Bio-Signal Analysis for Smoking

Agenda

**Importing the Libraries Loading the Data**

**02**

**01**

**Data Cleaning One Hot Encoding**

**04**

**03**

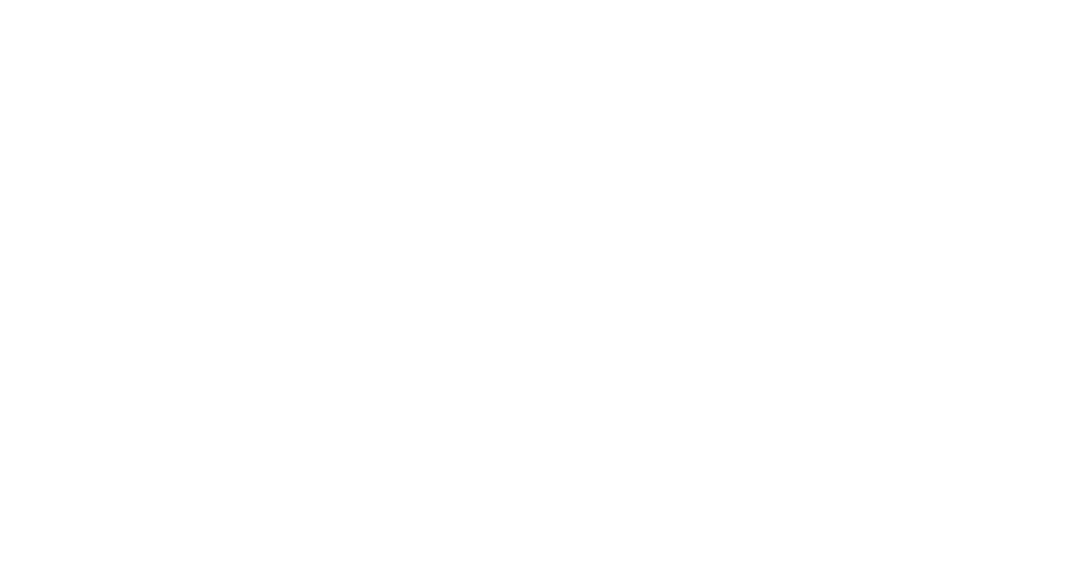
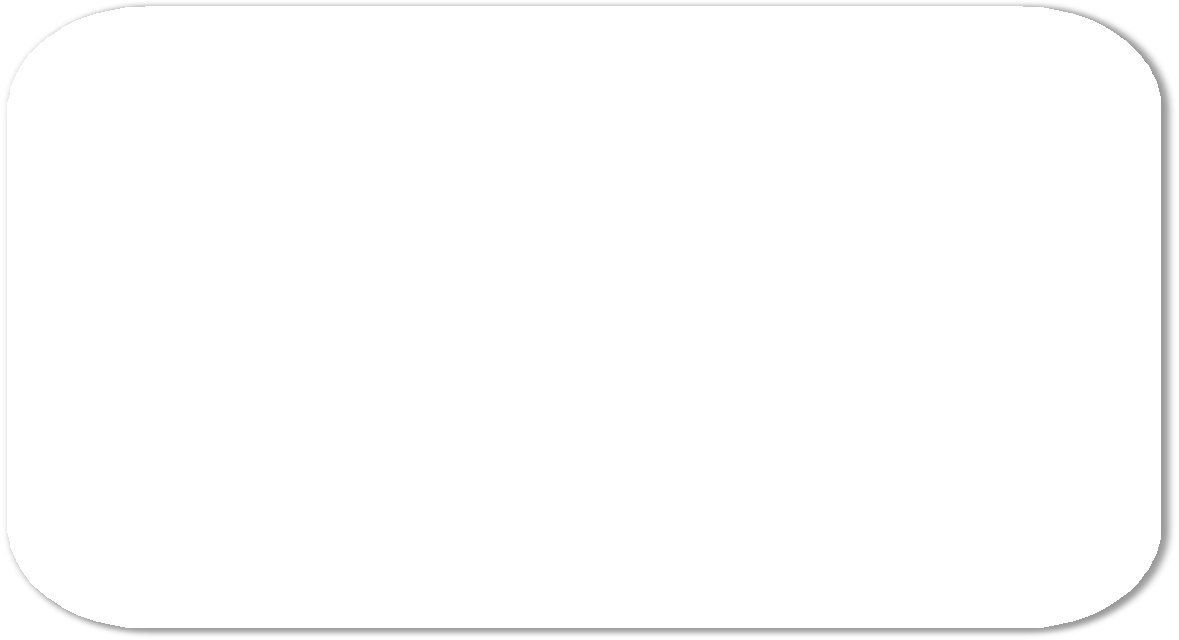
**Feature Selection Bagging Algorithms**

**06**

**05**

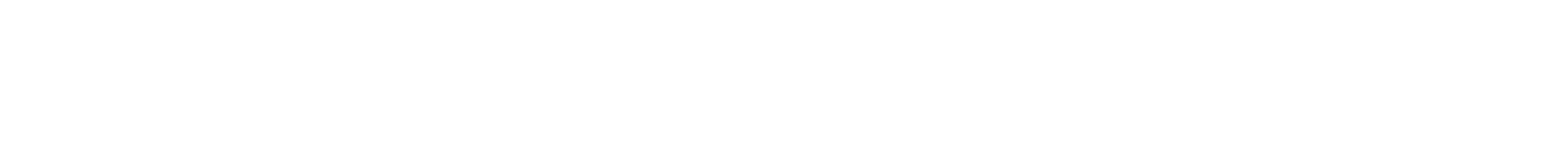
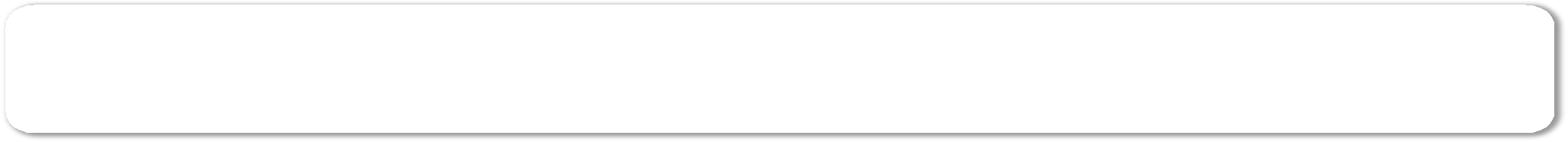
Problem Statement





I build an intelligent system from the data to determine the presence or absence of smoking in a person through bio-signals. Given a person’s information, build a machine learning model that can classify the presence or absence of smoking.

Dataset Information



This dataset is a collection of basic health biological signal data which

contains around 55K record with 27 attributes.



Dataset Information

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| ID | index |
| gender | gender of a person (M or F) |
| age | age of a person (5-years gap) |
| height(cm) | height of a person |
| weight(kg) | weight of a person |
| waist(cm) | waist circumference length |
| eyesight(left) | left eyesight |
| eyesight(right) | right eyesight |
| hearing(left) | hearing pulse in left ear |
| hearing(right) | hearing pulse in right ear |
| systolic | Blood pressure |

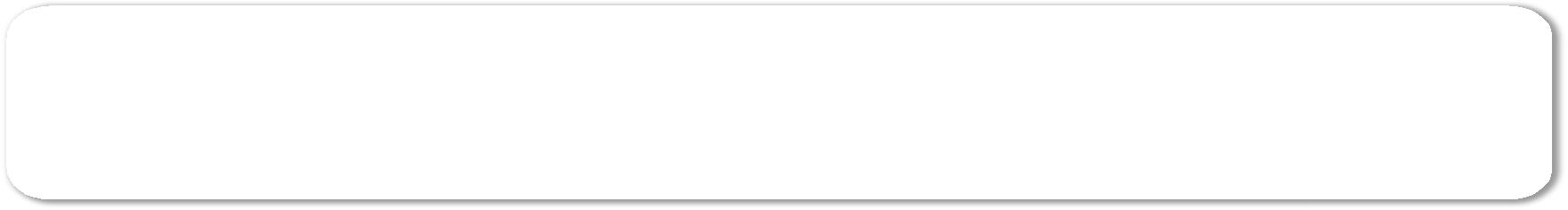
Dataset Information

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| relaxation | Blood pressure |
| fasting blood sugar | Blood test |
| Cholesterol | total |
| triglyceride | Lipid found in blood |
| HDL | cholesterol type |
| LDL | cholesterol type |
| hemoglobin | Transporting oxygen in blood |
| Urine protein | Excess of bloodborne proteins in urine |
| serum creatinine | Amount of creatinine in blood |
| AST | glutamic oxaloacetic transaminase type |
| ALT | glutamic oxaloacetic transaminase type |

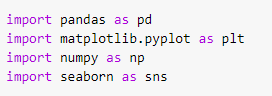
Dataset Information

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| Gtp | γ-GTP |
| oral | Oral Examination status |
| dental caries | Tooth decay |
| tartar | tartar status |
| smoking | Smoker (0 or 1) |

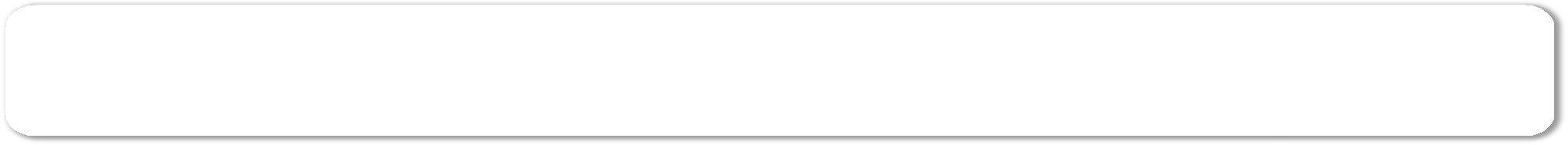
Importing the Libraries



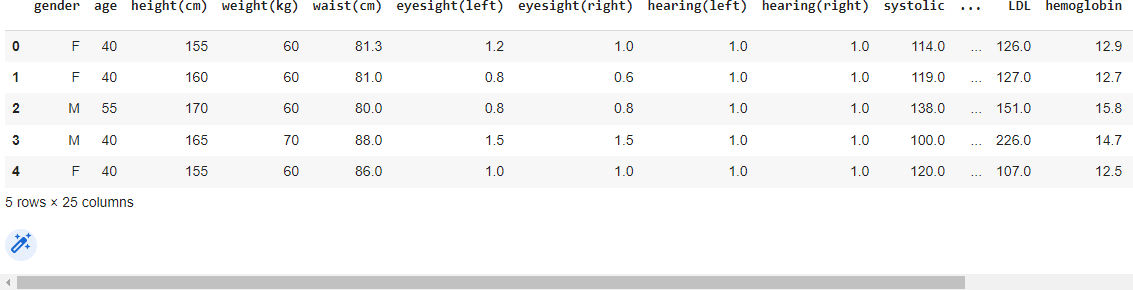
We start off this project by importing all the necessary libraries that will be required for the process.

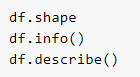
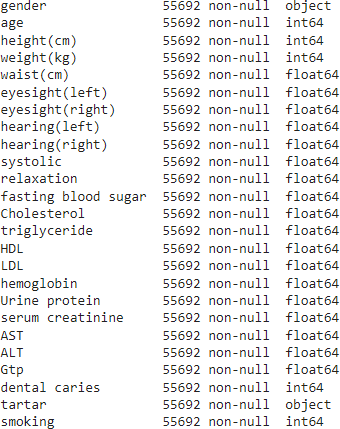


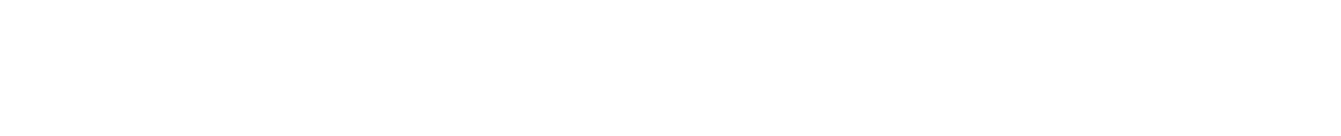
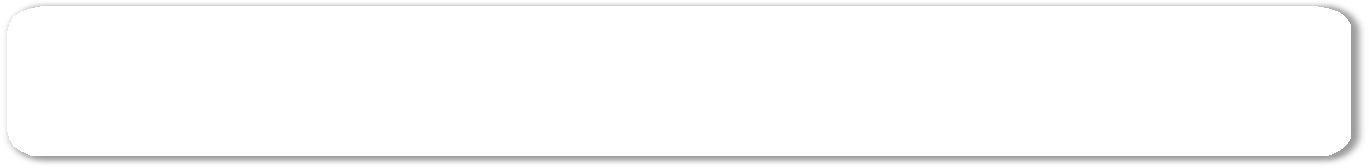
Loading the Data



Loading the data and removing the irrelevant columns.

C:\Users\HP\Downloads\1.png

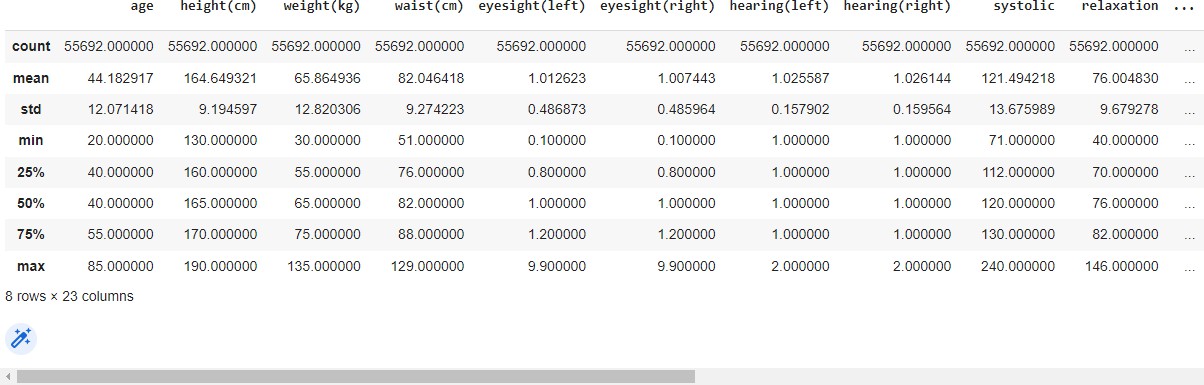
C:\Users\HP\Downloads\1.pngLoading the Data

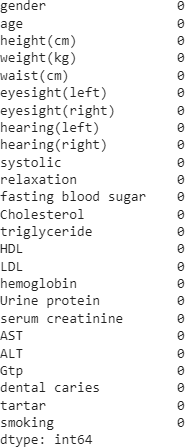


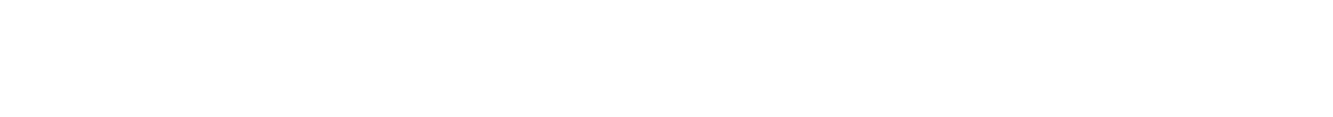
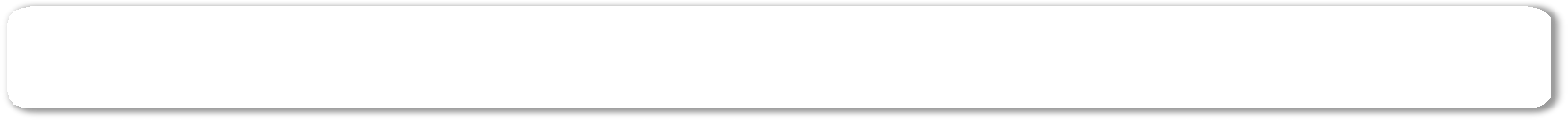
Checking the shape of a dataframe and datatypes of all columns

along with calculating the statistical data.

Loading the Data



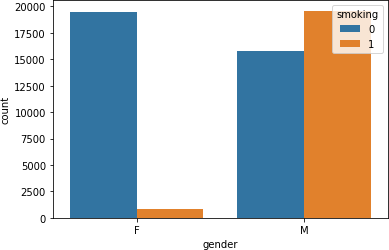
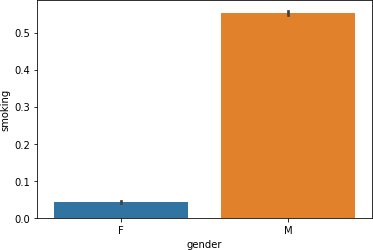
Missing Values

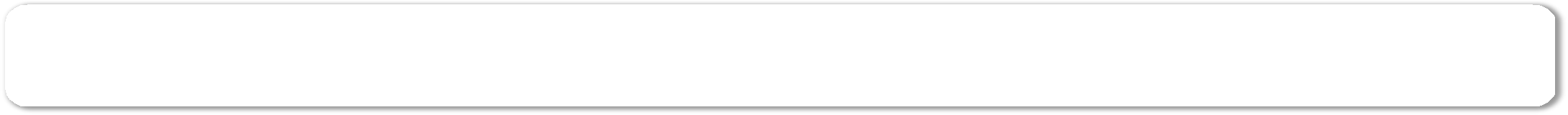


Checking out the missing values in a dataframe

C:\Users\HP\Downloads\8.png

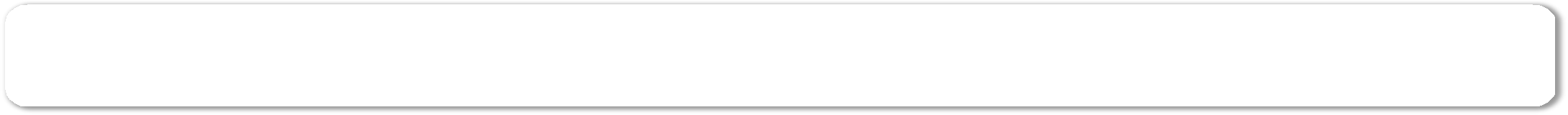
Data Visualization

C:\Users\HP\Downloads\1.pngC:\Users\HP\Downloads\1.png



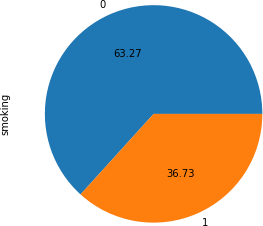
We can clearly see from the below graph that most smokers are men

Data Visualization

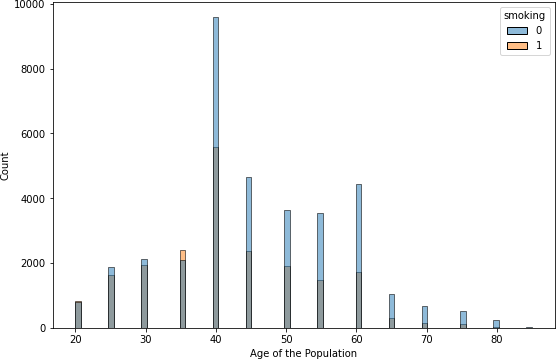
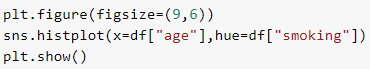


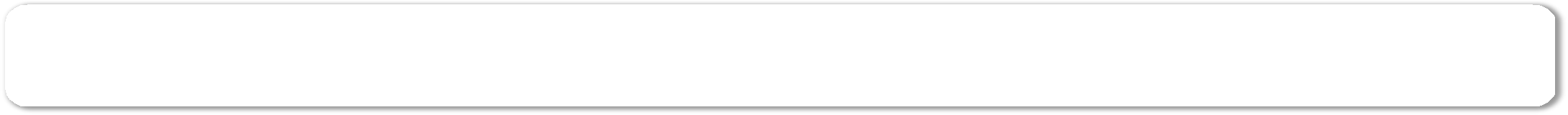
There are 36.73 percent of the people who are smoking ciggarette.

C:\Users\HP\Downloads\1.png



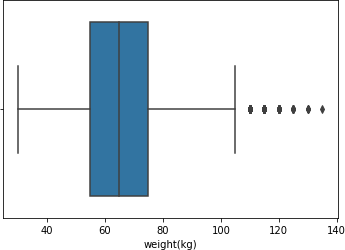
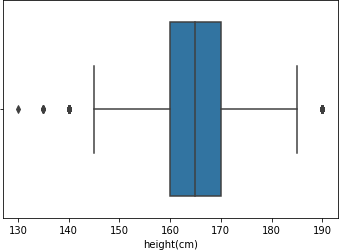
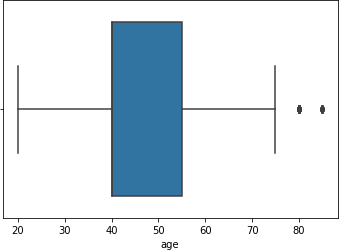
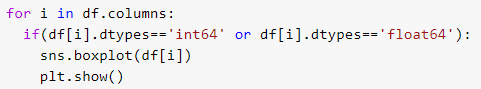
Data Visualization

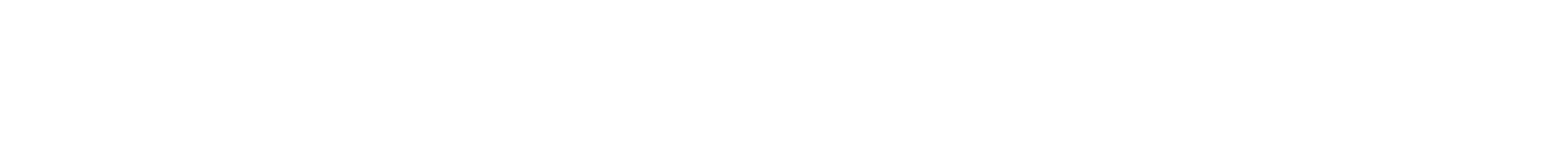
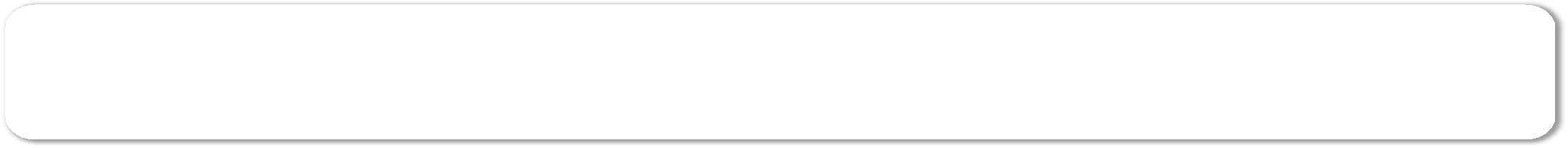




Most number of smokers are having the age 40

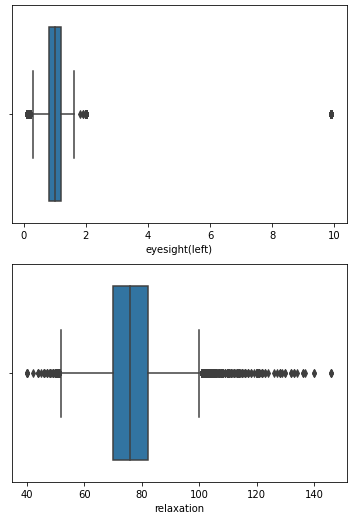
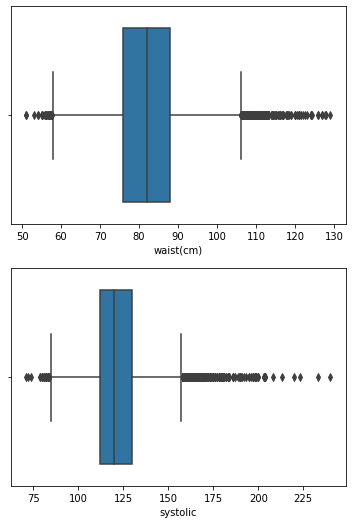
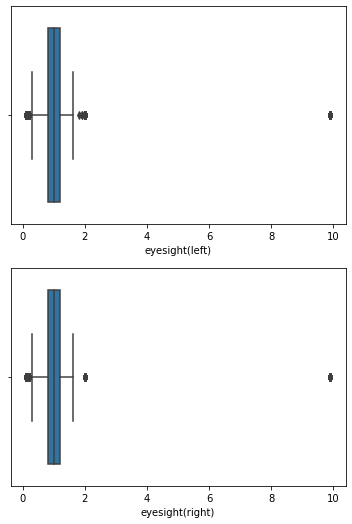
Data Visualization



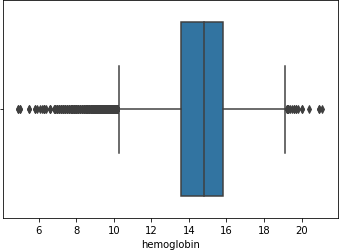
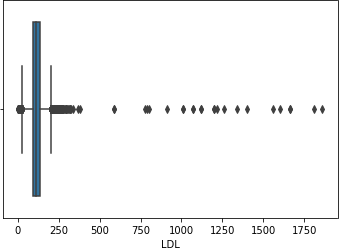
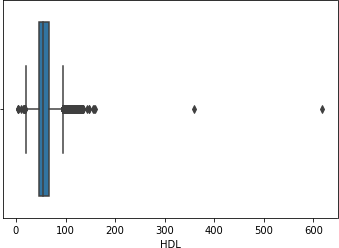
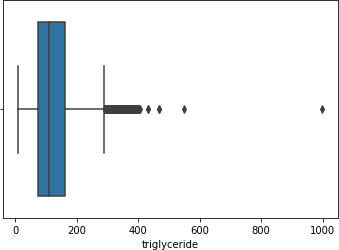
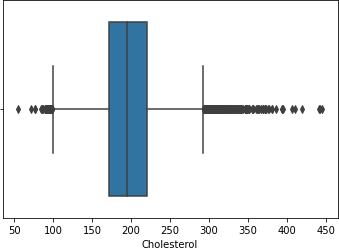
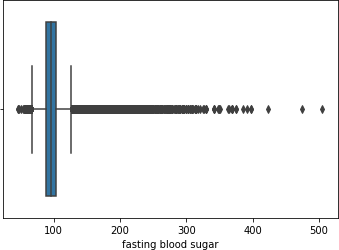


Representation of columns using boxplot to detect outliers. Here outliers represent natural variations in the population, and they should be left as is in the dataset. These are called true outliers. Therefore for this dataset we will not remove outliers.

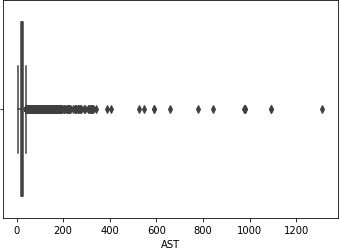
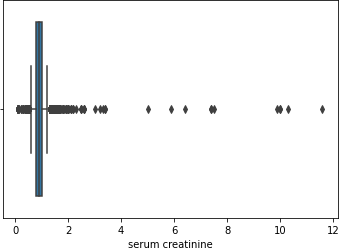
Data Visualization

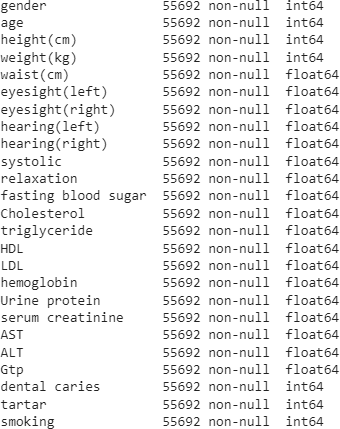
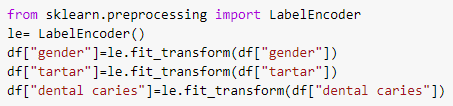


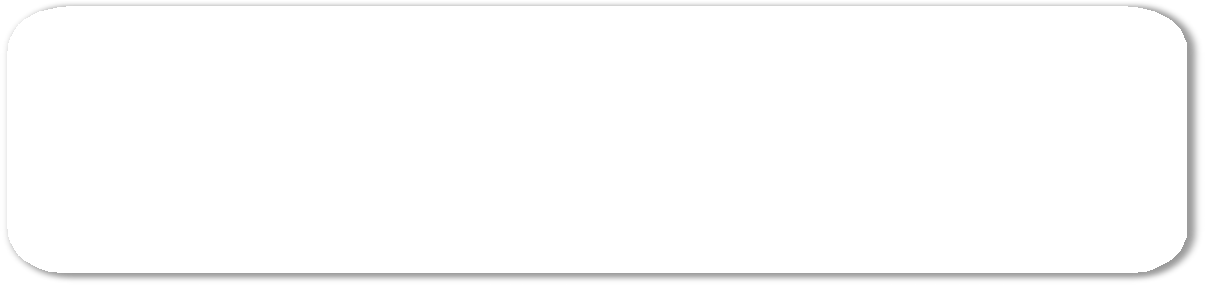
Data Visualization



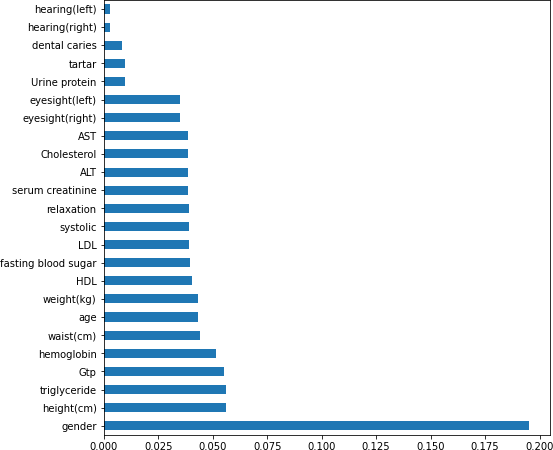
Data Visualization

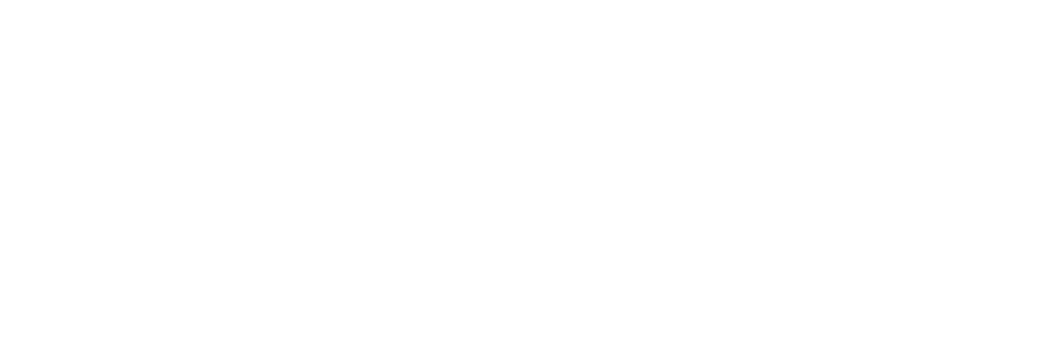
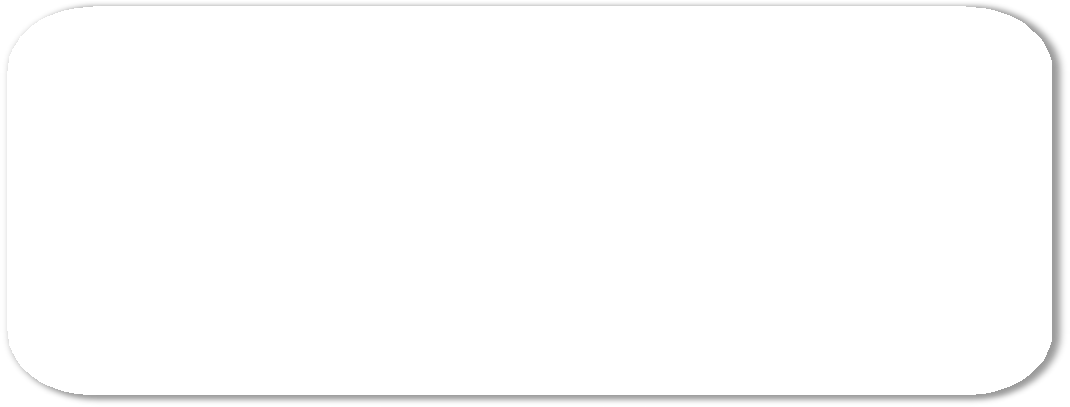


Data Cleaning



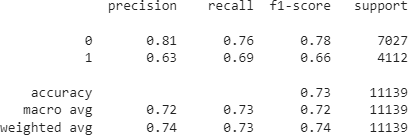
Performing One Hot Encoding for categorical features of a dataframe

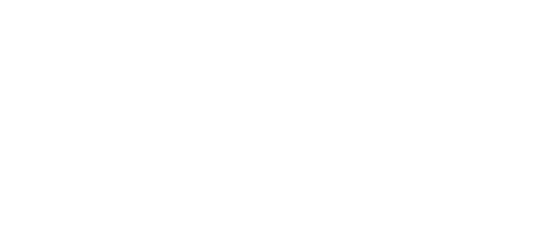
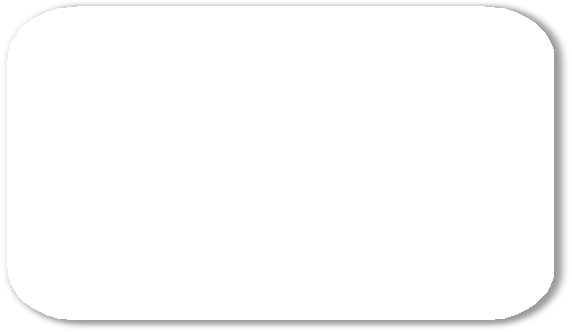
Feature selection using feature importance



Feature importance is a technique that calculate a score for all the input features for a given model. So out of 24 features we will select the top 15 features based on the score.

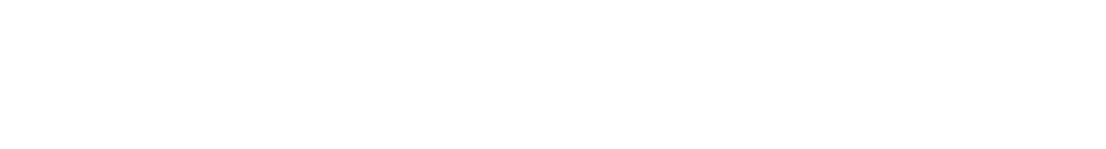
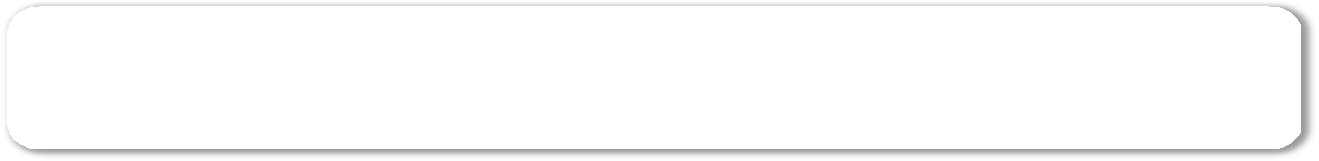
Logistic Regression

C:\Users\HP\Downloads\1.png



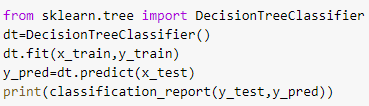
Calculating accuracy and generating the classification report of Logistic Regression

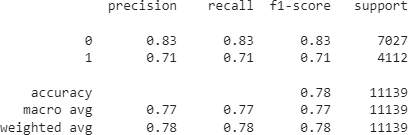
Decision Tree



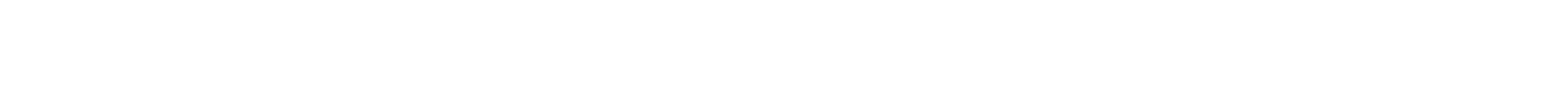
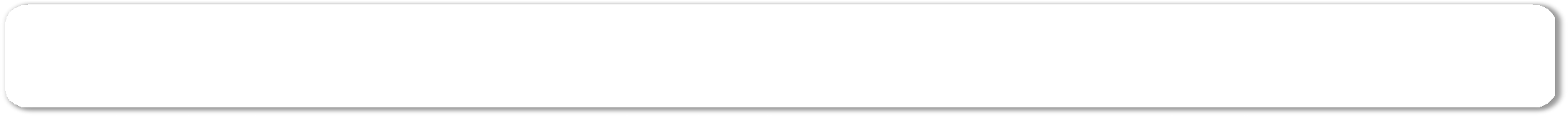
The accuracy of the logistic regression model is

78 percentage



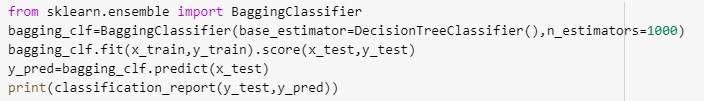


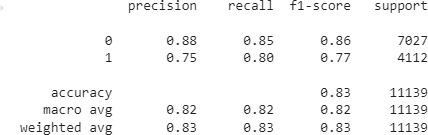
Bagging Algorithm – Bagging Classifier



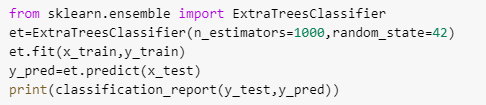
Bootstrap Aggregation or bagging involves taking multiple samples from the training dataset

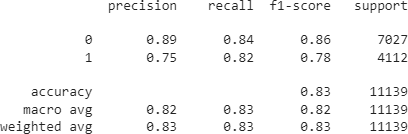
(with replacement) and training a model for each sample.





Bagging Algorithm – Extra Trees





Bagging Algorithm – Random Forest

