Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

MODELADO PREDICTIVO CON *RANDOM FOREST* PARA LA DETECCIÓN DE ENFERMEDADES CARDÍACAS

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestro en Ciencia de Datos

Presenta:
Hector Daniel Estrada Rodriguez

Director:
Dr. Jaime Emmanuel Alcalá Temores

Tlaquepaque, Jalisco, 13 de mayo de 2024

MODELADO PREDICTIVO CON *RANDOM FOREST* PARA LA DETECCIÓN DE ENFERMEDADES CARDÍACAS

Hector Daniel Estrada Rodriguez

Resumen

El diagnóstico precoz de enfermedades cardíacas permite mejorar la probabilidad de supervivencia de las personas, así como reducir los gastos adicionales al sistema de salud. Es por esto que el objetivo principal de este proyecto es aplicar un modelo de detección de patologías cardíacas, resolviendo de manera particular la necesidad de herramientas analíticas avanzadas que puedan procesar datos clínicos y biomédicos de manera efectiva. Se destaca la implementación y comparación de un modelo de *Random Forest* frente a la Regresión Logística, así como los procesos de limpieza, preparación de datos y la ingeniería de características realizada. Se exponen los resultados obtenidos, que demuestran la superioridad predictiva del *Random Forest* en comparación con la *Regresión Logística*. Finalmente, se presentan las conclusiones del trabajo, enfatizando la viabilidad del *Random Forest* para la aplicación clínica.

Tabla de Contenidos

Índice de figuras

Página

Índice de tablas

Página

Dedicado

Agradezco a mi madre, quien ha sabido formarme con buenos valores y principios, lo cual me permitió seguir adelante en momentos difíciles. También dedico este trabajo a la Maestra Cristina Sánchez, quien me ha apoyado durante el transcurso de la maestría. Asimismo, agradezco a la Doctora Rocío Carrasco y al Doctor Jaime Emmanuel Alcalá por su dedicación y tiempo para la redacción de este trabajo de obtención de grado.

Quiero agradecer a mis hermanas Debora y Ruth por estar siempre presentes a lo largo de este proyecto. Gracias a Fernanda Melchor por el apoyo y tiempo que dedicaste para lograr esta meta.

1 Introducción

En este capítulo se presenta una introducción al tema de la aplicación de la ciencia de datos para predecir fallas cardíacas mediante el modelo Random Forest *el cual utiliza datos clínicos y biomédicos como variables predictivas de una posible falla cardíaca. Siendo así, este proyecto reconoce la importancia de abordar el tema para proponer una herramienta que agilice la detección oportuna de fallas cardíacas que genere un impacto positivo en la calidad de vida de los pacientes y a su vez, reduzca los recursos económicos y humanos del sector salud [1].*

1.1 Contexto

La prevalencia de las fallas cardíacas ha ido en aumento en los últimos años, y se estima que afecta a 2 millones de personas en todo el mundo [2]. Además de la carga física y emocional que impone a los pacientes, las fallas cardíacas también generan una carga significativa en los sistemas de atención médica y tienen un impacto económico considerable debido a las hospitalizaciones frecuentes y los tratamientos costosos. En México, según datos de la Federación Mundial del Corazón, las enfermedades cardíacas afectan al 26 % de la población con un costo de $ 6.1 mil millones de dólares y ocupan el 4 % de todo el gasto en salud [3].

Históricamente, la detección y el manejo de las fallas cardíacas se han basado en enfoques clínicos, que comúnmente dependen de la identificación de síntomas físicos o mediciones de la función cardíaca, como la fracción de eyección[1]. Sin embargo, estas técnicas pueden no ser lo suficientemente completas ni específicas para determinar de manera concreta la aparición de fallas cardíacas en sus etapas iniciales.

En este sentido surge la aplicación de la ciencia de datos y el aprendizaje automático como herramientas prometedoras para la detección temprana y la gestión más efectiva de las fallas cardíacas. La recopilación de datos clínicos completos, la monitorización continua y la aplicación de algoritmos avanzados pueden ofrecer una visión más completa del estado físico de un individuo y permitir la identificación de riesgo y patrones de predicción.

Este enfoque innovador tiene el potencial de transformar la forma en que abordamos las fallas cardíacas, permitiendo intervenciones médicas más oportunas y estrategias de prevención más efectivas. A medida que avanzamos hacia una era de medicina personalizada, la capacidad de predecir y gestionar las fallas cardíacas de manera individualizada se presenta como una oportunidad emocionante para mejorar la atención médica y reducir la carga de esta enfermedad en la sociedad.

[1] La fracción de eyección es una medida del porcentaje de sangre que sale del corazón cada vez que se comprime

1.2 Justificación

Este trabajo encuentra su justificación en una serie de factores de tipo científica, económica y social que resaltan su necesidad y urgencia.

Según datos proporcionados por el gobierno de México, anualmente se registra una alarmante cifra de aproximadamente 220 mil decesos debido a enfermedades cardíacas, de los cuales 177 mil ocurrieron en el año 2021 y están directamente relacionados con infartos al miocardio [4].

Con base en los comunicados del gobierno de México; esta realidad es alarmante y resalta la urgencia de abordar de manera efectiva

la problemática de las enfermedades cardíacas en nuestra sociedad. Es fundamental reconocer que las afecciones cardíacas tienen raíces genéticas, lo que implica una progresión gradual en su desarrollo. Por esta razón, la gestión adecuada del colesterol acumulado en las arterias, el control de la presión arterial y la prevención de enfermedades como la diabetes adquieren un papel esencial en la lucha contra estas condiciones. Además, el tabaquismo, un factor de riesgo bien documentado, puede desencadenar complicaciones coronarias en cualquier etapa de la vida y contribuir significativamente a la elevada tasa de fatalidades asociadas a enfermedades cardíacas [5].

Dada la magnitud de este problema de salud pública y su impacto en la sociedad, este trabajo se presenta como un esfuerzo necesario para profundizar en la comprensión de las enfermedades cardíacas, identificar factores de riesgo y desarrollar estrategias efectivas de prevención y manejo. El conocimiento derivado de esta investigación contribuirá a la prevención y detección temprana de enfermedades cardiovasculares, a su vez podría tener un impacto positivo en la economía, al reducir los costos asociados a la atención médica de enfermedades cardíacas y disminuir la carga de enfermedad en el sistema de salud.

Por conclusión, este trabajo busca resaltar la importancia de abordar las enfermedades cardíacas con un enfoque interdisciplinar entre la medicina y la ciencia de datos.

1.3 *Problema*

En el campo de la atención médica existe una gran necesidad de mejorar la detección temprana de enfermedades cardíacas para proporcionar un tratamiento oportuno y mejorar el pronóstico de un paciente. Actualmente no se implementa un sistema de predicción oportuna que permita la identificación de pacientes de alto riesgo de desarrollar una enfermedad cardiaca en una etapa temprana debido a sus implicaciones legales, eticas, sociales y economicas.

El problema científico que se abordará en esta investigación se centra en desarrollar y evaluar un modelo de clasificación basado en el algoritmo *Random Forest* [6], para predecir la presencia o ausencia de enfermedades cardíacas en pacientes. Este problema científico involucra:

- Identificar qué atributos clínicos y biomédicos son más significativos en la predicción de enfermedades cardíacas.
- Determinar la eficacia y precisión del modelo de *Random Forest* en la detección temprana de enfermedades cardíacas en comparación con otros enfoques de aprendizaje automático.

1.4 Objetivos

1.4.1 Objetivo general

Emplear un modelo de clasificación basado en el algoritmo *Random Forest* que sea capaz de predecir con precisión la presencia o ausencia de enfermedades cardíacas mediante variables clínicas y biomédicas de los pacientes, con el objetivo de mejorar la detección temprana y contribuir a una atención médica más efectiva.

1.4.2 Objetivos específicos

1. Obtener y preparar un conjunto de datos clínicos y biomédicos que incluya información sobre factores de riesgo e incluyendo el estado del paciente.
2. Realizar un análisis exploratorio de datos para identificar variables significativas, incluyendo la existencia de correlaciones que puedan influir en la predicción de enfermedades cardíacas.
3. Diseñar un modelo de clasificación usando el algoritmo *Random Forest*, dando importancia a la selección de variables significativas.
4. Entrenar un modelo *Random Forest* utilizando un conjunto de datos óptimo y evaluar su rendimiento en las métricas de precisión, sensibilidad, especificidad.
5. Comparar el rendimiento del modelo *Random Forest* con otro modelo, como la regresión logística.

2 Metodología

En este capítulo se presenta en detalle el desarrollo metodológico que incluye pasos o procesos a seguir.

2.1 Descripción de los datos

El proceso de recopilación de datos comenzó con una meticulosa búsqueda de información de salud en entidades reconocidas como la Organización Mundial de la Salud (OMS). No obstante, debido a la naturaleza privada de las bases de datos y de los procesos requeridos para acceder a estos, se decidió utilizar los recursos de IEEE DataPort [7]. Esta plataforma proporciona acceso a una base de datos que agrupa cinco diferentes conjuntos de información sobre enfermedades cardíacas. Los datos comprenden variables clave significativas: la edad, el sexo del paciente, el tipo de dolor de pecho, la presión arterial en reposo, el nivel de colesterol sérico, el azúcar en sangre en ayunas, resultados del electrocardiograma en reposo, la frecuencia cardíaca máxima alcanzada, la angina inducida por ejercicio, y la depresión del segmento ST, cada variable esta codificada mediante: datos numéricos, binarios y nominales. Estos atributos forman una base esencial para el análisis y la interpretación de las variables cardíacas en la población estudiada, incluyendo la descripción de los atributos numéricos como se observa en la Tabla 2.1.

Atributo	**Unidad**	**Tipo de dato**
Edad	años	Numérico
Sexo	1, 0	Binario
Tipo de dolor de pecho	1,2,3,4	Nominal
Presión arterial en reposo	mm Hg	Numérico
Colesterol en suero	mg/dl	Numérico
Azúcar en sangre en ayunas (> 120 mg/dl)	1, 0	Binario
Resultados del electrocardiograma en reposo	0,1,2	Nominal
Frecuencia cardíaca máxima alcanzada	71–202	Numérico
Angina inducida por ejercicio	0,1	Binario
Depresión ST	mV	Numérico
La pendiente del segmento ST de ejercicio máximo	0,1,2	Nominal
Clase (Objetivo)	0,1	Binario

Tabla 2.1: Descripción de las variables del conjunto de datos.

En la preparación de los datos para el proceso de modelado de clasificación de enfermedades cardíacas, se realizaron diversas pasos para asegurar la confianza y utilidad de los datos. Este proceso incluyó un análisis descriptivo para obtener estadísticas básicas así como visualización de las distribuciones de variables numéricas que se observan en la Tabla 2.1, y de las variables nominales como se observa en la Tabla 2.2. Se detectaron y eliminaron valores que físicamente no pueden existir, como el caso de presiones arteriales y niveles de colesterol con valores de cero o negativos, lo que redujo el conjunto de datos de 1190 a 1018 registros. Además, se consideró la aplicación de transformaciones logarítmicas para corregir la distribución de la presión arterial en reposo y el colesterol debido a la presencia de outliers. A pesar de esto, no se encontraron valores faltantes que requirieran atención adicional.

Atributo	**Descripción**
Sexo	1 = masculino, 0 = femenino
Tipo de dolor de pecho	Valor 1: angina típica Valor 2: angina atípica Valor 3: dolor no anginoso Valor 4: asintomático
Azúcar en sangre en ayunas	(azúcar en sangre en ayunas > 120 mg/dl) (1 = verdadero; 0 = falso)
Resultados del electrocardiograma en reposo	Valor 0: normal Valor 1: con anormalidad en la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST de > 0.05 mV) Valor 2: mostrando probable o definitiva hipertrofia ventricular izquierda según los criterios de Estes
Angina inducida por ejercicio	1 = sí; 0 = no
La pendiente del segmento ST de ejercicio máximo	Valor 1: ascendente Valor 2: horizontal Valor 3: descendente
Clase	1 = enfermedad cardíaca, 0 = Normal

Tabla 2.2: Descripción de los Atributos Nominales.

2.2 *Análisis exploratorio*

La fase de exploración de los datos para el desarrollo del modelo de *Random Forest* arrancó con una revisión profunda de la matriz de correlación representada en la Figura 2.1, que reveló la existencia de relaciones lineales fuertes entre algunas variables numéricas. En particular, la frecuencia cardíaca máxima, la angina inducida por el ejercicio y la pendiente del segmento ST mostraron correlaciones altas, lo que implica que hay una fuerte dependencia lineal entre ellas. A pesar de estas correlaciones, se reconoció que el algoritmo de *Random Forest* está bien equipado para manejarlas debido a su capacidad

de construir múltiples árboles de decisión que individualmente consideran subconjuntos de variables, lo que reduce el efecto de la multicolinealidad.

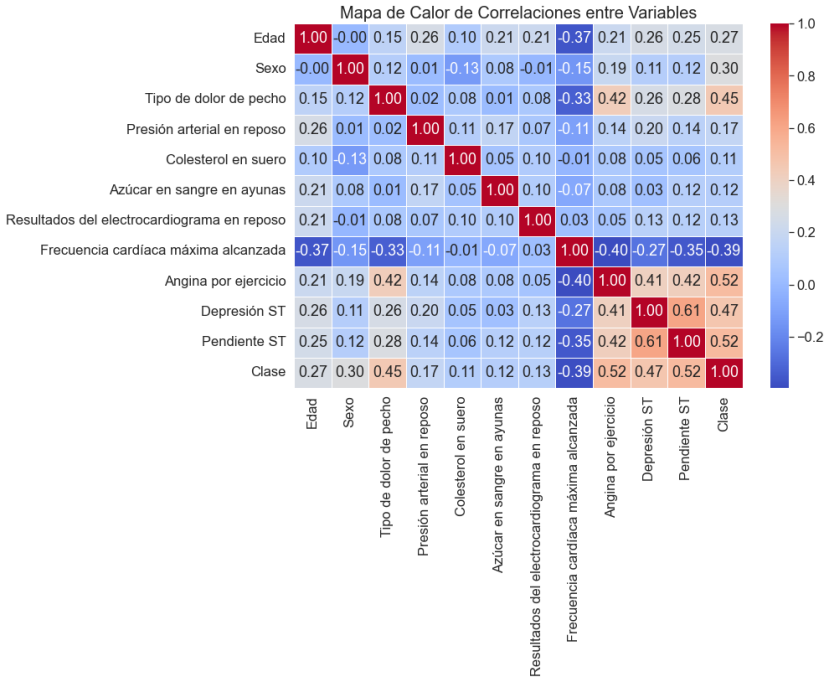

Figura 2.1: Mapa de Calor de Correlaciones entre Variables .

Al revisar la distribución de la variable objetivo, se constató un balance saludable entre las clases como se observa en la Figura 2.2, con 541 instancias sin presencia de enfermedad cardíaca (0) y 477 instancias con presencia de la misma (1). Lo cual es beneficioso ya que un desequilibrio significativo podría requerir técnicas de reequilibrio para prevenir un sesgo en las predicciones del modelo.

La exploración se extendió a las variables categóricas y su relación con la variable objetivo que se observa en la Figura 2.3. Se descubrieron patrones claros que indican que algunas categorías están más asociadas con la presencia o ausencia de enfermedades cardíacas. Estos patrones son particularmente útiles para el *Random Forest*, que es competente en la creación de divisiones basadas en categorías y puede, por tanto, utilizar esta información para mejorar la precisión de sus predicciones.

Se resalta la idoneidad de las variables categóricas para proporcionar un soporte estructural robusto para el *Random Forest*, ya que este modelo aprovecha la agrupación de observaciones en nodos de árbol basados en características similares. Este soporte estructural se traduce en una mayor capacidad de generalización del modelo y en la potencialidad de identificar con precisión los factores de riesgo asociados con las enfermedades cardíacas en pacientes, lo cual es el objetivo final del

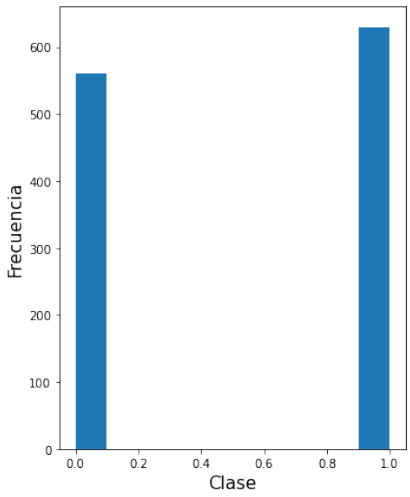

Figura 2.2: Distribución de la clase objetivo .

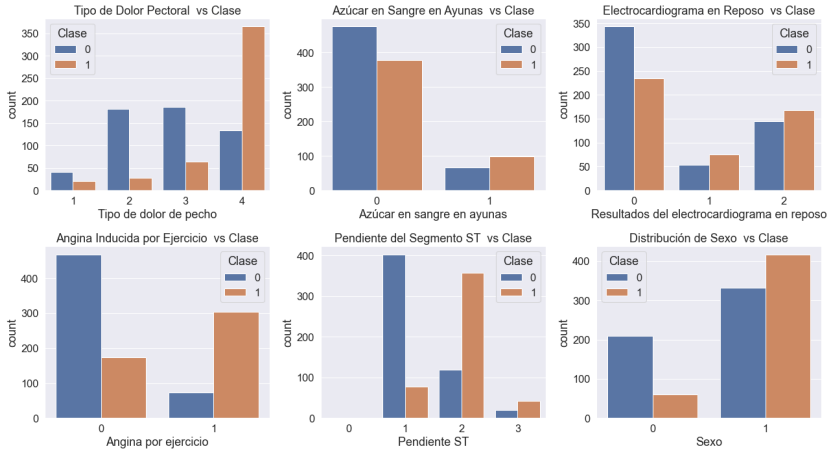

Figura 2.3: Distribución de la clase objetivo en relación a variables categóricas .

modelo predictivo en cuestión.

Se identificaron valores fisiológicamente improbables, como lecturas de cero o negativas para la presión arterial en reposo y los niveles de colesterol, y se procedió a su eliminación, lo cual resultó en una reducción del conjunto de datos esto se puede observar en la Tabla 2.3.

Tamaño de muestra original	**Tamaño de muestra post limpieza**
1190	1018

Tabla 2.3: Valores antes y después de limpieza de datos.

Se consideró la transformación logarítmica para el Colesterol en suero, la Presión arterial en reposo, Frecuencia cardíaca máxima alcanzada las cuales se pueden observar en la Figura 2.4, basándose en estas gráficas, se puede observar una mejora una vez aplicada la transformación, debido a la normalización de los datos y a una escala menor. Este cuidadoso proceso de transformación fue crucial

para preparar adecuadamente los datos, asegurando que el modelo de *Random Forest* pudiera operar con la mayor precisión posible en la predicción de enfermedades cardíacas.

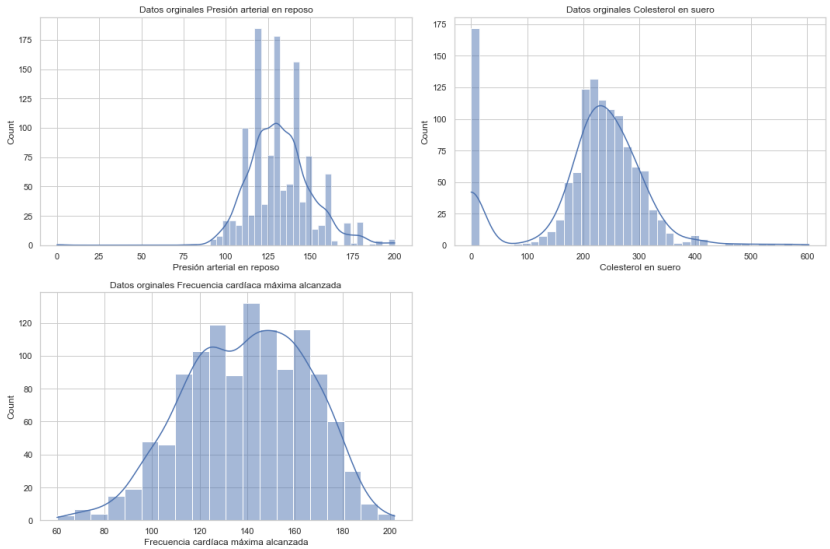

Figura 2.4: Distribución de datos previo a aplicación de transformación logarítmica .

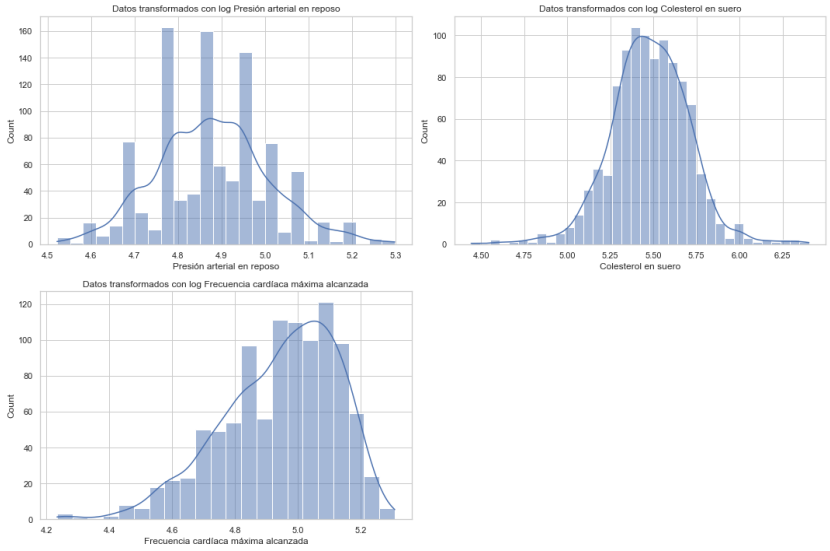

Figura 2.5: Distribución de datos post aplicación de transformación logarítmica .

Para el caso particular de la frecuencia cardíaca máxima alcanzada, se puede identificar que la distribución no se normaliza como se observa en la Figura 2.6. Sin embargo, en la escala se puede identificar una mejora observando la Figura 2.7, por lo cual se mantiene la transformación logarítmica para esta variable.

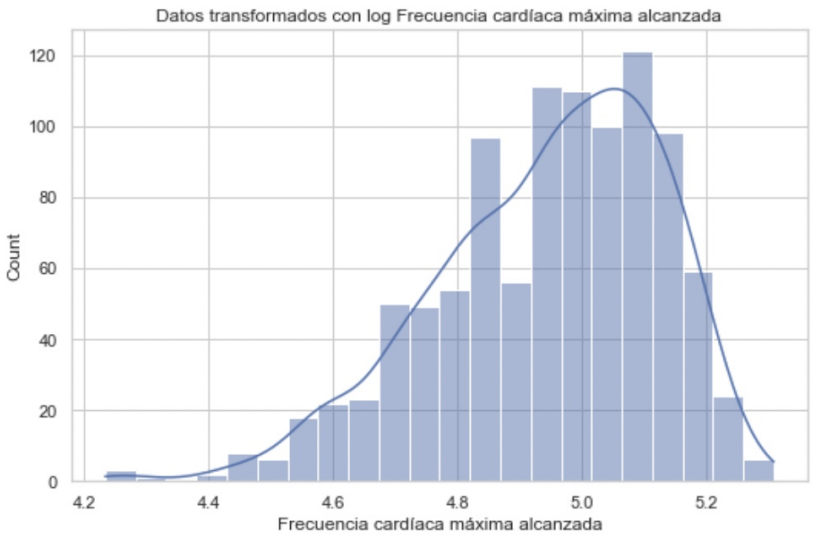

Figura 2.6: Distribución de datos post aplicación de transformación logarítmica para el caso de Frecuencia cardíaca .

2.3 *Descripción de los modelos*

En esta sección se incluye una descripción de los modelos utilizados para su comparación y el proceso de selección para justificar su eficacia.

Se implementarán dos modelos predictivos: un *Random Forest* [6] y una *Regresión Logística* [8]. El modelo de *Random Forest*, conocido por su capacidad para manejar grandes conjuntos de datos con múltiples variables y su robustez frente a la multicolinealidad, se configurará y entrenará primero. Paralelamente, se desarrollará un modelo de *Regresión Logística*, que es un método estadístico tradicionalmente utilizado para la clasificación binaria [8] y que servirá como punto de referencia para comparar la eficacia del *Random Forest*.

El objetivo es evaluar cuál de los dos modelos ofrece mejor rendimiento en la tarea de predecir la presencia o ausencia de enfermedades cardíacas, utilizando métricas estándares como la precisión, la sensibilidad (*recall* y la especificidad. La comparación directa permitirá determinar la viabilidad del modelo *Random Forest* frente a un enfoque más simple y lineal como la *Regresión Logística* en el contexto específico de los datos cardíacos procesados.

Un modelo de *Random Forest* se compone de un conjunto de árboles de decisión. Cada árbol, de manera individual, entrena una muestra aleatoria extraída de los datos de entrenamiento originales mediante el *bootstrapping*. Esto implica que cada árbol se entrena con una muestra ligeramente diferente, lo que contribuye a la diversidad del modelo y evita el sobreajuste [9].

Los modelos de *Random Forest* cuentan con una gran capacidad

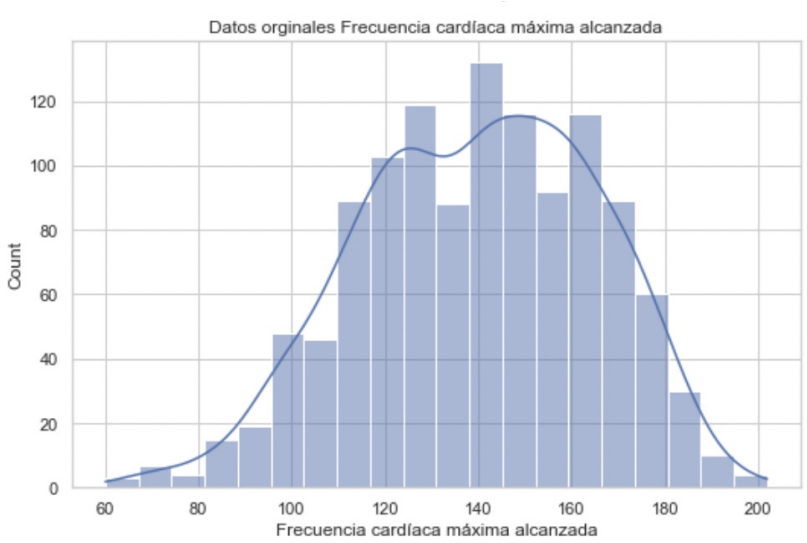

Figura 2.7: Distribución de datos previo aplicación de transformación logarítmica para el caso de Frecuencia cardíaca .

para manejar valores atípicos, gracias al entrenamiento de los árboles mediante muestras aleatorias dentro del mismo conjunto de datos. Además, tiene la capacidad manejar tanto variables categóricas como numéricas sin necesidad de aplicar un método de transformación a las variables categóricas. Esto se debe a que durante la construcción de cada árbol, en cada división de nodo, se consideran solo un subconjunto aleatorio de las características. Esta proceso aleatorio permite que el modelo se adapte a diferentes tipos de datos y aumenta su solidez frente a problemas de escala y etiquetas [6].

Otra ventaja de los *Random Forest*s es su capacidad para estimar la importancia de las variables en la predicción del resultado. Esto se logra mediante el cálculo de la precisión del modelo cuando se quita cada una de las variables, lo que proporciona información valiosa sobre qué variables son más significativas para el problema a predecir. En el caso particular de *Scikit-learn* este utiliza el método de importancia de características basado en *Gini impurity* para *Random Forest*s [9]. La importancia de una variable se calcula sumando la disminución de impureza ponderada sobre todos los nodos en los árboles de decisión que utilizan esa variable (2.1).

$$\text{Importancia de Característica}(j) = \frac{\sum_{t=1}^{T} \text{Impureza de Gini}_t(j) \times \text{Número de muestras en nodo}(j)}{\sum_{t=1}^{T} \text{Número total de muestras en nodo } j} \quad (2.1)$$

- T es el número total de árboles en el bosque.
- Impureza de Gini es la disminución de la impureza de Gini en el

nodo j del t-ésimo árbol.
- Número de muestras en nodo j es el número de muestras que llegan al nodo j.
- Número total de muestras en nodo j es el número total de muestras que llegan al nodo j en todos los árboles.

La *Regresión Logística* es un método estadístico ampliamente utilizado en análisis predictivo y modelado de datos. Su principal aplicación reside en la predicción de la probabilidad de ocurrencia de un evento binario, es decir, un evento que puede tener solo dos posibles resultados, como sí o no, éxito o fracaso, positivo o negativo [8].

Este método se utiliza comúnmente en problemas de clasificación binaria. Algunos ejemplos de uso en diferentes disciplinas son: la medicina para el diagnóstico de enfermedades; en finanzas como la evaluación de riesgos crediticios y en ciencias sociales como la predicción de comportamientos de los consumidores [10]. La *Regresión Logística* modela la relación entre las variables independientes y la variable dependiente utilizando la función sigmoide, que produce una salida en el rango de 0 a 1. Esta salida se interpreta como la probabilidad de que la variable de respuesta pertenezca a una de las dos categorías [8].

La *Regresión Logística* utiliza el concepto de logaritmo de odds, que es la transformación logarítmica de la probabilidad de un evento dividido por la probabilidad de que ese evento no ocurra(2.2). Al aplicar esta transformación, la *Regresión Logística* convierte la relación lineal entre las variables independientes y la probabilidad de ocurrencia del evento en una forma que puede ser modelada mediante una función logística.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

- x es la entrada.
- $f(x)$ está en el rango de (0,1)(0,1).

2.4 Descripción de las métricas

En el ámbito de la clasificación binaria, se asigna a cada instancia uno de dos posibles resultados: *Positivo* y *Negativo*. Un resultado "Positivoïndica que la instancia ha sido identificada acertadamente como perteneciente al grupo de interés. Por otro lado, un resultado *Negativo* sugiere que la instancia no forma parte del grupo objetivo.

Es crucial en este proceso diferenciar entre los Verdaderos Positivos, es decir, aquellas instancias que han sido correctamente identificadas como parte del grupo deseado, y los Falsos Positivos, que son aquellos casos erróneamente catalogados como pertenecientes al grupo cuando

no lo son. De manera similar, debemos reconocer los Verdaderos Negativos, que son las instancias adecuadamente clasificadas como ajenas al grupo de interés, y los Falsos Negativos, que son casos mal asignados fuera del grupo objetivo a pesar de que realmente sí pertenecen al mismo [11].

Exactitud (Accuracy): Esta métrica evalúa la proporción de clasificaciones correctas, incluyendo tanto Positivos como Negativos, en relación con el conjunto total de casos examinados como se puede observar en la fórmula [11] (2.3).

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

- TP son los verdaderos positivos.
- TN son los verdaderos negativos.
- FP son los falsos positivos.
- FN son los falsos negativos.

Precisión (Precisión): Esta medida se concentra en la proporción de instancias correctamente identificadas como Positivas frente al número total de instancias clasificadas como Positivas, correctas o no (2.4) [11].

$$\text{Precisión} = \frac{TP}{TP + FP} \tag{2.4}$$

Puntaje F1 (F1 *Score*): Representa la media entre la Precisión y la Sensibilidad (*Recall*), proporcionando un equilibrio entre ambas. Esta medida es particularmente valiosa en contextos donde existe un desequilibrio significativo en la distribución de las clases (2.5) [11].

$$F1 = 2 \times \frac{\text{Precisión} \times \textit{Recall}}{\text{Precisión} + \textit{Recall}} \tag{2.5}$$

Sensibilidad (*Recall*): Mide la proporción de Verdaderos Positivos identificados correctamente de entre todos los casos que son Verdaderos Positivos, en términos generales esta métrica busca que sean minimizados la cantidad de Falsos Negativos (2.6) [11].

$$\textit{Recall} = \frac{TP}{TP + FN} \tag{2.6}$$

En el marco de este proyecto, es esencial priorizar el uso de cuatro métricas clave para evaluar adecuadamente el desempeño del modelo. Sin embargo, dentro de estas destaca la sensibilidad ya que adquiere una importancia en el ámbito de la salud. Un Falso Negativo podría significar que a un paciente con una afección cardíaca no se le diagnostica correctamente, resultando en la falta de los cuidados y tratamientos necesarios. Esta falta no solo aumenta el riesgo de pérdida

de vidas sino que también puede conllevar un incremento en el uso de recursos que pueden ser invertidos en otras cuestiones. Por lo tanto, es vital reducir los Falsos Negativos para asegurar una atención médica oportuna y efectiva.

2.5 *Descripción de los experimentos o simulaciones*

Con los datos previamente preparados y listos para su uso, se procede a la etapa de modelado. Primero se procede a separar las variables predictoras X y la variable objetivo Y, con esto definido se adopta una división clásica de los datos en un conjunto de entrenamiento y un conjunto de prueba con una proporción de 80/20, lo que significa que el 0.80 de los datos se utilizará para entrenar los modelos y el 0.20 para evaluar su desempeño.

2.5.1 *Primer experimento con Random Forest:*

En este experimento, se procesarán las 11 variables sin aplicar ningún tipo de transformación, limitándose únicamente a eliminar los valores no plausibles, como aquellos de presión arterial y niveles de colesterol que presenten cifras negativas o iguales a cero. Se empleará un modelo de *Random Forest* sin optimización de parámetros. Dentro del estudio se utiliza una división de 80/20 y una semilla aleatoria de 42.

2.5.2 *Segundo experimento con Regresión Logística:*

Se repetirá el proceso con las mismas 11 variables sin aplicar transformaciones, realizando la misma limpieza de datos no plausibles. Se mantendrá la división de 80/20 y se fijará un máximo de 1000 iteraciones.

2.5.3 *Tercer experimento con Random Forest:*

En el tercer experimento, se trabajará con las 11 variables, aplicando esta vez una transformación logarítmica específicamente a las variables de presión arterial en reposo, colesterol en suero y frecuencia cardíaca máxima alcanzada. Para asegurar la replicabilidad, se adoptará una división de 80/20 y se utilizará la misma semilla aleatoria de 42.

2.5.4 *Cuarto experimento con Regresión Logística:*

Para el cuarto experimento, se conservarán las 11 variables, incluyendo la transformación logarítmica para las variables de presión arterial en reposo, colesterol en suero y frecuencia cardíaca máxima

alcanzada. Se seguirá el esquema de división de 80/20 y se establecerá un máximo de 1000 iteraciones.

3 Resultados y discusión

En este capítulo se presentan los resultados obtenidos del desarrollo de este trabajo y una discusión sobre el objeto de estudio.

3.1 Resultados

Durante el proceso de experimentación, se ejecuta la *Regresión Logística* como parámetro de comparación, obteniendo las métricas que se muestran en la Tabla 3.2. Se observó que la sensibilidad (*Recall*) fue del 80 %. Esto indica que el modelo falló en el 20 % de los casos, lo que no generó confianza suficiente para utilizar este modelo como herramienta inicial de predicción de enfermedades cardíacas. Es importante agregar que el valor *1* representa los casos de falla cardíaca, mientras que el valor "orepresenta los casos normales.

Se implementó el modelo *Random Forest*, obteniendo los resultados que se observan en la Tabla 3.1. Al tener una mayor capacidad para capturar el comportamiento de los datos, este modelo demostró una sensibilidad del 93 %, reduciendo considerablemente el margen de posibles fallas en la predicción de casos de enfermedades cardíacas. Esta comparativa es fundamental, ya que muestra una mejora significativa del 13 % con respecto al desempeño del modelo de *Regresión Logística*.

Un punto importante al comparar entre modelos es el costo computacional. Por esta razón, se decidió aplicar una transformación logarítmica a las variables de presión arterial en reposo, colesterol en suero y frecuencia cardíaca máxima alcanzada. Al comparar las métricas entre las Tabla 3.1 y Tabla 3.3 para el modelo de *Random Forest*, se observa que los resultados son iguales, pero el costo computacional es menor en cuestión de segundos. Esto es relevante, ya que puede marcar una gran diferencia al entrenar el modelo y ejecutarlo en aplicaciones reales. Este mismo fenómeno se puede observar en el caso de la Regresión Logística, entre las Tabla 3.2 y Tabla 3.4, donde el costo computacional se reduce significativamente manteniendo los mismos resultados en cuanto a las métricas.

Clase	Precisión	*Recall*	F1-*score*
0	0.93	0.92	0.92
1	0.92	0.93	0.93
Exactitud		0.93	
Costo computacional (en segundos)		0.1636	

Tabla 3.1: Informe de métricas Modelo Random Forest.

Clase	Precisión	*Recall*	F1-*score*
0	0.79	0.81	0.80
1	0.81	0.80	0.81
Exactitud		0.80	
Costo computacional (en segundos)		0.2088	

Tabla 3.2: Informe de métricas Modelo Regresión Logística.

Clase	Precisión	*Recall*	F1-*score*
0	0.93	0.92	0.92
1	0.92	0.93	0.93
Exactitud		0.93	
Costo computacional (en segundos)		0.1558 Seg	

Tabla 3.3: Informe de métricas Modelo *Random Forest.* con transformación logarítmica

Clase	Precisión	*Recall*	F1-*score*
0	0.79	0.81	0.80
1	0.81	0.80	0.81
Exactitud		0.80	
Costo computacional (en segundos)		0.1106	

Tabla 3.4: Informe de métricas Modelo Regresión Logística con transformación logarítmica.

3.2 *Discusión*

Al analizar los cuatro modelos, se destaca que la aplicación de transformaciones logarítmicas a las variables no compromete sus relaciones con la variable de interés, además de que simplifica el procesamiento de los modelos. En consecuencia, se optimiza significativamente el desempeño del modelo *Random Forest*, el cual demuestra una precisión del 92 %. Esta optimización se traduce en una gestión más eficiente de los recursos computacionales, logrando un balance óptimo entre la solidez del modelo y su capacidad de operar en sistemas con menor recurso computacional. Tal mejora abre la puerta a una implementación más extensiva y eficiente del modelo en diversos contextos.

4 Conclusiones y trabajo futuro

En este capítulo se presentan las conclusiones obtenidas del desarrollo de este trabajo y se plantean posibles líneas de investigación futuras.

4.1 Conclusiones

Al implementar el modelo *Random Forest* utilizando todas las variables disponibles y con la transformación logarítmica a los datos, alcanzó una exactitud del 93 %. Este resultado es notablemente superior al 80 % de precisión obtenido con la regresión logística para el mismo conjunto de datos.

Además de la mejora en la precisión, el rendimiento del modelo *Random Forest* se evaluó desde una perspectiva de eficiencia computacional. La transformación logarítmica aplicada no solo mejoró la calidad de los datos sino que también redujo el tiempo de computación, pasando de 0.16 segundos a 0.15 segundos, como se observa en las Tablas 3.1 y 3.3. Este aumento en la eficiencia demuestra la viabilidad del modelo de *Random Forest* para aplicaciones en tiempo real con el uso de dispositivos con bajos recursos computacionales, sin perder su exactitud al momento de predecir eventos de fallas cardíacas.

4.2 Trabajo futuro

El presente estudio demuestra una necesidad de expandir la base de datos, al mismo tiempo que se extiende la capacidad para almacenar, transformar y utilizar dichos datos de manera eficaz. La implementación de herramientas avanzadas como las ofrecidas por los servicios de software en la nube de Google, podrían proporcionar una solución a este dilema. Esto, a su vez, facilitaría el desarrollo de aplicaciones capaces de proporcionar resultados en tiempo real basándose en análisis clínicos fisiológicos y de nuestros dispositivos inteligentes. La idea de mejorar la detección temprana sienta el camino para la aplicación de esta metodología para la detección de otras enfermedades.

Sin embargo, es fundamental realizar investigaciones adicionales y

profundizar en el conocimiento médico para determinar la eficacia del modelo y de la base de datos utilizada en el desarrollo del mismo. Este punto resalta la importancia de la interseccionalidad, la cual es esencial para mitigar los sesgos que pueden surgir por el desconocimiento específico en el campo de estudio o de temas adyacentes. Por ende, este proyecto exige la colaboración de un equipo diverso de especialistas, que incluya no solo a médicos, sino también a profesionales farmacéuticos, expertos en ética y ciencias sociales.

Bibliografía

[1] N. Heart, "What is heart failure?." https://www.nhlbi.nih.gov/health/heart-failure, Mar. 2022.

[2] OPS, "Las enfermedades del corazón siguen siendo la principal causa de muerte en lasaméricas." https://www.paho.org/es/noticias/29-9-2021-enfermedades-corazon-siguen-siendo-principal-causa-muerte-americas, Sept. 2021.

[3] F. M. del Corazón, "El costo de las enfermedades cardiacas en américa latina supera los 30 mil millones de dólares." https://world-heart-federation.org/wp-content/uploads/2017/05/spanish-press-release.pdf, June 2016.

[4] G. D. México, "Cada año, 220 mil personas fallecen debido a enfermedades del corazón." https://www.gob.mx/salud/prensa/490-cada-ano-220-mil-personas-fallecen-debido-a-enfermedades-del-corazon, Sept. 2021.

[5] C. Reyes-Méndez, "Efectos cardiovasculares del tabaquismo." https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0028-37462019000100056, Nov. 2019.

[6] J. A. Rodrigo, "Random forest con python." https://cienciadedatos.net/documentos/py08_random_forest_python, Sept. 2023.

[7] M. SIDDHARTHA, "Heart disease dataset." https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive, Nov. 2020.

[8] J. A. Rodrigo, "Regresión logística con python." https://cienciadedatos.net/documentos/py17-regresion-logistica-python.html, Nov. 2020.

[9] scikit-learn developers, "Ensembles: Gradient boosting, random forests, bagging, voting, stacking." https://scikit-learn.org/stable/modules/ensemble.html#feature-importance-evaluation, Sept. 2021.

[10] F. Gonzales, "Regresión logística, ¿qué es y cuáles son sus aplicaciones?." https://es.linkedin.com/pulse/regresi%C3%B3n-log%C3%ADstica-qu%C3%A9-es-y-cu%C3%A1les-son-sus-francisco-gonz%C3%A1lez, Feb. 2023.

[11] T. Kanstrén, "A look at precision, recall, and f1-score." https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec, Sept. 2020.