Explicación del problema

Problema:

Enfermedades cardiovasculares son la causa N°1 de muertes alrededor del mundo, tomando un estimado de 17.9 millones de vidas cada año, el cual corresponde al 31% del total de muertes globales. Las fallas al corazón son eventos muy comunes causados por problemas cardiovasculares. Las personas con enfermedades cardiovasculares o aquellos quienes están en riesgo (debido a la presencia de diversos factores de riesgo como hipertensión, diabetes, hiperlipidemias u otra condición establecida) necesitan detección temprana, por lo que un manejo en el cual un modelo de Machine Learning pueda apoyar a la evaluación de diversas variables asociadas puede ser de gran ayuda.

Meta:

Clasificar/predecir si es que un paciente es propenso a tener fallas al corazón dependiendo de múltiples atributos.

Es una clasificación binaria con múltiples características numéricas y categóricas.

Atributos del Dataset:

Age: age of the patient [years]

Sex: sex of the patient [M: Male, F: Female]

ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]

RestingBP: resting blood pressure [mm Hg]

Cholesterol: serum cholesterol [mm/dl]

FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]

RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]

MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]

ExerciseAngina: exercise-induced angina [Y: Yes, N: No]

Oldpeak: oldpeak = ST [Numeric value measured in depression]

ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

HeartDisease: output class [1: heart disease, 0: Normal]

Conclusiones

- El dataset presentado sirve como un buen entendimiento de las clasificaciones binarias con la combinación de variables tanto numéricas como categóricas.
- La asesoría profesional es clave para poder entender la naturaleza del dataset, puesto que hay tecnicismos dentro de las variables que son entendibles por personal de salud y ellos tienen el detalle acerca de cómo poder traducirlo a la salud de los pacientes.
- El promedio de las edades de los pacientes que tienen enfermedades al corazón redondea los 56 años, mientras que las que son sanas están cercanas a los 50.
- Las relaciones entre ST_Slope y HeartDisease son proporcionales, ya que la mayoría de los casos en que las personas presentaron complicaciones al corazón tienen su ST_Slope a nivel alto.
- En su gran mayoría, la cantidad de personas que han sufrido de afecciones cardíacas tienen tipos de dolor asintomáticos. El MaxHR no tiene que ser necesariamente alto para mostrar afecciones en caso de tener un dolor asintomático, redondea valores de entre 100 y 140.
- Para los dolores que no provienen de anginas (NAP), normalmente las personas con más de 55 años son aquellas que presentan mayores afecciones cardíacas.
- Por algún motivo, hay algunos valores atípicos menores que 0 en la variable de Oldpeak.
- La gran mayoría de las personas que sufren afecciones cardíacas sufren anginas en el ejercicio.
- Se evaluaron diversos modelos de predicción con % de Accuracy bastante aceptables, entre ellos destacan los algoritmos de Regresión Logística y SVM, con un F1 Score de 89.1% y 88.3% respectivamente.