Análisis de Riesgo Cardiovascular - Informe del Proyecto

Introducción

Este proyecto tiene como objetivo analizar un conjunto de datos médicos para predecir el riesgo de enfermedad cardiovascular (CVD). Se utilizan variables demográficas, antropométricas y de estilo de vida para construir modelos de machine learning que puedan anticipar la presencia de enfermedad cardiovascular.

Descripción del Dataset

El dataset contiene registros de salud de miles de individuos, con las siguientes variables clave:
- Edad (en años)
- Género (1: Femenino, 2: Masculino)
- Altura (cm)
- Peso (kg)
- Presión arterial sistólica (ap_hi) y diastólica (ap_lo)
- Colesterol (1: normal, 2: por encima de lo normal, 3: muy por encima de lo normal)
- Glucosa
- Fumar (0: no, 1: sí)
- Consumo de alcohol (0: no, 1: sí)
- Actividad física (0: no, 1: sí)
- Variable objetivo: enfermedad cardiovascular (0: ausente, 1: presente)

Preprocesamiento de Datos

Se realizó limpieza de datos para eliminar valores extremos y errores (como presión arterial negativa o invertida). Se normalizaron variables numéricas y se aplicó codificación One-Hot a variables categóricas.

Modelos Entrenados

Se entrenaron tres modelos de clasificación:
- Regresión Logística
- Random Forest
- XGBoost (modelo recomendado)

📊 Evaluación de Modelos

Se utilizaron métricas como Accuracy, Precision, Recall y F1-score.

Resultados aproximados:
- Regresión Logística: Accuracy 68%, F1-score 68%
- Random Forest: Accuracy 72%, F1-score 72%
- XGBoost: Accuracy 73%, F1-score 73%

XGBoost demostró mejor rendimiento general, por lo cual se seleccionó como modelo final.

🔁 Validación Cruzada

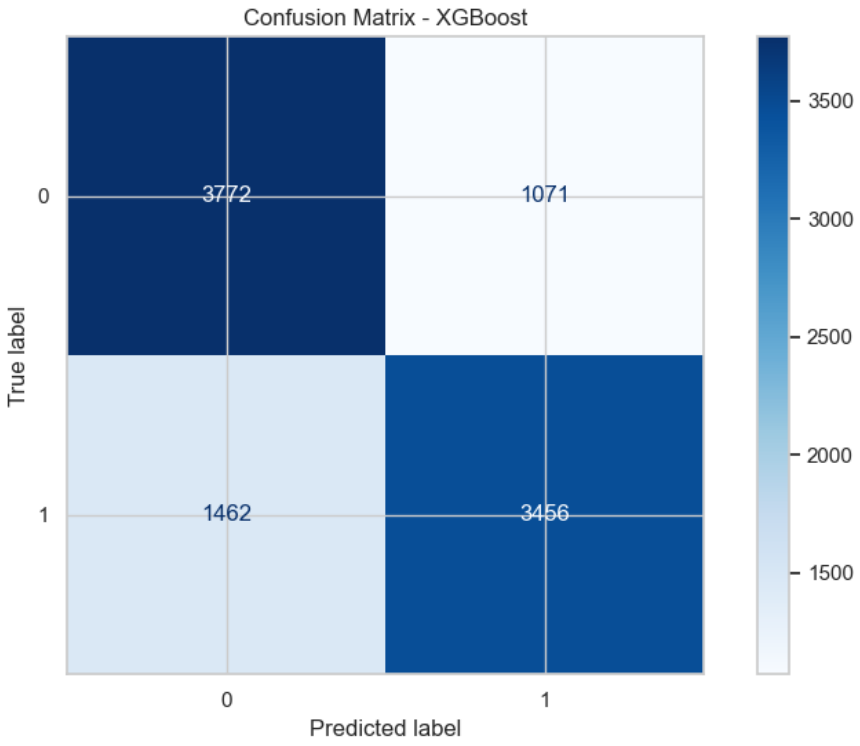
Se aplicó validación cruzada (cross-validation) con 10 particiones. El F1-score promedio fue de aproximadamente 0.7273. Esto demuestra una buena estabilidad del modelo ante nuevos datos.

📌 Interpretación y Feature Importance

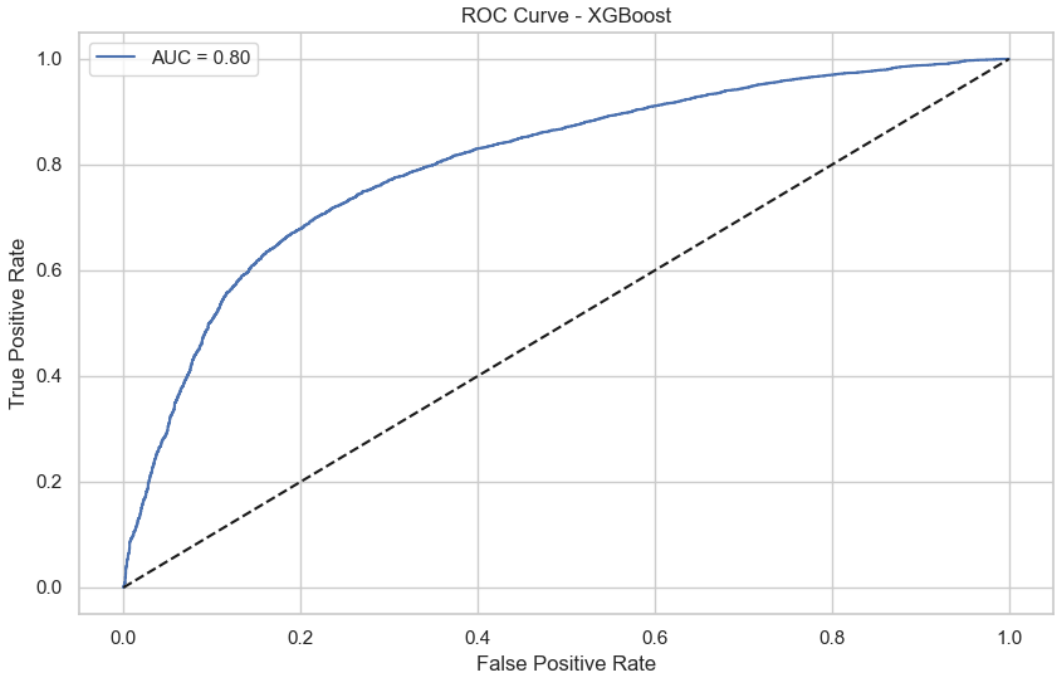
Se analizó la importancia de las variables con `feature_importances_`. La presión sistólica fue la más influyente, aunque su correlación lineal con la variable objetivo fue baja (0.05). Esto indica que el modelo capta relaciones no lineales que no se observan con correlación simple.

Visualizaciones

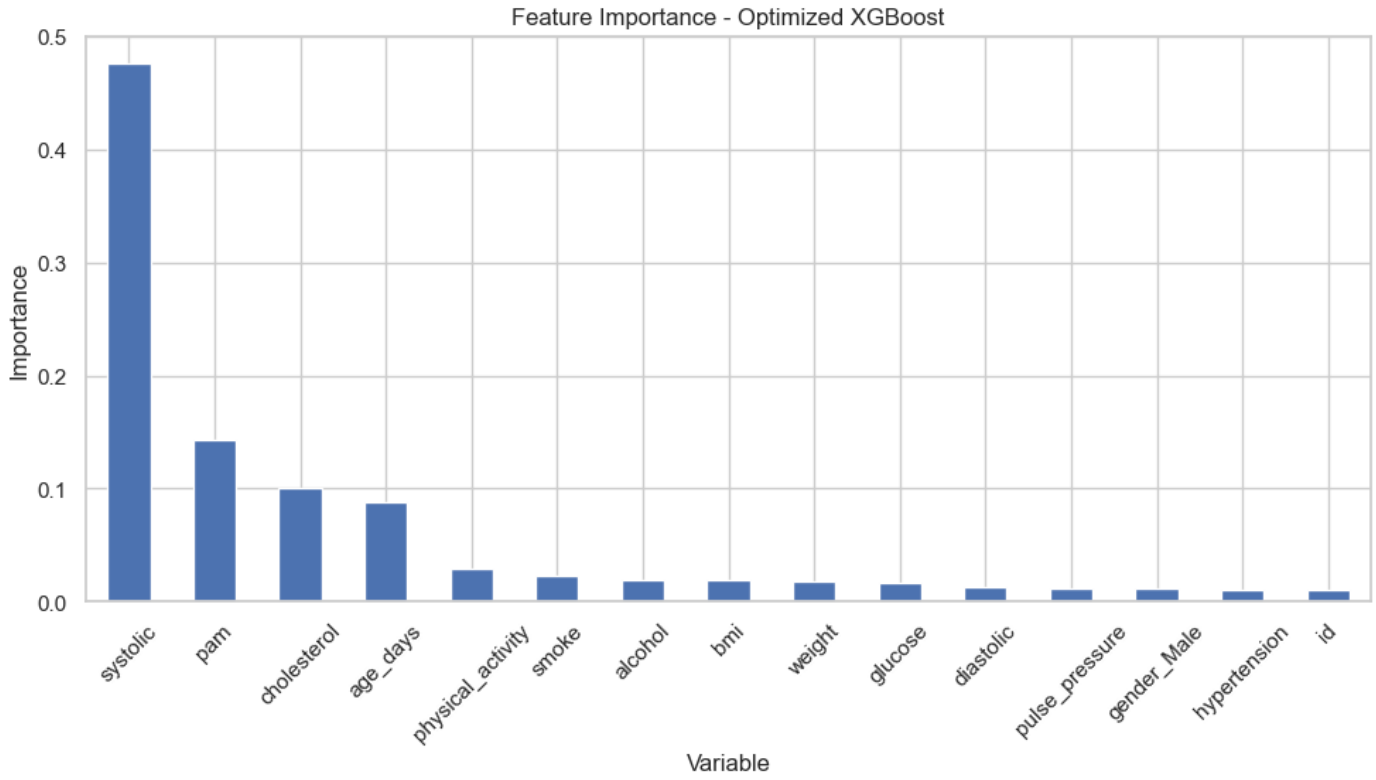
- Matriz de confusion



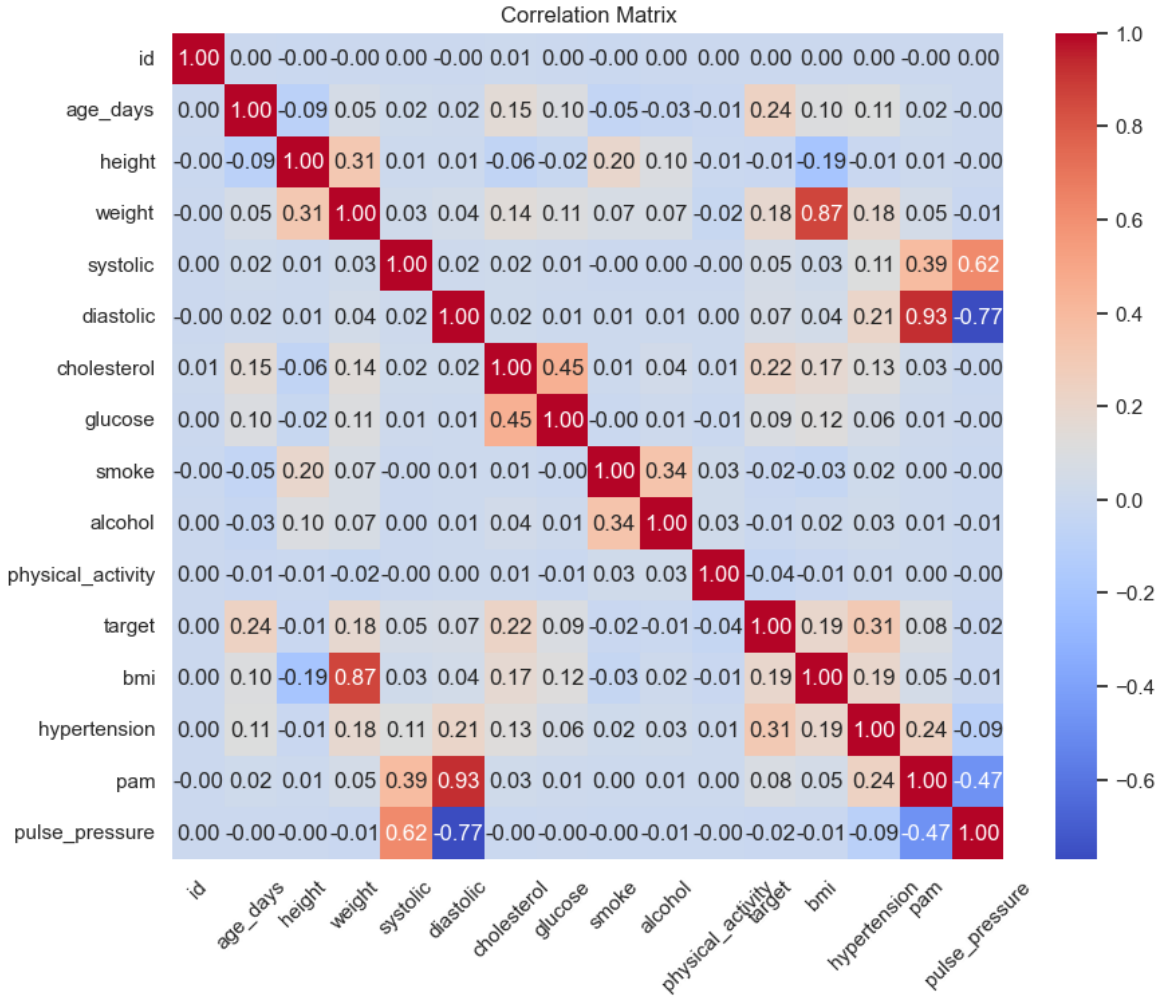
- Curva ROC-AUC

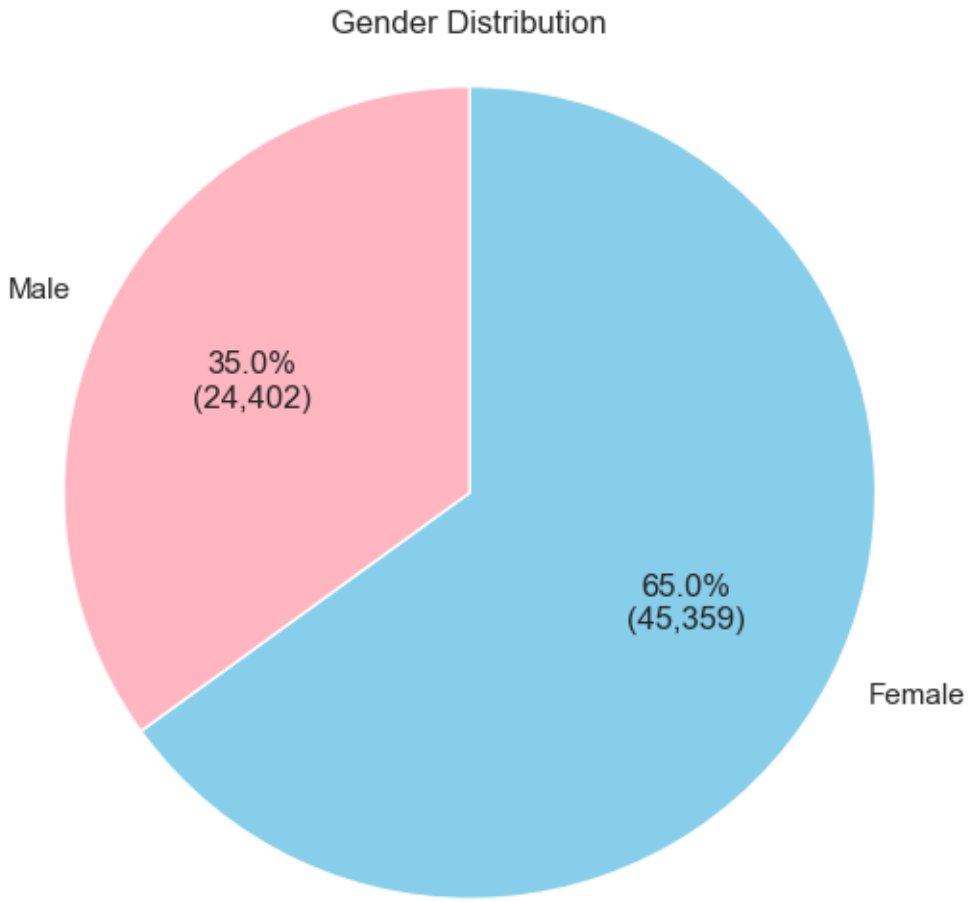


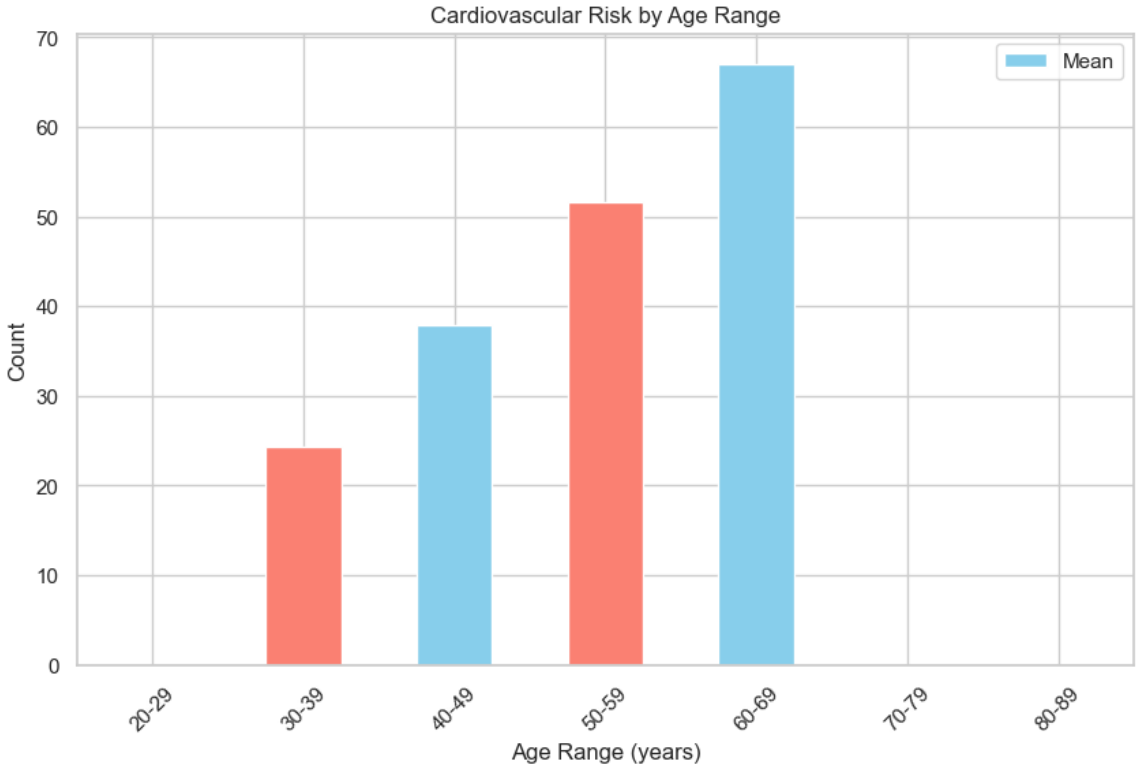
- Importancia de variables

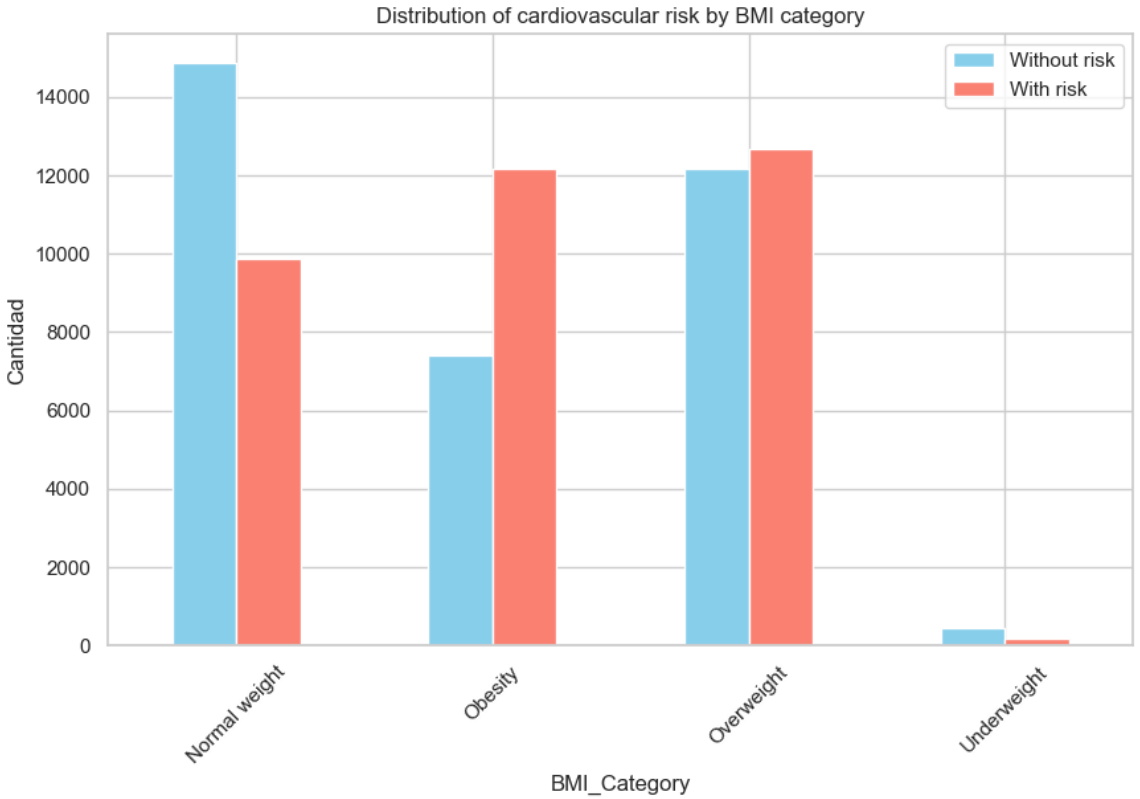


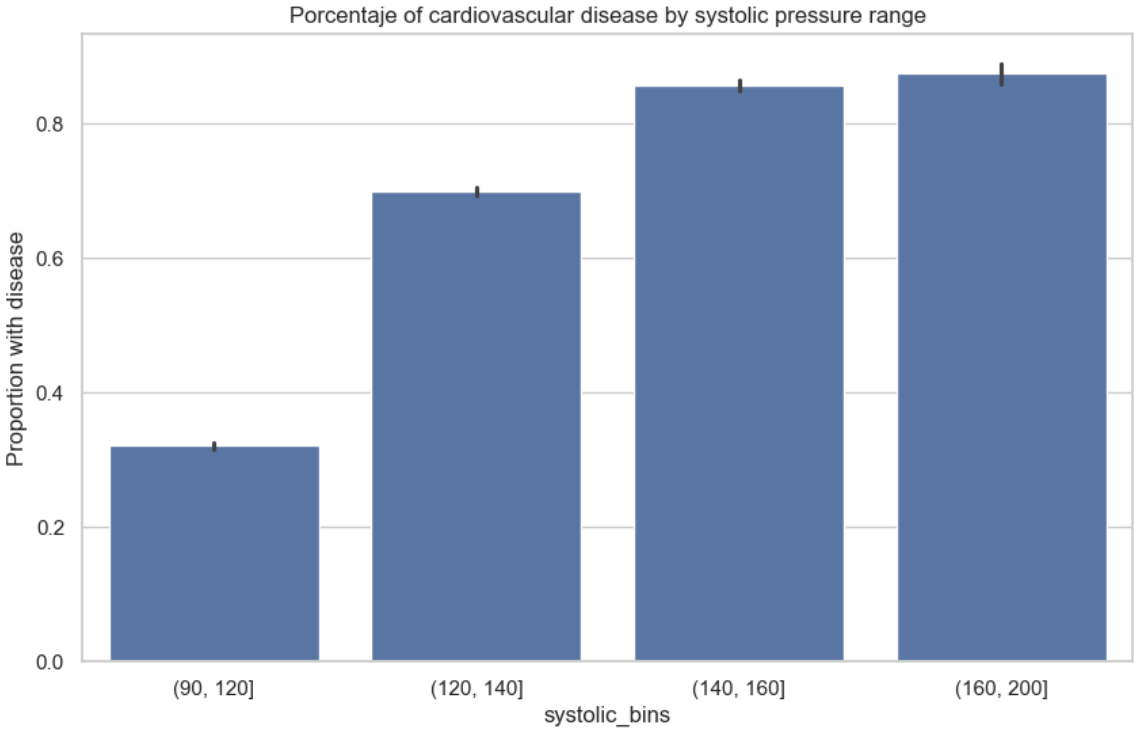
- Histogramas y boxplots de variables destacadas

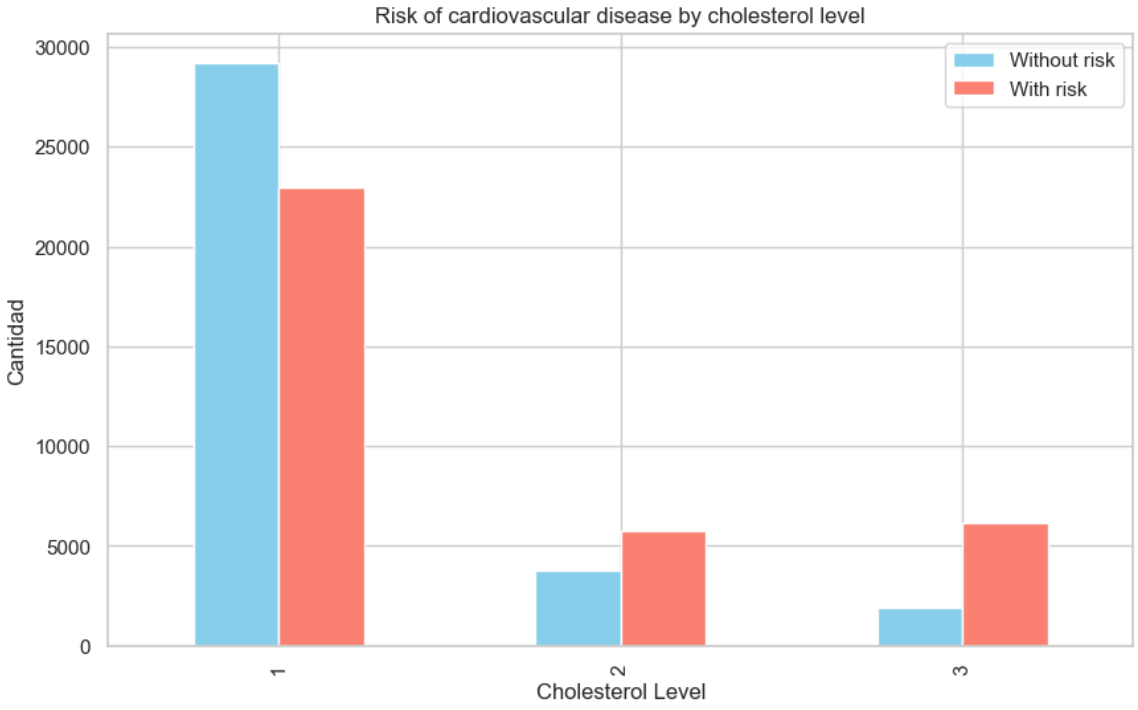


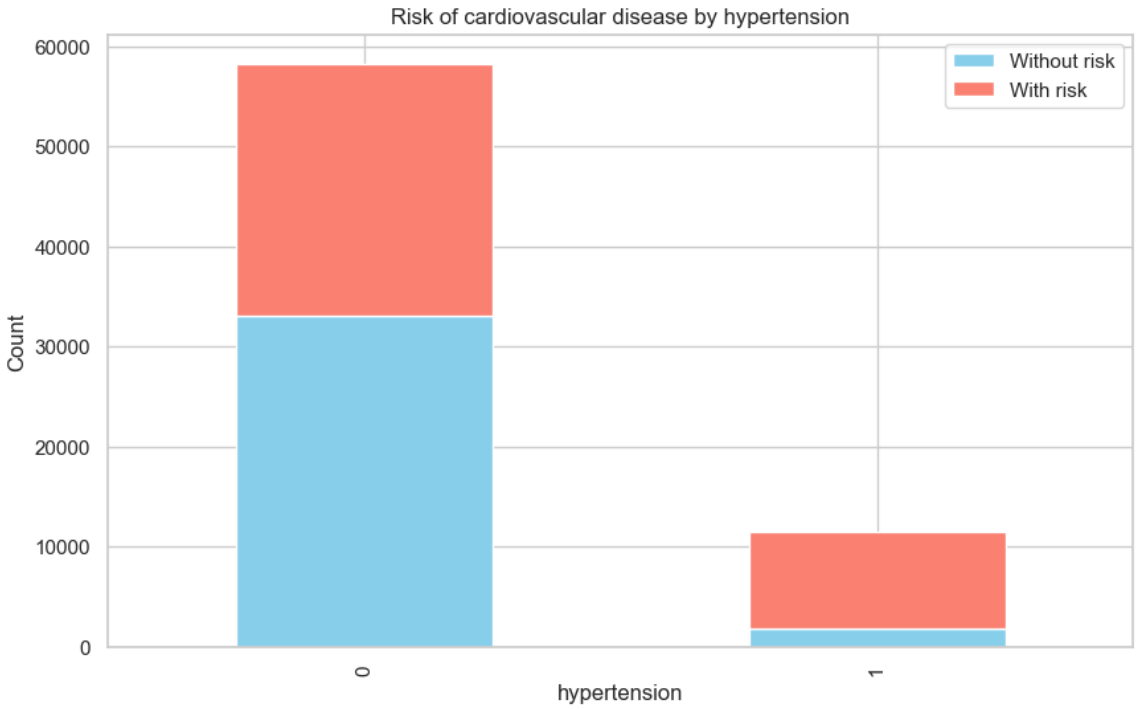

Gender Distribution
Male
35.0%
(24,402)
65.0%
(45,359)
Female


Cardiovascular Risk by Age Range
Mean
Count
70
60
50
40
30
20
10
0
20-29
30-39
40-49
50-59
60-69
70-79
80-89
Age Range (years)


Distribution of cardiovascular risk by BMI category
Without risk
With risk
Cantidad
14000
12000
10000
8000
6000
4000
2000
0
Normal weight
Obesity
Overweight
Underweight
BMI_Category


Porcentaje of cardiovascular disease by systolic pressure range
Proportion with disease
0.0
0.2
0.4
0.6
0.8
(90, 120]
(120, 140]
(140, 160]
(160, 200]
systolic_bins


Risk of cardiovascular disease by cholesterol level
Without risk
With risk
Cantidad
0
5000
10000
15000
20000
25000
30000
1
2
3
Cholesterol Level



☑ Conclusiones

El modelo XGBoost ofrece un buen equilibrio entre precisión y estabilidad. El proyecto demuestra cómo los datos médicos pueden ser aprovechados para predecir condiciones críticas como enfermedades cardiovasculares, ofreciendo un recurso valioso para decisiones clínicas basadas en datos.