MODELO DE PREDICCIÓN DE DIABETES TIPO 2 (DMT2) A PARTIR DE VARIABLES NO CLÍNICAS EN UNA POBLACIÓN ASEGURADA DEL SUROCCIDENTE COLOMBIANO PERTENECIENTE AL RÉGIMEN SUBSIDIADO

LARRY FARID CASTRO SALAMANCA

JUAN ESTEBAN LÓPEZ

UNIVERSIDAD ICESI

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

MODELO DE PREDICCIÓN DE DIABETES TIPO 2 (DMT2) A PARTIR DE VARIABLES NO CLÍNICAS EN UNA POBLACIÓN ASEGURADA DEL SUROCCIDENTE COLOMBIANO PERTENECIENTE AL RÉGIMEN SUBSIDIADO

LARRY FARID CASTRO SALAMANCA

JUAN ESTEBAN LÓPEZ

Trabajo de Grado presentado como requisito para optar al

Título de Magister en Ciencia de Datos

Director: JOSÉ ARMANDO ORDÓÑEZ, PhD

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

FACULTAD DE INGENIEÍA

MAESTRÍA EN CIENCIA DE DATOS

AUTORES:

LARRY FARID CASTRO SALAMANCA

JUAN ESTEBAN LÓPEZ

TITULO:

MODELO DE PREDICCIÓN DE DIABETES TIPO 2 (DMT2) A PARTIR DE VARIABLES NO CLÍNICAS EN UNA POBLACIÓN ASEGURADA DEL SUROCCIDENTE COLOMBIANO PERTENECIENTE AL RÉGIMEN SUBSIDIADO

TEMAS O PALABRAS CLAVES DE LA TESIS:

Diabetes Mellitus, Variables No Clínicas, Machine Learning, Balance de Clases, Modelo de predicción

NOTA DE ACEPTACIÓN:

_____
_____
_____

_____
PRESIDENTE DEL JURADO

_____
JURADO

_____
JURADO

SANTIAGO DE CALI, JUNIO DE 2023

TABLA DE CONTENIDO

Pág.

LISTA DE TABLAS

Pág.

LISTA DE ILUSTRACIONES

Pág.

LISTA DE ANEXOS

Pág.

RESUMEN

La detección de la diabetes es el desafío de los sistemas de Salud debido a la complejidad de tener en cobertura su población asegurada con respecto a los programas de promoción y prevención, y los de gestión del riesgo, los cuales, tienen altos costos asociados en el proceso de la toma de pruebas clínicas, dado que requieren una logística compleja, sobre todo en zonas rurales o de difícil acceso.

El presente trabajo de grado estudió información de una población vulnerable del régimen subsidiado la cual está asignada a una EPS que tiene cobertura en el suroccidente colombiano. Esta información consignada en una base de datos tiene características no clínicas como sociales, ambientales, modo de vida, entorno en el que viven y una etiqueta asignada por el área de riesgo de la EPS, la cual indica si un afiliado tiene o no diabetes mellitus tipo 2 (DMT2). De acuerdo con lo anterior, se le aplicó a la información, técnicas para entrenar modelos supervisados con el fin de detectar y predecir la enfermedad DMT2 mediante la metodología CRISP-DM, donde consideramos las fases de entendimiento del negocio, entendimiento de los datos, preparación y modelado.

En la fase de modelado se utilizaron técnicas de *machine learning* para aprendizaje supervisado como son: Naive Bayes, Random Ferest, Árboles de decisión, KNN, Regresión Logística, Gradient Boosting, eXtreme Gradient Boosting y Multilayer perceptrón (Aplicado en el escenario 2). Además, estos fueron utilizados junto con técnicas de muestreo como *Undersampling, Oversampling, Smote y Smote-tomek* para afrontar el problema de balance de clases entre pacientes sanos y enfermos, los cuales son la minoría (6.3%). La principal métrica para optimizar en el proyecto fue la ***sensibilidad o recall***, dado que es más importante identificar correctamente un paciente enfermo sobre el que no padece la enfermedad. Sin embargo, se buscó la estabilidad entre la ***exactitud (accuracy) y la sensibilidad (recall)*** para obtener un modelo congruente.

Para la validación del modelo se utilizó la técnica de ***hold-out*** estratificado para obtener una base de datos de prueba (test) del 20% del tamaño original y el 80% restante se utilizó para el entrenamiento (train) mediante un ***k-fold=10*** junto con una ***optimización bayesiana*** para buscar los mejores hiperparámetros de cada algoritmo. Se plantean dos escenarios, uno con un procedimiento adicional de selección de variables mediante regularización Lasso y el otro solo con la selección inicial dada la calidad de la base de datos y la revisión del estado del arte.

En primera instancia, la aplicación de los modelos supervisados sin técnicas de muestreo, generaron métricas de ***accuracy*** por encima del 90% en los algoritmos de *KNN, Gradient Boosting, extreme Gradient Boosting, Árbol de decisión, Random Forest y Multilayer Perceptron* (solamente utilizado en el escenario 2) pero con métricas ***recall*** por debajo del 19% en general. Naive Bayes y Regresión logística, bajo esta instancia, generaron métricas de ***recall*** entre el 70% y el 77% con ***accuracy*** más bajos, entre el 67% y el 80%, generalizando mejor los resultados y sin afectación del problema de desbalance de clases. Al aplicar técnicas de muestreo, ***KNN y eXtreme Gradient Boosting***, mejoraron las métricas de ***recall*** por encima del 70% a coste de disminución del ***accuracy*** entre el 15% y 19%, generando un modelo congruente en resultados.

1. PROBLEMA DE INVESTIGACIÓN

1.1 Contexto, Antecedentes y Justificación

La diabetes es una enfermedad metabólica crónica con causas multifactoriales, caracterizada por niveles elevados de glucosa en sangre y en los que interactúan elementos genéticos, sociodemográficos y ambientales. Actualmente, más de 451 millones de personas tienen diabetes, es decir, el 5.64% de la población mundial, y en la Región de las Américas, 62 millones de personas la padecen, representada en el 6.1% de la población de la región (Pan American Health Organization, 2022). Se estima para el 2045, que el 6.7% de la población mundial padezca de diabetes (Villalobos, et al., 2019). La mortalidad por diabetes ha aumentado un 70% desde el año 2000, situándose entre las 10 principales causas de muerte en todo el mundo. En la región de las Américas, 284 mil personas murieron debido a la diabetes en 2019, con una tasa de mortalidad estandarizada por edad de 20.9 por cada 100 mil habitantes (Pan American Health Organization, 2022).

La Organización Mundial de la Salud (OMS) reconoce los siguientes tipos de diabetes (Gómez-Encino, Cruz-León, Zapata-Vázquez, & Morales- Ramón, 2015):

- Tipo I (DM1): autoinmune que destruye las células que producen insulina del páncreas, con más frecuencia en niños y jóvenes adultos.

- Tipo II (DM2): no insulinodependiente, resistencia a la insulina. Suele manifestarse en edad adulta, después de los 40 años.

- Gestacional: usualmente se desarrolla en la segunda mitad del embarazo con intolerancia a la glucosa.

Actualmente, no es posible prevenir la diabetes de tipo 1 con los conocimientos que se tienen. Por el contrario, existen maneras eficaces de prevenir la diabetes de tipo 2, con la

aplicación de políticas y prácticas (aplicables en las poblaciones enteras y en contextos particulares como la escuela, hogar, entorno laboral, etc.) que fomenten el bienestar de todos (Organización Mundial de la Salud, 2016).

Esta enfermedad, se concentra principalmente en países de bajos y medianos ingresos (Pan American Health Organization, 2022), tres de cada cuatro personas viven con diabetes en países con estas características (Gómez-Encino, Cruz-León, Zapata-Vázquez, & Morales- Ramón, 2015). Además, dadas sus complicaciones, generan grandes pérdidas económicas tanto para los pacientes diabéticos y sus familias, como para los sistemas de salud y economías nacionales, representados en gastos médicos en el ámbito hospitalario y ambulatorio, y aumento del costo de los análogos de la insulina (Organización Mundial de la Salud, 2016). En el sistema de salud colombiano, las personas diabéticas, representan un aumento en los costos de la atención ambulatoria del orden de 3 o 4 veces con respecto a la Unidad de Pago por Capitación (UPC). Si el paciente tiene complicaciones agudas, puede elevarse de 4 a 5 veces; pero si tiene complicaciones macrovasculares (diálisis), el incremento puede ser de 10 a 12 veces la UPC (Ministerio de Salud y Protección Social, 2021).

De acuerdo con lo anterior, para minimizar el impacto de la carga del sistema de salud, "*es importante que se realicen diagnósticos tempranos, ya que según la Organización Mundial de la Salud (OMS), alrededor del 45 % de la población no sabe que la padece, y es necesario para controlar las comorbilidades, brindar una adecuada prescripción farmacológica, de alimentación y así mejorar la calidad de vida"* (Ministerio de Salud y Protección Social, 2021).

1.2 Planteamiento del Problema

Actualmente, la mayoría de estudios de predicción de la diabetes utilizan pruebas clínicas, como el de Pérez-Gandía (2014), el cual propone algoritmos de predicción de glucosa en pacientes diabéticos por medio de monitorización continua, la utilización de Clasificadores Bayesianos para detectar personas con diabetes con tasa de acierto entre el 81.53% y 95.38% (Castrillón, Sarache, & Castaño, 2017), redes neuronales para predicción (Song, Mitnitski, Cox, & Rockwood, 2004), series de tiempo (Saria, Rajani, Gould, Koller, & Penn, 2010) y aplicación de minería de datos en bioinformática con la herramienta RapidMiner (Han, Rodriguez, & Beheshti, 2008).

Sin embargo, en el contexto latinoamericano, escenario de desigualdades, el 7% de las personas de la región tienen la enfermedad y los métodos de evaluación varían, lo que puede subestimar el número real de personas con diabetes al no tener en cuenta factores adicionales como los sociales, ambientales y políticas nacionales o locales (Avilés-Santa, Monroig-Rivera, Soto-Soto, & Lindberg, 2020). Además, acceder a los datos arrojados por las pruebas clínicas no es tarea fácil para las Instituciones Prestadoras de Salud (IPS), por los costos asociados a la toma y la distribución poblacional, como zonas rurales, de difícil acceso, hogares indígenas y de alta vulnerabilidad social (Villalobos, et al., 2019), por ejemplo, la población del régimen subsidiado en Colombia.

De acuerdo con lo anterior, surge la necesidad de revisar estudios sobre el uso de parámetros no clínicos para la generación de modelos predictivos como apoyo al método clínico. En el estudio de Mathew & Sherly (2018), obtuvieron buenos resultados (recall y precisión mayor al 80%) con la utilización de parámetros no clínicos y un modelo de clasificación Random Forest, implementado con éxito a bajo costo médico (Mathew & Sherly, 2018).

Por otro lado, se encuentra el estudio de predicción de los científicos de datos colombianos Mejía, Oviedo, Ordonez, & Valencia (2022), basado en información medioambiental y socioeconómica con la aplicación de Extreme Gradient Boosting Classifier en población del régimen subsidiado (Mejía, Oviedo, Ordonez, & Valencia, 2022), y aunque no tuvieron los

resultados esperados (recall de 48.44% y precisión 73.39%), sientan bases para explorar y mejorar las predicciones bajo estas condiciones.

1.3 Pregunta de investigación

Para concluir lo expuesto anteriormente, se sintetiza el problema mediante la formulación del siguiente interrogante:

¿Cómo podemos apoyar en la detección temprana de la diabetes mellitus tipo 2 (DMT2) mediante un modelo supervisado que involucre variables no clínicas identificadas a partir de las características de una población perteneciente al régimen subsidiado?

2. OBJETIVOS

2.1 Objetivo General

Proponer un modelo de predicción de Diabetes tipo 2 (DMT2) a partir de variables no clínicas en una población del régimen subsidiado de una EPS del Suroccidente Colombiano.

2.2 Objetivos Específicos

- Caracterizar la población con enfermedad de Diabetes, consignada en una base de datos de una EPS del suroccidente colombiano.

- Identificar variables no clínicas que puedan influir en la detección DMT2.

- Seleccionar los modelos más relevantes para la predicción de la Diabetes con las variables no clínicas identificadas en la base de datos.

- Evaluar el desempeño de los modelos de predicción propuestos, teniendo en cuenta las variables no clínicas identificadas para predecir la DMT2.

3. REVISIÓN BIBLIOGRÁFICA

3.1 Marco Teórico

3.1.1 Dominio del Problema.

- **Diabetes mellitus**

La diabetes mellitus, generalmente conocida solo como “diabetes” o “diabetes sacarina”, es un grupo de trastornos metabólicos caracterizados por la presencia de hiperglucemia si no se recibe tratamiento. Se presenta por diferentes causas y comprende deficiencias en la secreción de insulina. Las complicaciones específicas de la diabetes a largo plazo son la retinopatía, la nefropatía y la neuropatía, también corren un mayor riesgo de sufrir otros trastornos, como cardiopatías, arteriopatía periférica, afecciones cerebrovasculares, cataratas. También son más propensas a ciertas enfermedades infecciosas, como la tuberculosis, con un pronóstico más desfavorable (Organización Panamericana de la Salud, 2020).

- **Factores de Riesgo de la Diabetes**

De acuerdo con el tipo de diabetes, se presentan los siguientes factores de riesgo (Organización Mundial de la Salud, 2016):

Tipo 1. La creencia general es que este tipo de diabetes obedece a una interacción compleja entre los genes y factores ambientales, aunque no se ha demostrado que ningún factor ambiental en particular haya causado un número de casos importante. La mayoría de los casos de diabetes de tipo 1 se producen en niños y adolescentes.

Tipo 2. El riesgo de diabetes de tipo 2 (DMT2) se ve determinado por la interacción de factores genéticos y metabólicos. Dicho riesgo se eleva cuando factores étnicos, un antecedente de diabetes en la familia y un episodio anterior de diabetes gestacional se

combinan con la presencia de edad avanzada, sobrepeso y obesidad, alimentación malsana, falta de actividad física y tabaquismo.
La diabetes gestacional. Entre los factores y marcadores del riesgo de DG figuran la edad (mientras más años tiene una mujer en edad reproductiva, más alto es su riesgo de padecer DG); el sobrepeso o la obesidad; el aumento de peso excesivo durante el embarazo; la presencia de antecedentes familiares de diabetes; el haber padecido DG durante un embarazo previo; el haber tenido un hijo mortinato o con una anomalía congénita; y el exceso de glucosa en la orina durante el embarazo.

- **Variables No Clínicas**

Srivastava, Kumar, Fore, & Tomar (2021), relacionan las *variables no clínicas*, a la elimicación o dependencia correspondiente ***al análisis de sangre y exámenes que se deriven de éste***. Los autores proponen el uso de variables como la edad, grosor de la sintura, índice de masa corporal, presión arterial sistólica, duastólica, antecedentes familiares y sexo. Hajjaj, Salek, Basra, & Finlay (2010), además de mencionar las características anteriores, relacionan variables como el estado socioeconómico, raza del paciente, estilo de vida, deseos y actitud del paciente.

- **Costos médicos asociados a la Diabetes**

Entre los costos médicos directos que se asocian con la diabetes figuran los de prevenir y tratar la enfermedad y sus complicaciones. Dichos costos comprenden los de la atención ambulatoria y de urgencias; los de la atención intrahospitalaria; los de los medicamentos e insumos médicos, tales como los dispositivos de inyección y los que se utilizan para el control de la glucemia por el propio paciente; y los de la atención médica prolongada (Organización Mundial de la Salud, 2016).

- **Unidad de Pago por Capitación (UPC)**

De acuerdo con la Ley 100 de 1993, marco de la organización y funcionamiento del Sistema, establece, que se debe garantizar el acceso a los servicios de salud en todos los

niveles de atención definidos en el Plan Obligatorio de Salud (POS, en su momento; ahora financiación con recursos de la UPC), y que todo ciudadano debe participar en el Sistema General de Seguridad Social en Salud (SGSSS). Por dichos servicios se reconoce a las Entidades Promotoras en Salud (EPS) un valor de prima llamado Unidad de Pago por Capitación (UPC), que debe ser definido por el Ministerio de Salud y Protección Social.

3.1.2 Dominio de la Solución

- **Machine Learning / aprendizaje automático**

"*En esta era de la tecnología moderna, hay un recurso en abundancia: una gran cantidad de datos estructurados y no estructurados. En la segunda mitad del siglo veinte, el aprendizaje automático evolucionó como un subcampo de la Inteligencia Artificial (IA), que involucraba algoritmos de autoaprendizaje que derivaban el conocimiento a partir de los datos para crear predicciones. En lugar de requerir que los humanos deriven manualmente reglas y construyan modelos de análisis de grandes cantidades de datos, el aprendizaje automático ofrece una alternativa más eficiente para capturar el conocimiento de los datos, mejorar gradualmente el rendimiento de los modelos predictivos y realizar predicciones basadas en datos*" (Raschka & Mirjalili, 2019).

- **Tipos de aprendizaje automático**

Existen tres tipos de aprendizaje automático, se observan a continuación en la siguiente ilustración:

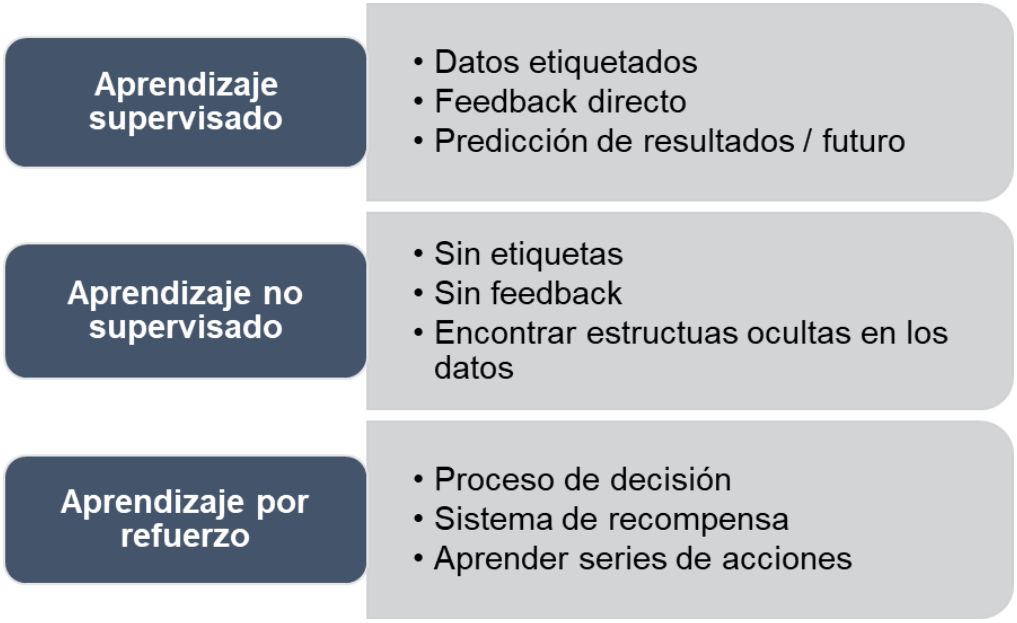

Ilustración 1. Tipos de aprendizaje automático - Machine Learning. Fuente: (Raschka & Mirjalili, 2019)

- **Predicciones con el aprendizaje supervisado**

El objetivo principal del aprendizaje supervisado es aprender un modelo, a partir de datos de entrenamiento etiquetados, que permite hacer predicciones sobre datos futuros o no vistos. El término supervisado se refiere a un conjunto de muestras donde las señales de salida deseadas (etiquetas) ya se conocen (Raschka & Mirjalili, 2019).

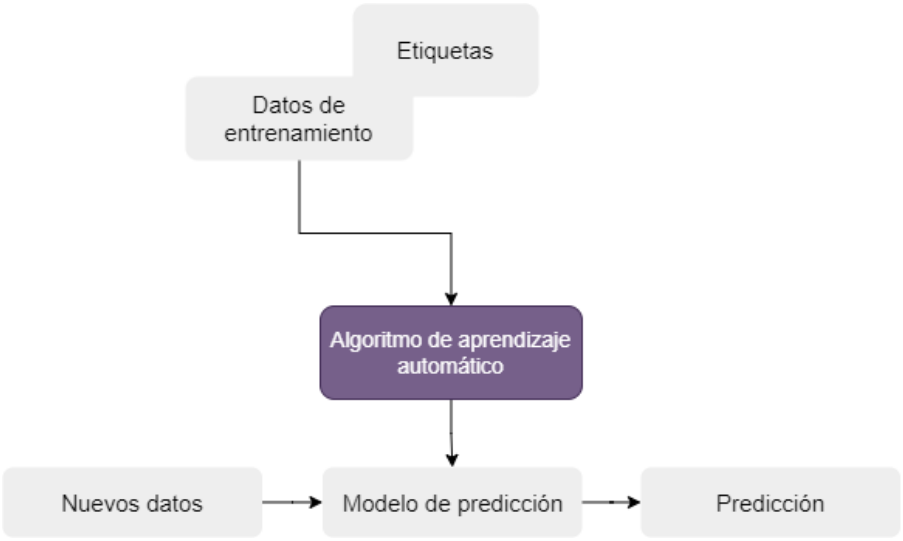

Ilustración 2. Flujo General del Aprendizaje automático supervisado. Fuente: (Raschka & Mirjalili, 2019)

De acuerdo con la ilustración anterior, una tarea de aprendizaje supervisado con etiquetas de clase discreta, también se conoce como **tarea de clasificación**.

- **Clasificación para predecir etiquetas de clase**

La clasificación es una categoría del aprendizaje supervisado cuyo objetivo es predecir las etiquetas de la clase categórica de nuevas instancias, basadas en observaciones pasadas. Estas etiquetas de clase son discretas, valores desordenados que se pueden entender como membresías grupales de las instancias. Las etiquetas de clase no tienen que ser de naturaleza binaria. El modelo predictivo aprendido mediante un algoritmo de aprendizaje supervisado puede asignar cualquier etiqueta de clase que se presente en el

conjunto de datos de entrenamiento a una nueva instancia sin etiqueta. En la siguiente ilustración se observa el concepto de una tarea de clasificación binaria (Raschka & Mirjalili, 2019).

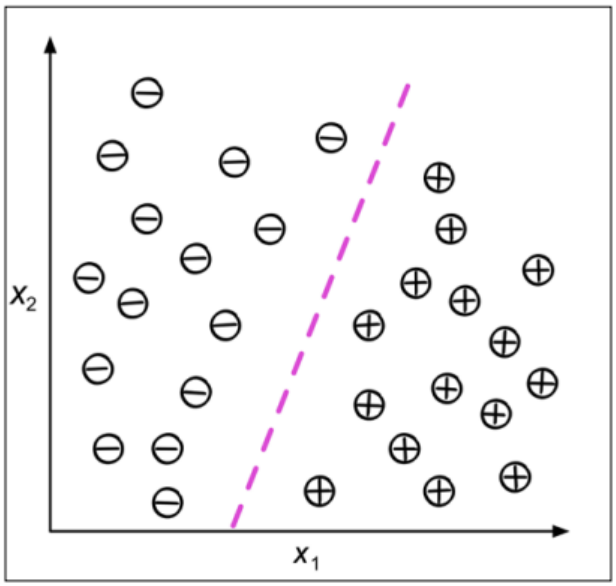

Ilustración 3. utilización de algoritmo de aprendizaje automático supervisado en una tarea de clasificación. Fuente: (Raschka & Mirjalili, 2019)

3.2 Estado del Arte

3.2.1 Trabajos seleccionados.

En la literatura, se presentan una gran cantidad de proyectos y/o estudios que han abordado el problema de la predicción de la Diabetes y de Enfermedades Crónicas no transmisibles. Principalmente, éstos se han enfocado en el uso de Machine Learning para la predicción de la enfermedad. Además, en estos estudios se han enfocado en variables clínicas y/o no clínicas como insumo de ayuda para la toma de decisiones médicas y diagnósticos de enfermedades. En este sentido, se ha realizado un Top de proyectos, los cuales se han considerado que serán de aporte al proyecto de grado que se pretende realizar.

3.2.1.1 Diabetes Detection and Prediction Using Machine Learning/IoT: A Survey, (Sharma & Singh, 2019)

Este Proyecto explora el uso de modelos de machine learning aplicado a variables clínicas entre las que se encuentran el nivel de glucosa, presión arterial e insulina. Además, explora el uso de tecnologías IoT (Internet of Things) para un monitoreo de variables que pueden ser afectadas por los hábitos del paciente, y aumentar el riesgo de padecer diabetes. Entre los modelos expuestos con mejores resultados se encuentran Support Vector Machine, K-Nearest Neighbor, J48 algorithm y las Redes neuronales. Con una precisión de 98% para el Support Vector Machine, y un 94.44% para las Redes Neuronales.

3.2.1.2 Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice (Hajjaj, Salek, Basra, & Finlay, 2010)

El objetivo de este estudio consistió en revisar la influencia de variables no clínicas en la toma de decisiones médicas. Estas variables no clínicas incluyen factores como el Nivel socioeconómico, calidad de vida, expectativas y deseos del paciente, características físicas,

y características profesionales. Es importante resaltar que, de estas variables no clínicas, principalmente las preferencias del paciente y su estado económico tienen un gran potencial para perjudicar la gestión de los pacientes en la prevención de una enfermedad, debido a que el paciente decida no seguir las recomendaciones médicas debido a desinterés, o por el impacto que implicaría en sus finanzas personales o familiares.

3.2.1.3 Analysis of Supervised Learning Techniques for Cost Effective Disease Prediction Using Non-Clinical Parameters (Mathew & Sherly, 2018)

El objetivo de este proyecto consistió en utilizar variables no clínicas para la predicción de enfermedades como la diabetes para mejorar el costo-efectividad de la gestión de enfermedades, ya que, al contar con un primer pronóstico temprano, se le puede brindar un manejo al paciente que reduzca el riesgo de complicaciones. En el estudio, aplican los modelos de Random forest, Naive Bayes, Regresión logística y Reduced error pruning tree, donde el mejor resultado lo obtuvo con Random Forest, con las métricas de evaluación recall y precision mayor al 80%, seguido por Naive Bayes con precision del 80.37%. Para estudios futuros, los autores sugieren utilizar una versión con pesos del modelo de Naive Bayes, para obtener mejores resultados.

3.2.1.4 Low-Cost Method for Multiple Disease Prediction (Bayati, Bhaskar, & Montanari, 2015)

El Proyecto consiste en la detección temprana de diversas enfermedades, entre ellas la diabetes, a través de métodos estadísticos, con el fin de evitar los altos costos asociados a las complicaciones de estas enfermedades mediante la gestión de estos pacientes en etapas tempranas de la enfermedad. Utiliza modelos como el LR (Logistic Regression), STL (Single Task Learning Model), MTL (Multitask Learning Model) y el OLR-M (que deriva del MTL). En los resultados, los modelos STL, MTL y OLR-M tuvieron un rendimiento similar, pero el modelo OLR-M es el menos costoso de implementar debido a que necesita menos información para llegar a buenos resultados.

3.2.1.5 A Study of Five Models Based on Non-clinical Data for the Prediction of Diabetes Onset in Medically Under-Served Populations (Srivastava, Kumar, Fore, & Tomar, 2021)

Este estudio consiste en la comparación de modelos de Machine Learning para identificar una estrategia de minería de datos que realice una mejor detección de la diabetes utilizando variables no clínicas, debido a los altos costos de los exámenes de sangre que actualmente se utilizan para su diagnóstico. En el proyecto se tienen en cuenta variables como la Edad, Grosor de la cintura, IMC (Índice de masa corporal), Presión arterial sistólica, Presión arterial diastólica, Antecedentes familiares y Género. Estas variables se usan para comparar el rendimiento de modelos como Decision tree, Multilayer Perceptron (Red neuronal), Bagging, Simple logistics y Support vector machine. De acuerdo con lo anterior, los mejores resultados los tuvo el modelo Bagging con una precisión de 88.56% y Multilayer perceptron con una precisión del 88.21%.

3.2.1.6 Prediction of Diabetes based on environmental and socioeconomic information (Mejía, Oviedo, Ordonez, & Valencia, 2022)

Este proyecto consiste en la predicción de la diabetes, de una población colombiana asegurada perteneciente a una EPS, con la utilización de variables ambientales y socioeconómicas como alternativas a las pruebas clínicas. El proyecto usa información medioambiental y socioeconómica con la aplicación de 3 modelos de machine learning, Voting classifier, Gradient Boosting, y Extreme Gradient Boosting, donde el mejor resultado se obtuvo con el Extreme Gradient Boosting Classifier. Aunque no tuvieron los resultados esperados (recall de 48.44% y precisión 73.39%), sientan las bases para realizar estudios complementarios y mejorar las predicciones.

3.2.2 Matriz de comparación.

	7.1.1 (Sharma, N., Singh, A, 2019)	**7.1.2** (Hajjaj, F. M. et al., 2010)	**7.1.3** (Mathew & Sherly, 2018)	**7.1.4** (Bayati M. et al., 2015)	**7.1.5** (Srivastava, R. et al., 2021)	**7.1.6** (Mejía et al., 2022)	**Proyecto de Grado**
Año de publicación	2019	2010	2018	2015	2021	2022	-
Enfermedades en la que se enfoca	Diabetes	General	Diabetes	Enfermedades crónicas	Diabetes	Diabetes	Diabetes
País	India	Reino Unido	India	EE.UU	India	Colombia	Colombia
Tipo de variables	Clínicas	No Clínicas	No Clínicas	Biométricas / Clínicas	No Clínicas	No Clínicas	No Clínicas
Origen de los datos	Privado / Público	-	Privado	Privado / Publico	Privado	Privado	Privado
Modelos propuestos	SVM KNN J48 CNN	-	Random forest Naive Baye Regresión logística Reduced error pruning tree	LR STL MTL OLR-M	Decision tree CNN Bagging Simple logistics SVM	Voting classifier Gradient Boosting Extreme Gradient Boosting	Por definir

Tabla 1. Resumen de los criterios de comparación entre los artículos seleccionados y el proyecto de grado

3.2.3 Conclusiones del estado del arte.

De acuerdo con los proyectos presentados se puede concluir que: a) se ha realizado un esfuerzo importante para encontrar factores de predicción temprana de la diabetes que sirvan de apoyo al concepto médico. b) Los modelos de machine learning son los llamados a utilizar para el problema de predicción y clasificación de las enfermedades. c) Es posible la utilización de variables no clínicas que sirvan de insumo a los modelos para la predicción, por ende, el reto está en definir las variables críticas de acuerdo con la información poblacional con la que se cuente. d) Aunque existen muchos trabajos en los cuales se han empezado a utilizar variables no clínicas para decisiones médicas, las diferencias se encuentran principalmente en el procesamiento de datos, algo que se deberá tener en cuenta definitivamente en el proyecto que se abordará antes de abordar la mejor técnica de predicción.

3.3 Modelos Predictivos / Clasificación

De acuerdo con los trabajos relacionados en el estado del arte, tomados como base para el desarrollo del presente trabajo de grado, se seleccionaron una serie de modelos de referencia, los cuales cumplen con características y están alineados con el objetivo general planteado en el desarrollo del presente proyecto.

3.3.1 KNN – K-Nearest-Neighbor

El algoritmo KNN es un enfoque no paramétrico utilizado para el problema de clasificación. El enfoque sobre la información de sus puntos vecinos para la clasificación de las etiquetas de salida. En 1951, se introduce el algoritmo KNN para la aplicación de problemas de clasificación y reconocimiento de patrones. KNN tiene la capacidad de predecir la clase objetivo con mayor precisión de una manera más sencilla, además, pertenece al tipo de algoritmo de aprendizaje "*perezoso*" cuya función se estima solo localmente y el cálculo completo se "*atrasa"* hasta el proceso de clasificación (Rajaguru & Chakravarthy, 2019).

3.3.2 Árboles de Decisión – Decision Tree

La metodología del Árbol de Decisión es un método de minería de datos de uso común para establecer sistemas de clasificación basados en múltiples covariables o para desarrollar algoritmos de predicción para una variable objetivo. Clasifica una población en segmentos similares por medio de ramas, que construyen un árbol invertido con un nodo raíz, nodos internos y nodos hoja. El algoritmo no es paramétrico y maneja eficientemente conjuntos de datos grandes y complejos sin imponer una estructura paramétrica compleja (Song & Lu, 2015).

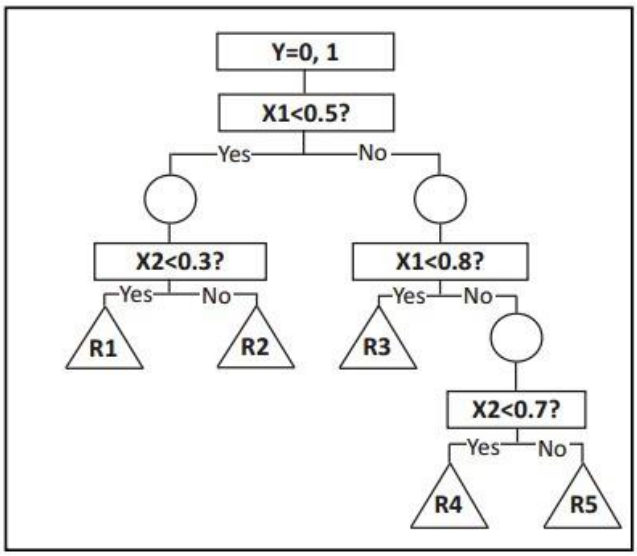

Ilustración 4. Ejemplo de Árbol de Decisión con variable objetivo binaria Y. Fuente: (Song & Lu, 2015)

El método del árbol de decisión es una poderosa herramienta estadística para la clasificación, predicción, interpretación y manipulación de datos que tiene varias aplicaciones potenciales en la investigación médica. El uso de modelos de árboles de decisión para describir los resultados de la investigación tiene las siguientes ventajas (Song & Lu, 2015):

- Simplifica las relaciones complejas entre las variables de entrada y las variables objetivo al dividir las variables de entrada originales en subgrupos significativos.
- Fácil de entender e interpretar.
- Enfoque no paramétrico sin supuestos distributivos.
- Fácil manejo de valores perdidos sin necesidad de recurrir a la imputación.
- Fácil de manejar datos pesados sesgados sin necesidad de recurrir a la transformación de datos.
- Robusto frente a valores atípicos.

3.3.3 Bosques Aleatorios – Random Forest

El clasificador Random Forest toma un vector de entrada y hace crecer muchos árboles de clasificación a partir de él. Cada árbol vota por una determinada clasificación y se elige el que obtiene la mayoría de los votos. Puede tolerar grandes conjuntos de datos y grandes conjuntos de parámetros de entrada. Este clasificador también puede averiguar la importancia de las variables (Mathew & Sherly, 2018). Son una herramienta eficaz en la predicción. Además, debido a la ley de los grandes números, no se sobreajustan a medida que se agregan más árboles (overfitting) (Breiman, 2001).

3.3.4 Naive Bayes

El algoritmo de clasificación Naive Bayes es usado ampliamente en el análisis de big data y otros campos debido a su estructura simple y rápida (Hong Chen, Songhua Hu, Rui Hua, & Xiuju Zhao, 2021). Se ha convertido en uno de los métodos más atractivos debido a su forma única de expresión de conocimiento incierto, su capacidad en términos de expresar probabilidades y poseer características de aprendizaje incremental, además de eficiencia computacional. La ventaja de este algoritmo es que sólo necesita estimar los parámetros como media y varianza de las variables, en función de una pequeña cantidad de datos de entrenamiento. Dada la suposición de variables independientes, no necesita la utilización de toda la matriz de covarianza (Hong Chen, Songhua Hu, Rui Hua, & Xiuju Zhao, 2021). Sin embargo, en la práctica, el supuesto de independencia de atributos puede ser violado, como resultado, sus estimaciones de probabilidad a menudo se ven afectadas, por lo tanto, se realiza selección aleatoria de los diferentes parámetros para identificar una combinación adecuada (Mathew & Sherly, 2018).

De acuerdo con las propiedades mencionadas, el Clasificador Naive Bayes, tiene amplia gama de aplicaciones, como en medicina clínica, telecomunicaciones, inteligencia artificial, lingüística, tecnología genética, instrumentos de precisión y otros campos (Hong Chen, Songhua Hu, Rui Hua, & Xiuju Zhao, 2021).

3.3.5 Regresión Logística

La regresión logística es un método estadístico en forma de algoritmo de clasificación. Analiza un conjunto de datos con variables independientes para determinar un resultado. El resultado se mide con solo dos resultados posibles, 1 y 0. La regresión logística explica las relaciones entre una variable dependiente y una o más variables independientes nominales, ordinales o categóricas. Tiene aplicaciones en el análisis predictivo (Mathew & Sherly, 2018).

3.3.6 Gradient Boosting

Pertenece a la familia de algoritmos usados tanto en clasificación con en regresión, se basan en la combinación de modelos predictivos débiles, por lo general árboles de decisión, entrenados de forma secuencial. La idea principal detrás de este algoritmo es construir los *nuevos aprendices* base para que se correlacionen al máximo con el gradiente negativo de la función de pérdida, asociado con todo el conjunto. Las funciones de pérdida aplicadas pueden ser arbitrarias. Si la función de error es la clásica pérdida de error cuadrático, el procedimiento de aprendizaje daría como resultado errores de ajuste consecutivos (Natekin & Knoll, 2013).

3.3.7 eXtreme Gradient Boosting

Xgboost es una de las implementaciones de máquinas de aumento de gradiente (Gradient Boosting), que se conoce como uno de los algoritmos de mejor rendimiento utilizados para el aprendizaje supervisado. Se puede utilizar tanto para problemas de regresión como de clasificación. Xgboost preferido por los científicos de datos debido a su alta velocidad de ejecución fuera del cómputo central (Ahmed Osman, Ahmed, Chow, & Huang, 2021). Es fácil de implementar y entrenar, lo que significa que a medida que haya más datos disponibles, mejorará el rendimiento predictivo. Se ha utilizado en campo de Salud en la clasificación de pacientes con Covid, teniendo mejor rendimiento que el algoritmo Random Forest (Ramón, y otros, 2022).

3.3.8 Multi Layer Perceptron

Los *perceptrones multicapa o MLP*, son un miembro poderoso de la familia de redes neuronales artificiales usados para resolver problemas complejos que un solo perceptrón no puede. Los MLP son complejos y tienen una colección de perceptrones individuales interconectados, también conocidos como neuronas o nodos, que trabajan juntos para procesar y analizar datos. Propuesta por Goeffrey Hinton en la década de 1980, Los perceptrones multicapa son ese tipo de ***red Feed-forward***, en la que los datos se transmiten solo en una dirección. A diferencia de otras redes como las redes neuronales recurrentes, donde los datos pasan en ambas direcciones y forman un ciclo (Gardner & Dorling, 1998).

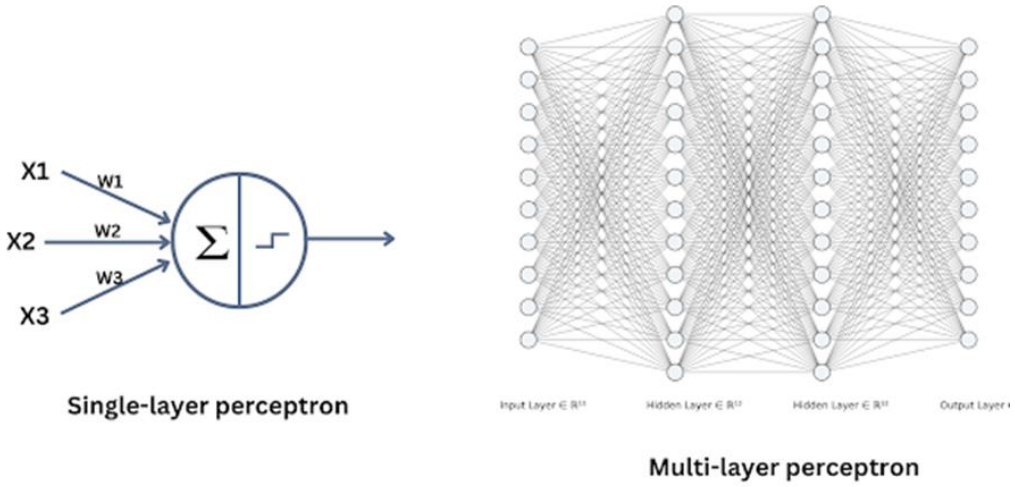

Ilustración 5. Perceptrones de una capa frente a multicapa. Fuente: https://www.pycodemates.com/2023/01/multi-layer-perceptron-a-complete-overview.html

4. METODOLOGÍA

La metodología elegida para desarrollar el presente proyecto es CRISP-DM (Cross-Industry Standard Process for Data Mining), desarrollada en 1996 por un consorcio de empresas conformado por DaimlerChrysler (entonces Daimler-Benz), SPSS Inc. (entonces ISL) y NCR Systems Engineering Copenhagen, y publicada en el año 2000 (Chapman, et al., 2000).

La metodología CRISP-DM, se describe en términos de un modelo jerárquico, que consta de conjunto de tareas descritas en cuatro niveles de abstracción (general y específico) (Chapman, et al., 2000). La siguiente ilustración muestra el desglose de los niveles y le modelo de proceso para minería de datos:

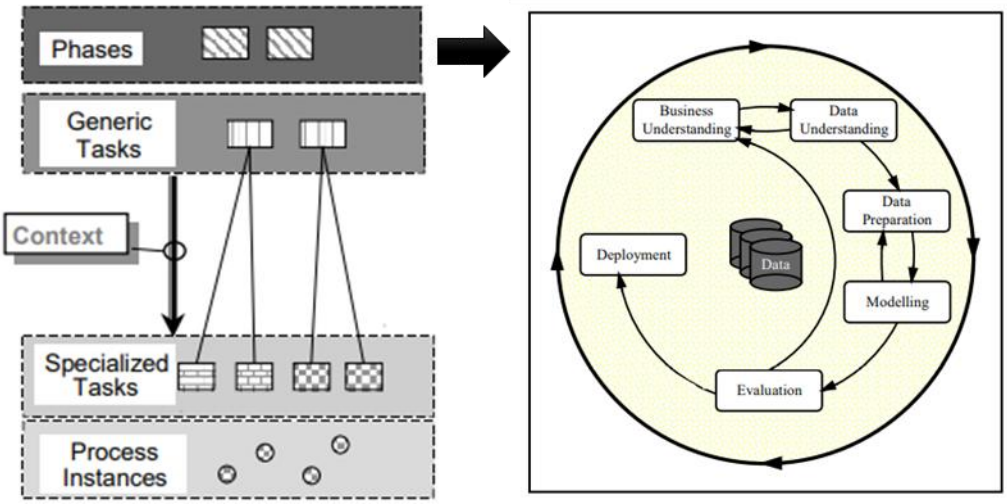

Ilustración 6. Fases del Modelo CRISP-DM. Fuente: Wirth, 2000

En el nivel superior, el proceso de minería de datos se organiza en varias fases; cada fase consta de varias tareas genéricas de segundo nivel. Este segundo nivel se llama genérico, porque pretende ser lo suficientemente general para cubrir todas las situaciones posibles de minería de datos. Las tareas genéricas pretenden ser lo más completas y estables posible. Completo significa que cubre tanto el proceso completo de minería de datos como

todas las posibles aplicaciones de minería de datos. Estable significa que el modelo debería ser válido para desarrollos aún imprevistos, como nuevas técnicas de modelado (Chapman, et al., 2000).

El ciclo de vida de un proyecto de minería de datos se divide en seis fases que se muestran en la figura derecha de la ilustración 4. La secuencia de las fases no es estricta. Las flechas indican solo las dependencias más importantes y frecuentes entre las fases, pero en un proyecto en particular, depende del resultado de cada fase qué fase, o qué tarea particular de una fase, debe realizarse a continuación (Chapman, et al., 2000). Se describe brevemente a continuación:

- **Fase 1: Comprensión del negocio**

Fase inicial del ciclo de vida de un proyecto de analítica. En esta fase se busca conocer los requerimientos y objetivos desde el punto de vista del negocio, con el fin de realizar la formulación del problema a resolver y desarrollar un plan.

- **Fase 2: Comprensión de los datos**

Se busca la familiarización con los datos que van a ser objeto de análisis, por medio de una adquisición inicial de datos, la revisión de calidad y posibles hallazgos que permitan iniciar con una declaración de hipótesis para la información oculta. Existe un vínculo estrecho entre la comprensión empresarial y la comprensión de datos.

- **Fase 3: Preparación de los datos**

Se busca trabajar con los datos crudos hasta convertirlos en datos que puedan ser leídos por los modelos. Incluye todas las actividades de selección, limpieza y transformación de datos.

- **Fase 4: Modelado**

Durante esta etapa se seleccionan y ejecutan diversas técnicas de modelado sobre los datos preparados. Adicionalmente, los parámetros de cada modelo son calibrados para obtener valores óptimos. Es probable que, dependiendo de los modelos, los datos deban tener una preparación específica.

- **Fase 5: Evaluación**

En esta fase, se busca la construcción y evaluación de modelos con el fin de validar si cumplen con los objetivos del negocio. Es importante evaluar a fondo el modelo y revisar los pasos ejecutados para construirlo, asegurando que éste logre cumplir con los objetivos del negocio.

- **Fase 6: Despliegue**

La creación del modelo generalmente no es el final del proyecto. Por lo general, el conocimiento adquirido deberá organizarse y presentarse de manera que el cliente pueda utilizarlo. Según los requisitos, la fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de extracción de datos repetible. En muchos casos será el usuario, quien llevará a cabo los pasos de despliegue. En cualquier caso, es importante comprender de antemano qué acciones deberán llevarse a cabo para poder utilizar realmente los modelos creados.

5. PRESENTACIÓN DEL TRABAJO DE INVESTIGACIÓN (METODOLOGÍA PROPUESTA)

Este trabajo de grado tiene como finalidad proponer un modelo de Machine Learning que permita realizar la predicción de diabetes mellitus tipo 2 (DMT2) a partir de variables analizadas y seleccionadas desde el histórico de caracterización poblacional suministrado por la EPS. La metodología presentada a continuación, muestra la descripción de los datos y sus fuentes, el sustento para la selección de variables, los algoritmos de aprendizaje utilizados para el entrenamiento, métricas de evaluación a implementar y los hiperparámetros de los modelos.

5.1 Entendimiento de los datos

5.1.1 Recolección y descripción de los datos

La EPS, dentro de su enfoque de aseguramiento orientado a la política de atención integral en salud, tiene constituido el proceso de Gestión del Riesgo en Salud, en la cual, junto con las Instituciones Prestadoras de Salud (IPS) pertenecientes a su red primaria, en algunos municipios, realizaron una encuesta denominada *Caracterización Poblacional*. En esta encuesta, está consignada información asociada a los usuarios asegurados con respecto al modo, hábitos y condiciones de vida, antecedentes familiares, nivel socioeconómico y condiciones ambientales. La consolidación de las encuestas y homologación de la información contenida fue realizada por el equipo de trabajo de la EPS. Adicionalmente, dado el seguimiento que le realizan a las diferentes cohortes o grupos de riesgo, identificaron en la base de datos consolidada, los usuarios que tienen diagnóstico confirmado DMT2 (Variable *AC_DIABETES=1*). Esta etiqueta es la variable de respuesta con la que se trabajó. Por último, la entidad mediante una hoja de cálculo de ***Google Drive***, suministraron la respectiva base de datos para el desarrollo del trabajo de grado. La base de datos entregada por la entidad comprende información recolectada de usuarios únicos en los años 2018, 2019, 2020 y 2021 respectivamente. Además, tiene la siguiente estructura general (**Tabla 1**):

Tabla 1. Estructura inicial base de datos EPS. Fuente: Elaboración Propia

Características	Valores
Nombre base de datos	Datos_Consolidados_ConVarObjetivo_sinDatosSensibles.xlsx
Cantidad registros	130.166
Cantidad variables	132
Periodos de información	2018-2019-2020-2021
Cantidad de variables cuantitativas	6
Cantidad de variables cualitativas	123
Cantidad de variables tipo fecha	3

De acuerdo con la tabla anterior, en la base de datos inicial, primó las variables cualitativas (94%). La descripción de las 132 variables se muestra en el **Anexo 1**.

5.1.2 Análisis Exploratorio de los datos

Se realizó el proceso de análisis exploratorio de los datos, con el fin de comprender su contenido. Se identificaron varios aspectos para tener en cuenta en la depuración de ésta. El primer análisis se realizó con la herramienta de minería de datos IBM SPSS Modeler 18.3[1], la cual, permite crear rutas de datos, manejar valores perdidos, leer, procesar y dar salida datos en diferentes formatos, y desarrollar modelos predictivos de una manera eficiente.

5.1.2.1 Usuarios no afiliados a la EPS

De acuerdo con la revisión de la base de datos inicial, se identificaron usuarios encuestados que no están afiliados a la EPS, por ende, se excluyeron del desarrollo del trabajo de grado.

Tabla 2. Caracterización de las personas afiliadas y no afiliadas que realizaron la encuesta. Fuente: Elaboración Propia

id_afiliado	No. Personas
Afiliado	128.504
No Afiliado	1.662
Total	130.166

[1] https://www.ibm.com/docs/es/SS3RA7_18.3.0/pdf/ModelerApplications.pdf

Con la exclusión de los usuarios no afiliados, la base de datos se reduce a 128.504 afiliados.

5.1.2.2 Caracterización de la población objeto de estudio

El alcance del trabajo de grado se situó en la población del suroccidente colombiano (Valle del Cauca, Nariño y Cauca), por ende, se excluyó la población de las demás regiones donde tiene cobertura la EPS.

Tabla 3. Distribución por departamentos de los usuarios afiliados encuestados. Fuente: Elaboración propia

Departamento	No. Afiliados
Cauca	59.152
Caquetá	24.394
Nariño	9.555
Risaralda	7.305
Valle Del Cauca	6.738
Quindío	6.672
Cesar	4.656
Huila	4.502
Caldas	4.028
Santander	1.502
Total	128.504

Departamento	No. Afiliados
Cauca	59.152
Nariño	9.555
Valle Del Cauca	6.738
Total	75.445

Como se mencionó anteriormente, dado el alcance del trabajo de grado, se extrae el 59% de los usuarios afiliados, es decir, 75.445 afiliados.

- **Perfil poblacional de la enfermedad**

En el informe del Fondo Colombiano de Enfermedades de Alto Costo, emitido el 14 de noviembre del 2022, sobre el *día mundial de la Diabetes*, en su caracterización poblacional, el 77.70% de las personas con diagnóstico de diabetes mellitus en Colombia se encuentran en edades de 55 y más años (Fondo Colombiano de Enfermedades de Alto Costo, 2022), es decir, que la enfermedad se concentra en adultos, principalmente en el género femenino, como se muestra a continuación:

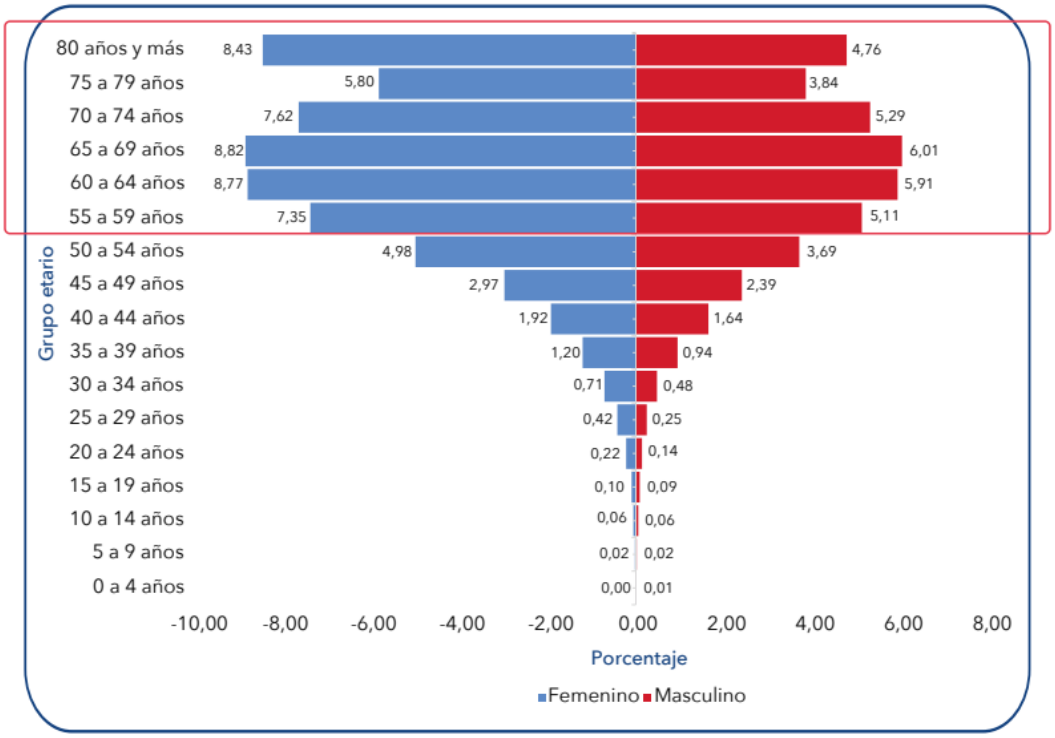

Ilustración 7. Pirámide poblacional de los casos prevalentes de diabetes mellitus, Colombia 2021. Fuente: (Fondo Colombiano de Enfermedades de Alto Costo, 2022)

De acuerdo con la pirámide poblacional de prevalencia de diabetes (***Ilustración 5***), el 95.3% de las personas diagnosticadas, están entre la edad de los 40 y más años (Fondo Colombiano de Enfermedades de Alto Costo, 2022). En contraste con la información suministrada por la EPS, se realizó la caracterización por grupo etario de los usuarios diagnosticados con diabetes (1.913 usuarios):

Tabla 4. Caracterización por grupo etario de los afiliados diagnosticados con diabetes. Fuente: Elaboración propia

Grupo Etario	No. Afiliados Diagnosticados	% frente al total diagnosticados	
80 y más años	241,0	12,6%	
75-79 años	174,0	9,1%	
70-74 años	195,0	10,2%	
65-69 años	259,0	13,5%	
60-64 años	307,0	16,0%	**95,7%**
55-59 años	244,0	12,8%	
50-54 años	199,0	10,4%	
45-49 años	121,0	6,3%	
40-44 años	91,0	4,8%	
35-39 años	28,0	1%	
30-34 años	27,0	1%	
25-29 años	21,0	1%	
20-24 años	3,0	0%	
15-19 años	2,0	0%	
10-14 años	1,0	0%	
Total	1913,0	100%	

Con base en la información presentada sobre las edades de prevalencia de la enfermedad, en el que se contrasta la información nacional de acuerdo a la Cuenta de Alto Costo versus la información de la EPS, caracterizando el suroccidente colombiano, podemos concluir inicialmente, que la muestra representa el comportamiento nacional, dado que de acuerdo a los resultados descritos en la Tabla 4, el 95.7% de los afiliados con diagnóstico de diabetes están entre las edades de 40 años y más (versus el 95.3% del país), además, que el 74.2% de los usuarios enfermos están entre 75 años y más (versus el 77.70% del país). Por tal motivo, con base en este análisis y con estudios previos realizados por autores como *Leiva, y otros – 2018* (Leiva, y otros, 2018), en el que identificaron que *"las personas que tienen mayor riesgo de DMT2 son aquellas que tienen una edad superior a 45 años*". Para efectos del presente trabajo de grado, se decidió tomar la población de más de 40 años.

Tabla 5. Caracterización de la población objeto de estudio por grupo etario. Fuente: Elaboración propia

Grupo Etario	No. Afiliados sin DMT2	No. Afiliados con DMT2	Total General
40-44 años	5.000	91	5.091
45-49 años	4.191	121	4.312
50-54 años	3.779	199	3.978
55-59 años	3.343	244	3.587
60-64 años	3.082	307	3.389
65-69 años	2.407	259	2.666
70-74 años	1.866	195	2.061
75-79 años	1.348	174	1.522
80 y más años	2.271	240	2.511
Total	**27.287**	**1.830**	**29.117**
% de afiliados dignosticados frente al total		**6,3%**	

Con la segmentación observada previamente, la población objeto con la que se trabajó el presente proyecto, fue de 29.117 afiliados, de los cuales, 1.830 afiliados tienen la etiqueta de DMT2, es decir, el 6.3%. De acuerdo con lo anterior, desde el punto de vista de línea base de la etiqueta “AC_DIABETES”, se evidencia un desbalance de clases entre la población enferma y sana.

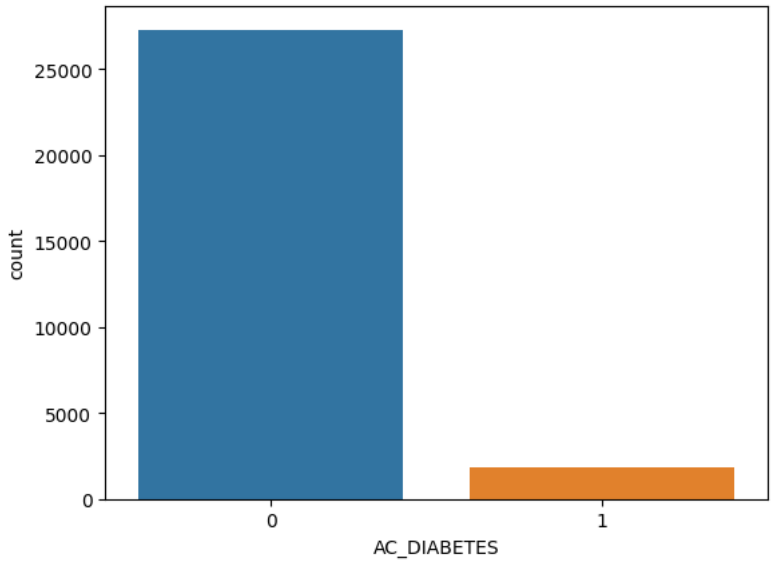

Ilustración 8. Desbalance de clases entre población con DMT2 y Sana. Fuente: Elaboración propia.

- **Distribución poblacional por Género**

De acuerdo con la población objeto del proyecto, se observa a continuación, que el 57% corresponde al género “femenino” y el 43% al “masculino”. Adicionalmente, de la población con DMT2, el 71% corresponde al género “femenino” y el 29% al “masculino”, lo cual concuerda con lo mencionado en la sección de **perfil poblacional de la enfermedad**.

Tabla 6. Distribución poblacional de la población por género. Fuente: Elaboración propia.

AC_DIABETES	Femenino	Masculino	Total por Clase
Sin DMT2	15.176	12.111	**27.287**
Con DMT2	1.306	524	**1.830**
Total por Género	**16.482**	**12.635**	**29.117**
% Part. por género total	**57%**	**43%**	
% Part. Por género con DMT2	**71%**	**29%**	

- **Distribución poblacional por Estrato**

Desde la caracterización por estrato, la mayor proporción de población que está bajo el aseguramiento de la EPS se encuentra en el *estrato 1*. Así mismo, la población con DMT2 se concentra en este estrato. De acuerdo con lo anterior, dentro del alcance del estudio, la prevalencia de la enfermedad está en población vulnerable.

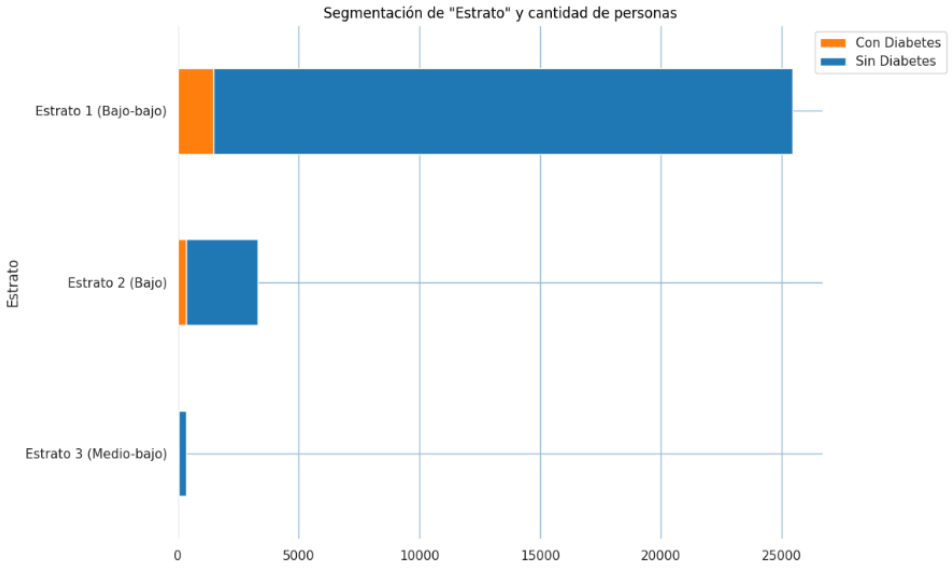

Ilustración 9. Segmentación de la población por estrato y concentración de la enfermedad. Fuente: Elaboración propia.

- **Distribución poblacional por Ingresos**

Desde la caracterización por ingreso, la mayor proporción de población con DMT2 tiene ingresos inferiores a 1 SMLV.

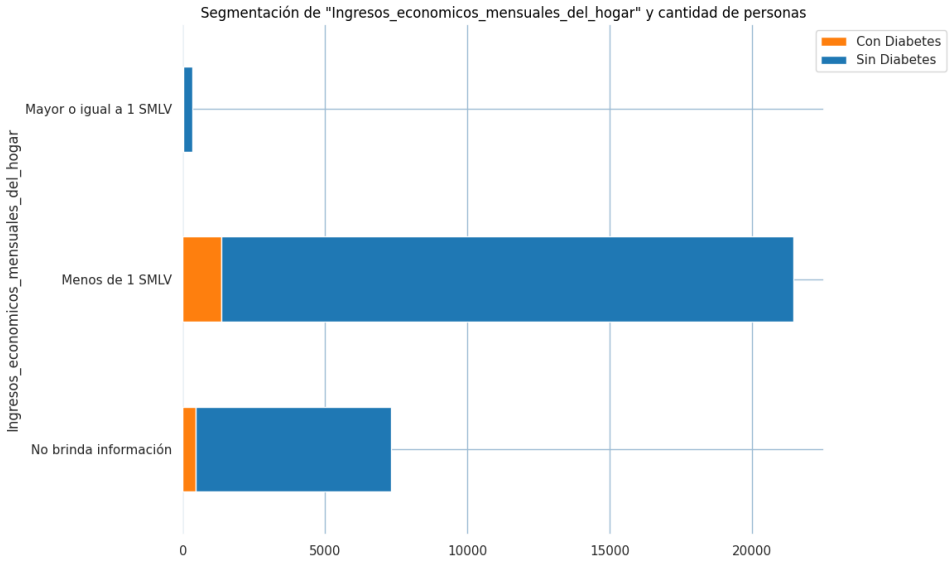

Ilustración 10. Segmentación de la población por ingreso y concentración de la enfermedad. Fuente: Elaboración propia.

- **Distribución poblacional por nivel de estudio**

Desde la caracterización por nivel de estudio, la mayor proporción de población con DMT2, tienen estudio entre básica primaria y secundaria.

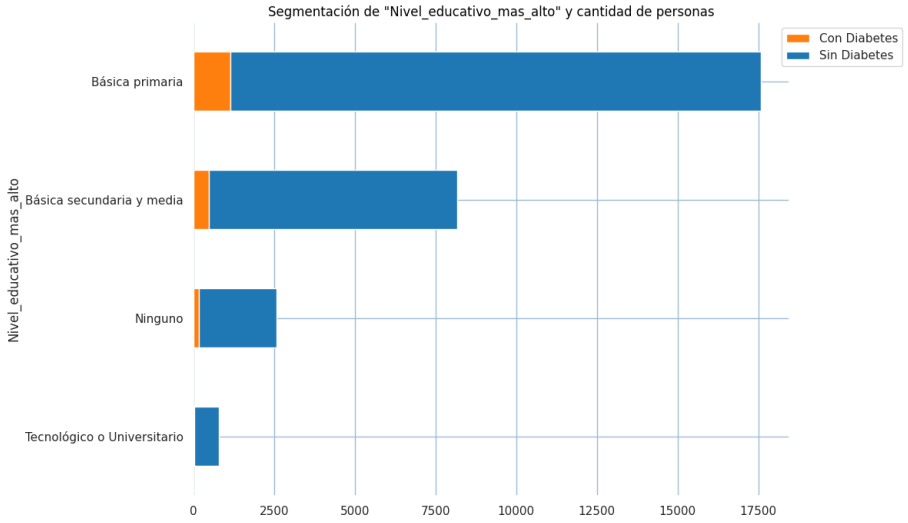

Ilustración 11. Segmentación de la población por nivel de estudio y concentración de la enfermedad. Fuente: Elaboración propia.

- **Distribución poblacional por frecuencia de consumo de fruta**

Desde la caracterización por frecuencia de consumo de fruta, la mayor proporción de población con DMT2 se distribuyen en aquellos que consumen frutas.

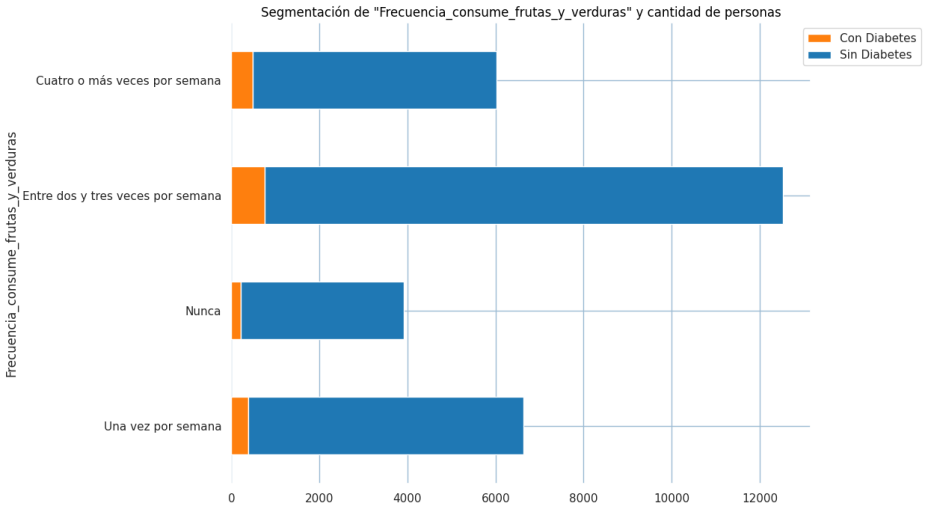

Ilustración 12. Segmentación de la población por consumo de frutas y concentración de la enfermedad. Fuente: Elaboración propia.

- **Distribución poblacional de acuerdo con consanguinidad**

Desde la caracterización en referencia con familiares que tienen diabetes mellitus, la mayor proporción (30%) está en aquellos afiliados con DMT2 que tienen familiares con diabetes. Prácticamente duplica en comparación con afiliados sanos que tienen familiares con diabetes.

Tabla 7. Influencia de consanguinidad con respecto a los afiliados que tienen o no DMT2. Fuente: Elaboración propia.

AC_DIABETES	Familiar sin DM	Familiar con DM	Total por Clase	% part. Familiar con DM
Sin DMT2	23.050	4.237	**27.287**	**16%**
Con DMT2	1.273	557	**1.830**	**30%**
Total por Consanguinidad	**24.323**	**4.794**	**29.117**	

- **Distribución poblacional por influencia de Hipertensión Arterial**

Desde la caracterización de los afiliados con DMT2 frente a otra enfermedad de base como la Hipertensión, se observa que aquellos que tienen Hipertensión, el 22% tienen DMT2. De acuerdo con lo anterior, este segmento representa el 68% del total de afiliados con DMT2.

Tabla 8. Influencia de Hipertensión con respecto a los afiliados que tienen o no DMT2. Fuente: Elaboración propia.

AC_DIABETES	Sin Hipertensión	Con Hipertensión	Total por Clase	% part. Familiar con DM
Sin DMT2	22.869	4.418	**27.287**	**16%**
Con DMT2	591	1.239	**1.830**	**68%**
Total	**23.460**	**5.657**	**29.117**	
% de afiliados con DMT2 de acuerdo a categoría	**3%**	**22%**		

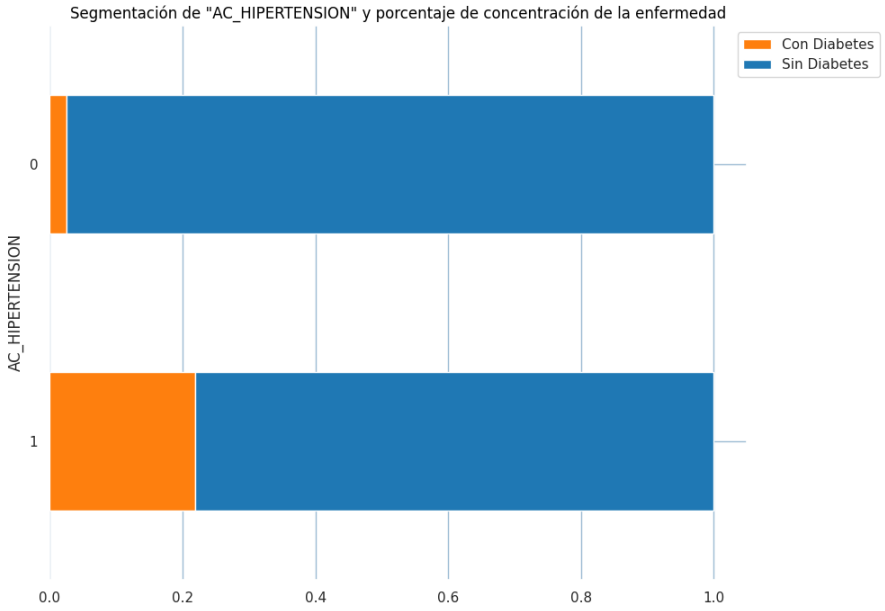

Ilustración 13. Segmentación de la población contrastando la hipertensión versus la concentración de DMT2. Fuente: Elaboración propia.

- **Distribución poblacional por Grupo Etario**

Desde la caracterización por grupo etario, la mayor proporción de población con DMT2 se distribuyen después de los 60 años.

Tabla 9. Segmentación de la población por grupo etario y concentración de la enfermedad. Fuente: Elaboración propia.

AC_DIABETES	40-44 años	45-49 años	50-54 años	55-59 años	60-64 años	65-69 años	70-74 años	75-79 años	80 y más años	Total por Clase
Sin DMT2	5.000	4.190	3.779	3.343	3.082	2.407	1.865	1.347	2.274	**27.287**
Con DMT2	91	121	199	244	307	259	195	174	240	**1.830**
Total por Grupo Etario	**5.091**	**4.311**	**3.978**	**3.587**	**3.389**	**2.666**	**2.060**	**1.521**	**2.514**	**29.117**
% part. de afiliados con DMT2 por cada categoría	**2%**	**3%**	**5%**	**7%**	**9%**	**10%**	**9%**	**11%**	**10%**	

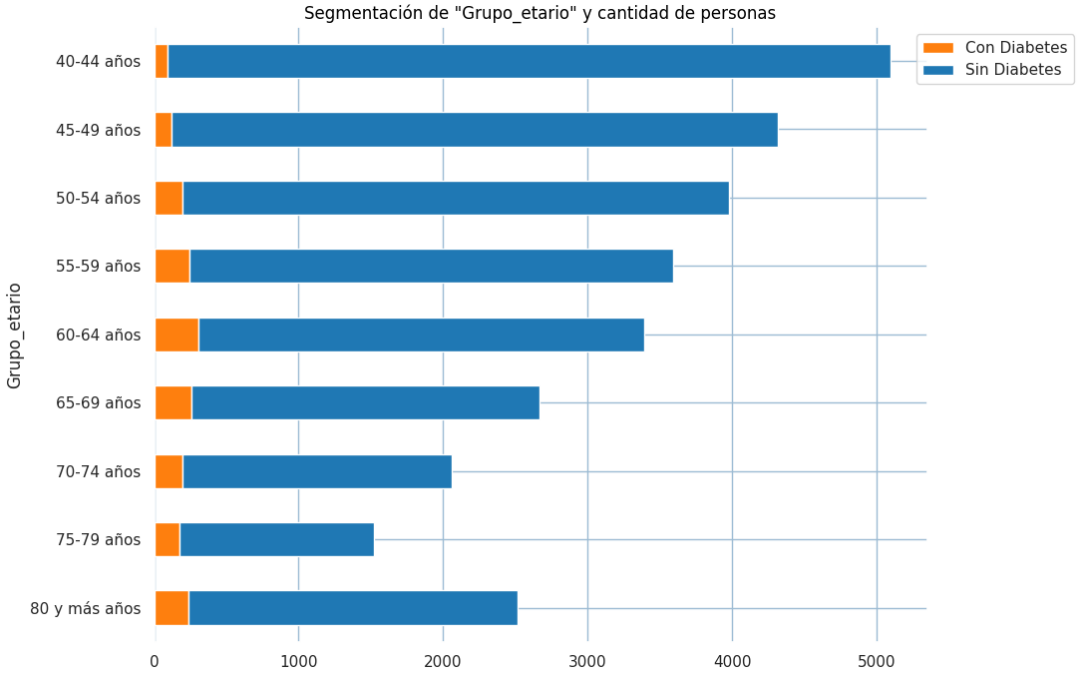

Ilustración 14. Caracterización de la población por grupo Etario - Concentración de DMT2 después de los 60 años. Fuente: Elaboración propia.

5.1.2.3 Revisión e identificación de variables no clínicas que pueden influir en la DMT2

Una vez realizada la caracterización de la base de datos con la que se estableció la población objetivo de estudio y sus características, se revisaron las variables, con el fin de identificar en primera instancia, aquellas que no tienen información relevante, están contenidas en otras y/o están fuera de alcance del proyecto de grado. De acuerdo con los trabajos e investigaciones realizados por *Leiva y otros (2018), Mathew & Sherly (2018) y Hajjaj, Salek, Basra, & Finlay (2010)*, mencionan el estado socioeconómico, raza del paciente, estilo de vida, deseos y actitud del paciente, entre muchos factores, con el desarrollo de la DMT2, además: "*algunos no modificables como edad, sexo e historia familiar de DMT2. El conocimiento de dichos factores es la clave para su prevención y tratamiento*" (Leiva, y otros, 2018). En concordancia con lo anterior, uno de los objetivos del presente trabajo de grado, fue determinar los factores no clínicos que se asocian con el riesgo de padecer DMT2, por ende, las variables identificadas para exclusión inicial fueron:

Tabla 10. Variables identificadas para excluir en primera instancia. Fuente: Elaboración propia

Exclusión de variables iniciales		
FILA	tiene perros gatos	Residente Habitual
id_afiliado	Perros gatos están vacunados	Estado civil
id_person_historic	Con quien permanece el menor	Sabe leer y escribir
Fecha de nacimiento	Menor asiste atención de primera infancia	Tiene celular
Tipo de sangre	Tipo cancel familiar 1 grado	lava manos antes de comer
Factor RH	Padecimiento enfermedad	Familiar con Cáncer
Fecha caracterización	signos o síntomas respiratorios Año	Ha sido mordido animales
Fecha de fin de caracterización	signos niño	Recibió vacuna antirrábica
acepta la toma de medidas	Departamento Afiliación	Síntomas después mordedura
Código municipio	Municipio Afiliación	Vida sexual activa
Código departamento	Familiar con enfermedad	relaciones sexuales1
Buscar barrio vereda	Iluminación natural o artificial suficiente	relaciones sexuales2
Barrio vereda	Presencia de insectos vectores	IPS
Dirección con nomenclatura	Síntomas últimos 30 días	Tienen servicio de teléfono fijo
AC_DIALISIS	AC_TRASPLANTE RENAL	Tipo de Cáncer padece o padecido
Vía de acceso	Fuente principal del agua para alimentos	Energía combustible utilizan para cocinar
Tipo de Vivienda	Tratamiento del agua para beber	Grupos poblacionales se identifica
Material paredes exteriores	En donde preparan los alimentos	Material de los pisos
Tipo de servicio sanitario en el hogar		

En la tabla 10 se observan las primeras variables identificadas a ser excluidas. El *tipo de sangre*, *factor rh* del afiliado, *si tiene mascotas*, *actividad sexual*, *estado civil*, *lavado de manos, teléfono celular o fijo y saber leer o escribir, vías de acceso, tipo y material de vivienda, fuente y tratamiento del agua*, etc., en la revisión del estado del arte, no se observó la influencia de éstas en los diversos estudios además que algunas ya están contenidas en otras variables. Las concernientes a la *historia de enfermedades familiares*, solamente se contempló la información de familiares que tuvieron DMT2, dado que las personas que tienen familiares con antecedentes presentan mayor riesgo de desarrollar DMT2 (Leiva, y otros, 2018). De la *fecha de nacimiento* se construye la variable *edad* del afiliado. Por último, de la información de enfermedades que padece el afiliado; se trabajó con la etiqueta asignada por el área de gestión del riesgo de la EPS para aquellos enfermos de DMT2 y la etiqueta de hipertensión arterial (AC_HIPERTENSION), dado que esta enfermedad es un factor de riesgo significativo en el desarrollo de DMT2 (Leiva, y otros, 2018). Como estudios

a futuro se podría realizar el análisis de los afiliados con otras comorbilidades versus la presencia de DMT2.

En segunda instancia, se realizó el análisis de calidad, en el que se encontraron variables con valores faltantes con más de 14%, por lo cual, generar un proceso de imputación puede incurrir en sesgo en el momento de entrenar los modelos posteriormente. Se relaciona a continuación, el siguiente corte de identificación para exclusión de variables:

Tabla 11. Variables excluidas por datos faltantes en segunda instancia. Fuente: Elaboración propia

Variables	Registros válidos	Cadena vacía	Espacio en blanco	% Valores nulos / vacíos	% Completo
Tipo de zona	24.161	4.960	4.960	17,0%	83,0%
disposición final de sobrantes	3.300	25.821	25.821	88,7%	11,3%
Fumo en los últimos 12 meses	2.658	26.463	26.463	90,9%	9,1%
Patrón de inyección	9	29.112	29.112	100,0%	0,0%
Medicamento de enfermedad cardiaca	284	28.837	28.837	99,0%	1,0%
Medicamento de tensión o presión alta	3.866	25.255	25.255	86,7%	13,3%
Asiste control de hipertensión arterial	3.871	25.250	25.250	86,7%	13,3%
Medicamento enfermedad renal	92	29.029	29.029	99,7%	0,3%
Medicamentos enfermedad de diabetes	1.137	27.984	27.984	96,1%	3,9%
Asiste a control de diabetes	1.142	27.979	27.979	96,1%	3,9%
Medicamentos control cáncer	205	28.916	28.916	99,3%	0,7%
Medicamentos control artritis	380	28.741	28.741	98,7%	1,3%
Medicamentos control EPOC	92	29.029	29.029	99,7%	0,3%
Medicamentos control VIH-SIDA	12	29.109	29.109	100,0%	0,0%
Medicamentos control Enf Mental	102	29.019	29.019	99,6%	0,4%
Medicación hipertensión regularmente	3.223	25.898	25.898	88,9%	11,1%
Recibe insulinas	1.137	27.984	27.984	96,1%	3,9%
Asiste control Artritis Reumatoidea	380	28.741	28.741	98,7%	1,3%
Recibió tratamiento para tuberculosis	39	29.082	29.082	99,9%	0,1%
Ha utilizado oxigeno medicinal en casa	24.810	4.311	4.311	14,8%	85,2%
baciloscopia para tuberculosis	751	28.370	28.370	97,4%	2,6%
Expuesto a humo de tabaco	2.608	26.513	26.513	91,0%	9,0%
signos o síntomas respiratorios	9.319	19.802	19.802	68,0%	32,0%

De acuerdo con el análisis de revisión de variables, la base de datos resultante tiene dimensiones de ***29.117 registros*** por ***52 variables***. En el siguiente capítulo se mostrará las variables entrantes y el proceso para completar el preprocesamiento.

5.2 Preparación de los datos

De acuerdo con la comprensión del negocio y los datos, se describe a continuación, las actividades realizadas, *adicionales a las mencionadas en la sección* ***5.1.2.1****,* ***5.1.2.2*** *y* ***5.1.2.3***, para transformar, ajustar tipos de datos, crear y eliminar variables, permitiendo obtener la base de datos limpia insumo para los modelos propuestos. Las secciones ***5.2.1*** y ***5.2.2***, se realizaron en el software de minería de datos IBM SPSS Modeler 18.3. A partir de este punto, todo el desarrollo del trabajo de grado con la base de datos obtenida se trasladó a ***Google Colab***[2], bajo el lenguaje de programación ***Python***.

5.2.1 Ajuste de tipo de datos y valores de variables

En primera instancia, se realizó un proceso de análisis de cada variable categórica, evaluando la posibilidad de reducir el número de categorías por la cuales estaban desagregadas. La finalidad de lo anterior corresponde que al momento de aplicar métodos de codificación (*enconding)*, como *One-hot*, aumentarían el número de variables. A continuación, se relacionan las variables objeto del análisis para disminución de categorías.

- **Actividades básicas diarias se ven afectadas:** agrupación de actividades afectadas semejantes con el fin de disminuir el número de categorías. Pasó de tener 190 a 8 categorías.

- **Ocupación de la persona:** agrupación de acuerdo con el tipo de ocupación, pasando de 211 categorías a 5 categorías.

- **Ingresos económicos mensuales del hogar:** agrupación de acuerdo con el rango del salario, pasando de 6 a 3 categorías.

- **Grupo Étnico:** agrupación de etnias de acuerdo con la cantidad de afiliados, las minorías se agrupan bajo una misma etiqueta, pasando de 8 a 4 categorías.

[2] https://colab.research.google.com/?hl=es#scrollTo=NZDxgaDWCFqI

- **Consumo de sustancia:** agrupación de acuerdo con la cantidad de afiliados asociados a la categoría, pasando de 19 a 3 categorías.

- **Frecuencia consume frutas y verduras:** agrupación de acuerdo con la cantidad de afiliados asociados al tipo de categoría, pasando de 6 a 4 categorías.

- **Consumo entre 4 y 8 vasos agua día:** agrupación de categorías, pasando de 6 a 5 categorías.

- **Familiar con Diabetes Mellitus:** la variable se transformó a dicotómica (SI/NO), dado que aquellos familiares que tenían la enfermedad correspondían al primer grado de consanguinidad. Pasa de 5 a 2 categorías.

- **Inflación en articulaciones:** la variable se transformó a dicotómica (SI/NO), dado que realizan combinaciones en la cual se encasillan varias partes del cuerpo, además que se repiten, pasando de 114 a 2 categorías.

- **Lesión en piel últimos 15 días:** la variable se transformó a dicotómica (SI/NO).

- **Estrato:** agrupación de acuerdo con categorías similares, reduciendo de 8 a 4 categorías.

- **Nivel educativo más alto:** agrupación de acuerdo con categorías similares, reduciendo de 9 a 4 categorías.

- **Estado de salud últimos 30 días:** agrupación de acuerdo con categorías similares, reduciendo de 6 a 3 categorías.

- **Tratamiento de variables categóricas dicotómicas / binarias:** se realizó el cambio a escalares las variables dicotómicas, en la cual, se le asignó a la etiqueta de clase “SI” el “1” y la etiqueta de clase “NO” el “0”. Lo anterior, quiere decir que la dimensión del vector de codificación es 1, puesto que solo aparece una variable de codificación por cada variable categórica original del problema, por lo cual, no se altera la dimensión

del espacio de características (Rocha Íñigo, 2020). Se realizó este cambio a 33 variables de la base de datos.

Tabla 12. Variables Dicotómicas tratadas en la base de datos. Fuente: Elaboración propia.

Variables Dicotómicas / Binarias		
Energía eléctrica	sordera total	Expuesto a humo de lena
Alcantarillado	Actividad física 30 min	Perdida capacidad del habla
Recolección de basuras	Adiciona sal a las comidas	debilidad entumecimiento cuerpo
Acueducto	Adiciona azúcar comidas	Palpitaciones en el pecho
Ventilación adecuada	Familiar con Diabetes Mellitus	Dolor opresivo en el pecho
Reservorios de agua	Valores de glucosa altos	Dificultad para respirar o sensación de ahogo
Aseo adecuado de la vivienda	Tiene dificultad visual	Perdida de la fuerza en manos pies
Actividad agropecuaria	Dificultad para oír	sudoración fría y palidez
Utiliza químicos en actividad agropecuaria	Tos con expectoración	AC DIABETES
Discapacidad física o mental	Piel blancas o rojizas	AC HIPERTENSION
ceguera total	Lesión en piel últimos 15 días	Inflamación en articulaciones

5.2.2 Creación de variables y escalado de variables numéricas

A partir de la fecha de nacimiento entregada en la base de datos, se realizó la construcción de la variable ***Edad*** del afiliado a corte 2023. Por otro lado, se aplicó El escalador ***sklearn.preprocessing.MinMaxScaler***, el cual transforma las características de las variables numéricas dejándolas en un rango de 0 a 1. Este proceso de escalamiento se realizó para tareas específicas como aplicación de *regresión logística con penalización Lasso para selección de variables*, dentro de la *optimización bayesiana para encontrar parámetros óptimos* y en la *base de datos de prueba (test) para obtener los resultados definitivos con los modelos entrenados bajo parámetros óptimos*.

5.2.3 Variables de entrada identificadas para la detección de DMT2

Con base en los preprocesamientos iniciales realizados y descritos en el capítulo de Entendimiento de los datos, secciones ***5.2.1*** y ***5.2.2*** del presente capítulo, se constituye una base de datos con dimensiones ***29.117*** registros por ***52*** variables. Dentro éstas, nuestra variable de respuesta será “AC_DIABETES”. A continuación, las variables elegidas como insumo para la predicción de la DMT2 con los modelos de Machine Learning, fueron las siguientes:

Tabla 13. Variables seleccionadas para el desarrollo del trabajo de grado. Fuente: Elaboración propia.

Variables	Descripción
Edad	Edad del afiliado al 2023
Sexo	Sexo
Latitud y longitud	Coordenadas donde el afiliado reside
Energía eléctrica	Si, si tiene energía eléctrica - No, lo contrario
Estrato	Estrato socioeconómico
Alcantarillado	Si, cuenta con alcantarillado - No, lo contrario
Recolección de basuras	Si, cuenta con recolección de basura - No, lo contrario
Acueducto	Si, cuenta con acueducto - No, lo contrario
Ventilación adecuada	Si, cuenta con ventilación adecuada - No, lo contrario
Reservorios de agua	Si, cuenta con reservorio de agua - No, lo contrario
Aseo adecuado de la vivienda	Si, aseo adecuado - No, lo contrario
Cuantos hogares alimentos por separado	No. De hogares que conviven en la vivienda
Cuantos cuartos en total dispone hogar	No. De cuartos que tiene la vivienda
Cuantos cuartos para dormir	No. De cuartos destinados para dormir
Ingresos económicos mensuales del hogar	Rango de ingresos que tiene el hogar (SMLV)
Actividad agropecuaria	Si, realiza actividad agropecuaria - No, lo contrario
Utiliza químicos en actividad agropecuaria	Si, utiliza químicos en la actividad agropecuaria - No, lo contrario
Cantidad de personas en el hogar	Cantidad de personas que conviven en el hogar
Grupo étnico	Grupo étnico al que pertenece el afiliado
Nivel educativo mas alto	Nivel educativo del afiliado
Condición de trabajo actual	Descripción general del tipo de oficio del afiliado
Ocupación de la persona	Descripción detallada del trabajo que realiza
Discapacidad física o mental	Si, tiene discapacidad - No, lo contrario
Actividades básicas diarias se ven afectadas	Tipo de actividad en la que se ve afectado por discapacidad
ceguera total	Si, tiene ceguera total - No, lo contrario
sordera total	Si, tiene sordera total - No, lo contrario
Consumo de sustancia	Tipo de sustancia que consume el afiliado
Actividad física 30 min	Si, realiza actividad física - No, lo contrario
Frecuencia consume frutas y verduras	Descripción de frecuencia de consumo de frutas
Consumo entre 4 y 8 vasos agua día	Descripción de frecuencia de consumo de agua
Adiciona sal a las comidas	Si, adiciona sal a la comida - No, lo contrario
Adiciona azúcar comidas	Si, adiciona azúcar a la comida - No, lo contrario
Familiar con Diabetes Mellitus	Descripción de acuerdo al familiar que tiene o no Diabetes Mellitus
Valores de glucosa altos	Si, tiene valor alto de glucosa - No, lo contrario
Inflamación en articulaciones	Descripción sobre en qué articulación tiene inflamación o si no ha presentado
Estado salud últimos 30 días	Descripción sobre como el afiliado ha estado en términos de salud en los últimos 30 días
Comparación estado salud	Descripción sobre como el afiliado ha estado en términos de salud
Tiene dificultad visual	Si, el afiliado tiene dificultad visual - No, lo contrario
Dificultad para oír	Si, el afiliado tiene dificultad para escuchar - No, lo contrario
Tos con expectoración	Si, el afiliado tiene Tos - No, lo contrario
Piel blancas o rojizas	Si, el afiliado tiene piel blanca o rojiza - No, lo contrario
Lesión en piel últimos 15 días	Si, el afiliado ha tenido una lesión - No, lo contrario
Expuesto a humo de leña	Si, el afiliado ha sido expuesto a humo de leña - No, lo contrario
Perdida capacidad del habla	Si, el afiliado ha perdido capacidad del habla - No, lo contrario
debilidad entumecimiento cuerpo	Si, el afiliado tiene entumecimiento del cuerpo - No, lo contrario
Palpitaciones en el pecho	Si, el afiliado ha tenido palpitaciones en el pecho - No, lo contrario
Dolor opresivo en el pecho	Si, el afiliado ha tenido dolor opresivo en el pecho - No, lo contrario
Dificultad para respirar o sensación de ahogo	Si, el afiliado ha tenido dificultad para respirar o sensación de ahogo - No, lo contrario
Perdida de la fuerza en manos pies	Si, el afiliado ha perdido fuerza en las manos / pies - No, lo contrario
sudoración fría y palidez	Si, el afiliado ha tenido sudoración fría y palidez - No, lo contrario
AC_DIABETES	Indicador de Diabetes / 1, el afiliado la padece - 0, lo contrario
AC_HIPERTENSION	Indicador de Hipertensión / 1, el afiliado la padece - 0, lo contrario

5.2.4 Codificación de las variables categóricas con más de una categoría

Dentro de la base de datos, se encuentran variables categóricas que están compuestas con varias clases. Inicialmente, éstas estaban como tipo de dato "***object***" y se transformaron a "***category***". De acuerdo con lo anterior, antes de incluirlas en los modelos propuestos, se realizaron procesos de codificación:

5.2.4.1 Codificación one-hot

También conocida como variables dummy, es una de las técnicas más utilizadas tanto para la resolución de problemas como en la investigación, con buenos resultados que normalmente produce cuando las categorías de la variable a codificar son mutuamente excluyentes (Rocha Íñigo, 2020). El método se basa, que a partir de una variable categórica X_i se crean p_i nuevas variables, las cuales toman valores de "0" a excepción de la categoría de referencia que toma el valor de "1".

Tabla 14. Ejemplo de codificación one-hot. Fuente: (Rocha Íñigo, 2020)

Vector de codificación	Categorías				
	Azul	Rojo	Verde	Blanco	Negro
$_{Azul}\Phi_{Color}$	1	0	0	0	0
$_{Rojo}\Phi_{Color}$	0	1	0	0	0
$_{Verde}\Phi_{Color}$	0	0	1	0	0
$_{Blanco}\Phi_{Color}$	0	0	0	1	0
$_{Negro}\Phi_{Color}$	0	0	0	0	1

Autores como J. Cohen, P. Cohen, West, & Aiken, 2013, proponen el uso de $p_i - 1$ variables cuando se codifique, con el fin de evitar redundancias y mejorar algunas propiedades estadísticas. Con base en lo anterior, para el trabajo de grado, se optó por el enfoque one-hot enconding tomando $p_i - 1$ variables. La nueva dimensión de la base de datos es de ***29.117 registros con 73 variables predictoras***.

5.2.5 Selección de Variables aplicando Regresión Logística con penalización Lasso

En primero instancia, se realizó una primera depuración de variables predictoras de acuerdo con el estado del arte y calidad de los datos como se describió en la sección *5.1.2.3 Revisión e identificación de variables no clínicas que pueden influir en la DMT2*. Sin embargo, puede haber variables que no estén relacionadas con la variable de respuesta "AC_DIABETES", afectando en términos de complejidad el modelo resultante. De acuerdo a lo anterior, se utilizó un modelo de Regresión Logística (librería *sklearn.linear_model*) en la cual se aplicó dentro de sus hiperparámetros, la penalización de acuerdo a Lasso (*penalty='l1'*), con el fin de reducir a cero los coeficientes de menor significancia. *Adicionalmente, se planteó una regla de selección de acuerdo con los resultados de los coeficientes, aplicando un umbral, en valor absoluto, por encima a 0,25*. Lo anterior, se aplicó sobre la base de datos de entrenamiento (80%), reduciendo a 24 variables más significativas. A continuación, se observa gráficamente, los valores de los coeficientes de las 73 variables predictoras.

Valores de coeficientes de las variables predictoras

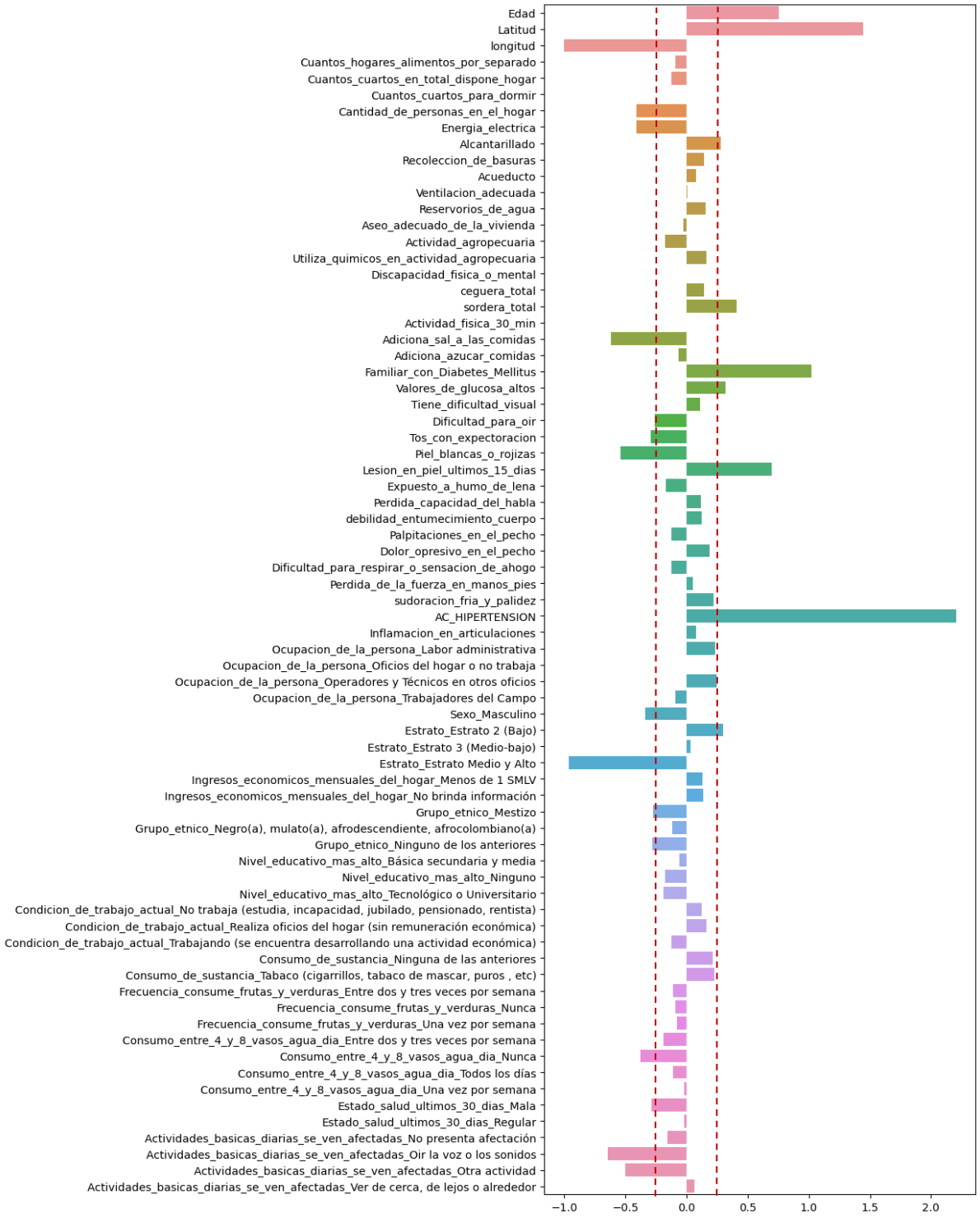

Ilustración 15. Resultado de los coeficientes de las variables predictoras. Fuente: Elaboración propia.

Las variables resultantes después de realizar la selección de variables son las siguientes:

Tabla 15. Variables resultantes del proceso de selección. Fuente: Elaboración propia.

Variables	Coeficiente
Edad	0,753609
Latitud	1,446411
longitud	-1,005417
Cantidad_de_personas_en_el_hogar	-0,408246
Energia_electrica	-0,411173
Alcantarillado	0,277205
sordera_total	0,411294
Adiciona_sal_a_las_comidas	-0,619472
Familiar_con_Diabetes_Mellitus	1,018631
Valores_de_glucosa_altos	0,316172
Dificultad_para_oir	-0,261571
Tos_con_expectoracion	-0,294232
Piel_blancas_o_rojizas	-0,541605
Lesion_en_piel_ultimos_15_dias	0,693479
AC_HIPERTENSION	2,204951
Sexo_Masculino	-0,339875
Estrato_Estrato 2 (Bajo)	0,301820
Estrato_Estrato Medio y Alto	-0,966169
Grupo_etnico_Mestizo	-0,272225
Grupo_etnico_Ninguno de los anteriores	-0,283031
Consumo_entre_4_y_8_vasos_agua_dia_Nunca	-0,376437
Estado_salud_ultimos_30_dias_Mala	-0,289329
Actividades_basicas_diarias_se_ven_afectadas_Oir la voz o los sonidos	-0,644189
Actividades_basicas_diarias_se_ven_afectadas_Otra actividad	-0,500384

Con base en el proceso realizado de selección de variables aplicando *regularización Lasso*, la variable de mayor influencia es AC_HIPERTENSIÓN, seguidas por las de ubicación (latitud y longitud), consanguinidad (Familiar con diabetes mellitus), estrato y edad. De acuerdo con lo anterior, es coincidente con las conclusiones y resultados que autores como *Leiva, y otros (2018)* encontraron en sus respectiva investigación. De manera complementaria, también tienen influencia las variables de estilo de vida y afectaciones en funciones físicas, contempladas también en el estudio desarrollado por *Mathew & Sherly (2018)*.

5.3 Modelado

Con la utilización de Python en Google Colab, los modelos empleados para predecir nuestra variable de respuesta "***AC_DIABETES***" fueron los ***Árboles de decisión***, ***KNN***, ***Random Forest, Gradient Boosting, eXtreme Gradient Boosting, MultiLayer Perceptron, Regresión Logística*** y ***Naive Bayes***.

5.3.1 Descripción de la función de optimización bayesiana

Después de realizar el preprocesamiento mencionado en los capítulos anteriores, con la base de datos limpia, se realizó la partición de ésta, asegurando inicialmente (***hold-out***) el **80%** de datos para entrenamiento (***training***) y el ***20%*** para prueba (***test***). De la partición de entrenamiento, se utilizaron 10 agrupaciones (***k-fold=10***), las cuales, se iteraron, con el fin de que cada grupo se utilizara como un conjunto de validación. Los datos restantes se utilizaron como nuevos conjuntos de entrenamiento para para buscar los mejores hiperparámetros de cada modelo propuesto, utilizando una función de ***optimización bayesiana*** de acuerdo con el trabajo desarrollado por los autores (Snoek, Larochelle, & Adams, 2012).

Para las métricas a optimizar, se realizó, una combinación entre el ***recall***, el ***f1-score*** y el ***ROC-AUC***, siendo la variable más importante el ***recall*** dada la finalidad del desarrollo del trabajo de grado, respecto a la identificación de personas con DMT2. La mejor combinación, de acuerdo con las métricas planteadas, arrojó una distribución de pesos del ***50% para el valor del recall, 20% para el f1_score y 30% para el roc_auc***.

Como se explicó al inicio de la sección, se utilizaron 10 agrupaciones (***K-fold=10***), para realizar las iteraciones de entrenamiento y validación de cada modelo. Por ende, dado que el entrenamiento se realizó 10 veces, se utilizó la media de sus valores para la métrica final.

5.3.2 Hiperparámetros empleados

Con base en los modelos utilizados, previamente, se realizó un proceso de evaluación para hallar los mejores valores de hiperparámetros propuestos en cada modelo, mediante la utilización una función de optimización bayesiana. A continuación, se observan los hiperparámetros a optimizar con su respectivo rango de valores.

Nota: Los modelos seleccionados para utilizar en la ***optimización bayesiana***, pertenecen a la librería ***Scikit-learn*** en Python, la cual contiene un amplio catálogo de modelos de machine learning para aprendizaje supervisado (Pedregosa, y otros, s.f.).

Tabla 16. Combinación de hiperparámetros en los modelos de machine learning para la predicción de DMT2. Fuente: Elaboración propia.

Algoritmos	Hiperparámetros y rango de valores a evaluar
KNN	n_neighbors: (3, 33)
Árboles de Decisión	max_depth: (10, 100)
	max_features: (0.3, 1)
Random Forest	n_estimators: (20, 300)
	max_depth: (10, 100)
	max_samples: (0.5, 1)
	max_features: (0.3, 1)
Naive Bayes	alpha: (0, 1)
Gradient Boosting	n_estimators: (100, 150)
	max_depth: (1, 20)
Regresión Logística	C = (0.01,1)
	class_weight:{0: 1, 1: w}, donde w: (1, 20)
Regresión Logística	class_weight: 'Balanced'
eXtreme Gradient Boosting	n_estimators: (100, 200)
	max_depth: (1, 20)
Multi Layer Perceptron	activation = ["identity", "logistic", "tanh", "relu"]
	solver = ["lbfgs", "sgd", "adam"]
	alpha = (0.0001, 1)
	learning_rate = ["constant", "invscaling", "adaptive"]
	learning_rate_init = (0.001, 1)
	hidden_layer_sizes = (hl_neurons = (1, 100))* (hl_layers = (1, 10))

5.3.3 Desbalance de clases

Un conjunto de datos está en desbalance, si las categorías de clasificación no están representadas aproximadamente por igual. Los problemas del "mundo real", se caracterizan por el desequilibrio de datos, que pueden desenlazar en costos de clasificación erróneo los cuales son desconocidos durante el momento de aprendizaje. El ***"accuraccy" (exactitud)***, la cual es una métrica popular para evaluar el rendimiento de un clasificador, puede no ser adecuada cuando hay desbalance con un costo de error de clasificación que varía notablemente (Chawla, 2010). Con base en lo anterior, es obvio que los clasificadores tradicionales que buscan un rendimiento preciso en una amplia gama de instancias no son adecuados para hacer frente a tareas de aprendizaje desequilibradas, ya que tienden a clasificar todos los datos en la clase mayoritaria, que suele ser la clase menos importante (Kotsiantis, Kanellopoulos, & Pintelas, 2006). En el área de salud, se han aplicado métodos de muestreo para el balance de clases con el fin de mejorar la predicción de los algoritmos de clasificación. La aplicación de muestreo con SMOTE en el trabajo realizado por Srivastava, Kumar, Fore, & Tomar (2021), les permitió mejorar la exactitud de sus modelos para identificar pacientes con *diabetes*. En esta sección se presentan algunas técnicas de muestreo que se utilizaron para equilibrar los conjuntos de datos.

5.3.3.1 Submuestreo Aleatorio (Undersampling)

"*Es un método no heurístico que tiene como objetivo equilibrar la distribución de clases mediante la eliminación aleatoria de ejemplos de clases mayoritarias. La razón detrás de esto es tratar de equilibrar el conjunto de datos en un intento de superar las idiosincrasias del algoritmo de aprendizaje automático. El principal inconveniente del submuestreo aleatorio es que este método puede descartar datos potencialmente útiles que podrían ser importantes para el proceso de inducción*" (Kotsiantis, Kanellopoulos, & Pintelas, 2006). Otro problema con la aplicación de este método es la pérdida de aleatoriedad de las muestras para estimar la distribución de la población, el cual es uno de los propósitos de clasificadores de aprendizaje automático (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

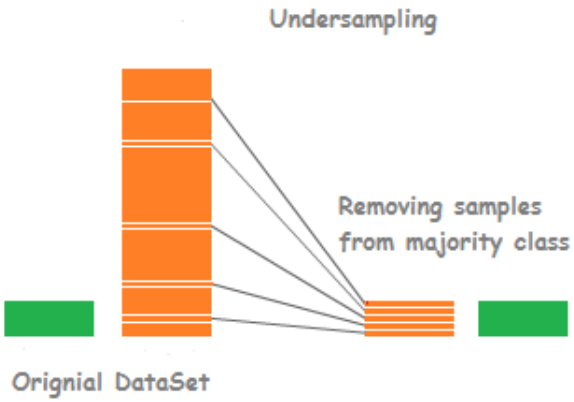

Ilustración 16. Random undersampling. Fuente: (Mohammed, Rawashdeh, & Abdullah, 2020)

5.3.3.2 Sobremuestreo Aleatorio (Oversampling)

Es un método no heurístico que tiene como objetivo equilibrar la distribución de clases a través de la replicación aleatoria de ejemplos de *clases minoritarias*. Sin embargo, varios autores coinciden en que el sobremuestreo aleatorio puede aumentar la probabilidad de que ocurra un sobreajuste (overfitting), ya que hace copias exactas de los ejemplos de las clases minoritarias. Además, puede introducir a una tarea computacional adicional si el conjunto de datos es bastante grande, pero está desequilibrado (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

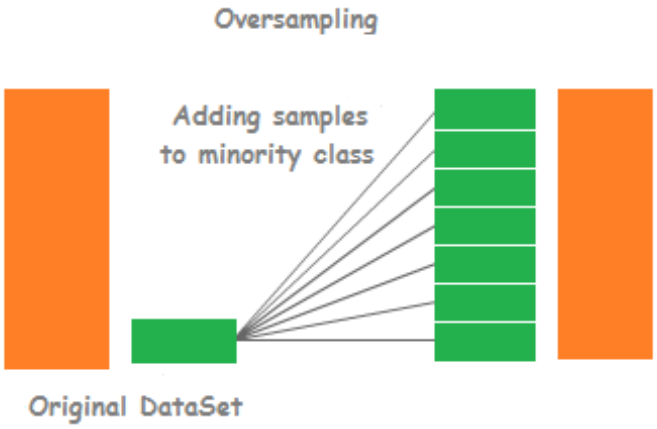

Ilustración 17. Random Oversampling. Fuente: (Mohammed, Rawashdeh, & Abdullah, 2020)

5.3.3.3 Smote

Este método consiste en generar ejemplos minoritarios sintéticos para sobremuestrear la clase minoritaria. Su idea principal es formar nuevos ejemplos de clases minoritarias mediante la interpolación entre varios ejemplos de clases minoritarias que se encuentran juntos. Para cada uno de éstos, se calculan sus k (*que se establece en 5 en SMOTE*) vecinos más cercanos de la misma clase, luego se seleccionan aleatoriamente algunos ejemplos de ellos de acuerdo con la tasa de sobremuestreo. Después, se generan nuevos ejemplos sintéticos a lo largo de la línea entre el ejemplo minoritario y sus vecinos más cercanos seleccionados. Por lo tanto, se evita el problema del sobreajuste y hace que los límites de decisión de la clase minoritaria se extiendan más hacia el espacio de la clase mayoritaria (Kotsiantis, Kanellopoulos, & Pintelas, 2006).

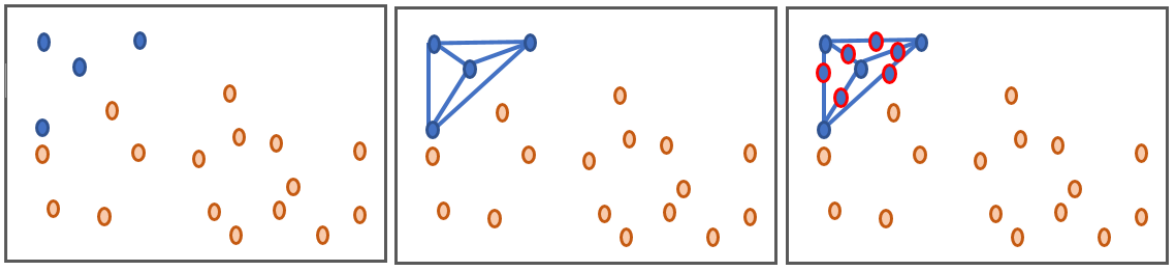

Ilustración 18. Funcionamiento Smote. Fuente: https://datasciencecampus.github.io/balancing-data-with-smote/

5.3.3.4 Smote-tomek

Método híbrido introducido por primera vez por Batista et al. (2003), en la cual, combina la capacidad de ***SMOTE*** para generar datos sintéticos para la clase minoritaria y la capacidad de ***Tomek-Links***, el cual funciona eliminando las instancias de clase mayoritaria que están más cerca de la clase minoritaria mediante la aplicación de la regla del vecino más cercano para seleccionar instancias (Fezeka Swana, Doorsamy, & Bokoro, 2022).

5.4 Evaluación

En la sección **5.3.1**, se mencionó la utilización de métricas como la ***sensibilidad*** (***recall), f1-score y el ROC-AUC*** dentro del proceso de la función de optimización bayesiana, siendo el ***recall*** la métrica que primó en todo el enfoque del trabajo de grado. Adicional a éstas, se tuvo en cuenta la métrica ***exactitud*** (***accuracy)*** para efectos de complementariedad a la hora de la evaluación de la bondad de los modelos propuestos.

5.4.1 Exactitud (Accuracy)

Es la métrica que ayuda a saber que tan exacto o cercano es el resultado al valor verdadero, brindando, además, información de los posibles errores en el proceso de clasificación (Zapeta Hernández, Galindo Rosales, Juan Santiago, & Martínez Lee, 2022). La ecuación es la siguiente:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \quad (\mathbf{1})$$

Donde TP: Verdadero positivo; FP: Falso positivo; TN: Verdadero negativo y FN: Falso negativo.

5.4.2 Sensibilidad (Recall)

Métrica de rendimiento equivalente a la tasa de verdaderos positivos. La sensibilidad es igual a todos los verdaderos positivos divididos por la suma de falso negativos y verdaderos positivos (Zapeta Hernández, Galindo Rosales, Juan Santiago, & Martínez Lee, 2022). De acuerdo con lo anterior, esta métrica da respuesta a la pregunta ¿***Qué porcentaje de afiliados que tienen DMT2 es capaz de identificarlos los modelos***? La ecuación es la siguiente:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \quad (\mathbf{2})$$

5.4.3 F1-Score

Métrica que combina la precisión y la sensibilidad en un solo valor. Es de gran utilidad para en clases desbalanceadas (Zapeta Hernández, Galindo Rosales, Juan Santiago, & Martínez Lee, 2022). La ecuación es la siguiente:

$$\boldsymbol{F1} = \mathbf{2} * \frac{\boldsymbol{Pre} * \boldsymbol{Rec}}{\boldsymbol{Pre} + \boldsymbol{Rec}} \quad (\mathbf{3})$$

Donde Pre: métrica de ***Precisión***.

5.4.4 ROC-AUC

Desde el punto de vista en Salud, "*la curva ROC es una herramienta estadística que se utiliza para evaluar la capacidad discriminativa de una prueba diagnóstica dicotómica. Se trata de curvas en las que se presenta la sensibilidad (recall) en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte. Son útiles para elegir el punto de corte más adecuado de una prueba, conocer el rendimiento global de ésta y comparar la capacidad discriminativa de 2 o más pruebas diagnósticas"* (Martínez Pérez & Pérez Martin, 2022). En general, mide el rendimiento de un clasificar binario.

6. PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

En el presente capítulo, se observa, cómo se comportaron los distintos algoritmos propuestos al intentar predecir/clasificar los datos de la variable de respuesta "***AC_DIABETES***". Se plantearon dos escenarios en el proceso de experimentación:

- Aplicación de los algoritmos con los hiperparámetros óptimos con la base de datos limpia obtenida, *sin ningún proceso adicional de selección de variables*, es decir con las 73 variables que surgen después de la aplicación del ***one-hot enconding*** *en la* ***sección 5.2.4***. Adicionalmente, se presentan los resultados en primera instancia sin la influencia de técnicas de muestreo. Después, se observan los rendimientos con la aplicación de las técnicas de muestreo explicadas en la ***sección 5.3.3***.

- Aplicación de las técnicas con los hiperparámetros óptimos con la base de datos limpia obtenida, aplicando ***Regresión Logística con penalización de Lassi (sección 5.2.5)*** a las 73 variables predictoras obtenidas a partir de ***one-hot enconding***. De acuerdo con lo anterior, se obtuvieron 24 variables predictoras con los valores de coeficientes (betas) que tuvieron un umbral de importancia igual o superior de ***2.5*** (en valor absoluto). Igual que en el primer escenario, se presentan los resultados sin la influencia de técnicas de muestreo. Después, se observan los rendimientos con la aplicación de las técnicas de muestreo explicadas en la ***sección 5.3.3***.

Flujo General del desarrollo metodológico

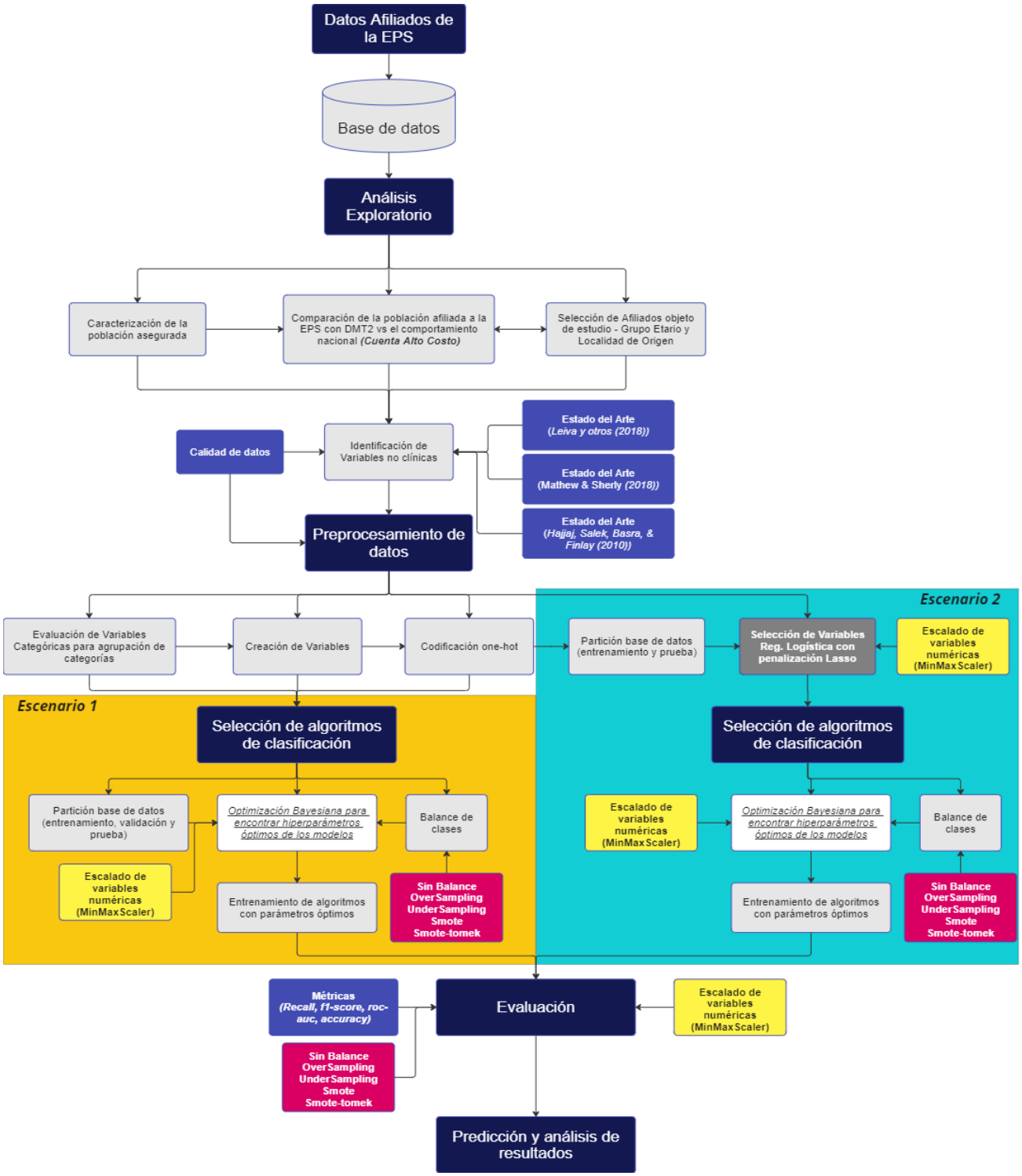

Ilustración 19. Flujo general del desarrollo de la metodología para la obtención del Modelo de Predicción de DMT2. Fuente: Elaboración propia en miro.

6.1 Escenario 1: Aplicación de algoritmos con métodos de muestreo sobre el total de la base de datos

6.1.1 Resultados con datos limpios sin balance de clases

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos con el desbalanceo original, sin la aplicación de alguna técnica de muestreo, sin embargo, se hace refinamiento en los pesos de las clases en la técnica ***regresión logística*** para afrontar esta problemática.

Tabla 17. Resultado de las métricas de evaluación para cada algoritmo con desbalance de clases. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,9371±2,2e-05	0,9371±0,0002	0,94	0,00±0,00	0,00±0,00	0,00
Árbol decisión	max_depth	59,00	0,9994±7,4e-05	0,8907±0,0060	0,89	0,9911±0,0013	0,1831±0,0417	0,18
	max_features	0,73						
Random Forest	n_estimators	213,00	0,9994±8,4e-05	0,9338±0,0028	0,93	0,9914±0,0017	0,0478±0,0159	0,04
	max_depth	73,00						
	max_samples	0,93						
	max_features	0,74						
Naive Bayes	***alpha***	***0,87***	***0,6601±0,0015***	***0,6593±0,0084***	***0,67***	***0,7777±0,0046***	***0,7711±0,0274***	***0,77***
Gradient Boosting	max_depth	20,00	0,9993±7,9e-05	0,9240±0,0034	0,92	0,9921±0,0014	0,1045±0,0330	0,10
	n_estimators	141,00						
Regresión Logística	***C***	***0,01***	***0,7714±0,0022***	***0,7698±0,0090***	***0,77***	***0,7794±0,0042***	***0,7657±0,0363***	***0,74***
	class_weight	***{0: 1, 1:20}***						
Regresión Logística	***class_weight***	***balanced***	***0,8034±0,0016***	***0,8021±0,0085***	***0,80***	***0,7337±0,0025***	***0,7220±0,0339***	***0,70***
eXtreme Gradient Boosting	max_depth	19,00	1,00±0,00	0,9300±0,0027	0,93	1,00±0,00	0,0915±0,0198	0,08
	n_estimators	119,00						

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	3,00	0,00±0,00	0,00±0,00	0,00	0,8518±0,0018	0,7554±0,0263	0,76
Árbol decisión	max_depth	59,00	0,9949±0,0006	0,1731±0,0353	0,17	0,9999±8,6e-07	0,5626±0,0203	0,56
	max_features	0,73						
Random Forest	n_estimators	204,00	0,9952±0,0006	0,0829±0,0255	0,07	0,9999±5,0e-06	0,8000±0,0149	0,78
	max_depth	83,00						
	max_samples	1,00						
	max_features	0,97						
Naive Bayes	***alpha***	***0,43***	***0,2234±0,0015***	***0,2216±0,0095***	***0,23***	***0,7855±0,0019***	***0,7802±0,0197***	***0,78***
Gradient Boosting	max_depth	20,00	0,9951±0,0006	0,1468±0,0044	0,14	0,9999±8,1e-07	0,7686±0,0201	0,78
	n_estimators	141,00						
Regresión Logística	***C***	***0,01***	***0,3000±0,022***	***0,2950±0,0016***	***0,29***	***0,8362±0,0019***	***0,8254±0,017***	***0,82***
	class_weight	***{0: 1, 1:20}***						
Regresión Logística	***class_weight***	***balanced***	***0,3193±0,0024***	***0,3147±0,0175***	***0,31***	***0,8371±0,0020***	***0,8233±0,180***	***0,82***
eXtreme Gradient Boosting	max_depth	19,00	1,00±0,00	0,1406±0,0280	0,13	1,00±3,5e-17	0,7913±0,0140	0,79
	n_estimators	119,00						

De acuerdo con los resultados anteriores, en cuanto a la métrica ***recall***, el algoritmo de mejor desempeño fue *Naive Bayes* con el 77%, seguido por *Regresión logística* con pesos óptimos (*class_weight {0:1, 1:20}*) en un 74%. Estos resultados son congruentes con respecto a los valores obtenidos en entrenamiento y validación de la métrica. Sin embargo, desde la perspectiva *general en rendimiento de los modelos*, el algoritmo de Regresión Logística refinado con pesos óptimos de las clases, aunque tiene un menor *recall* con respecto a Naive Bayes (-3 puntos porcentuales), supera de manera importante en las demás métricas; *accuracy* en 77% (10 puntos porc. por encima), *f1-score* en 29% (6 puntos porc. por encima) y *roc-auc* en 82% (4 puntos porc. por encima).

6.1.2 Resultados con datos limpios con balanceo de clases aplicando Random Oversampling

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos aplicando la técnica de muestreo *random oversampling*. Para este caso, con respecto a la Regresión Logística, no se realizó ningún tipo de refinamiento al modelo, dado que el muestreo aplicado ya realizó un balance de clases.

Tabla 18. Resultado de las métricas de evaluación para cada algoritmo con Oversampling. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,7985±0,0019	0,7071±0,0121	0,71	0,8725±0,0034	0,6298±0,0514	0,65
Árbol decisión	max_depth	10,00	0,8427±0,0056	0,7585±0,0109	0,78	0,9042±0,0117	0,6919±0,0423	0,68
	max_features	0,97						
Random Forest	n_estimators	181,00	0,8754±0,0015	0,8228±0,0113	0,82	0,9086±0,0045	0,6946±0,0349	0,66
	max_depth	10,00						
	max_samples	0,50						
	max_features	0,30						
Naive Bayes	***alpha***	***1,00***	***0,7143±0,0019***	***0,6574±0,0072***	***0,67***	***0,7780±0,0034***	***0,7718±0,0301***	***0,78***
Gradient Boosting	***max_depth***	***1,00***	***0,7716±0,0022***	***0,8118±0,0082***	***0,81***	***0,7256±0,0045***	***0,7240±0,0346***	***0,69***
	n_estimators	***150,00***						
Regresión Logística	***class_weight***	***{0: 1, 1:1}***	***0,7709±0,0021***	***0,8032±0,0094***	***0,80***	***0,7332±0,0037***	***0,7213±0,0338***	***0,70***
eXtreme Gradient Boosting	***max_depth***	***1,00***	***0,7778±0,0026***	***0,8070±0,0079***	***0,81***	***0,7436±0,0054***	***0,7281±0,0369***	***0,70***
	n_estimators	***100,00***						

Tabla 17-Continuación. Resultado de las métricas de evaluación para cada algoritmo con Oversampling. Fuente: Elaboración propia.

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8124±0,0018	0,2130±0,0191	0,22	0,8927±0,0017	0,7273±0,0261	0,73
Árbol decisión	max_depth	10,00	0,8515±0,0051	0,2649±0,0124	0,28	0,9064±0,0048	0,7406±0,0293	0,74
	max_features	0,97						
Random Forest	n_estimators	181,00	0,8794±0,0015	0,3307±0,0217	0,31	0,9506±0,0009	0,8253±0,0149	0,82
	max_depth	10,00						
	max_samples	0,50						
	max_features	0,30						
Naive Bayes	***alpha***	***1,00***	***0,7314±0,0021***	***0,2207±0,0093***	***0,23***	***0,7854±0,0019***	***0,7797±0,0201***	***0,78***
Gradient Boosting	***max_depth***	***1,00***	***0,7606±0,0028***	***0,3262±0,0185***	***0,31***	***0,8401±0,0015***	***0,8332±0,0124***	***0,83***
	n_estimators	***150,00***						
Regresión Logística	***class_weight***	***{0: 1, 1:1}***	***0,7619±0,0026***	***0,3157±0,0187***	***0,31***	***0,8368±0,0023***	***0,8228±0,0178***	***0,82***
eXtreme Gradient Boosting	***max_depth***	***1,00***	***0,7699±0,0032***	***0,3218±0,0180***	***0,31***	***0,8451±0,0015***	***0,8334±0,0130***	***0,83***
	n_estimators	***100,00***						

De acuerdo con los resultados anteriores, en cuanto a la métrica ***recall***, el algoritmo de mejor desempeño fue *Naive Bayes* con el 78%, seguido por *Regresión logística, eXtreme Gradient Boosting* y *Gradient Boosting* con el 70%, 70% y 69% respectivamente. Estos resultados son congruentes con respecto a los valores obtenidos en entrenamiento y validación de la métrica. Sin embargo, desde la perspectiva *general en rendimiento de los modelos*, los algoritmos de Regresión Logística, eXtreme Gradient Boosting y Gradient Boosting, los cuales tienen resultados similares, aunque tiene un menor *recall* con respecto a Naive Bayes (-8 puntos porcentuales aprox.), superan de manera importante en las demás métricas; *accuracy* en 80%-81% (13 puntos porc. por encima aprox.), *f1-score* en 31% (7 puntos porc. por encima) y *roc-auc* en 82%-83% (5 puntos porc. por encima aprox.). En este punto, es importante definir junto con la EPS, hasta qué punto la detección de personas enfermas prima sobre clasificar personas sanas como posibles enfermas. Este análisis quedará como trabajo futuro, dado que se tendrá que incluir otros tipos de conceptos técnicos en salud y hasta económicos, que no hacen parte del trabajo de grado desarrollado.

6.1.3 Resultados con datos limpios con balanceo de clases aplicando UnderSampling

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos aplicando la técnica de muestreo *random undersampling*. Para este caso, no se realizó ningún tipo de refinamiento con respecto al balance de clase a los modelos, dado que el muestreo aplicado ya realizó el respectivo balance.

Tabla 19. Resultado de las métricas de evaluación para cada algoritmo con Undersampling. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	19,00	0,7507±0,0078	0,7153±0,0162	0,71	0,7489±0,01105	0,7328±0,0385	0,72
Árbol decisión	max_depth	10,00	0,8794±0,0074	0,7095±0,0157	0,72	0,8900±0,0095	0,7425±0,0265	0,74
	max_features	1,00						
Random Forest	n_estimators	219,00	0,9863±0,0017	0,7702±0,0110	0,77	0,9869±0,0018	0,7650±0,0268	0,74
	max_depth	38,00						
	max_samples	0,63						
	max_features	0,54						
Naive Bayes	***alpha***	***1,00***	***0,7199±0,0072***	***0,6502±0,0173***	***0,65***	***0,7787±0,0071***	***0,7677±0,0247***	***0,78***
Gradient Boosting	max_depth	1,00	0,7736±0,0087	0,8085±0,0072	0,80	0,7285±0,0035	0,7260±0,0350	0,69
	n_estimators	136,00						
Regresión Logística	class_weight	{0: 1, 1:1}	0,7753±0,0092	0,7918±0,0097	0,80	0,7420±0,0089	0,7301±0,0398	0,71
eXtreme Gradient Boosting	***max_depth***	***1,00***	***0,7877±0,0103***	***0,7898±0,0119***	***0,78***	***0,7646±0,0130***	***0,7465±0,0373***	***0,72***
	n_estimators	***189,00***						

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	19,00	0,7502±0,0069	0,2436±0,0143	0,25	0,8218±0,0078	0,7712±0,0187	0,77
Árbol decisión	max_depth	10,00	0,8807±0,0065	0,2436±0,0143	0,25	0,9446±0,0061	0,7224±0,02610	0,74
	max_features	1,00						
Random Forest	n_estimators	219,00	0,9863±0,0017	0,2954±0,0150	0,29	0,9993±0,0001	0,8173±0,0166	0,82
	max_depth	38,00						
	max_samples	0,63						
	max_features	0,54						
Naive Bayes	***alpha***	***1,00***	***0,7355±0,0061***	***0,2166±0,0113***	***0,22***	***0,7903±0,0091***	***0,7768±0,0199***	***0,78***
Gradient Boosting	max_depth	1,00	0,7629±0,0075	0,3228±0,0149	0,31	0,8445±0,0081	0,8322±0,0125	0,83
	n_estimators	136,00						
Regresión Logística	class_weight	{0: 1, 1:1}	0,7676±0,0090	0,3062±0,0185	0,30	0,8429±0,0084	0,8203±0,0188	0,82
eXtreme Gradient Boosting	***max_depth***	***1,00***	***0,7827±0,0107***	***0,3089±0,0155***	***0,30***	***0,8568±0,0080***	***0,8292±0,0138***	***0,82***
	n_estimators	***189,00***						

De acuerdo con los resultados anteriores, en cuanto a la métrica ***recall***, el algoritmo de mejor desempeño fue *Naive Bayes* con el 78%, seguido por *eXtreme Gradient Boosting*,

KNN, *Regresión logística* y *Gradient Boosting* con el 72%, 71% y 69% respectivamente. Estos resultados son congruentes con respecto a los valores obtenidos en entrenamiento y validación de la métrica. Sin embargo, desde la perspectiva *general en rendimiento de los modelos*, los algoritmos de Regresión Logística y eXtreme Gradient Boosting, los cuales tienen resultados similares, aunque tiene un menor *recall* con respecto a Naive Bayes (-7 y -6 puntos porcentuales aprox.), superan de manera importante en las demás métricas; *accuracy* en 80% y 78% (15 y 13 puntos porc. por encima aprox.), *f1-score* en 30% (8 puntos porc. por encima) y *roc-auc* en 82% (5 puntos porc. por encima aprox.).

6.1.4 Resultados con datos limpios con balanceo de clases aplicando SMOTE

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos aplicando la técnica de muestreo *SMOTE*. Para este caso, no se realizó ningún tipo de refinamiento con respecto al balance de clase a los modelos, dado que el muestreo aplicado ya realizó el respectivo balance.

Tabla 20. Resultado de las métricas de evaluación para cada algoritmo con SMOTE. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8217±0,0023	0,6505±0,0105	0,65	0,9806±0,0019	0,7035±0,0462	0,69
Árbol decisión	max_depth	11,00	0,8540±0,0012	0,7893±0,0199	0,82	0,8779±0,022	0,5028±0,0960	0,45
	max_features	0,30						
Random Forest	n_estimators	199,00	0,90164±0,0038	0,8299±0,0078	0,83	0,9390±0,0041	0,5621±0,041	0,56
	max_depth	10,00						
	max_samples	1,00						
	max_features	0,30						
Naive Bayes	alpha	1,00	0,7507±0,0026	0,6731±0,0047	0,68	0,8239±0,0031	0,6236±0,0529	0,62
Gradient Boosting	***max_depth***	***1,00***	***0,8062±0,0026***	***0,8150±0,0105***	***0,81***	***0,7871±0,0076***	***0,6612±0,0418***	***0,64***
	n_estimators	***100,00***						
Regresión Logística	class_weight	{0: 1, 1:1}	0,8364±0,0027	0,8125±0,0067	0,80	0,8367±0,0034	0,4767±0,0489	0,45
eXtreme Gradient Boosting	max_depth	1,00	0,8654±0,0019	0,8490±0,0085	0,84	0,8582±0,0036	0,5191±0,0366	0,52
	n_estimators	104,00						

Tabla 20 - Continuación. Resultado de las métricas de evaluación para cada algoritmo con SMOTE. Fuente: Elaboración propia.

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8462±0,0018	0,2019±0,0125	0,20	0,9566±0,0007	0,7269±0,0259	0,73
Árbol decisión	max_depth	11,00	0,8574±0,0119	0,2293±0,0267	0,24	0,9322±0,0121	0,6872±0,0430	0,69
	max_features	0,30						
Random Forest	n_estimators	199,00	0,9052±0,0036	0,2935±0,0200	0,29	0,9358±0,0015	0,7960±0,0149	0,80
	max_depth	10,00						
	max_samples	1,00						
	max_features	0,30						
Naive Bayes	alpha	1,00	0,7677±0,0025	0,1933±0,0141	0,19	0,8322±0,0023	0,6974±0,0264	0,70
Gradient Boosting	***max_depth***	***1,00***	***0,8024±0,0036***	***0,3103±0,0224***	***0,30***	***0,8883±0,0008***	***0,8026±0,0135***	***0,79***
	n_estimators	***100,00***						
Regresión Logística	class_weight	{0: 1, 1:1}	0,8365±0,0028	0,2421±0,0239	0,22	0,9152±0,0015	0,7172±0,02671	0,71
eXtreme Gradient Boosting	max_depth	1,00	0,8644±0,0019	0,3022±0,0252	0,29	0,9411±0,0010	0,7917±0,0127	0,78
	n_estimators	104,00						

De acuerdo con los resultados anteriores, desde la perspectiva *general en rendimiento de los modelos*, *Gradient Boosting* fue el de mejor desempeño bajo las condiciones ya mencionadas. Sin embargo, su métrica ***recall*** en los datos de *prueba (test)*, presenta brechas frente a los valores del *entrenamiento (training)* (diferencias alrededor de 14 puntos porc. aprox.). En síntesis, el muestreo con datos sintéticos bajo *SMOTE*, no influyó de manera positiva en los algoritmos de clasificación propuestos.

6.1.5 Resultados con datos limpios con balanceo de clases aplicando SMOTE-Tomek

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos aplicando la técnica de muestreo *SMOTE-tomek*. Para este caso, no se realizó ningún tipo de refinamiento con respecto al balance de clase a los modelos, dado que el muestreo aplicado ya realizó el respectivo balance.

Tabla 21. Resultado de las métricas de evaluación para cada algoritmo con SMOTE-tomek. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8228±0,0023	0,6502±0,0103	0,65	0,9806±0,0020	0,7035±0,0467	0,68
Árbol decisión	max_depth	10,00	0,8380±0,0168	0,7755±0,0368	0,80	0,8670±0,0213	0,5436±0,0788	0,45
	max_features	0,30						
Random Forest	n_estimators	299,00	0,9033±0,0049	0,8279±0,0064	0,82	0,9412±0,0042	0,5532±0,0373	0,54
	max_depth	10,00						
	max_samples	0,83						
	max_features	0,42						
Naive Bayes	alpha	0,98	0,7519±0,0026	0,6726±0,0047	0,68	0,8243±0,0030	0,6243±0,0534	0,62
Gradient Boosting	***max_depth***	***1,00***	***0,8063±0,0023***	***0,8152±0,0092***	***0,81***	***0,7849±0,0068***	***0,6639±0,0393***	***0,63***
	n_estimators	***100,00***						
Regresión Logística	class_weight	{0: 1, 1:1}	0,8378±0,0027	0,8117±0,0065	0,81	0,8378±0,0033	0,4774±0,0483	0,46
eXtreme Gradient Boosting	max_depth	1,00	0,8643±0,0018	0,8457±0,0119	0,84	0,8576±0,0018	0,5334±0,0336	0,54
	n_estimators	100,00						

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8469±0,0018	0,2018±0,0124	0,20	0,9277±0,0007	0,7275±0,0261	0,73
Árbol decisión	max_depth	10,00	0,8428±0,0134	0,2342±0,0167	0,22	0,9171±0,0157	0,7105±0,0231	0,71
	max_features	0,30						
Random Forest	n_estimators	299,00	0,9068±0,0046	0,2878±0,0171	0,27	0,9700±0,0009	0,7934±0,0141	0,79
	max_depth	10,00						
	max_samples	0,83						
	max_features	0,42						
Naive Bayes	alpha	0,98	0,7686±0,0024	0,1932±0,0142	0,19	0,8337±0,0023	0,6979±0,0263	0,70
Gradient Boosting	***max_depth***	***1,00***	***0,8021±0,0032***	***0,3114±0,0199***	***0,29***	***0,8894±0,0008***	***0,8028±0,0132***	***0,79***
	n_estimators	***100,00***						
Regresión Logística	class_weight	{0: 1, 1:1}	0,8378±0,0028	0,2416±0,0236	0,23	0,9162±0,0015	0,7177±0,0266	0,71
eXtreme Gradient Boosting	max_depth	1,00	0,8634±0,0018	0,3030±0,0183	0,30	0,9404±0,0009	0,7914±0,0119	0,79
	n_estimators	100,00						

De acuerdo con los resultados anteriores, desde la perspectiva *general en rendimiento de los modelos*, *Gradient Boosting* fue el de mejor desempeño bajo las condiciones ya mencionadas. Sin embargo, su métrica ***recall*** en los datos de *prueba (test)*, presenta brechas frente a los valores del *entrenamiento (training)* (diferencias alrededor de 15 puntos porc. aprox.). En síntesis, el muestreo con datos sintéticos bajo *SMOTE-tomek*, no influyó de manera positiva en los algoritmos de clasificación propuestos.

6.2 Escenario 2: Aplicación de algoritmos con métodos de muestreo sobre la base de datos después de selección de variables con regularización Lasso.

Como se mencionó al principio del presente capítulo, el desarrollo de este escenario se caracterizó por la utilización de selección de variables con la aplicación de regularización Lasso. Se incluyó dentro de este desarrollo un algoritmo de redes neuronales multicapa denominada ***Multi Layer Perceptron***.

6.2.1 Resultados con datos limpios sin balance de clases

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos con el desbalanceo original, sin la aplicación de alguna técnica de muestreo, sin embargo, se hace refinamiento en los pesos de las clases en la técnica ***regresión logística*** para afrontar esta problemática.

Tabla 22. Resultado de las métricas de evaluación para cada algoritmo con desbalance de clases. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	32,00	0,9370±0,0002	0,9367±0,0010	0,94	0,0100±0,0049	0,0068±0,0061	0,003
Árbol decisión	max_depth	98,00	0,9982±0,0002	0,8932±0,0059	0,90	0,9735±0,0033	0,2104±0,0362	0,19
	max_features	0,78						
Random Forest	n_estimators	159,00	0,9980±0,0001	0,9295±0,0030	0,93	0,9724±0,0021	0,0799±0,0233	0,07
	max_depth	88,00						
	max_samples	0,84						
	max_features	0,92						
Naive Bayes	***alpha***	***0,33***	***0,7348±0,0020***	***0,7335±0,0075***	***0,74***	***0,7673±0,0053***	***0,7636±0,0362***	***0,75***
Gradient Boosting	max_depth	3,00	0,9395±0,0002	0,9365±0,0013	0,94	0,0497±0,0065	0,0239±0,0130	0,03
	n_estimators	117,00						
Regresión Logística	***C***	***1,00***	***0,7711±0,0024***	***0,7707±0,0053***	***0,77***	***0,7684±0,0028***	***0,7636±0,0263***	***0,73***
	class_weight	***{0: 1, 1:20}***						
Regresión Logística	***class_weight***	***balanced***	***0,8020±0,0018***	***0,8019±0,0098***	***0,80***	***0,7329±0,0029***	***0,7301±0,0344***	***0,70***
eXtreme Gradient Boosting	max_depth	19,00	0,9999±1,43e-05	0,9250±0,0050	0,93	0,9995±0,0003	0,1059±0,0217	0,07
	n_estimators	168,00						
Multi Layer Perceptron	activation	relu	0,9416±0,0004	0,9336±0,0025	0,94	0,1139±0,0109	0,0519±0,0134	0,04
	solver	lbfgs						
	alpha	1,00						
	learning_rate	constant						
	learning_rate_init	0,01						
	hidden_layer_sizes	78,00						

Tabla 22 - Continuación. Resultado de las métricas de evaluación para cada algoritmo con desbalance de clases. Fuente: Elaboración propia.

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	32,00	0,01954±0,0095	0,01336±0,0119	0,01	0,8753±0,0016	0,8085±0,0173	0,78
Árbol decisión	max_depth	98,00	0,9858±0,0016	0,1983±0,0321	0,19	0,9999±7,74-06	0,5781±0,0167	0,57
	max_features	0,78						
Random Forest	n_estimators	159,00	0,9841±0,0009	0,1248±0,0357	0,11	0,9999±1,15e-05	0,7871±0,0149	0,78
	max_depth	88,00						
	max_samples	0,84						
	max_features	0,92						
Naive Bayes	***alpha***	***0,33***	***0,2667±0,0013***	***0,2649±0,0134***	***0,26***	***0,8155±0,0019***	***0,8130±0,0187***	***0,79***
Gradient Boosting	max_depth	3,00	0,0935±0,0116	0,0448±0,0239	0,05	0,8591±0,0016	0,8334±0,0138	0,83
	n_estimators	117,00						
Regresión Logística	***C***	***1,00***	***0,2968±0,0223***	***0,2951±0,0110***	***0,29***	***0,8326±0,0019***	***0,8286±0,0174***	***0,82***
	class_weight	***{0: 1, 1:20}***						
Regresión Logística	***class_weight***	***balanced***	***0,3176±0,0026***	***0,3169±0,0189***	***0,30***	***0,8325±0,0019***	***0,8288±0,0174***	***0,82***
eXtreme Gradient Boosting	max_depth	19,00	0,9996±0,0001	0,1511±0,0321	0,10	0,9999±5,79e-09	0,7790±0,0176	0,77
	n_estimators	168,00						
Multi Layer Perceptron	activation	relu	0,1969±0,0162	0,0894±0,0222	0,07	0,8837±0,0021	0,8051±0,0194	0,79
	solver	lbfgs						
	alpha	1,00						
	learning_rate	constant						
	learning_rate_init	0,01						
	hidden_layer_sizes	78,00						

De acuerdo con los resultados anteriores, desde la perspectiva general en rendimiento de los modelos, el algoritmo que mejoró su desempeño fue *Naive Bayes*, aunque penaliza el ***recall*** en un -2.14% con respecto al ***escenario 1***, mejora en un 10% el accuracy, pasando del 67% de exactitud al 74%. *Regresión logística* no tiene un impacto significativo importante frente al escenario 1, sin embargo, siguen generando buenos resultados, aunque tenga un ***recall*** menor Naive Bayes.

6.2.2 Resultados con datos limpios con balanceo de clases aplicando Random Oversampling

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos aplicando la técnica de muestreo *random oversampling*.

Tabla 23. Resultado de las métricas de evaluación para cada algoritmo con Oversampling. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8188±0,0017	0,7193±0,0119	0,72	0,9089±0,0043	0,7472±0,0401	0,70
Árbol decisión	max_depth	10,00	0,8133±0,0038	0,7660±0,0184	0,76	0,8493±0,0177	0,7226±0,0244	0,70
	max_features	0,30						
Random Forest	n_estimators	132,00	0,8481±0,0024	0,8121±0,0096	0,80	0,8684±0,004	0,7124±0,0322	0,67
	max_depth	10,00						
	max_samples	0,50						
	max_features	0,30						
Naive Bayes	***alpha***	***0,99***	***0,7500±0,0019***	***0,7342±0,0078***	***0,74***	***0,7662±,0038***	***0,7623±0,0349***	***0,74***
Gradient Boosting	***max_depth***	***2,00***	***0,7864±0,0025***	***0,7984±0,0103***	***0,80***	***0,7708±0,0086***	***0,7438±0,0393***	***0,70***
	n_estimators	***127,00***						
Regresión Logística	***class_weight***	***{0: 1, 1:1}***	***0,7691±0,0020***	***0,8032±0,0102***	***0,80***	***0,7305±0,0028***	***0,7295±0,0351***	***0,70***
eXtreme Gradient Boosting	***max_depth***	***1,00***	***0,7781±0,0020***	***0,8062±0,0080***	***0,81***	***0,7450±0,0041***	***0,7336±0,0326***	***0,70***
	n_estimators	***111,00***						
Multi Layer Perceptron	***activation***	***relu***	***0,7731±0,0024***	***0,7859±0,0115***	***0,79***	***0,7571±0,0088***	***0,7506±0,0305***	***0,71***
	solver	***sgd***						
	alpha	***0,74***						
	learning_rate	***invscaling***						
	learning_rate_init	***0,06***						
	hidden_layer_sizes	***92x9***						

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8338±0,0018	0,2510±0,0177	0,24	0,9079±0,0014	0,7895±0,0181	0,76
Árbol decisión	max_depth	10,00	0,8197±0,0052	0,2805±0,0164	0,27	0,8811±0,0038	0,7652±0,0220	0,77
	max_features	0,30						
Random Forest	n_estimators	132,00	0,8511±0,0025	0,3232±0,0187	0,30	0,9273±0,0010	0,8270±0,0147	0,82
	max_depth	10,00						
	max_samples	0,50						
	max_features	0,30						
Naive Bayes	***alpha***	***0,99***	***0,7540±0,0022***	***0,2651±0,0133***	***0,26***	***0,8147±0,0018***	***0,8126±0,0190***	***0,79***
Gradient Boosting	***max_depth***	***2,00***	***0,7830±0,0037***	***0,3172±0,0196***	***0,31***	***0,8509±0,0012***	***0,8336±0,0133***	***0,83***
	n_estimators	***127,00***						
Regresión Logística	***class_weight***	***{0: 1, 1:1}***	***0,7598±0,0022***	***0,3182±0,0196***	***0,30***	***0,8317±0,0021***	***0,8288±0,0175***	***0,82***
eXtreme Gradient Boosting	***max_depth***	***1,00***	***0,7705±0,0025***	***0,3226±0,0165***	***0,31***	***0,8446±0,0013***	***0,8344±0,0134***	***0,83***
	n_estimators	***111,00***						
Multi Layer Perceptron	***activation***	***relu***	***0,7693±0,0033***	***0,3063±0,0165***	***0,30***	***0,8330±0,0022***	***0,8273±0,0189***	***0,82***
	solver	***sgd***						
	alpha	***0,74***						
	learning_rate	***invscaling***						
	learning_rate_init	***0,06***						
	hidden_layer_sizes	***92x9***						

De acuerdo con los resultados anteriores, desde la perspectiva general en rendimiento de los modelos, el algoritmo que mejoró su desempeño fue *Naive Bayes*, aunque penaliza el recall en un -5.59% con respecto al ***escenario 1***, mejora en un 10.2% el ***accuracy***, pasando del 67% de exactitud al 74%. *Regresión logística*, *Gradient Boosting, eXtreme GB* no tiene un impacto significativo importante frente al escenario 1, sin embargo, siguen generando buenos resultados junto con la inclusión del modelo *Multi Layer Perceptron*, aunque tenga un ***recall*** menor que *Naive Bayes*.

6.2.3 Resultados con datos limpios con balanceo de clases aplicando UnderSampling

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos aplicando la técnica de muestreo *random undersampling*.

Tabla 24. Resultado de las métricas de evaluación para cada algoritmo con Undersampling. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	***n_neighbors***	***29,00***	***0,7787±0,0081***	***0,7557±0,0131***	***0,76***	***0,7877±0,0090***	***0,7787±0,0329***	***0,73***
Árbol decisión	max_depth	13,00	0,9181±0,0085	0,6940±0,0158	0,71	0,9326±0,0119	0,7391±0,0422	0,69
	max_features	0,92						
Random Forest	n_estimators	148,00	0,9429±0,0024	0,7613±0,0095	0,77	0,9459±0,0051	0,7677±0,0326	0,75
	max_depth	29,00						
	max_samples	0,52						
	max_features	0,91						
Naive Bayes	***alpha***	***0,55***	***0,7573±0,0059***	***0,7306±0,0166***	***0,73***	***0,7727±0,0081***	***0,7657±0,0379***	***0,74***
Gradient Boosting	***max_depth***	***1,00***	***0,7730±0,0082***	***0,8111±0,0083***	***0,80***	***0,7247±0,0049***	***0,7233±0,0329***	***0,69***
	n_estimators	***120,00***						
Regresión Logística	***class_weight***	***{0: 1, 1:1}***	***0,7712±0,0078***	***0,7995±0,0097***	***0,79***	***0,7354±0,0035***	***0,7342±0,0340***	***0,70***
eXtreme Gradient Boosting	***max_depth***	***1,00***	***0,7824±0,0102***	***0,7954±0,0103***	***0,79***	***0,7550±0,0140***	***0,7424±0,0310***	***0,71***
	n_estimators	***100,00***						
Multi Layer Perceptron	activation	logistic	0,7699±0,0079	0,7704±0,0325	0,81	0,7627±0,0358	0,7636±0,0454	0,68
	solver	sgd						
	alpha	0,68						
	learning_rate	adaptive						
	learning_rate_init	0,74						
	hidden_layer_sizes	15x2						

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	***n_neighbors***	***29,00***	***0,7807±0,0063***	***0,2863±0,0128***	***0,28***	***0,8452±0,0082***	***0,8205±0,0180***	***0,80***
Árbol decisión	max_depth	13,00	0,9193±0,0074	0,2332±0,0146	0,23	0,9754±0,0069	0,7065±0,0257	0,71
	max_features	0,92						
Random Forest	n_estimators	148,00	0,9431±0,0025	0,2881±0,0147	0,29	0,9924±0,0004	0,8138±0,0159	0,81
	max_depth	29,00						
	max_samples	0,52						
	max_features	0,91						
Naive Bayes	***alpha***	***0,55***	***0,7610±0,0057***	***0,2636±0,0151***	***0,26***	***0,8180±0,0092***	***0,8117±0,0188***	***0,79***
Gradient Boosting	***max_depth***	***1,00***	***0,7615±0,0073***	***0,3251±0,0159***	***0,31***	***0,8420±0,0081***	***0,8331±0,0124***	***0,82***
	n_estimators	***120,00***						
Regresión Logística	***class_weight***	***{0: 1, 1:1}***	***0,7628±0,0066***	***0,3155±0,0157***	***0,30***	***0,8353±0,0086***	***0,8286±0,0166***	***0,82***
eXtreme Gradient Boosting	***max_depth***	***1,00***	***0,7763±0,0108***	***0,3135±0,0131***	***0,30***	***0,8495±0,0081***	***0,8327±0,0135***	***0,82***
	n_estimators	***100,00***						
Multi Layer Perceptron	activation	logistic	0,7679±0,0111	0,2971±0,0231	0,31	0,8320±0,0092	0,8276±0,0171	0,81
	solver	sgd						
	alpha	0,68						
	learning_rate	adaptive						
	learning_rate_init	0,74						
	hidden_layer_sizes	15x2						

De acuerdo con los resultados anteriores, desde la perspectiva general en rendimiento de los modelos, los algoritmos que mejoraron su desempeño fue *Naive Bayes y KNN*. El primero, aunque penaliza el recall en un -5% con respecto al ***escenario 1***, mejora en un 12% el ***accuracy***, pasando del 65% de exactitud al 73%. Así mismo, KNN mejora su recall en 1% y en 7% el accuracy. *Regresión logística*, *Gradient Boosting y eXtreme GB* no tiene un impacto significativo importante frente al escenario 1, sin embargo, siguen generando buenos resultados, aunque tenga un ***recall*** menor a *Naive Bayes*.

6.2.4 Resultados con datos limpios con balanceo de clases aplicando SMOTE

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos aplicando la técnica de muestreo *SMOTE*.

Tabla 25. Resultado de las métricas de evaluación para cada algoritmo con SMOTE. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8266±0,0008	0,7393±0,0099	0,73	0,9067±0,0021	0,7657±0,0302	0,70
Árbol decisión	max_depth	10,00	0,8216±0,0088	0,7840±0,0158	0,77	0,8391±0,0201	0,6577±0,0310	0,67
	max_features	0,32						
Random Forest	n_estimators	169,00	0,9804±0,0005	0,8792±0,0072	0,87	0,9924±0,0004	0,3818±0,0398	0,32
	max_depth	26,00						
	max_samples	0,51						
	max_features	0,38						
Naive Bayes	alpha	1,00	0,7626±0,0028	0,7438±0,0088	0,74	0,7787±0,0047	0,7028±0,0314	0,68
Gradient Boosting	***max_depth***	***1,00***	***0,7810±0,0021***	***0,8247±0,0079***	***0,82***	***0,7290±0,0044***	***0,7008±0,0330***	***0,66***
	n_estimators	***109,00***						
Regresión Logística	class_weight	{0: 1, 1:1}	0,7766±0,0018	0,8118±0,0082	0,80	0,7324±0,0026	0,6858±0,0294	0,65
eXtreme Gradient Boosting	max_depth	1,00	0,8201±0,0018	0,8350±0,0069	0,83	0,7955±0,0034	0,6749±0,0277	0,62
	n_estimators	100,00						
Multi Layer Perceptron	***activation***	***logistic***	***0,7744±0,0027***	***0,8020±0,0178***	***0,80***	***0,7425±0,0160***	***0,7247±0,0318***	***0,68***
	solver	***sgd***						
	alpha	***0,50***						
	learning_rate	***invscaling***						
	learning_rate_init	***0,53***						
	hidden_layer_sizes	***18x2***						

Tabla 25 - Continuación. Resultado de las métricas de evaluación para cada algoritmo con SMOTE. Fuente: Elaboración propia.

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8394±0,0008	0,2699±0,0138	0,25	0,9157±0,0012	0,8035±0,0166	0,77
Árbol decisión	max_depth	10,00	0,8245±0,0105	0,2774±0,0149	0,27	0,8935±0,0083	0,7632±0,0136	0,77
	max_features	0,32						
Random Forest	n_estimators	169,00	0,9806±0,0005	0,2842±0,0245	0,24	0,9987±6,18e-05	0,7966±0,0184	0,79
	max_depth	26,00						
	max_samples	0,51						
	max_features	0,38						
Naive Bayes	alpha	1,00	0,7663±0,0032	0,2566±0,0131	0,25	0,8333±0,00158	0,7793±0,0205	0,76
Gradient Boosting	***max_depth***	***1,00***	***0,7690±0,0028***	***0,3347±0,0190***	***0,32***	***0,8666±0,0013***	***0,8269±0,0114***	***0,82***
	n_estimators	***109,00***						
Regresión Logística	class_weight	{0: 1, 1:1}	0,7663±0,0020	0,3144±0,0174	0,29	0,8464±0,0017	0,8131±0,0181	0,79
eXtreme Gradient Boosting	max_depth	1,00	0,8155±0,0021	0,3398±0,0159	0,31	0,9011±0,0016	0,8246±0,0146	0,82
	n_estimators	100,00						
Multi Layer Perceptron	***activation***	***logistic***	***0,7669±0,0050***	***0,3162±0,0226***	***0,30***	***0,8406±0,0023***	***0,8252±0,0187***	***0,80***
	solver	***sgd***						
	alpha	***0,50***						
	learning_rate	***invscaling***						
	learning_rate_init	***0,53***						
	hidden_layer_sizes	***18x2***						

De acuerdo con los resultados anteriores, desde la perspectiva *general en rendimiento de los modelos*, *Gradient Boosting* y *Multi Layer Perceptron* fueron los de mejor desempeño bajo las condiciones ya mencionadas. La selección de variables introdujo una mejora en los modelos. Sin embargo, su métrica ***recall*** no supera los resultados obtenidos en la aplicación de los métodos de muestreo. En síntesis, el muestreo con datos sintéticos bajo *SMOTE*, no influyó de manera importante en los algoritmos de clasificación propuestos.

6.2.5 Resultados con datos limpios con balanceo de clases aplicando SMOTE-Tomek

En la presente sección, se observan los resultados correspondientes al comportamiento de los distintos algoritmos, al intentan predecir o clasificar los datos aplicando la técnica de muestreo *SMOTE-tomek*.

Tabla 26. Resultado de las métricas de evaluación para cada algoritmo con SMOTE-tomek. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8369±0,0013	0,7381±0,0099	0,73	0,9086±0,0024	0,7677±0,0271	0,70
Árbol decisión	max_depth	10,00	0,8265±0,0115	0,7632±0,0262	0,78	0,8591±0,0118	0,6769±0,0362	0,60
	max_features	0,30						
Random Forest	n_estimators	276,00	0,8626±0,0018	0,8168±0,0091	0,81	0,8713±0,0042	0,6701±0,0359	0,67
	max_depth	10,00						
	max_samples	0,50						
	max_features	0,30						
Naive Bayes	alpha	1,00	0,7724±0,0031	0,7422±0,0085	0,74	0,7857±0,0052	0,7083±0,0318	0,69
Gradient Boosting	***max_depth***	***1,00***	***0,7875±0,0020***	***0,8235±0,0082***	***0,82***	***0,7287±0,0043***	***0,7008±0,0341***	***0,66***
	n_estimators	***100,00***						
Regresión Logística	class_weight	{0: 1, 1:1}	0,7870±0,0020	0,8108±0,0080	0,80	0,7399±0,0029	0,6899±0,0306	0,65
eXtreme Gradient Boosting	max_depth	1,00	0,8297±0,0016	0,8332±0,0076	0,83	0,8017±0,0028	0,6776±0,0271	0,64
	n_estimators	100,00						
Multi Layer Perceptron	activation	relu	0,8422±0,0065	0,7675±0,0224	0,79	0,8784±0,0187	0,6735±0,0311	0,62
	solver	adam						
	alpha	0,0001						
	learning_rate	constant						
	learning_rate_init	0,001						
	hidden_layer_sizes	21x10						

			F1-score			Roc-Auc		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	n_neighbors	33,00	0,8478±0,0013	0,2695±0,0125	0,25	0,9261±0,0015	0,8046±0,0162	0,77
Árbol decisión	max_depth	10,00	0,8321±0,0089	0,2658±0,0221	0,25	0,9004±0,0076	0,7549±0,0243	0,73
	max_features	0,30						
Random Forest	n_estimators	276,00	0,8638±0,0021	0,3153±0,0199	0,31	0,9416±0,0008	0,8203±0,0131	0,81
	max_depth	10,00						
	max_samples	0,50						
	max_features	0,30						
Naive Bayes	alpha	1,00	0,7754±0,0035	0,2569±0,0132	0,25	0,8449±0,0016	0,7815±0,0205	0,76
Gradient Boosting	**max_depth**	**1,00**	**0,7743±0,0026**	**0,3333±0,0195**	**0,31**	**0,8749±0,0012**	**0,8267±0,0113**	**0,82**
	n_estimators	**100,00**						
Regresión Logística	class_weight	{0: 1, 1:1}	0,7764±0,0023	0,3146±0,0175	0,29	0,8573±0,0018	0,8134±0,0181	0,79
eXtreme Gradient Boosting	max_depth	1,00	0,8248±0,0018	0,33838±0,0169	0,32	0,9101±0,0018	0,8255±0,0137	0,81
	n_estimators	100,00						
Multi Layer Perceptron	activation	relu	0,8477±0,0070	0,2680±0,0179	0,27	0,9119±0,0064	0,7679±0,0214	0,75
	solver	adam						
	alpha	0,0001						
	learning_rate	constant						
	learning_rate_init	0,001						
	hidden_layer_sizes	21x10						

De acuerdo con los resultados anteriores, desde la perspectiva general en rendimiento de los modelos, *Gradient Boosting* fue el de mejor desempeño bajo las condiciones ya mencionadas. La selección de variables introdujo una mejora en los modelos. Sin embargo, su métrica ***recall*** no supera los resultados obtenidos en la aplicación de los métodos de muestreos (Sin contar SMOTE). En síntesis, el muestreo con datos sintéticos bajo SMOTE-tomek, no influyó de manera importante en los algoritmos de clasificación propuestos.

7. CONCLUSIONES

El proceso de detección de enfermedades de alto costo como la Diabetes Mellitus, además de impactar en la salud de la población, es un problema de manejo logístico, social y económico que afrontan los sistemas actuales de salud en el mundo. Los altos costos del manejo y detección de la enfermedad, ha llevado a ser un campo objeto de estudio de la Ciencia de Datos, que ha implicado el uso de técnicas de análisis y mejoramiento de procesos, de simulación y optimización donde se han desarrollado métodos de solución exactos, heurísticas y metaheurísticas aplicadas bajo el concepto del *Machine Learning*.

De acuerdo con lo anterior, el trabajo de grado desarrollado analiza varios modelos o técnicas que podrían predecir/clasificar la diabetes dentro del umbral de características ***No Clínicas***, posibilitando la reducción sobre la necesidad de realizar pruebas de sangre en toda la población objeto de aseguramiento, es decir, la carga del costo del diagnóstico de diabetes se puede reducir al adoptar estos métodos para complementar la prueba de diagnóstico. Los mejores Clasificadores, teniendo en cuenta la métrica de ***sensibilidad o recall*** (en concordancia o balance o con las demás métricas) fueron ***Naive Bayes, Regresión Logística, KNN y eXtreme Gradient Boosting***.

El modelo de ***Naive Bayes*** arrojó buenos resultados en los dos escenarios planteados, con la excepción en las que se aplicó con las técnicas de muestreo Smote y Smote-tomek. De acuerdo con lo anterior, la mejor combinación fue siendo en el escenario 2 (con selección de variables) y sin técnicas de balanceo de clases, obteniendo un *recall* del 75% y un *accuracy* del 74%, acompañado de un roc-auc del 79%. Es importante resaltar, que el algoritmo reaccionó eficientemente frente al problema de desbalance de clases.

Tabla 27. Mejores resultados con la aplicación del algoritmo Naive Bayes. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
Naive Bayes	*alpha*	*0,331*	*0,7348±0,0020*	*0,7335±0,0075*	*0,74*	*0,7673±0,0053*	*0,7636±0,0362*	*0,75*
			F1-score			**Roc-Auc**		
			Entrenam.	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
			0,2667±0,0013	*0,2649±0,0134*	*0,26*	*0,8155±0,0019*	*0,8130±0,0187*	*0,79*

El modelo de ***Regresión Logística*** arrojó buenos resultados en los dos escenarios planteados, con la excepción en las que se aplicó con las técnicas de muestreo Smote y Smote-tomek. De acuerdo con lo anterior, la mejor combinación fue siendo en el escenario 2 (con selección de variables) y sin técnicas de balanceo de clases, obteniendo un *recall* del 73% y un *accuracy* del 77%, acompañado de un roc-auc del 82%. Es importante resaltar, que el algoritmo reaccionó eficientemente frente al problema de desbalance de clases dado que se realizó refinamiento de los pesos dadas la clase a predecir/clasificar. Adicionalmente, aunque éste es uno de los algoritmos básicos de clasificación, tienen alta eficiencia en su implementación con buenos resultados.

Tabla 28. Mejores resultados con la aplicación del algoritmo Regresión Logística. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
Regresión Logística	*C*	*1,000*	*0,7711±0,0024*	*0,7707±0,0053*	*0,77*	*0,7684±0,0028*	*0,7636±0,0263*	*0,73*
			F1-score			**Roc-Auc**		
	class_weight	*{0: 1, 1:20}*	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
			0,2968±0,0223	*0,2951±0,0110*	*0,29*	*0,8326±0,0019*	*0,8286±0,0174*	*0,82*

El modelo de ***KNN*** arrojó buenos resultados con la combinación del escenario 2 (con selección de variables) con la aplicación de muestreo ***Undersampling*** para el balance de clases, obteniendo un *recall* del 73% y un *accuracy* del 76%, acompañado de un roc-auc del 80%.

Tabla 29. Mejores resultados con la aplicación del algoritmo KNN. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
KNN	*n_neighbors*	*29,000*	*0,7787±0,0081*	*0,7557±0,0131*	*0,76*	*0,7877±0,0090*	*0,7787±0,0329*	*0,73*
			F1-score			**Roc-Auc**		
			Entrenam.	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
			0,7807±0,0063	*0,2863±0,0128*	*0,28*	*0,8452±0,0082*	*0,8205±0,0180*	*0,80*

El modelo ***eXtreme Gradient Boosting*** tuvo influencia en su rendimiento al usar técnicas de muestreo, los resultados fueron similares con la aplicación de *Underdampling* y *Oversampling* en los dos escenarios, es decir, no tuvo impacto al realizar o no selección de variables. De acuerdo con lo anterior, postulamos como mejor resultado, el correspondiente al escenario 2 (con selección de variables) con aplicación de *Undersampling*, obteniendo un *recall* del 71% y un *accuracy* del 79%, acompañado de un roc-auc del 82%.

Tabla 30. Mejores resultados con la aplicación del algoritmo XGBoosting. Fuente: Elaboración propia.

			Exactitud (Accuracy)			Sensibilidad (Recall)		
Algoritmo	**Hiperparám.**	**Valor óptimo**	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
eXtreme Gradient Boosting	***max_depth***	***1,000***	***0,7824±0,0102***	***0,7954±0,0103***	***0,79***	***0,7550±0,0140***	***0,7424±0,0310***	***0,71***
			F1-score			**Roc-Auc**		
	n_estimators	***100,000***	**Entrenam.**	**Validación**	**Prueba**	**Entrenam.**	**Validación**	**Prueba**
			0,7763±0,0108	***0,3135±0,0131***	***0,30***	***0,8495±0,0081***	***0,8327±0,0135***	***0,82***

En el desarrollo del trabajo de grado, se probaron múltiples escenarios y combinaciones en flujo de trabajo con el objetivo de mejorar las métricas propuestas para la detección de DMT2. Sin embargo, al confrontar los casos de éxitos de trabajos realizados como los de Mathew & Sherly (2018) y Srivastava, Kumar, Fore, & Tomar (2021), se observa que dentro del *pull* de variables utilizadas como ***no clínicas,*** están las concernientes a las características físicas de la persona como el grosor de la cintura, IMC (índice de grasa corporal), presión arterial sistólica y diastólica, tipo de comida preferida y horas de sueño. Variables con las cuales no contábamos en la base de datos relacionada por el EPS. Por ende, surge la hipótesis que los modelos propuestos en el presente trabajo de grado realizado se pueden mejorar con la inclusión de éstas. Lo anterior, quedará como trabajo a futuro.

Por último, con toda la propuesta y desarrollo mostrado, es nuestro primer aporte como futuros Científicos de Datos que responde a la pregunta planteada en ***la sección 1.3***:

¿Cómo podemos apoyar en la detección temprana de la diabetes mellitus tipo 2 (DMT2) mediante un modelo supervisado que involucre variables no clínicas identificadas a partir de las características de una población perteneciente al régimen subsidiado?

8. RECOMENDACIONES Y ESTUDIOS FUTUROS

En cuanto al uso del modelo en el entorno de producción, una opción que permitirá mejorar la recopilación de datos y a la detección temprana de la enfermedad sería una aplicación de M-Health, la cual consiste en aplicaciones que normalmente hacen seguimiento a indiciadores de salud de una persona tales como el peso, la dieta, el nivel de ejercicio que realizan permitiendo al usuario poder monitorear sus hábitos. Algunas de estas aplicaciones pueden compartir estos datos con entidades prestadoras de salud (El-Sappagh, y otros, 2019).

Como recomendación para la encuesta de caracterización poblacional que realiza la EPS, importante incluir información concerniente a las características físicas del afiliado, como el grosor de la cintura, IMC (índice de grasa corporal), presión arterial sistólica y diastólica, tipo de comida preferida y horas de sueño. Dado que los casos de éxito de diversas investigaciones contemplan este tipo de información.

Surge incluir en este tipo de investigaciones, el apoyo de un pull de profesionales médicos o junta multidisciplinaria médica, con el fin de revisar todas las variables de la encuesta desde el punto de vista técnico médico. En el trabajo de grado desarrollado, no se incluyó ningún apoyo con expertos médicos, con el objetivo de contemplar la mayor cantidad de información posible que junto al estado del arte, nos permitió obtener información para modelar los distintos algoritmos sin la inducción de un “sesgo médico”. Sería interesante realizar selección de variables con connotación médica y entrenar todos los algoritmos nuevamente. Además, de contemplar otras enfermedades de base que pueden tener los afiliados, dado que para el trabajo de grado se tomó como segunda enfermedad de base la Hipertensión.

Para estudios posteriores se debería considerar la aplicación de nuevos hiperparámetros, sobre todo los concernientes a la penalización de clases (para los que aplique) para afrontar el problema de balance entre el número de pacientes sanos vs enfermos, como se realizó con el algoritmo de Regresión Logística. Adicionalmente, explorar ensambles de algoritmos.

Por último, considerar explorar la distancia entre la ubicación del afiliado y la ciudad con IPS de nivel medio o alta a la cual podría acceder y ser atendido. La idea de obtener esta información es si la localización tiene alguna relación con la variable de respuesta “AC_DIABETES”, además que podría captar información adicional como, el coste de viajar a la ciudad para hacerse pruebas de prevención o seguimiento versus el costo de no tratamiento para identificación temprana.

BIBLIOGRAFÍA

Ahmed Osman, A., Ahmed, A., Chow, M., & Huang, Y. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater. *Ain Shams Engineering Journal*. doi:https://doi.org/10.1016/j.asej.2020.11.011

Avilés-Santa, M. L., Monroig-Rivera, A., Soto-Soto, A., & Lindberg, N. M. (2020). Current State of Diabetes Mellitus Prevalence, Awareness, Treatment, and Control in Latin America: Challenges and Innovative Solutions to Improve Health Outcomes Across the Continent. *Springer Nature*. doi:https://doi.org/10.1007/s11892-020-01341-9

Bayati, M., Bhaskar, S., & Montanari, A. (2015). A Low-Cost Method for Multiple Disease Prediction. *AMIA Annu Symp Proc*.

Breiman, L. (2001). Random Forests. *Kluwer Academic Publishers. Manufactured in The Netherlands*.

Castrillón, O. D., Sarache, W., & Castaño, E. (2017, Diciembre). Sistema bayesiano para la predicción de la diabetes. *Inf. Tecnol, vol 28*, 161-168.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0. Step-by-step data mining guide.* SPSS Inc. doi:https://doi.org/10.1017/CBO9781107415324.004

Chawla, N. (2010). *Data Mining and Knowledge Discovery Handbook - Data Mining for Imbalanced Datasets: An Overview.* Boston: Springer, Boston, MA. doi:https://doi.org/10.1007/978-0-387-09823-4_45

Cohen, J., Cohen, P., West, S., & Aiken, L. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences.* New York: Routledge. doi:https://doi.org/10.4324/9780203774441

El-Sappagh, S., Ali, F., El-Masri, S., Kim, K., Ali, A., & Kwak, S. (2019). Mobile Health Technologies for Diabetes Mellitus: Current State and Future Challenges. doi:10.1109/ACCESS.2018.2881001

Fezeka Swana, E., Doorsamy, W., & Bokoro, P. (2022). Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. *MDPI - Academic Open Access Publishing*. doi:https://doi.org/10.3390/s22093246

Fondo Colombiano de Enfermedades de Alto Costo. (2022). *Infografía Día Mundial de la Diabetes.* Obtenido de https://cuentadealtocosto.org/site/general/dia-mundial-de-la-diabetes-2022/

Gardner, M., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*. doi:https://doi.org/10.1016/S1352-2310(97)00447-0

Gómez-Encino, G. d., Cruz-León, A., Zapata-Vázquez, R., & Morales- Ramón, F. (2015). Nivel de conocimiento que tienen los pacientes con Diabetes Mellitus tipo 2 en relación a su enfermedad. *Salud en Tabasco*, 17-25.

Hajjaj, F., Salek, M., Basra, M., & Finlay, A. (2010). Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice. *In Journal of the Royal Society of Medicine*, (Vol. 103, Issue 5, pp. 178–187).

Han, J., Rodriguez, J. C., & Beheshti, M. (2008). Diabetes data analysis and prediction model discovery using rapidminer. *2008 Second International Conference on Future Generation Communication and Networking*, 96-99. doi:10.1109/FGCN.2008.226

Hong Chen, Songhua Hu, Rui Hua, & Xiuju Zhao. (2021). Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*. doi:https://doi.org/10.1186/s13634-021-00742-6

Jacobs-Basadien, M., Pather, S., & Petersen, F. (2022). The role of culture in the adoption of mobile applications for the self-management of diabetes in low resourced urban communities. Obtenido de https://nebulosa.icesi.edu.co:2144/10.1007/s10209-022-00951-2

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*.

Leiva, A. M., Martínez, M. A., Petermann, F., Garrido Méndez, A., Poblete Valderrama, F., Díaz Martínez, X., & Celis Morales, C. (2018). Risk factors associated with type 2 diabetes in Chile. *Nutrición Hospitalaria, 35*(2), 400-407. doi:https://dx.doi.org/10.20960/nh.1434

Martínez Pérez, J., & Pérez Martin, P. (2022). La curva ROC. *Elservier*. doi:10.1016/j.semerg.2022.101821

Mathew, T. J., & Sherly, E. (2018). Analysis of Supervised Learning Techniques for Cost Effective Disease Prediction Using Non-clinical Parameters. *2018 International CET Conference on Control, Communication, and Computing (IC4)*.

Mejía, J., Oviedo, M., Ordonez, A., & Valencia, J. F. (2022). Prediction of Diabetes based on environmental and socioeconomic information.

Ministerio de Salud y Protección Social. (2021, 10 18). *Ministerio de Salud y Protección Social*. Retrieved from Ministerio de Salud y Protección Social: https://www.minsalud.gov.co/Paginas/Prevenir-la-diabetes-clave-desde-los-habitos-saludables.aspx

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *ResearchGate*. doi:10.1109/ICICS49469.2020.239556

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. doi:doi: 10.3389/fnbot.2013.00021

Organización Mundial de la Salud. (2016). *Informe Mundial sobre la Diabetes.* Ginebra: Se reservan todos los derechos.

Organización Panamericana de la Salud. (2020). *Diagnóstico y manejo de la diabetes de tipo 2 (HEARTS-D).*

Pan American Health Organization. (2022). *Panorama of Diabetes in the Americas.* Washington D.C.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (s.f.). *Scikit-learn: Machine Learning in Python*. Obtenido de Journal of Machine Learning Research: http://jmlr.org/papers/v12/pedregosa11a.html

Pérez-Gandía, C. (Marzo de 2014). Propuesta de algoritmos de predicción de glucosa en pacientes diabéticos. Madrid.

Rajaguru, H., & Chakravarthy, S. (2019). Analysis of Decision Tree and K-Nearest Neighbor Algorithm in the Classification of Breast Cancer. *Asian Pac J Cancer Prev*. doi:10.31557/APJCP.2019.20.12.3777

Ramón, A., Torres, A., Milara, J., Cascón, J., Blasco, P., & Mateo, J. (2022). eXtreme Gradient Boosting-based method to classify patients with COVID-19. *Journal of Investigative Medicine*. doi:http://dx.doi.org/10.1136/jim-2021-002278

Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning. Aprendizaje automático y aprendizaje profundo con Python, scikit-learn y TensorFlow.* MARCOMBO, S.A.

Rocha Íñigo, A. (2020). *Codificación de variables categóricas en aprendizaje automático.* Tesis Máster, Universidad de Sevilla, Depto. de Ingeniería de Sistemas y Automática, Sevilla.

Obtenido de https://idus.us.es/bitstream/handle/11441/108887/M1909%20Rocha%20%c3%8d%c3%b1igo%2c%20Adri%c3%a1n.pdf?sequence=1&isAllowed=y

Saria, S., Rajani, A. K., Gould, J., Koller, D., & Penn, A. A. (2010). Integration of early physiological responses predicts later illness severity in preterm infants. *Science Translational Medicine.*, 2(48):48ra65. doi:10.1126/scitranslmed.3001304

Sharma, N., & Singh, A. (2019). *Diabetes Detection and Prediction Using Machine Learning/IoT: A Survey.* Springer Singapore.

Snoek, J., Larochelle, H., & Adams, R. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*. doi:https://doi.org/10.48550/arXiv.1206.2944

Song, X., Mitnitski, A., Cox, J., & Rockwood, K. (2004). Comparison of machine learning techniques with classical statistical models in predicting health outcomes. Stud Health Technol Inform. *Stud Health Technol Inform.*, 107(Pt 1):736-40.

Song, Y.-y., & Lu, Y. (2015). Decision tree methods: applications for classification. *Shanghai Arch Psychiatry*. doi:10.11919/j.issn.1002-0829.215044

Srivastava, R., Kumar, S., Fore, V., & Tomar, R. (2021). A Study of Five Models Based on Non-clinical Data for the Prediction of Diabetes Onset in Medically Under-Served Populations. *Springer Nature Switzerland AG*, 116–124. doi:https://doi.org/10.1007/978-3-030-88244-0_12

Villalobos, A., Rojas-Martínez, R., Aguilar-Salinas, C. A., Romero-Martínez, M., Mendoza-Alvarado, L. R., Flores-Luna, M. d., . . . Ávila-Burgos, L. (2019). Atención médica y acciones de autocuidado en personas que viven con diabetes, según nivel socioecnómico. *Salud Publica Mex.*, 876-887.

Zapeta Hernández, A., Galindo Rosales, G., Juan Santiago, H., & Martínez Lee, M. (2022). Métricas de rendimiento para evaluar el aprendizaje automático en la clasificación de imágenes petroleras utilizando redes neuronales convolucionales. *Ciencia Latina Revista Científica Multidisciplinar*. doi:https://doi.org/10.37811/cl_rcm.v6i5.3420

ANEXOS

ANEXO 1: Descripción de las variables iniciales de la base de datos

Variables	Descripción
FILA	No. de registro
id_afiliado	1, si el afiliado pertenece a la EPS - 0, lo contrario
id_person_historic	No. De identificación histórica en la EPS
Fecha de nacimiento	Fecha de nacimiento del afiliado
Edad	Edad del afiliado al 2023
Grupo_etario	Grupo etario
Sexo	Sexo
Tipo de sangre	Tipo de sangre
Factor RH	Factor RH
Fecha caracterización	Fecha de inicio de la encuesta
Fecha de fin de caracterización	Fecha fin de la encuesta
acepta la toma de medidas	Si, si el afiliado acepta la toma de medidas - No, lo contrario
Latitud y longitud	Coordenadas donde el afiliado reside
Código municipio	Identificación del municipio
Nombre municipio	Nombre municipio donde reside
Código departamento	Identificación del departamento
Nombre departamento	Nombre departamento donde reside
Buscar barrio vereda	Si, si reside en vereda - No, lo contrario
Barrio vereda	Nombre de la vereda
Tipo de zona	Tipo de zona en la que reside
Dirección con nomenclatura	Dirección de la residencia
Vía de acceso	Infraestructura de la vía de acceso
Tipo de Vivienda	Tipo de vivienda
Material paredes exteriores	Infraestructura de la vivienda
Material de los pisos	Material de los pisos de la vivienda
Energía eléctrica	Si, si tiene energía eléctrica - No, lo contrario
Estrato	Estrato socioeconómico
Alcantarillado	Si, cuenta con alcantarillado - No, lo contrario
Recolección de basuras	Si, cuenta con recolección de basura - No, lo contrario
Acueducto	Si, cuenta con acueducto - No, lo contrario
Iluminación natural o artificial suficiente	Si, cuenta con iluminación suficiente - No, lo contrario
Ventilación adecuada	Si, cuenta con ventilación adecuada - No, lo contrario
Reservorios de agua	Si, cuenta con reservorio de agua - No, lo contrario
Presencia de insectos vectores	Si, hay presencia de insectos - No, lo contrario
Aseo adecuado de la vivienda	Si, aseo adecuado - No, lo contrario
Cuantos hogares alimentos por separado	No. De hogares que conviven en la vivienda
Cuantos cuartos en total dispone hogar	No. De cuartos que tiene la vivienda
Cuantos cuartos para dormir	No. De cuartos destinados para dormir
Síntomas últimos 30 días	Afectaciones de la vivienda en los últimos 30 días
Tipo de servicio sanitario en el hogar	Infraestructura del sistema sanitario de la vivienda

Variables	Descripción
Fuente principal del agua para alimentos	Tipo de fuente de acceso a agua para preparar alimentos
Tratamiento del agua para beber	Tratamiento del agua que realizan para el consumo
En donde preparan los alimentos	Lugar donde preparan los alimentos
Energía combustible utilizan para cocinar	Tipo de combustible utilizado para cocinar
Tienen servicio de teléfono fijo	Si, tiene servicio de teléfono fijo - No, lo contrario
Ingresos económicos mensuales del hogar	Rango de ingresos que tiene el hogar (SMLV)
Actividad agropecuaria	Si, realiza actividad agropecuaria - No, lo contrario
Utiliza químicos en actividad agropecuaria	Si, utiliza químicos en la actividad agropecuaria - No, lo contrario
disposición final de sobrantes	Tratamiento que realizan con los sobrantes
tiene perros gatos	Si, tiene mascotas (perro / gatos) - No, lo contrario
Perros gatos están vacunados	Si, las mascotas están vacunadas - No, lo contrario
Residente Habitual	Residente Habitual
Cantidad de personas en el hogar	Cantidad de personas que conviven en el hogar
Grupo étnico	Grupo étnico al que pertenece el afiliado
Grupos poblacionales se identifica	Tipo de grupo poblacional
Estado civil	Estado civil del afiliado
Con quien permanece el menor	Acompañante del menor en el hogar
Menor asiste atención de primera infancia	Si, el menor asiste a atención de primera infancia - No, lo contrario
Sabe leer y escribir	Si, sabe leer - No, lo contrario
Nivel educativo mas alto	Nivel educativo del afiliado
Condición de trabajo actual	Descripción general del tipo de oficio del afiliado
Ocupación de la persona	Descripción detallada del trabajo que realiza
Tiene celular	Si, tiene celular - No, lo contrario
Discapacidad física o mental	Si, tiene discapacidad - No, lo contrario
Actividades básicas diarias se ven afectadas	Tipo de actividad en la que se ve afectado por discapacidad
ceguera total	Si, tiene ceguera total - No, lo contrario
sordera total	Si, tiene sordera total - No, lo contrario
Consumo de sustancia	Tipo de sustancia que consume el afiliado
Fumo en los últimos 12 meses	Si, el afiliado fumó en los últimos 12 meses - No, lo contrario
Consumo droga por vía inyectada	Si, ha consumido - No, lo contrario
Patrón de inyección	Descripción de frecuencia de consumo
Actividad física 30 min	Si, realiza actividad física - No, lo contrario
Frecuencia consume frutas y verduras	Descripción de frecuencia de consumo de frutas
Consumo entre 4 y 8 vasos agua día	Descripción de frecuencia de consumo de agua
Adiciona sal a las comidas	Si, adiciona sal a la comida - No, lo contrario
Adiciona azúcar comidas	Si, adiciona azúcar a la comida - No, lo contrario
lava manos antes de comer	Si, se lava las manos - No, lo contrario
Familiar con enfermedad	Descripción del tipo de enfermedad que padece el familiar
Familiar con Diabetes Mellitus	Descripción de acuerdo al familiar que tiene o no Diabetes Mellitus
Familiar con Cáncer	Descripción de acuerdo al familiar que tiene cáncer

Variables	Descripción
Tipo cáncer familiar 1 grado	Tipo de cáncer que tiene el familiar en grado 1
Padecimiento enfermedad	Tipo de enfermedad que padece el afiliado
Medicamento de enfermedad cardiaca	Si, el afiliado toma medicamento de enfermedad cardiaca - No, lo contrario
Medicamento de tensión o presión alta	Si, el afiliado toma medicamento de tensión o presión alta - No, lo contrario
Asiste control de hipertensión arterial	Si, el afiliado asiste a control de hipertensión arterial - No, lo contrario
Medicamento enfermedad renal	Si, el afiliado toma medicamento de enfermedad renal - No, lo contrario
Medicamentos enfermedad de diabetes	Si, el afiliado toma medicamento de enfermedad de diabetes - No, lo contrario
Asiste a control de diabetes	Si, asiste a control de diabetes - No, lo contrario
Medicamentos control cáncer	Si, el afiliado toma medicamento de control de cáncer - No, lo contrario
Medicamentos control artritis	Si, el afiliado toma medicamento de control de artritis - No, lo contrario
Medicamentos control EPOC	Si, el afiliado toma medicamento de control de EPCO - No, lo contrario
Medicamentos control VIH-SIDA	Si, el afiliado toma medicamento de control de VIH-SIDA - No, lo contrario
Medicamentos control Enf Mental	Si, el afiliado toma medicamento de control de Enf Mental - No, lo contrario
Medicación hipertensión regularmente	Si, el afiliado toma medicamento de hipertensión regularmente - No, lo contrario
Recibe insulinas	Si, el afiliado recibe insulina - No, lo contrario
Asiste control Artritis Reumatoidea	Si, asiste a control artritis reuma - No, lo contrario
Recibió tratamiento para tuberculosis	Descripción sobre bajo que condiciones recibe tratamiento o no
Valores de glucosa altos	Si, tiene valor alto de glucosa - No, lo contrario
Ha utilizado oxigeno medicinal en casa	Si, el afiliado ha utilizado oxígeno medicinal en casa - No, lo contrario
Inflamación en articulaciones	Descripción sobre en qué articulación tiene inflamación o si no ha presentado
Tipo de Cáncer padece o padecido	Tipo de cáncer que padece o ha padecido el afiliado
Estado salud últimos 30 días	Descripción sobre como el afiliado ha estado en términos de salud en los últimos 30 días
Comparación estado salud	Descripción sobre como el afiliado ha estado en términos de salud
Tiene dificultad visual	Si, el afiliado tiene dificultad visual - No, lo contrario
Dificultad para oír	Si, el afiliado tiene dificultad para escuchar - No, lo contrario
Tos con expectoración	Si, el afiliado tiene Tos - No, lo contrario
baciloscopia para tuberculosis	Si, el afiliado le han realizado baciloscopia para tuberculosis - No, lo contrario
Piel blancas o rojizas	Si, el afiliado tiene piel blanca o rojiza - No, lo contrario
Lesión en piel últimos 15 días	Descripción sobre la lesión de piel que tiene o no en los últimos 15 días
Ha sido mordido animales	Si, el afiliado ha sido mordido por animales - No, lo contrario
Recibió vacuna antirrábica	Descripción sobre si el afiliado recibió o no vacuna antirrábica
Síntomas después mordedura	Descripción sobre los síntomas que tuvo el afiliado después de la mordedura
Expuesto a humo de tabaco	Si, el afiliado ha sido expuesto a humo de tabaco - No, lo contrario
Expuesto a humo de leña	Si, el afiliado ha sido expuesto a humo de leña - No, lo contrario
signos o síntomas respiratorios Año	Descripción sobre los síntomas respiratorios que tuvo el afiliado o no en el año
signos o síntomas respiratorios	Descripción sobre los síntomas respiratorios que tuvo el afiliado o no
signos niño	Descripción sobre los síntomas que tuvo el afiliado menor de edad
Vida sexual activa	Si, el afiliado tiene vida sexual activa - No, lo contrario
Perdida capacidad del habla	Si, el afiliado ha perdido capacidad del habla - No, lo contrario
debilidad entumecimiento cuerpo	Si, el afiliado tiene entumecimiento del cuerpo - No, lo contrario

Variables	Descripción
Palpitaciones en el pecho	Si, el afiliado ha tenido palpitaciones en el pecho - No, lo contrario
Dolor opresivo en el pecho	Si, el afiliado ha tenido dolor opresivo en el pecho - No, lo contrario
Dificultad para respirar o sensación de ahogo	Si, el afiliado ha tenido dificultad para respirar o sensación de ahogo - No, lo contrario
Perdida de la fuerza en manos pies	Si, el afiliado ha perdido fuerza en las manos / pies - No, lo contrario
sudoración fría y palidez	Si, el afiliado ha tenido sudoración fría y palidez - No, lo contrario
relaciones sexuales1	Si, el afiliado ha tenido relaciones sexuales - No, lo contrario
relaciones sexuales2	Si, el afiliado ha tenido relaciones sexuales - No, lo contrario
IPS	IPS de asignación o atención primaria
AC_DIABETES	Indicador de Diabetes / 1, el afiliado la padece - 0, lo contrario
AC_DIALISIS	Indicador de Diálisis / 1, el afiliado ha sido dializado - 0, lo contrario
AC_HIPERTENSION	Indicador de Hipertensión / 1, el afiliado la padece - 0, lo contrario
AC_TRASPLANTE RENAL	Indicador de Trasplante renal / 1, el afiliado ha tenido trasplante - 0, lo contrario
Departamento Afiliación	Descripción del departamento donde se realizó la afiliación
Municipio Afiliación	Descripción del municipio donde se realizó la afiliación