Name: SUKITH DE SILVA
Subject : Data Science Bootcamp SLASSCOM
Exam: Assignment 01

Description of Data Set:The UCI Heart Disease dataset is a widely used dataset in cardiovascular research and machine learning. It provides valuable information for predicting the presence or absence of heart disease in patients. The dataset consists of various clinical and demographic attributes that are commonly associated with heart conditions.

The dataset contains a total of 303 instances or records.

The dataset consists of 14 attributes (features) that provide information about the patients.

Features :input variables/Independent variables

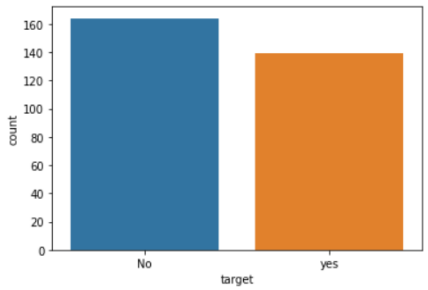
age: The person age in years
sex: The person sex (1 = male, 0 = female)
cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
trestbps: The person resting blood pressure (mm Hg on admission to the hospital)
chol: The person cholesterol measurement in mg/dl
fbs: The person fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes criteria)
thalach: The person 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: The number of major vessels (0-3)
thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)

Target Variable/Dependent Variable

num/target: (0-4)

based on researches we can assigned
0: Absence of heart disease
1: Presence of heart disease (value 1-4 indicates increasing severity)

Based on the count plot of target (dependent variable) identify absence of heart diseases around 160 patients and around 140 patients having heart diseases.from this count we can assume that the data set is a well balanced data set.



Data Loading and Exploration:

The UCI Heart Disease dataset was loaded into a pandas DataFrame.
Librarie used for this analysis:Matplotlib,Seaborn,Numpy,sklearn (logistic regression)

We examined the structure of the dataset, checking for missing values and gaining a basic understanding of the features and target variable.In the data set we identify features ‘ca’ and ‘thal’ has values with ‘?’ (6 data points).due to less no of missing values count we drop these values.

Data Preprocessing:

We performed necessary data preprocessing steps, checking missing values, scaling numerical features, and encoding categorical variables

Exploratory Data Analysis:

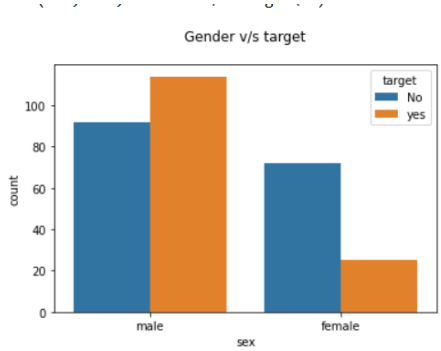
We conducted exploratory analysis to uncover trends and patterns within the dataset.

We visualized the distribution of features and examined their relationships with the target variable using plots such as count plots,displot,bar plot,jointplot,violinplot ,pairplot and correlation matrices.

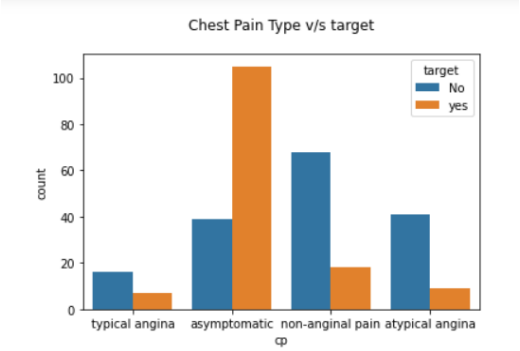
We investigated any notable patterns, outliers, or significant correlations between variables.

Key highlights

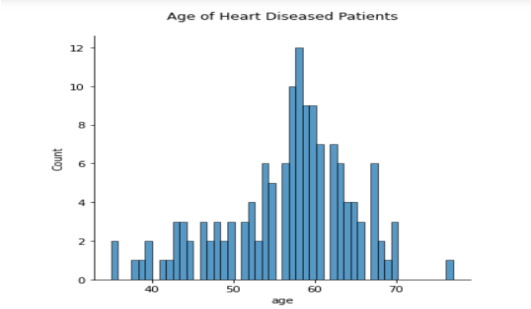
Compare to gender higher number of males are positive for having heart diseases compared to females



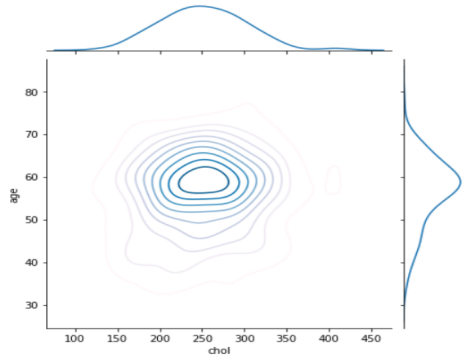
There are four types of chest pain, asymptomatic, atypical angina, non-anginal pain and typical angina. Most of the Heart Disease patients are found to have asymptomatic chest pain.



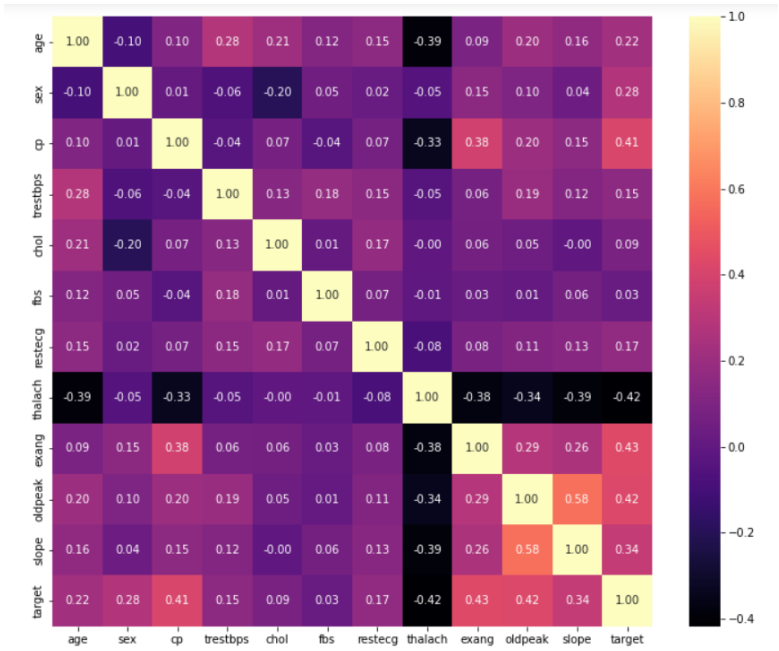
Heart Disease is very common in the seniors who are of age group 60 and above and common among adults which belong to the age group of 50 to 60. But it's rare among the age group of 19 to 40 and very rare among the age group of 0 to 18.



Most Heart diseased patients in their age of upper 50s or lower 60s tend to have Cholesterol between 200mg/dl to 300mg/dl.



Correlation shows whether the characteristics are related to each other or to the target variable. Correlation can be positive (increase in one value, the value of the objective variable increases) or negative (increase in one value, the value of the target variable decreased). From this heatmap we can observe that the 'cp' chest pain is highly related to the target variable.



Features such as exang, oldpeak, and maximum heart rate showed variations between with and without heart disease.

Some features displayed moderate correlations with the target variable, indicating their potential predictive power.

Do you think whether you can use a logistic regression model to predict heart disease?

Yes we can use logistic regression in this scenario

logistic regression can be applied to the Heart Disease dataset for binary classification tasks.Here dependent variable 'num'/'target' is converted into binary values (0: Absence of heart disease) & 1: Presence of heart disease (value 1-4 indicates increasing severity).Using Logistic regression algorithm used for binary classification we can predict the outcome whether patient has heart disease or not.

Accuracy: 0.87%
Precision: 0.81%
Recall: 0.875
F1 Score: 0.84

Accuracy: An accuracy of 0.87 (87%) indicates that the model is correctly predicting the presence or absence of heart disease for approximately 87% of the cases in the testing set. Generally, an accuracy above 80% is considered acceptable, but the specific context and requirements of your application should determine whether this level of accuracy is sufficient.

Precision: A precision of 0.81 (81%) suggests that out of all the instances predicted as having heart disease, 81% of them are correct. In other words, when the model predicts a positive outcome (presence

of heart disease), it is accurate around 81% of the time. Precision is an important metric, especially in medical contexts, as it measures the model's ability to avoid false positives.

Recall: A recall of 0.875 (87.5%) indicates that the model correctly identifies around 87.5% of the actual cases of heart disease in the testing set. Recall, also known as sensitivity or true positive rate, measures the model's ability to capture positive instances correctly. In this case, the model has a relatively high recall, suggesting that it performs well in detecting instances of heart disease.

F1 Score: The F1 score combines precision and recall into a single metric that balances both metrics. With an F1 score of 0.84, the model achieves a good balance between precision and recall. It indicates that the model's performance is favorable, considering both the ability to minimize false positives and false negatives.

Overall, the provided evaluation metrics suggest that the logistic regression model performs reasonably well on the UCI Heart Disease dataset.