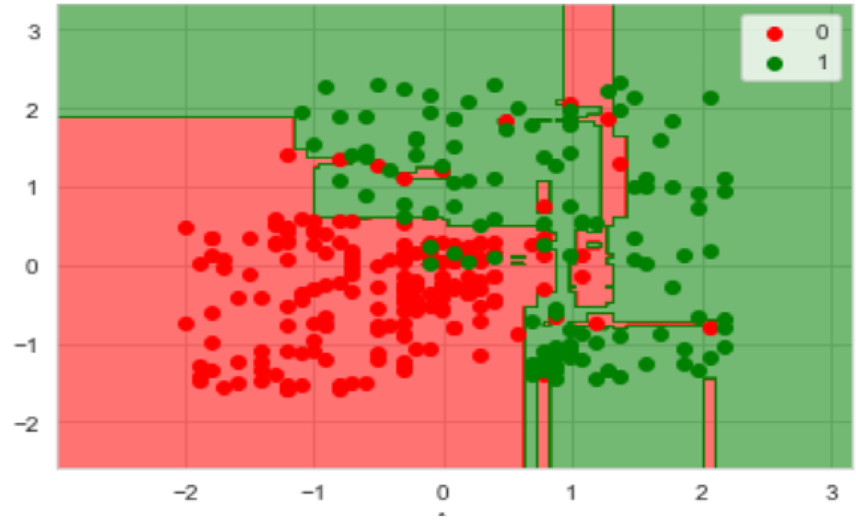
PROBLEM STATEMENT 2

# SOCIAL NETWORK AD ANALYSIS

## PROBLEM STATEMENT

Try to understand the dataset of Social_Network_Ads.csv and try to find the best suitable ML algorithm and write the code in python for algorithm from scratch and try to achieve the below output.



## OBJECTIVE

- UNDERSTANDING THE DATA SET
- DATA PREPROCESSING
- MODEL BUILDING
- RESULTS VISUALIZATION

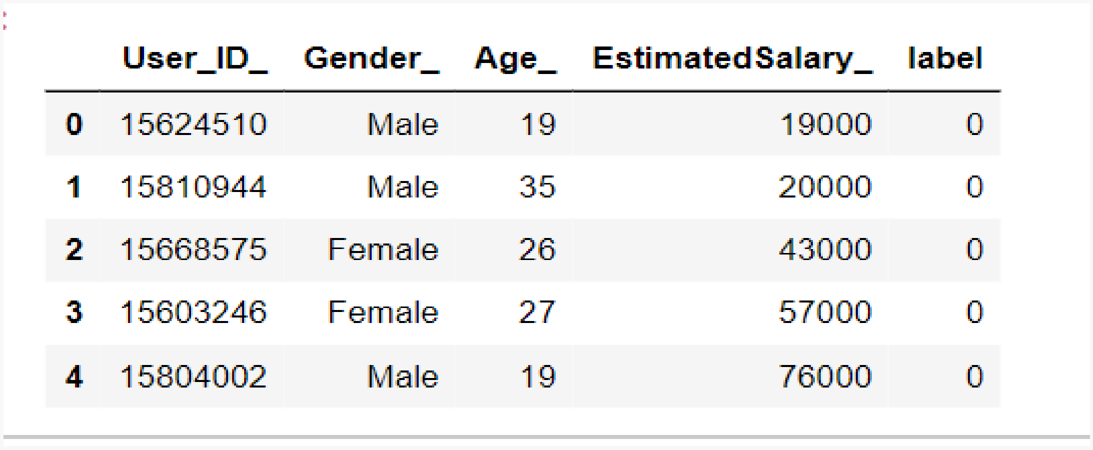
## DATASET

Dataset on Social media ads describes users, whether users have purchased a product by clicking on the advertisements shown to them.

ANALYSIS

UNDERSTANDING THE DATA SET

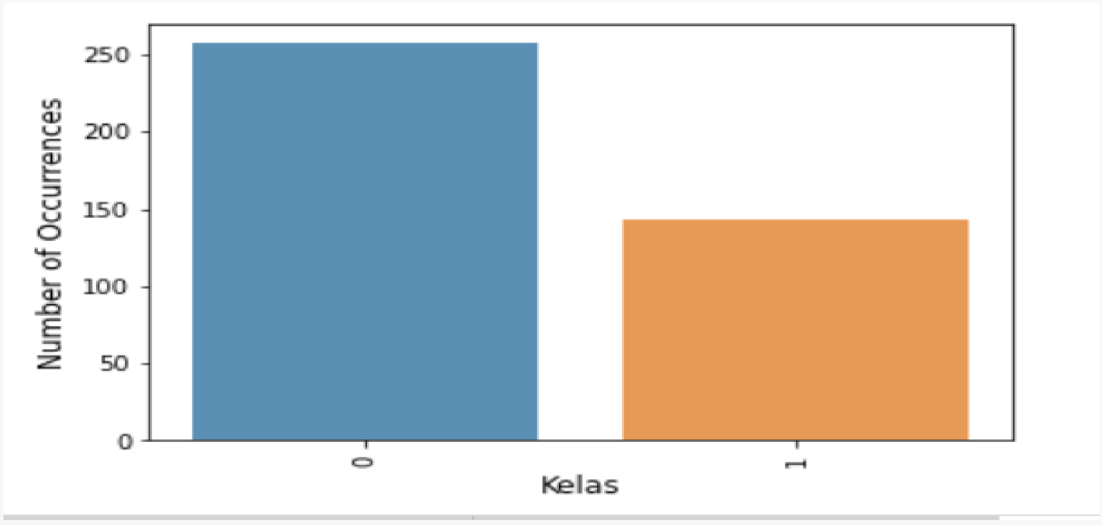
The dataset consists of 400 rows and 5 columns without any missing values. All the columns contains the int datatype except the gender column with the object datatype.



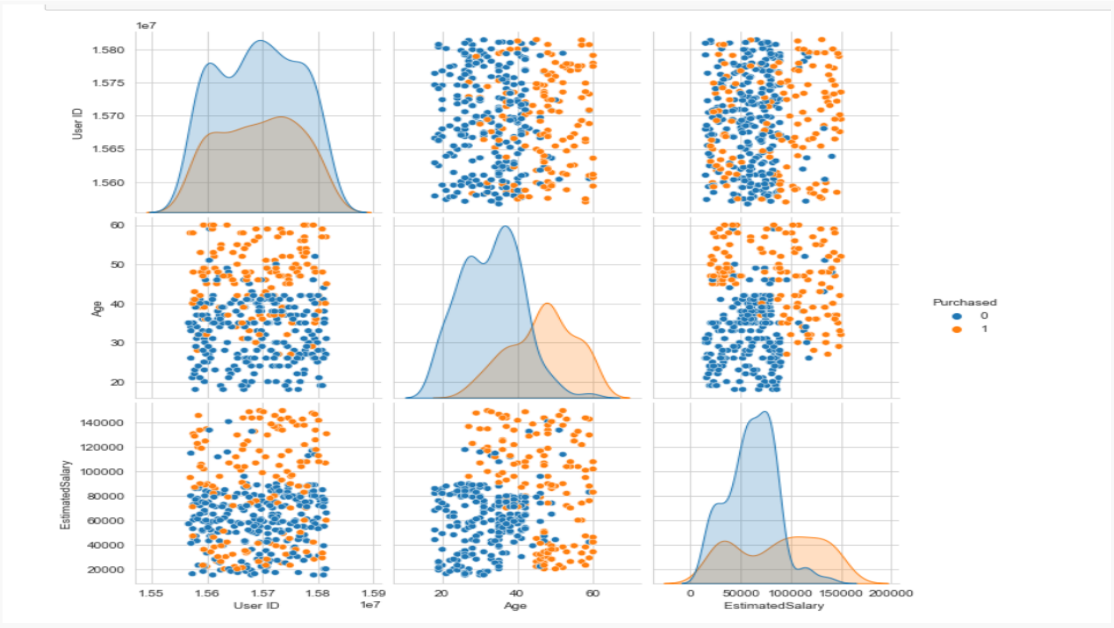
|   | User_ID_ | Gender_ | Age_ | EstimatedSalary_ | label |
|---|---|---|---|---|---|
| 0 | 15624510 | Male | 19 | 19000 | 0 |
| 1 | 15810944 | Male | 35 | 20000 | 0 |
| 2 | 15668575 | Female | 26 | 43000 | 0 |
| 3 | 15603246 | Female | 27 | 57000 | 0 |
| 4 | 15804002 | Male | 19 | 76000 | 0 |

VISUALIZATION OF DATASET

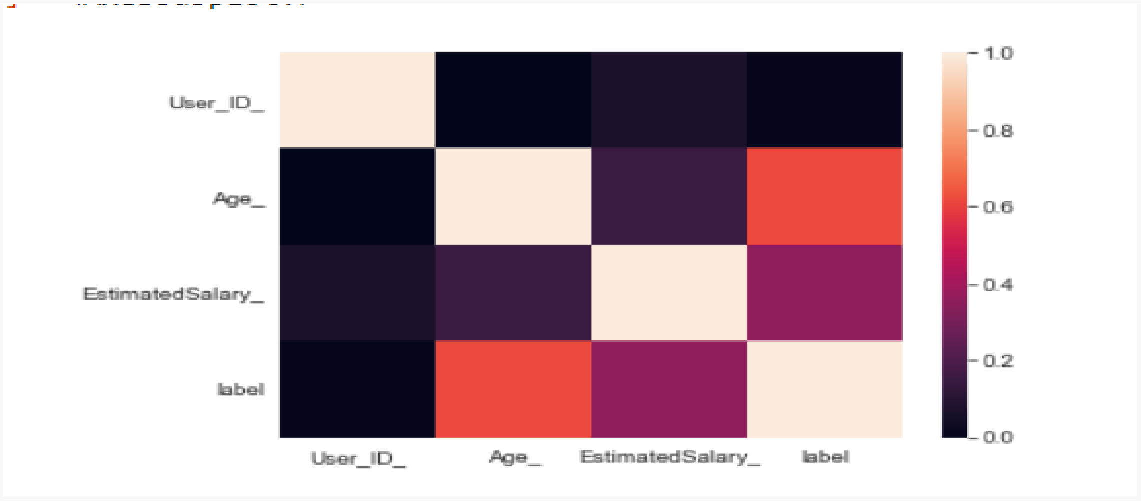
Here, barplot is fit to analyse the number of occurrence of people purchased vs non purchased.



The pairplot gives the complete understanding of data purchased with all other attributes in the dataset.



Heatmap shows the correlation between all the attributes.



RANDOMFOREST CLASSIFIER

Why is random forest used?

Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees. As the data set consists of clusters of purchases and non purchases, I chose random forest over decision trees as it classifies more than 1 tree to get more accuracy.

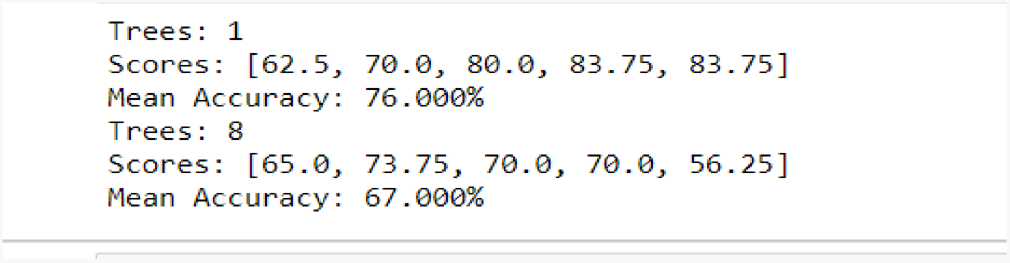
The random forest counted the no. of trees that counted YES(user buys suv) and no. of trees that counted NO (user doesn't buy SUV) and then takes the prediction that was voted the most times

DATA PREPARATION :

SPLITTING INTO TRAIN AND TEST DATA

x_train and x_test are trained using standard scalar.

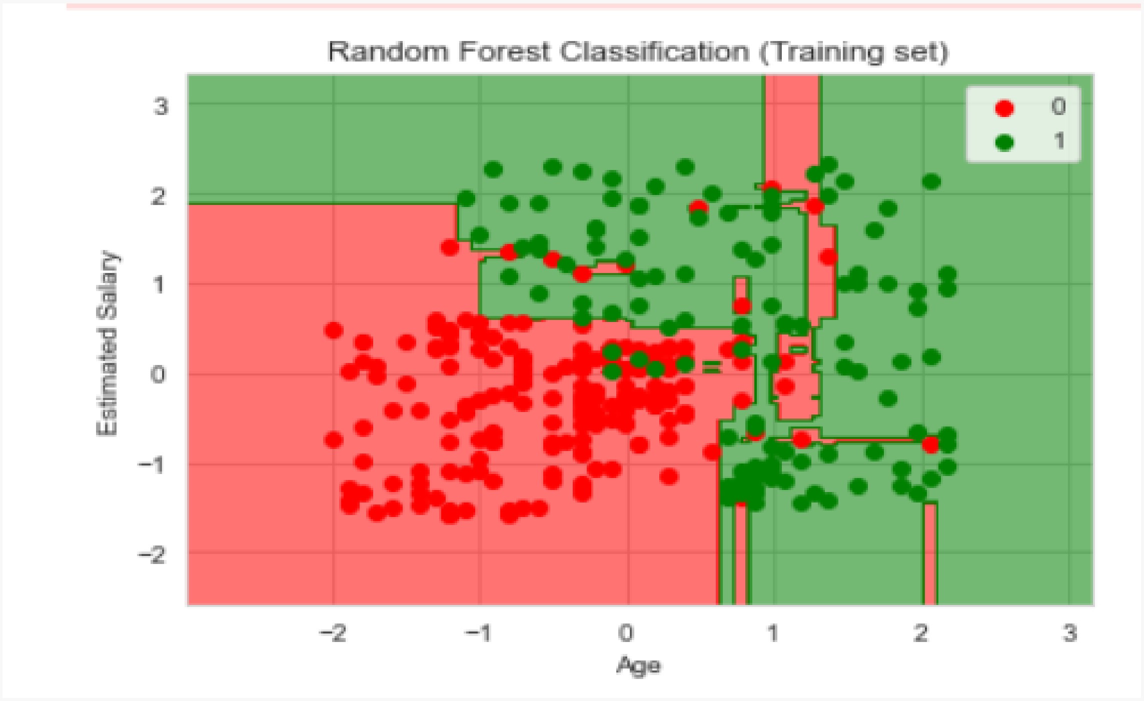
MODEL ACCURACY:



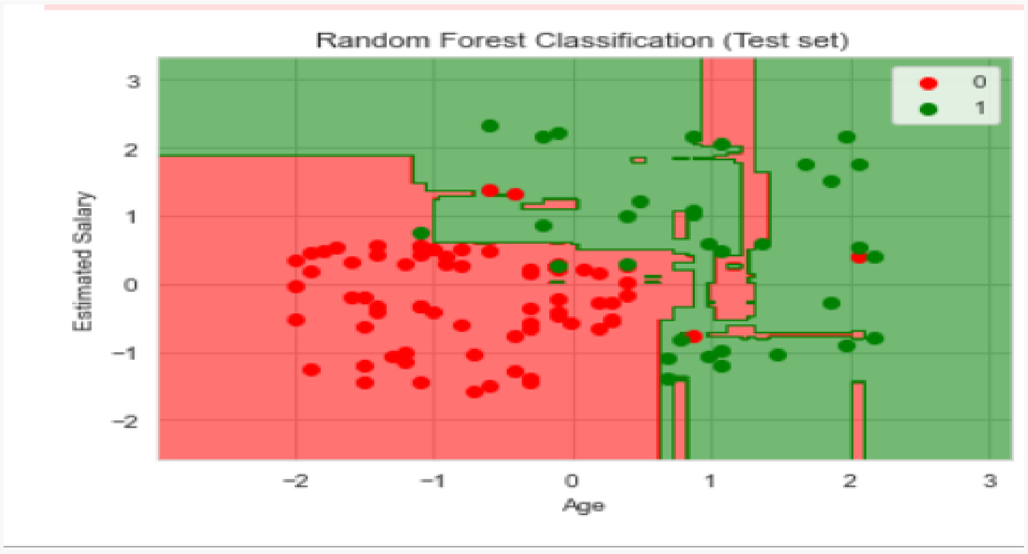
```
Trees: 1
Scores: [62.5, 70.0, 80.0, 83.75, 83.75]
Mean Accuracy: 76.000%
Trees: 8
Scores: [65.0, 73.75, 70.0, 70.0, 56.25]
Mean Accuracy: 67.000%
```

VISUALIZATION OF TRAINING AND TESTING SET

TRAINING SET



TESTING SET


Random Forest Classification (Test set)
Estimated Salary
Age
3
2
1
0
−1
−2
−2
−1
0
1
2
3
0
1