DETECCIÓN TEMPRANA DEL DIAGNOSTICO DE DIABETES TIPO II A PARTIR DE VARIABLES NO CLÍNICAS UTILIZANDO TÉCNICAS DE MACHINE LEARNING

PROYECTO DE GRADO

LEIDY TATIANA OME NARVÁEZ

DANNY GUILLERMO ORDÓÑEZ QUINTERO

Asesor
JAVIER DIAZ CELY, PhD

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2023

DETECCIÓN TEMPRANA DEL DIAGNOSTICO DE DIABETES TIPO II A PARTIR DE VARIABLES NO CLÍNICAS UTILIZANDO TÉCNICAS DE MACHINE LEARNING

LEIDY TATIANA OME NARVÁEZ

DANNY GUILLERMO ORDÓÑEZ QUINTERO

Trabajo de grado para optar al título de
Magister en Ciencia de Datos

Asesor
JAVIER DIAZ CELY, PhD

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2023

CONTENIDO

pág.

RESUMEN **8**

1. INTRODUCCIÓN **8**

1.1 Contexto y Antecedentes *8*

1.2 Planteamiento del Problema *10*

1.3 Objetivo General *12*

1.4 Objetivos Específicos *13*

2. ANTECEDENTES **13**

2.1 Estado del arte *13*

2.1.1 Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study 13

2.1.2 Diagnóstico de la diabetes mediante el uso de técnicas de aprendizaje automático 14

2.1.3 Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por Insuficiencia Cardiaca con datos clínicos 14

2.1.4 Predicción del diagnóstico de diabetes a partir de perfiles clínicos de pacientes utilizando aprendizaje automático 15

2.2 Marco Teórico *16*

2.2.1 Dominio del problema 17

2.2.1.1 Diabetes 17

2.2.1.2 Tipos de Diabetes 18

2.2.1.3 Prevención contra la diabetes tipo II 19

2.2.1.4 Impactos en la salud 19

2.2.2 Dominio de la solución 20

2.2.2.1 Aprendizaje Automático (Machine Learning) 20

2.2.2.2 Clasificación 21

2.2.2.3 XG-Boost 21

2.2.2.4 SVM 23

2.2.2.5 Redes Neuronales (ANN) 24

2.2.2.6 Métricas de evaluación: Matriz de confusión y métricas de clasificación 27

3. METODOLOGÍA **30**

4. PRESENTACIÓN DE LA PROPUESTA **34**

4.1 Recolección de los datos *34*

4.2 Limpieza y preparación de los datos *34*

4.3 Conjunto de datos final *35*

4.4 Definición de variables *36*

4.5 Análisis Exploratorio *37*

4.5.1 Asociación de la Diabetes con las variables predictoras cualitativas 44

4.6 Modelamiento *46*

4.6.1 Pre-procesamiento 46

4.7 Hiperparámetros empleados *48*

4.8 Evaluación del modelo *51*

5. DISEÑO DE EXPERIMENTO DE VALIDACIÓN **52**

6. RESULTADOS OBTENIDOS **53**

6.1 Benchmarking de modelos *53*

6.1 Definición del punto de corte para las probabilidades obtenidas *57*

6.2 Resultados del Modelo *59*

7. CONCLUSIONES Y FUTURO TRABAJO **61**

8. BIBLIOGRAFÍA **64**

LISTA DE TABLAS

Tabla 1 Resumen de los criterios de comparación (trabajos relacionados) **16**
Tabla 2 Esquema de Matriz de confusión **28**
Tabla 3 Estadísticas Descriptivas **38**
Tabla 4 Conteo de datos faltantes **38**
Tabla 5 Prevalencia de diabetes tipo II por regional **43**
Tabla 6 Prueba Chi-Cuadrado **45**
Tabla 7 Combinación de hiperparámetros utilizados en el algoritmo XGBoost **49**
Tabla 8 Combinación de hiperparámetros utilizados en el algoritmo MLP Clasiffier **50**
Tabla 9 Combinación de hiperparámetros utilizados en el algoritmo SVM Clasiffier **51**
Tabla 10 Hiperpárametros modelos XGBoost **54**
Tabla 11 Hiperpárametros modelos MLP-Classifier **56**
Tabla 12 Métricas de desempeño de los mejores modelos **59**

LISTA DE FIGURAS

Figura 1 Cantidad calculada de adultos entre 20 y 79 años con diabetes, Fuente: Atlas de la Diabetes de la FID - Novena edición 2019 ..... **10**
Figura 2 Prevalencia de Diabetes Mellitus en Colombia -Fuente: Cuenta de Alto Costo **11**
Figura 3 Incidencia de Diabetes Mellitus en Colombia, Fuente: Cuenta de Alto Costo **12**
Figura 4 Casos de separación de clases (Elaboración propia) **23**
Figura 5 Estructura general de una red neuronal (imagen tomada de internet) **27**
Figura 6 Singh, N. (2020). Ejemplo de curva ROC. **30**
Figura 7 Comprensión del negocio **31**
Figura 8 Comprensión de los datos **31**
Figura 9 Preparación de los datos **32**
Figura 10 Modelado **32**
Figura 11 Evaluación **33**
Figura 12 Proceso de estructuración del set de datos **36**
Figura 13 Distribución de Diabetes Tipo II **39**
Figura 14 Distribución de Diabetes Tipo II de acuerdo con la Edad, IMC y Superficie Corporal **39**
Figura 15 Distribución de diabetes por sexo **40**
Figura 16 Distribución de diabetes por nivel educativo **41**
Figura 17 Distribución de diabetes por rango salarial **42**
Figura 18 Distribución de antecedentes familiares de diabetes **44**
Figura 19 Partición de datos en entrenamiento y prueba. **48**
Figura 20 Protocolo de evaluación del modelo **52**
Figura 21 Iteraciones del modelo XGBoost **54**
Figura 22 Iteraciones del modelo MLP-Classifier **55**
Figura 23 Iteraciones del modelo SVM **57**
Figura 24 Puntos de corte para los mejores modelos seleccionados **58**
Figura 25 Importancia de variables del modelo seleccionado **60**

RESUMEN

La diabetes tipo II es una enfermedad crónica y grave que se caracteriza por niveles elevados de glucosa en la sangre debido a la incapacidad del cuerpo para producir o utilizar eficazmente la insulina. Si no se controla a largo plazo, esta deficiencia de insulina puede causar daño en varios órganos del cuerpo, lo que conduce a complicaciones y potencialmente mortales como enfermedades cardiovasculares, neuropatía, nefropatía y problemas oculares. Por lo tanto, el diagnóstico temprano y el tratamiento de la diabetes tipo II son de vital importancia para prevenir el desarrollo y las complicaciones de enfermedades cardiovasculares y renales. Por tal motivo en este trabajo de grado se formuló la elaboración de un modelo predictivo para el diagnóstico temprano de esta enfermedad.

Para el desarrollo de este modelo se tuvo en cuenta información no clínica de los pacientes tales como: datos demográficos, sociodemográficos, actividad física y antecedentes familiares. Se utilizaron datos de 204.572 usuarios afiliados a una IPS con presencia a nivel nacional, donde el 20,4% están diagnosticados con diabetes tipo II. Se entrenaron modelos supervisados de clasificación como: Maquina de Vectores de Soporte (SVM), Refuerzo Extremo de Gradiente (XGBoost) y Perceptrón Multicapa (MLP), donde se encontró como mejor modelo para la predicción de diabetes tipo II el XGBoost con una métrica de desempeño de ROC-AUC del 77%.

1. INTRODUCCIÓN

1.1 Contexto y Antecedentes

La diabetes es una condición crónica en la que los niveles elevados de glucosa en la sangre de una persona son causados por la incapacidad de su cuerpo para producir suficiente insulina o utilizarla de manera efectiva. La insulina es una hormona vital que se sintetiza en el páncreas y posibilita el ingreso de la glucosa sanguínea a las células del organismo para su transformación en energía. La carencia de insulina o la ineficacia de las células para responder a ella resulta en niveles elevados de glucosa en la sangre, lo cual se conoce como hiperglucemia. Estos hechos han sido respaldados por pruebas clínicas y son reconocidos por la Federación Internacional de Diabetes en 2019.

> De acuerdo con la Federación Internacional de Diabetes, en adelante se nombrará como (FID), si no se controla el déficit de insulina a largo plazo, muchos de los órganos del cuerpo pueden resultar dañados, lo que derivará en complicaciones de la salud incapacitantes y potencialmente mortales, como las enfermedades cardiovasculares (ECV), lesión de los nervios (neuropatía), enfermedad renal (nefropatía) y afección ocular (causante de la retinopatía, la pérdida de visión e incluso la ceguera). Sin embargo, si se logra un tratamiento apropiado de la diabetes, estas graves complicaciones se pueden retrasar o prevenir totalmente.

Los tipos de diabetes más prevalentes son la diabetes tipo I y la diabetes tipo II; la diabetes tipo I se caracteriza por una reacción autoinmune en la cual el sistema inmunitario del cuerpo ataca las células beta del páncreas responsables de la producción de insulina. Como resultado, el cuerpo presenta una producción insuficiente de insulina o ninguna producción en absoluto. Este estado puede manifestarse en cualquier etapa de la vida, aunque es más común en niños y

jóvenes. Por otro lado, la diabetes tipo II se origina debido a la incapacidad de las células corporales para responder de manera adecuada a la acción de la insulina, fenómeno conocido como "resistencia a la insulina". Durante esta fase, la hormona no logra ejercer su efecto de forma eficiente, lo que conlleva a un aumento en la producción de insulina. Con el tiempo, la producción de insulina puede volverse insuficiente debido a que las células del páncreas no pueden satisfacer la demanda requerida. Anteriormente, la diabetes tipo II suele ser más frecuente en personas mayores, pero en la actualidad se observan casos con mayor frecuencia en niños y adultos jóvenes debido al incremento de la obesidad, la falta de actividad física y una alimentación inapropiada (FID, 2019).

El diagnóstico temprano y el tratamiento de la diabetes tipo II se encuentran entre las acciones más relevantes para prevenir un mayor desarrollo y complicaciones de enfermedades cardiovasculares y renales, por tal motivo el objetivo de este trabajo consiste en la elaboración de un modelo predictivo para el diagnóstico temprano de esta patología. Para el desarrollo de este modelo se tendrá en cuenta información de los pacientes tales como: datos demográficos, hábitos alimenticios, actividad física y antecedentes familiares.

La importancia de este trabajo radica en la contribución de predecir el diagnóstico temprano de diabetes, lo cual es de gran relevancia, porque permitirá al paciente disminuir el riesgo de desarrollar otro tipo de enfermedades, como lo son las cardiovasculares, renales, enfermedades del sistema nervioso, oculares entre otras. Lo cual, influye directamente en la disminución de la tasa de mortalidad causada por esta patología.

Desde el punto de vista económico, beneficia tanto al paciente como al sistema de salud, dado que si la enfermedad de diabetes no se trata a tiempo puede generar complicaciones en el paciente como: cirugías, amputaciones, hospitalizaciones o incurrir en comorbilidades de mayor complejidad ocasionando un mayor costo para

el Sistema de Salud. Por consiguiente, construir modelos utilizando herramientas de Machine Learning que permita predecir en una fase temprana si una persona va a padecer o no diabetes, es de gran ayuda para el paciente y para el sistema de salud, dadas las descripciones anteriormente hechas.

1.2 Planteamiento del Problema

De acuerdo con la información proporcionada por la Federación Internacional de Diabetes (FID, 2021), en el año 2021, aproximadamente 537 millones de adultos entre 20 y 79 años padecían diabetes, y el 79,4% de ellos residían en países de ingresos bajos y medios. Si esta tendencia se mantiene, se estima que para el año 2045 habrá un aumento significativo, llegando a un total de 783 millones de adultos afectados por la enfermedad (FID, 2021).

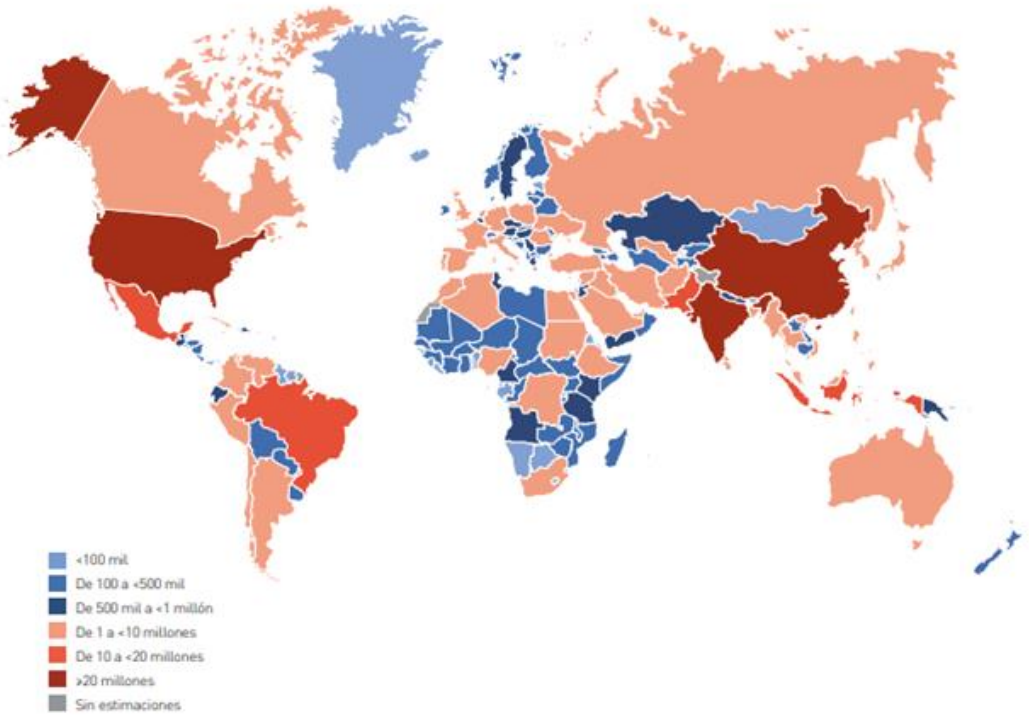

Figura 1 Cantidad calculada de adultos entre 20 y 79 años con diabetes, Fuente: Atlas de la Diabetes de la FID - Novena edición 2019

Según los datos registrados en la cuenta de alto costo CAC en Colombia, se identificó que hasta junio de 2020 había un total de 1.426.574 personas diagnosticadas con diabetes mellitus (DM). Esto representa aproximadamente el 3% de la población colombiana que sufre de esta enfermedad. La diabetes es reconocida como uno de los principales factores que contribuyen al desarrollo de la enfermedad renal crónica (ERC), lo que subraya la importancia de un control especializado para prevenir complicaciones en la salud de los afectados (Cuenta de Alto Costo [CAC], 2021).

En la región Central es donde se presenta el mayor número de personas con diagnóstico de diabetes, seguido se encuentra Bogotá y la región Pacífica, por último, la región Amazónica es donde se presenta el menor número de personas con diagnóstico de diabetes (Cuenta de Alto Costo [CAC], 2021).

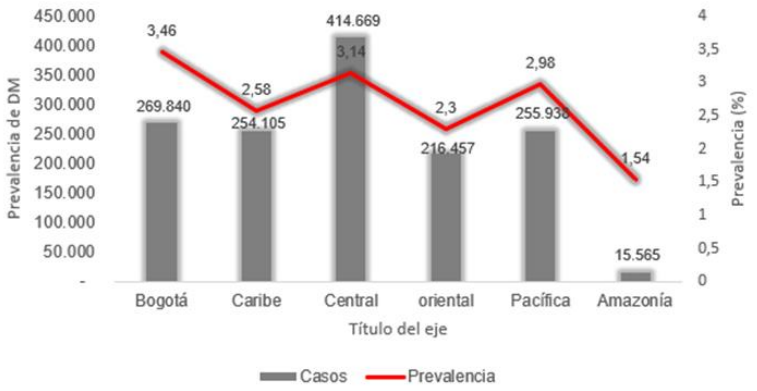

Figura 2 Prevalencia de Diabetes Mellitus en Colombia -Fuente: Cuenta de Alto Costo

En los últimos años se ha observado un aumento en la incidencia de la diabetes mellitus, excepto en el año 2020, en el cual, “los casos nuevos reportados disminuyeron en 9,54% con respecto al año anterior”. Esta disminución se debe a que el año 2020 se declaró la emergencia sanitaria por COVID-19 y los servicios

básicos en las clínicas y centros de atención se suspendieron para dar prioridad a la emergencia. Comparando la incidencia por sexo se evidencia que las mujeres presentan mayor frecuencia de diabetes tipo II (Cuenta de Alto Costo [CAC], 2021).

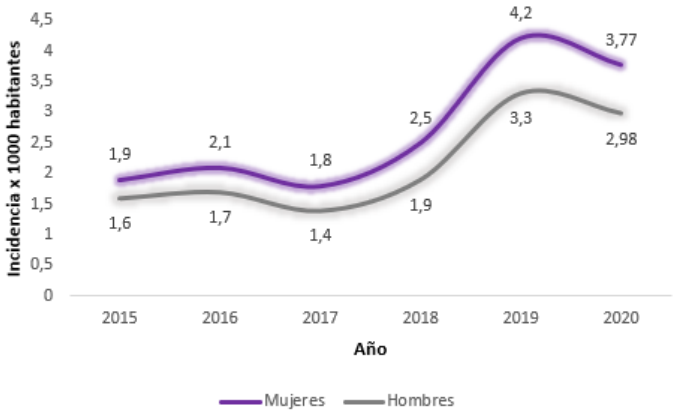

Figura 3 Incidencia de Diabetes Mellitus en Colombia, Fuente: Cuenta de Alto Costo

Respecto a la mortalidad, durante los años 2020 y 2021 la diabetes está ubicada entre las enfermedades con mayor frecuencia de muertes reportadas, donde una de las causas para que esta enfermedad presenta alta tasa de mortalidad es que el diagnóstico es tardío y se detecta en etapas avanzadas, debido a que muchos de los pacientes son asintomáticos (Ministerio de Salud y Protección Social [MinSalud], 2022).

1.3 Objetivo General

Desarrollar diferentes modelos de Machine Learning para predecir el riesgo de sufrir diabetes tipo II, utilizando datos no clínicos de pacientes recopilados de una red de Instituciones Prestadoras de Salud con presencia en varios departamentos de Colombia.

1.4 Objetivos Específicos

- Caracterizar la población de usuarios diagnosticados con diabetes tipo II.
- Determinar el conjunto de variables relevantes para la predicción del diagnóstico de diabetes tipo II.
- Aplicar diferentes metodologías de machine learning para la predicción de enfermedades, evaluando el modelo que presente la mejor capacidad predictiva teniendo en cuenta las métricas de desempeño.

2. ANTECEDENTES

2.1 Estado del arte

En la siguiente sección se exponen las ideas y/o trabajos relacionados con este proyecto; cabe mencionar que las metodologías de ML han abordado aplicaciones de todos los sectores, donde la parte médica no es la excepción, se realizaron búsquedas sobre la predicción de diagnósticos haciendo uso de metodologías de ML bajo criterios: cronológicos, geográficos, Diagnóstico clínico y herramientas utilizadas. El orden en que se muestran corresponde a criterios de similitud contra este estudio.

2.1.1 Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study

En este artículo utilizaron información de una cohorte de 36.652 pacientes de la zona rural de Henan en China tomados entre Julio 2015 y septiembre de 2017 para probar la capacidad de los algoritmos de aprendizaje automático para predecir el riesgo de diabetes mellitus tipo II, donde utilizaron seis algoritmos de aprendizaje automático, que incluyen regresión logística (RL), árbol de clasificación y regresión (CART), redes neuronales artificiales (ANN), máquina de vectores de soporte

(SVM), bosque aleatorio (RF) y aumento de gradiente máquina (GBM). “Para el rendimiento de los modelos se midieron: área bajo la curva característica operativa del receptor, sensibilidad, especificidad, valor predictivo positivo, valor predictivo negativo y área bajo la curva de recuperación de precisión”. Todos los modelos para predecir el riesgo de diabetes tipo II demostraron un fuerte rendimiento predictivo, con AUC que oscilaban entre 0,811 y 0,872 con datos de laboratorio y entre 0,767 y 0,817 sin datos de laboratorio. Entre ellos, el modelo GBM funcionó mejor (AUC: 0,872 con datos de laboratorio y 0,817 sin datos de laboratorio) (Zhang, l., et. al,. 2020).

2.1.2 Diagnóstico de la diabetes mediante el uso de técnicas de aprendizaje automático

En este proyecto se desarrolló una serie de modelos basados en aprendizaje automático con el objetivo de poder detectar si una persona padece o no diabetes (cualquier tipo de diabetes), donde se utilizaron los modelos: GaussianNB (GNB), GradientBoostingClassifier (GB), RandomForestClassifier (RF), KNeighborsClassifier (KNN) y SVC, Adicional se implementó una red neuronal (ANN). Se hizo uso de una base de datos correspondientes al pueblo Pima, el cual es un grupo indígena localizado en Arizona (Estados Unidos) y en Sonora (México), la cual contiene información de variables físicas y clínicas de los pacientes. Finalmente se obtuvieron resultados similares en los modelos planteados a excepción de RF, donde de acuerdo con el porcentaje de Accuracy (80.7%), el mejor modelo se obtuvo del GradientBoostingClassifier (GB) utilizando un Train/Test de 75-25. (Martínez Leal, A. 2021).

2.1.3 Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por Insuficiencia Cardiaca con datos clínicos

En este proyecto desarrollado en el año 2020 por estudiantes de la Universidad Antonio Nariño de Colombia, se evaluaron modelos de clasificación como: Máquina

de vectores se soporte (SVM), Random Forest (RF) y Redes Neuronales (ANN) para realizar “la predicción del riesgo de fallecimiento en pacientes con insuficiencia cardiaca, evaluando datos clínicos, por ejemplo: la capacidad de bombeo del corazón, sodio en la sangre, plaquetas, entre otros”. Para el alcance, se utilizaron 299 datos recolectados de un hospital de Pakistán que cuenta con 13 variables, de las cuales seis eran variables clínicas de seguimiento y el resto se dividen en variables demográficas y de estilos de vida. En cuanto a las mediciones realizadas del rendimiento de los modelos se detalla un promedio para el clasificador MLP de 87.38%, 86.51% para SVM y 83.87% para RF (Gallego Valcárcel, D. A., & Lucas Monsalve, D. F. 2021).

2.1.4 Predicción del diagnóstico de diabetes a partir de perfiles clínicos de pacientes utilizando aprendizaje automático

Este proyecto fue realizado por estudiantes de la Universidad Antonio Nariño de Bogotá. con el fin de tener una herramienta que permita realizar la predicción del riesgo de desarrollar diabetes (Tipo I, II y gestacional) utilizando herramientas de machine learning como: Redes neuronales artificiales, random forest y máquinas de vector de soporte (SVM). Para llevar a cabo la investigación utilizaron datos de carácter público de pacientes del Hospital de Sylhet, Bangladesh, la información utilizada contiene información demográfica de los pacientes y de síntomas relacionados con la diabetes.

Finalmente, este estudio utiliza métricas de rendimiento como precisión, recall, F-measure y ROC-AUC para elegir el modelo con el mejor resultado. Donde se concluye que el modelo de random forest respecto a las métricas de precisión y ROC-AUC es el que presenta el mejor resultado (Pérez Leal, L. E., & Buitrago Cardenas, J. A. 2021).

En la siguiente tabla, se muestra el resumen de los criterios de comparación entre los artículos seleccionados y el proyecto de grado. Las celdas en color verde denotan las coincidencias con lo planteado en el trabajo de grado.

Tabla 1 Resumen de los criterios de comparación (trabajos relacionados)

Referencia	Criterios			
	Diagnóstico clínico	**Geografía**	**Modelos ML utilizados**	**Dimensión temporal**
7.1	Diabetes Tipo II	Los datos corresponden a una cohorte de 36.652 pacientes de la población rural de Henan/China	Modelos ML: Regresión Logística (LR), Arboles de Clasificación (CART), Redes Neuronales (ANN), Máquina de Vectores de soporte (SVM), Random Forest (RF) y GBM. Cabe mencionar que utilizaron segmentos de variables con y sin datos de laboratorio.	Se reclutaron pacientes entre Julio 2015 y Septiembre 2017. Desarrollo 2020
7.2	Diabetes en General	Base de datos, correspondientes al pueblo Pima, el cual es un grupo Indigena localizado en Arizona (Estados Unidos) y en Sonora (México). Desarrollo realizado por estudiantes de la Universidad Politécnica de Valencia, España.	Rede Neuronal Artificial (ANN)	Contiene información de 21 años, pero no se especifica los periodos. Desarrollo 2021
7.3	Insuficiencia cardíaca	Utilizaron datos recolectados de un Hospital de Pakistan, sin embargo, la implementación de los modelos la realizaron estudiantes de Colombia (UAN).	Modelos ML: Máquinas de vectores de soporte (SVM), redes neuronales y random forest.	Pacientes con insuficiencia cardíaca que ingresaron en el Instituto de Cardiología y el hospital aliado Faisalabad-Pakistán durante abril-diciembre (2015)
7.4	Diabetes tipo I, tipo II y gestacional.	Conjunto de datos recopilados por el Hospital de Sylhet, Bangladesh. Sin embargo, la implementación de los modelos la realizaron estudiantes de Colombia (UAN).	Modelos ML: Máquinas de vectores de soporte (SVM), redes neuronales y random forest.	

2.2 Marco Teórico

Esta sección se enfoca en la descripción de los conceptos que se utilizan en el trabajo, los cuales se dividen en los conceptos que se abordan desde el problema que en este caso es médico y los conceptos de la solución que están aplicados a las metodologías de ciencia de datos, específicamente Machine Learning.

2.2.1 Dominio del problema

En la siguiente sección se describen los conceptos médicos que se abordan en el contexto del proyecto.

2.2.1.1 Diabetes

El Instituto Nacional de Salud (NIH por sus siglas en inglés) es la agencia principal del gobierno de los Estados Unidos responsable de la biomedicina e investigación en la salud pública. La cual tiene un segmento para temas relacionados sobre la diabetes, enfermedades digestivas y renales. De esta forma, los siguientes conceptos se exponen de la información brindada por esta fuente (Instituto Nacional de Salud [NIH], 2016).

> La diabetes es una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en sangre (o azúcar en sangre). La glucosa en la sangre es la principal fuente de energía y proviene de los alimentos. La insulina, una hormona que produce el páncreas, ayuda a que la glucosa de los alimentos ingrese en las células para usarse como energía. Algunas veces, el cuerpo no produce suficiente o no produce nada de insulina o no la usa adecuadamente y la glucosa se queda en la sangre y no llega a las células (Instituto Nacional de Salud [NIH], 2016).

Con el transcurso del tiempo, la presencia elevada de glucosa en la sangre puede generar complicaciones en la salud. A pesar de que la diabetes carece de cura, las personas con esta condición pueden adoptar medidas para controlar su enfermedad y preservar su bienestar.

2.2.1.2 Tipos de Diabetes

La diabetes tipo I, tipo II y diabetes gestacional son las que se escuchan más frecuentemente y presentan la siguiente descripción:

- **Diabetes tipo I:** Se produce una ausencia de insulina en el cuerpo debido a la respuesta del sistema inmunitario que ataca y destruye las células productoras de insulina en el páncreas. Esta forma de diabetes suele diagnosticarse en niños y adultos jóvenes, aunque puede surgir/presentarse en cualquier etapa de la vida. Las personas con diabetes tipo I requieren el uso diario de insulina para asegurar su supervivencia (Instituto Nacional de Salud [NIH], 2016).

- **Diabetes tipo II:** En este tipo de diabetes, existe una deficiencia en la producción o el uso adecuado de la insulina por parte del cuerpo. Este tipo de diabetes puede desarrollarse en cualquier etapa de la vida, incluso durante la infancia, aunque es más común en personas de mediana edad y ancianos. Se considera el tipo más prevalente de diabetes (Instituto Nacional de Salud [NIH], 2016).

- **Diabetes gestacional:** Durante el embarazo, algunas mujeres pueden experimentar diabetes gestacional, la cual generalmente desaparece después del parto. No obstante, las mujeres que han tenido diabetes gestacional tienen un mayor riesgo de desarrollar diabetes tipo II en etapas posteriores de sus vidas. “En ocasiones, la diabetes diagnosticada durante el embarazo puede ser en realidad diabetes tipo II” (Instituto Nacional de Salud [NIH], 2016).

- **Otros tipos de diabetes: “**Existen otros tipos menos frecuentes de esta enfermedad, como la diabetes monogénica, que es una forma de diabetes hereditaria, y la diabetes asociada a la fibrosis quística” (Instituto Nacional de Salud [NIH], 2016).

2.2.1.3 Prevención contra la diabetes tipo II

Bajo el concepto de la Organización Panamericana de Salud (OPS), “se ha demostrado que las medidas sencillas de estilo de vida son eficaces para prevenir o retrasar la aparición de la diabetes tipo II. Para ayudar a prevenir la diabetes tipo II y sus complicaciones, las personas deben: (OPS, 2019).

- Lograr y mantener un peso corporal saludable.
- Ser físicamente activo: realizar al menos 30 minutos de actividad regular de intensidad moderada la mayoría de los días. Se requiere actividad para controlar el peso.
- Seguir una dieta saludable, evitando el azúcar y grasas saturadas.
- Evitar el consumo de tabaco, ya que este es un inductor para el riesgo de diabetes y enfermedades cardiovasculares”.

2.2.1.4 Impactos en la salud

Según la Organización Panamericana de la Salud (OPS, 2019), la diabetes tipo II puede incurrir en los siguiente impactos en el cuerpo humano:

- “Con el tiempo, la diabetes puede dañar el corazón, los vasos sanguíneos, los ojos, los riñones y los nervios.
- Los adultos con diabetes tienen un riesgo dos o tres veces mayor de sufrir ataques cardíacos y accidentes cerebrovasculares.
- Combinado con un flujo sanguíneo reducido, la neuropatía (daño a los nervios) en los pies aumenta la posibilidad de úlceras en el pie, infección y eventual necesidad de amputación de una extremidad.
- La retinopatía diabética es una causa importante de ceguera y se produce como resultado del daño acumulado a largo plazo en los pequeños vasos

sanguíneos de la retina. Cerca de 1 millón de personas son ciegas debido a la diabetes.

- La diabetes es una de las principales causas de insuficiencia renal".

2.2.2 Dominio de la solución

En la siguiente sección se describen los conceptos de ciencia de datos que se utilizan para dar solución a los objetivos planteados en el proyecto.

2.2.2.1 Aprendizaje Automático (Machine Learning)

El Aprendizaje Automático (ML, por sus siglas en inglés) es una disciplina del campo de la Inteligencia Artificial que se encuentra basada en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con mínima intervención humana (Mahesh, B. 2020). El ML se ha aplicado en diferentes áreas como: Salud, Finanzas, Procesos comerciales, redes sociales, ingeniería, reconocimiento de patrones, entre otros. Los algoritmos de Machine Learning se dividen en tres categorías, siendo las dos primeras las más comunes:

- **Aprendizaje supervisado**: estos algoritmos parten de una base de datos donde los datos se encuentran etiquetados y los cuales permiten tomar decisiones o realizar predicciones. entrenamiento. En esta categoría se encuentran varios algoritmos como: Árboles de Decisión, Regresión Logística, Máquinas de Vectores de Soporte - SVM, Redes Neuronales, XG-Boost, entre otros (Mahesh, B. 2020).

- **Aprendizaje no supervisado:** estos algoritmos no cuentan con un conocimiento previo, el fin en estos algoritmos es buscar patrones o encontrar relaciones (Mahesh, B. 2020).

- **Aprendizaje por refuerzo**: “su objetivo es que un algoritmo aprenda a partir de la propia experiencia. Esto es, que sea capaz de tomar la mejor decisión ante diferentes situaciones de acuerdo con un proceso de prueba y error en el que se recompensan las decisiones correctas” (Mahesh, B. 2020).

2.2.2.2 Clasificación

El algoritmo de aprendizaje supervisado se divide en dos tipos: clasificación y regresión. Respecto a clasificación, el objetivo del algoritmo es predecir la etiqueta a la cual corresponde un individuo, para realizar esta etiqueta se utiliza la información aprendida de los datos. Dependiendo del tipo de etiqueta, se puede decir que la clasificación es binaria o multiclase. Cuando solo existen dos clases se trata de clasificación binaria, y si existen más de dos clases se denomina clasificación multiclase (Roman, V., 2019).

En este trabajo se requiere utilizar un modelo de clasificación con el fin objetivo de predecir el riesgo de sufrir diabetes tipo II. Teniendo en cuenta la literatura consultada, a continuación, se describen los algoritmos supervisados que se van a utilizar.

2.2.2.3 XG-Boost

El algoritmo XGBoost es un método de aprendizaje supervisado el cual se utiliza tanto para clasificación como regresión. Este algoritmo propuesto por Chen y Guestrin (2016), es un algoritmo altamente efectivo y ampliamente utilizado en diversos campos dado que ha producido rendimientos excelentes debido a las ventajas de regularización y el procesamiento paralelo efectivo de los árboles.

Este algoritmo utiliza el residual como calibración del predictor anterior en cada iteración; esto corresponde al proceso de optimización de la función de pérdida. Adicional, con el fin de reducir el riesgo de sobreajuste en el proceso de calibración, XGBoost agrega regularización a la función objetivo, que se puede describir mediante la siguiente ecuación.

$$J(\theta) = L(\theta) + \Omega(\theta)$$

Donde:

θ: Corresponde al parámetro entrenado.
Ω: Denota regularización, que está destinada a evitar el sobreajuste ya que puede controlar la complejidad del modelo.
L: Indica las funciones de pérdida de entrenamiento (es decir, pérdida cuadrada / logística). la cual determina qué tan bien se ajusta el modelo a los datos de entrenamiento.

De acuerdo con la teoría del árbol de decisión, la salida del modelo depende del promedio de una colección F de árboles (Zhou et al., 2020).

$$\widehat{y_i} = \sum_{i=1}^{n} f_k(x_i)$$

Donde:
$f_k \epsilon F$
$\widehat{y_i}$ es la función de predicción.

El algoritmo XGBoost presenta diversas ventajas destacadas:

- Capacidad para manejar grandes conjuntos de datos con múltiples variables.

- Habilidad para tratar con valores perdidos.
- Producción de resultados altamente precisos.

2.2.2.4 SVM

Máquinas de Soporte Vectorial (SVM de la expresión en inglés Support Vector Machines) es un método de clasificación supervisada que se desarrolló dentro de la ciencia computacional en la década de 1990 y ha crecido en popularidad desde entonces. SVM se ha demostrado que funciona muy bien en varios entornos y a menudo se considera uno de los mejores clasificadores (James, et al., 2013).

El propósito de SVM es determinar la frontera óptima entre dos clases que se pueden escalar a un número mayor. En el caso más simple, linealmente separable, existe una distancia positiva entre dos clases. Por tanto, se puede determinar un hiperplano que maximice la distancia de cada clase al mismo. En el caso más complejo, que no es linealmente separable, las categorías se superponen espacialmente de modo que no pueden separarse por un hiperplano.

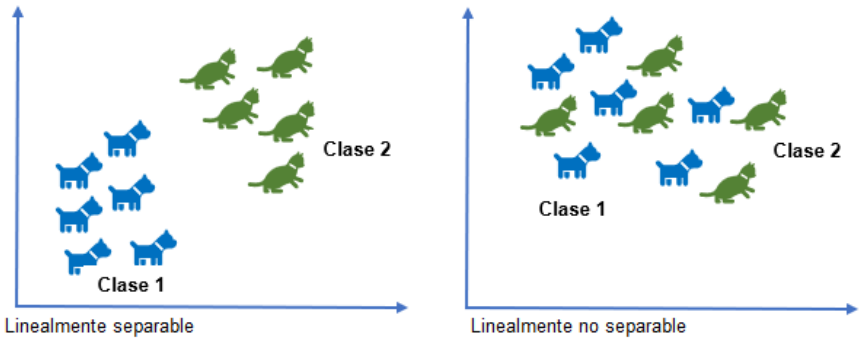

Figura 4 Casos de separación de clases (Elaboración propia)

SVM considera ampliar el espacio de características usando funciones de los predictores, como términos cuadráticos y cúbicos, para abordar esta no linealidad. En el caso del clasificador, se puede abordar el problema de límites posiblemente no lineales entre clases de una manera similar, ampliando el espacio de

características utilizando cuadráticas, cúbicas e incluso de orden superior funciones polinómicas de los predictores. La máquina de vectores de soporte (SVM) es una extensión de la máquina de vector clasificador que resulta de ampliar el espacio de características de una manera específica, utilizando kernels. La principal idea del kernel es ampliar el espacio de características con el fin de acomodar un límite no lineal entre las clases (James, et al., 2013). Los tipos de funciones del kernel son lineal, polinomial, RBF, el sigmoide, entre otros.

2.2.2.5 Redes Neuronales (ANN)

Las redes neuronales artificiales (ANN) son programas informáticos inspirados en la biología diseñados para simular la forma en que el cerebro humano procesa la información. Las ANN reúnen su conocimiento al detectar patrones y relaciones en los datos y aprenden (o son entrenadas) a través de la experiencia, no de la programación (Agatonovic-Kustrin, S., & Beresford, R. 2000).

> Una ANN se compone de múltiples unidades individuales, conocidas como neuronas artificiales o elementos de procesamiento (PE), los cuales están interconectados mediante coeficientes denominados pesos (wk). Estos pesos son responsables de la organización en capas y la estructura neuronal de la ANN. “El poder de los cálculos neuronales proviene de conectar las neuronas en una red. Cada PE tiene entradas ponderadas, función de transferencia y una salida. El comportamiento de una red neuronal está determinado por las funciones de transferencia de sus neuronas, por la regla de aprendizaje y por la propia arquitectura. Los pesos son los parámetros ajustables y, en ese sentido, una red neuronal es un sistema parametrizado. La suma ponderada de las entradas constituye la activación de la neurona Ω. La señal de activación pasa a través de la función de transferencia para

producir una salida única de la neurona y_ky agrega un término umbral, b_k." Adaptado de (Agatonovic-Kustrin, S., & Beresford, R. 2000).

$$y_k = \Omega * (x * w_k + b_k)$$

La función de transferencia desempeña un papel fundamental al introducir la no linealidad en la red. Durante el proceso de entrenamiento, las conexiones entre las unidades se ajustan y optimizan con el objetivo de minimizar el error en las predicciones y lograr el nivel de precisión especificado para la red. Una vez que la red ha sido entrenada y evaluada, es posible utilizarla con nuevos datos de entrada para realizar predicciones sobre la salida correspondiente. "Ya se han diseñado muchos tipos de redes neuronales y cada semana se inventan nuevas, pero todas pueden describirse por las funciones de transferencia de sus neuronas, por la regla de aprendizaje y por la fórmula de conexión. En términos de especificación del modelo, las redes neuronales artificiales no requieren conocimiento de la fuente de datos, pero dado que a menudo contienen muchos pesos que deben estimarse, requieren grandes conjuntos de entrenamiento. Además, las ANN pueden combinar e incorporar datos experimentales y basados en la literatura para resolver problemas. Las diversas aplicaciones de las ANN se pueden resumir en clasificación o reconocimiento de patrones, predicción y modelado" (Agatonovic-Kustrin, S., & Beresford, R. 2000).

"Muchas variaciones de redes neuronales existen en la literatura, pero todas tienen una estructura particular: una capa de entradas, una capa de salida, y varias capas intermedias (capas ocultas), a este tipo de redes neuronales se les llama Perceptrones Multicapa (MLP). Otro tipo de red neuronal bastante usada en la literatura son las redes neuronales de base radial (RBF), las cuales utilizan funciones de activación en los nodos ocultos radialmente simétricas. Se dice que una función es radialmente simétrica (o es una Función de Base Radial) si su salida depende de la distancia entre un vector

que almacena los datos de entrada y un vector de pesos sinápticos, que recibe el nombre de centro o centroide. Las redes RBF tienen una estructura de tres capas de conexión hacia adelante (es decir, que no presentan conexiones laterales ni hacia neuronas de capas anteriores sino solamente con neuronas de la siguiente capa): la capa de entrada, la capa oculta o intermedia y la capa de salida. Las neuronas de la capa de entrada tienen la función de enviar la información a la capa intermedia. Las neuronas de la capa oculta se activan en función de la distancia que separa cada patrón de entrada con respecto al centroide que cada neurona oculta almacena, a la que se le aplica una función radial con forma gaussiana" (Hemati et,al, 2019).

Las neuronas ubicadas en la capa de salida operan de manera lineal, realizando simplemente la suma ponderada de las salidas provenientes de la capa oculta. La red neuronal tiene como objetivo establecer una relación entre un conjunto de entradas y las salidas esperadas, para lograr esto emplea métodos de aprendizaje como Levenberg-Marquardt, retropropagación, entre otros. Estos métodos calculan el error a lo largo de la red y ajustan los pesos w_ky el umbral en cada iteración, con el propósito de acercar cada vez más la salida calculada por la red a la salida real. (Galindo, E. A., et. al.,2020).

En la Figura 5 se muestra la estructura general que presenta una red neuronal.

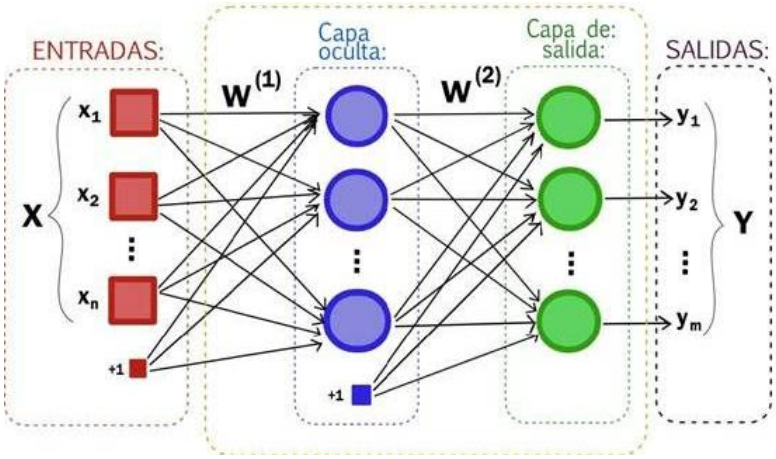

Figura 5 Estructura general de una red neuronal (imagen tomada de internet)

2.2.2.6 Métricas de evaluación: Matriz de confusión y métricas de clasificación

Uno de los pasos más importante es diagnosticar o medir el rendimiento del modelo de aprendizaje automático, ya que este paso es el que finalmente dice que tan bien generalizada el modelo sobre datos no vistos y poder determinar si el modelo de aprendizaje seleccionado es adaptable a otras observaciones para poder ponerlo en producción.

Al ser este un estudio relacionado con la salud de las personas se esperan errores mínimos en las predicciones, en este caso se exponen todas las métricas que se pueden evaluar desde la matriz de confusión.

2.2.2.6.1 Matriz de confusión

La matriz de confusión se utiliza como una tabla para evaluar el rendimiento de un modelo de clasificación supervisada en los datos de prueba, donde se dispone de los valores reales conocidos. Este término se emplea debido a que permite

identificar de manera sencilla las situaciones en las que el sistema confunde dos clases. (sitiobigdata,2019)

Tabla 2 Esquema de Matriz de confusión

		Predicción	
		Negativos	**Positivos**
Valores Reales	**Negativos**	Verdaderos Negativos (VN)	Falsos Positivos (FP)
	Positivos	Falsos Negativos (FN)	Verdaderos Positivos (VP)

2.2.2.6.2 Exactitud

La precisión se refiere al porcentaje de elementos que son correctamente clasificados. Esta métrica proporciona una evaluación directa de la calidad de los clasificadores, representada por un valor que oscila entre 0 y 1. (sitiobigdata,2019)

$$Exactitud = \frac{VN + VP}{Total}$$

2.2.2.6.3 Sensibilidad (Recall)

Se define como la proporción de elementos correctamente identificados como positivos con respecto al total de positivos reales. (sitiobigdata,2019)

$$Recall = \frac{VP}{VP + VN}$$

2.2.2.6.4 Precisión

“Es el número de elementos identificados correctamente como positivo de un total de elementos identificados como positivos”. (sitiobigdata,2019)

$$Precisión = \frac{VP}{VP + FP}$$

2.2.2.6.5 Puntuación F1

La puntuación F1 se calcula como la media armónica de la precisión y la exhaustividad. La puntuación F1 alcanza su máximo valor de 1 cuando tanto la precisión como la exhaustividad son perfectas, y su valor mínimo de 0 cuando ambas métricas son nulas.

$$F1 = 2 * \frac{Precisión * Recall}{Precisión + Recall}$$

La media armónica de una lista de números tiende a ser influenciada en mayor medida por los últimos elementos de la lista, en comparación con la media aritmética. Esto resulta en una reducción del impacto de los valores atípicos más grandes y un aumento del impacto de los valores más pequeños. (sitiobigdata,2019)

2.2.2.6.6 Curva de características operativas del receptor (ROC)

La evaluación de un modelo puede hacerse de manera efectiva midiendo el área bajo la curva ROC. Al trazar la sensibilidad (tasa positiva verdadera) en función de la especificidad (tasa de falsos positivos), obtenemos la curva de Característica Operativa del Receptor (ROC). Esta curva proporciona una representación visual

del equilibrio entre la detección de verdaderos positivos y la aparición de falsos positivos (Singh, N. 2020).

En la siguiente Figura se observa “ejemplos de buenas curvas ROC. La línea discontinua sería una suposición aleatoria (sin valor predictivo) y se utiliza como línea de base; cualquier cosa por debajo se considera peor que una suposición. El deber es estar hacia la esquina superior izquierda”:

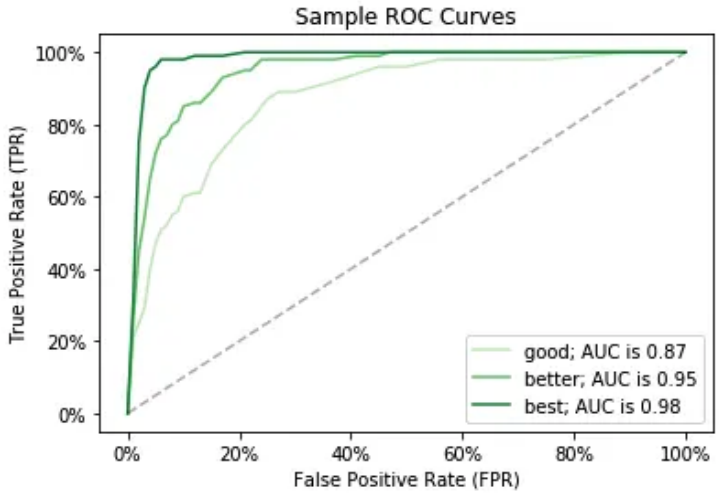

Figura 6 Singh, N. (2020). Ejemplo de curva ROC.

3. METODOLOGÍA

Para el desarrollo de esta investigación se adaptará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), la cual es una de las más utilizadas en proyectos de esta dimensión. Se abordarán diferentes fases que nos permitirá desarrollar los objetivos propuestos, iniciando con el entendimiento del problema de negocio y finalizando con el despliegue de un modelo de IA que permita predecir el riesgo de sufrir diabetes tipo II.

En el siguiente flujo se muestra cómo se va a realizar la ejecución del proyecto, el cual será dividido en tareas, las cuales serán agrupadas en una estructura de desglose de trabajo (EDT):

Fase 1 Comprensión del negocio: En la etapa inicial, se dedica atención a comprender los objetivos del proyecto. A partir de este conocimiento de los datos, se procede a establecer una definición del problema de minería de datos y se crea un plan preliminar que busca alcanzar los objetivos establecidos.

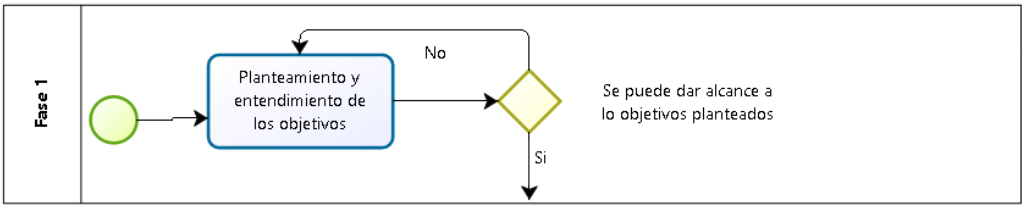

Figura 7 Comprensión del negocio

Fase 2 Comprensión de los datos: Esta se inicia con la recopilación inicial de datos y se desarrolla a través de actividades que buscan familiarizarse con los datos, identificar problemas de calidad, descubrir conocimiento preliminar sobre los datos y/o encontrar subconjuntos relevantes para formular hipótesis sobre información oculta.

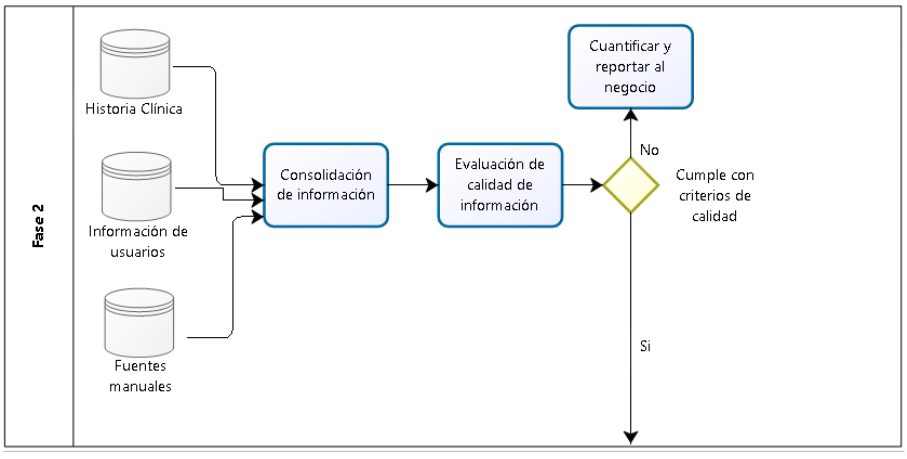

Figura 8 Comprensión de los datos

Fase 3 Preparación de los datos: La etapa de preparación de datos engloba todas las acciones requeridas para construir el conjunto definitivo de datos que será utilizado en las herramientas de modelado, partiendo de los datos brutos iniciales. Estas tareas abarcan la selección de tablas, registros y atributos pertinentes, así como la transformación y limpieza de los datos en preparación para su utilización en las herramientas de modelado.

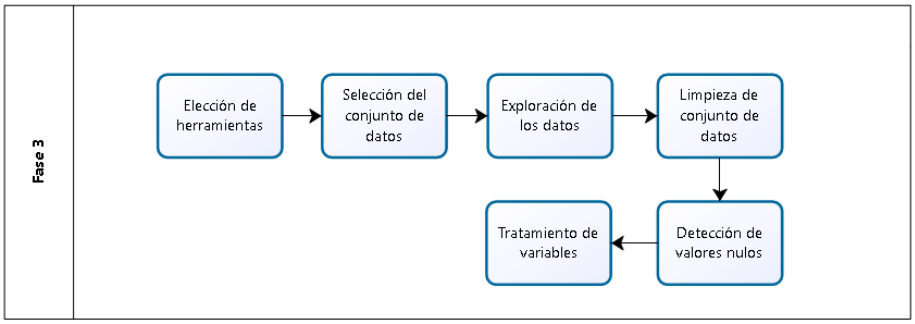

Figura 9 Preparación de los datos

Fase 4: Modelado: En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema, y se calibran sus parámetros a valores óptimos. Es probable que, dependiendo de los modelos, los datos deban tener una preparación específica, es por eso que existe una relación con la etapa previa de preparación de datos.

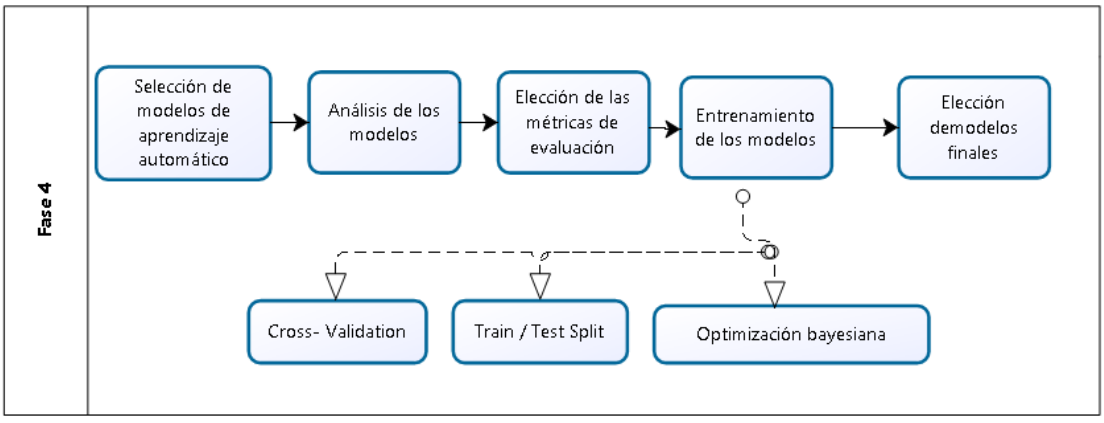

Figura 10 Modelado

Fase 5 Evaluación: En esta fase del proyecto, se han creado uno o varios modelos que aparentan exhibir una calidad apropiada en términos del análisis de datos. Antes de avanzar con la implementación final del modelo, es esencial llevar a cabo una evaluación minuciosa y revisar detalladamente los pasos seguidos para su desarrollo, comparando el modelo logrado con los objetivos empresariales establecidos.

Uno de los objetivos clave consiste en determinar si existen aspectos significativos relacionados con el negocio que no se hayan considerado de manera suficiente. Al concluir esta etapa, se aguarda obtener una conclusión acerca de la utilización de los resultados obtenidos del proceso de análisis de datos. En caso de no existir pruebas suficientes para definir el modelo final, será necesario regresar a la etapa inicial para realizar ajustes necesarios.

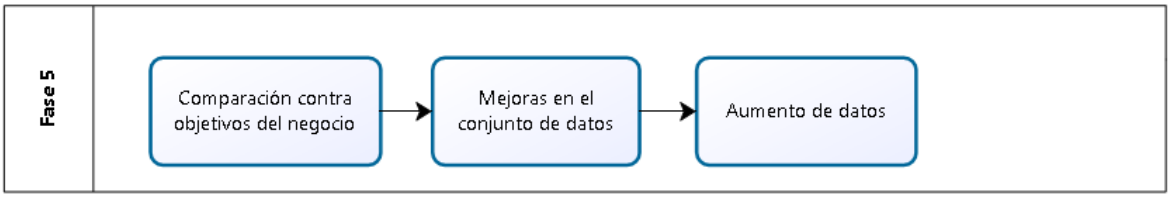

Figura 11 Evaluación

Fase 6 Despliegue: Por lo general, la creación del modelo no marca el final del proyecto. Incluso si el objetivo del modelo es mejorar la comprensión de los datos, el conocimiento adquirido deberá organizarse y presentarse de manera que el cliente pueda utilizarlo. El alcance de la fase de desarrollo dependerá de los requisitos específicos, pudiendo ser tan sencillo como la generación de un informe o tan complejo como la implementación periódica y posiblemente automatizada de un proceso de análisis de datos en la organización.

4. PRESENTACIÓN DE LA PROPUESTA

En este capítulo, se presenta la metodología utilizada para lograr el objetivo de entrenar un modelo de clasificación con el fin de predecir de manera temprana el riesgo de desarrollar diabetes tipo II. Para el ajuste del modelo, se utilizará como variable de respuesta la identificación de usuarios (pacientes) clasificados en riesgo de diabetes tipo II, donde se considerará "SI" si el afiliado ha sido diagnosticado con dicha enfermedad.

4.1 Recolección de los datos

Partiendo del objetivo principal y de evaluar el sistema de información que administra la Institución Prestadora de Servicios de Salud (IPS) se identificaron las fuentes de información que se requerían, de esta forma se consolidó la información de **Historia clínica** (HC) de los pacientes atendidos entre los años 2021 y 2022; la HC contiene información desde la identidad del paciente hasta antecedentes médicos, diagnósticos, tratamientos y procedimientos. De esta fuente se recolectó 3 GB de información que corresponde a todas las atenciones o transacciones que se realizaron los pacientes en la IPS en el periodo establecido. Dentro del sistema de información también se tiene la fuente de **Población,** la cual corresponde a los usuarios que tiene asignados la IPS para ser atendidos, dentro de este conjunto de información se cuenta con variables de identificación, demográficas y sociales. Por último, también se tiene **fuentes manuales**, las cuales corresponden a reportes normativos de que se envían al Ministerio de Salud.

4.2 Limpieza y preparación de los datos

En el proceso de depuración y preparación de datos se realizó un trabajo articulado con el área de Tecnología e Información (TI) de la IPS, de acuerdo con lo descrito en la metodología, en la fase comprensión de los datos (Ver Figura 8), se observa que posterior a la consolidación de las fuentes de datos se realiza un control de

calidad a la información de los campos de interés y se retroalimentaba al área de TI para realizar ajustes con búsquedas activas[1] y de esta manera mejorar la calidad del dato.

4.3 Conjunto de datos final

Partiendo de la base de datos de historia clínica (HC), se encontró un total de 1.367.166 registros[2] y 75 variables dentro del periodo enero 2021 a noviembre 2022. Dentro de este set de datos se encontró información no relevante para el desarrollo de este proyecto, como información clínica y administrativa, razón por la cual estas variables son eliminadas del set de datos. Por el contrario, se conservaron las variables de interés para el desarrollo del proyecto como: variables sociodemográficas, demográficas, estilo de vida, entre otras del paciente. Con esta información se lograron obtener un total de 330.744 pacientes únicos.

Posteriormente y con ayuda de un experto se seleccionan las variables **no clínicas** que contribuyen a la predicción temprana del diagnóstico de diabetes tipo II, conservando un total de 16 variables. Respecto a la cantidad de pacientes únicos, se conservan 204.572, dado que la recomendación del experto fue realizar el estudio sobre pacientes mayores a 18 años, además se excluyeron registros de pacientes que no tenían completitud en la mayoría de los campos de interés. El proceso descrito anteriormente se puede evidenciar en la Figura 12.

[1] La búsqueda activa es un proceso en el que se realiza una búsqueda sistemática y planificada de personas que pueden estar en riesgo de una enfermedad o condición de salud específica, con el fin de identificar a aquellos que puedan necesitar atención médica o medidas preventivas.
[2] Un registro corresponde a cada una de las atenciones que se le realizó a un paciente en la IPS durante el periodo establecido.

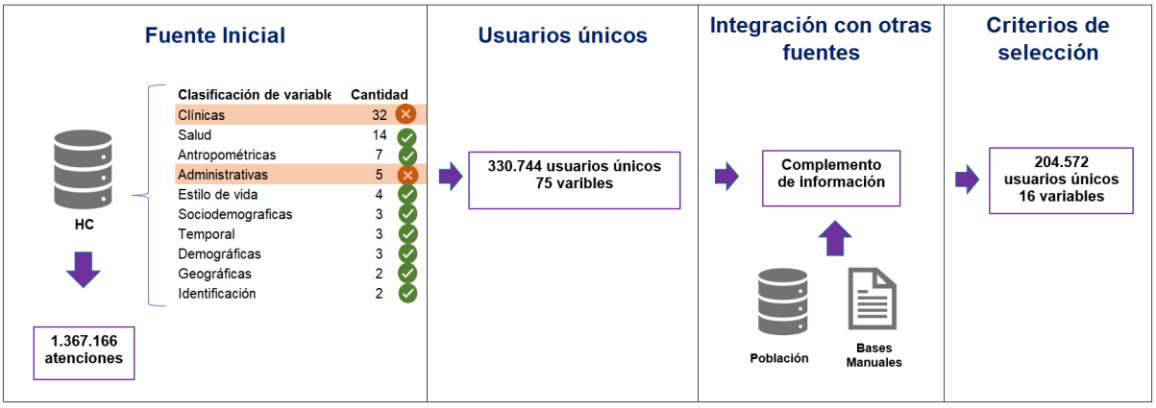

Figura 12 Proceso de estructuración del set de datos
Fuente: Elaboración propia

4.4 Definición de variables

En esta sección, se proporciona una descripción de las variables predictoras que se utilizan en el entrenamiento de los modelos. Estas variables abarcan aspectos sociodemográficos y geográficos con el objetivo de establecer asociaciones con la identificación del riesgo de desarrollar diabetes tipo II, permitiendo así caracterizar a los pacientes.

- Edad
- Sexo
- Nivel educativo
- Rango salarial
- Regional
- Zona

El siguiente conjunto de variables está asociado con los estilos de vida de los pacientes.

- Índice de Masa Corporal (IMC)
- Superficie Corporal
- Actividad física
- Consulta nutrición

Por último, se incluyen variables relacionadas con antecedentes familiares de los usuarios.

- Enfermedades Coronarias
- Dislipidemia
- Obesidad
- Hipotiroidismo
- Insuficiencia renal crónica (IRC)
- Enfermedad pulmonar obstructiva crónica (EPOC)

4.5 Análisis Exploratorio

En esta sección, se lleva a cabo el análisis exploratorio de los datos, realizando limpieza de los datos respecto a datos faltantes, inconsistencias en valores u outliers. Antes de empezar con los análisis se verifica los siguientes aspectos:

- Datos faltantes.
- Valores inconsistentes u outliers.

Respecto a datos faltantes, posterior de la estructuración del set de datos final y los procesos realizados para completar la información de los usuarios, no fue posible encontrar información de la variable superficie corporal para 5.616 usuarios, los cuales representan aproximadamente el 2.7% del total de registros.

En relación con los valores inconsistentes o atípicos, se han identificado algunos datos sospechosos en las variables IMC y superficie corporal. Estos valores se validan con las áreas medicas de la IPS, con las cuales se determinan reglas de negocio para clasificar un dato en inconsistente u outliers. Para definir estas reglas de negocio, se consideró el comportamiento de las variables, en el caso del IMC se observó un valor mínimo de 10, mientras que para la superficie corporal se evidenció

valor mínimo de 0, estos casos al validarlo con los expertos del negocio indicaron que son valores que no cumplen con las referencias y estándares médicos.

Tabla 3 Estadísticas Descriptivas

Descriptivas	IMC	Superficie corporal
Mínimo	10,00	0,00
Percentil 25	24,47	1,66
Media	28,12	1,81
Mediana	27,56	1,80
Percentil 75	31,20	1,95
Máximo	60,00	3,19

Fuente: Elaboración propia

Teniendo en cuenta lo anterior se definen las siguientes reglas de negocio:

- Valores de IMC inferiores a 16.
- Valores de superficie corporal inferiores a 1.2 y superiores a 2.4.

Los registros que cumplan con las condiciones anteriores, dentro del proceso se asignaran como valores faltantes para posterior realizar un método de imputación de datos. Una vez aplicada las reglas de negocio, se obtiene que del total de registros las variables IMC y superficie corporal presentan 0.22% y 4,12% de faltantes, respectivamente.

Tabla 4 Conteo de datos faltantes

Variable	Registros vacíos	%
IMC	445	0,22%
Superficie corporal	8.437	4,12%

Fuente: Elaboración propia

A continuación, se presenta el análisis de las variables en relación con el problema propuesto. En la Figura 13 se observa la distribución de pacientes identificados con

el riesgo de diabetes tipo II, donde se tiene que el 20.4% de los usuarios presentan la enfermedad de diabetes tipo II. En este caso, se considera que el conjunto de datos presenta un desbalanceo moderado, por lo tanto, en el presente trabajo no se van a aplicar técnicas de balanceo en los datos para la modelación.

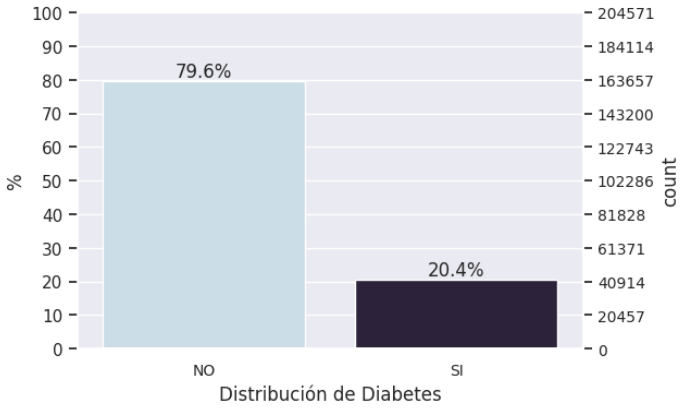

Figura 13 Distribución de Diabetes Tipo II
Fuente*:* Elaboración propia

Se observa una fuerte relación entre los casos de diabetes tipo II y la edad, lo que indica que a medida que aumenta la edad, también aumenta el riesgo de padecer la enfermedad. Mientras que, en el IMC y la superficie corporal se observa una leve relación con el riesgo de padecer la patología.

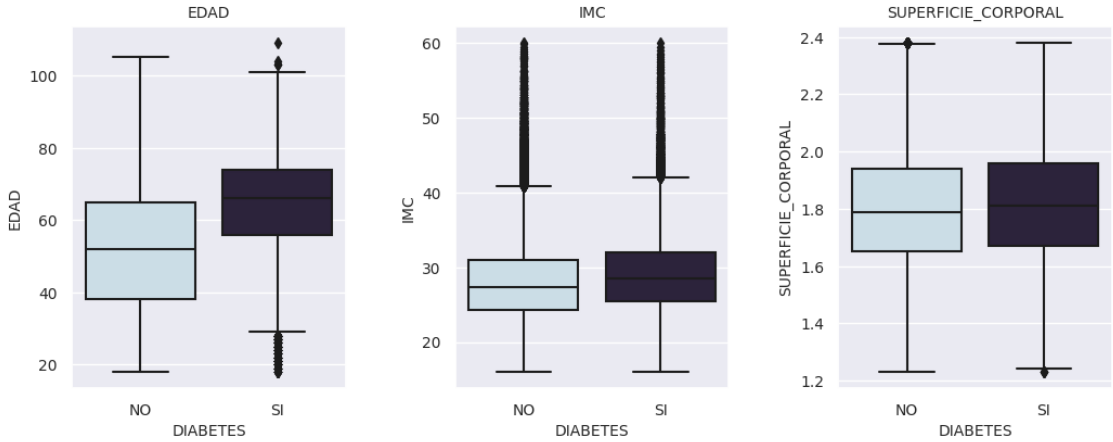

Figura 14 Distribución de Diabetes Tipo II de acuerdo con la Edad, IMC y Superficie Corporal
Fuente*:* Elaboración propia

En la caracterización de las variables sociodemográficas de los pacientes, se observa en la Figura 15 que los pacientes de sexo femenino presentan una menor tasa de los pacientes con la enfermedad de diabetes tipo II, mientras que de los pacientes de género masculino tienen una proporción más alta (23.3%). A pesar de este comportamiento se menciona que las mujeres presentan mayores factores de riesgo que las exponen a esta patología, como el síndrome de ovario poliquístico y la diabetes gestacional, los cuales son factores que aumentan el riesgo de desarrollar diabetes tipo II en el futuro.

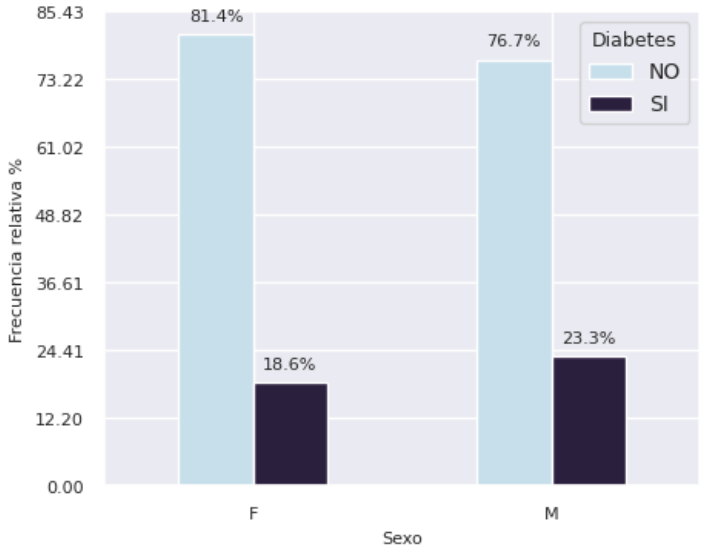

Figura 15 Distribución de diabetes por sexo
Fuente*:* Elaboración propia

Al evaluar la distribución de pacientes con diabetes tipo II en función del nivel educativo, se observa en la Figura 16 que, en el grupo de personas con educación básica primaria, el 77% no presentan diabetes, mientras que el 23% sí la tiene. Por otro lado, en el grupo de personas con educación posgrado, se encontró que el 80.7% no presenta diabetes, mientras que el 19.3% sí la tiene. De lo anterior se puede identificar una relación inversa entre el nivel educativo y la prevalencia de diabetes tipo II, dado que a medida que se incrementa el nivel educativo, disminuye la proporción de casos positivos de diabetes. Esto se debe a que las personas con

niveles educativos bajos pueden tener menos acceso a información de hábitos saludables o atención médica.

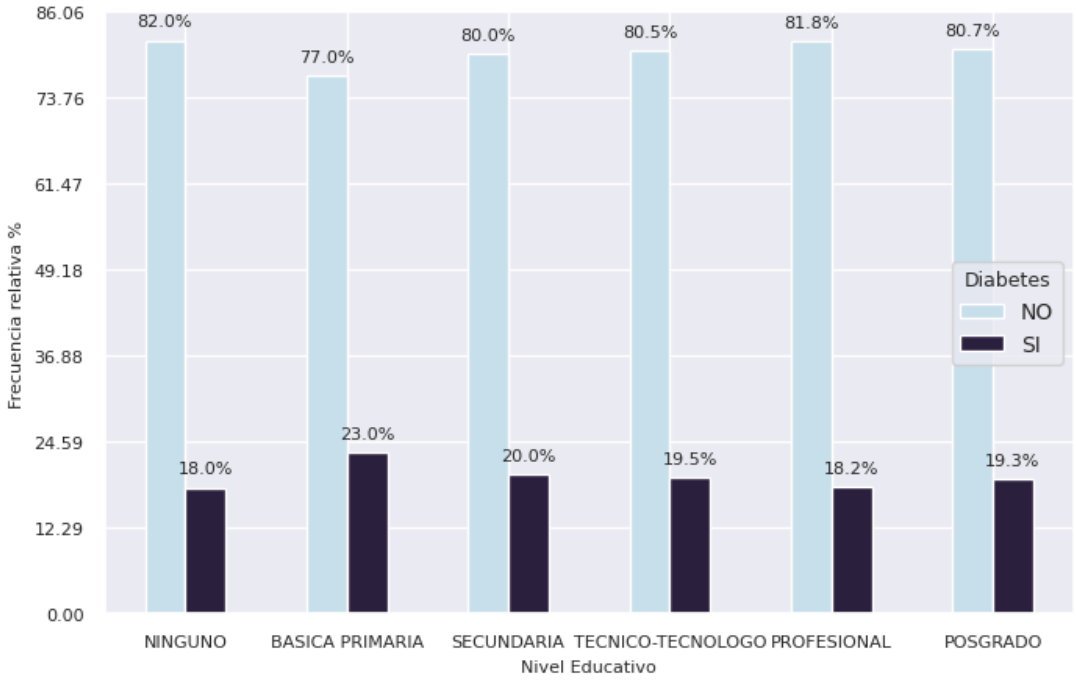

Figura 16 Distribución de diabetes por nivel educativo
Fuente*:* Elaboración propia

Al igual que con el nivel educativo, se observa una relación entre el rango salarial y el riesgo de desarrollar diabetes tipo II, a medida que el rango salarial aumenta, se aprecia una disminución en la proporción de casos positivos de diabetes. En la Figura 17 se observa en el rango salarial de 0 a 1, el 22.20% presenta diabetes, mientras que, en los rangos salariales más altos, como de 2.1 a 3 y mayor a 8, la proporción de casos positivos de diabetes se reduce al 17.30%.

Estos hallazgos sugieren una posible asociación entre el nivel de ingresos y la presencia de diabetes, indicando que niveles salariales más altos podrían estar relacionados con una menor prevalencia de esta enfermedad.

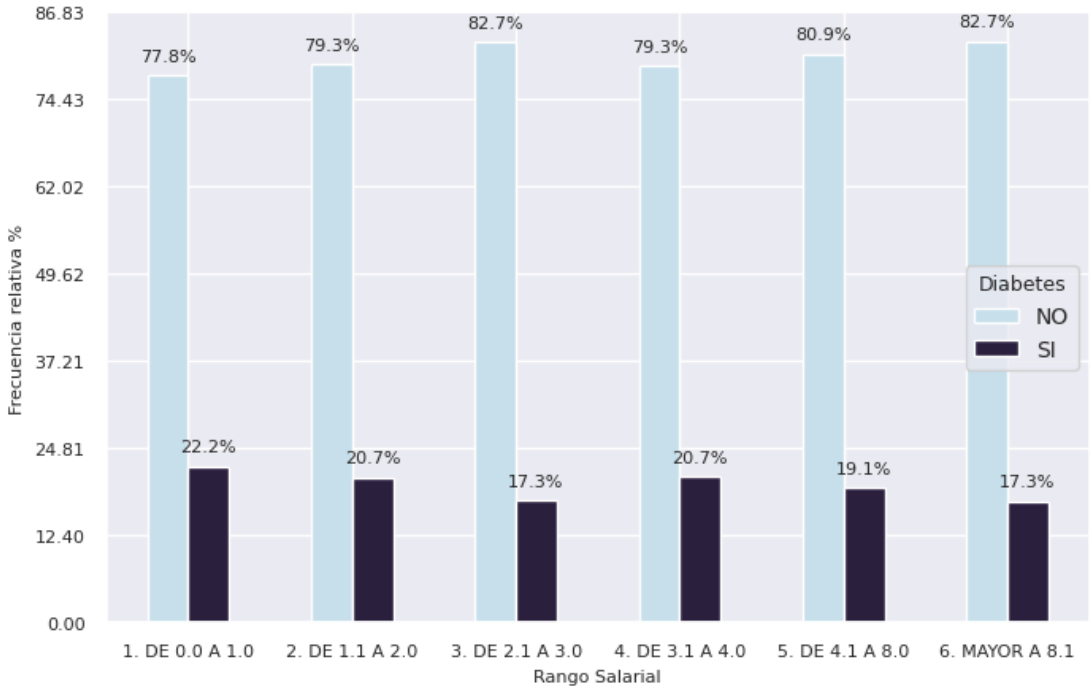

Figura 17 Distribución de diabetes por rango salarial
Fuente*:* Elaboración propia

Respecto al análisis de las variables geográficas, en la Tabla 5 Prevalencia de diabetes tipo II por regionalTabla 5 se presenta la prevalencia de la diabetes tipo II por regional, donde se evidencia que la regional caribe presenta una prevalencia de 28.85, es decir, por cada 100 pacientes, 29 padecen diabetes. Mientras que, la regional suroccidente a pesar de que es la regional donde la IPS tiene una mayor cobertura, es la que presenta la menor prevalencia, por cada 100 pacientes, 19 padecen diabetes.

Geográficamente se puede decir que la región suroccidente, nororiente, noroccidente y centroriente presenta una prevalencia similar respecto a la diabetes tipo II.

Tabla 5 Prevalencia de diabetes tipo II por regional

Regional	Diabetes Tipo II	Usuarios	Participación Regional	Prev. Regional
Suroccidente	No	42.393	25,39%	18,39
	Si	9.553		
Noroccidente	No	39.423	23,81%	19,07
	Si	9.289		
Nororiente	No	25.427	15,53%	19,95
	Si	6.338		
Centroriente	No	24.291	14,59%	18,60
	Si	5.549		
Caribe	No	17.042	11,71%	28,85
	Si	6.909		
Eje Cafetero	No	14.224	8,97%	22,51
	Si	4.133		

Fuente*:* Elaboración propia

Dentro del análisis es importante contemplar variables asociadas con los antecedentes familiares de enfermedades crónicas, dado que estas pueden ser inductores o factores de riesgo que aumentan la probabilidad de desarrollar la enfermedad de diabetes tipo II. Para la inclusión de estas variables se realizó la búsqueda de la información en la historia clínica de los pacientes y se seleccionaron aquellas variables disponibles que de acuerdo con los expertos pueden tener relación con la diabetes. Dentro de esta selección, se incluyeron antecedentes como la dislipidemia, obesidad, hipotiroidismo, enfermedad pulmonar crónica (EPOC), insuficiencia renal crónica (IRC) y enfermedades coronarias.

En la Figura 18 se observa la relación que existe con los antecedentes familiares de enfermedades crónicas y la diabetes tipo II, donde las personas con antecedentes de enfermedades podrían tener una mayor propensión a presentar diabetes tipo II en comparación con aquellos que no presentan antecedentes.

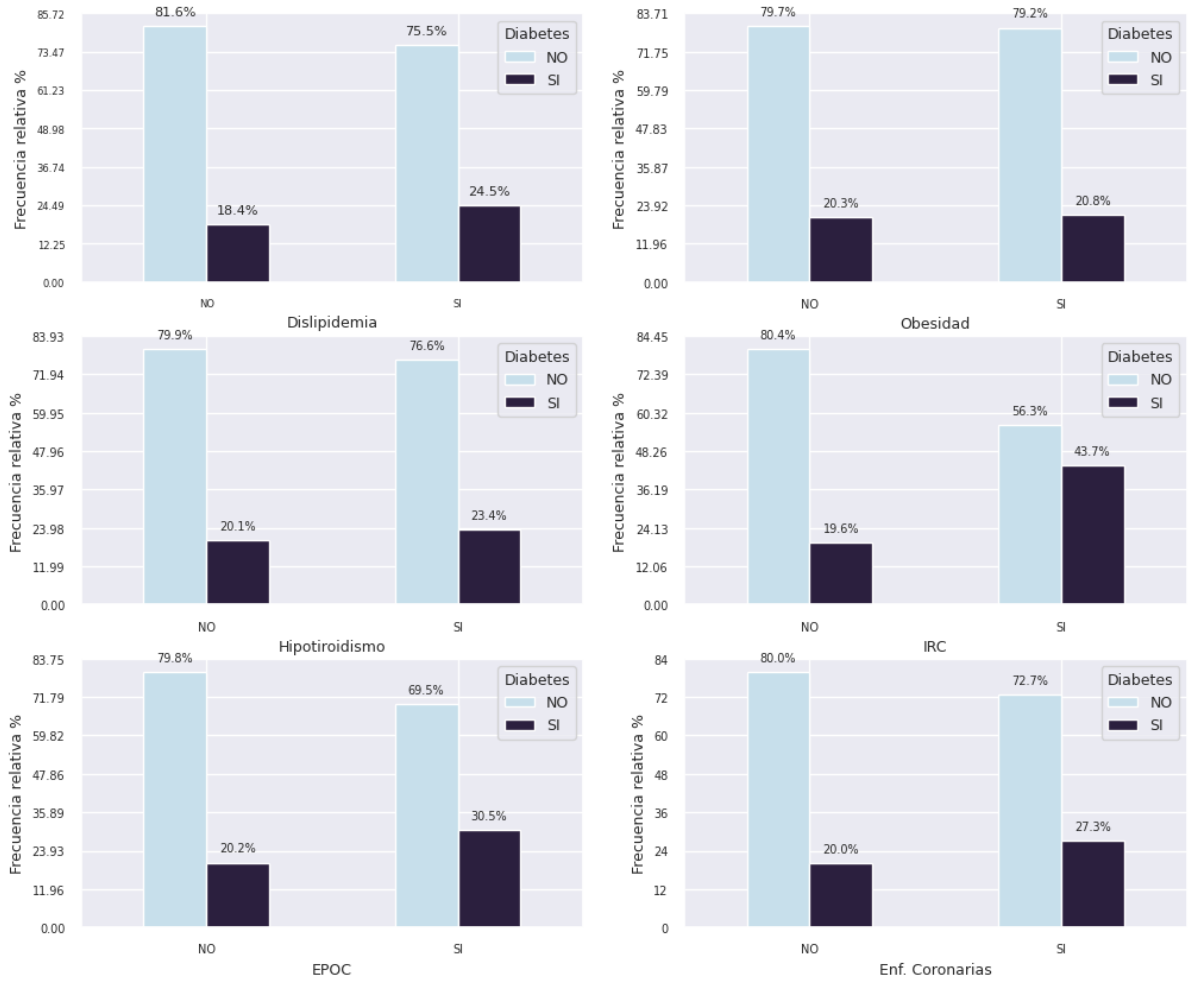

Figura 18 Distribución de antecedentes familiares de diabetes
Fuente*:* Elaboración propia

Cabe mencionar que dentro del análisis se consideraron variables como el grupo poblacional, antecedente familiar de enfermedad mental y grupo étnico, sin embargo, no se observó relación con la diabetes tipo II, por lo tanto, se excluyeron del análisis.

4.5.1 Asociación de la Diabetes con las variables predictoras cualitativas

En esta sección, se realiza prueba de asociación chi-cuadrado para evaluar la correlación entre la variable objetivo (riesgo de diabetes tipo II) y las variables

predictoras del presente estudio. El objetivo es determinar si existe una relación estadísticamente significativa entre estas variables. Para lo cual se plantea la siguiente prueba de hipótesis:

H_o: $No\ existe\ asociación\ entre\ las\ variables\ \ vs\ \ H_a$: $Existe\ asociación\ entre\ las\ variables$

En la Tabla 6 se presentan los resultados de la prueba chi-cuadrado, donde se observa que el valor-p es menor al nivel de significancia del 5%, este resultado indica que existe suficiente evidencia para rechazar la hipótesis nula. En otras palabras, se puede concluir que existe una asociación estadísticamente significativa entre la variable objetivo y las variables predictoras.

Tabla 6 Prueba Chi-Cuadrado

	Chi-squared statistic	**Valor-p**	**Grados de libertad**
Nivel Educativo	457,06	0,00000	5
Sexo	635,31	0,00000	1
Actividad Física	1242,90	0,00000	2
Regional	1347,90	0,00000	5
Rango Salarial	338,16	0,00000	5
Dislipidemia	1042,41	0,00000	1
Obesidad	7,0923	0,00773	1
Hipotiroidismo	129,674	0,00000	1
IRC	2484,35	0,00000	1
EPOC	228,718	0,00000	1
Enfer_Coronaria	356,851	0,00000	1

Fuente: Elaboración propia

El anterior resultado respalda la idea de que las variables predictoras analizadas están relacionadas con el riesgo de diabetes tipo II y son relevantes para la predicción de la enfermedad.

4.6 Modelamiento

Para dar respuesta a problemas de clasificación, existen diferentes modelos, en este caso, para predecir el riesgo de sufrir diabetes tipo II utilizando datos no clínicos, se utilizan algoritmos de machine learning tales como: Máquinas de Soporte Vectorial (SVM), XGBoost y Perceptrón Multicapa (MLP).

4.6.1 Pre-procesamiento

El preprocesamiento abarca todas las transformaciones aplicadas a los datos con el objetivo de que puedan ser interpretados de manera eficiente por el algoritmo de aprendizaje automático. Con el fin de cumplir con la condición de que ninguna información proveniente de las observaciones de prueba participe o influya en el ajuste del modelo, todas las operaciones se llevan a cabo durante el proceso de entrenamiento utilizando las observaciones de entrenamiento. A continuación, se describen los pasos utilizados en el preprocesamiento:

- **Imputación de datos faltantes**

Teniendo en cuenta la presencia de datos faltantes en algunas variables, se emplean las siguientes técnicas de imputación de datos. En primer lugar, se utiliza **SimpleImputer** para imputar los valores faltantes utilizando la mediana como valor estadístico de referencia. Posteriormente, se aplica **KNNImputer,** que utiliza el algoritmo de k-Nearest Neighbors para la imputación de datos.

Estas técnicas de imputación permiten estimar y reemplazar los valores faltantes en las variables, utilizando tanto la información estadística de la mediana como la similitud entre los registros vecinos. Al combinar estas estrategias, se busca obtener una imputación más precisa y robusta de los datos faltantes.

- **Estandarización de variables numéricas**

Si las variables predictoras son de tipo numérico, la escala en la que se miden tiene un impacto significativo en el modelo. Algunos algoritmos de machine learning, como SVM y redes neuronales, son sensibles a la escala de las variables. Esto implica que si no se igualan de alguna manera los predictores, aquellos que se midan en una escala mayor o tengan una mayor variabilidad dominarán el modelo, incluso si no tienen una relación más fuerte con la variable objetivo. Para evitar este problema, se utiliza la siguiente estrategia: **StandardScaler**. Con esta estrategia, los datos se transforman estandarizando cada variable predictora mediante la substracción de la media y la posterior escalada a una varianza unitaria.

- **Dummificación de variables categóricas**

En este caso, para las variables predictoras cualitativas se utiliza el proceso de **one-hot-encoding** el cual consiste en crear nuevas variables dummy con cada uno de los niveles de las variables cualitativas, donde se garantiza eliminar uno de los niveles de cada variable para evitar redundancias o multicolinealidad.

Finalmente, para preprocesar los datos se utiliza de la librería **scikit-learn** las funciones de **ColumnTransformer** y **pipeline**. La función **ColumnTransformer** posibilita la combinación de múltiples transformaciones de preprocesamiento, indicando qué columnas se aplican a cada una. Al igual que cualquier otro transformador, esta función cuenta con un método de entrenamiento (fit) y un método de transformación (transform). Esto posibilita que las transformaciones se aprendan exclusivamente utilizando las observaciones del conjunto de entrenamiento y luego se puedan aplicar a cualquier conjunto de datos.

4.7 Hiperparámetros empleados

Se utilizó la herramienta de Python bajo el entorno de Google Colab para el entrenamiento y ajuste del mejor modelo. Realizando una partición aleatoria de los datos, con un 80% para entrenamiento y un 20% para prueba, teniendo en cuenta la estratificación de la variable objetivo (Diabetes).

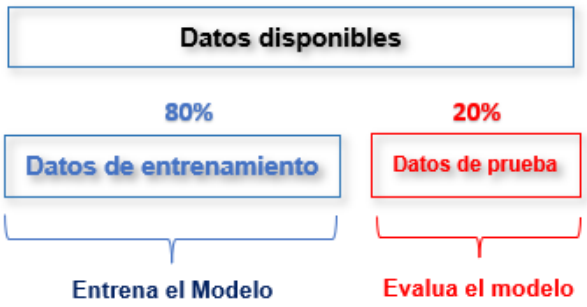

Figura 19 Partición de datos en entrenamiento y prueba.

Respecto a la búsqueda optima de hiperparametros, se utiliza la **optimización bayesiana** de la librería scikit-learn. En líneas generales, la optimización bayesiana de hiperparámetros implica la construcción de un modelo probabilístico en el que la métrica de validación del modelo, en este caso el ROC AUC, es la función objetivo. Mediante esta estrategia, la búsqueda se dirige iterativamente hacia las regiones de mayor interés, permitiendo un enfoque más efectivo en la exploración de los hiperparámetros.

En las siguientes tablas se describen los hiperparámetros empleados en cada uno de los modelos con el propósito de encontrar la mejor combinación:

Las definiciones de cada uno de los hiperparámetros para el modelo XGBoost se tomaron de (Xgboost- 2022).

Tabla 7 Combinación de hiperparámetros utilizados en el algoritmo XGBoost

Hiperparámetros	Definición	Rango o valor de Hiperparámetros
learning_rate	"Tasa de aprendizaje utilizada para controlar el tamaño de los ajustes de los pesos del modelo en cada paso durante el entrenamiento. Este parámetro se utiliza para regular la influencia de cada árbol en el modelo final y prevenir el sobreajuste del modelo"	(0 - 1)
min_child_weight	"Se utiliza para controlar la cantidad mínima de peso que se requiere para crear un nuevo nodo en el árbol de decisión".	(1 - 15)
max_depth	"Se utiliza para controlar la profundidad máxima de cada árbol de decisión en el modelo. Este parámetro se utiliza para evitar el sobreajuste del modelo al limitar la cantidad de divisiones que se pueden hacer en un árbol de decisión".	(2 - 6)
subsample	"Se utiliza para controlar la fracción de observaciones que se utilizan para construir cada árbol de decisión en el modelo. Este parámetro se utiliza para reducir el sobreajuste del modelo y mejorar la generalización del modelo".	(0.1 - 0.9)
colsample_bytree	"Es un hiperparámetro que controla la fracción de características que se utilizan para construir cada árbol de decisión en el modelo, y se utiliza para reducir el sobreajuste del modelo y mejorar la generalización del modelo".	(0.1 - 0.8)
reg_lambda	"Controla la cantidad de regularización L2 que se aplica a los pesos del modelo, y se utiliza para evitar el sobreajuste del modelo al penalizar los valores grandes de los pesos del modelo".	(0.1 - 0.5)
reg_alpha	"Controla la cantidad de regularización L1 que se aplica a los pesos del modelo, y se utiliza para evitar el sobreajuste del modelo al penalizar los valores grandes y no importantes de los pesos del modelo".	(0.1 - 0.5)
gamma	"Controla la cantidad de reducción mínima necesaria en la función de pérdida para que se produzca una partición adicional en el árbol de decisión, y se utiliza para evitar el sobreajuste del modelo al reducir el número de particiones en el árbol de decisión".	(1 - 10)
n_estimators	"Es un hiperparámetro que controla el número de árboles de decisión que se deben construir durante el entrenamiento del modelo, y se utiliza para controlar la complejidad del modelo y el tiempo de entrenamiento".	2000

Fuente: Elaboración propia

Las definiciones de cada uno de los hiperparámetros para la red neurobal se tomaron de la documentación de la librería scikit learn (Scikitlearn- MLP (2013).

Tabla 8 Combinación de hiperparámetros utilizados en el algoritmo MLP Clasiffier

Hiperparámetros	Definición	Rango o valor de Hiperparámetros
hidden_layer_sizes	"Se refiere al número de neuronas en cada capa oculta de la red neuronal".	(5 - 7)
activation	"Función de activación de la capa oculta"	**relu**: función de unidad lineal rectificada. **logistic**: función logística sigmoidea. **tanh**: función tan hiperbólica.
solver	"Solucionador para la optimización del peso"	**lbfgs**: "es un optimizador en la familia de métodos cuasi-Newton". **Sgd**: "se refiere al descenso de gradiente estocástico". **adam**: "se refiere a un optimizador estocástico basado en gradientes propuesto por Kingma, Diederik y Jimmy Ba".
batch_size	"Tamaño de minilotes para optimizadores estocásticos. Si el solucionador es 'lbfgs', el clasificador no usará minibatch".	(6 - 12)
learning_rate_init	"La tasa de aprendizaje inicial utilizada. Controla el tamaño del paso en la actualización de los pesos. Solo se usa cuando solver='sgd' o 'adam'".	(0.0001 - 1)
alpha	"Fuerza del término de regularización L2. El término de regularización L2 se divide por el tamaño de la muestra cuando se suma a la pérdida".	(0.0001 - 1)
max_iter	"Número máximo de iteraciones. El solucionador itera hasta la convergencia. Para los solucionadores estocásticos ('sgd', 'adam'), determina la cantidad de épocas (cuántas veces se usará cada punto de datos), no la cantidad de pasos de gradiente".	(100-500)
early_stopping	"Si se debe utilizar la interrupción anticipada para finalizar el entrenamiento cuando la puntuación de validación no mejora".	True

Fuente: Elaboración propia

Las definiciones de cada uno de los hiperparámetros para el modelo de SVM Classifier se tomaron de la documentación de la librería (Scikitlearn- SVC (2013).

Tabla 9 Combinación de hiperparámetros utilizados en el algoritmo SVM Clasiffier

Hiperparámetros	Definición	Rango o valor de hyperparametros
C	“Parámetro de regularización. La fuerza de la regularización es inversamente proporcional a C. Debe ser estrictamente positiva. La penalización es una penalización de l2 al cuadrado”.	(0.00001 - 1)
gamma	“Coeficiente kernel para 'rbf', 'poly' y 'sigmoid'”.	(0.00001 - 1)
tol		(0.00001 - 1)
kernel	“Especifica el tipo de kernel que se utilizará en el algoritmo. Si no se proporciona ninguno, se utilizará 'rbf'. Si se proporciona un invocable, se utiliza para calcular previamente la matriz del kernel a partir de matrices de datos”	linear, sigmoid, rbf
max_iter	“Límite estricto en las iteraciones dentro del solucionador. -1 para ningun límite”	(100 - 500)

Fuente: Elaboración propia

4.8 Evaluación del modelo

Para evaluar el modelo se utilizó el protocolo de evaluación K-Fold Cross-Validation, “el cual consiste en dividir los datos de forma aleatoria en *k* grupos de aproximadamente el mismo tamaño, *k-1* grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite *k* veces utilizando un grupo distinto como validación en cada iteración, este protocolo se caracteriza por permitir un balance entre sesgo y varianza, una característica ideal para alcanzar el objetivo del presente trabajo” (Raschka, S., & Mirjalili, V. 2017).

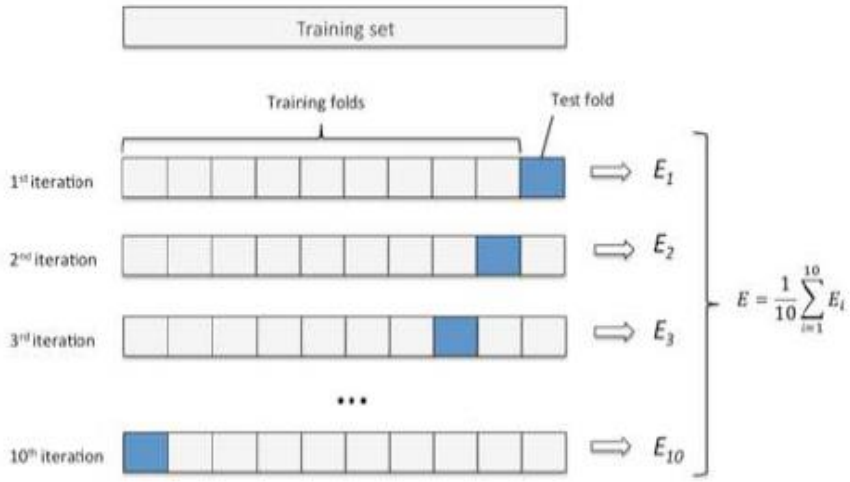

Figura 20 Protocolo de evaluación del modelo
Fuente: Tomado del libro Python Machine Learning de Sebastian Raschka.

Respecto a la evaluación de los resultados del modelo se tiene en cuenta el comportamiento de la curva ROC y las métricas de evaluación recall, precisión y puntuación F1, las cuales se encuentran definidas en la sección 2.2.2.6.

5. DISEÑO DE EXPERIMENTO DE VALIDACIÓN

En las investigaciones relacionadas con predicción de enfermedades se esperan errores mínimos en las predicciones dado que se encuentra de por medio la salud de las personas, en este trabajo, se establece con la IPS predecir correctamente al menos un 70% de los casos, lo cual es un resultado inicial que se puede ir incrementando con el mejoramiento continuo del modelo y de los procesos de información de la entidad.

Con el fin de obtener una mejor estimación de la variable respuesta, se recurrió a la optimización bayesiana para encontrar los mejores hiperparámetros minimizando la función de pérdida. Debido a la gran cantidad de combinaciones de hiperparámetros que se pueden obtener, se definieron en la sección 48 los rangos de búsqueda para

cada hiperpárametro en cada familia de modelo. Adicionalmente, en este proyecto el protocolo de evaluación a utilizar será la técnica de K-fold Cross-Validation, en el cual se selecciona un k=3.

6. RESULTADOS OBTENIDOS

Para exponer los resultados obtenidos se va a realizar un benchmarking de los modelos planteados con el objetivo de establecer comparaciones de los indicadores de desempeño de la predicción y poder seleccionar el mejor modelo para predecir el diagnostico de diabetes tipo II.

6.1 Benchmarking de modelos

De acuerdo con la metodología planteada se muestra el comportamiento de cada iteración, donde en cada una de ellas se están variando los hiperpárametros definidos en la sección 487 para cada familia de modelo, con el objetivo de buscar los dos de mejor desempeño con la métrica ROC AUC.

En la Figura 21, se observa el desempeño de cada iteración realizada para la familia de modelos XGBoost en función de la métrica ROC AUC. En general, se observa que el desempeño del modelo en el conjunto de entrenamiento está por encima del desempeño comparado con el conjunto de validación, es decir que el modelo está mejorando su capacidad de generalización en cada iteración, lo que sugiere que se está logrando un buen equilibrio entre la capacidad de ajuste del modelo y su capacidad de generalización.

Figura 21 Iteraciones del modelo XGBoost
Fuente: Elaboración propia

De acuerdo con el optimizador de la familia XGboost se seleccionaron los siguientes dos mejores modelos, en los cuales se puede evidenciar que el ROC AUC presenta resultados similares entre los conjuntos de datos de train y de test, dado que ambas métricas son similares nos da un indicio de que no se presenta sobreajuste en el modelo seleccionado, en la Tabla 10 se muestran los hiperparámetros seleccionados para cada modelo.

Tabla 10 Hiperpárametros modelos XGBoost

Hiperpárametro	Modelo 1	Modelo 2
“colsample_bytree	0.8	0.8
gamma	10	8
learning_rate	0.1	0.1
max_depth	2	2
min_child_weight	15	15
n_estimators	2000	2000
n_jobs	-1	-1
objective	binary logistic	binary logistic
random_state	42	42
reg_alpha	0.9	0.9

reg_lambda	0.5	0.1
Subsample”	0.1	0.9

Validation ROC AUC	0.7661	0.7668
Train ROC AUC	0.7693	0.7707

Fuente: Elaboración propia

En la Figura 22, se observa el desempeño de cada iteración realizada para la familia de modelos MLP-Classifier en función de la métrica ROC AUC, donde se evidencia en cada iteración resultados muy similares en los conjuntos de datos de train y de test, lo cual es una señal de que no se presenta sobreajuste en los modelos seleccionados.

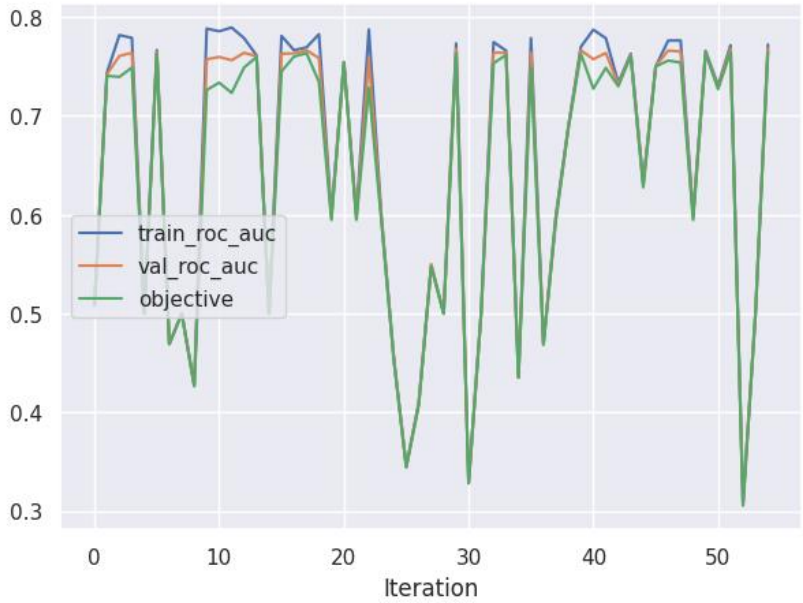

Figura 22 Iteraciones del modelo MLP-Classifier
Fuente: Elaboración propia

De acuerdo con el optimizador de la familia MLP-Classifier se seleccionaron los siguientes dos mejores modelos, en la Tabla 11Tabla 10 se muestran los hiperparámetros seleccionados para cada modelo.

Tabla 11 Hiperpárametros modelos MLP-Classifier

Hiperpárametro	Modelo 1	Modelo 2
“activation	logistic	logistic
alpha	0.6624	0.9507
batch_size	64	128
early_stopping	True	True
hidden_layer_sizes	[32, 16, 8]	[16, 8]
learning_rate_init	0.1115	0.2299
max_iter	133	135
random_state	42	42
solver”	lbfgs	lbfgs
Validation ROC AUC	0.7682	0.7684
Train ROC AUC	0.7716	0.7723

Fuente: Elaboración propia

Adicional, se realizó el entrenamiento para la familia de modelos Support Vector Machine, en la Figura 23 se observa el desempeño de cada iteración en función de la métrica ROC AUC, donde se puede evidenciar que el valor máximo que alcanza el ROC AUC es del 60%, dado que este estudio está relacionado con la predicción de una enfermedad se esperan errores mínimos en las predicciones, por lo tanto, se define que el modelo debe predecir correctamente al menos un 70% de los casos. Teniendo en cuenta lo anterior, se descarta el modelo SVM al no cumplir con el criterio establecido respecto a la métrica ROC AUC.

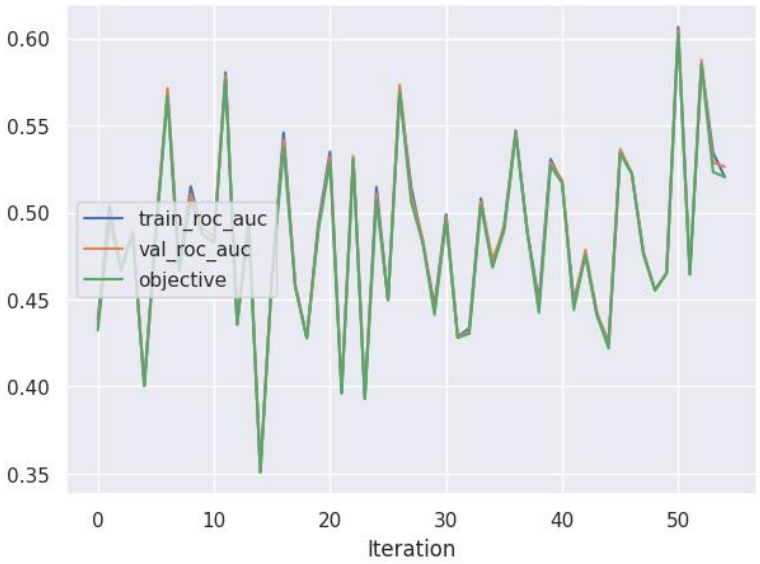

Figura 23 Iteraciones del modelo SVM
Fuente: Elaboración propia

6.1 Definición del punto de corte para las probabilidades obtenidas

Posterior de seleccionar los mejores modelos de cada familia, surge un paso muy importante que implica combinar los resultados con el contexto del negocio, donde es importante socializar los resultados obtenidos con los expertos y de esta manera definir el punto corte para las probabilidades. El punto de corte se utiliza para determinar el umbral a partir del cual se clasificará una muestra como positiva o negativa en función de la probabilidad estimada por el modelo.

En este caso, se decide seleccionar dos puntos de corte, uno donde la tasa de positivos sea del 10% y otro donde sea del 20%, al establecer el punto de corte con una tasa de positivos del 10%, se está adoptando un enfoque más conservador para minimizar los falsos positivos, es decir, se prefiere clasificar erróneamente a menos personas como positivas, incluso si eso implica perder algunos casos verdaderos positivos.

Por otro lado, al seleccionar el punto de corte con una tasa de positivos del 20%, se está optando por un umbral más bajo y menos restrictivo, buscando capturar un mayor número de casos positivos, aunque esto pueda conllevar una proporción más alta de falsos positivos.

En la Figura 24 se observa el comportamiento de las métricas recall, precisión, F1 y tasa de positivos (Rate_Positive) versus el umbral para cada uno de los mejores modelos seleccionados, donde la línea punteada corresponde a los dos puntos de corte seleccionados, se evidencia que los puntos de cortes seleccionados para los modelos son menores a 0.5, es decir, se está aumentando la probabilidad de clasificar una muestra como positiva; lo que implica en una mayor tasa de verdaderos positivos.

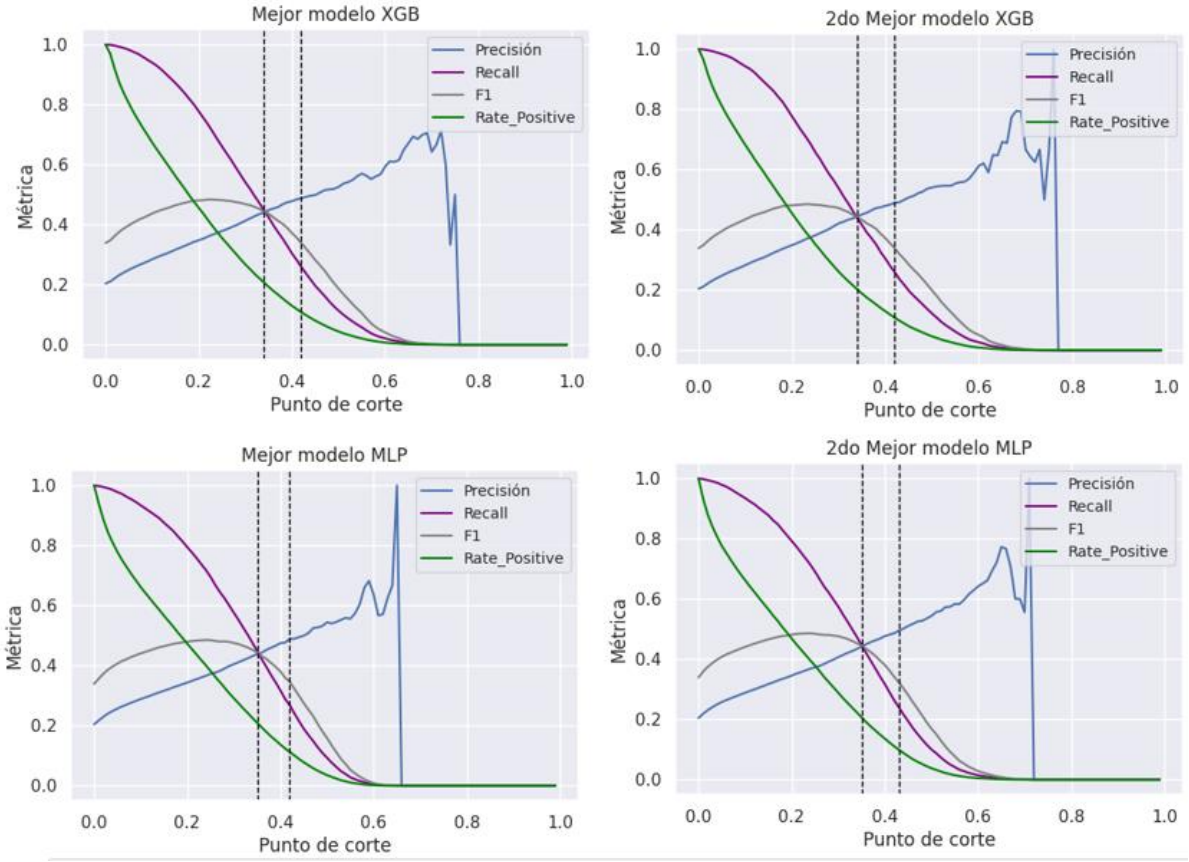

Figura 24 Puntos de corte para los mejores modelos seleccionados
Fuente: Elaboración propia

6.2 Resultados del Modelo

Una vez seleccionado el punto de corte para cada uno de los modelos, el paso posterior es evaluar el desempeño y determinar qué tan bien puede generalizar para clasificar nuevas muestras. Para este objetivo se muestra en la Tabla 12 las métricas utilizadas para evaluar el desempeño de los modelos, donde se evidencia resultados similares en las métricas.

En este caso, es de interés seleccionar un modelo que presente un buen recall, dado que se desea clasificar correctamente a los pacientes que realmente tienen riesgo de desarrollar la enfermedad diabetes tipo II. A su vez, es importante tener una alta precisión, ya que con esto se asegura que el modelo prediga correctamente los pacientes que van a padecer el diagnostico. Por lo tanto, se decide seleccionar el modelo que muestre el mejor desempeño en la métrica F1-Score y que tenga el menor sobreajuste entre los conjuntos de datos de train y test. De acuerdo con los criterios anteriores, se elige el mejor modelo de la familia XGBoost con los parámetros establecidos en la Tabla 7.

Tabla 12 Métricas de desempeño de los mejores modelos

Modelo	Punto de corte	Rate-positive	Precisión	F1	recall	kappa	ROC- AUC (Train)	ROC- AUC (Test)
Mejor modelo XgBoost	0,43	0,1	0,492	0,321	0,238	0,217	0,766	0,769
	0,34	0,2	0,440	0,444	0,447	0,300		
2do Mejor modelo XgBoost	0,42	0,1	0,490	0,338	0,258	0,230	0,767	0,771
	0,34	0,2	0,445	0,441	0,438	0,299		
Mejor modelo MLP1	0,42	0,1	0,488	0,346	0,268	0,235	0,768	0,772
	0,35	0,2	0,438	0,444	0,449	0,299		
2do Mejor modelo MLP2	0,43	0,1	0,495	0,323	0,241	0,219	0,768	0,772
	0,35	0,2	0,441	0,444	0,447	0,301		

Fuente: Elaboración propia

Para el modelo seleccionado se obtuvo un recall de 44.7%, lo cual indica que el modelo tiene la capacidad de detectar aproximadamente el 45% de los casos de diabetes presentes en el conjunto de datos. Mientras que, se está obteniendo una precisión del 44% lo que indica que de todos los casos que el modelo etiqueta como

diabetes, aproximadamente el 44% son realmente casos de diabetes, mientras que el resto son falsos positivos.

Partiendo del modelo seleccionado, en la Figura 25 se observa las variables que más contribuyen en la predicción temprana de diabetes tipo II. Donde el IMC, edad y superficie corporal son factores determinantes para llegar a clasificar un paciente con el riesgo de sufrir diabetes, dado que un IMC elevado puede afectar la capacidad del cuerpo para producir y utilizar la insulina de manera eficiente aumentando el riesgo de padecer el diagnostico. La edad, también es un factor importante, ya que a mayor edad el cuerpo humano se vuelve menos sensible a la insulina que se produce. (Federación Internacional de Diabetes [FID], 2019)

Se valida estos hallazgos con las direcciones medicas de la IPS donde resaltan la importancia de monitorear y controlar el IMC, la edad y la superficie corporal como medidas preventivas para la diabetes tipo II. Identificar y abordar estos factores de riesgo de manera temprana puede ser fundamental para prevenir o retrasar el desarrollo de la enfermedad.

Figura 25 Importancia de variables del modelo seleccionado
Fuente: Elaboración propia

7. CONCLUSIONES Y FUTURO TRABAJO

En este trabajo se ha abordado el reto de predecir la diabetes tipo II mediante el uso de metodología de Machine Learning, donde su principal objetivo es contar con el diagnóstico temprano para mejorar la calidad de vida de las personas y reducir los costos asociados a su tratamiento, en este trabajo se ha validado la predicción del diagnóstico a partir de datos no clínicos y genéticos de los pacientes que han consultado en una Institución Prestadora de Salud durante los años 2021 y 2022. Los resultados obtenidos muestran lo siguiente:

- Dentro del análisis descriptivo se obtuvo como resultado en las variables demográficas que de los pacientes que estaban identificados con el riesgo de diabetes tipo II, el 56% eran mujeres, mientras que el 44% son hombres, mostrando una mayor prevalencia del riesgo para las mujeres. En cuanto a la distribución de la edad se demuestra que los pacientes con edades mayores a 50 años presentan más prevalencia, no obstante, se encuentran casos diagnosticados para todas las edades.

- Dentro de las variables sociodemográficas, se observa que de los diagnosticados el 69 % se distribuyen entre usuarios que ganan menos de un salario mínimo y los que se encuentran entre 1.6 y 2 salarios. También se evidenció que el 35.6% de los casos son usuarios que alcanzaron un nivel educativo de básica primaria y el 37% son usuarios que culminaron la secundaria.

- Respecto al análisis de las variables geográficas, se evidencia que la regional caribe presenta una prevalencia de 28.85, es decir, por cada 100 pacientes, 29 padecen diabetes. Mientras que, la región suroccidente a pesar de que es

la regional donde la IPS tiene una mayor cobertura, es la que presenta la menor prevalencia, por cada 100 pacientes, 19 padecen diabetes.

- En el análisis de las variables antropométricas, se observa en la distribución del IMC que el 38.3% de los usuarios con diabetes tipo II presentan una medida mayor o igual a 30, donde para esta medida se clasifican como usuarios con obesidad. Por otra parte, la medida de superficie corporal el 55.48% de los casos con diabetes presentan una medida mayor o igual a 1.8 la cual se considera como una medida alta para esta métrica.

- De acuerdo con el grupo de variables genéticas o de antecedentes familiares que se consultaron: Dislipidemia, Obesidad, Hipotiroidismo, IRC, EPOC y Enfermedades coronarias, se observó que dentro de estos perfiles de riesgo se obtuvo participación de casos de diabetes tipo II, pero los riesgos que más prevalencia tuvieron con diabetes fueron la Dislipidemia, Obesidad y enfermedades coronarias.

- Dentro de los resultados obtenidos de la aplicación de metodologías de Machine Learning, se determinó para el conjunto de datos evaluados que los modelos entrenados con SVM no alcanzan a obtener un resultado óptimo en el desempeño de las predicciones obteniendo como máximo un ROC-AUC del 60%.

- Para el conjunto de datos analizados en este estudio, los modelos entrenados utilizando el algoritmo XGBoost y MLP presentan métricas de desempeño muy similares en las predicciones. Sin embargo, el modelo basado en XGBoost presentó un menor grado de sobreajuste, obteniendo como resultado un ROC-AUC del 77%.

- XGBoost tiene la propiedad de proporcionar información sobre la importancia de variables en la predicción de la diabetes tipo II, es decir que el modelo

seleccionado identifica las variables más relevantes que inciden en la probabilidad de desarrollar la enfermedad, como resultado para este trabajo no indica que las variables IMC, edad y la superficie corporal son identificadas como factores de riesgos asociados para padecer el riesgo. Estos resultados destacan la importancia de adoptar hábitos saludables como mantener un peso adecuado y llevar una alimentación equilibrada y para reducir el riesgo de desarrollar esta enfermedad crónica.

- Los resultados obtenidos en el trabajo son positivos y puede ser tomado como punto de partida, ya que la no inclusión de variables clínicas hace que los actores del sistema de salud en Colombia puedan llegar a simplificar el proceso y poder obtener una detección temprana de la diabetes tipo II previniendo comorbilidades a los pacientes y disminuir el costo médico.

Para trabajos futuros, se sugiere incluir variables no consideradas en el desarrollo de este trabajo como datos genéticos, antecedentes familiares de diabetes, hábitos de estilo de vida y etnia las cuales pueden estar asociadas con la diabetes tipo II y podrían ayudar a mejorar la precisión y el recall obtenido en los modelos de predicción considerados. Estas variables no se incluyeron en el presente trabajo debido a problemas de calidad de los datos, dado que presentaron una alta proporción de valores faltantes, por lo tanto, su inclusión en el modelamiento podía generar sesgos o distorsiones en los resultados.

8. BIBLIOGRAFÍA

Chen, T., y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. En Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794).

Federación Internacional de Diabetes (2019). ATLAS DE LA DIABETES DE LA FID, Novena Edición. URL: https://www.diabetesatlas.org/upload/resources/material/20200302_133352_2406-IDF-ATLAS-SPAN-BOOK.pdf

Federación Internacional de Diabetes (12 de diciembre de 2021).Diabetes facts & figures. https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Galindo, E. A., Perdomo, J. A., & Figueroa-García, J. C. (2020). Estudio comparativo entre máquinas de soporte vectorial multiclase, redes neuronales artificiales y sistema de inferencia neuro-difuso auto organizado para problemas de clasificación. Información tecnológica, 31(1), 273-286.

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 22(5), 717-727.

Hemati, S., Beiranvand, P., & Sharafi, M. (2019). ellipse perimeter estimation using nonparametric regression of rbf neural network based on elliptic integral of the second type. Investigación Operacional, 39(4), 639-646.

Instituto Nacional de Salud [NIH], (16 de noviembre de 2016). Información general sobre la Diabetes. https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/que-es

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

Martínez Leal, A. (2021). Diagnóstico de la diabetes mediante el uso de técnicas de aprendizaje automático (Doctoral dissertation, Universitat Politècnica de València).

Ministerio de Salud y Protección Social (14 de febrero de 2022). Mortalidad en Colombia.
https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/ED/GCFI/mortalidad-colombia-periodo-2020-2021.pdf

Organización Panamericana de la Salud [OPS], 2019. Información general sobre la Diabetes. https://www.paho.org/es/temas/diabetes

Pérez Leal, L. E. (2021). Predicción del diagnostico de diabetes a partir de perfiles clínicos de pacientes utilizando aprendizaje automático.

Roman, V. (2019, Marzo 27). Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos. Obtenido de Medium: https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducci%C3%B3n-a-la-clasificaci%C3%B3n-y-principales-algoritmos-dadee99c9407

Vargas, J., Conde, M. B., Paccapelo, M. V., & Zingaretti, M. L. (2012, August). Máquinas de soporte vectorial: metodología y aplicación en R. In Décimo Congreso Latinoamericano de Sociedades de Estadística.

Zhou, J., Qiu, Y., Zhu, S., Armaghani, D. J., Khandelwal, M., y Mohamad, E. T. (2020). Estimation of the tbm advance rate under hard rock conditions using xgboost and bayesian optimization. Underground Space.

Zhang, L., Wang, Y., Niu, M., Wang, C., & Wang, Z. (2020). Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan Rural Cohort Study. Scientific reports, 10(1), 1-10.

Gallego Valcárcel, D. A., & Lucas Monsalve, D. F. (2021). Modelos de aprendizaje automático para la predicción del riesgo de fatalidad por insuficiencia cardiaca con datos clínicos.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” Neural Computation 9(8):1735–80.

Cuenta de Alto Costo (2021). Día mundial de la diabetes 2021. Recuperado el día 02 de diciembre de 2022 tomado de: https://cuentadealtocosto.org/site/general/dia-mundial-de-la-diabetes-2021/#:~:text=M%C3%A1s%20de%20un%20mill%C3%B3n%20seiscientos%20mil%20colombianos%20tienen%20diabetes&text=De%20acuerdo%20con%20la%20informaci%C3%B3n,mellitus%20(DM)%20en%20Colombia.

Raschka, S., & Mirjalili, V. (2017). Python machine learning: Machine learning and deep learning with python. Scikit-Learn, and TensorFlow. Second edition ed, 3

Sitiobigdata.com https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/#

Scikitlearn- MLP (2013)
https://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

(Scikitlearn- SVC (2013) https://scikit-learn.org/stable/about.html#citing-scikit-learn

(Xgboost- 2022) https://xgboost.readthedocs.io/en/stable/parameter.html