

Trabajo Final - Análisis y Modelado de Datos con Orange Data Mining

Joaquín Llapur, Nicolás Cozzo, Santiago Lazaroni, Juan Bourel y Juan Cruz Vidal.

Definición del Problema

El objetivo de este trabajo es desarrollar un modelo predictivo que permita determinar la presencia de enfermedad cardíaca en un paciente, utilizando diversas variables clínicas y demográficas.

Este problema se sitúa en el ámbito de la salud preventiva, dado que las enfermedades cardíacas son una de las principales causas de mortalidad a nivel global. La capacidad de anticipar el riesgo a partir de datos médicos básicos, como la edad, la presión arterial, los niveles de colesterol o el tipo de dolor de pecho, facilita diagnósticos tempranos y una optimización de los recursos en la atención médica.

En este contexto, la pregunta central que guía este estudio es:

¿Es posible predecir la presencia de enfermedad cardíaca en un paciente basándose en sus características clínicas y hábitos de vida?

El propósito es construir un modelo de aprendizaje automático que, a partir del **Heart Failure Prediction Dataset**, sea capaz de **clasificar** a los pacientes en dos categorías:

- **1 → Tiene enfermedad cardíaca**
- **0 → No tiene enfermedad cardíaca**

Además, se busca evaluar qué variables tienen **mayor influencia** en la predicción y comparar distintos modelos de clasificación para determinar cuál ofrece el mejor rendimiento.

Análisis Exploratorio de Datos:

El objetivo de esta etapa es identificar patrones y relaciones entre las variables y la presencia de enfermedad cardíaca (*HeartDisease*), utilizando estadísticas descriptivas y gráficos obtenidos con *Orange Data Mining*.

Descripción de las variables del conjunto de datos:

El conjunto de datos proviene de un estudio de predicción de enfermedades cardíacas. Incluye 918 observaciones y 12 variables que describen las características clínicas de los pacientes.

A continuación, se detallan las variables más relevantes y su interpretación:

- La variable **Age** (edad del paciente en años) se asocia directamente con el riesgo de desarrollar enfermedades cardíacas a medida que aumenta.
- **Sex** identifica el género del paciente (*M* para masculino y *F* para femenino). Históricamente, se observa una mayor incidencia de enfermedad coronaria en hombres.
- **ChestPainType** clasifica el tipo de dolor torácico: *ATA* (angina típica), *NAP* (angina no típica), *ASY* (asintomático) y *TA* (dolor no anginoso). Estos tipos reflejan distintos grados de riesgo cardiovascular.
- **RestingBP** indica la presión arterial en reposo (en mmHg). Valores elevados pueden sugerir hipertensión, un factor de riesgo importante.
- **Cholesterol** muestra el nivel de colesterol total (en mg/dL). Niveles altos incrementan el riesgo de enfermedad cardiovascular.
- **FastingBS** indica si el nivel de glucosa en sangre en ayunas supera los 120 mg/dL, lo que puede estar relacionado con patologías metabólicas que afectan el sistema cardiovascular.
- **RestingECG** describe el resultado del electrocardiograma en reposo, que ayuda a detectar irregularidades en la actividad eléctrica del corazón.
- **MaxHR** es la frecuencia cardíaca máxima alcanzada durante el ejercicio. Valores bajos pueden indicar bajo rendimiento cardíaco o limitaciones en la capacidad de esfuerzo.
- **ExerciseAngina** registra si el paciente presenta angina durante el ejercicio, siendo un indicador directo de obstrucciones coronarias.
- **Oldpeak** refleja la depresión del segmento ST durante el esfuerzo físico, expresada en milímetros. Su incremento suele asociarse con anomalías cardíacas.
- **ST_Slope** muestra la pendiente del segmento ST durante el ejercicio (*Up*, *Flat* o *Down*), una variable útil para analizar la recuperación del corazón después del esfuerzo.

Finalmente, **HeartDisease** es la variable objetivo del estudio, donde *1* indica presencia de enfermedad cardíaca y *0* su ausencia.

Estadísticas generales del conjunto de datos:

Antes de analizar las variables más relevantes, se observaron las estadísticas descriptivas generales mediante el módulo *Feature Statistics*.

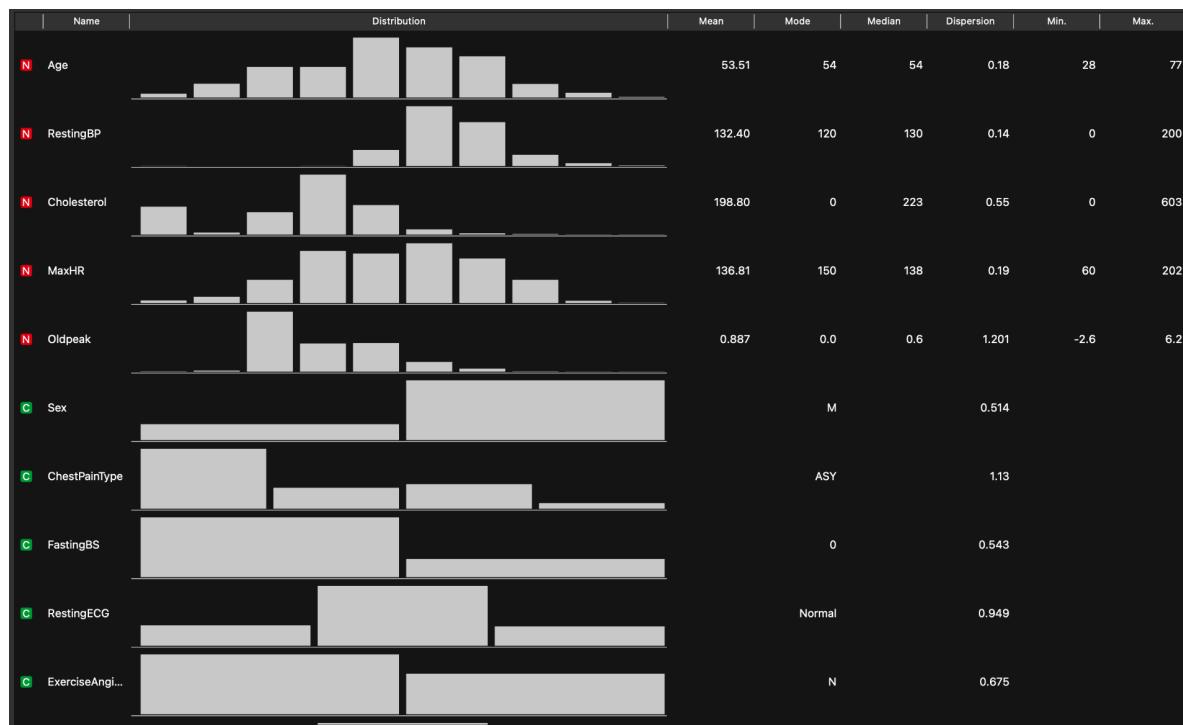
Esto permitió identificar la media, mediana, desviación estándar y rango de cada variable, además de confirmar que no existían valores faltantes.

Entre los principales resultados, la edad promedio de los pacientes es de aproximadamente 53.5 años, con un rango entre 28 y 77 años.

La presión arterial en reposo tiene una media de 132 mmHg, sin valores faltantes.

El colesterol muestra una alta variabilidad, con valores que van desde 0 hasta más de 600 mg/dL, lo que sugiere la presencia de posibles casos atípicos.

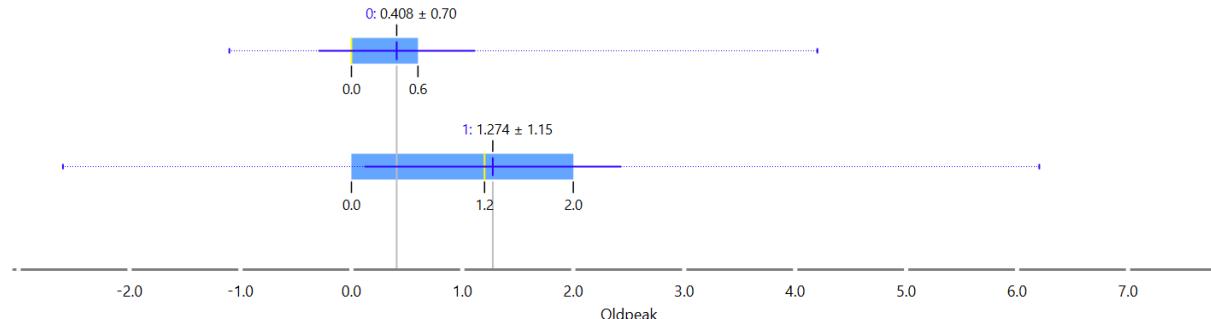
El valor medio de *Oldpeak* (descenso del segmento ST) es 0.88, con valores extremos que llegan hasta 6.2.



Relación entre variables numéricas y enfermedad cardíaca:

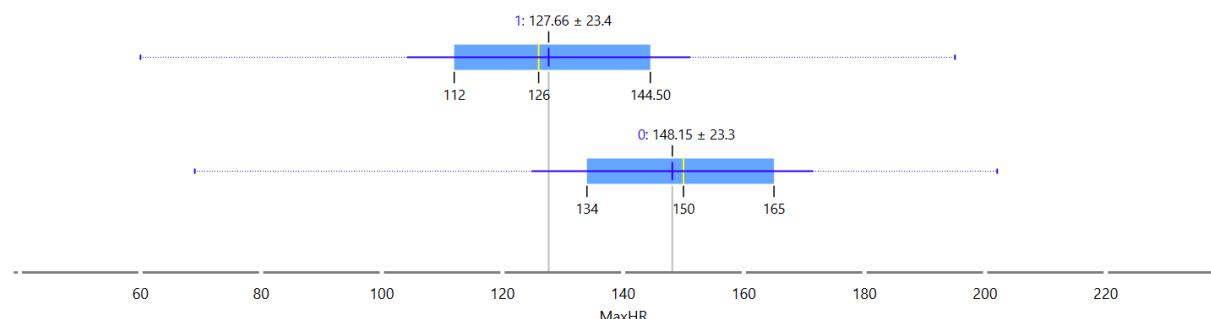
Para observar cómo varían las variables cuantitativas entre pacientes con y sin enfermedad cardíaca, se utilizaron gráficos de tipo *Box Plot*, tomando *HeartDisease* como variable de referencia.

1. **Oldpeak** mostró una clara diferencia. Los pacientes con enfermedad cardíaca presentan un valor medio de 1.27 ± 1.15 , mientras que los pacientes sin enfermedad tienen un promedio menor, de 0.40 ± 0.70 . Esto evidencia que una mayor depresión del segmento ST se asocia con la presencia de enfermedad cardíaca.



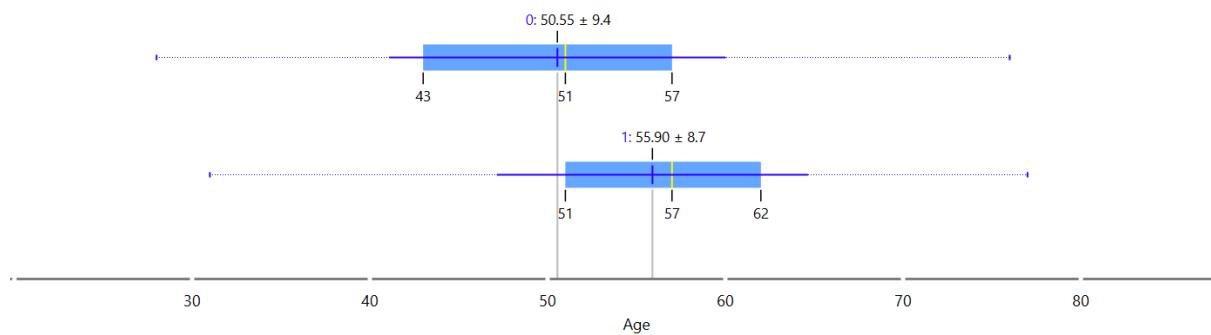
Student's t: 14.055 ($p=0.000$, N=918)

2. En cuanto a **MaxHR** (frecuencia cardíaca máxima alcanzada), el comportamiento es opuesto: los pacientes sin enfermedad alcanzan frecuencias cardíacas más altas (media 148.1 ± 23.3), mientras que los que presentan enfermedad cardíaca alcanzan valores más bajos (media 127.6 ± 23.4). Esto sugiere que los individuos con enfermedad cardíaca tienen una menor capacidad de respuesta cardiovascular al esfuerzo físico.



Student's t: 13.246 ($p=0.000$, N=918)

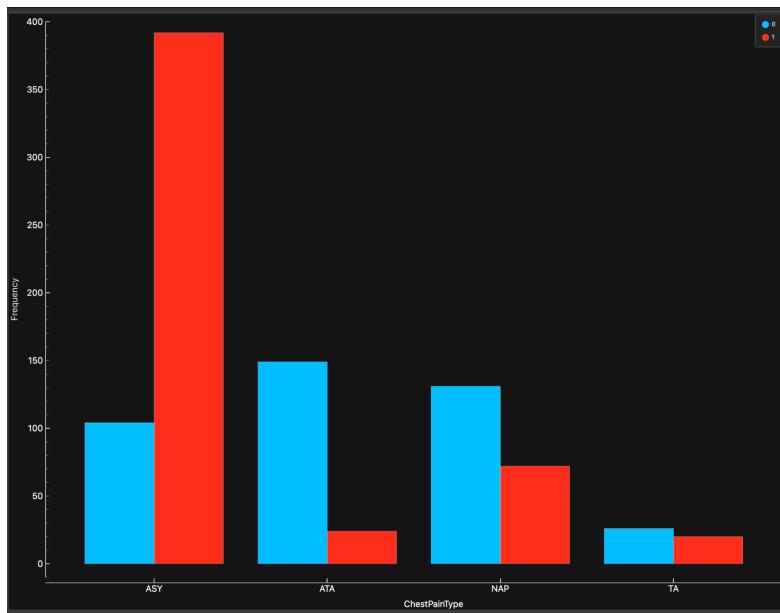
3. Finalmente, **Age** también mostró diferencias significativas. Los pacientes con enfermedad cardíaca tienen una edad promedio de 55.9 años, mientras que los pacientes sin diagnóstico tienen una media de 50.6 años. Esto confirma que la edad es un factor de riesgo relevante, mostrando una tendencia positiva entre mayor edad y mayor probabilidad de enfermedad.



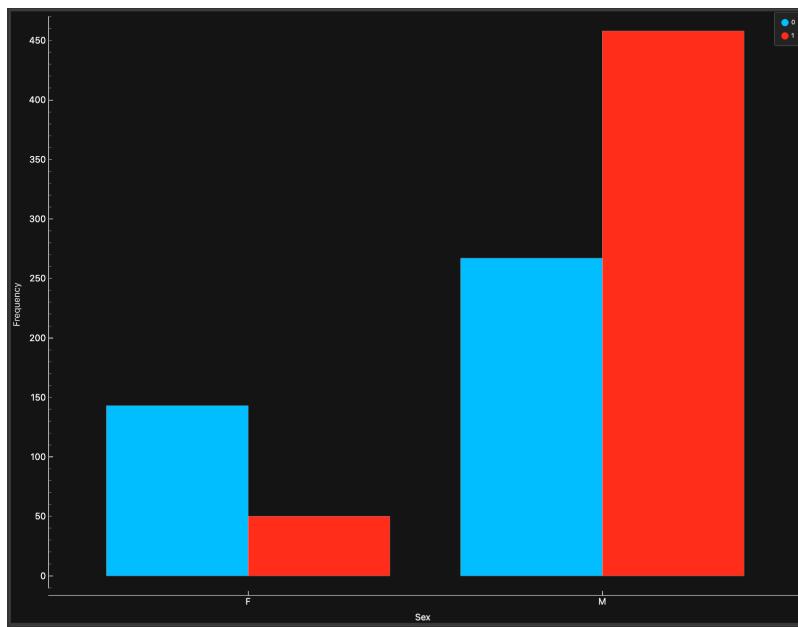
Relación entre variables categóricas y enfermedad cardíaca:

Las variables categóricas se analizaron mediante gráficos de distribución para identificar asociaciones entre las categorías y el diagnóstico de enfermedad cardíaca.

En el caso de **ChestPainType**, se observó que el tipo de dolor ASY (asintomático) predomina entre los pacientes con enfermedad cardíaca, mientras que los tipos ATA y NAP son más frecuentes en los pacientes sin diagnóstico. Esto sugiere que la ausencia de dolor típico o la presencia de síntomas atípicos se asocia con una mayor probabilidad de enfermedad cardíaca.



Respecto a **Sex**, se evidenció que los hombres presentan una mayor proporción de enfermedad cardíaca en comparación con las mujeres, lo que coincide con tendencias epidemiológicas reconocidas.



Conclusiones del análisis exploratorio

El análisis exploratorio permitió detectar patrones claros entre las variables y la presencia de enfermedad cardíaca.

Se concluye que existen diferencias significativas entre pacientes con y sin diagnóstico, destacándose las variables *Oldpeak*, *MaxHR*, *Age*, *ChestPainType* y *Sex* como las más relevantes.

Los pacientes de mayor edad, con menor frecuencia cardíaca máxima, mayor depresión del segmento ST, síntomas asintomáticos o atípicos, y de sexo masculino, presentan una mayor probabilidad de padecer enfermedad cardíaca.

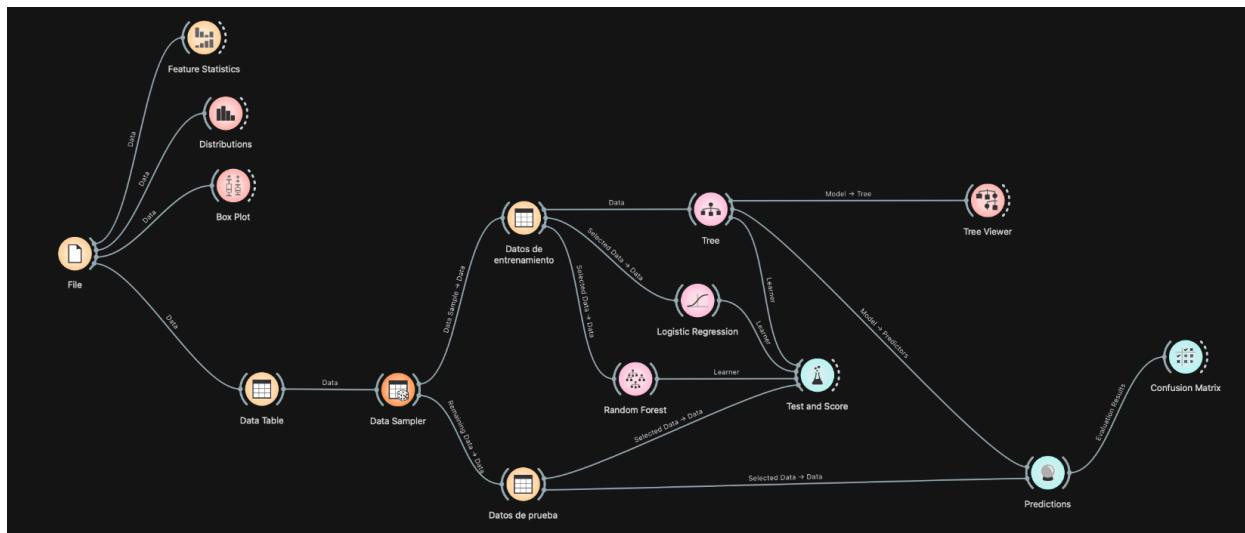
Estos resultados confirman las hipótesis iniciales y orientan la selección de variables predictoras clave para la construcción de los modelos de clasificación en las siguientes etapas.

Preprocesamiento y Selección de Variables

Antes de usar los modelos predictivos, preprocesamos y seleccionamos las variables. Esto aseguró que los datos estuvieran bien estructurados y listos para el entrenamiento.

Primero, revisamos que no hubiera valores faltantes y confirmamos que los tipos de datos (numéricos y categóricos) fueran consistentes. Luego, usamos un Data Sampler para dividir el conjunto de datos en un 70% para entrenamiento y un 30% para prueba. Nos aseguramos de que la distribución fuera representativa aplicando un muestreo estratificado según la variable objetivo, *HeartDisease*.

Para la selección de variables, mantuvimos todas las del conjunto de datos original porque un análisis previo mostró que la mayoría estaban relacionadas significativamente con la presencia de enfermedad cardíaca. Sin embargo, consideramos que `Age`, `Sex`, `ChestPainType`, `MaxHR`, `Oldpeak` y `ExerciseAngina` eran las más relevantes, ya que mostraron diferencias claras entre los grupos con y sin la enfermedad. Estas variables fueron las principales que usaron los modelos de clasificación (Árbol de Decisión, Regresión Logística y Random Forest).



Modelado

En esta etapa, usamos tres modelos distintos de aprendizaje automático en *Orange Data Mining* para intentar predecir si una persona tiene una enfermedad cardíaca (HeartDisease).

Los modelos que utilizamos fueron:

1. **Árbol de Decisión (Tree)**
2. **Regresión Logística (Logistic Regression)**
3. **Bosque Aleatorio (Random Forest)**

Configuración de los modelos:

- Usamos el **Data Sampler** para dividir los datos:
 - El 70% se usó para entrenar los modelos.
 - El 30% restante se usó para probarlos, manteniendo la misma proporción de casos con y sin enfermedad cardíaca (muestreo estratificado).
- En el módulo **Test & Score**, aplicamos una **validación cruzada de 5 pliegues (5-fold Cross Validation)** para que las métricas fueran más confiables.
- Todos los modelos se entrenaron con las mismas variables, que son las que se

usaron para predecir: Age, Sex, ChestPainType, MaxHR, Oldpeak, ExerciseAngina, RestingBP, Cholesterol, FastingBS, RestingECG y ST_Slope.

Breve descripción de los modelos

- **Árbol de Decisión:** Este modelo divide los datos basándose en distintas variables para clasificar los casos. Es útil porque podemos entender fácilmente qué condiciones llevan a predecir la enfermedad cardíaca.
- **Regresión Logística:** Es un modelo estadístico lineal que calcula la probabilidad de que ocurra un evento (en este caso, tener enfermedad cardíaca) a partir de las variables que le damos.
- **Random Forest:** Este es un modelo que combina varios árboles de decisión (es un "ensamble"). Al juntar los resultados de muchos árboles, mejora la precisión y ayuda a evitar que el modelo se ajuste demasiado a los datos de entrenamiento.

Evaluación del Modelo

Para evaluar cómo funcionan los tres modelos que entrenamos (Árbol de Decisión, Regresión Logística y Random Forest), usamos el módulo Test & Score de Orange Data Mining.

La evaluación se hizo con una validación cruzada de 5 pliegues (5-Fold Cross Validation). Esto significa que cada parte de los datos se usó tanto para entrenar como para probar los modelos, lo que hace que las métricas obtenidas sean más fiables.

Métricas utilizadas

- **AUC (Área bajo la curva ROC):** Esta métrica nos dice qué tan bien el modelo puede diferenciar entre las distintas clases (1 es excelente, 0.5 es como adivinar al azar).
- **CA (Classification Accuracy):** Mide el porcentaje de predicciones correctas que hizo el modelo sobre el total.
- **Precision:** Indica, de todas las veces que el modelo dijo que algo era positivo, cuántas veces acertó realmente.
- **Recall (Sensibilidad):** Muestra cuántos de los casos positivos reales logró detectar el modelo → De todos los pacientes que realmente estaban enfermos, ¿a cuántos logramos detectar?
- **F1:** Es un promedio que combina precision y recall, buscando un equilibrio entre ambos.

- **MCC (Matthews Correlation Coefficient):** Evalúa la calidad general de la clasificación, tomando en cuenta tanto los aciertos como los errores (positivos y negativos).

Resultados obtenidos:

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.781	0.753	0.753	0.756	0.753	0.505
Logistic Regression	0.921	0.858	0.858	0.858	0.858	0.713
Random Forest	0.927	0.858	0.858	0.858	0.858	0.713

Comparación de desempeño de modelos

El Árbol de Decisión tuvo el rendimiento más bajo, con un AUC de 0.781, lo que indica que no fue muy efectivo para distinguir entre las clases.

La Regresión Logística mostró una mejora notable, alcanzando un AUC de 0.921. Esto sugiere una fuerte relación lineal entre las variables utilizadas para predecir y el diagnóstico de enfermedad cardíaca.

El modelo con mejor rendimiento fue el Random Forest, con un AUC de 0.927. Esto se debe a que combina varios árboles de decisión y promedia sus resultados, lo que ayuda a reducir el sobreajuste y aumenta la precisión. Por lo tanto, este modelo es el que tiene mayor capacidad de generalización.

Considerando las métricas, el modelo Random Forest es el más adecuado para este problema por varias razones:

- Tiene el AUC más alto (0.927).
- Mantiene una alta precisión y un buen *recall*, lo que significa que identifica correctamente la mayoría de los casos positivos sin generar muchos falsos positivos.
- Es un modelo robusto y no lineal, capaz de capturar relaciones complejas entre distintas variables clínicas.

La Regresión Logística es una buena segunda opción debido a su desempeño similar y a que sus resultados son fáciles de interpretar clínicamente. El Árbol de Decisión, aunque útil para visualizar explicaciones, no es tan eficaz en la predicción.

Punto 7 – Interpretación de las Predicciones

Después de seleccionar y evaluar los modelos de aprendizaje automático, analizamos las **variables más importantes** para predecir la enfermedad cardíaca. Queríamos entender qué factores influyen más en el modelo final y comparar esos resultados con nuestras hipótesis iniciales.

Nos enfocamos principalmente en el modelo **Random Forest**, ya que fue el que tuvo el mejor rendimiento general según las métricas de evaluación ($AUC = 0.927$). Este modelo nos permite ver y entender la **importancia de cada variable** en la clasificación, lo que nos da una idea clara de qué atributos son más decisivos para el diagnóstico.

Variables más relevantes según el modelo:

1. **Oldpeak (Depresión del segmento ST)**

Esta variable fue muy influyente. Los pacientes con enfermedad cardíaca mostraron valores de *Oldpeak* significativamente más altos, lo que indica una mayor respuesta anormal del segmento ST durante el ejercicio o el estrés cardíaco.

2. **MaxHR (Frecuencia cardíaca máxima alcanzada)**

Observamos que las personas con menor *MaxHR* tendían a tener enfermedad cardíaca, mientras que quienes tenían valores más altos pertenecían al grupo sin enfermedad. Esto confirma que una menor capacidad cardíaca al esfuerzo puede ser un indicador de riesgo.

3. **Age (Edad)**

La edad también mostró una relación clara con la enfermedad cardíaca. Los pacientes diagnosticados eran, en promedio, mayores que los que no la padecían. Esto coincide con la evidencia médica que indica que el riesgo cardiovascular aumenta con la edad.

4. **ChestPainType (Tipo de dolor de pecho)**

El tipo de dolor *ASY* (*asintomático*) se relacionó directamente con la enfermedad cardíaca, mientras que los tipos *ATA* y *NAP* fueron más comunes en pacientes sin diagnóstico. Esta variable es un indicador clínico relevante para la detección temprana de afecciones cardíacas.

5. **Sex (Sexo)**

Los resultados mostraron una mayor proporción de hombres con enfermedad cardíaca en comparación con las mujeres, lo que se alinea con las tendencias epidemiológicas conocidas.

Relación con los hallazgos previos:

Los resultados de esta etapa confirman lo observado en el análisis exploratorio de datos (punto 3). Las variables que inicialmente mostraron diferencias notables entre los grupos con y sin enfermedad, como Oldpeak, MaxHR y Edad, fueron identificadas por el modelo como las más relevantes para la predicción. También se verificó que el tipo de dolor de pecho y el sexo contribuyen significativamente al diagnóstico, reforzando nuestras hipótesis iniciales basadas en los boxplots y las distribuciones.

Conclusión:

El modelo **Random Forest** no solo presentó el mejor rendimiento predictivo, sino que además reflejó patrones coherentes con los fundamentos clínicos del problema. Las variables que detectamos como más influyentes mantienen una relación médica plausible, lo que brinda confianza en la validez de las predicciones y demuestra que el modelo logra captar correctamente las características más importantes para diagnosticar la enfermedad cardíaca.

Además, este modelo podría aplicarse en la práctica para ayudar a identificar pacientes con alto riesgo y priorizar su atención médica, optimizando los recursos disponibles y permitiendo una intervención más temprana. La coherencia entre los resultados exploratorios y predictivos indica que el modelo no se limita a ajustar datos, sino que aprende relaciones reales y útiles para la toma de decisiones clínicas.

Como trabajo futuro, se propone mejorar la calidad del dataset corrigiendo valores atípicos, ampliar la muestra del modelo y evaluar nuevas métricas o técnicas de calibración para obtener estimaciones más precisas del riesgo.

En conclusión, el modelo constituye una herramienta sólida y confiable con un gran potencial para su aplicación en entornos de salud preventiva.