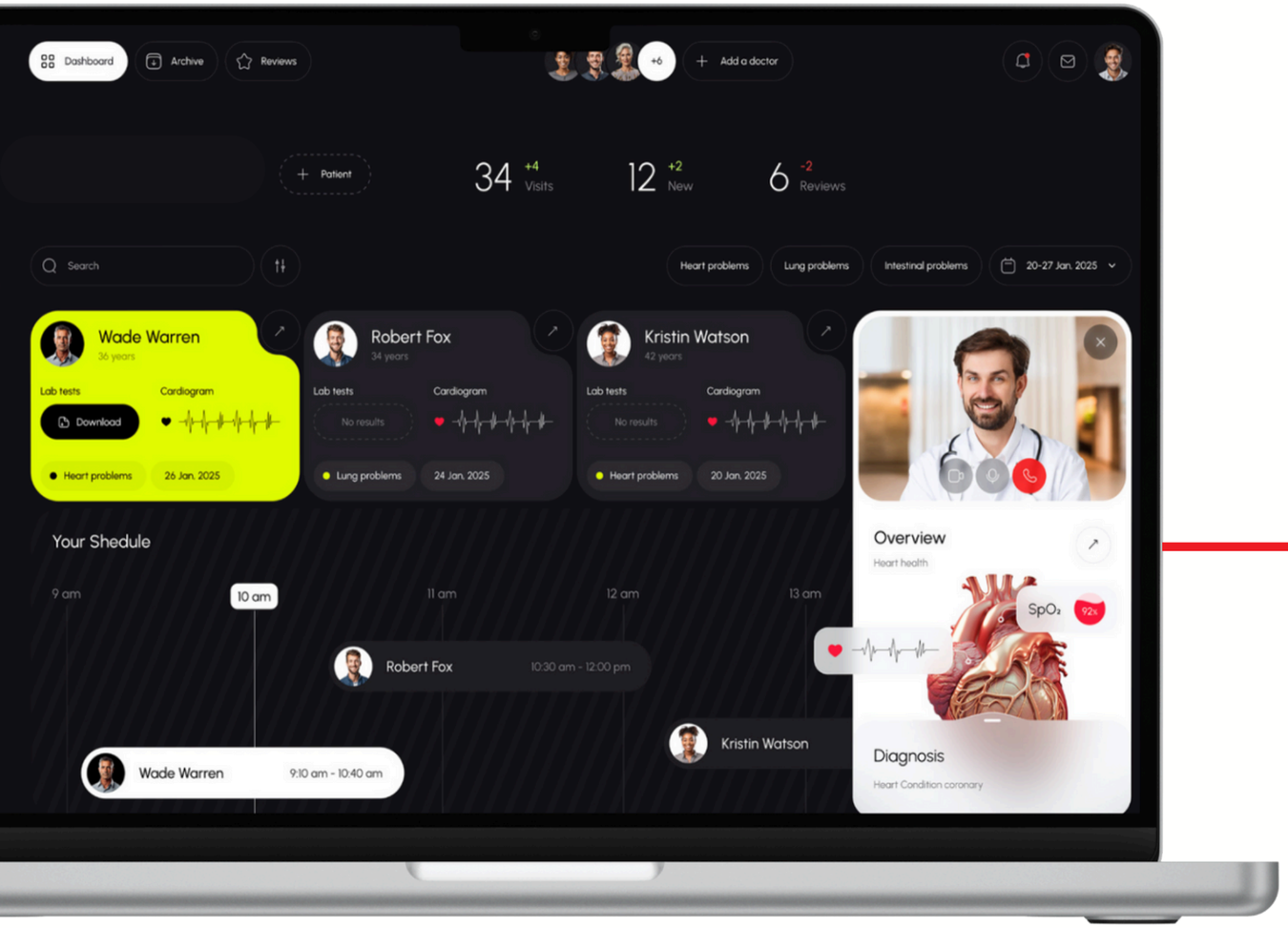


Heart Disease Classifier

Punto de Partida

DETECCIÓN DE ENFERMEDAD CARDIACA

Uno de los principales objetivos es conseguir la detección temprana de enfermedades cardíacas.

DATASET INCOMPLETO

trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
156	173	0	2	119	1	3	3	?	?
160.0	196.0	0.0	0.0	165.0	0.0	0.0	-9.0	-9.0	-9.0
100.0	-9.0	0.0	0.0	100.0	0.0	0.0	-9.0	-9.0	-9.0
115	0	?	0	128	1	2.5	3	?	?
110.0	175.0	0.0	0.0	123.0	0.0	0.6	1.0	0.0	3.0
...	...	...	...	...	...	...	...	...	...
200.0	198.0	0.0	0.0	142.0	1.0	2.0	2.0	-9.0	-9.0
110	214	1	1	180	0	?	?	?	?
152.0	212.0	0.0	2.0	150.0	0.0	0.8	2.0	0.0	7.0
170.0	288.0	0.0	2.0	159.0	0.0	0.2	2.0	0.0	7.0
?	203	1	0	?	?	?	?	?	?

El principal reto ha sido la calidad del dataset.

INFORMACIÓN SOBRE MÚLTIPLES PACIENTES

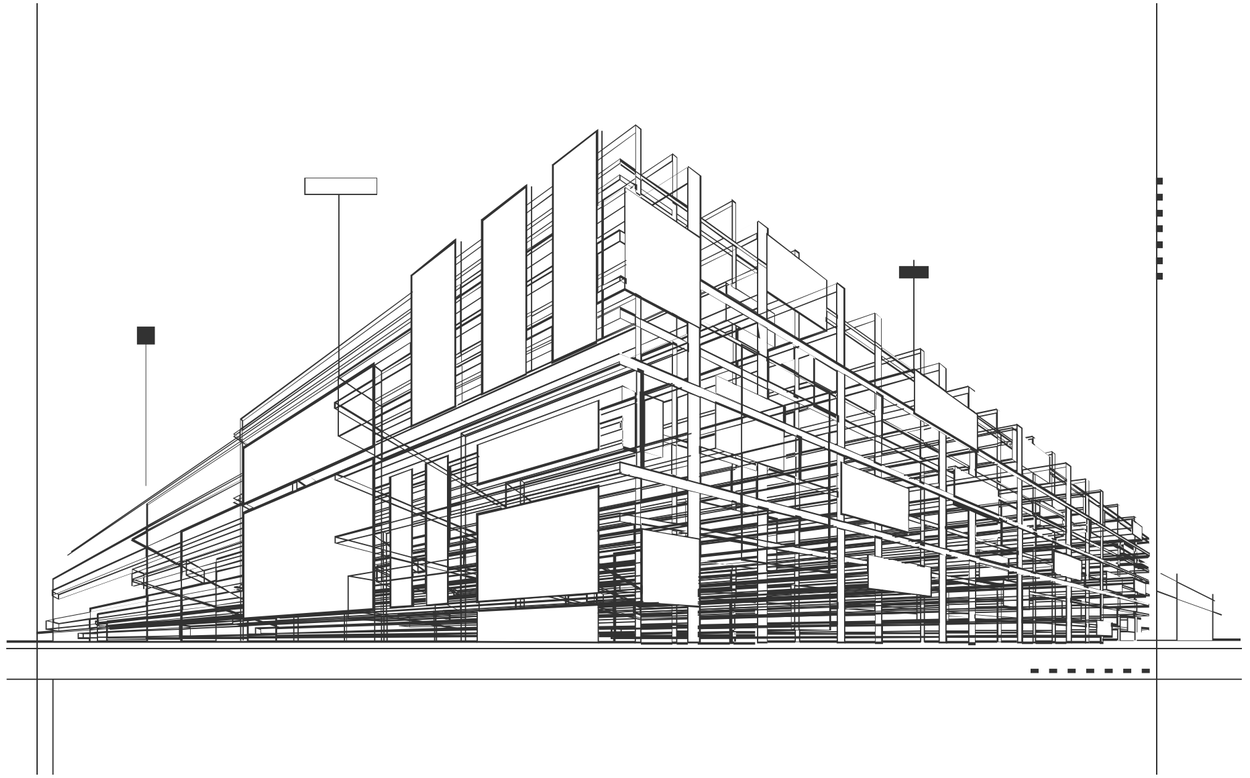
La cantidad de información detallada de varios pacientes representa una oportunidad y un desafío para la modelización.

CONSTRUCCIÓN EN PROGRESO

En este tipo de proyectos, el modelo y las estrategias se ajustan continuamente a medida que se analizan más datos y se obtienen nuevos insights.

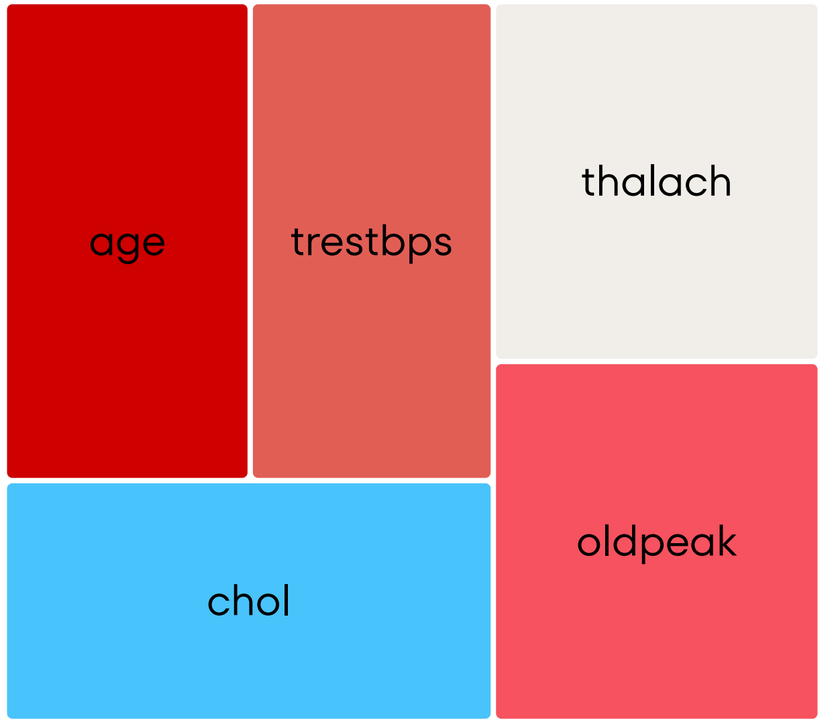
La Solución

Modelo de aprendizaje automático que pueda predecir la presencia de enfermedad en el corazón basándose en las características recogidas en diferentes pruebas médicas como variables clínicas.



Tipos de Variables

Numéricas



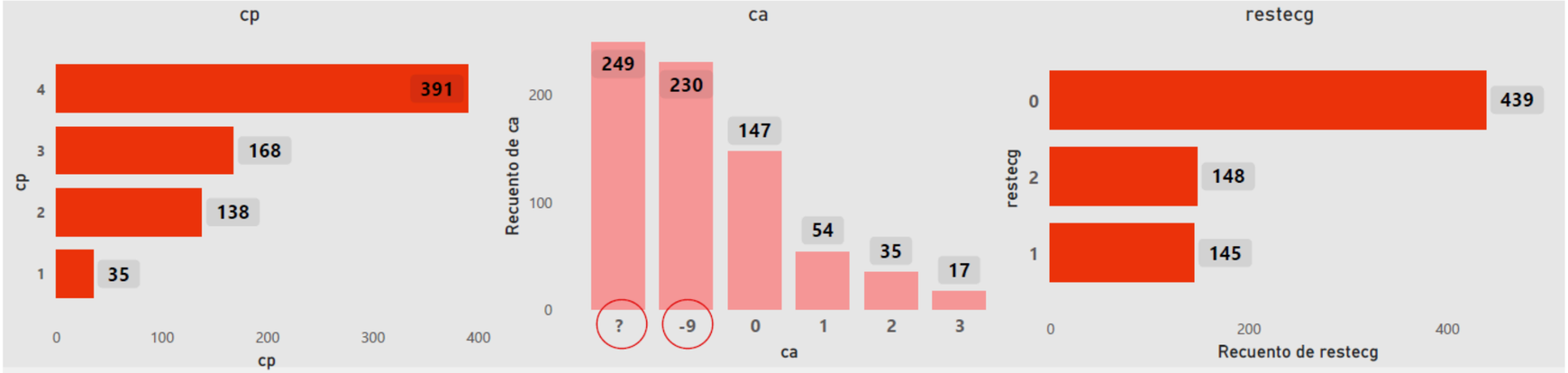
Categóricas

Análisis de la distribución de valores para las variables categóricas

Distribución de Tipos de Dolor de Pecho (cp)

Recuento de Vasos Sanguíneos Principales (ca)

Resultados de Electrocardiograma en Reposo (restecg)

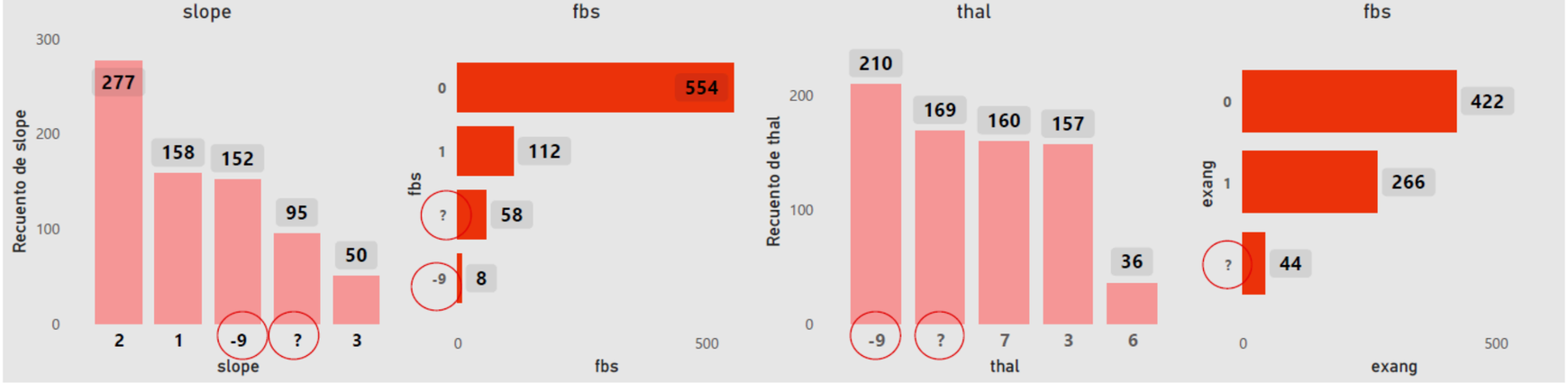


Para las variables categóricas como sex, cp, fbs, restecg, exang, slope, ca y thal, los valores NaN no pueden simplemente rellenarse con promedios. En estos caso valoraremos:

· Eliminar las filas si el número de casos faltantes no es significativo en comparación con el tamaño del dataset.

· Imputar los NaN basándote en la moda (el valor más frecuente) de la columna, o mediante un modelo de imputación más sofisticado.

Presencia de Angina Inducida por Ejercicio (exang)



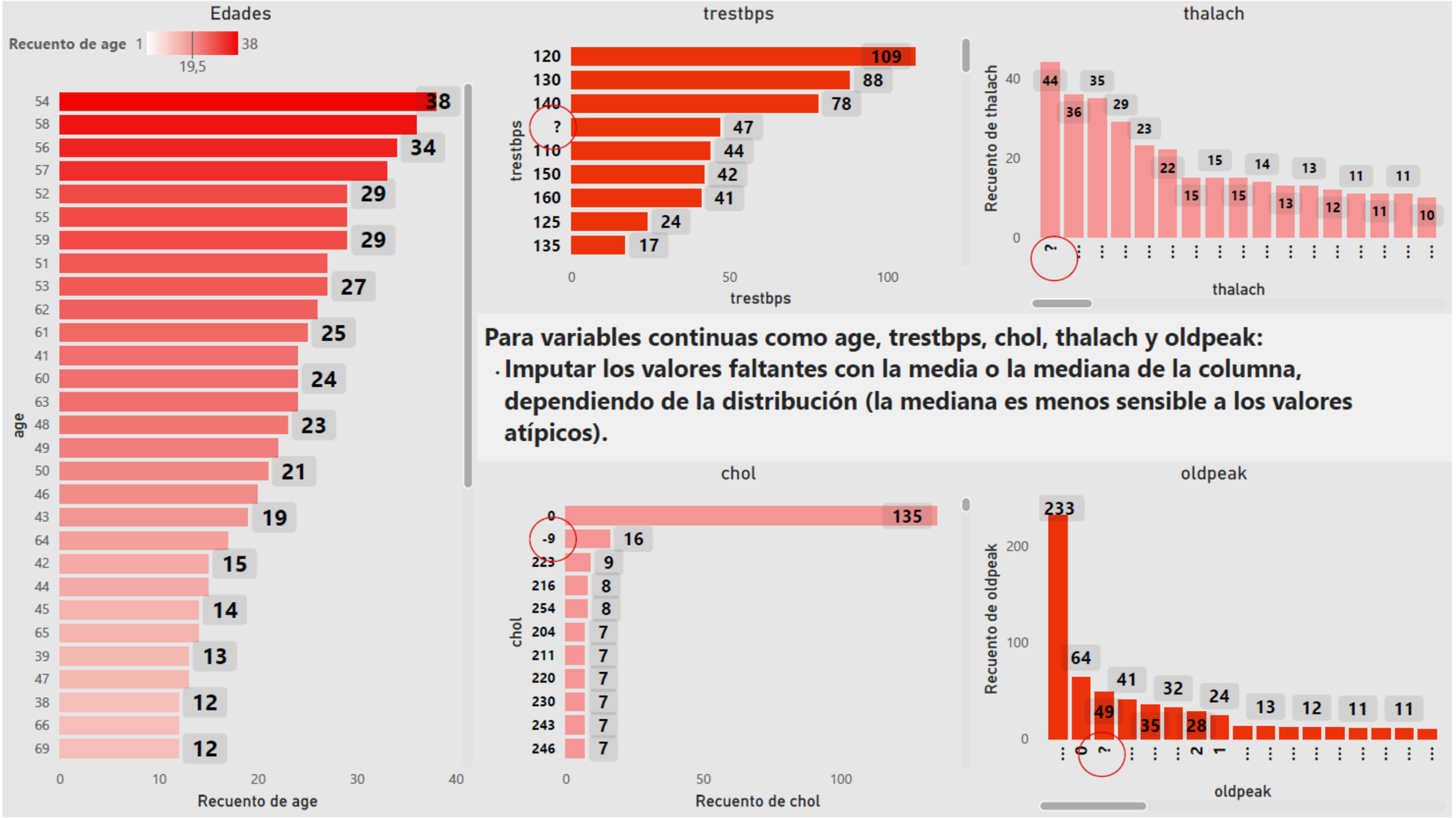
Pendiente del Segmento ST en Ejercicio (slope)

Presencia de Dolor Provocado por el Esfuerzo (fbs)

Frecuencia Cardiaca Máxima Alcanzada (thalach) y Trastornos de Thal (thal)

Distribución de Edades en el Estudio

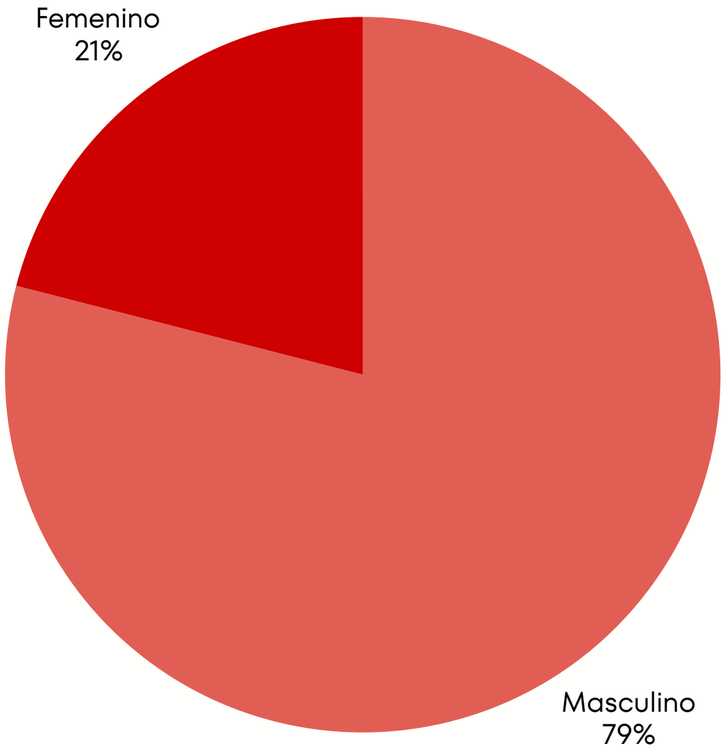
Análisis de la Presión Arterial en Reposo



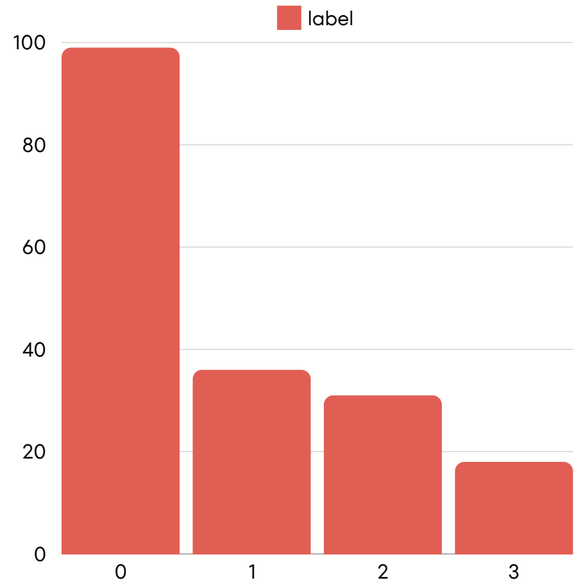
Distribución de Niveles de Colesterol

Frecuencia Cardíaca Máxima (thalach) y Depresión del Segmento ST (oldpeak)

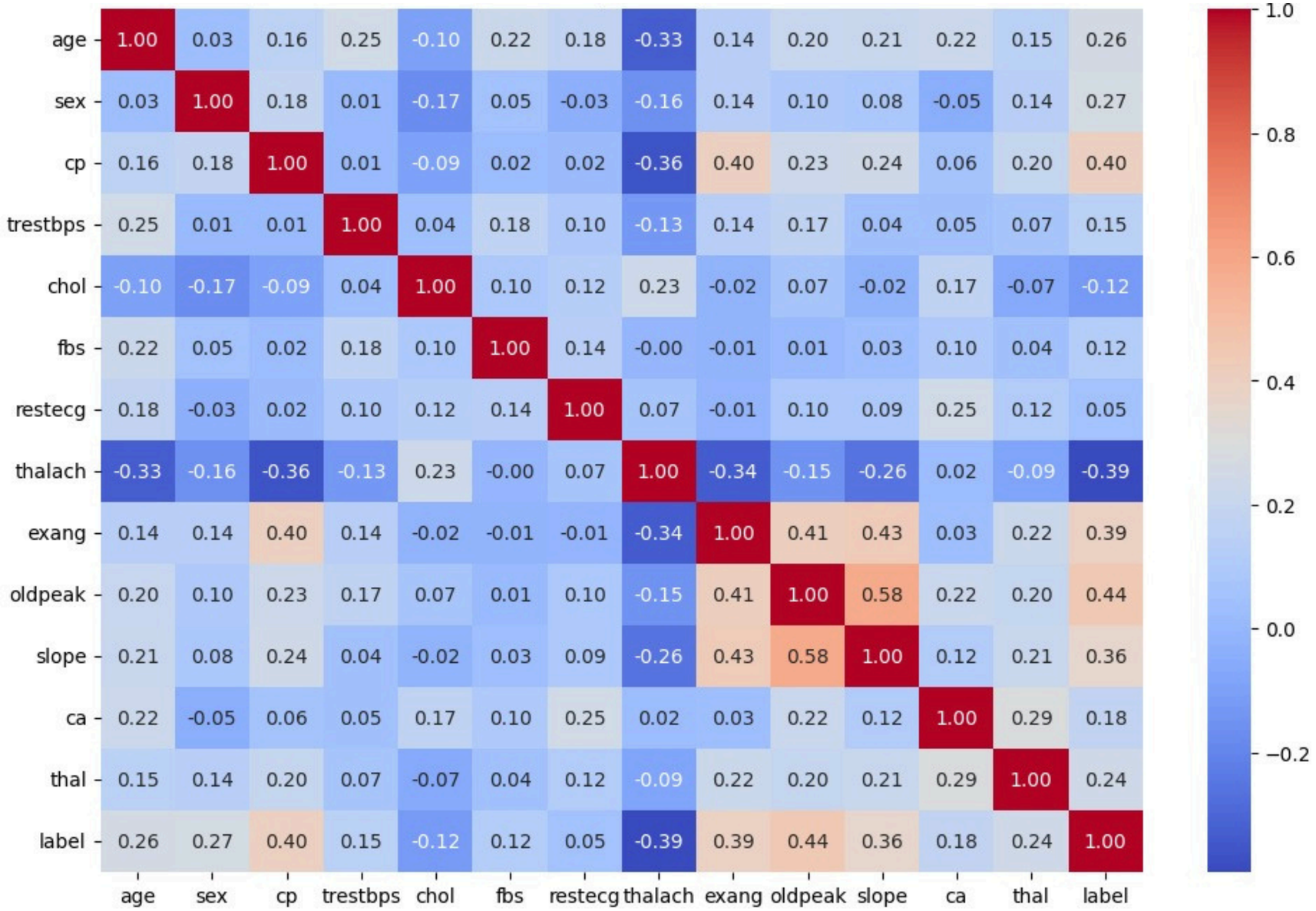
Distribución por sexo



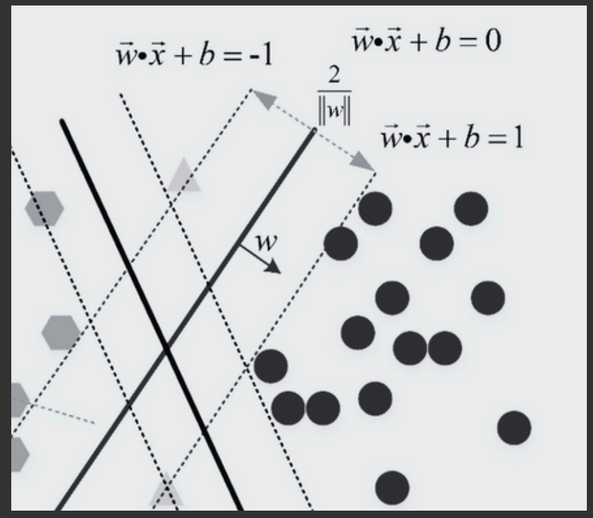
Distribución de Recuentos por Etiqueta



Matriz de Correlación del Conjunto de Datos de Entrenamiento

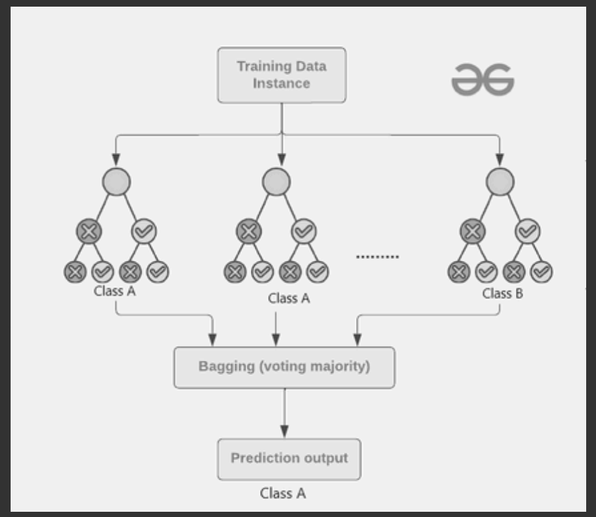


Primera aproximación



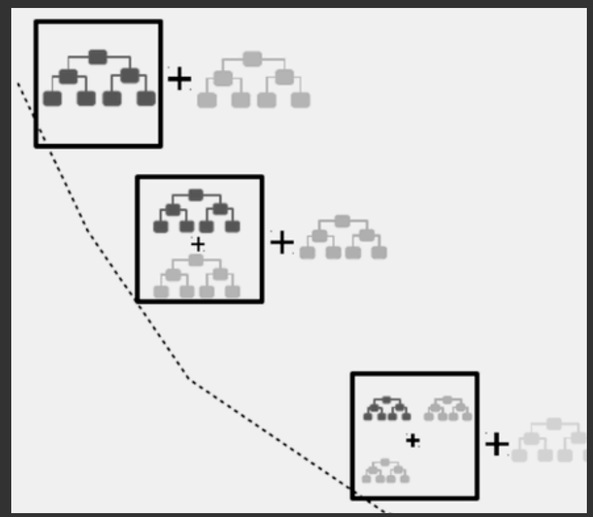
SVC

Es efectivo en espacios de alta dimensión , lo que lo hace ideal para conjuntos de datos con múltiples variables clínicas, logrando distinguir eficientemente entre clases mediante la maximización del margen de separación.



RANDOM FORREST

Consideradamos de ayuda para características categóricas y numéricas al ser capaz de modelar interacciones complejas entre variables.



GRADIENT BOOSTING

Mejora progresivamente el rendimiento a través de la optimización de un conjunto de predictores débiles, permitiendo un ajuste fino y mejorado del modelo sobre los datos, especialmente útil para maximizar el F1-score, que balancea precisión y sensibilidad.

Score: 0.55434

JUSTIFICACIÓN

Modelación Flexible y Robusta

Ofrece modelación flexible y robusta sin suposiciones estrictas sobre la distribución de los datos y la relación entre variables.

Manejo de Variables

Maneja variables categóricas como continuas

Capacidad de Paralelización

Es capaz de ejecutar entrenamientos en paralelo a través de múltiples árboles

Resumen: el modelo que mejor desempeño ha mostrado es el Random Forest con un umbral de importancia de 0.03. A través de varias evaluaciones y comparaciones con otros modelos, se ha observado que el Random Forest obtiene las métricas más favorables en términos de precisión, exhaustividad y puntuación F1.

MÉTRICAS

Métrica	Random Forest	Gradient Boosting
Precisión promedio	0.5171	0.5075
Desv. estándar de precisión	0.0272	0.0268
Precisión promedio (macro)	0.3604	0.3436
Desv. estándar de prec. (macro)	0.0832	0.0483
Recall promedio	0.3263	0.3357
Desv. estándar de recall	0.0315	0.0247
Puntuación F1 promedio	0.3438	0.3374
Desv. estándar de F1	0.0456	0.0351

Random Forest **ligeramente superior** a Gradient Boosting.

Menor en Random Forest, indicando mayor consistencia.

Gradient Boosting tiene una **ligera ventaja**.

Menor en Random Forest, mostrando mayor consistencia en la precisión entre clases.

Ligeramente más alto en Random Forest.

Similar para ambos modelos.

Ligeramente **más alta** en **Gradient Boosting.**

M**enor** en **Random Forest**.

Conclusión

1

ELECCIÓN DEL MODELO ÓPTIMO (RF)

2

IMPORTANCIA DEL PREPROCESAMIENTO DE DATOS

3

EFECTIVIDAD DE LA NORMALIZACIÓN Y LIMPIEZA DE DATOS

Pese a no haber obtenido unos resultados competitivos en un entorno real de producción, vemos la fiabilidad de los modelos empleados y llegamos a la conclusión de que los datos empleados no tienen la riqueza adecuada para ser empleados en un entorno clínico tan crítico.