INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

# IISC Bengaluru CCE 2022
# Computing for Artificial Intelligence and Machine Learning

## Project - Transportation Choice Prediction for office employees using Machine Learning

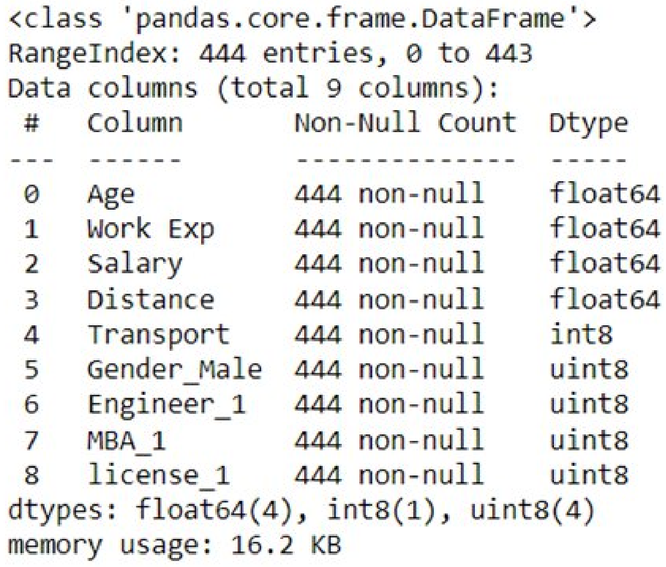
**Submitted by**
**Mr. Prashant Kumar Lanka**
**Mr. Ravikumar**

# Project Summary

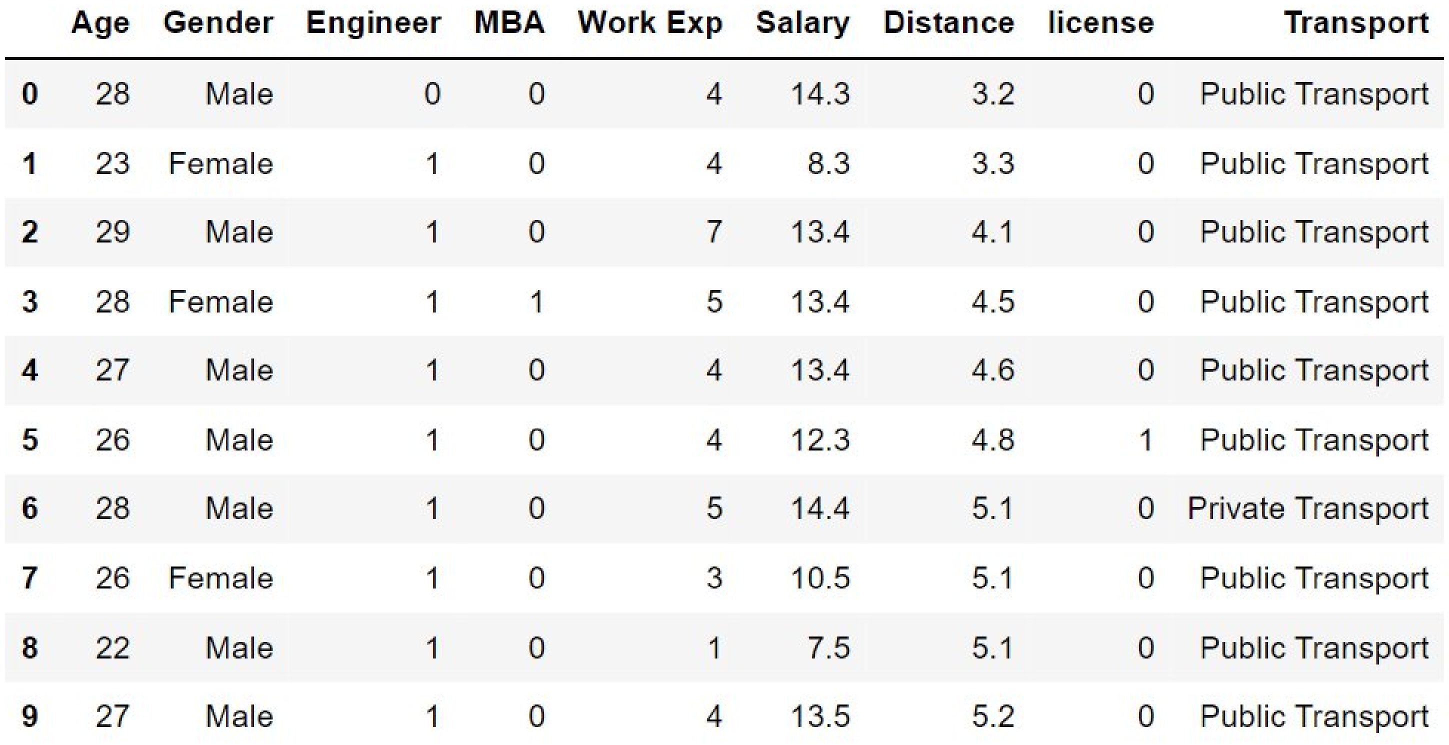
**Objective :**
Using different machine learning models to predict whether an employee will choose public transport or private transport basis on the features like Age, Gender, Education, Salary & license possession

Shape & Info of the dataset:

- number of rows: 444
- number of columns: 9

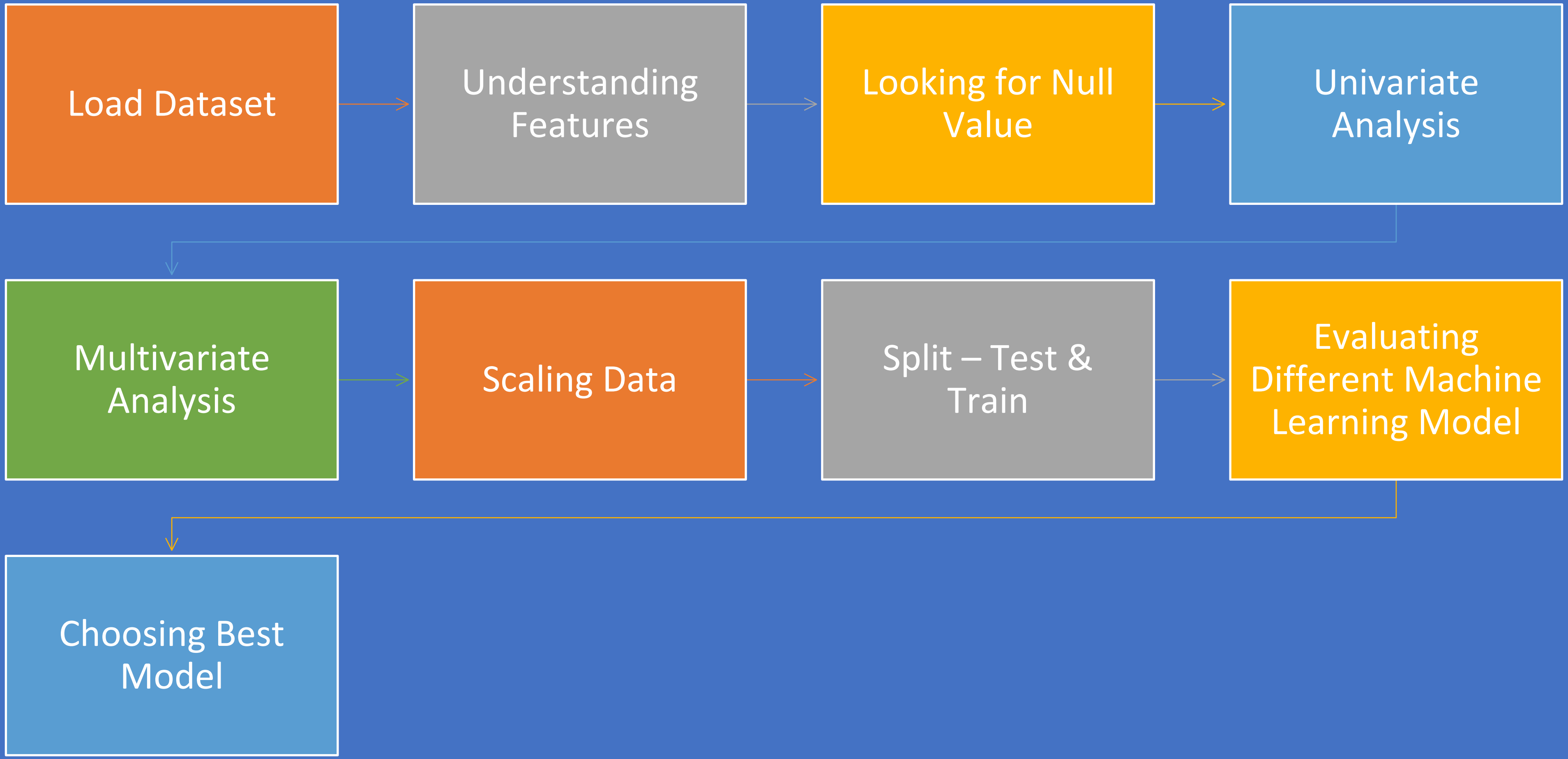


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 444 entries, 0 to 443
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Age          444 non-null    float64
 1   Work Exp     444 non-null    float64
 2   Salary       444 non-null    float64
 3   Distance     444 non-null    float64
 4   Transport    444 non-null    int8
 5   Gender_Male  444 non-null    uint8
 6   Engineer_1   444 non-null    uint8
 7   MBA_1        444 non-null    uint8
 8   license_1    444 non-null    uint8
dtypes: float64(4), int8(1), uint8(4)
memory usage: 16.2 KB
```



| | 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 |
| **5** | 26 | Male | 1 | 0 | 4 | 12.3 | 4.8 | 1 | Public Transport |
| **6** | 28 | Male | 1 | 0 | 5 | 14.4 | 5.1 | 0 | Private Transport |
| **7** | 26 | Female | 1 | 0 | 3 | 10.5 | 5.1 | 0 | Public Transport |
| **8** | 22 | Male | 1 | 0 | 1 | 7.5 | 5.1 | 0 | Public Transport |
| **9** | 27 | Male | 1 | 0 | 4 | 13.5 | 5.2 | 0 | Public Transport |

# Work Flow -



Load Dataset

Understanding Features

Looking for Null Value

Univariate Analysis

Multivariate Analysis

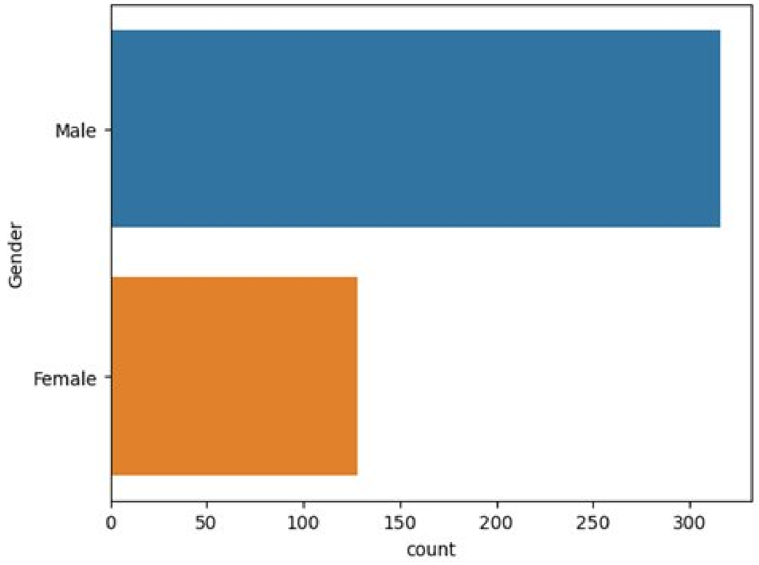
Scaling Data

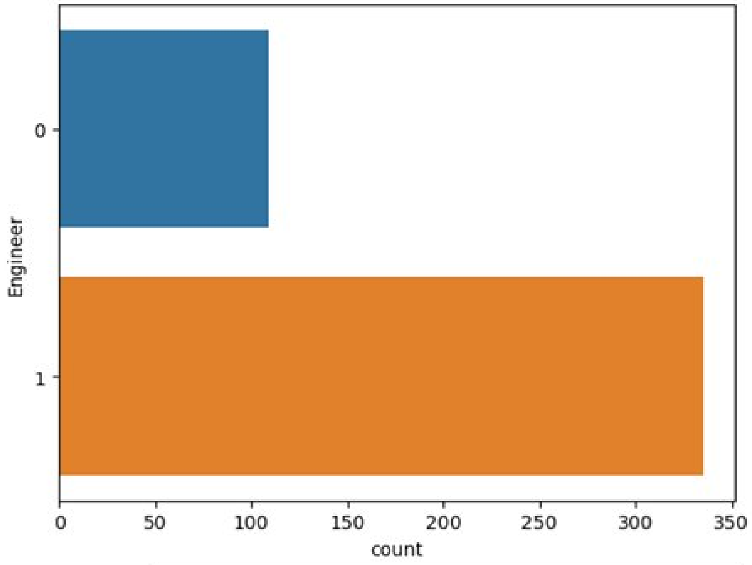
Split – Test & Train

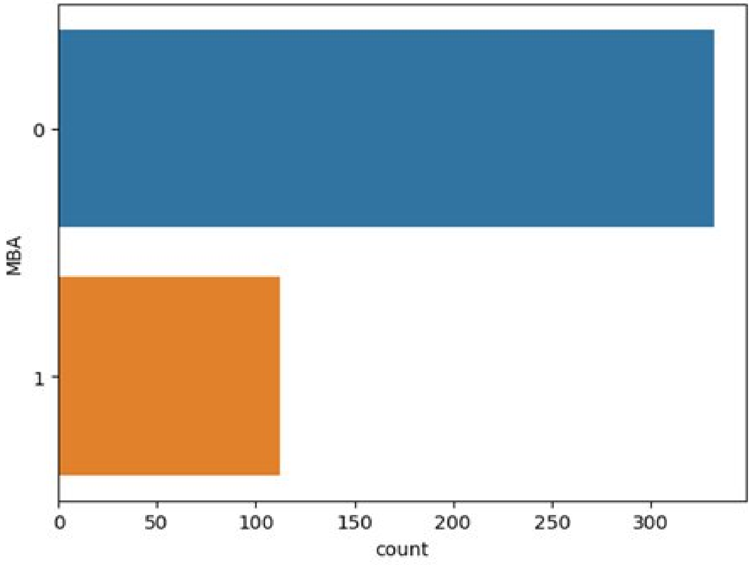
Evaluating Different Machine Learning Model

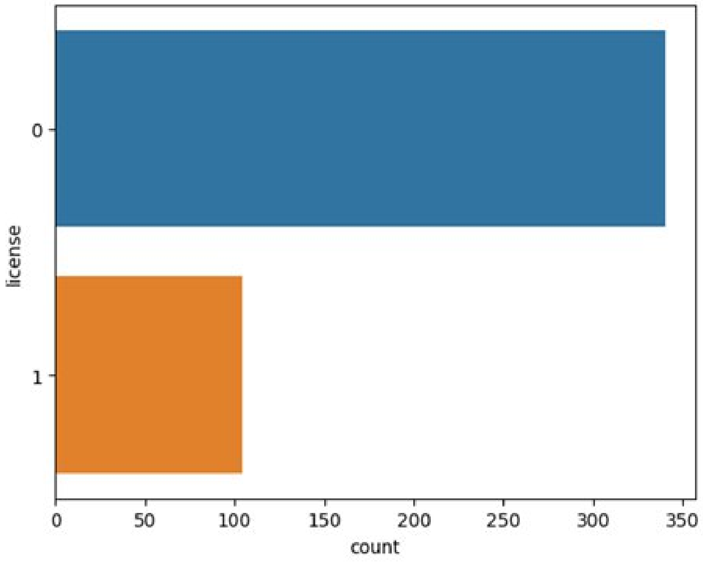
Choosing Best Model

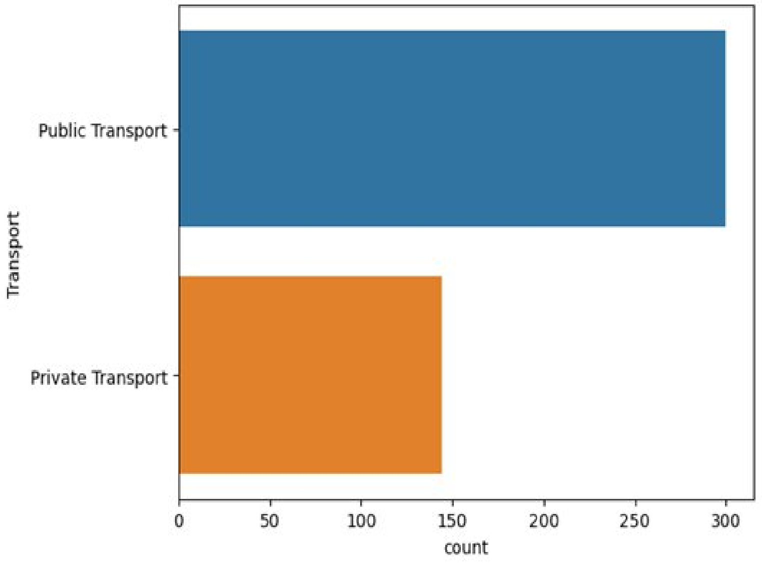
# Univariate Analysis:



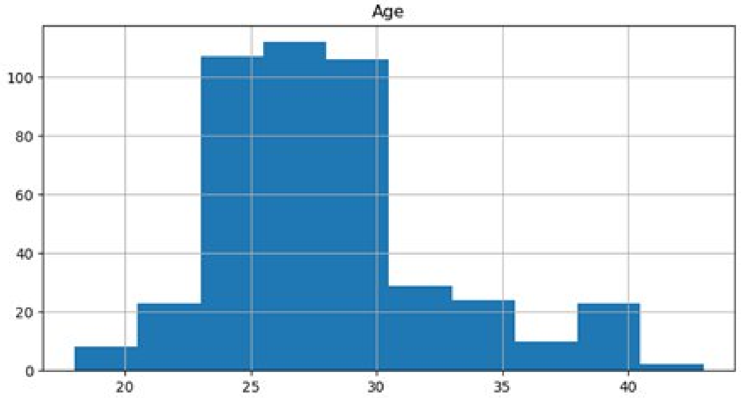


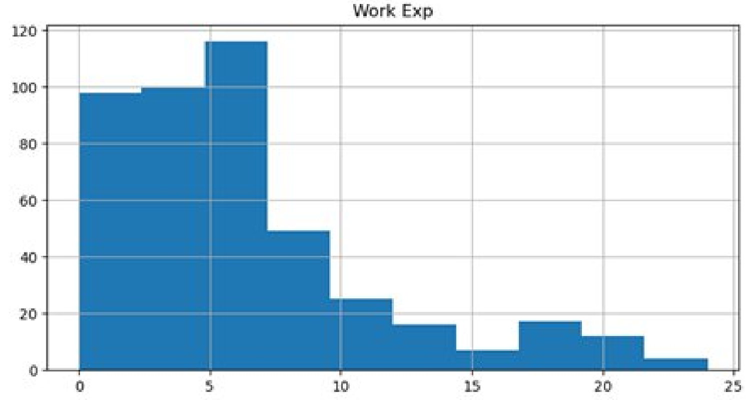


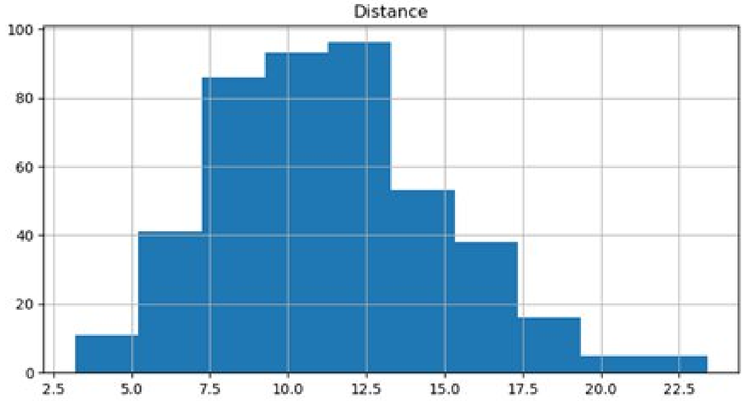


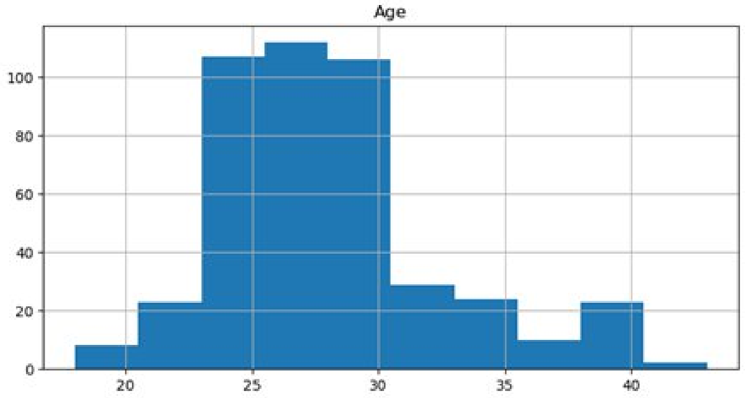


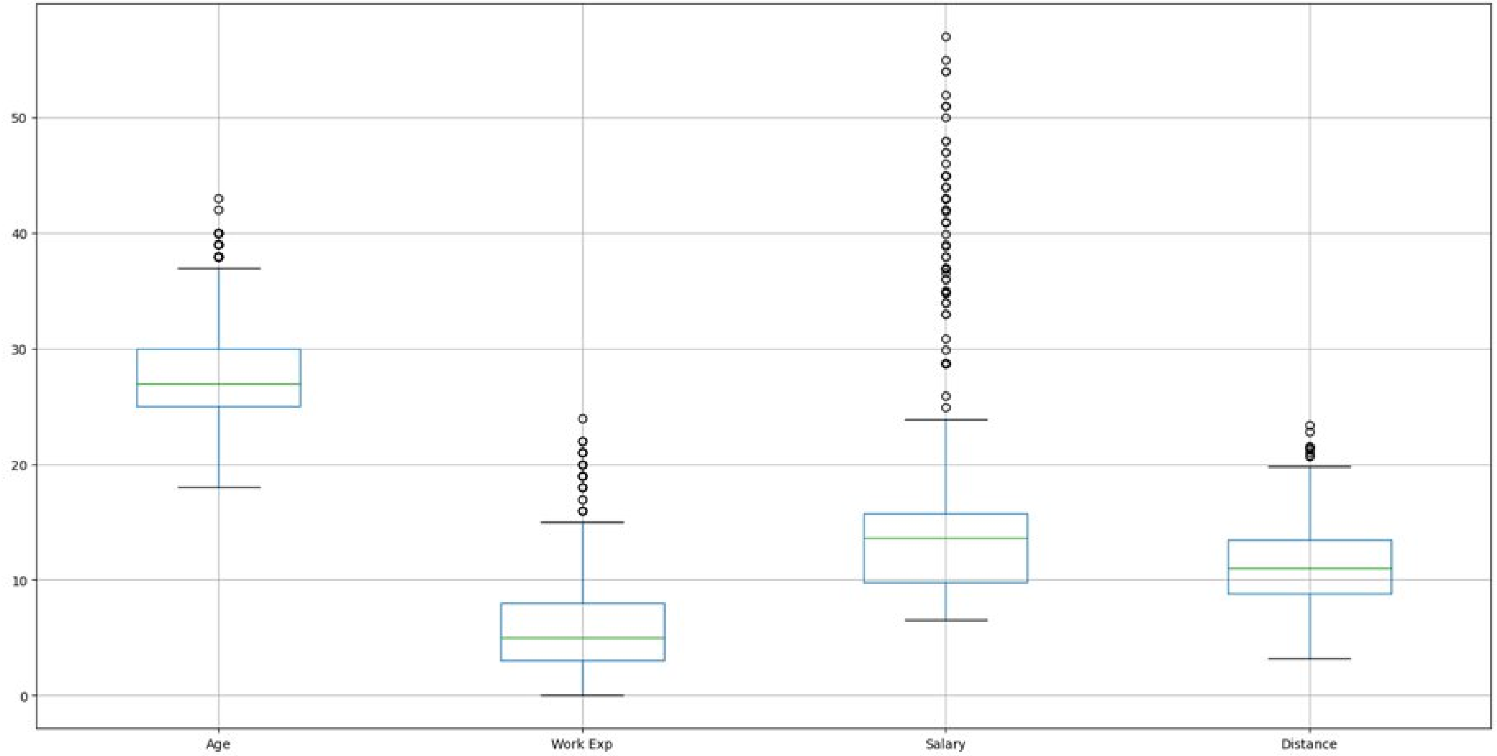
# Univariate Analysis:



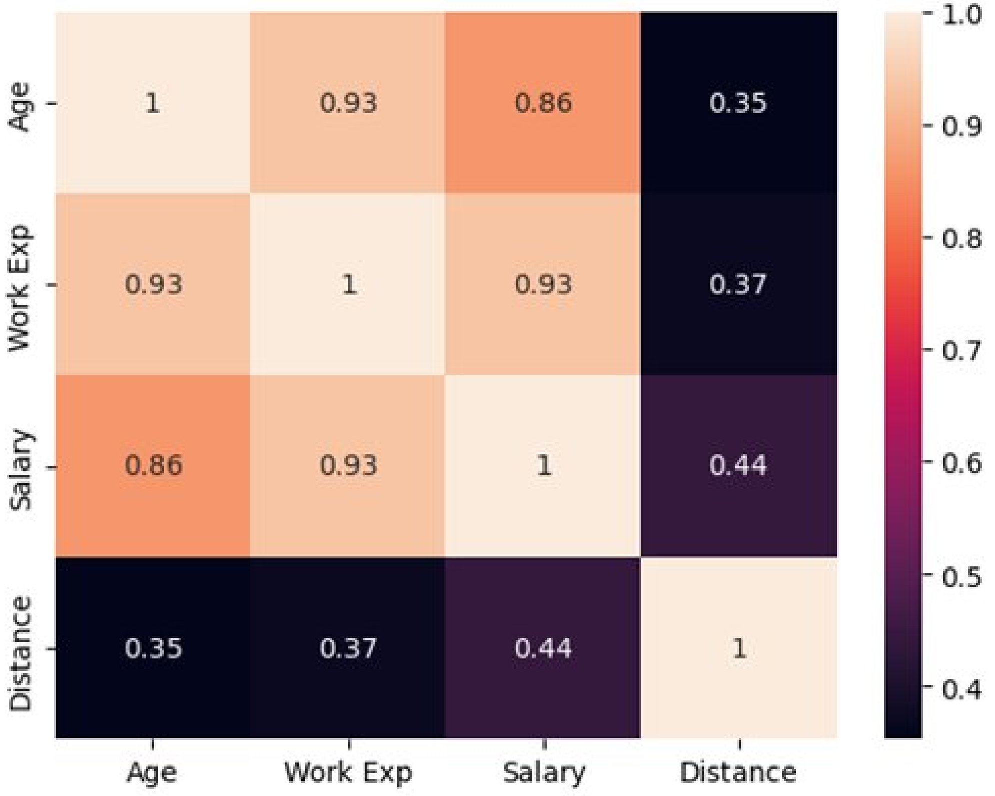


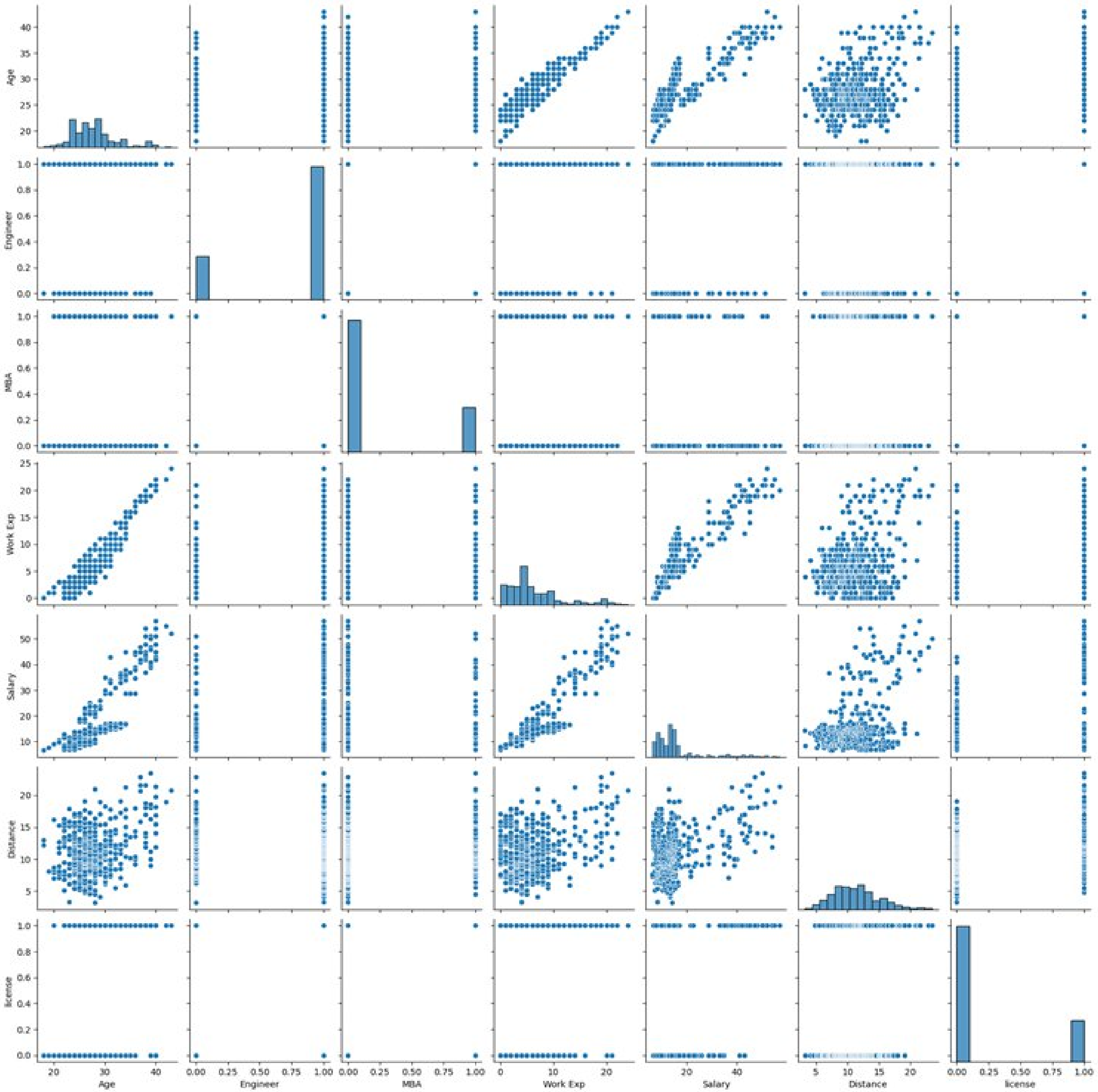






# Multivariate Analysis

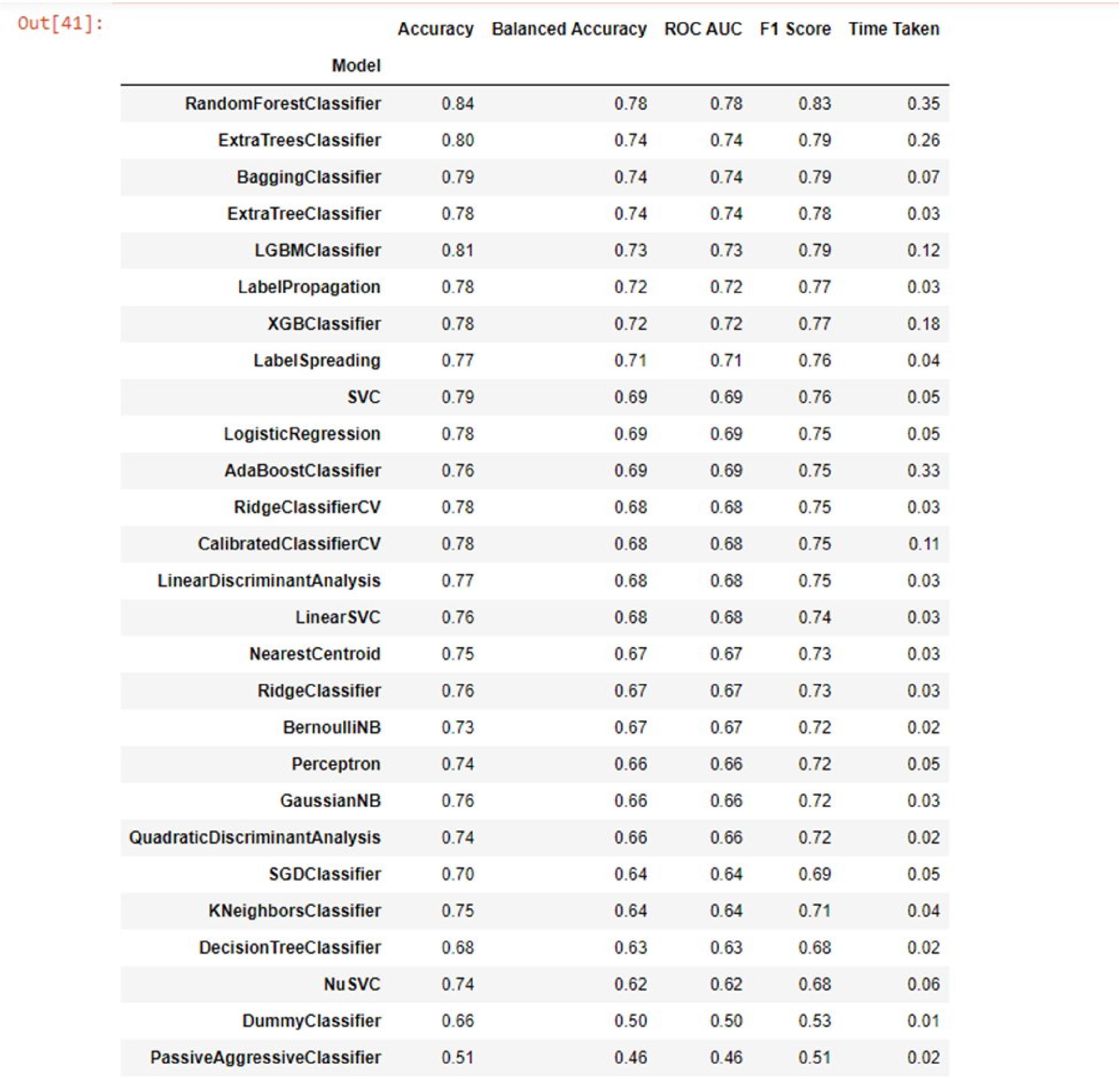




It is found that Age & Work Ex. And Age and Salary are highly correlated. There is a poor or less correlation between distance and other feature types.

# Machine Learning Model

Machine Learning Models are evaluated and compared using Lazy predict code. This checks for the 26 different machine learning model and gives comparable data for it.



Out[41]:

| | Accuracy | Balanced Accuracy | ROC AUC | F1 Score | Time Taken |
|---|---|---|---|---|---|
| **Model** | | | | | |
| **RandomForestClassifier** | 0.84 | 0.78 | 0.78 | 0.83 | 0.35 |
| **ExtraTreesClassifier** | 0.80 | 0.74 | 0.74 | 0.79 | 0.26 |
| **BaggingClassifier** | 0.79 | 0.74 | 0.74 | 0.79 | 0.07 |
| **ExtraTreeClassifier** | 0.78 | 0.74 | 0.74 | 0.78 | 0.03 |
| **LGBMClassifier** | 0.81 | 0.73 | 0.73 | 0.79 | 0.12 |
| **LabelPropagation** | 0.78 | 0.72 | 0.72 | 0.77 | 0.03 |
| **XGBClassifier** | 0.78 | 0.72 | 0.72 | 0.77 | 0.18 |
| **LabelSpreading** | 0.77 | 0.71 | 0.71 | 0.76 | 0.04 |
| **SVC** | 0.79 | 0.69 | 0.69 | 0.76 | 0.05 |
| **LogisticRegression** | 0.78 | 0.69 | 0.69 | 0.75 | 0.05 |
| **AdaBoostClassifier** | 0.76 | 0.69 | 0.69 | 0.75 | 0.33 |
| **RidgeClassifierCV** | 0.78 | 0.68 | 0.68 | 0.75 | 0.03 |
| **CalibratedClassifierCV** | 0.78 | 0.68 | 0.68 | 0.75 | 0.11 |
| **LinearDiscriminantAnalysis** | 0.77 | 0.68 | 0.68 | 0.75 | 0.03 |
| **LinearSVC** | 0.76 | 0.68 | 0.68 | 0.74 | 0.03 |
| **NearestCentroid** | 0.75 | 0.67 | 0.67 | 0.73 | 0.03 |
| **RidgeClassifier** | 0.76 | 0.67 | 0.67 | 0.73 | 0.03 |
| **BernoulliNB** | 0.73 | 0.67 | 0.67 | 0.72 | 0.02 |
| **Perceptron** | 0.74 | 0.66 | 0.66 | 0.72 | 0.05 |
| **GaussianNB** | 0.76 | 0.66 | 0.66 | 0.72 | 0.03 |
| **QuadraticDiscriminantAnalysis** | 0.74 | 0.66 | 0.66 | 0.72 | 0.02 |
| **SGDClassifier** | 0.70 | 0.64 | 0.64 | 0.69 | 0.05 |
| **KNeighborsClassifier** | 0.75 | 0.64 | 0.64 | 0.71 | 0.04 |
| **DecisionTreeClassifier** | 0.68 | 0.63 | 0.63 | 0.68 | 0.02 |
| **NuSVC** | 0.74 | 0.62 | 0.62 | 0.68 | 0.06 |
| **DummyClassifier** | 0.66 | 0.50 | 0.50 | 0.53 | 0.01 |
| **PassiveAggressiveClassifier** | 0.51 | 0.46 | 0.46 | 0.51 | 0.02 |

# Conclusion -

The bivariate analysis indicate that based on gender and availability of license, both males and females having license prefer private transport. Therefore, apart from gender, license is another deciding factor in determining the use the private or public transport.

Based on the accuracy of the various models, it can be concluded that the Random Forest classifier model is the best model for the given dataset as this model gives the highest accuracy of 84%.