DECISION TREE ASSIGNMENT REPORT

- **The libraries** that are needed are imported
- Loaded the **Dataset**
- Understood the meaning of the attributes and the **Data Types of The Attributes**
- Obtain the **Statistical Summary** of the numerical values
- Found the **Unique Values** of the categorical data
- Checked for **Duplicate Values And Null Values**
- Found null values and visualized them using heatmap
- And filled the null values with the mean method and confirmed it with the heat map again
- Plotted the **Histogram** and the **Density Plots** to get the distribution of the data
- Also plotted the **Box Plot** and identified the visualization
- Tried the **Scatterplot** between attributes and found **No Significant Relation**
- Using **Pie Chart** found out the percent of people respective to their stage of heart disease
- Applied **One Hot Encoding** for the categorical features with unique values less than 3
- There are some **misplaced values** in the dataset under the attribute “Exang” and they were as follows, FALSE for False and TURE for True and they were rectified
- And after that applied **Label Encoding** to categorical features with unique values more than 3
- Found out the **Correlation Matrix** as now every attribute is numerical value
- And used the correlation matrix for **Heat Mapping Visualization**
- Found out the **skewness** of the data
- Applied **both Standard Scaler and Min Max Scaling** for the data
- Applied **The Log Transformation** for the skewed data and added that as new feature into the dataset
- And some **null values appeared after log transformation** so removed them
- Time for **Decision Tree Classification**
- Applied **The Train Test split** and initialized the model
- **Trained the model** and predicted the test set
- Calculated the **Accuracy** of the model and as the it was multi class, the **Average Method** was changed for the metrics **Precision, Recall, F1 Score, Roc_Auc**
- And plot the roc curve, and as the roc curve takes str as input converted the datatype of the target while plotting the curve
- Went on with **hyper parameter tuning** with **Grid Search CV** to identify the best of all and decided the best parameter to be `{'criterion': 'entropy', 'max_depth': 10, 'min_samples_split': 10}`

- And found that the metrics apart from accuracy, precision, recall, F1 score were good with the initial metrics and roc_auc was better off with the grid search method

- And finally, the **Decision Tree Was Plotted**