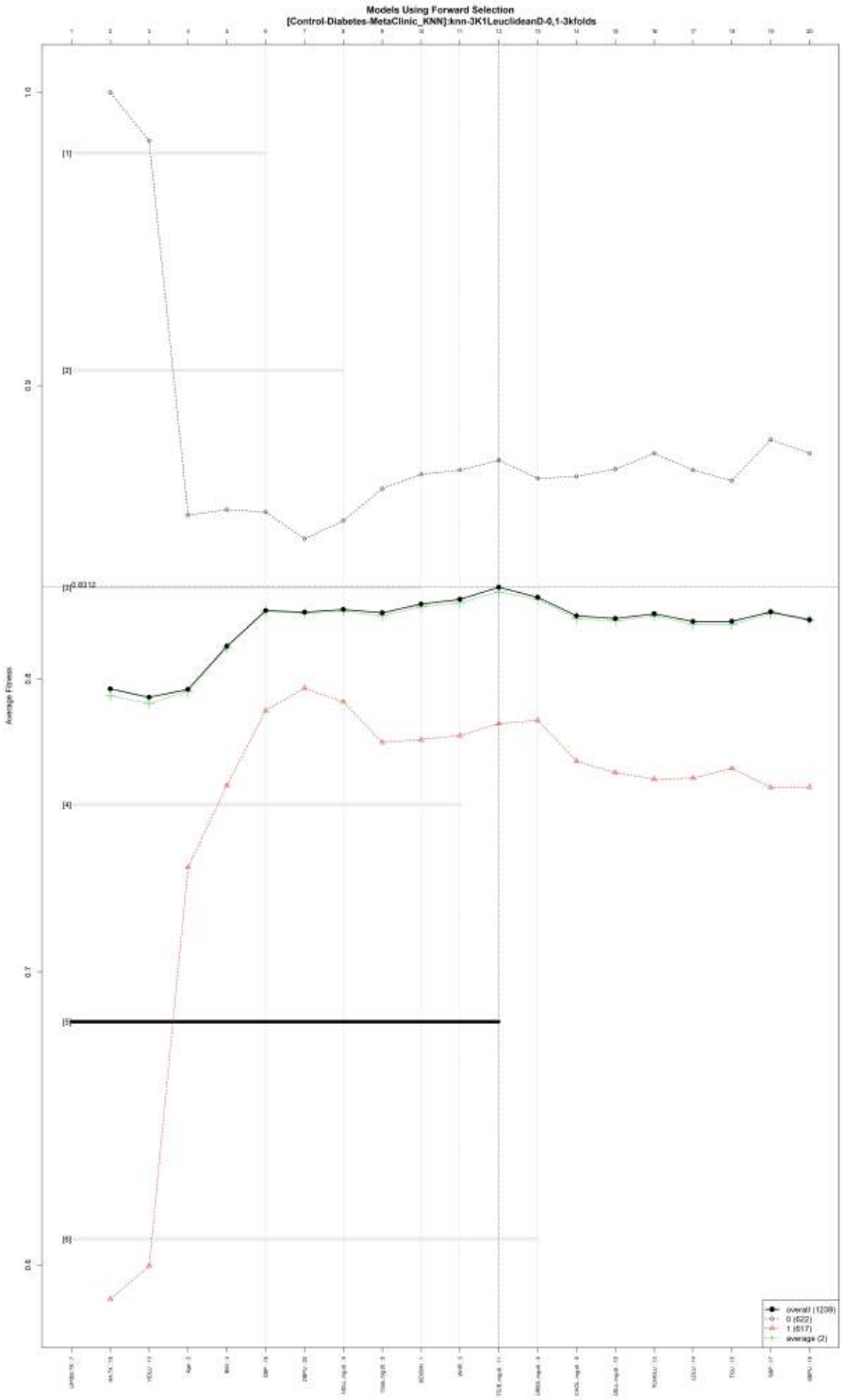


Figura 5.17: dataset SigloXXI_Control-Diabetes FSM-KNN

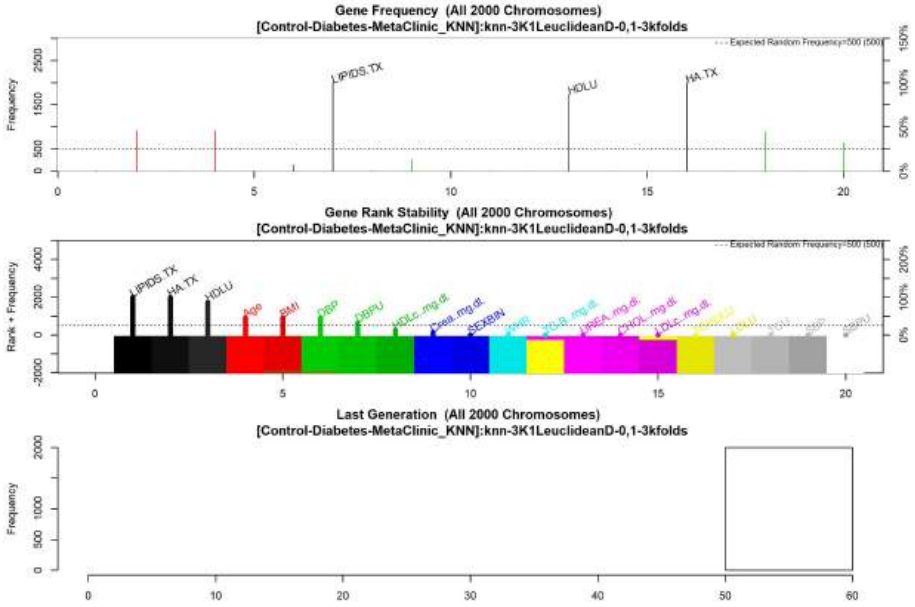


Figura 5.18: Frecuencia y estabilidad del rango genético dataset Siglo XXI_Control-Diabetes

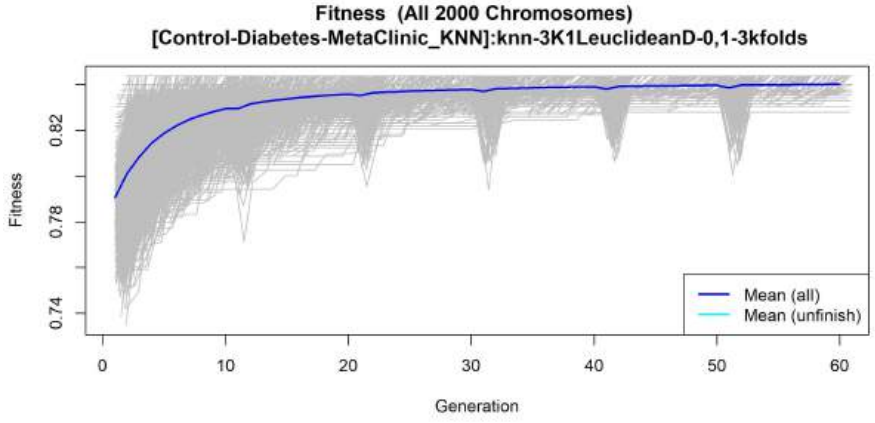


Figura 5.19: Evolución del ajuste en Siglo XXI_Control-Diabetes-KNN

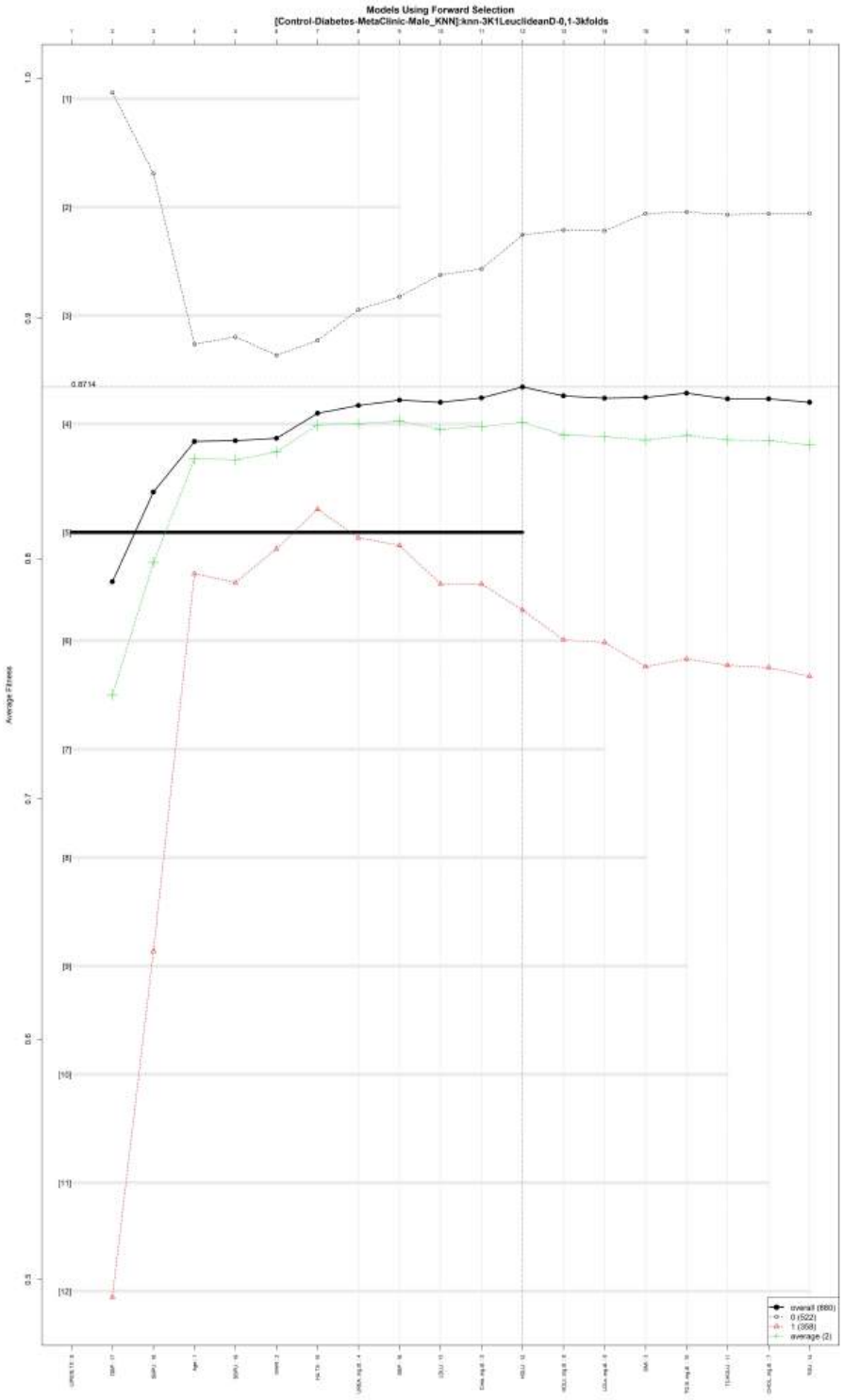


Figura 5.20: SigloXXI_Control-Diabetes-Hombres FSM-KNN.

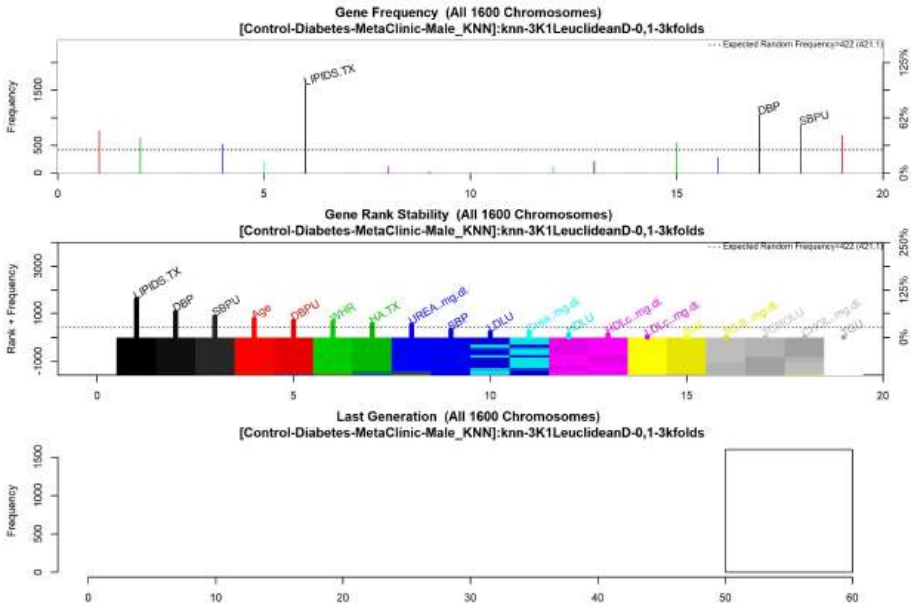


Figura 5.21: Frecuencia y estabilidad del rango genético en el conjunto de datos Siglo XXI_Control-Diabetes-Hombres

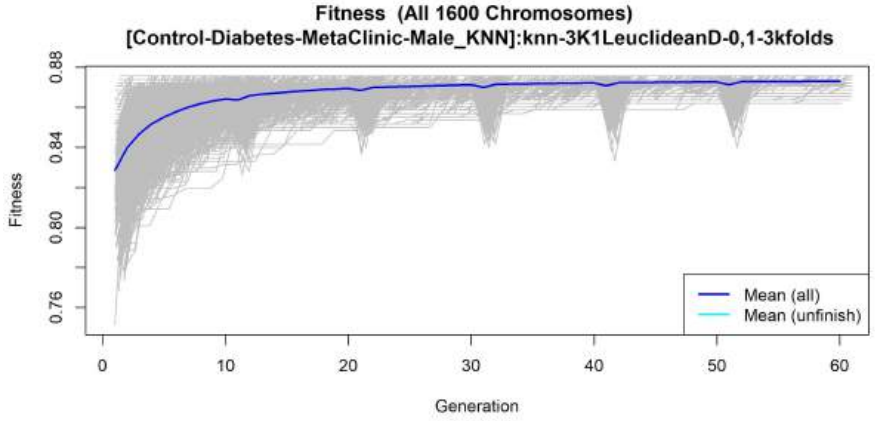


Figura 5.22: Evolución del ajuste en Siglo XXI_Control-Diabetes-Hombres-KNN

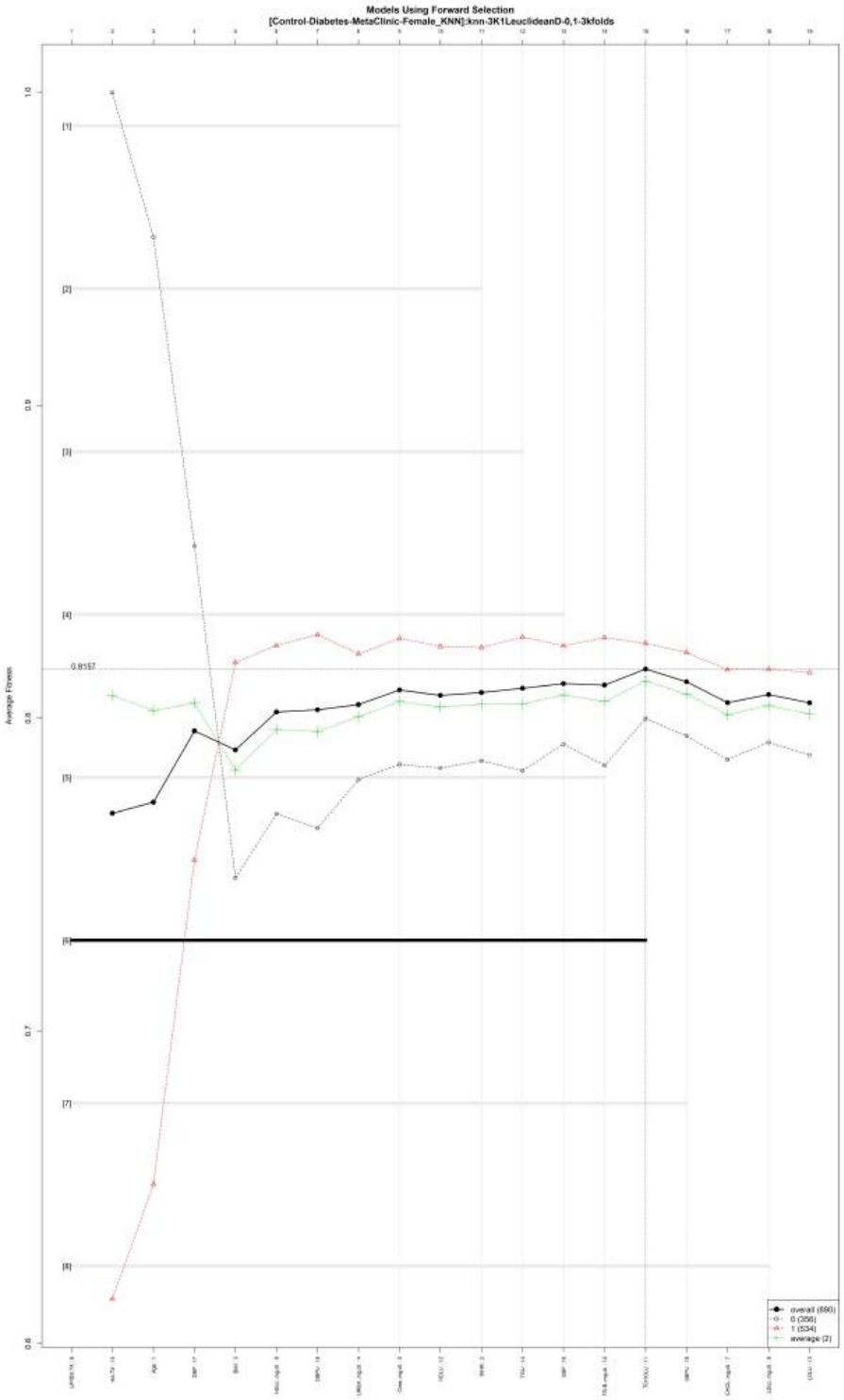


Figura 5.23: SigloXXI_Control-Diabetes-Mujeres FSM-KNN.

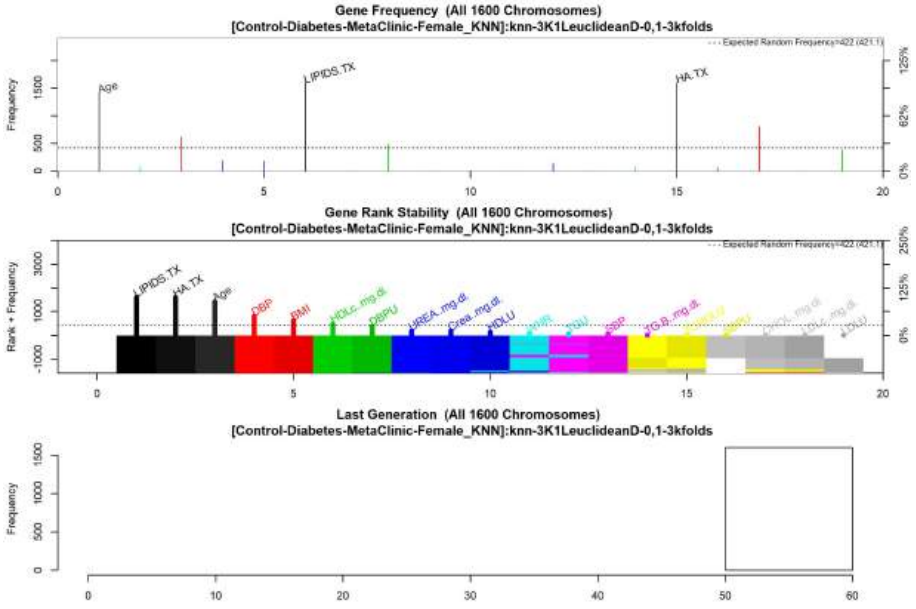


Figura 5.24: Frecuencia y estabilidad del rango genético en el conjunto de datos Siglo XXI_Control-Diabetes-Mujeres

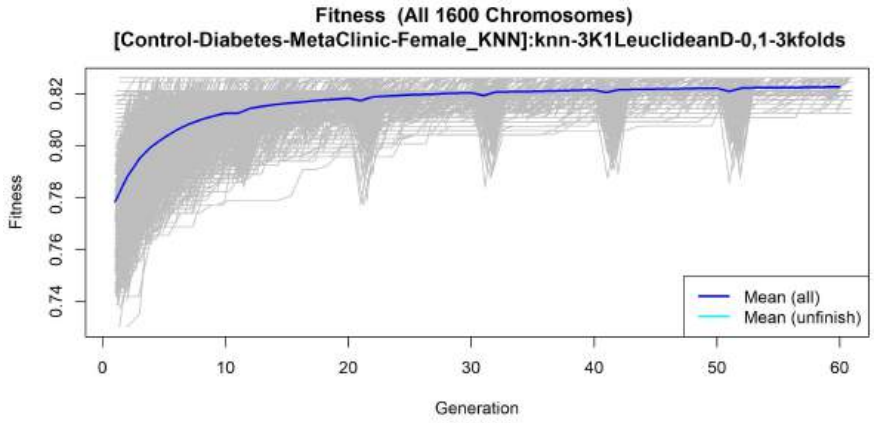


Figura 5.25: Evolución del ajuste en Siglo XXI_Control-Diabetes-Mujeres-KNN

Tabla 5.16: Características del resultado KNN Galgo Siglo XXI General.

Siglo XXI General
“Creatinine”, “Triglycerides (sin corregir)”, “Colesterol total (sin corregir)”, “Sexo”, “Proporción cintura cadera”, “Presión arterial sistólica”, “Presión arterial sistólica (sin corregir)”, “Cholesterol”, “Urea”

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML KNN, con 2000 Big Bangs y 60 generaciones.

Tabla 5.17: Características del resultado KNN Galgo Masculino/Femenino.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
“Creatinine”	“Creatinine”
“Presión arterial sistólica”	“Triglycerides (sin corregir)”
“Presión arterial diastólica”	“Presión arterial sistólica”
“Presión arterial sistólica (sin corregir)”	“Relación cintura-cadera”
“Age”	“Cholesterol”
“Triglycerides (sin corregir)”	“Presión arterial sistólica (sin corregir)”
”Índice de masa corporal”	”Índice de masa corporal”
“Tratamiento de la hipertensión”	“Urea”
“Lipoproteína de alta densidad (sin corregir)”	“Colesterol total (sin corregir)”
“Urea”	“Age”
“Colesterol total (sin corregir)”	“Lipoproteínas de baja densidad (sin corregir)”
	“Tratamiento de la hipertensión”
	“Lipoproteínas de baja densidad”
	“Triglycerides”

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML KNN, con 1600 Big Bangs y 60 generaciones.

Características obtenidas por el método GALGO Nearcent

Las características obtenidas por GALGO y el mejor modelo de selección directa en el conjunto de datos general, como se muestra en la tabla 5.18, 16 de 21 características se obtuvieron con una ACC promedio de 0.8503, como se muestra en la figura 5.26. Debido al bajo número de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.27 y la figura de ajuste 5.28.

La frecuencia genética y la estabilidad del rango genético de los modelos se determinan utilizando GA con Nearcent para seleccionar las características principales en el conjunto de datos. El gráfico superior muestra la frecuencia genética de veces que aparece una característica en los modelos. El gráfico del medio, ordenado por rango, muestra la estabilidad del rango genético y la frecuencia de cada característica en los modelos. El gráfico inferior muestra el número de generaciones empleadas, que en este caso es 2000.

Las características obtenidas por GALGO para el conjunto de datos masculino y el mejor modelo de selección directa, como se muestra en la tabla 5.19, 4 de 20 características se obtuvieron con una ACC promedio de 0.8799, como se muestra en la figura 5.29. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.30 y la figura de ajuste 5.31.

Las características obtenidas por GALGO para el conjunto de datos femenino y el mejor modelo de selección directa, como se muestra en la tabla 5.19, 7 de 20 características se obtuvieron con una ACC promedio de 0.8503, como se muestra en la figura 5.32. Debido al bajo número de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.33 y la figura de ajuste 5.34.

Tabla 5.18: Características del resultado general de Nearcent Galgo Siglo XXI.

Siglo XXI General
“Creatinine”, “TGU”, “Sex”, “SBPU”, “LDLc”, “SBP”, “BMI”, “WHR”, “Age”, “Triglycerides”, “Cholesterol”, “LDLU”, “Urea”, “TCHOLU”, “HDLU”, “LIPIDS-TX”

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML Nearcent, con 2000 Big Bangs y 60 generaciones.

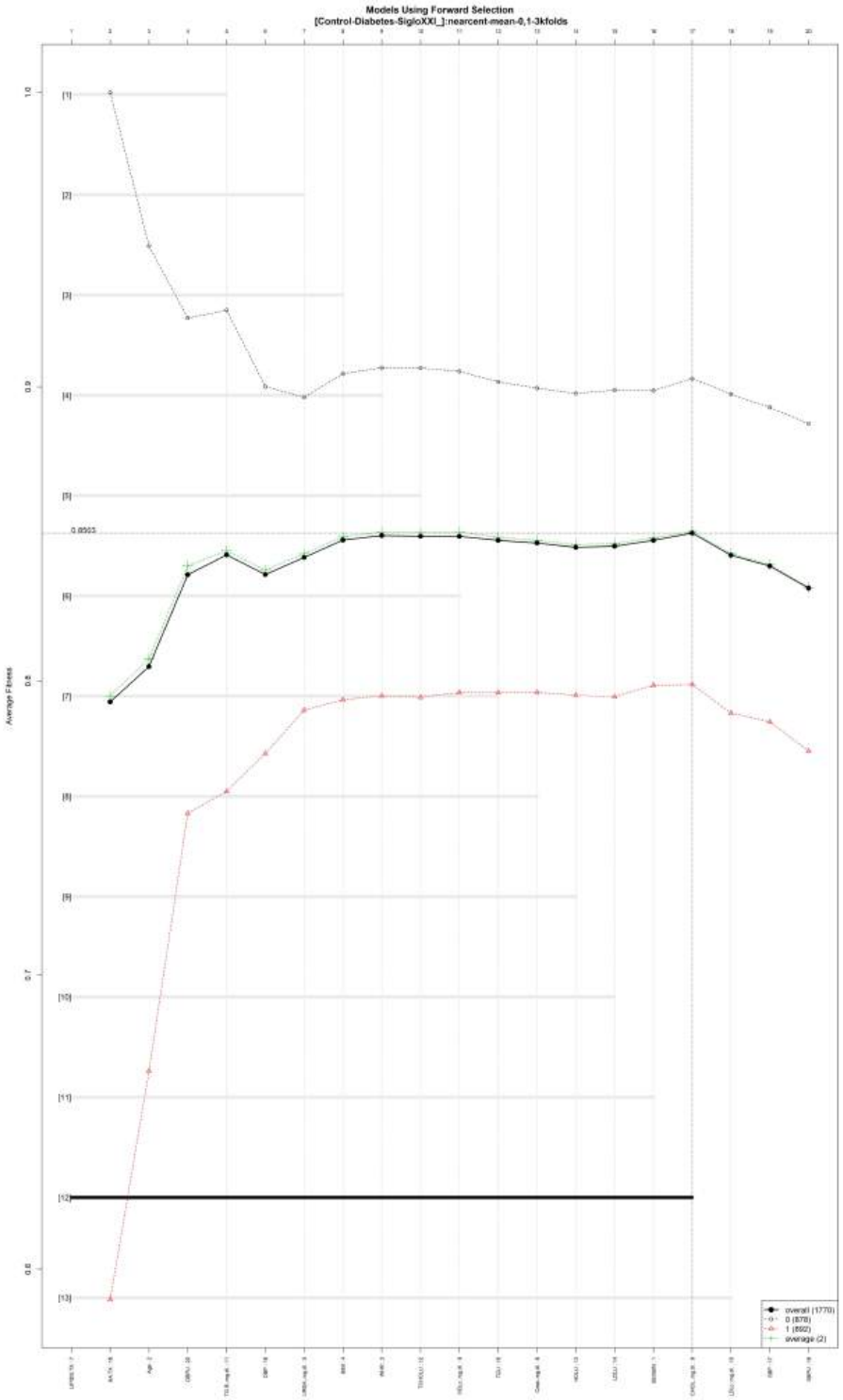


Figura 5.26: SigloXXI_Control-Diabetes FSM-Nearcent.

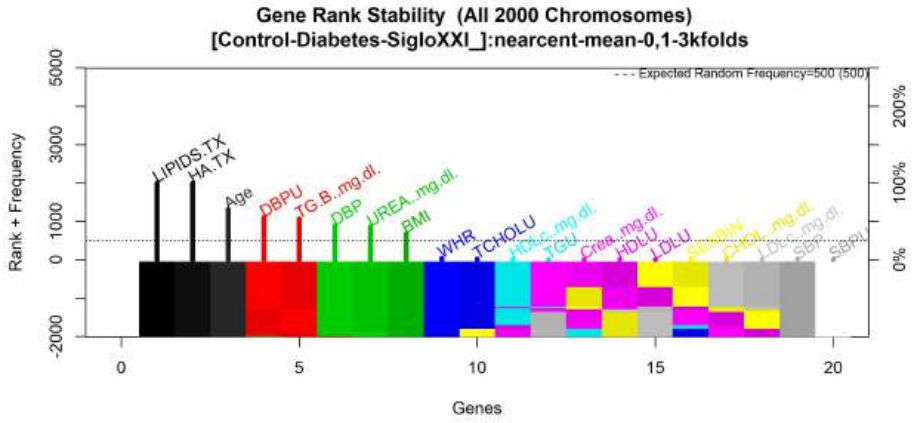


Figura 5.27: La frecuencia genética y la estabilidad del rango genético En el conjunto de datos SigloXXI_Control-Diabetes.

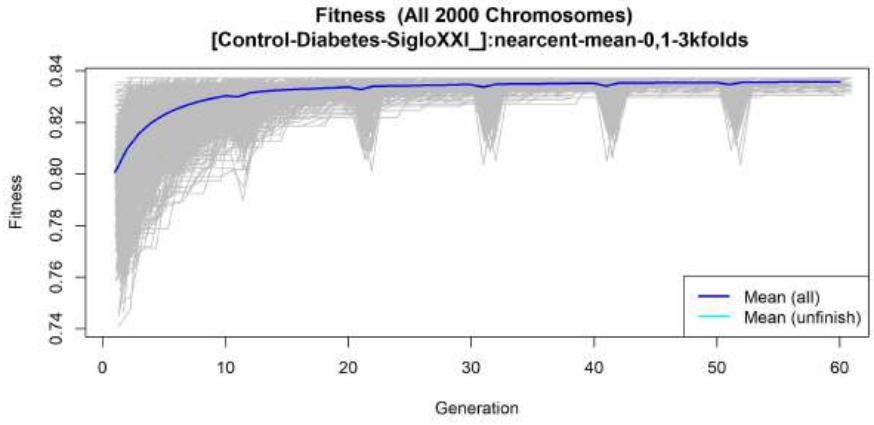


Figura 5.28: Evolución del ajuste en Siglo XXI_Control-Diabetes-Nearcent

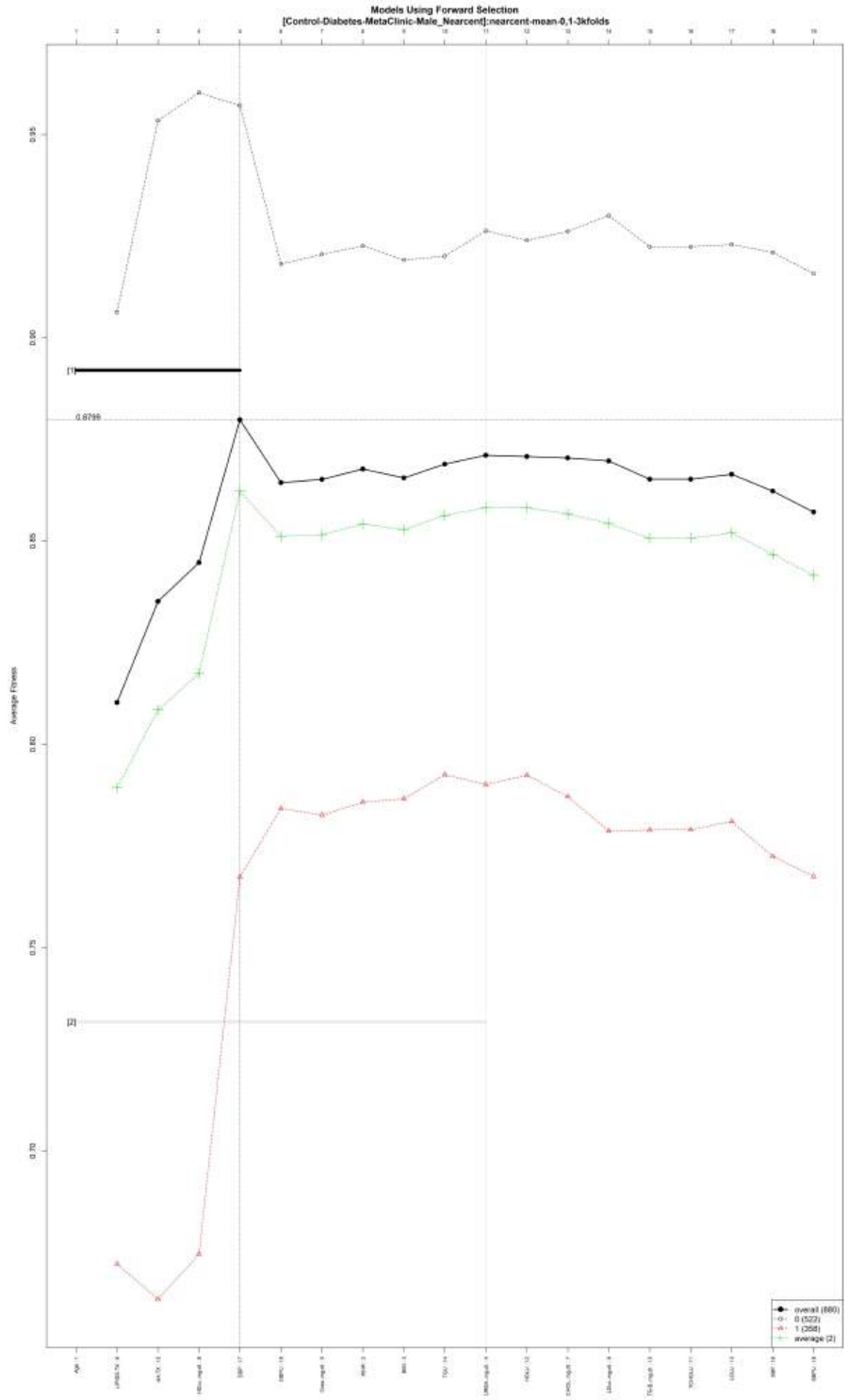


Figura 5.29: SigloXXI_Control-Diabetes-Hombres FSM-Nearcent.

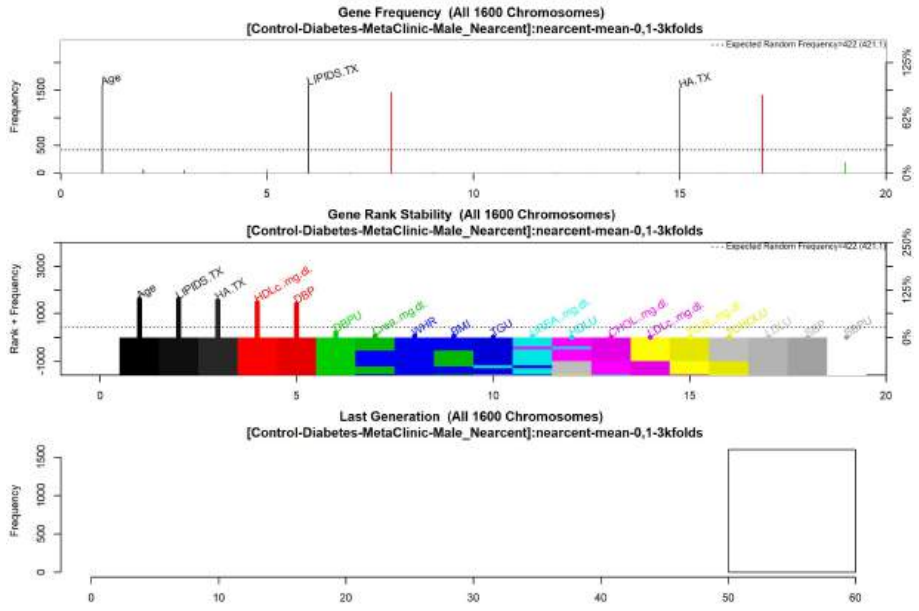


Figura 5.30: La frecuencia genética y la estabilidad del rango genético En el conjunto de datos SigloXXI_Control-Diabetes-Hombres.

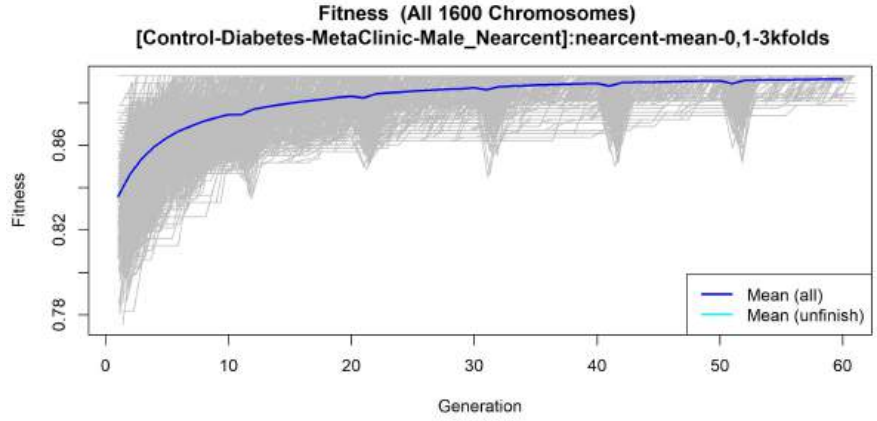


Figura 5.31: Evolución del ajuste en Siglo XXI_Control-Diabetes-Nearcent

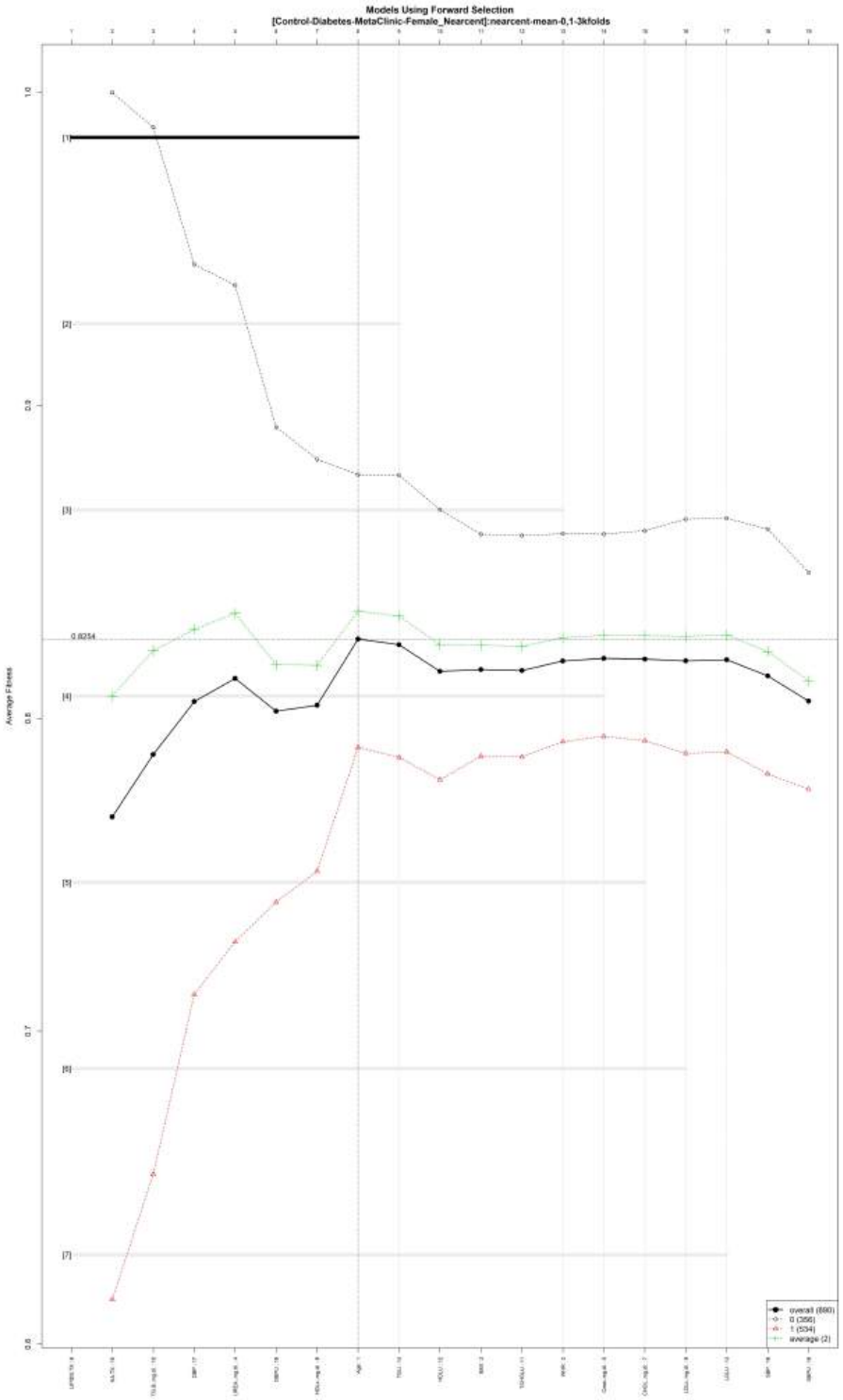


Figura 5.32: SigloXXI_Control-Diabetes-Mujeres FSM-KNN.

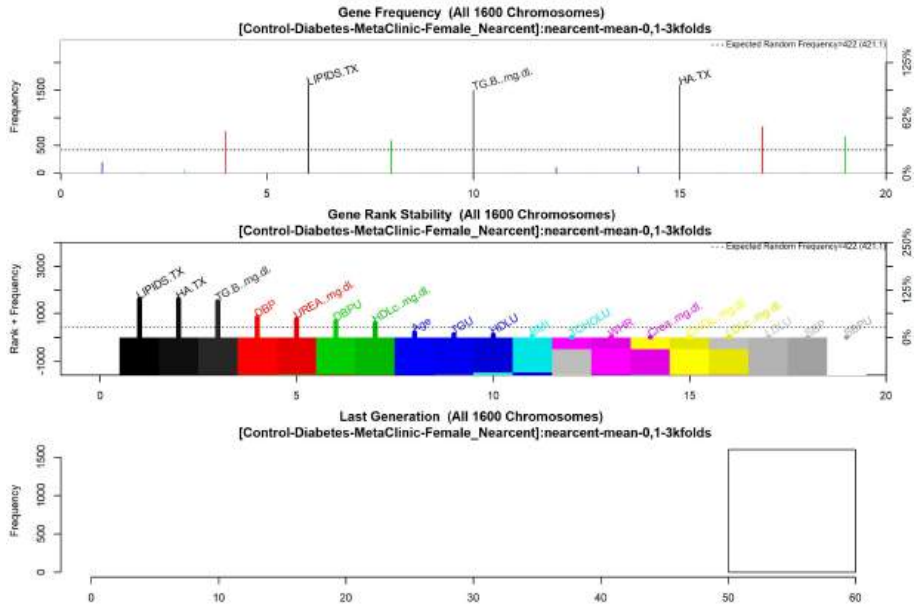


Figura 5.33: La frecuencia genética y la estabilidad del rango genético En el conjunto de datos SigloXXI_Control-Diabetes-Mujeres.

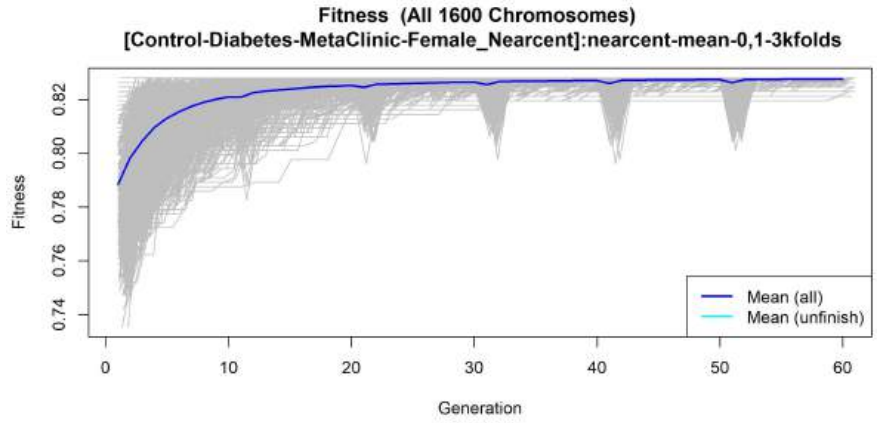


Figura 5.34: Evolución del ajuste en Siglo XXI_Control-Diabetes-Nearcent.

Tabla 5.19: Características del resultado Nearcent Galgo Masculino/Femenino.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
"Creatinine"	"Creatinine"
"TGU"	"TGU"
"Cholesterol"	"LDLc"
"SBP"	"SBP"
	"BMI"
	"SBPU"
	"Cholesterol"

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML Nearcent, con 1600 Big Bangs y 60 generaciones.

Características obtenidas por el método GALGO SVM

Las características obtenidas por GALGO y el mejor modelo de selección directa en el conjunto de datos general, como se muestra en la tabla 5.20, 18 de 21 características se obtuvieron con una ACC promedio de 0.8705, como se muestra en la figura 5.35. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.36 y la figura de ajuste 5.37.

La frecuencia genética y la estabilidad del rango genético de los modelos se determinan utilizando GA con SVM para seleccionar las características principales en el conjunto de datos. El gráfico superior muestra la frecuencia genética de veces que aparece una característica en los modelos. El gráfico del medio, ordenado por rango, muestra la estabilidad del rango genético y la frecuencia de cada característica en los modelos. El gráfico inferior muestra el número de generaciones empleadas, que en este caso es 2000.

Las características obtenidas por GALGO para el conjunto de datos masculino y el mejor modelo de selección directa, como se muestra en la tabla 5.21, 8 de 20 características se obtuvieron con una ACC promedio de 0.8923, como se muestra en la figura 5.38. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.39 y la figura de ajuste 5.40.

Las características obtenidas por GALGO para el conjunto de datos femenino

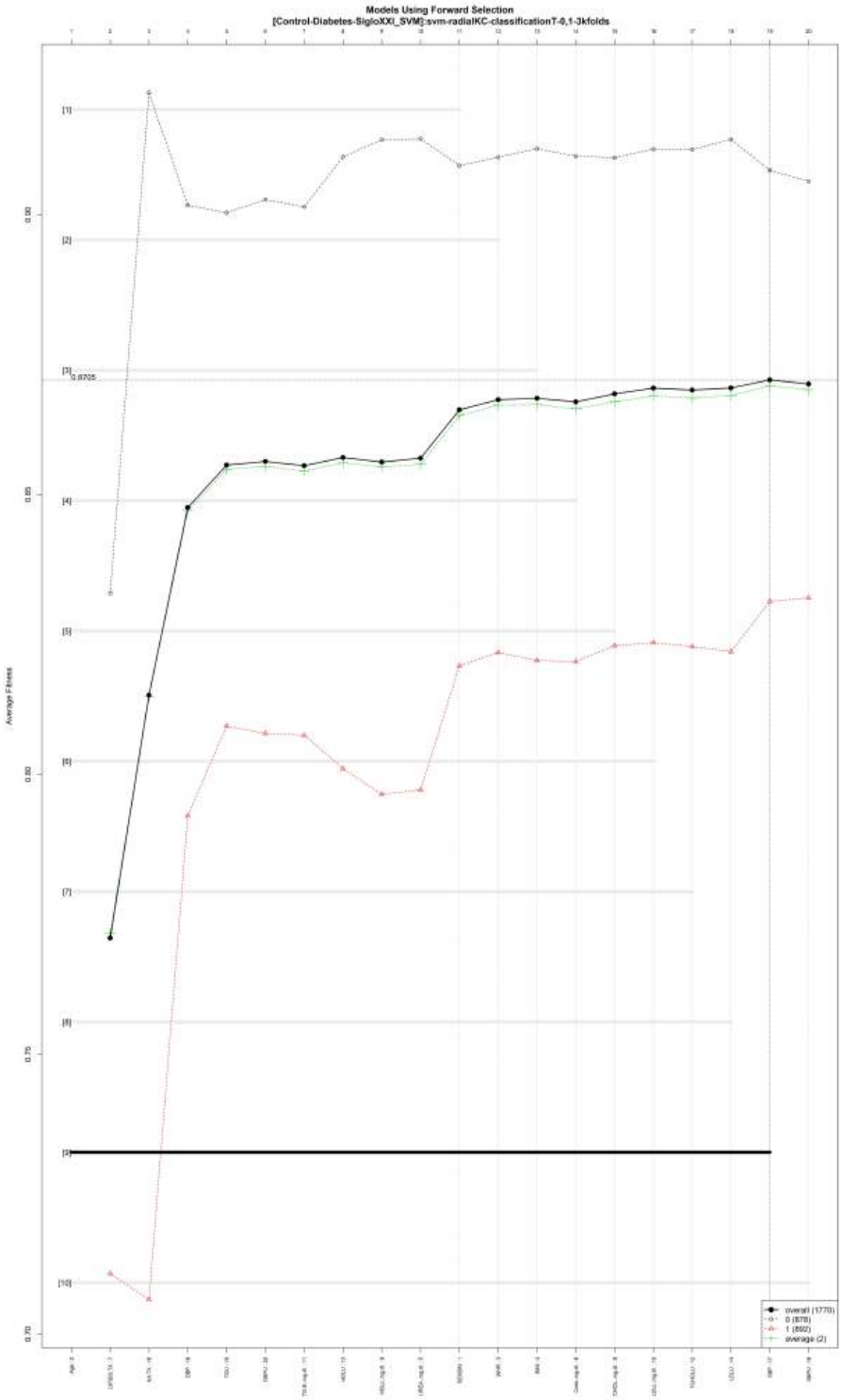


Figura 5.35: SigloXXI_Control-Diabetes FSM-SVM.

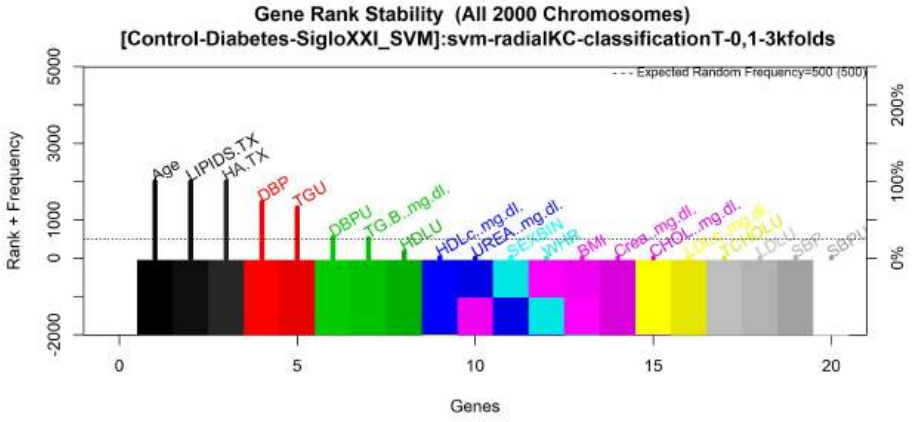


Figura 5.36: Frecuencia genética y la estabilidad del rango genético Control-Diabetes-SVM.

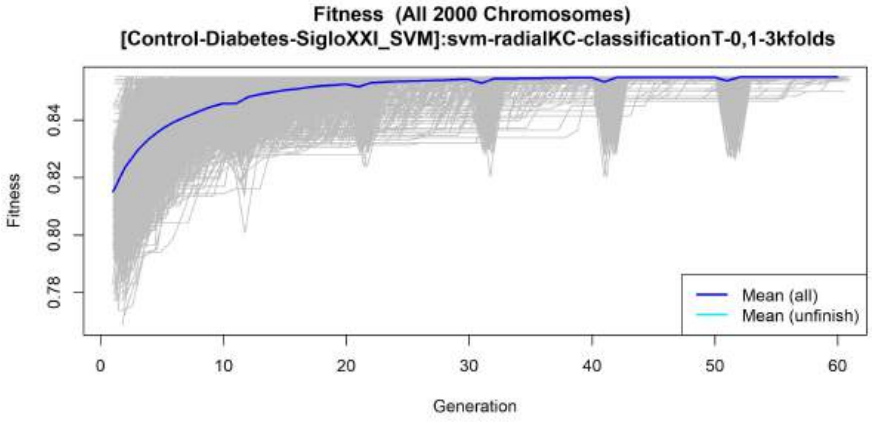


Figura 5.37: Evolución del ajuste en Siglo XXI_Control-Diabetes-.

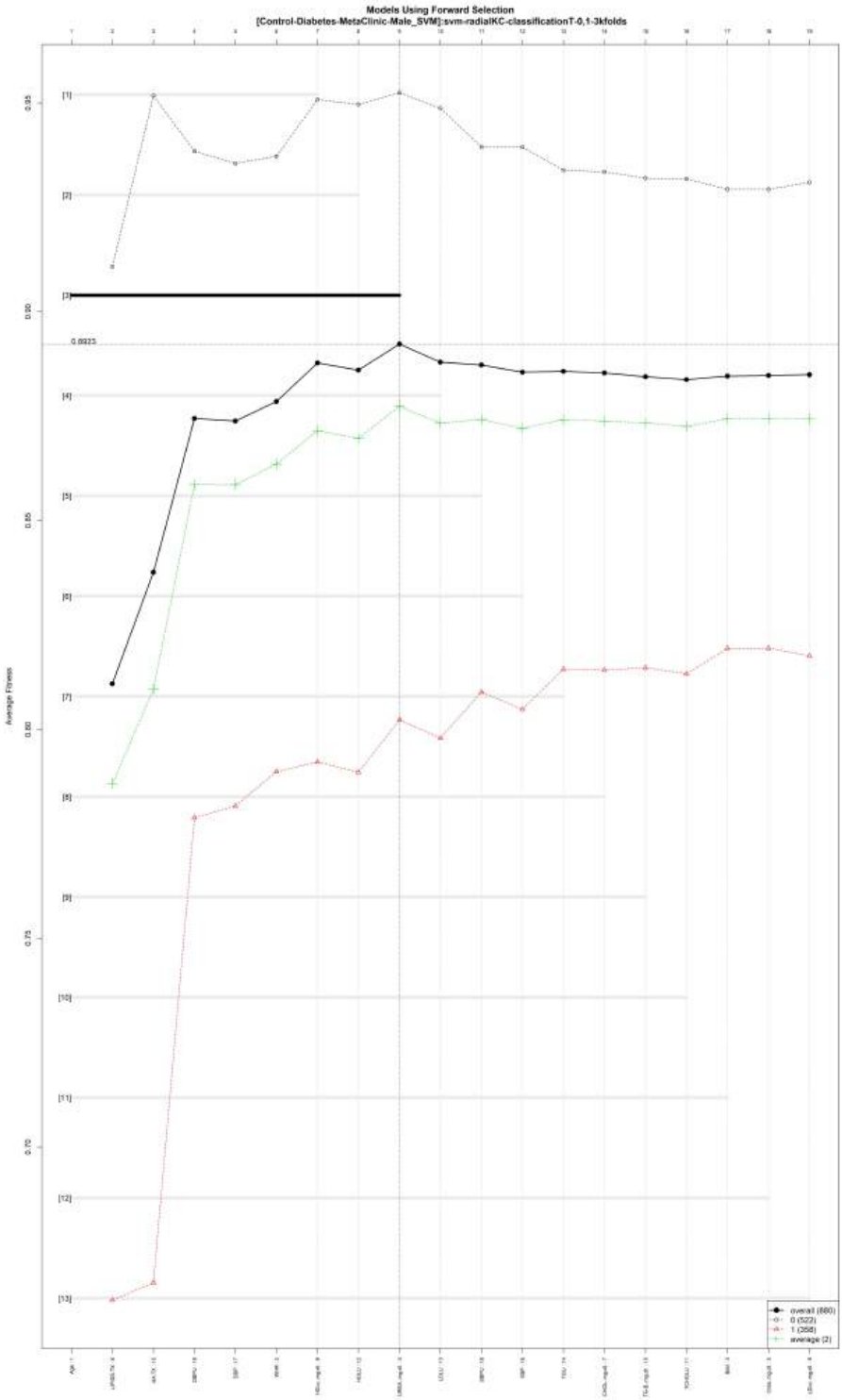


Figura 5.38: SigloXXI_Control-Diabetes-Hombres FSM-SVM.

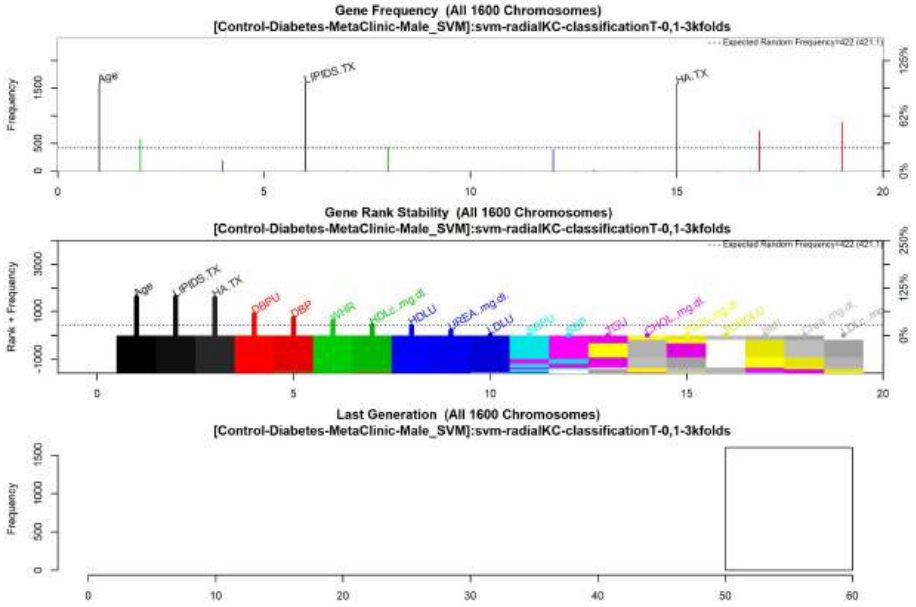


Figura 5.39: Frecuencia genética y la estabilidad del rango genético Control-Diabetes-Hombres-SVM.

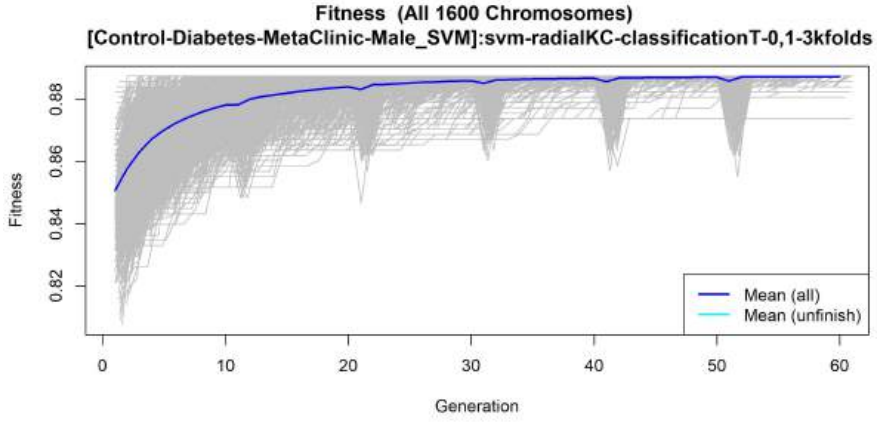


Figura 5.40: Evolución del ajuste en Siglo XXI_Control-Diabetes-Hombres-SVM

y el mejor modelo de selección directa, como se muestra en la tabla 5.21, 9 de 20 características se obtuvieron con una ACC promedio de 0.8364, como se muestra en la figura 5.41. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.42 y la figura de ajuste 5.43.

Tabla 5.20: Características del resultado SVM Galgo Siglo XXI General.

Siglo XXI General
"Sex", "Creatinine", "TGU", "SBP", "LDLU", "SBPU", "LDLc", "TCHOLU", "Cholesterol", "BMI", "Age", "WHR", "Urea", "LIPIDS-TX", "HDL", "Triglycerides", "HDLU", "HA-TX"

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML SVM, con 2000 Big Bangs y 60 generaciones.

Tabla 5.21: Características del resultado SVM Galgo Masculino/Femenino.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
"Creatinine"	"Creatinine"
"TGU"	"TGU"
"SBPU"	"SBP"
"SBP"	"LDLc"
"Age"	"LDLU"
"Cholesterol"	"SBPU"
"TCHOLU"	"BMI"
"BMI"	"TCHOLU"
	"Cholesterol"

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML SVM, con 1600 Big Bangs y 60 generaciones.

Características obtenidas por el método GALGO LR

Las características obtenidas por GALGO y el mejor modelo de selección directa en el conjunto de datos general, como se muestra en la tabla 5.22, 18 de 21 características se obtuvieron con una ACC promedio de 0.8652, como se muestra en la figura 5.44. Debido al bajo número de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados

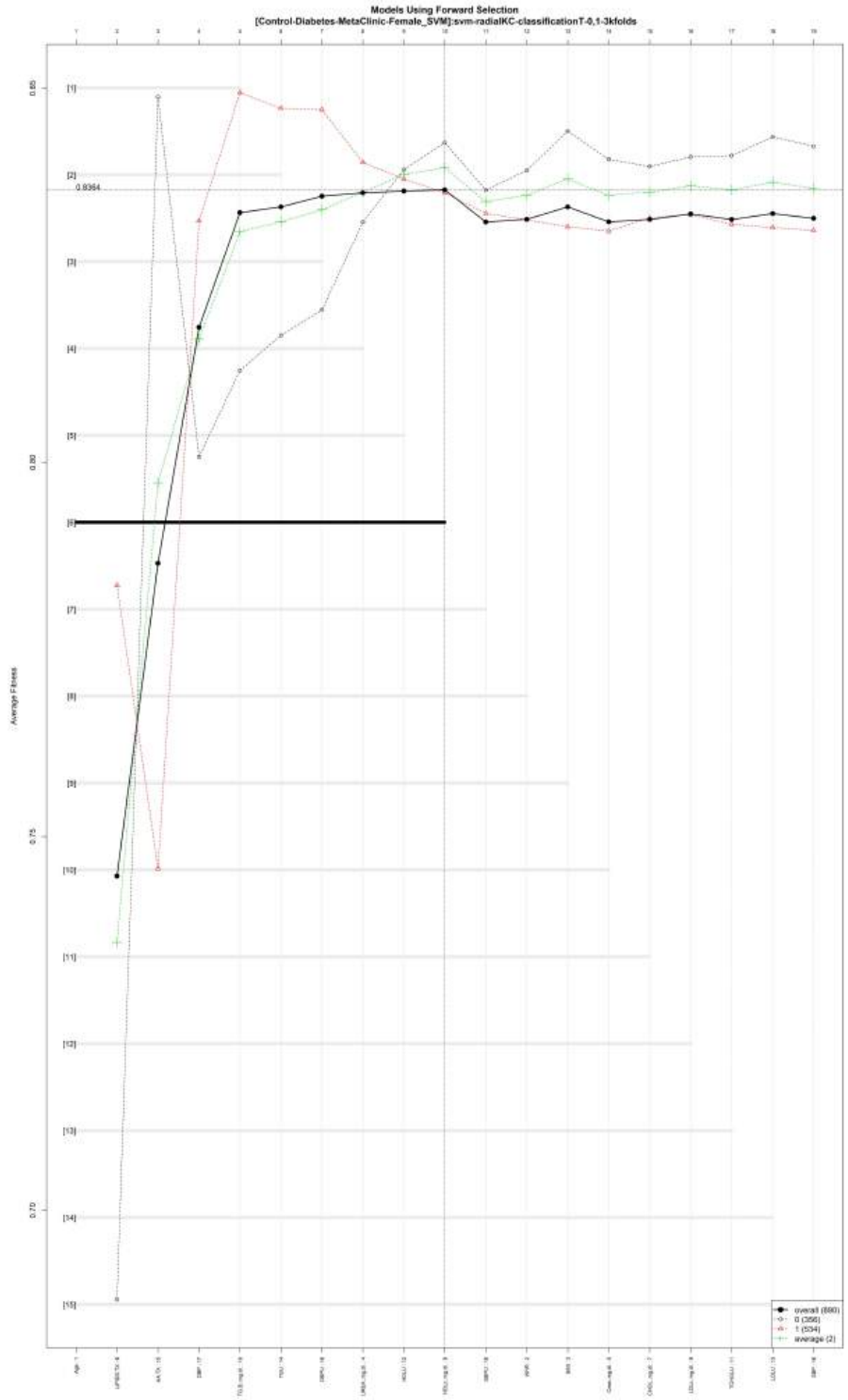


Figura 5.41: SigloXXI_Control-Diabetes-Mujeres FSM-SVM.

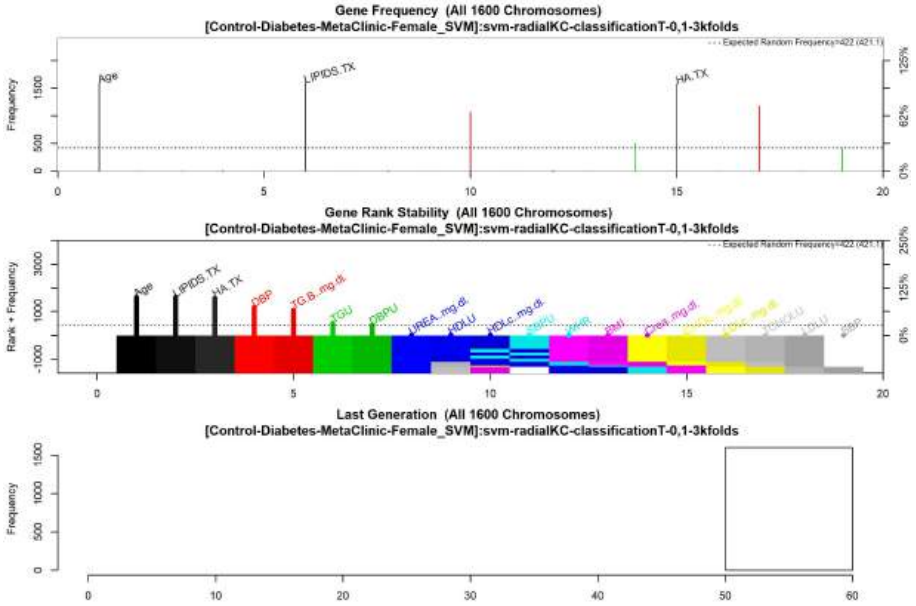


Figura 5.42: Frecuencia genética y la estabilidad del rango genético Control-Diabetes-Mujeres-SVM.

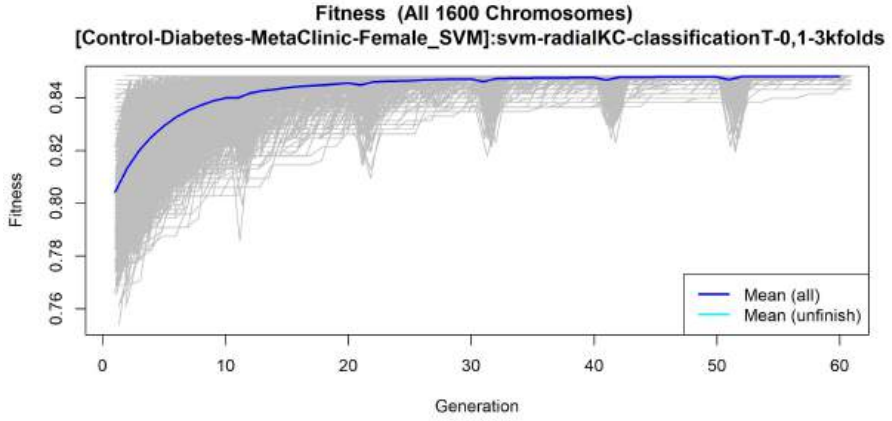


Figura 5.43: Evolución del ajuste en Siglo XXI_Control-Diabetes-Mujeres-SVM.

muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.45 y la figura de ajuste 5.46.

La frecuencia genética y la estabilidad del rango genético de los modelos se determinan utilizando GA con LogReg para seleccionar las características principales en el conjunto de datos. El gráfico superior muestra la frecuencia genética de veces que aparece una característica en los modelos. El gráfico del medio, ordenado por rango, muestra la estabilidad del rango genético y la frecuencia de cada característica en los modelos. El gráfico inferior muestra el número de generaciones empleadas, que en este caso es 2000.

Las características obtenidas por GALGO para el conjunto de datos masculino y el mejor modelo de selección directa, como se muestra en la tabla 5.23, 4 de 20 características se obtuvieron con una ACC promedio de 0.883, como se muestra en la figura 5.47. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.48 y la figura de ajuste 5.49.

Las características obtenidas por GALGO para el conjunto de datos femenino y el mejor modelo de selección directa, como se muestra en la tabla 5.23, 13 de 20 características se obtuvieron con una ACC promedio de 0.8456, como se muestra en la figura 5.50. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.51 y la figura de ajuste 5.52.

Tabla 5.22: Características del resultado LR Galgo Siglo XXI General.

Siglo XXI General
“Sex”, “Creatinine”, “SBP”, “TGU”, “TCHOLU”, “Cholesterol”, “SBPU”, “LDLc”, “LDLU”, “HA-TX”, “DBP”, “WHR”, “LIPIDS-TX”, “HDL”, “Triglycerides”, “HDLU”, “Age”, “BMI”

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML LR, con 2000 Big Bangs y 60 generaciones.

Características obtenidas por el método GALGO NNET

Las características obtenidas por GALGO y el mejor modelo de selección directa en el conjunto de datos general, como se muestra en la tabla 5.24, 2 de 21

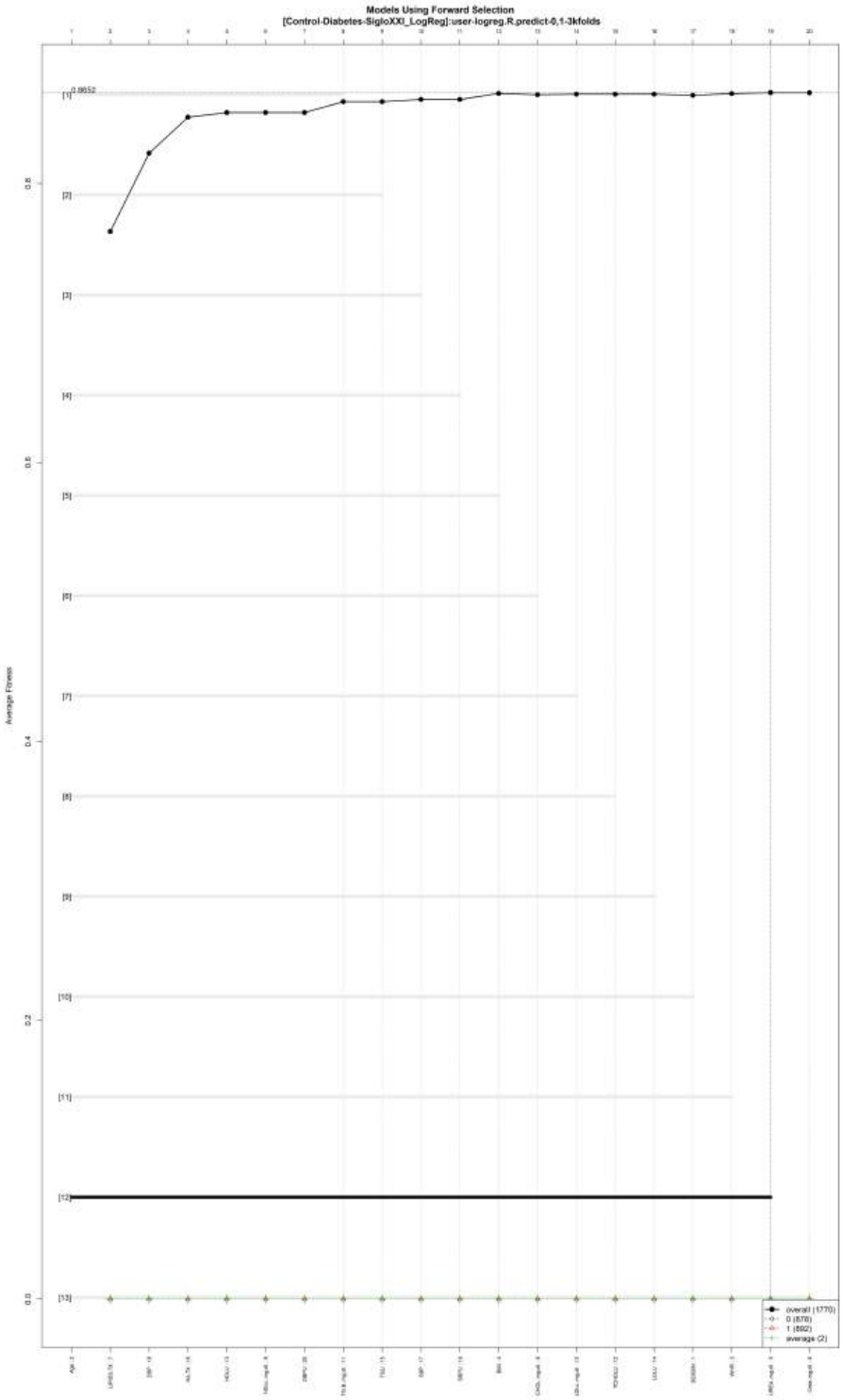


Figura 5.44: SigloXXI_Control-Diabetes FSM-LogReg.

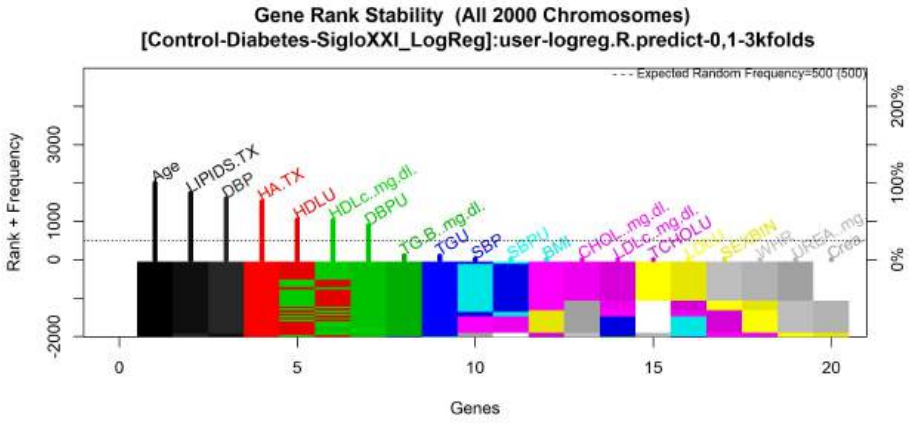


Figura 5.45: La frecuencia genética y la estabilidad del rango genético Control-Diabetes-LogReg.

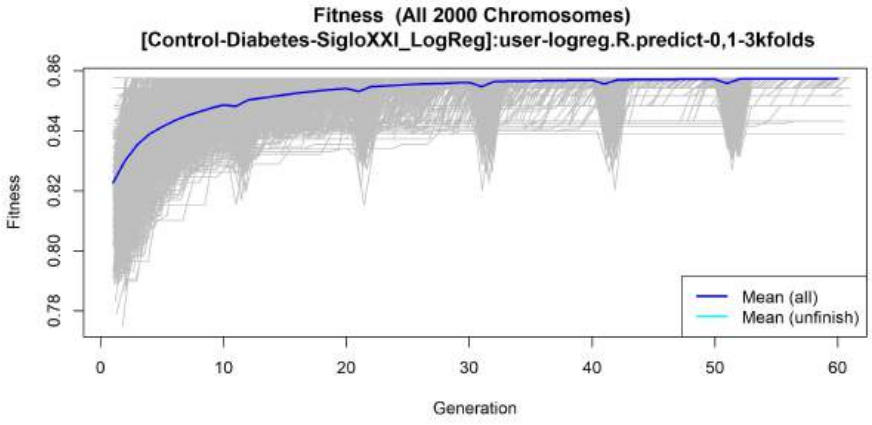


Figura 5.46: Evolución del ajuste en Siglo XXI_Control-Diabetes-LogReg.

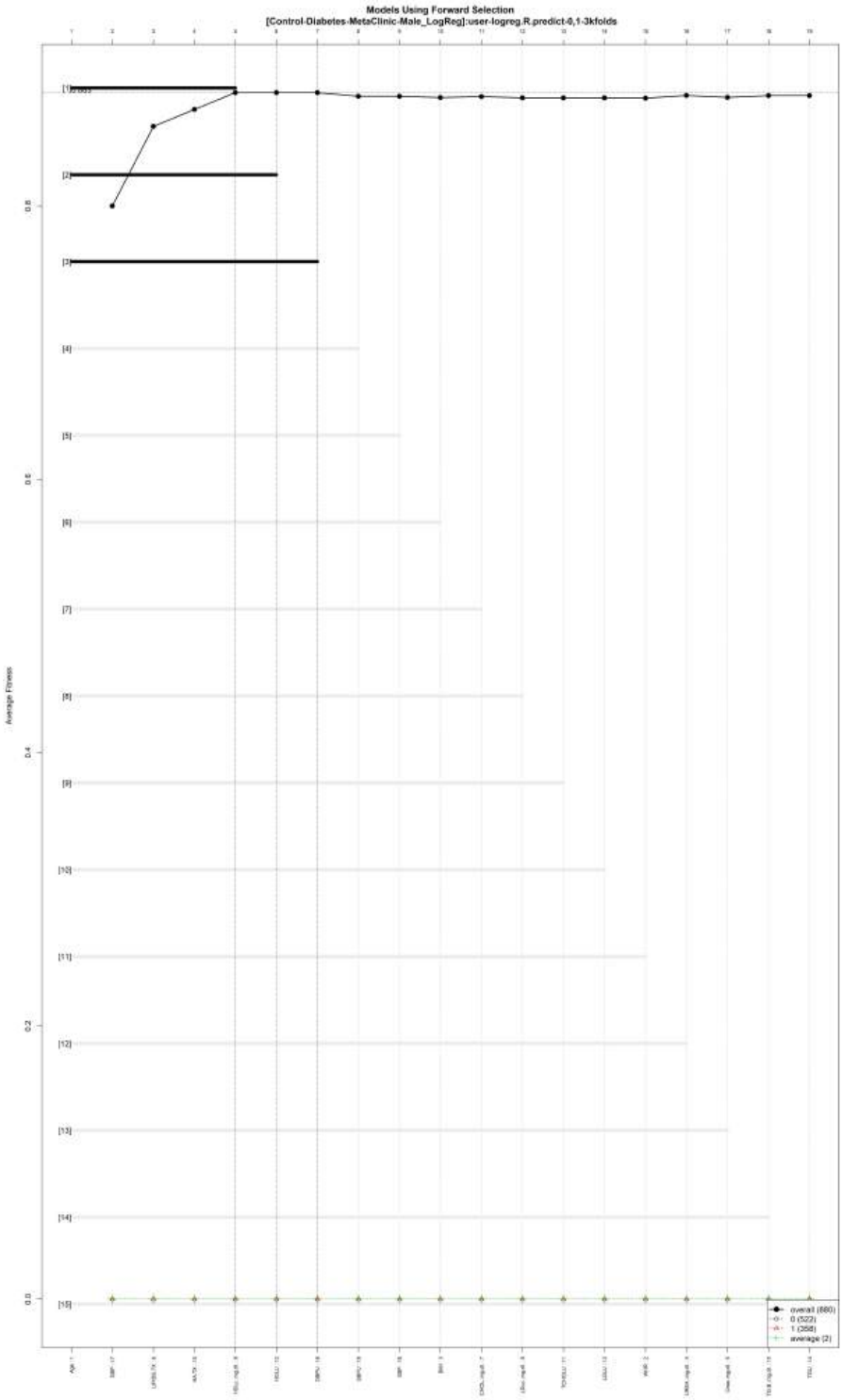


Figura 5.47: SigloXXI_Control-Diabetes-Hombres FSM-LogReg.

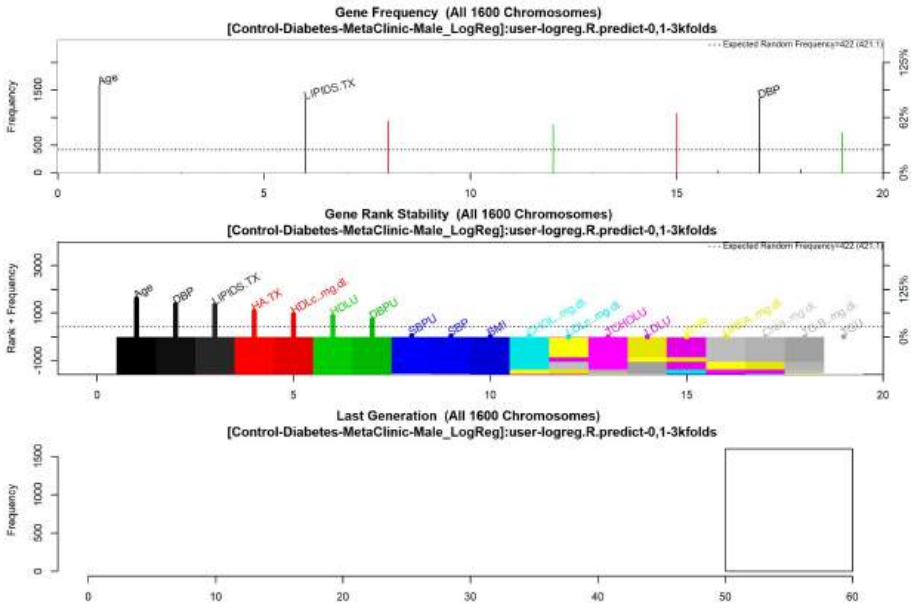


Figura 5.48: La frecuencia genética y la estabilidad del rango genético Control-Diabetes-Hombres-LogReg.

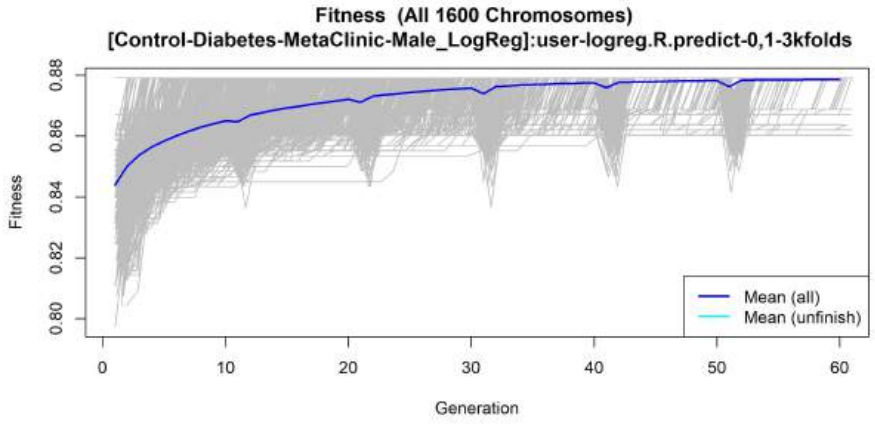


Figura 5.49: Evolución del ajuste en Siglo XXI_Control-Diabetes-Hombres-LogReg.

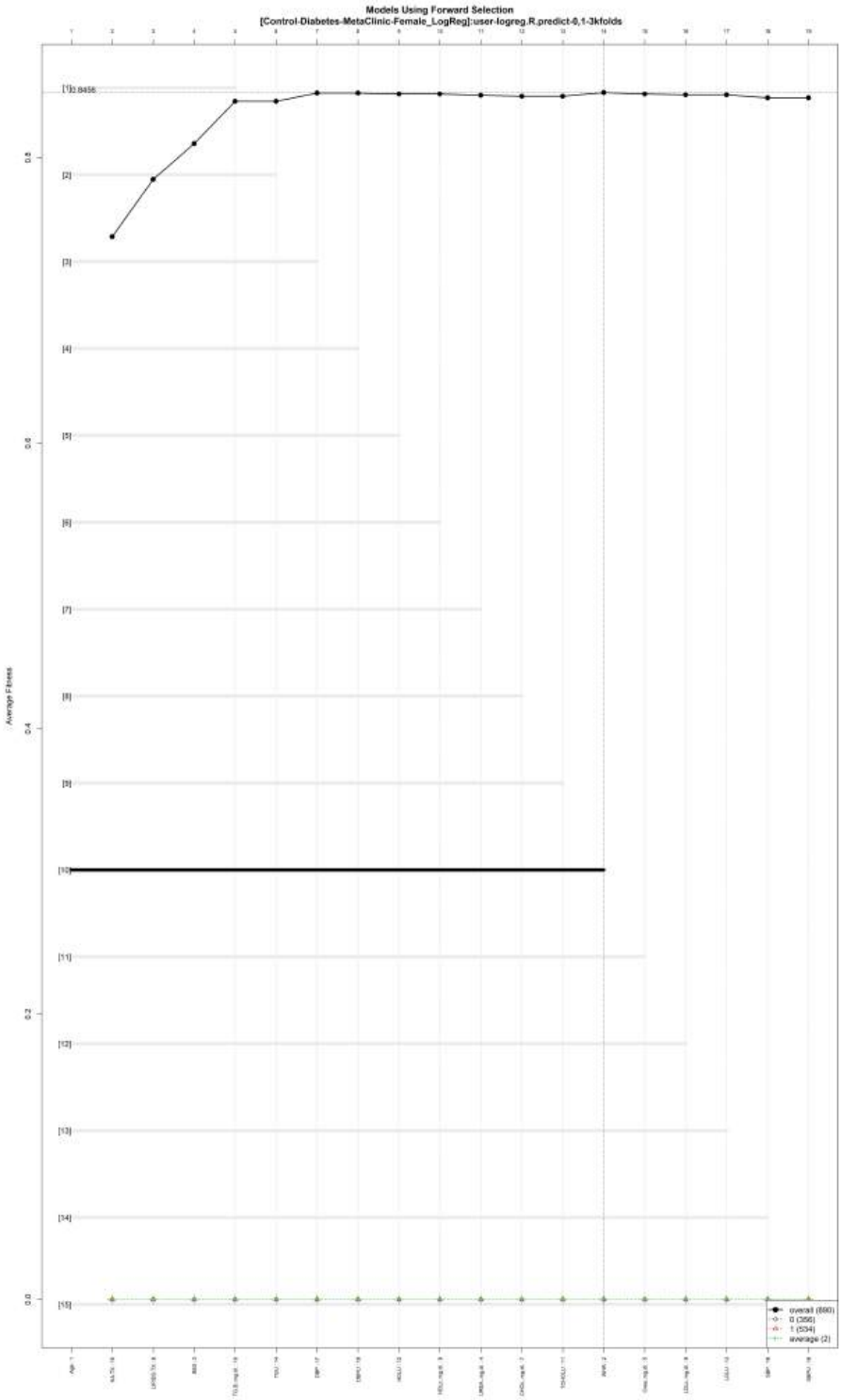


Figura 5.50: SigloXXI_Control-Diabetes-Mujeres FSM-LogReg.

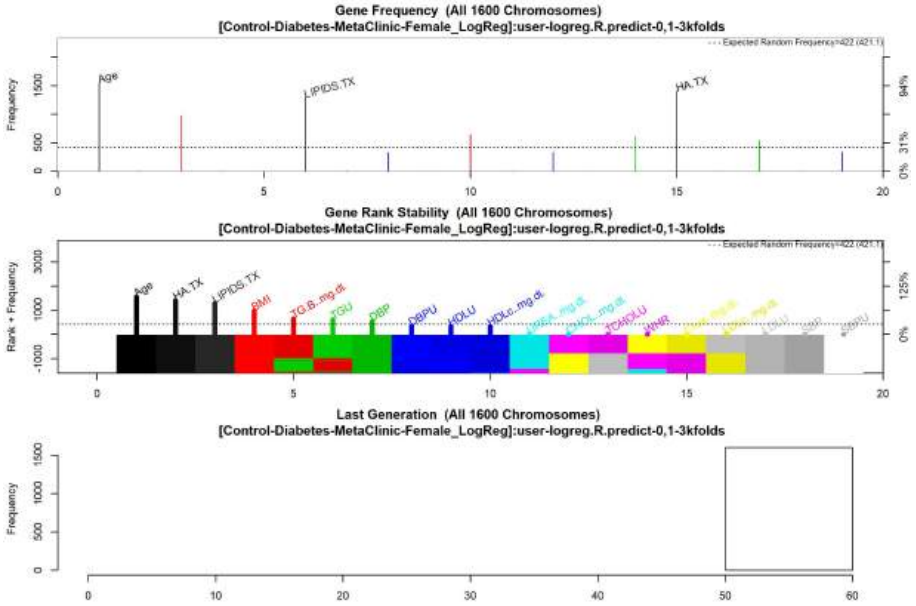


Figura 5.51: La frecuencia genética y la estabilidad del rango genético Control-Diabetes-Mujeres-LogReg.

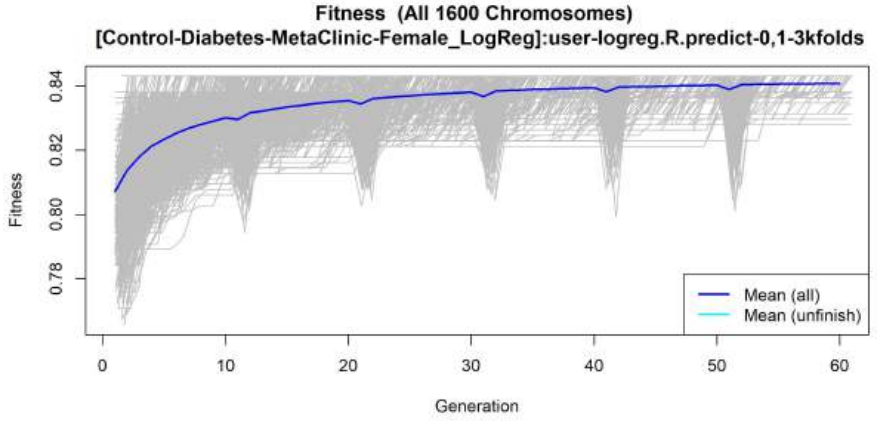


Figura 5.52: Evolución del ajuste en Siglo XXI_Control-Diabetes-Mujeres-LogReg.

Tabla 5.23: Características del resultado LR Galgo Masculino/Femenino.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
"SBP"	"TGU"
"Creatinine"	"Creatinine"
"TGU"	"WHR"
"Cholesterol"	"LDLc"
	"LDLU"
	"SBP"
	"SBPU"
	"TCHOLU"
	"Cholesterol"
	"BMI"
	"LIPIDS-TX"
	"Triglycerides"
	"Age"

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML LR, con 1600 Big Bangs y 60 generaciones.

características se obtuvieron con una ACC promedio de 0.793, como se muestra en la figura 5.53. Debido al bajo número de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.54 y la figura de ajuste 5.55.

La frecuencia genética y la estabilidad del rango genético de los modelos se determinan utilizando GA con NNET para seleccionar las características principales en el conjunto de datos. El gráfico superior muestra la frecuencia genética de veces que aparece una característica en los modelos. El gráfico del medio, ordenado por rango, muestra la estabilidad del rango genético y la frecuencia de cada característica en los modelos. El gráfico inferior muestra el número de generaciones empleadas, que en este caso es 1600.

Las características obtenidas por GALGO para el conjunto de datos masculino y el mejor modelo de selección directa, como se muestra en la tabla 5.25, 15 de 20 características se obtuvieron con una ACC promedio de 0.8074, como se muestra en la figura 5.56. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.57 y la figura de ajuste 5.58.

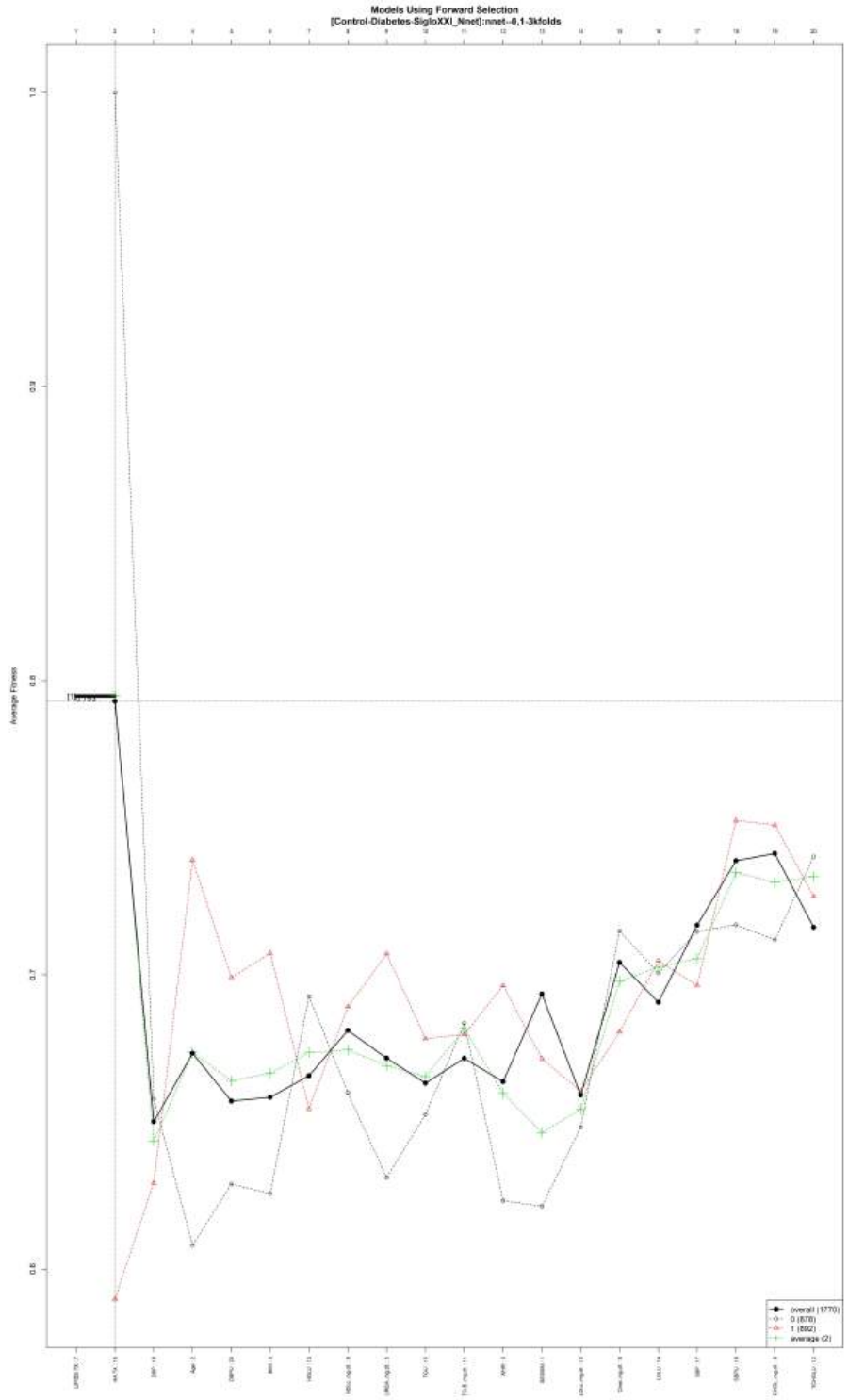


Figura 5.53: SigloXXI_Control-Diabetes FSM-NNET.

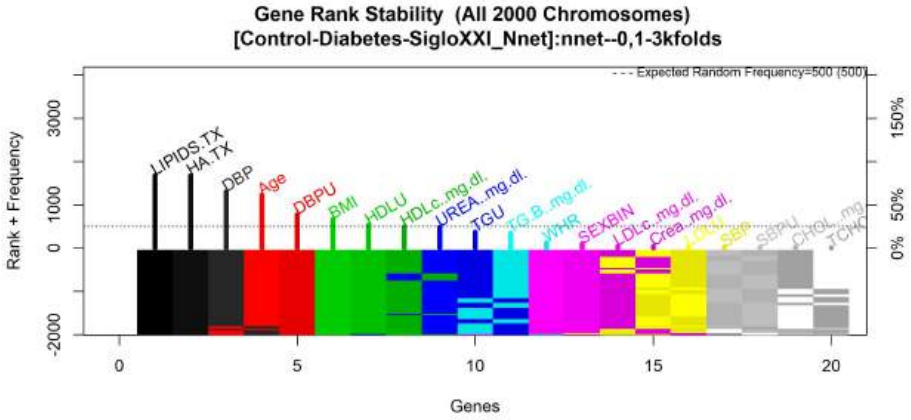


Figura 5.54: La frecuencia genética y la estabilidad del rango genético Control-Diabetes-NNET.

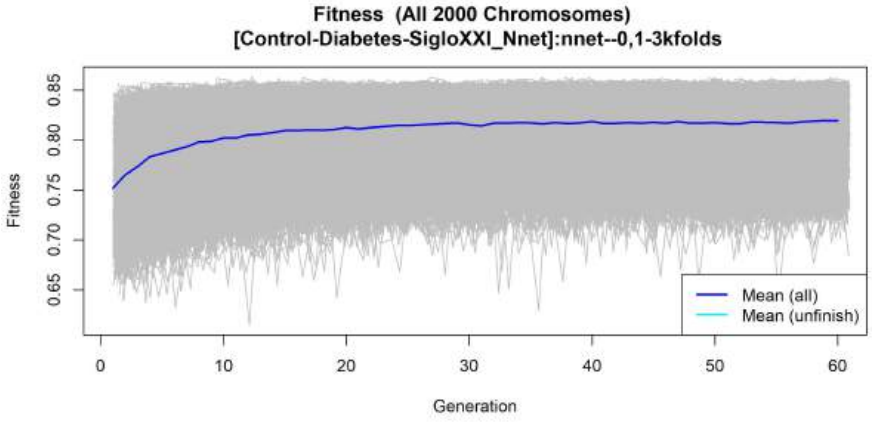


Figura 5.55: Evolución del ajuste en Siglo XXI_Control-Diabetes-NNET.

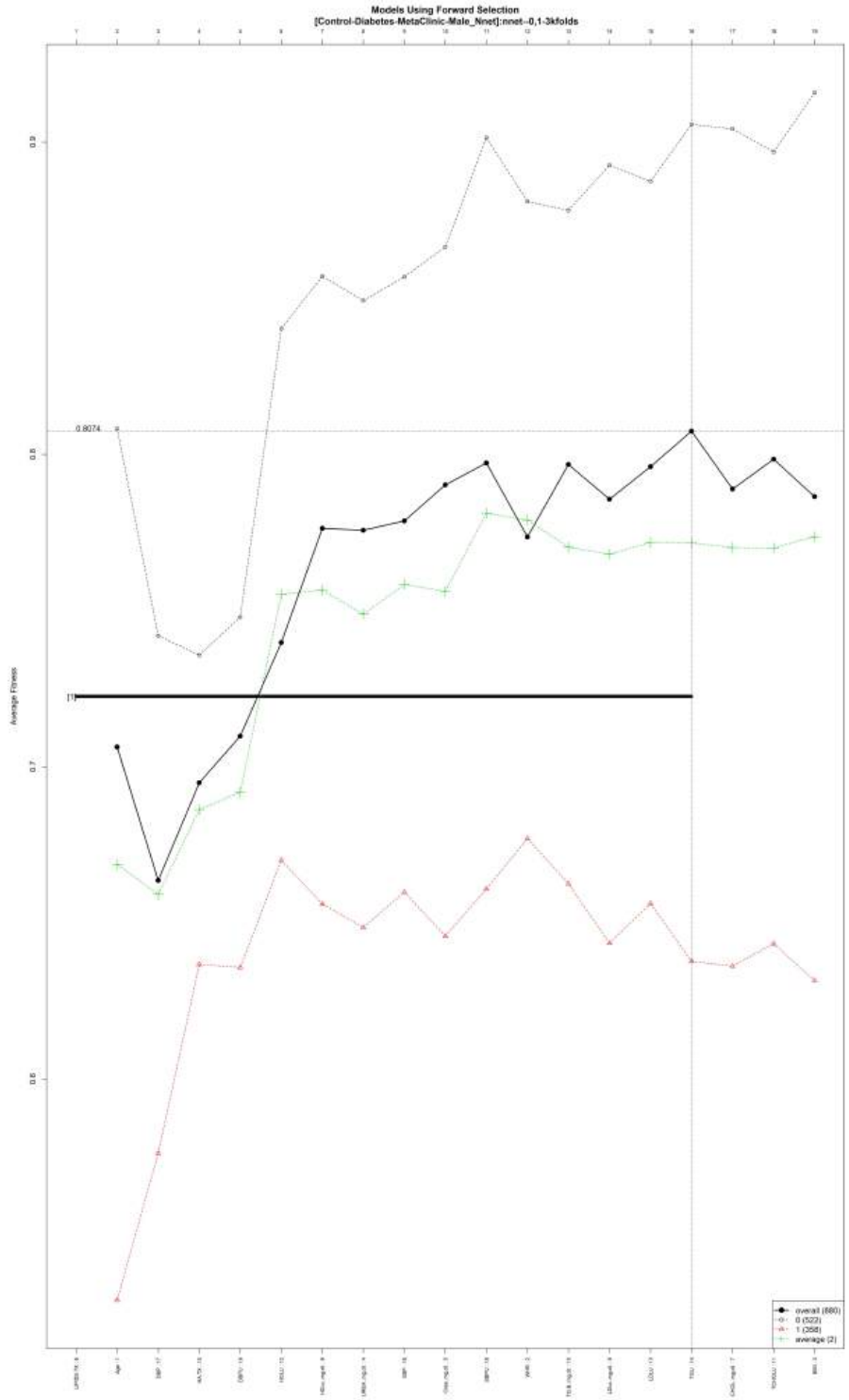


Figura 5.56: SigloXXI_Control-Diabetes-Hombres FSM-NNET.

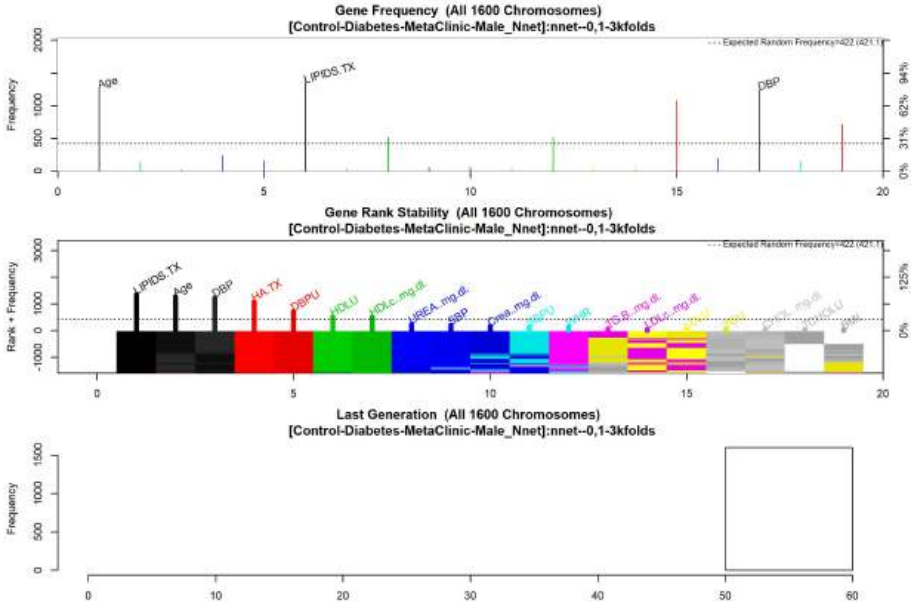


Figura 5.57: La frecuencia genética y la estabilidad del rango genético Control-Diabetes-Hombres-NNET.

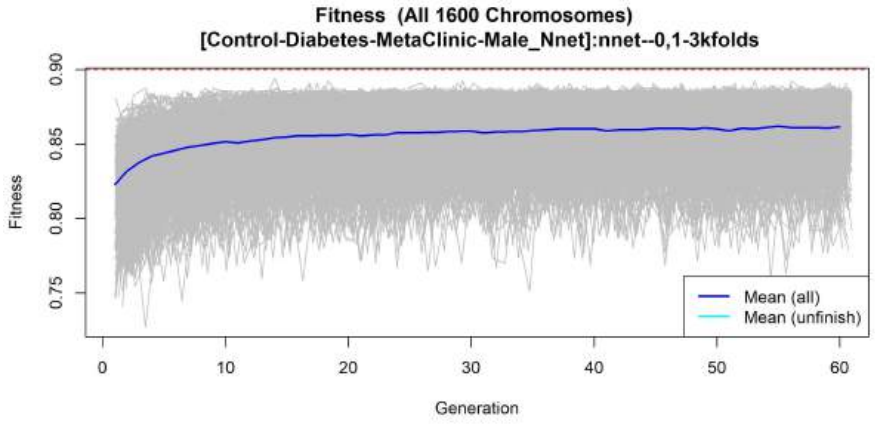


Figura 5.58: Evolución del ajuste en Siglo XXI_Control-Diabetes-Hombres-NNET.

Las características obtenidas por GALGO para el conjunto de datos femenino y el mejor modelo de selección directa, como se muestra en la tabla 5.25, 13 de 20 características se obtuvieron con una ACC promedio de 0.782, como se muestra en la figura 5.59. Debido a la baja cantidad de características en comparación con la gran cantidad de muestras incluidas en el modelo, estos resultados muestran que las características incluidas son extremadamente significativas, como se muestra en la figura de estabilidad de rango genético 5.60 y la figura de ajuste 5.61.

Tabla 5.24: Características del resultado NNET Galgo Siglo XXI General.

“Creatinine”, “TGU”

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML NNET, con 2000 Big Bangs y 60 generaciones.

Tabla 5.25: Características del resultado NNET Galgo Masculino/Femenino.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
“Creatinine”	“TGU”
“SBP”	“Creatinine”
“TGU”	“SBP”
“SBPU”	“WHR”
“TCHOLU”	“LDLU”
“Cholesterol”	“LDLc”
“BMI”	“SBPU”
“HA-TX”	“Cholesterol”
“Urea”	“BMI”
“DBP”	“TCHOLU”
“Age”	“Age”
“LDLc”	“LIPIDS-TX”
“HDL”	“Triglycerides”
“HDLU”	
“LDLU”	

Todas las características de esta tabla se obtuvieron utilizando GALGO y el modelo ML NNET, con 1600 Big Bangs y 60 generaciones.

LASSO

Este selector proporcionó un enfoque más simple:

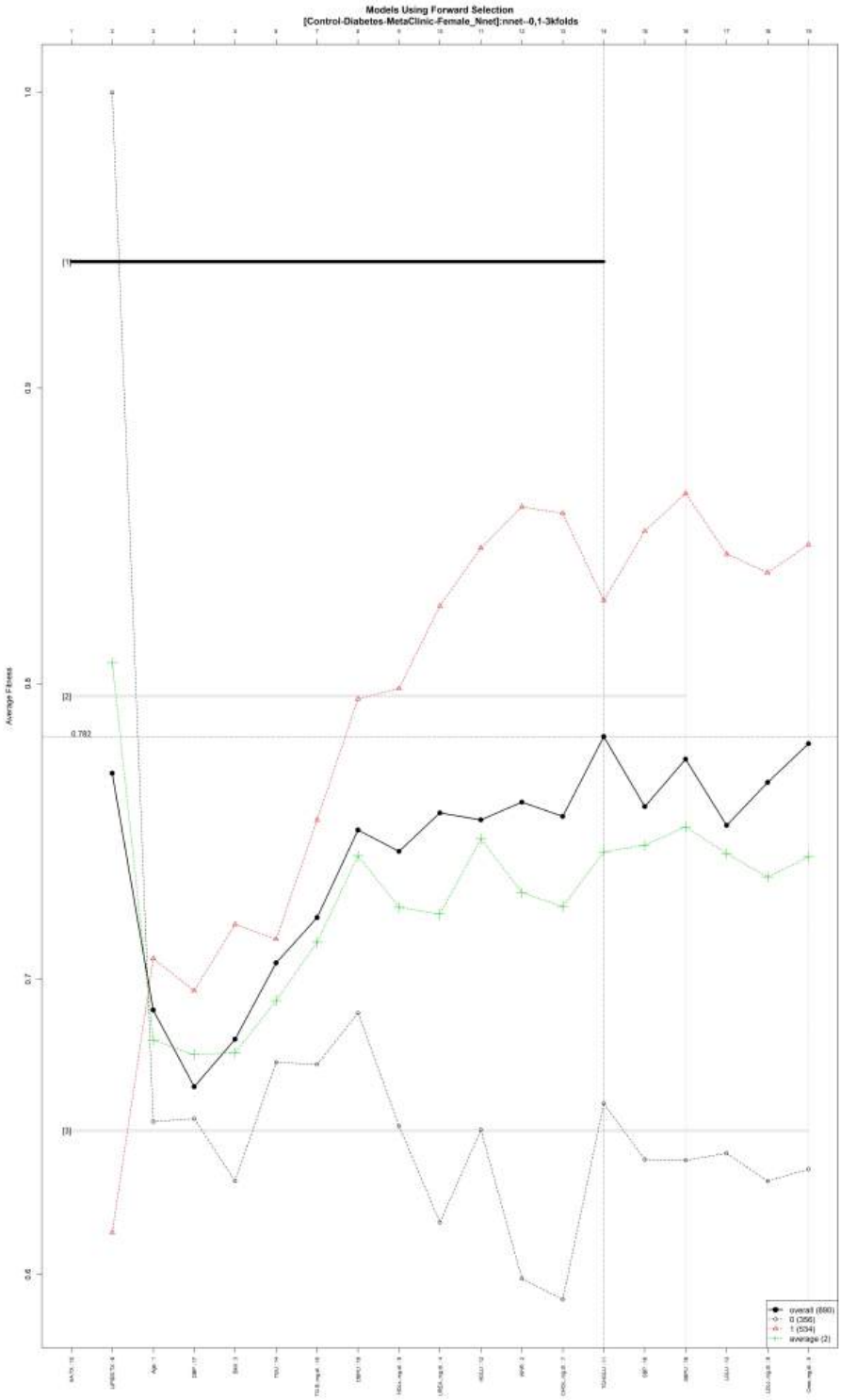


Figura 5.59: SigloXXI_Control-Diabetes-Mujeres FSM-NNET.

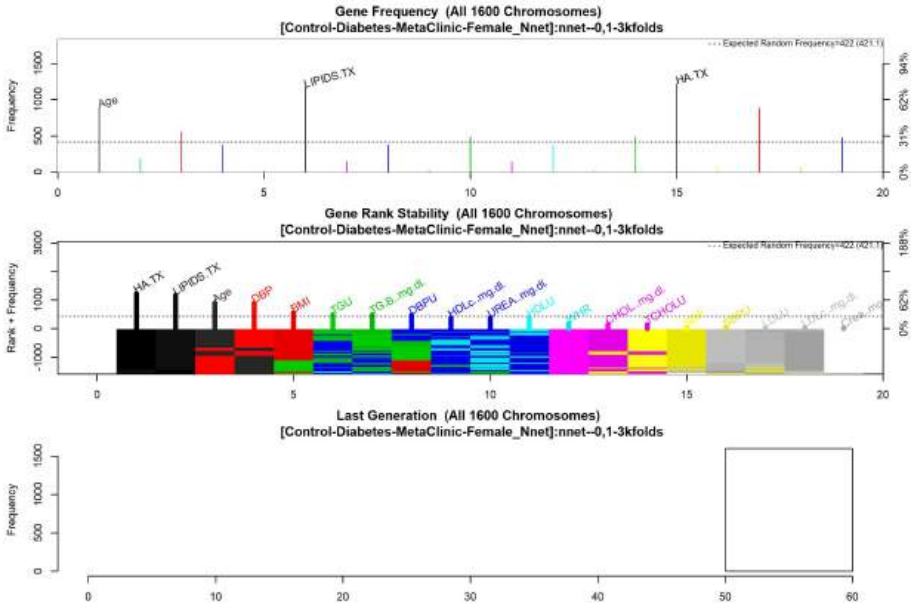


Figura 5.60: La frecuencia genética y la estabilidad del rango genético Control-Diabetes-Mujeres-NNET.

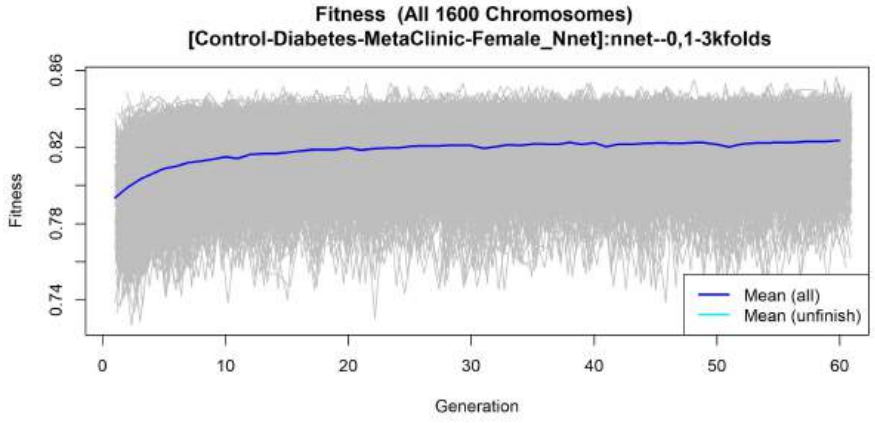


Figura 5.61: Evolución del ajuste en Siglo XXI_Control-Diabetes-Mujeres-NNET.

Tabla 5.26: Características del Resultado LASSO Siglo XXI General.

Siglo XXI General
"Sex", "Age", "WHR", "BMI", "Urea", "LIPIDS-TX", "HDL", "Triglycerides", "HA-TX", "DBP", "SBPU"

Todas las características de esta tabla se obtuvieron utilizando el método de red LASSO Elastic en el conjunto de datos general.

Tabla 5.27: Características del resultado LASSO Masculino/Femenino.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
“Age”	“Age”
“WHR”	“BMI”
“Urea”	“Urea”
“LIPIDS-TX”	“LIPIDS-TX”
“HDL”	“HDL”
“Triglycerides”	“Triglycerides”
“HA-TX”	“HA-TX”
“DBP”	“DBP”
“SBPU”	

Todas las características de esta tabla se obtuvieron utilizando el método de red LASSO Elastic en los conjuntos de datos masculinos y femeninos.

RFE

RFE se implementa con 3 métodos de selección diferentes: LR, SVM con kernel lineal y RF.

Tabla 5.28: Resultados de RFE Siglo XXI en general con LR, SVM y RF.

Siglo XXI General		
LR	SVM	RF
“Sex”	“Sex”	“Age”
“WHR”	“WHR”	“LIPIDS-TX”
“CHOL (mg/dl)”	“Crea (mg/dl)”	“TG-B (mg/dl)”
“TCHOLU”	“CHOL (mg/dl)”	“DBP”
“HA-TX”	“TCHOLU”	“DBPU”
“SBP”	“SBP”	
“DBP”	“DBP”	
“SBPU”	“SBPU”	
“DBPU”	“DBPU”	

Todas las características de esta tabla se obtuvieron utilizando RFE con LR, SVM y RF como estimadores en el conjunto de datos general.

5.0.2. Resultados de métricas del modelo de conjunto

Según los resultados de la implementación de métricas de rendimiento para los diferentes modelos (LR, SVM, RF), el modelo RF funcionó mejor en múltiples métricas en comparación con LR y SVM. RF tiene la *Sensitivity* más alta

Tabla 5.29: Características del resultado RFE Masculino/Femenino LR.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
“CHOL (mg/dl)”	“WHR”
“TCHOLU”	“CHOL (mg/dl)”
“SBP”	“TG-B (mg/dl)”
“DBP”	“TCHOLU”
“SBPU”	“TGU”
	“SBP”
	“DBP”
	“SBPU”

Todas las características de esta tabla se obtuvieron utilizando RFE con LR como estimador en el conjunto de datos masculino y femenino.

Tabla 5.30: Características de resultados de SVM Masculino/Femenino de RFE.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
“Age”	“Age”
“WHR”	“WHR”
“Crea (mg/dl)”	“BMI”
“SBP”	“Crea (mg/dl)”
“DBP”	“CHOL (mg/dl)”
“SBPU”	“TCHOLU”
“DBPU”	“SBP”
	“DBP”
	“SBPU”
	“DBPU”

Todas las características de esta tabla se obtuvieron utilizando RFE con SVM con kernel lineal como estimador en el conjunto de datos masculino y femenino.

Tabla 5.31: Características del resultado RFE Masculino/Femenino.

Conjunto de datos masculino Siglo XXI	Conjunto de datos femenino Siglo XXI
“Age”	“Age”
“BMI”	“WHR”
“LIPIDS-TX”	“BMI”
“DBP”	“TG-B (mg/dl)”
“DBPU”	“DBP”

Todas las características de esta tabla se obtuvieron utilizando RFE con RF como estimador en el conjunto de datos masculino y femenino.

Tabla 5.32: Matriz de confusión resultante del modelo Ensamble con la Selección de características RFE en el conjunto de datos Overall Siglo XXI incluyendo prueba ciega.

Métrica	LR	SVM	RF
Sensitivity	0.8127	0.8137	0.8682
Specificity	0.8640	0.8543	0.8958
Precision	0.8645	0.8526	0.8924
Negative Predictive Value	0.8120	0.8158	0.8722
False Positive Rate	0.1360	0.1457	0.1042
False Discovery Rate	0.1355	0.1474	0.1076
False Negative Rate	0.1873	0.1863	0.1318
Accuracy	0.8375	0.8337	0.8820
F1-Score	0.8378	0.8327	0.8802

de 0.8682, lo que indica una mejor capacidad para identificar correctamente los casos positivos en comparación con LR 0.8127 y SVM 0.8137. En *Specificity* con 0.8958 supera a LR 0.8640 y SVM 0.8543, demostrando una mejor capacidad para identificar correctamente los casos negativos. RF tiene la *Precision* más alta de 0.8924, lo que indica una tasa más baja de falsos positivos en comparación con LR 0.8645 y SVM 0.8526. RF logró la ACC más alta de 0.8820, lo que sugiere un rendimiento general superior en la clasificación correcta de casos positivos y negativos, en comparación con LR 0.8375 y SVM 0.8337. Por último, RF tiene la *F1-Score* más alta: 0.8802, que es la media armónica de *Precision* y *Sensitivity*. Esto sugiere un mejor equilibrio entre *Precision* y *Sensitivity* en comparación con LR 0.8378 y SVM 0.8327.

El modelo RF demuestra un rendimiento superior en términos de *Sensitivity*, *Specificity*, *Precision*, ACC y *F1-Score*. En este caso, parece ser un modelo más eficaz para la tarea o conjunto de datos determinado en comparación con los modelos LR y Support Vector Machine. Las características contenidas en RF son: “Age”, “LIPIDS-TX’, “Triglycerides”, “DBP”, “DBPU”.

En este caso, tanto el modelo LR como el de RF funcionaron igual de bien en todas las métricas. *Sensitivity* de 0.8314 que indica una capacidad igual para identificar correctamente los casos positivos. SVM 0.8452 tuvo un desempeño ligeramente mejor en este aspecto. LR y RF superaron a SVM con 0.9167 sobre 0.8119 en *Specificity*, lo que demuestra una mejor capacidad para identificar correctamente los casos negativos. LR y RF lograron la *Precision* más alta de 0.9533, lo que indica una tasa más baja de falsos positivos en comparación con SVM 0.8733. LR y RF tienen la misma ACC de 0.8594, que es mayor que SVM 0.8320, lo que sugiere un mejor rendimiento general en la clasificación correcta

Tabla 5.33: Matriz de confusión resultante del modelo Ensamble con la Selección de características RFE en el conjunto de datos Masculino Siglo XXI incluyendo prueba ciega.

Métrica	LR	SVM	RF
Sensitivity	0.8314	0.8452	0.8314
Specificity	0.9167	0.8119	0.9167
Precision	0.9533	0.8733	0.9533
Negative Predictive Value	0.7264	0.7736	0.7264
False Positive Rate	0.0833	0.1881	0.0833
False Discovery Rate	0.0467	0.1267	0.0467
False Negative Rate	0.1686	0.1548	0.1686
Accuracy	0.8594	0.8320	0.8594
F1-Score	0.8882	0.8590	0.8882

de casos positivos y negativos. LR y RF tienen la misma *F1-Score* de 0.8882, que es superior a SVM 0.8590, lo que indica un mejor equilibrio entre *Precision* y *Sensitivity*.

Según las métricas presentadas, tanto el modelo LR como el de RF funcionan igualmente bien en el conjunto de datos masculino, lo que demuestra alta *Precision*, *Sensitivity*, *Specificity*, ACC y *F1-Score*. Estos modelos parecen ser más efectivos para la tarea o conjunto de datos determinado en comparación con el modelo de SVM. Las características del modelo en estos casos son: para RFE-LR “Cholesterol”, “TCHOLU”, “SBP”, “DBP”, “SBPU” y para RFE-RF “Age”, “BMI”, “LIPIDS-TX”, “DBP”, “DBPU”.

En el conjunto de datos femenino, el modelo LR parece funcionar mejor en general. LR tiene la *Sensitivity* más alta entre los modelos 0.7545, lo que indica una mejor capacidad para identificar correctamente los casos positivos. SVM 0.7455 y RF 0.7333 tienen ambas el valor de *Sensitivity* ligeramente más bajo. LR logró la *Specificity* más alta, 0.8800, lo que indica una mejor capacidad para identificar correctamente los casos negativos. SVM 0.8733 le sigue de cerca, mientras que RF 0.7941 tiene una *Specificity* menor. LR tiene la *Precision* más alta, 0.8218, lo que indica una tasa más baja de falsos positivos. Le sigue SVM 0.8119 y RF 0.6535 tiene una *Precision* menor. LR tiene la ACC más alta, 0.8269, lo que indica un mejor rendimiento general en la clasificación correcta de casos positivos y negativos. Le sigue SVM 0.8192 y RF 0.7731 tiene una ACC menor. LR tiene la *F1-Score* más alta, 0.7867, que equilibra la *Precision* y la *Sensitivity* de forma eficaz. Le sigue SVM 0.7773 y RF 0.6911 tiene una *F1-Score* más baja.

Las características implementadas con RFE-LR en este conjunto de datos son: “WHR”, “Cholesterol”, “Triglycerides”, “TCHOLU”, “TGU”, “SBP”, “DBP”, “SBPU”.

Tabla 5.34: Matriz de confusión resultante del modelo Ensamble con la Selección de características RFE en el conjunto de datos Femenino Siglo XXI incluyendo prueba ciega.

Métrica	LR	SVM	RF
Sensitivity	0.7545	0.7455	0.7333
Specificity	0.8800	0.8733	0.7941
Precision	0.8218	0.8119	0.6535
Negative Predictive Value	0.8302	0.8239	0.8491
False Positive Rate	0.1200	0.1267	0.2059
False Discovery Rate	0.1782	0.1881	0.3465
False Negative Rate	0.2455	0.2545	0.2667
Accuracy	0.8269	0.8192	0.7731
F1-Score	0.7867	0.7773	0.6911

Tabla 5.35: Matriz de confusión resultante del modelo Ensamble con la Selección de características Galgo en el conjunto de datos Overall Siglo XXI incluyendo prueba ciega.

Métrica	KNN	Nearcent	LR	SVM	ANN
Sensitivity	0.8028	0.8377	0.8690	0.8511	0.5781
Specificity	0.9167	0.8849	0.8792	0.8902	0.6054
Precision	0.9243	0.8845	0.8725	0.8884	0.5896
Negative Predictive Value	0.7857	0.8383	0.8759	0.8534	0.5940
False Positive Rate	0.0833	0.1151	0.1208	0.1098	0.3946
False Discovery Rate	0.0757	0.1155	0.1275	0.1116	0.4104
False Negative Rate	0.1972	0.1623	0.1310	0.1489	0.4219
Accuracy	0.8530	0.8607	0.8743	0.8704	0.5919
F1-Score	0.8593	0.8605	0.8708	0.8694	0.5838

En el conjunto de datos general, el modelo LR parece tener el mejor rendimiento. LR tiene la *Sensitivity* más alta, 0.8690, lo que indica una mejor capacidad para identificar correctamente los casos positivos. Le siguen SVM 0.8511, Nearcent 0.8377, KNN 0.8028 y ANN 0.5781 en orden descendente. KNN logró la *Specificity* más alta, 0.9167, lo que indica una mejor capacidad para identificar correctamente los casos negativos. Le siguen SVM 0.8902, Nearcent 0.8849, LR 0.8792 y ANN 0.6054 en orden descendente. KNN tiene la *Precision* más alta, 0.9243, lo que indica una tasa más baja de falsos positivos. SVM 0.8884, Nearcent 0.8845, LR 0.8725 y ANN 0.5896 siguen en orden descendente. LR tiene la ACC más alta, 0.8743, lo que indica un mejor rendimiento general en la clasificación correcta de casos positivos y negativos. Le siguen SVM 0.8704, Nearcent 0.8607, KNN 0.8530 y ANN 0.5919 en orden descendente. LR tiene la *F1-Score* más alta, 0.8708, que equilibra la *Precision* y la *Sensitivity* de forma eficaz. Le siguen SVM 0.8694, Nearcent 0.8605, KNN 0.8593 y ANN 0.5838 en orden descendente.

El modelo LR supera a los modelos KNN, Nearcent, SVM y ANN con las siguientes características: “Sexo”, “Creatinine”, “SBP”, “TGU”, “TCHOLU”, “Cholesterol”, “SBPU”, “LDLc”, “LDLU”, “HA-TX”, “DBP”, “WHR”, “LIPIDS-TX”, “HDL”, “Triglycerides”, “HDLU”, “Age”, “BMI”.

Tabla 5.36: Matriz de confusión resultante del modelo Ensamble con la Selección de características Galgo en el conjunto de datos Masculino Siglo XXI incluyendo prueba ciega.

Métrica	KNN	Nearcent	LR	SVM	ANN
Sensitivity	0.8373	0.6856	0.6859	0.8235	0.8443
Specificity	0.8778	0.7258	0.7077	0.8837	0.8989
Precision	0.9267	0.8867	0.8733	0.9333	0.9400
Negative Predictive Value	0.7453	0.4245	0.4340	0.7170	0.7547
False Positive Rate	0.1222	0.2742	0.2923	0.1163	0.1011
False Discovery Rate	0.0733	0.1133	0.1267	0.0667	0.0600
False Negative Rate	0.1627	0.3144	0.3141	0.1765	0.1557
Accuracy	0.8516	0.6953	0.6914	0.8438	0.8633
F1-Score	0.8797	0.7733	0.7683	0.8750	0.8896

En el conjunto de datos masculino, el modelo ANN parece tener el mejor rendimiento en general. Aquí hay una explicación basada en métricas clave: ANN tiene la *Sensitivity* más alta 0.8443, lo que indica una mejor capacidad para identificar correctamente los casos positivos. KNN (0.8373), SVM (0.8235), LR (0.6859) y Nearcent (0.6856) le siguen en orden descendente. ANN logró la *Specificity* más

alta, 0.8989, lo que indica una mejor capacidad para identificar correctamente los casos negativos. SVM 0.8837, KNN 0.8778, LR 0.7077 y Nearcent 0.7258 le siguen en orden descendente. ANN tiene la *Precision* más alta, 0.9400, lo que indica una tasa más baja de falsos positivos. SVM 0.9333, KNN 0.9267, Nearcent 0.8867 y LR 0.8733 siguen en orden descendente. ANN tiene la ACC más alta, 0.8633, lo que indica un mejor rendimiento general en la clasificación correcta de casos positivos y negativos. KNN 0.8516, SVM 0.8438, Nearcent 0.6953 y LR 0.6914 siguen en orden descendente. ANN tiene la *F1-Score* más alta, 0.8896, que equilibra la *Precision* y la *Sensitivity* de forma eficaz. KNN 0.8797, SVM 0.8750, Nearcent 0.7733 y LR 0.7683 siguen en orden descendente.

El modelo ANN supera a los modelos KNN, Nearcent, SVM y LR en el conjunto de datos masculino con estas características: “Creatinine”, “SBP”, “TGU”, “SBPU”, “TCHOLU”, “Cholesterol”, “BMI”, ” HA-TX”, “Urea”, “DBP”, “Age”, “LDLc”, “HDL”, “HDLU”, “LDLU”.

Tabla 5.37: Matriz de confusión resultante del modelo Ensamble con la Selección de características Galgo en el conjunto de datos Femenino Siglo XXI incluyendo prueba ciega.

Métrica	KNN	Nearcent	LR	SVM	ANN
Sensitivity	0.7658	0.6804	0.7719	0.7000	0.7748
Specificity	0.8926	0.7853	0.9110	0.8786	0.8993
Precision	0.8416	0.6535	0.8713	0.8317	0.8515
Negative Predictive Value	0.8365	0.8050	0.8365	0.7736	0.8428
False Positive Rate	0.1074	0.2147	0.0890	0.1214	0.1007
False Discovery Rate	0.1584	0.3465	0.1287	0.1683	0.1485
False Negative Rate	0.2342	0.3196	0.2281	0.3000	0.2252
Accuracy	0.8385	0.7462	0.8500	0.7962	0.8462
F1-Score	0.8019	0.6667	0.8186	0.7602	0.8113

En el conjunto de datos femenino, el modelo LR parece tener el mejor rendimiento en general. LR tiene la *Sensitivity* más alta 0.7719, lo que indica una mejor capacidad para identificar correctamente los casos positivos. KNN 0.7658, ANN 0.7748, SVM 0.7000 y Nearcent 0.6804 le siguen en orden descendente. LR logró la *Specificity* más alta, 0.9110, lo que indica una mejor capacidad para identificar correctamente los casos negativos. ANN 0.8993, KNN 0.8926, SVM 0.8786 y Nearcent 0.7853 le siguen en orden descendente. LR tiene la *Precision* más alta, 0.8713, lo que indica una tasa más baja de falsos positivos. ANN 0.8515, KNN 0.8416, SVM 0.8317 y Nearcent 0.6535 le siguen en orden descendente. LR 0.8500 tiene la mayor ACC, lo que indica un mejor rendimiento general en la

clasificación correcta de casos positivos y negativos. ANN 0.8462, KNN 0.8385, SVM 0.7962 y Nearcent 0.7462 le siguen en orden descendente. LR tiene la *F1-Score* más alta, 0.8186, que equilibra la *Precision* y la *Sensitivity* de forma eficaz. ANN 0.8113, KNN 0.8019, SVM 0.7602 y Nearcent 0.6667 le siguen en orden descendente.

El modelo LR supera a los modelos KNN, Nearcent, SVM y ANN en el conjunto de datos femenino con estas características: "TGU", "Creatinine", "WHR", "LDLc", "LDLU", "SBP", "SBPU", "TCHOLU", "Cholesterol", "BMI", "LIPIDS-TX", "Triglycerides", "Age".

Tabla 5.38: Matriz de confusión resultante del modelo Ensamble con la Selección de características Galgo en el conjunto de datos Overall / Masculino / Femenino Siglo XXI incluyendo prueba ciega.

Métrica	General	Masculino	Femenino
Sensitivity	0.8760	0.8443	0.7500
Specificity	0.8801	0.8989	0.8851
Precision	0.8725	0.9400	0.8317
Negative Predictive Value	0.8835	0.7547	0.8239
False Positive Rate	0.1199	0.1011	0.1149
False Discovery Rate	0.1275	0.0600	0.1683
False Negative Rate	0.1240	0.1557	0.2500
Accuracy	0.8781	0.8633	0.8269
F1-Score	0.8743	0.8896	0.7887

El modelo de conjunto con selección de características LASSO funciona bien en todos los conjuntos de datos, con un rendimiento particularmente notable en el conjunto de datos masculino.

Para el conjunto de datos general, el modelo muestra un rendimiento equilibrado con una *Sensitivity*, *Specificity*, *Precision*, ACC y *F1-Score* relativamente altas. Funciona bien en la identificación de casos tanto positivos como negativos. Para el conjunto de datos masculino, el modelo funciona excepcionalmente bien, especialmente en términos de *Sensitivity* y *Precision*. Tiene una alta capacidad para identificar correctamente los casos positivos (*Sensitivity*) y una baja tasa de falsos positivos (*Precision*). Para el conjunto de datos femenino, el modelo tiene menor *Sensitivity* y *Precision* en comparación con el conjunto de datos masculino, lo que da como resultado una ACC general y una *F1-Score* ligeramente inferiores. Puede que valga la pena explorar formas de mejorar el rendimiento específicamente en el conjunto de datos femenino.

La implementación del conjunto de datos masculino tiene estas características: "Age", "WHR", "Urea", "LIPIDS-TX", "HDL", "Triglycerides", "HA-TX", "DBP",

"SBPU".

El uso del AIC sobre el AICc se debe a que ha demostrado ser una mejor herramienta en modelos con más observaciones que características. Se representa de la siguiente manera:

$$AIC = 2k - 2\ln(\hat{L}). \tag{5.9}$$

Estos conjuntos fueron seleccionados por el AICc más bajo implementado en el conjunto:

Conjunto de datos General RFE-RF: "Age", "LIPIDS-TX", "Triglycerides", "DBP", "DBPU" con una ACC 0.8820. Conjunto de datos LASSO: "Age", "WHR", "Urea", "LIPIDS-TX", "HDL", "Triglycerides", "HA-TX", "DBP", "SBPU", *Accuracy* 0.8633. Conjunto de datos Femenino RFE-LR: "WHR", "Cholesterol", "Triglycerides", "TCHOLU", "TGU", "SBP", "DBP", "SBPU", *Accuracy* 0.8269.

5.0.3. Rangos

Los rangos presentados en esta sección sirven como base para establecer variaciones fisiológicas normales e identificar posibles valores atípicos o anormales dentro de diferentes conjuntos de datos. Comprender estos rangos es esencial para que los médicos, investigadores y profesionales de la salud interpreten las evaluaciones de salud, tomen decisiones de diagnóstico informadas y adapten las intervenciones a las necesidades individualizadas de los pacientes. El carácter integral de estos rangos garantiza una caracterización exhaustiva de los indicadores de salud investigados.

El rango de edad representa el lapso de edades observadas dentro de cada conjunto de datos. Incluye los valores mínimo y máximo, así como cuartiles, lo que proporciona información sobre la tendencia central y la dispersión de la edad entre los casos. Los rangos de WHR abarcan la variabilidad en la relación entre la circunferencia de la cintura y la circunferencia de la cadera. Los cuartiles y los valores extremos ofrecen una visión integral de la distribución de esta medida antropométrica, crucial para evaluar la distribución de la grasa corporal. Los rangos de BMI representan la distribución de los índices de masa corporal dentro de cada conjunto de datos. Incluyen cuartiles y valores extremos, lo que permite comprender la variabilidad del peso corporal en relación con la altura entre los casos. Los rangos de niveles de urea y creatinina brindan información sobre la distribución de estos indicadores de la función renal. Los valores mínimo, máximo y cuartil ofrecen una visión general completa de la variabilidad de estos marcadores bioquímicos. Los rangos del perfil de colesterol abarcan los niveles de colesterol total, colesterol HDL, colesterol LDL y triglicéridos. Estos rangos ofrecen una perspectiva detallada sobre la distribución de los marcadores lipídicos, cruciales para las evaluaciones de la salud cardiovascular. Los rangos de

presión arterial incluyen valores sistólicos (SBP, SBPU) y diastólicos (DBP, DBPU). Los cuartiles y los valores extremos proporcionan una comprensión integral de la variabilidad de la presión arterial dentro de cada conjunto de datos, crucial para la evaluación de la salud cardiovascular.

Tabla 5.39: Rangos en el conjunto de datos Overall (controles y casos).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	30	45	52	60	82
WHR	0.72	0.87	0.92	0.97	1.11
BMI	17.00	25.26	27.90	31.24	40.20
Urea	6	24	28	36	54
Creatinine	0.36	0.68	0.79	0.93	1.30
Cholesterol	75.00	167.00	195.05	230.00	324.10
HDL	10	35	43	52	77
LDLc	45.0	112.0	136.0	162.9	239.0
Triglycerides	32	116	156	219	371
TCHOLU	82	161	187	215	296
HDLU	12	36	44	52	76
LDLU	45	107	127	149	212
TGU	32	110	148	207	350
SBP	80	110	120	130	160
SBPU	80	110	120	130	160
DBP	50	70	80	85	105
DBPU	55	70	80	80	95

Cada tabla representa un conjunto de datos específico (general, masculino, femenino) con controles y casos, y los rangos de las variables difieren entre estos conjuntos de datos. Por ejemplo, el conjunto de datos masculino tiende a tener valores más altos para variables como BMI, Crea y LDLc en comparación con el conjunto de datos femenino. Los conjuntos de datos masculinos y femeninos incluyen variables específicas de género como TCHOLU, HDLU, LDLU y TGU, que están relacionadas con los niveles de Cholesterol en la urea. Estas variables no están presentes en el conjunto de datos general. Las variables de presión arterial (SBP, SBPU, DBP, DBPU) exhiben variabilidad entre los conjuntos de datos, lo que refleja diferencias relacionadas con el género en las distribuciones de la presión arterial. El conjunto de datos de Mujeres generalmente muestra valores más altos de CHOL, HDLc, LDLc, TG.B y TCHOLU en comparación con el conjunto de datos de Hombres. La distribución por edad entre los conjuntos de datos varía: el conjunto de datos masculino tiene valores de mediana y cuartil superior ligera-

Tabla 5.40: Rangos en el conjunto de datos masculino (controles y casos).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	30	44	50	59	81
WHR	0.82	0.92	0.95	0.99	1.09
BMI	18.400	25.245	27.400	29.950	36.810
Urea	6	24	30	36	54
Creatinine	0.44	0.77	0.88	1.01	1.37
Cholesterol	75.00	160.50	186.00	219.05	303.10
HDL	10.1	32.0	40.0	47.0	69.1
LDLc	45.0	109.0	131.0	157.0	225.9
Triglycerides	43.00	115.20	155.00	221.55	380.00
TCHOLU	82	155	180	207	285
HDLU	15	33	40	48	70
LDLU	45	102	125	145	203
TGU	43	112	150	210	357
SBP	90.0	110.5	120.0	130.0	158.0
SBPU	80	110	120	130	160
DBP	50	69	76	85	105
DBPU	55	69	76	80	95

mente más altos en comparación con el conjunto de datos femenino. El conjunto de datos masculino tiende a tener un WHR ligeramente más alto en comparación con el conjunto de datos femenino.

En los pacientes de control, los valores de presión arterial sistólica y diastólica siguen un patrón similar en los conjuntos de datos generales, masculinos y femeninos, con rangos superpuestos para SBP, SBPU, DBP y DBPU. Si bien las edades mínima y máxima son consistentes en todos los conjuntos de datos, existen diferencias sutiles en los valores de la mediana y del cuartil superior. El conjunto de datos femenino generalmente muestra una mediana y una edad del cuartil superior ligeramente más altas en comparación con el conjunto de datos masculino. Los valores de la relación cintura-cadera (WHR) varían entre los conjuntos de datos, y el conjunto de datos femenino muestra una mediana más baja y un cuartil superior en comparación con el conjunto de datos masculino. Los rangos del BMI difieren: el conjunto de datos femenino tiende a tener valores de mediana y cuartil superior más bajos en comparación con el conjunto de datos masculino. El conjunto de datos generales se encuentra entre los conjuntos de datos masculinos y femeninos con respecto al BMI. Existe variabilidad en los perfiles de colesterol y lípidos, el conjunto de datos femenino generalmente

Tabla 5.41: Rangos en el conjunto de datos femenino (controles y casos).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	30	47	54	60	79
WHR	0.71	0.84	0.89	0.93	1.06
BMI	17.20	25.28	28.60	32.46	43.21
Urea	9	24	28	34	49
Creatinine	0.360	0.625	0.720	0.810	1.080
Cholesterol	100.0	177.0	205.0	237.0	325.1
HDL	12.00	38.05	46.70	56.00	82.00
LDLc	60.0	116.0	140.0	167.9	243.9
Triglycerides	32.00	116.00	156.00	216.05	364.00
TCHOLU	96	168	193	220	296
HDLU	16	40	48	57	82
LDLU	49	109	130	153	219
TGU	32	110	144	204	341
SBP	80	110	120	130	160
SBPU	80	110	120	130	160
DBP	50	70	80	85	105
DBPU	50	70	80	85	100

Tabla 5.42: Rangos en el conjunto de datos general (solo controles).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	34	43	47	53	68
WHR	0.74	0.87	0.92	0.96	1.09
BMI	18.00	24.90	27.10	29.65	36.40
Urea	12	24	28	32	43
Creatinine	0.42	0.70	0.79	0.91	1.22
Cholesterol	87.0	162.0	186.0	214.5	291.0
HDL	17	38	45	53	75
LDLc	45	108	128	151	214
Triglycerides	32	101	134	181	299
TCHOLU	87.0	162.0	186.0	214.5	291.0
HDLU	17	38	45	53	75
LDLU	45	108	128	151	214
TGU	32	101	134	181	299
SBP	90	110	119	127	152
SBPU	90	110	119	127	152
DBP	46	66	70	80	100
DBPU	46	66	70	80	100

muestra valores más bajos de CHOL, HDLc, LDLc, TG.B y TCHOLU en comparación con el conjunto de datos masculino. Los niveles de creatinina, representados por “Creatinine”, varían entre los conjuntos de datos, y el conjunto de datos femenino muestra valores de mediana y cuartil superior ligeramente más bajos en comparación con el conjunto de datos masculino.

Por último, en los casos, los valores de edad mínima y máxima son consistentes en todos los conjuntos de datos, lo que indica una distribución de edad uniforme dentro de los casos. Los casos presentan valores de presión arterial sistólica (SBP, SBPU) y diastólica (DBP, DBPU) que exhiben patrones comparables en los conjuntos de datos generales, masculinos y femeninos. Los niveles de urea y creatinina (“Urea”, “Creatinine”) muestran patrones similares entre los conjuntos de datos, con rangos superpuestos. El colesterol total (Cholesterol), el colesterol HDL (HDL), el colesterol LDL (LDLc) y los triglicéridos no corregidos (TGU) siguen tendencias de distribución similares entre géneros y en la población general. Los valores de WHR varían entre los conjuntos de datos, y el conjunto de datos femenino generalmente muestra valores de mediana y cuartil superior más bajos en comparación con el conjunto de datos masculino. Los rangos del BMI son diferentes: el conjunto de datos femenino tiende a tener valores de mediana

Tabla 5.43: Rangos en el conjunto de datos masculino (solo controles).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	34	42	46	51	64
WHR	0.83	0.91	0.94	0.97	1.05
BMI	19.10	25.25	27.30	29.45	35.60
Urea	13	24	28	32	43
Creatinine	0.440	0.750	0.855	0.965	1.270
Cholesterol	87	159	181	208	281
HDL	15	34	41	49	71
LDLc	45.0	105.0	126.0	146.5	203.0
Triglycerides	44	107	140	192	318
TCHOLU	87	159	181	208	281
HDLU	15	34	41	49	71
LDLU	45.0	105.0	126.0	146.5	203.0
TGU	44	107	140	192	318
SBP	90	111	120	128	152
SBPU	90	111	120	128	152
DBP	55	65	70	78	95
DBPU	55	65	70	78	95

y cuartil superior más bajos en comparación con el conjunto de datos masculino. El conjunto de datos generales se encuentra entre los conjuntos de datos masculinos y femeninos con respecto al BMI. Los niveles de creatinina (“Creatinine”) en el conjunto de datos femenino muestran valores de mediana y cuartil superior ligeramente más bajos en comparación con el conjunto de datos masculino. Existe variabilidad en el colesterol total, y el conjunto de datos femenino generalmente muestra valores más altos en comparación con el conjunto de datos masculino.

Tabla 5.44: Rangos en el conjunto de datos femenino (solo controles).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	34.0	43.0	49.0	54.5	71.0
WHR	0.65	0.81	0.87	0.93	1.11
BMI	17.20	24.50	26.80	29.95	38.10
Urea	11	21	26	32	47
Creatinine	0.42	0.62	0.72	0.80	1.06
Cholesterol	100	167	195	220	291
HDL	19.0	43.0	51.0	59.5	84.0
LDLc	60	111	132	155	221
Triglycerides	32	96	125	158	244
TCHOLU	100	167	195	220	291
HDLU	19.0	43.0	51.0	59.5	84.0
LDLU	60	111	132	155	221
TGU	32	96	125	158	244
SBP	90	109	113	124	145
SBPU	90	109	113	124	145
DBP	50	68	71	80	95
DBPU	50	68	71	80	95

Tabla 5.45: Rangos en el conjunto de datos general (sólo casos).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	31	50	57	63	82
WHR	0.75	0.88	0.92	0.97	1.10
BMI	17.48	25.81	28.95	32.68	42.96
Urea	9	26	30	39	58
Creatinine	0.26	0.67	0.79	0.95	1.35
Cholesterol	91.00	176.00	207.10	242.05	340.00
HDL	10	33	41	50	75
LDLc	51.0	118.0	143.0	175.0	256.9
Triglycerides	41.00	139.00	185.00	253.05	417.10
TCHOLU	91	161	187	215	296
HDLU	12.0	35.5	42.0	52.0	76.0
LDLU	49	106	127	148	209
TGU	41	123	164	230	387
SBP	70	110	130	140	180
SBPU	80	110	120	130	160
DBP	65	80	85	90	105
DBPU	70	80	80	90	100

Tabla 5.46: Rangos en el conjunto de datos masculino (sólo casos).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	30	50	58	65	86
WHR	0.84	0.94	0.97	1.01	1.11
BMI	17.480	25.245	27.810	31.015	39.340
Urea	6	26	32	43	66
Creatinine	0.54	0.79	0.93	1.08	1.51
Cholesterol	91.00	165.05	194.10	237.10	340.00
HDL	10.10	30.10	36.70	44.05	63.00
LDLc	51.0	113.0	138.0	168.5	250.9
Triglycerides	43.0	134.2	182.0	257.7	433.0
TCHOLU	91.0	151.0	180.0	204.5	284.0
HDLU	16	31	39	45	66
LDLU	51.0	99.0	122.0	143.5	201.0
TGU	43.0	122.0	166.0	230.5	377.0
SBP	80	110	130	140	180
SBPU	80	110	120	130	160
DBP	65	80	85	90	105
DBPU	70	80	80	90	100

Tabla 5.47: Rangos en el conjunto de datos femenino (sólo casos).

Característica	Mínimo	Cuartil inferior	Mediana	Cuartil superior	Máximo
Age	34.0	50.5	56.0	63.0	79.0
WHR	0.73	0.85	0.89	0.93	1.05
BMI	17.84	26.53	29.77	33.61	43.96
Urea	6	24	30	36	54
Creatinine	0.36	0.63	0.72	0.82	1.10
Cholesterol	108.00	184.05	213.55	245.60	333.10
HDL	12.0	36.0	44.1	53.0	78.0
LDLc	62.00	123.00	147.95	179.90	264.90
Triglycerides	41.00	141.05	188.25	248.90	407.00
TCHOLU	96.0	168.5	192.0	220.0	296.0
HDLU	16	38	46	54	78
LDLU	49.0	108.0	129.5	150.0	209.0
TGU	41.0	123.0	164.0	229.5	387.0
SBP	90	115	130	140	175
SBPU	80	110	120	130	160
DBP	65	80	85	90	105
DBPU	70	80	80	90	100

Capítulo 6
Discusión

La detección temprana y la clasificación de la DMT2 son esenciales para prevenir complicaciones y mejorar los resultados de los pacientes, como se enfatiza en estos tres casos de estudio. El caso de estudio 1 se centra en un enfoque conjunto que combina GLM, SVM y ANN mediante votación directa para mejorar la precisión de la detección. El proceso implica equilibrar, estandarizar, imputar e integrar datos en tres modelos para clasificar a los pacientes con DMT2 de forma dicotómica. La selección de características se realiza utilizando LASSO, seguida de una validación cruzada y evaluación diez veces con AUC. El estudio informa un AUC de 0.92 para SVM y 0.90 para el modelo conjunto, lo que demuestra la eficacia de combinar múltiples modelos. El caso de estudio 2 destaca la ineficiencia en el uso de insulina como causa principal de DMT2 y atribuye una proporción significativa de los casos de diabetes a DMT2. Propone una selección inteligente de características de metabolitos relacionados con diferentes etapas de la diabetes utilizando GA e implementa SVM, KNN y Nearcent para la clasificación. El conjunto de datos se obtiene del Instituto Mexicano del Seguro Social y el estudio muestra la efectividad de GA con tasas de ACC que oscilan entre 0.74 y 0.96, lo que indica un rendimiento sólido en múltiples implementaciones. El caso de estudio 3 tiene como objetivo identificar biomarcadores específicos de género y diferencias en los perfiles clínicos y antropométricos para la detección de DMT2, excluyendo los biomarcadores relacionados con la glucosa. Emplea un enfoque integral de selección de características utilizando RFE, LASSO y GA, seguido de una comparación de modelos de ML como regresión logística, ANN, SVM, KNN y Nearcent. Este estudio encuentra que RFE con RF y AIC produce la mayor ACC de 0.8820, lo que revela posibles biomarcadores como la presión arterial sistólica, los triglicéridos, el colesterol y la presión arterial diastólica, con notables diferencias de género. Los tres casos de estudio subrayan la importancia crítica de la selección de características, el uso de diversos modelos de ML y la combinación de modelos para mejorar la precisión de la predicción. El énfasis en la detección temprana a través de métodos menos invasivos, incluyendo los datos

antropométricos, tales como, la presión arterial sistólica, el peso, el sexo, la edad, la altura u otras características no relacionadas directamente con los niveles de glucosa, el establecimiento de rangos específicos por sexo orientados a mejorar los resultados de los pacientes a través de intervenciones más personalizadas y oportunas, además de la confirmación que nos dan los biomarcadores potenciales derivados de la metabolómica estableciendo relaciones sumamente relevantes como el ácido ganodérico C2 y el riesgo a padecer DMT2 o su progresión hacia la ND, nos proporcionan indicadores de posibles tratamientos como ejemplo, un extracto de los triterpenoides contenidos en los ganoderma puede reducir las concentraciones de glucosa plasmática(Ma *et al.*, 2015). Los hallazgos colectivos de estos 3 estudios resaltan la necesidad de realizar más investigaciones para validar estos biomarcadores y metodologías en otras poblaciones y entornos clínicos más amplios, garantizando su aplicabilidad y eficacia en escenarios del mundo real.

La metodología propuesta muestra potencial para resolver un problema de clasificación dicotómica y se puede determinar que las características incluidas en los experimentos realizados están directamente relacionadas para la detección de DMT2, tal como se presenta en los resultados. La partición de datos del 75% para entrenamiento y del 25% para pruebas ciegas con una validación cruzada de 10 (Moreno-Torres *et al.*, 2012) es suficiente para proporcionar coherencia en la implementación de LASSO como método de selección de características y todos los modelos implementados, dio como resultado un AUC en todas las pruebas de implementación consistentes a lo largo de todos los datos analizados, con el conjunto de 12 características del producto como hiperparámetros utilizados para entrenar y probar los modelos, las características producidas por esta metodología son:

Con los datos analizados se puede deducir que las personas con ingresos mayores a 5,000 pesos mexicanos tienden a tener datos antropométricos más saludables, niveles de lípidos más bajos y mejor presión arterial, teniendo menos riesgo de padecer DMT2 que las personas con ingresos menos de 5000 pesos mexicanos. Esto se relaciona estadísticamente en América Latina, ya que la clase de ingresos bajos y medios tiene más probabilidades de correr riesgo de DMT2 (Manne-Goehler *et al.*, 2019). De los datos de la característica Sexo, se da a entender que los pacientes masculinos clasificados como positivos son más que los femeninos, logrando una relación directa con las estadísticas mostradas en la introducción y la muestra, sin embargo esta consistencia no presenta interacción significativa con las otras características como las presenta Gou *et al.* (2020) en la construcción de un microbioma con características similares comparándolo con el rasgo sexual. La característica Edad con mayor riesgo de desarrollar DMT2 según los datos analizados, es entre 44 y 58 años. WHR y BMI derivados de los

datos muestran que estas características están directamente relacionadas con la obesidad (Chatterjee *et al.*, 2020), si los valores están entre las clases de factores de riesgo de obesidad como se muestra en la clasificación de la obesidad dado por el Centro para Control y Prevención de Enfermedades (CDC) (Centers for Disease Control and Prevention, 2022), estos casos aumentarán el riesgo de desarrollar la enfermedad de DMT2 a medida que aumenta la edad y el BMI (Xie *et al.*, 2019). Los niveles de Urea están directamente asociados con un mayor riesgo de DMT2, otra característica con uso potencial para la detección, y presentada como factor clave en la detección de diabetes por Dinh *et al.* (2019). Los lípidos en el tratamiento están de alguna manera relacionados con la DMT2, ya que su variabilidad puede usarse en el seguimiento de la diabetes como lo respalda Lee *et al.* (2021), pero esta característica se puede descartar en este experimento ya que los niveles de lípidos no tratados y el tipo de medicamento para el tratamiento no se muestran en los datos, lo que hace imposible hacer comparaciones e identificar diferencias o variaciones, solo muestra si Se utilizó o no medicación para corregir los niveles. La característica HDL es un predictor significativo de DMT2, como lo confirma Lai *et al.* (2019) muestra correlación directa con Triglycerides, SBP y BMI. La Hipertensión bajo tratamiento, la DBP o presión arterial diastólica y la SBP no corregida o presión arterial sistólica no corregida con medicamentos están altamente correlacionadas con la DMT2. En el caso de este experimento el factor diferente es la corrección de niveles mediante medicación en la hipertensión, esta correlación fue revisada y comparada sin diferencias significativas con el estudio de Zheng *et al.* (2021) donde la presión arterial y la hipertensión no fueron controladas ni corregidas con medicamentos. La presión arterial diastólica es la característica más relevante en los 3 modelos implementados en este experimento y es un biomarcador potencial para la detección de DMT2 digno de futuros estudios.

Como principales hallazgos se puede establecer que un modelo sin características relacionadas con la glucosa tiene un rendimiento similar a los modelos con características relacionadas con la glucosa como se muestra en la tabla 6.1:

Las tasas de falsos positivos y falsos negativos del conjunto propuesto, claramente presenta excelentes resultados, como se muestra en la tabla 5.8, estas bajas tasas explican que el modelo es sólido y utilizable para respaldar diagnósticos precisos en situaciones del mundo real como los escenarios presentados en los datos de cada paciente analizados.

AUC del SVM es ligeramente mejor que el resto de los modelos incluyendo el conjunto final, ni mucho menos, esto podría cambiar a medida que cambie la naturaleza aleatoria de los conjuntos de particiones, funcionando peor en algunos casos, el escenario presentado en este trabajo se incluye ya que proporcionó un resultado que alcanzó más del 90 % AUC.

Se ha demostrado que el uso de ML es una herramienta eficaz de clasificación y selección de características, sin embargo, la comunidad metabolómica tiene algunas preocupaciones por la falta de explicación sobre de dónde proviene la importancia de estos biomarcadores. Hay algunos métodos que develan estas dudas, la validación estadística por ejemplo, tiene el validador más conocido, el AUC (Barberis *et al.*, 2022). Sin embargo, en esta propuesta no se utilizará esta métrica, sino que se propone el uso de ACC promedio estrictamente en la selección de características con algoritmos genéticos, ya que proviene de un validador proporcionado por el método de selección directa proporcionado por GALGO.

La selección propuesta y la implementación del modelo muestran potencial para establecer metabolitos significativos en cada etapa de las enfermedades descritas. La afirmación de la progresión se puede hacer a medida que las características chocan y los modelos producto de las características y el modelo ML SVM encajan. Como la metabolómica obtenida se establece principalmente a partir de muestras de suero, las familias de lípidos implicadas en las características obtenidas revelan metabolitos de riesgo potencial y, más concretamente, proponen una base para crear una herramienta de apoyo al diagnóstico personalizado.

Para establecer una conexión en cada etapa, se crearon los 5 conjuntos diferentes dentro de una clasificación previa y transición directa para poder aseverar que habrá una relación estadística entre cada etapa. La propuesta proporciona una herramienta útil para seleccionar características independientemente del tipo de datos seleccionados como en este caso la metabolómica, siendo un campo amplio a cubrir pero un poderoso aliado para diagnosticar una enfermedad o identificar su progresión.

La identificación de metabolitos como biomarcadores potenciales o candidatos de incidencia de DN en sujetos hiperglucémicos se puede realizar con selección inteligente de características (Huang *et al.*, 2020), sin embargo en el caso de LASSO los biomarcadores obtenidos pueden ser diversos dependiendo no solo del tipo de metabolómica en el conjunto de datos, dado que los algoritmos genéticos demuestran superioridad y diversidad de modelos que LASSO, se pueden hacer nuevas aproximaciones. La predicción de la progresión de T2DM a DN sigue siendo difícil, incluso con biomarcadores potenciales, para detectar uno u otro, se necesitan nuevos biomarcadores para resistir la progresión de la enfermedad, sin embargo, los métodos de ML pueden detectar biomarcadores potenciales que de otro modo podrían escapar a la identificación mediante el método estadístico convencional. Incluso para identificar un biomarcador potencial como el 1-metilpiridin-1-io (NMP) urinario (Hirakawa *et al.*, 2022) con metabolitos no objetivo, aún se necesitan otras características como complemento para marcarlo como clínicamente utilizable, en lugar de un grupo de biomarcadores significativos. Los metabolitos pueden ser útiles para la detección de cada etapa de la

enfermedad. Limitar el número de características para encontrar predictores sólidos podría resaltar diferencias en las vías que conducen a predicciones precisas en poblaciones más específicas y proporcionar pistas novedosas para conducir a características de gran importancia (Frohnert *et al.*, 2019), sin embargo, es necesario tener suficiente Al realizar observaciones para hacer un modelo que no se sobreadapte, esto se vuelve difícil en la detección de la progresión de la DMT2, ya que no hay certeza de cuándo se desarrollarán las etapas de esta enfermedad, que puede tardar décadas en desarrollarse. En el caso de un número pequeño de observaciones e incluso con una validación cruzada que omite una, parece sobreajustarse, la selección de características puede permanecer estable con buenos valores de ACC promedio (más de 0.88 en este estudio), cuando hay muchas características (más de 700 en este caso).

La identificación de biomarcadores de metabolitos puede ayudar ampliamente a comprender la progresión y obtener una suposición más cercana sobre cómo funciona una enfermedad en una parte específica del cuerpo humano, o cómo puede progresar aumentando el riesgo de desarrollar una comorbilidad, por ejemplo en la diabetes gestacional. Una serie de metabolitos pueden identificar el riesgo potencial de desarrollar DMT2, simplemente percibiendo las variaciones y encontrando patrones con ML y técnicas de minería de datos (Kumar *et al.*, 2022).

En comparación con los datos clínicos o antropométricos, los datos de metabolitos se presentan como complemento de los modelos, sin embargo, los metabolitos presentados en este estudio pueden funcionar mejor que los datos clínicos y antropométricos (Dritsas and Trigka, 2022), este comportamiento se puede explicar con el uso de algoritmos genéticos como selectores, y una adecuada preselección de las más importantes o familiares relacionadas con la enfermedad que se va a predecir.

El modelo ML KNN en la implementación GALGO, tuvo los modelos más pequeños en comparación con los otros dos presentados en este trabajo. Nearcent y SVM tuvieron una mejor ACC promedio, y para establecer un modelo comparable en la misma cantidad con el otro trabajo relacionado que tenía menos de 10 características como resultado final (Huang *et al.*, 2020; Hirakawa *et al.*, 2022; Frohnert *et al.*, 2019), esto a excepción del conjunto de datos Control-Prediabetes y la T2DM-Nefropatía, que incluso con KNN produjo un modelo de más de 10 características.

La distribución y variabilidad de varias métricas relacionadas con la salud en diferentes conjuntos de datos, incluidos conjuntos de datos generales, masculinos y femeninos, tanto para controles como para casos, tiene variables como TCHOLU, HDLU, LDLU y TGU relacionadas con los niveles de Cholesterol en la urea que están presentes en el grupo masculino y conjuntos de datos femeni-

nos, pero no en el conjunto de datos general. La inclusión de estas variables en conjuntos de datos específicos de género indica un enfoque en las variaciones relacionadas con el género en los niveles de colesterol (Kanaya *et al.*, 2002), lo que potencialmente enfatiza la importancia de estas variables para comprender los riesgos para la salud específicos de cada género al planificar la seguridad por encima de las recetas o los medicamentos orales (Bolen *et al.*, 2007). Las variables de presión arterial (SBP, SBPU, DBP, DBPU) exhiben variabilidad entre conjuntos de datos, lo que refleja diferencias relacionadas con el género en las distribuciones de la presión arterial. La coherencia en los patrones de presión arterial sistólica y diastólica en los conjuntos de datos generales, masculinos y femeninos sugiere que las tendencias de la presión arterial siguen siendo relativamente similares, independientemente del género (Siransy-Balayssac *et al.*, 2020). Si bien las edades mínima y máxima son consistentes en todos los conjuntos de datos, existen diferencias sutiles en los valores de la mediana y del cuartil superior. El conjunto de datos femenino generalmente muestra una mediana y un cuartil superior de edad ligeramente más altos en comparación con el conjunto de datos masculino, lo que sugiere una posible diferencia relacionada con la edad entre los géneros.

Los valores de WHR y BMI varían entre los conjuntos de datos, el conjunto de datos femenino generalmente muestra valores de mediana y cuartil superior más bajos en comparación con el conjunto de datos masculino. Estas diferencias resaltan posibles variaciones relacionadas con el género en la composición corporal que tienden a afectar el equilibrio energético y aumentan el riesgo de desarrollar resistencia a la insulina y complicaciones relacionadas con la obesidad (Geer and Shen, 2009), y las mujeres tienden a tener WHR e BMI más bajos en comparación con los hombres. Existe variabilidad en los perfiles de colesterol y lípidos, el conjunto de datos femenino generalmente muestra valores más bajos de CHOL, HDLc, LDLc, TG.B y TCHOLU en comparación con el conjunto de datos masculino, esta comparación está alineada para buscar factores clave en las comorbilidades de DMT2 y enfermedades cardiovasculares (Lam *et al.*, 2015). Esta variabilidad relacionada con el género en los perfiles de lípidos enfatiza la importancia de considerar las diferencias específicas de género en los factores de riesgo cardiovascular. Los niveles de creatinina varían entre los conjuntos de datos, el conjunto de datos femenino muestra valores de mediana y cuartil superior ligeramente más bajos en comparación con el conjunto de datos masculino, lo que proporciona información útil para encontrar factores de riesgo de ND, esta diferencia puede reflejar variaciones relacionadas con el género en la función renal o masa muscular (Halbesma *et al.*, 2008). Los casos exhiben una distribución de edad uniforme en todos los conjuntos de datos, lo que indica coherencia en la representación de la edad dentro de los casos. Esta uniformidad simplifica

las comparaciones y análisis relacionados con la edad dentro de los casos. Los valores de presión arterial sistólica y diastólica, así como los perfiles de colesterol y lípidos, exhiben patrones comparables entre géneros y la población general en los casos. Esta coherencia sugiere que ciertas métricas de salud mantienen tendencias similares en diferentes conjuntos de datos al considerar casos.

Este análisis proporciona información valiosa sobre las variaciones y distribuciones de métricas relacionadas con la salud en diferentes conjuntos de datos, destacando posibles diferencias relacionadas con el género y patrones consistentes dentro de los casos. Estos hallazgos pueden guiar futuras investigaciones sobre los factores subyacentes que contribuyen a estas variaciones e informar estrategias de atención médica personalizadas.

Para determinar la importancia de los "TGU" (triglicéridos) y la "DBP" (presión arterial diastólica) en el conjunto de datos general, podemos analizar sus rangos y su importancia en el contexto de la relevancia clínica y de salud. Se sabe que los niveles elevados de triglicéridos están asociados con un mayor riesgo de enfermedades cardiovasculares, los triglicéridos son cruciales para evaluar el metabolismo de los lípidos y la salud cardiovascular en general y el rango de triglicéridos en el conjunto de datos abarca de 32 a 371, lo que indica una variabilidad considerable en los niveles de triglicéridos. Los valores más altos, especialmente más allá del cuartil superior, pueden sugerir un posible riesgo cardiovascular y anomalías metabólicas. La presión arterial diastólica (DBP) es un componente clave en la evaluación de la presión arterial, que es un indicador crítico de la salud cardiovascular. La presión arterial diastólica elevada se asocia con un mayor riesgo de enfermedad cardíaca y otros eventos cardiovasculares. El rango de DBP en el conjunto de datos varía de 50 a 105, cubriendo un espectro de niveles de presión arterial. Valores más altos de DBP pueden indicar hipertensión y mayor riesgo cardiovascular.

Tanto los trigliceridos como la presión arterial distólica son indicadores vitales de la salud cardiovascular y se utilizan comúnmente en evaluaciones clínicas. Los amplios rangos observados en el conjunto de datos enfatizan la diversidad de condiciones de salud entre los individuos, monitorear estos parámetros es crucial para identificar posibles problemas de salud, especialmente en el contexto de las enfermedades cardiovasculares. En el conjunto de datos general, los "Triglycerides" y la "DBP" destacan como características importantes debido a su importancia clínica en la evaluación de la salud cardiovascular como comorbilidades o enfermedades directamente relacionadas con la DMT2. Monitorear y controlar los niveles de triglicéridos y la presión arterial, especialmente la presión diastólica, son esenciales para la salud general y la prevención de enfermedades cardiovasculares.

El rango de triglicéridos en el conjunto de datos de controles abarca de 32 a

299, lo que indica una variabilidad considerable en los niveles de triglicéridos entre personas sin DMT2. Los niveles elevados de triglicéridos, especialmente más allá del cuartil superior, pueden sugerir un riesgo cardiovascular potencial, incluso en ausencia de diabetes. El rango de triglicéridos en el conjunto de datos de casos abarca de 41.00 a 417.10, lo que indica niveles más altos, especialmente más allá de la media, que sugieren un mayor riesgo cardiovascular y complicaciones en personas con DMT2.

El rango de DBP en el conjunto de datos de controles generales varía de 46 a 100, cubriendo un espectro de niveles de presión arterial en personas sin DMT2. El rango de DBP en el conjunto de datos de casos varía de 65 a 105 y muestra valores de DBP más altos que los controles, especialmente más allá del cuartil superior, lo que puede indicar hipertensión y un mayor riesgo cardiovascular en personas con DMT2.

Tabla 6.1: Comparación de trabajos relacionados.

Autor	**Modelo ML**	**Conjunto de datos**	**Métricas**
El-Sappagh *et al.* (2019)	Ensemble of: k-nearest neighbors, naïve Bayes, decision tree, support vector machine, fuzzy decision tree, artificial neural network, and logistic regression	Electronic health records of Mansura University Hospitals (Mansura, Egypt)	90 % de *Accuracy*, 90.2 % de recall, and 94.9 % de *Precision*
Kumari *et al.* (2021)	Ensemble of: random forest, logistic regression, and Naïve Bayes	PIMA Indians diabetes dataset	79.04 % de *Accuracy*, 73.48 % de *Precision*, 71.45 % de recall, and 80.6 % de $F1$score
Singh and Singh (2020)	stacking-based evolutionary ensemble learning system "NSGA-II-Stacking"	PIMA Indians diabetes dataset	*Accuracy* de 83.8 %, *Sensitivity* de 96.1 %, *Specificity* de 79.9 %, f-measure de 88.5 % y AUC de 85.9 %
Liu *et al.* (2019)	Majority voting Ensemble: Support vector machine, tree-based methods and neural networks	REACTION study (Risk Evaluation of Cancers in Chinese Diabetic Individuals: A Longitudinal Study)	Majority voting with model selection results: AUC de 0.802 (80.2 %), *Sensitivity* de 0.662 (66.2 %), *Specificity* de 0.702 (70.2 %)
This Work Authors	**Hard voting Ensemble of: generalized linear regression, support vector machines and artificial neural networks**	**Centro Médico Nacional Siglo XXI dataset**	***Sensitivity* de 0.8788 (87.88 %) *Specificity* de 0.9242 (92.42 %) *Precision* de 0.9269 (92.69 %) *Area under the ROC curve* 90.5 %**

Tabla 6.2: Selección de características - trabajos relacionados.

Título	Técnica de Selección de características	Métricas de Validación	Resultado
Machine Learning Approaches Reveal Metabolic Signatures of Incident Chronic Kidney Disease in Individuals With Prediabetes and Type 2 Diabetes (Huang *et al.*, 2020)	LASSO	AUC	0.857
Potential progression biomarkers of diabetic kidney disease determined using comprehensive machine learning analysis of non-targeted metabolomics (Hirakawa *et al.*, 2022)	NON	AUC	0.775
Predictive Modeling of Type 1 Diabetes Stages Using Disparate Data Sources (Frohnert *et al.*, 2019)	Repeated Optimization for Feature Interpretation	AUC	0.91
Machine Learning–Derived Prenatal Predictive Risk Model to Guide Intervention and Prevent the Progression of Gestational Diabetes Mellitus to Type 2 Diabetes: Prediction Model Development Study (Kumar *et al.*, 2022)	CatBoost tree ensembles	AUC	0.86
Data-Driven Machine-Learning Methods for Diabetes Risk Prediction (Dritsas and Trigka, 2022)	Pearson Correlation, Gain Ratio, Naïve Bayes and Random Forest	AUC	0.942

Título	Técnica de Selección de características	Métricas de Validación	Resultado
Interpretable machine learning-derived nomogram model for early detection of diabetic retinopathy in type 2 diabetes mellitus: a widely targeted metabolomics study (Li *et al.*, 2022)	Classification and Regression Tree	AUC	0.95
Environmental chemical exposure dynamics and machine learning-based prediction of diabetes mellitus (Wei *et al.*, 2022)	Lasso	AUC	0.78
This Work in Control-Prediabetes	**Genetic Algorithm with GALGO-nearcent**	***Accuracy***	**0.925**
This Work in Control-T2DM	**Genetic Algorithm with GALGO-SVM**	***Accuracy***	**0.962**
This Work in Control-Diabetic Nephropathy	**Genetic Algorithm with GALGO-nearcent**	***Accuracy***	**0.962**
This Work in Prediabetes-T2DM	**Genetic Algorithm with GALGO-nearcent**	***Accuracy***	**0.934**
This Work in T2DM-Diabetic Nephropathy	**Genetic Algorithm with GALGO-nearcent**	***Accuracy***	**0.925**

Tabla 6.3: Continuación de la tabla 6.2.

Capítulo 7
Conclusiones

Los resultados obtenidos fueron 90 % ± 3 % en el modelo de conjunto GLM - SVM - ANN usando LASSO para la selección de características y dividiendo los datos en 75 % para entrenamiento y 25 % para pruebas, validado por 10 La validación cruzada y el AUC son satisfactorios pero concluyentes. El porcentaje de AUC se acerca al objetivo del 95 % para ser utilizado en un entorno clínico, la necesidad de implementar nuevas formas de procesar los datos y aún evitar la característica de Glucosa o usarla solo como referencia, como se propone, asegura la obtención características con potencial predictivo.

Como desventaja, se obtuvieron inconsistencias en la implementación de la técnica LASSO, en el primer caso de estudio, arrojando una característica decimotercera que variaba alternando entre: educación, LDL y SBPU. Estas características se mostraron por la aleatoriedad producto de la validación cruzada y no fueron tomadas como parte del modelo definitivo. Por esta razón, de lo contrario, las características que fueron seleccionadas persistieron en cada corrida incluso con el cambio de datos utilizados en los diferentes muestreos que realizó cada separación generada por los pliegues.

Todos los modelos de matrices de confusión incluyendo el conjunto final obtuvieron una *Sensitivity* entre un 3 % y un 4 % superior a la *Specificity*, comportamiento que se considera dentro de los parámetros aceptados para la detección de una enfermedad.

Los algoritmos de ML y los datos clínicos utilizados para esta metodología mostraron potencial para identificar relaciones, predicciones y patrones de comportamiento que clasifican con ACC a más del 90 % de los pacientes con o sin DMT2, utilizando biomarcadores extraídos mediante métodos no invasivos con la misma o superior precisión que los análisis de sangre de laboratorio invasivos: FPG, OGTT o el HbA1c.

Los resultados obtenidos en el método de selección directa respaldado por GALGO fueron muy precisos (0.86-0.96 u 86 %-96 %) en todos los conjuntos de datos analizados. Nearcent como método de clasificación para GALGO demuestra

ser efectivo y el más preciso en 4 de 5 conjuntos de datos, sin embargo, en este estudio, se seleccionó el KNN para ser presentado incluso cuando era menos preciso que svm o nearcent, ya que produjo los modelos con menos características. Las características producto de la selección pueden estudiarse en detalle como trabajo futuro, dada la amplia variedad de metabolitos encontrados como herramienta potencial para el diagnóstico en cada fase o estadio de la enfermedad.

Como desventaja, el limitado número de observaciones presentadas en este estudio, deriba en el uso de implementaciones de generación de datos sintéticos, que si bien sirven para dar balance a los datos, no podrían tomarse como parte de un modelo de ML para una mayor validación, ya que esto podría presentar un posible sobreajuste en el rendimiento (entre 95 % y 100 % de AUC) o salidas con un sesgo o error. Más observaciones reales con los mismos metabolitos pueden validar adecuadamente o complementar esta propuesta.

El análisis integral de métricas relacionadas con la salud en diferentes conjuntos de datos, incluidos conjuntos de datos generales, masculinos y femeninos, proporciona información valiosa sobre las diferencias relacionadas con el género y los patrones consistentes dentro de los casos. Las diferencias relacionadas con el género en los perfiles de colesterol y lípidos presentan variaciones relacionadas con el género en variables como TCHOLU, HDLU, LDLU y TGU, lo que enfatiza la importancia de considerar las diferencias específicas de género en los niveles de colesterol. Las mujeres generalmente exhiben valores más bajos de colesterol y lípidos en comparación con los hombres, lo que resalta posibles variaciones relacionadas con el género en los factores de riesgo cardiovascular. Los perfiles de presión arterial y lípidos, incluidos los valores de presión arterial sistólica y diastólica, así como los perfiles de colesterol y lípidos, exhiben patrones comparables entre géneros y la población general en los casos. Esta coherencia sugiere que ciertas métricas de salud mantienen tendencias similares en diferentes conjuntos de datos al considerar casos. La importancia de los “Triglycerides” y la “DBP” como características principales (presentes en todos los métodos finales de selección de características), tienen amplios rangos en los controles y los casos indican una variabilidad significativa en los niveles de triglicéridos. Los niveles elevados, especialmente más allá del cuartil superior, pueden sugerir un riesgo cardiovascular potencial, lo que enfatiza la importancia de monitorear los niveles de triglicéridos para la salud general como ocurre con la presión arterial diastólica (DBP), mostrar un rango diverso de valores de DBP tanto en los controles como en los casos subraya su importancia en evaluar la salud cardiovascular. Los valores más altos de DBP, especialmente más allá del cuartil superior, pueden indicar hipertensión y un mayor riesgo cardiovascular.

La importancia clínica y las implicaciones de los triglicéridos y la DBP en el conjunto de datos general hacen que tanto los “Triglycerides” como la “DBP” sur-

jan como indicadores cruciales de la salud cardiovascular. Sus amplios rangos enfatizan la diversidad de condiciones de salud entre los individuos, destacando su importancia clínica en la evaluación de la salud cardiovascular, especialmente en el contexto de comorbilidades asociadas con la DMT2. Los controles de triglicéridos muestran una variabilidad considerable en los niveles de triglicéridos, lo que indica un riesgo cardiovascular potencial incluso en ausencia de diabetes. En algunos casos, los niveles más altos sugieren un mayor riesgo cardiovascular y complicaciones en personas con DMT2. Finalmente, los casos de DBP muestran valores de DBP más altos que los controles, particularmente más allá del cuartil superior, lo que indica hipertensión y mayor riesgo cardiovascular en individuos con DMT2. Comprender estas diferencias y similitudes relacionadas con el sexo es crucial para adaptar las estrategias e intervenciones de atención médica. Controlar los niveles de triglicéridos y la presión arterial, especialmente la presión diastólica, es esencial para la salud general y la prevención de enfermedades cardiovasculares, particularmente en el contexto de la diabetes. Los hallazgos proporcionan una base para futuras investigaciones sobre los factores subyacentes que contribuyen a estas variaciones y respaldan el desarrollo de enfoques de atención médica personalizados.

Referencias

(2017). WMA - The World Medical Association-Declaración de Helsinki de la AMM – Principios éticos para las investigaciones médicas en seres humanos.

(2022a). Idf diabetes atlas — tenth edition.

(2022b). Idf news 2022.

Adiwinoto, R. D., Pranoto, A., Sugihartono, T., Soelistijo, S. A., and Adiwinoto, R. P. (2024). Triglyceride to high-density lipoprotein cholesterol ratio as a marker of non-alcoholic fatty liver disease in type 2 diabetes. *International Journal of Public Health Science (IJPHS)*, 13(3):1039.

Agliata, A., Giordano, D., Bardozzo, F., Bottiglieri, S., Facchiano, A., and Tagliaferri, R. (2023). Machine learning as a support for the diagnosis of Type 2 diabetes. *International Journal of Molecular Sciences*, 24(7):6775.

Akhtar, T., Gilani, S. O., Mushtaq, Z., Arif, S., Jamil, M., Ayaz, Y., Butt, S. I., and Waris, A. (2021). Effective Voting Ensemble of Homogenous Ensembling with Multiple Attribute-Selection Approaches for Improved Identification of Thyroid Disorder. *Electronics*, 10(23):3026.

Albrizio, A. (2007). Biometry and anthropometry: from galton to constitutional medicine. *Journal of Anthropological Sciences*, 85:101–123.

Alieva, A., Alimov, A., Khaidarova, F., Ismailov, S., Rakhimova, G., Nazhmutdinova, D., Shagazatova, B., and Tsareva, V. (2022). Assessing the effectiveness of type 2 diabetes mellitus screening in the republic of uzbekistan. *International Journal of Endocrinology and Metabolism*, 20(4).

Allen, A., Barnes, G. L., Green-Saxena, A., Hurtado, M., Hoffman, J., Mao, Q., and Das, R. (2022). Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. *BMJ open diabetes research & care*, 10(1):e002560.

Alur, V., Raju, V., Vastrad, B., Vastrad, C., Kavatagimath, S., and Kottur-shetti, S. (2023). Bioinformatics Analysis of Next Generation Sequencing Data Identifies Molecular Biomarkers Associated With Type 2 Diabetes Mellitus. *Clinical Medicine Insights Endocrinology and Diabetes*, 16:117955142311556.

Arneth, B., Arneth, R., and Shams, M. (2019). Metabolomics of type 1 and type 2 diabetes. *International Journal of Molecular Sciences*, 20(10):2467.

Aruna D., P. (2014). Sex differences in the metabolic syndrome: Implications for cardiovascular health in women. *Clinical Chemistry*, 60(1):44–52.

Bailey, C. J. (2024). Metformin: Therapeutic profile in the treatment of type 2 diabetes. *Diabetes Obesity and Metabolism*, 26(S3):3–19.

Barberis, E., Khoso, S., Sica, A., Falasca, M., Gennari, A., Dondero, F., Afantitis, A., and Manfredi, M. (2022). Precision medicine approaches with metabolomics and artificial intelligence. *International Journal of Molecular Sciences*, 23(19):11269.

Barda, N., Yona, G., Rothblum, G. N., Greenland, P., Leibowitz, M., Balicer, R., Bachmat, E., and Dagan, N. (2020). Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, 28(3):549–558.

Bergman, R. N. (1989). Toward Physiological Understanding of Glucose Tolerance: Minimal-Model Approach. *Diabetes*, 38(12):1512–1527.

Bi, Y., Yang, Y., Yuan, X., Wang, J., Wang, T., Liu, Z., Tian, S., and Sun, C. (2024). Association between liver enzymes and type 2 diabetes: a real-world study. *Frontiers in Endocrinology*, 15.

Bolen, S., Feldman, L., Vassy, J. L., Wilson, L. M., Yeh, H. C., Marinopoulos, S. S., Wiley, C., Selvin, E., Wilson, R. F., Bass, E. B., and Brancati, F. L. (2007). Systematic Review: Comparative Effectiveness and Safety of oral medications for Type 2 Diabetes Mellitus. *Annals of Internal Medicine*, 147(6):386.

Boutilier, J. J., Chan, T. C. Y., Ranjan, M., and Deo, S. (2021). Risk Stratification for Early Detection of Diabetes and Hypertension in Resource-Limited Settings: Machine Learning Analysis. *Journal of Medical Internet Research*, 23(1):e20123.

Braffett, B. H., Ghormli, L. E., Albers, J. W., Feldman, E. L., Herman, W. H., Gubitosi-Klug, R. A., Martin, C. L., Orchard, T. J., White, N. H., Lachin, J. M., Perkins, B. A., and Pop-Busui, R. (2024). Neuropathic Pain With and Without Diabetic Peripheral Neuropathy in Type 1 Diabetes. *Diabetes Care*.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Buyrukoğlu, S. and Akbaş, A. (2022). Machine Learning based early prediction of Type 2 diabetes: a new hybrid feature selection approach using correlation matrix with HeatMap and SFS. *Balkan journal of electrical & computer engineering*, 10(2):110–117.

Cahn, A., Shoshan, A., Sagiv, T., Yesharim, R., Goshen, R., and Shalev, V. (2020). Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. *Diabetes/metabolism research and reviews*, 36(2).

Carrillo-Larco, R. M., Castillo-Cara, M., Anza-Ramirez, C., and Bernabé-Ortiz, A. (2021). Clusters of people with type 2 diabetes in the general population: unsupervised machine learning approach using national surveys in Latin America and the Caribbean. *BMJ Open Diabetes Research Care*, 9(1):e001889.

Castañé, H., Iftimie, S., Baiges-Gayà, G., Rodríguez-Tomàs, E., Jiménez-Franco, A., López-Azcona, A. F., Garrido, P., Castro, A., Camps, J., and Joven, J. (2022). Machine learning and semi-targeted lipidomics identify distinct serum lipid signatures in hospitalized COVID-19-positive and COVID-19-negative patients. *Metabolism*, 131:155197.

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics probability letters*, 33(2):201–208.

Cavanaugh, J. E. and Neath, A. A. (2019). The Akaike Information Criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, 11(3).

Cefalu, W. T. (2001). Insulin Resistance: Cellular and Clinical Concepts. *Experimental Biology and Medicine*, 226(1):13–26.

Celaya-Padilla, J., Villagrana-Bañuelos, K., Oropeza-Valdez, J., Monárrez-Espino, J., Castañeda-Delgado, J., Sofía, H.-V. O. A., Fernández-Ruiz, J., Ochoa-González, F., Borrego, J., Enciso-Moreno, J., López, J., López-Hernández, Y., and Galván-Tejada, C. (2021). Kynurenine and hemoglobin as sex-specific variables in covid-19 patients: A machine learning and genetic algorithms approach. *Diagnostics*, 11(12):2197.

Centers for Disease Control and Prevention (2022). Defining adult overweight and obesity.

Chai, J., Chen, G.-C., Yu, B., Xing, J., Li, J., Khambaty, T., Perreira, K., Perera, M., Vidot, D., Castaneda, S., Selvin, E., Rebholz, C., Daviglus, M., Cai, J., Horn, L., Isasi, C., Sun, Q., Hawkins, M., Xue, X., Boerwinkle, E., Kaplan, R., and Qi, Q. (2022). Serum metabolomics of incident diabetes and glycemic changes in a population with high diabetes burden: The hispanic community health study/study of latinos. *Diabetes*, 71(6):1338–1349.

Chan, L., Nadkarni, G. N., Fleming, F., McCullough, J. R., Connolly, P., Mosoyan, G., Salem, F., Kattan, M. W., Vassalotti, J. A., Murphy, B., Donovan, M., Coca, S. G., and Damrauer, S. M. (2021). Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease. *Diabetologia*, 64(7):1504–1515.

Chatterjee, A., Gerdes, M. W., and Martinez, S. G. (2020). Identification of risk factors associated with obesity and overweight—a machine learning overview. *Sensors*, 20(9).

Chatterjee, S., Khunti, K., and Davies, M. (2017). Type 2 diabetes. *The Lancet*, 389(10085):2239–2251.

Chen, Z.-Z. and Gerszten, R. (2020). Metabolomics and proteomics in type 2 diabetes. *Circulation Research*, 126(11):1613–1627.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98(464):900–916.

Cohen, S., Dagan, N., Cohen-Inger, N., Ofer, D., and Rokach, L. (2021). ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models. *IEEE Access*, 9:91584–91592.

Collins, G. S., Mallett, S., Omar, O., and Yu, L.-M. (2011). Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, 9(1).

Collyns, O. J., Meier, R. A., Betts, Z. L., Chan, D. S., Frampton, C., Frewen, C. M., Hewapathirana, N. M., Jones, S. D., Roy, A., Grosman, B., Kurtz, N., Shin, J., Vigersky, R. A., Wheeler, B. J., and De Bock, M. I. (2021). Improved Glycemic Outcomes With Medtronic MiniMed Advanced Hybrid Closed-Loop Delivery: Results From a Randomized Crossover Trial Comparing Automated Insulin Delivery With Predictive Low Glucose Suspend in People With Type 1 Diabetes. *Diabetes Care*, 44(4):969–975.

Courcoulas, A. P., Patti, M. E., Hu, B., Arterburn, D. E., Simonson, D. C., Gourash, W. F., Jakicic, J. M., Vernon, A. H., Beck, G. J., Schauer, P. R., Kashyap, S. R., Aminian, A., Cummings, D. E., and Kirwan, J. P. (2024). Long-Term Outcomes of Medical Management vs Bariatric Surgery in Type 2 Diabetes. *JAMA*, 331(8):654.

De Cervantes, B. V. M. (2024). El cuerpo humano : Oriente y Grecia antigua.

Deberneh, H. M. and Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *International Journal of Environmental Research and Public Health*, 18(6):3317.

DeFronzo, R. A., Ferrannini, E., Groop, L., Henry, R. R., Herman, W. H., Holst, J. J., Hu, F. B., Kahn, C. R., Raz, I., Shulman, G. I., Simonson, D. C., Testa, M. A., and Weiss, R. (2015). Type 2 diabetes mellitus. *Nature Reviews Disease Primers*, 1(1).

Diamanti, K., Cavalli, M., Pan, G., Pereira, M., Kumar, C., Skrtic, S., Grabherr, M., Risérus, U., Eriksson, J., Komorowski, J., and Wadelius, C. (2019). Intra- and inter-individual metabolic profiling highlights carnitine and lysophosphatidylcholine pathways as key molecular defects in type 2 diabetes. *Scientific Reports*, 9(1).

Dieterle, F., Ross, A., Schlotterbeck, G., and Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in ¡sup¿1¡/sup¿h nmr metabonomics. *Analytical Chemistry*, 78(13):4281–4290.

Dinh, A., Miertschin, S., Young, A., and Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making*, 19(1).

Dritsas, E. and Trigka, M. (2022). Data-driven machine-learning methods for diabetes risk prediction. *Sensors*, 22(14):5304.

Drupad, K., TrivediKatherine, A., and Royston, G. (2017). Metabolomics for the masses: The future of metabolomics in a personalized world. *European Journal of Molecular amp; Clinical Medicine*, 3(6):294.

Du, G., Ma, L., Hu, J.-S., Zhang, J., Xiang, Y., Shao, D., and Wang, H. (2019). Prediction of 30-Day Readmission: An Improved Gradient Boosting Decision Tree Approach. *Journal of Medical Imaging and Health Informatics*, 9(3):620–627.

Dubin, R. F. and Rhee, E. P. (2019). Proteomics and metabolomics in kidney disease, including insights into etiology, treatment, and prevention. *Clinical Journal of the American Society of Nephrology*, 15(3):404–411.

DVEENT (2021). Boletines Diabetes Mellitus Tipo 2. Boletines Diabetes Mellitus Tipo 2.

El-Sappagh, S., Elmogy, M., Ali, F., ABUHMED, T., Islam, S. M. R., and Kwak, K.-S. (2019). A Comprehensive Medical Decision–Support Framework Based on a Heterogeneous Ensemble Classifier for Diabetes Prediction. *Electronics*, 8(6):635.

Emiliano, P. C., Vivanco, M. J., and De Menezes, F. S. (2014). Information criteria: How do they behave in different models? *Computational statistics data analysis*, 69:141–153.

Fitriyani, N. L., Syafrudin, M., Alfian, G., and Rhee, J. (2019). Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension.

Fonti, V. and Belitser, E. (2017). Feature Selection using LASSO.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Frimpong, E., Oluwasanmi, A., Baagyere, E. Y., and Qin, Z. (2021). A feedforward artificial neural network model for classification and detection of Type 2 diabetes. *Journal of physics*, 1734(1):012026.

Frohnert, B., Webb-Robertson, B., Bramer, L., Reehl, S., Waugh, K., Steck, A., Norris, J., and Rewers, M. (2019). Predictive modeling of type 1 diabetes stages using disparate data sources. *Diabetes*, 69(2):238–248.

Fujihara, K., Matsubayashi, Y., Yamada, M. H., Yamamoto, M., Iizuka, T., Miyamura, K., Hasegawa, Y., Maegawa, H., Kodama, S., Yamazaki, T., and Sone, H. (2021). Machine Learning Approach to Decision Making for Insulin Initiation in Japanese Patients With Type 2 Diabetes (JDDM 58): Model Development and Validation Study. *JMIR Medical Informatics*, 9(1):e22148.

Gao, L. Q., Xue, C. C., Cui, J., Xu, J., Zhang, C., Chen, D. N., Jonas, J. B., and Wang, Y. X. (2023). Diabetic Retinopathy and Chronic Kidney Disease: Associations and Comorbidities in a Large Diabetic Population – The Tongren Health Care Study. *American Journal of Nephrology*, 55(2):175–186.

García-Carretero, R., Vigil-Medina, L., and Barquero-Pérez, O. (2021). The Use of Machine Learning Techniques to Determine the Predictive Value of Inflammatory Biomarkers in the Development of Type 2 Diabetes Mellitus. *Metabolic syndrome and related disorders*, 19(4):240–248.

García-Domínguez, A., Galván-Tejada, C. E., Magallanes-Quintanar, R., Gamboa-Rosales, H., González-Curiel, I., Peralta-Romero, J., and Cruz-López, M. (2023). Diabetes detection models in Mexican patients by combining machine learning algorithms and feature selection techniques for clinical and paraclinical attributes: a Comparative evaluation. *Journal of diabetes research*, 2023:1–19.

Geer, E. B. and Shen, W. (2009). Gender differences in insulin resistance, body composition, and energy balance. *Gender Medicine*, 6:60–75.

Ghamisi, P. and Benediktsson, J. A. (2015). Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization.

Gonzalez, J. L. (2020a). Tipos de aprendizaje automático.

Gonzalez, L. (2020b). Aprendizaje Supervisado: Logistic Regression.

Gou, W., Ling, C.-w., He, Y., Jiang, Z., Fu, Y., Xu, F., Miao, Z., Sun, T.-y., Lin, J.-s., Zhu, H.-l., Zhou, H., Chen, Y.-m., and Zheng, J.-S. (2020). Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes. *Diabetes Care*, 44(2):358–366.

Halbesma, N., Brantsma, A. H., Bakker, S. J. L., Jansen, D., Stolk, R. P., De Zeeuw, D., De Jong, P. E., and Gansevoort, R. (2008). Gender differences in predictors of the decline of renal function in the general population. *Kidney International*, 74(4):505–512.

Hallberg, S. J., Gershuni, V. M., Hazbun, T. L., and Athinarayanan, S. J. (2019). Reversing Type 2 Diabetes: A Narrative Review of the Evidence.

Hamdi, T., Ali, J. B., Di Costanzo, V., Fnaiech, F., Moreau, E., and Ginoux, J.-M. (2018). Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. *Biocybernetics and Biomedical Engineering*, 38(2):362–372.

Han, J.-L. and Lin, H.-L. (2014). Intestinal microbiota and type 2 diabetes: From mechanism insights to therapeutic perspective. *World journal of gastroenterology*, 20(47):17737–17745.

Hasan, M. K., Alam, M. A., Das, D., Hossain, E., and Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access*, 8:76516–76531.

Hernández-Sampieri, R. and Mendoza, C. (2019). Metodología de la investigación. las rutas cuantitativa, cualitativa y mixta. *Revista Universitaria Digital de Ciencias Sociales (RUDICS)*, 10(18):92–95.

Hirakawa, Y., Yoshioka, K., Kojima, K., Yamashita, Y., Shibahara, T., Wada, T., Nangaku, M., and Inagi, R. (2022). Potential progression biomarkers of diabetic kidney disease determined using comprehensive machine learning analysis of non-targeted metabolomics. *Scientific Reports*, 12(1).

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems*.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators.

Huang, J., Huth, C., Covic, M., Troll, M., Adam, J., Zukunft, S., Prehn, C., Wang, L., Nano, J., F. Scheerer, M., Neschen, S., Kastenmüller, G., Suhre, K., Laxy, M., Schliess, F., Gieger, C., Adamski, J., Hrabe de Angelis, M., Peters, A., and Wang-Sattler, R. (2020). Machine learning approaches reveal metabolic signatures of incident chronic kidney disease in individuals with prediabetes and type 2 diabetes. *Diabetes*, 69(12):2756–2765.

Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Identifying Key Algorithm Parameters and Instance Features Using Forward Selection.

Huynh-Thu, V. A., Saeys, Y., Wehenkel, L., and Geurts, P. (2012). Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics*, 28(13):1766–1774.

I., L. (2005). Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics*, 6(1):68.

Iglay, K., Hannachi, H., Howie, P. J., Xu, J., Li, X., Engel, S. S., Moore, L. M., and Rajpathak, S. (2016). Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitus. *Current medical research and opinion*, 32(7):1243–1252.

INEGI (2021). Estadísticas a propósito del Día mundial de la diabetes (14 de noviembre).

INEGI (2022). Defunciones por diabetes mellitus por entidad federativa de residencia habitual de la persona fallecida y grupo quinquenal de edad según sexo, serie anual de 2010 a 2021.

Ivanescu, A., Popescu, S., Ivanescu, R., Potra, M., and Timar, R. (2024). Predictors of Diabetic Retinopathy in Type 2 Diabetes: A Cross-Sectional Study. *Biomedicines*, 12(8):1889.

Jain, A., Duin, R., and ao, J. M. (2000). Statistical pattern recognition: a review.

Jiang, G. and Zhang, B. B. (2003). Glucagon and regulation of glucose metabolism. *AJP Endocrinology and Metabolism*, 284(4):E671–E678.

Jones, A. G. and Hattersley, A. T. (2013). The clinical utility of C-peptide measurement in the care of patients with diabetes. *Diabetic Medicine*, 30(7):803–817.

Julia, H.-C. and Carol, C. (2016). Diabetes treatments and risk of amputation, blindness, severe kidney failure, hyperglycaemia, and hypoglycaemia: open cohort study in primary care. *BMJ*, page i1450.

Kanaya, A. M., Grady, D., and Barrett-Connor, E. (2002). Explaining the sex difference in coronary heart disease mortality among patients with type 2 diabetes mellitus. *Archives of internal medicine*, 162(15):1737.

Kanda, E., Suzuki, A., Miyamoto, M., Tsubota, H., Kanemata, S., Shirakawa, K., and Yajima, T. (2022). Machine learning models for prediction of HF and CKD development in early-stage type 2 diabetes patients. *Scientific Reports*, 12(1).

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15:104–116.

Klaine, P. V., Imran, M., Onireti, O., and Souza, R. (2017). A Survey of Machine Learning Techniques Applied to Self-Organizing Cellular Networks.

Klein, M. S. and Shearer, J. (2015). Metabolomics and Type 2 Diabetes: Translating Basic Research into Clinical Application. *Journal of Diabetes Research*, 2016:1–10.

Klonoff, D. C. and Price, W. N. (2016). The Need for a Privacy Standard for Medical Devices That Transmit Protected Health Information Used in the Precision Medicine Initiative for Diabetes and Other Diseases. *Journal of Diabetes Science and Technology*, 11(2):220–223.

Kocbek, S., Kocbek, P., Gosak, L., Fijačko, N., and Štiglic, G. (2022). Extracting New Temporal Features to Improve the Interpretability of Undiagnosed Type 2 Diabetes Mellitus Prediction Models. *Journal of Personalized Medicine*, 12(3):368.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial intelligence in medicine*, 23(1):89–109.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., and Stiglic, G. (2020a). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, 10(1):11981.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., and Štiglic, G. (2020b). Early detection of Type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, 10(1).

Kotronen, A., Juurinen, L., Tiikkainen, M., Vehkavaara, S., and Yki–Järvinen, H. (2008). Increased Liver Fat, Impaired Insulin Clearance, and Hepatic and Adipose Tissue Insulin Resistance in Type 2 Diabetes. *Gastroenterology*, 135(1):122–130.

Kukova, L., Munir, K. M., Sayeed, A., and Davis, S. N. (2024). Assessing the therapeutic and toxicological profile of novel GLP-1 receptor agonists for type 2 diabetes. *Expert Opinion on Drug Metabolism Toxicology*, pages 1–14.

Kumar, M., Ang, L., Ho, C., Soh, S., Tan, K., Chan, J., Godfrey, K., Chan, S., Chong, Y., Eriksson, J., Feng, M., and Karnani, N. (2022). Machine learning–derived prenatal predictive risk model to guide intervention and prevent the progression of gestational diabetes mellitus to type 2 diabetes: Prediction model development study. *JMIR Diabetes*, 7(3):e32366.

Kumari, S., Kumar, D., and Mittal, M. (2021). An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, 2:40–46.

Lai, H., Huang, H., Keshavjee, K., Guergachi, A., and Gao, X. (2019). Predictive models for diabetes mellitus using machine learning techniques. *BMC Endocrine Disorders*, 19(1).

Lam, B. C. C., Koh, G. C.-H., Chen, C., Wong, M., and Fallows, S. (2015). Comparison of body mass Index (BMI), Body adiposity Index (BAI), waist circumference (WC), Waist-To-Hip Ratio (WHR) and Waist-To-Height Ratio (WHTR) as predictors of cardiovascular disease risk factors in an adult population in Singapore. *PLOS ONE*, 10(4):e0122985.

Leclercq, M., Vittrant, B., Martin-Magniette, M. L., Boyer, M. P. S., Périn, O., Bergeron, A., Fradet, Y., and Droit, A. (2019). Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. *Frontiers in genetics*, 10.

Lee, S., Zhou, J., Wong, W. T., Liu, T., Wu, W. K. K., Wong, I. C. K., Zhang, Q., and Tse, G. (2021). Glycemic and lipid variability for predicting complications and mortality in diabetes mellitus using machine learning. *BMC Endocrine Disorders*, 21(1).

Li, J., Guo, C., Wang, T., Xu, Y., Peng, F., Zhao, S., Li, H., Jin, D., Xia, Z., Che, M., Zuo, J., Zheng, C., Hu, H., and Mao, G. (2022). Interpretable machine learning-derived nomogram model for early detection of diabetic retinopathy in type 2 diabetes mellitus: a widely targeted metabolomics study. *Nutrition amp; Diabetes*, 12(1).

Li, L., Krznar, P., Erban, A., Agazzi, A., Martin-Levilain, J., Supale, S., Kopka, J., Zamboni, N., and Maechlerr, P. (2019). Metabolomics identifies a biomarker revealing in vivo loss of functional -cell mass before diabetes onset. *Diabetes*, 68(12):2272–2286.

Li, Y., Tian, J., Hou, T., Gu, K., Yan, Q., Sun, S., Zhang, J., Sun, J., Liu, L., Sheng, C.-S., Pang, Y., Cheng, M., Wu, C., Harris, K., Shi, Y., Bloomgarden, Z. T., Chalmers, J., Fu, C., and Ning, G. (2024). Association Between Age at Diabetes Diagnosis and Subsequent Incidence of Cancer: A Longitudinal Population-Based Cohort. *Diabetes Care*, 47(3):353–361.

Liao, P.-C., Chen, M.-S., Jhou, M.-J., Chen, T., Yang, C.-T., and Lu, C.-J. (2022). Integrating Health Data-Driven Machine Learning Algorithms to Evaluate Risk Factors of Early Stage Hypertension at Different Levels of HDL and LDL Cholesterol. *Diagnostics*, 12(8):1965.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Lin, C. C., Li, C. I., Liu, C. S., Lin, W. Y., Yang, S. Y., and Li, T. (2017). Development and validation of a risk prediction model for end-stage renal disease in patients with Type 2 diabetes. *Scientific Reports*, 7(1).

Lin, X., Wang, Q., Yin, P., Tang, L., Tan, Y.-X., Li, H., Yan, K. K., and Xu, G. (2011). A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature

elimination, genetic algorithm and random forest for feature selection. *Metabolomics*, 7(4):549–558.

Lipscombe, L. L., Hwee, J., Webster, L., Shah, B. R., Booth, G. L., and Tu, K. (2018). Identifying diabetes cases from administrative data: a population-based validation study. *BMC Health Services Research*, 18(1).

Liu, J., Semiz, S., Lee, S., Spek, A., Verhoeven, A., Klinken, J., Sijbrands, E., Harms, A., Hankemeier, T., Dijk, K., Duijn, C., and Demirkan, A. (2017). Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study. *Metabolomics*, 13(9).

Liu, Y., Ye, S., Xiao, X., Sun, C., Wang, G., Wang, G., and Zhang, B. (2019). Machine Learning For Tuning, Selection, And Ensemble Of Multiple Risk Scores For Predicting Type 2 Diabetes. *Risk Management and Healthcare Policy*, Volume 12:189–198.

Lock, E. A. and Bonventre, J. V. (2008). Biomarkers in translation; past, present and future. *Toxicology*, 245(3):163–166.

Lund, A., Bagger, J. I., Christensen, M., Knop, F. K., and Vilsbøll, T. (2014). Glucagon and Type 2 Diabetes: the Return of the Alpha Cell. *Current Diabetes Reports*, 14(12).

Ma, H.-T., Hsieh, J.-F., and Chen, S.-T. (2015). Anti-diabetic effects of Ganoderma lucidum. *Phytochemistry*, 114:109–113.

Maisueche Cuadrado, A. (2019). Utilización del Machine Learning en la industria 4.0.

Makroum, M. A., Adda, M., Bouzouane, A., and Ibrahim, H. (2022). Machine Learning and Smart Devices for Diabetes Management: Systematic Review. *Sensors*, 22(5):1843.

Manne-Goehler, J., Geldsetzer, P., Agoudavi, K., Andall-Brereton, G., Aryal, K. K., Bicaba, B. W., Bovet, P., Brian, G., Dorobantu, M., Gathecha, G., Singh Gurung, M., Guwatudde, D., Msaidie, M., Houehanou, C., Houinato, D., Jorgensen, J. M. A., Kagaruki, G. B., Karki, K. B., Labadarios, D., Martins, J. S., Mayige, M. T., McClure, R. W., Mwalim, O., Mwangi, J. K., Norov, B., Quesnel-Crooks, S., Silver, B. K., Sturua, L., Tsabedze, L., Wesseh, C. S., Stokes, A., Marcus, M., Ebert, C., Davies, J. I., Vollmer, S., Atun, R., Bärnighausen, T. W., and Jaacks, L. M. (2019). Health system performance for people with diabetes in 28 low- and middle-income countries: A cross-sectional study of nationally representative surveys. *PLOS Medicine*, 16(3):e1002751.

Mathworks (2024). Support Vector Machine (SVM).

McIntyre, H. D., Catalano, P., Zhang, C., Desoye, G., Mathiesen, E. R., and Damm, P. (2019). Gestational diabetes mellitus. *Nature Reviews Disease Primers*, 5(1).

Medina, F. and Galván, M. (2007). Imputación de datos: teoría y práctica.

Mejía, J. A., Oviedo-Benálcazar, M. A., Ordoñez, J. A., and Valencia, J. F. (2023). Aprendizaje automático aplicado a la predicción de diabetes mellitus, utilizando información socioeconómica y ambiental de usuarios del sistema de salud. *Revista de la Escuela Nacional de Salud Pública*, 41(2):e351168.

Microsoft (2021). Glosario de aprendizaje automático - ML.NET.

Misra, P. and Yadav, A. S. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation. *International Journal on Emerging Technologies*, 11(3):659–665.

Misra, S., Wagner, R., Özkan, B., Schön, M., Sevilla-González, M., Prystupa, K., Wang, C. C., Kreienkamp, R. J., Cromer, S. J., Rooney, M. R., Duan, D., Thuesen, A. C. B., Wallace, A. S., Leong, A., Deutsch, A. J., Andersen, M. K., Billings, L. K., Eckel, R. H., Sheu, W. H. H., Hansen, T., Stefan, N., Goodarzi, M. O., Ray, D., Selvin, E., Florez, J. C., Meigs, J. B., and Udler, M. S. (2023). Systematic review of precision subclassification of type 2 diabetes. *medRxiv (Cold Spring Harbor Laboratory)*.

Mohammad, I. J., Kashanian, S., Rafipour, R., Aljwaid, H., and Hashemi, S. (2023). Evaluation of the relationship of cytokines concentrations tumor necrosis factor-alpha, interleukin-6, and C-reactive protein in obese diabetics and obese non-diabetics: A comparative study. *Biotechnology and Applied Biochemistry*, 71(2):272–279.

Molina, N. (2024). El glucagón y sus funciones.

Moosaie, F., Fatemi Abhari, S. M., Deravi, N., Karimi Behnagh, A., Esteghamati, S., Dehghani Firouzabadi, F., Rabizadeh, S., Nakhjavani, M., and Esteghamati, A. (2021). Waist-to-height ratio is a more accurate tool for predicting hypertension than waist-to-hip circumference and bmi in patients with type 2 diabetes: A prospective study. *Frontiers in Public Health*, 9:726288.

Moreno-Torres, J. G., Saez, J. A., and Herrera, F. (2012). Study on the Impact of Partition-Induced Dataset Shift on -Fold Cross-Validation. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1304–1312.

Morgan-Benita, J., Sánchez-Reyna, A. G., Espino-Salinas, C. H., Oropeza-Valdez, J. J., Luna-García, H., Galván-Tejada, C. E., Galván-Tejada, J. I., Gamboa-Rosales, H., Enciso-Moreno, J. A., and Celaya-Padilla, J. (2022a). Metabolomic Selection in the Progression of Type 2 Diabetes Mellitus: A Genetic Algorithm Approach. *Diagnostics*, 12(11):2803.

Morgan-Benita, J. A., Celaya-Padilla, J. M., Luna-García, H., Galván-Tejada, C. E., Cruz, M., Galván-Tejada, J. I., Gamboa-Rosales, H., Sánchez-Reyna, A. G., Rondon, D., and Villalba-Condori, K. O. (2024). Setting Ranges in Potential Biomarkers for Type 2 Diabetes Mellitus Patients Early Detection By Sex—An Approach with Machine Learning Algorithms. *Diagnostics*, 14(15):1623.

Morgan-Benita, J. A., Galván-Tejada, C. E., Cruz, M., Galván-Tejada, J. I., Gamboa-Rosales, H., Arceo-Olague, J. G., Luna-García, H., and Celaya-Padilla, J. M. (2022b). Hard Voting Ensemble Approach for the Detection of Type 2 Diabetes in Mexican Population with Non-Glucose Related Features. *Healthcare*, 10(8):1362.

Moszczuk, B., Krata, N., Rudnicki, W. R., Foroncewicz, B., Cysewski, D., Paczek, L., Kaleta, B., and Mucha, K. (2022). Osteopontin—A potential biomarker for IGA nephropathy: Machine learning application. *Biomedicines*, 10(4):734.

Mucherino, A., Papajorgji, P. J., and Pardalos, P. M. (2009). k-nearest neighbor classification. *Data Mining in Agriculture*, pages 83–106.

Na8, N. (2020). Algoritmo k-Nearest Neighbor.

Nagaraj, S. B. and Kieneker, L. M. (2021). Kidney Age Index (KAI): a novel age-related biomarker to estimate kidney function in patients with diabetic kidney disease using machine learning. *Computer Methods and Programs in Biomedicine*, 211:106434.

Nath, T. (2021). Body fat predicts exercise capacity in persons with Type 2 Diabetes Mellitus: A machine learning approach.

Nedyalkova, M. (2021). Combinatorial K-Means Clustering as a Machine Learning Tool Applied to Diabetes Mellitus Type 2.

Nedyalkova, M., Madurga, S., Ballabio, D., Robeva, R., Romanova, J., Kichev, I., Elenkova, A., and Simeonov, V. (2020). Diabetes mellitus type 2: Exploratory data analysis based on clinical reading. *Open Chemistry*, 18(1):1041–1053.

Ngiam, K. Y. and Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *Lancet oncology/Lancet. Oncology*, 20(5):e262–e273.

Oh, E., Yoo, T. K., and Park, S. (2013). Diabetic Retinopathy Risk Prediction for FUNDUS Examination Using Sparse Learning: A Cross-sectional study. *BMC Medical Informatics and Decision Making*, 13(1).

Ou, Q., Jin, W., Lin, L., Lin, D., Chen, K., and Quan, H. (2023a). LASSO-based machine learning algorithm to predict the incidence of diabetes in different stages. *The Aging Male*, 26(1).

Ou, S., Tsai, M.-J., Lee, K., Tseng, W., Yang, C., Chen, T.-H., Bin, P.-J., Chen, T. J., Lin, Y., Sheu, W. H., Lee, F.-Y., and Tarng, D. (2023b). Prediction of the risk of developing end-stage renal diseases in newly diagnosed type 2 diabetes mellitus using artificial intelligence algorithms. *BioData Mining*, 16(1).

Ouyang, A., Hu, K., and Chen, L. (2024). Trends and risk factors of diabetes and prediabetes in US adolescents, 1999–2020. *Diabetes Research and Clinical Practice*, 207:111022.

Park, A. and Nam, S. (2023). MIRDM-RFGA: genetic algorithm-based identification of a MIRNA set for detecting Type 2 diabetes. *BMC Medical Genomics*, 16(1).

Pascoe, M. K., Low, P. A., Windebank, A. J., and Litchy, W. J. (1997). Subacute diabetic proximal neuropathy. *Mayo Clinic Proceedings*, 72(12):1123–1132.

Peddinti, G., Cobb, J., Yengo, L., Froguel, P., Kravić, J., Balkau, B., Tuomi, T., Aittokallio, T., and Groop, L. (2017). Early metabolic markers identify potential targets for the prevention of type 2 diabetes. *Diabetologia*, 60(9):1740–1750.

Pedrero, V., Reynaldos-Grandón, K., Ureta-Achurra, J., and Cortez-Pinto, E. (2021). Generalidades del Machine Learning y su aplicación en la gestión sanitaria en Servicios de Urgencia. *Revista médica de Chile*, 149(2):248–254.

Prandi, F. R., Lecis, D., Illuminato, F., Milite, M., Celotto, R., Lerakis, S., Romeo, F., and Barillà, F. (2022). Epigenetic Modifications and Non-Coding RNA in Diabetes-Mellitus-Induced Coronary Artery Disease: Pathophysiological Link and New Therapeutic Frontiers. *International journal of molecular sciences*, 23(9):4589.

Pérez-Lozano, D. L., Camarillo-Nava, V. M., Juárez-Zepeda, T. E., Andrade-Pineda, J. E., Pérez-López, D., Reyes-Pacheco, J. A., Lucho-Gutiérrez, Z. M.,

and Carmona-Aparicio, L. (2023). Costo-efectividad del tratamiento de la diabetes mellitus tipo 2 en México.

Rahbar, S. (2005). The Discovery of Glycated Hemoglobin: A Major Event in the Study of Nonenzymatic Chemistry in Biological Systems. *Annals of the New York Academy of Sciences*, 1043(1):9–19.

Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1).

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology advances*, 49:107739.

RisC, N. (2022). Ranking (

Rocio Diaz, E., Pedram, H., Gabriela Delevati, C., Hilda, A., Lucy, C., Shivanki, J., Glenda, T., Guadalupe J, O., James, S., Natalie, T., Bhanu Priya, G., H. Alex, C., Fudong, L., Louise D., M., and Andrey S., T. (2022). Sex differences in global metabolomic profiles of covid-19 patients. *Cell Death and Disease*, 13(5).

Rodrigo, J. A. (2020). Validación de modelos predictivos (machine learning): Cross-validation, OneLeaveOut, Bootstraping.

Rodríguez-Romero, V., Bergstrom, R. F., Decker, B. S., Lahu, G., Vakilynejad, M., and Bies, R. R. (2019). Prediction of nephropathy in Type 2 diabetes: An analysis of the ACCORD Trial applying machine learning techniques. *Clinical and Translational Science*, 12(5):519–528.

Rojas-García, M., Vázquez, B., Torres-Poveda, K., and Madrid-Marina, V. (2023). Lethality Risk markers by sex and Age-group for COVID-19 in Mexico: a cross-sectional study based on machine learning approach. *BMC Infectious Diseases*, 23(1).

Roointan, A., Gheisari, Y., Hudkins, K. L., and Gholaminejad, A. (2021). Non-invasive metabolic biomarkers for early diagnosis of diabetic nephropathy: Meta-analysis of profiling metabolomics studies. *Nutrition, Metabolism and Cardiovascular Diseases*, 31(8):2253–2272.

Sabitha, E. and Durgadevi, M. (2022). Improving the diabetes diagnosis prediction rate using data preprocessing, data augmentation and recursive feature elimination method. *International Journal of Advanced Computer Science and Applications*, 13(9).

Sadhasivam, J., Muthukumaran, V., Raja, J. T., Joseph, R. B., Munirathanam, M., and Balajee, J. (2021). Diabetes Disease Prediction using Decision Tree for feature selection. *Journal of physics*, 1964(6):062116.

Sadoughi, F., Valinejadi, A., and Salehi, M. (2016). Diabetes knowledge translation status in developing countries: A mixed method study among diabetes researchers in case of Iran. *International Journal of Preventive Medicine*, 7(1):33.

Saigusa, D., Matsukawa, N., Hishinuma, E., and Koshiba, S. (2021). Identification of biomarkers to diagnose diseases and find adverse drug reactions by metabolomics. *Drug metabolism and pharmacokinetics*, 37:100373.

Salihovic, S., Broeckling, C. D., Ganna, A., Prenni, J. E., Sundström, J., Berne, C., Lind, L., Ingelsson, E., Fall, T., Ärnlöv, J., and Nowak, C. (2020). Non-targeted urine metabolomics and associations with prevalent and incident type 2 diabetes. *Scientific Reports*, 10(1).

Sánchez-Reyna, A., Celaya-Padilla, J., Galván-Tejada, C., Luna-García, H., Gamboa-Rosales, H., Ramirez-Morales, A., and Galván-Tejada, J. (2021). Multimodal Early Alzheimer's Detection, a Genetic Algorithm Approach with Support Vector Machines. *Healthcare*, 9(8):971.

Sasako, T. (2023). Exploring mechanisms underlying diabetes comorbidities and strategies to prevent vascular complications. *Diabetology International*, 15(1):34–40.

Satheesh, G., Ramachandran, S., and Jaleel, A. (2020). Metabolomics-based prospective studies and prediction of type 2 diabetes mellitus risks. *Metabolic Syndrome and Related Disorders*, 18(1):1–9.

Savolainen, O., Fagerberg, B., Lind, M. V., Sandberg, A.-S., Ross, A. B., and Bergström, G. (2017). Biomarkers for predicting type 2 diabetes development—can metabolomics improve on existing biomarkers? *PLOS ONE*, 12(7):e0177738.

Schaffner, H., Wiener, J., DeLuca, A., Genovese, A., Deeb, A., Deeb, W., Sheikh-Ali, M., Sutton, D., Gore, A., Berner, J., Huston, J., and Goldfaden, R. (2024). Insulin icodec: A novel once-weekly treatment for diabetes. *Diabetic Medicine*.

Scheffer, J. (2013). Dealing with missing data.

Schena, F. and Gesualdo, L. (2005). Pathogenetic mechanisms of diabetic nephropathy. *Journal of the American Society of Nephrology*, 16(3 suppl 1):S30–S33.

Sevil, M., Rashid, M., Hajizadeh, I., Park, M., Quinn, L., and Cinar, A. (2021). Physical Activity and Psychological Stress Detection and Assessment of Their Effects on Glucose Concentration Predictions in Diabetes Management. *IEEE Transactions on Biomedical Engineering*.

Shanik, M. H., Xu, Y., Skrha, J., Dankner, R., Zick, Y., and Roth, J. (2008). Insulin Resistance and Hyperinsulinemia. *Diabetes Care*, 31(Supplement$_2$) : $S262 - -S268$.

Shen, Z., Wu, Q., Wang, Z., Chen, G., and Lin, B. (2021). Diabetic retinopathy prediction by ensemble learning based on biochemical and physical data. *Sensors*, 21(11):3663.

Singh, N. and Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering*, 40(1):1–22.

Singh, Y. and Tiwari, M. (2022). A novel hybrid approach for detection of type-2 diabetes in women using lasso regression and artificial neural network. *International journal of intelligent systems and applications*, 14(4):11–20.

Siransy-Balayssac, E., Ouattara, S., Yéo, T. A., Kondo, A. L., Touré, M., Dah, C., and Bogui, P. (2020). Physiological variations of blood pressure according to gender and age among healthy young Black Africans aged between 18 and 30 years in Côte d'Ivoire, West Africa. *Physiological Reports*, 8(18).

Slieker, R. C., Van Der Heijden, A. A. W. A., Siddiqui, M. K., Langendoen-Gort, M., Nijpels, G., Herings, R. M. C., Feenstra, T., Moons, K. G., Bell, S., Elders, P., Hart, L. M., and Beulens, J. W. (2021). Performance of prediction models for nephropathy in people with Type 2 Diabetes: Systematic review and external validation study. *BMJ*, page n2134.

Spurr, S., Bally, J., Bullin, C., Allan, D., and McNair, E. (2020). The prevalence of undiagnosed prediabetes/type 2 diabetes, prehypertension/hypertension and obesity among ethnic groups of adolescents in western canada. *BMC pediatrics*, 20:1–9.

Syed, A. H. and Khan, T. (2020). Machine learning-based application for predicting risk of type 2 diabetes mellitus (t2dm) in saudi arabia: A retrospective cross-sectional study. *IEEE Access*, 8:199539–199561.

Thyde, D. N., Mohebbi, A., Bengtsson, H., Jensen, M. L., and Mørup, M. (2021). Machine learning-based adherence detection of type 2 diabetes patients on once-daily basal insulin injections. *Journal of Diabetes Science and Technology*.

Tiwari, P. and Singh, V. B. (2021). Diabetes Disease Prediction using significant Attribute selection and Classification approach. *Journal of physics*, 1714(1):012013.

Trujillo, E., Davis, C., and Milner, J. (2006). Nutrigenomics, Proteomics, Metabolomics, and the Practice of Dietetics. *Journal of the American Dietetic Association*, 106(3):403–413.

Tuppad, A. and Patil, S. D. (2022). Machine learning for diabetes clinical decision support: a review. *Advances in Computational Intelligence*, 2(2).

V., given=Victor, T. and F., given=Francesco, F. (2006). Galgo: an r package for multivariate variable selection using genetic algorithms. *Bioinformatics*, 22(9):1154–1156.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferrán, E. A., Lee, G., Li, B., Madabhushi, A., Shah, P. K., Spitzer, M., and Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature reviews. Drug discover/Nature reviews. Drug discovery*, 18(6):463–477.

Van Der Greef, J., Hankemeier, T., and McBurney, R. N. (2006). Metabolomics-Based Systems Biology and Personalized Medicine: Moving Towards n = 1 Clinical Trials? *Pharmacogenomics*, 7(7):1087–1094.

Verrotti, A., Prezioso, G., Scattoni, R., and Chiarelli, F. (2014). Autonomic Neuropathy in Diabetes Mellitus. *Frontiers in Endocrinology*, 5.

Vinik, A., Mehrabyan, A., Colen, L., and Boulton, A. (2004). Focal Entrapment Neuropathies in Diabetes. *Diabetes Care*, 27(7):1783–1788.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, 17(2):228–243.

Wang, Y. (2021). Genetic Risk Score Increased Discriminant Efficiency of Predictive Models for Type 2 Diabetes Mellitus Using Machine Learning: Cohort Study.

Watanabe, S. (1985). Pattern recognition: human and mechanical.

Wei, H., Sun, J., Shan, W., Xiao, W., Wang, B., Ma, X., Hu, W., Wang, X., and Xia, Y. (2022). Environmental chemical exposure dynamics and machine learning-based prediction of diabetes mellitus. *Science of The Total Environment*, 806:150674.

WHO (2022). Diabetes.

Wiesen, J. P. ((2018)). Benefits, Drawbacks, and Pitfalls of z-Score Weighting.

William H., H., Wen, Y., Simon J., G., Rebecca A., S., Melanie J., D., Kamlesh, K., Guy E.H.M., R., Annelli, S., Torsten, L., Knut, B.-J., Morton B., B., and Nicholas J., W. (2015). Early detection and treatment of type 2 diabetes reduce cardiovascular morbidity and mortality: A simulation of the results of the anglo-danish-dutch study of intensive treatment in people with screen-detected diabetes in primary care (addition-europe). *Diabetes Care*, 38(8):1449–1455.

Wishart, D., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V., Varshavi, D., Varshavi, D., Allen, D., Arndt, D., Khetarpal, N., Sivakumaran, A., Harford, K., Sanford, S., Yee, K., Cao, X., Budinski, Z., Liigand, J., Zhang, L., Zheng, J., Mandal, R., Karu, N., Dambrova, M., Schiöth, H., Greiner, R., and Gautam, V. (2021). Hmdb 5.0: the human metabolome database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631.

World Health Organization (2022). Diabetes.

World Population Review (2021). Diabetes Rates By Country 2021.

Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., and Ma, S. (2019). A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. *High-throughput*, 8(1):4.

XERIDIA (2021). Redes Neuronales artificiales: Qué son y cómo se entrenan.

Xia, J., Broadhurst, D., Wilson, M., and Wishart, D. S. (2012). Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, 9(2):280–299.

Xie, Z., Nikolayeva, O., Luo, J., and Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Preventing Chronic Disease*, 16.

Yang, L., Xue, Y., Wei, J., Dai, Q., and Li, P. (2021). Integrating metabolomic data with machine learning approach for discovery of Q-markers from Jinqi Jiangtang preparation against type 2 diabetes. *Chinese Medicine*.

Ye, J. (2013). Mechanisms of insulin resistance in obesity. *Frontiers of Medicine*, 7(1):14–24.

You, Y., Doubova, S. V., Pinto-Masis, D., Pérez-Cuevas, R., Borja-Aburto, V. H., and Hubbard, A. (2019). Application of machine learning methodology to assess the performance of DIABETIMSS program for patients with type 2 diabetes in family medicine clinics in Mexico.

Zeng, Q., Zhao, M., Wang, F., Li, Y., Li, H., Zheng, J., Chen, X., Zhao, X., Ji, L., Gao, X., Liu, C., Wang, Y., Cheng, S., Xu, J., Pan, B., Sun, J., Li, Y., Li, D., He, Y., and Zheng, L. (2022). Integrating choline and specific intestinal microbiota to classify type 2 diabetes in adults: A machine learning based metagenomics study. *Frontiers in Endocrinology*, 13.

Zhang, J., Fuhrer, T., Ye, H., Kwan, B., Montemayor, D., Tumova, J., Darshi, M., Afshinnia, F., Scialla, J, J., Anderson, A., Porter, A, C., Taliercio, J, J., Rincon-Choles, H., Rao, P., Xie, D., Feldman, H., Sauer, U., Sharma, K., and Natarajan, L. (2022). High-throughput metabolomics and diabetic kidney disease progression: Evidence from the chronic renal insufficiency (cric) study. *American Journal of Nephrology*, 53(2-3):215–225.

Zhang, X.-D. (2020). A Matrix Algebra Approach to Artificial Intelligence.

Zhang, Z., Treviño, V., Hoseini, S. S., Belciug, S., Boopathi, A. M., Zhang, P., Gorunescu, F., Velappan, S., and Dai, S. (2018). Variable selection in logistic regression model with genetic algorithm. *Annals of Translational Medicine*, 6(3):45.

Zheng, H., Ryzhov, I. O., Xie, W., and Zhong, J. (2021). Personalized Multimorbidity Management for Patients with Type 2 Diabetes Using Reinforcement Learning of Electronic Health Records. *Drugs*, 81(4):471–482.

Zhu, K., Qian, F., Lu, Q., Li, R., Qiu, Z., Li, L., Li, R., Yu, H., Deng, Y., Yang, K., Pan, A., and Liu, G. (2024). Modifiable Lifestyle Factors, Genetic Risk, and Incident Peripheral Artery Disease Among Individuals With Type 2 Diabetes: A Prospective Study. *Diabetes Care*, 47(3):435–443.

Zierath, J., Krook, A., and Wallberg-Henriksson, H. (2000). Insulin action and insulin resistance in human skeletal muscle. *Diabetologia*, 43(7):821.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Apéndice A
Publicaciones y reconocimientos

A lo largo del periodo como estudiante de doctorado, se han realizado tres publicaciones en revistas JCR como primer autor relacionadas con el tema de investigación. Asimismo, se ha colaborado en siete publicaciones adicionales, de las cuales tres son en revistas JCR y cuatro en congresos, todas en calidad de coautor. Finalmente, se llevaron a cabo doce actividades enfocadas en la contribución al conocimiento.

Recientemente, se presentó la tesis de Julio César Morales Hernández para obtener el título de ingeniero en electrónica industrial, en la cual participé como secretario y miembro del comité sinodal (ver figura A.1). Además, actualmente estoy codirigiendo dos tesis más: una de maestría en ciencias del procesamiento de la información, de Ciro Edgar Robles Muro, y otra de ingeniería electrónica industrial, de Yedid Vianey Herrera Trejo.

Adicionalmente a las actividades de contribución al conocimiento, formamos parte del equipo encargado del desarrollo de una patente para una urna electrónica, un proyecto en colaboración con profesores y estudiantes de la Universidad Autónoma de Zacatecas, en el que participamos en la conceptualización de las interfaces y los flujos de información.

Título	**Revista**
Hard Voting Ensemble Approach for the Detection of Type 2 Diabetes in Mexican Population with Non-Glucose Related Features. DOI: 10.3390/healthcare10081362 (vea figura A.2)	MDPI-Healthcare
Metabolomic Selection in the Progression of Type 2 Diabetes Mellitus: A Genetic Algorithm Approach. DOI: 10.3390/diagnostics12112803 (vea figura A.3)	MDPI-Diagnostics
Setting Ranges in Potential Biomarkers for Type 2 Diabetes Mellitus Patients Early Detection By Sex—An Approach with Machine Learning Algorithms. DOI: 10.3390/diagnostics14151623 (vea figura A.4)	MDPI-Diagnostics

Tabla A.1: Publicaciones como primer autor

Título	**Revista**
Driver Identification Using Statistical Features of Motor Activity and Genetic Algorithms. DOI: 10.3390/s23020784 (vea figura A.5)	MDPI-Sensors
Feature Selection of Motor Activity in Intervals of Time with Genetics Algorithms for Depression Detection. DOI: 10.17488/RMIB.44.4.3 (vea figura A.6)	Revista Mexicana de Ingeniería Biomédica
Driver Identification Using Machine Learning and Motor Activity as Data Source (vea figura A.7)	SpringerLink-Human-Computer Interaction (HCI-COLLAB 2022)
Selección de metabolitos como características de un modelo de bosques aleatorios para el diagnóstico del COVID-19 (vea figura A.8)	Institulo Politécnico Nacional-Research in Computing Science
Synthetic Data in the Detection of States of Cognitive Progression to Alzheimer's through Neuropsychological Assessments and Machine Learning Models (vea figura A.9)	CLAIB-CLASD 2024
Synthetic data analysis for early detection of Alzheimer progression through machine learning algorithms	Peerj 2024
Synthetic Data in the Detection of States of Cognitive Progression to Alzheimer's through Neuropsychological Assessments and Machine Learning Models	CLAIB 2024

Tabla A.2: Publicaciones como coautor

Título
Congreso RELEEM (vea figura A.10)
Instructor del taller de Flutter y Dart en LABSOL (vea figura A.11)
Acceso universal al conocimiento (vea figura A.12)
Intel Challenge (A.13)
Instructor del taller de Python en La Maestría en Ciencias del Procesamiento de la Información (vea figura A.14)
Estancia internacional en Colombia (vea figura A.15)
Conferencia Internacional (vea figura A.16)
Participación en foro y mesas de trabajo (vea figura A.17)
Participación en la Jornada Estatal de Ciencia y Tecnología (vea figura A.18)
Jurado del Concurso nacional de prototipado (vea figura A.19)
Speaker en Talent Land (vea figura A.20)
Participación en la impartición de examen de ingreso CENEVAL (vea figura A.21)

Tabla A.3: Actividades extracurriculares para la contribución al conocimiento

Universidad Autónoma de Zacatecas
"Francisco García Salinas"

COPIA SIMPLE DEL
ACTA DE OTORGAMIENTO DE NIVEL LICENCIATURA

UNIDAD ACADÉMICA DE INGENIERÍA ELECTRICA

ACTA NO. 84635

ACTA DE OTORGAMIENTO DE NIVEL LICENCIATURA EN INGENIERÍA ELECTRÓNICA INDUSTRIAL

Al C. MORALES HERNANDEZ JULIO CESAR.

En la ciudad de Zacatecas, Zac., a veintiséis días del mes de septiembre del año dos mil veinticuatro, reunidos en Jurado de Examen los CC. DR. ANTONIO BALTAZAR RAIGOSA, DRA. ANA GABRIELA SANCHEZ REYNA, M. EN I. CLAUDIA REYES RIVAS, MTRO. JORGE ALEJANDRO MORGAN BENITA, M.C.T.E. NADIA GABRIELA GARIBAY RENDÓN. Bajo la presidencia del último y en cumplimiento del Acuerdo de la Rectoría de la Universidad Autónoma de Zacatecas, "Francisco García Salinas", de fecha trece del mes de septiembre del año dos mil veinticuatro, se procedió a practicar EXAMEN PROFESIONAL DE TESIS DE LA LICENCIATURA EN INGENIERÍA ELECTRÓNICA INDUSTRIAL.

Al C. MORALES HERNANDEZ JULIO CESAR.

De conformidad con lo dispuesto por los artículos 188 fracción I del Estatuto General Vigente; 149 al 165 del Reglamento Escolar General; y una vez que se cumplió con la evaluación, el jurado tuvo a bien:

APROBARLO POR UNANIMIDAD

Otorgando al C. MORALES HERNANDEZ JULIO CESAR el nivel de LICENCIATURA EN INGENIERÍA ELECTRÓNICA INDUSTRIAL.

Lo que hizo saber al sustentante, firmando las personas que formaron parte del jurado:

M.C.T.E. NADIA GABRIELA GARIBAY RENDÓN
CED. PROF. 4903276
PRESIDENTE

MTRO. JORGE ALEJANDRO MORGAN BENITA
CED. PROF. 12444456
SECRETARIO

M. EN I. CLAUDIA REYES RIVAS
CED. PROF. 8836221
VOCAL

DRA. ANA GABRIELA SANCHEZ REYNA
CED. PROF. 13036294
VOCAL

DR. ANTONIO BALTAZAR RAIGOSA
CED. PROF. 14566[illegible]
VOCAL

Leída que fue el acta anterior y habiendo sido protestado al sustentante en forma, para el fiel y leal desempeño del nivel, cuyo título en esta fecha adquiere, habiendo aceptado las responsabilidades inherentes a éste, los miembros del jurado acordaron hacer entrega de la documentación respectiva. Así mismo, en esta fecha se entregó copia certificada de la presente acta.

DR. ANGEL ROMAN GUTIERREZ
SECRETARIO GENERAL DE LA U.A.Z.

COPIA.- Coordinación de Unidad Académica

Figura A.1: Tesis: Desarrollo de un punto de venta web en php para la optimización de procesos comerciales.

healthcare MDPI

Article

Hard Voting Ensemble Approach for the Detection of Type 2 Diabetes in Mexican Population with Non-Glucose Related Features

Jorge A. Morgan-Benita [1], Carlos E. Galván-Tejada [1], Miguel Cruz [2], Jorge I. Galván-Tejada [1], Hamurabi Gamboa-Rosales [1], Jose G. Arceo-Olague [1], Huizilopoztli Luna-García [1,*] and José M. Celaya-Padilla [1,*]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, Zacatecas 98000, Mexico; alejandro.morgan@uaz.edu.mx (J.A.M.-B.); ericgalvan@uaz.edu.mx (C.E.G.-T.); gatejo@uaz.edu.mx (J.I.G.-T.); hamurabigr@uaz.edu.mx (H.G.-R.); arceojg@uaz.edu.mx (J.G.A.-O.)
[2] Unidad de Investigación Médica en Bioquímica, Hospital de Especialidades, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Av. Cuauhtémoc 330, Col. Doctores, Del. Cuauhtémoc, Mexico City 06720, Mexico; miguel.cruzlo@imss.gob.mx
* Correspondence: hlugar@uaz.edu.mx (H.L.-G.); jose.celaya@uaz.edu.mx (J.M.C.-P.)

Abstract: Type 2 diabetes mellitus (T2DM) represents one of the biggest health problems in Mexico, and it is extremely important to early detect this disease and its complications. For a noninvasive detection of T2DM, a machine learning (ML) approach that uses ensemble classification models with dichotomous output that is also fast and effective for early detection and prediction of T2D can be used. In this article, an ensemble technique by hard voting is designed and implemented using generalized linear regression (GLM), support vector machines (SVM) and artificial neural networks (ANN) for the classification of T2DM patients. In the materials and methods as a first step, the data is balanced, standardized, imputed and integrated into the three models to classify the patients in a dichotomous result. For the selection of features, an implementation of LASSO is developed, with a 10-fold cross-validation and for the final validation, the Area Under the Curve (AUC) is used. The results in LASSO showed 12 features, which are used in the implemented models to obtain the best possible scenario in the developed ensemble model. The algorithm with the best performance of the three is SVM, this model obtained an AUC of 92% ± 3%. The ensemble model built with GLM, SVM and ANN obtained an AUC of 90% ± 3%.

Keywords: ensemble model; machine learning; logistic regression; support vector machine; neural networks; type 2 diabetes mellitus detection

Citation: Morgan-Benita, J.A.; Galván-Tejada, C.E.; Cruz, M.; Galván-Tejada, J.I.; Gamboa-Rosales, H.; Arceo-Olague, J.G.; Luna-García, H.; Celaya-Padilla, J.M. Hard Voting Ensemble Approach for the Detection of Type 2 Diabetes in Mexican Population with Non-Glucose Related Features. *Healthcare* **2022**, *10*, 1362. https://doi.org/10.3390/healthcare10081362

Academic Editor: Tin-Chih Toly Chen

Received: 16 June 2022
Accepted: 15 July 2022
Published: 22 July 2022

1. Introduction

According to the World Health Organization (WHO), diabetes is "a chronic metabolic disease characterized by high levels of glucose in the blood, which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves" [1], being Type 2 diabetes mellitus (T2DM) the most common type. T2DM develops primarily due to an inactive lifestyle, lack of exercise, and body weight [2]. According to the International Diabetes Federation (IDF), in 2021 there were 537 million cases in the world with ages between 20 and 79 years, 541 million adults are at increased risk of developing T2DM [3]. For more than 60 million people living with T2DM, insulin is essential to reduce the risk of kidney failure, blindness and limb amputation [4]. The proportion of people affected with T2DM that is related to medical complications is alarming, either because of the lack of these drugs or because of their cost. In Mexico, according to the 2020 mortality data given by the Instituto Nacional de Estadística y Geografía INEGI, in Mexico, 1,086,743 deaths were reported, of which 14% (151,019) correspond to deaths from diabetes mellitus. Of

Figura A.2: Hard Voting Ensemble Approach for the Detection of Type 2 Diabetes in Mexican Population with Non-Glucose Related Features

diagnostics

MDPI

Article

Metabolomic Selection in the Progression of Type 2 Diabetes Mellitus: A Genetic Algorithm Approach

Jorge Morgan-Benita [1,†], Ana G. Sánchez-Reyna [1,†], Carlos H. Espino-Salinas [1,†], Juan José Oropeza-Valdez [2], Huizilopoztli Luna-García [1], Carlos E. Galván-Tejada [1], Jorge I. Galván-Tejada [1], Hamurabi Gamboa-Rosales [1], Jose Antonio Enciso-Moreno [2] and José Celaya-Padilla [1,*]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Mexico
[2] Metabolomics and Proteomics Laboratory, Autonomous University of Zacatecas, Zacatecas 98000, Mexico
* Correspondence: jose.celaya@uaz.edu.mx
† These authors contributed equally to this work.

Abstract: According to the World Health Organization (WHO), type 2 diabetes mellitus (T2DM) is a result of the inefficient use of insulin by the body. More than 95% of people with diabetes have T2DM, which is largely due to excess weight and physical inactivity. This study proposes an intelligent feature selection of metabolites related to different stages of diabetes, with the use of genetic algorithms (GA) and the implementation of support vector machines (SVMs), K-Nearest Neighbors (KNNs) and Nearest Centroid (NEARCENT) and with a dataset obtained from the Instituto Mexicano del Seguro Social with the protocol name of the following: "Análisis metabolómico y transcriptómico diferencial en orina y suero de pacientes pre diabéticos, diabéticos y con nefropatía diabética para identificar potenciales biomarcadores pronósticos de daño renal" (differential metabolomic and transcriptomic analyses in the urine and serum of pre-diabetic, diabetic and diabetic nephropathy patients to identify potential prognostic biomarkers of kidney damage). In order to analyze which machine learning (ML) model is the most optimal for classifying patients with some stage of T2DM, the novelty of this work is to provide a genetic algorithm approach that detects significant metabolites in each stage of progression. More than 100 metabolites were identified as significant between all stages; with the data analyzed, the average accuracies obtained in each of the five most-accurate implementations of genetic algorithms were in the range of 0.8214–0.9893 with respect to average accuracy, providing a precise tool to use in detections and backing up a diagnosis constructed entirely with metabolomics. By providing five potential biomarkers for progression, these extremely significant metabolites are as follows: "Cer(d18:1/24:1) i2", "PC(20:3-OH/P-18:1)", "Ganoderic acid C2", "TG(16:0/17:1/18:1)" and "GPEtn(18:0/20:4)".

Keywords: genetic algorithm; machine learning; metabolites; type 2 diabetes

Citation: Morgan-Benita, J.; Sánchez-Reyna, A.G.; Espino-Salinas, C.H.; Oropeza-Valdez, J.J.; Luna-García, H.; Galván-Tejada, C.E.; Galván-Tejada, J.I.; Gamboa-Rosales, H.; Enciso-Moreno, J.A.; Celaya-Padilla, J. Metabolomic Selection in the Progression of Type 2 Diabetes Mellitus: A Genetic Algorithm Approach. *Diagnostics* **2022**, *12*, 2803. https://doi.org/10.3390/diagnostics12112803

Academic Editors: Yuli Huang, Yong Yuan and Peisong Chen

Received: 15 October 2022
Accepted: 7 November 2022
Published: 15 November 2022

1. Introduction

Diabetes is a chronic and progressive disease that occurs in the pancreas when it is no longer able to make a hormone known as insulin or when the body is unable to use it properly [1]. Adults numbering 537 million (20–79 years) currently live with diabetes in the world, and over 6.7 million deaths in 2021 are reported (approximately one death every 5 s) [2]. Type 2 diabetes mellitus (T2DM) is a progressive condition that is produced by relative insulin deficiencies caused by pancreatic β-cell (cells that synthesize and secrete insulin and amylin) dysfunction and insulin resistance [3]. The International Diabetes Federation (IDF) presented in 2021 that 541 million people in adulthood present a higher risk of developing T2DM [4]. As T2DM progresses, the comorbidities associated with hyperglycemia that induces renal damage directly or via hemodynamic modifications appear, which cause

Figura A.3: Metabolomic Selection in the Progression of Type 2 Diabetes Mellitus: A Genetic Algorithm Approach

diagnostics

MDPI

Article

Setting Ranges in Potential Biomarkers for Type 2 Diabetes Mellitus Patients Early Detection By Sex—An Approach with Machine Learning Algorithms

Jorge A. Morgan-Benita [1], José M. Celaya-Padilla [1,*], Huizilopoztli Luna-García [1], Carlos E. Galván-Tejada [1], Miguel Cruz [2], Jorge I. Galván-Tejada [1], Hamurabi Gamboa-Rosales [1], Ana G. Sánchez-Reyna [1], David Rondon [3] and Klinge O. Villalba-Condori [4]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Zacatecas 98000, Mexico; alejandro.morgan@uaz.edu.mx (J.A.M.-B.); hlugar@uaz.edu.mx (H.L.-G.); ericgalvan@uaz.edu.mx (C.E.G.-T.); gatejo@uaz.edu.mx (J.I.G.-T.); hamurabigr@uaz.edu.mx (H.G.-R.); agsreyna@uaz.edu.mx (A.G.S.-R.)
[2] Unidad de Investigación Médica en Bioquímica, Hospital de Especialidades, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, Ciudad de México 06720, Mexico; mcruzl@yahoo.com
[3] Departamento de Estudios Generales, Universidad Continental, Arequipa 04001, Peru; drondon@continental.edu.pe
[4] Vicerrectorado de Investigación, Universidad Católica de Santa María, Yanahuara 04013, Peru; kvillalba@ucsm.edu.pe
* Correspondence: jose.celaya@uaz.edu.mx

Abstract: Type 2 diabetes mellitus (T2DM) is one of the most common metabolic diseases in the world and poses a significant public health challenge. Early detection and management of this metabolic disorder is crucial to prevent complications and improve outcomes. This paper aims to find core differences in male and female markers to detect T2DM by their clinic and anthropometric features, seeking out ranges in potential biomarkers identified to provide useful information as a pre-diagnostic tool whie excluding glucose-related biomarkers using machine learning (ML) models. We used a dataset containing clinical and anthropometric variables from patients diagnosed with T2DM and patients without TD2M as control. We applied feature selection with three different techniques to identify relevant biomarker models: an improved recursive feature elimination (RFE) evaluating each set from all the features to one feature with the Akaike information criterion (AIC) to find optimal outputs; Least Absolute Shrinkage and Selection Operator (LASSO) with glmnet; and Genetic Algorithms (GA) with GALGO and forward selection (FS) applied to GALGO output. We then used these for comparison with the AIC to measure the performance of each technique and collect the optimal set of global features. Then, an implementation and comparison of five different ML models was carried out to identify the most accurate and interpretable one, considering the following models: logistic regression (LR), artificial neural network (ANN), support vector machine (SVM), k-nearest neighbors (KNN), and nearest centroid (Nearcent). The models were then combined in an ensemble to provide a more robust approximation. The results showed that potential biomarkers such as systolic blood pressure (SBP) and triglycerides are together significantly associated with T2DM. This approach also identified triglycerides, cholesterol, and diastolic blood pressure as biomarkers with differences between male and female actors that have not been previously reported in the literature. The most accurate ML model was selection with RFE and random forest (RF) as the estimator improved with the AIC, which achieved an accuracy of 0.8820. In conclusion, this study demonstrates the potential of ML models in identifying potential biomarkers for early detection of T2DM, excluding glucose-related biomarkers as well as differences between male and female anthropometric and clinic profiles. These findings may help to improve early detection and management of the T2DM by accounting for differences between male and female subjects in terms of anthropometric and clinic profiles, potentially reducing healthcare costs and improving personalized patient attention. Further research is needed to validate these potential biomarkers ranges in other populations and clinical settings.

check for updates

Citation: Morgan-Benita, J.A.; Celaya-Padilla, J.M.; Luna-García, H.; Galván-Tejada, C.E.; Cruz, M.; Galván-Tejada, J.I.; Gamboa-Rosales, H.; Sánchez-Reyna, A.G.; Rondon, D.; Villalba-Condori, K.O. Setting Ranges in Potential Biomarkers for Type 2 Diabetes Mellitus Patients Early Detection By Sex—An Approach with Machine Learning Algorithms. *Diagnostics* **2024**, *14*, 1623. https://doi.org/10.3390/diagnostics14151623

Academic Editor: Aw Tar-Choon

Received: 19 June 2024
Revised: 20 July 2024
Accepted: 24 July 2024
Published: 27 July 2024

Figura A.4: Setting Ranges in Potential Biomarkers for Type 2 Diabetes Mellitus Patients Early Detection By Sex—An Approach with Machine Learning Algorithms

sensors

MDPI

Article

Driver Identification Using Statistical Features of Motor Activity and Genetic Algorithms

Carlos H. Espino-Salinas [1,†], Huizilopoztli Luna-García [1,*], José M. Celaya-Padilla [2], Jorge A. Morgan-Benita [1,†], Cesar Vera-Vasquez [3], Wilson J. Sarmiento [4], Carlos E. Galván-Tejada [1], Jorge I. Galván-Tejada [1], Hamurabi Gamboa-Rosales [1] and Klinge Orlando Villalba-Condori [5]

[1] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Mexico
[2] CONACYT, Universidad Autónoma de Zacatecas, Jardín Juarez 147, Centro, Zacatecas 98000, Mexico
[3] Ingeniería Mecanica, Universidad Continental, Arequipa 04002, Peru
[4] Ingeniería en Multimedia, Universidad Militar de Nueva Granada, Cra 11, Bogotá 101-80, Colombia
[5] Vicerrectorado de Investigación, Universidad Católica de Santa María, Arequipa 04002, Peru
* Correspondence: hlugar@uaz.edu.mx; Tel.: +52-492-124-5533
† These authors contributed equally to this work.

Abstract: Driver identification refers to the process whose primary purpose is identifying the person behind the steering wheel using collected information about the driver him/herself. The constant monitoring of drivers through sensors generates great benefits in advanced driver assistance systems (ADAS), to learn more about the behavior of road users. Currently, there are many research works that address the subject in search of creating intelligent models that help to identify vehicle users in an efficient and objective way. However, the different methodologies proposed to create these models are based on data generated from sensors that include different vehicle brands on routes established in real environments, which, although they provide very important information for different purposes, in the case of driver identification, there may be a certain degree of bias due to the different situations in which the route environment may change. The proposed method seeks to intelligently and objectively select the most outstanding statistical features from motor activity generated in the main elements of the vehicle with genetic algorithms for driver identification, this process being newer than those established by the state-of-the-art. The results obtained from the proposal were an accuracy of 90.74% to identify two drivers and 62% for four, using a Random Forest Classifier (RFC). With this, it can be concluded that a comprehensive selection of features can greatly optimize the identification of drivers.

Keywords: driver identification; genetic algorithms; feature extraction; ADAS; random forest

Citation: Espino-Salinas, C.H.; Luna-García, H.; Celaya-Padilla, J.M.; Morgan-Benita, J.A.; Vera-Vasquez, C.; Sarmiento, W.J.; Galván-Tejada, C.E.; Galván-Tejada, J.I.; Gamboa-Rosales, H.; Villalba-Condori, K.O. Driver Identification Using Statistical Features of Motor Activity and Genetic Algorithms. *Sensors* **2023**, *23*, 784. https://doi.org/10.3390/s23020784

Academic Editor: Arturo de la Escalera Hueso

Received: 9 December 2022
Revised: 31 December 2022
Accepted: 5 January 2023
Published: 10 January 2023

1. Introduction

Currently, there are systems capable of improving the quality and experience of vehicle users; an example are the Advanced Driver-Assistance Systems (ADAS). These systems have gained considerable attention in the automotive field as enablers for vehicle energy consumption, safety, and comfort enhancement [1]. These systems exploit a large number of sensors available in some vehicles today and serve the driver, alerting him in cases of potential problems. However, their generated data can also be useful for ensuring vehicle safety by recognizing the driver's identity. This is advantageous for ADAS because it guarantees vehicle security by warning the owner in the event of an unauthorized driver or even theft, as in recent years, vehicle theft has increased around the world. According to the FBI, vehicle theft in 2017 is estimated at 773,139 in the US, a 10.4% increase compared to the 2013 report, which gives a general overview of this problem. In the case of the ADAS

Figura A.5: Driver Identification Using Statistical Features of Motor Activity and Genetic Algorithms

RESEARCH ARTICLE
VOL. 44 | NO. 3 | **SPECIAL ISSUE 2023** | PP 38 - 52

Revista Mexicana de Ingeniería Biomédica

https://doi.org/10.17488/RMIB.44.4.3

E-LOCATION ID: 1364

Feature Selection of Motor Activity in Intervals of Time with Genetics Algorithms for Depression Detection

Selección de Características de la Actividad Motora en Intervalos de Tiempo con Algoritmos Genéticos para la Detección de Depresión

Carlos H. Espino-Salinas [1], *Carlos E. Galván-Tejada*[1], *Ana G. Sánchez-Reyna*[1], *Huizilopoztli Luna-García*[1], *Hamurabi Gamboa-Rosales*[1], *Jorge A. Morgan-Benita*[1], *José M. Celaya-Padilla*[1], *Jorge I. Galván-Tejada*[1],

[1]Universidad Autónoma de Zacatecas, Zacatecas - México

ABSTRACT

It is estimated that depression affects more than 300 million people in worldwide. Unfortunately, the current method of psychiatric evaluation requires a great effort on the part of clinicians to collect complete information. The aim of this paper is determine the optimal time intervals to detect depression using genetic algorithms and machine learning techniques; from motor activity readings of 55 participants during a week at one-minute intervals. The time intervals with the best performance in detecting depression in individuals were selected by applying Genetic Algorithms (GA). Methodology. 385 observations of the study participants were evaluated, obtaining an accuracy of 83.0 % with Logistic Regression (LR). Conclusion. There is a relationship between motor activity and people with depression since it is possible to detect it using machine learning techniques. However, the changes in the variables of the time intervals could be established as key factors since, at different times, they could give good or bad results because the motor activity in the patients could vary. However, the results present a first approximation for developing tools that help the opportune and objective diagnosis of depression.

KEYWORDS: artificial intelligence, depression, feature selection, genetic algorithm, motor activity

Figura A.6: Feature Selection of Motor Activity in Intervals of Time with Genetics Algorithms for Depression Detection

Figura A.7: Driver Identification Using Machine Learning and Motor Activity as Data Source

ISSN 1870-4069

Selección de metabolitos como características de un modelo de bosques aleatorios para el diagnóstico del COVID-19

Hugo Alexis Torres-Pasillas, José María Celaya-Padilla, Yamilé López-Hernández, Carlos Erick Galván-Tejada, Alejandra García-Hernández, Pedro Daniel Alaniz-Lumbreras, José Alejandro Morgan-Benita

Universidad Autónoma de Zacatecas, Unidad Académica de Ingeniería Eléctrica, México

ylopezher@conacyt.mx {hugo.tpasillas, jose.celaya, ericgalvan, alegarcia, dalaniz, alejandro.morgan}@uaz.edu.mx

Resumen. El COVID-19 es una enfermedad reciente que surgió a finales de 2019 causado por un nuevo tipo de coronavirus. A pesar de los avances en la investigación del virus y el desarrollo tanto de vacunas como de posibles tratamientos, el diagnóstico de la enfermedad, especialmente de forma temprana, continúa siendo una de las mejores herramientas para combatir la enfermedad y su transmisión. El objetivo de este estudio es seleccionar el mejor conjunto de metabolitos como potenciales biomarcadores para el diagnóstico, que son utilizados como características de un modelo de bosques aleatorios. Para ello, se utilizaron 4 diferentes técnicas de selección de características que son utilizadas con frecuencia dentro del Aprendizaje Automático, y un conjunto de datos que contiene mediciones de 110 metabolitos de 158 pacientes sospechosos de COVID-19 (121 enfermos y 37 sanos confirmados por pruebas rt-PCR). Los resultados muestran cuatro distintos conjuntos de metabolitos capaces de diagnosticar el COVID-19 con un alto desempeño en 6 distintas métricas utilizadas. El conjunto con mejor rendimiento en el conjunto de entrenamiento consta de 15 metabolitos y logra tener un desempeño alto en la validación a ciegas (f1=0.921, exactitud balanceada=0.875, AUC=0.910), mientras que el conjunto con menor número de características (5) obtiene el segundo mejor rendimiento en el conjunto de entrenamiento pero el mejor desempeño en la validación a ciegas (f1=0.931, exactitud balanceada=0.896, AUC=0.858).

Palabras clave: COVID-19, aprendizaje automático, metabolitos, selección de características, diagnóstico.

Selection of Metabolites as Features of a Random Forest Model for COVID-19 Diagnosis

Abstract. COVID-19 is a recent disease that emerged in late 2019 caused by a new type of coronavirus. Despite advances in virus research and the development of both vaccines and potential treatments, early and accurate

pp. 161–174; rec. 2023-04-17; acc. 2023-05-01 161 *Research in Computing Science* 152(6), 2023

Figura A.8: Selección de metabolitos como características de un modelo de bosques aleatorios para el diagnóstico del COVID-19

Universidad Francisco de Paula Santander
Ocaña - Colombia
FACULTAD DE INGENIERÍA

La Universidad Francisco de Paula Santander seccional Ocaña y la Facultad de Ingeniería

Certifican que:

Jorge Alejandro Morgan Benita

Pasaporte. N06197129 de Sinaloa, México

Participó como ponente oral en el

11th International Conference of Technological Innovation

Dado en los días 11, 12 y 13 de septiembre de 2024

Titulo de la ponencia: *"Plataforma web para el diagnóstico temprano de la enfermedad de alzheimer por medio de modelos de aprendizaje automático"*

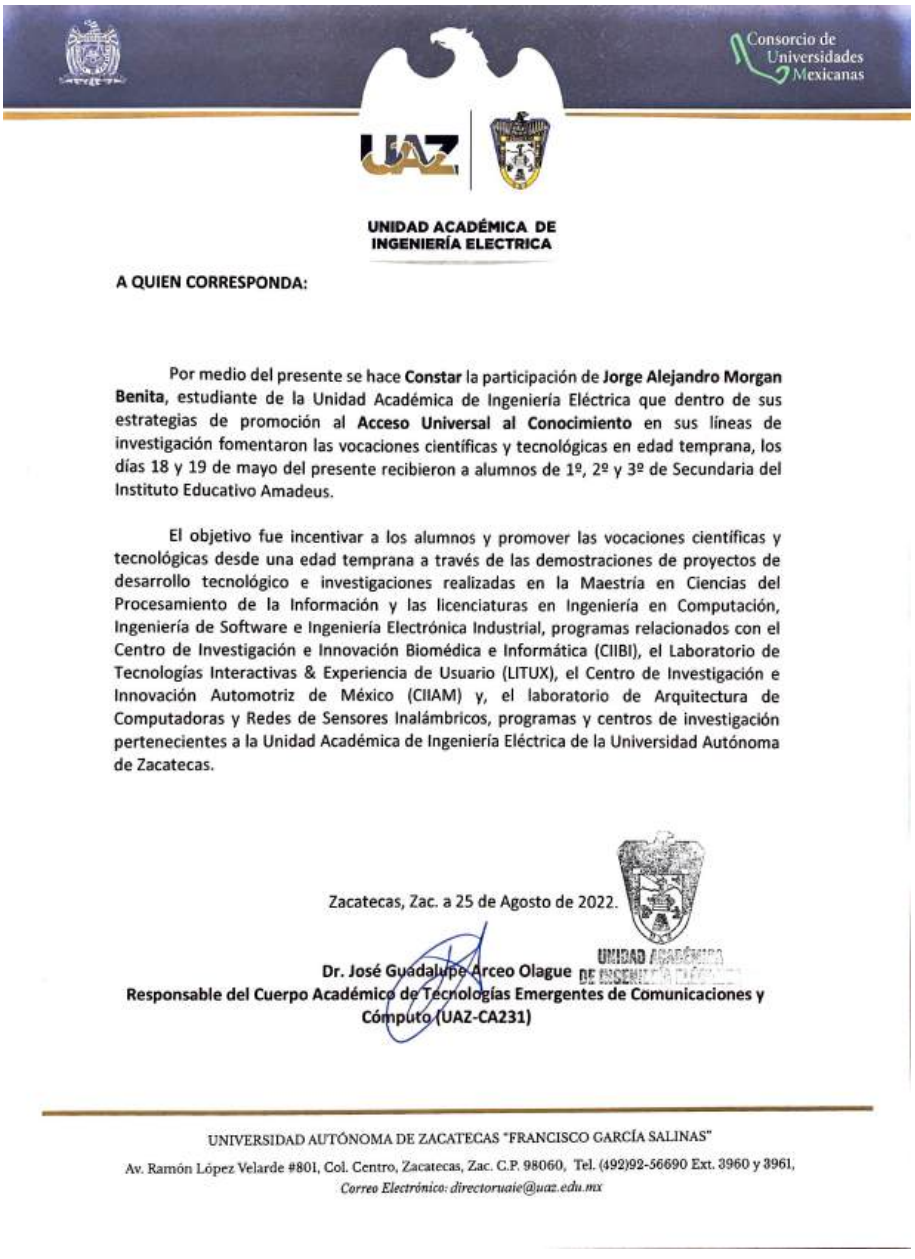
Édgar Antonio Sánchez Ortiz
Director UFPS Ocaña

Phd. Nelson Afanador García
Decano Facultad de Ingeniería

Figura A.9: Synthetic Data in the Detection of States of Cognitive Progression to Alzheimer's through Neuropsychological Assessments and Machine Learning Models, Trabajo aceptado en CLAIB-CLASD

Figura A.10: Congreso RELEEM

Figura A.11: Instructor del taller de Flutter y Dart en LABSOL



Consorcio de Universidades Mexicanas

UAZ

UNIDAD ACADÉMICA DE INGENIERÍA ELECTRICA

A QUIEN CORRESPONDA:

Por medio del presente se hace **Constar** la participación de **Jorge Alejandro Morgan Benita**, estudiante de la Unidad Académica de Ingeniería Eléctrica que dentro de sus estrategias de promoción al **Acceso Universal al Conocimiento** en sus líneas de investigación fomentaron las vocaciones científicas y tecnológicas en edad temprana, los días 18 y 19 de mayo del presente recibieron a alumnos de 1º, 2º y 3º de Secundaria del Instituto Educativo Amadeus.

El objetivo fue incentivar a los alumnos y promover las vocaciones científicas y tecnológicas desde una edad temprana a través de las demostraciones de proyectos de desarrollo tecnológico e investigaciones realizadas en la Maestría en Ciencias del Procesamiento de la Información y las licenciaturas en Ingeniería en Computación, Ingeniería de Software e Ingeniería Electrónica Industrial, programas relacionados con el Centro de Investigación e Innovación Biomédica e Informática (CIIBI), el Laboratorio de Tecnologías Interactivas & Experiencia de Usuario (LITUX), el Centro de Investigación e Innovación Automotriz de México (CIIAM) y, el laboratorio de Arquitectura de Computadoras y Redes de Sensores Inalámbricos, programas y centros de investigación pertenecientes a la Unidad Académica de Ingeniería Eléctrica de la Universidad Autónoma de Zacatecas.

Zacatecas, Zac. a 25 de Agosto de 2022.

UNIDAD ACADÉMICA DE INGENIERÍA ELÉCTRICA

Dr. José Guadalupe Arceo Olague
Responsable del Cuerpo Académico de Tecnologías Emergentes de Comunicaciones y Cómputo (UAZ-CA231)

UNIVERSIDAD AUTÓNOMA DE ZACATECAS "FRANCISCO GARCÍA SALINAS"
Av. Ramón López Velarde #801, Col. Centro, Zacatecas, Zac. C.P. 98060, Tel. (492)92-56690 Ext. 3960 y 3961,
Correo Electrónico: directoruaie@uaz.edu.mx

Figura A.12: Acceso universal al conocimiento

Figura A.13: Intel Challenge

Figura A.14: Instructor del taller de Python en La Maestría en Ciencias del Procesamiento de la Información

UNIVERSIDAD MILITAR NUEVA GRANADA

40 AÑOS
UNIDOS POR LA EXCELENCIA

UMNG-VICACD-FACING

Bogotá, Junio de 2023

Cordial saludo

Me permito certificar que el estudiante **Jorge Alejandro Morgan Benita**, del Doctorado en Ciencias de la Ingeniería de la Universidad Autónoma de Zacatecas; realizó una estancia de investigación entre el 23 de mayo y el 09 de junio, en el programa de Doctorado de Ingeniería, bajo la dirección y acompañamiento del profesor Ing. Wilson Javier Sarmiento, M.Sc., Ph.D, líder del Grupo de Investigación en Multimedia -GIM-.

En nombre de la Comunidad Neogranadina, queremos manifestar que estamos muy complacidos de que se haya elegido nuestra Alma Mater para realizar la estancia de investigación en el período académico descrito.

Atentamente,

Nancy Olarte

ING. NANCY ESPERANZA OLARTE LÓPEZ, M.Sc.
Decana, Facultad de Ingeniería- Sede Bogotá
Universidad Militar Nueva Granada
Bogotá, Colombia

Sede Bogotá, Carrera 11 n.° 101-80,
Sede Campus Nueva Granada, kilómetro 2 vía Cajicá-Zipaquirá
PBX (571) 650 00 00 - 634 32 00
www.umng.edu.co
Colombia-Sur América

Figura A.15: Estancia internacional en Colombia

Figura A.16: Conferencia Internacional

Zacatecas
GOBIERNO DEL ESTADO
2021-2027

CONSEJO ZACATECANO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN
ESTADO DE ZACATECAS

Bienestar y Progreso

El Consejo Zacatecano de Ciencia, Tecnología e Innovación
otorga la presenta

C O N S T A N C I A

A:

Jorge Alejandro Morgan B.

Por su participación en el Foro y mesas de trabajo de personas, pueblos y comunidades Indígenas Originarios y comunidades afromexicanos del Estado de Zacatecas, cuidado del medio ambiente, para la Nueva Ley de Humanidades, Ciencias, Tecnologías e Innovación.

Zacatecas, Zacatecas, a 20 de octubre de 2023.

Dr. en C. Hamurabi Gamboa Rosales
Director General
Consejo Zacatecano de Ciencia, Tecnología e Innovación

MTI. Luis Fernando Araiz Morales
Director de Innovación y Desarrollo Regional
Consejo Zacatecano de Ciencia, Tecnología e Innovación

Figura A.17: Participación en foro y mesas de trabajo

Figura A.18: Participación en la Jornada Estatal de Ciencia y Tecnología

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Subsecretaría de Educación Media Superior
Dirección General de Educación Tecnológica Industrial y de Servicios
Oficina Estatal de la DGETI en el Estado de Zacatecas
Área académica

XXVI CONCURSO NACIONAL DE PROTOTIPOS Y PROYECTOS DE EMPRENDIMIENTO

LA OFICINA ESTATAL DE LA
DIRECCIÓN GENERAL DE EDUCACIÓN TECNOLÓGICA
INDUSTRIAL Y DE SERVICIOS EN ZACATECAS

OTORGA LA PRESENTE

CONSTANCIA

A: JORGE ALEJANDRO MORGAN BENITA

POR HABER PARTICIPADO COMO JURADO EN EL XXVI
CONCURSO NACIONAL DE PROTOTIPOS Y PROYECTOS
DE EMPRENDIMIENTO 2024, ETAPA ESTATAL, EN LA
MODALIDAD DE DESARROLLO DE SOFTWARE

VICENTE SÁNCHEZ MÁRQUEZ
COMISIONADO RESPONSABLE DE LA DGETI
EN EL ESTADO DE ZACATECAS

S.E.P. S.E.M.S.
DGETI
OFICINA ESTATAL
DE LA DGETI
EN ZACATECAS
C.C.T. 32FTI0001R

Jerez, Zacatecas, a 29 Febrero 2024

correo electrónico:oficinaestatal.zacatecas@dgeti.sems.gob.mx

2024 Felipe Carrillo PUERTO

Figura A.19: Jurado del Concurso nacional de prototipado

Figura A.20: Speaker en Talent Land

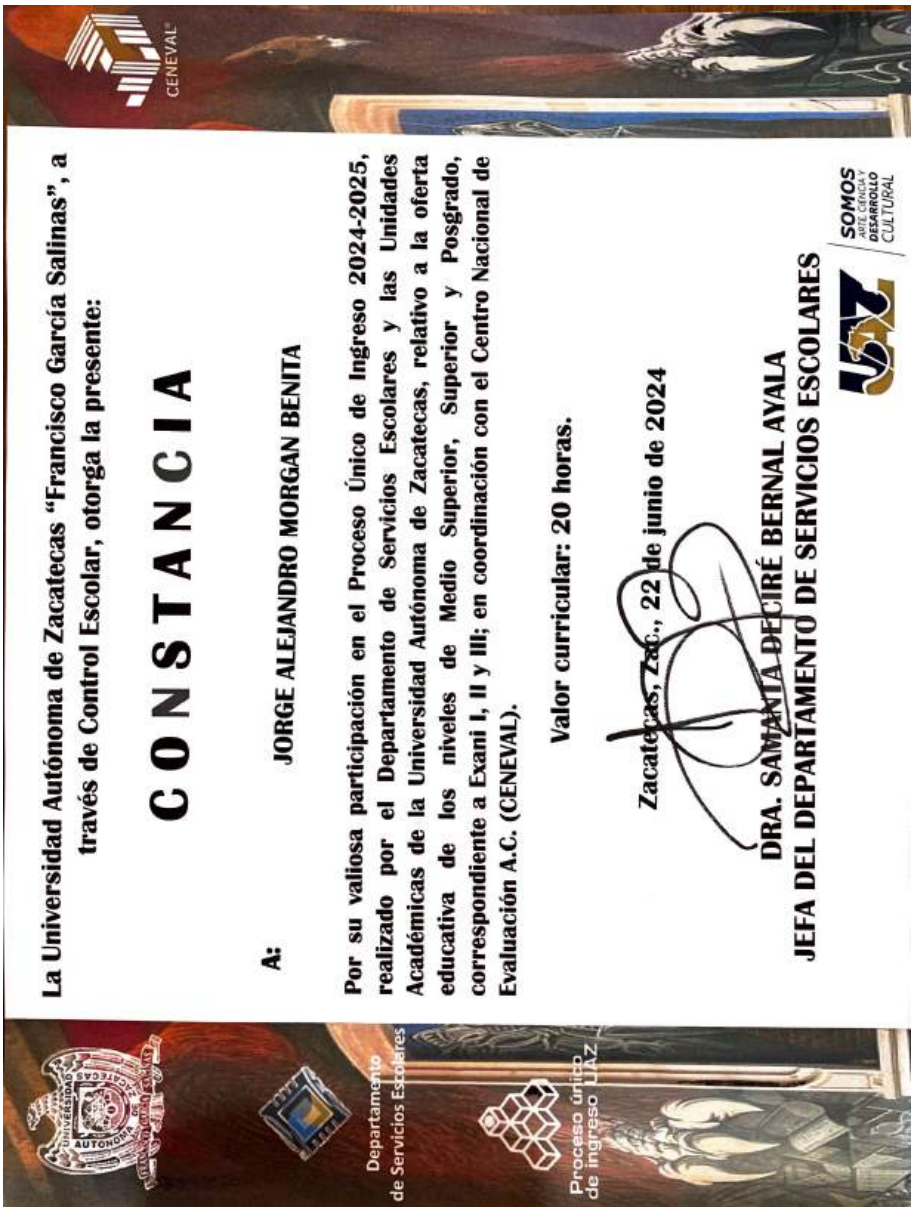


Figura A.21: Participación en la impartición de examen de ingreso CENEVAL