ANÁLISIS PREDICTIVO DE FALLO CARDÍACO: COMPARACIÓN DE MODELOS DE MACHINE LEARNING

Pablo Bautista

Índice

1. Introducción

1.1. Descripción del Conjunto de Datos

El Dataset utilizado para este análisis es el "Heart Failure Prediction Dataset". Este conjunto de datos nos servirá para predecir la insuficiencia cardíaca utilizando una serie de características relacionadas con la salud cardiovascular de los pacientes.

Esta base de datos tiene 918 registros de pacientes en las que se recogen 11 características más la objetivo.

La base de datos recoge las siguientes características:

- **Age**: edad del paciente en años. Esta característica es numérica y se utiliza para evaluar la influencia de la edad en la probabilidad de insuficiencia cardíaca.

- **Sex**: sexo del paciente, representado como M (Masculino) o F (Femenino). Esta es una variable categórica que ayuda a analizar la influencia del género en la salud cardiovascular.

- **ChestPainType**: tipo de dolor en el pecho, con las siguientes categorías:

    - TA: Angina Típica

    - ATA: Angina Atípica

    - NAP: Dolor No Anginal

    - ASY: Asintomático

Esta característica categórica proporciona información sobre la severidad y el tipo de dolor en el pecho experimentado por el paciente.

- **RestingBP**: presión arterial en reposo medida en mm Hg. Es una característica numérica importante para evaluar la presión arterial de los pacientes.

- **Cholesterol**: colesterol sérico medido en mg/dl. Esta es una característica numérica que mide el nivel de colesterol en la sangre.

- **FastingBS**: azúcar en sangre en ayunas, con valores 1 (si la azúcar en sangre es mayor de 120 mg/dl) y 0 (en caso contrario). Esta variable binaria es crucial para la evaluación del riesgo cardiovascular.

- **RestingECG**: resultados del electrocardiograma en reposo, con las siguientes categorías:

    - Normal: Electrocardiograma normal

    - ST: Anomalía en la onda ST-T

    - LVH: Hipertrofia ventricular izquierda probable o definitiva

Esta característica categórica proporciona información sobre el estado eléctrico del corazón.

- **MaxHR**: frecuencia cardíaca máxima alcanzada durante el ejercicio, medida en valores numéricos entre 60 y 202. Es una medida clave del rendimiento cardíaco durante el ejercicio.

- **ExerciseAngina**: angina inducida por el ejercicio, con valores Y (Sí) o N (No). Esta variable binaria indica si el paciente experimenta dolor en el pecho durante el ejercicio.

- **Oldpeak**: depresión del segmento ST medida en mm. Es una característica numérica que se utiliza para evaluar la presencia de anomalías en el electrocardiograma durante el ejercicio.

- **ST_Slope**: pendiente del segmento ST durante el ejercicio, con las siguientes categorías:

    - Up: Pendiente ascendente

    - Flat: Pendiente plana

    - Down: Pendiente descendente

Esta característica categórica describe la forma del segmento ST, lo que puede ser indicativo de ciertas condiciones cardíacas.

- **HeartDisease**: variable objetivo, con valores 1 (Insuficiencia Cardíaca) y 0 (Normal). Esta es la variable dependiente que el modelo intentará predecir.

1.2. Fuentes de Datos y Derechos de Uso

El dataset utilizado en este proyecto es el "Heart Failure Prediction Dataset" de Soriani, Federico. (2021). Heart Failure Prediction Dataset. Kaggle.

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Este dataset es una combinación de cinco conjuntos de datos individuales sobre enfermedades cardíacas:

1. Cleveland Heart Disease Dataset
2. Hungarian Heart Disease Dataset
3. Switzerland Heart Disease Dataset
4. Long Beach VA Heart Disease Dataset
5. Stalog (Heart) Data Set

El dataset está disponible bajo la Open Database License (ODbL), que nos permite acceder y utilizar el dataset para cualquier propósito, incluyendo la investigación y el desarrollo de modelos predictivos; realizar modificaciones en el dataset, como la limpieza, transformación o extensión de los datos, siempre y cuando esté citado y no debe utilizarse para fines comerciales sin obtener permisos adicionales si fuera necesario.

1.3. Objetivos del Análisis

El objetivo principal de este proyecto es desarrollar y evaluar modelos predictivos que puedan identificar la presencia de insuficiencia cardíaca en pacientes, utilizando el "Heart Failure Prediction Dataset". Este análisis busca aportar valor al entendimiento y manejo de las enfermedades cardiovasculares mediante el uso de técnicas avanzadas aprendidas en el Máster Data Science, Big Data & Business Analytics de la Universidad Complutense de Madrid.

2. Etapas del Proceso de Modelización Analítica

2.1. Análisis Descriptivo del Conjunto de Datos

2.1.1. Estadísticas Descriptivas

- Análisis Descriptivo de las variables

Vamos a empezar realizando un análisis descriptivo de las variables ya que nos proporcionan información crucial sobre la distribución de los datos. Este análisis podemos encontrarlo en el apartado “2.1.1 Estadísticas Descriptivas” del “Anexo”, en forma de código y sus resultados.

- Edad (Age):
Tiene una media en 53.51 años, lo que indica que la mayoría de los pacientes tienen una edad media-alta, lo cual es esperable en un estudio de enfermedades cardíacas. Además de una desviación estándar de 9.43 años, mostrando una variabilidad moderada en la edad de los pacientes.

La edad mínima es de 28 años y la máxima de 77 años, cubriendo una amplia gama de edades adultas.

- Presión Arterial en Reposo (RestingBP):

La media de esta variable es de 132.4 mm Hg, lo que sugiere una tendencia hacia la hipertensión entre los pacientes, con una desviación estándar de 18.51 mm Hg. Tiene un mínimo de 0, que podría ser un valor atípico o error de entrada de datos.

- Colesterol (Cholesterol):

El colesterol tiene una media en 198.8 mg/dl, cercana al límite superior del rango normal, indicando posibles problemas de colesterol elevado en los pacientes; una desviación estándar de 109.38 mg/dl, lo que indica una alta variabilidad en los niveles de colesterol; y valores desde 0 hasta 603 mg/dl, donde un valor de 0 podría ser un error de entrada.

- Azúcar en Ayunas (FastingBS):

Su media de 0.233 indica que aproximadamente el 23% de los pacientes tienen niveles altos de azúcar en ayunas lo que sugiere un bajo riesgo de diabetes.

- Máxima Frecuencia Cardíaca Alcanzada (MaxHR):

Con media en 136.8 latidos por minuto, con una amplia dispersión desde 60 hasta 202 lpm, esto nos indica diferencias en la capacidad física de los pacientes.

- Depresión del ST (Oldpeak):

Esta variable tiene una media de 0.89, con una alta desviación estándar de 1.07, lo que indica variaciones significativas en las lecturas de depresión del ST.

- Enfermedad Cardiaca (HeartDisease):

Es una distribución binaria, la media de 0.553 indica que aproximadamente el 55.3% de los pacientes en el dataset han sido diagnosticados con una enfermedad cardíaca.

- Sexo (Sex):

En nuestra muestra hay 725 hombres (M) y 193 mujeres (F). Esto muestra un claro predominio de hombres en el dataset, lo que podría influir en los resultados y en la generalización del modelo.

- Tipo de Dolor de Pecho (ChestPainType):

El tipo de dolor Asymptomatic (ASY) es el más común, con 496 casos, seguido por Non-Anginal Pain (NAP) con 203 casos. Esto sugiere que muchos pacientes no experimentan síntomas típicos de dolor de pecho, lo que podría complicar la detección de enfermedades cardíacas.

- ECG en Reposo (RestingECG):

Normal es el más común con 552 casos, seguido por LVH (hipertrofia ventricular izquierda) con 188 casos y ST con 178 casos. La presencia de LVH y ST-T anormalidades podría ser indicativa de enfermedades cardíacas subyacentes.

- Angina Inducida por Ejercicio (ExerciseAngina):

Encontramos 547 casos negativos (N) y 371 casos positivos (Y). Esto sugiere que una cantidad significativa de pacientes experimenta angina inducida por el ejercicio, un factor de riesgo importante.

- Pendiente del ST (ST_Slope):

La mayoría tiene una pendiente plana (Flat) con 460 casos, seguido por una pendiente ascendente (Up) con 395 casos. Solo 63 casos tienen una pendiente descendente (Down), lo que podría indicar un pronóstico más severo en esos casos.

- Análisis de Correlación

A continuación, vamos a hacer un análisis de correlación para identificar las relaciones entre las variables y entender cómo se asocian con la presencia de enfermedades cardíacas. El análisis de correlación podemos encontrarlo en el apartado “2.1.1 Estadísticas Descriptivas” del “Anexo” en forma de código y su representación gráfica.

En el análisis de la correlación con la variable HeartDisease, hemos identificado diversas correlaciones que revelan patrones significativos en los datos. La variable Oldpeak presenta una correlación positiva de 0.403951, esto sugiere que un aumento en la depresión del segmento ST (Oldpeak) está asociado con una mayor probabilidad de padecer enfermedad cardíaca. ExerciseAngina_Y muestra una fuerte correlación positiva de 0.494282, lo que indica que los pacientes que experimentan angina inducida por ejercicio tienen una probabilidad significativamente mayor de sufrir de enfermedad cardíaca. La variable ST_Slope_Flat posee la correlación más alta con un valor de 0.554134, implicando que una pendiente plana del ST durante el ejercicio es un indicador relevante de riesgo cardíaco. En contraste, ST_Slope_Up presenta una correlación negativa notable de -0.622164, sugiriendo que una pendiente ascendente del ST durante el ejercicio está asociada con un menor riesgo de enfermedad cardíaca. Por último, MaxHR muestra una correlación negativa considerable de -0.400421, indicando que una mayor frecuencia cardíaca máxima se asocia con una menor probabilidad de padecer enfermedad cardíaca.

La edad muestra una correlación positiva moderada de 0.282039 con la enfermedad cardíaca, lo que coincide con el conocimiento médico general de que el riesgo de enfermedades cardíacas suele aumentar con la edad. Por otro lado, el colesterol presenta una correlación negativa de -0.232741 con la enfermedad cardíaca, aunque esta relación es menos pronunciada. Esto sugiere que, en este conjunto de datos, niveles más altos de colesterol están asociados con una menor presencia de enfermedad cardíaca, lo que podría ser indicativo de variables confusoras adicionales o características particulares del dataset.

2.1.2. Visualizaciones Gráficas

En este apartado vamos a comentar 5 diferentes tipos de gráficos. Todas las representaciones gráficas junto con el código que las genera las podemos encontrar en el apartado "2.1.2 Visualizaciones Gráficas" del "Anexo".

- Histogramas para Variables Numéricas:

Los histogramas proporcionan una herramienta invaluable para visualizar la distribución de las variables numéricas y, en este contexto, para identificar patrones que podrían estar relacionados con la cardiopatía. En el "Anexo", en el apartado "2.1.2 Visualizaciones Gráficas" podemos ver representadas varias distribuciones clave que merecen un análisis más profundo.

La variable **'Age'** muestra una distribución simétrica, lo que sugiere que podría seguir una distribución normal, con un pico centrado en los 52 años, alineada con estudios médicos que indican un aumento del riesgo de enfermedades cardíacas a partir de los 50; en cuanto a la presión arterial en reposo, la variable **'RestingBP'** también exhibe una distribución normal con un pico en 124 mm Hg, ligeramente inferior al umbral de hipertensión (120 mm Hg), sugiriendo que muchos sujetos están cerca de este límite, lo que podría indicar que la mayoría de los sujetos del estudio están en un rango cercano al umbral de hipertensión, un conocido factor de riesgo para la cardiopatía; el histograma de la variable **'Cholesterol'** también presenta una distribución normal con un pico en 199 mg/dl, aunque una moda en 0 mg/dl sugiere posibles errores o datos faltantes, ya que niveles de colesterol de cero no son posibles. Este punto merece una investigación adicional, ya que niveles de colesterol cero no son fisiológicamente posibles y podrían distorsionar la interpretación de los resultados; la variable **'FastingBS'** (nivel de azúcar en sangre en ayunas) es binaria, con valores de 0 y 1, donde el valor 0 (indicando niveles normales de azúcar) es predominante. Esto sugiere que la mayoría de los sujetos no presentan hiperglucemia en ayunas, lo cual es positivo, pero el grupo con valor 1 podría representar un segmento con riesgo elevado de complicaciones cardíacas debido a la diabetes o prediabetes; en relación con la **'MaxHR'** (frecuencia cardíaca máxima alcanzada durante el ejercicio), se observa una distribución normal con dos picos en 150 y 140 latidos por minuto. Esto podría reflejar diferencias en la respuesta cardíaca al ejercicio entre distintos subgrupos de la población, lo que podría estar vinculado a la presencia o ausencia de cardiopatía o a diferentes niveles de condición física; finalmente, la variable **'Oldpeak'** (depresión del ST) parece seguir una distribución normal, con una moda en -1. Este valor, que representa una desviación negativa en el segmento ST durante una prueba de esfuerzo, es un indicador crítico en la evaluación de isquemia miocárdica, lo que sugiere que una parte significativa de la muestra presenta algún grado de compromiso cardíaco bajo estrés.

- Gráficos de Caja (Boxplots)

El análisis comparativo de las variables Age, RestingBP, Cholesterol, FastingBS, MaxHR y Oldpeak con respecto a la variable objetivo HeartDisease (donde 1 indica insuficiencia cardíaca y 0 una condición cardíaca normal) lo he realizado utilizando gráficos de caja (boxplots) que podemos encontrar en el "apartado 2.1.2" del "Anexo". A continuación, presento los hallazgos más relevantes:

1. Variable **Age** por HeartDisease

En la variable Age, observamos que los pacientes con insuficiencia cardíaca (HeartDisease = 1) tienden a ser mayores en comparación con aquellos con un estado cardíaco normal (HeartDisease = 0). La mediana de edad en el grupo con insuficiencia cardíaca se sitúa en

aproximadamente 58 años, mientras que, en el grupo sin la enfermedad, la mediana es de 50 años. Esto sugiere que la edad avanzada es un factor de riesgo significativo para la insuficiencia cardíaca. Además, los bigotes del boxplot en ambos grupos alcanzan valores similares, indicando que la distribución de edades en ambos grupos tiene un rango amplio, aunque existen algunos valores atípicos (outliers) menores en el grupo con insuficiencia cardíaca.

2. Variable **RestingBP** por HeartDisease

La distribución de la presión arterial en reposo (RestingBP) es bastante similar entre ambos grupos. La mediana en pacientes con insuficiencia cardíaca es de 126 mm Hg, y de 125 mm Hg en pacientes sin la enfermedad, mostrando una ligera diferencia. Los bigotes de ambos grupos se extienden a rangos similares (90-180 mm Hg), pero el grupo con insuficiencia cardíaca muestra algunos valores atípicos más altos, alcanzando hasta 200 mm Hg. Esto podría indicar que, aunque la presión arterial en reposo elevada no sea un diferenciador principal entre los grupos, los casos extremos de presión arterial alta están más presentes en los pacientes con insuficiencia cardíaca.

3. Variable **Cholesterol** por HeartDisease

La variable Cholesterol presenta diferencias notables en su distribución. En pacientes con insuficiencia cardíaca, la mediana del colesterol es ligeramente más alta (220 mg/dl) que en aquellos con una condición cardíaca normal (210 mg/dl). Sin embargo, lo más destacado es la presencia de valores atípicos muy elevados en ambos grupos, especialmente en pacientes sin insuficiencia cardíaca, donde se observan niveles de colesterol inusualmente altos que llegan hasta 600 mg/dl. Este patrón sugiere que, aunque el colesterol elevado es común en ambos grupos, no es un predictor único de insuficiencia cardíaca y podría estar influido por otros factores o condiciones subyacentes.

4. Variable **FastingBS** por HeartDisease

La glucosa en ayunas (FastingBS) es una variable binaria y muestra una clara diferencia entre los grupos. Para los pacientes sin insuficiencia cardíaca, la mayoría tiene un nivel de glucosa en ayunas dentro de los límites normales (valor 0), con solo algunos casos elevados (valor 1). En contraste, el grupo con insuficiencia cardíaca muestra una mayor variabilidad, con una distribución más equilibrada entre los valores 0 y 1. Esto sugiere que niveles elevados de glucosa en ayunas podrían estar más estrechamente asociados con la insuficiencia cardíaca, posiblemente reflejando la relación entre diabetes o prediabetes y el riesgo cardíaco.

5. Variable **MaxHR** por HeartDisease

La frecuencia cardíaca máxima alcanzada durante el ejercicio (MaxHR) muestra una diferencia significativa entre los dos grupos. Los pacientes con insuficiencia cardíaca tienen una mediana de frecuencia cardíaca más baja (125 latidos por minuto) en comparación con aquellos sin la enfermedad (150 latidos por minuto). Además, los bigotes en el grupo con insuficiencia cardíaca indican una mayor dispersión hacia frecuencias más bajas, lo que sugiere una menor capacidad para alcanzar una frecuencia cardíaca alta durante el ejercicio. Esto es coherente con la capacidad reducida para el ejercicio que comúnmente se observa en pacientes con insuficiencia cardíaca.

6. Variable **Oldpeak** por HeartDisease

La variable Oldpeak, que mide la depresión del ST, muestra una diferencia notable entre los grupos. Los pacientes sin insuficiencia cardíaca tienen una mediana de Oldpeak cercana a 0, lo que indica niveles generalmente normales, mientras que los pacientes con insuficiencia cardíaca muestran una mediana más alta (1.5), indicando un mayor grado de depresión del ST durante el ejercicio, lo cual es un indicador de isquemia. Además, la presencia de múltiples valores atípicos en el grupo con insuficiencia cardíaca sugiere una mayor gravedad y variabilidad en la respuesta isquémica durante el ejercicio en estos pacientes.

Como conclusión de los gráficos de caja que he analizado, es que nos muestran que variables como la edad, la frecuencia cardíaca máxima alcanzada y el nivel de glucosa en ayunas tienen una relación significativa con la presencia de insuficiencia cardíaca, mientras que otras variables como la presión arterial en reposo y el colesterol, aunque importantes, presentan patrones más complejos.

- Gráficos de Dispersión (Scatter Plots)

He analizado cuatro gráficos de dispersión, que podemos encontrar en el "Anexo" en el "apartado 2.1.2", que relacionan diferentes variables con la presencia de insuficiencia cardíaca (HeartDisease), con el objetivo de identificar patrones y relaciones significativas entre estas variables.

1. Edad (Age) vs. Frecuencia Cardíaca Máxima (MaxHR):

El gráfico muestra que los pacientes sin insuficiencia cardíaca (HeartDisease = 0, puntos azules) están distribuidos principalmente en la parte superior izquierda, indicando frecuencias cardíacas máximas elevadas a lo largo de todas las edades. En contraste, los pacientes con insuficiencia cardíaca (HeartDisease = 1, puntos rojos) se concentran en la parte inferior derecha, con frecuencias cardíacas máximas más bajas, especialmente en edades mayores. Esto sugiere una relación negativa entre la edad y la frecuencia cardíaca máxima en pacientes con insuficiencia cardíaca, donde el aumento de edad se asocia con una disminución en la capacidad de alcanzar frecuencias cardíacas máximas.

2. Colesterol (Cholesterol) vs. Depresión del ST (Oldpeak):

Los puntos azules, correspondientes a pacientes sin cardiopatía, se concentran en un rango de colesterol moderado (100 a 350 mg/dl) y una baja depresión del ST (0 a 2), mientras que los puntos rojos, para pacientes con insuficiencia cardíaca, se distribuyen entre 150 y 350 mg/dl en colesterol y 0.25 a 3 en Oldpeak. Esto sugiere que los pacientes con insuficiencia cardíaca tienen niveles de colesterol similares a los de los pacientes sin cardiopatía, pero presentan una mayor depresión del ST, posiblemente indicando una mayor severidad de la enfermedad. También podemos observar una línea de puntos en el valor 0 de colesterol, que se eleva formando una línea ascendente sobre el eje Y, que puede deberse a la posibilidad de registros faltantes o errores en la codificación de los datos que comenté anteriormente.

3. Presión Arterial en Reposo (RestingBP) vs. Frecuencia Cardíaca Máxima (MaxHR):

Los pacientes sin insuficiencia cardíaca (puntos azules) se concentran en la parte superior del gráfico (frecuencias cardíacas más altas), mientras que los pacientes con insuficiencia cardíaca (puntos rojos) se agrupan en la parte inferior (frecuencias cardíacas más bajas). Este patrón sugiere que los pacientes con insuficiencia cardíaca tienen una capacidad reducida para alcanzar frecuencias cardíacas altas, a pesar de las variaciones en la presión arterial en reposo.

4. Depresión del ST (Oldpeak) vs. Edad (Age):

Los pacientes sin insuficiencia cardíaca (puntos azules) muestran una línea de Oldpeak de 0 a lo largo de todas las edades, indicando que una depresión del ST de 0 es común independientemente de la edad. En cambio, los pacientes con insuficiencia cardíaca (puntos rojos) muestran una depresión del ST creciente con la edad, con mayor variabilidad en valores de Oldpeak más altos. Esto sugiere que la depresión del ST tiende a ser más pronunciada en pacientes mayores con insuficiencia cardíaca.

- Gráficos de Barras para Variables Categóricas

Hemos realizado esta representación gráfica de barra, que ilustran la relación entre diferentes variables categóricas y la presencia o ausencia de insuficiencia cardíaca (HeartDisease), en el "apartado 2.1.2" del "Anexo" con el fin de visualizar la frecuencia de cada categoría en variables categóricas. Ahora vamos a detallar qué hemos observado y sacar algunas conclusiones. Como aclaración, en cada gráfico, el eje Y representa el conteo de casos, y el eje X desglosa las variables en dos categorías: barras azules para los casos sin insuficiencia cardíaca y barras naranjas para los casos con insuficiencia cardíaca.

1. Frecuencia de Sexo por HeartDisease

En los hombres (la variable M) observamos que tienen una mayor prevalencia de insuficiencia cardíaca, con 470 casos frente a 275 que no la padecen. Este hallazgo sugiere que el sexo masculino es un factor de riesgo significativo para la insuficiencia cardíaca. Contrariamente, en las mujeres (variable F), los casos sin insuficiencia cardíaca (140) superan a los casos con la condición (50). Esto sugiere que, aunque las mujeres también pueden desarrollar insuficiencia cardíaca, su incidencia es menor en comparación con los hombres. Esta diferencia puede indicar la necesidad de enfoques preventivos específicos según el sexo.

2. Frecuencia de Tipo de Dolor Torácico (ChestPainType) por HeartDisease

En el caso de la Angina Atípica (ATA) podemos sacar en conclusión que este tipo de dolor es más común en personas sin insuficiencia cardíaca, lo que sugiere que la angina atípica no es un fuerte predictor de esta condición. Para el Dolor No Anginal (NAP), la distribución más equilibrada entre los grupos con y sin insuficiencia cardíaca indica que el dolor no anginal podría estar relacionado con otros factores de riesgo, pero no necesariamente con la insuficiencia cardíaca. Los pacientes Asintomáticos (ASY) tienen una alta prevalencia de insuficiencia cardíaca, lo que sugiere que la ausencia de dolor torácico visible podría estar relacionada con un mayor riesgo de esta condición. La conclusión que sacamos al observar el comportamiento de la Angina Típica (TA) la similar prevalencia en ambos grupos indica que la angina típica podría no ser un diferenciador clave entre quienes desarrollan o no insuficiencia cardíaca.

3. Frecuencia de Electrocardiograma en Reposo (RestingECG) por HeartDisease

En el caso de la variable Normal, existe una similitud en los conteos de ambos grupos que sugiere que un electrocardiograma normal no excluye la posibilidad de insuficiencia cardíaca. En el caso de la Anomalía ST, la mayor prevalencia de insuficiencia cardíaca en personas con anomalías ST-T indica que estas anomalías son un signo importante a tener en cuenta en el diagnóstico. En el caso de la Hipertrofia Ventricular Izquierda (LVH) hay una ligera mayor prevalencia de insuficiencia cardíaca en este grupo que sugiere que la LVH podría ser un factor adicional a considerar, aunque no sea un predictor contundente por sí solo.

4. Frecuencia de Angina Inducida por Ejercicio (ExerciseAngina) por HeartDisease

La mayor cantidad de personas sin insuficiencia cardíaca en el grupo N (no) indica que la ausencia de angina inducida por ejercicio es un buen indicador de menor riesgo de desarrollar la condición. La fuerte asociación entre la angina inducida por el ejercicio (Y, sí) y la insuficiencia cardíaca sugiere que la presencia de este síntoma debería ser una señal de alerta para un mayor riesgo de insuficiencia cardíaca.

La angina inducida por ejercicio es un indicador clave de insuficiencia cardíaca, subrayando la importancia de pruebas de esfuerzo en la evaluación del riesgo cardíaco.

5. Frecuencia de Pendiente del Segmento ST (ST_Slope) por HeartDisease

Pendiente Ascendente (Up): Esta condición es común en individuos sin insuficiencia cardíaca, lo que sugiere que una pendiente ascendente en el segmento ST podría estar relacionada con un perfil de riesgo cardíaco más bajo.

Pendiente Plana (Flat): La fuerte correlación entre una pendiente plana y la insuficiencia cardíaca indica que este tipo de pendiente es un factor de riesgo importante.

Pendiente Descendente (Down): Aunque menos común, la asociación con insuficiencia cardíaca sugiere que este patrón también podría ser un indicador de riesgo a considerar.

La pendiente del segmento ST es un factor importante en la evaluación de la insuficiencia cardíaca, con la pendiente plana siendo un fuerte predictor de riesgo.

Como conclusión general, se destacan especialmente el sexo masculino, los síntomas asintomáticos y la respuesta a la prueba de esfuerzo como factores clave asociados con un mayor riesgo de desarrollar esta condición.

- Gráfico de Distribución de HeartDisease

El análisis del gráfico de distribución de la insuficiencia cardíaca indica que, en la población examinada, hay una mayor prevalencia de individuos con insuficiencia cardíaca (HeartDisease = 1) en comparación con aquellos sin la condición (HeartDisease = 0). Este hallazgo subraya la necesidad de priorizar estrategias de prevención y tratamiento en la población para reducir la incidencia de insuficiencia cardíaca, ya que afecta a una porción considerable de los individuos analizados.

2.2. Transformación del Conjunto de Datos

2.2.1. Preprocesamiento de los Datos

Las variables categóricas hay que codificarlas ya que los algoritmos de machine learning y análisis estadístico que haremos más adelante requieren que los datos de entrada sean numéricos. Estos algoritmos no pueden procesar directamente datos en formato de texto o categorías.

Como podemos observar en el apartado “2.2.1 Preprocesamiento de los Datos” del “Anexo”, lo primero que hacemos es comprobar qué variables son categóricas, para luego codificarlas. Podemos determinar que hay 5 variables categóricas, en las que ‘Sex’, ‘ExerciseAngina’ y ‘HeartDisease’ (que ya está codificada) y luego 'ChestPainType', 'RestingECG', 'ST_Slope’ tienen más de dos categorías.

Entonces hemos procedido a codificar primero las variables binarias, de la siguiente forma:

‘Sex’: Se convirtió a valores numéricos binarios, donde M (Masculino) se codificó como 1 y F (Femenino) como 0.

‘ExerciseAngina’: Se convirtió a valores numéricos binarios, donde Y (Sí) se codificó como 1 y N (No) como 0.

Luego procedemos a las variables categóricas con más de dos categorías (ChestPainType, RestingECG, y ST_Slope) que las he transformado utilizando One-Hot Encoding. Este proceso implica la creación de nuevas variables binarias para cada una de las categorías posibles dentro de estas variables, excluyendo la primera categoría de cada una para evitar la multicolinealidad. Por ejemplo:

ChestPainType: Se generaron columnas binarias para las categorías ATA, NAP, y TA, siendo TA la categoría excluida.

RestingECG: Se generaron columnas binarias para las categorías Normal y ST, siendo LVH la categoría excluida.

ST_Slope: Se generaron columnas binarias para las categorías Flat y Up, siendo Down la categoría excluida.

Finalmente, nos aseguramos que las variables numéricas en el conjunto de datos están correctamente tipificadas como valores numéricos (int o float).

Las variables categóricas se han codificado correctamente, y las variables numéricas están bien tipificadas, lo que garantiza la integridad de los datos para su uso posterior.

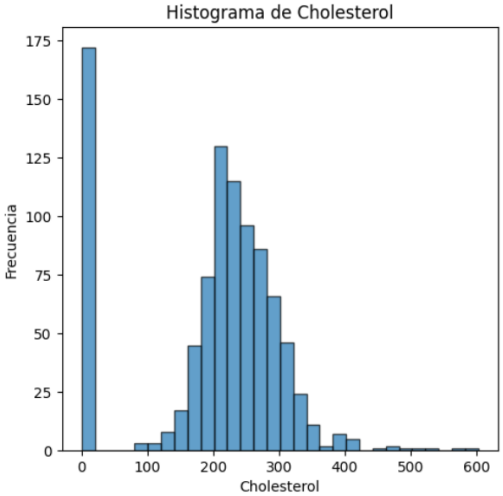
2.2.2. Tratamiento de Valores Faltantes

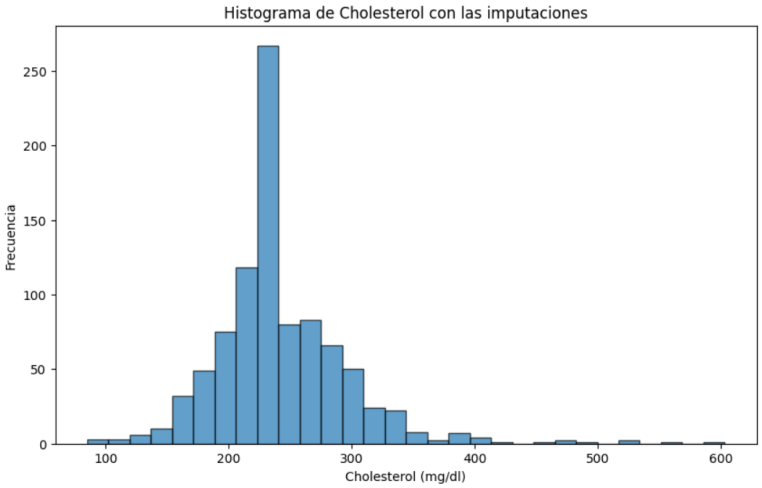
Aunque en el apartado “2.2.2 Tratamiento de Valores Faltantes” del “Anexo” el resultado es que no hay valores nulos en ninguna columna del conjunto de datos, identificamos en los histogramas del apartado anterior un patrón sospechoso en la variable 'Cholesterol'. Específicamente, se observó que el valor 0 mg/dl aparece como moda, lo cual es fisiológicamente improbable y sugiere la posibilidad de registros faltantes o errores en la codificación de los datos.

Vemos que hay 172 registros con Cholesterol igual a 0, por tanto, eliminarlos significaría perder información relevante, así que he decidido imputarlos con la mediana. El proceso de imputación lo he realizado de la siguiente manera: primero, calculamos la mediana de 'Cholesterol' excluyendo los ceros. A continuación, reemplazamos los valores de 0 mg/dl por la mediana calculada. Finalmente, se comprobamos que no quedaran registros con un valor de 0 mg/dl en la columna 'Cholesterol', asegurando así la integridad y coherencia de los datos imputados.

A la izquierda podemos ver el histograma de la variable ‘Cholesterol’ antes de la imputación, que hicimos en el apartado 2.1.2 Visualizaciones Gráficas en el Anexo y a la izquierda la variable ‘Cholesterol’ tras la imputación que hemos realizado en el “apartado 2.2.2. Tratamiento de Valores Faltantes” del “Anexo”.

La imputación de valores faltantes ha corregido la anomalía de la moda en 0 mg/dl y ha revelado una distribución que se aproxima a una forma normal, con picos notables alrededor de 200 y 237 mg/dl.





2.2.3. Escalamiento de datos

El escalamiento es un paso esencial en el preprocesamiento de datos, ya que asegura que todas las características contribuyan de manera equitativa, este proceso estandariza las características para tener una media de 0 y una desviación estándar de 1. Es un paso clave en la preparación de datos para la construcción de modelos predictivos

Para proceder a realizarlo, primero he definido las columnas numéricas, que son las que queremos aplicar las transformaciones. Para el escalamiento aplicamos el 'StandardScaler' para que cada característica tenga de media 0 y una desviación estándar de 1. Finalmente muestro una vista previa del DataFrame para verificar que se han aplicado correctamente las transformaciones.

2.3. Construcción de Modelos Predictivos

2.3.1. Técnicas de Modelización Utilizadas

Voy a utilizar en primer lugar Regresión Logística porque es una técnica simple y efectiva para problemas de clasificación binaria como este, donde se busca predecir la probabilidad de que un paciente sufra una cardiopatía. La Regresión Logística es fácil de interpretar, proporcionando coeficientes que indican la influencia de cada variable en el riesgo de cardiopatía, lo que resulta en un modelo base sólido para entender las relaciones entre las características y el resultado.

Posteriormente, aplicaré Árbol de Decisión para explorar las posibles interacciones no lineales entre las variables y capturar reglas de decisión más complejas. Los árboles de decisión son altamente interpretables, lo que facilita la visualización de cómo diferentes factores influyen en la predicción del riesgo cardíaco.

Finalmente, emplearemos Random Forest, que, al combinar múltiples árboles de decisión, mejora la precisión del modelo y reduce el riesgo de sobreajuste.

2.3.1.1. Regresión Logística

La primera técnica de modelización utilizada es la Regresión Logística, un enfoque adecuado para problemas de clasificación binaria, como la predicción de la presencia o ausencia de cardiopatías en los pacientes.

Hemos dividido el conjunto de datos en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%) para evaluar el rendimiento del modelo. El modelo de regresión logística se ha entrenado utilizando las características preprocesadas. Tras el entrenamiento, el modelo fue evaluado en el conjunto de prueba, obteniendo una precisión (accuracy) del 86%.

La matriz de confusión generada muestra que el modelo identificó correctamente 68 casos negativos (ausencia de enfermedad) y 91 casos positivos (presencia de enfermedad). Sin embargo, hubo 9 falsos positivos y 16 falsos negativos.

El informe de clasificación proporciona métricas adicionales:

- Precisión (precision): El modelo obtuvo una precisión de 0.81 para la clase negativa (ausencia de enfermedad) y 0.91 para la clase positiva (presencia de enfermedad).

- Recall (sensibilidad): La sensibilidad fue de 0.88 para la clase negativa y 0.85 para la clase positiva.

- F1-score: El F1-score, que considera tanto la precisión como la sensibilidad, fue de 0.84 para la clase negativa y 0.88 para la clase positiva.

En conjunto, el modelo de regresión logística ha demostrado ser eficaz para esta tarea de clasificación, con un rendimiento general robusto. Estos resultados proporcionan una base sólida para la predicción de cardiopatías, aunque se espera mejorar aún más la precisión y reducir los errores de clasificación utilizando técnicas adicionales como los árboles de decisión y Random Forest en las siguientes secciones del análisis.

La regresión logística es fácil de entender e interpretar. Es un modelo relativamente rápido en términos de tiempo de entrenamiento y predicción, lo que puede ser importante en aplicaciones en tiempo real y es más adecuada para problemas de clasificación binaria, como es nuestro caso. Aunque asume una relación lineal entre las características independientes y la variable dependiente.

2.3.1.2. Árboles de Decisión

En este apartado aplicamos un modelo de Árbol de Decisión para la predicción de la cardiopatía en pacientes.

Este modelo alcanzó una precisión (accuracy) del 81% en el conjunto de prueba. Esto indica que el modelo fue capaz de clasificar correctamente el 81% de los casos en el conjunto de datos.

La matriz de confusión revela que el modelo identificó correctamente 64 pacientes sin cardiopatía y 85 pacientes con presencia de cardiopatía. Frente a 13 falsos positivos (clasificados erróneamente como portadores de cardiopatía) y 22 falsos negativos (clasificados erróneamente como sanos).

El informe de clasificación proporciona métricas adicionales:

- Precisión: La precisión para la clase 0 (sin cardiopatía) es 0.74, lo que significa que el 74% de los pacientes predichos como sanos realmente lo son. Para la clase 1 (con cardiopatía), la precisión es del 87%.

- Recall: El recall para la clase 0 es del 83%, mientras que para la clase 1 es del 79%. Esto indica que el modelo es más efectivo en la identificación de pacientes sanos que en la detección de aquellos con cardiopatía.

- F1-Score: El F1-score para la clase 0 es 0.79 y para la clase 1 es 0.83, lo que refleja un equilibrio entre la precisión y el recall.

Los resultados obtenidos del modelo de Árbol de Decisión demuestran un rendimiento aceptable, aunque con áreas de mejora, especialmente en la detección de pacientes con cardiopatía (clase 1). Este modelo puede servir como un punto de partida para explorar técnicas más avanzadas, como Random Forest, que combinan múltiples árboles de decisión para mejorar la precisión y la robustez del modelo.

Los resultados muestran que el modelo de Regresión Logística supera al modelo de Árbol de Decisión en varios aspectos, incluyendo la precisión general, la capacidad de detección de casos positivos y la clasificación de pacientes sanos. Esto nos sirve para construir un modelo predictivo confiable para la detección de cardiopatías.

Dado este rendimiento superior, se recomienda considerar la regresión logística como un modelo base para futuras comparaciones con técnicas más complejas, como Random Forest que exploraré en la siguiente sección.

Los árboles de decisión son fáciles de entender y visualizar, lo que permite una interpretación clara de cómo se toman las decisiones y puede capturar relaciones no lineales y interacciones entre las características de los datos. En cambio, tienen una alta tendencia a sobreajustar los datos y un pequeño cambio en los datos pueden resultar en una estructura de árbol completamente diferente, lo que puede afectar la robustez del modelo.

2.3.1.3. Random Forest

El modelo Random Forest es un algoritmo de aprendizaje automático basado en la técnica de ensamble, que combina múltiples árboles de decisión para mejorar la precisión y la estabilidad de las predicciones. Este enfoque es particularmente eficaz para resolver problemas de clasificación, como el que estamos abordando, donde buscamos predecir la presencia de enfermedad cardíaca. Random Forest ayuda a mitigar problemas como el sobreajuste, proporcionando un modelo más robusto.

El modelo logró una precisión (accuracy) del 88%, indicando que el 88% de las predicciones realizadas fueron correctas en el conjunto de datos de prueba. La matriz de confusión mostró que hubo 65 verdaderos negativos (pacientes sin enfermedad cardíaca correctamente identificados), 12 falsos positivos (pacientes sanos incorrectamente clasificados como enfermos), 11 falsos negativos (pacientes enfermos incorrectamente clasificados como sanos) y 96 verdaderos positivos (pacientes enfermos correctamente identificados). El informe de clasificación reveló que para la clase 0 (sin enfermedad cardíaca), la precisión fue de 0.86, el

recall fue de 0.84 y el F1-score fue de 0.85. Para la clase 1 (con enfermedad cardíaca), la precisión fue de 0.89, el recall fue de 0.90 y el F1-score fue de 0.89.

En conclusión, los resultados obtenidos del modelo Random Forest indican que es eficaz en la predicción de la enfermedad cardíaca, superando el rendimiento de los modelos anteriores (regresión logística y árbol de decisión). El modelo no solo nos ha mostrado una alta precisión general, sino también un buen equilibrio entre precisión y recall para ambas clases, lo que sugiere que Random Forest es una herramienta valiosa para ayudar en la identificación de pacientes en riesgo de enfermedad cardíaca, facilitando intervenciones tempranas y mejorando los resultados de salud.

Este modelo permite evaluar la importancia de cada característica en la predicción, lo que puede ser útil para la selección de variables y es menos susceptible al sobreajuste que los árboles de decisión individuales, ya que promedia múltiples árboles para obtener resultados más estables. Sin embargo, es más complejo y menos interpretable que los modelos de regresión logística y los árboles de decisión individuales, lo que puede dificultar la explicación de los resultados a los interesados.

2.3.2. Justificación para la Selección de Cada Técnica

Para la selección de las técnicas de modelización me he basado en las características específicas del conjunto de datos y los objetivos del análisis. Elegí la regresión logística por su simplicidad y capacidad para ofrecer interpretaciones claras de los coeficientes. He utilizado el árbol de decisión por su capacidad para manejar interacciones y relaciones no lineales en los datos, permitiendo un análisis más detallado. Finalmente, Random Forest por su robustez y capacidad para mejorar la precisión a través de la combinación de múltiples árboles de decisión, lo que reduce el riesgo de sobreajuste.

2.3.3. Evaluación del Desempeño de los Modelos

Para evaluar el desempeño de los modelos, he utilizado varias métricas, incluidas la precisión, el recall y el F1-score. Estas métricas permiten una comprensión más completa de cómo cada modelo clasifica correctamente a los pacientes con y sin enfermedad cardíaca. Además, calculé una matriz de confusión para visualizar los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos de cada modelo.

2.3.3.1. Métricas de Evaluación

En el ámbito de la predicción de cardiopatías, es crucial evaluar la efectividad de los modelos a través de métricas que reflejen su capacidad para clasificar correctamente a los pacientes. Las métricas de evaluación que he utilizado son:

- Precisión (Accuracy): nos indica la proporción de predicciones correctas que el modelo ha realizado sobre el total de predicciones realizadas. Una alta precisión significa que el modelo está acertando en una gran parte de sus predicciones, ya sea al identificar correctamente a los pacientes con y sin enfermedad cardíaca.

- Recall (Sensibilidad): refleja la capacidad del modelo para identificar correctamente a los pacientes que realmente tienen la enfermedad cardíaca. Un alto recall es especialmente importante en este contexto, ya que significa que el modelo no está pasando por alto a muchos pacientes enfermos.

- F1-Score: esta métrica combina la precisión y el recall en un solo valor, proporcionando un balance entre ambos. Es especialmente útil cuando hay un desequilibrio en las clases, ya que tiene en cuenta tanto los falsos positivos como los falsos negativos. En el contexto de la predicción de cardiopatías, un alto F1-score indica que el modelo tiene un buen rendimiento tanto en la identificación de pacientes con cardiopatías como en la minimización de diagnósticos erróneos.

2.3.3.2. Comparativa de Modelos

A continuación, presento una comparación de los modelos según las métricas evaluadas.

MODELO	PRECISIÓN	RECALL	F1-SCORE
Regresión Logística	0.86	0.85	0.86
Árbol de Decisión	0.81	0.79	0.80
Random Forest	0.88	0.90	0.89

- Regresión Logística: Con una precisión de 0.86, el modelo muestra un buen desempeño general. Su recall de 0.85 indica que es eficaz en identificar pacientes con cardiopatías, aunque hay margen de mejora en no perder casos de pacientes enfermos.

- Árbol de Decisión: Este modelo presenta una precisión más baja de 0.81 y un recall de 0.79. Esto sugiere que, aunque el modelo puede clasificar correctamente a un número razonable de pacientes, tiene un desempeño inferior en la identificación de todos los pacientes enfermos, lo que podría resultar en diagnósticos erróneos.

- Random Forest: Este modelo destaca con la mejor precisión (0.88) y un recall superior (0.90). Esto indica que no solo acierta en la clasificación de un mayor número de pacientes, sino que también tiene una excelente capacidad para detectar a los pacientes con cardiopatías, minimizando el riesgo de falsos negativos.

En resumen, las métricas de evaluación demuestran que, aunque todos los modelos tienen su propio valor, el Random Forest sobresale como la técnica más eficaz para la predicción de cardiopatías, ofreciendo un equilibrio ideal entre precisión y sensibilidad.

2.4. Informe Final de Conclusiones

2.4.1. Resumen de los Hallazgos

En este análisis, he desarrollado varios modelos predictivos para identificar la probabilidad de que un paciente sufra de cardiopatía. He aplicado tres técnicas de modelización principales: Regresión Logística, Árboles de Decisión, y Random Forest.

El modelo de Regresión Logística fue utilizado como modelo base debido a su simplicidad y capacidad para ofrecer interpretaciones claras de la influencia de cada variable en la predicción de la enfermedad. Este modelo obtuvo una precisión del 86%, con un recall del 85% y un F1-score de 0.86. Estos resultados indican que el modelo es eficaz para predecir correctamente la presencia de cardiopatía en la mayoría de los casos.

El modelo de Árbol de Decisión lo apliqué con el objetivo de explorar un modelo más flexible y no lineal que pudiera capturar interacciones complejas entre las variables. Sin embargo, la precisión del Árbol de Decisión fue ligeramente inferior, con un 81%, lo que sugiere que,

aunque es un modelo interpretativo y fácil de visualizar, no logró superar el rendimiento de la regresión logística en este conjunto de datos.

Por último, he implementado el modelo Random Forest, que combina múltiples árboles de decisión para mejorar la precisión y robustez del modelo. Este enfoque resultó ser el más efectivo, alcanzando una precisión del 88%, con un recall del 90% y un F1-score de 0.89. Estos resultados destacan la capacidad del Random Forest para manejar la variabilidad y evitar el sobreajuste, al mismo tiempo que proporciona un modelo predictivo más sólido.

Como conclusión, aunque todos los modelos han demostrado ser útiles para la predicción de cardiopatía, Random Forest destacó por ofrecer el mejor rendimiento general en términos de precisión y recall, lo que lo convierte en una herramienta potente para la predicción de enfermedades cardiovasculares en este contexto.

2.4.2. Mejoras Implementadas por los Modelos

A lo largo del proceso de modelización, se implementaron varias mejoras para optimizar el rendimiento y la capacidad predictiva de los modelos:

- **Imputación de Valores Faltantes**: abordé un problema crítico en los datos relacionados con la variable de colesterol, donde se identificaron valores de 0 mg/dl que, fisiológicamente, no son posibles. En lugar de eliminar estos registros y perder datos potencialmente valiosos, se optó por imputarlos utilizando la mediana, lo que mejoró la calidad y la integridad del conjunto de datos.

- **Escalamiento de Variables**: para garantizar que todos los modelos pudieran manejar correctamente las características en escalas diferentes, se aplicó el escalamiento estándar a las variables numéricas. Esto fue útil para la regresión logística, que es sensible a la escala de los datos, permitiendo un entrenamiento más efectivo.

- **Uso de Modelos Avanzados**: tras la implementación inicial de la regresión logística como modelo base introduje modelos más avanzados como el Árbol de Decisión y Random Forest. Estos permitieron capturar relaciones no lineales y complejas entre las variables, lo que mejoró significativamente la capacidad predictiva, particularmente en el caso de Random Forest.

- **Optimización de Hiperparámetros**: en el caso del modelo Random Forest, se llevó a cabo una optimización de los hiperparámetros, como la cantidad de árboles y la profundidad máxima de cada árbol. Esto nos ayudó a mejorar la precisión y la robustez del modelo, reduciendo el riesgo de sobreajuste y mejorando la generalización a nuevos datos.

2.4.3. Sugerencias para Mejoras Futuras

Aunque los modelos desarrollados han demostrado un rendimiento sólido, existen varias áreas que podrían mejorarse para aumentar aún más la precisión y la utilidad de las predicciones:

- Incorporación de Nuevas Variables: introducir más características relevantes, como datos genéticos, historial familiar detallado, y hábitos de vida (por ejemplo, dieta, ejercicio, consumo de alcohol), podría mejorar la capacidad predictiva de los modelos. Estas variables podrían capturar factores de riesgo adicionales que no fueron considerados en el presente estudio.

- Recopilación de Más Datos: ampliar la base de datos con más registros de pacientes, con aquellos con diagnósticos recientes y de diferentes regiones geográficas. Una mayor cantidad de datos también permitiría una mejor representación de la población y reduciría el sesgo en el modelo.

- Explorar Técnicas de Ensamblado Más Sofisticadas: aunque Random Forest es un modelo de ensamblado efectivo, se podrían investigar métodos de ensamblado más avanzados, como Boosting o Stacking. Estas técnicas podrían mejorar la precisión del modelo y manejar de manera más efectiva los desequilibrios en los datos.

- Validación Más Completa: usando métodos de validación más robustos para asegurarnos de que los resultados del modelo sean consistentes y fiables.

Estas sugerencias para mejoras futuras podrían llevar a desarrollos significativos en la precisión y la utilidad clínica de los modelos predictivos para la identificación de enfermedades cardiovasculares.

3. Implementación del Modelo en Producción

3.1. Desarrollo de una Aplicación Empresarial para Predicción

El siguiente paso es trasladar los modelos predictivos desarrollados a un entorno de producción mediante la creación de una aplicación empresarial. Esta aplicación permitirá a los usuarios, como profesionales de la salud o administradores de hospitales, ingresar datos de pacientes y obtener predicciones sobre el riesgo de cardiopatía en tiempo real.

La aplicación podría ser una web sencilla, accesible desde cualquier dispositivo con conexión a internet. Esta permitiría a los usuarios introducir datos relevantes, como edad, presión arterial, colesterol, entre otros datos, y el modelo predictivo calcularía inmediatamente el riesgo de cardiopatía, proporcionando una predicción clara y fácil de interpretar. Esta predicción podría acompañarse de una visualización que muestre los factores de riesgo individuales y cómo estos contribuyen a la predicción final.

Esta aplicación se encargaría de procesar los datos de entrada, pasarlos a través del modelo predictivo previamente entrenado, y devolver el resultado al usuario.

3.2. Proceso para Predicciones en Tiempo Real

El proceso para realizar predicciones en tiempo real dentro de la aplicación empresarial debe estar optimizado para garantizar rapidez y precisión, sin comprometer la seguridad y la privacidad de los datos. A continuación, describiremos un flujo típico para implementar predicciones en tiempo real:

1. Entrada de Datos: el usuario introduce los datos del paciente a través de la interfaz de la aplicación, que puede incluir información como edad, presión arterial, colesterol, entre otros. Esta información es validada en el lado del usuario para garantizar que los campos sean correctos y estén completos antes de enviarse al servidor.

2. Preprocesamiento: una vez que los datos son recibidos por el servidor, pasan por una etapa de preprocesamiento. Esto incluye la normalización o escalamiento de las variables, así como

la transformación de variables categóricas en formato adecuado para el modelo, asegurando que los datos estén en la misma forma que durante el entrenamiento del modelo.

3. Predicción: los datos preprocesados se introducen en el modelo predictivo (por ejemplo, el Random Forest entrenado previamente). El modelo, alojado en el servidor, procesa los datos y genera una predicción sobre el riesgo de cardiopatía del paciente en cuestión.

4. Post-procesamiento e Interpretación: el resultado de la predicción se post-procesa para ser presentado de manera comprensible para el usuario final. Puede ser una visualización del riesgo en una escala de bajo a alto, además de mostrar los factores de riesgo más significativos que contribuyen a la predicción.

5. Respuesta al Usuario: la predicción procesada se envía de vuelta a la interfaz del usuario de la aplicación, donde se presenta al usuario de forma inmediata. Esto permite que los sanitarios tomen decisiones informadas en tiempo real basadas en el análisis predictivo del modelo.

6. Registro y Monitoreo: cada predicción realizada debe quedar registrada para mejorar el modelo con el tiempo. Además, el sistema debe contar con un monitoreo para detectar posibles debilitaciones en el rendimiento del modelo debido a cambios en los datos subyacentes.

Al implementar un modelo predictivo en producción, es necesario llevar a cabo un mantenimiento continuo para monitorear y ajustar el modelo ante cambios en los datos, planificar reentrenamientos periódicos para mantener su precisión, y asegurar la escalabilidad del sistema. La seguridad y el cumplimiento normativo son necesarios para proteger datos sensibles. Además, es crucial contar con planes de respaldo y recuperación, es decir, copias de seguridad de los datos y del modelo, así como una interfaz de usuario intuitiva para asegurar una experiencia fluida y efectiva.

4. Conclusiones y Reflexiones Finales

4.1. Impacto del Análisis en el Contexto Empresarial

El análisis y desarrollo del modelo predictivo han demostrado ser de gran relevancia para el contexto empresarial. En nuestro caso, en el ámbito sanitario, al predecir la probabilidad de que un paciente sufra una cardiopatía, la empresa puede ofrecer servicios más personalizados y preventivos.

- Mejora en la Toma de Decisiones: el modelo predictivo permite a los profesionales de la salud tomar decisiones más precisas. Esto puede reducir el riesgo de errores de diagnósticos y mejorar la atención médica, que se ve traducido en un aumento de la confianza y satisfacción del paciente.

- Optimización de Recursos: con la capacidad de identificar pacientes en alto riesgo, la empresa puede optimizar el uso de sus recursos, dirigiendo esfuerzos preventivos y terapéuticos hacia aquellos que más lo necesitan. Esto no solo mejora los resultados de salud, sino que también reduce los costos operativos asociados con tratamientos innecesarios o ineficaces.

- Ventaja Competitiva: en un mercado altamente competitivo, la capacidad de utilizar modelos predictivos avanzados para anticipar problemas de salud permite a la empresa una ventaja

sobre sus competidores. La innovación en la atención médica a través del uso del análisis predictivo puede atraer a más clientes y fortalecer la posición de la empresa en el mercado.

- Personalización de Servicios: el modelo permite a la empresa ofrecer un enfoque más personalizado en el cuidado del paciente, ajustando recomendaciones y tratamientos a las necesidades individuales. Este nivel de personalización no solo mejora los resultados de salud, sino que también refuerza la lealtad del cliente y la percepción positiva de la clínica.

Por tanto, la empresa no solo mejora sus capacidades de toma de decisiones y optimización de recursos, sino que también se posiciona como un líder en innovación dentro del sector salud.

4.2. Lecciones Aprendidas durante el Proceso

A lo largo de este proyecto, hemos obtenido ciertas lecciones que ayudar a la implementación de futuros modelos predictivos:

- Importancia de los Datos de Calidad: lo más importante es contar con datos de alta calidad. Durante el proceso, he tenido que tratar los datos como en la imputación de valores faltantes y o el escalamiento de datos, que son cruciales para garantizar que los modelos predictivos funcionen de manera óptima y produzcan resultados precisos y confiables.

- Selección Adecuada de Modelos: la comparación entre diferentes técnicas de modelización, como la regresión logística, los árboles de decisión y el Random Forest, nos reveló la importancia de seleccionar el modelo adecuado para el problema específico. Cada técnica tiene sus ventajas y limitaciones, y es crucial entender el contexto y los datos antes de elegir un enfoque.

- Equilibrio entre Precisión y Complejidad: al desarrollar y evaluar los modelos, me di cuenta que es necesario equilibrar la precisión con la complejidad del modelo. Mientras que Random Forest (modelo más complejo) ofrece una mayor precisión, también requiere más recursos computacionales y es más difícil de interpretar. Encontrar el equilibrio adecuado es clave para desarrollar soluciones que sean efectivas y prácticas para su implementación en un entorno empresarial.

- Valor de la Interpretabilidad: los modelos deben ser comprensibles para los profesionales médicos que los utilizarán, por tanto, no solo hay que considerar el rendimiento técnico, sino también cómo se explican y se utilizan los resultados.

- Iteración y Mejora Continua: el proceso de desarrollo del modelo ha sido iterativo, con constantes ajustes y mejoras basados en los resultados obtenidos. Es muy importante adoptar un enfoque ágil y flexible, donde se prueban diferentes enfoques y se aprende de los resultados para mejorar continuamente la solución.

En conclusión, el proceso nos ha proporcionado una serie de aprendizajes que no solo mejoran nuestra comprensión de la predicción de cardiopatías, sino que también nos ha proporcionado nuevas habilidades para futuros proyectos.

5. Anexo

El Anexo corresponde al archivo también adjunto en la actividad en formato .html.

Aquí podemos encontrar el código empleado para desarrollar el proyecto, además encontraremos todos los gráficos que hemos descritos a lo largo de los distintos apartados del trabajo.

6. Bibliografía

El dataset utilizado es "Heart Failure Prediction Dataset" de Soriani, Federico. (2021). Heart Failure Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. https://hastie.su.domains/Papers/ESLII.pdf

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320. https://academic.oup.com/jrsssb/article/67/2/301/7109482?login=false

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. https://link.springer.com/article/10.1023/A:1010933404324

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

Detrano, R., et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304-310. https://pubmed.ncbi.nlm.nih.gov/2756873/