Análisis y Modelado de Datos con Orange Data Mining

Predicción de Enfermedades Cardíacas

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

Definición del Problema

Objetivos del Proyecto

Objetivo Principal: Desarrollar un modelo predictivo que permita determinar la presencia de enfermedad cardíaca en un paciente.

Pregunta Central: ¿Es posible predecir la enfermedad cardíaca basándose en características clínicas y hábitos de vida?

Meta de Clasificación: Clasificar pacientes en dos categorías: **1 (Tiene enfermedad)** o **0 (No tiene enfermedad)**.

Meta Secundaria: Evaluar qué variables tienen mayor influencia en la predicción y comparar el rendimiento de distintos modelos.

Análisis Exploratorio de Datos (EDA)

Variables:

1	Age	N numeric	feature	
2	Sex	C categorical	feature	F, M
3	ChestPainType	C categorical	feature	ASY, ATA, NAP, TA
4	RestingBP	N numeric	feature	
5	Cholesterol	N numeric	feature	
6	FastingBS	C categorical	feature	0, 1

7	RestingECG	C categorical	feature	LVH, Normal, ST
8	MaxHR	N numeric	feature	
9	ExerciseAngina	C categorical	feature	N, Y
10	Oldpeak	N numeric	feature	
11	ST_Slope	C categorical	feature	Down, Flat, Up
12	**HeartDisease**	C **categorical**	**target**	**0, 1**

Descripción del Dataset (Heart Failure Prediction)

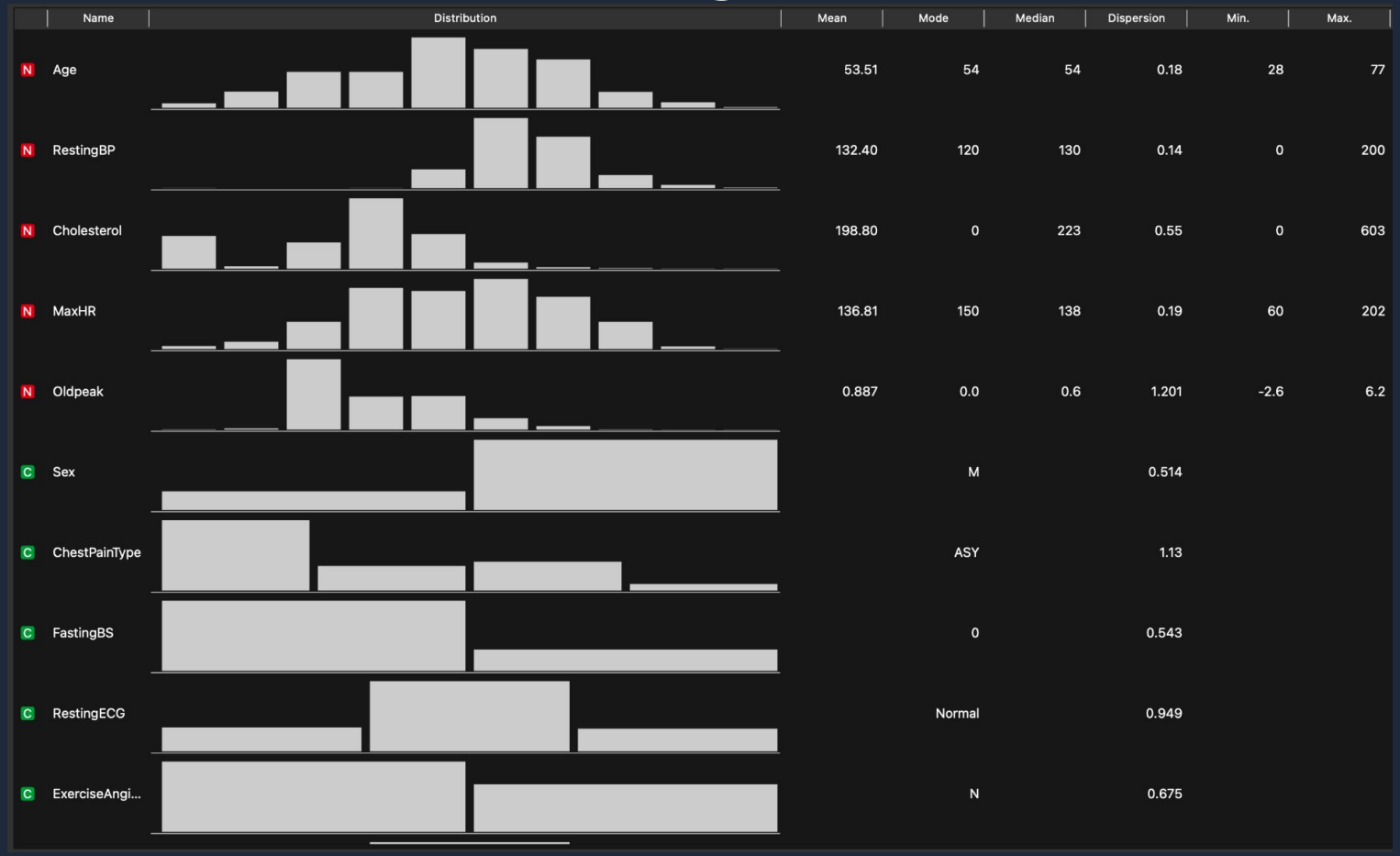
Visión General

El estudio se basa en un conjunto de datos que incluye **918 observaciones** y **12 variables** que describen las características clínicas y demográficas de los pacientes.

Variables Clave

Se analizan predictores como **Age**, **Sex**, **ChestPainType**, **MaxHR** (Frec. cardíaca máx.), **Oldpeak** (Depresión ST) y **ExerciseAngina** para predecir la variable objetivo: **HeartDisease**.

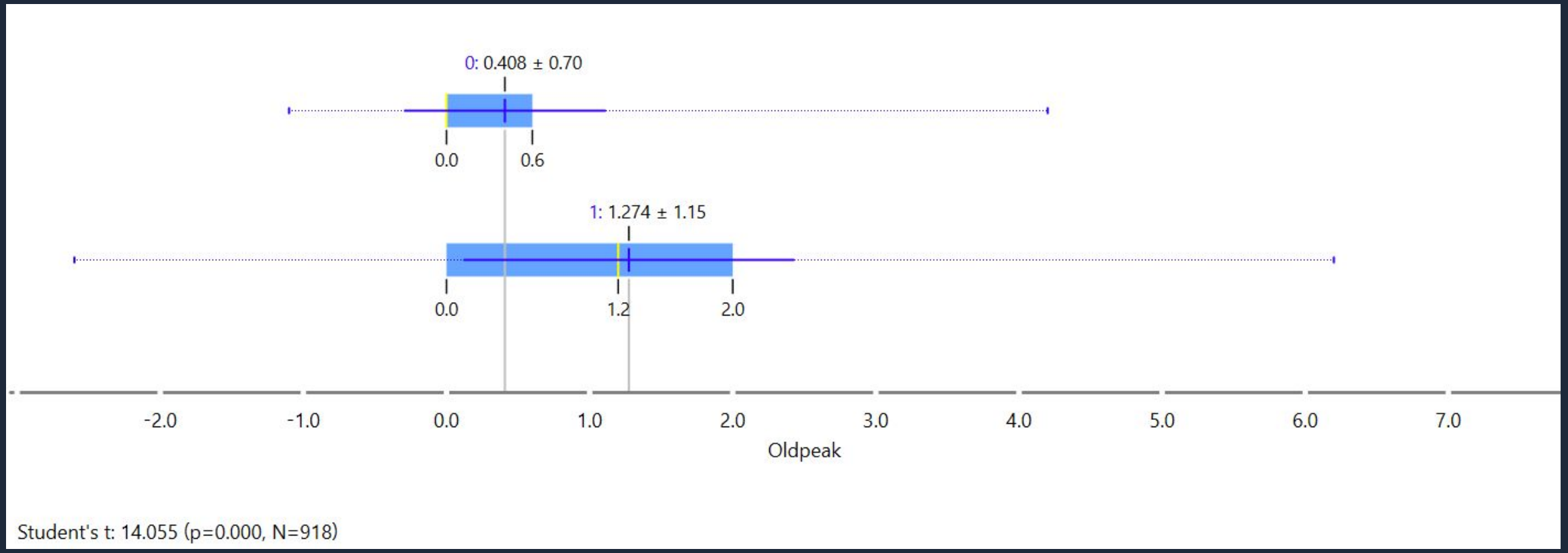
Estadísticas Generales (Widget: Feature Statistics)



	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.
N	Age		53.51	54	54	0.18	28	77
N	RestingBP		132.40	120	130	0.14	0	200
N	Cholesterol		198.80	0	223	0.55	0	603
N	MaxHR		136.81	150	138	0.19	60	202
N	Oldpeak		0.887	0.0	0.6	1.201	-2.6	6.2
C	Sex			M		0.514		
C	ChestPainType			ASY		1.13		
C	FastingBS			0		0.543		
C	RestingECG			Normal		0.949		
C	ExerciseAngi...			N		0.675		

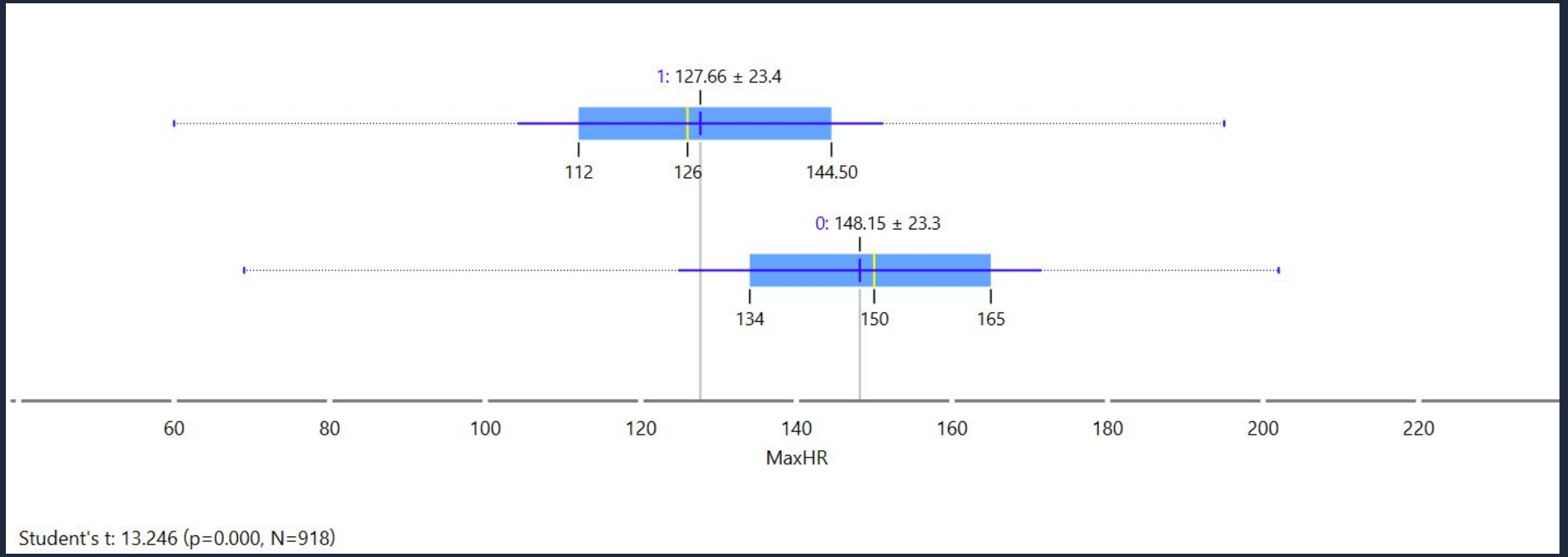
- **Edad Promedio:** La edad promedio de los pacientes es de 53.5 años.
- **Valores Atípicos (Outliers):** La variable **Cholesterol** muestra alta variabilidad (rango de 0 a 603 mg/dL), sugiriendo la presencia de valores atípicos o datos nulos codificados como 0.
- **Oldpeak Media:** El valor medio de la depresión del segmento ST (`Oldpeak`) es de 0.88.

EDA: Variables Numéricas (Widget: Box Plot)



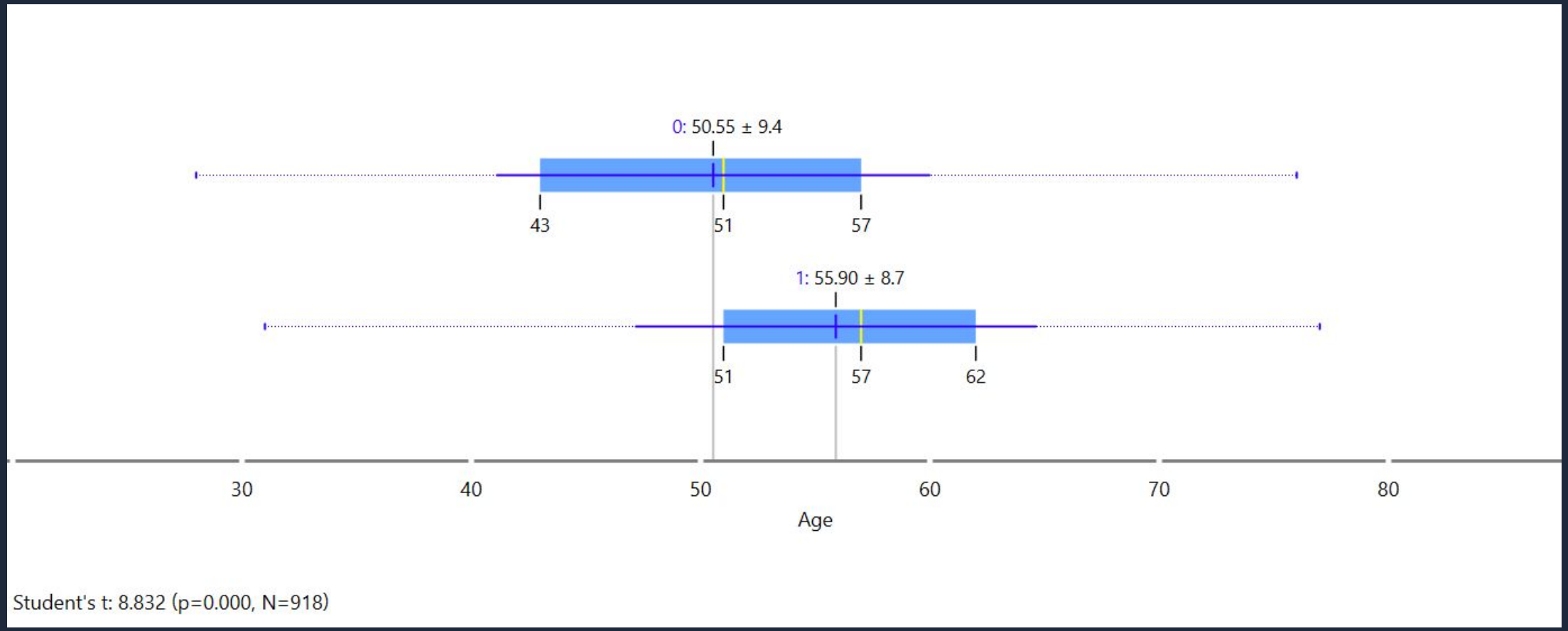
Oldpeak: Pacientes con enfermedad (1) muestran valores más altos (Media 1.27) vs. sanos (0) (Media 0.40).

EDA: Variables Numéricas (Widget: Box Plot)



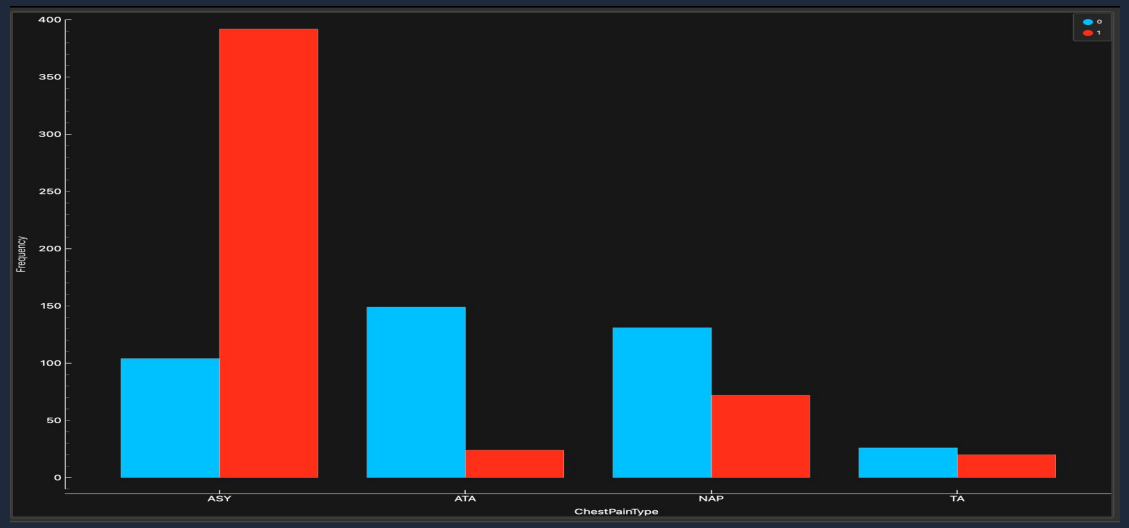
MaxHR: Pacientes sanos (0) alcanzan frecuencias cardíacas máximas más altas (Media 148.1) vs. enfermos (1) (Media 127.6).

EDA: Variables Numéricas (Widget: Box Plot)

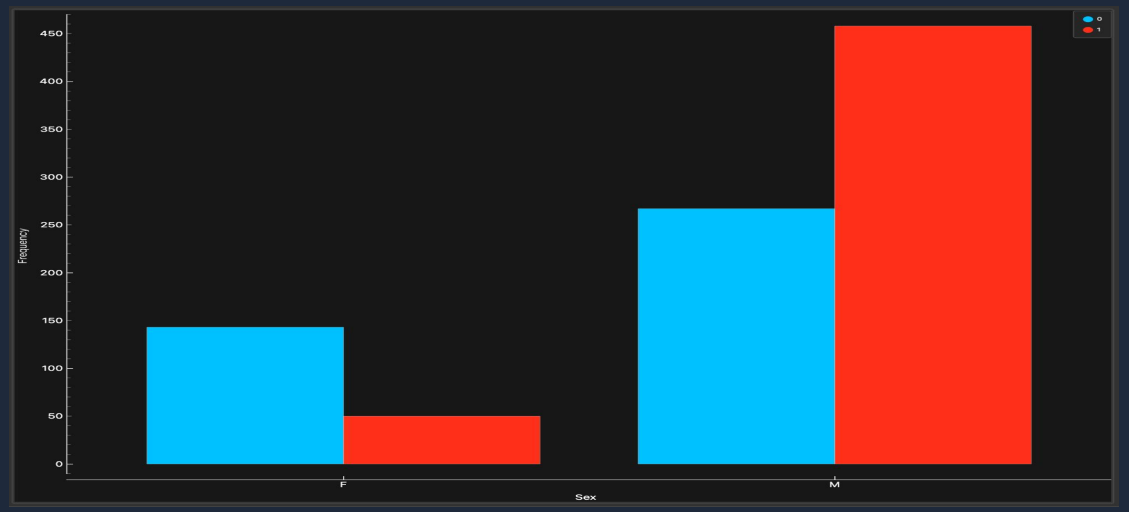


Age: Pacientes con enfermedad (1) tienen una edad promedio mayor (55.9 años) vs. sanos (0) (50.6 años).

EDA: Variables Categóricas (Widget: Distributions)



ChestPainType: El tipo 'ASY' (Asintomático) predomina fuertemente en pacientes con enfermedad cardíaca, mientras 'ATA' y 'NAP' son más comunes en pacientes sanos.

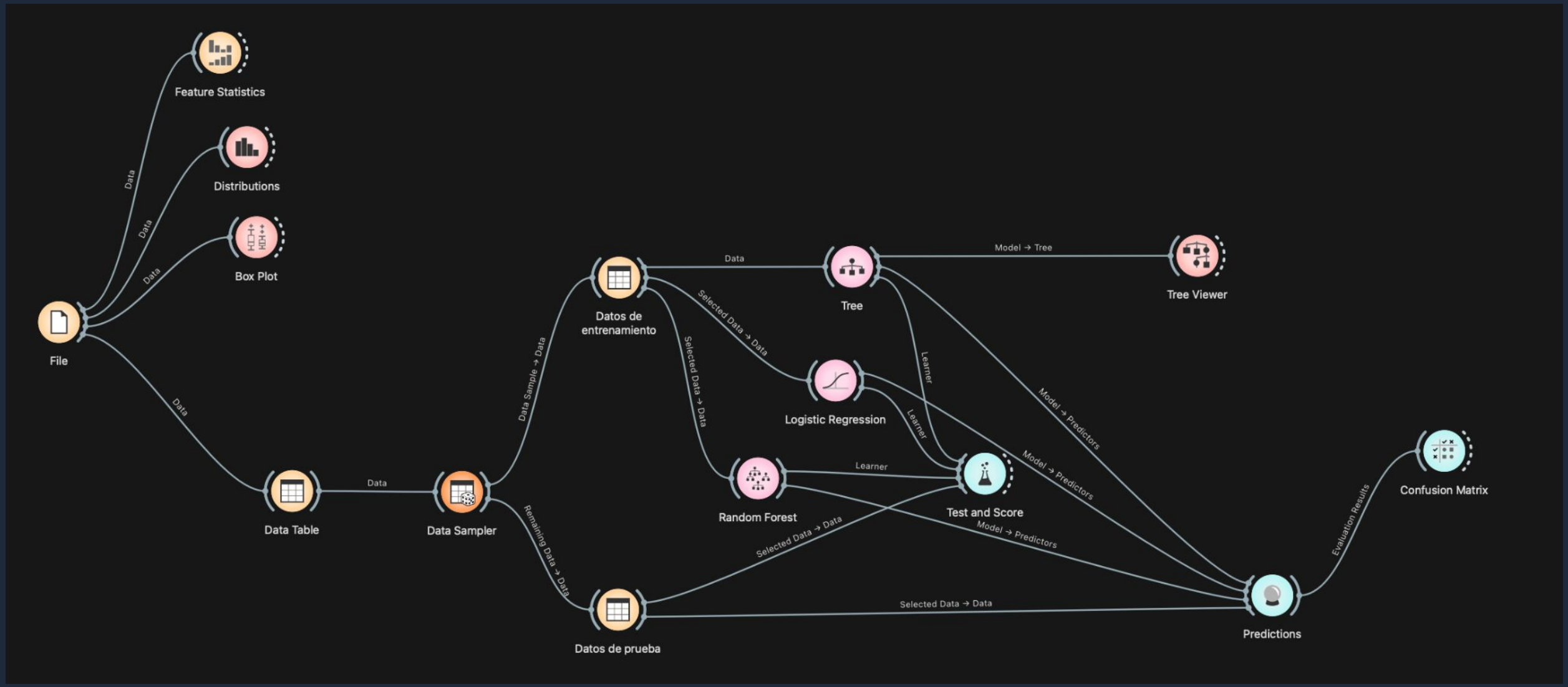


Sex: Los hombres ('M') presentan una proporción significativamente mayor de enfermedad cardíaca en el dataset, confirmando tendencias epidemiológicas.

Conclusiones del Análisis Exploratorio

"El análisis exploratorio detectó patrones claros. Pacientes de **mayor edad, sexo masculino, con menor MaxHR, mayor Oldpeak** y **síntomas ASY** presentan **mayor probabilidad** de enfermedad cardíaca."

Preprocesamiento y Workflow del Modelo



Preprocesamiento y Workflow del Modelo

- **División (Data Sampler):** El dataset se dividió en **70% Entrenamiento** y **30% Prueba**.
- **Muestreo Estratificado:** Se utilizó muestreo estratificado (basado en `HeartDisease`) para asegurar que la proporción de clases (enfermos/sanos) fuera idéntica en ambos conjuntos.
- **Selección de Variables:** Se mantuvieron todas las variables, destacando las identificadas en el EDA (`Oldpeak`, `MaxHR`, `Age`, `ChestPainType`, `Sex`) como las más relevantes para los modelos.

Modelado y Evaluación

Modelos de Clasificación Implementados

Árbol de Decisión (Tree)

Divide los datos basándose en variables para clasificar. Es útil por su facilidad para interpretar visualmente las reglas de decisión.

Regresión Logística

Modelo estadístico lineal que calcula la probabilidad de que un paciente tenga la enfermedad, basándose en las variables de entrada.

Random Forest

Modelo de ensamble que combina múltiples árboles de decisión para mejorar la precisión, reducir el sobreajuste y aumentar la robustez.

Resultados de Evaluación (Validación Cruzada 5-Pliegues)

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

Comparativa y Selección de Modelo

0.927

AUC - Random Forest

Random Forest es el Mejor Modelo

El **Random Forest** fue el modelo con mejor rendimiento (AUC 0.927), superando marginalmente a la Regresión Logística (0.921) y significativamente al Árbol de Decisión (0.781).

Se selecciona por su alta capacidad de generalización y su habilidad para manejar relaciones no lineales complejas entre variables clínicas, siendo el modelo más robusto.

Interpretación del Modelo Ganador (Random Forest)

- **Oldpeak:** Identificada como la variable más influyente. Valores altos de depresión ST se asocian fuertemente con la enfermedad.
- **MaxHR:** Alta influencia. Frecuencias cardíacas máximas (MaxHR) bajas durante el esfuerzo son un claro indicador de riesgo.
- **Age:** La edad es un factor clave y consistente; a mayor edad, mayor riesgo cardiovascular detectado por el modelo.
- **ChestPainType:** El tipo 'ASY' (Asintomático) es un predictor decisivo para el diagnóstico de enfermedad.
- **Coherencia del Modelo:** Estos hallazgos confirman al 100% los patrones descubiertos durante el análisis exploratorio (EDA).

Conclusión y Trabajo Futuro

Conclusión: El modelo **Random Forest** es robusto, preciso (0.927 AUC) y sus predicciones son coherentes con los fundamentos clínicos. Se ha creado una herramienta con gran potencial para la salud preventiva, capaz de identificar pacientes de alto riesgo.

Trabajo Futuro: Se propone mejorar la calidad del dataset (ej. corrigiendo valores atípicos de Colesterol), ampliar la muestra del modelo y evaluar nuevas técnicas de calibración para estimaciones de riesgo más precisas.

Gracias