

CIS 593

Special Topics in CIS

Group 5

Name: Sai Shanmuka Teja Bagadi

Name: Varshitha Boddu

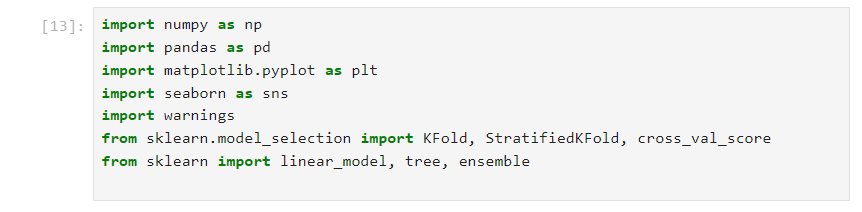
Professor: Apostolos Kalatzis

**Introduction**

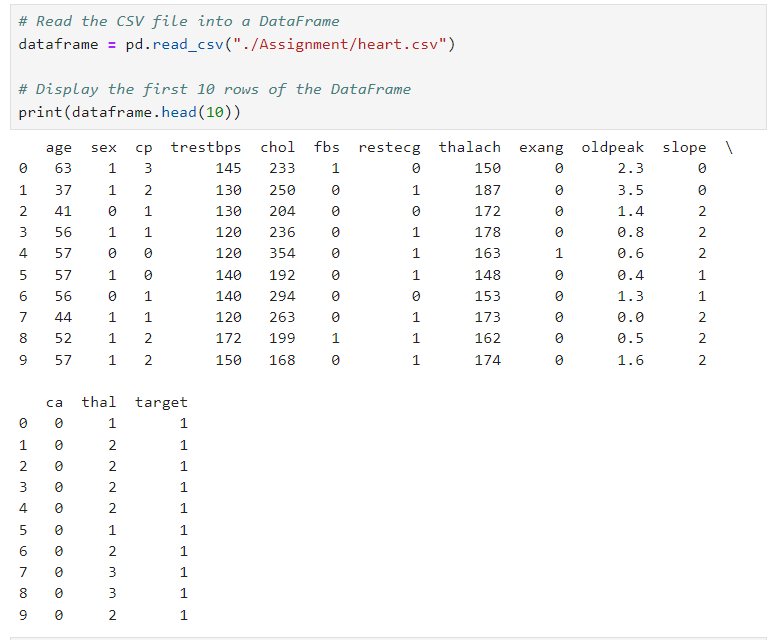
The project starts by downloading [the Heart Dataset](https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset) from the Kaggle Dataset. The dataset comprises several health parameters that correspond to a person’s heart health. The next process will be to process the dataset so that it can fit a machine learning algorithm to learn. The data, therefore, will be split into training data and test data, as the algorithm will need them to evaluate the model performance. Then, it will be possible to feed the training data to the machine learning model including Logistic Regression, Random Forest Classifier, Support Vector Machine, K Nearest Neighbours, and XG Boost. After training the model, the project will have a trained model, and it will be possible to predict whether a person has heart disease or not.

**What factors influence the likelihood of heart disease in patients according to the dataset?**

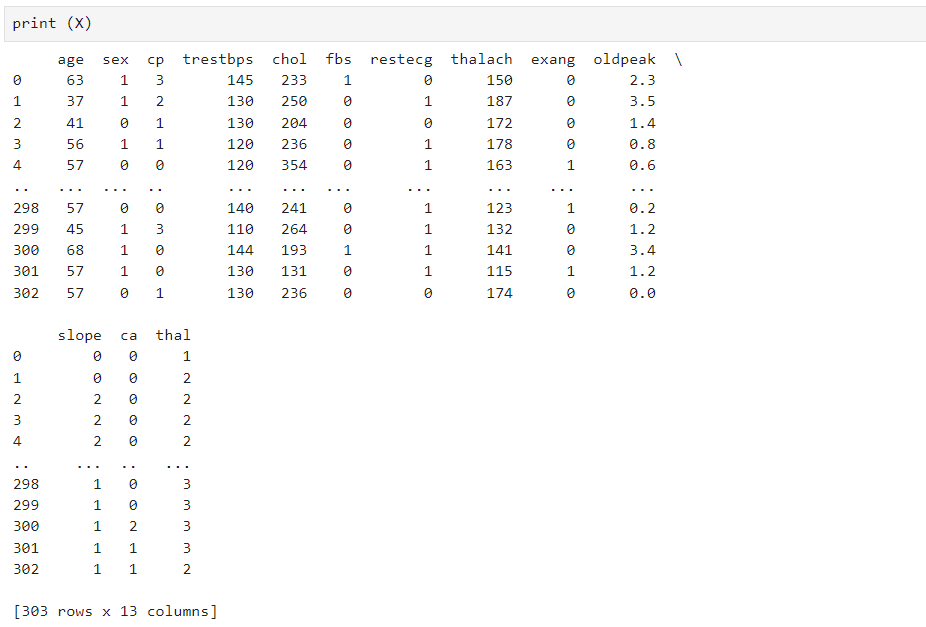
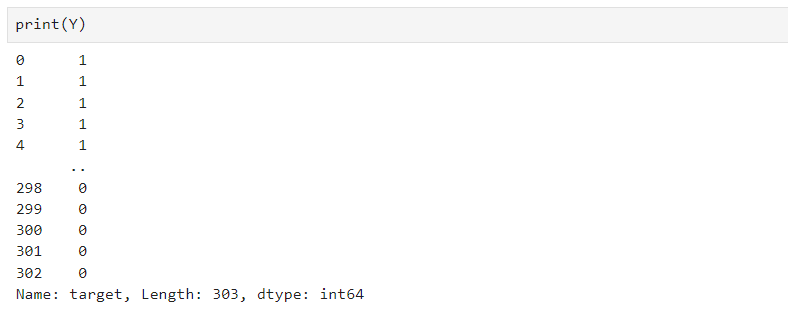
Import the libraries and dependencies.



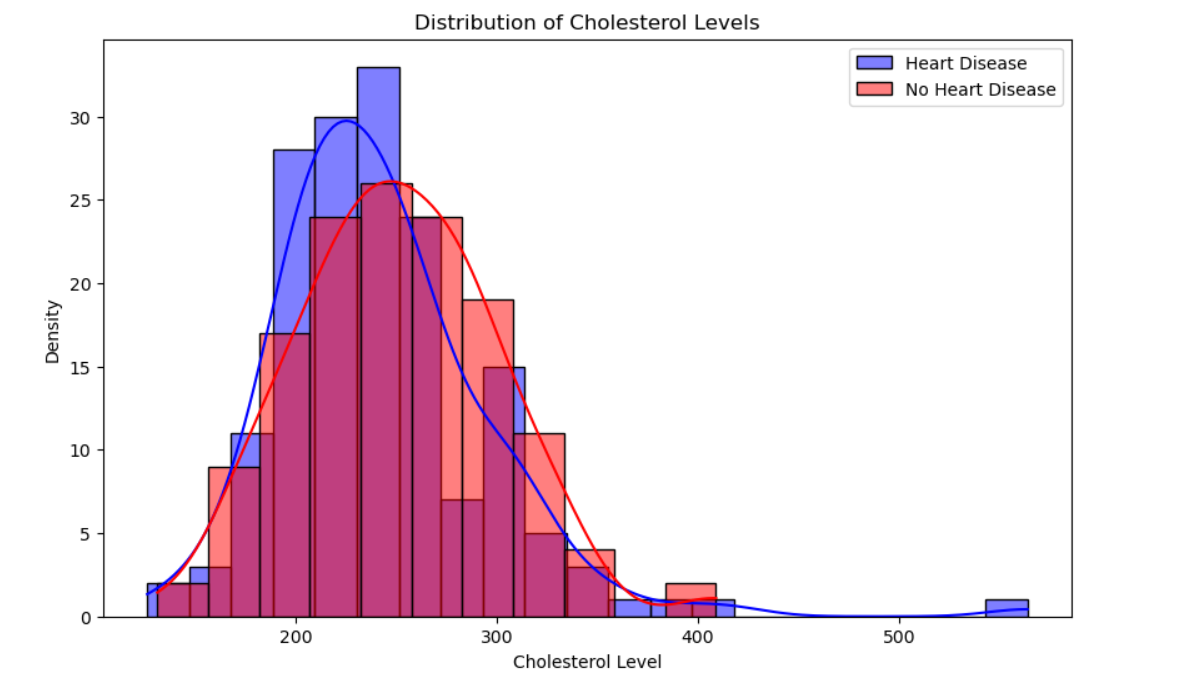
In data collection and preprocessing, the project starts by uploading the CSV dataset to the Pandas DataFrame so that it can be possible to determine factors that influence the likelihood of heart disease in patients according to the dataset. See the first ten rows of the parameters in the data frame below.



More so, to understand the basic information of the data frame, it is possible to find the number of rows and columns, which was (303, 14). The next step was to check for missing values, which was conducted using the dataframe.isna().sum() function. The data didn’t show any null values or duplicates, so the data is good for further analysis. Also, the distribution of the data was found to be fine after checking the defective and healthy heart values (1, 0), respectively, which shows there was no need to process the data.

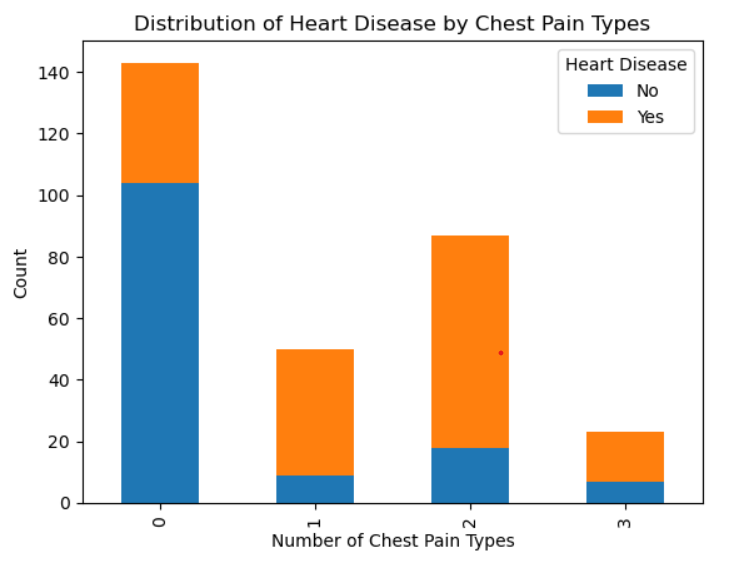
The next important step was to split the data into Features and Targets, as shown below.

**What is the distribution of cholesterol levels among patients with heart disease, and how does it compare to those without heart disease? Please visualize this using a histogram or a density plot.**

After filtering the dataset to separate patients with heart disease (target = 1) and without heart disease (target = 0). The next step was to set up the figure and axis for the plot using plt.figure(figsize=(width, height)). Then, it was time to plot the density plot for cholesterol levels for each group using sns.histplot() from the Seaborn library. Specify the column containing cholesterol levels for each group and set the color parameter to differentiate between the groups. Also, set kde=True to overlay a kernel density estimate on the histogram bars. The last was to set the labels for the x-axis, y-axis, and title of the plot using plt.xlabel(), plt.ylabel(), and plt.title() functions, respectively. When the cholesterol level is low, the density of heart disease is high compared to that of those without heart disease.

**Is there a relationship between the number of chest pain types experienced by patients and the likelihood of heart disease?**

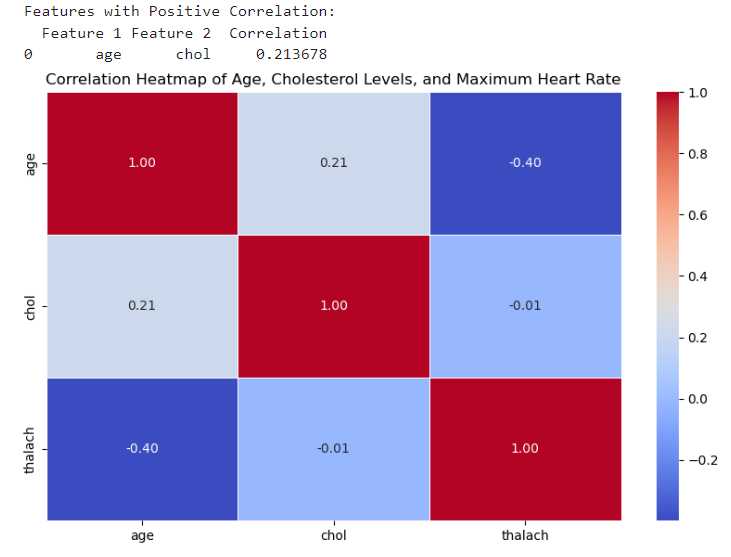
The project used a combination of visualization and statistical analysis to understand the relationship between the number of chest pain types experienced by patients and the likelihood of heart disease. One approach was to create a cross-tabulation or a bar plot to see how the distribution of heart disease varies with the number of chest pain types. The project used pd.crosstab() to create a cross-tabulation between the number of chest pain types (cp) and the presence or absence of heart disease (target). Then, it was possible to visualize this cross-tabulation using a stacked bar plot to see the relationship between the number of chest pain types and the likelihood of heart disease.



This shows that Type 0 had the highest count of people with no heart disease while cp 2 had the most counts of people with heart disease.

**Can you use a heatmap to depict the correlation between different variables in the dataset, such as age, cholesterol levels, and maximum heart rate achieved?**

This is an important figure that can be determined to identify the correlations between different variables in this dataset. The project will then generate a heat map to show the correlation coefficients existing between age, cholesterol levels, and maximum heart rate achieved. From the chart, it is evident that the highest positive correlation is between age and chol, with a 0.2137 positive correlation.

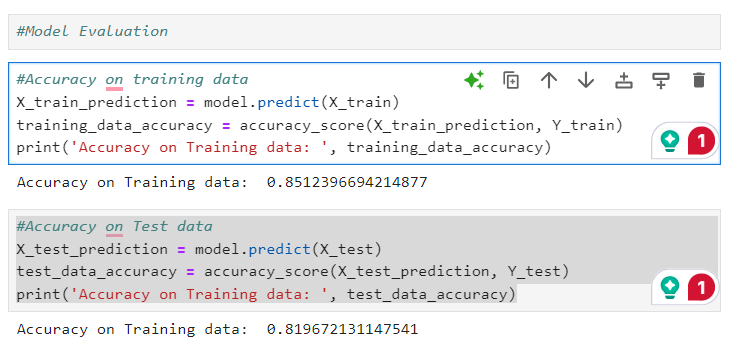


The next step will be to conduct a train-test split where the whole dataset will be split into trainset and testset, which contains 80% train and 20% test. This will be important since the project will use train set of classifiers to train the model, and a test set will be used to predict the performance of the model by different classifiers. The project ended up having (303, 13) (242, 13) (61, 13) for X.shape, X\_train.shape, and X\_test.shape respectively.

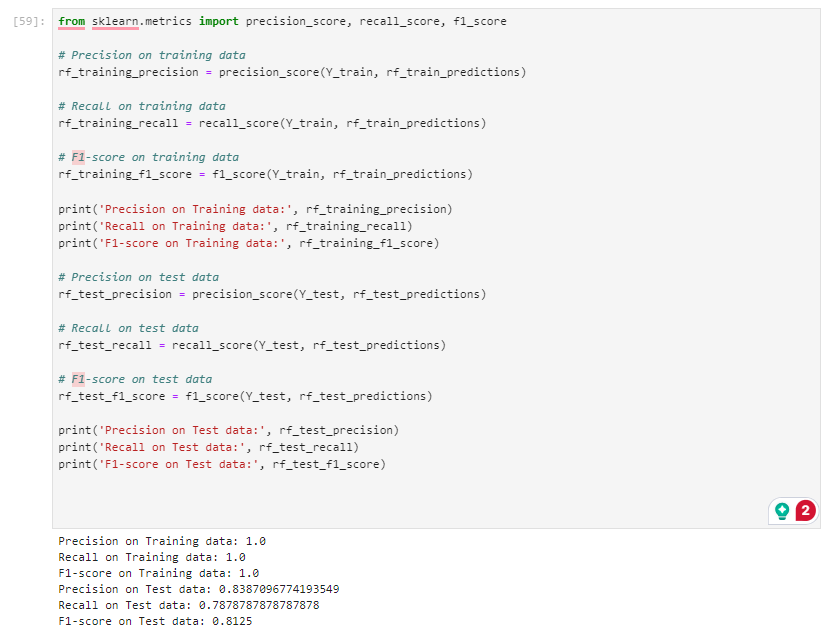
**Model Evaluation**

The promise of machine learning in forecasting and handling heart disease underscores the importance of ongoing research and innovation to enhance precision, versatility, and practicality. Enhancing prevention and treatment methods through cutting-edge technologies like AI and machine learning is essential for mitigating the disease's burden on both individuals and communities. The utilization of Scikit-Learn's model facilitated the estimation of the probability of heart disease occurrence by considering a range of risk factors. This project considered how best this algorithm can be tuned to provide the best outcome of the models. The algorithms with the highest accuracy levels are regarded as the best to make predictions as shown below.

**Logistic Regression**



**Random Forest Classifier**



**Support Vector Machine**



**K Nearest Neighbours**



**XG Boost**



**Conclusion**

After closely evaluating different machine learning algorithms, Random Forest and Xg Boost showed high accuracy, precision, and F1-score. This project has considered Random Forest as the best model to consider to predict heart disease among patients. After determining this, the algorithm determined the model was integrated with Random Forest, and it correctly predicted the entered data. This led to the creation of the best model. This project enables the prediction of real-time heart disease by analyzing patient data through the Random Forest Algorithm, resulting in precise machine learning-based predictions for heart disease.

