

Factores sociodemográficos, clínicos y conductuales presentes que caracterizan el riesgo cardiovascular en una muestra de pacientes de la ESE Oriente de Cali entre 2016 a 2023

Jovany Cely Ospina
Víctor Alfonso Peña Ocampo

Trabajo de grado II

Director
Nelson Andrés Andrade Bonilla

Universidad Icesi
Facultad de ingeniería y ciencias aplicadas
Maestría en Ciencia de datos
Santiago de Cali
2024

Resumen	1
Introducción	1
1. Contexto y antecedentes	2
2. Planteamiento del problema	3
3. Alcance del proyecto	3
4. Objetivos	4
4.1. Objetivo general	4
4.2. Objetivos específicos	4
5. Metodología	4
6. Marco Teórico	5
7. Estado del arte	11
8. Contribución a la ciencia de datos	15
8.1. Selección y estudio de viabilidad del conjunto de datos	15
8.2. Análisis exploratorio de datos	17
8.3. Identificación del Tipo de Problema y Paradigma de IA Adecuado	24
8.4. Proceso de selección y descarte de características. Elementos de preprocesamiento	24
8.5. Algoritmos de entrenamiento	27
8.6. Generación de modelos	28
Modelo 1: Conjunto Completo de Variables	28
Modelo 2: Supresión de Variables Relevantes	29
Modelo 3: Selección de Variables Framingham	29
8.7. Resultados	30
Modelo 1: Evaluación de Modelos con Datos Antiguos y Recientes	30
Modelo 2: Comparación con la Supresión de Variables Relevantes	31
Modelo 3: Evaluación de Modelos en Variables Seleccionadas de Framingham	31
9. Conclusiones	35
10. Recomendaciones y futuras investigaciones	36
Anexos	37
Anexo 1. Campos generados	37
Anexo 2. Porcentaje de nulidad	39
Anexo 3. Nombre de variables seleccionadas de cada conjunto de datos	41
Anexo 4. Nombre de variables seleccionadas para el modelo dos en ambos conjuntos de datos	42
Referencias	43

Resumen

La enfermedad cardiovascular es la principal causa de morbimortalidad a nivel mundial, y su desarrollo está relacionado con diversos factores de riesgo. Por ello, la prevención depende de identificar y controlar estos factores para influir en el curso clínico de la enfermedad.

En este proyecto se propuso la creación de modelos de predicción del riesgo cardiovascular a partir de un dataset de pacientes que son atendidos en una empresa social del estado (ESE) de Cali. El conjunto de datos se dividió en dos subconjuntos, según la fecha de toma de la data, y se realizó una selección de variables en cada uno con el objetivo de analizar cómo esta elección afecta el desempeño de los diferentes modelos.

Finalmente, con base en métricas de desempeño, se definió el modelo de random forest como el mejor modelo, identificando las variables clave que influyen en la variable objetivo y las variables sociodemográficas de mayor peso con el fin de que estas puedan ser consideradas por las entidades de salud en los procesos de prevención.

Introducción

El riesgo cardiovascular es la probabilidad de padecer o desarrollar enfermedades del corazón y los vasos sanguíneos en un plazo determinado, usualmente entre 5 y 10 años. Es un riesgo que cada vez afecta a un mayor número de personas causando incapacidades permanentes y muerte prematura.

Es de interés de salud pública conocer las causas y los factores que promueven y/o aceleran el padecimiento de este tipo de enfermedades. Para medirlo y controlarlo se han desarrollado diferentes modelos de pronóstico basado en factores clínicos y conductuales presentes en poblaciones determinadas, sin embargo, poder determinar aquellas variables más relevantes para focalizar los esfuerzos de prevención y control es importante.

En este sentido, para dar respuesta a lo anterior se buscó desarrollar un modelo de predicción cardiovascular que identifique y prediga el riesgo en una población de la ciudad de Cali. Para lo cual, en primera instancia se organizó la información de control y seguimiento que se ha venido evaluando desde la entidad utilizando herramientas de análisis exploratorio de datos, luego se realizó el proceso de modelamiento con base en la selección de variables y se evaluaron los modelos. Con las variables de desempeño obtenidas, se llevó a cabo la selección del mejor modelo predictivo y se discutió acerca de los resultados obtenidos.

1. Contexto y antecedentes

Las enfermedades cardiovasculares (ECV) incluyen diversas afecciones del corazón y los vasos sanguíneos, como enfermedad coronaria, enfermedad cerebrovascular, enfermedad arterial periférica, enfermedad cardíaca reumática, enfermedad cardíaca congénita, trombosis venosa profunda y embolia pulmonar y para el año 2019, fueron la principal causa de muerte a nivel mundial con aproximadamente 17.9 millones de personas, lo que representó el 32% de todas las muertes a nivel global (Organización Mundial de la Salud, 2021).

Sin embargo, en los países de ingresos bajos y medios, las ECV representan un doble desafío debido a la falta de acceso a servicios de salud de manera efectiva y equitativa. Tal como se puede observar a nivel regional, según la Organización Panamericana de la Salud (OPS), la cardiopatía isquémica y el accidente cerebrovascular son las principales causas de muerte y discapacidad, especialmente, en Haití, Guyana y Honduras; asimismo, las ECV causaron 40.8 millones de años ajustados por discapacidad (AVAD)¹ 36.4 millones de años de vida perdidos por muerte prematura² (YLLs)³, (89% del total de AVAD por ECV) y 4.5 millones de años vividos con discapacidad (AVD)⁴ (Organización Panamericana de la Salud, 2021)

A nivel nacional se encuentra que de acuerdo con cifras preliminares, estas patologías representaron la primera causa de mortalidad en 2022 (175,73 por 100.000 habitantes), principalmente a expensas de la enfermedad isquémica coronaria (96,57 por 100.000 habitantes), la enfermedad cerebro vascular (33,53 por 100.000 habitantes) y las enfermedades hipertensivas (21 por 100.000 habitantes). Además, se presentaron, aproximadamente 24.395 defunciones en personas entre 30 a 70 años y los territorios que reportaron mayor número de muertes fueron: Tolima, Norte de Santander,

¹ Los años de vida ajustados por discapacidad (AVAD) son un indicador sintético de salud utilizado para medir la carga de enfermedad a nivel poblacional, que proporciona información conjunta de las consecuencias mortales y no mortales de las enfermedades, lesiones y factores de riesgo (Fernández de Larrea-Baz, Morant-Ginestar, Catalá-López, & Gênova-Maleras, 2015, pág. 969)

² Aquella que ocurre antes de alcanzar la esperanza máxima de vida potencial observada a la edad de la persona que falleció (Martinez, Soliz, Caixeta, & Ordunez, 2019, pág. 2)

³ Es una métrica utilizada en epidemiología y salud pública para cuantificar la carga de enfermedad y mortalidad prematura en una población, a través de calcular la cantidad de años que una persona habría vivido si no hubiera fallecido prematuramente debido a una causa específica. Este cálculo se base en la diferencia entre la edad al morir y la esperanza de vida estándar a esa edad (Martinez, Soliz, Caixeta, & Ordunez, 2019)

⁴ Es una medida utilizada en epidemiología y salud pública para cuantificar la carga de enfermedad relacionada con la discapacidad en una población. Se calcula a través de la suma de los años de vida perdidos por muerte prematura y los años vividos con discapacidad (Organización Panamericana de la Salud, s.f.)

Risaralda, Caldas, Guaviare, Archipiélago de San Andrés, Valle del Cauca, Atlántico, Huila y Quindío (Ministerio de Salud y Protección Social, 2023).

A nivel distrital, durante el período de 2010 a 2022, se observa que las ECV representaron la principal causa de consultas médicas tanto en hombres como en mujeres (Alcaldía de Santiago de Cali, 2024). Sin embargo, este aumento, ha sido sectorizado, ya que el análisis espacial del riesgo cardiovascular llevado a cabo por Pico, Hernández & Muñoz encontraron que el riesgo cardiovascular alto según la residencia de los individuos se concentraba en las comunas 2, 6, 7, 13, 14, 16 y 19, mientras que los casos de riesgo moderado se distribuía en las comunas 4, 8, 9, 10, 12, 15 y 21 (2022, pág. 137)

Ante este panorama y teniendo en cuenta que el desarrollo de estas enfermedades se encuentra influenciado, principalmente, por el tabaquismo, la dieta no saludable, la obesidad, la inactividad física y el consumo nocivo de alcohol, y que además puede conllevar a presentar manifestaciones como presión arterial, glucosa y lípidos en sangre elevados, hace importante la prevención de su aparición, por medio de la identificación y control de estos factores de riesgo (Donado Gómez J. H., 2017). De esta manera se busca impactar en el curso clínico o en la historia natural de la enfermedad de acuerdo con los síntomas, hábitos y diagnósticos que se han presentado en un individuo específico.

2. Planteamiento del problema

Con base en lo anterior, el presente trabajo pretende responder la siguiente pregunta de investigación:

¿Cuáles son los factores sociodemográficos, clínicos y conductuales que caracterizan el riesgo cardiovascular de pacientes que asisten a la ESE Oriente de Cali entre 2016 a 2023?

3. Alcance del proyecto

El alcance de este proyecto se centra en el desarrollo de un modelo predictivo del riesgo cardiovascular, utilizando como variables predictoras un conjunto de condiciones sociodemográficas, clínicas y conductuales en una población de la ESE Oriente, durante el período 2016-2023.

Este trabajo constituye la continuación del proyecto realizado en el semestre anterior, en el cual se llevó a cabo un análisis exploratorio de datos, con el objetivo de caracterizar y consolidar la información relevante para la presente investigación.

4. Objetivos

4.1. Objetivo general

- Desarrollar un modelo predictivo del riesgo cardiovascular alto en pacientes de la ESE Oriente de Cali, tomando como base características sociodemográficas, clínicas y conductuales.

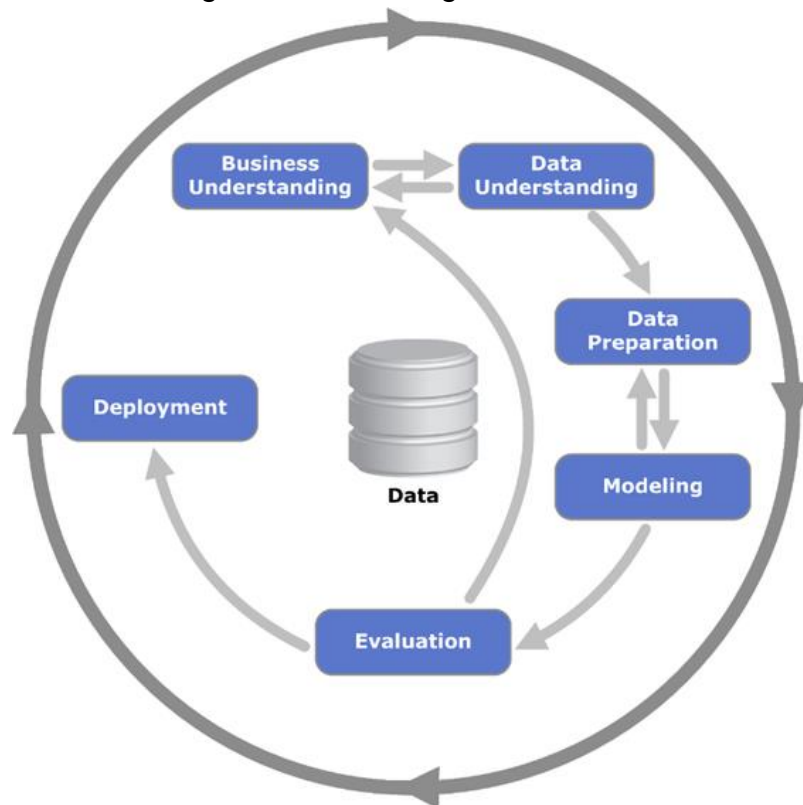
4.2. Objetivos específicos

- Realizar un análisis exploratorio de los datos de los pacientes de la ESE Oriente de Cali para identificar y seleccionar las variables sociodemográficas, clínicas y conductuales relevantes del estudio y caracterizar la muestra de individuos.
- Desarrollar y entrenar modelos predictivos utilizando algoritmos de aprendizaje automático para estimar el riesgo cardiovascular alto en los pacientes de la ESE Oriente.
- Identificar las variables clave que influyen en la predicción del riesgo cardiovascular alto, con el fin de proporcionar información útil para la toma de decisiones en procesos de prevención y atención de la salud.

5. Metodología

El siguiente proyecto de grado utilizará la metodología CRISP-DM (Ver figura 1) para dar respuesta a la pregunta y lograr los objetivos del proyecto, debido a que proporciona un marco estructurado para identificar, analizar y modelar los factores asociados al riesgo cardiovascular en pacientes de la ESE Oriente de Cali. A través de este proceso, es posible identificar variables explicativas del riesgo, diferentes a las utilizadas en la escala de medición Framingham (escala de medición usada por la entidad para la medición del riesgo).

Figura 1. Metodología CRISP-DM



Fuente. <https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/>

6. Marco Teórico

Entender el riesgo cardiovascular (RCV) implica conocer, previamente, los factores de riesgo (FR) que pueden predecir el desarrollo de eventos mórbidos futuros, es decir, cuáles son las características biológicas, hábitos o estilos de vida que aumentan la probabilidad de padecer o de morir a causa de una enfermedad (Lobos & Brotons, 2011), en este caso puntual padecer o desarrollar una enfermedad cardiovascular. Tal como afirma Lira (2022, pág. 535), el manejo integral del riesgo cardiovascular se fundamenta en la comprensión de que los factores de riesgo clásicos, como la edad, el sexo, los lípidos anormales, el tabaquismo, la hipertensión y la diabetes, interactúan de manera sinérgica, produciendo un efecto sumatorio que aumenta la probabilidad de desarrollar enfermedades cardiovasculares.

No obstante, es importante destacar que la incidencia de enfermedad coronaria y la carga de riesgo atribuible a cada factor varían significativamente entre las poblaciones, lo que requiere un enfoque personalizado y adaptado a las características específicas de cada grupo poblacional. Además, que a través del tiempo los factores de riesgo han ido adaptándose a las poblaciones obteniendo una clasificación de estos (Ver tabla 1).

Tabla 1. Tipo de factores de riesgo de enfermedad cardiovascular

Factores no modificables	Factores modificables	Otros factores (psicosociales)
<ul style="list-style-type: none"> • Edad • Sexo • Factores genéticos • Historia familiar 	<ul style="list-style-type: none"> • Hipertensión arterial (HTA) • Tabaquismo • Hipercolesterolemia • Diabetes mellitus (DM) • Sobrepeso/obesidad 	<ul style="list-style-type: none"> • Bajo nivel socioeconómico • Aislamiento social • Depresión • Estrés laboral o familiar

Fuente. Elaboración propia con base en (Lobos & Brotons, 2011)

Teniendo en cuenta la complejidad del riesgo cardiovascular, según Roselo et al (2019), el riesgo cardiovascular (RCV) absoluto, también conocido como riesgo total o global, se define como la probabilidad de desarrollar enfermedades del corazón y los vasos sanguíneos en un plazo determinado, usualmente entre 5 y 10 años. Esta probabilidad abarca una amplia gama de condiciones, incluyendo la enfermedad coronaria, la enfermedad cerebrovascular, la insuficiencia cardíaca y la enfermedad arterial periférica, entre otras. Para evaluar el impacto de cada factor de riesgo en esta probabilidad, se utiliza una ecuación aritmética que asigna un peso específico a cada factor, lo que permite calcular un resultado global de riesgo, según Lira (2022, pág. 535), este resultado se clasifica generalmente en tres categorías: bajo, intermedio o alto, lo que permite a los profesionales de la salud tomar decisiones informadas sobre la prevención y el tratamiento de las enfermedades cardiovasculares.

Con base en lo anterior, la pregunta a responder es ¿cómo se mide el riesgo cardiovascular? Para responder lo anterior, Donado Gómez, Higueta-Duque, & Castro-Palacio (2017) a través de la revisión sistemática de información en revistas alojadas en PubMed encuentran que hay 11 escalas de evaluación del pronóstico del riesgo cardiovascular (Ver tabla 2), las cuales tienen sus propias limitaciones y debilidades, y que la elección del algoritmo adecuado dependerá de la población y el contexto clínico en el que se va a utilizar.

Tabla 2. Elementos característicos de los modelos de pronóstico de enfermedad cardiovascular

#	Escala de riesgo	Población estudiada	Variables que evalúan	Desenlace que se evalúa a 10 años
1	FRAMINGHAM SCORE 1998	Descendientes de caucásicos de origen europeo	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Diabetes ● Lípidos ● Tratamiento anti-HTA ● Presión arterial ● Historia familiar prematura de enfermedad coronaria cardíaca (CHD) o enfermedad cardiovascular (ECV) 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Infarto Agudo de Miocardio (IAM) fatal o no fatal ● Enfermedad Cerebro Vascular (ECV) o Accidente Isquémico Transitorio (TIA) ● Angina estable o inestable
2	PROCAM	Población alemana	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Diabetes ● Lípidos ● Tratamiento anti-HTA ● Presión arterial ● Historia familiar a cualquier edad 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Infarto Agudo de Miocardio (IAM) fatal o no fatal ● Angina estable o inestable
3	MESA SCORE	Población de Estados Unidos	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Diabetes ● Raza/Etnicidad ● Lípidos ● Tratamiento anti-HTA ● Presión arterial ● Historia familiar a cualquier edad 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Infarto Agudo de Miocardio (IAM) fatal o no fatal ● Revascularización
4	POOLED COHORT	Múltiples grupos raciales de Estados Unidos	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Raza/Etnicidad ● Lípidos 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Infarto Agudo de Miocardio (IAM) fatal o no fatal

#	Escala de riesgo	Población estudiada	Variables que evalúan	Desenlace que se evalúa a 10 años
			<ul style="list-style-type: none"> ● Tratamiento anti-HTA ● Presión arterial 	<ul style="list-style-type: none"> ● Enfermedad Cerebro Vascular (ECV) o Accidente Isquémico Transitorio (TIA)
5	SCORE	Poblaciones de nueve países europeos	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Lípidos ● Presión arterial 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Enfermedad Cerebro Vascular (ECV) o Accidente Isquémico Transitorio (TIA)
6	REYNOLDS SCORE	Mujeres de Estados Unidos	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Diabetes ● Lípidos ● Presión arterial ● Historia familiar prematura de enfermedad coronaria cardíaca (CHD) o enfermedad cardiovascular (ECV) ● Proteína C reactiva (PCR) 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Infarto Agudo de Miocardio (IAM) fatal o no fatal ● Enfermedad Cerebro Vascular (ECV) o Accidente Isquémico Transitorio (TIA) ● Revascularización
7	QRISK2	Grupos poblaciones de Gales e Inglaterra	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Diabetes ● Raza/Etnicidad ● Lípidos ● Tratamiento anti-HTA ● Presión arterial ● Historia familiar prematura de enfermedad coronaria cardíaca (CHD) o enfermedad cardiovascular (ECV) ● Enfermedad renal crónica (ERC) ● Fibrilación auricular (FA) 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Enfermedad Cerebro Vascular (ECV) o Accidente Isquémico Transitorio (TIA)

#	Escala de riesgo	Población estudiada	Variables que evalúan	Desenlace que se evalúa a 10 años
			<ul style="list-style-type: none"> ● Artritis reumatoide (AR) ● Índice de masa corporal (IMC) 	
8	QRISK	Grupos poblaciones de Gales e Inglaterra	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Diabetes ● Lípidos ● Tratamiento anti-HTA ● Presión arterial ● Índice de masa corporal (IMC) 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Infarto Agudo de Miocardio (IAM) fatal o no fatal ● Enfermedad Cerebro Vascular (ECV) o Accidente Isquémico Transitorio (TIA) ● Angina estable o inestable ● Enfermedad vascular periférica ● Revascularización
9	INTERHEART RISK SCORE	Poblaciones de 52 países, mayoría europeos	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Diabetes ● Tratamiento anti-HTA ● Presión arterial ● Historia familiar a cualquier edad ● Estrés y/o depresión ● Actividad física ● Dieta ● Índice cintura/cadera 	<ul style="list-style-type: none"> ● Infarto Agudo de Miocardio (IAM) fatal o no fatal
10	JBS3	Población del Reino Unido	<ul style="list-style-type: none"> ● Edad ● Sexo ● Tabaquismo ● Diabetes ● Raza/Etnicidad ● Estrato socioeconómico ● Lípidos ● Tratamiento anti-HTA ● Presión arterial 	<ul style="list-style-type: none"> ● Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal ● Infarto Agudo de Miocardio (IAM) fatal o no fatal ● Enfermedad Cerebro Vascular (ECV) o Accidente Isquémico Transitorio (TIA) ● Angina estable o inestable ● Enfermedad vascular periférica

#	Escala de riesgo	Población estudiada	Variables que evalúan	Desenlace que se evalúa a 10 años
			<ul style="list-style-type: none"> • Historia familiar prematura de enfermedad coronaria cardíaca (CHD) o enfermedad cardiovascular (ECV) • Enfermedad renal crónica (ERC) • Fibrilación auricular (FA) • Artritis reumatoide (AR) • Índice de masa corporal (IMC) 	<ul style="list-style-type: none"> • Revascularización
11	ASSIGN	Población de Escocia	<ul style="list-style-type: none"> • Edad • Sexo • Tabaquismo • Diabetes • Estrato socioeconómico • Lípidos • Presión arterial • Historia familiar prematura de enfermedad coronaria cardíaca (CHD) o enfermedad cardiovascular (ECV) • Artritis reumatoide (AR) 	<ul style="list-style-type: none"> • Enfermedad Cardiovascular Aterosclerótica (ECA) fatal o no fatal • Revascularización

Fuente. Elaboración propia con base en (Donado Gómez J. H., 2017, págs. 155-156)

7. Estado del arte

La predicción del riesgo cardiovascular es un tema de gran interés en la medicina actual, ya que las enfermedades cardiovasculares son la principal causa de muerte a nivel mundial. En este sentido, el contar con datos confiables, permite lograr usar de manera eficiente técnicas de aprendizaje automático (Machine Learning, ML) se ha reconocido como una herramienta poderosa para mejorar la atención médica y tomar decisiones basadas en datos. A continuación, se enlistan alguno de los trabajos que han abordado esta temática.

El artículo realizado por Daza, Castañeda, & Castaño (2022) en el que presentan el desarrollo de una plataforma de analítica de datos para caracterizar la población y evaluar el riesgo cardiovascular en pacientes de la región centro-occidente de Colombia, atendidos por la Unidad Vasculár Cardiológica y Neurológica. La plataforma integra datos sociodemográficos, clínicos, hábitos de vida y resultados de laboratorio de los pacientes, permitiendo un análisis integral y automatizado del riesgo cardiovascular.

El desarrollo de la plataforma se llevó a cabo bajo el modelo de desarrollo evolutivo exploratorio con modelado de prototipo. Para lo cual se realizó un levantamiento de requerimientos mediante mesas de trabajo con personal clínico, administrativo y desarrolladores. La arquitectura sigue un modelo vista-controlador con Flask de Python como servidor principal, MongoDB como base de datos NoSQL y librerías gráficas interactivas. Lo anterior, dio como resultado una plataforma con dos módulos principales.

Uno de formularios que permite recolectar información de los pacientes no registrada en el sistema de historia clínica, como resultados de laboratorio, hábitos de vida, entre otros. Y el segundo de reportes que ofrece estadísticas e informes interactivos sobre datos demográficos, clínicos, atenciones, laboratorios y perfil del paciente, incluyendo la evaluación del riesgo cardiovascular. En este sentido, la plataforma permite articular información clínica y demográfica para el estudio y seguimiento de patologías, centralizando y depurando datos previamente disgregados. Además, brinda la posibilidad de diseñar nuevos módulos de caracterización y análisis, inspirando protocolos de intervención y planes administrativos. Como conclusión, los autores sostienen que a través de la plataforma de analítica de datos se crearon estrategias de asignación de citas y control de patologías, mediante la caracterización de asignación de perfiles de riesgo poblacional (2022, pág. 11).

Igualmente, el trabajo realizado por Sarraju A, Ward A, Chung S, et al (2021) en el que se pretende identificar a los pacientes de alto riesgo para llevar a cabo estrategias de prevención eficaz de las ECV. Sin embargo, sostienen que no hay la certeza acerca de

la aplicación de modelos de aprendizaje automático (ML) basados en historias clínicas electrónicas (EHR) puedan mejorar la estratificación del riesgo de ECV en comparación con una puntuación de riesgo de prevención secundaria desarrollada a partir de ensayos clínicos aleatorios (Thrombolysis in Myocardial Infarction Risk Score for Secondary Prevention, TRS 2°P)⁵.

Desde esta perspectiva, los autores identificaron pacientes con ECV, en el que se incluía la enfermedad cardiovascular aterosclerótica (ASCVD) para lo cual se tenía en cuenta los siguientes criterios mayores de 18 años que recibieron atención en el norte de California entre el 1 de enero de 2009 y el 31 de diciembre de 2018. Asimismo, debían tener al menos dos visitas ambulatorias que fueran con al menos 1 año de diferencia. En este sentido, serían excluidos aquellos pacientes que tuvieran menos de 5 años de seguimiento total y no tenían un evento de resultado, dando como cohorte de estudio un total de 32.192 pacientes (Sarraj, y otros, 2021, pág. 2). Posterior a ello, se procedía a extraer características consignadas en las EHR para entrenar modelos de ML⁶ y evaluarlos a través del área bajo la curva ROC (AUC).

Con base en lo anterior, los resultados obtenidos fue que el mejor modelo para realizar la predicción fue el XG Boost demostró un AUC de 0,70 (IC 95% 0,68 a 0,71) en la cohorte completa de ECV y un AUC de 0,71 (IC 95% 0,69 a 0,73) para la cohorte de ASCVD. En comparación, el TRS 2°P su desempeño no fue adecuado ECV (AUC 0,51, IC 95% 0,50 a 0,53) y ASCVD (AUC 0,50, IC 95% 0,48 a 0,52), ante estos resultados los autores afirman que los modelos de ML entrenados en datos de EHR predijeron el riesgo de eventos de ECV a 5 años y superaron significativamente la puntuación TRS 2°P (2021, pág. 4). Además, los modelos ML identificaron variables predictivas no tradicionales como lo fue el nivel educativo y las visitas de atención.

Sin embargo, la limitación encontrada por ellos fue que debido a la composición de la cohorte los resultados no pueden llegar a ser generalizables a todo el territorio estadounidense. Sino únicamente con la población de estudio, por lo que se recomienda realizar una validación externa con otra población para evaluar la generalización del modelo.

Asimismo, el trabajo de Cho, S., Kim, S., Kang, S. et al (2021) tiene como objetivo evaluar la calibración y discriminación de algoritmos de riesgo de ECV preexistentes, como la

⁵ Es un sistema de puntuación utilizado para evaluar el riesgo de eventos cardiovasculares adversos en pacientes que han sufrido un infarto de miocardio. Este puntaje se basa en varios factores de riesgo modificables, como la edad, el sexo, el tabaquismo, la presión arterial, el colesterol y la diabetes, entre otros. Es una herramienta útil para estratificar el riesgo a largo plazo y guiar las decisiones de tratamiento preventivo secundario en pacientes post-infarto (Puymirat, y otros, 2019)

⁶ Los modelos que se entrenaron fueron random forest (RF), gradient boosted machines (GBM), extreme gradient boosted models (XG Boost) y regresión logística con penalización L1 y L2.

puntuación de riesgo de Framingham (FRS), la ecuación de cohorte agrupada (PCE), la evaluación sistemática del riesgo coronario (SCORE) y QRISK3. Sin embargo, estas herramientas aún tienen margen de mejora en cuanto a su precisión. La literatura científica ha demostrado que el área bajo la curva (AUC) de estas calculadoras varía entre 0,65 y 0,85, por lo que evalúan modelos de predicción de riesgos basados en ML para comparar el rendimiento con los algoritmos preexistentes.

Este proceso de evaluación implicó la subdivisión de tres cohortes poblacionales según el tipo de algoritmo tradicional, la cohorte de análisis principal, se incluyeron 222.998 individuos sin antecedentes de ECV aterosclerótica con una edad media de 58 años de los cuales el 58,1% eran hombres. Además, de la cohorte el 5,5% tenía diabetes mellitus y el 21,1% recibía tratamiento antihipertensivo. Durante el seguimiento de 5 años, 7819 sujetos experimentaron eventos de ECV aterosclerótico (tasa de eventos: 3,51%). Por su parte, en Las cohortes FRS, SCORE y QRISK3 tenían diferentes cantidades de individuos y perfiles de riesgo similares, con distinciones como la ausencia de ciertas condiciones en cada cohorte. Los criterios de valoración y las tasas de eventos a 5 años variaron entre cada una de las cohortes, en un rango que iba desde 0,30% hasta 3,51% (Cho, y otros, 2021, pág. 2).

En relación a los modelos aplicados por los autores, se evaluaron la regresión logística, TreeBag, Random Forest, AdaBoost y redes neuronales. De estos, se destacó el modelo de redes neuronales, que exhibió el estadístico C^7 más alto (0.751) y notablemente superior a la PCE, además, demostró una mejor alineación entre el riesgo predicho y los resultados observados en comparación con la PCE. Sin embargo, a pesar de resaltar las ventajas del empleo de algoritmos de ML, los autores identificaron ciertas limitaciones. En primer lugar, los algoritmos basados en ML suelen requerir una amplia gama de variables, algunas de las cuales no se registran de manera rutinaria en la práctica clínica.

En segundo lugar, aunque se observó una mejora, esta fue relativamente modesta en términos absolutos, ya que el incremento absoluto no sobrepasó el 1.3%. Asimismo, en cuanto a las limitaciones del estudio, señalan que el seguimiento de los datos se limitó a 5 años, mientras que la mayoría de los modelos de predicción de riesgos tienen como objetivo predecir resultados a 10 años. Igualmente, afirman que el estudio podría estar sujeto a sesgo de selección, ya que la población de estudio se seleccionó entre los participantes del programa de detección de salud general. Y por último, existe un riesgo

⁷ El estadístico C conocido como área bajo la curva (AUC, por sus siglas en inglés), es una medida de la capacidad de un modelo de clasificación para distinguir entre clases, especificando la probabilidad de que el modelo asigna una probabilidad más alta a una instancia positiva que a una instancia negativa. Lo que se traduce en el contexto clínico como la probabilidad de que un paciente seleccionado al azar que experimentó un evento tuviera una puntuación de riesgo más alta que un paciente que no experimentó el evento.

potencial de sesgo de clasificación errónea, ya que muchas covariables y resultados se definieron utilizando información de reclamaciones (2021, págs. 5-6)

Por otro lado, el trabajo de Andaur, Damen, Takada, Nijman, et al (2020) buscan evaluar la calidad de reporte, la conducta metodológica y el riesgo de sesgo de los estudios de modelos de predicción que aplicaron técnicas de ML para el desarrollo y/o validación del modelo. En gran medida, a que los estudios que abordan el desarrollo y/o validación de modelos de predicción diagnóstica y pronóstica son abundantes en la mayoría de los dominios clínicos. Sin embargo, las revisiones sistemáticas han demostrado que la calidad metodológica y de reporte de estos estudios es subóptima. Debido a la creciente disponibilidad de datos médicos complejos y de mayor tamaño, así como a la creciente aplicación de técnicas de Inteligencia Artificial (IA) o aprendizaje automático (ML).

Desde esta perspectiva el estudio realiza una búsqueda en PubMed⁸, para el período entre enero de 2018 a diciembre de 2019, que predigan resultados relacionados a los pacientes, utilicen cualquier diseño de estudio o fuente de datos y se encuentren disponibles en idioma inglés. Con esta delimitación, se pretende identificar aquellos estudios que desarrollaron y/o hayan validado modelos de predicción utilizando cualquier metodología de ML. Con base en ello, se pretende verificar si los estudios analizados se adhieren a las pautas de informes transparentes de un modelo de predicción multivariable para pronóstico o diagnóstico individual⁹ (TRIPOD por sus siglas en inglés), y si el riesgo de sesgo en dichos estudios ha sido evaluado mediante la Herramienta de Evaluación del Riesgo de Sesgo en Modelos de Predicción Clínica¹⁰ (PROBAST, por sus siglas en inglés)

Los resultados arrojados posterior a la revisión sistemática de los estudios encontró que los estudios cumplían la guía de informes TRIPOD así como la del riesgo de sesgo evaluado mediante PROBAST. Sin embargo, la discusión se centra en que las investigaciones sobre modelos predictivos basados en ML suelen carecer de información

⁸ Es un motor de búsqueda gratuito que proporciona acceso a la base de datos bibliográfica en el campo de la medicina y las ciencias de la vida, incluye artículos de investigación, revisiones sistemáticas, ensayos clínicos, guías de práctica clínica y otros tipos de literatura científica.

⁹ Es una iniciativa internacional diseñada para mejorar la transparencia y la calidad de la presentación de modelos de predicción multivariable utilizados en el pronóstico individual o diagnóstico, desarrollada para abordar la necesidad de estándares claros en la publicación de modelos de predicción clínica, teniendo en cuenta elementos como la selección de variables, el manejo de datos faltantes, la evaluación del desempeño del modelo y la validación externa, entre otros (Collins, Reitsma, Altman, & Moons, 2015).

¹⁰ Es una herramienta diseñada que consta de cuatro dominios principales para evaluar el riesgo de sesgo en modelos de predicción, desarrollada para abordar la necesidad de estándares claros en la evaluación crítica de modelos de predicción clínica. A continuación, se detallan los dominios que aborda la herramienta, el primero relacionado a la selección de los participantes; el segundo a la definición y medición de las variables predictoras; el tercero a la definición y medición del resultado y el cuarto al análisis y consideraciones adicionales (Wolff, y otros, 2019)

adecuada acerca del modelo final, sus estimaciones y rendimiento. Asimismo, es escasa la investigación en la que el modelo de predicción sea accesible tanto para pacientes como para profesionales de la salud. Esta falta de claridad hace que los estudios de modelos predictivos no se consideren interpretables, generando así un problema, principalmente, en el ámbito del diagnóstico y pronóstico médico (Andaur Navarro, y otros, 2020, pág. 5).

En resumen, los estudios revisados destacan el potencial de los modelos de ML para mejorar la precisión en la predicción del riesgo cardiovascular y la importancia de evaluar la calidad metodológica y de reporte de los estudios de modelos de predicción que utilizan técnicas de ML. Además, la creación de plataformas de analítica de datos puede ser fundamental para caracterizar la población y evaluar el riesgo cardiovascular, lo que puede contribuir a mejorar la calidad asistencial y eficiencia de los servicios.

El presente trabajo se enmarca dentro de la propuesta de Sarraju A., Ward A., Chung S. et al. (2021), ya que busca explicar el fenómeno del riesgo cardiovascular mediante la identificación de variables distintas a las utilizadas en las escalas de riesgo tradicionalmente establecidas. De esta manera, se pretende descubrir factores predictivos no convencionales que podrían ser clave para una comprensión más integral del riesgo cardiovascular. Al incorporar estas variables adicionales, el estudio ofrece una perspectiva novedosa que podría enriquecer las estrategias de prevención y tratamiento, permitiendo una intervención más personalizada y efectiva para los pacientes.

8. Contribución a la ciencia de datos

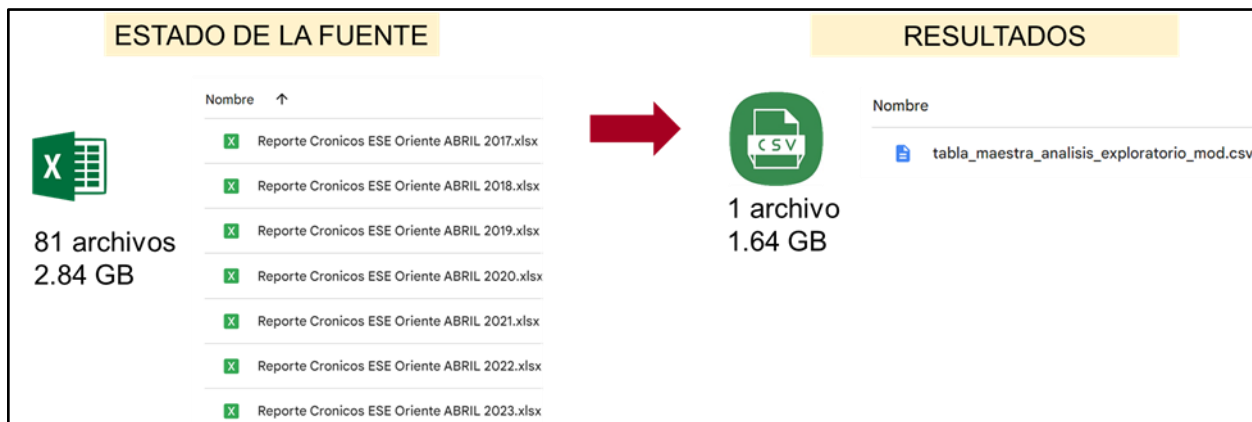
8.1. Selección y estudio de viabilidad del conjunto de datos

Se recibieron 81 archivos en formato xlsx correspondientes a los datos registrados entre los meses de septiembre 2016 a abril 2023 que en conjunto pesaban 2.84 GB los cuales presentaban una hoja de base con los datos, una hoja de tabla dinámica y otras hojas vacías o con extracción segmentada de los datos. Se extrajo de cada archivo la hoja de la base de datos y se le eliminó las primeras 4 filas que no contenían registros, pero ocupaban memoria con el logo de la Empresa. Así mismo se eliminaron los últimos registros que ocupaban espacio, pero no contenían data.

Al término de este proceso se generaron 81 archivos que en conjunto pesaban 1.10 GB. Para este procedimiento se utilizó un almacenamiento en la nube de Google y un procesamiento por medio de Google Colab. Entre los 81 archivos se obtuvieron 139 campos, en los que se encontraron variables duplicadas (con distinto nombre) y también campos nulos. Después de analizar estos casos y realizar las modificaciones

correspondientes se obtuvieron 130 variables, las cuales se pueden categorizar en información de identificación del paciente, campos demográficos, variables clínicas y conductuales. Con los archivos modificados manualmente para eliminar o renombrar archivos duplicados, se generó una tabla con el consolidado de los datos (Ver figura 2).

Figura 2. Proceso de consolidación de la información



Fuente. Elaboración propia (2024)

Con el archivo consolidado de 130 variables se realizó el análisis de aquellos campos que tuvieran el 100% de sus datos nulos. Se revisaron desde la fuente origen y al validar que desde el origen son nulos, se eliminó la variable. Así mismo se revisaron campos que se utilizaron como anotación y formulación del profesional de la salud y que no contenía mediciones propias de los pacientes. A continuación, en la tabla 3 se sintetiza el resultado de la información obtenida posterior al proceso de consolidación, al cabo de esta fase se redujo la cantidad de campos a 104 variables (Ver Anexo 1) en las que se pudo observar que a con el paso de los años se iban anexando variables al conjunto de datos.

Tabla 3. Resultado del proceso de unificación de campos y eliminación de columnas vacías

Estado de la fuente	Resultado
139 campos 2.874.477 registros	104 campos 2.867.535
Campos vacíos de columnas en blanco Unnamed:0, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88	Eliminación de los registros
Campos de ruido: 1900-01-02 00:00:00, 760010395712, 1, 2, 2019	
Campos vacíos de columnas nombradas: Clasificación DM, 40_FechaDxERC, 41_ProgramAteERC (NEFROPROTECCION), 15_FechaAfilePS, 17_FechaInnefro	Aunque la nulidad es alta, en primera instancia se decidió dejar los registros en el conjunto de datos unificados

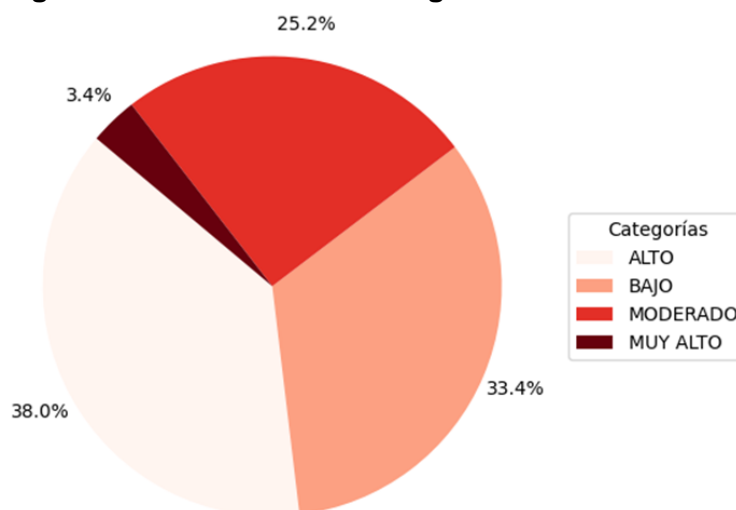
Fuente. Elaboración propia (2024).

8.2. Análisis exploratorio de datos

De acuerdo a la construcción del archivo unificado con la información de los pacientes desde el año 2016 a 2023 se encuentran las siguientes características de las variables analizadas del conjunto de datos y que en primera instancia son de interés.

- Respecto a la variable objetivo (CLASIFICACIÓN FINAL DEL RIESGO) en primer lugar se ubica riesgo alto con un 38.01%, en segundo lugar, riesgo bajo con un 33.44%, tercer lugar riesgo moderado con 25.17% y por último riesgo muy alto con un 3.38%

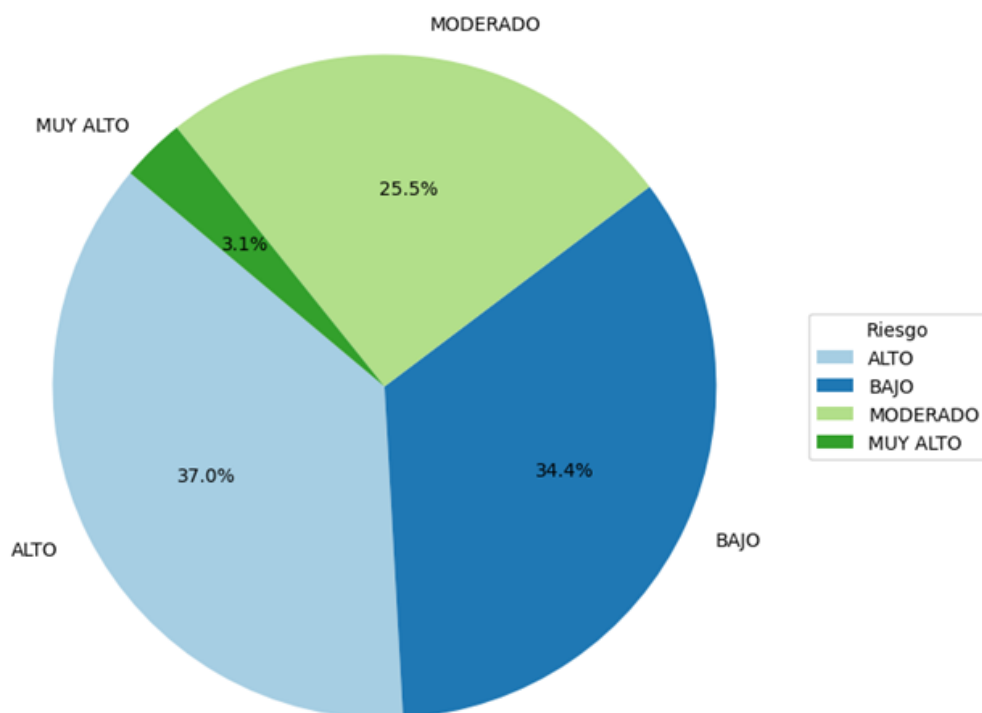
Figura 3. Clasificación del riesgo cardiovascular



Fuente. Elaboración propia (2024)

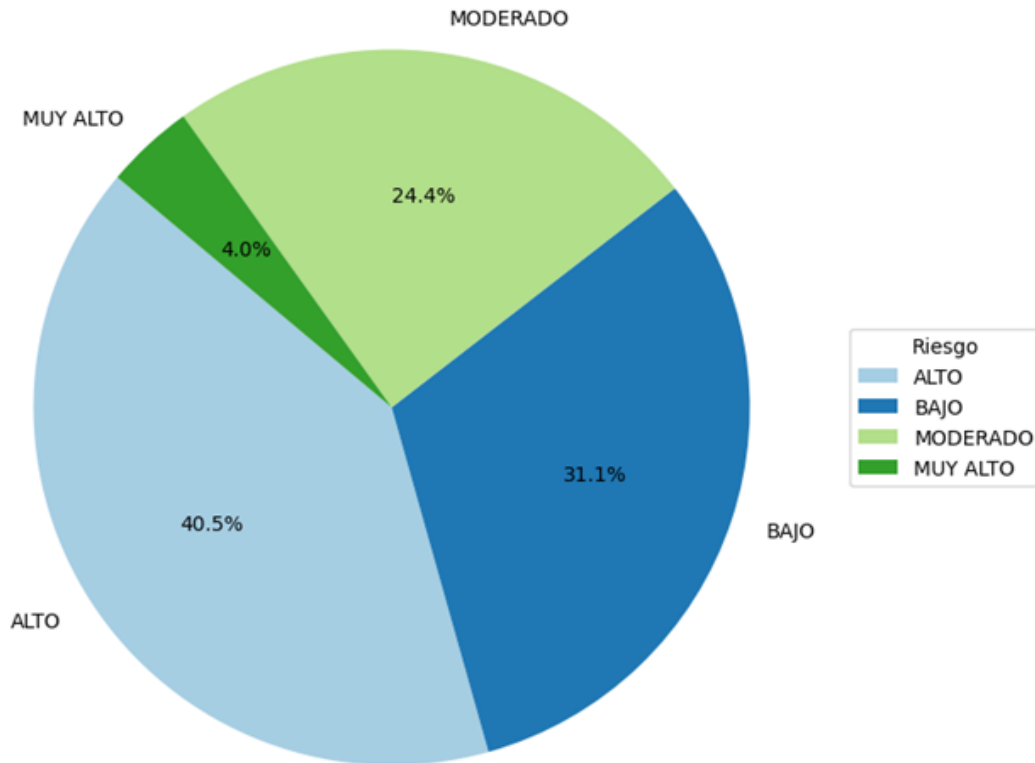
2. En relación a la variable género se encuentra la siguiente distribución: el 71.13% de los datos corresponden a Femenino y el 28.87% a Masculino. Asimismo, al cruzar esta variable con la de clasificación final del riesgo (ver figura 4) se concluye que la distribución del riesgo en el género Femenino predomina el riesgo alto con un 37%. Por su parte en el género Masculino (ver figura 5) predomina el riesgo alto con un 40.5%

Figura 4. Distribución del riesgo cardiovascular en el género femenino



Fuente. Elaboración propia (2024)

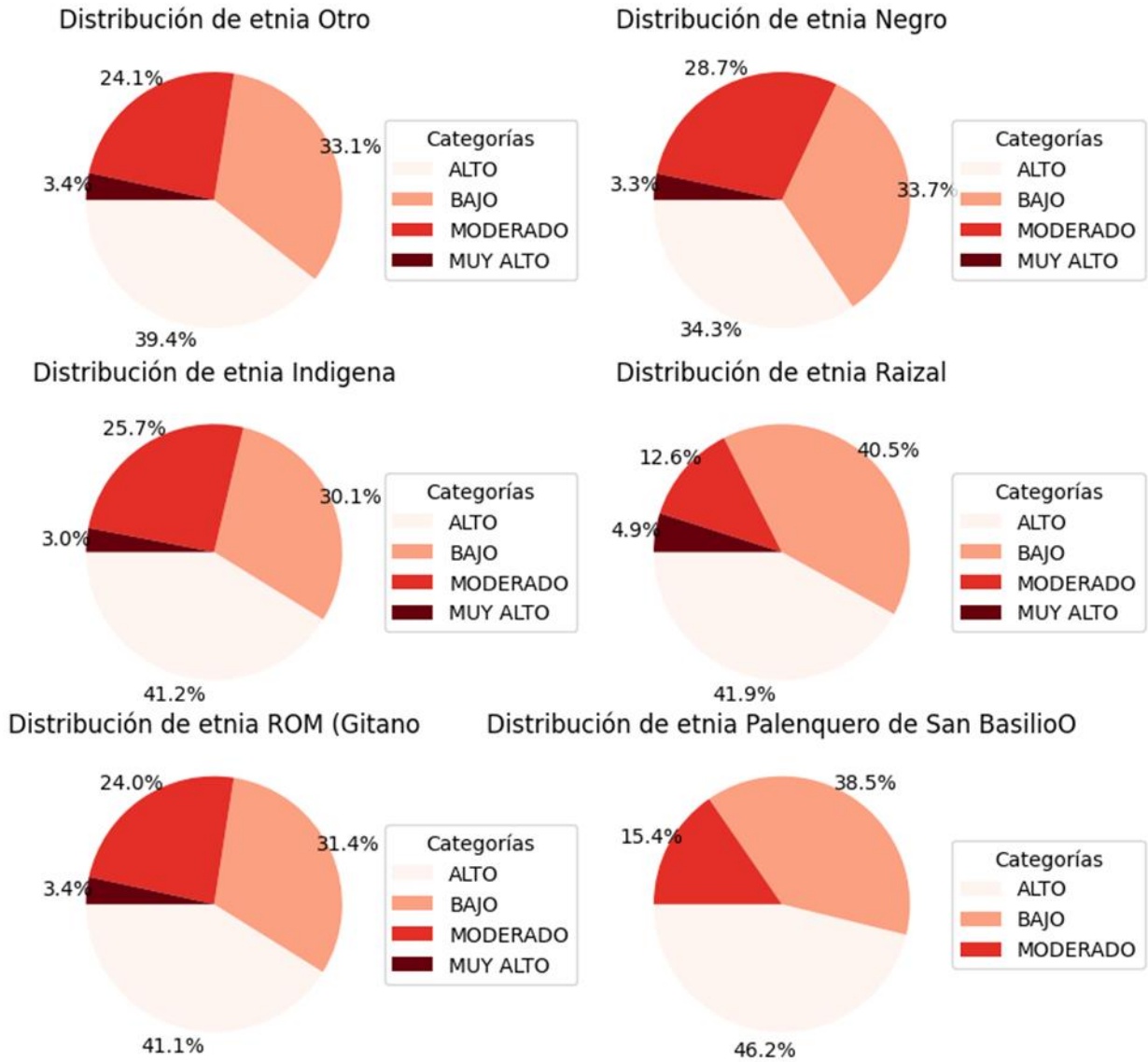
Figura 5. Distribución del riesgo cardiovascular en el género masculino



Fuente. Elaboración propia (2024)

- Respecto a la variable etnia, en primer lugar, se ubica el valor Otro con un 69.8%; en segundo lugar, la negra con un 28.3%; en tercer lugar, la indígena con un 1.7 y en cuarto lugar con menos de 1% se encuentran raizal, ROM (Gitano) y Palenquero de San Basilio. Al realizar el análisis bivariado entre la variable etnia y clasificación final del riesgo (ver figura 6), se encuentra que el riesgo cardiovascular alto se distribuye en igual proporción entre cada una de las diferentes etnias con una media de 40.6%.

Figura 6. Distribución del riesgo frente a etnia

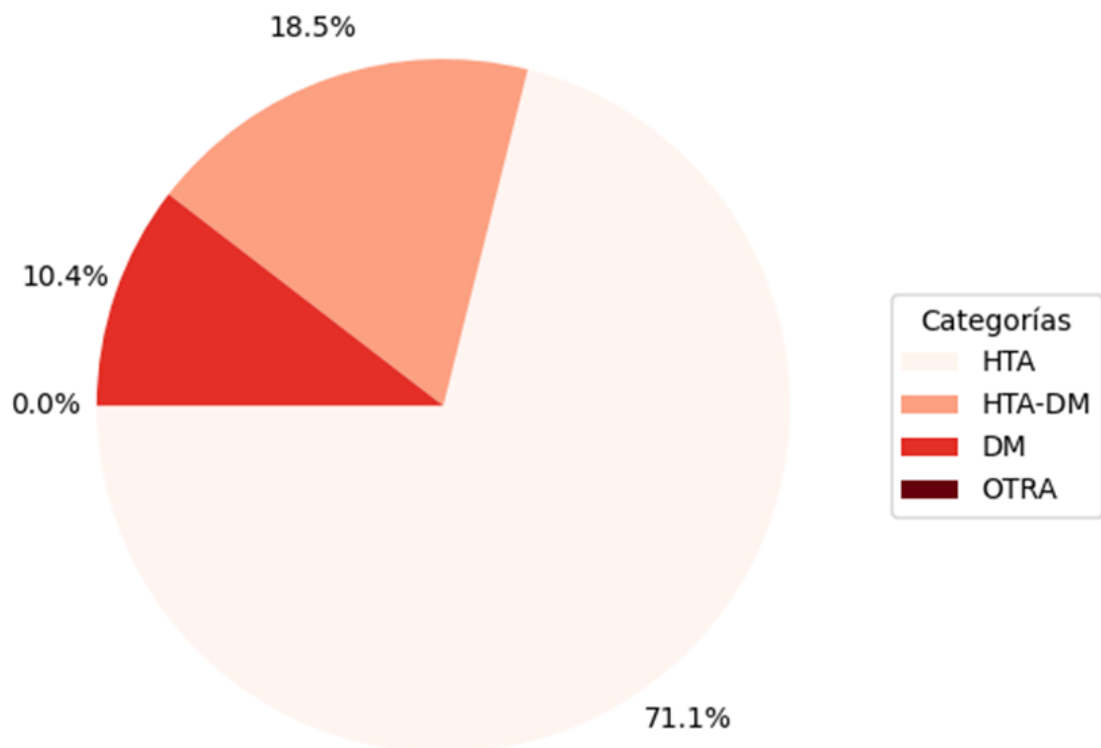


Fuente. Elaboración propia (2024)

4. En correspondencia a la clasificación de edad se aprecia que el grupo mayoritario se encuentra entre los 45 y 65 años con un 45.37%, seguido de los mayores de 65 con un 44.7% y por último los menores de 45 con un 9.91%
5. En concordancia al régimen de salud se encuentra que la mayoría pertenece al régimen subsidiado con un 97.92%, en segundo lugar, los no asegurados con un 1.54%, y en último lugar el contributivo con un 0.54% (se realizó unificación de campo ya que aparecía CONTRIBUTIVO y contributivo).
6. Relacionado a diagnósticos de patologías en la población se encuentra:

- a. Diabetes: El 71.1% de la población no presenta dicha patología frente a un 28.9% que está diagnosticada.
- b. Enfermedad renal crónica¹¹: Se encuentra relacionada a ciertos factores que pueden aumentar el riesgo entre los que se encuentran la hipertensión arterial (HTA), diabetes mellitus (DM), enfermedades del corazón, tabaquismo y obesidad. En la figura 7, se relacionan los factores asociados a la enfermedad renal crónica, en primer lugar, hipertensión arterial con un 71.7%, en segundo lugar hipertensión arterial y diabetes mellitus con un 18.5%, en tercer lugar diabetes mellitus con un 10.4%. De lo anterior, se puede inferir que una patología prevalente entre la población es la hipertensión arterial, variable explicativa del riesgo cardiovascular.

Figura 7. Factores asociados a la enfermedad renal crónica



Fuente. Elaboración propia (2024)

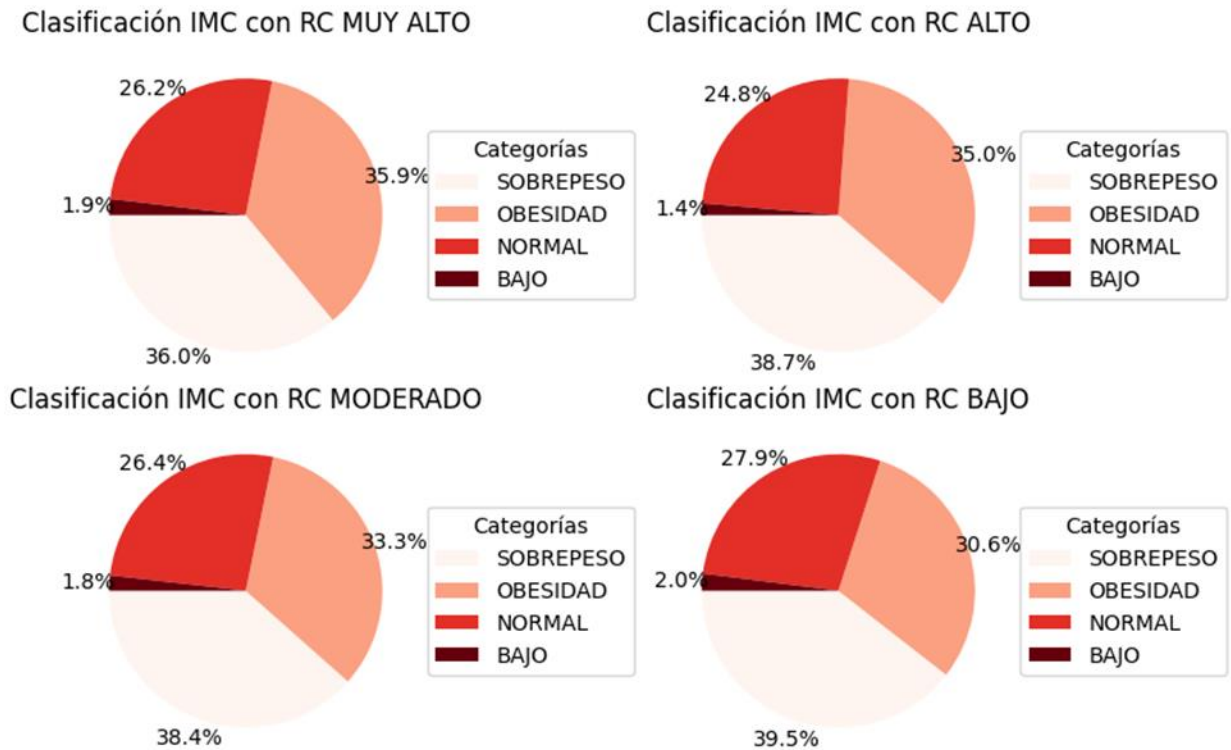
- 7. Respecto al índice de masa corporal¹² (IMC) se encuentra que el 38.8% tiene sobrepeso, seguido con un 33.1% con obesidad; asimismo un 26.3% se encuentra

¹¹ Según la Organización Panamericana de la Salud (OPS) la enfermedad renal crónica describe la pérdida gradual de la función renal, cuando la enfermedad renal crónica alcanza una etapa avanzada, niveles peligrosos de líquidos, electrolitos y los desechos pueden acumularse en el cuerpo (Organización Panamericana de la Salud, 2024)

¹² Relación entre la masa corporal de una persona y su estatura. Según los valores propuestos por la Organización Mundial de la Salud (OMS), el IMC es uno de los principales recursos para evaluar el estado

con un índice normal y, por último, con un 1.7% que se encuentra con un índice bajo. De lo anterior se puede inferir que 71.9% presenta problemas crónicos respecto al peso, variable directamente relacionada con problemas cardiovasculares, tal como se observa en la figura 8, en la que se aprecia una relación directa entre las diferentes clasificaciones del riesgo cardiovascular con la clasificación del IMC.

Figura 8. Clasificación del riesgo vs IMC

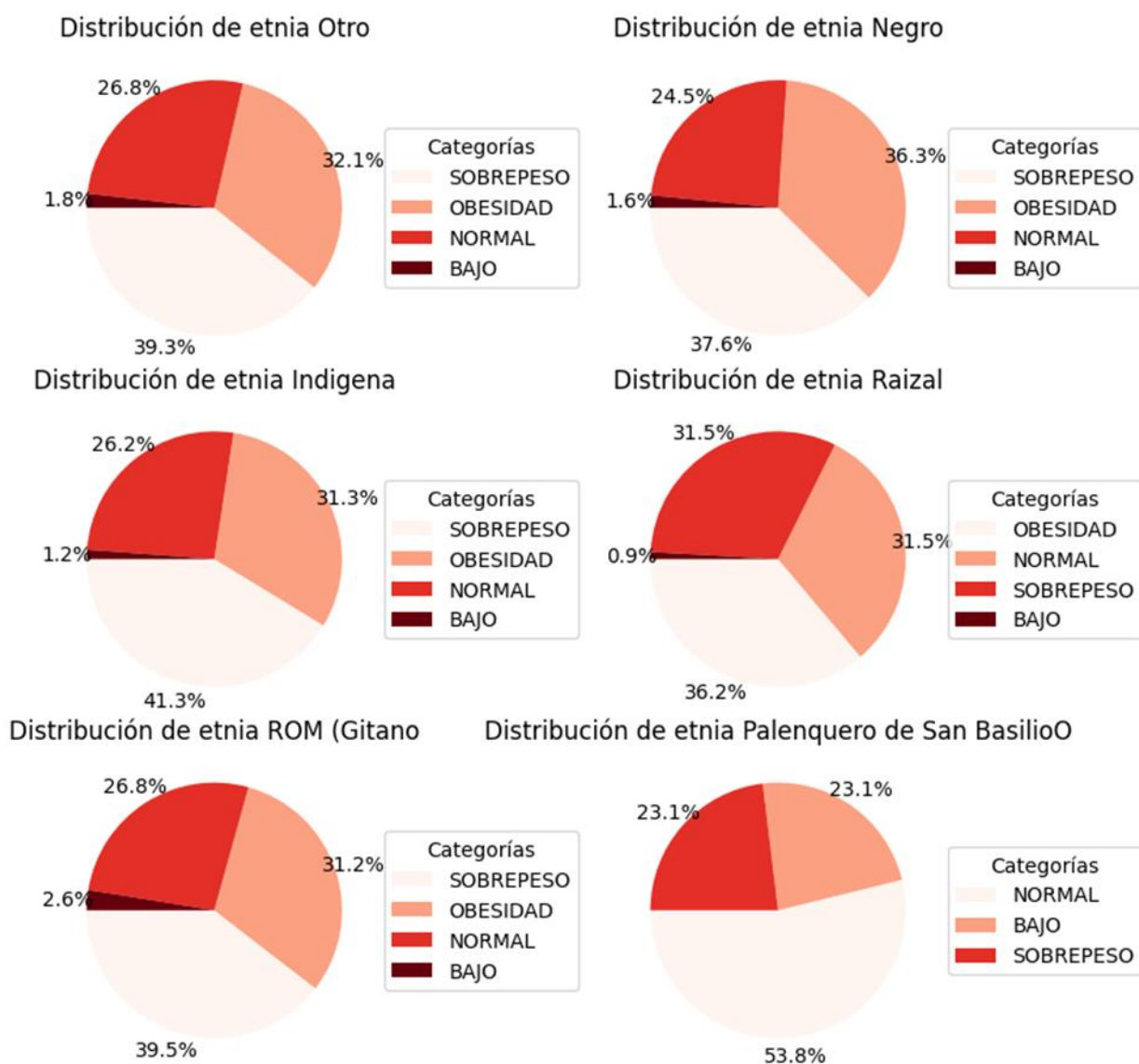


Fuente. Elaboración propia (2024)

- Al cruzar la información relacionada de etnia con el IMC se encuentra que el sobrepeso se distribuye en igual proporción en las distintas etnias con una media de 35.3%. Siendo la etnia con mayor prevalencia de sobrepeso la indígena con un 41.3% y con menor prevalencia los Palenqueros de San Basilio con un 23.1% (Ver figura 9).

Figura 9. Distribución del IMC entre las diferentes etnias

nutricional. Además, dichos valores varían, en función de la edad y el género, para lactantes, niños y adolescentes (Organización Mundial de la Salud, 2024).



Fuente. Elaboración propia (2024)

9. Relacionado a variables de hábitos¹³ se encuentran los siguientes resultados:
- Sedentarismo. Solamente el 11.31% de la población es clasificada como sedentaria frente al 88.69% que no se ubican en esta categoría.
 - Cigarrillo. El 2.84% de la población es categorizada como fumadora frente a un 97.16% que no se encuentran en esta clasificación.
 - Alcohol. El 5.27% de la población es clasificada como consumidora de alcohol frente un 94.73% que no consume.

¹³ Tener en cuenta que estas variables son recientes y no se encuentran desde el inicio en el conjunto de datos. Por lo que el porcentaje de nulidad llega a valores por encima del 70%.

- d. Alimentación inadecuada. El 23.92% de la población es categorizada como consumidora de alimentación inadecuada, es decir, que consume alta cantidad de sal, pocas frutas y verduras. Frente a un 76.08% que tiene una alimentación adecuada.

8.3. Identificación del Tipo de Problema y Paradigma de IA Adecuado

El análisis del riesgo cardiovascular se enmarca como un **problema de clasificación supervisada**, ya que se tiene un conjunto de datos de tipo cualitativo y etiquetados donde cada paciente está clasificado en función de su riesgo cardiovascular y principalmente de tipo **binario**, ya que se va a hacer énfasis en si un paciente puede desarrollar riesgo cardiovascular alto o bajo. En este sentido, el objetivo es predecir esta clasificación teniendo en cuenta en primer lugar el conjunto de variables seleccionadas y en segundo lugar clasificar con base en aquellas variables no contempladas en el modelo de medición Framingham Score (Ver tabla 2). Aprendiendo la relación entre las características y las clases para poder predecir la clase de nuevos datos desconocidos.

8.4. Proceso de selección y descarte de características. Elementos de preprocesamiento

El conjunto de datos, que se encontraba distribuido en varias columnas, presenta duplicidad en sus registros, particularmente en las variables categóricas, donde había redundancia en la información. Este error podría atribuirse a equivocaciones de digitación, ya que se encontraron valores como "Si", "si", "No", "no", "SI", "NO", que deberían haber sido estandarizados a una única forma, lo que generó la repetición innecesaria de los datos. En otras variables, se optó por realizar un agrupamiento de valores, ya que los valores obtenidos eran tan pequeños que no aportan información relevante o útil para el análisis. Al agruparlos, se busca simplificar la interpretación de los datos y mejorar la precisión de los resultados, eliminando aquellas categorías que no tienen un impacto significativo en el comportamiento general de las variables, ejemplo de ello fue la variable "Nivel de escolaridad".

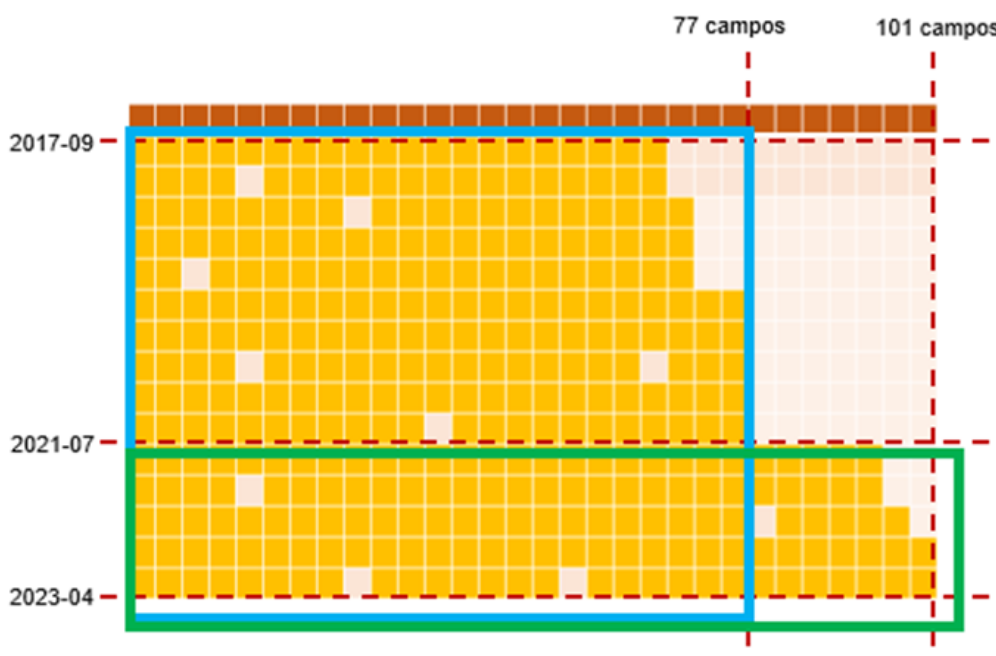
Además de lo mencionado, se aplicó la codificación one-hot encoding, cuyo propósito es convertir las variables categóricas, que contienen valores no numéricos, en un formato adecuado para ser utilizado por modelos matemáticos y algoritmos computacionales. Con la codificación, cada categoría se maneja de forma independiente, evitando interpretaciones incorrectas en el proceso de modelado.

Asimismo, a las variables numéricas¹⁴ se procedió a revisar la concordancia del tipo de dato con los valores registrados en estas variables, encontrándose que si había una correspondencia. Después del proceso anterior, se revisó el porcentaje de nulidad (Ver anexo 2) de los 104 campos con los que se realizó el análisis exploratorio y con base en ello se delimitaron las variables con las cuales se procederá a realizar el modelo de predicción del riesgo cardiovascular, teniendo en cuenta la evolución del número de columnas añadidas al conjunto de datos a lo largo del tiempo. A continuación, se presenta el número de columnas en cada periodo:

- Hasta 2017-09 hay 71 columnas
- Desde 2017-10 hasta 2019-12 hay 75 columnas
- Desde 2020-01 hasta 2021-06 hay 77 columnas
- Desde 2021-07 hasta 2022-08 hay 100 columnas
- Desde 2022-09 hasta 2023-04 hay 101 columnas

Con base en estos datos, se decidió realizar una partición en 2021-07 (ver figura 10). A partir de esta división, se consolidaron dos conjuntos de datos denominados: *Data Antigua* (con hasta 77 campos) y *Data Reciente* (con hasta 101 campos).

Figura 10. Partición del conjunto de datos



¹⁴ Variables de tipo numérico que fueron revisadas: 25_TenArtSis; 26_TenArtDitlica; 27_Creatinina; 28_HemoGlicosilada; 29_Albuminuria; 30_Creatinuria; 31_ColesterolTotal; 32_HDL; 33_LDL; 82_Glicemia; 83_Trigliceridos; 84_Cociente albuminuria/creatinuria; Resultado Penúltima Creatinina; Resultado Penúltima RAC

Fuente. Elaboración propia (2024)

A continuación, se detallan las estrategias de selección de variables para cada conjunto de datos, teniendo en cuenta los valores de correlación (ver tabla 4) y el porcentaje de nulidad. En primer lugar, el ítem de correlación se encuentra que en ambas datas las etiologías, especialmente, DM y ERC HTA-DM presentan una correlación positiva frente a la etiología ERC HTA. Además, con relación a los diagnósticos (Dx) la correlación positiva se presenta con el de diabetes mellitus y en la data reciente una variable que presenta una correlación positiva fue la glicemia.

Tabla 4. Análisis de correlación para ambos conjuntos de datos

Data Antigua	Data reciente
<ul style="list-style-type: none">• Etiología DM (0.39)• Etiología ERC HTA (- 0.72)• Etiología ERC HTA-DM (0.53)• Dx HTA (-0.39)• Dx DM (0.72)	<ul style="list-style-type: none">• Etiología DM (0.42)• Etiología ERC HTA (-0.82)• Etiología ERC HTA-DM (0.61)• Dx HTA (-0.42)• Dx DM (0.82)• Glicemia (0.39)

Fuente. Elaboración propia (2024)

Con base en estos resultados, se definieron las siguientes estrategias de selección de variables para cada conjunto de datos:

Data Antigua:

- Análisis con las variables sin nulidad (17 variables).
- Análisis con las variables con nulidad hasta el 6% (31 variables) y eliminación posterior de registros nulos.
- Análisis con las variables con nulidad hasta el 37% (39 variables) y eliminación posterior de registros nulos.

Data Reciente:

- Análisis con las variables sin nulidad (34 variables).
- Análisis con las variables con nulidad hasta el 21% (47 variables) y eliminación posterior de registros nulos.
- Análisis con las variables con nulidad hasta el 37% (58 variables) y eliminación posterior de registros nulos.
- Análisis con las variables sin nulidad + glicemia (valores nulos imputados a través de la media) (35 variables).

De acuerdo con lo mencionado, la estrategia de selección de variable para ambos conjuntos de datos fue:

- Para la data antigua trabajar con un total de 31 variables, es decir, manejando un 6% de nulidad.
- Para la data reciente se optó por trabajar con 35 variables de las cuales 34 no presentan nulidad y la última -glicemia- tenía un porcentaje bajo, los cuales fueron imputados a través de la media (Ver anexo 3).

Para la variable objetivo (CLASIFICACIÓN FINAL DEL RIESGO) se realizó un proceso de unificación de los campos con el propósito de dejar dos (2) variables y así estar coherente con el tipo de problema a solucionar planteado en ambos conjuntos de datos. De lo anterior, se puede concluir que las clases no presentan un gran desbalance entre ellas con una media de 59.325% con riesgo bajo y un 40.675% con riesgo alto (Ver tabla 5). Sin embargo, para evitar problemas de sesgos, se opta por utilizar k fold stratification, en este caso con un valor de k=5, asegurando una distribución de ambas clases en cada uno de los k subconjuntos, que sea proporcional a la distribución original. Además, de permitir a los modelos una mejor capacidad de generalización, es decir, la capacidad de aplicar lo aprendido en el conjunto de entrenamiento a datos nuevos o no observados, de manera efectiva y con buen rendimiento.

Tabla 5. Balanceo de clases

Conjunto de datos	Variable. Riesgo	Cantidad de datos	Proporción
Data antigua	0	1.629.273	58.54%
	1	1.153.635	41.46%
Data reciente	0	546.868	60.11%
	1	362.815	39.89%

Fuente. Elaboración propia (2024)

Por último, a las variables numéricas se llevó a cabo el proceso de normalización de sus valores para que todas tuvieran un rango comparable en este caso [0 y 1], evitando que las variables de mayor escala dominen el proceso de modelado y permitiendo no alterar la distribución de los datos, es decir, manteniendo las relaciones entre los valores dentro de cada característica

8.5. Algoritmos de entrenamiento

Para la predicción del riesgo cardiovascular, se eligieron los siguientes modelos de machine learning en función de su complejidad creciente y con base en las investigaciones consultadas en el numeral 7: regresión logística, árboles de decisión, random forest y gradient boosting. La regresión logística se seleccionó como el modelo más simple, ideal para establecer una base de comparación debido a su capacidad para

modelar relaciones lineales entre las variables predictoras y la probabilidad de riesgo. Los árboles de decisión se eligieron por su capacidad para capturar relaciones no lineales de manera interpretable. El random forest, un conjunto de árboles de decisión fue considerado por su robustez y capacidad para reducir el sobreajuste, mejorando la precisión en datos más complejos. Finalmente, el modelo de gradient boosting se implementó por ser uno de los más avanzados, reconocido por su alto rendimiento en tareas de clasificación complejas al combinar múltiples modelos débiles en uno fuerte.

Para evaluar y comparar el rendimiento de cada modelo, se calcularán varias métricas de desempeño clave: la exactitud (accuracy), que mide la proporción de predicciones correctas sobre el total de predicciones realizadas; la precisión (precisión), que evalúa la proporción de verdaderos positivos entre todos los elementos clasificados como positivos; la sensibilidad (recall), que indica la proporción de verdaderos positivos sobre todos los elementos que realmente son positivos, lo cual es crucial en el contexto de la predicción de enfermedades; el F1-Score, que proporciona un balance entre precisión y sensibilidad; y, finalmente, el área bajo la curva ROC (AUC-ROC), que mide la capacidad del modelo para distinguir entre clases positivas y negativas, reflejando su capacidad discriminativa. Estas métricas permiten identificar el modelo que ofrece el mejor equilibrio entre precisión y capacidad de detección, ayudando a seleccionar el mejor clasificador para el riesgo cardiovascular.

8.6 Generación de modelos

Para abordar el problema de predicción del riesgo cardiovascular con dos conjuntos de datos¹⁵, se diseñaron tres modelos principales, aplicando los mismos algoritmos (regresión logística, árboles de decisión, random forest y gradient boosting), pero con diferentes enfoques respecto a la selección y supresión de variables. Cada uno de estos modelos tuvo como objetivo evaluar cómo la inclusión o eliminación de variables impacta el desempeño general del modelo y su capacidad predictiva en ambos conjuntos de datos. A continuación, se describen los modelos utilizados:

Modelo 1: Conjunto Completo de Variables

En el primer modelo, se utilizó el conjunto completo de variables disponibles de ambos conjuntos de datos (Ver anexo 3), exceptuando la variable a predecir, en este caso denominada **RIESGO**. Este modelo sirvió como referencia para evaluar el desempeño de los algoritmos con la información más completa posible, sin realizar ningún tipo de supresión de variables. La variable **RIESGO** se mantuvo

¹⁵ Data Antigua y Data Reciente

como el objetivo a predecir, mientras que todas las demás variables fueron utilizadas como predictores.

Modelo 2: Supresión de Variables Relevantes

En el segundo modelo, se aplicó un proceso de selección de variables centrado en **descartar aquellas características con mayor influencia** en el desenlace del riesgo cardiovascular, en ambos conjuntos de datos. Para ello, se utilizó un análisis de **correlación** y se consideraron las **variables explicativas del riesgo** según la literatura analizada (Ver tabla 2). El análisis resultó en la selección de variables, principalmente **sociodemográficas** (Ver anexo 4), en ambos conjuntos de datos.

Al igual que en el Modelo 1, se mantuvo la variable **RIESGO** como el objetivo a predecir, asegurando que esta fuera el foco de las predicciones. Este enfoque permitió explorar si, al eliminar las variables de mayor influencia, otros tipos de variables, especialmente sociodemográficas, podrían proporcionar una explicación más adecuada del riesgo cardiovascular.

Modelo 3: Selección de Variables Framingham

El tercer modelo se centró exclusivamente en la **selección de las variables de la escala Framingham** para la predicción del riesgo cardiovascular, en ambos conjuntos de datos. En este caso, solo se consideraron las características definidas por la **escala Framingham** que incluyen las siguientes variables: peso, tensión arterial sistólica, tensión arterial diastólica, edad, sexo, diagnóstico de hipertensión, diagnóstico de diabetes mellitus, tratamiento IECA y ARA II, es decir, utilización de fármacos para tratar hipertensión arterial y otras patologías cardiovasculares como la insuficiencia cardíaca.

Para cada uno de los modelos, se utilizó una partición de los datos en conjuntos de entrenamiento (train) y prueba (test), con una distribución de 70% para entrenamiento y 30% para prueba. Esta división permitió evaluar el desempeño de los modelos de manera más precisa y generalizable, asegurando que el modelo se entrenará con una porción suficiente de los datos y se probará con un conjunto independiente para medir su capacidad predictiva.

Además, se llevó a cabo un proceso de identificación de las variables más importantes en cada modelo, tanto en el conjunto de datos de entrenamiento como en el de prueba,

utilizando técnicas específicas de selección de características. Este análisis se realizó en ambos conjuntos de datos, con el objetivo de identificar aquellas variables que contribuyen más significativamente a la predicción del riesgo cardiovascular.

El propósito de la generación de los modelos fue doble: por un lado, evaluar el desempeño predictivo de los modelos en cuanto a su capacidad para clasificar correctamente el riesgo cardiovascular, y por otro, analizar cómo la selección de variables influye en el rendimiento de los modelos. Esto permitió entender no sólo cuáles variables eran más relevantes para la predicción, sino también cómo la inclusión o eliminación de ciertas características podía afectar la precisión, la capacidad de generalización y el tiempo de entrenamiento de los modelos.

8.7 Resultados

Los resultados obtenidos de los tres modelos aplicados a los dos conjuntos de datos se presentan a continuación, con el objetivo de evaluar el desempeño de los algoritmos de *regresión logística*, *árboles de decisión*, *random forest* y *gradient boosting*. Cada uno de estos modelos fue entrenado con enfoques diversos en cuanto a la selección de variables, lo que permitió identificar cómo las distintas configuraciones de variables impactan la precisión y la efectividad de las predicciones en relación al riesgo cardiovascular.

De acuerdo con los resultados obtenidos a través de la aplicación de diversos modelos predictivos sobre los conjuntos de datos *Data Antigua* y *Data Reciente*, se observa una tendencia clara en cuanto a los modelos más efectivos y las variables clave que influyen en la predicción del riesgo cardiovascular. En general, el modelo *Random Forest* sobresale en ambos conjuntos de datos, destacándose por su alta capacidad para distinguir entre las clases positivas y negativas, como lo indica el AUC-ROC, que alcanza valores cercanos a 0.99. Este desempeño sugiere que *Random Forest* es el mejor modelo para ambas bases de datos, gracias a su habilidad para manejar grandes cantidades de información y capturar complejidades no lineales en los datos.

Modelo 1: Evaluación de Modelos con Datos Antiguos y Recientes

De acuerdo con los resultados del **Modelo 1**, el *Random Forest* es el modelo más robusto en términos de AUC-ROC, alcanzando un valor de 0.99 tanto en el conjunto de datos antiguo como en el reciente. Sin embargo, al examinar otras métricas como el *F1-Score*, se observa que el rendimiento es aún mejor cuando se utiliza el conjunto de datos reciente, con un *F1-Score* de 0.97. Esto indica una mayor eficiencia del modelo en la distinción de clases positivas (riesgo alto) y negativas (riesgo bajo), mejorando no solo

la capacidad de clasificación sino también la identificación precisa de los casos de alto riesgo.

Por otro lado, los modelos de *Regresión Logística* y *Gradient Boosting*, aunque presentaron buenos valores en términos de AUC-ROC (0.90 en ambos modelos para la *Data Antigua*), mostraron una disminución considerable en el *F1-Score* (0.82). Este descenso en el *F1-Score* señala que, aunque estos modelos son capaces de diferenciar entre clases, no logran detectar de manera óptima los casos de alto riesgo, lo que podría tener implicaciones negativas en aplicaciones prácticas donde la identificación precisa de los casos positivos es crucial.

Modelo 2: Comparación con la Supresión de Variables Relevantes

En el **Modelo 2**, que analiza el impacto de la supresión de variables relevantes, se observa un fuerte deterioro en el desempeño de todos los modelos evaluados, especialmente en el AUC-ROC. En los conjuntos de datos tanto antiguos como recientes, el AUC-ROC cae en un rango entre 0.52 y 0.79, lo que indica una significativa pérdida de capacidad predictiva al eliminar variables claves. A pesar de que otras métricas, como la precisión y el *recall*, siguen mostrando valores constantes, los modelos no logran mantener su efectividad sin las variables más informativas, lo que resalta la importancia de la correcta selección de variables. La inclusión de variables como *CodEtnia* (etnia) y *ENDOSALUD_SI* (régimen de salud) disminuye drásticamente la capacidad de los modelos para predecir de manera precisa el riesgo cardiovascular.

Modelo 3: Evaluación de Modelos en Variables Seleccionadas de Framingham

En el **Modelo 3**, que examina los modelos basados en la selección de variables de Framingham, se observa que, una vez más, *Random Forest* es el modelo que sobresale en ambos conjuntos de datos, con un AUC-ROC de 0.97 en los datos antiguos y 0.96 en los datos recientes. La *Regresión Logística* y *Gradient Boosting* nuevamente muestran un desempeño más limitado, con un AUC-ROC de 0.93 y 0.94 respectivamente. Sin embargo, el modelo de *Árboles de Decisión* continúa demostrando una sólida capacidad predictiva, especialmente cuando se considera la *DxHTA* y el *DxDM* como variables clave, lo que subraya la importancia de estas variables clínicas en la predicción del riesgo cardiovascular.

A continuación, en la Tabla 6, se sintetizan los resultados de las métricas de desempeño de cada uno de los modelos para cada algoritmo. Las métricas utilizadas incluyen exactitud (accuracy), precisión (precision), sensibilidad (recall), F1-score y AUC-ROC. Estas métricas permiten una evaluación exhaustiva de la capacidad predictiva de los

modelos, ofreciendo una visión integral de su rendimiento en distintos aspectos, como la correcta clasificación de casos positivos y negativos, la capacidad de detección de casos positivos reales, y la balanceada combinación de precisión y exhaustividad en las predicciones.

Tabla 6. Media de las puntuaciones de las diferentes métricas de desempeño de cada modelo en ambos conjuntos de datos

Conjunto de datos	Subconjunto de datos	Modelo	Exactitud	Precisión	Recall	F1-Score	AUC-ROC	Variables de mayor peso
Data Antigua	Conjunto completo de variables	Regresión logística	0.873	0.984	0.705	0.821	0.901	DxDM_Si, EtiologiaERC_HTA-DM, EtiologiaERC_DM, Edad
		Árboles de decisión	0.957	0.953	0.942	0.947	0.969	DxDM_Si, Índice de masa corporal, Edad, Peso
		Random forest	0.963	0.970	0.941	0.955	0.992	EtiologiaERC_HTA, DxDM_Si, Edad, Índice de masa corporal
		Gradient boosting	0.873	0.981	0.709	0.823	0.906	DxDM_Si, EtiologiaERC_HTA, TenArtSis, Edad
	Supresión de variables relevantes	Regresión logística	0.585	0.462	0.00011	0.00022	0.523	CodEtnia_Otro, ENDOSALUD_SI, CodEtnia_Indigena, CodEtnia_Raizal
		Árboles de decisión	0.585	0.539	0.0029	0.0057	0.523	CodEtnia_Negro, ENDOSALUD_SI, Regimen_NO_ASEGURADO, Regimen_SUBSIDIADO
		Random forest	0.585	0.538	0.0029	0.0058	0.523	CodEtnia_Negro, CodEtnia_Otro, ENDOSALUD_SI, Regimen_NO_ASEGURADO
		Gradient boosting	0.585	0.539	0.0029	0.0058	0.523	CodEtnia_Negro, ENDOSALUD_SI, Regimen_NO_ASEGURADO, Regimen_SUBSIDIADO
	Selección de variables Framingham	Regresión logística	0.873	0.984	0.705	0.821	0.900	DxDM_SI, Edad, TenArtSis, RcbeARA2_SI
		Árboles de decisión	0.928	0.947	0.877	0.911	0.962	DxDM_SI, Peso, Edad, TenArtSis
		Random forest	0.931	0.950	0.880	0.913	0.975	DxDM_SI, Peso, Edad, DxHTA_SI
		Gradient boosting	0.873	0.980	0.709	0.823	0.906	DxDM_SI, TenArtSis, Edad, RcbeARA2_SI

Conjunto de datos	Subconjunto de datos	Modelo	Exactitud	Precisión	Recall	F1-Score	AUC-ROC	Variables de mayor peso
Data Reciente	Conjunto completo de variables	Regresión logística	0.908	0.987	0.780	0.872	0.938	DxDM_Si, EtiologiaERC_HTA-DM, EtiologiaERC_DM, Peso
		Árboles de decisión	0.967	0.961	0.957	0.959	0.967	DxDM_Si, Glicemia, Índice de masa corporal, Edad
		Random forest	0.977	0.991	0.951	0.970	0.996	DxDM_Si, EtiologiaERC_HTA, EtiologiaERC_HTA-DM, Glicemia
		Gradient boosting	0.910	0.984	0.787	0.874	0.948	DxDM_Si, EtiologiaERC_HTA, TenArtSis, Edad
	Supresión de variables relevantes	Regresión logística	0.745	0.907	0.402	0.558	0.737	Glicemia, CodEtnia_Indigena, Regimen_CONTRIBUTIVO, CodEtnia_Otro
		Árboles de decisión	0.767	0.857	0.499	0.631	0.792	Glicemia, Escolaridad_Basica, CodEtnia_Negro, Escolaridad_Ninguno
		Random forest	0.767	0.855	0.500	0.631	0.792	Glicemia, CodEtnia_Negro, CodEtnia_Otro, Regimen_NO ASEGURADO
		Gradient boosting	0.756	0.839	0.481	0.612	0.767	Glicemia, CodEtnia_Negro, Regimen_NO ASEGURADO, Escolaridad_Ninguno
	Selección de variables Framingham	Regresión logística	0.908	0.989	0.779	0.871	0.936	DxDM_Si, TenArtSis, RcbeARA2_SI, Edad
		Árboles de decisión	0.933	0.972	0.857	0.911	0.960	DxDM_Si, TenArtSis, Edad, TenArtDitlica
		Random forest	0.934	0.973	0.858	0.912	0.967	DxDM_Si, DxHTA_SI, Edad, TenArtSis
		Gradient boosting	0.909	0.989	0.782	0.873	0.942	DxDM_Si, TenArtSis, Edad, RcbeARA2_SI

Fuente. Elaboración propia (2024)

9. Conclusiones

Los datos obtenidos de la población estudiada evidencian en su caracterización que el riesgo cardiovascular se distribuye de forma similar entre las mediciones baja, moderada y alta, siendo poco representativo el valor muy alto. En cuanto a género e IMC, no se evidencian distribuciones diferentes del riesgo. A nivel de etnia, la población raizal y palenquera presentan mayor cantidad de riesgos bajos que el resto de las etnias y una disminución del riesgo moderado, manteniendo valores similares en las categorías de riesgo alto y muy alto. La población estudiada está integrada principalmente por personas mayores de 45 años, con régimen subsidiado, con hábitos relativamente saludables y sin otros diagnósticos como hipertensión o diabetes.

Asimismo, el análisis muestra que, en general, los modelos basados en random forest y árboles de decisión son los más efectivos para predecir las condiciones de salud en los conjuntos de datos tanto antiguos como recientes, destacando la relevancia de variables relacionadas con la hipertensión, la diabetes, la edad, la etiología de enfermedad renal crónica y el IMC. La supresión de variables clave reduce significativamente el rendimiento de los modelos, lo que subraya la importancia de las características seleccionadas para mantener la precisión y efectividad del modelo. Además, la selección de variables específicas, como las basadas en el modelo de Framingham, puede ser útil para enfocar el análisis en factores de riesgo bien establecidos, sin comprometer la capacidad predictiva.

En contraste, explicar el riesgo cardiovascular sólo con variables sociodemográficas reduce el desempeño de los modelos por lo que es necesario incluir variables clínicas clásicas, tal como se evidenció con la incorporación de la glicemia en el modelo, pasando de un ROC-AUC de 0.52 a 0.79. Esto implica que, para obtener resultados robustos en tareas predictivas de alto impacto como la evaluación del riesgo cardiovascular, es crucial incluir las variables relevantes que puedan contribuir significativamente a la precisión y eficiencia del modelo.

Incorporar algunas variables conductuales y sociodemográficas nuevas fue el aporte de esta investigación donde se definieron la etnia, la escolaridad y el régimen como variables con algún peso de significancia media a alta, que pueden ayudar a predecir el riesgo cardiovascular y que su medición no implica altos costos de exámenes médicos o evolución del estado de salud con presencia de diagnósticos preexistentes.

Al final de esta investigación se concluye que se alcanzaron los objetivos planteados, caracterizando la población de estudio, identificando el mejor modelo predictivo y las variables que más peso presentan en la predicción del riesgo cardiovascular.

10. Recomendaciones y futuras investigaciones

1. La muestra de la población no presenta en su mayoría hábitos saludables pudiendo presentar un desbalance de datos en este tipo de clasificación. Se recomienda analizar la forma en que se obtiene la información y si a los pacientes les queda claro el concepto con el que se mide si una persona tiene tabaquismo, sedentarismo y buenos hábitos alimenticios o si presenta algún tipo de sesgo por parte de los pacientes (dar información imprecisa al personal de salud).
2. En el conjunto total de registros se encontró una tasa de 35 valoraciones por persona, lo que permitirá en futuras investigaciones evaluar la evolución de los pacientes a lo largo del estudio, principalmente aquellos que cambiaron el nivel de riesgo o aquellos que murieron y se conoce la causa de muerte.
3. Algunos campos evaluados son redundantes como la obesidad medida por el IMC y el campo Obesidad SÍ/NO. En estos casos se plantea dejar solo un campo, el que sea más fiable técnicamente y no por percepción.
4. Algunos campos se pueden simplificar si se estandariza y se reduce la cantidad de opciones presentadas como la escolaridad.
5. Hay campos muy abiertos que presentan opciones de diferentes categorías, las cuales, al seleccionar una opción excluyen otras características que no deberían ser tratadas en la misma categoría. Por ejemplo, la variable grupo poblacional presenta entre sus opciones estados de embarazo, estados de vulnerabilidad social, etnia, enfermedades mentales, entre otros.

Anexos

Anexo 1. Campos generados

Nombre del campo	N° Columna	Nombre del campo	N° Columna
File Name	1	31_ColesterolTotal	53
COMUNA	2	31_1_FechaCT	54
Consecutivo	3	32_HDL	55
Fehca de Inscripcion	4	32_1_FechaHDL	56
AÑO	5	33_LDL	57
Fecha Ultima Atención	6	33_1_FechaLDL	58
IPS	7	35_TFG	59
1_PrimerNombre	8	36_RcbeIECA	60
2_SegundoNombre	9	37_RcbeARA2	61
3_PrimerApellido	10	38_DxERC	62
4_SegundoApellido	11	ENDOSALUD	63
5_Tipoidentificacion	12	79_Novedad	64
7_FechaNac	13	81_FechaMuerte	65
EDAD (Años cumplidos)	14	80_CausaMuerte	66
GRUPO EDAD	15	82_Glicemia	67
8_Sexo	16	82_1_Fecha Glicemia	68
9_Regimen	17	83_Trigliceridos	69
10_EPS	18	83_1_Fecha Trigliceridos	70
11_CodEtnia	19	84_Cociente albuminuria/creatinuria	71
12_GrupoPob	20	84_1_Cociente albuminuria/creatinuria	72
METODO DE PLANIFICACION	21	Proteinas 0-15	73
13_CodMun	22	Eritrocitos 0-18 mgdl	74
14_Telefono	23	Cilindros patologicos	75
DIRECCION	24	Analisis Uroanalysis	76
15_Escolaridad	25	Fecha Uroanalysis	77
16_CodIPSsegto	26	Fecha Electrocardiograma	78
Zona	27	Fecha consulta medicina familiar	79
18_DxHTA	28	Fecha consulta medicina interna	80
Tipo de Consulta y profesional	29	Fecha consulta psicologia	81
20_DxDM	30	Fecha consulta nutricion	82
22_EtiologiaERC	31	Perimetro abdominal	83
23_Peso	32	Perimetro de cadera	84
24_Talla	33	Adherencia a tratamiento (Morisky green	85
Indice Masa Corporal = Peso/talla Al Cuadrado	34	Conocimiento de la enfermedad(Test de batalla)	86
Clasificación IMC	35	Consumo sustancias psicoactivas	87
Sedentario	36	Fecha proxima cita	88
Cigarrillo	37	Profesional que atendera (Medico o enfermera)	89
Sobrepeso	38	Resultado Penultima Creatinina	90
Obesidad	39	Fecha resultado penultima creatinina	91
Alcohol	40	Resultado Penultima RAC	92
Aliment inadecuada (sal, pocas frutas y verduras)	41	Fecha resultado penultima RAC	93
25_TenArtSis	42	Resultado Penultima RAC.1	94
26_TenArtDitlica	43	Fecha resultado penultima RAC.1	95
CLASIFICACION FINAL DEL RIESGO	44	ESTATINA	96

Nombre del campo	N° Columna	Nombre del campo	N° Columna
27_Creatinina	45	Clasificación HTA	97
27_1_FechaCrea	46	Resultado Antepenultima RAC	98
28_HemoGlicosilada	47	Fecha resultado antepenultima RAC	99
28_1_FechaHemoGlico	48	Eritrocitos 0-18 mg/dl	100
29_Albuminuria	49	Eritrocitos 0-18 mgd	101
29_1_FechaAlbnuria	50	Cruce Poblacion Asignada	102
30_Creatinuria	51	FINAL UROANALISIS	103
30_FechaCreatinuria	52	6_ID	104

Fuente. Elaboración propia (2024)

Anexo 2. Porcentaje de nulidad

Nombre de la Columna	Cantidad de Nulos	Tipo de Datos	Nulidad
File Name	0	object	0
COMUNA	2286125	object	0.7972439744
Consecutivo	0	int64	0
Fecha de Inscripción	1068145	object	0.3724958893
AÑO	0	float64	0
Fecha Última Atención	0	object	0
IPS	0	object	0
1_PrimerNombre	0	object	0
2_SegundoNombre	1869	object	0.0006517793157
3_PrimerApellido	0	object	0
4_SegundoApellido	820	object	0.0002859598924
5_Tipoidentificacion	0	object	0
7_FechaNac	0	object	0
EDAD (Años cumplidos)	0	float64	0
GRUPO EDAD	0	object	0
8_Sexo	0	object	0
9_Regimen	0	object	0
10_EPS	0	object	0
11_CodEtnia	66235	object	0.02309823594
12_GrupoPob	14203	float64	0.004953034575
METODO DE PLANIFICACION	2705101	object	0.9433541352
13_CodMun	235	float64	8.20E-05
14_Telefono	384794	object	0.1341898181
DIRECCION	576935	object	0.2011954518
15_Escolaridad	1483176	object	0.5172303041
16_CodIPSsegto	0	float64	0
Zona	1483176	object	0.5172303041
18_DxHTA	0	object	0
Tipo de Consulta y profesional	2028837	object	0.7075195246
20_DxDM	0	object	0
22_EtiologiaERC	0	object	0
23_Peso	10914	float64	0.003806056421

Nombre de la Columna	Cantidad de Nulos	Tipo de Datos	Nulidad
31_ColesterolTotal	385706	object	0.1345078613
31_1_FechaCT	385706	object	0.1345078613
32_HDL	401422	object	0.1399885267
32_1_FechaHDL	401422	object	0.1399885267
33_LDL	415247	object	0.1448097408
33_1_FechaLDL	415247	object	0.1448097408
35_TFG	271993	float64	0.09485254757
36_RcbeECA	132	object	4.60E-05
37_RcbeARA2	132	object	4.60E-05
38_DxERC	2867463	object	0.9999748913
ENDOSALUD	14907	object	0.005198541605
79_Novedad	2762024	object	0.9632049827
81_FechaMuerte	2855752	object	0.9958908958
80_CausaMuerte	2856136	object	0.9960248088
82_Glicemia	873933	object	0.3047680325
82_1_Fecha Glicemia	873933	object	0.3047680325
83_Trigliceridos	886281	object	0.30907417
83_1_Fecha Trigliceridos	886271	object	0.3090706827
84_Cociente albuminuria/creatinuria	2334558	object	0.8141340908
84_1_Cociente albuminuria/creatinuria	2334558	object	0.8141340908
Proteinas 0-15	2157986	object	0.7525578589
Eritrocitos 0-18 mgdl	2257906	object	0.787403118
Cilindros patologicos	2158005	object	0.7525644848
Analisis Uroanalisis	2193855	object	0.7650665118
Fecha Uroanalisis	2158101	object	0.7525979631
Fecha Electrocardiograma	2166170	object	0.7554118781
Fecha consulta medicina familiar	2524509	object	0.8803760024
Fecha consulta medicina interna	2524368	object	0.8803268312
Fecha consulta psicologia	2710518	object	0.9452432141
Fecha consulta nutricion	2632084	object	0.9178908017
Perimetro abdominal	2144946	float64	0.7480103992
Perimetro de cadera	2170946	float64	0.7570774202

Nombre de la Columna	Cantidad de Nulos	Tipo de Datos	Nulidad
24_Talla	11046	float64	0.003852088989
Indice Masa Corporal = Peso/talla Al Cuadrado	11558	float64	0.004030639556
Clasificación IMC	11558	object	0.004030639556
Sedentario	2143702	object	0.7475765771
Cigarrillo	2054821	object	0.7165809659
Sobrepeso	2681623	object	0.9351666152
Obesidad	2694934	object	0.9398085812
Alcohol	2143765	object	0.7475985472
Aliment inadecuada (sal, pocas frutas y verduras)	2865620	object	0.999332179
25_TenArtSis	6700	object	0.00233650156
26_TenArtDitlica	6719	object	0.00234312746
CLASIFICACION FINAL DEL RIESGO	0	object	0
27_Creatinina	270716	object	0.09440721735
27_1_FechaCrea	270704	object	0.09440303257
28_HemoGlicosilada	2023419	object	0.7056300969
28_1_FechaHemoGlic o	2023419	object	0.7056300969
29_Albuminuria	759765	object	0.2649540459
29_1_FechaAlbnuria	759765	object	0.2649540459
30_Creatinuria	1514862	object	0.5282802128
30_FechaCreatinuria	1514398	object	0.5281184013

Nombre de la Columna	Cantidad de Nulos	Tipo de Datos	Nulidad
Adherencia a tratamiento (Morisky green)	2162990	object	0.7543029117
Conocimiento de la enfermedad(Test de batalla)	2263765	object	0.7894463363
Consumo sustancias psicoactivas	2134597	object	0.7444013761
Fecha proxima cita	2186005	object	0.762328969
Profesional que atendera (Medico o enfermera)	2186135	object	0.7623743041
Resultado Penultima Creatinina	2239221	object	0.7808870685
Fecha resultado penultima creatinina	2239221	object	0.7808870685
Resultado Penultima RAC	2845525	object	0.992324418
Fecha resultado penultima RAC	2845459	object	0.9923014017
Resultado Penultima RAC.1	2852911	float64	0.9949001494
Fecha resultado penultima RAC.1	2852889	object	0.9948924773
ESTATINA	2775230	object	0.9678103319
Clasificación HTA	2685941	object	0.9366724382
Resultado Antepenultima RAC	2867434	float64	0.9999647781
Fecha resultado antepenultima RAC	2867434	object	0.9999647781
Eritrocitos 0-18 mg/dl	2836042	object	0.9890173965
Eritrocitos 0-18 mgd	2799096	object	0.9761331597
Cruce Poblacion Asignada	2774847	float64	0.9676767677
FINAL UROANALISIS	2837111	object	0.9893901905
6_ID	0	object	0

Fuente. Elaboración propia (2024)

Anexo 3. Nombre de variables seleccionadas de cada conjunto de datos

Conjunto de datos	Data Antigua	Data Reciente
Variables seleccionadas	11_CodEtnia_Indigena	11_CodEtnia_Indigena
	11_CodEtnia_Negro	11_CodEtnia_Negro
	11_CodEtnia_Otro	11_CodEtnia_Otro
	11_CodEtnia_Palenquero de San Basilio	11_CodEtnia_Raizal
	11_CodEtnia_Raizal	11_CodEtnia_ROM (Gitano)
	11_CodEtnia_ROM (Gitano)	15_Escolaridad_Basica
	18_DxHTA_NO	15_Escolaridad_Media tecnica
	18_DxHTA_SI	15_Escolaridad_Ninguno
	20_DxDM_SI	15_Escolaridad_Profesional
	22_EtiologiaERC_DM	18_DxHTA_NO
	22_EtiologiaERC_HTA	18_DxHTA_SI
	22_EtiologiaERC_HTA-DM	20_DxDM_SI
	23_Peso	22_EtiologiaERC_DM
	24_Talla	22_EtiologiaERC_HTA
	25_TenArtSis	22_EtiologiaERC_HTA-DM
	26_TenArtDitlica	23_Peso
	36_RcbeIECA_SI	24_Talla
	37_RcbeARA2_SI	25_TenArtSis
	8_Sexo_F	26_TenArtDitlica
	9_Regimen_CONTRIBUTIVO	36_RcbeIECA_SI
	9_Regimen_NO ASEGURADO	37_RcbeARA2_SI
	9_Regimen_SUBSIDIADO	8_Sexo_F
	CLASIFICACION FINAL DEL RIESGO_ALTO	82_Glicemia
	CLASIFICACION FINAL DEL RIESGO_BAJO	9_Regimen_CONTRIBUTIVO
	CLASIFICACION FINAL DEL RIESGO_MODERADO	9_Regimen_NO ASEGURADO
	CLASIFICACION FINAL DEL RIESGO_MUY ALTO	9_Regimen_SUBSIDIADO
	EDAD (Años cumplidos)	CLASIFICACION FINAL DEL RIESGO_ALTO
	ENDOSALUD_SI	CLASIFICACION FINAL DEL RIESGO_BAJO
	Indice Masa Corporal = Peso/talla Al Cuadrado	CLASIFICACION FINAL DEL RIESGO_MODERADO
		CLASIFICACION FINAL DEL RIESGO_MUY ALTO
		EDAD (Años cumplidos)
		Indice Masa Corporal = Peso/talla Al Cuadrado

Fuente. Elaboración propia (2024)

Anexo 4. Nombre de variables seleccionadas para el modelo dos en ambos conjuntos de datos

Conjunto de datos	Data Antigua	Data Reciente
Variables seleccionadas	11_CodEtnia_Indigena 11_CodEtnia_Negro 11_CodEtnia_Otro 11_CodEtnia_Palenquero de San Basilio 11_CodEtnia_Raizal 11_CodEtnia_ROM (Gitano) 9_Regimen_CONTRIBUTIVO 9_Regimen_NO ASEGURADO 9_Regimen_SUBSIDIADO ENDOSALUD_SI	11_CodEtnia_Indigena 11_CodEtnia_Negro 11_CodEtnia_Otro 11_CodEtnia_ROM (Gitano) 11_CodEtnia_Raizal 15_Escolaridad_Basica 15_Escolaridad_Media tecnica '82_Glicemia' 15_Escolaridad_Ninguno 15_Escolaridad_Profesional 9_Regimen_CONTRIBUTIVO 9_Regimen_NO ASEGURADO 9_Regimen_SUBSIDIADO

Fuente. Elaboración propia (2024)

Referencias

- Alcaldía de Santiago de Cali. (2024). *Análisis situacional de salud participativo - ASIS 2023*. Cali.
- Andaur Navarro, C. L., Damen, J. A., Takada, T., Nijman, S. W., Dhiman, P., Ma, J., . . . Hooft, L. (2020). Protocol for a systematic review on the methodological and reporting quality of prediction model studies using machine learning techniques. *BMJ Open*. doi:<https://doi.org/10.1136/bmjopen-2020-038832>
- Cho, S. Y., Kim, S. H., Kang, S. H., Lee, K. J., Choi, D., Kang, S., . . . Chae, I. H. (2021). Pre-existing and machine learning-based models for cardiovascular risk prediction. *Scientific Report*. doi:<https://doi.org/10.1038/s41598-021-88257-w>
- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Annals of Internal Medicine*, 55-63. doi:<https://doi.org/10.7326/M14-0697>
- Daza, G., Castañeda, J., & Castaño, J. I. (2022). Plataforma de analítica de datos para la caracterización poblacional y la evaluación del riesgo cardiovascular en pacientes del centro-occidente de Colombia. *Salud UIS*. doi:<https://doi.org/10.18273/saluduis.54.e:22042>
- Donado Gómez, J. H. (2017). Modelos de predicción de riesgo cardiovascular. *Medicina UPB*, 153-160.
- Donado Gómez, J. H., Higueta-Duque, L. N., & Castro-Palacio, J. J. (2017). Modelos de predicción de riesgo cardiovascular. *Modelos de predicción de riesgo cardiovascular*, 153-160.
- Fernández de Larrea-Baz, N., Morant-Ginestar, C., Catalá-López, F., & Gènova-Maleras, R. &.-M. (2015). Años de vida ajustados por discapacidad perdidos por cardiopatía isquémica en España. *Revista Española de Cardiología*, 968-975.
- Lira, M. T. (2022). Estratificación de riesgo cardiovascular: conceptos, análisis crítico, desafíos e historia de su desarrollo en Chile. *Revista Médica Clínica Las Condes*, 534-544. doi:<https://doi.org/10.1016/j.rmcl.2022.08.003>
- Lobos, B. J., & Brotons, C. C. (2011). Factores de riesgo cardiovascular y atención primaria: evaluación e intervención. *Atención Primaria*, 668-677. doi:<https://doi.org/10.1016/j.aprim.2011.10.002>
- Martinez, R., Soliz, P., Caixeta, R., & Ordunez, P. (2019). Años de vida perdidos por muerte prematura: una medida versátil y abarcadora para el monitoreo de la mortalidad por enfermedades no transmisibles. *Pan American Journal of Public Health*, 1-10. doi:<https://doi.org/10.1093/ije/dyy254>
- Ministerio de Salud y Protección Social. (29 de Septiembre de 2023). *Ministerio de Salud y Protección Social*. Obtenido de Ministerio de Salud y Protección Social: <https://onx.la/ae4ea>

- Organización Mundial de la Salud. (2019). *Global Health Estimates 2019: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019*.
- Organización Mundial de la Salud. (11 de junio de 2021). *Organización Mundial de la Salud*. Recuperado el 9 de marzo de 2024, de [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- Organización Panamericana de la Salud. (2021). OPS. Obtenido de OPS: <https://www.paho.org/es/enlace/carga-enfermedades-cardiovasculares>
- Organización Panamericana de la Salud. (s.f.). *Organización Panamericana de la Salud*. Obtenido de paho.org: <https://www.paho.org/en/enlace/technical-notes>
- Pico Fonseca, S. M., Hernández Carrillo, M., & Muñoz Orozco, L. C. (2022). Descripción espacial del riesgo cardiovascular en población adulta mayor: Caso de Cali-Colombia. *Nutrición Clínica y Dietética Hospitalaria*, 133-141. doi:10.12873/422pico
- Puymirat, E., Bonaca, M., Fumery, M., Tea, V., Aissaoui, N., Lemesles, G., . . . investigators, F.-M. (2019). Atherothrombotic risk stratification after acute myocardial infarction: The Thrombolysis in Myocardial Infarction Risk Score for Secondary Prevention in the light of the French Registry of Acute ST Elevation or non-ST Elevation. *Clinical cardiology*, 227-234. doi:<https://doi.org/10.1002/clc.23131>
- Rossello, X., Dorresteijn, J. A., Janssen, A., Lambrinou, E., Scherrenberg, M., Bonnefoy-Cudraz, E., . . . Nursing, E. J. (2019). Risk prediction tools in cardiovascular disease prevention: A report from the ESC Prevention of CVD Programme led by the European Association of Preventive Cardiology (EAPC) in collaboration with the Acute Cardiovascular Care Association (ACCA) and the As. *European journal of preventive cardiology*, 1534-1544. doi:<https://doi.org/10.1177/2047487319846715>
- Sarraj, A., Ward, A., Chung, S., Li, J., Scheinker, D., & Rodríguez, F. (2021). Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open Heart*. doi:<https://doi.org/10.1136/openhrt-2021-001802>
- Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., . . . Group†, P. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of internal medicine*, 51-58. doi:<https://doi.org/10.7326/M18-1376>