# ENGINEERING GRADUATE SALARY PREDICTION

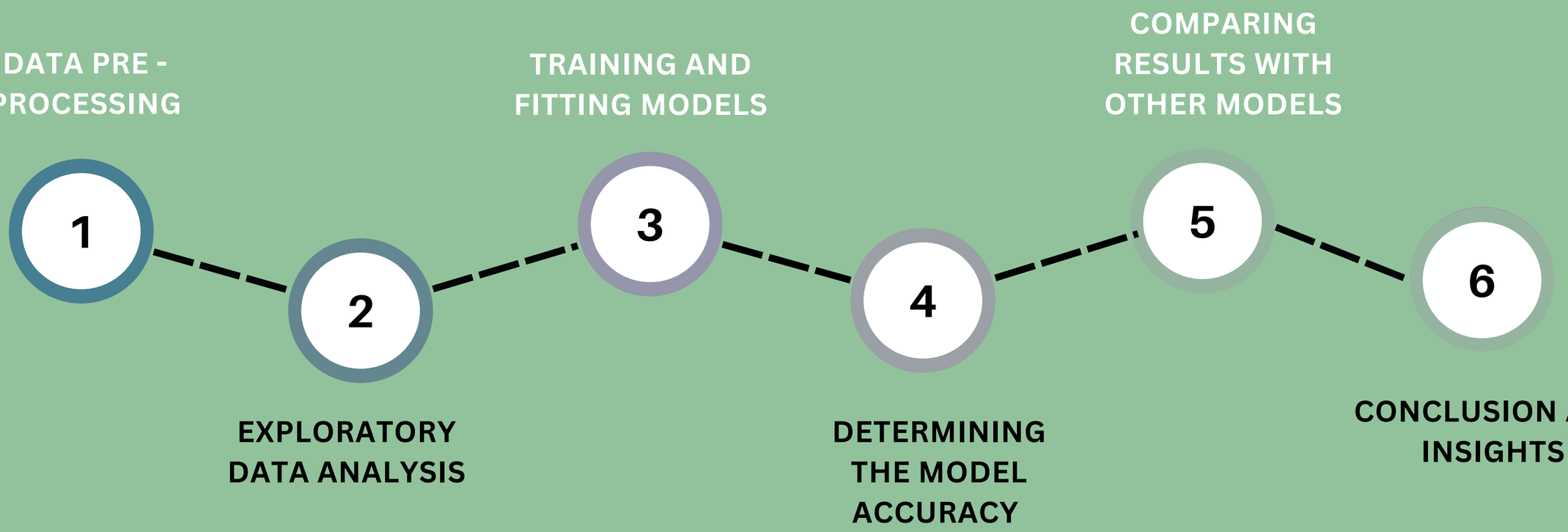
CLASSIFICATION PROJECT

## GROUP 12


- SHAIK ISHAQ
- SRI RAM
- SANDYA
- MANOGNA

BHAVANS VIVEKANANDA COLLEGE I BSC HONOS DATA SCIENCE

# MACHINE LEARNING ANALYSIS TIMELINE



# ABOUT THE DATA SET

- The data was aggregated from a number of sources including the salaries of  Employees. The data has been collected from recent graduates entering the workforce and includes a range of factors influencing salary levels.
- The term "engineering graduate salary" refers to the average or typical salary earned by individuals who have graduated with a degree in engineering.

# OBJECTIVE

- This project aims to develop a classification model that predicts salary ranges based on various features related to individuals' education, experience, skills, and other relevant factors.
- Developing an accurate and robust predictive model for engineering graduate salaries, utilizing a diverse set of machine learning algorithms.

# PATH

- Buliding multiple machine learning models to fit the best model for the dataset.

# DATA

| | |
|---|---|
| **ID** | • A unique identifier for each individual in the dataset. |
| **Gender** | • The gender of the individual (e.g., Male or Female). |
| **DOB** | • Date of Birth of the individual. |
| **10percentage** | • Percentage of marks obtained in the 10th grade |
| **12board** | • The educational board for the 10th grade examination. |
| **College ID** | • Unique identifier for the college. |
| **CollegeTier** | • The tier of the college (e.g., Tier 1 or Tier 2). |
| **Agreeableness** | • Personality trait indicating how cooperative and helpful an individual is. |
| **Extraversion** | • Personality trait indicating how outgoing and social an individual is. |

| | |
|---|---|
| **Degree** | • The type of degree pursued by the individual (e.g., B.Tech). |
| **Specialization** | • The field of specialization in the degree. |
| **collegeGPA** | • Grade Point Average (GPA) obtained in college. |
| **CollegeCityID** | • Unique identifier for the city where the college is located. |
| **CollegeCityTier** | • The tier of the city where the college is located. |
| **CollegeState** | • The state in which the college is located. |
| **Graduation year** | • Year of graduation from college. |
| **English** | • Score in the English section of an aptitude test |
| **CivilEngg** | • Gradutes from the streams of Civil Engg |
| **Conscientiousness** | • Personality trait indicating how responsible and organized an individual is. |

| | |
|---|---|
| **Logical** | • Score in the Logical section of an aptitude test. |
| **Quant** | • Score in the Quantitative section of an aptitude test. |
| **Domain** | • A measure of domain knowledge or skills. |
| **ComputerProgramming** | • Score in computer programming skills. |
| **ElectronicsAndSemicon** | • Score in electronics and semiconductor skills. |
| **Computer Science** | • people from stream of computer science |
| **MechanicalEngg** | • Gradutes from the fieldof Mechanical Engineering |
| **TelecomEngg** | • Gradutes from the field of Telecom Engg |
| **nueroticism** | • Personality trait indicating emotional stability. |
| **Openess_to_experience** | • These traits are considered broad dimensions of personality that encompass a range of more specific traits and behaviors |
| **Salary** | • Salaries can vary widely depending on factors such as the industry, job role, level of experience, and geographical location. |

# DATA QUALITY

## MISSING VALUES

- There were 3 Categorical variables , 34 total variables.There are no missing values and null values
- There are 33 input variables and 1 output variables

## DROPPED COLUMNS

- ID
- DOB
- 10board
- 12graduation
- 12board
- College ID
- College City ID
- CollegeCityTier
- College State
- Graduation Year
- Specialization
- Degree
- Salary

- The following columns are dropped from the data set due to their presences won’t effect the target variable ....

# DATA PREPROCESSING

- There are **2998 rows** and **34 columns**
- There are no null values
- we have dropped **13 columns**
- After removing the outliers, from **collegeGPA column** the new data has **2989 rows and 21 columns**

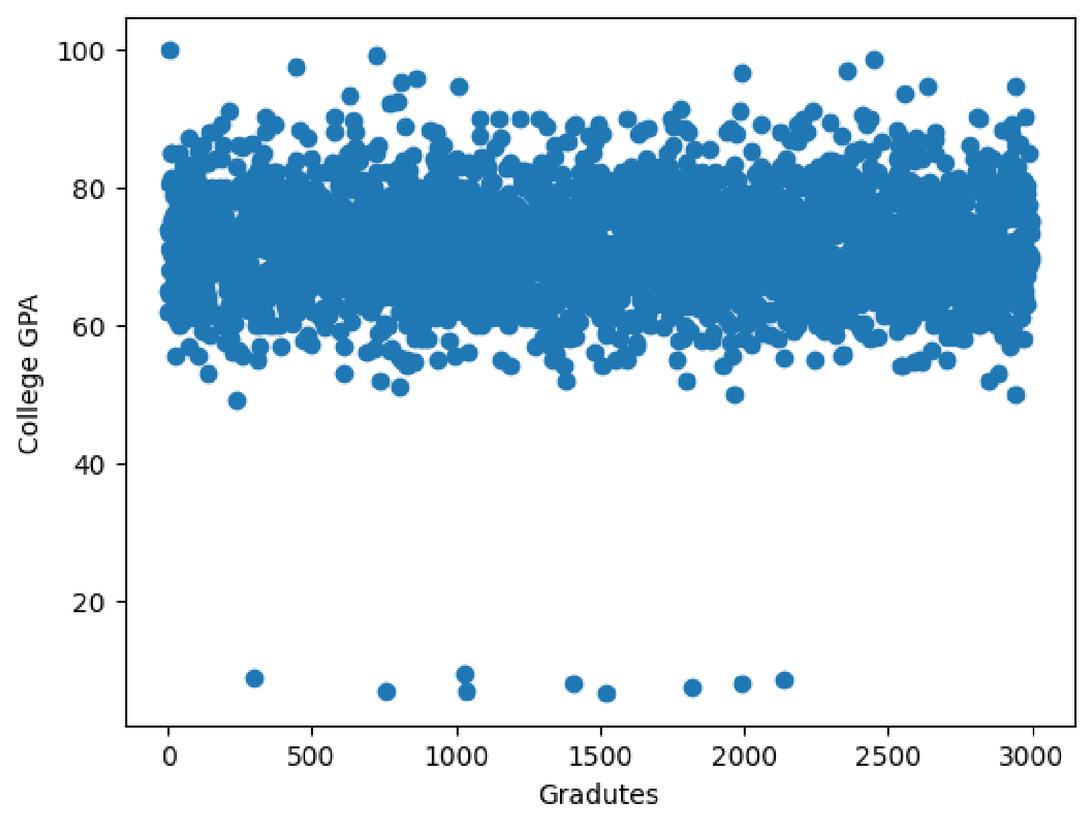
# DUMMY VARIABLE ENCODING

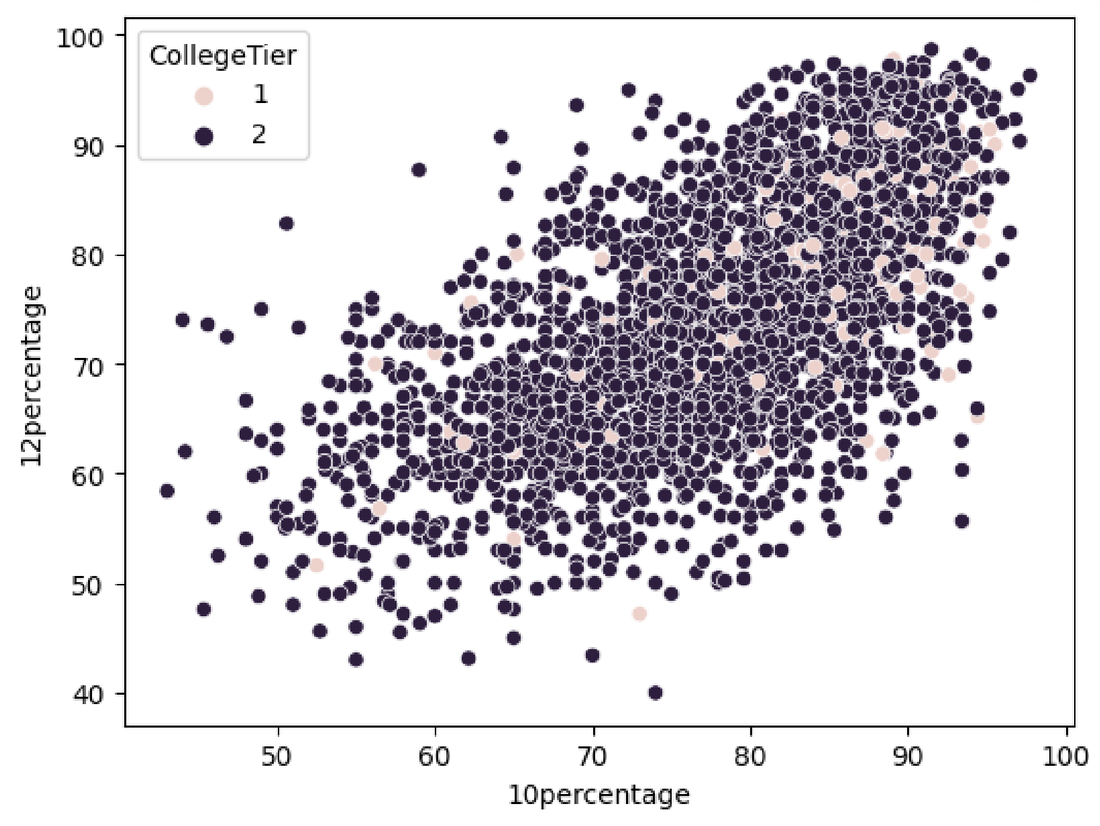
- The Gender column in the date set is being dropped down after performing the dummy variable encoding and the dummies which are classified in gender is noted as a Male and Female

- The Degree column in the date set is being dropped down after performing the dummy variable encoding and the dummies which are classified into respecteive degree fields

- The Salary column in the date set is being dropped down after categorizing it in to its respective Salary Range column

# EDA(EXPLORATORY DATA ANALYSIS)

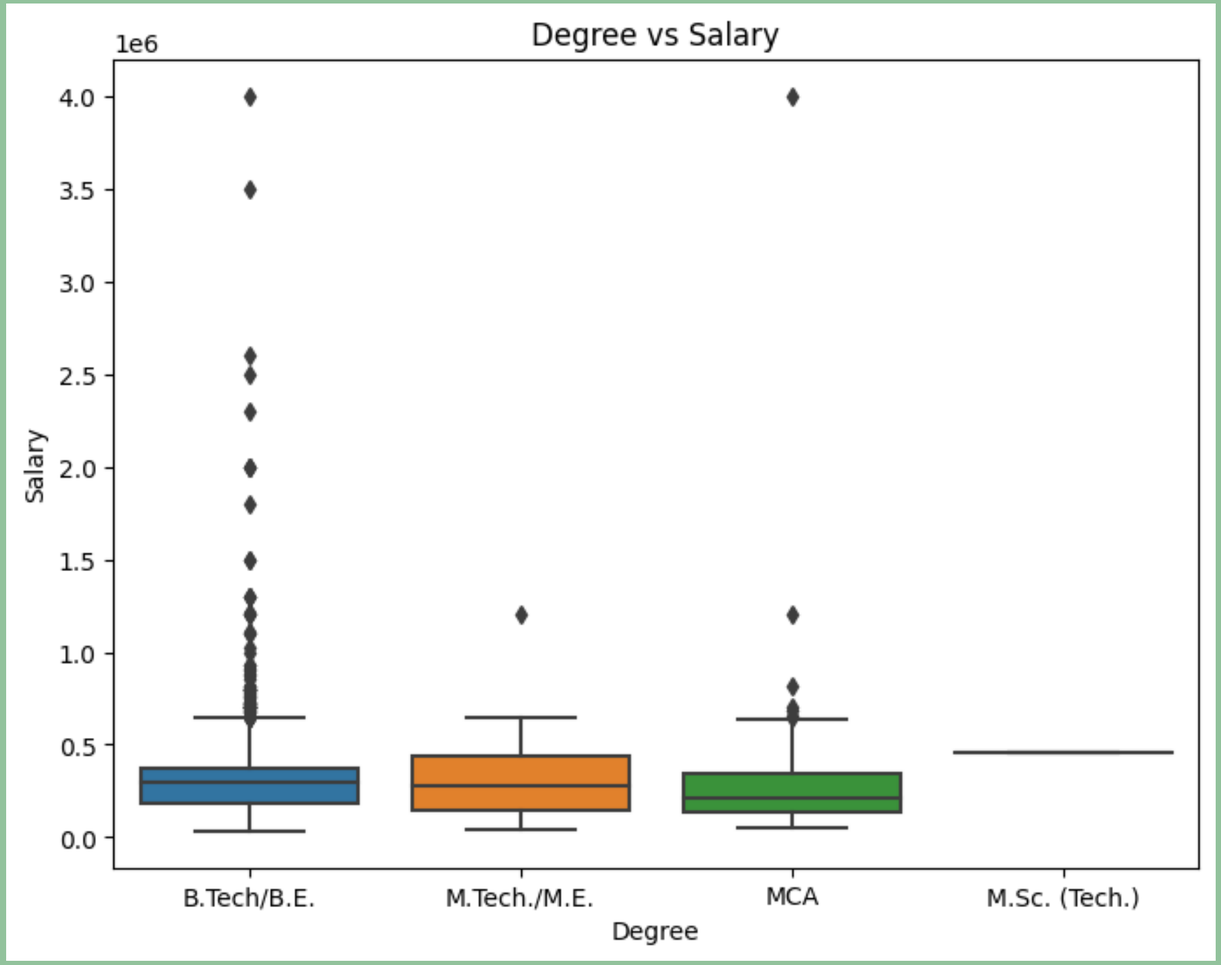


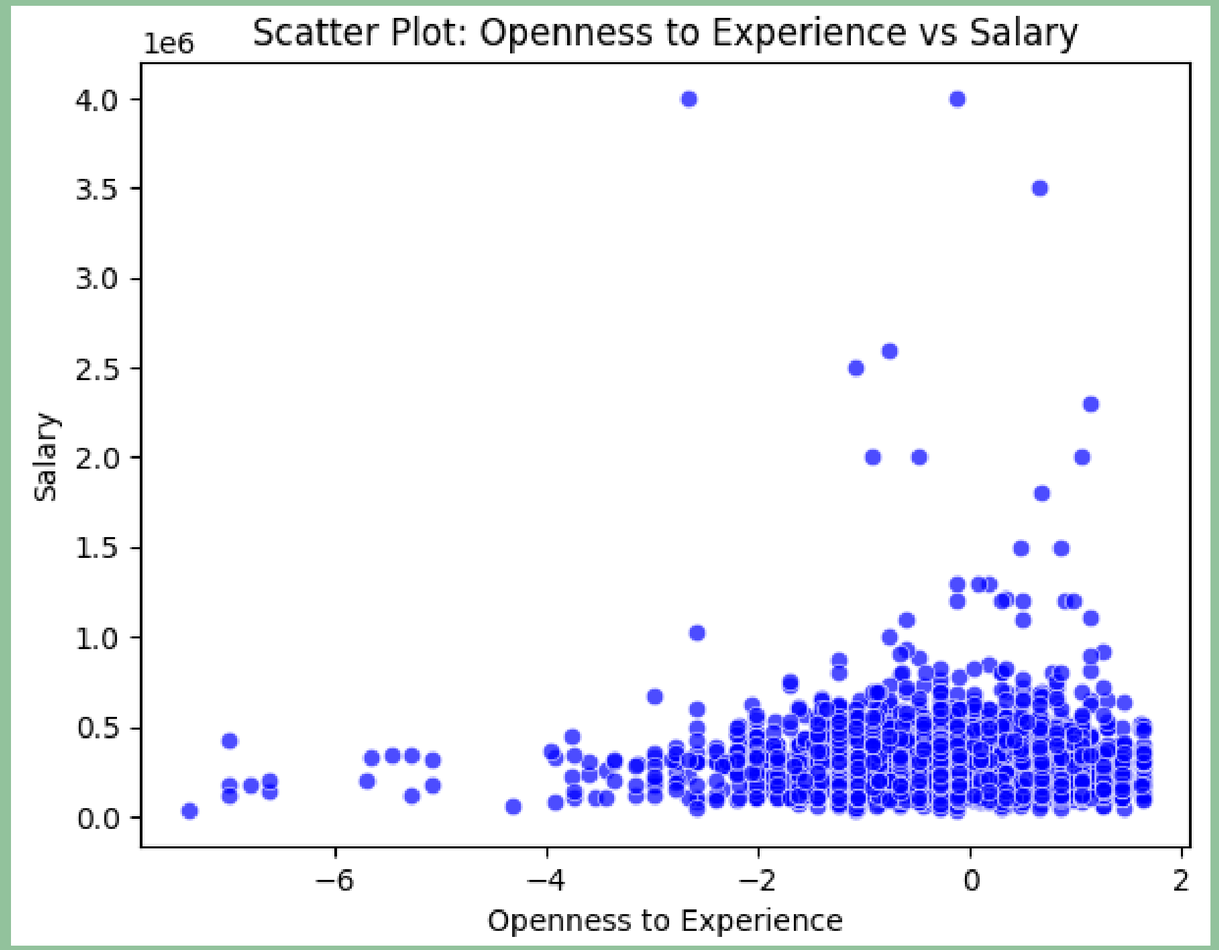




- The above scatter plot shows the College GPAs of the Engineering Gradutes

- The above scatter plot shows the ranking of the Collge Tier based on the 10,12percentages of Engineering Gradutes

- The Below Barplot shows the Salary of Engineering Gradutes with respect to their Degrees
- The below plot shows who has B.tech/B.E degrees that particular individuals has high Salaries

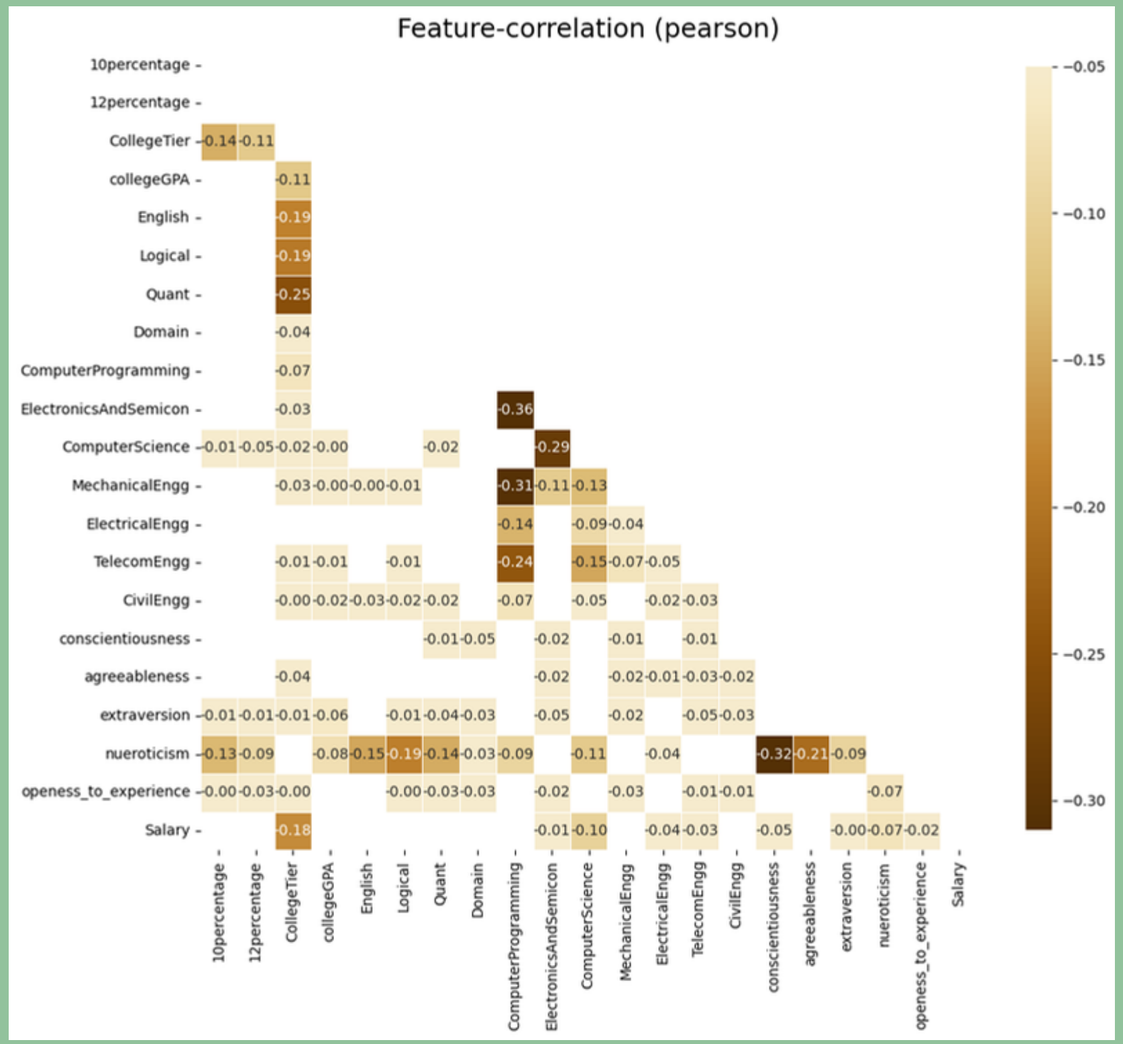


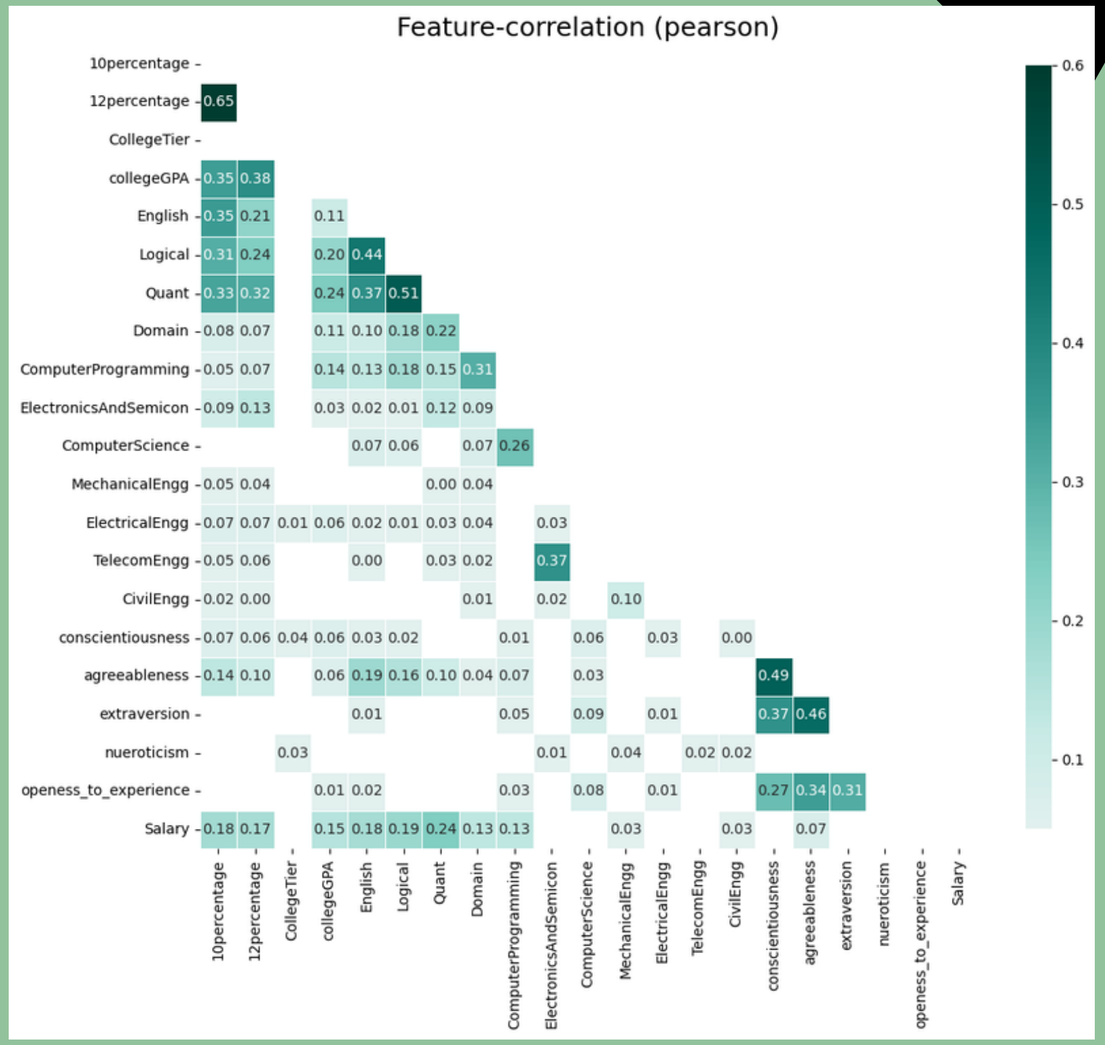

- The Above Scatterplot shows the Salary of Engineering Gradutes with respect to their openness to Experience
- Those who has the high Openness to Experience then particular individual has high Salary

# CORRELATION PLOT

- These are correlation plots shows how the variables are correlated to each other




- for example if u take **Computer scienceand** **Mechanical Engg** they are negatively correlatede




- for example if u take **12percentage** and **CollegeGPA** they are positively correlated

# VIF(Variance Inflation Factor)

- **Multicollinearity occurs when multiple independent variables in a multiple regression model have high correlation,making it difficult to distinguish individual effects on the dependent varable**

- **Typically, a VIF value above 10 is considered high, indicating potential multicollinearity. Based on this criterion, the variables with high VIF values are:**

- **10percentage (VIF = 113.768820)**
- **12percentage (VIF = 83.923687)**
- **collegeGPA (VIF = 75.480933)**

| Variable | VIF | |
|---|---|---|
| 0 | 10percentage | 113.768820 |
| 1 | 12percentage | 83.923687 |
| 2 | CollegeTier | 33.702093 |
| 3 | collegeGPA | 75.480933 |
| 4 | English | 32.540193 |
| 5 | Logical | 51.274446 |
| 6 | Quant | 29.588888 |
| 7 | Domain | 2.850794 |
| 8 | ComputerProgramming | 7.301853 |
| 9 | ElectronicsAndSemicon | 2.143264 |
| 10 | ComputerScience | 1.516794 |
| 11 | MechanicalEngg | 1.405631 |
| 12 | ElectricalEngg | 1.117559 |
| 13 | TelecomEngg | 1.323248 |
| 14 | CivilEngg | 1.025903 |
| 15 | conscientiousness | 1.544124 |
| 16 | agreeableness | 1.986241 |
| 17 | extraversion | 1.408017 |
| 18 | nueroticism | 1.217530 |
| 19 | openess_to_experience | 1.724807 |

# MODEL FITTING

- DECISION TREE CLASSIFIER
- NAIVE BAYIES CLASSIFIER
- K-NEAREST NEIGHBORS
- LOGISTIC REGRESSION
- NEURAL NETWORKS
- RANDOM FOREST CLASSIFIER

# DECISION TREE CLASSIFIER

| TRAIN-TEST RATIO | ACCURACY CALCULATION |
|---|---|
| 70 - 30 | **0.628** |
| 80- 20 | **0.59** |
| 65 - 35 | **0.6047** |
| 75 - 25 | **0.5866** |

- **The Highest Accuarcy recorded as 0.628**

# NAIVE BAYIES CLASSIFIER

| TRAIN-TEST RATIO | ACCURACY CALCULATION |
|---|---|
| 70 - 30 | 0.6466 |
| 80 - 20 | 0.6566 |
| 75 - 25 | 0.6666 |
| 65 - 35 | 0.6866 |

- The Highest accuracy of the model is recorrded as 0.686

# K-NEARST NEIGHBORS CLASSIFIER

| TRAIN-TEST RATIO | ACCURACY CALCULATION | N_NEIGHBORS |
|---|---|---|
| 70 - 30 | 0.6477 | 5 |
| 80 - 20 | 0.6533 | 5 |
| 75 - 25 | 0.6428 | 5 |
| 65 - 35 | 0.652 | 5 |

- The Highest accuracy of the model is recorded as 0.653

# LOGISTIC REGRESSION

| TRAIN-TEST RATIO | ACCURACY CALCULATION |
|---|---|
| 70 - 30 | 0.6366 |
| 80 - 20 | 0.6516 |
| 75 - 25 | 0.6413 |
| 65 - 35 | 0.6428 |

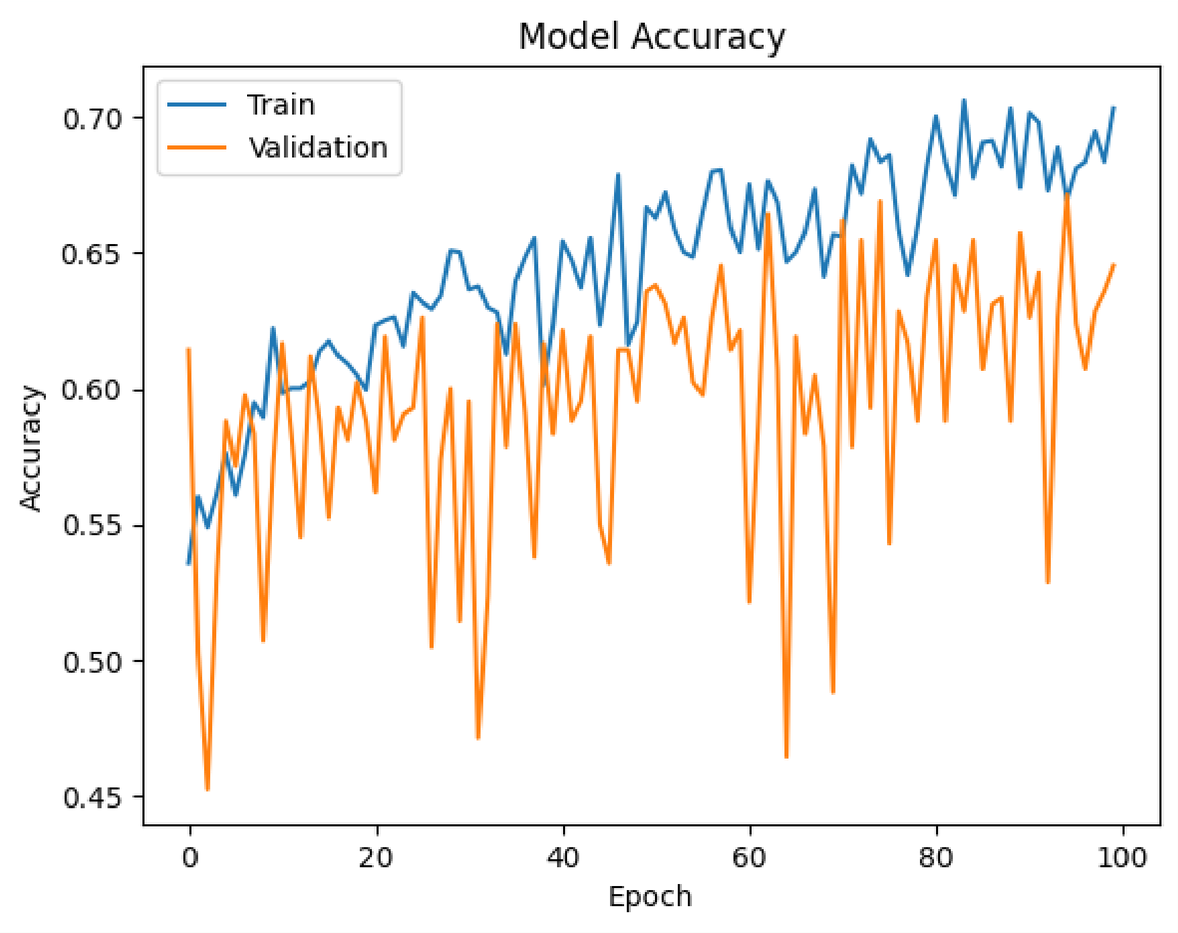
- The Highest accuracy of the model is recorrded as 0.65

# NEURAL NETWORKS ARCHITECTURE

| TRAIN-TEST RATIO | ARCHITECTURE | EPOCHS | OPTIMIZER | ACCURACY |
|---|---|---|---|---|
| 70 - 30 | 64 - 32 - 16-14-1 | 100 | adam | 0.6211 |
| 70 - 30 | 54 - 32 - 4 - 2- 1 | 350 | SGD | 0.554 |
| 75 - 25 | 62- 54 - 34 - 4 - 2 -1 | 300 | SGD | 0.549 |
| 75 - 25 | 45-37-23-14-7 | 500 | SGD | 0.549 |
| 80 - 20 | 64 - 32 - 16-14-1 | 200 | adam | 0.566 |
| 80 - 20 | 64-32-2-1 | 200 | adam | 0.576 |
| 65-35 | 28-14-12-6-1 | 200 | RMSprop | 0.548 |

- The Highest Accuracy is recorded as 0.6255

# NEURAL NETWORKS OBSERVATION PLOT




- **In the above plot neural network algorithm able to predict on trained  data**
- **certainly the validations and train  classes intersect at the accuracy value at around 100 Epochs which has 70-30 ratio**

# RANDOM FOREST ALGORITHM

| TRAIN-TEST RATIO | N_ESTIMATORS | ACCURACY |
|---|---|---|
| 70 - 30 | 100 | 0.685 |
| 80 - 20 | 150 | 0.678 |
| 75 - 25 | 200 | 0.692 |
| 65 - 35 | 350 | 0.7152 |

- The Highest Accuarcy recorded as 0.71

# COMPARISON OF IMPLEMTED MODELS

| ALGORITHMS | TEST - TRAIN RATIO | ACCURACY CALCULATION |
|---|---|---|
| DECISION TREE | **65-35** | **0.604** |
| NAVIE BAYIES | **65-35** | **0.686** |
| NEURAL NETWORKS | **65-35** | **0.548** |
| K -NEARST NEIGHBOR | **65-35** | **0.604** |
| LOGISTIC REGRESSION | **65-35** | **0.642** |
| **RANDOM FOREST TREE** | **65-35** | **0.715** |

# SUMMARY (OR) CONCLUSION

- For the Engineering Graduate Salary dataset, we performed six types of machine learning algorithms: Decision Tree ,Naive Bayies Classifier,Logistic Regression, Neural Network, Random Forest.

- From all the analysis and implementations of the algorithm, we found the best results in RANDOM FOREST algorithm i.e. for 65-35 train-test ratio with accuracy of 71% among all the other accuracy calaculations.

- So, here we can conclude that RANDOM FOREST Algorithm can be considered the best model for the given Engineering Graduate Salary dataset.

- The Higher Accuarcy represents that the model is performing good on data set

# INSIGHTS (OR) FUTURE IMPLEMETATION

- By performing the analysis we found to know that predictions of salary is used in such real world activities :

- Predictions of salary can be applied in various fields and contexts, providing valuable insights for individuals, organizations, and policymakers. Here are some applications of salary predictions:

- Recruitment and Hiring, Human Resources Management, Workforce Planning ,Education and Career Guidance, Economic Research, Financial Planning

# THANK YOU

**collab link:**


**PRESENTED BY**

SHAIK ISHAQ

SRI RAM

SANDHYA

MANOGNA

# APPENDIX

# TRAINING AND TESTING

Traning and Testing

```
[688] df = df.drop('Salary', axis=1)
```

```
[689] X = df.drop(['salary Range'], axis=1)
      y = df['salary Range']
```

```
X_train,X_test,y_train,y_test = train_test_split(X,y, test_size =0.3,random_state =42)
```

```
print(X_test.shape)
print(X_train.shape)
```

```
(900, 20)
(2098, 20)
```

# DECISION TREE CLASSIFIER

```
[735] from sklearn.tree import DecisionTreeClassifier
      from sklearn import metrics
      X_train,X_test,y_train,y_test = train_test_split(X,y, test_size =0.3,random_state =42)
      clf = DecisionTreeClassifier()
      clf = clf.fit(X_train,y_train)
      y_predict = clf.predict(X_test)
      print("the accuracy is :",metrics.accuracy_score(y_predict,y_test))

      the accuracy is : 0.6144444444444445
```

# NAIVE BAYIES CLASSIFIER

Naive bayes Classifier

```
[698] from sklearn.naive_bayes import GaussianNB
      gnb = GaussianNB()
      X_train,X_test,y_train,y_test = train_test_split(X,y, test_size =0.35,random_state = 34)
      gnb.fit(X_train,y_train)
      y_predict = gnb.predict(X_test)
      print("the accuracy is : ",metrics.accuracy_score(y_predict,y_test))

      the accuracy is :  0.6866666666666666
```

# K-NEARST NEIGHBORS CLASSIFIER

KNN CLASSIFIER

```
from sklearn.neighbors import KNeighborsClassifier
model = KNeighborsClassifier(n_neighbors=5)
X_train,X_test,y_train,y_test = train_test_split(X,y, test_size =0.25,random_state = 34)
model.fit(X_train,y_train)
y_predict = model.predict(X_test)
print("the accuracy is :",metrics.accuracy_score(y_predict,y_test))
```

```
the accuracy is : 0.652
```

# LOGISTIC REGRESSION

```
from sklearn.linear_model import LogisticRegression
X_train,X_test,y_train,y_test = train_test_split(X,y, test_size =0.35,random_state = 34)
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

# Make predictions on the test set
y_predict = logreg.predict(X_test)
print("the accuracy is :",metrics.accuracy_score(y_predict,y_test))
```

```
the accuracy is : 0.6428571428571429
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed to converg
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
 n_iter_i = _check_optimize_result(
```

# NEURAL NETWORKS

```python
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense

model = Sequential([
    Dense(28, activation='relu', input_shape=(train_data.shape[1],)),
    Dense(14, activation='relu'),
    Dense(12, activation='relu'),
    Dense(6, activation='relu'),
    Dense(target.shape[1], activation='sigmoid')  # Output layer with softmax for multi-class classificatio
])
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
history = model.fit(train_data, train_targets, epochs=1000, validation_split=0.2, verbose=2)
test_loss, test_accuracy = model.evaluate(test_data, test_targets)
print(f"Test Accuracy: {test_accuracy}")
```

```
Epoch 999/1000
60/60 - 0s - loss: 0.4475 - accuracy: 0.7878 - val_loss: 1.1666 - val_acc
Epoch 1000/1000
60/60 - 0s - loss: 0.4124 - accuracy: 0.7847 - val_loss: 1.1250 - val_acc
19/19 [==============================] - 0s 3ms/step - loss: 1.0968 - acc
Test Accuracy: 0.6116666793823242
```

# RANDOM FOREST ALGORITHM