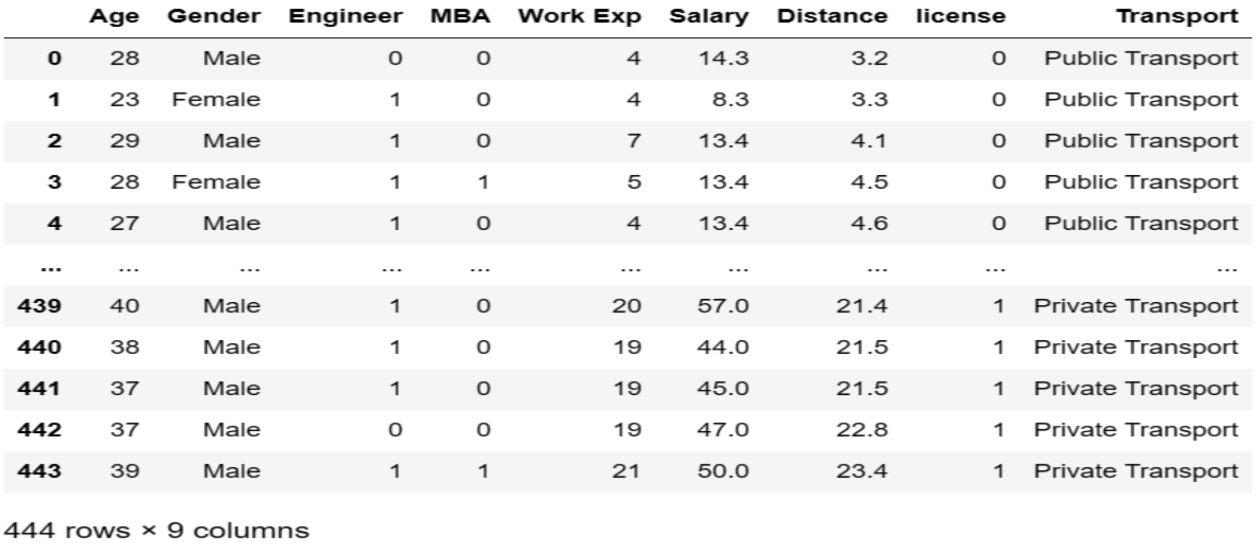
1.1 Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.

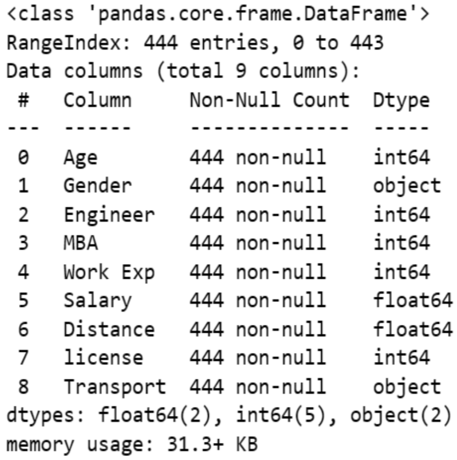
In this Dataset has 444 rows and 9 columns in transport dataset.



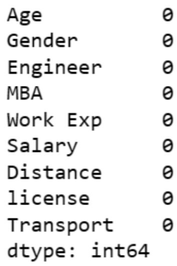
	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28	Male	0	0	4	14.3	3.2	0	Public Transport
1	23	Female	1	0	4	8.3	3.3	0	Public Transport
2	29	Male	1	0	7	13.4	4.1	0	Public Transport
3	28	Female	1	1	5	13.4	4.5	0	Public Transport
4	27	Male	1	0	4	13.4	4.6	0	Public Transport
...	...	...	...	...	...	...	...	...	...
439	40	Male	1	0	20	57.0	21.4	1	Private Transport
440	38	Male	1	0	19	44.0	21.5	1	Private Transport
441	37	Male	1	0	19	45.0	21.5	1	Private Transport
442	37	Male	0	0	19	47.0	22.8	1	Private Transport
443	39	Male	1	1	21	50.0	23.4	1	Private Transport

444 rows × 9 columns

In the dataset have 5 integers, 2 float and 2 objects in transportation data and there is no null in this data.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 444 entries, 0 to 443
Data columns (total 9 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Age        444 non-null    int64  
 1   Gender     444 non-null    object 
 2   Engineer   444 non-null    int64  
 3   MBA        444 non-null    int64  
 4   Work Exp   444 non-null    int64  
 5   Salary     444 non-null    float64
 6   Distance   444 non-null    float64
 7   license    444 non-null    int64  
 8   Transport  444 non-null    object 
dtypes: float64(2), int64(5), object(2)
memory usage: 31.3+ KB
```



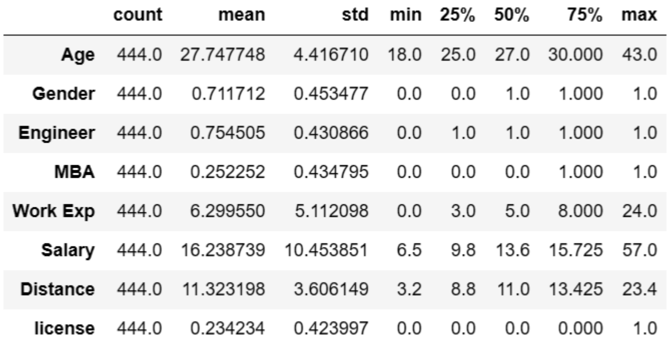
```
Age          0
Gender       0
Engineer     0
MBA          0
Work Exp     0
Salary       0
Distance     0
license      0
Transport    0
dtype: int64
```

Summary Statistics:

The minimum and maximum age of employee is 18 to 43 in our Transport data.

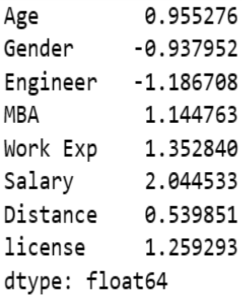
As per the summary statistics, the standard deviation of **age(4.416)** has **less variation** and **low dispersion** with respect to the **mean (27.747).**
The spread of the employees age are highly clustered that means the employee categories are young to middle age persons and the distance has less variation and low dispersion.

We can see that **work experience** column has **more variation** and **high dispersion** so the most of the experienced person are unique from one another in **ABC Company**.



	count	mean	std	min	25%	50%	75%	max
Age	444.0	27.747748	4.416710	18.0	25.0	27.0	30.000	43.0
Gender	444.0	0.711712	0.453477	0.0	0.0	1.0	1.000	1.0
Engineer	444.0	0.754505	0.430866	0.0	1.0	1.0	1.000	1.0
MBA	444.0	0.252252	0.434795	0.0	0.0	0.0	1.000	1.0
Work Exp	444.0	6.299550	5.112098	0.0	3.0	5.0	8.000	24.0
Salary	444.0	16.238739	10.453851	6.5	9.8	13.6	15.725	57.0
Distance	444.0	11.323198	3.606149	3.2	8.8	11.0	13.425	23.4
license	444.0	0.234234	0.423997	0.0	0.0	0.0	0.000	1.0

Univariate Analysis:

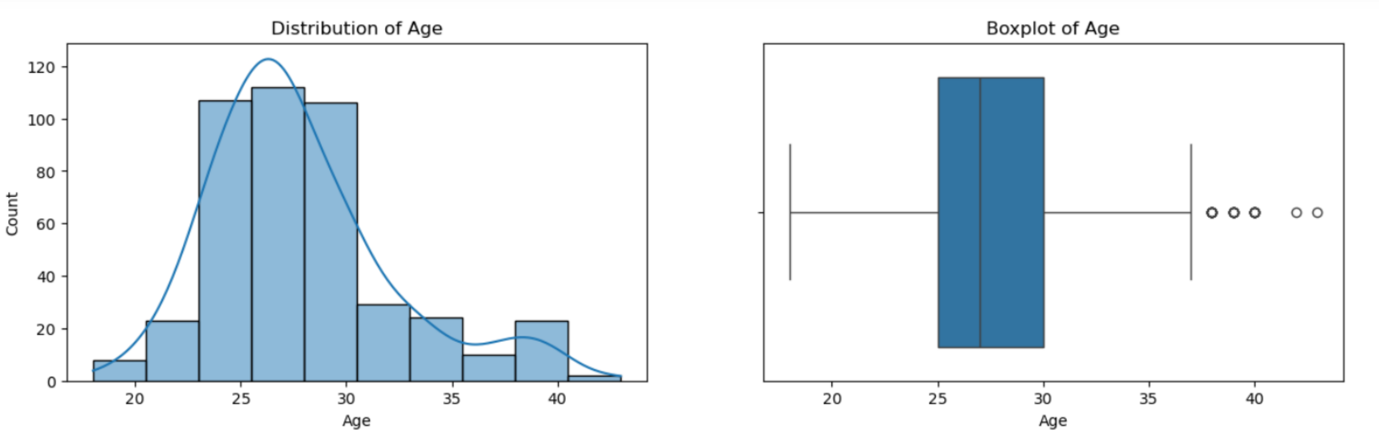


```
Age         0.955276
Gender     -0.937952
Engineer   -1.186708
MBA         1.144763
Work Exp    1.352840
Salary      2.044533
Distance    0.539851
license     1.259293
dtype: float64
```

Age Column:

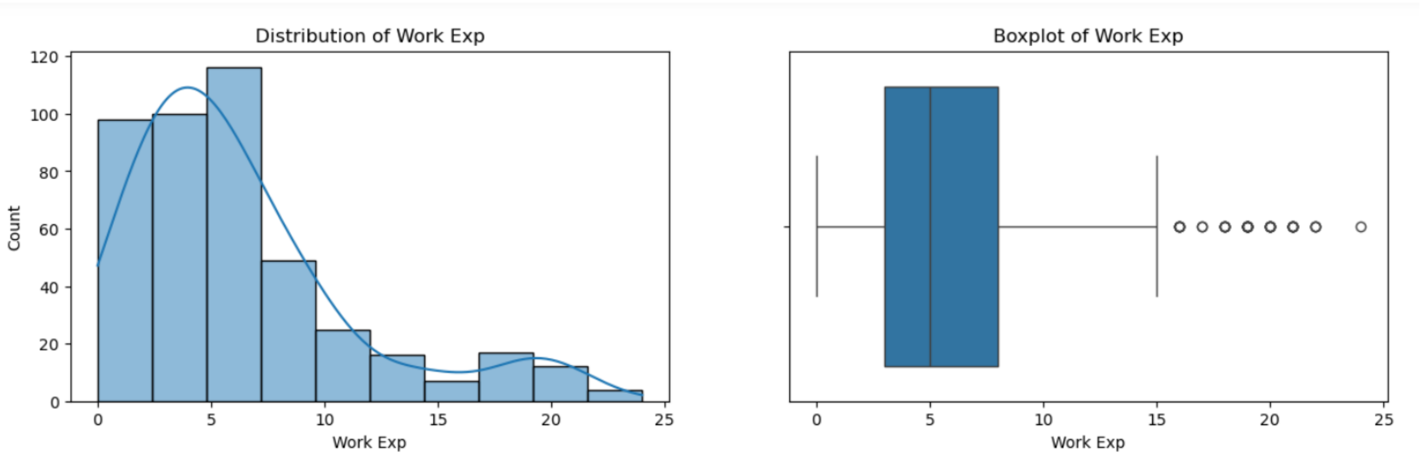
As we can see in histogram and box plot, the age has positive skewed distribution. In this positive skewness that the data points with lower age values and tail extend to the higher age values.

The distribution is not symmetrical**.** positive skewness may imply that the sample has higher concentration of younger individuals with fewer individuals. Outliers are also presented in the box plot**.**



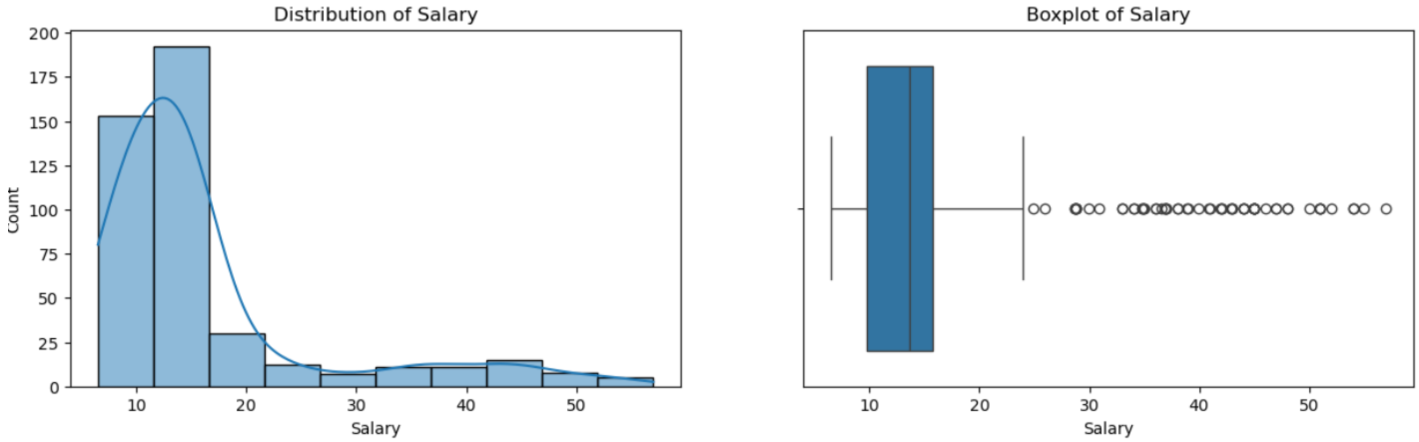
Work experience column:

There is a right skewed distribution in experience column and it's not symmetrical distribution.



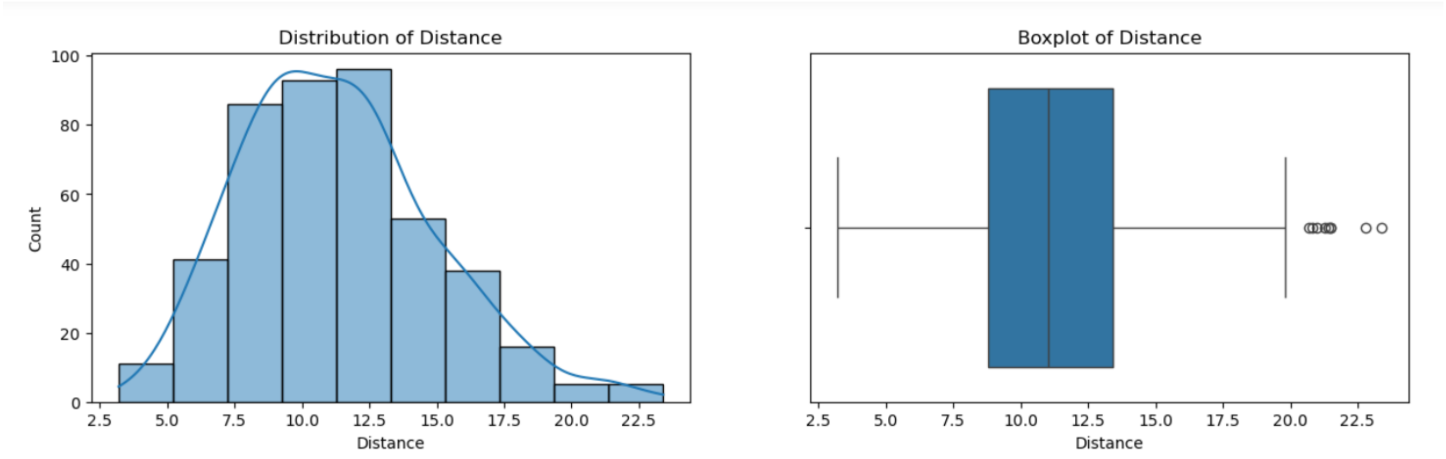
Salary Column:

The salary column have right skewed distribution and it's a positive skew. In **boxplot** have more outliers presented in salary feature.



Distance Column:

In this distance column have normal distribution and it’s almost have a symmetrical distribution.



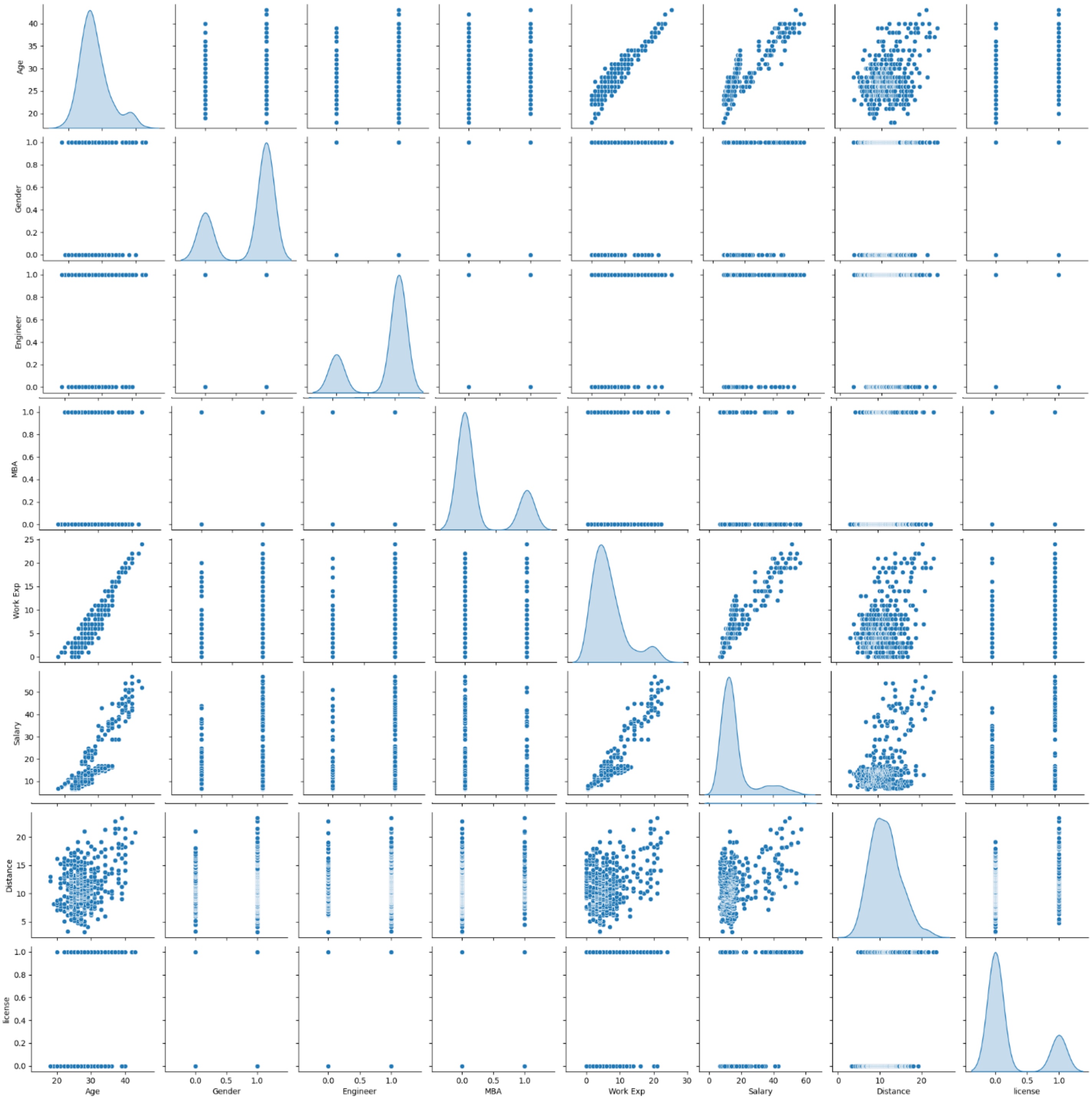
Bivariate Analysis – Pairplot

We can see the most of features are presented in a outlier except distance column. Similarly, the age and work experience column has positive correlation, age and salary also positive correlation.

There is a cloud formation happened in age and distance we can’t clearly tell positive or negative.

Positive correlation happened in work experience and salary feature.

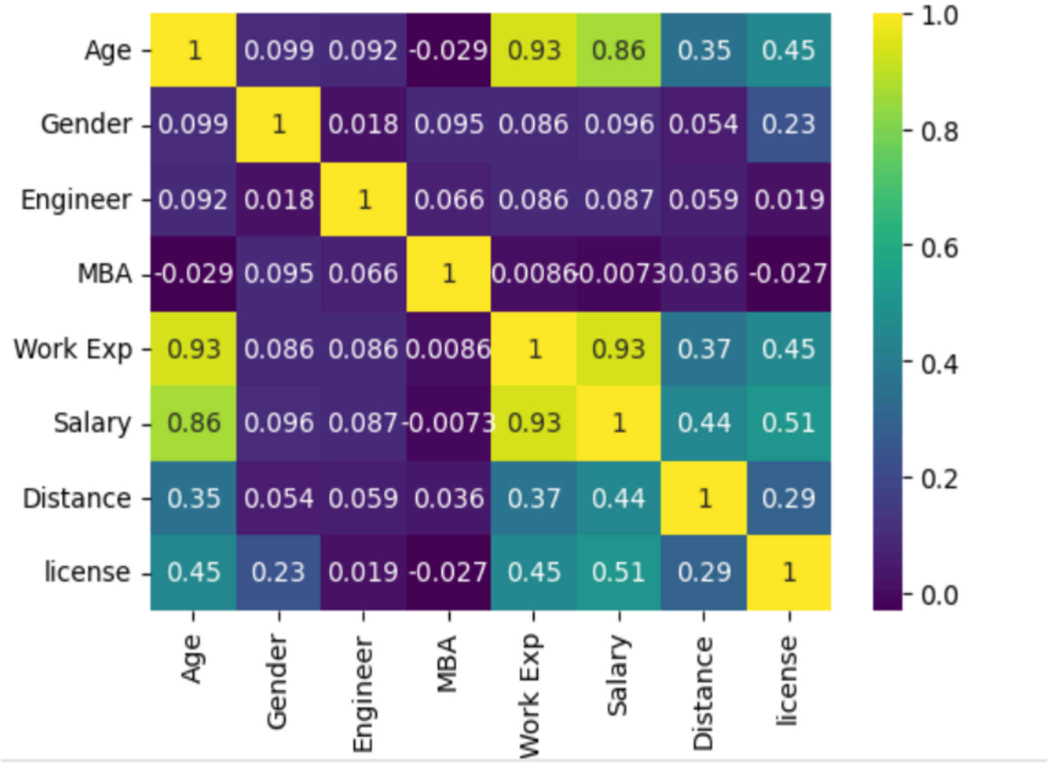
Cloud formation on distance and work experience.


Age
20
25
30
35
40
Gender
0.0
0.2
0.4
0.6
0.8
1.0
Engineer
0.0
0.2
0.4
0.6
0.8
1.0
MBA
0.0
0.2
0.4
0.6
0.8
1.0
Work Exp
0
5
10
15
20
25
Salary
10
20
30
40
50
Distance
5
10
15
20
license
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
Age
0.0
0.5
1.0
Gender
0.0
0.5
1.0
Engineer
0.0
0.5
1.0
MBA
0
10
20
30
Work Exp
0
20
40
60
Salary
0
10
20
Distance
0.0
0.5
1.0
license

Heat map – Correlation

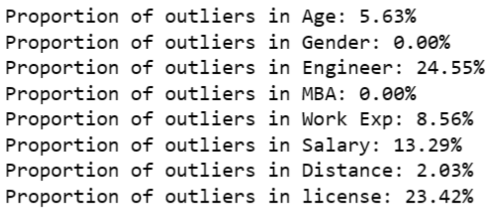
Strong correlation: Work experience and age, salary and age and salary and work experience.

Weak Correlation: Salary and MBA, MBA and License, MBA and age, Engineer and license.

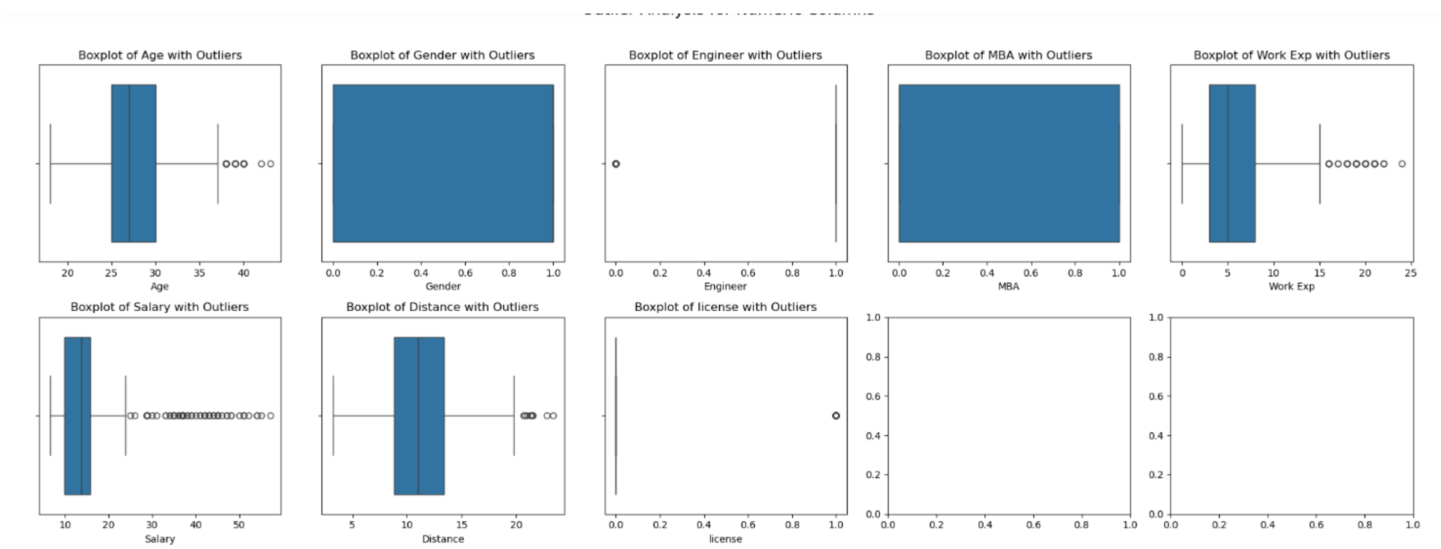


Outlier Detection Analysis -

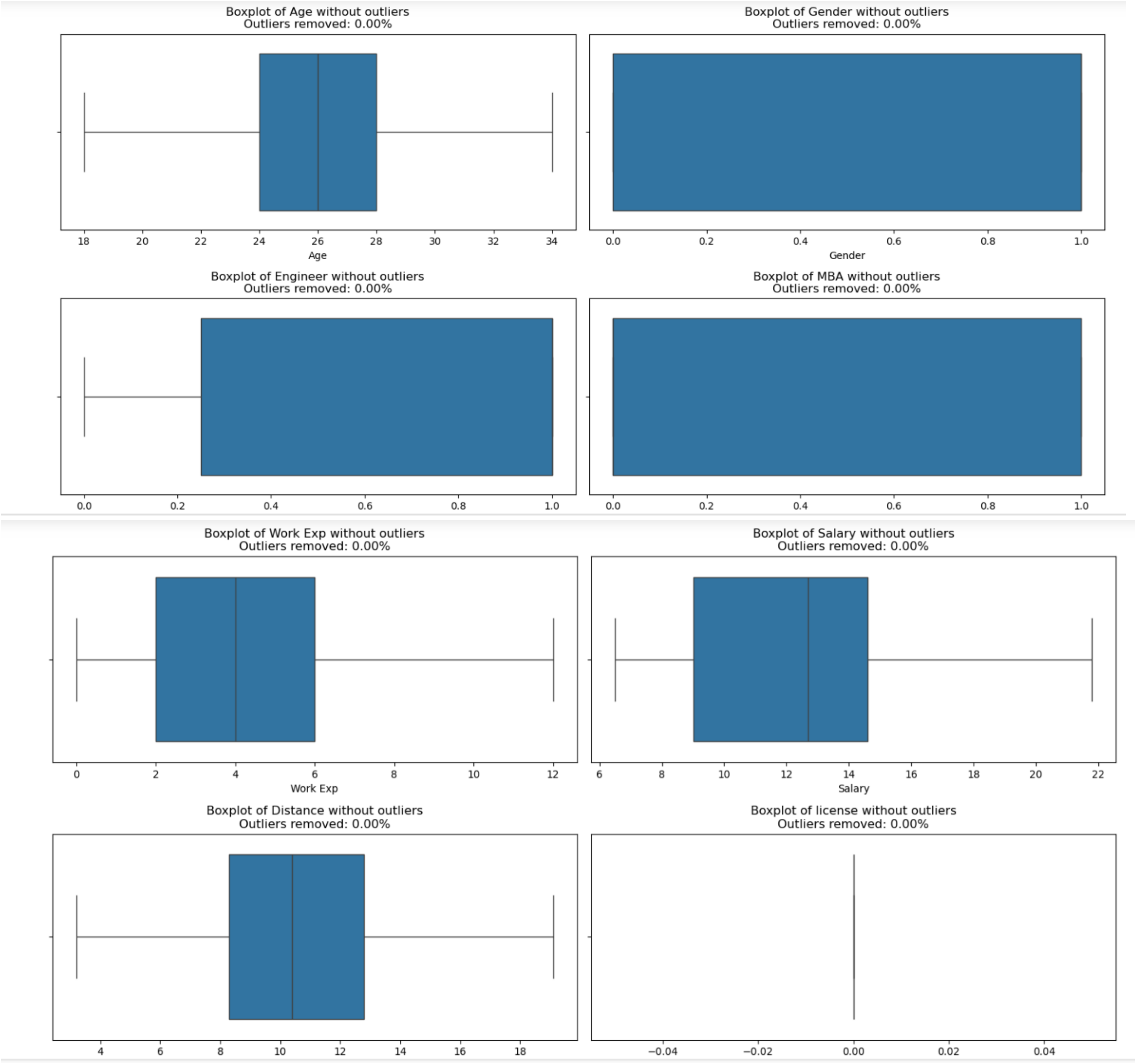
There is outlier for multiple columns except MBA and Gender.



```
Proportion of outliers in Age: 5.63%
Proportion of outliers in Gender: 0.00%
Proportion of outliers in Engineer: 24.55%
Proportion of outliers in MBA: 0.00%
Proportion of outliers in Work Exp: 8.56%
Proportion of outliers in Salary: 13.29%
Proportion of outliers in Distance: 2.03%
Proportion of outliers in license: 23.42%
```



Boxplot – Outlier Removal



Train test Split Method and scaling is necessary because, scaling ensures that all features contribute equally to the model training process.

It helps optimization algorithms to converge faster.

We can use the standard scalar method.

In this feature to have a **mean 0 and standard deviation 1.**

1.3 Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance.: a. Logistic Regression Model b. Linear Discriminant Analysis c. Decision Tree Classifier – CART model d. Naïve Bayes Model e. KNN Model f. Random Forest Model g. Boosting Classifier Model using Gradient boost.

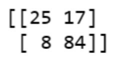
1.Logistic Regression:

Let's we look into recall function.

Generally the **public is 1 and private is 0**

The algorithm clearly tells us the 91 % of the employees came in public transport and the f1 score also 87% will fall on public transport.

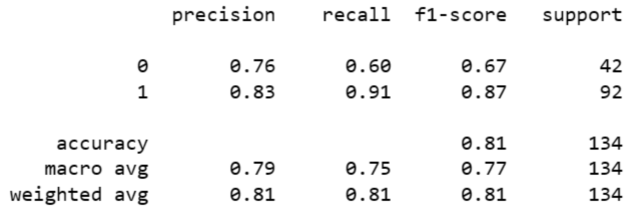
Confusion Matrix:
Here the correct predictions are 25,84 and 17,8 are predicted as incorrect.



```
[[25 17]
 [ 8 84]]
```

Classification Report:

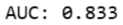
The accuracy of the model is 81% and performed well.



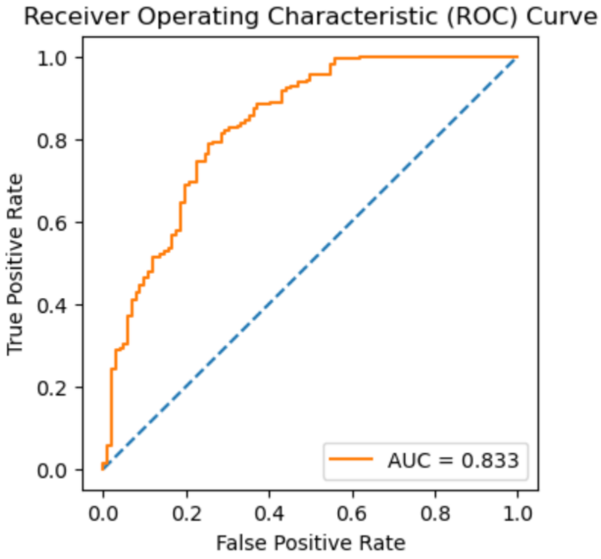
```
              precision    recall  f1-score   support

           0       0.76      0.60      0.67        42
           1       0.83      0.91      0.87        92

    accuracy                           0.81       134
   macro avg       0.79      0.75      0.77       134
weighted avg       0.81      0.81      0.81       134
```

ROC-AUC Curve:

For this model has AUC-ROC rate is 83% so the TPR and FPR rate is well.



```
AUC: 0.833
```

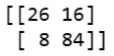


In this model 83 %came by public transport

2.Linear Discriminant analysis:

Confusion Matrix:

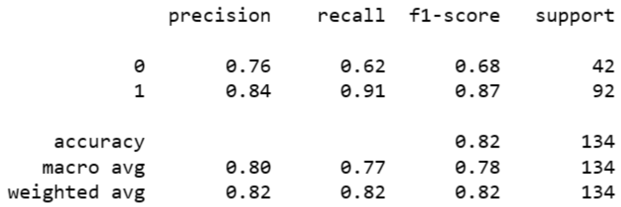
It describes about TP, TN and FP and FN
26 and 84 are correct prediction and 16 and 8 are incorrect prediction



```
[[26 16]
 [ 8 84]]
```

Classification report:

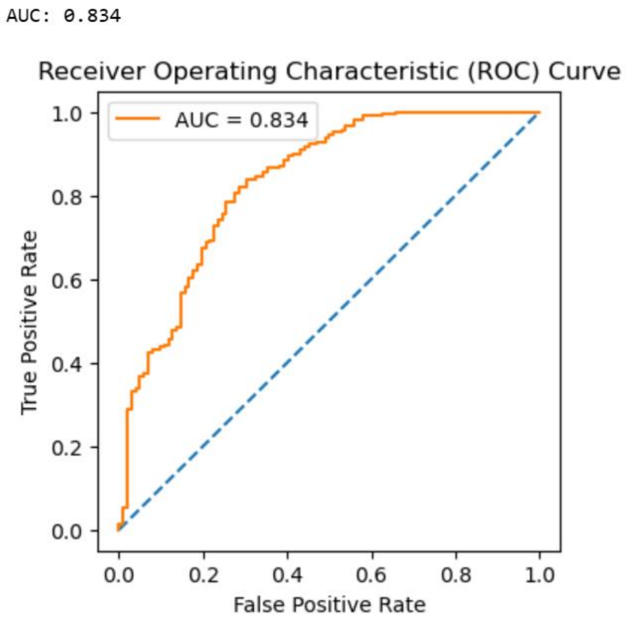
 The Classification describes the model predicted as the 87% of public transport and the over all accuracy socre is 82%



```
              precision    recall  f1-score   support

           0       0.76      0.62      0.68        42
           1       0.84      0.91      0.87        92

    accuracy                           0.82       134
   macro avg       0.80      0.77      0.78       134
weighted avg       0.82      0.82      0.82       134
```

ROC –AUC Curve:

The 83.3% ROC suggest the strong ability have this model for private and public transport

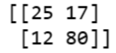


```
AUC: 0.834
```

3.KNN Naive Bayes:

Confusion Matrix:

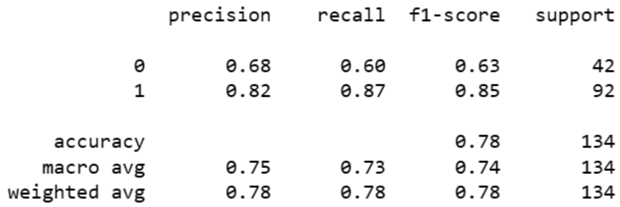
The 25 and 80 are correct prediction and 17 and 12 are incorrect prediction.



```
[[25 17]
 [12 80]]
```

Classification Report:

Model predicted 87% of predicted the employee came by public transport.Overall score is 78% percentage

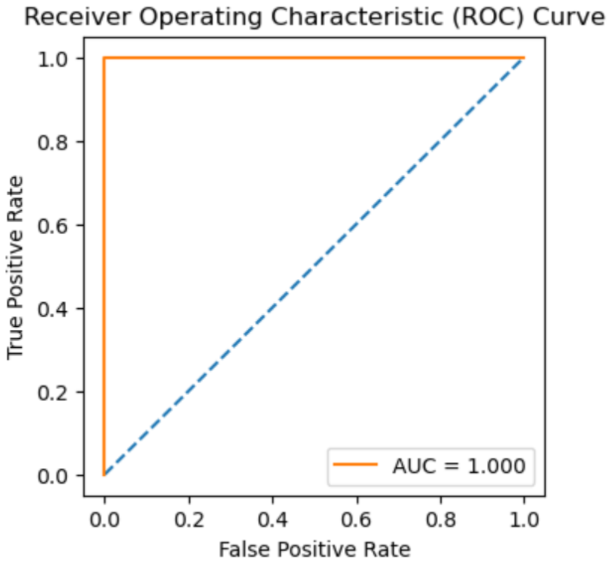


```
              precision    recall  f1-score   support

           0       0.68      0.60      0.63        42
           1       0.82      0.87      0.85        92

    accuracy                           0.78       134
   macro avg       0.75      0.73      0.74       134
weighted avg       0.78      0.78      0.78       134
```

ROC- AUC Curve:

This model suggest strong classification of public and private transport.The AUC and ROC score 100%

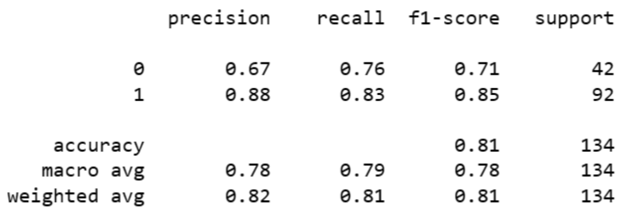
```
AUC: 1.000
```



4.Decision Tree Classifier – CART Model:

Classification report:

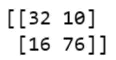
This model has 83% correctly predicted that the employee came in public transport. The accuracy is 78%.



```
              precision    recall  f1-score   support

           0       0.67      0.76      0.71        42
           1       0.88      0.83      0.85        92

    accuracy                           0.81       134
   macro avg       0.78      0.79      0.78       134
weighted avg       0.82      0.81      0.81       134
```

Confusion Matrix:

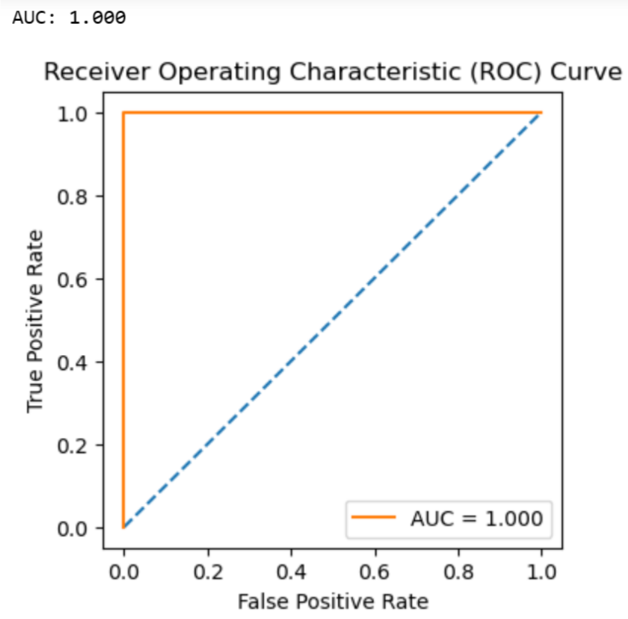
32 and 76 are correctt prediction and 10 and 126 are incorrect prediction



```
[[32 10]
 [16 76]]
```

ROC-AUC Curve:

This model has strong classification positive and negative classes.

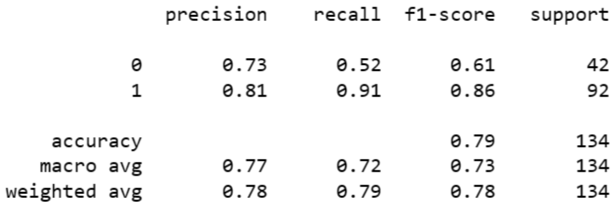


```
AUC: 1.000
```

5.Naïve Bayes Model:

Classification report:

The model has correctly predicted as 91% came by puiblic transport and accuracy is 79%

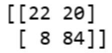


```
              precision    recall  f1-score   support

           0       0.73      0.52      0.61        42
           1       0.81      0.91      0.86        92

    accuracy                           0.79       134
   macro avg       0.77      0.72      0.73       134
weighted avg       0.78      0.79      0.78       134
```

Confusion Matrix:
Correct prediction:
22 and 84

Incorrect Prediction:

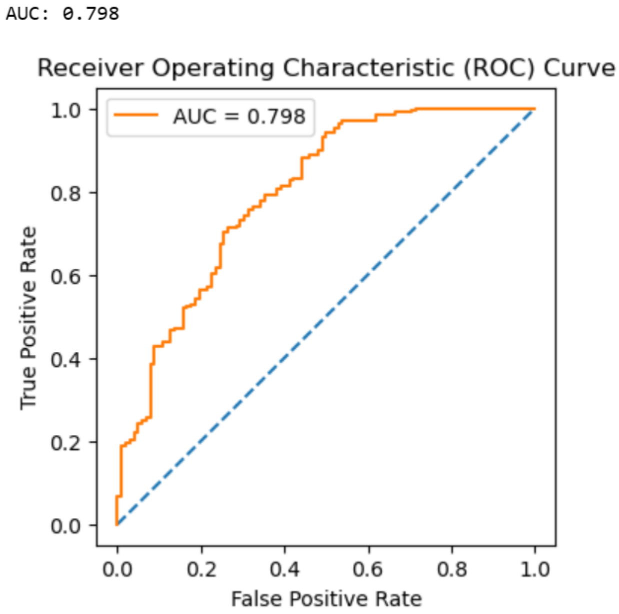
20 and 8



```
[[22 20]
 [ 8 84]]
```

ROC-AUC Curve:

This model is fair distinguish when compare to other model

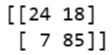


```
AUC: 0.798
```

6.Random Forest Classifier:

This model has 24 and 85 are correct prediction and 18 and 78 are incorrect prediction.

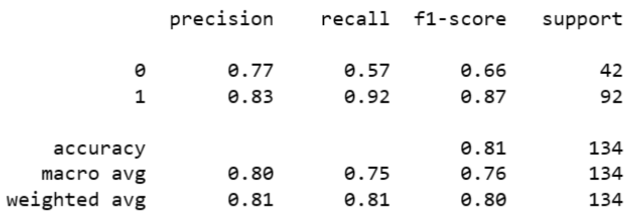
Confusion matrix:



```
[[24 18]
 [ 7 85]]
```

Classification Report:

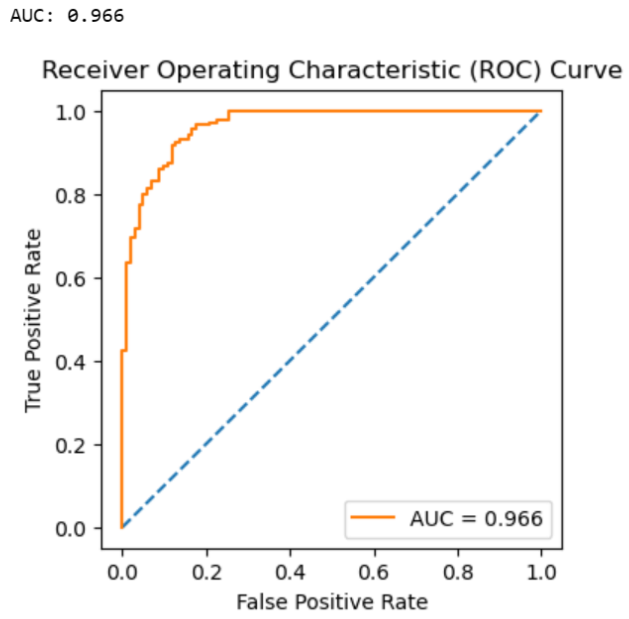
The recall percentage is 92 so the employees came by public transport



```
              precision    recall  f1-score   support

           0       0.77      0.57      0.66        42
           1       0.83      0.92      0.87        92

    accuracy                           0.81       134
   macro avg       0.80      0.75      0.76       134
weighted avg       0.81      0.81      0.80       134
```

ROC- AUC Curve:

This model has a strong discriminant ability between public and private transport.
Overall percentage is 96%

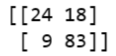


```
AUC: 0.966
```

7.Gradient Boositng Classifier:

Confusion Matrix:

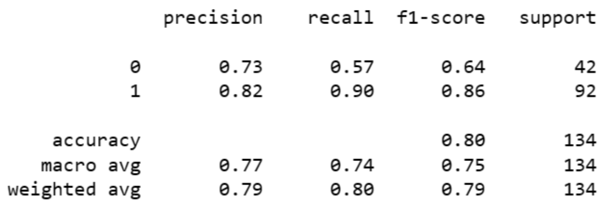
The model has 24 and 83 are correct and 18,9 are incorrect prediction



```
[[24 18]
 [ 9 83]]
```

Classification report:

90% of the prediction has public transport

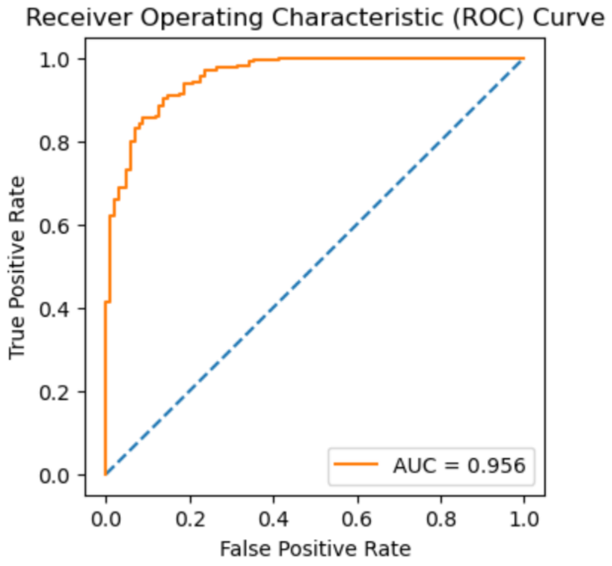


```
              precision    recall  f1-score   support

           0       0.73      0.57      0.64        42
           1       0.82      0.90      0.86        92

    accuracy                           0.80       134
   macro avg       0.77      0.74      0.75       134
weighted avg       0.79      0.80      0.79       134
```

ROC_AUC Curve:

This model has strong distinguish between public and private transport

```
AUC: 0.956
```



1.4 Which model performs the best?

KNN Naïve bayes and Decision Tree Classifier has performed well among all the model

The 2 model score is 100

1.5 What are your business insights?

For all the model describes, people came by public transport. We need to give our ABC can provide the transport to employees. Because each and every model clearly tells the employees travelled from home to office through bus.