Summary of the project

In this project the steps taken are

->Missing values

->Outlier handling

->EDA

->.Feature Engineering

->Model Training

->Model Evaluation

->Model Comparision

1.Missing values :

First the count of missing values is observed in each column then the each column is filled with their approporiate values

For example for any numerical columns missing values are filled with their median and for the categorical values it is filled with the first mode value

Some time if the values are not known then they are filled with none or unknown

After filling the columns are checked again for the number of missing values in each column

2. Handling outliers :

In this the columns most of them doesn't have the outliers the columns which has the ouliers are dealt and scaled using the robust scaler which also handles outliers

3.EDA(EXploaratory data analysis):

In this data is plotted as graphs and observed in different types based on the target variable

4.Feature engineering :

In this first the feature selection is done i.e using the correlation method the features are selected

After that the columns are scaled for numerical columns using the standard scaler and for the categorical columns the one hot encoding is used

Then the data is separated as the X containing the features and Y containing the target variable

The x and y are divided as the train and test data in a particular percentage as xtrain,xtest and vice versa

Model Selection :

For this project the classification models are used mainly for this project 5 classification models are used

1.random forest

2.SVC

3.Logistic

4.Multinomial NB

5.XGBM

And all the models are evaluated and tested in which random forest and xgbm did a great job with high accuracy