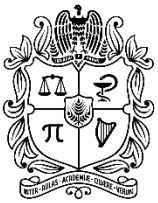
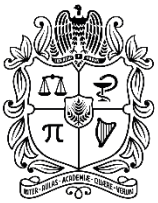


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Definición de un modelo de clasificación de riesgo cardiovascular para una población de adultos mayores usando técnicas de aprendizaje de máquinas

Manuela Londoño Ocampo

Universidad Nacional de Colombia
Facultad de Minas, Área Curricular de Sistemas e Informática
Medellín, Colombia
2020



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Definition of a cardiovascular risk classification model for a population of older adults using machine learning techniques

Manuela Londoño Ocampo

National University of Colombia
Faculty of Mines, Systems and Informatics Curriculum Area
Medellín, Colombia
2020

Definición de un modelo de clasificación de riesgo cardiovascular para una población de adultos mayores usando técnicas de aprendizaje de máquinas

Manuela Londoño Ocampo

Trabajo de grado presentado como requisito parcial para optar al título de:

Magíster en Ingeniería – Ingeniería Analítica

Director:

PhD. Fernán Alonso Villa Garzón

Universidad Nacional de Colombia

Facultad de Minas, Área Curricular de Sistemas e Informática

Medellín, Colombia

2020

Resumen

Según cifras reportadas por la organización mundial de la salud (OMS), las enfermedades cardiovasculares son la principal causa de muerte en el mundo, el riesgo de padecerlas aumenta en adultos mayores y ante la presencia de factores de riesgo como la hipertensión, diabetes, obesidad y tabaquismo; en Colombia la situación es similar. Para la estimación del riesgo cardiovascular se han creado diferentes métodos que analizan el desarrollo de este tipo de enfermedades según el grado de exposición a algunos factores de riesgo, pero estos no suelen ser precisos en todas las poblaciones. El aprendizaje de máquinas ha demostrado su valor de aplicación en contextos médicos, siendo una herramienta novedosa y alternativa que apoya tareas complejas como el diagnóstico de enfermedades. En el presente trabajo se consideran tres modelos de aprendizaje de máquinas, usados en tareas de clasificación que, según la revisión de la literatura desarrollada, pueden ser óptimos en información clínica, con lo que se pretende identificar para una población colombiana de adultos mayores el riesgo asociado al padecimiento de enfermedades cardiovasculares. Para la evaluación del desempeño de los modelos propuestos se utilizan criterios de bondad de ajuste como exactitud, sensibilidad, precisión y f1-score, donde el modelo óptimo se obtiene con el algoritmo de máquina de aumento de gradiente, con un rendimiento mejor a los reportados por estudios similares referenciados.

Palabras clave: Aprendizaje de máquinas, riesgo cardiovascular, clasificación, informática de la salud.

Abstract

Figures of the World Health Organization (WHO) demonstrated that cardiovascular disease (CVD) is the first cause of death worldwide. The risk of suffering is higher in older people and increases with some preview's pathologies like hypertension, diabetes, obesity, and smoking. The tendency is equal in Colombia. To estimate cardiovascular risk, different methods have been created that analyze the development of this type of disease according to the degree of exposure to different risk factors, but these are not usually accurate in all populations. Machine learning has proven its application value in medical contexts, being a novel and alternative tool that supports complex tasks such as disease diagnosis. For the present work, three models of machine learning are considered, used in classification tasks that, according to the review of the developed literature may be optimal for clinical information with which it is intended to identify risk for a Colombian population of older adults associated with the suffering of cardiovascular diseases. For the evaluation of the performance of the proposed models, goodness of fit criteria such as accuracy, sensitivity, precision and f1-score are used, where the optimal model is obtained with the gradient boosting machine algorithm, with a better performance than those reported by similar studies referenced.

Keywords: Machine Learning, cardiovascular risk, classification, health informatics.

Contenido

Resumen ..... **IV**

Contenido ..... **VI**

Lista de tablas ..... **VIII**

Lista de figuras ..... **IX**

Introducción **1**

1. Aprendizaje de máquinas para información médica **4**

1.1 Exploración de la literatura **5**

2.* Entendimiento del problema abordado** ***13

2.1 Contextualización del problema **14**

2.1.1 Envejecimiento de la población 15

2.1.2 Riesgo cardiovascular 16

2.1.3 Estimación del riesgo cardiovascular 17

2.2 Recursos disponibles **21**

2.3 Inventario de modelos **21**

2.3.1 Regresión logística: 22

2.3.2 Máquinas de Aumento de Gradiente (Gradient Boosting Machines) 22

2.3.3 Bosque Aleatorio 23

2.4 Requerimientos y restricciones **23**

2.4.1 Confidencialidad de la información 23

2.4.2 Ética de la investigación en salud 24

2.4.3 Transformación de los datos 24

2.5 Objetivos **25**

2.5.1 Objetivo General 25

2.5.2 Objetivos Específicos 25

2.6 Metodología CRISP-DM **25**

2.6.1 Comprensión del negocio 26

2.6.2 Comprensión de los datos 27

2.6.3 Preparación de los datos 27

2.6.4 Modelado 27

2.6.5 Evaluación 27
2.6.6 Implementación 28

3. Entendimiento de los datos 29

3.1 Información disponible 31

3.2 Exploración de los datos 31
3.2.1 Proceso de limpieza 31
3.2.2 Variables Derivadas 37
3.2.3 Análisis descriptivo de las variables 38
3.2.4 Análisis de correlación 43

4. Modelado y evaluación 47

4.1 Preparación de los datos 47

4.2 Selección de variables 48

4.3 Partición de los datos y métricas de evaluación 49

4.4 Optimización de hiperparámetros 51

4.5 Resultados 52

5. Discusión 57

6. Conclusiones 60

6.1 Objetivo 1. Caracterizar variables de evaluación físicas y antropométricas para una población de estudio de adultos mayores 60

6.2 Objetivo 2. Seleccionar variables de evaluación físicas y antropométricas para una población de estudio de adultos mayores, según su relevancia para el modelo de clasificación 60

6.3 Objetivo 3. Construir un modelo híbrido para la clasificación del riesgo asociado al padecimiento de enfermedades cardiovasculares que pueda ser aplicado en una población de adultos mayores. 61

6.4 Objetivo 4. Validar el modelo de clasificación de grupos poblacionales de adultos mayores según el nivel de riesgo asociado al padecimiento de enfermedades cardiovasculares mediante técnicas de validación cruzada. 62

7. Referencias 63

Lista de tablas

Tabla 1. Exploración de la literatura 5
Tabla 2. Variables predictoras identificadas en la exploración de la literatura referente a riesgo cardiovascular 11
Tabla 3. Indicadores de calidad para la base de datos de estudio 32
Tabla 4. Hiperparámetros óptimos para el modelo de regresión logística de mejor desempeño 51
Tabla 5. Hiperparámetros óptimos para el modelo de bosque aleatorio de mejor desempeño 52
Tabla 6. Hiperparámetros óptimos para el modelo de máquina de aumento de gradiente de mejor desempeño 52
Tabla 7. Resultado de las corridas de los modelos base 53
Tabla 8. Resultado de las métricas de desempeño para el algoritmo de regresión logística con diferentes estrategias de SMOTE 54
Tabla 9. Resultado de las métricas de desempeño para el algoritmo de bosque aleatorio con diferentes estrategias de SMOTE 54
Tabla 10. Resultado de las métricas de desempeño para el algoritmo de máquina de aumento de gradiente con diferentes estrategias de SMOTE 54
Tabla 11. Resultado de las métricas de desempeño para los algoritmos de regresión logística y bosque aleatorio con hiperparámetro class weight ajustado 55
Tabla 12. Resultado de las métricas de desempeño para los algoritmos de clasificación considerados con ajuste de hiperparámetros 55
Tabla 13. Resultado de las métricas de desempeño por clase para el algoritmo de máquina de aumento de gradiente 56

Lista de figuras

Figura 1. Tabla de predicción de riesgo cardiovascular para las Américas (AMR) grupo B de la OMS/ISH, para los contextos en que no se puede medir el colesterol sanguíneo. 19
Figura 2 Tabla de predicción de riesgo cardiovascular para las Américas (AMR) grupo B de la OMS/ISH, para los contextos en que se puede medir el colesterol sanguíneo 20
Figura 3. Fases de la metodología CRIP-DM 26
Figura 4. Flujo de Knime® de la etapa 1: Limpieza de datos 29
Figura 5. Flujo de Knime® de la etapa 2: Tratamiento de valores ausentes y valores atípicos 30
Figura 6. Flujo de Knime® de la etapa 3: Transformación de datos y variables derivadas 30
Figura 7. Diagramas de dispersión y Diagramas de caja y bigotes para la Talla (cm) 35
Figura 8. Diagramas de dispersión y Diagramas de caja y bigotes para el Peso 35
Figura 9. Diagramas de dispersión y Diagramas de caja y bigotes para el Perímetro de la cintura 36
Figura 10. Diagramas de dispersión y Diagramas de caja y bigotes para el Perímetro de la cadera máxima 36
Figura 11. Diagramas de dispersión y Diagramas de caja y bigotes para la Frecuencia Cardíaca 37
Figura 12. Número de muestras para cada nivel de riesgo cardiovascular 39
Figura 13. (1) Número de muestras por Género - (2) Riesgo cardiovascular por género. 39
Figura 14. Distribución de la variable respuesta según la edad 40

Figura 15. (1) Número de muestras por padecimiento de diabetes - (2) Riesgo cardiovascular por padecimiento de diabetes. 41
Figura 16. (1) Número de muestras por padecimiento de hipertensión - (2) Riesgo cardiovascular por padecimiento de hipertensión. 42
Figura 17. (1) Número de muestras según el hábito de fumar - (2) Riesgo cardiovascular según el hábito de fumar 42
Figura 18. Medidas antropométricas para cada nivel de riesgo cardiovascular 43
Figura 19. Matriz de correlación entre variables numéricas usando coeficiente de Spearman 44
Figura 20. Gráficos de dispersión entre pares de variables correlacionadas 45
Figura 21. Asociación entre variables categóricas usando V de Cramer 46
Figura 22. Exactitud del modelo estimador para cada conjunto de características consideradas. 48
Figura 23. Importancia de las variables seleccionadas por el RFE con validación cruzada 49
Figura 24. Matriz de confusión para un problema de clasificación multiclase 51

Introducción

La adopción de tecnologías de información y comunicación en diferentes sectores e industrias ha generado cambios en la manera como se hacen las cosas. Para el caso de los sistemas de atención en salud, dichas tecnologías junto con herramientas avanzadas de procesamiento permiten la transformación de sistemas de información hospitalarios en sistemas de información en salud, donde la información se genera de manera constante y es almacenada en registro clínicos electrónicos. Esta información digitalizada presenta nuevas oportunidades referente al uso de técnicas avanzadas de análisis como el aprendizaje de máquinas.

En los últimos cinco años, diferentes estudios demuestran el potencial que tiene la incorporación del aprendizaje de máquinas en tareas médicas, como el diagnóstico de enfermedades o el suministro de tratamientos personalizados. Entre las principales razones para aplicar técnicas de aprendizaje de máquinas en información clínica se encuentran el procesamiento de conjuntos de datos con alta dimensionalidad, la posibilidad de analizar información en diferentes formatos como imágenes y sonidos, y la capacidad de los algoritmos para detectar patrones y relaciones complejas entre los datos.

Por lo anterior, en este trabajo se pretende abordar el desarrollo de un modelo híbrido de clasificación para una población de adultos mayores pertenecientes a un municipio de Antioquia - Colombia, que permita la identificación del riesgo asociado al padecimiento de

enfermedades cardiovasculares, reconociendo que tanto estas como el envejecimiento de la población se incorporan como elementos de particular interés en investigaciones relacionadas con el cuidado de la salud.

Así mismo, se hace uso de la metodología CRISP-DM usada en proyectos de minería de datos. Esta metodología establece seis etapas que componen todo el ciclo de minería, en ellas se consideran actividades que permiten la comprensión del contexto del negocio o problema abordado, el entendimiento de los datos y su preparación para la etapa de modelado, la aplicación de técnicas de aprendizaje de máquinas, la evaluación del desempeño de dichas técnicas y su implementación en sistemas productivos.

Este trabajo explora tres modelos de aprendizaje de máquinas usados ampliamente en tareas de clasificación, entre ellos se encuentra la regresión logística, que es de uso común en investigaciones médicas, además, se consideran los modelos de bosque aleatorio y máquina de aumento de gradiente. Los tres modelos se comparan mediante métricas de desempeño que permiten la correcta interpretación de los modelos de clasificación evaluados en términos de exactitud, precisión y sensibilidad. Después de la evaluación, se hace la elección del modelo óptimo con el que se aborda la clasificación de grupos poblacionales de adultos mayores según los niveles de riesgo cardiovascular.

El presente documento se estructura de la siguiente manera: en el capito uno se realiza una exploración de la literatura existente, referente a la aplicación de técnicas de aprendizaje de máquinas en información médica para la detección de enfermedades, haciendo énfasis en el diagnóstico de enfermedades cardiovasculares o el riesgo asociado a su padecimiento. En el capítulo dos se desarrolla todo el entendimiento del problema abordado, la explicación de la

metodología empleada, los objetivos del trabajo, así como aspectos relevantes para la comprensión del problema de analítica. El capítulo tres presenta una visión general de los datos disponibles y la explicación de las técnicas usadas para el tratamiento de problemas de calidad existentes. En el capítulo cuatro se desarrolla la etapa de modelado, donde se aplican técnicas de selección de variables, técnicas de muestreo de clases minoritarias y optimización de hiperparámetros, finalizando con la evaluación de los diferentes modelos propuestos. Por último en los capítulos cinco y seis se exponen la discusión y las conclusiones del trabajo, respectivamente.

1. Aprendizaje de máquinas para información médica

La creciente generación de información digitalizada, consecuencia del uso de las tecnologías de información, la adopción de dispositivos móviles e inteligentes y la mejora de los recursos de hardware, ha impulsado la aplicación de análisis avanzados, aprendizaje de máquinas (ML *por sus siglas en inglés*) e inteligencia artificial en muchos contextos industriales y académicos. La diversidad de esta información ha permitido la exploración de muchos casos de uso y diferentes estudios demuestran que los avances obtenidos son significativos. Entre los casos de aplicación más populares, se encuentran aquellos relacionados con el cuidado de la salud y la medicina (Basu et al., 2020).

Actualmente, es común encontrar muchos sistemas de atención en salud, por ejemplo, clínicas, hospitales, establecimientos de atención primaria, entre otros; que han adoptado el uso generalizado de registros electrónicos para centralizar la información de la historia clínica de los pacientes, puesto que el formato digital facilita el acceso y uso de dicha información, lo cual conlleva a tener conjuntos de datos de gran dimensionalidad. Esta disponibilidad de grandes conjuntos de datos al combinarse con herramientas de ML puede expandir la base de evidencia médica y apoyar de manera importante la toma de decisiones. El ML es clave para mejorar la eficiencia y calidad de la asistencia en salud, por esto la inversión en el desarrollo de plataformas que integran datos y aplicaciones analíticas cada vez es más frecuente en organizaciones médicas (Ashfaq & Nowaczyk, 2019; López-Martínez et al., 2020).

Según Basu *et al.* (2020), en el ámbito médico, el ML se ha aplicado para el diagnóstico de enfermedades, la optimización de procesos y el suministro de tratamientos personalizados, pero no se limita sólo a estos casos. Con el fin de abordar de manera correcta el problema de

interés, se realiza una exploración de la literatura que compendia los aportes teórico-prácticos más relevantes respecto al desarrollo de modelos de ML usando datos médicos.

1.1 Exploración de la literatura

Se realiza la exploración de la literatura existente referente al uso de modelos de aprendizaje automático en el ámbito de la salud o en sistemas similares, con el objetivo de enmarcar el trabajo en la evidencia teórico-práctica de los ultimo cinco años. Los artículos considerados, resumidos en la *tabla 1*, abordan aspectos relevantes de la aplicación de métodos de ML en información médica, específicamente para tareas de clasificación y diagnóstico de enfermedades. La mayoría de los artículos referenciados están relacionados con la detección de enfermedades cardiovasculares (ECV) o con el riesgo asociado a su padecimiento.

Se identifican, además, varios artículos que recopilan estudios de casos aplicados hasta el año 2020, brindando información relevante para comprender el alcance de las técnicas de ML más usadas, pues contrastan los resultados obtenidos en diferentes contextos. Entre estos, destacan los artículos de Alizadehsani *et al.* (2020), Kalantari *et al*. (2018), Bellamy, Celi y Beam (2020) y Basu *et al.* (2020).

Tabla 1. Exploración de la literatura

Título del estudio (referencia)	Enfoque Medico	Datos	Método de ML	Objetivo
Khan *et al.* (2019) An e-Health care services framework for the detection and classification of breast cancer in breast cytology images as an IoMT application	Diagnóstico de cáncer de mama	Imágenes de citología mamaria	Máquinas de vectores de soporte (SVM) – Naïve Bayes (NB) – Bosque aleatorio (RF)	Aplicar un enfoque basado en ML e inteligencia computacional para detectar y clasificar células malignas de cáncer de mama, incorporando algoritmos genéticos para la selección optima de variables

Tabla 1. Exploración de la literatura (Continuación)

Título del estudio (referencia)	Enfoque Medico	Datos	Método de ML	Objetivo
Hameed, Shabut, Ghosh, & Hossain, (2020) Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques	Clasificación de enfermedades de la piel	3.672 imágenes de la piel clasificadas como sanas, con eczema, benignas y malignas	Algoritmo de clasificación Multi-Clase y Multi-Nivel (MCML), implementado mediante ML tradicional y enfoques avanzados de aprendizaje profundo	Proponer un marco de diagnóstico inteligente para la clasificación de lesiones cutáneas, usando algoritmo de niveles múltiples con clases múltiples para mejorar la precisión
Das, (2010) A comparison of multiple classification methods for diagnosis of Parkinson disease	Diagnóstico de la enfermedad de Parkinson (EP)	Conjunto de datos compuesto por diferentes medidas biomédicas de voz de 31 personas, algunas con EP	Red neuronal – DM Neural – Regresión logística – Árbol de decisión	Estudio comparativo de diferentes métodos de clasificación para la identificación de individuos sanos de aquellos que tienen EP
Cosma *et al.* (2016) Prediction of Pathological Stage in Patients with Prostate Cancer: A Neuro-Fuzzy Model	Diagnóstico de cáncer de próstata	Registros recopilados de 399 pacientes diagnosticados con un tipo de cáncer de próstata	Método Hibrido: Neuro-fuzzy (NF)	Desarrollar un modelo de ML usando redes neuronales artificiales y algoritmos de lógica difusa para la predicción del cáncer de próstata
Loreto, Lisboa, & Moreira, (2020) Early prediction of ICU readmissions using classification algorithms	Clasificación de pacientes con probabilidad de reingreso a la UCI	11.805 registros clínicos de pacientes adultos con 185 atributos referentes a características basales e información de ingreso a la UCI	Naïve Bayes (NB) – Árbol de decisión – Bosque aleatorio (RF) – Sequential Minimal Optimization (SMO) – JRip – AdaBoost (AB) – Logit Iterative Classifier (ICO)	Crear modelos de predicción de alta calidad para la predicción del reingreso en UCI, basados atributos recogidos en el momento del ingreso del paciente
Bhatti, Kehar, & Memon, (2020) Prognosis of Diabetes by Performing Data Mining of HbA1c	Diagnóstico de diabetes	El conjunto de datos de 8.524 pacientes con resultados del test HbA1c	Árbol de decisión – J-48 – Clasificador bayesiano – Naïve Bayes (NB) – Perceptrón multicapa (MLP) – SVM – Bosque aleatorio (RF) –	Prever el padecimiento de diabetes mediante la aplicación estrategias de minería de datos

Tabla 1. Exploración de la literatura (Continuación)

Título del estudio (referencia)	Enfoque Medico	Datos	Método de ML	Objetivo
Castellanos Vázquez, Moreno, Herrera, & Sautto Vallejo, (2019) Valoración de riesgo cardiovascular mediante modelos de clasificación	Diagnóstico de riesgo de enfermedad cardiovascular	Datos relacionados con estudios de sobrepeso, obesidad y diabetes	Árbol de decisión – Regresión logística – Bosque aleatorio (RF)	Determinar el mejor modelo de clasificación, a partir de criterios de bondad, entre diferentes métodos de ML
Weng, Reps, Kai, Garibaldi, & Qureshi, (2017) ¿Can machine-learning improve cardiovascular risk prediction using routine clinical data?	Diagnóstico de riesgo de enfermedad cardiovascular	Datos clínicos de rutina de 378.256 pacientes de consultorios familiares del Reino Unido	Bosque aleatorio (RF) – Regresión logística – Máquina de aumento de gradiente (GBM) – Redes Neuronales	Evaluar si el ML puede mejorar la precisión de la predicción del riesgo cardiovascular en una gran población con atención primaria en salud
Bandyopadhyay *et al.* (2015) Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data	Diagnóstico de riesgo de enfermedad cardiovascular	Registros electrónicos de historias clínicas	Redes Bayesianas	Proponer una extensión general de redes bayesianas usando probabilidad inversa de censura de pesos (IPCW)
Ward *et al.* (2020) Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population	Diagnóstico de riesgo de enfermedad cardiovascular	Registros electrónicos de historias clínicas de 797.505 pacientes	Regresión logística con penalización – Bosque aleatorio (RF) – Máquina de aumento de gradiente (GBM) – Máquina de aumento de gradiente extremo (XGBM)	Desarrollar modelos de ML para la predicción del riesgo de enfermedad cardiovascular aterosclerótica (ASCVD) para pacientes multiétnicos utilizando una base de datos de historia clínica electrónica
Jamthikar *et al.* (2020) Cardiovascular/stroke risk prevention: A new machine learning framework integrating carotid ultrasound image-based phenotypes and its harmonics with conventional risk factors.	Diagnóstico de enfermedad cardiovascular	Datos de 212 pacientes que incluyen 13 variables asociadas a factores de riesgo convencionales y 35 de fenotipos de imágenes de ecografía carotidea.	Bosque Aleatorio (RF).	Predecir el riesgo de ECV/accidente cerebrovascular utilizando un framework de ML sobre datos retrospectivos asociados a factores de riesgo convencionales y fenotipos de imágenes de ecografía carotidea

Tabla 1. Exploración de la literatura (Continuación)

Título del estudio (referencia)	Enfoque Medico	Datos	Método de ML	Objetivo
Padmanabhan, Yuan, Chada, & Nguyen, (2019) Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction	Diagnóstico de riesgo de enfermedad cardiovascular	Conjuntos de datos cardiovasculares	Naïve Bayes (NB) – Regresión lineal – Análisis de discriminante cuadrático – Aprendizaje basado en instancias – Máquinas de vectores de soporte (SVM)	Demostrar la facilidad de construcción de clasificadores de aprendizaje automático en estudios biomédicos
Alaa, Bolton, Di Angelantonio, Rudd, & van der Schaar, (2019) Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants	Diagnóstico de riesgo de enfermedad cardiovascular	Registros electrónicos de historias clínicas de 423,604 pacientes	Máquinas de vectores de soporte (SVM) – Bosque Aleatorio (RF) – Regresión Logística – AdaBoost (AB) – Máquina de aumento de gradiente (GBM)	Desarrollar modelos de predicción de riesgo cardiovascular basados en ML y comparar sus desempeños predictivos en población general y las subpoblaciones clínicamente relevantes
Suzuki *et al*. (2019) Comparison of risk models for mortality and cardiovascular events between machine learning and conventional logistic regression analysis	Diagnóstico de enfermedad cardiovascular	Registros electrónicos de 15.933 pacientes para los que están disponibles muestras de sangre, ecocardiograma y registros de seguimiento clínico	Máquinas de vectores de soporte (SVM) – Bosque Aleatorio (RF) – Redes Neuronales	Examinar si existen diferencias entre el modelado por ML y el análisis de regresión logística utilizando una cohorte de un hospital cardiovascular
Wang *et al.* (2019) Using Machine Learning to Integrate Socio-Behavioral Factors in Predicting Cardiovascular-Related Mortality Risk	Diagnóstico de riesgo de enfermedad cardiovascular	Datos de *The Cardiovascular Disease Proyecto de agrupación de riesgos de vida* (LRPP)	Naïve Bayes (NB) – Regresión Logística – Máquinas de vectores de soporte (SVM) – Bosque Aleatorio (RF)	Aplicar diferentes modelos de aprendizaje automático para predecir el riesgo de mortalidad relacionado con las enfermedades cardiovasculares mediante la integración de factores de comportamiento social con la fisiopatología de los pacientes

Tabla 1. Exploración de la literatura (Continuación)

Título del estudio (referencia)	Enfoque Medico	Datos	Método de ML	Objetivo
(Melillo et al., 2015) Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis	Diagnóstico de enfermedad cardiovascular	Base de datos con registros holter electrocardiográficos (ECG) nominales de 24 h de 139 pacientes hipertensos	Naïve Bayes (NB) – Árbol de decisión C4.5 – Bosque Aleatorio (RF) – AdaBoost (AB) – Máquinas de vectores de soporte (SVM) – Perceptrón multicapa (MLP)	Aplicar métodos de análisis de RHV (*Heart rate variability*) lineales y no lineales y esquemas de reconocimiento de patrones para discriminar entre pacientes hipertensos de alto y bajo riesgo cardiovascular

Como principales fuentes de datos, los artículos referenciados en la *tabla 1* hacen uso de registros electrónicos de historias clínicas para la clasificación de individuos como sanos y no sanos, aunque esto puede variar según el enfoque médico aplicado, puesto que para enfermedades relacionadas con lesiones cutáneas, el uso de imágenes es predominante, como lo es también para la identificación del cáncer de mama, cáncer de colon e incluso la identificación de anomalías cerebrales (Kalantari et al., 2018). Es importante destacar que, por su naturaleza, los registros electrónicos de historias clínicas pueden presentar más problemas de calidad en comparación con aquellos conjuntos de datos que son recopilados por entidades médicas o académicas en el desarrollo de investigaciones, añadiendo un grado de dificultad mayor, dado que este tipo de problemas pueden afectar los resultados y el desempeño de los métodos aplicados.

Artículos como el de Ward *et al.* (2020), Suzuki *et al.* (2019), Alaa *et al.* (2019) y Weng *et al.* (2017), optan por la aplicación de métodos de imputación para mitigar el efecto de los valores ausentes, usando medidas de tendencia central como la mediana o algoritmos más sofisticados como *Miss Forest Imputation Algorithm* (MFIA) basado en bosques aleatorios*;* por el contrario, otros autores como Wang *et al*. (2019) optan por excluir dichos valores, esto en la práctica no suele ser recomendable ya que se incurriría en perdida de información y se introducirían sesgos en los análisis realizados (Amón Uribe, 2010).

Además, es común encontrar casos en los que se poseen muchas más muestras de un tipo de clase en particular, teniendo así una clase mayoritaria y otra u otras clases minoritarias. Este problema es común cuando se busca identificar llamadas o transacciones fraudulentas, detectar anomalías o para el caso médico, detectar enfermedades. Por lo general, las clases desbalanceadas suelen afectar a los algoritmos de ML en cuanto a su capacidad de generalizar, perjudicando a la clase minoritaria. Para tratar este problema, es común aplicar técnicas de muestreo que pueden eliminar muestras para la clase mayoritaria (*Under-sampling*) o crear muestras sintéticas de la clase minoritaria (*Over-sampling*) hasta alcanzar un nivel de equilibrio aceptable entre ellas (Cosma et al., 2016; Ma et al., 2011; Monroy-de-Jesús et al., 2018). El tratamiento de problemas de calidad y formato en los datos es abordado antes del modelado de las técnicas de ML, con lo que se espera mejorar el desempeño de los modelos.

Respecto a los estudios enfocados en el riesgo cardiovascular, se han aplicado diversos algoritmos de aprendizaje de máquinas, como regresión logística, bosque aleatorio, máquina de vectores de soporte, máquina de aumento de gradiente, redes neuronales y otras más, para la detección de ECV y la clasificación de pacientes en niveles de riesgo. En general, el algoritmo más usado, entre los artículos considerados, es el de bosque aleatorio, seguido de la regresión logística que, según Castellanos Vázquez *et al*. (2019), "… *es más recomendable para fines prácticos de interpretación, toda vez que con estos modelos es posible hacer predicciones de riesgo cardiovascular en términos probabilísticos.*" (p. 86), así mismo, los árboles de decisión, que son posibles de representar en forma gráfica, pueden contribuir en buena medida con la interpretación del problema.

En cuanto a las variables incluidas en los artículos citados, la mayoría son características demográficas y clínicas como la edad, el sexo, el nivel de colesterol o la presencia de diabetes, no obstante, también se encontraron estudios como el de Wang et al. (2019) y Ward et al. (2020), que buscan aproximar modelos de ML para predecir el riesgo cardiovascular incluyendo factores socio-conductuales, tales como la raza o si el paciente fuma o consume alcohol con frecuencia, cuyos resultados demuestran que la inclusión de estas variables contribuye en la precisión de los modelos propuestos por los autores. En la *tabla 2* se presenta un resumen de las variables usadas en los artículos referenciados.

Tabla 2. Variables predictoras identificadas en la exploración de la literatura referente a riesgo cardiovascular

Variables	Modelos Citados
Variables Demográficas Sexo, Edad,	(Alaa et al., 2019) (Bhatti et al., 2020) (Castellanos Vázquez et al., 2019) (Jamthikar et al., 2020) (Padmanabhan et al., 2019) (Wang et al., 2019) (Ward et al., 2020) (Weng et al., 2017)
Variables Clínicas Test HbA1c hipertensión arterial (HTA), Diabetes, RCFraming Edad, presión diastólica, presión sistólica, LDL - Colesterol, OB1 Presión arterial, Diabetes, obesidad Enfermedad mental severa,	(Alaa et al., 2019) (Bhatti et al., 2020) (Castellanos Vázquez et al., 2019) (Jamthikar et al., 2020) (Padmanabhan et al., 2019) (Wang et al., 2019) (Ward et al., 2020) (Weng et al., 2017)
Variables socio-conductuales Raza, Tabaquismo, Antecedentes familiares, Raza, Actividad física, Medicamentos prescritos	(Alaa et al., 2019) (Bhatti et al., 2020) (Jamthikar et al., 2020) (Padmanabhan et al., 2019) (Ward et al., 2020) (Wang et al., 2019), (Weng et al., 2017)
Medidas antropométricas IMC, Peso, Talla	(Alaa et al., 2019) (Padmanabhan et al., 2019) (Wang et al., 2019) (Weng et al., 2017)

Esta exploración proporciona un marco de referencia para la aplicación y evaluación de nuevas metodologías, así como formas novedosas de abordar el problema. Con ella se determinan los algoritmos que se aplican en este trabajo considerando el desempeño obtenido en términos de exactitud con datos similares a los disponibles. Algunos aspectos referentes al tratamiento de

problemas de calidad abordados por los artículos referenciados, también se consideran, haciendo énfasis en la identificación de valores atípicos. Finalmente, la evidencia teórico-practica revisada no arrojó estudios similares a este para la población de la que se tienen datos, por lo tanto, se espera contribuir en el campo de la informática de la salud, en el contexto colombiano. En el siguiente capítulo se desarrolla la primera parte del proyecto de minería expuesto, donde se abarcan aspectos importantes para el desarrollo de los objetivos establecidos, se define la metodología aplicada y se describen los materiales y métodos utilizados.

2. Entendimiento del problema abordado

En el ámbito del cuidado de la salud, las tecnologías de información y comunicación (TICS) junto con técnicas avanzadas de procesamiento, posibilitan la transformación de los sistemas de información hospitalarios en sistemas de información en salud, donde la generación, almacenamiento y acceso a la información clínica desempeñan un papel clave para la atención de la salud, al contribuir significativamente en su calidad y eficiencia con la incorporación de sistemas menos propensos a errores y que soportan la toma de decisiones (Plazzotta et al., 2015). En este contexto, la manera en que son usados los datos clínicos puede ser un factor de alto impacto para la generación de nuevo conocimiento; así mismo, la aplicación de nuevas técnicas de análisis como el ML (Wiens & Shenoy, 2018).

Como se expuso en el capítulo anterior, el ML extiende una amplia gama de aplicaciones que pueden apoyar diversas tareas médicas, algunas más complejas que otras, pero que en últimas sirven de referencia para el desarrollo de nuevos enfoques de investigación, cuyos hallazgos y contribuciones fundamentan la transformación de los métodos usados para abordar problemas relacionados con la atención en salud. Uno de estos problemas se refiere a la detección de enfermedades, en donde el tiempo se inserta como factor clave para su tratamiento, esta es una de las razones por las que incluir técnicas de ML en problemas como este tiene gran potencial, dado que la entrega de diagnósticos se haría de manera efectiva desde fases tempranas de la enfermedad.

En este trabajo se aplica una metodología basada en proyectos de minería de datos para la identificación del riesgo asociado al padecimiento de enfermedades cardiovasculares para una población de adultos mayores de un municipio de Antioquia, Colombia; entendiendo que tanto las enfermedades cardiovasculares como la vejez de la población son relevantes para el ámbito de la salud pública. En este capítulo se aborda de manera concreta los aspectos considerados

para la comprensión del contexto en el que se aplican las diferentes técnicas de ML consideradas, los objetivos trazados y la metodología aplicada para el desarrollo de estos, así como los materiales y métodos utilizados.

2.1 Contextualización del problema

Las enfermedades cardiovasculares son una de las principales causas de muerte en el mundo, de acuerdo con la Organización Mundial de la Salud (OMS, 2014) en su Informe sobre la situación mundial de las enfermedades no transmisibles, 56 millones de muertes ocurrieron en todo el mundo durante el año 2012, de ellas, aproximadamente el 67% se debieron a enfermedades no transmisibles (ENT), principalmente ECV, cáncer y enfermedades respiratorias crónicas.

En Colombia la situación es similar, según el Instituto Nacional de Salud de Colombia (INS, 2013), entre los años 1998 y 2011, las ECV causaron 23,5% de defunciones, siendo la principal causa de muerte y con mayor presencia en personas mayores de 65 años, el INS también asegura que "*aunque la tasa de mortalidad por este evento se ha incrementado durante el periodo 1998-2011, esto se debe al envejecimiento poblacional, pues las tasas de mortalidad ajustadas por edad muestra una tendencia hacia el descenso*" (p. 1).

Con respecto al envejecimiento, según información publicada por la OMS (2015), la proporción de personas mayores está aumentando de forma notable en las poblaciones de todo el mundo, por esto, el envejecimiento se ha convertido en una cuestión política y de salud pública. Para el caso colombiano, se ha demostrado que el envejecimiento de la población también es un fenómeno actual, la esperanza de vida es mucho mayor en comparación con años atrás, lo que ha llevado a un proceso de transformación demográfica que supone nuevos retos para el país (Ministerio de Salud y Protección Social de Colombia, 2013).

La población de adultos mayores por su grado de dependencia puede estar más expuesta a eventos que atenten contra su integridad; por ejemplo "*las mujeres mayores de hoy en su mayoría no cuentan con un alto nivel educativo, carecen de pensión y se hacen muchas veces dependientes por no contar con un ingreso seguro que les dignifique su vejez*" (p.28), asegura el ministerio de salud y protección social (2013).

2.1.1 Envejecimiento de la población

Colombia ha transitado por un rápido proceso de cambio demográfico en las últimas décadas, hasta alcanzar en la actualidad la etapa de transición demográfica avanzada, afirma el Ministerio de Salud y Protección Social (2013) en su informe de "Envejecimiento demográfico: Colombia 1951-2020". Este cambio se transfiere en la modificación de la estructura por edad de la población y es visible si se revisan las cifras reportadas en dicho informe, que demuestran el aumento de la población mayor en Colombia y como su tasa de crecimiento es superior que la tasa de crecimiento poblacional total.

En este informe, se presenta que la población de 60 años o más, pasa de ser aproximadamente 3.8 millones de personas en el año 2005, a ser aproximadamente 4.4 millones en el año 2010 según proyecciones realizadas; esto equivale a un ritmo de crecimiento aproximado del 3,18% anual y estimado del 3.76% para el año 2020. Cabe mencionar que los departamentos con mayores índices de envejecimiento en Colombia corresponden a: Bogotá D.C, Caldas, Risaralda, Quindío, Valle, Antioquia y Santander (Ministerio de Salud y Protección Social de Colombia, 2013).

El fenómeno de la vejez no es vivenciado solo en el ámbito regional, Gómez (2011) afirma que la población mundial mayor de 60 años aumentará en los años próximos, llegando a proporciones de 16,6% en el año 2030 y a 21,4% en el año 2050, según las tendencias de envejecimiento actuales. La vejez y la presencia de ENT son factores que se relacionan de manera directa, de acuerdo con Llibre Guerra, Guerra Hernández y Perera Miniet (2008) por su importante aumento, las ENT constituyen en la mayoría de países latinoamericanos nuevas prioridades de salud, donde la vejez se inserta como uno de los factores de riesgo más importantes para el padecimiento de dichas enfermedades, de ahí la importancia de conocer, prevenir o retrasar estas enfermedades, no solo por su alto costo sino por su velocidad y la carga que conllevan.

Es innegable que estos cambios demográficos repercuten en ámbitos sociales, culturales y económicos y dan surgimiento a la importancia de abordar el tema del cuidado y la protección de la salud de los adultos mayores como una prioridad de las políticas públicas, así como la

necesidad de crear redes sociales que fortalezcan los factores protectores para una vejez digna, activa y saludable.

2.1.2 Riesgo cardiovascular

De las enfermedades cardiovasculares se puede afirmar que su presencia en el mundo ha suscitado particular interés en el estudio y análisis de los factores asociados a sus causas y sus efectos sobre la salud humana; según la OMS (2014), las ENT fueron la causa de aproximadamente el 67% de muertes ocurridas en el mundo en el año 2012. Dentro de las cuatro principales enfermedades no transmisibles, las ECV fueron las causantes del 46.2% de las muertes, equivalente a 17.5 millones de defunciones para ese mismo año. Igualmente, en Colombia las causas de mortalidad por ENT están encabezadas por las ECV, con aproximadamente un 24% de ocurrencia dentro del total de defunciones para el periodo 1998–2011. Entre las ECV con mayor ocurrencia se encuentran la enfermedad cardiaca isquémica, enfermedad cerebrovascular y enfermedad hipertensiva (INS, 2013).

Para el departamento de Antioquia, en los municipios que componen su área metropolitana, las enfermedades del sistema circulatorio causaron el 26,6% de las muertes entre 1998 y 2014, de las cuales la principal causa de muerte fue la Enfermedad Isquémica del corazón (EIC) con una tasa de mortalidad correspondiente al 51,36% dentro del total de defunciones por enfermedades del sistema circulatorio. El grupo poblacional con mayor casos de muertes fue el de 80 años y más (Bedoya-Mejía et al., 2019).

Existen una serie de factores, biológicos o de hábitos de vida, que hacen que la probabilidad de padecer ECV aumente para los individuos que los presentan. Estos factores de riesgo pueden ser no modificables, como la edad, el sexo o antecedentes familiares relacionados con ECV; o modificables, como la hipertensión arterial, el tabaquismo, la diabetes mellitus (DM) y el sobrepeso/obesidad (Lobos Bejarano & Brotons Cuixart, 2011). En estudios realizados en poblaciones colombianas, se sostiene que entre los factores de riesgo cardiovascular con más prevalencia, se encuentran la hipertensión, las dislipidemias y la obesidad, relacionada con hábitos sedentarios (Díaz-Realpe et al., 2007; Fernando & Arrieta, 2005; Lobos Bejarano & Brotons Cuixart, 2011; Patiño-Villada et al., 2011).

La mayoría de las ENT, en especial las cardiovasculares, pueden ser prevenidas y tratadas a tiempo si se reducen los factores de riesgo modificables asociados a su padecimiento mediante la implementación de políticas y programas de salud pública enfocados a reducir la prevalencia de estos en diferentes grupos poblacionales (Patiño-Villada et al., 2011). Enfrentar este tipo de enfermedades debe ser una prioridad dentro de la agenda nacional, en tanto que su costo puede ser tan elevado que pone en riesgo la capacidad económica de las personas que las padecen. Esto se expresa como obstáculo en el desarrollo socioeconómico de las comunidades, agregándole una alta carga al tratamiento de la enfermedad.

2.1.3 Estimación del riesgo cardiovascular

Como se ha mencionado anteriormente, las principales enfermedades que padece la población mayor están asociadas a enfermedades cardiovasculares. Según Álvarez Ceballos *et al.* (2017), el riesgo cardiovascular puede definirse como la probabilidad que tiene un individuo de sufrir un evento de ECV dentro de un periodo de tiempo determinado.

En los años cincuenta se pusieron en marcha varios estudios epidemiológicos para aclarar las causas de la ECV, entre estos se encuentra el *Framingham Heart Study* (FHS) que tenía como finalidad estudiar la epidemiología y los factores de riesgo de la ECV. Este estudio ha permitido el desarrollo de nuevos métodos estadísticos multivariables, donde se analiza el desarrollo de enfermedades complejas mediante la estimación del riesgo individual según el grado de exposición a diferentes factores de riesgo (O'Donnell & Elosua, 2008).

Existen diversos métodos de estimación para el riesgo cardiovascular, como los métodos cualitativos que se basan en la suma de factores de riesgo para clasificar al individuo en un riesgo leve, moderado o alto y los métodos cuantitativos que dan la probabilidad de presentar un evento cardiovascular en un determinado periodo de tiempo. La forma de estimación más común del riesgo cardiovascular es a través de tablas de riesgo, que son el resultado de los diversos estudios realizados en el ámbito de la epidemiología cardiovascular. Entre las tablas existentes, destacan (Álvarez Cosmea, 2001):

- Framingham clásica y por categorías
- Tabla de riesgo SCORE (Sociedades Europeas)
- Tabla de riesgo de Sociedades Británicas
- Tablas de riesgo de Nueva Zelanda

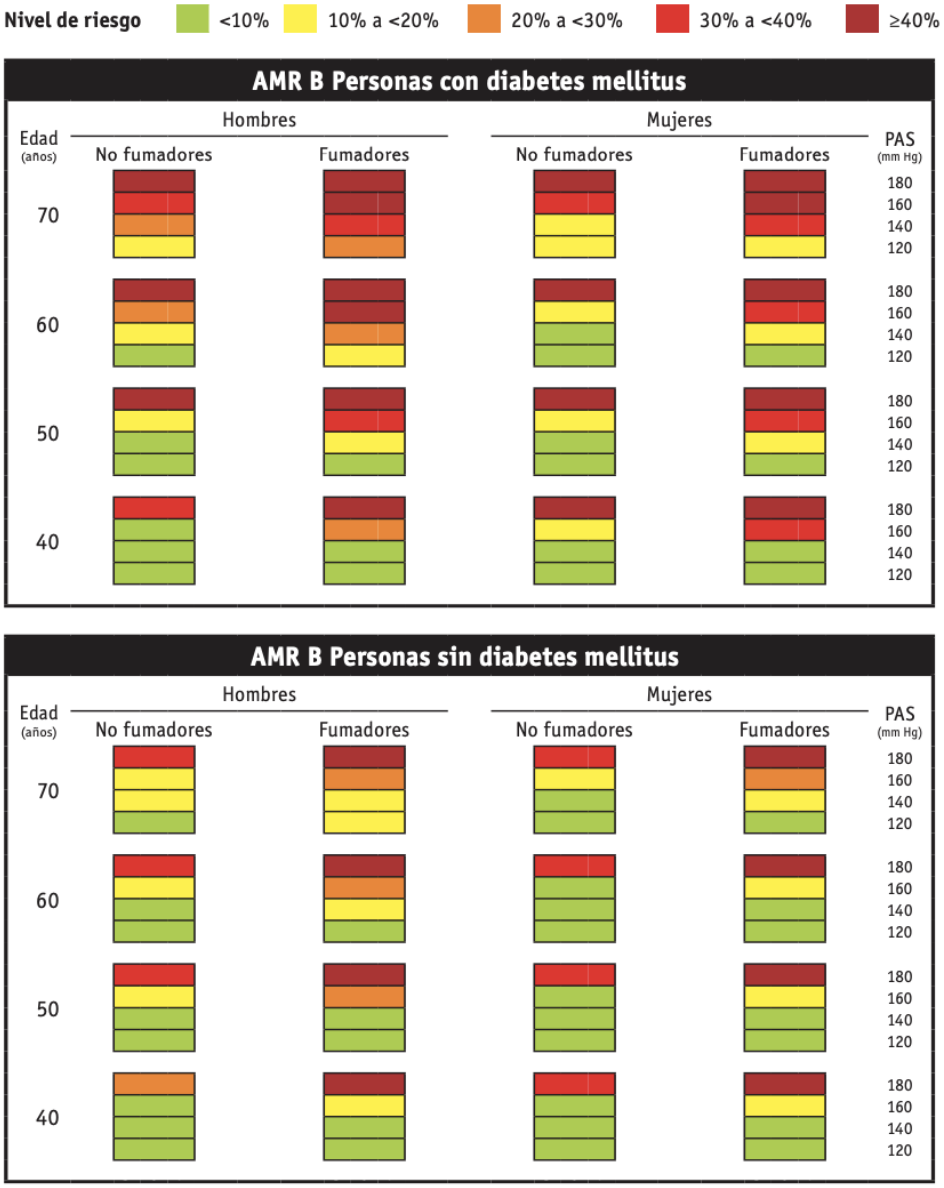
- Tabla de Sheffield

En Colombia, se han realizado varios estudios que buscan validar modelos de estimación de gran aceptación a nivel mundial, como el Framingham o el modelo PROCAM *(Prospective Cardiovascular Munster*) en poblaciones colombianas (Muñoz et al., 2014, 2015; Muñoz V et al., 2017), cuyos resultados sugieren el uso de modelos recalibrados en el país, sin embargo, la guía para la detección de riesgo cardiovascular del ministerio de salud aplica la escala de Framingham (Martínez et al., 2016) abalada por la OMS, de ella se reconocen diferentes variantes basadas en:

- *Subregiones epidemiológicas de la OMS*. Las principales regiones del mundo (África, las Américas, Mediterráneo Oriental, Europa, Asia sudoriental, Pacifico Occidental) se dividen en subregiones para llegar a un total de 14 categoría, es decir, una tabla de riesgo para cada subregión.
- *Modelos de tablas*. Uno de ellos (14 tablas) es válido para los contextos en los que se puede determinar el colesterol en sangre, mientras que el otro (14 tablas) se ha concebido para los contextos en que no es posible (OMS, 2008).

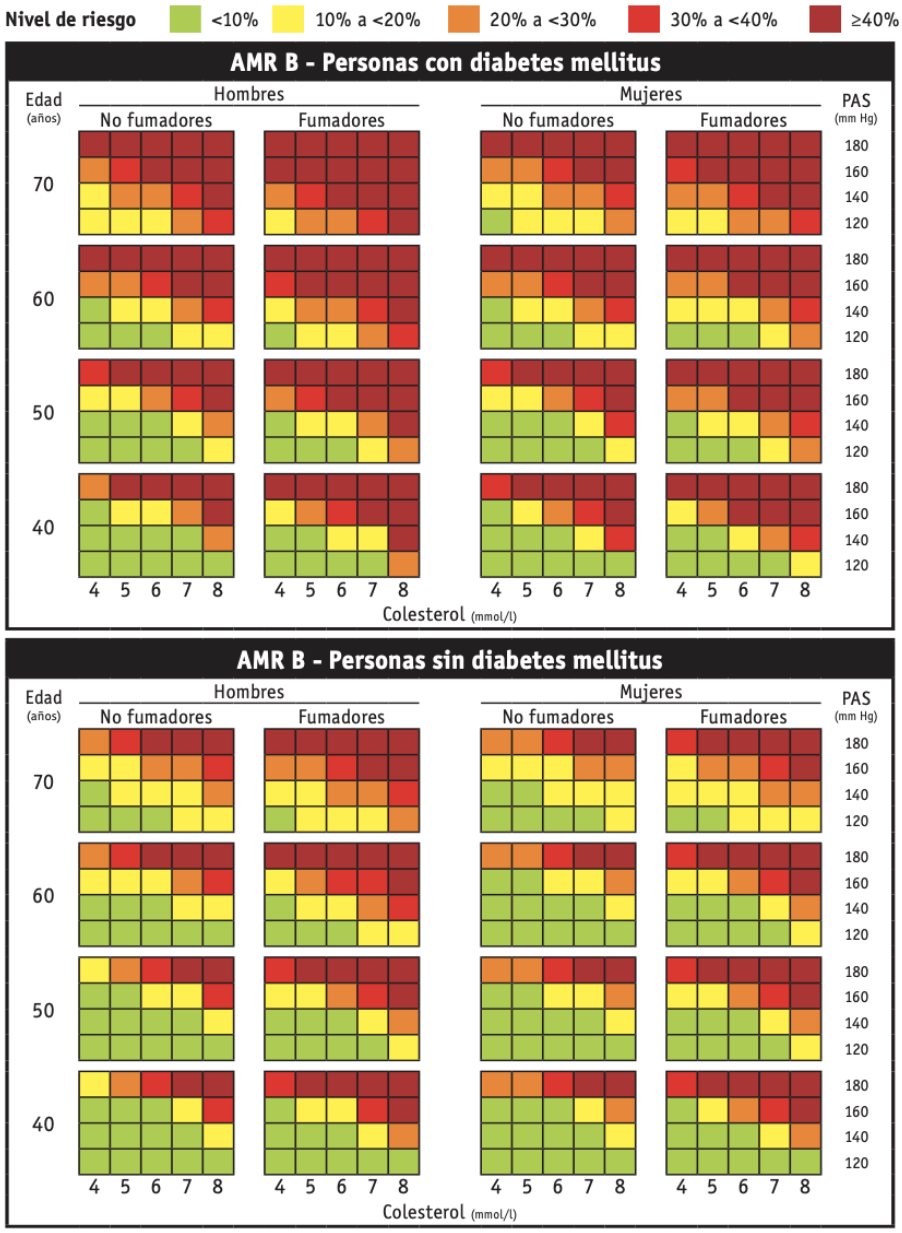
La escala de riesgo cardiovascular de la OMS se define por categorías de riesgo que van desde riesgo menor del 10% hasta riesgo mayor del 30% en un periodo de tiempo de 10 años, según las variables de: sexo, edad, presión arterial sistólica, consumo de tabaco y presencia o ausencia de diabetes mellitus (OMS, 2008). Colombia pertenece al grupo B de la región de las Américas, por tal motivo, las tablas asociadas a esta región son las que deben aplicarse en el contexto colombiano, estas pueden verse en las *figuras 1 y 2*.

Figura 1. Tabla de predicción de riesgo cardiovascular para las Américas (AMR) grupo B de la OMS/ISH, para los contextos en que no se puede medir el colesterol sanguíneo.



Tomado de (OMS, 2018).

Figura 2 Tabla de predicción de riesgo cardiovascular para las Américas (AMR) grupo B de la OMS/ISH, para los contextos en que se puede medir el colesterol sanguíneo



Tomado de (OMS, 2018).

Una vez planteados los principales aspectos que componen el entendimiento del problema abordado, es importante mencionar los recursos informáticos y de datos a los que se tiene acceso, los métodos de ML aplicados y los requisitos y restricciones bajo los cuales se desarrolla este trabajo.

2.2 Recursos disponibles

En este estudio se cuenta con los datos de evaluaciones físicas y antropométricas para una población de adultos mayores pertenecientes a un municipio de Antioquia. Se trabaja con una fuente de datos secundaria. Los datos disponibles contienen características que logran describir, con su análisis, aspectos relevantes de la salud de la población mayor de ese municipio. Con la información a la que se ha tenido acceso, no es posible la identificación particular de los individuos a los que se les hicieron dichas evaluaciones, con lo que se garantiza la protección de la confidencialidad de la información.

Como recursos informáticos se consideran lenguajes de programación que permiten la aplicación de algoritmos de ML con librerías especializadas y de código abierto, como:

- Python: lenguaje de programación interpretado que tiene estructuras de datos eficientes de alto nivel y un enfoque simple pero efectivo para la programación orientada a objetos, también soporta programación imperativa y, en menor medida, programación funcional, por esto es un lenguaje de programación multiparadigma (van Rossum, 1995).
- R: entorno de software libre para gráficos y computación estadística. Se compila y se ejecuta en una amplia variedad de plataformas (R Core Team, 2013).

También se utilizan herramientas analíticas especializadas, como:

- Knime®: entorno modular que permite la creación visual y ejecución interactiva de canalización de datos para ciencia de datos avanzada (Berthold et al., 2009).
- RapidMiner: entorno de código abierto gratuito para KDD y aprendizaje automático que proporciona una amplia variedad de métodos que permiten la creación rápida de prototipos (Mierswa et al., 2006).

2.3 Inventario de modelos

Se consideran los modelos de regresión logística, máquina de aumento de gradiente y bosque aleatorio para la identificación del nivel de riesgo cardiovascular. El modelo de bosque aleatorio también se considera como modelo estimador en la selección de variables. Para la selección de variables se opta por el uso de un método empaquetado o tipo *wrapper*, en el cual el algoritmo de selección aplica primero un modelo de estimación y mide el desempeño que

tienen las predicciones que este genera, luego con base en estas predicciones selecciona un subconjunto de variables sobre el cual se aplica nuevamente el modelo de estimación. El procedimiento anterior se repite iterativamente hasta obtener un subconjunto óptimo de variables (Cardona Álzate, 2019). En este trabajo se usa el método de eliminación recursiva de características (RFE) que en esencia es una forma de *wrapper.* El método RFE emplea las medidas de importancia generadas por el modelo de bosque aleatorio como criterio base para realizar la selección de las variables predictoras.

2.3.1 Regresión logística:

"*Es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia se ha puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables* dummy*, es decir variables simuladas*" (Chitarroni, 2002).

Generalmente lo que se busca con la aplicación de este modelo es poder predecir la probabilidad de ocurrencia del evento analizado a un cierto sujeto, así como la determinación del peso que cada una de las variables dependientes en esta probabilidad; ambos elementos se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos (Chitarroni, 2002).

2.3.2 Máquinas de Aumento de Gradiente (Gradient Boosting Machines)

El enfoque principal de los algoritmos de aumento es combinar iterativamente varios modelos simples, llamados "aprendices débiles", para obtener un "aprendiz fuerte" con una precisión de predicción mejorada. Hace algunos años en estos algoritmos se introdujeron los conceptos de funciones de pérdida dando origen al método de máquinas de aumento de gradiente (*Gradient Boosting Machine*, GBM). Este puede verse como un algoritmo de optimización numérica que tiene como objetivo encontrar un modelo aditivo que minimice la función de pérdida. De manera más precisa, en la regresión, el algoritmo comienza inicializando el modelo mediante una primera conjetura, que suele ser un árbol de decisión que reduce al máximo la función de pérdida. Así, el algoritmo GBM agrega iterativamente en cada paso un nuevo árbol

de decisión (es decir, "aprendiz débil") que reduce mejor la función de pérdida (Touzani et al., 2018).

2.3.3 Bosque Aleatorio

Un bosque aleatorio es un clasificador que consta de una colección de clasificadores estructurados en árbol *{h (x, k)}*, donde *{k}* son vectores aleatorios independientes distribuidos de manera idéntica y cada árbol emite un voto unitario por la clase más popular en la entrada *x* (Breiman, 2001).

El algoritmo de bosque aleatorio puede manejar datos de alta dimensionalidad y utilizar una gran cantidad de árboles en el conjunto. Esto combinado con el hecho de que la selección aleatoria de variables para una división busca minimizar la correlación entre los árboles en el conjunto y es computacionalmente mucho más liviano, hace que el clasificador final sea más cercano al modelo real (Gislason et al., 2005).

2.4 Requerimientos y restricciones

Se identifican los siguientes requerimientos y restricciones asociados al uso de la información médica disponible para el desarrollo del trabajo expuesto:

2.4.1 Confidencialidad de la información

El conjunto de información disponible está compuesto por datos anonimizados. Al ser información referente a la salud, está protegida bajo la Ley Colombiana de Habeas Data 1581 de 2012 en la que se establece que esta información es de carácter sensible e involucra el derecho a la privacidad e intimidad de las personas (Ley 1581 de 2012 - Ley Estatutaria de Hábeas Data, 2012)

Dicha ley también establece en su *artículo 6* que el tratamiento de este tipo de datos puede hacerse cuando se cumpla alguna de las excepciones definidas en él; una de estas excepciones se refiere al tratamiento que tenga una finalidad histórica, estadística o científica. Para ese caso, deberán adoptarse las medidas conducentes a la supresión de identidad de los titulares. De igual manera, establece que la autorización previa e informada del titular para el caso expuesto no es requerida.

2.4.2 Ética de la investigación en salud

En Colombia la Resolución 8430 de 1993 es la norma marco para la investigación en salud, esta ofrece las condiciones mínimas y los aspectos formales a considerar en el desarrollo de dichas investigaciones (Lopera, 2017). Tomando en cuenta las categorías de investigación en salud definidas en el *artículo 11* de la resolución, es posible afirmar que las investigaciones sin riesgo son:

> *Estudios que emplean técnicas y métodos de investigación documental retrospectivos y aquellos en los que no se realiza ninguna intervención o modificación intencionada de las variables biológicas, fisiológicas, sicológicas o sociales de los individuos que participan en el estudio, entre los que se consideran: revisión de historias clínicas, entrevistas, cuestionarios y otros en los que no se le identifique ni se traten aspectos sensitivos de su conducta* (Ministerio de Salud de Colombia, 1993, p.3).

Bajo este concepto, la misma resolución establece en su *artículo 16*, parágrafo primero que, tratándose de investigaciones sin riesgo, se podrá dispensar al investigador de la obtención del consentimiento informado. El presente estudio, al trabajar con una fuente de datos secundaria (estudio documental) y cuyo alcance se limita a la aplicación de técnicas de análisis sobre los datos y no a un estudio poblacional o epidemiológico, pueden aplicarse los conceptos legislativos descritos anteriormente.

2.4.3 Transformación de los datos

La base de datos disponible está conformada en su mayoría por evaluaciones físicas y antropométricas, para la aplicación de transformaciones a los datos se consideraron las escalas de medición de cada variable, así como sus rangos promedio dentro de investigaciones relacionadas. De esta manera, se establecen valores de referencia con los que se evalúa la posibilidad de que ciertos datos sean producto de mediciones erróneas.

2.5 Objetivos

2.5.1 Objetivo General

- Definir un modelo de clasificación que permita la identificación del nivel de riesgo asociado al padecimiento de enfermedades cardiovasculares para una población de adultos mayores en un municipio de Antioquia.

2.5.2 Objetivos Específicos

- Caracterizar variables de evaluación físicas y antropométricas para una población de estudio de adultos mayores.
- Seleccionar variables de evaluación físicas y antropométricas para una población de estudio de adultos mayores, según su relevancia para el modelo de clasificación.
- Construir un modelo híbrido para la clasificación del riesgo asociado al padecimiento de enfermedades cardiovasculares que pueda ser aplicado en una población de adultos mayores.
- Validar el modelo de clasificación de grupos poblacionales de adultos mayores según el nivel de riesgo asociado al padecimiento de enfermedades cardiovasculares mediante técnicas de validación cruzada.

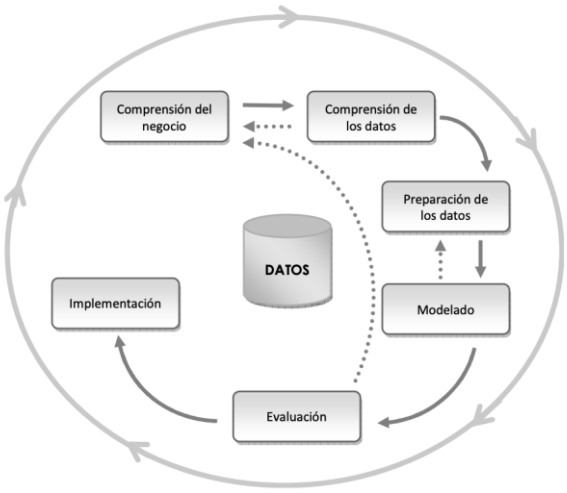
2.6 Metodología CRISP-DM

En la actualidad, una de las metodologías utilizadas en contextos académicos e industriales para el desarrollo de proyectos de minería de datos es el modelo CRISP-DM (*Cross Industry Standard Process for Data Mining*), el cual propone un modelo de proceso jerárquico que da un enfoque estándar a los proyectos, proporcionándoles medios para evaluar la efectividad de los resultados y documentar la experiencia (Wirth & Hipp, 2000). Esta metodología es aplicada para el desarrollo de este trabajo, sin embargo, se limita a la aplicación de determinadas técnicas de ML.

La metodología CRISP-DM está estructurada en un proceso jerárquico compuesto por tareas descritas en cuatro niveles diferentes de abstracción, que van desde lo general a lo específico. CRISP-DM propone en el nivel más alto de abstracción seis fases para el proceso de minería de datos: entendimiento del negocio, entendimiento de los datos, preparación de los datos,

modelado, evaluación e implementación (Moine, 2013). Cada una de estas fases componen el ciclo representado en la *figura 3:*

Figura 3. Fases de la metodología CRIP-DM



Tomado de (Moine, 2013).

Según Moine (2013), las fases del modelo CRISP-DM están compuesta por diferentes actividades que varían en función de los objetivos de cada una, estas actividades se describen a continuación:

2.6.1 Comprensión del negocio

En esta etapa se determinan los objetivos y requerimientos del proyecto desde una perspectiva de negocio, definiendo el problema de minería y el plan de trabajo a ejecutar. Son actividades comunes de esta etapa, la evaluación de la situación actual del problema, la definición de los objetivos de minería y la determinación del plan del proyecto (Moine, 2013). Los entregables de esta etapa corresponden a la definición del problema y su desarrollo teórico-practico dentro del aprendizaje de máquinas, abordados en los capítulos uno y dos.

2.6.2 Comprensión de los datos

Este es el primer acercamiento que se tiene con los datos disponibles para el proyecto y en el cual se identifican características sobre ellos, por esto es común el uso de la estadística descriptiva. También, se determina si los datos son relevantes y suficientes para construir un modelo (Moine, 2013). En esta etapa del trabajo se aplican técnicas de análisis descriptivo y estadístico sobre los datos, cuyos resultados se exponen en el tercer capítulo.

2.6.3 Preparación de los datos

En esta etapa se desarrollan aquellas actividades relacionadas con el tratamiento y corrección de errores presentes en los datos, así como la aplicación de transformaciones para construir el conjunto de datos final con el que se pretende modelar (Moine, 2013). La preparación de los datos abarca tareas, tales como selección de los datos, su limpieza y estandarización, así como la generación de variables derivadas si se consideran necesarias. En el tercer capítulo de este trabajo se describen las estrategias y transformaciones que se aplican para mitigar el efecto de los problemas de calidad de datos identificados en la etapa anterior.

2.6.4 Modelado

En esta etapa se aplican las diversas técnicas y algoritmos de ML sobre el conjunto de datos final, estos aprenderán automáticamente con los datos históricos preparados en las etapas anteriores para dar un resultado óptimo (Moine, 2013). Para este trabajo, se evalúan tres tipos de algoritmos de clasificación ampliamente usados en el análisis de datos clínicos, estos algoritmos corresponden a regresión logística, máquina de aumento de gradiente y bosque aleatorio. El entregable de esta etapa comprende el entrenamiento de los algoritmos antes mencionados y la optimización de hiperparámetros. Dicho entregable se desarrolla en el capítulo cuatro de este trabajo.

2.6.5 Evaluación

Etapa en la que se analizan los resultados obtenidos en función de los objetivos definidos. En esta etapa se debería determinar si se ha omitido algún objetivo importante y si el modelo será implementado, es decir, si se pasará a la próxima etapa (Moine, 2013). Para este trabajo, se evalúan métricas de exactitud, precisión, sensibilidad y f1-score, como principales insumos

para la interpretación del desempeño de los modelos entrenados en la etapa anterior. El entregable de esta etapa se desarrolla en el capítulo cuatro.

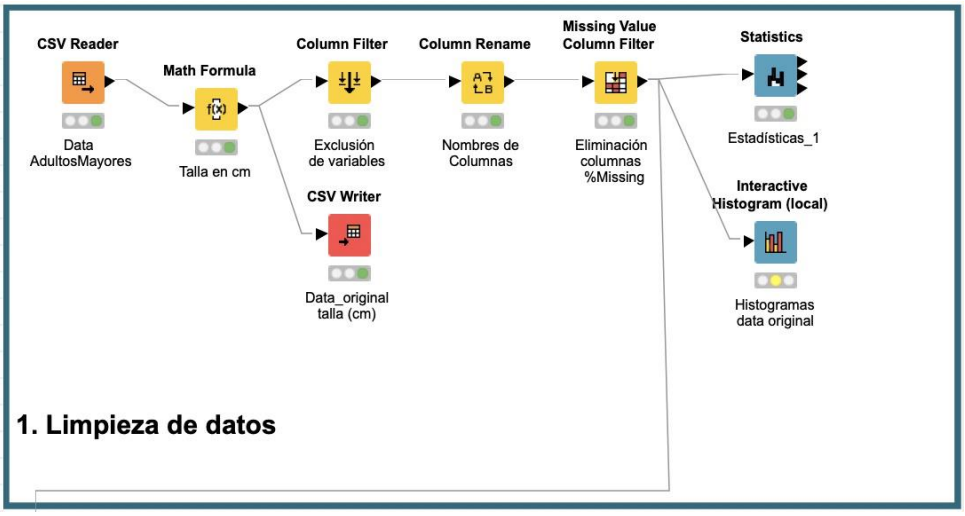
2.6.6 Implementación

Consiste en la comunicación e implementación del nuevo conocimiento, el cual debe ser representado de forma entendible para el usuario (Moine, 2013). En esta etapa del trabajo se exponen las conclusiones y la discusión como parte de la divulgación de los resultados obtenidos.

3. Entendimiento de los datos

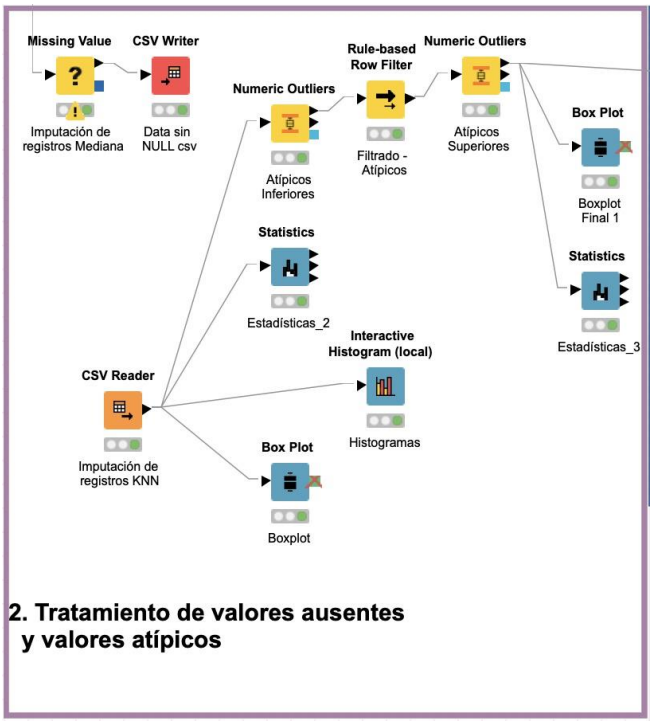
Para el desarrollo de esta fase se usan dos aplicaciones de código abierto especializadas en minería de datos. Ambas funcionan bajo el concepto de flujos, por lo que cada nodo representa un estado de los datos en el proceso de preparación. El flujo principal de todo el proceso se crea en Knime®, pero se recurre al uso de ciertas funcionalidades de RapidMiner para nutrir y complementar los análisis abordados. Dicho flujo se divide en 3 etapas, la primera de ellas corresponde a la limpieza de datos. Como se observa en la *figura 4*, se aplica un filtro a las variables, excluyendo aquellas con porcentajes mayores al 80% de valores ausentes.

Figura 4. Flujo de Knime® de la etapa 1: Limpieza de datos



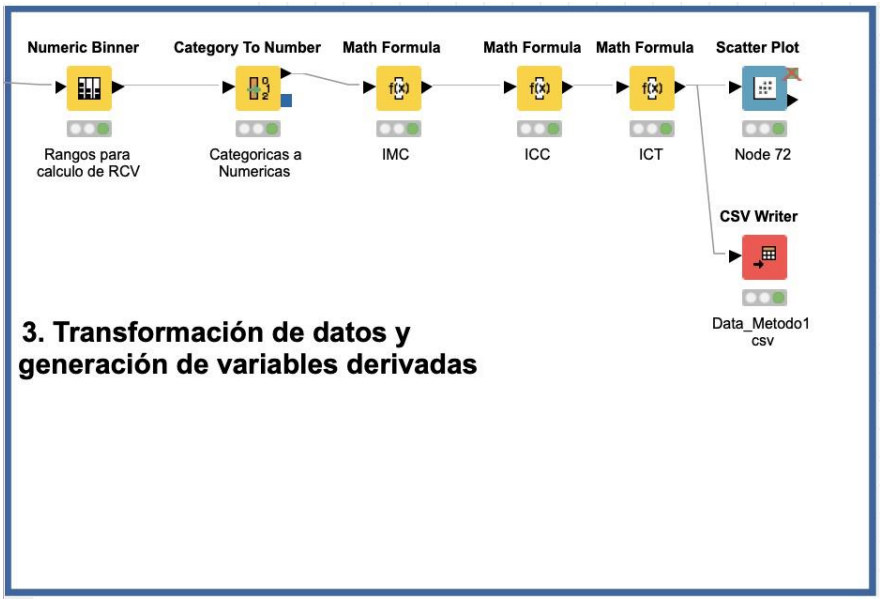
La segunda etapa corresponde al tratamiento de valores ausentes y valores atípicos, cada una de las actividades que la conforman pueden identificarse en la *figura 5*.

Figura 5. Flujo de Knime® de la etapa 2: Tratamiento de valores ausentes y valores atípicos



En la última etapa, se calculan tres variables derivadas y se transforman a binarias las variables categóricas (*ver figura 6*). A continuación, se expone cada uno de los aspectos abordados para el entendimiento y la preparación de los datos.

Figura 6. Flujo de Knime® de la etapa 3: Transformación de datos y variables derivadas



3.1 Información disponible

La base de datos a la que se tiene acceso está conformada por 1.524 registros y 70 atributos correspondientes a evaluaciones físicas e índices antropométricos de adultos mayores pertenecientes a un municipio de Antioquia. Como primer filtro, solo se incluyeron en los análisis posteriores, variables de las cuales se tiene evidencia que han sido usadas en otros estudios de riesgo cardiovascular o pueden tener influencia en este (Alaa et al., 2019; Castellanos Vázquez et al., 2019; Luengo Pérez et al., 2009; Martínez et al., 2016; Suzuki et al., 2019; Wang et al., 2019; Weng et al., 2017).

Se consideran las variables de: sexo, presencia de hipertensión, enfermedad coronaria, insuficiencia cardiaca, diabetes, EPOC (Enfermedad pulmonar obstructiva crónica), obesidad, enfermedad osteomuscular y osteoporosis; peso, talla, frecuencia cardiaca, presión arterial sistólica y diastólica; perímetro de la cintura y de la cadera; antecedentes familiares de diabetes mellitus, enfermedades cardiovasculares y otros antecedentes; variables que permiten identificar aspectos del estilo de vida del adulto mayor como el tabaquismo, los días de actividad física, las porciones de frutas y verduras por día, el número de cigarrillos consumidos por día, la prescripción de medicamentos, entre otras. En total se seleccionan como atributos relevantes 29 de los 70 originales.

3.2 Exploración de los datos

3.2.1 Proceso de limpieza

El primer análisis que se realiza sobre los datos consiste en la determinación de sus métricas de calidad. Las métricas de calidad pueden apoyar la decisión de usar ciertas transformaciones sobre los datos ante la presencia de un determinado problema. Las transformaciones que se aplican obedecen a la necesidad de contar con datos limpios de ruido y completos, puesto que los problemas de calidad pueden influir de manera incorrecta sobre los algoritmos y métodos aplicados a los datos, sesgando los resultados y dificultando el proceso de modelado. Los indicadores que se analizan para determinar el nivel de calidad de los datos se describen en la *tabla 3*.

Tabla 3. Indicadores de calidad para la base de datos de estudio

Métrica de calidad	Descripción
ID-ness	Mide el grado en que el atributo se asemeja a un identificador: $Número\ de\ valores\ diferentes \Big/ Número\ total\ de\ filas\ de\ datos$
Estabilidad	Mide cuán estable o constante es la columna evaluada. Identifica si es una constante: $\begin{matrix} número\ de\ filas\ con\ valores \\ no\ faltantes\ más\ frecuente \end{matrix} \Big/ \begin{matrix} número\ total\ de\ filas \\ con\ valores\ no\ faltante \end{matrix}$
Faltantes	Número de valores faltantes en la columna evaluada como una fracción del número total de filas de datos.
Text-ness	Mide si la columna evaluada parece contener texto libre.

Indicadores calculados con el software RapidMiner sobre la base de datos de estudio.

Entre los principales resultados que se obtienen al evaluar los indicadores de calidad, se determina que ningún atributo se asemeja a un identificador, las variables relacionadas con identificadores de procesos y actividades fueron retiradas previamente. Para el caso del indicador de estabilidad, se identifica que variables como el sexo, tabaquismo o la presencia de diabetes presentan índices superiores a 0.90, lo que indica la existencia de característica más frecuentes que otras para la población de estudio.

En la evaluación del indicador de valores ausentes, algunos atributos obtienen porcentajes superiores al 50% del total de registros, estos atributos no se incluyen en los análisis posteriores. En total se retiran cinco atributos que no tienen un impacto directo en la estimación del riesgo cardiovascular y que describen factores de contexto clínico, como la edad de inicio de consumo de cigarrillo, otros antecedentes familiares o la edad de diagnóstico de enfermedades. Los atributos con porcentajes de valores ausentes inferiores al 20% se conservan.

Los valores faltantes son más comunes de lo que en realidad se piensa, muchas bases de datos industriales y de investigaciones sufren de este problema. Según Amón (2010) "… *entre las razones para que esto suceda, se encuentran procedimientos imperfectos de captura de datos en forma manual, mediciones incorrectas, errores en los equipos y migraciones entre diferentes aplicaciones*" (p. 43). Existen diferentes estrategias para el tratamiento de valores ausentes, la de mayor facilidad de aplicación es la denominada *Listwise Deletion* cuyo principio es el de trabajar con datos completo, es decir, eliminar los registros que contengan algún atributo faltante. Para Amón (2010), esta práctica en realidad no es la más apropiada, pues si las observaciones completas no son submuestras al azar de los datos originales, se podrían introducir sesgos en los análisis posteriores realizados, por esto una de las primeras labores que se deben abordar en el tratamiento de valores ausentes, es la identificación del patrón que siguen dichos valores.

Se pueden identificar principalmente tres tipos de patrones de datos ausentes:

- *MCAR (Missing Completely At Random)*: Datos ausentes completamente al azar. Este patrón se puede identificar cuando los registros con los datos completos son similares a los registros que tienen valores ausentes, es decir, ambos constituyen una muestra aleatoria simple de todos los sujetos que conforman la muestra (Amón Uribe, 2010).

- *MAR (Missing At Random):* Datos ausentes al azar. El patrón MAR se presenta cuando los registros con información completa son diferentes de aquellos con información incompleta. Los patrones de los datos faltantes se pueden predecir a partir de la información contenida en otras variables y no de la variable que está incompleta (Amón Uribe, 2010).

- MNAR *(Missing Not at Random)*: Datos ausentes no al azar. El patrón de los datos ausentes no es aleatorio y no se puede obtener a partir de la información contenida en otras variables. Bajo este patrón, el proceso de ausencia de los datos sólo se explica por los datos que están ausente (Amón Uribe, 2010).

Las variables tratadas por valores ausentes corresponden a: presión arterial sistólica, presión arterial diastólica, frecuencia cardiaca en reposo, peso, talla, perímetro de la cintura, perímetro de la cadera máxima, días de actividad física moderada, porciones de verduras por día, y

porciones de frutas por día. No se tiene evidencia de que los datos ausentes presentes en estas variables puedan predecirse usando los datos completos, por tal motivo siguen un patrón de ausentes completamente aleatorios.

Una vez identificado el tipo de patrón de datos ausentes, se procede con la aplicación de imputación como estrategia para el tratamiento de dichos datos. Siguiendo la metodología propuesta por Amón (2010), se usa la imputación con mediana en atributos con índices inferiores al 10% de ausentes; el autor expone que esta es una buena estrategia para los datos ausentes con un patrón MCAR. Para valores superiores del 10% se realiza la imputación con el algoritmo de KNN (*K-Nearest-Neighbors*).

Otro de los problemas comunes que se encuentran en todo tipo de bases de datos y que de igual manera impacta e introduce sesgos en los análisis, son los atributos con registros atípicos. Un registro atípico o anómalo según Prasad, Almanza-Garcia, y Lu (2009), se refiere a un patrón en los datos que no se ajusta a una noción bien definida de comportamiento normal. Por otra parte, Abellana Sangra y Farran Codina (2015), definen que un dato atípico es una observación claramente diferente del resto de datos, es una observación extrema. Para el caso del presente trabajo, la identificación de valores atípicos presentes en las variables numéricas de la base de datos de estudio se realiza mediante la revisión de los diagramas de dispersión y diagramas de caja y bigotes. Se analizan en detalle las siguientes variables:

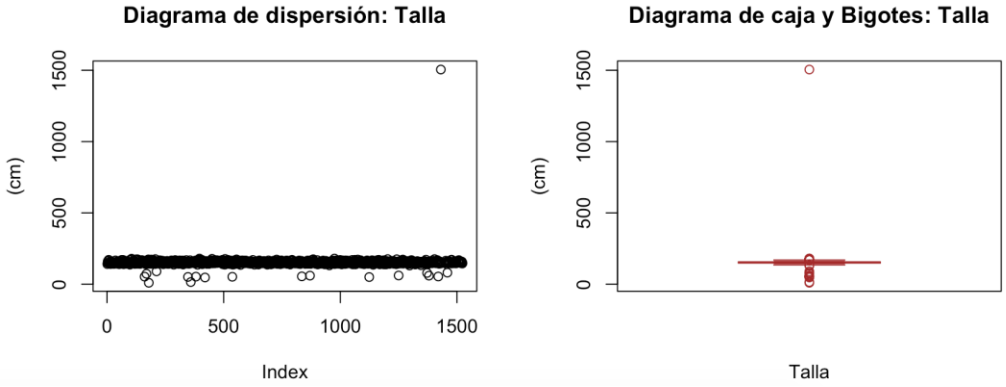
- Talla: Como se observa en la *figura 7,* existe un valor extremo en el atributo talla correspondiente a 1.550 cm, al hacer una revisión de la base de datos se determina que el valor mínimo registrado es de 10 cm. Estos datos no son válidos, pues no existen personas que midan 15 metros o 10 centímetros. También se pudo identificar que algunos valores son inferiores a 50 cm, lo que es consistente con lo observado en la gráfica de dispersión.

  Para el análisis de los rangos en que pueden oscilar las variables antropométricas consideradas, como la talla o el perímetro de la cintura, se toma como referencia los parámetros antropométricos presentados por Estrada, Camacho, Restrepo y Parra (1998), Ramírez-Vélez, Agredo, Jerez y Chapal (2008) y Avila-Chaurand, Prado-León y

González-Muñoz (2001) en sus investigaciones realizadas a la población de adultos en Colombia, una de ellas hace énfasis en la población de adultos mayores.

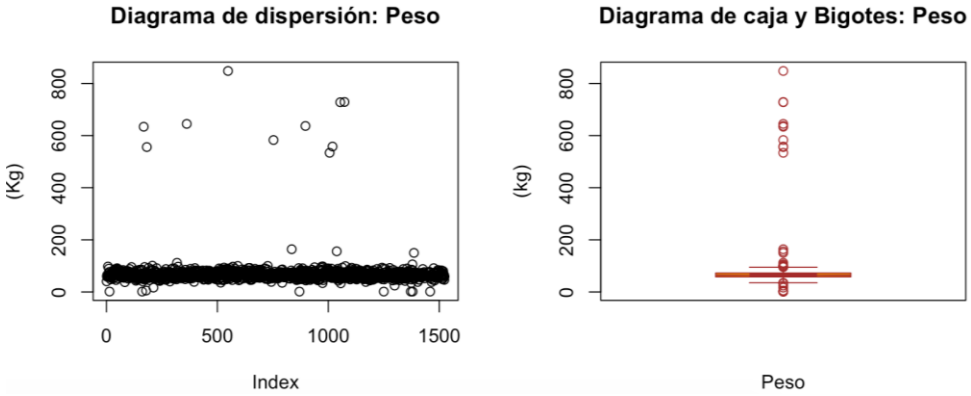
En promedio la talla para adultos en Colombia es de 165.6 cm y puede oscilar entre los 130 cm y los 185 cm. Con este margen de referencia, es posible que el valor extremo superior sea producto de un ingreso erróneo en la base datos al igual que el valor mínimo de 10 cm.

Figura 7. Diagramas de dispersión y Diagramas de caja y bigotes para la Talla



- Peso: En la *figura 8* se observa que el peso tiene varios registros superiores a los 600 Kg, al igual que registros muy cercanos a cero. Nuevamente al revisar la base de datos, se determina que el valor mínimo para esta variable es de 1.48 kg y su valor máximo de 848 kg.

Figura 8. Diagramas de dispersión y Diagramas de caja y bigotes para el Peso



Considerando los estudios antropométricos de referencia, es posible afirmar que el peso de la población adulta de Colombia oscila entre los 40 Kg y los 100 Kg. Asumiendo este rango de referencia, se determina que los valores extremos identificados, son valores atípicos para el peso.

- Perímetro de la cintura y perímetro de la cadera: Se identifican, en ambos casos, valores atípicos superiores que sobrepasan de manera extrema los márgenes de referencia definidos a partir de los estudios ya mencionados, dichos valores son visibles en las *figuras 9 y 10*.

Figura 9. Diagramas de dispersión y Diagramas de caja y bigotes para el Perímetro de la cintura

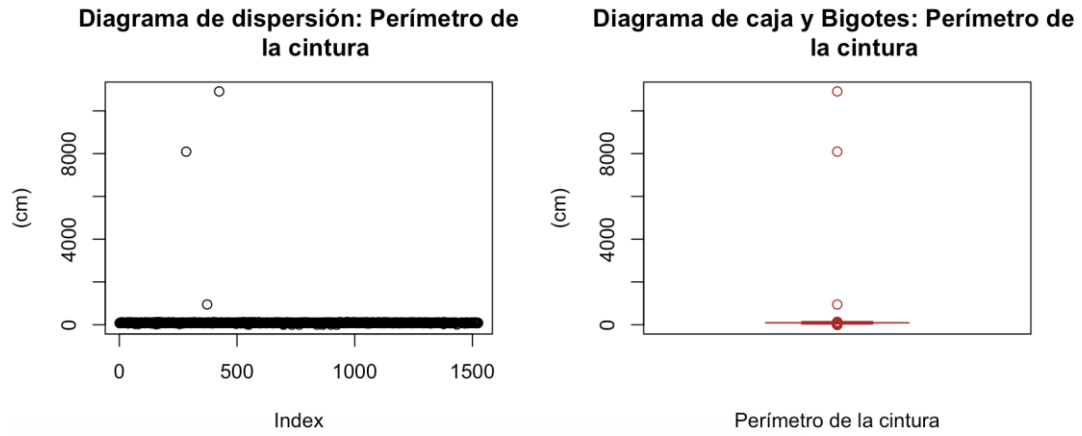
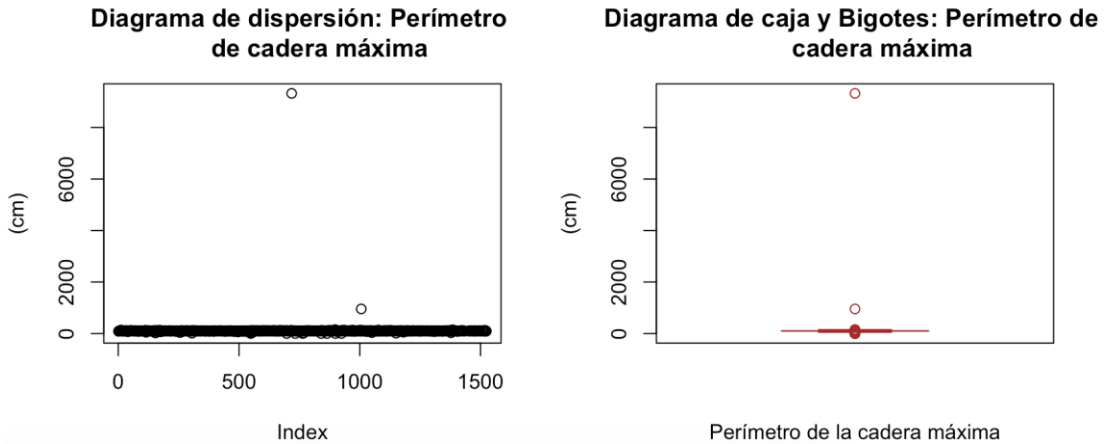
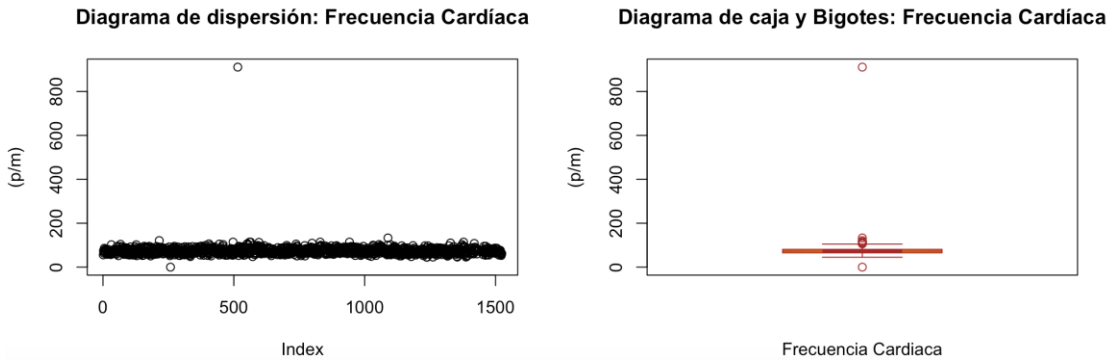


Figura 10. Diagramas de dispersión y Diagramas de caja y bigotes para el Perímetro de la cadera máxima



- Frecuencia Cardíaca: el diagrama de dispersión representado en la *figura 11* demuestra la existencia de un valor extremo superior de 911 pulsaciones por minuto y valores cercanos o iguales a cero pulsaciones por minuto. Se asume que estos valores son producto del ingreso erróneo de información a la base de datos, pues son datos que para el presente trabajo y contexto no son válidos.

Figura 11. Diagramas de dispersión y Diagramas de caja y bigotes para la Frecuencia Cardíaca



Por medio del rango intercuartílico, se identifican los valores atípicos para las variables expuestas anteriormente. Se decide eliminar dichos registros de la base de datos de estudio pues no representan valores válidos para ninguna de las variables analizadas. En total se descartaron 175 registros de los 1.524 originales.

3.2.2 Variables Derivadas

Al indagar sobre la relación de algunos parámetros antropométricos con el riesgo cardiovascular, se encuentran artículos que revelan que el Índice cintura/talla (ICT), índice cintura/cadera (ICC) y la circunferencia de la cintura (CC) son predictores precisos a la hora de discriminar el riesgo cardiovascular. Según Luengo Pérez, Urbano Gálvez y Pérez Miranda (2009), en un meta análisis realizado para determinar cuál es el mejor discriminador de enfermedades cardiovasculares como la hipertensión, diabetes mellitus tipo 2 y dislipidemia; entre los índices antropométricos IMC, CC, ICC y ICT, el ICT obtuvo mayor correlación estadística por separado para varones y mujeres; los autores expresan, además, que la CC también es un buen indicador del riesgo cardiovascular. Para este caso, se calculan los índices antropométricos IMC, ICC e ICT

Por último, mediante el uso de las tablas para la estimación del riesgo cardiovascular propuestas por la OMS para las regiones de las Américas (grupo B), se calcula el riesgo cardiovascular asociado a cada una de las observaciones que hacen parte de la base de datos de estudio. Las variables que intervienen en el cálculo son: edad, presión arterial sistólica, presencia de diabetes y tabaquismo. La OMS divide los niveles de riesgo cardiovascular en cuatro categorías, según el porcentaje obtenido con la aplicación de dichas tablas. Las categorías propuestas son:

- Bajo (< 10%)
- Medio (10% - <20%)
- Alto (20% - < 30%)
- Muy alto (≥ 30%).

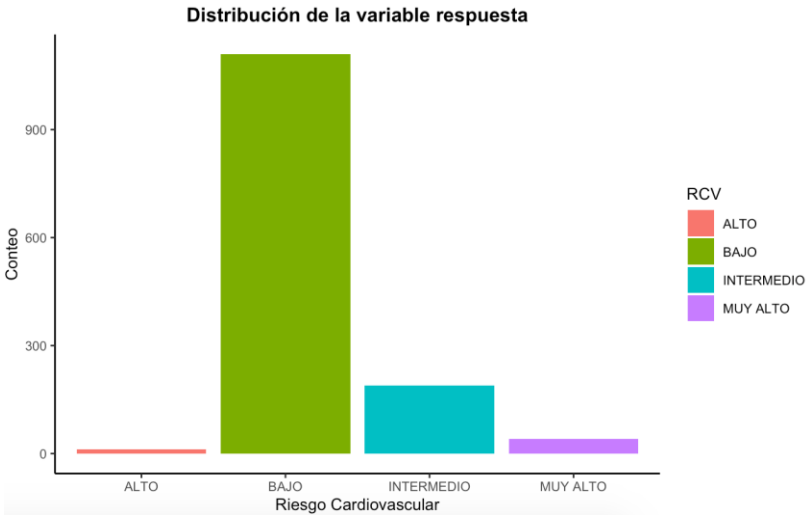
Estas categorías componen la variable respuesta para el problema de clasificación abordado.

3.2.3 Análisis descriptivo de las variables

Una vez preprocesada la base de datos de estudio, se procede con la descripción y análisis de algunas características de los datos para un mejor entendimiento de su comportamiento. Primero se verifica como está distribuida la variable respuesta, en la *figura 12* se observa que hay niveles de riesgo cardiovascular, como el alto y muy alto, cuya cantidad de muestras es inferior en comparación con el nivel bajo. Esto es evidentemente un problema de clases desbalanceadas y será abordado en el capítulo cuatro.

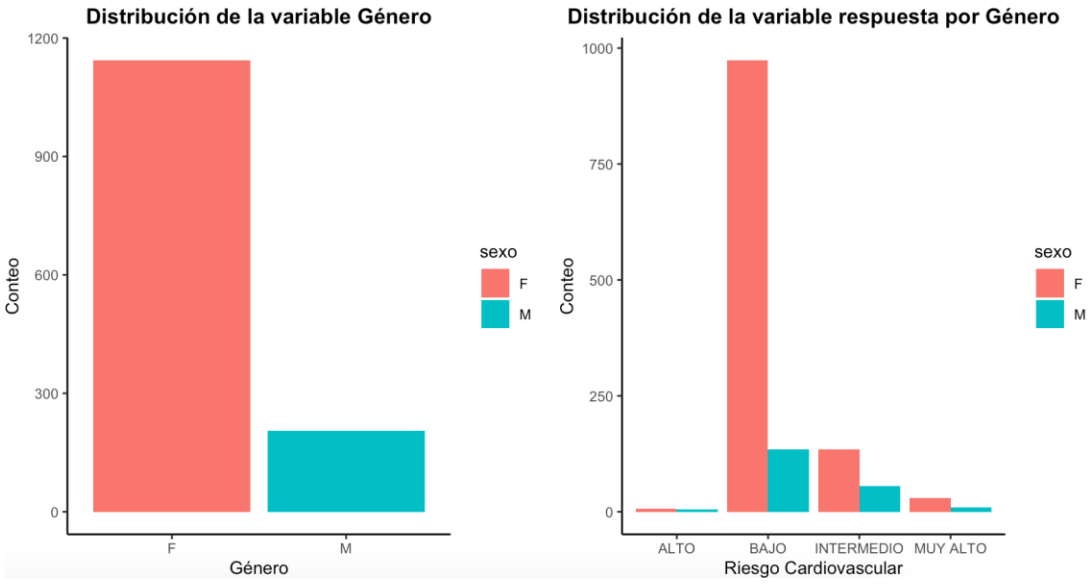
El 82,20% de la población de estudio presenta un riesgo cardiovascular bajo, mientras que el 3,78% posee un nivel de riesgo cardiovascular entre alto y muy alto. Respecto al nivel intermedio se pudo determinar que el 14,01% de los adultos mayores evaluados se sitúan en este nivel, estos adultos pueden estar en riesgo de pasar a un nivel alto si no se toman medidas oportunas que mitiguen los factores de riesgo a los que están expuestos.

Figura 12. Número de muestras para cada nivel de riesgo cardiovascular



En los estudios citados sobre el riesgo cardiovascular, variables como la edad, el sexo y algunos hábitos como fumar pueden influir en la propensión de tener un riesgo cardiovascular mayor. En la *figura 13*, se puede observar que, para todos los niveles de riesgo cardiovascular, las mujeres superan en cantidad a los hombres, representando el 84,80% de las muestras totales.

Figura 13. (1) Número de muestras por Género - (2) Riesgo cardiovascular por género.

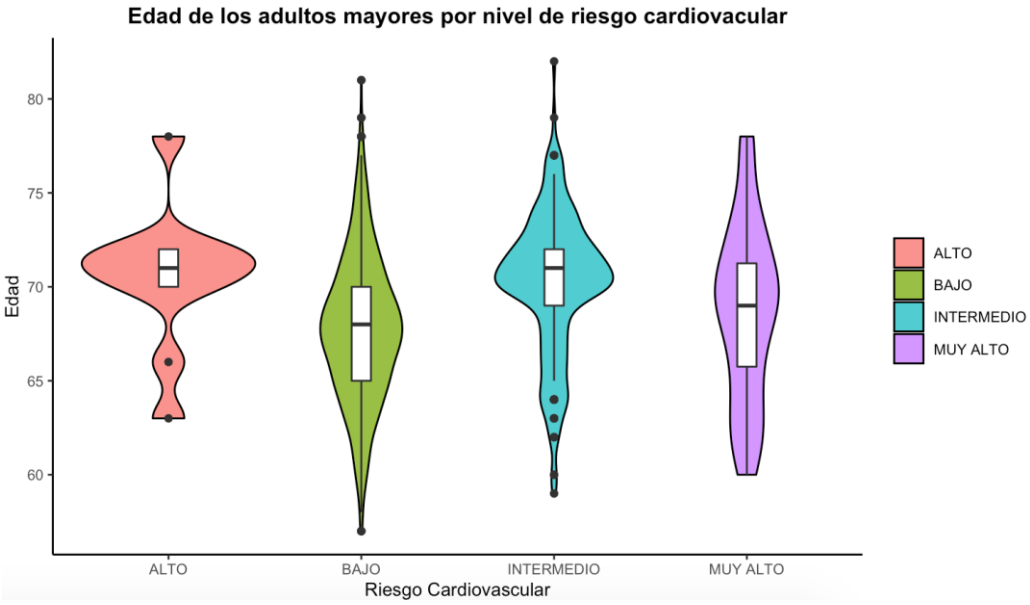


En la *figura 14* se identifica que, para la población de estudio con un nivel alto de riesgo cardiovascular, la edad es en promedio 72 años y en comparación con los demás niveles su

distribución se concentra en edades superiores a los 68 años. Lo anterior se debe a que, en edades cercanas a los 70 años, el riesgo de padecer ECV es mayor; aunque al observar la distribución del nivel muy alto, se evidencia la presencia de individuos más jóvenes, en este caso el nivel asociado estaría determinado por otros factores de riesgo como el peso, el padecimiento de diabetes o de hipertensión.

Además, en la *figura 14* para el nivel bajo de riesgo cardiovascular la edad de los adultos mayores presenta una distribución uniforme, con mayor frecuencia de edades que oscilan entre los 65 y 70 años, de igual manera se pueden observar casos en los que la edad es inferior a los 60 años, siendo la población más joven registrada para los diferentes niveles evaluados. Es importante recordar que dicho nivel presenta mayor número de muestras y por ende su rango de edades es más amplio.

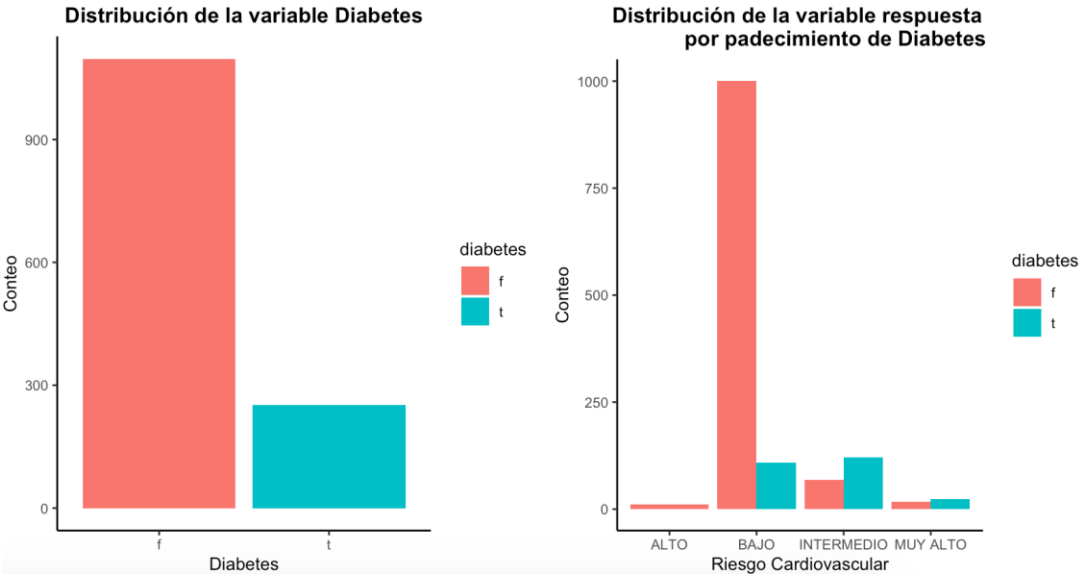
Figura 14. Distribución de la variable respuesta según la edad



Con respecto al padecimiento de diabetes, en la *figura 15* se evidencia que la mayoría de la población no tiene diagnosticada dicha enfermedad, pero al observar la distribución de la variable respuesta para los niveles de riesgo cardiovascular intermedio y muy alto, es notable que los adultos mayores con padecimiento de diabetes son más en comparación con aquellos que no la padecen, de manera puntual, el 72,60% de los individuos que registran un nivel de riesgo intermedio tienen diabetes, así mismo para el nivel muy alto este porcentaje es de

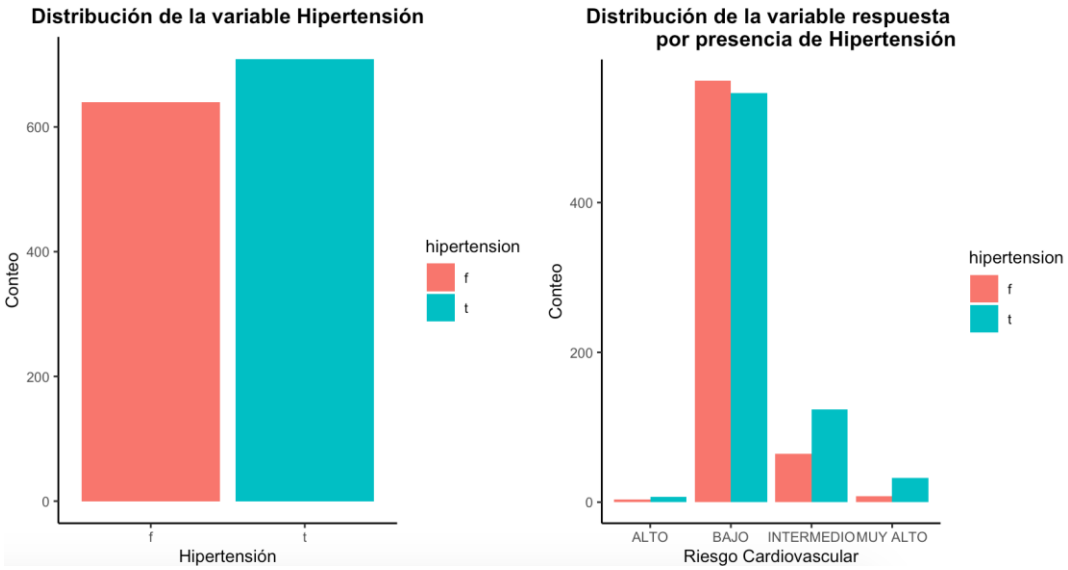
57,5%; esto corrobora que en los individuos que registran un nivel muy alto de riesgo cardiovascular, la edad no es una característica tan influyente como la diabetes.

Figura 15. (1) Número de muestras por padecimiento de diabetes - (2) Riesgo cardiovascular por padecimiento de diabetes.



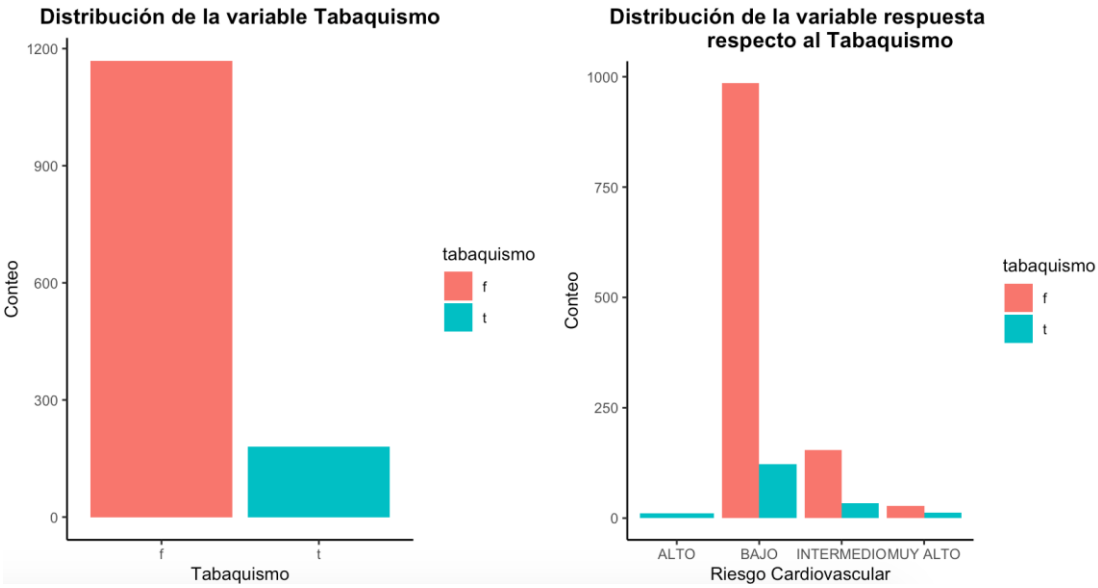
Otro factor determinante para la estimación del nivel de riesgo cardiovascular es la presencia de hipertensión; según la *figura 16*, la mayoría de los adultos mayores padecen de esta enfermedad (52,16%) y para los niveles altos de riesgo los individuos con hipertensión son mayoritarios (76.47%); para el nivel intermedio los adultos con hipertensión representan el 65,60% y para el nivel muy alto representan el 80%.

Figura 16. (1) Número de muestras por padecimiento de hipertensión - (2) Riesgo cardiovascular por padecimiento de hipertensión.



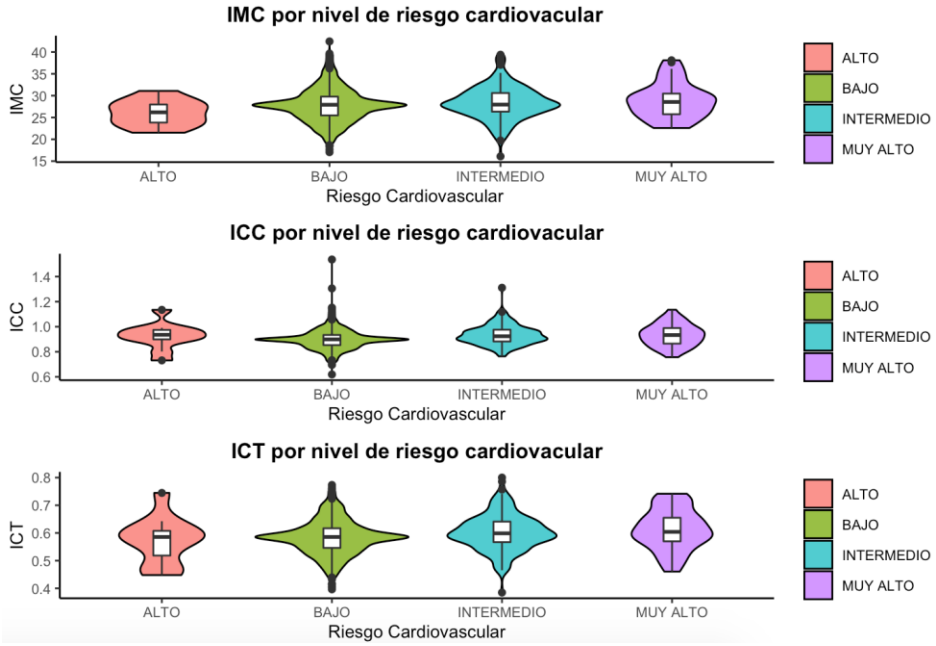
Siguiendo con el análisis de los factores de riesgo más influyentes, se observa que la mayoría de los adultos mayores no tienen el hábito de fumar y solo para un nivel alto de riesgo cardiovascular esta característica es predominante (*ver figura 17)*. Así mismo, los adultos mayores que padecen de obesidad (6.03%) son menos a los que no son obesos.

Figura 17. (1) Número de muestras según el hábito de fumar - (2) Riesgo cardiovascular según el hábito de fumar



También, en la *figura 17* se identifica la distribución de las medidas antropométricas calculadas: IMC – Índice de masa corporal -, ICC – Índice cintura/cadera e ICT – Índice cadera/talla - para cada nivel de riesgo cardiovascular. Dichas medidas presentan distribuciones muy similares para los niveles de riesgo intermedio y bajo; las medianas de cada nivel no presentan variaciones significativas.

Figura 18. Medidas antropométricas para cada nivel de riesgo cardiovascular

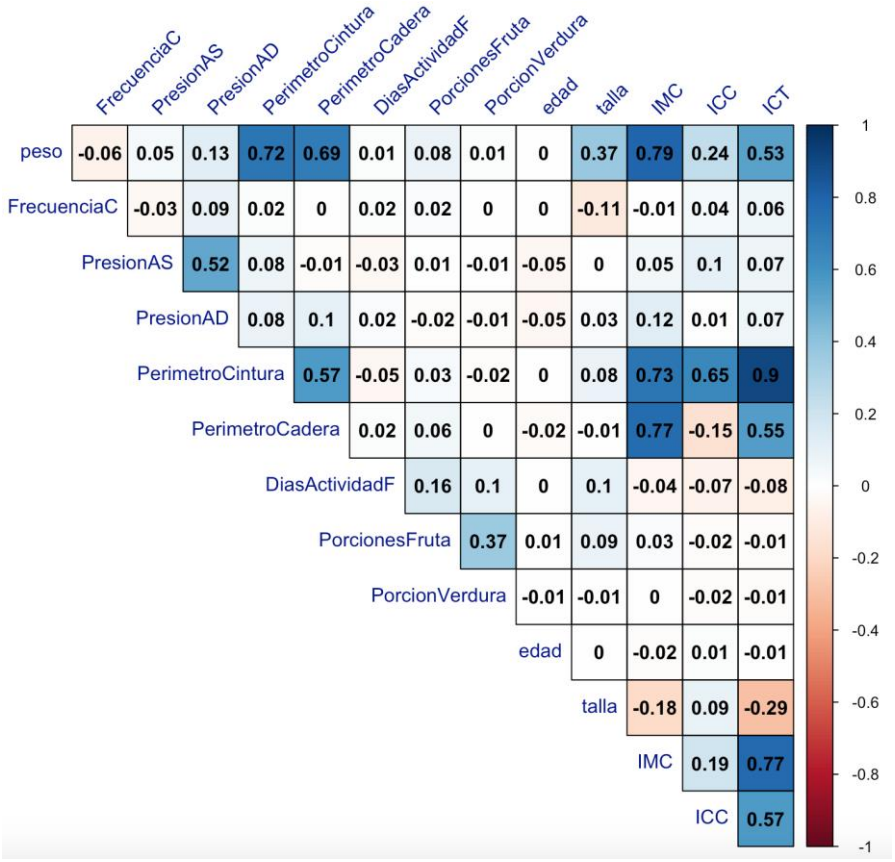


3.2.4 Análisis de correlación

Según Yu & Liu (2003), una característica es buena si es relevante para el concepto de clase pero no es redundante para ninguna de las otras características relevantes. Adoptando la correlación entre dos variables como medidas de asociación, según lo anterior, una característica es relevante si está altamente correlacionada con la clase, pero no con ninguna de las otras características.

En la *figura 19* se observa la matriz de correlación, calculada usando el coeficiente de correlación de Spearman, que se aplica en escenarios donde no es posible asumir un distribución normal sobre los datos (Restrepo & González, 2007). Solo se incluyeron para este análisis las variables numéricas de la base de datos de estudio.

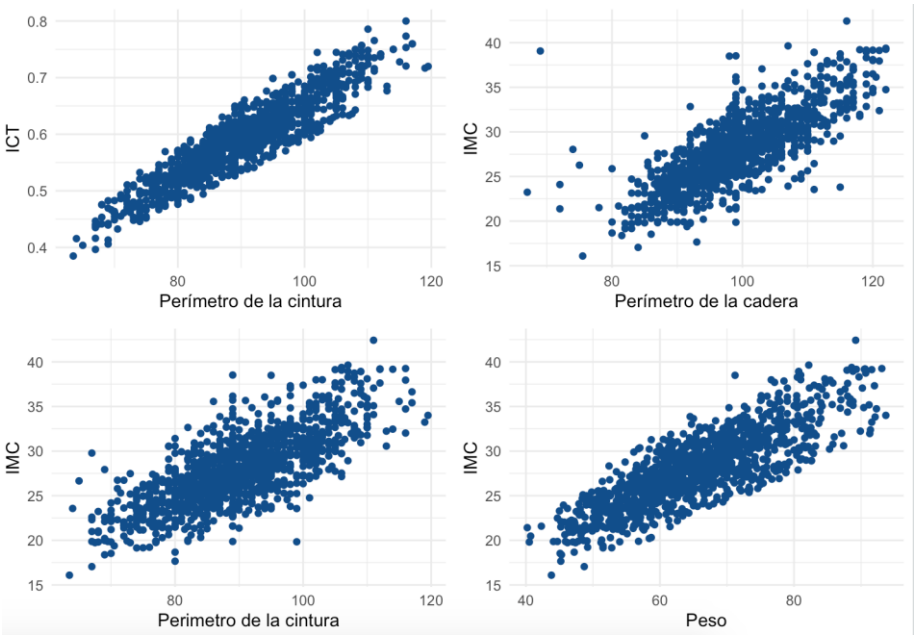
Figura 19. Matriz de correlación entre variables numéricas usando coeficiente de Spearman



Abreviaturas: *FrecuenciaC – Frecuencia Cardiaca. PresionAS – Presión Arterial sistólica. PresionAD – Presión arterial diastólica. DiasActividadF: Días de actividad física moderada*.

Según la figura 19, existe entre algunas variables una correlación positiva significativa, mayor a 0.70. El coeficiente de correlación de Spearman tal y como lo explica Restrepo y González (2007), mide la tendencia de *X, Y* a relacionarse en forma monótona creciente o decreciente. Al medir el grado de asociación entre las variables de forma monótona, el coeficiente no se limita a descubrir sólo una asociación lineal. Como se observa en la *figura 20*, las variables con altos coeficientes de correlación presentan un patrón lineal entre ellas, haciendo más evidente el tipo de asociación existente.

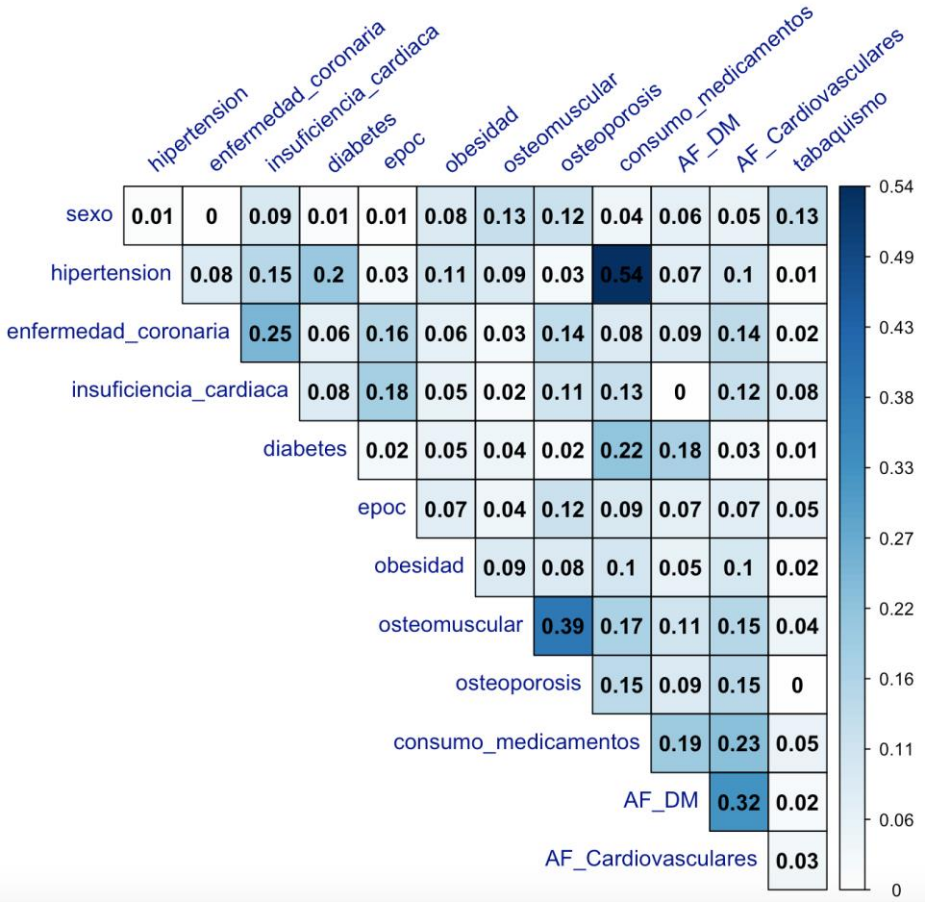
Figura 20. Gráficos de dispersión entre pares de variables correlacionadas



Como primer filtro, antes de proceder con la selección de variables, se establece que dos variables están altamente correlacionadas si tienen un coeficiente de Spearman mayor a 0.80. Se descarta, entre el par de variables evaluadas, aquella que tenga una mayor correlación absoluta media con el resto de las variables. En este caso, se excluye la variable calculada *ICT*.

Para determinar la asociación entre variables categóricas, se calcula el estadístico V de Cramer. Este es una alternativa a *Phi* de la prueba Chi-cuadrado en tablas de tabulación de cualquier dimensión, que varía entre 0 y 1 sin valores negativos. Similar al coeficiente de Pearson, un valor cercano a 0 indica que no hay asociación entre las variables. Sin embargo, un valor superior a 0,25 se denomina una relación muy fuerte para la V de Cramer (Akoglu, 2018). Las medidas basadas en el test de Chi-cuadrado generalmente son difíciles de interpretar, por tanto, es recomendable utilizar la V de Cramer porque, a diferencia de las demás medidas, su carácter estandarizado permite al menos comparar la "estrechez de la asociación" entre tablas diferentes; pero esta estrechez o fortaleza de la asociación comparada, no responde a ningún concepto intuitivo claro de asociación (Navarro Céspedes, 2008). La matriz de asociación con el estadístico V de Cramer para variables categóricas se representa en la *figura 21*.

Figura 21. Asociación entre variables categóricas usando V de Cramer



Abreviaturas: *AF_DM: Antecedentes Familiares de Diabetes mellitus. AF_Cardiovasculares – Antecedentes Familiares Cardiovasculares.*

En la *figura 21* se evidencia que existe una fuerte asociación entre el consumo de medicamentos y el padecimiento de hipertensión; siguiendo el criterio de evaluación definido para las variables numéricas, se opta por descartar una de ellas, en este caso, se conserva la variable dicotómica de hipertensión.

4. Modelado y evaluación

La fase de modelado se desarrolla en Python, donde utiliza principalmente la librería *scikit-learn* para la aplicación de transformaciones y algoritmos de aprendizaje automático. Esta librería es de código abierto y posibilita la aplicación de aprendizaje supervisado y no supervisado. También proporciona varias herramientas para el ajuste de modelos, el preprocesamiento de datos, la selección y evaluación de modelos y variables, entre muchas otras utilidades (Pedregosa et al., 2012). En este capítulo se exponen las actividades desarrolladas como parte del proceso de modelado y sus respectivos resultados.

4.1 Preparación de los datos

Esta fase comienza con la preparación de los datos para la aplicación de los algoritmos de clasificación propuestos. Para ello, es necesario garantizar que los datos estén en una misma escala y formato; y que además sean los correctos. Parte de esta fase se desarrolló en el capítulo tres, donde se convirtieron las variables categóricas de entrada en variables *dummy* o dicotómicas. Dichas variables son incorporadas en la modelación para medir el efecto de determinada característica de los individuos en la muestra o para capturar algún cambio estructural durante el periodo de estudio; una variable *dummy* toma el valor de uno cuando la característica de interés está presente y cero en caso contrario (Alonso & Munoz, 2014).

La variable respuesta, por su parte, está compuesta por cuatro clases: bajo, intermedio, alto y muy alto, que indican el nivel de riesgo cardiovascular. Esta propiedad hace que la tarea de clasificación sea multiclase. Se utilizan dos enfoques diferentes para la incorporación de la variable respuesta en los algoritmos de clasificación evaluados; para la regresión logística y el algoritmo de máquina de aumento de gradiente se utiliza el enfoque uno contra todos (*one against all o one vs rest*); este enfoque heurístico se usa para el tratamiento de variables

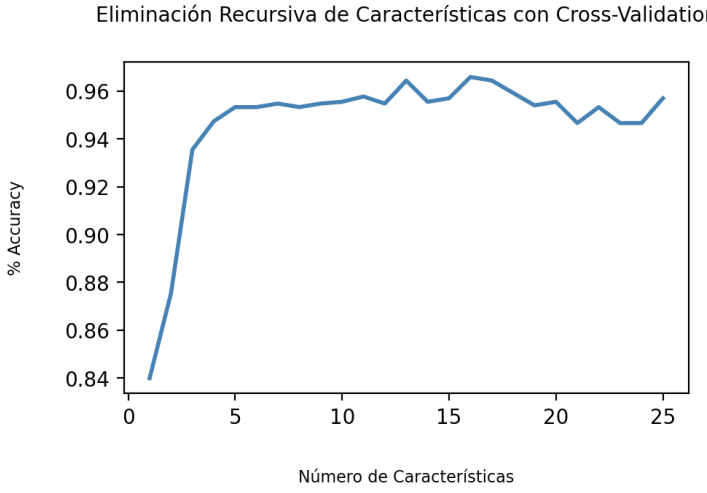
multiclases, donde se divide el conjunto de datos de clases múltiples en varios problemas de clasificación binaria enfocados a predecir una sola clase (Bengio et al., 2010). Para el algoritmo de bosque aleatorio se codifica la variable numéricamente, conservando el orden entre los niveles de riesgo.

Posterior a estas transformaciones, se estandarizan todas las variables de entrada con el fin de mitigar el efecto de escalas diferentes, aunque para la aplicación del método de eliminación recursiva de características, con el que se definió el set de características para el modelo de clasificación, no se usaron datos normalizados, dado que el algoritmo de bosque aleatorio no es sensible al escalado de las variables.

4.2 Selección de variables

Se aplica el método de eliminación recursiva de características (RFE) con el algoritmo de bosque aleatorio como modelo estimador. Además, se incorpora el método de validación cruzada para la evaluación del desempeño del modelo estimador por cada número diferente de características consideradas. En *la figura 22* se observa la exactitud obtenida con cada conjunto de características utilizadas.

Figura 22. Exactitud del modelo estimador para cada conjunto de características consideradas.

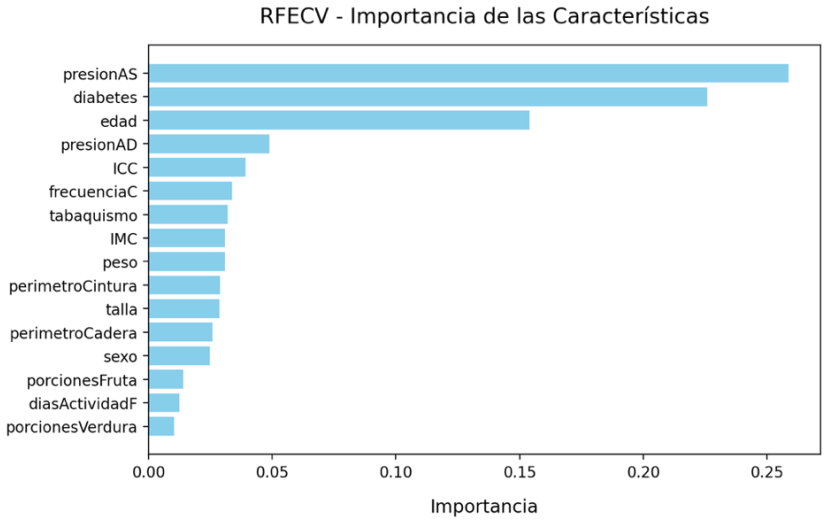


La cantidad óptima de características a considerar es 16 y según el modelo de selección, no se excluyen de los análisis posteriores las variables de: hipertensión, enfermedad coronaria, insuficiencia cardiaca, enfermedad pulmonar obstructiva crónica, obesidad, enfermedad

osteomuscular, osteoporosis, antecedentes familiares de diabetes y antecedentes familiares cardiovasculares. El modelo estimador genera la importancia relativa o contribución de cada característica en la predicción, calculando automáticamente la puntuación de relevancia de cada una en la fase de entrenamiento. El algoritmo de bosque aleatorio utiliza como criterio de importancia la impureza de *Gini*.

En la *figura 23* se identifica el nivel de importancia para las variables seleccionadas, dentro de las cuales la presión arterial sistólica, la diabetes y la edad son las de mayor puntaje y la variable derivada ICC se encuentra entre las cinco principales. Con solo estas cinco variables el modelo estimador podría obtener un resultado cercano al obtenido con todas las variables, dado que las variaciones de la exactitud del modelo no son grandes en magnitud al adicionar más variables, esto puede comprobarse en la *figura 22.*

Figura 23. Importancia de las variables seleccionadas por el RFE con validación cruzada



Abreviaturas: *AF_DM: Antecedentes Familiares de Diabetes mellitus. AF_Cardiovasculares – Antecedentes Familiares Cardiovasculares.*

4.3 Partición de los datos y métricas de evaluación

Una vez definido el conjunto de datos final, se utiliza el método de validación cruzada para el entrenamiento y evaluación de los modelos propuestos; este método particiona de manera aleatoria los datos disponibles en dos conjuntos, uno de entrenamiento y otro de validación, lo que permite validar el algoritmo sobre un conjunto de datos diferente del empleado para estimar sus parámetros (Otero & Sánchez, 2007). En este caso la partición de los datos se hace

conservando una proporción de 70% para el conjunto de entrenamiento (944 muestras) y 30% para el conjunto de validación (405 muestras).

Para el tratamiento de las clases desbalanceadas, identificadas en el capítulo anterior, se recurre al uso de métodos basados en técnicas de muestreo, los cuales duplican o eliminan patrones o muestras de entrenamiento hasta alcanzar un relativo equilibrio entre el número de muestras de las distintas clases (Monroy-de-Jesús et al., 2018); específicamente, se aplica la técnica de sobre muestreo de minorías sintéticas (SMOTE), este es un enfoque que crea muestras sintéticas interpoladas entre patrones de la clase minoritaria. Potencialmente, funciona mejor que el sobre muestreo simple y es usado ampliamente (Blagus & Lusa, 2013).

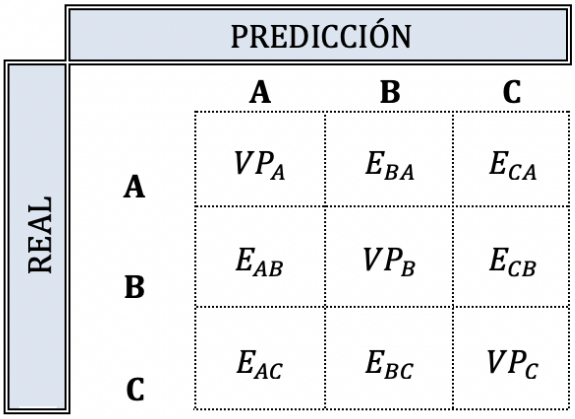
En cuanto a la medición del rendimiento de los modelos, se eligen cuatro métricas usadas con frecuencia en problemas de clasificación: precisión, exactitud, sensibilidad y f1-score. Como la tarea de clasificación es multiclase, se calcula el promedio de las métricas listadas anteriormente para cada clase o macro promedio, dado que permite realizar una aproximación general de la calidad de la clasificación (Sokolova & Lapalme, 2009; Tharwat, 2020).

Según Sokolova y Lapalme (2009), estas métricas pueden definirse como:

- Precisión: promedio por clase de los ejemplos positivos clasificados correctamente dividido el número total de ejemplos que el modelo clasificó como positivos.
- Exactitud: efectividad (clasificaciones correctas) promedio por clase del modelo.
- Sensibilidad: promedio por clase de los ejemplos positivos clasificados correctamente dividido por el número total de ejemplos positivos.
- F1-score*:* relación entre las etiquetas positivas de los datos y las dadas por un clasificador según un promedio por clase.

Estas métricas se derivan del análisis de la matriz de confusión, que es una manera gráfica de representar el desempeño del clasificador. Cada columna de la matriz se conforma por el número de predicciones de cada clase realizadas por el modelo y las filas son las clases reales, con lo que se puede identificar los aciertos y errores para cada clase. En la *figura 24* se puede encontrar la estructura de una matriz de confusión para un problema de clasificación multiclase, donde *VP* se refiere a los verdaderos positivos (ejemplos correctamente clasificados) y *E* a los errores o ejemplos mal clasificados.

Figura 24. Matriz de confusión para un problema de clasificación multiclase



Representación adaptada de (Sokolova & Lapalme, 2009)

4.4 Optimización de hiperparámetros

Con el uso de la búsqueda en grilla se optimizan algunos hiperparámetros de los algoritmos de clasificación propuestos, en las *tablas 4, 5 y 6* se detalla el valor de referencia en el modelo base (antes del cambio del hiperparámetro) y el valor óptimo definido. En el apartado de resultados se encuentra el detalle del desempeño de los modelos antes y después de la optimización de hiperparámetros.

Tabla 4. Hiperparámetros óptimos para el modelo de regresión logística de mejor desempeño

Hiperparámetro	Valor inicial (Referencia)	Valor óptimo
Número máximo de iteraciones	100	100
Penalización	L2	L2
C	1	1
Peso de las clases	ninguno	balanced

Tabla 5. Hiperparámetros óptimos para el modelo de bosque aleatorio de mejor desempeño

Hiperparámetro	Valor inicial (Referencia)	Valor óptimo
Número de estimadores	100	500
Mínimo de muestras para división del nodo	2	8
Max Feature	auto	auto
Profundidad Máxima del árbol	ninguno	10
Criterio	Gini	Entropy

Tabla 6. Hiperparámetros óptimos para el modelo de máquina de aumento de gradiente de mejor desempeño

Hiperparámetro	Valor inicial (Referencia)	Valor óptimo
Número de árboles en el bosque	100	100
mínimo de muestras para división del nodo	2	10
Max Feature	auto	auto
Profundidad Máxima de los estimadores	ninguno	6
Submuestras	1	0.9

4.5 Resultados

La fase de modelado inicia con el entrenamiento de un modelo base para cada uno de los algoritmos de clasificación considerados, en el que no se realiza ningún cambio o adición de hiperparámetros. En la *tabla 7* se encuentran los resultados obtenidos en el cálculo de las métricas de desempeño; estos resultados sirven como punto de referencia para determinar si las técnicas de optimización aplicadas, como el sobre muestreo y el ajuste de hiperparámetros, aportaron al mejoramiento de las métricas calculadas.

Tabla 7. Resultado de las corridas de los modelos base

Modelo Base	Exactitud	Precisión	Sensibilidad	F1-score
	Validación	Validación	Validación	Validación
Regresión logística	0.8987	0.8961	0.6477	0.7234
Bosque aleatorio	0.9555	0.9691	0.8166	0.8808
Máquina de aumento de gradiente	0.9753	0.8178	0.7609	0.7864

Los modelos base obtienen un desempeño bueno en términos de exactitud, aunque para los algoritmos de regresión logística y máquina de aumento de gradiente, la métrica de sensibilidad es baja (inferior al 80%), esto no es un buen indicador para el problema de clasificación, ya que esta métrica se relaciona directamente con el número de falsos negativos que el modelo obtiene, es decir, es más probable que no se clasifiquen, por ejemplo, individuos en niveles altos de riesgo cardiovascular, cuando en realidad si lo presentan. Este resultado está asociado al problema de clases desbalanceadas.

Dicho problema, se trata con la aplicación de técnicas de muestreo, puntualmente con la técnica de SMOTE. En las *tablas 8, 9 y 10* se presentan los resultados de los algoritmos de clasificación propuestos bajo cuatro diferentes estrategias de SMOTE: SMOTE (sobre muestreo sintético de las clases minoritarias), ADASYN (sobre muestreo de las clases minoritarias con un enfoque sintético adaptativo), BorderlineSMOTE (sobre muestreo de las clases minoritarias que se centra en muestras cercanas al límite de la función de decisión), SMOTETomek (combinación de sobre muestreo de las clases minoritarias y sub muestreo de la clase mayoritaria) (Lemaître et al., 2017).

Tabla 8. Resultado de las métricas de desempeño para el algoritmo de regresión logística con diferentes estrategias de SMOTE

Estrategia	Exactitud	Precisión	Sensibilidad	F1-score
	Validación	Validación	Validación	Validación
SMOTE	0.8864	0.6467	0.7516	0.6353
ADASYN	0.8839	0.6448	0.7479	0.6319
BorderlineSMOTE	0.8543	0.5598	0.7613	0.6226
SMOTETomek	0.8666	0.5810	0.7924	0.6479

Tabla 9. Resultado de las métricas de desempeño para el algoritmo de bosque aleatorio con diferentes estrategias de SMOTE

Estrategia	Exactitud	Precisión	Sensibilidad	F1-score
	Validación	Validación	Validación	Validación
SMOTE	0.9432	0.8231	0.7495	0.7620
ADASYN	0.9506	0.8825	0.8828	0.8627
BorderlineSMOTE	0.9382	0.8841	0.7810	0.8233
SMOTETomek	0.9530	0.8372	0.8158	0.8256

Tabla 10. Resultado de las métricas de desempeño para el algoritmo de máquina de aumento de gradiente con diferentes estrategias de SMOTE

Estrategia	Exactitud	Precisión	Sensibilidad	F1-score
	Validación	Validación	Validación	Validación
SMOTE	0.9654	0.8568	0.9081	0.8682
ADASYN	0.9703	0.8586	0.9156	0.8733
BorderlineSMOTE	0.9604	0.7950	0.8883	0.8201
SMOTETomek	0.9629	0.8013	0.9133	0.8362

Es evidente que la aplicación de técnicas de muestreo antes del entrenamiento de los modelos propuestos incrementa, en la mayoría de los casos, las métricas de desempeño y en particular mejora de manera significativa la métrica de sensibilidad en los algoritmos de regresión logística y máquina de aumento de gradiente. Adicionalmente, los algoritmos de regresión logística y bosque aleatorio tienen la opción de especificar por medio del hiperparámetro *class*

weight, que los pesos de las clases que componen la variable dependiente sean inversamente proporcionales a las frecuencias de estas, en cuyo caso se estaría haciendo énfasis en las clases minoritarias y tratando de manera implícita el desequilibro de clases sin usar técnicas de sobre muestreo. En la *tabla 11* se encuentran los resultados de los modelos con el ajuste de este hiperparámetro.

Tabla 11. Resultado de las métricas de desempeño para los algoritmos de regresión logística y bosque aleatorio con hiperparámetro *class weight* ajustado

Modelo	**Exactitud**	**Precisión**	**Sensibilidad**	**F1-score**
	Validación	Validación	Validación	Validación
Regresión logística	0.8839	0.6811	0.8150	0.7342
Bosque aleatorio	0.9308	0.9657	0.7763	0.8528

Para la búsqueda de hiperparámetros óptimos, se considera la estrategia con la que se obtuvieron mejores resultados en el tratamiento de clases desbalanceadas, con excepción del modelo de bosque aleatorio, en el que las estrategias aplicadas no mejoran el desempeño de este. Para el caso del algoritmo de regresión logística el ajuste del hiperparámetro *class weight* obtiene mejores resultados que las estrategias de SMOTE y para el algoritmo de máquina de aumento de gradiente la estrategia de ADASYN obtiene mejores métricas. En la *tabla 12* se encuentran los resultados de las métricas de exactitud, precisión, sensibilidad y f1-score de los modelos con los hiperparámetros óptimos.

Tabla 12. Resultado de las métricas de desempeño para los algoritmos de clasificación considerados con ajuste de hiperparámetros

Modelo optimizado	**Exactitud**	**Precisión**	**Sensibilidad**	**F1-score**
	Validación	Validación	Validación	Validación
Regresión logística	0.8839	0.6811	0.8150	0.7342
Bosque aleatorio	0.9530	0.9710	0.8099	0.8775
Máquina de aumento de gradiente	0.9728	0.9231	0.9406	0.9260

El algoritmo de máquina de aumento de gradiente con hiperparámetros optimizados y sobre muestreo es el que mejores métricas obtiene con los datos de validación y por ende es el modelo final con el que se pretende abordar el problema de clasificación planteado en este trabajo. Evaluando el desempeño por clase del modelo, se pudo identificar que para las clases minoritarias de alto y muy alto riesgo cardiovascular, el algoritmo en la mayoría de los casos puede identificar de manera correcta los ejemplos correspondientes a estas clases (ver *tabla 13*), dando evidencia que corrobora la efectividad de la estrategia de muestreo aplicada.

Haciendo énfasis en el nivel alto de riesgo cardiovascular, la precisión obtenida con el modelo es del 75% pero la sensibilidad es del 100%, esto indica que el modelo identifica la clase muy bien, pero incluye en algunos casos muestras de otras clases, es decir, existen falsos positivos. Para los casos en que la presencia de falsos negativos tiene un alto impacto, como en el diagnóstico de enfermedades, la sensibilidad es el indicador que mejor podría determinar la calidad del modelo, dado que no predecir correctamente un caso positivo de un nivel de riesgo como el alto o el muy alto tendría un alto costo. Si se observan los resultados para la métrica de f1-score, que es una medida de equilibrio entre la precisión y la sensibilidad, para todas las clases, esta métrica es superior o igual al 85%, siendo un resultado relativamente alto para una base de datos con desequilibrio de clases.

Tabla 13. Resultado de las métricas de desempeño por clase para el algoritmo de máquina de aumento de gradiente

Clase	**Exactitud**	**Precisión**	**Sensibilidad**	**F1-score**
	Validación	Validación	Validación	Validación
0: Bajo	0.97	0.98	1.0	0.99
1: Intermedio		0.97	0.87	0.91
2: Alto		0.75	1.0	0.86
3: Muy alto		1.0	0.90	0.95
Macro promedio	**0.95**	**0.92**	**0.94**	**0.93**

5. Discusión

En este estudio se evalúan tres modelos de ML para la clasificación de adultos mayores pertenecientes a un municipio de Antioquia en cuatro niveles de riesgo cardiovascular. Los modelos se entrenan con datos de 1.524 evaluaciones físicas y antropométricas, de las cuales se reconocen variables clínicas, socio-conductuales, de contexto clínico e índices antropométricos, por lo anterior se dice que el modelo de clasificación es hibrido. Del análisis descriptivo de los datos, se identifica que el 82.20% de los adultos mayores tienen un riesgo bajo de ECV, lo que posteriormente se trata como un problema de clases desbalanceadas; el 84,8% de las evaluaciones son de mujeres y del total de la población, aproximadamente el 52% padece de hipertensión. Otros factores de riesgo cardiovascular como la diabetes y el tabaquismo no son predominantes.

Hasta donde se sabe, este es el primer estudio que aplica técnicas de ML a datos médicos para una población de adultos mayores en Colombia, demostrando la factibilidad de incorporar herramientas como las presentadas en campos de alto impacto, como lo es el de la salud. El estudio presentado ilustra de manera detallada aspectos importantes para abordar problemas de calidad presentes en los datos y algunas alternativas para su tratamiento, también expone como el ML puede ser aplicado para la selección de variables, además, de diferentes métricas de evaluación que facilitan la interpretación del desempeño de los modelos de ML considerados.

El modelo con mejores resultados en las métricas de exactitud (97.28%), sensibilidad (94.06%) y f1-score (92.60%), se obtiene con el algoritmo de máquina de aumento de gradiente con hiperparámetros optimizados y la aplicación de estrategias de muestreo para el balanceo de clases. La precisión de este algoritmo (92.31%) no supera la precisión obtenida con el algoritmo de bosque aleatorio (97.10%), pero como criterio de selección se les da mayor

importancia a las métricas de sensibilidad y f1-score. El modelo óptimo de clasificación propuesto obtiene un desempeño alto en comparación con estudios similares, donde logra superar los resultados obtenidos por Weng *et al.* (2017), que utiliza el mismo algoritmo de clasificación, con una precisión y sensibilidad corresponden a 67.5% y 70.7% respectivamente.

También, el modelo óptimo puede contrastarse con otros estudios, que aunque no obtienen los mejores resultados con el mismo algoritmo, si pretenden abordar la predicción del riesgo cardiovascular, por ejemplo Castellanos Vázquez, Moreno, Bouza Herrera, *et al.* (2019), mediante el uso de bosque aleatorio obtienen una precisión del 97% y métrica de sensibilidad de 97.1%, superando por poco al modelo óptimo. Otros estudios como el de Ward *et al.* (2020) y A. M. Alaa *et al.* (2019) utilizan para la evaluación de los modelos de clasificación la AUC de la curva ROC, para este trabajo dicha métrica no fue considerada como indicador de calidad o criterio de selección del mejor modelo, pero para efectos comparativos se calcula usando la estrategia de *one vs rest*. El resultado de esta métrica para el modelo óptimo es de 98.46%, que es significativamente mayor a los reportados en los estudios mencionados.

Se destaca de este trabajo la efectividad del modelo óptimo para la predicción de clases de mayor impacto para la salud correspondientes a niveles altos de riesgo cardiovascular. Dicha efectividad se identifica con el cálculo de diferentes métricas de desempeño derivadas del análisis de la matriz de confusión. La importancia de las métricas depende del problema de clasificación, puesto que para algunos casos se espera reducir la porción de falsos positivos y en otros se espera reducir la porción de falsos negativos, por tanto, se recomienda complementar los resultados de métricas estándar como la exactitud con métricas más específicas como la precisión, la sensibilidad o la AUC de la curva ROC.

La calidad de los resultados arrojados por el modelo puede ser mejor si se incluyen variables como pruebas de laboratorio que en este caso no estuvieron disponibles en la base de datos de estudio y que son incorporadas en varios modelos de estimación de riesgo cardiovascular como parte de los predictores base (Martínez et al., 2016) . El método de estimación aplicado a pesar de ser avalado por la OMS y el Ministerio de salud de Colombia, presenta ciertas limitaciones que son expuestas por autores como Muñoz et al. (2014) en su estudio de validación de modelos predictivos de riesgo cardiovascular en una población colombiana, donde establecen que el modelo de Framingham sobrestima el riesgo y demuestra una baja

capacidad de discriminación. Por esta razón los autores recomiendan usar con precaución dicho modelo y más en poblaciones colombianas de riesgo bajo e intermedio sin historia previa de ECV. El modelo PROCAM ajustado por sexo resulta ser una mejor opción para estimar el riesgo de ECV según los resultados expuestos por los autores.

Como futuros estudios, se plantea la oportunidad de validar este modelo y las técnicas de preprocesado usadas en conjuntos de datos que tengan un grado de control mayor en su recolección y que incluyan variables que permitan utilizar otros métodos de estimación de riesgo cardiovascular que hayan sido recalibrados para la población colombiana. De igual manera, se puede explorar el uso de los modelos evaluados en conjuntos de datos más grandes, que incluya poblaciones de diferentes regiones del país y con más variabilidad en cuanto a las comorbilidades existentes, lo que permitiría identificar la importancia de diferentes predictores ante muestras más heterogéneas a las disponibles para este estudio. Se debe tener en cuenta que expandir la evidencia actual permitiría evaluar la viabilidad de las aplicaciones de ML en la práctica clínica, lo que en ultimas impulsaría el desarrollo de la capacidad tecnológica de los sistemas de salud del país, ante la necesidad de que se centralice la información de los pacientes en sistemas eficientes para su acceso y análisis.

6. Conclusiones

6.1 Objetivo 1. Caracterizar variables de evaluación físicas y antropométricas para una población de estudio de adultos mayores

En el capito tres de este trabajo final, se aprecian los principales resultados del análisis exploratorio aplicado a los datos disponibles para el desarrollo del modelo de clasificación de riesgo cardiovascular. Además, se exponen las diferentes técnicas consideradas para tratar los problemas de calidad identificados en los datos como parte de las estrategias definidas para mejorar el desempeño de los modelos evaluados. Como principales características encontradas se destacan la cantidad de adultos mayores con riesgo cardiovascular bajo, que representa más del 80% de la población y la prevalencia de la hipertensión con más del 50% de la población diagnosticada con esta enfermedad; caso contrario ocurre con la diabetes (16.53%) y la obesidad (6.03%) que también son factores importantes de riesgo cardiovascular.

6.2 Objetivo 2. Seleccionar variables de evaluación físicas y antropométricas para una población de estudio de adultos mayores, según su relevancia para el modelo de clasificación

En este trabajo final se aplica un método de selección de variables tipo *wrapper* conocido como eliminación recursiva de características (RFE). En el capítulo cuatro se aprecian los resultados obtenidos por el algoritmo de bosque aleatorio usado como modelo estimador de la importancia de las variables, cuya selección optima es de 16 características que representan una reducción de la dimensionalidad de aproximadamente el 55%. Dentro de las cinco variables con mayor importancia para el modelo de selección se encuentran la presión arterial

sistólica, la diabetes, la edad, la presión arterial diastólica y el índice antropométrico cintura/cadera.

6.3 Objetivo 3. Construir un modelo híbrido para la clasificación del riesgo asociado al padecimiento de enfermedades cardiovasculares que pueda ser aplicado en una población de adultos mayores.

En este trabajo se desarrolla un modelo de clasificación aplicando una metodología propia de proyectos de minería de datos, dicho modelo se selecciona después de evaluar y comparar su desempeño con otros modelos de ML considerados. El modelo de ML óptimo es obtenido mediante la optimización de hiperparámetros del algoritmo de máquina de aumento de gradiente y la aplicación de una estrategia de SMOTE en clases minoritarias de riesgo cardiovascular. El desempeño de los modelos de regresión logística y bosque aleatorio es alto en comparación con estudios similares, pero el modelo óptimo logra discriminar mejor entre los niveles altos de riesgo cardiovascular. Los resultados de la etapa de modelado se pueden encontrar en el capito cuatro de este trabajo.

La incorporación de modelos de ML como el que se expone a lo largo de este trabajo puede apoyar la identificación preliminar de pacientes con propensión a desarrollar una ECV en el mediano plazo, esto puede ser beneficioso si se piensa en la prevención y reducción de los factores de riesgo cardiovasculares modificables, que puede lograrse con la incorporación de hábitos de vida más saludables. En especial, para la población de adultos mayores, es primordial la atención y seguimiento constante de su salud, dado que, por sus razones de dependencia tienen mayor exposición a situaciones que pueden vulnerar su bienestar y herramientas como la desarrollada en este trabajo, pueden contribuir con diagnósticos tempranos relacionados a ECV.

6.4 Objetivo 4. Validar el modelo de clasificación de grupos poblacionales de adultos mayores según el nivel de riesgo asociado al padecimiento de enfermedades cardiovasculares mediante técnicas de validación cruzada.

Los modelos considerados para la clasificación de adultos mayores según el riesgo de padecer ECV se evaluaron con cuatro métricas de desempeño usadas en tareas de clasificación, estas corresponden a: exactitud, precisión, sensibilidad y f1-score. Para el cálculo de las métricas de desempeño se aplica la técnica de validación cruzada, que divide el conjunto de datos disponible en dos subconjuntos, uno para el entrenamiento de los modelos y otro para la validación. La proporción establecida en este caso es de 70% para el subconjunto de entrenamiento y 30% para el subconjunto de validación. Los resultados de las métricas en los diferentes escenarios de modelado propuestos se exponen en el capito cuatro de este estudio.

7. Referencias

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93. https://doi.org/https://doi.org/10.1016/j.tjem.2018.08.001

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS ONE*, *14*(5), e0213653. https://doi.org/10.1371/journal.pone.0213653

Alizadehsani, R., Roshanzamir, M., Hussain, S., Khosravi, A., Koohestani, A., Zangooei, M. H., Abdar, M., Beykikhoshk, A., Shoeibi, A., Zare, A., Panahiazar, M., Nahavandi, S., Srinivasan, D., Atiya, A. F., & Acharya, U. R. (2020). *Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991-2020)*. http://ezproxy.unal.edu.co/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsarx&AN=edsarx.2008.10114&lang=es&site=eds-live

Alonso, J. C., & Munoz, A. F. (2014). Interpretación de variables Dummy en modelos loglin. *Apuntes de Economıa No. 40*.

Álvarez Ceballos, J. C., Álvarez Múñoz, A., Carvajal Gutiérrez, W., González, M. M., Duque, J. L., & Nieto Cárdenas, O. A. (2017). Determinación del riesgo cardiovascular en una población. *Revista Colombiana de Cardiologia*, *24*(4), 334–341. https://doi.org/10.1016/j.rccar.2016.08.002

Álvarez Cosmea, A. (2001). Las tablas de riesgo cardiovascular: Una revisión crítica. *MEDIFAM*, *11*(3), 122–139. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1131-57682001000300002

Amón Uribe, I. (2010). *Guía metodológica para la selección de técnicas de depuración de datos [Tesis de maestría]*. Universidad Nacional de Colombia.

Ashfaq, A., & Nowaczyk, S. (2019). Machine learning in healthcare - a system's perspective. In

A. V. B. Aditya Prakash Shweta Bansal, Adam Sadelik (Ed.), *Proceedings of the ACM SIGKDD Workshop on Epidemiology meets Data Mining and Knowledge Discovery (epiDAMIK)* (pp. 14–17). http://hh.diva-portal.org/smash/get/diva2:1342677/FULLTEXT01.pdf

Avila-Chaurand, R., Prado-León, L. R., & González-Muñoz, E. L. (2001). Dimensiones antropométricas de la población latinoamericana: México, Cuba, Colombia, Chile. *Centro Universitario de Arte, Arquitectura y Diseño, UDG. 1ª Ed. Guadalajara Jalisco*.

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P. E., & O'Connor, P. J. (2015). Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, *29*(4), 1033–1069. https://doi.org/10.1007/s10618-014-0386-6

Basu, T., Engel-Wolf, S., & Menzer, O. (2020). The ethics of machine learning in medical sciences: Where do we stand today? *Indian Journal of Dermatology*, *65*(5), 358–364. http://10.0.16.7/ijd.IJD_419_20

Beard, J., Officer, A., Cassels, A., Bustreo, F., Worning, A. M., & Asamoa-Baah, A. (2015). Informe mundial sobre el envejecimiento y la salud. *OMS*.

Bedoya-Mejía, S., Henao-Valencia, C., & Cardona-Arango, D. (2019). Mortalidad por enfermedades del sistema circulatorio, en los municipios del área metropolitana, Antioquia, 1998-2014. *Revista Facultad Nacional de Salud Pública*, *37*(1), 96–105.

Bellamy, D., Celi, L., & Beam, A. L. (2020). *Evaluating Progress on Machine Learning for Longitudinal Electronic Healthcare Data*. http://ezproxy.unal.edu.co/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsarx&AN=edsarx.2010.01149&lang=es&site=eds-live

Bengio, S., Weston, J., & Grangier, D. (2010). Label Embedding Trees for Large Multi-Class Tasks. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 23, pp. 163–171). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2010/file/06138bc5af6023646ede0e1f7c1eac75-Paper.pdf

Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kötter, T., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2009). KNIME - The Konstanz information miner : Version 2.0 and Beyond. *SIGKDD Explorations*, *11*, 26–31.

Bhatti, S., Kehar, V., & Memon, M. A. (2020). Prognosis of Diabetes by Performing Data Mining of HbA1c. *International Journal of Computer Science and Information Security (IJCSIS)*,

18(1).

Blagus, R., & Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, *14*, 106. https://doi.org/10.1186/1471-2105-14-106

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Cardona Alzate, N. I. (2019). *Predicción y selección de variables con bosques aleatorios en presencia de variables correlacionadas* [Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/bitstream/handle/unal/75561/8063120.2019.pdf?sequence=1

Castellanos Vázquez, J., Moreno, A. S., Herrera, C. B., & Sautto Vallejo, J. M. (2019). Valoración de riesgo cardiovascular mediante modelos de clasificación. *Investigación Operacional*, *40*(1), 80–87. http://ezproxy.unal.edu.co/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=133858008&lang=es&site=eds-live

Chitarroni, H. (2002). *La regresión logística*. http://www.salvador.edu.ar/csoc/idicso

Ley 1581 de 2012 - Ley Estatutaria de Hábeas Data, (2012). http://bibliotecadigital.ccb.org.co/bitstream/handle/11520/13629/Ley 1581 de 2012.pdf?sequence=1

Cosma, G., Acampora, G., Brown, D., Rees, R. C., Khan, M., & Pockley, A. G. (2016). Prediction of Pathological Stage in Patients with Prostate Cancer: A Neuro-Fuzzy Model. *PLOS ONE*, *11*(6), e0155856. https://doi.org/10.1371/journal.pone.0155856

Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*, *37*(2), 1568–1572. https://doi.org/10.1016/j.eswa.2009.06.040

Díaz-Realpe, J. E., Muñoz-Martínez, J., & Sierra-Torres, C. H. (2007). Factores de riesgo para enfermedad cardiovascular en trabajadores de una institución prestadora de servicios de salud, Colombia. *Revista de Salud Pública*, *9*, 64–75.

Estrada, J., Camacho, J. A., Restrepo, M. T., & Parra, C. M. (1998). Parámetros antropométricos de la población laboral colombiana, 1995. *Revista Facultad Nacional de Salud Pública*, *32*, 64–78.

Fernando, M., & Arrieta, C. (2005). Estudio sociológico y del conocimiento de los factores de riesgo de las enfermedades cardiovasculares en la Costa Caribe Colombiana (Estudio Caribe). *Revista Colombiana de Cardiología*, *12*, 122–128.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2005). Random Forests for land cover

classification. *Pattern Recognition Letters*, *27*, 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Gómez, L. A. (2011). Las enfermedades cardiovasculares: un problema de salud pública y un reto global. *Biomédica*, *31*(4).

Hameed, N., Shabut, A. M., Ghosh, M. K., & Hossain, M. A. (2020). Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. *Expert Systems with Applications*, *141*, 112961. https://doi.org/10.1016/j.eswa.2019.112961

Instituto Nacional de Salud de Colombia (INS). (2013). *Observatorio Nacional de Salud (ONS) Boletín N0.1*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/IA/INS/Boletin-tecnico-1-ONS.pdf

Jamthikar, A., Gupta, D., Khanna, N. N., Saba, L., Laird, J. R., & Suri, J. S. (2020). Cardiovascular/stroke risk prevention: A new machine learning framework integrating carotid ultrasound image-based phenotypes and its harmonics with conventional risk factors. *Indian Heart Journal*, *72*(4), 258–264. https://doi.org/https://doi.org/10.1016/j.ihj.2020.06.004

Kalantari, A., Kamsin, A., Shamshirbandb, S., Gani, A., Alinejad Rokny, H., & Chronopoulos, A. T. (2018). Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions. *Neurocomputing*, *276*.

Khan, S. U., Islam, N., Jan, Z., Din, I. U., Khan, A., & Faheem, Y. (2019). An e-Health care services framework for the detection and classification of breast cancer in breast cytology images as an IoMT application. *Future Generation Computer Systems*, *98*, 286–296. https://doi.org/10.1016/j.future.2019.01.033

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, *18*(1), 559–563.

Llibre Guerra, J. C., Guerra Hernández, M. A., & Perera Miniet, E. (2008). Comportamiento de las enfermedades crónicas no transmisibles en adultos mayores. *Revista Cubana de Medicina General Integral*, *24*(4), 0.

Lobos Bejarano, J. M., & Brotons Cuixart, C. (2011). Factores de riesgo cardiovascular y atención primaria: evaluación e intervención. *Atención Primaria*, *43*(12), 668–677. https://doi.org/https://doi.org/10.1016/j.aprim.2011.10.002

Lopera, M. M. (2017). Revisión comentada de la legislación colombiana en ética de la investigación en salud. *Biomedica*, *37*(4), 1–44. https://doi.org/10.7705/biomedica.v37i4.3333

López-Martínez, F., Núñez-Valdez, E. R., García-Díaz, V., & Bursac, Z. (2020). A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management. *Algorithms*, *13*(4), 102. http://10.0.13.62/a13040102

Loreto, M., Lisboa, T., & Moreira, V. P. (2020). Early prediction of ICU readmissions using classification algorithms. *Computers in Biology and Medicine*, *118*, 103636. https://doi.org/10.1016/j.compbiomed.2020.103636

Luengo Pérez, L. M., Urbano Gálvez, J. M., & Pérez Miranda, M. (2009). Validación de índices antropométricos alternativos como marcadores del riesgo cardiovascular. *Endocrinologia y Nutricion*, *56*(9), 439–446. https://doi.org/10.1016/S1575-0922(09)72964-X

Ma, H., Wang, L., & Shen, B. (2011). A new fuzzy support vector machines for class imbalance learning. *2011 International Conference on Electrical and Control Engineering*, 3781–3784. https://doi.org/10.1109/ICECENG.2011.6056838

Martínez, E. A. B., Ramírez, A. F., & Villamil, E. S. (2016). Modelos predictivos de riesgo cardiovascular. *Revista Cuarzo*, *22*(2), 80–91.

Melillo, P., Izzo, R., Orrico, A., Scala, P., Attanasio, M., Mirra, M., De Luca, N., & Pecchia, L. (2015). Automatic Prediction of Cardiovascular and Cerebrovascular Events Using Heart Rate Variability Analysis. *PLoS ONE*, *10*(3), 1–14. http://10.0.5.91/journal.pone.0118504

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Vol. 2006). https://doi.org/10.1145/1150402.1150531

Resolucion 8430 de 1993, (1993).

Ministerio de Salud y Protección Social de Colombia. (2013). *Envejecimiento demográfico. Colombia 1951-2020 dinámica demográfica y estructuras poblacionales*.

Moine, J. M. (2013). *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo*.

Monroy-de-Jesús, J., Guadalupe-Ramírez, A., Ambriz-Polo, J., & López-González, E. (2018). Algoritmo de aprendizaje eficiente para tratar el problema del desbalance de múltiples clases. *Research in Computing Science*, *147*, 143–157. https://doi.org/10.13053/rcs-147-

5-11

Muñoz, O. M., García, Á. A., Fernández-Ávila, D., Higuera, A., Ruiz, Á. J., Aschner, P., Toro, J. M., Arteaga, J. M., Merchán, A., Sánchez, G., & Villalba, Y. (2015). Guía de práctica clínica para la prevención, detección temprana, diagnóstico, tratamiento y seguimiento de las dislipidemias: evaluación del riesgo cardiovascular. *Revista Colombiana de Cardiología*, *22*(6), 263–269. https://doi.org/https://doi.org/10.1016/j.rccar.2015.04.009

Muñoz, O. M., Rodríguez, N. I., Ruiz, Á., & Rondón, M. (2014). Validación de los modelos de predicción de Framingham y PROCAM como estimadores del riesgo cardiovascular en una población colombiana. *Revista Colombiana de Cardiología*, *21*(4), 202–212. https://doi.org/https://doi.org/10.1016/j.rccar.2014.02.001

Muñoz V, O. M., Ruiz Morales, Á. J., Mariño Correa, A., & Bustos C., M. M. (2017). Concordancia entre los modelos de SCORE y Framingham y las ecuaciones AHA/ACC como evaluadores de riesgo cardiovascular. *Revista Colombiana de Cardiología*, *24*(2), 110–116. https://doi.org/https://doi.org/10.1016/j.rccar.2016.06.013

Navarro Céspedes, J. M. (2008). *Análisis de Componentes Principales y Análisis de Regresión para Datos Categóricos. Aplicación en HTA.* . Universidad Central "Marta Abreu" de Las Villas.

O'Donnell, C. J., & Elosua, R. (2008). Cardiovascular risk factors. Insights from framingham heart study. *Revista Espanola de Cardiologia*, *61*(3), 299–310. https://doi.org/10.1157/13116658

OMS. (2008). *Prevención de las enfermedades cardiovasculares: guía de bolsillo para la estimación y el manejo del riesgo cardiovascular*. Organización Mundial de la Salud.

OMS. (2014). *Global status report on noncommunicable diseases 2014* (Issue WHO/NMH/NVI/15.1). World Health Organization.

Otero, J., & Sánchez, L. (2007). Disenos experimentales y tests estadısticos, tendencias actuales en Machine Learning. *V Congreso Espanol Sobre Metaheurısticas, Algoritmos Evolutivos y Bioinspirados (MAEB'07). Universidad de La Laguna. Puerto de La Cruz (Espana, 2007)*, 295–302.

Padmanabhan, M., Yuan, P., Chada, G., & Nguyen, H. Van. (2019). Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction. *Journal of Clinical Medicine*, *8*(7), 1050. https://doi.org/10.3390/jcm8071050

Patiño-Villada, F. A., Arango-Vélez, E. F., Quintero-Velásquez, M. A., & Arenas-Sosa, M. M. (2011). Cardiovascular risk factors in an urban Colombia population. *Revista de Salud Pública*, *13*(3), 433–445.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*.

Plazzotta, F., Luna, D., & González Bernaldo de Quirós, F. (2015). Sistemas de Información en Salud: Integrando datos clínicos en diferentes escenarios y usuarios . In *Revista Peruana de Medicina Experimental y Salud Publica* (Vol. 32, pp. 343–351). scielo .

Prasad, N. R., Almanza-Garcia, S., & Lu, T. T. (2009). Anomaly detection. *Computers, Materials and Continua*, *14*(1), 1–22. https://doi.org/10.1145/1541880.1541882

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org/

Ramírez-Vélez, R., Agredo, R. A., Jerez, A. M., & Chapal, L. Y. (2008). Calidad de vida y condiciones de salud en adultos mayores no institucionalizados en Cali, Colombia. *Revista de Salud Pública*, *10*, 529–536.

Restrepo, L. F., & González, J. (2007). From Pearson to Spearman. *Revista Colombiana de Ciencias Pecuarias*, *20*(2). http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-06902007000200010

Sangra, R. A., & Codina, A. F. (2015). Identificación, impacto y tratamiento de datos perdidos y atípicos en epidemiología nutricional. *Rev Esp Nutr Comunitaria*, *21*(Supl 1), 188–194.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/https://doi.org/10.1016/j.ipm.2009.03.002

Suzuki, S., Yamashita, T., Sakama, T., Arita, T., Yagi, N., Otsuka, T., Semba, H., Kano, H., Matsuno, S., Kato, Y., Uejima, T., Oikawa, Y., Matsuhama, M., & Yajima, J. (2019). Comparison of risk models for mortality and cardiovascular events between machine learning and conventional logistic regression analysis. *PLOS ONE*, *14*(9), e0221911. https://doi.org/10.1371/journal.pone.0221911

Tharwat, A. (2020). Classification assessment methods. In *Applied Computing and Informatics: Vol. ahead-of-p* (Issue ahead-of-print). https://doi.org/10.1016/j.aci.2018.08.003

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*. https://doi.org/10.1016/j.enbuild.2017.11.039

van Rossum, G. (1995). *Python tutorial* (Issue R 9526). CWI.

Wang, H., Li, Y., Ning, H., Wilkins, J., Lloyd-Jones, D., & Luo, Y. (2019). Using Machine Learning to Integrate Socio-Behavioral Factors in Predicting Cardiovascular-Related Mortality Risk. *Studies in Health Technology and Informatics*, *264*, 433–437. https://doi.org/10.3233/SHTI190258

Ward, A., Sarraju, A., Chung, S., Li, J., Harrington, R., Heidenreich, P., Palaniappan, L., Scheinker, D., & Rodriguez, F. (2020). Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. *Npj Digital Medicine*, *3*(1), 125. https://doi.org/10.1038/s41746-020-00331-1

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS One*, *12*(4), e0174944–e0174944. https://doi.org/10.1371/journal.pone.0174944

Wiens, J., & Shenoy, E. S. (2018). Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. *Clinical Infectious Diseases*, *66*(1), 149–153. https://doi.org/10.1093/cid/cix731

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–39.

Yu, L., & Liu, H. (2003). *Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*.