greatlearning
Power Ahead

MACHINE LEARNING
PROJECT
BUSINESS REPORT

VAISHNAV U
PGP-DSBA ONLINE
19/03/2023

Table of Contents

Content

Problem-1

List of tables

List of figures

Transport data

Summary

You work for an office transport company. You are in discussions with ABC Consulting company for providing transport for their employees. For this purpose, you are tasked with understanding how do the employees of ABC Consulting prefer to commute presently (between home and office). Based on the parameters like age, salary, work experience etc. given in the data set 'Transport.csv', you are required to predict the preferred mode of transport. The project requires you to build several Machine Learning models and compare them so that the model can be finalized.

Introduction

The data consists of 444 rows and 9 columns. The objective is to build several Machine Learning models and compare them so that we can find the choice of transport of employees in Abc company.

Data description

1	Age	Employee age.
2	Gender	Employee gender.
3	Engineer	Employee education.
4	MBA	Employee education.
5	Work Exp	Employee work experience.
6	Salary	Employee salary.
7	Distance	Distance travelled by employees to reach office.
8	License	Driving license.
9	Transport	Transport method used by employees.

Sample of dataset

Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
28	Male	0	0	4	14.3	3.2	0	Public Transport
23	Female	1	0	4	8.3	3.3	0	Public Transport
29	Male	1	0	7	13.4	4.1	0	Public Transport
28	Female	1	1	5	13.4	4.5	0	Public Transport
27	Male	1	0	4	13.4	4.6	0	Public Transport

Table 1.1

Problem-1

1.1 **Read the dataset. Do the descriptive statistics and do the null value condition check. Write an inference on it.**

Exploratory data analysis

1	Age	444 non-null	int64
2	Gender	444 non-null	object
3	Engineer	444 non-null	int64
4	MBA	444 non-null	int64
5	Work Exp	444 non-null	int64
6	Salary	444 non-null	float 64
7	Distance	444 non-null	float 64
8	License	444 non-null	int64
9	Transport	444 non-null	object

Descriptive statistics of data

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	444.0	NaN	NaN	NaN	27.747748	4.41671	18.0	25.0	27.0	30.0	43.0
Gender	444	2	Male	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Engineer	444.0	NaN	NaN	NaN	0.754505	0.430866	0.0	1.0	1.0	1.0	1.0
MBA	444.0	NaN	NaN	NaN	0.252252	0.434795	0.0	0.0	0.0	1.0	1.0
Work Exp	444.0	NaN	NaN	NaN	6.29955	5.112098	0.0	3.0	5.0	8.0	24.0
Salary	444.0	NaN	NaN	NaN	16.238739	10.453851	6.5	9.8	13.6	15.725	57.0
Distance	444.0	NaN	NaN	NaN	11.323198	3.606149	3.2	8.8	11.0	13.425	23.4
license	444.0	NaN	NaN	NaN	0.234234	0.423997	0.0	0.0	0.0	0.0	1.0
Transport	444	2	Public Transport	300	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 1.2

Five number summaries of columns

Five number summary of - Age
Minimum: 18.0
25%: 25.0
50% or Median: 27.0
75%: 30.0
Maximum: 43.0
IQR: 5.0

Five number summary of - Work Exp
Minimum: 0.0
25%: 3.0
50% or Median: 5.0
75%: 8.0
Maximum: 24.0
IQR: 5.0

Five number summary of - Salary
Minimum: 6.5
25%: 9.8
50% or Median: 13.6
75%: 15.725
Maximum: 57.0
IQR: 5.924999999999999

Five number summary of - Distance
Minimum: 3.2
25%: 8.8
50% or Median: 11.0
75%: 13.425
Maximum: 23.4
IQR: 4.625

```
Age          0
Gender       0
Engineer     0
MBA          0
Work Exp     0
Salary       0
Distance     0
license      0
Transport    0
dtype: int64
```

Table 1.3

From the above table it can be observed that there are no null values in the dataset.

1.2 **Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers (5 pts). Interpret the inferences for each (3 pts)**

Univariate analysis

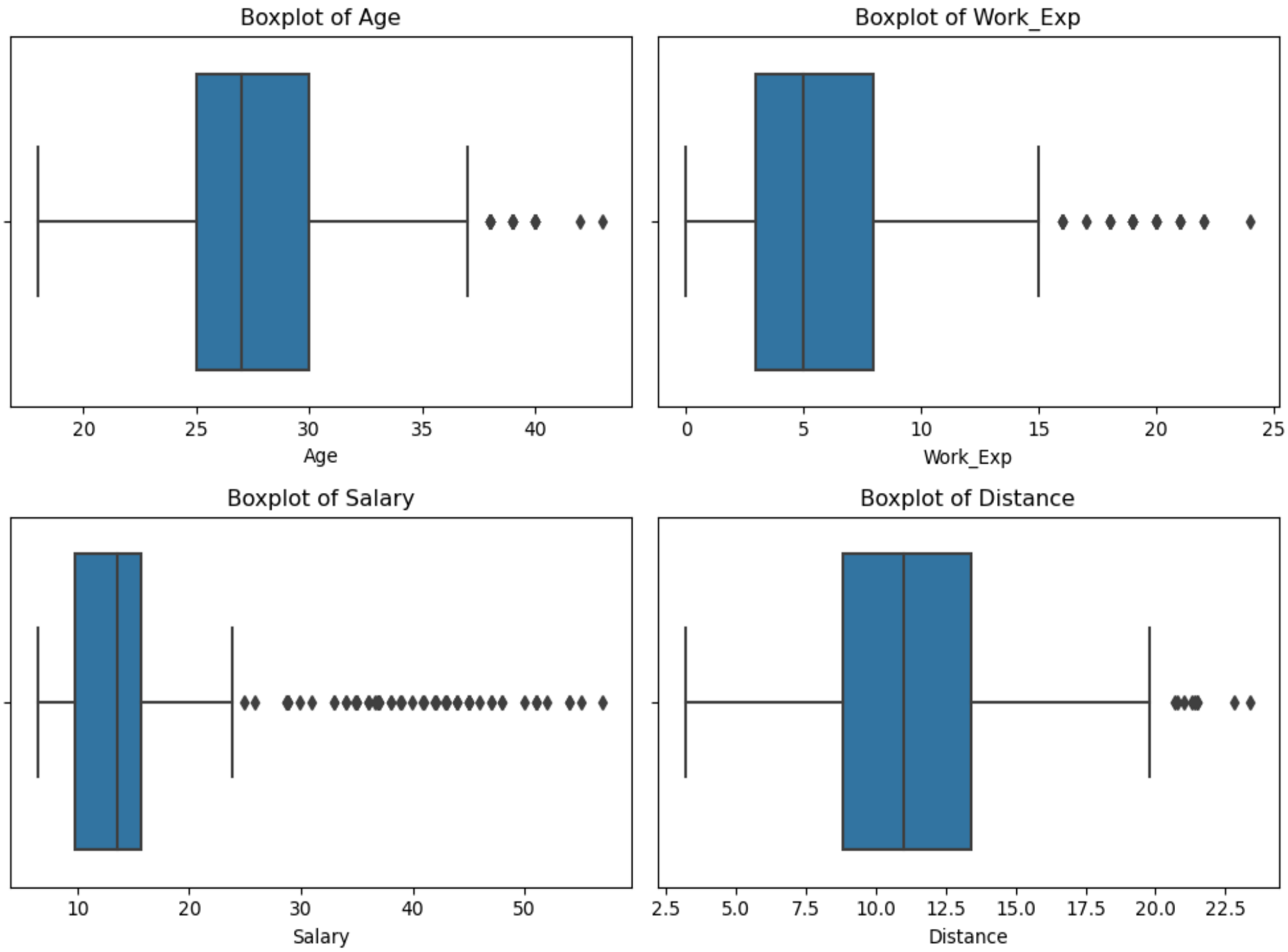


Fig 1.1

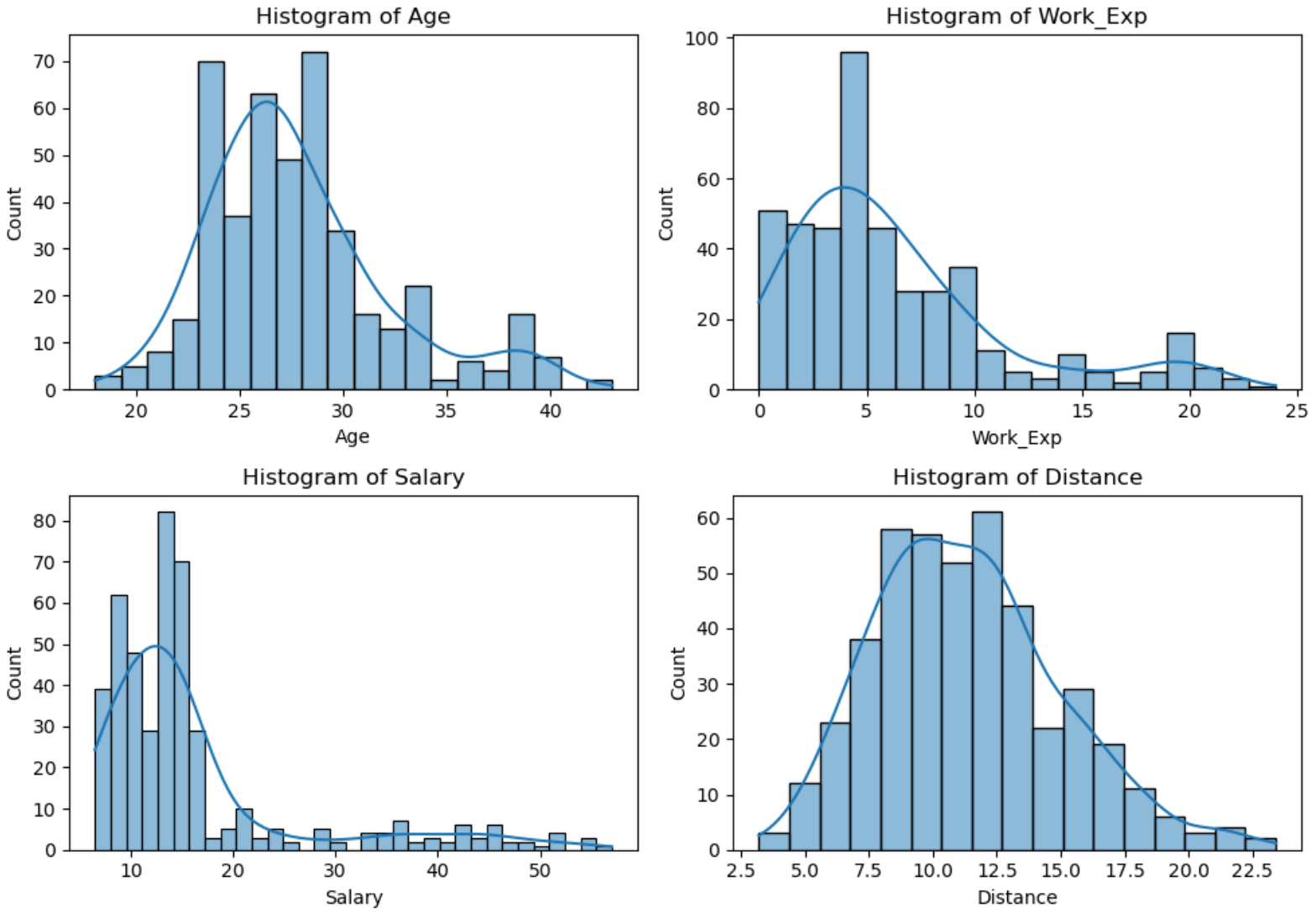


Fig 1.2

It can be observed from the box plot and histogram that there are ouliers in the dataset. Also the columns work experience and salary is right skewed.The other columns follow a normal distribution.

Bivariate analysis

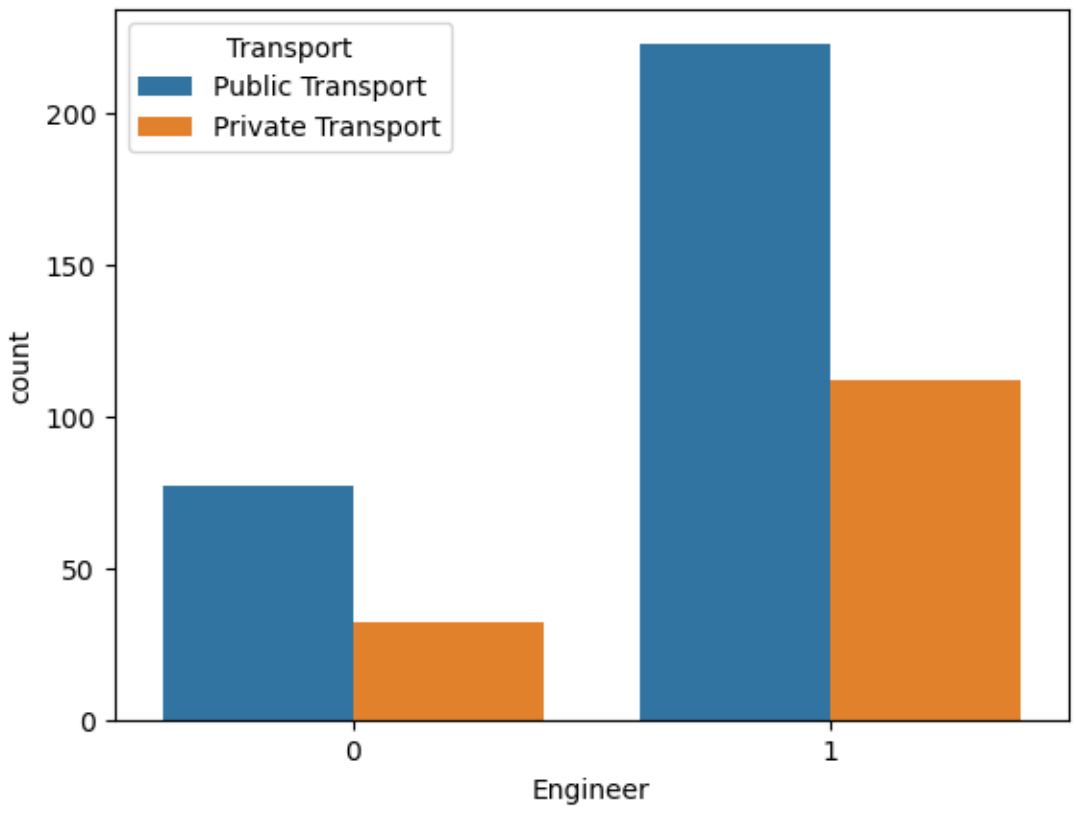


Fig 1.3

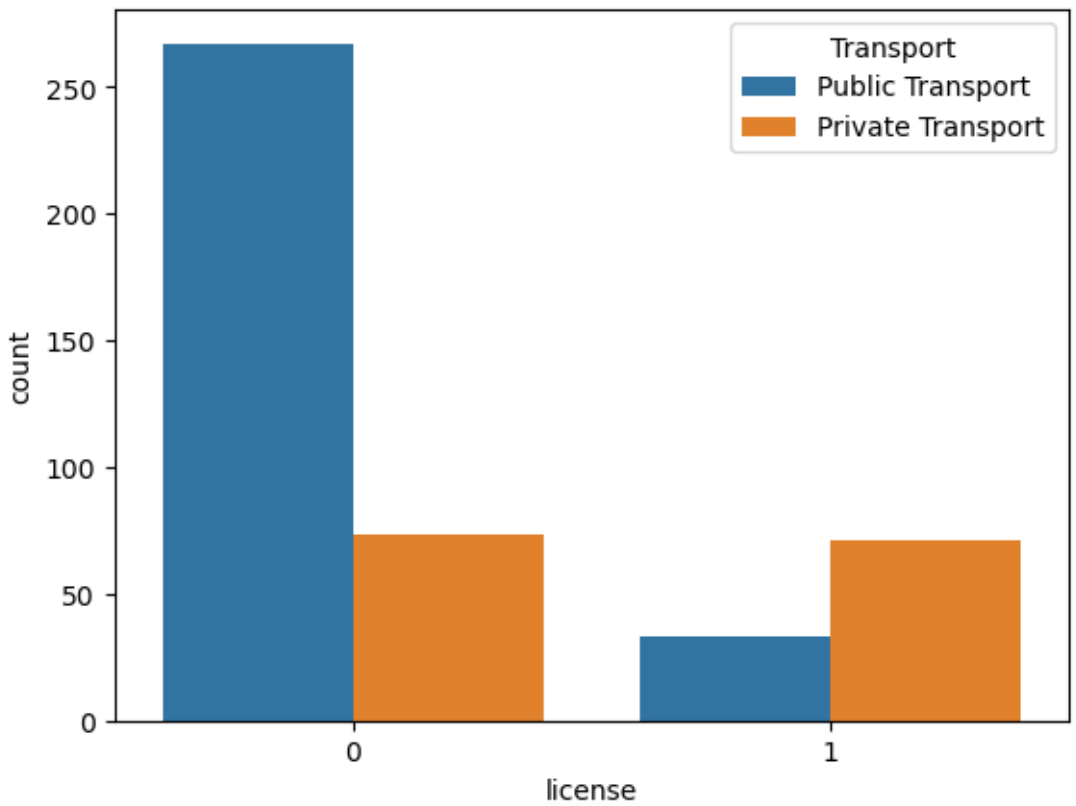


Fig 1.4

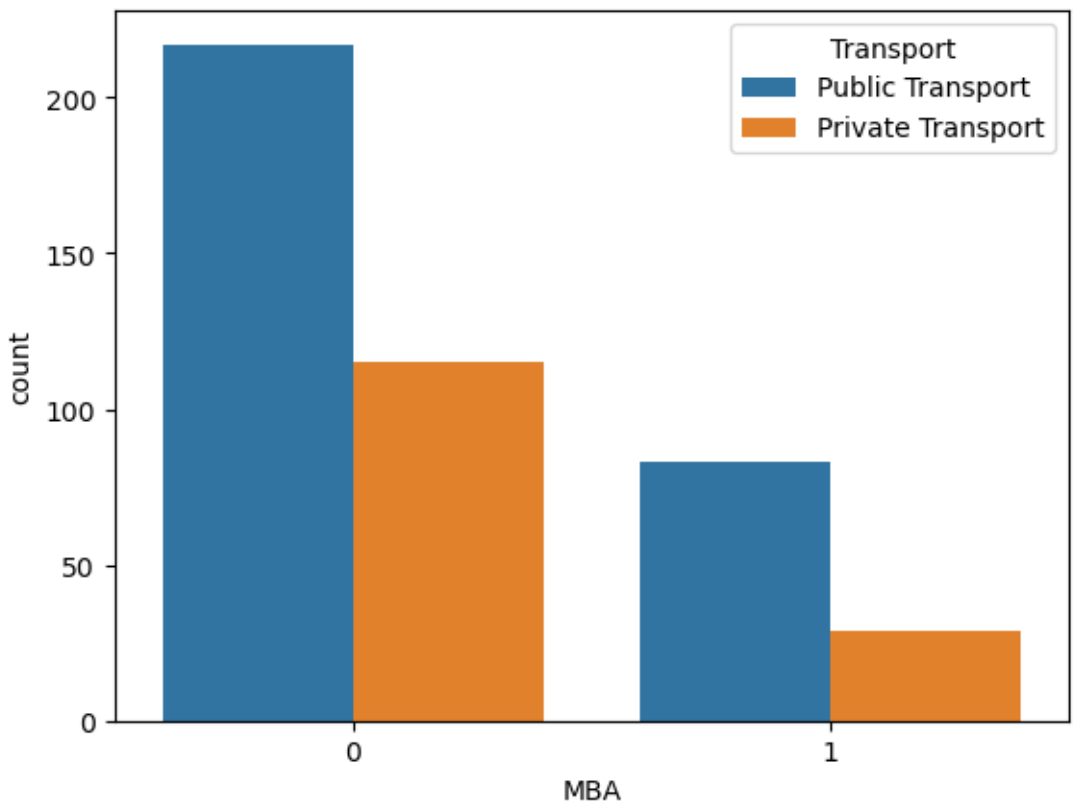


Fig 1.5

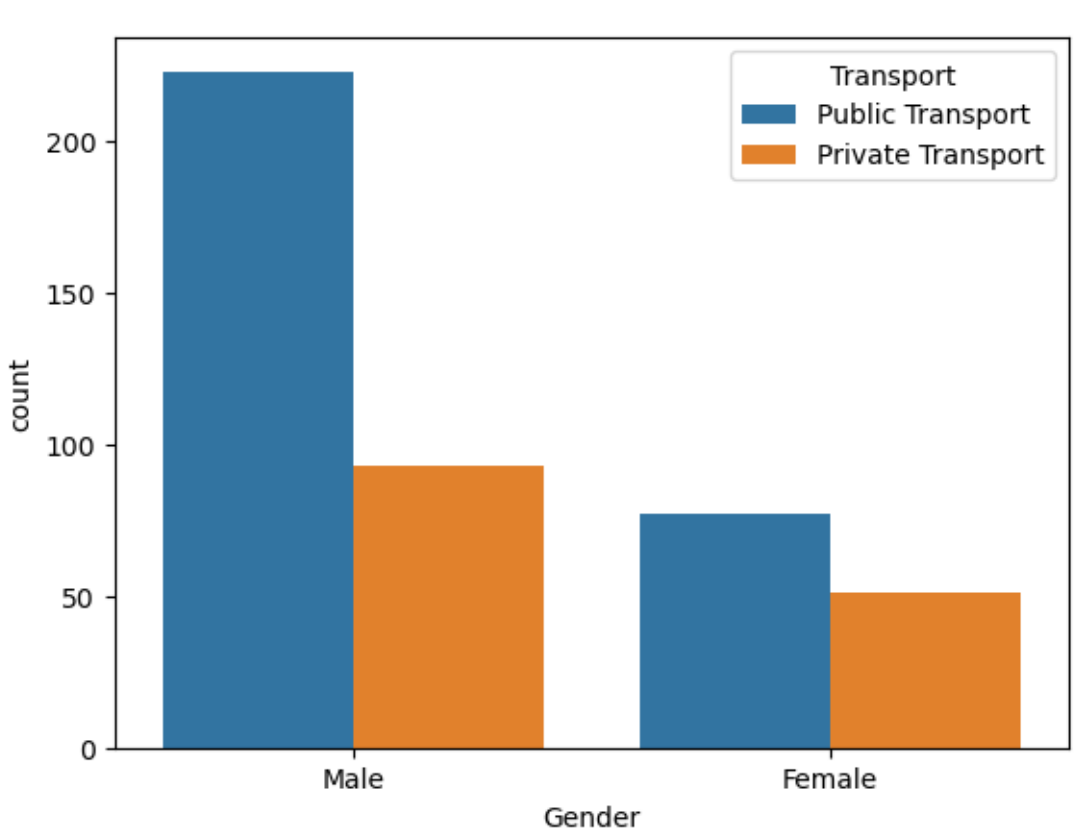


Fig 1.6


Age
Engineer
MBA
Work_Exp
Salary
Distance
license
Transport
Public Transport
Private Transport

Fig 1.7

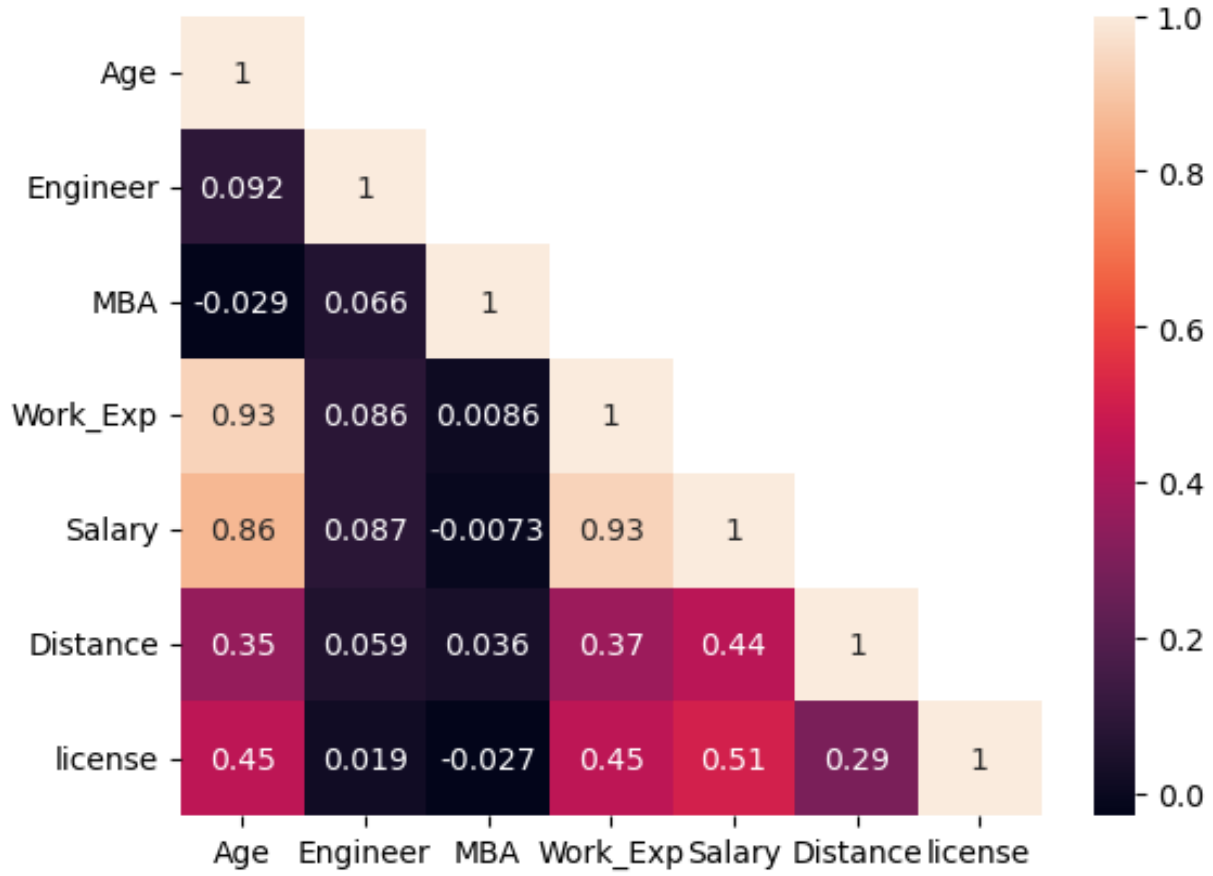


Fig 1.8

We will check the skewness of the columns provided.

```
Age          0.955276
Engineer    -1.186708
MBA          1.144763
Work_Exp     1.352840
Salary       2.044533
Distance     0.539851
license      1.259293
dtype: float64
```

Table 1.4

If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.

If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.

If the skewness is less than -1 or greater than 1, the data are highly skewed.

From that skewness table it can be observed that most of the columns are right skewed.

Outlier percentage in the data set.

	% OUTLIERS
Salary	13.2883
Work_Exp	8.5586
Age	5.6306
Distance	2.0270

Table 1.5

There are outliers in the dataset. We choose not to treat the outliers because an industry expert can only evaluate the relevance of the outlier values. Replacing the outlier values without prior knowledge in the industry either by capping or imputing with median can change the outcome of predictions..

1.3 **Encode the data (having string values) for Modelling(2 pts). Is Scaling necessary here or not?(2 pts), Data Split: Split the data into train and test (70:30) (2 pts).**

Encoded data set

	Age	Engineer	MBA	Work_Exp	Salary	Distance	license	Gender_Male	Transport_Public Transport
0	28	0	0	4	14.3	3.2	0	1	1
1	23	1	0	4	8.3	3.3	0	0	1
2	29	1	0	7	13.4	4.1	0	1	1
3	28	1	1	5	13.4	4.5	0	0	1
4	27	1	0	4	13.4	4.6	0	1	1

Table 1.6

Scaling can be necessary for distance-based algorithms depending on the specific algorithm and the data being used.

Distance-based algorithms, such as k-nearest neighbors (KNN), use the distance between data points to make predictions or group data points together. If the data being used has features with vastly different scales, the distances calculated between data points can be dominated by the features with the largest scales. This can lead to inaccurate results and bias in the algorithm.

Therefore, in some cases, scaling the data to have similar scales for all features can be necessary for distance-based algorithms to work effectively. Scaling can help ensure that each feature is equally weighted in the distance **calculation** and prevent the algorithm from being biased towards features with larger scales.

However, not all distance-based algorithms require scaling, and the decision to scale the data should be made on a case-by-case basis, depending on the algorithm and the data being used.

Here we will use scaled data for all the algorithms except Logistic regression.

Train-Test split

First five rows of the train data

	Age	Engineer	MBA	Work_Exp	Salary	Distance	license	Gender_Male
201	29	0	0	5	15.9	10.5	0	1
386	27	1	1	6	12.9	15.6	0	1
329	27	1	0	6	12.9	13.3	0	1
249	23	1	0	0	6.9	11.7	0	1
349	30	1	0	7	14.9	14.0	0	1

Table 1.7

First five rows of the test data

	Age	Engineer	MBA	Work_Exp	Salary	Distance	license	Gender_Male
247	26	1	0	8	14.6	11.6	0	0
179	27	0	1	5	13.9	10.0	0	1
186	35	1	0	16	28.7	10.2	0	0
31	24	1	1	2	8.6	6.4	0	1
218	33	1	0	11	16.7	10.9	1	1

Table 1.8

1.4 Apply Logistic Regression(4 pts). Interpret the inferences of both model s (2 pts)

LOGISTIC REGRESSION USING SCIKIT LEARN - BASE MODEL

The base model is created with all the default parameters for logistic regression.

Base model intercept = -2.17671219

Base model co-efficient = [-0.26711503, -0.28302822, 0.3672301 , -0.15968492, -0.06568729, -0.22325126, -1.88535722, 1.09452195**]**

VIF values of predictors:

```
VIF values:

Age            22.264997
Engineer        4.022232
MBA             1.384646
Work_Exp       20.027645
Salary         30.250094
Distance       12.785748
license         1.910177
Gender_Male     3.531363
dtype: float64
```

Table 1.9

Variance inflation factor measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables. Variance inflation factors allow a quick measure of how much a variable is contributing to the standard error in the regression. The default VIF cut-off value is 5, only variables with a VIF less than 5 will be included in the model.

But here we will not remove the columns having high VIF.

Some of the default parameters used in Logistic regression are:

“Penalty”: The default penalty used in scikit-learn logistic regression is L2 regularization, which helps prevent overfitting by adding a penalty to the magnitude of the coefficients. This penalty is controlled by the regularization parameter “C”, which is set to 1.0 by default.

“solver”: The default solver used in scikit-learn logistic regression is 'lbfgs'.

LOGISTIC REGRESSION USING SCIKIT LEARN – HYPER PARAM TUNED MODEL

model intercept = -0.40457013
model co-efficient = [0.19422328, -0.29567806, 0.33719647, -0.09015534, -0.22727713, -1.82835552, 1.07870214]

Hyper parameters used in Logistic regression are:

```
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'penalty':['l1', 'l2', 'elasticnet','none'],
    'solver' : ['lbfgs', 'liblinear', 'newton-cg', 'newton-cholesky', 'sag', 'saga'],
}
```

From the above parameters model chooses the best params. They are:
{'C': 1, 'penalty': 'l2', 'solver': 'liblinear'}

C=1 means that the regularization strength is moderate, and the model is penalized for large coefficient values, but not too strongly.
'l2' penalty is the default parameter.
For small datasets, ‘liblinear’ is a good choice.

The optimal model is found by drawing conclusions from the classification report and confusion matrix of the hyper-tuned model and base model. Question 1.7 and Question 1.8 will address this.

1.5 **Apply KNN Model(4 pts). Interpret the inferences of each model (2 pts)**

KNN USING SCIKIT LEARN – BASE MODEL

Knn is a distance-based algorithm. So we need to scale the data to make the values in same scale
Some of the default parameters for this algorithm are:

“n_neighbors**”** : 5 (default value): This parameter specifies the number of nearest neighbors to consider when making a prediction. The default value is 5.

“weights**”** : 'uniform' (default value): This parameter specifies the weight function used in prediction. 'uniform' means that all points in each neighborhood are weighted equally, while 'distance' means that closer neighbors have a greater influence on the prediction.

“algorithm**”** : 'auto' (default value): This parameter specifies the algorithm used to compute the nearest neighbors. 'auto' means that the algorithm will choose the most appropriate algorithm based on the values of the input data. Other possible values are 'ball_tree', 'kd_tree', and 'brute'.

“leaf_size**”** : 30 (default value): This parameter specifies the leaf size of the tree that will be used for querying the neighbors. The value of this parameter affects the speed and accuracy of the algorithm.

“p” : 2 (default value): This parameter specifies the distance metric used for the tree. The default value of 2 corresponds to the Euclidean distance, but other distance metrics such as Manhattan distance (p=1) can also be used.

KNN USING SCIKIT LEARN – HYPER PARAM TUNED MODEL

The following parameters are passed to find the best parameters.

```
param_grid_knn = {
    'n_neighbors': [3, 5, 7, 9],
    'weights': ['uniform', 'distance'],
    'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute'],
    'leaf_size': [10, 30, 50],
    'metric': ['euclidean','minkowski','mahalanobis','cityblock'],
    'n_jobs' : [-1],
    'p' : [1,2,3] }
```

From the above parameters model chooses the best params. They are:

{'algorithm': 'auto',
 'leaf_size': 10,
 'metric': 'minkowski',
 'n_jobs': -1,
 'n_neighbors': 9,
 'p': 1,
 'weights': 'distance'}

The optimal model is found by drawing conclusions from the classification report and confusion matrix of the hyper-tuned model and base model. Question 1.7 and Question 1.8 will address this.

1.6 **Bagging (4 pts) and Boosting (4 pts), Model Tuning (4 pts).**

BAGGING USING SCIKIT LEARN – BASE MODEL

Bagging is a machine learning technique used to improve the accuracy and stability of models by combining the predictions of multiple models trained on different subsets of the data.

Some of the default parameters for this algorithm are:

“base_estimator” : None (uses a decision tree by default)
“n_estimators” :10 (the number of base estimators to use)
“max_samples” :1.0 (the fraction of samples to use for training each base estimator)
“max_features” :1.0 (the fraction of features to use for training each base estimator)
“bootstrap” : True (whether to use bootstrapping for sampling the training data)
“bootstrap_features” **:** False (whether to use bootstrapping for sampling the features)

BAGGING USING SCIKIT LEARN – HYPER PARAM TUNED MODEL

The following parameters are passed to find the best parameters.

```
param_grid = {
    'n_estimators': [10, 50, 100],
    'max_features' :[0,1],
}
```

“n_estimators” : parameter determines the number of bootstrapped datasets that are created from the original dataset, and thus the number of base models that are trained.

“max_features” : This parameter determines the number of features to consider when splitting each node of the decision tree. The value ranges from 0 to 1.
From the above parameters model chooses the best params. They are:
{'max_features': 1, 'n_estimators': 50}

ADA BOOSTING USING SCIKIT LEARN – BASE MODEL

Ada boosting is a popular ensemble learning technique that combines multiple weak learners into a strong learner.
Some of the default parameters for this algorithm are:
“base_estimator” : The default base estimator for AdaBoost is usually a decision tree with maximum depth 1.
“n_estimators” : The default number of weak learners to use in AdaBoost is usually 50, which is often sufficient for achieving good performance.
“learning_rate” : The default learning rate for AdaBoost is usually 1.0, which means that the weights of the misclassified samples are updated by the full amount at each round of boosting. A smaller learning rate can be used to reduce the impact of each weak learner and prevent overfitting.

ADA BOOSTING USING SCIKIT LEARN – HYPER PARAM TUNED MODEL

The following parameters are passed to find the best parameters.
param_grid = {
'n_estimators': [50, 100, 200],
'learning_rate': [0.01, 0.1, 1.0] }
From the above parameters model chooses the best params. They are:
{'learning_rate': 0.1, 'n_estimators': 100}
The optimal model is found by drawing conclusions from the classification report and confusion matrix of the hyper-tuned model and base model. Question 1.7 and Question 1.8 will address this.

MODEL TUNING

Model tuning is applied in each algorithm from questions 1.4 ,1.5 and 1.6. Also the parameters used for model tuning is noted .
Here model overfitting or underfitting issues are not shown.

1.7 **Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model (5 pts) Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner. Describe on which model is best/optimized (3 pts)**

LOGISTIC REGRESSION

CLASSIFICATION REPORT FOR BASE MODEL TRAIN DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.77	0.54	0.64	102
1	0.80	0.92	0.85	208
ACCURACY			0.80	310
MACRO AVG	0.79	0.73	0.75	310
WEIGHTED AVG	0.79	0.80	0.79	310

Table 1.10

CLASSIFICATION REPORT FOR BASE MODEL TEST DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.71	0.57	0.63	42
1	0.82	0.89	0.85	92
ACCURACY			0.79	134
MACRO AVG	0.76	0.73	0.74	134
WEIGHTED AVG	0.78	0.79	0.78	134

Table 1.11

CONFUSION MATRIX FOR BASE MODEL TRAIN DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	55(TN)	47(FP)
1 (ACTUAL POSTIVE)	16(FN)	192(TP)

Table 1.12

CONFUSION MATRIX FOR BASE MODEL TEST DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	24(TN)	18(FP)
1 (ACTUAL POSTIVE)	10(FN)	82(TP)

Table 1.13

ROC_AUC_SCORE_TRAIN DATA : 0.834

ROC_AUC_SCORE_TEST DATA : 0.816

ROC_CURVE PLOT_TRAIN DATA :

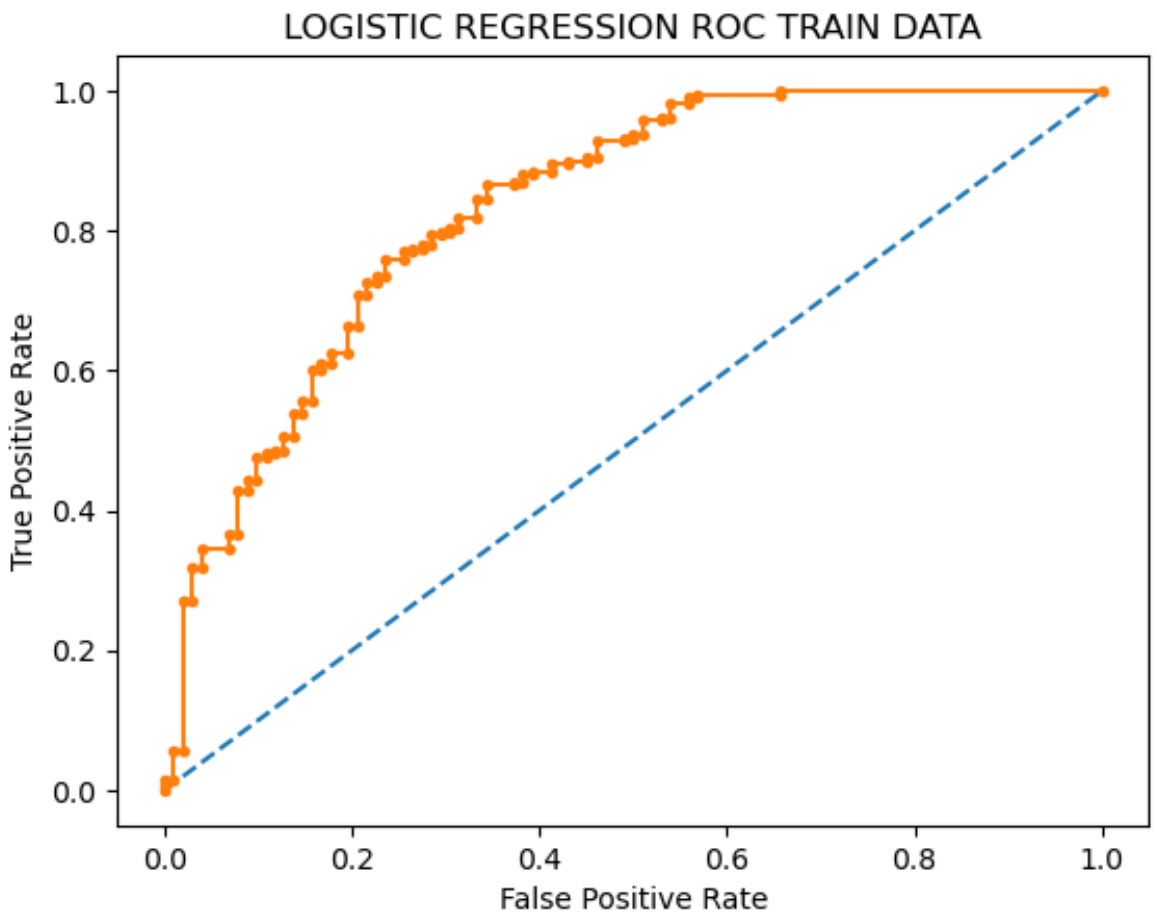


Fig 1.9

ROC_CURVE PLOT_TEST DATA :

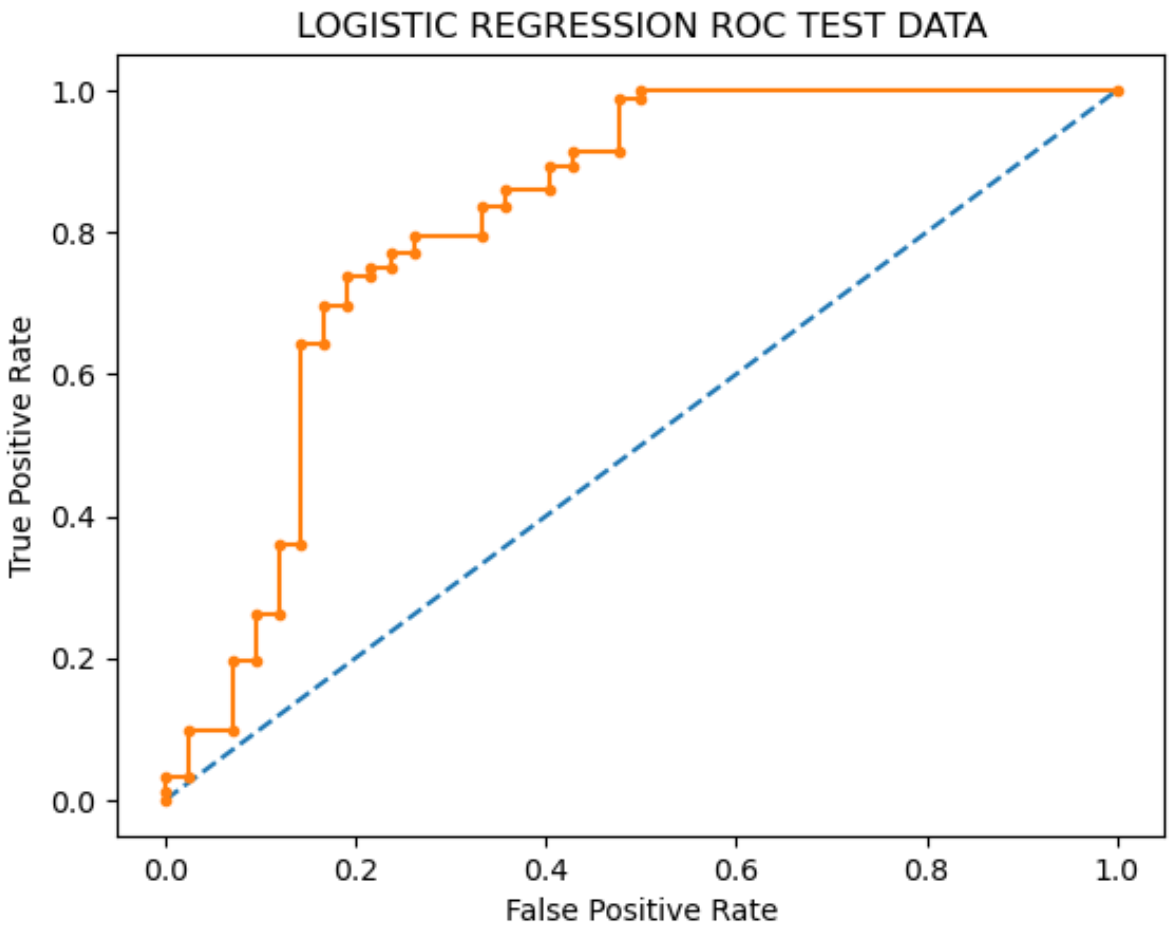


Fig 1.10

LOGISTIC REGRESSION -HYPER TUNED MODEL

CLASSIFICATION REPORT FOR HYPER TUNED MODEL TRAIN DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.76	0.52	0.62	102
1	0.80	0.92	0.85	208
ACCURACY			0.79	310
MACRO AVG	0.78	0.72	0.73	310
WEIGHTED AVG	0.78	0.79	0.77	310

Table 1.14

CLASSIFICATION REPORT FOR HYPER TUNED MODEL TEST DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.78	0.60	0.68	42
1	0.83	0.92	0.88	92
ACCURACY			0.82	134
MACRO AVG	0.81	0.76	0.78	134
WEIGHTED AVG	0.82	0.82	0.81	134

Table 1.15

CONFUSION MATRIX FOR HYPER TUNED MODEL TRAIN DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	53(TN)	49(FP)
1 (ACTUAL POSTIVE)	17(FN)	191(TP)

Table 1.16

CONFUSION MATRIX FOR HYPER TUNED MODEL TEST DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	25(TN)	17(FP)
1 (ACTUAL POSTIVE)	7(FN)	85(TP)

Table 1.17

ROC_AUC_SCORE_TRAIN DATA : 0.829

ROC_AUC_SCORE_TEST DATA : 0.820

ROC_CURVE PLOT_TRAIN DATA :

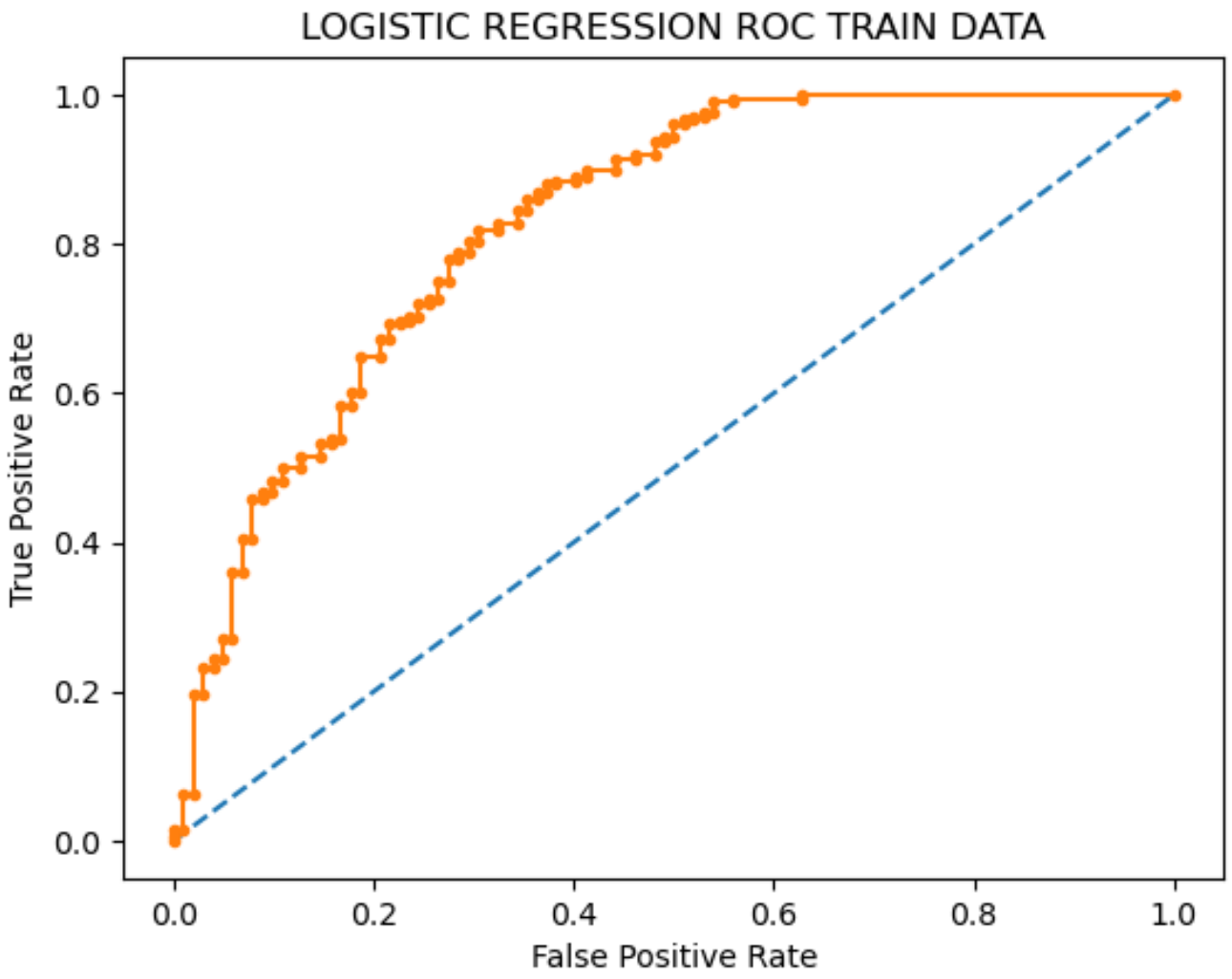


Fig 1.11

ROC_CURVE PLOT_TEST DATA :

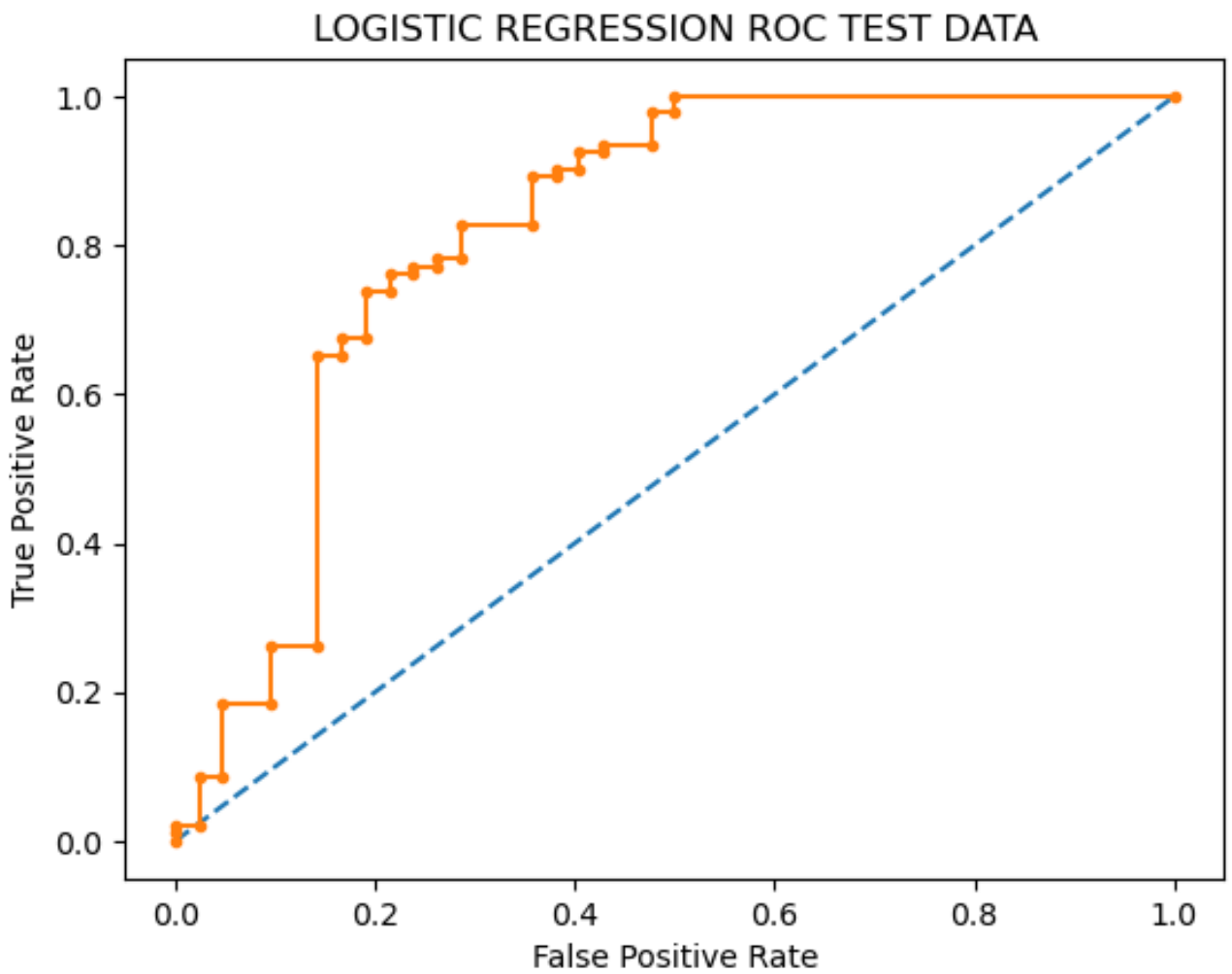


Fig 1.12

KNN(K-NEAREST NEIGHBOUR)

CLASSIFICATION REPORT FOR BASE MODEL TRAIN DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.85	0.67	0.75	102
1	0.85	0.94	0.89	208
ACCURACY			0.85	310
MACRO AVG	0.85	0.80	0.82	310
WEIGHTED AVG	0.85	0.85	0.85	310

Table 1.18

CLASSIFICATION REPORT FOR BASE MODEL TEST DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.63	0.52	0.57	42
1	0.80	0.86	0.83	92
ACCURACY			0.75	134
MACRO AVG	0.71	0.69	0.7	134
WEIGHTED AVG	0.74	0.75	0.785	134

Table 1.19

CONFUSION MATRIX FOR BASE MODEL TRAIN DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	68(TN)	34(FP)
1 (ACTUAL POSTIVE)	12(FN)	196(TP)

Table 1.20

CONFUSION MATRIX FOR BASE MODEL TEST DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	22(TN)	20(FP)
1 (ACTUAL POSTIVE)	13(FN)	79(TP)

Table 1.21

ROC_AUC_SCORE_TRAIN DATA : 0.929

ROC_AUC_SCORE_TEST DATA : 0.780

ROC_CURVE PLOT_TRAIN DATA :

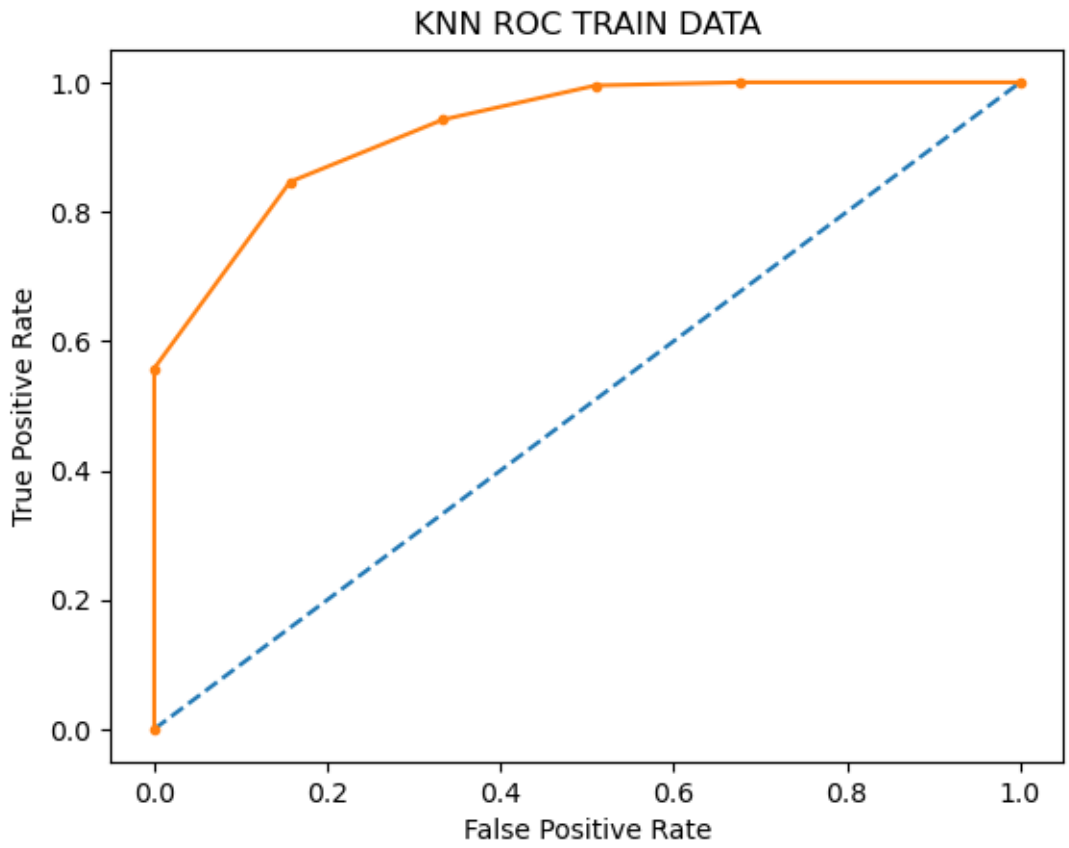


Fig 1.13

ROC_CURVE PLOT_TEST DATA :

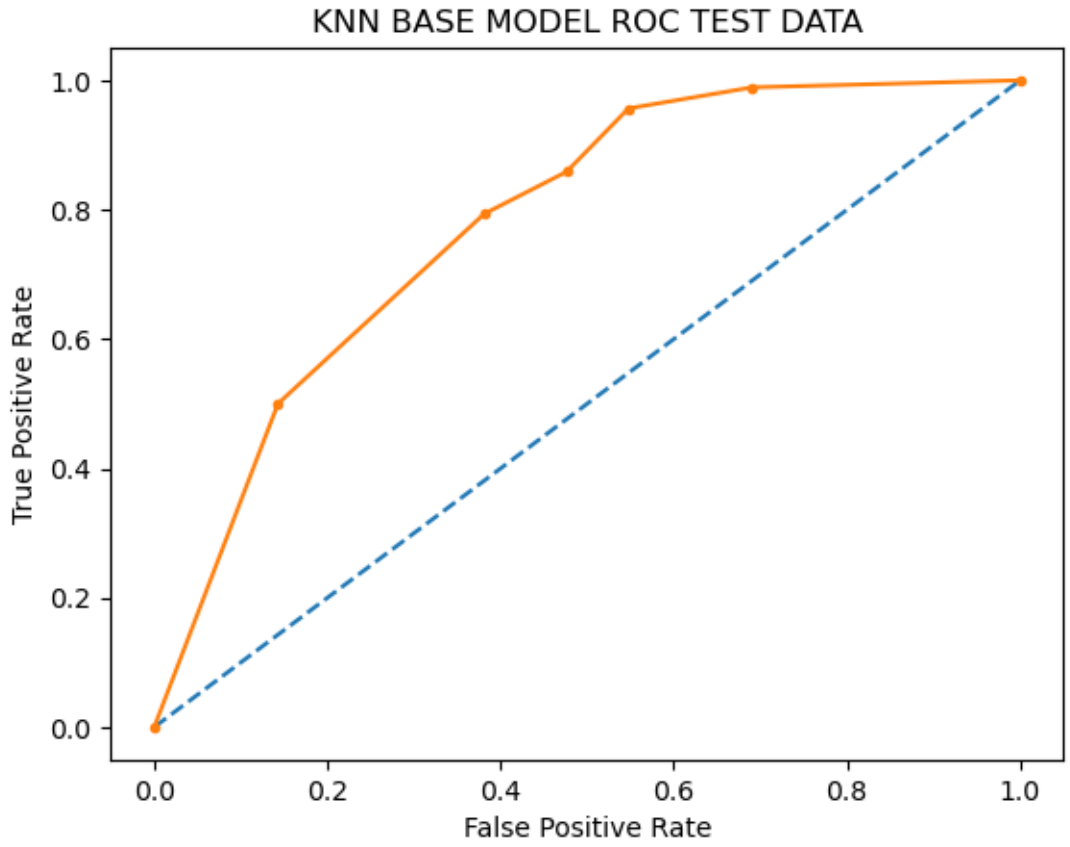


Fig 1.14

KNN -HYPER TUNED MODEL

CLASSIFICATION REPORT FOR HYPER TUNED MODEL TRAIN DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	1.00	1.00	1.00	102
1	1.00	1.00	1.00	208
ACCURACY			1.00	310
MACRO AVG	1.00	1.00	1.00	310
WEIGHTED AVG	1.00	1.00	1.00	310

Table 1.22

CLASSIFICATION REPORT FOR HYPER TUNED MODEL TEST DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.78	0.50	0.61	42
1	0.83	0.92	0.86	92
ACCURACY			0.80	134
MACRO AVG	0.79	0.72	0.74	134
WEIGHTED AVG	0.80	0.80	0.7	134

Table 1.23

CONFUSION MATRIX FOR HYPER TUNED MODEL TRAIN DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	102(TN)	0(FP)
1 (ACTUAL POSTIVE)	0(FN)	280(TP)

Table 1.24

CONFUSION MATRIX FOR HYPER TUNED MODEL TEST DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	21(TN)	21(FP)
1 (ACTUAL POSTIVE)	6(FN)	86(TP)

Table 1.25

ROC_AUC_SCORE_TRAIN DATA : 1.00

ROC_AUC_SCORE_TEST DATA : 0.792

ROC_CURVE PLOT_TRAIN DATA :

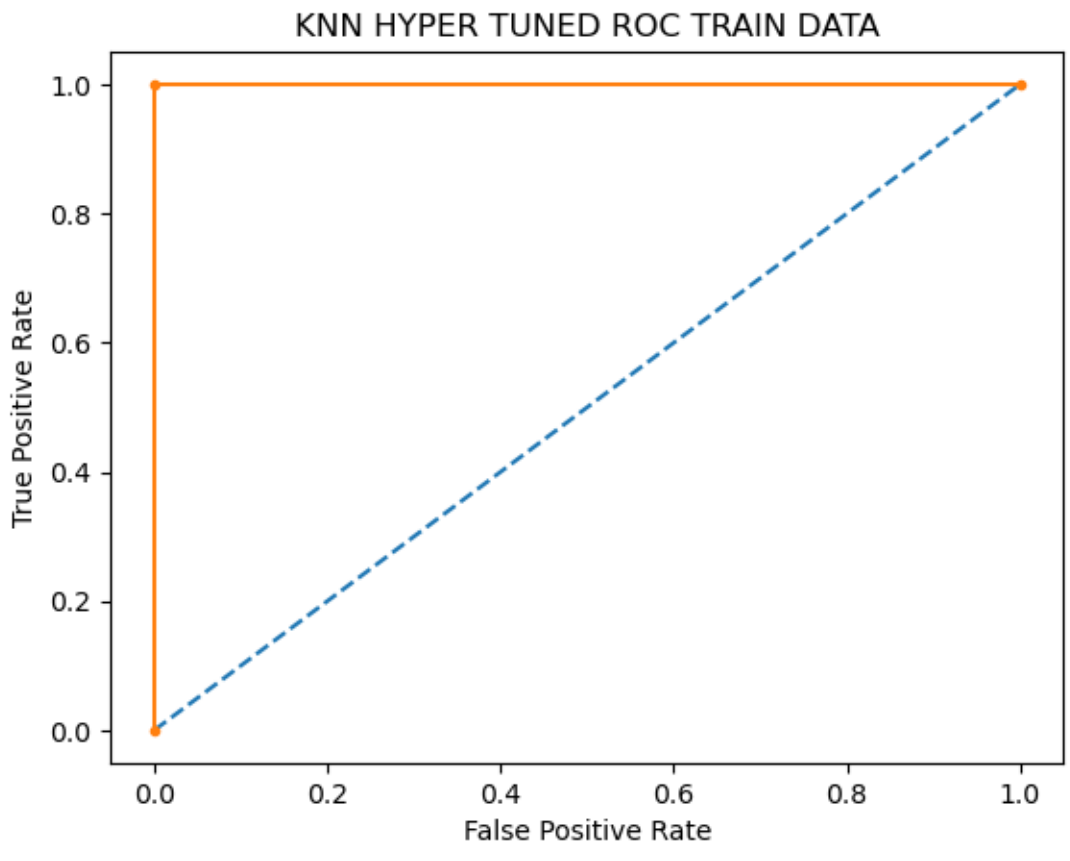


Fig 1.15

ROC_CURVE PLOT_TEST DATA :

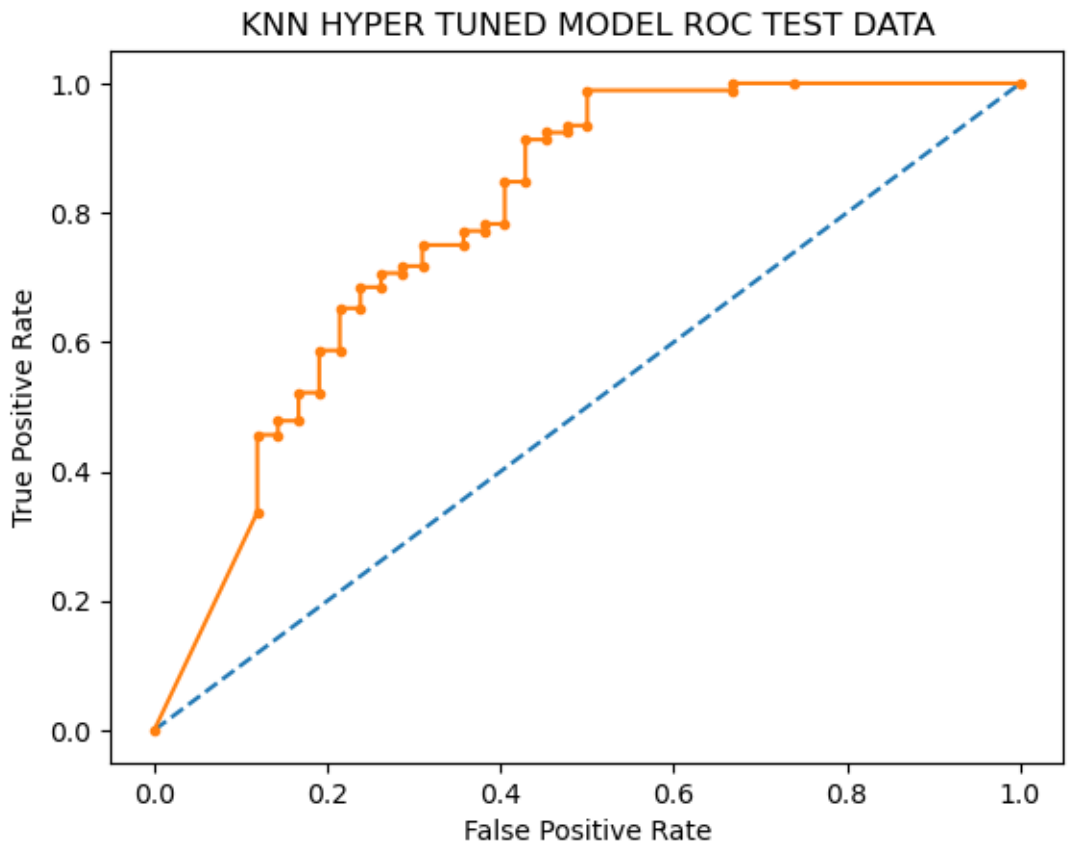


Fig 1.16

BAGGING

CLASSIFICATION REPORT FOR BASE MODEL TRAIN DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	1.00	0.90	0.95	102
1	0.95	1.00	0.98	208
ACCURACY			0.97	310
MACRO AVG	0.98	0.98	0.96	310
WEIGHTED AVG	0.97	0.97	0.97	310

Table 1.26

CLASSIFICATION REPORT FOR BASE MODEL TEST DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.70	0.55	0.61	42
1	0.81	0.89	0.85	92
ACCURACY			0.78	134
MACRO AVG	0.75	0.72	0.73	134
WEIGHTED AVG	0.78	0.78	0.78	134

Table 1.27

CONFUSION MATRIX FOR BASE MODEL TRAIN DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	92(TN)	10(FP)
1 (ACTUAL POSTIVE)	0(FN)	208(TP)

Table 1.28

CONFUSION MATRIX FOR BASE MODEL TEST DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	23(TN)	19(FP)
1 (ACTUAL POSTIVE)	10(FN)	82(TP)

Table 1.29

ROC_AUC_SCORE_TRAIN DATA : 0.999

ROC_AUC_SCORE_TEST DATA : 0.822

ROC_CURVE PLOT_TRAIN DATA :

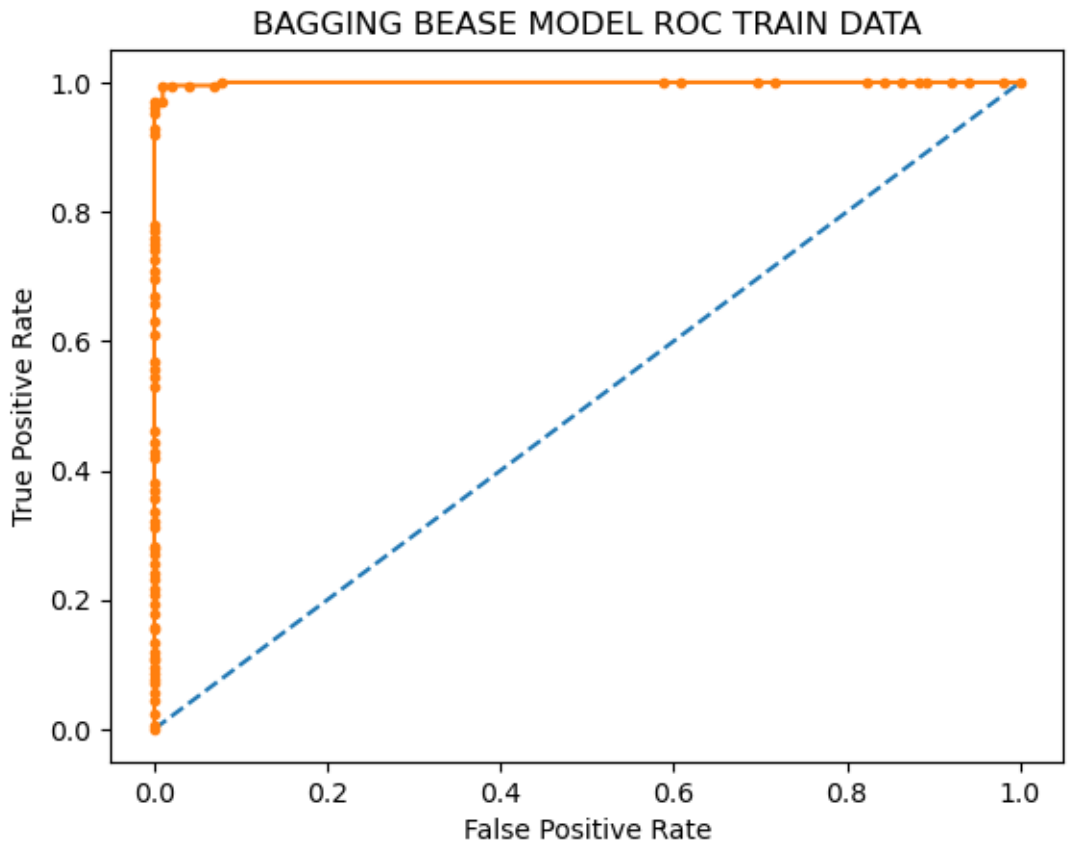


Fig 1.17

ROC_CURVE PLOT_TEST DATA :

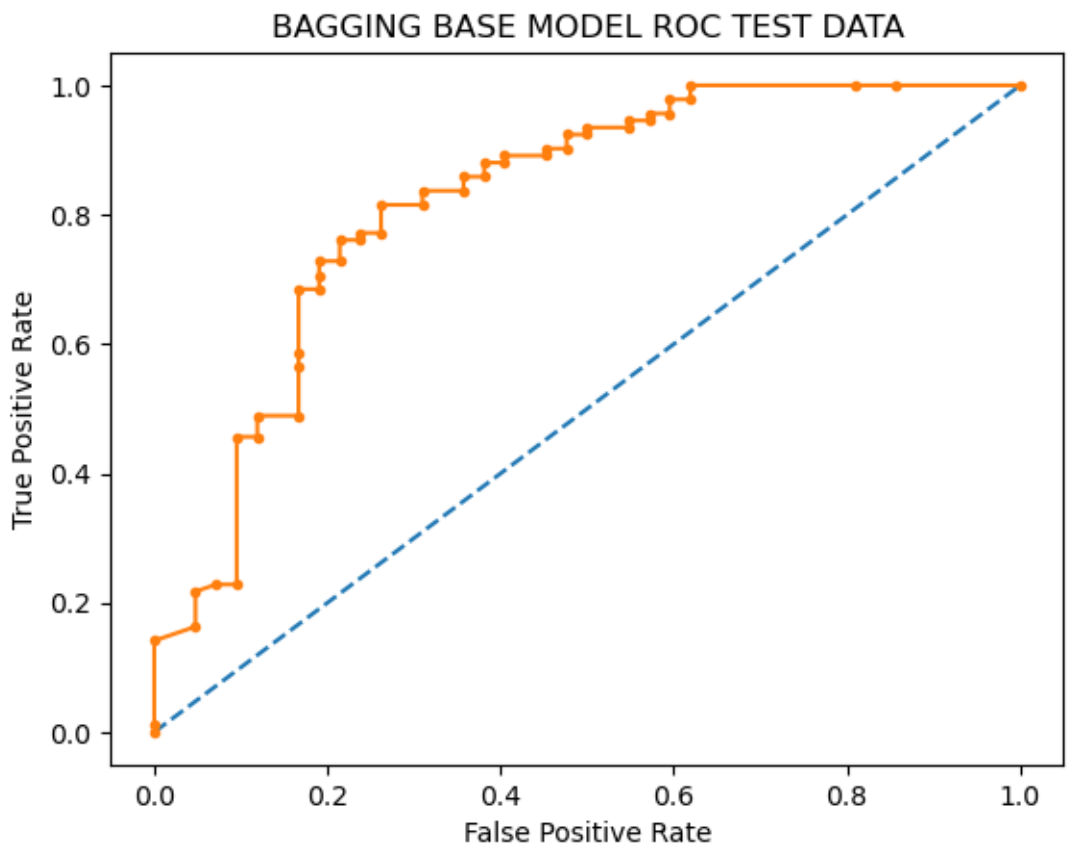


Fig 1.18

BAGGING -HYPER TUNED MODEL

CLASSIFICATION REPORT FOR HYPER TUNED MODEL TRAIN DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	1.00	0.46	0.63	102
1	0.79	1.00	0.88	208
ACCURACY			0.82	310
MACRO AVG	0.90	0.73	0.76	310
WEIGHTED AVG	0.86	0.82	0.80	310

Table 1.30

CLASSIFICATION REPORT FOR HYPER TUNED MODEL TEST DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	1.00	0.36	0.53	42
1	0.77	1.00	0.87	92
ACCURACY			0.80	134
MACRO AVG	0.89	0.68	0.70	134
WEIGHTED AVG	0.84	0.80	0.76	134

Table 1.31

CONFUSION MATRIX FOR HYPER TUNED MODEL TRAIN DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	47(TN)	55(FP)
1 (ACTUAL POSTIVE)	0(FN)	208(TP)

Table 1.32

CONFUSION MATRIX FOR HYPER TUNED MODEL TEST DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	15(TN)	27(FP)
1 (ACTUAL POSTIVE)	0(FN)	92(TP)

Table 1.33

ROC_AUC_SCORE_TRAIN DATA : 0.944

ROC_AUC_SCORE_TEST DATA : 0.752

ROC_CURVE PLOT_TRAIN DATA :

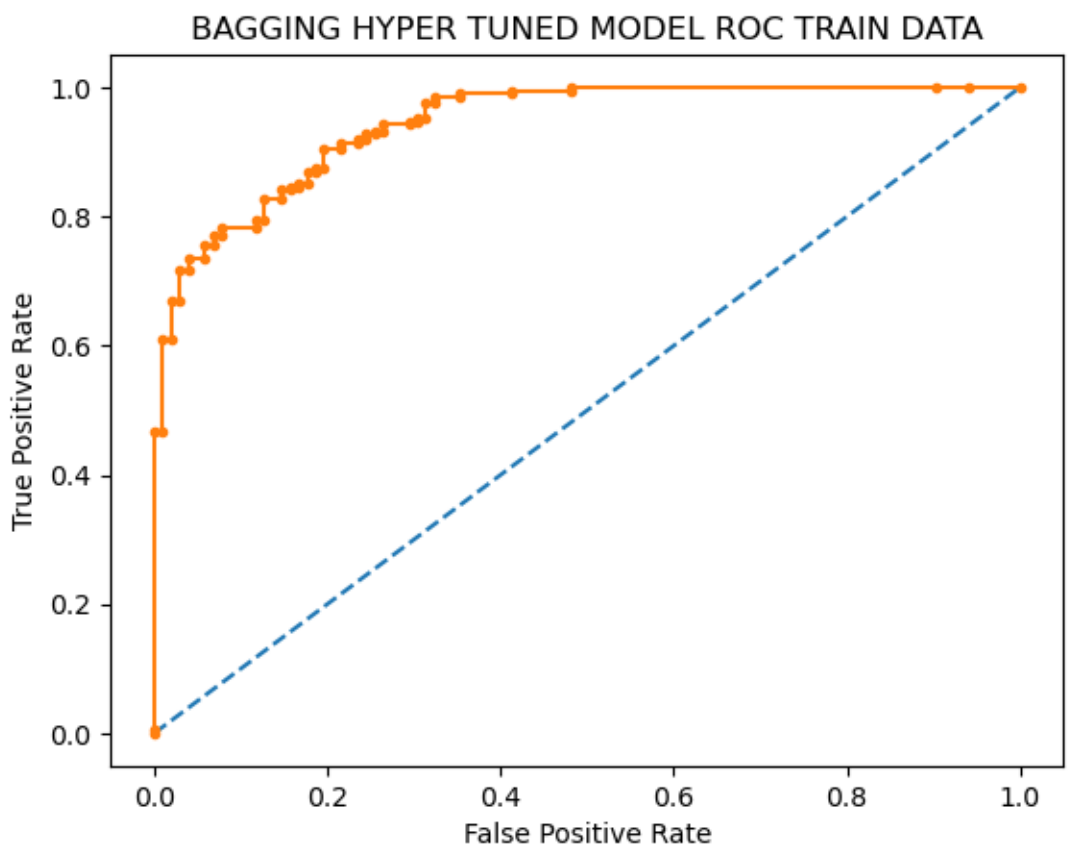
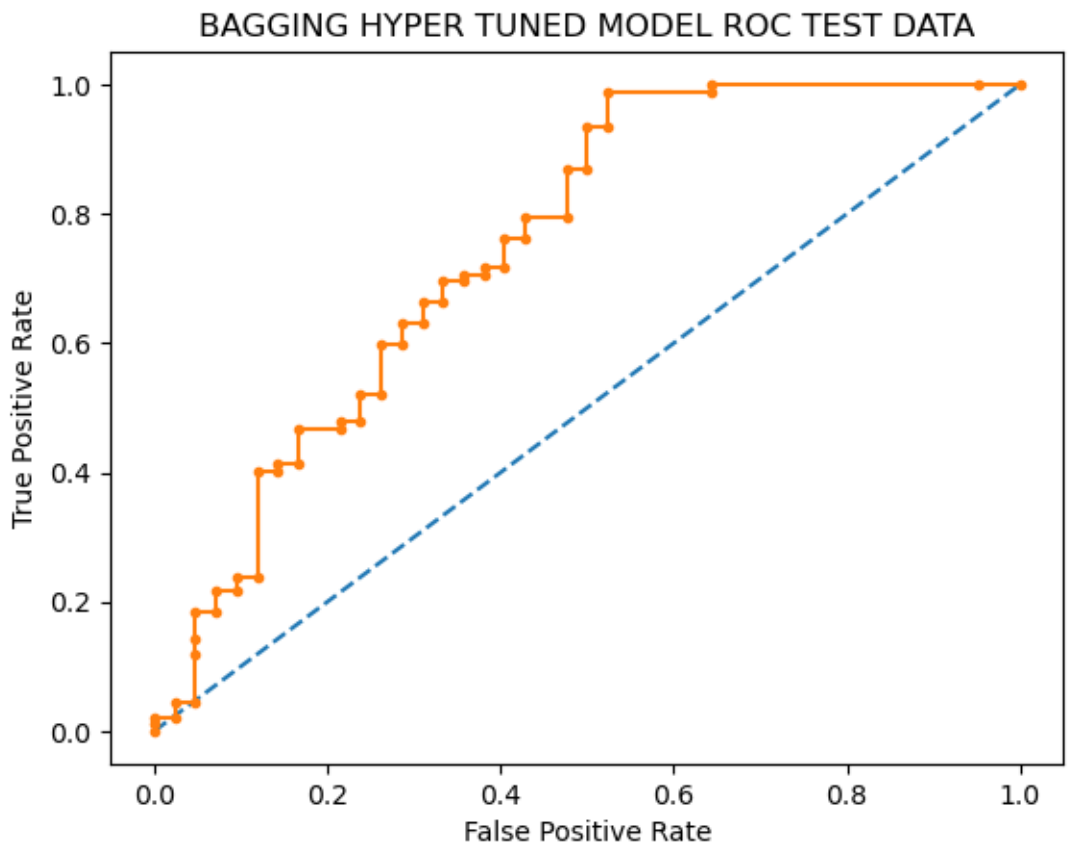


Fig 1.19

ROC_CURVE PLOT_TEST DATA :



-

Fig 1.20

ADA-BOOSTING

CLASSIFICATION REPORT FOR BASE MODEL TRAIN DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.86	0.75	0.80	102
1	0.88	0.94	0.91	208
ACCURACY			0.88	310
MACRO AVG	0.87	0.84	0.86	310
WEIGHTED AVG	0.88	0.88	0.87	310

Table 1.34

CLASSIFICATION REPORT FOR BASE MODEL TEST DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.70	0.67	0.68	42
1	0.85	0.87	0.86	92
ACCURACY			0.81	134
MACRO AVG	0.78	0.77	0.77	134
WEIGHTED AVG	0.80	0.81	0.80	134

Table 1.35

CONFUSION MATRIX FOR BASE MODEL TRAIN DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	76(TN)	26(FP)
1 (ACTUAL POSTIVE)	12(FN)	196(TP)

Table 1.36

CONFUSION MATRIX FOR BASE MODEL TEST DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	28(TN)	14(FP)
1 (ACTUAL POSTIVE)	12(FN)	80(TP)

Table 1.37

ROC_AUC_SCORE_TRAIN DATA : 0.940

ROC_AUC_SCORE_TEST DATA : 0.809

ROC_CURVE PLOT_TRAIN DATA :

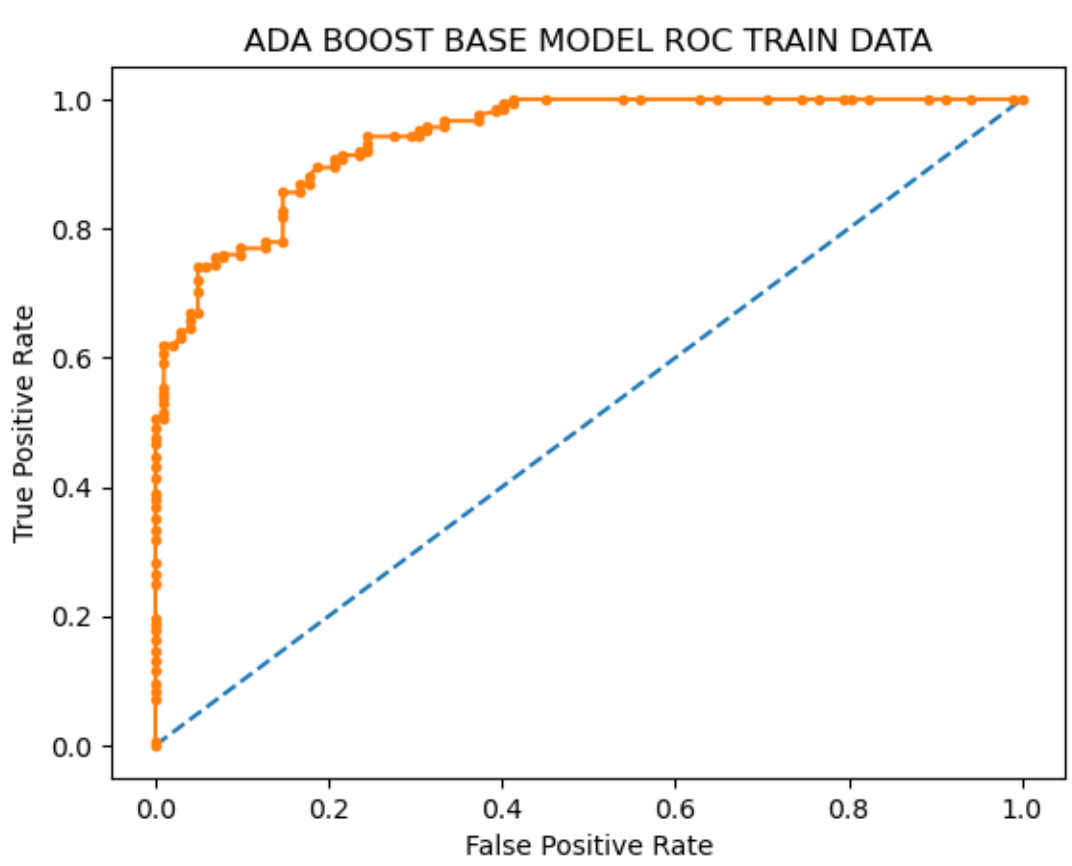


Fig 1.21

ROC_CURVE PLOT_TEST DATA :

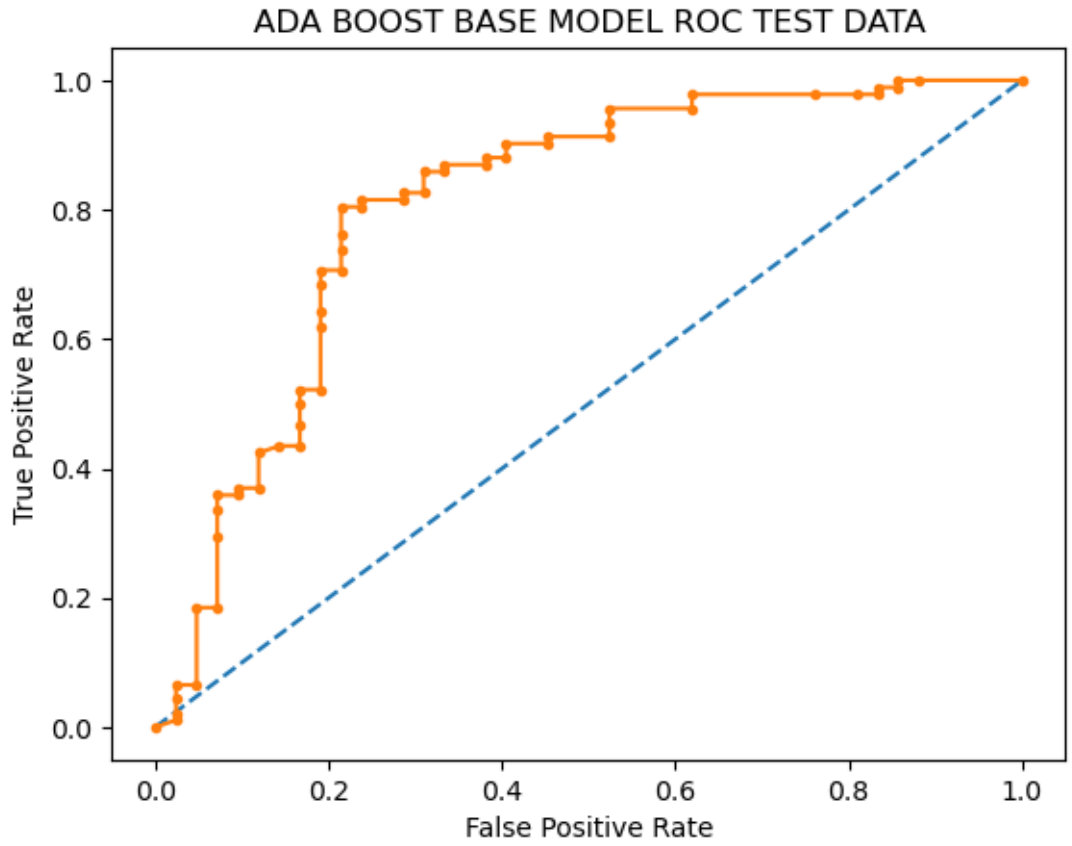


Fig 1.22

ADA BOOSTING -HYPER TUNED MODEL

CLASSIFICATION REPORT FOR HYPER TUNED MODEL TRAIN DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.88	0.57	0.69	102
1	0.82	0.96	0.88	208
ACCURACY			0.83	310
MACRO AVG	0.85	0.77	0.79	310
WEIGHTED AVG	0.84	0.83	0.82	310

Table 1.38

CLASSIFICATION REPORT FOR HYPER TUNED MODEL TEST DATA

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.85	0.55	0.67	42
1	0.82	0.96	0.88	92
ACCURACY			0.83	134
MACRO AVG	0.84	0.75	0.78	134
WEIGHTED AVG	0.83	0.83	0.82	134

Table 1.39

CONFUSION MATRIX FOR HYPER TUNED MODEL TRAIN DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	58(TN)	44(FP)
1 (ACTUAL POSTIVE)	8(FN)	200(TP)

Table 1.40

CONFUSION MATRIX FOR HYPER TUNED MODEL TEST DATA

	PREDICTION	
ACTUAL	**0** (PREDICTED NEGATIVE)	**1** (PREDICTED POSITIVE)
0 (ACTUAL NEGATIVE)	23(TN)	19(FP)
1 (ACTUAL POSTIVE)	4(FN)	88(TP)

Table 1.41

ROC_AUC_SCORE_TRAIN DATA : 0.904

ROC_AUC_SCORE_TEST DATA : 0.825

ROC_CURVE PLOT_TRAIN DATA :

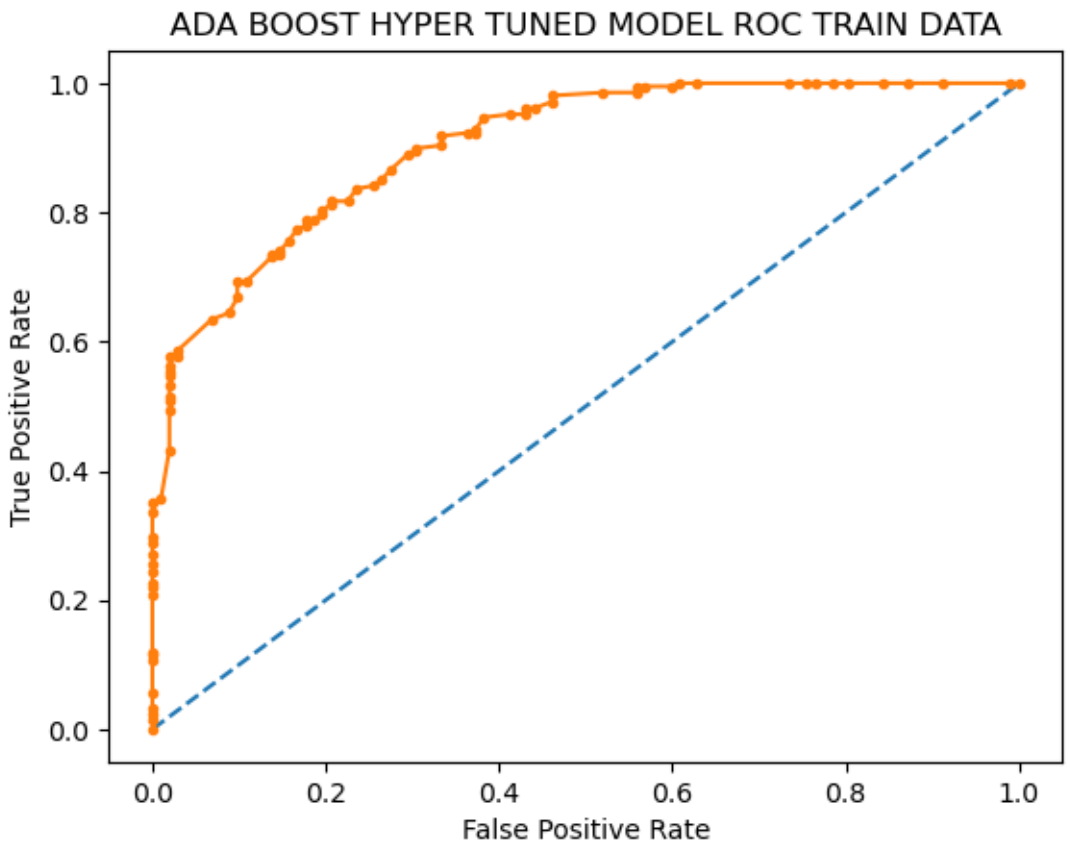


Fig 1.23

ROC_CURVE PLOT_TEST DATA :

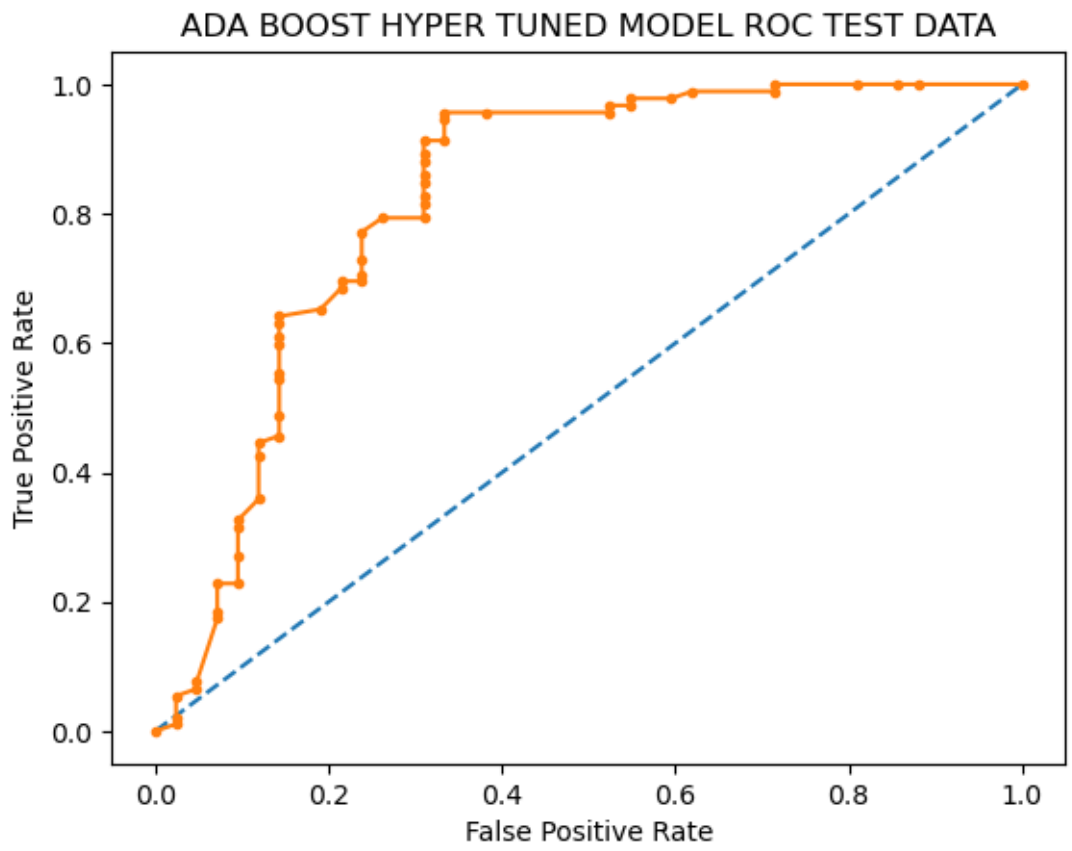


Fig 1.24

We have created different models using Logistic regression,Knn,bagging and boosting techniques. Based on the classification reports provided, it seems like the AdaBoosting model performs the best out of the four models in terms of accuracy and F1-score on the test data, followed by the Logistic Regression model and then the Bagging model. However, it's worth noting that the hyper-tuned Bagging model has a perfect precision score for class 0, but a low recall score, which may indicate overfitting on the training data. It's important to consider not only accuracy, but also precision, recall, and F1-score for each class when evaluating a classification model.It can also be observed that hyper tuned knn model has overfitted the data.Except for ada boost model all models shows the tendency to overfit.This might be due to the lack of data. Increasing the amount of training data can be a solution for the performace values obtained from different models.

1.8 **Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective.**

444 rows and 9 columns make up the dataset. The goal is to foresee the employees' preferred method of transportation from their place of employment to their homes. The impact of each column on the employee's chosen mode of transportation will be assessed.

Age : There are several reasons why age can influence the choice of transport mode:
1.Income: Younger employees may have lower incomes and may not be able to afford private transport, while older employees may have higher incomes and can afford to buy and maintain their own vehicles.
2.Accessibility: Older employees may have reduced mobility and may require private transport options to accommodate their needs.

Gender : Gender can also be a factor that influences the choice of transport mode for employees. While gender may not be as significant a factor as other variables such as income or distance from home to workplace, it can still play a role in determining transport mode preferences.
Some of the reasons why gender may influence transport mode choices include:
1.Safety concerns: Women may feel more vulnerable when using public transport at night, especially in areas with higher crime rates. As a result, they may be more likely to prefer private transport options.
2.Commute distance: Women may be more likely to choose public transport options for shorter commutes, while men may be more willing to drive longer distances to work.

Engineer & MBA : We have separate columns for employees who have MBA or the employee is a engineer . The preferred mode of transportation for an employee who is an engineer or has an MBA may depend on various factors such as personal preference, distance of commute etc.

Work experience: If the employee has a lot of work experience and is in a more senior position, they may have more flexibility and autonomy in choosing their preferred mode of transportation. They may prioritize comfort, convenience, and time efficiency, and may prefer to use private transportation such as a car or a taxi. This can provide more control over their schedule and route and may allow them to work on the go.

Salary: Salary can have a significant impact on an employee's choice of transportation mode. Generally, employees with higher salaries may be more likely to use private transportation options such as cars or taxis, while those with lower salaries may be more likely to use public transportation. This is because private transportation options tend to be more expensive, and employees with higher salaries may have more disposable income to spend on transportation. They may also prioritize convenience, comfort, and flexibility over cost savings, and may prefer to use private transportation options such as a car or a taxi.

On the other hand, employees with lower salaries may have limited financial resources and may prioritize cost savings over other factors such as comfort or convenience. They may be more likely to use public transportation such as buses, trains, or subways, which tend to be more affordable and cost-effective.

Distance: Distance of commute can also have a significant impact on an employee's choice of transportation mode. Generally, employees who live closer to their workplace may be more likely to walk, bike, or use public transportation, while those who live further away may be more likely to use private transportation such as cars or taxis.

License: Having a valid driver's license can also have an impact on an employee's preferred mode of transportation. Employees who have a driver's license may be more likely to use private transportation such as cars or motorcycles, while those who do not have a license may be more likely to use public transportation or other alternative modes of transportation.

Recommendations :

1. Conduct a survey regarding transportation needs and preferences of the employees. The survey should focus on the primary reasons why employees choose a particular mode of transportation, such as convenience, cost-effectiveness, comfort, reliability, or environmental impact. The survey can also include questions about the distance of commute, license status, salary, age, work experience, and other factors that may influence transportation choices.
2. Provide a range of transportation options based on the analysis, provide a range of transportation options that cater to the needs of all employees, regardless of their preferences and needs.

3. Consider the cost-effectiveness of each transportation option, both for the employees and for the company. While private transportation options may be preferred by some employees, they can also be more expensive and may not be feasible for all employees. Therefore, it is important to consider cost-effective options that provide the best value for both the employees and the company.
4. Monitor and evaluate the transportation services provided by the company on an ongoing basis to ensure that they are meeting the needs of the employees and are aligned with the overall business objectives of the company. This can help to identify any areas for improvement and to continuously improve the transportation services provided to the employees.
5. Offering long-distance commuting workers transportation options. The company's transportation services can reduce their expenditures since long-distance commuters often use private vehicles.