**UNIVERSIDAD NACIONAL DE CAAGUAZÚ**
**FACULTAD DE CIENCIAS Y TECNOLOGÍAS**
**CARRERA DE INGENIERÍA EN INFORMATICA**

PROYECTO FINAL DE GRADO

# Modelo de Predicción para alerta temprana de Eventos Cardiovasculares mediante técnicas de Supervised Machine Learning en el Instituto de Previsión Social (IPS) de Coronel Oviedo, 2024

## AUTORES:

José Antonio Espinoza Franco
Luz Melina Vázquez Cáceres

## TUTOR:

Prof. Mg. Ing. Víctor Manuel Melgarejo

CORONEL OVIEDO, DICIEMBRE DEL 2024

# DERECHO DE AUTOR

Quien/es suscribe/n, José Antonio Espinoza Franco y Luz Melina Vázquez Cáceres, autor/a/autores del trabajo de investigación titulado **"Modelo de Predicción para alerta temprana de Eventos Cardiovasculares mediante técnicas de Supervised Machine Learning en el Instituto de Previsión Social (IPS) de Coronel Oviedo, 2024"**, declara/n que voluntariamente cede/n a título gratuito en forma pura y simple ilimitada e irrevocablemente a favor de la Facultad de Ciencias y Tecnologías – UNCA, el derecho de autor de contenido patrimonial, que le corresponde sobre el trabajo de referencia. Conforme a lo anteriormente expresado, esta sesión le otorga a la FCyT la Facultad de comunicar la obra divulgarla, publicarla y reproducirla en soportes analógicos o digitales en la oportunidad que así lo estime conveniente. La FCyT deberá indicar qué autoría o creación del trabajo corresponde a mi persona y hará referencia al autor y a las personas que hayan colaborado en la realización del presente trabajo de investigación.

En la ciudad de Coronel Oviedo a los …. , del mes de …….. del 2024

…… ……

José Antonio Espinoza Franco Luz Melina Vázquez Cáceres

# PÁGINA DE APROBACIÓN

Trabajo de fin de grado para la obtención del Título de Ingeniero en Sistemas Informáticos, aprobado en representación de la Facultad Ciencias y Tecnologías de la Universidad Nacional de Caaguazú, por el Tribunal Examinador constituido por los siguientes profesores y con la siguiente nota final:

CALIFICACIÓN FINAL: _____

ACTA N°:_____

FECHA :_____

---
Prof. Ing.

---
Prof. Ing.

---
Prof. Ing.

# DEDICATORIA

En primer lugar, dedico este trabajo final de grado a Dios, por sostenerme y darme fortalezas en todo momento, por ayudarme a perseverar y seguir adelante a pesar de los obstáculos de la vida, y por sobre todo por guiarme y darme sabiduría en todo momento.

A mis queridos padres, Héctor y Myrta, quienes me han dado su apoyo y entrega incondicional para lograr cada objetivo en la vida. Su sabiduría y paciencia me han enseñado el valor del sacrificio y de afrontar con calma las situaciones adversas de la vida.

A mis hermanos, Ana, Karen e Igort, por brindarme su amor y paciencia incondicional, gracias por ayudarme en cada paso de mi vida y ser un pilar en mi día a día.

A mi compañera de vida, Belén González, Gracias por creer en mí siempre y por llenar mis días de amor y alegría. Este logro es también tuyo.

A mis tíos Guido Espinoza, Elinora Figueredo Valdés, Sonia Franco, Marlene Franco por brindarme su apoyo y cariño incondicional.

A mis abuelos, padrinos, tíos, sobrinos, por todo el cariño y amor, por animarme y aconsejarme para mi crecimiento personal y profesional, ustedes han sido pilares fundamentales en mi vida y agradezco inmensamente su apoyo.

A todos aquellos que, de una forma u otra me han brindado su apoyo incondicional.

José Antonio Espinoza Franco.

Dedico este Trabajo Final de Grado a Dios, quien ha sido mi guía, mi fortaleza a lo largo de toda mi carrera, su presencia y su sabiduría han sido mi mayor inspiración recordándome en todo momento que Dios tiene un propósito en cada paso que damos.

A mis padres, Alcides y Celsa, a quienes les debo mi gratitud infinita, ambos han sido mi apoyo incondicional en este arduo camino. Han secado mis lágrimas, han compartido las madrugadas de estudio y han sido mi motivación constante. Gracias a los dos, este esfuerzo es por ustedes y para ustedes.

A mi hermano Alcides, por su amor, apoyo y complicidad a lo largo de los años, su presencia en mi vida ha sido un regalo invaluable y una compañía hermosa.

A mis abuelos, padrinos, tíos, mis sobrinas, por todo el amor, alegría y apoyo incondicional que me han inspirado a no rendirme.

Pero, sobre todo, este logro es para mí, un testimonio de mi determinación, perseverancia y capacidad para alcanzar mis metas, a pesar de todas las adversidades. Me enorgullece el trabajo realizado y estoy emocionada por lo que está por venir.

¡Gracias a todos por ser parte de este viaje, su amor, sabiduría y su apoyo incondicional!
Esfuérzate y Sé Valiente.
Josué 1:9.

Luz Melina Vázquez Cáceres.

# AGRADECIMIENTOS

A mi tutor, Prof. Mg. Ing. Víctor Manuel Melgarejo Riveros por su incondicional apoyo, dedicación y voluntad, quien me ha orientado y guiado con sus conocimientos en la realización de este trabajo.

A todos los docentes de la carrera de Ingeniería en Sistemas Informáticos por sus enseñanzas y experiencias compartidas.

A mi compañera de trabajo Luz Vázquez por el excelente trabajo realizado y la dedicación puesta.

A los profesionales de la Salud Dr. Denis Figueredo (Director del Instituto de Previsión Social de Coronel Oviedo) y la Dra. Natalia Zorrilla (Cardióloga del Instituto de Previsión Social de Coronel Oviedo) por guiarnos con sus conocimientos en el área médica para la realización de este trabajo.

Al Programa de Becas de Grado de la Entidad Binacional Itaipú, por el apoyo financiero otorgado a través de la beca obtenida en la convocatoria 2018.

A todos mis familiares, amigos, hermanos de iglesia, compañeros y todas las personas que de alguna manera estuvieron a mi lado, muchas gracias.

José Antonio Espinoza Franco.

A mi tutor, Prof. Mg. Ing. Víctor Manuel Melgarejo Riveros por su paciencia, dedicación y apoyo, quien me guio con todos sus conocimientos en cada paso de este trabajo.

A todos los docentes de la carrera de Ingeniería en Sistemas Informáticos por haber compartido conmigo sus conocimientos.

A mi compañero de trabajo José Espinoza por el excelente trabajo realizado y la dedicación puesta.

A los profesionales de la Salud Dr. Denis Figueredo (Director del Instituto de Previsión Social de Coronel Oviedo) y la Dra. Natalia Zorrilla (Cardióloga del Instituto de Previsión Social de Coronel Oviedo) por guiarnos con sus conocimientos en el área médica para la realización de este trabajo.

A mis compañeras y amigas Mariela y Guadalupe por todo el apoyo y la amistad en todos estos años de constancia y lucha.

A mis mejores amigos Clara y Joel que me apoyaron, incentivaron y motivaron para seguir adelante.

A todos mis amigos y todas las personas que de alguna manera estuvieron a mi lado, muchas gracias.

Luz Melina Vázquez Cáceres.

# RESUMEN

Las enfermedades cardiovasculares están relacionado a una serie de problemas con el corazón y los vasos sanguíneos y es la principal causa de muerte en todo el mundo. Según el Ministerio de Salud Pública y Bienestar Social, las enfermedades cardiovasculares son responsables del 30% de las muertes a nivel mundial y del 27% de las muertes a nivel nacional. En el Instituto de Previsión social de la ciudad de Coronel Oviedo se registró 82.991 consultas de pacientes entre 2021 y 2024, muchos de los cuales presentaban factores de riesgo importantes como la hipertensión esencial.

Para abordar esta problemática, se desarrollaron diversos modelos utilizando técnicas de Supervised Machine Learning. Se entrenaron modelos de Regresión Logística, Arboles de Decisión y Random Forest, de los cuales el Random Forest ha sido seleccionado como modelo de alerta temprana para la aplicación dentro de la interfaz de predicción de eventos cardiovasculares ya que el mismo ha arrojado un 73% de exactitud y una cantidad mínima de errores.

Además, se planteó la creación de una base de datos que permitirá almacenar información sobre los pacientes, sus consultas médicas y generar un historial clínico para futuras predicciones, Por lo que esta herramienta contribuirá a mejorar la gestión y el seguimiento de la salud cardiovascular de los pacientes.

**Palabras clave:** Supervised Machine Learning, Eventos Cardiovasculares, Modelos predictivos, Interfaz de predicción.

# ABSTRACT

Cardiovascular disease is related to a number of problems with the heart and blood vessels and is the leading cause of death worldwide. According to the Ministry of Public Health and Social Welfare, cardiovascular diseases are responsible for 30% of deaths worldwide and 27% of deaths nationally. At the Social Welfare Institute of the city of Coronel Oviedo, 82,991 patient consultations were recorded between 2021 and 2024, many of whom had major risk factors such as essential hypertension.

To address this problem, several models were developed using Supervised Machine Learning techniques. Logistic regression, Decision Trees and Random Forest models were trained, of which the Random Forest has been selected as the early warning model for application within the cardiovascular event prediction interface as it has yielded 73% accuracy and a minimal amount of errors.

In addition, the creation of a database was proposed to store information about patients, their medical consultations and generate a clinical history for future predictions, so this tool will contribute to improve the management and monitoring of patients' health.

**Keywords:** Supervised Machine Learning, Cardiovascular events, Predictive models, Prediction interface.

# ÍNDICE

Introducción ..... 1

1 Modelo de Predicción para alerta temprana de Eventos Cardiovasculares mediante técnicas de Supervised Machine Learning en el Instituto de Previsión Social (IPS) de Coronel Oviedo, 2024 ..... 2

2 Marco Teórico ..... 3

2.1 Modelos de Clasificación ..... 3

2.1.1 Regresión Logística (RL) ..... 3

2.1.2 Arboles de Decisión (DT) ..... 4

2.1.3 Random Forest (RF) ..... 4

2.2 Métricas de evaluación ..... 5

2.2.1 Matriz de Confusión ..... 5

2.2.2 Exactitud o Accuracy ..... 6

2.2.3 Precisión (Presicion) ..... 6

2.2.4 Sensibilidad (Recall) ..... 6

2.2.5 F1 Score ..... 6

3 Objetivos ..... 7

3.1 Objetivo General ..... 7

3.2 Objetivos Específicos ..... 7

4 METODOLOGÍA ..... 8

4.1 Recolección y procesamiento de datos ..... 8

4.2 Análisis y Descripción de Datos. ..... 8

4.2.1 Presión Arterial Sistólica ..... 10

4.2.2 Presión Arterial Diastólica ..... 10

4.2.3 Colesterol ..... 10

4.2.4 Glucosa ..... 10

4.2.5 Hábitos de estilos de vida ..... 11
4.2.6 Índice de Masa Corporal ..... 11
4.3 Preprocesamiento de datos ..... 11
4.4 Organización de los datos ..... 18
4.5 Entrenamiento de modelos ..... 19
4.5.1 Entrenamiento inicial ..... 19
4.5.2 Entrenamiento con ajuste de hiperparametros - GridSearchCV ..... 20
4.6 Evaluación de modelos ..... 22
4.7 Comparativa de resultados ..... 23
5 RESULTADOS Y ANÁLISIS ..... 25
6 CONCLUSIONES Y RECOMENDACIONES ..... 27
7 BIBLIOGRAFIA ..... 29

# ÍNDICE DE FIGURAS

Gráfico 1- Historial de consultas entre 2021 al 2024 ..... 1
Gráfico 2 - Regresión Logística [4] ..... 3
Gráfico 3 - Árboles de Decisión ..... 4
Gráfico 4 - Random Forest ..... 5
Gráfico 5 - Matriz de Confusión ..... 5
Gráfico 6 - Rango máximo de la variable Edad ..... 13
Gráfico 7 - Gráfica de dispersión Presión Arterial Sistólica sin tratamiento de outliers ..... 14
Gráfico 8 - Gráfica de dispersión Presión Arterial Sistólica con tratamiento de outliers ..... 14
Gráfico 9 - Grafica de dispersión Presión Arterial Diastólica sin tratamiento de outliers ..... 15
Gráfico 10 - Gráfica de dispersión Presión Arterial Diastólica con tratamiento de outliers ..... 15
Gráfico 11 - Gráfica de dispersión de Pesos sin tratamiento de outliers ..... 16
Gráfico 12 - Gráfica de dispersión de Pesos con tratamiento de outliers ..... 16
Gráfico 13 - Gráfica de dispersión de Altura sin tratamiento de outliers ..... 17
Gráfico 14 - Gráfica de dispersión de Altura con tratamiento de outliers ..... 17

Gráfico 15 - Funcionamiento de validación cruzada [14] ..... 18
Gráfico 16 - Matriz de Confusión con ajustes de hiperparametros (RL) ..... 23
Gráfico 17 - Matriz de Confusión, sin ajustes de hiperparametros (RL) ..... 23
Gráfico 18 - Matriz de Confusión, sin ajustes de hiperparametros (DT) ..... 23
Gráfico 19 - Matriz de Confusión, con ajustes de hiperparametros (DT) ..... 23
Gráfico 20 - Matriz de Confusión, sin ajustes de hiperparametros (RF) ..... 24
Gráfico 21 - Matriz de Confusión, con ajustes de hiperparametros (RF) ..... 24
Gráfico 22 - Arquitectura de la Interfaz de predicción de eventos cardiovasculares ..... 26

# ÍNDICE DE TABLAS

Tabla 1 - Características obtenidas del dataset ..... 8
Tabla 2 - Caracteristicas con recomendaciones de la profesional en cardiologia ..... 9

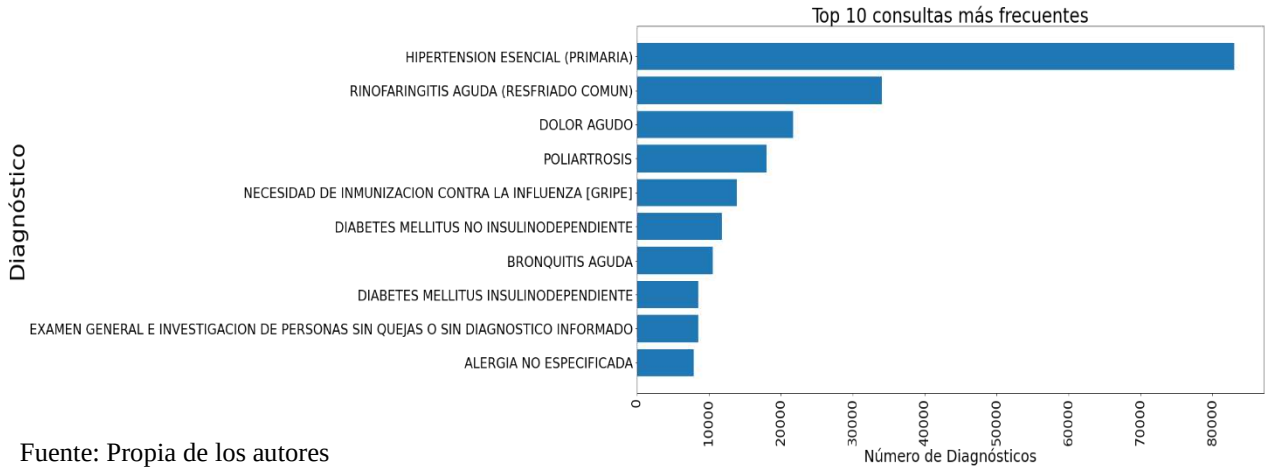
# Introducción

La enfermedad cardíaca o cardiopatía es un término general que incluye muchos tipos de problemas cardíacos. También se le llama enfermedad cardiovascular, es decir, enfermedad del corazón y de los vasos sanguíneos.

Las enfermedades cardiovasculares representan un gran porcentaje de muertes prematuras y el causante número uno de decesos en el mundo. Un estudio realizado por el Ministerio de Salud Pública y Bienestar Social reveló que aproximadamente 17.5 millones de personas murieron por enfermedades cardiacas, lo que representa el 30% de las muertes mundiales y 27% de las muertes nacionales. [1]

El Instituto de Previsión Social de la ciudad de Coronel Oviedo cuenta con un banco de datos de todas las consultas realizadas diariamente. Estos registros históricos fueron proporcionados por el Director de dicha Institución, el Dr. Denis Figueredo, de entre el 2021 al 2024, y el análisis nos reveló un significativo número de consultas relacionadas con uno de los factores esenciales considerados para la detección de enfermedades cardiovasculares (hipertensión esencial) dándonos así una cantidad de 82.991 registros de pacientes.

Gráfico 1- Historial de consultas entre 2021 al 2024




Fuente: Propia de los autores

Esta alarmante tendencia no solo plantea un desafío para la salud pública, sino que también demuestra la necesidad de implementar estrategias de prevención. Para reducir significativamente la mortalidad asociada a estas enfermedades y mejorar la salud general de la población, es fundamental la detección temprana de los factores de riesgo y la detección de posibles eventos cardiovasculares.

Al día de hoy, disponemos de tecnologías necesarias para poder aplicar técnicas que puedan

colaborar con los profesionales de la salud. Este trabajo propone la realización de un modelo de predicción para alerta temprana de eventos cardiovasculares mediante técnicas de Machine Learning Supervised, que también contará con una base de datos que permitirá almacenar los datos del paciente y de las consultas realizadas, como también generar un historial médico, ya que el área de Cardiología del Instituto de Previsión Social de Coronel Oviedo no cuenta con registros de historiales clínicos de los pacientes.

# 1 Modelo de Predicción para alerta temprana de Eventos Cardiovasculares mediante técnicas de Supervised Machine Learning en el Instituto de Previsión Social (IPS) de Coronel Oviedo, 2024

Para llevar a cabo este trabajo, realizamos un análisis profundo de un conjunto de datos y factores que pueden contribuir a diferentes tipos de eventos cardiovasculares. Los datos se obtuvieron de la plataforma Kaggle con el objetivo de entrenar varios modelos de machine learning y optar por el de mejor resultado al momento de predecir eventos cardiovasculares.

El Machine Learning (ML), es un subconjunto de la inteligencia artificial (IA) que se centra en el desarrollo de algoritmos informáticos que pueden mejorarse automáticamente a través de la experiencia y el uso de datos, lo que permite a los ordenadores aprender de los datos y tomar decisiones o hacer predicciones sin ser programadas explícitamente para cada tarea, mejorando su rendimiento con el tiempo, haciéndose más precisos y eficaces a medida que procesan más datos. [2]

El conjunto de datos obtenidos de la comunidad Kaggle [3] contiene datos de pacientes, y factores a considerar a la hora de predecir eventos cardiovasculares. Estos datos se listan a continuación:

- Edad: El riesgo de sufrir enfermedades cardiovasculares aumenta con la edad.
- Estatura y peso.
- Género: Ciertos factores pueden afectar el riesgo de enfermedades del corazón de manera diferente en las mujeres que en hombres.
- Hábitos de estilo de vida: Con el tiempo, los hábitos de estilo de vida poco saludables pueden aumentar su riesgo de enfermedades cardiovasculares.
- Ejercicio físico insuficiente.
- Beber demasiado alcohol.
- Fumar y exponerse al humo de segunda mano.

Tener otras afecciones médicas puede aumentar el riesgo de enfermedades cardiovasculares. Estos problemas incluyen:

- Presión arterial alta.
- Niveles de colesterol altos.
- Diabetes.
- Obesidad.

# 2 Marco Teórico

## 2.1 Modelos de Clasificación

Para la realización de este trabajo exploramos diversas opciones de modelos de aprendizaje supervisado, lo que resultó en la selección de tres modelos diferentes, entrenando cada modelo en nuestro conjunto de datos, con el objetivo de encontrar el más adecuado y con los mejores resultados.

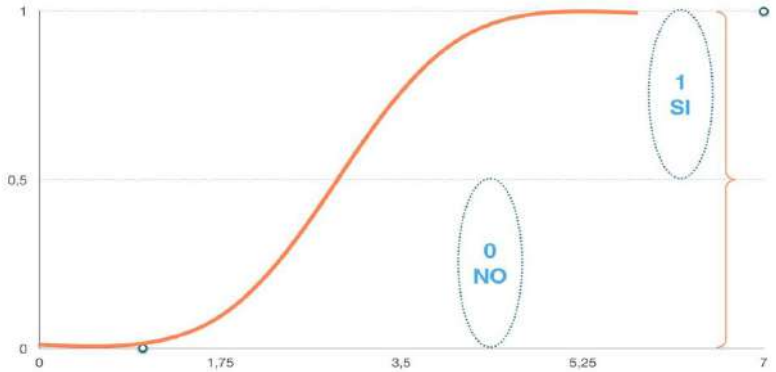
### 2.1.1 Regresión Logística (RL)

La Regresión Logística se basa en una función denominada sigmoide. Esta función tiene la forma de una curva en S y puede tomar cualquier número y lo transforma entre 1 y 0.

Estima la probabilidad de que ocurra un evento, calculada en un conjunto de datos determinados de variables independientes, encontrando así una ecuación que mejor estime las probabilidades. En términos más sencillos, se busca encontrar una fórmula que nos permita estimar la probabilidad de un evento en función del conjunto de datos que se posee.

En medicina, este tipo de análisis se puede utilizar para predecir la probabilidad de determinadas enfermedades en determinados grupos de personas. [4]

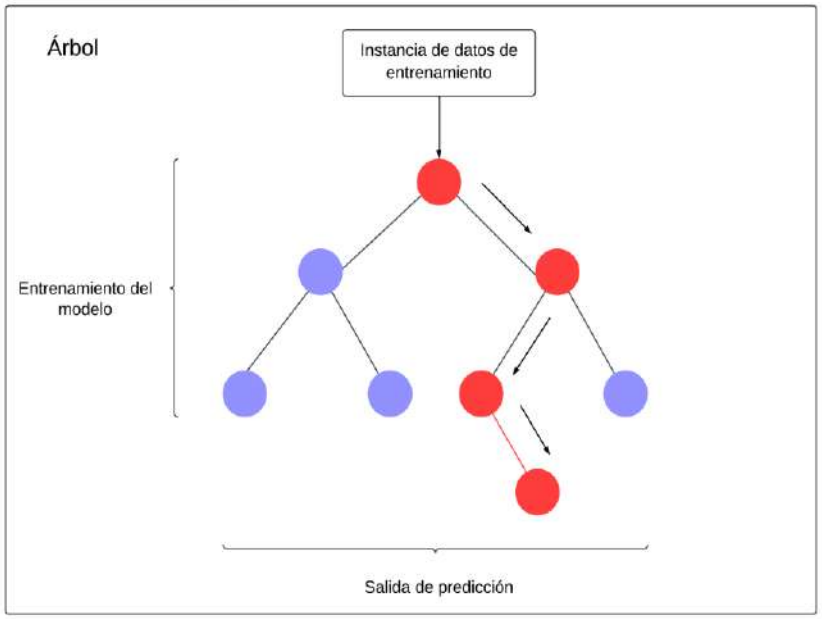
Gráfico 2 - Regresión Logística [4]



### 2.1.2 Arboles de Decisión (DT)

Son una herramienta de aprendizaje supervisado no paramétrico que es utilizado para la clasificación y la regresión. El objetivo es construir un modelo que prediga el valor de una variable objetivo mediante la construcción y el aprendizaje de reglas de decisión simples derivadas de las características de los datos. Un árbol puede considerarse como una aproximación constante por partes. [5]

Gráfico 3 - Árboles de Decisión

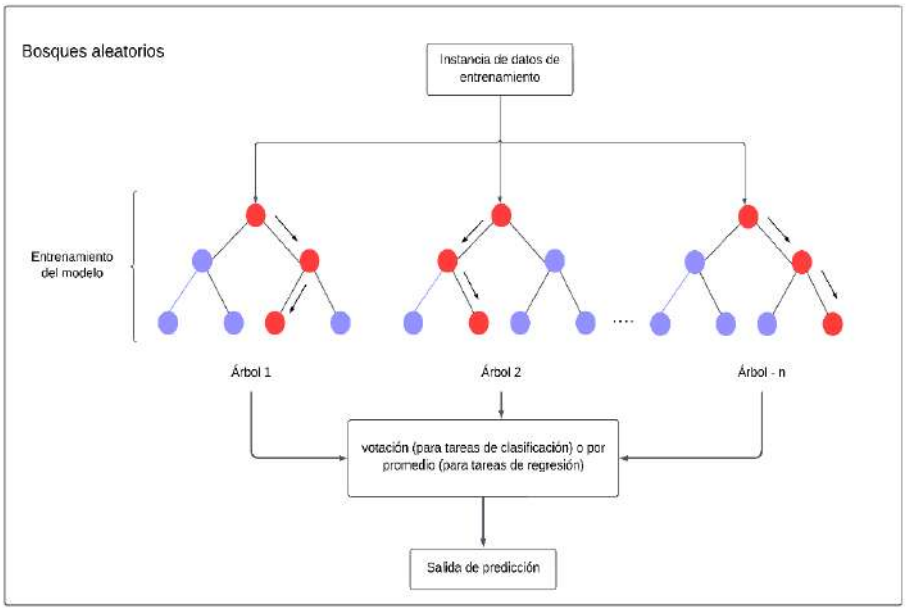



Fuente: Propia de los autores

### 2.1.3 Random Forest (RF)

Un bosque aleatorio es un metaestimador que funciona generando una serie de clasificadores de árboles de decisión durante la fase de entrenamiento. Cada árbol se construye utilizando un subconjunto del conjunto de datos y utiliza el promedio para mejorar la precisión de la predicción y controlar el sobreajuste. Este algoritmo combina los resultados de los árboles y puede proporcionar resultados sólidos y precisos mediante votación (para las tareas de clasificación) o promediación (para tareas de regresión). [6]

Gráfico 4 - Random Forest




Fuente: Propia de los autores

## 2.2 Métricas de evaluación

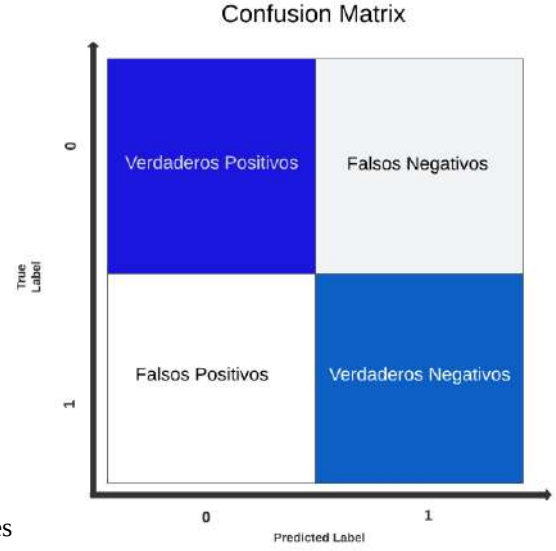
Para la evaluación de los modelos fueron utilizadas las siguientes métricas:

### 2.2.1 Matriz de Confusión

Una vez que hemos entrenado un modelo, debemos analizar los resultados obtenidos, para ello se utilizó la matriz de confusión.

La matriz de confusión es una herramienta fundamental que permite visualizar el desempeño de un modelo clasificador y obtener predicciones sobre los datos. En el caso de que un problema sea binario, nuestro modelo lo clasificara el conjunto de datos como 0 o 1. [7]

Gráfico 5 - Matriz de Confusión




Fuente: Propia de los autores

Observamos que los valores de la diagonal principal corresponden a los valores considerados de forma correcta por el modelo, tanto como los Verdaderos Positivos (VP), como los Verdaderos Negativos (VN). Por otro lado, la diagonal secundaria muestra los casos en los que el modelo "se ha equivocado" siendo así Falsos Positivos (FP) y Falsos Negativos (FN).

### 2.2.2 Exactitud o Accuracy

Cuando nuestras clases son aproximadamente iguales en tamaño, podemos utilizar esta métrica, ya que esto nos dará valores clasificados correctamente. Se define como el número de predicciones correctas hechas como una proporción de todas las predicciones hechas. [8]

$$\text{Exactitud} = \frac{(\text{Verdaderos Positivos} + \text{Verdaderos Negativos})}{\text{Total de Muestras}}$$

### 2.2.3 Precisión (Presicion)

Nos indica que proporción de las predicciones positivas del modelo son realmente positivas. Se define como el número de verdaderos positivos dividido por el número total de todos los positivos. [7]

$$\text{Presición} = \frac{\text{Verdadero Positivo}}{(\text{Verdadero Positivo} + \text{Falso Positivo})}$$

### 2.2.4 Sensibilidad (Recall)

Este modelo es utilizado para determinar cuántos valores positivos fueron clasificados correctamente. Se define como el número de verdaderos positivos dividido por la suma de los números de verdaderos positivos y falsos negativos. [7]

$$\text{Sensibilidad} = \frac{\text{Verdaderos Positivos}}{(\text{Verdaderos Positivos} + \text{Falsos Negativos})}$$

### 2.2.5 F1 Score

Es una métrica versátil y muy útil que puede ayudarnos a analizar conjuntos de datos que se encuentran desbalanceados. Combina de manera equilibrada la precisión y el recall al calcular la media armónica de ambas métricas y esto nos indica qué tan bien está funcionando el modelo. [7]

$$\text{F1 Score} = 2 \times \frac{\text{Presición} \times \text{Recall}}{\text{Presición} + \text{Recall}}$$

# 3 Objetivos

## 3.1 Objetivo General

Desarrollar un modelo de predicción para alerta temprana de eventos cardiovasculares mediante técnicas de supervised machine learning en el Instituto de Previsión Social (IPS) de Coronel Oviedo, 2024.

## 3.2 Objetivos Específicos

- Analizar datos clínicos con factores de riesgo para eventos cardiovasculares en el Instituto de Previsión Social (IPS) de Coronel Oviedo.
- Entrenar diferentes modelos de machine learning para la predicción del riesgo de eventos cardiovasculares, utilizando técnicas de aprendizaje automático.
- Evaluar el rendimiento de los modelos de machine learning mediante métricas de desempeño adecuadas y seleccionar el modelo con mejor exactitud.

# 4 METODOLOGÍA

## 4.1 Recolección y procesamiento de datos

Para obtener los resultados de este trabajo, realizamos un análisis de un dataset (conjunto de datos) obtenidos a través de la plataforma Kaggle [9], una plataforma de Data Science que permite a los usuarios extraer y publicar conjunto de datos. Este dataset consta con 70.000 registro de pacientes, 11 características + objetivo, lo que lo convierte en el conjunto de datos más grande sobre eventos cardiovasculares disponible hasta el momento para fines de investigación.

Las características se listan a continuación:

| **Variable** | **Campo** | **Característica** | **Tipo de Dato** |
|---|---|---|---|
| Edad | Edad | Objetiva | Int (días) |
| Altura | Altura | Objetiva | Entero (cm) |
| Peso | Peso | Objetiva | Flotante (kg) |
| Genero | Genero | Objetiva | Código categórico (1: hombre, 2: Mujer) |
| Presión arterial sistólica | PAS | Función de examen | Entero |
| Presión arterial diastólica | PAD | Función de examen | Entero |
| Colesterol | colesterol | Función de examen | Código categórico (1: normal, 2: por encima de lo normal, 3: muy por encima de lo normal) |
| Glucosa | Glucosa | Función de examen | Código categórico (1: normal, 2: por encima de lo normal, 3: muy por encima de lo normal) |
| Fumar | Fumar | Subjetiva | Binario (1: fumador, 0: no fumador) |
| Consumo de alcohol | Alcohol | Subjetiva | Binario (1: activo, 0: no inactivo) |
| Actividad física | Actividad_Física | Subjetiva | Binario (1: activo, 0: inactivo) |

Tabla 1 - Características obtenidas del dataset

## 4.2 Análisis y Descripción de Datos.

Hay tres tipos de funciones de entrada:

- Objetivo: Información fática.
- Examen: Resultados del examen médico.
- Subjetivo: Información proporcionada por el paciente.

La función de entrada “objetivo” hace referencia a información concreta y verídica del paciente.

La función de entrada "examen" está centrada en los resultados que se obtuvieron a partir de análisis clínicos.

La función de entrada "subjetivo" se obtuvo directamente del paciente.

En una conversación con la doctora especialista en cardiología del Instituto de Previsión Social de Coronel Oviedo, Natalia Zorrilla, identificamos los factores más relevantes para nuestro estudio, y también nos sugirió incluir el Índice de Masa Corporal (IMC) ya que es un factor esencial que debía incluirse a nuestro análisis.

El IMC es una característica crucial que juega un papel muy importante en lo que respecta a eventos cardiovasculares.

Siguiendo las recomendaciones de la profesional, la tabla queda de la siguiente manera:

| **Variable** | **Campo** | **Característica** | **Tipo de Dato** |
|---|---|---|---|
| Edad | Edad | Objetiva | Int (días) |
| Altura | Altura | Objetiva | Entero (cm) |
| Peso | Peso | Objetiva | Flotante (kg) |
| Genero | Genero | Objetiva | Código categórico (1: hombre, 2: Mujer) |
| Presión arterial sistólica | PAS | Función de examen | Entero |
| Presión arterial diastólica | PAD | Función de examen | Entero |
| Colesterol | colesterol | Características del examen | Código categórico (1: normal, 2: por encima de lo normal, 3: muy por encima de lo normal) |
| Glucosa | Glucosa | Características del examen | Código categórico (1: normal, 2: por encima de lo normal, 3: muy por encima de lo normal) |
| Fumar | Fumar | Subjetiva | Binario (1: fumador, 0: no fumador) |
| Consumo de alcohol | Alcohol | Subjetiva | Binario (1: activo, 0: no inactivo) |
| Actividad física | Actividad_Física | Subjetiva | Binario (1: activo, 0: inactivo) |
| Índice de Masa Corporal | IMC | Objetiva | Código categórico (1: bajo peso, 2: normal, 3: sobrepeso, 4: obesidad) |

Tabla 2 - Características con recomendaciones de la profesional en cardiología

Teniendo en cuenta estos factores a partir de lo recomendado por la especialista en cardiología, quién también nos brindó información de ciertos rangos que se debe tener en cuenta para detectar

algunas afecciones considerados factores importantes.

Consultando también con [10] [11], pudimos observar que son rangos estándares, utilizados a nivel mundial. Por ende, se tuvo en cuenta los siguientes rangos:

### 4.2.1 Presión Arterial Sistólica

Se refiere a la presión que ejerce el corazón cuando este se contrae y bombea sangre. Si la presión es de 120/80 mmHg, el 120 representa a la presión sistólica y se mide teniendo en cuenta el siguiente rango:

- Normal: 120 a 139.
- Medio: 140 a 159.
- Alto: 160 para adelante.

### 4.2.2 Presión Arterial Diastólica

Se refiere a la presión mínima del corazón cuando este está en reposo entre latidos. Si la presión es de 120/80 mmHg, el 80 representa a la presión diastólica y se mide teniendo en cuenta el siguiente rango:

- Normal: 80 a 89.
- Medio: 90 a 99.
- Alto: 100 para adelante.

### 4.2.3 Colesterol

Nos indica sobre las partículas de grasa que transitan por todo nuestro organismo, se encuentra presente en todas las células del cuerpo humano y es necesaria para el funcionamiento normal del organismo.

Se tuvo en cuenta el colesterol total, que sería la suma del colesterol LDL y el colesterol HDL y se mide teniendo en cuenta el siguiente rango:

- Normal: menos de 200 mg/dl.
- Por encima de lo normal: entre 200 y 240 mg/dl.
- Muy por encima de lo normal: por encima de 240 mg/dl.

### 4.2.4 Glucosa

Se refiere al azúcar que es esencial para nuestro cuerpo, que es una fuente importante para nuestras células y se mide teniendo en cuenta el siguiente rango.

- Normal: entre 70 y 99 mg/dl.
- Por encima de lo normal: entre 100 y 125 mg/dl.
- Muy por encima de lo normal: por encima de 125 mg/dl.

### 4.2.5 Hábitos de estilos de vida

Existen hábitos de estilos de vida poco saludables que son factores fundamentales a la hora de detectar un evento cardiovascular, como son: fumar, beber y la falta de actividad física.

De acuerdo con lo mencionado por la especialista en cardiología, una persona es considerada fumadora cuando consume como mínimo 3 veces por semana, una persona es considerada bebedora cuando consume alcohol como mínimo 3 veces por semana y una persona es considera activa físicamente cundo practica ejercicios como mínimo 3 veces por semana.

### 4.2.6 Índice de Masa Corporal

Según [12], el índice de masa corporal (IMC) es el peso de una persona en kilogramos dividido por el cuadrado de la estatura en metros y se mide teniendo en cuenta el siguiente rango:

- Bajo peso: por debajo de 18.5.
- Normal: entre 18.5 y 24.9.
- Sobrepeso: entre 25.0 y 29.9.
- Obesidad: 30.0 o más.

El IMC se calcula de la misma manera tanto para adultos como para niños. El cálculo se basa en la siguiente fórmula:

$$\mathrm{IMC} = \frac{\mathrm{P}}{(\mathrm{E})^2}$$

P: Peso (kg).
E: Estatura (m).

La característica que nos interesa predecir es la denominada Target_Variable_Cardio que nos indica si es probable que el paciente pueda padecer algún evento cardiovascular. Presencia de evento cardiovascular = 1 y Ausencia de evento cardiovascular = 0.

## 4.3 Preprocesamiento de datos

En esta etapa, Se llevó a cabo el preprocesamiento de los datos mediante la librería pandas, orientado principalmente en la manipulación y análisis de datos. Se han aplicado diversas técnicas

que se consideró crucial para el entrenamiento de los datos. A continuación, se detallan específicamente las técnicas aplicadas:

**a) Instancias Faltantes:**

Se comprobó si en el conjunto de datos de estudio existen campos faltantes, considerar este aspecto es importante ya que puede generarse errores en los resultados si no se trata correctamente.
Se aplicó la siguiente función de pandas para la detección de valores faltantes:

$$\text{pandas.dtf.isna()}$$ [13]

Este método ha retornado FALSE para cada campo de todas las columnas del conjunto de datos indicando así la inexistencia de campos vacíos.

**b) Conversión edad-días:**

Originalmente la variable Edad se expresaba en formato de días, por lo tanto, se realizó una conversión de edad a días para un mejor manejo estándar de las edades que equivale a años.
Para ello se aplicó la siguiente conversión de datos:

$$dtf['Edad'] \,/\, 365).round().astype('int')$$

Se toma cada valor de la columna edad en donde es dividida por 365 haciendo referencia a la cantidad de días anualmente, se procede luego a realizar el redondeo al número entero más cercano y con el método astype convierte los resultados redondeados a enteros.

**c) Establecimiento del rango etario:**

De acuerdo con la especialista en cardiología, aquellas personas que manifiestan algún problema de corazón antes de los 30-40 años, es porque probablemente ya hayan nacido con algún problema congénito del corazón. Por lo tanto, se ha delimitado el rango de edades aplicando una función que ofrece pandas que se muestra a continuación:

$$dtf['Edad'].between(40, 65)$$

El estudio estadístico de dicha variable nos ha revelado que el máximo valor de estudio se encuentra por los 65 años de edad:

Gráfico 6 - Rango máximo de la variable Edad

| | id | Edad |
|---|---|---|
| count | 69587.000000 | 69587.000000 |
| mean | 49977.387745 | 53.424303 |
| std | 28849.687875 | 6.692816 |
| min | 0.000000 | 40.000000 |
| 25% | 25008.500000 | 49.000000 |
| 50% | 50016.000000 | 54.000000 |
| 75% | 74890.500000 | 58.000000 |
| max | 99999.000000 | 65.000000 |

Fuente: Propia de los autores

Por lo tanto, el rango superior seleccionado fue de 65 años de edad.

**d) Ingeniería de atributos:**

En la Tabla N° 2, se presenta una nueva variable creada a partir de la aplicación de una ingeniería de atributos a la variable altura y peso.

**e) Detección y tratamiento de outliers o valores atípicos:**

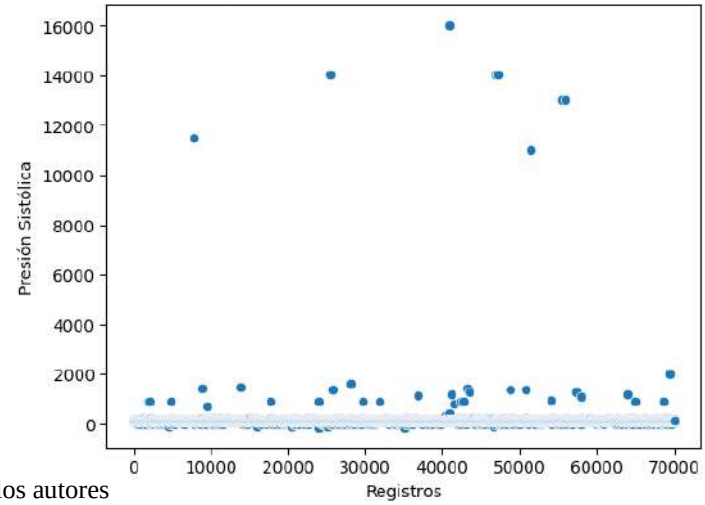
En la etapa de preprocesamiento de datos se han identificado valores atípicos o anormales en algunas variables de estudio. Para la identificación de outliers se ha utilizado gráficas de dispersión en donde se ha observado una importante cantidad de valores que pueden sesgar el rendimiento de los modelos. Estos gráficos han permitido además de visualizar, poder realizar una limpieza de valores atípicos.

A continuación, se muestra las variables que se han realizado una limpieza con el objetivo de mejorar la calidad de los datos:

- **Presión Arterial Sistólica (PAS)**

En el gráfico de dispersión se identificó que existen valores demasiado alejados del conjunto en donde se concentra la mayoría de los datos.

Gráfico 7 - Gráfica de dispersión Presión Arterial Sistólica sin tratamiento de outliers



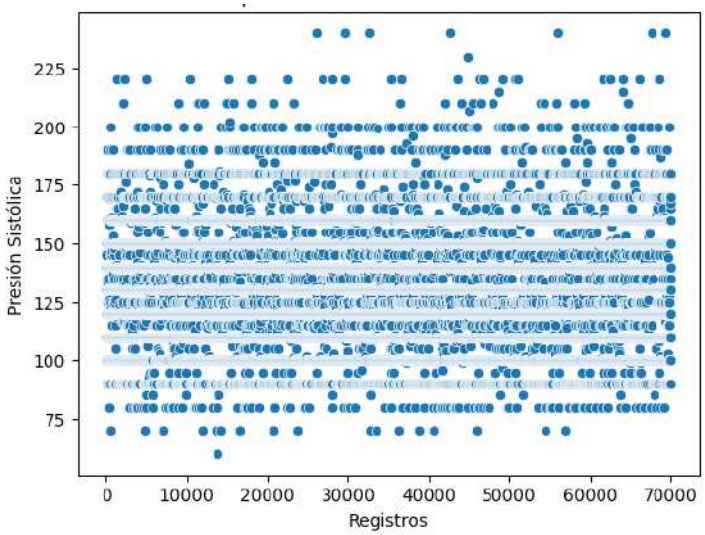

Fuente: Propia de los autores

Con el objetivo de abarcar las posibilidades de valores que puede tomar la presión arterial sistólica se estableció el rango de entre de 50 a 250 mmHg, expresado en la siguiente ecuación:

$$dtf[(df\_tratamiento['PAS'] >= 50) \,\&\, (df\_tratamiento['PAS'] <= 250)]$$

En la siguiente figura de dispersión se aprecia como el conjunto de datos se concentra de tal manera que ya no existen valores atípicos extremos:

Gráfico 8 - Gráfica de dispersión Presión Arterial Sistólica con tratamiento de outliers

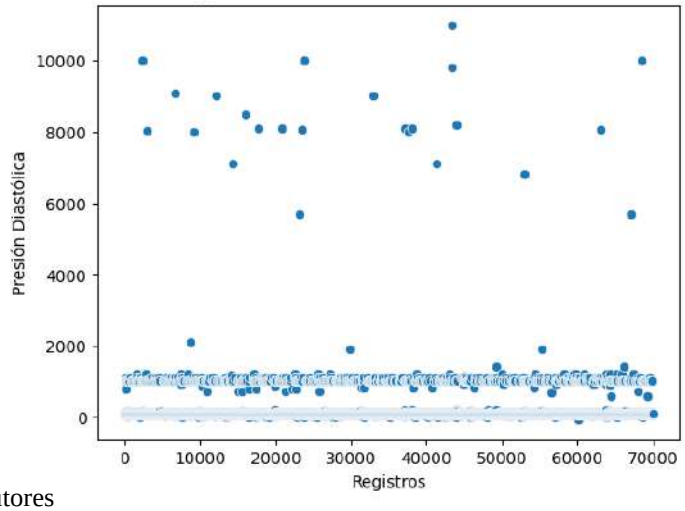



Fuente: Propia de los autores

- **Presión Arterial Diastólica (PAD)**

En el gráfico de dispersión se identificó valores atípicos de presión diastólica a partir de 2000 mmHg y por encima del mismo, los cuales se encuentran muy alejados del conjunto en donde se concentra la mayoría de los datos.

Gráfico 9 - Grafica de dispersión Presión Arterial Diastólica sin tratamiento de outliers



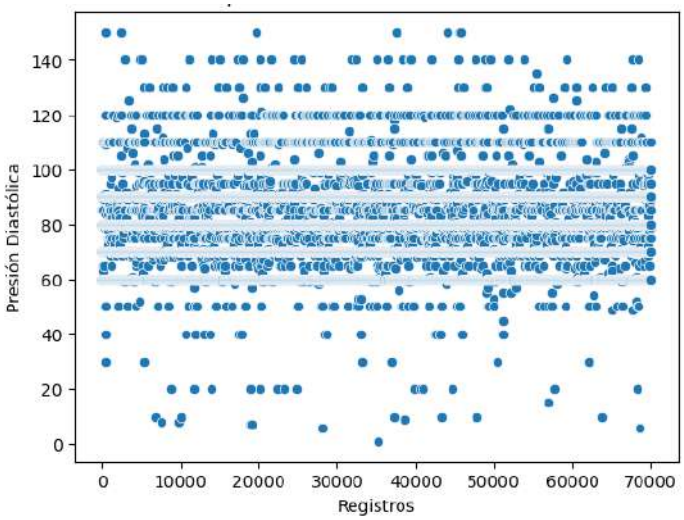

Fuente: Propia de los autores

Con el objetivo de abarcar las posibilidades de valores que puede tomar la presión arterial diastólica se estableció el rango de entre de 1 a 150 mmHg, expresado en la siguiente ecuación:

$$dtf[(df\_tratamiento['PAD'] >= 1) \,\&\, (df\_tratamiento['PAD'] <= 150)]$$

En la siguiente figura de dispersión se aprecia como el conjunto de datos se concentra de tal manera que ya no existen valores atípicos extremos:

Gráfico 10 - Gráfica de dispersión Presión Arterial Diastólica con tratamiento de outliers

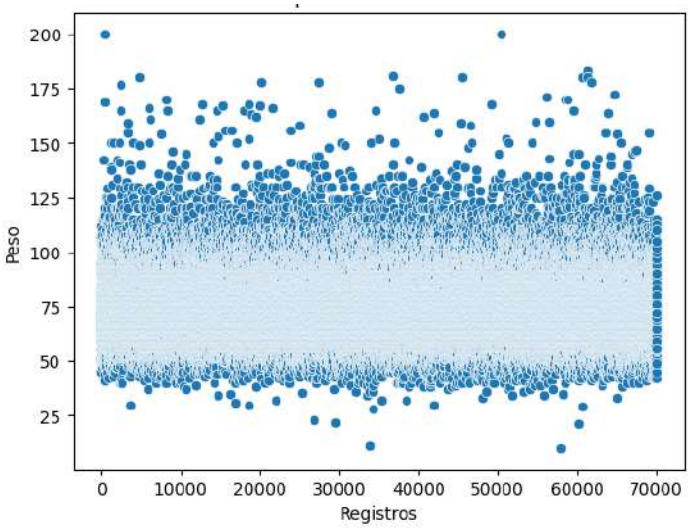



Fuente: Propia de los autores

- **Peso**

La siguiente figura representa la gráfica de dispersión de la variable peso. Se ha percibido una gran cantidad de valores que se alejan de lo normal.

Gráfico 11 - Gráfica de dispersión de Pesos sin tratamiento de outliers



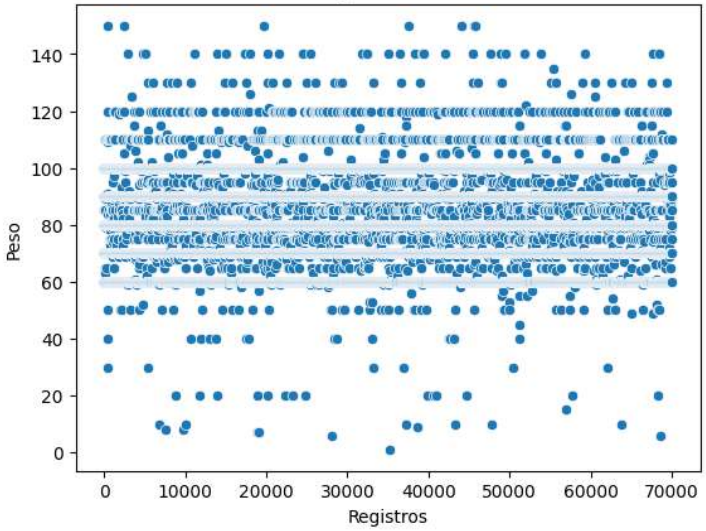

Fuente: Propia de los autores

Con el objetivo de abarcar las posibilidades de valores que puede tomar la variable peso se estableció el rango de entre de 50 a 200 Kg, expresado en la siguiente ecuación:

$$dtf[(df\_tratamiento['Peso'] >= 50) \,\&\, (df\_tratamiento['Peso'] <= 200)]$$

En la siguiente figura de dispersión se aprecia como el conjunto de datos se concentra de tal manera que ya se reduce significativamente los valores atípicos extremos:

Gráfico 12 - Gráfica de dispersión de Pesos con tratamiento de outliers

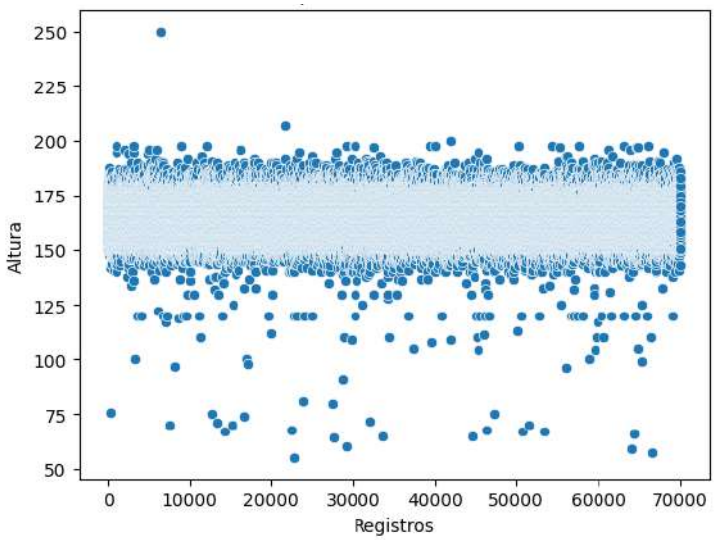



Fuente: Propia de los autores

- **Altura**

La siguiente figura representa la gráfica de dispersión de la variable altura. Se ha percibido una gran cantidad de valores que se alejan de lo normal.

Gráfico 13 - Gráfica de dispersión de Altura sin tratamiento de outliers



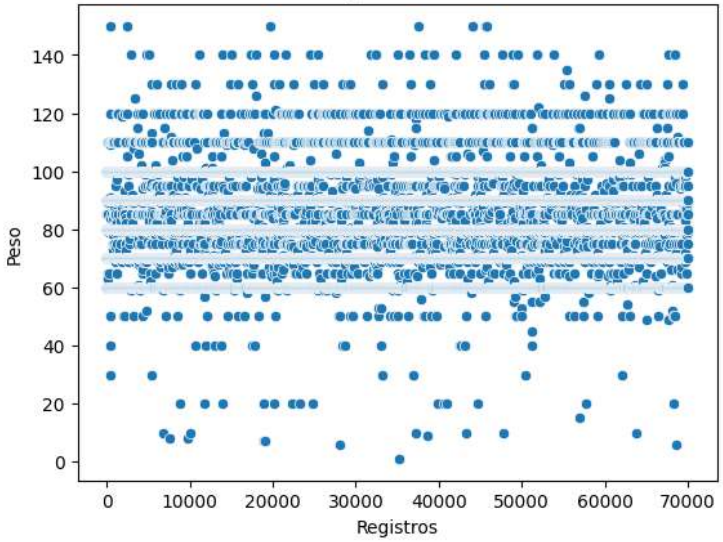

Fuente: Propia de los autores

Con el objetivo de minimizar el impacto de valores atípicos, se procedió a establecer la altura dentro del rango entre 100 a 200 cm, expresado en la siguiente ecuación:

$$dtf[(df\_tratamiento['Altura'] >= 100) \,\&\, (df\_tratamiento['Altura'] <= 200)]$$

La siguiente figura de dispersión revela como dentro de conjunto de datos se reduce significativamente los valores atípicos extremos:

Gráfico 14 - Gráfica de dispersión de Altura con tratamiento de outliers




Fuente: Propia de los autores

## 4.4 Organización de los datos

Para el proceso de organización de datos, se dividió el dataset en dos conjuntos. La primera división de datos se tomó un 80% para el conjunto de entrenamiento y el 20% para el conjunto de prueba.

| Conjunto de entrenamiento 80% | Conjunto de prueba 20% |
|---|---|

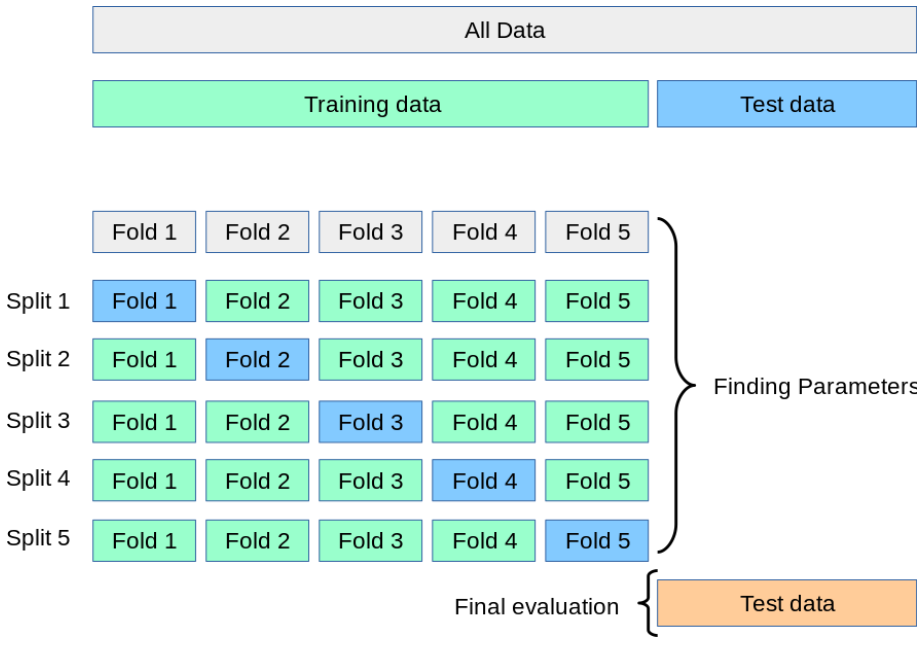
El conjunto total de datos final consta de 67.386 registros, El 80% de los datos corresponden a la cantidad de 53.908 registros y el 20% corresponde a una cantidad de 13.477 registros.

Con el fin de obtener una evaluación sólida de los modelos entrenados con ajuste de hiperparametros y evitar el sobreajuste, se ha empleado la técnica de validación cruzada utilizando la función de GridSearchCV. La validación cruzada consiste en dividir los datos en k subconjuntos iguales llamados folds, se selecciona el primer fold como conjunto de prueba y los restantes, es decir, k-1 como conjunto de entrenamiento y el modelo que se entrenó se evalúa en el conjunto de prueba y este proceso se repite k veces. [14]

Para el proceso de entrenamiento se tomó como valor 5 para el parámetro cv (cross-validation) esto implica una cantidad de 5 entrenamientos.

A continuación, se muestra la figura del funcionamiento de la validación cruzada:

Gráfico 15 - Funcionamiento de validación cruzada [14]



## 4.5 Entrenamiento de modelos

Se realizó el entrenamiento de los modelos utilizando la librería de scikit-learn teniendo en cuenta los parámetros por defectos del estimador. Posteriormente, se aplicó la técnica de GridSearchCV en búsqueda de optimizar el modelo mediante el ajuste hiperparametros con el fin de mejorar el desempeño de los modelos.

### 4.5.1 Entrenamiento inicial

**Regresión Logística (RL)**

Para el entrenamiento del modelo de regresión logística se creó el estimador y se utilizó los parámetros por defecto de la función. A continuación, se procedió a realizar el ajuste del modelo.

```
from sklearn.linear_model import LogisticRegression


regr_model = LogisticRegression()
regr_model.fit(X_train, y_train)
```

**Arboles de Decisión (DT)**

Se instanció un estimador de árboles de decisión utilizando los parámetros preestablecidos de la función y se ajustó el modelo a los datos de entrenamiento.

```
[ ] from sklearn.tree import DecisionTreeClassifier


    clf = DecisionTreeClassifier()
    clf.fit(X_train, y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier()
```

**Random Forest (RF)**

Se instanció un estimador de bosque aleatorio básico y se realizó el entrenamiento con parámetros preestablecidos dentro del modelo.

```
[ ]  from sklearn.ensemble import RandomForestClassifier

     # Entrenamiento del modelo con Random Forest
     random_forest_model=RandomForestClassifier()
     random_forest_model.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier()
```

### 4.5.2 Entrenamiento con ajuste de hiperparametros - GridSearchCV

**Regresión Logística (RL)**

Se creó un diccionario de datos en donde se definen los hiperparametros que se utilizaron para el ajuste del modelo y también se aplicó la técnica de validación cruzada.

```
from sklearn.model_selection import  GridSearchCV

# Crear un diccionario con los hiperparámetros y sus posibles valores
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear', 'saga']
}

# Crear el modelo base
logistic_regression_gs = LogisticRegression()

# Crear el objeto GridSearchCV
grid_search_regr = GridSearchCV(estimator=logistic_regression_gs,
                                param_grid=param_grid, cv=5)

# Ajustar el modelo
grid_search_regr.fit(X_train, y_train)

# Obtener los mejores parámetros
best_params = grid_search_regr.best_params_
print(best_params)
```

Luego del entrenamiento del modelo se obtuvo los siguientes hiperparamteros: {'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}.

- 1: sugiere un equilibro entre ajustar los datos evitando el sobreajuste
- L2: reduce la complejidad del modelo

- Liblinear: Determina el algoritmo de optimización que se utilizará

**Arboles de Decisión (DT)**

Se creó un diccionario de datos, se definió un conjunto de hiperparámetros para el ajuste del modelo, luego se creó un estimador o instancia de modelo para aplicar la técnica GridSearchCV y se seleccionaron 5 folds para su posterior entrenamiento.

```
import pandas as pd
from sklearn.model_selection import  GridSearchCV



# Definir la cuadrícula de parámetros para GridSearchCV
param_grid = {
    'max_depth': range(2, 10),
    'min_samples_split': range(2, 20),
    'criterion': ['gini', 'entropy']
}

# Creando el estimador
clf_gsearch = DecisionTreeClassifier()

# Realizando GridSearchCV para encontrar los mejores hiperparámetros
# 5-fold Validacion Cruzada
grid_search_arbol_decision = GridSearchCV(clf_gsearch,
                                          param_grid, cv=5)

grid_search_arbol_decision.fit(X_train, y_train)
```

Al finalizar el proceso de entrenamiento del modelo, se obtiene la siguiente combinación óptima de hiperparámetros: {'criterion': 'entropy', 'max_depth': 5, 'min_samples_split': 2}.

- 'entropy': Encuentra la distribución que maximice la pureza del nodo.
- 5: Profundidad máxima del árbol.
- 2: Mínimo de muestras requeridas para dividir un nodo interno.

En el entrenamiento se ha aplicado la combinación más óptima para el entrenamiento, maximizando la pureza del nodo y minimizando la entropía, una profundidad máxima de 5 árboles y una cantidad mínima de 2 muestras para la división de un nodo.

**Random Forest (RF)**

Se creó un diccionario de datos que definió un conjunto de hiperparámetros para el ajuste del modelo y también se aplicó la técnica de validación cruzada para el entrenamiento de los modelos a fin de tener un óptimo desempeño.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV

random_forest_model_gs=RandomForestClassifier()

parametros_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [5, 10, 15],
    'min_samples_split': [2, 5, 10]
}


# Crear y ajustar GridSearchCV
rf_grid_search = GridSearchCV(estimator=random_forest_model_gs,
                              param_grid=parametros_grid, cv=5)

#Ajustamos: el objeto GridSearchCV en los datos
rf_grid_search.fit(X_train, y_train)
```

Al finalizar el proceso de entrenamiento del modelo, se obtiene la siguiente combinación óptima de hiperparámetros: {'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 100}

- 10: profundidad máxima de cada árbol de decisión
- 2: mínimo de muestras requeridas para dividir un nodo interno
- 100: Número total de árboles de decisión en el bosque

En una búsqueda exhaustiva de las mejores combinaciones de hiperparamteros se ha aplicado la siguiente configuración para el entrenamiento del modelo: una profundidad máxima de árbol de 10 niveles, un mínimo de 2 muestras para la división de nodos y un total de 100 árboles en el bosque.

## 4.6 Evaluación de modelos

Para evaluar el rendimiento de los modelos de regresión logística, árboles de decisión y bosques aleatorios, creamos una tabla de comparación para el entrenamiento inicial y el ajuste de hiperparámetros utilizando métricas de evaluación: acurracy, recall, precision y f1-score. Las matrices de confusión han permitido visualizar cuantas observaciones han sido correctamente clasificadas como presencia de eventos cardiovasculares (verdaderos positivos), cuantos fueron clasificados erróneamente como presencia de eventos cardiovasculares (falsos positivos), cuantos han sido clasificados correctamente como ausencia de eventos cardiovasculares (verdaderos negativos) y cuantos fueron clasificados erróneamente como ausencia de eventos cardiovasculares (falsos negativos).

## 4.7 Comparativa de resultados

### Regresión Logística (RL)

Gráfico 17 - Matriz de Confusión, sin ajustes de hiperparametros (RL)

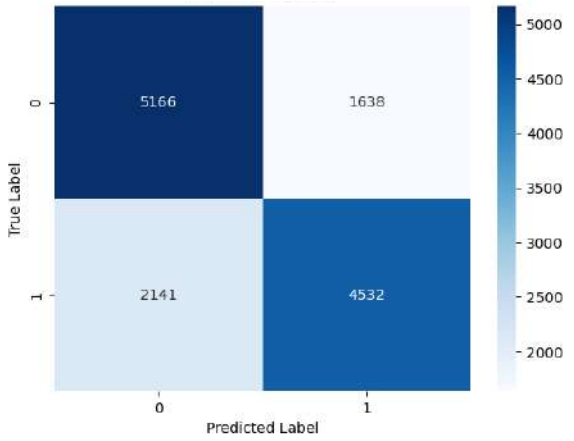
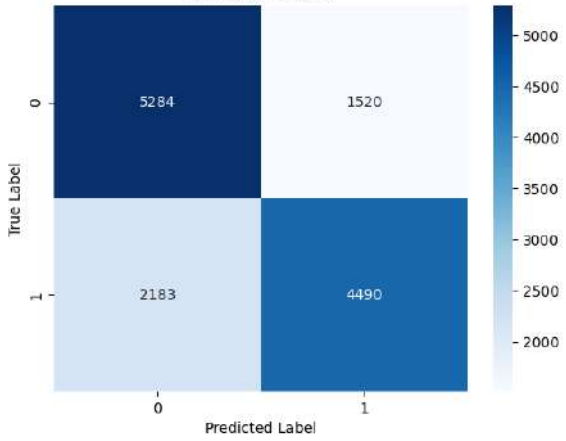



Gráfico 16 - Matriz de Confusión con ajustes de hiperparametros (RL)




Fuente: Propia de los autores

En el entrenamiento inicial, el modelo de regresión logística logró una exactitud 72%, con la matriz de confusión que representa el número de Verdaderos Positivos (VP): 5166, Verdaderos Negativos (VN): 4532, Falsos Positivos (FP): 2141 y Falsos Negativos (FN): 1638.

Después de la evaluación mediante el ajuste de hiperparámetros, se logró una exactitud del 73% y se obtuvieron mejores resultados en términos de clasificación de las observaciones: Verdaderos Positivos (TP): 5284, Verdaderos Negativos (VN): 4490, Falsos Positivos (FP): 2183 y Falsos Negativos (FN): 1520.

### Arboles de Decisión (DT)

Gráfico 18 - Matriz de Confusión, sin ajustes de hiperparametros (DT)

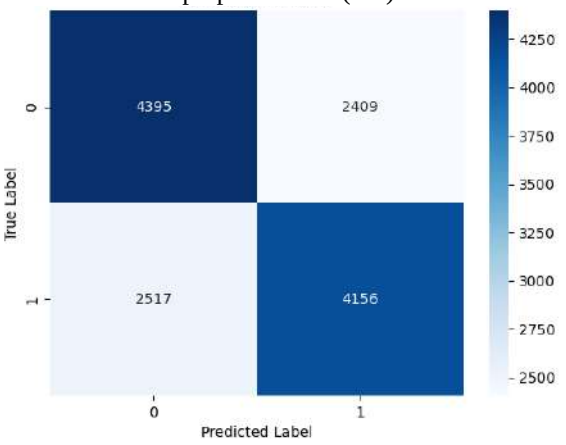
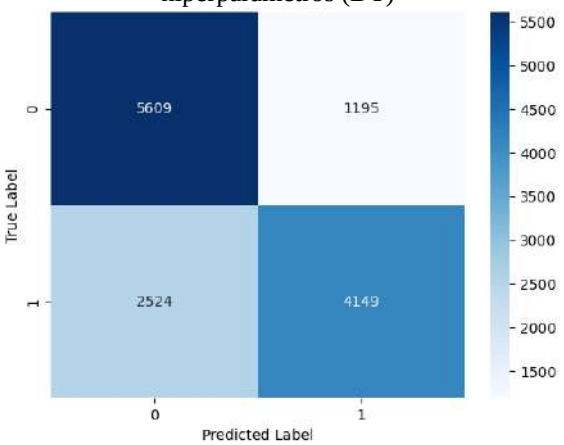



Gráfico 19 - Matriz de Confusión, con ajustes de hiperparametros (DT)




Fuente: Propia de los autores

En el entrenamiento inicial, el modelo de árboles de decisión logró una exactitud del 63% con la matriz de confusión que representa el número de Verdaderos Positivos (VP): 4395, Verdaderos Negativos (VN): 4156, Falsos Positivos (FP): 2517 y Falsos Negativos (FN): 2409.

Después de la evaluación mediante el ajuste de hiperparámetros, se logró una exactitud del 72% y se obtuvieron mejores resultados en términos de clasificación de las observaciones: Verdaderos Positivos (TP): 5609, Verdaderos Negativos (VN): 4149, Falsos Positivos (FP): 2524 y Falsos Negativos (FN): 1195.

**Random Forest (RF)**

Gráfico 20 - Matriz de Confusión, sin ajustes de hiperparametros (RF)

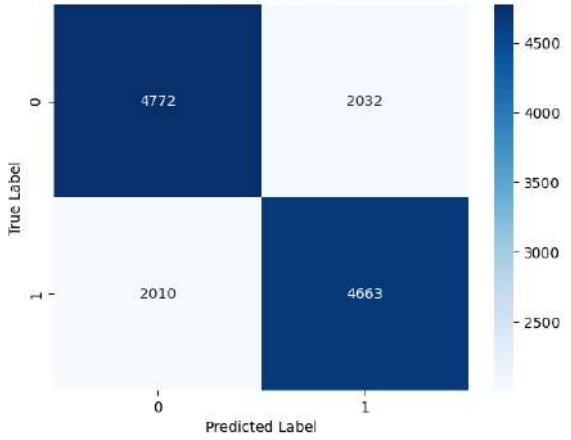
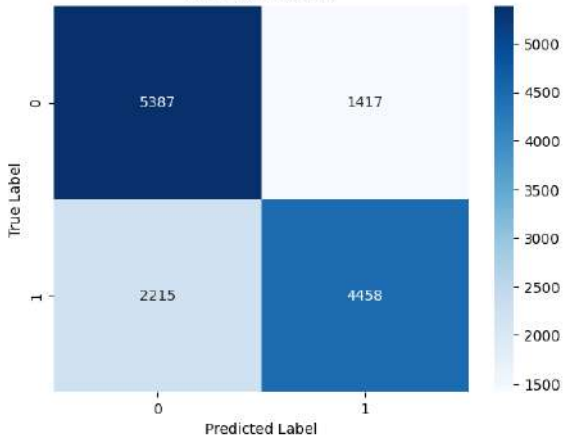



Gráfico 21 - Matriz de Confusión, con ajustes de hiperparametros (RF)




Fuente: Propia de los autores

En el entrenamiento inicial, el modelo de random forest logró una exactitud del 70% con la matriz de confusión que representa el número de Verdaderos Positivos (VP): 4772, Verdaderos Negativos (VN): 4663, Falsos Positivos (FP): 2010 y Falsos Negativos (FN): 2032.

Después de la evaluación mediante el ajuste de hiperparámetros, se logró una exactitud del 73% y se obtuvieron mejores resultados en términos de clasificación de las observaciones: Verdaderos Positivos (TP): 5387, Verdaderos Negativos (VN): 4458, Falsos Positivos (FP): 2215 y Falsos Negativos (FN): 1417.

# 5 RESULTADOS Y ANÁLISIS

Al finalizar el proceso de preprocesamiento de datos, entrenamiento y evaluación de los tres modelos: Regresión Logística, Arboles de Decisión y Bosques Aleatorios, se ha seleccionado el modelo de Random Forest (Bosques aleatorios) como modelo predictivo ya que el mismo ha alcanzado un 73% de Acurracy (Exactitud) y también a diferencia de los demás modelos presentó una cantidad mínima de falsos negativos y falsos positivos siendo así el que obtuvo mejores resultados.

**Desarrollo de la Interfaz de usuario de predicción de eventos cardiovasculares**

Se ha desarrollado una interfaz de usuario para la implementación del modelo predictivo, se ha seleccionado el modelo Random Forest para realizar las predicciones, el cual ha demostrado mejores resultados.

A continuación, se describen las tecnologías que se han utilizado para la construcción y desarrollo de la Interfaz de usuario para la predicción de eventos cardiovasculares:

- **Frontend:**

Streamlit es un framework open source que permite la creación de aplicaciones webs interactivas. Está diseñado específicamente para el desarrollo de interfaces de usuario orientado a machine learning y ciencia de datos. Se ha utilizado el marco de trabajo para la creación de la entrada de datos, el filtrado y la visualización de los datos de las predicciones realizadas, de igual manera para el historial de registros de predicciones.

- **Backend:**

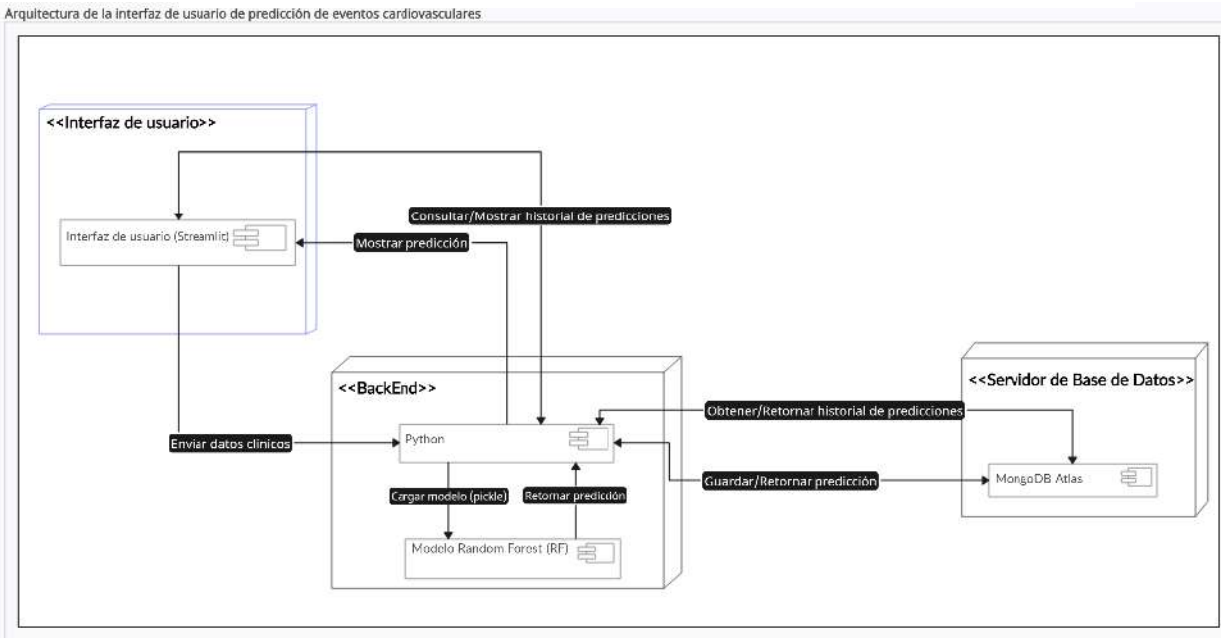
Python es un lenguaje de programación interpretado orientado especialmente para el análisis de datos y la creación de modelos de machine learning. Encargado principalmente del proceso de la codificación y construcción de la lógica, en donde se realizaron cálculos, importación y uso del modelo predictivo, así como también la gestión con la conexión con la base de datos.

- **Base de Datos:**

Se ha empleado una base de datos NoSQL específicamente MongoDB, para el almacenamiento y consulta de historiales de predicciones.

A continuación, se presenta la arquitectura de funcionamiento de la Interfaz de predicción de eventos cardiovasculares:

Gráfico 22 - Arquitectura de la Interfaz de predicción de eventos cardiovasculares




Fuente: Propia de los autores

Como se observa en el Gráfico N° 22, se presenta la arquitectura de la interfaz de usuario, el funcionamiento es el siguiente:

En primera instancia, el usuario inicia el proceso cargando los datos clínicos y de estilo de vida de los pacientes en la interfaz, los cuales son enviado al backend, a partir de realizar dicho procedimiento, el backend carga el modelo pre-entrenado, lo procesa y realiza la predicción en donde se muestra en consecuencia una ventana de la predicción realizada. Los datos, se almacena en una base de datos en la nube una vez que el usuario cargue los datos personales del paciente y presione el botón guardar.

Además, el usuario puede realizar consultas del historial de predicciones realizadas, realiza dicha petición, luego es procesado y consultado en la base de datos en donde al finalizar el proceso muestra el historial de predicciones del paciente.

# 6 CONCLUSIONES Y RECOMENDACIONES

**Conclusiones**

Al finalizar este trabajo de tesis de grado, nos hemos percatado de la gran importancia de tratar las enfermedades de manera temprana. Actualmente existe una alta prevalencia de enfermedades cardiovasculares, las cuales son la principal causa de mortalidad a nivel mundial. Un análisis exhaustivo de los datos clínicos de las consultas de pacientes del instituto de previsión social nos ha revelado el gran impacto de consultas realizadas por factores relacionados a enfermedades del corazón.

Hemos recopilado un conjunto de datos con factores específicos pertinentes referentes al ámbito cardiológico, lo que nos ayudó a poder realizar un análisis de los requerimientos necesarios para el entrenamiento de los modelos y a su vez evaluar el desempeño de cada uno de ellos.

Los modelos seleccionados para el entrenamiento fueron: regresión logística, arboles de decisión y bosques aleatorios. Dichos modelos fueron entrenados con y sin ajuste de hiperparametros, una técnica de machine learning que permite encontrar un conjunto de hiperparametros óptimos que mejoran los resultados de las predicciones.

Evaluamos el desempeño de los diferentes modelos entrenados. Entre los diversos modelos evaluados, el modelo de bosques aleatorios (Random Forest) demostró ser el más preciso, alcanzando un 73% de exactitud en la clasificación, con una mínima cantidad de falsos negativos (FN) y Falsos positivos (FP) siendo así el modelo selecto para la aplicación dentro de la interfaz gráfica de usuario (GUI). Una vez exportado dicho modelo en formato.PKL, se realizaron las predicciones con los mismos datos del dataset. Se tomaron 10 muestras de manera aleatoria de estos pacientes, arrojando así resultados positivos de acuerdo a nuestros objetivos.

**Recomendaciones**

1- Validación y evaluación de los modelos: Es importante realizar la validación y evaluación de los modelos con los nuevos datos que se estarán almacenando en la base de datos, de esta manera se podrá realizar predicciones en base a datos históricos de pacientes del Instituto de Previsión Social de la ciudad de Coronel Oviedo.

2- Aplicación e implementación de nuevos algoritmos de Inteligencia artificial: A partir del proceso de análisis de datos históricos almacenados en la base de datos se podría aplicar pronóstico de series de tiempo para realizar predicciones e identificar patrones y tendencias con respecto a eventos cardiovasculares.

3- Ampliación de las funcionalidades de la interfaz gráfica de usuario (GUI): Es posible escalar implementando gráficos en tiempo real, creación perfiles de usuario para que los pacientes puedan cerciorarse de su historial clínico cardiológico, entre otras funcionalidades.

4- Apoyo al proyecto para la toma de decisiones: En el estudio y análisis del historial clínico de las consultas realizadas en el Instituto de Previsión Social de Coronel Oviedo se observó un gran número de consultas por diversas enfermedades, por lo que dicha institución puede apoyar al proyecto para la toma de decisiones en el campo de la cardiología, como también puede ser útil para otros campos y profesionales de la salud.

# 7 BIBLIOGRAFIA

[1] M. d. S. P. y. B. Social, "Infarto y derrame, enfermedades cardiovasculares que producen mayor número de fallecidos," 2021. [Online]. Available: https://portal.mspbs.gov.py/infarto-y-derrame-enfermedades-cardiovasculares-que-producen-mayor-numero-de-fallecidos/. [Accessed 14 agosto 2024].

[2] Datacamp, "¿Que es el machine learning? Definición, tipos, herramientas y más”," 2024. [Online]. Available: https://www.datacamp.com/es/blog/what-is-machine-learning. [Accessed 26 julio 2024].

[3] Kaggle, "Cardiovascular Disease dataset," 2020. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. [Accessed 20 junio 2024].

[4] A. Amazon, "¿Que es la regresión logística," 2024. [Online]. Available: https://aws.amazon.com/es/what-is/logistic-regression/. [Accessed 26 julio 2024].

[5] S. Learn, "Árboles de Decisión," 2024. [Online]. Available: https://scikit-learn.org/stable/modules/tree.html. [Accessed 26 julio 2024].

[6] G. f. Geeks, "Algoritmo de bosque aleatorio en aprendizaje automático," 2024. [Online]. Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/. [Accessed 26 julio 2024].

[7] T. M. Learners, "Métricas de Clasificación," 2024. [Online]. Available: Métricas de Clasificación - Aprende a EVALUAR tu modelo. [Accessed 13 septiembre 2024].

[8] DataSource.ai, "Métricas de Evaluación De Modelos En El Aprendizaje Automático," 2023. [Online]. Available: Métricas De Evaluación De Modelos En El Aprendizaje Automático. [Accessed 13 septiembre 2024].

[9] I. DataPort, "Conjunto de datos de enfermedades cardiovasculares," 2024. [Online]. Available: https://ieee-dataport.org/documents/cardiovascular-disease-dataset. [Accessed 20 junio 2024].

[10] M. d. S. P. y. B. Social, "¿Qué es la Diabetes?," 2018. [Online]. Available: https://www.mspbs.gov.py/portal-16664/iquestque-es-la-diabetes.html. [Accessed 14 agosto 2024].

[11] F. E. d. Corazón, "Colesterol y Riesgo Cardiovascular," 2024. [Online]. Available: https://fundaciondelcorazon.com/prevencion/riesgo-cardiovascular/colesterol.html. [Accessed 14 agosto 2024].

[12] CDC, "Acerca del índice de masa corporal para adultos," 2022. [Online]. Available: https://www.cdc.gov/healthyweight/spanish/assessing/bmi/adult_bmi/index.html#:~:text=Con%20el%20sistema%20m%C3%A9trico%2C%20la,la%20estatura%20en%20metros%20cuadrados. [Accessed 14 agosto 2024].

[13] Pandas, "pandas.DataFrame.isna," 2024. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.isna.html. . [Accessed 15 octubre 2024].

[14] S. Learn, "Validación cruzada: evaluación del rendimiento del estimador," 2024. [Online]. Available: : https://scikit-learn.org/1.5/modules/cross_validation.html. [Accessed 13 septiembre 2024].

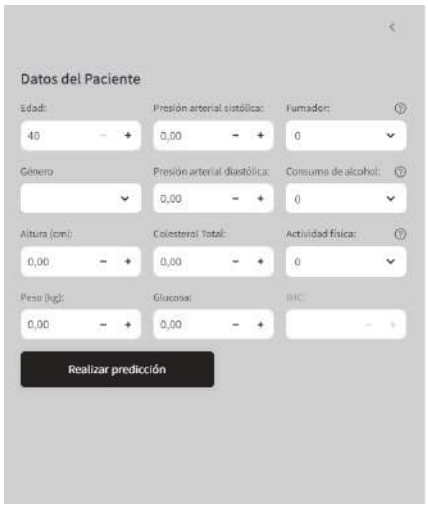
# ANEXOS

**Indicaciones para el uso de la Interfaz Gráfica para Predicciones de Eventos Cardiovasculares.**

Manual de uso de la Interfaz gráfica de usuario (GUI) para Predicciones de Eventos Cardiovasculares.

**Carga de datos del paciente:** En la sección ubicada en la barra lateral se encuentra los campos: edad (entero), género (Masculino, Femenino), altura (cm), peso (kg), presión arterial sistólica (entero), presión arterial diastólica (entero), colesterol total (entero), glucosa (entero), fumador (0: fumador inactivo, 1: fumador activo), consumo de alcohol (0: inactivo,1: activo), actividad física (0: Persona Inactivo Físicamente, 1: Persona Activo Físicamente) y el índice de masa corporal (IMC) que se calculara automáticamente una vez se haya introducido la altura y el peso del paciente. Estos campos deben ser rellenados cuidadosamente para iniciar con el proceso de predicción. Cabe mencionar que algunos campos poseen un a su lado un icono de información que al pasar por encima el cursor se detallara cuestiones médicas a tener en cuenta a la hora de cargado de dichos campos.

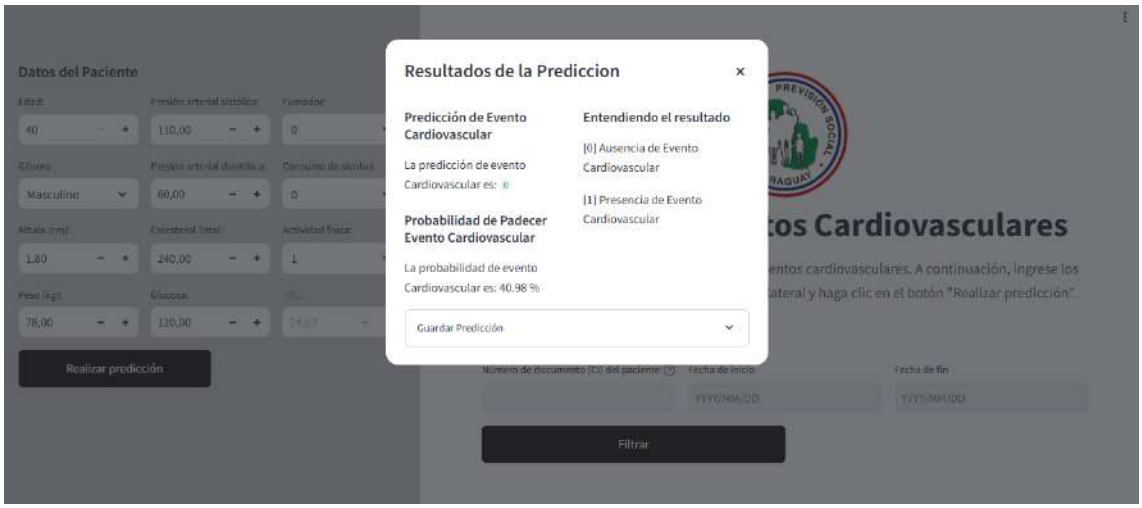
**Realización de predicciones:** continuando en la ubicación de la barra lateral se encuentra un botón de nombre “Realizar Predicción”. Al presionar el botón se realizará el proceso de predicción en base a los datos cargados en los campos.





**Visualización de resultado de la predicción:** una vez se haya concluido el proceso de predicción se desplegará una venta con el correspondiente resultado de la predicción del evento

cardiovascular, así como también la probabilidad de que el paciente padezca o no un evento cardiovascular.




**Guardar predicción:** En la misma ventana desplegada se encuentra un formulario para guardar las predicciones, los campos deben ser rellenados para guardar la predicción referente a un paciente. Los datos necesarios son: Documento (CI), Nombres y Apellidos. Al presionar el botón "Guardar Predicción", La información se almacenará en una base de datos en la nube.


**Historial de predicciones:** En esta sección se encuentra los campos de acceso de búsqueda de historial de registros de predicciones de pacientes. En el campo Documento CI se debe cargar la cedula de identidad del paciente para la búsqueda, además, se puede realizar un filtrado por rango de fecha de inicio y fecha fin, también es posible por fecha actual al seleccionar solo la fecha de

inicio. Una vez cargado los datos correspondientes se procede la búsqueda presionando el botón "Filtrar", se generará una tabla de los registros basados en los datos cargados.


| | CI | Nombre | Apellido | Edad | Genero | Altura | Peso | PAS | PAD | Colesterol_Total | Glucosa | Fumador | Consumo_Alcohol | Actividad_F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 66666666 | Juan | Perez | 40 | 1 | 180 | 78 | 110 | 60 | 3 | 2 | 0 | 0 | |
| 2 | 5669844 | Jose Antonio | Espinoza Franco | 40 | 1 | 140 | 68 | 170 | 70 | 2 | 2 | 0 | 0 | |
| 3 | 88888 | Milva | Perez | 40 | 2 | 134 | 40 | 120 | 120 | 1 | 2 | 0 | 1 | |
| 4 | 22222 | Sergio | Valdez | 40 | 1 | 190 | 74 | 240 | 120 | 1 | 0 | 0 | 0 | |
| 5 | 56565665 | Marcos | Ozuna | 40 | 1 | 170 | 47 | 200 | 200 | 2 | 3 | 0 | 0 | |
| 6 | 66666 | Juana | Lopez | 40 | 1 | 140 | 87 | 110 | 80 | 3 | 2 | 0 | 0 | |

**Exportar historial de predicciones:** luego de que se haya generado la tabla de historiales, tendrá la opción de exportar en formato CSV haciendo click en el icono de Download as CSV.

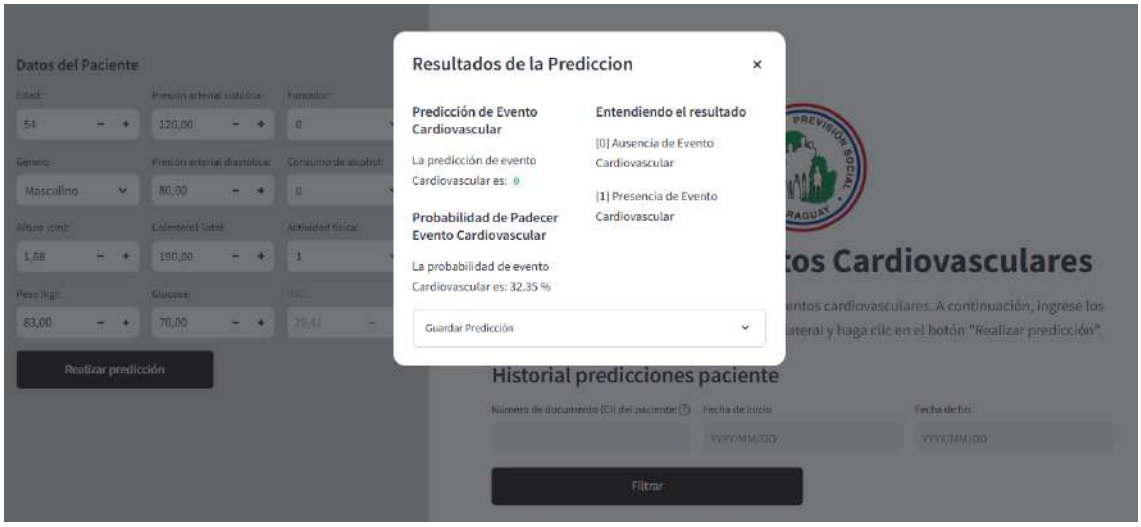
## Pruebas de validación de la interfaz de usuario para la predicción de eventos cardiovasculares

Subconjunto Aleatorio de 10 Muestras

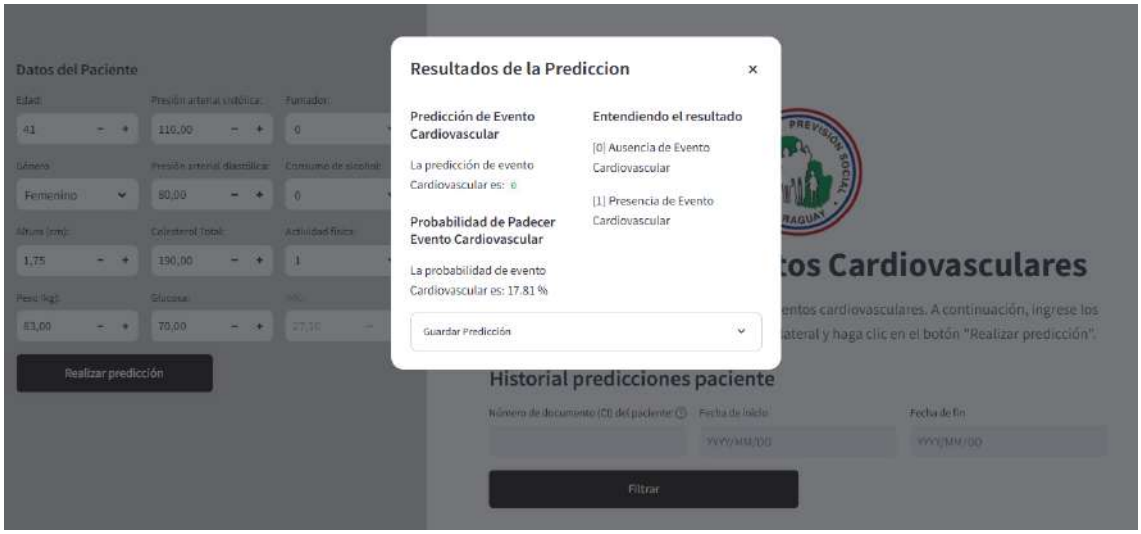
| | Edad | Genero | Altura | Peso | PAS | PAD | Colesterol | Glucosa | Fumar | Alcohol | Actividad_Fisica | Target_Variable_Cardio | imc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **30521** | 44 | 2 | 168 | 67.0 | 120 | 80 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| **26785** | 54 | 1 | 168 | 83.0 | 120 | 80 | 1 | 1 | 0 | 0 | 1 | 0 | 3 |
| **49471** | 41 | 2 | 175 | 83.0 | 110 | 80 | 1 | 1 | 0 | 0 | 1 | 0 | 3 |
| **21544** | 64 | 1 | 151 | 75.0 | 130 | 80 | 1 | 1 | 0 | 1 | 0 | 1 | 4 |
| **48871** | 46 | 2 | 163 | 63.0 | 120 | 80 | 3 | 1 | 0 | 0 | 0 | 1 | 2 |
| **5350** | 52 | 1 | 164 | 72.0 | 120 | 80 | 2 | 1 | 0 | 0 | 1 | 0 | 3 |
| **12642** | 40 | 1 | 156 | 58.0 | 110 | 70 | 1 | 1 | 0 | 0 | 1 | 0 | 2 |
| **30624** | 59 | 1 | 164 | 85.0 | 160 | 100 | 1 | 3 | 0 | 0 | 1 | 1 | 4 |
| **7342** | 52 | 1 | 156 | 76.0 | 120 | 80 | 1 | 1 | 0 | 0 | 1 | 0 | 4 |
| **18047** | 46 | 2 | 170 | 88.0 | 120 | 80 | 3 | 1 | 1 | 0 | 1 | 1 | 4 |

Primera Prueba de validación – Muestra 1

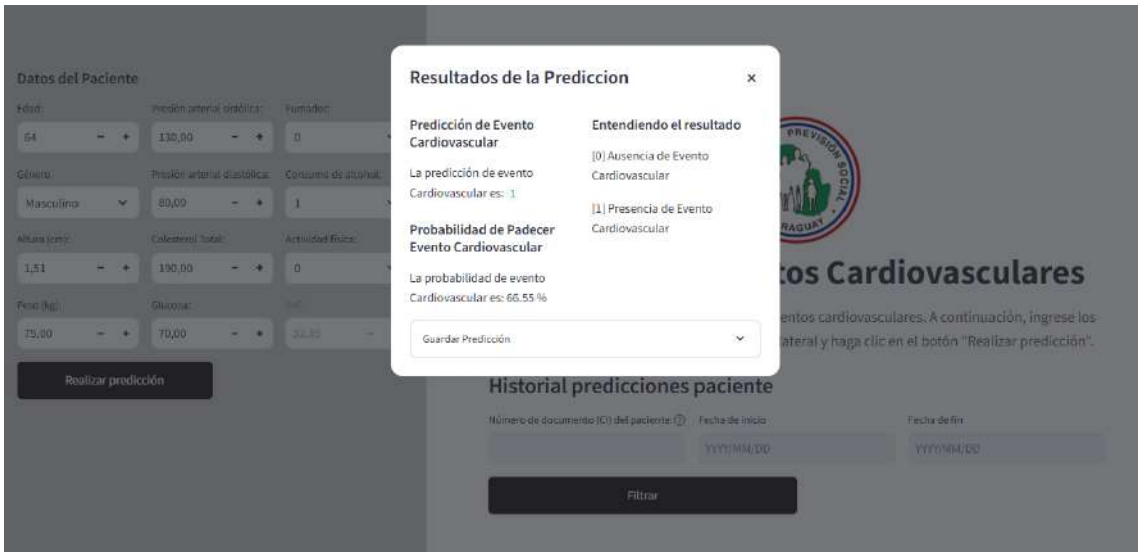
Segunda Prueba de validación – Muestra 2



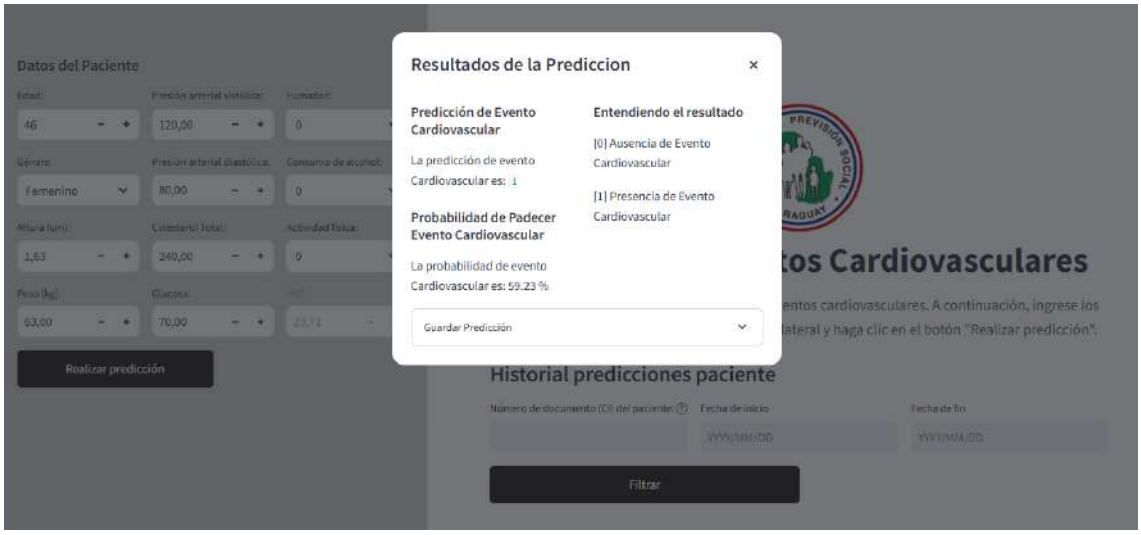

Tercera Prueba de validación – Muestra 3



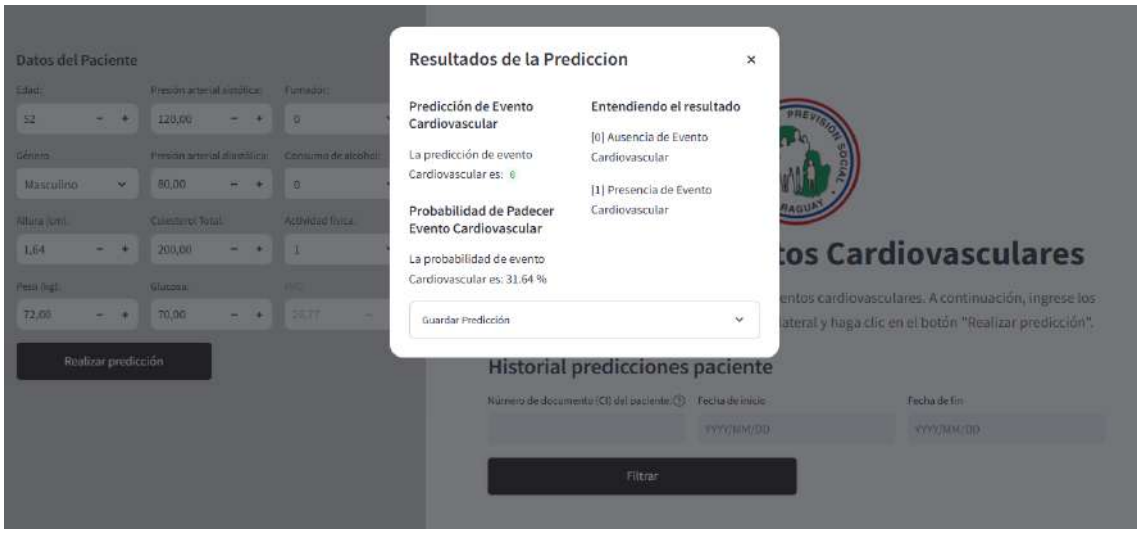
Cuarta Prueba de validación – Muestra 4



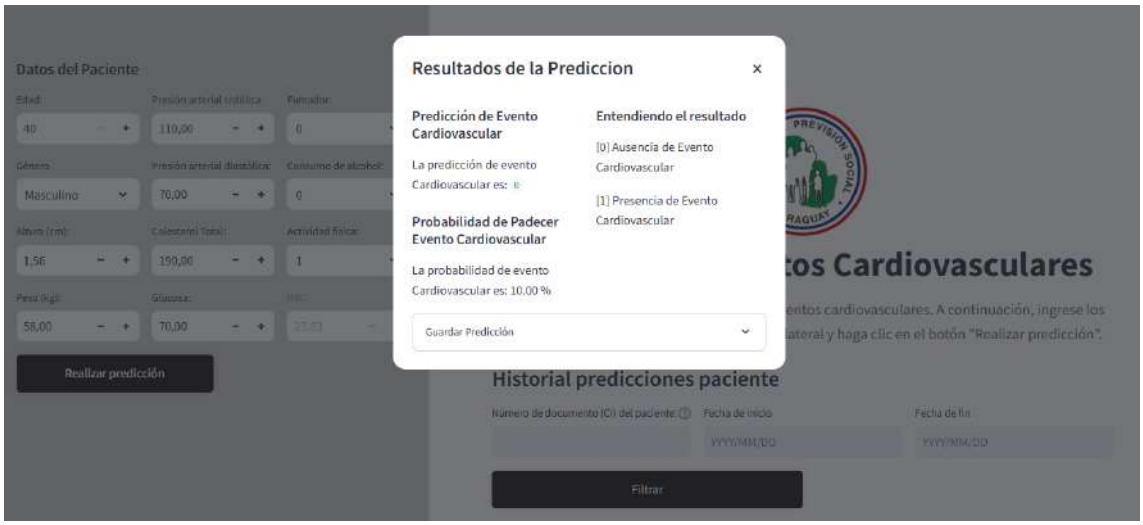

Quinta Prueba de validación – Muestra 5



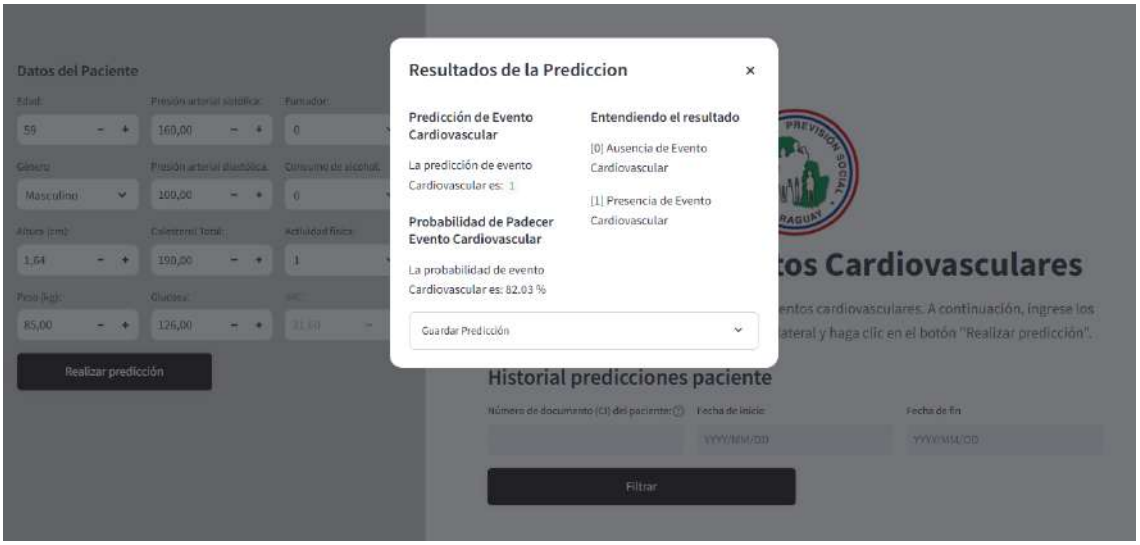
Sexta Prueba de validación – Muestra 6



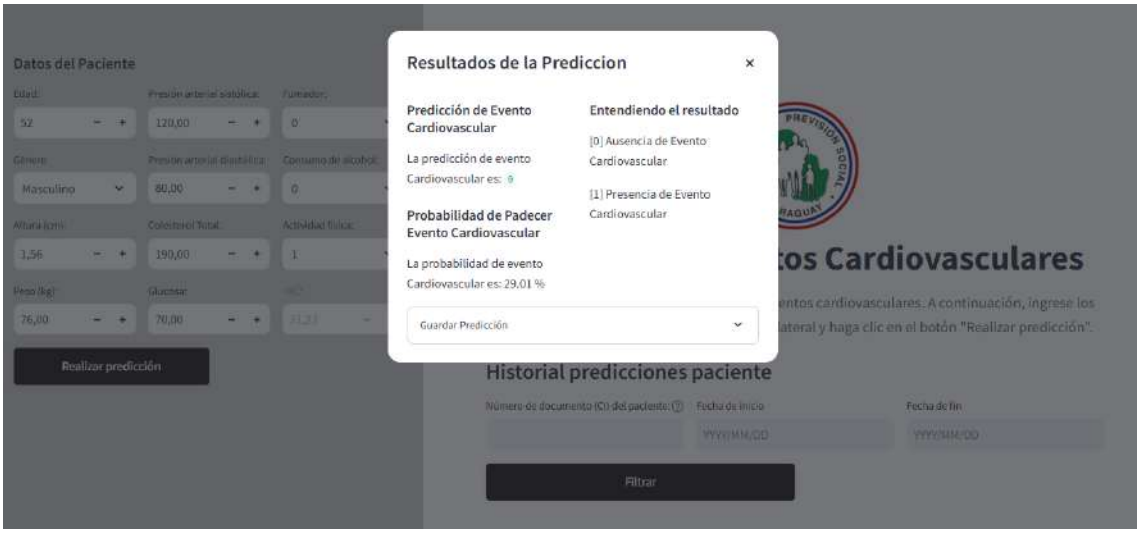

Séptima Prueba de validación – Muestra 7



Octava Prueba de validación – Muestra 8




Novena Prueba de validación – Muestra 9



Decima Prueba de validación – Muestra 10