Capstone Project : 07

Problem statement:

Heart disease has been a major problem in the recent times. It is no longer true that only older people are more prone to the disease. Your objective is to predict

Data Description:

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The ‘num’ feature indicates the presence of heart disease.

Attribute Information:

age: age in years

sex: sex (1 = male; 0 = female)

cp: chest pain type

-- Value 1: typical angina

-- Value 2: atypical angina

-- Value 3: non-anginal pain

-- Value 4: asymptomatic

trestbps: resting blood pressure (in mm Hg on admission to the hospital)

chol: serum cholestoral in mg/dl

fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg: resting electrocardiographic results

-- Value 0: normal

-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

thalach: maximum heart rate achieved

exang: exercise induced angina (1 = yes; 0 = no)

oldpeak = ST depression induced by exercise relative to rest

slope: the slope of the peak exercise ST segment

-- Value 1: upsloping

-- Value 2: flat

-- Value 3: downsloping

ca: number of major vessels (0-3) colored by flourosopy

thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

num: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing

Instructions:

1. Perform the required data pre-processing to treat for missing values and outliers.
2. Perform exploratory data analysis to visualise the spread of each of the X variables and the relationship between the various X variables and the Y variable
3. Divide the given data into train and test sets
4. Predict how presence of heart disease by building classification models
5. Interpret how each of the X variables influence the conversion propensity
6. Evaluate the model performance measures and choose the most optimum model
7. Enlist your key findings based on the most optimum model and the respective feature importance