# Analysis of graduate Student Academics and salary prediction using Big Data and Data Science

Team members

1. Divya Anusha Chandrupatla
2. Mani Venkata sai snehith Nadipally
3. Sannihitha kolipaka
4. Vamsi krishna chittyimadha

# Table of Contents

## Abstract

This project aims to predict the salaries of graduate students and explore the effect of different factors such as CGPA, board, specialization, location, and tier on their earnings using big data analytics tools. The primary motivation behind this project is to help students make informed decisions about their career paths by providing them with a fair estimation of their earning potential. Additionally, this project aims to assist recruiters in identifying high-performing students early in their academic careers to attract top talent and build relationships with them. The first increment of the project involved data preprocessing and analysis to remove null values, deal with missing data and outliers, and convert categorical variables into numerical ones. The project also included exploring the data to understand the distribution of features and their relationships with the target variable. This increment involved answering questions such as the effect of CGPA, board exams, specialization, location, and tier on salaries and which streams have the highest earning potential. The next increment of the project will involve building and evaluating machine learning models to accurately predict graduate salaries based on the available data. Overall, this project has significant implications for career planning, recruitment, salary negotiations, and guiding academic program development.

## Goals and Objectives:

The goals and objectives of this project are to predict the salaries of graduate students and analyze the effect of different factors such as grade, board, location, and tier on the results. The project aims to achieve this by employing big data analytics tools to explore and preprocess the data, then choosing a suitable machine learning model that can accurately predict graduate salaries based on the available data.

### Motivation:

The motivation behind this project is the fact that many students after graduation often struggle to find better job placements and higher salaries. This is because they lack knowledge about the practical scenario of the job market and their earning potential. Additionally, recruiters often face challenges identifying high-performing students beforehand. Therefore, this project aims to increase the earning potential of graduate students, help recruiters make better decisions on high-performing students, and enable students to negotiate fair salaries during the hiring process based on their education and qualifications. Furthermore, predicting graduate student salaries can also guide the development of academic programs by providing insights into which programs consistently produce graduates with high salaries.

### Significance:

The significance of this project lies in its potential to increase the earning potential of graduate students and facilitate fair salary negotiations during the hiring process. Additionally, it can help recruiters identify high-performing students and build relationships with them early in their careers. The predicted salaries can also aid students in making informed career decisions and guide the development of academic programs based on the success of certain programs in producing graduates with high salaries.

### Objectives:

The main objective of this project is to predict the salaries of graduate students based on various factors such as CGPA, board exams, specialization, stream, location, and tier. Additionally, the project aims to explore the relationships between these factors and graduate salaries and identify any trends or patterns.

Ultimately, the goal is to choose a suitable machine learning model that can accurately predict graduate salaries based on the available data.

### Features:

The project involves several features, including data cleaning and preprocessing, data analysis, and model building and evaluation. The data will be cleaned to remove null and inappropriate values, deal with missing values, and convert categorical variables into numerical ones. The data will be explored to understand the distribution of the features and their relationship with the target variable, graduate salaries. Various machine learning models will be evaluated to determine the most suitable model for accurately predicting graduate salaries.

## Related work

Several studies have been conducted on salary prediction and analysis using data science techniques. One such study by [1] utilized machine learning algorithms to predict salaries based on various factors such as education, job type, experience, and location. The study achieved high accuracy in predicting salaries for different job categories.

Another study by [2] used big data analytics to analyze the impact of education level, experience, and location on salary in China. The study found that education level and experience had a significant impact on salary, while the impact of location was relatively small.

In the field of education, a study by [3] analyzed the impact of academic performance on salary for college graduates in Taiwan. The study found that academic performance, as measured by GPA, had a positive impact on salary.

Similarly, a study by [4] examined the relationship between academic performance, major, and salary for graduates in Taiwan. The study found that major had a significant impact on salary, while academic performance had a relatively small impact.

Overall, these studies suggest that data science and big data analytics can be effective in predicting and analyzing salaries based on various factors, including education, experience, location, and academic performance.

## About Dataset

The dataset used in this study contains various attributes of engineering graduates in India, including their ID, annual CTC offered, gender, date of birth, overall marks obtained in grade 10 and 12 examinations, school board, year of graduation, college attended, college tier, degree obtained, specialization pursued, aggregate GPA at graduation, city and state of the college, and scores in various sections of AMCAT's job portal, including English, logical ability, quantitative ability, domain module, computer programming, electronics and semiconductor engineering, computer science, mechanical engineering, electrical engineering, telecommunications engineering, civil engineering, and personality test sections.

## Detail design of features

The detailed design of features involves determining which factors will be included in the analysis and how they will be represented. For this project, we plan to include features such as CGPA, board exams, specialization, location, and tier. These features will be transformed into numerical values using techniques such as one-hot encoding or label encoding.

Additionally, we will explore feature engineering techniques to create new features that may provide additional predictive power. For example, we may create a feature that represents the average salary of graduates from the same university or a feature that represents the percentage of students from a particular specialization who receive high salaries.

The design of the features will be informed by the results of the data analysis and domain knowledge of the factors that influence graduate salaries. We will also consider the potential impact of each feature on the accuracy of the predictive model.

Overall, the detailed design of features is a crucial step in developing an accurate and robust predictive model for graduate salaries.

## Data Cleansing and Preparation

The data has been cleansed by removing all the 'Na' values from the entire dataset. Irrelevant columns have been removed and the categorical values have been standardized using different types of schemes. Empty rows have also been removed to make the data more consistent. The relevant columns that could help us analyze the dataset for the graduate salary analysis and prediction have been selected. Overall, the data has been cleaned and processed to ensure that it is accurate and suitable for further analysis.

## Descriptive Analytics

The dataset provided contains information on 2998 individuals. The dataset consists of the following columns: ID, Gender, 10percentage, 10board, 12graduation, 12percentage, 12board, CollegeID, CollegeTier, Degree, Specialization, collegeGPA, CollegeCityID, CollegeCityTier, CollegeState, GraduationYear, English, Logical, Quant, Domain, ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg, conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience, and Salary.

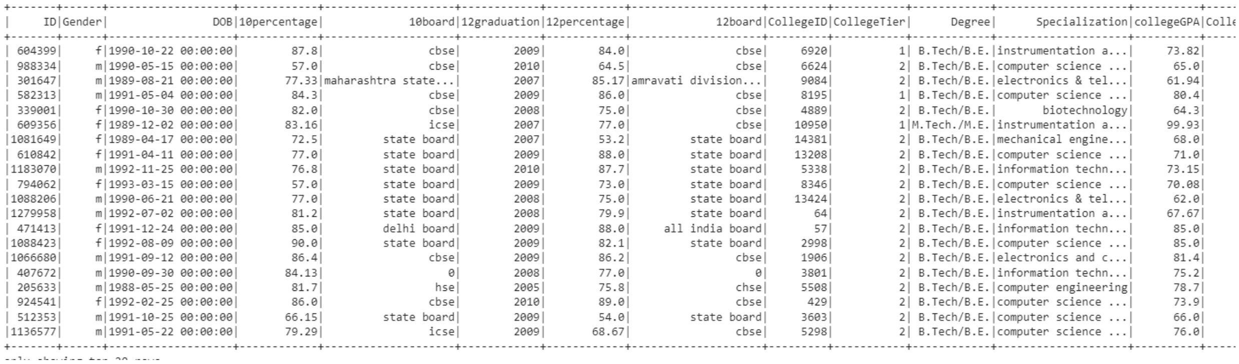
The mean value for the 10percentage column is 77.67, with a standard deviation of 10.00. The mean value for the 12percentage column is 74.34, with a standard deviation of 11.12. The mean value for the college GPA is 71.51, with a standard deviation of 8.12. The mean values for the test scores in English, Logical, and Quantitative abilities are 501.07, 500.43, and 514.14, respectively.

The dataset contains 2998 unique IDs with no missing values in any of the columns. The Gender column has no mean value since it is a categorical variable with two possible values: "f" for female and "m" for male.

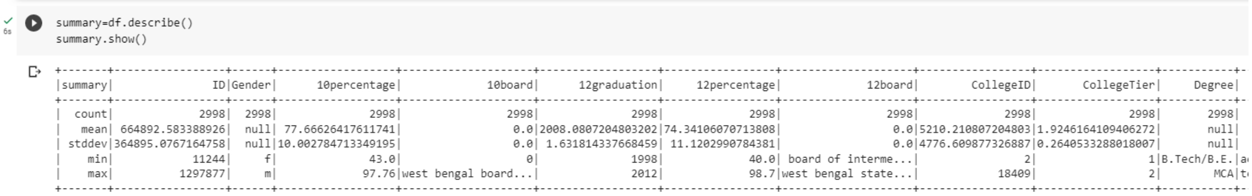
The dataset has 2998 unique CollegeIDs and 1492 unique CollegeCityIDs. The CollegeTier column is a binary categorical variable with two possible values: 1 for Tier 1 colleges and 2 for Tier 2 colleges.

The dataset contains individuals with different degrees and specializations, including B.Tech/B.E, M.Sc., M.Tech./M.E., B.Sc., MCA, and MBA. The Specialization column contains the area of specialization for each individual.

The dataset also includes information on the domain scores, computer programming skills, and electronics and semiconductor skills of individuals. Additionally, the dataset contains information on the individual's level of conscientiousness, agreeableness, extraversion, neuroticism, and openness to experience. Finally, the dataset includes the annual salary of each individual.



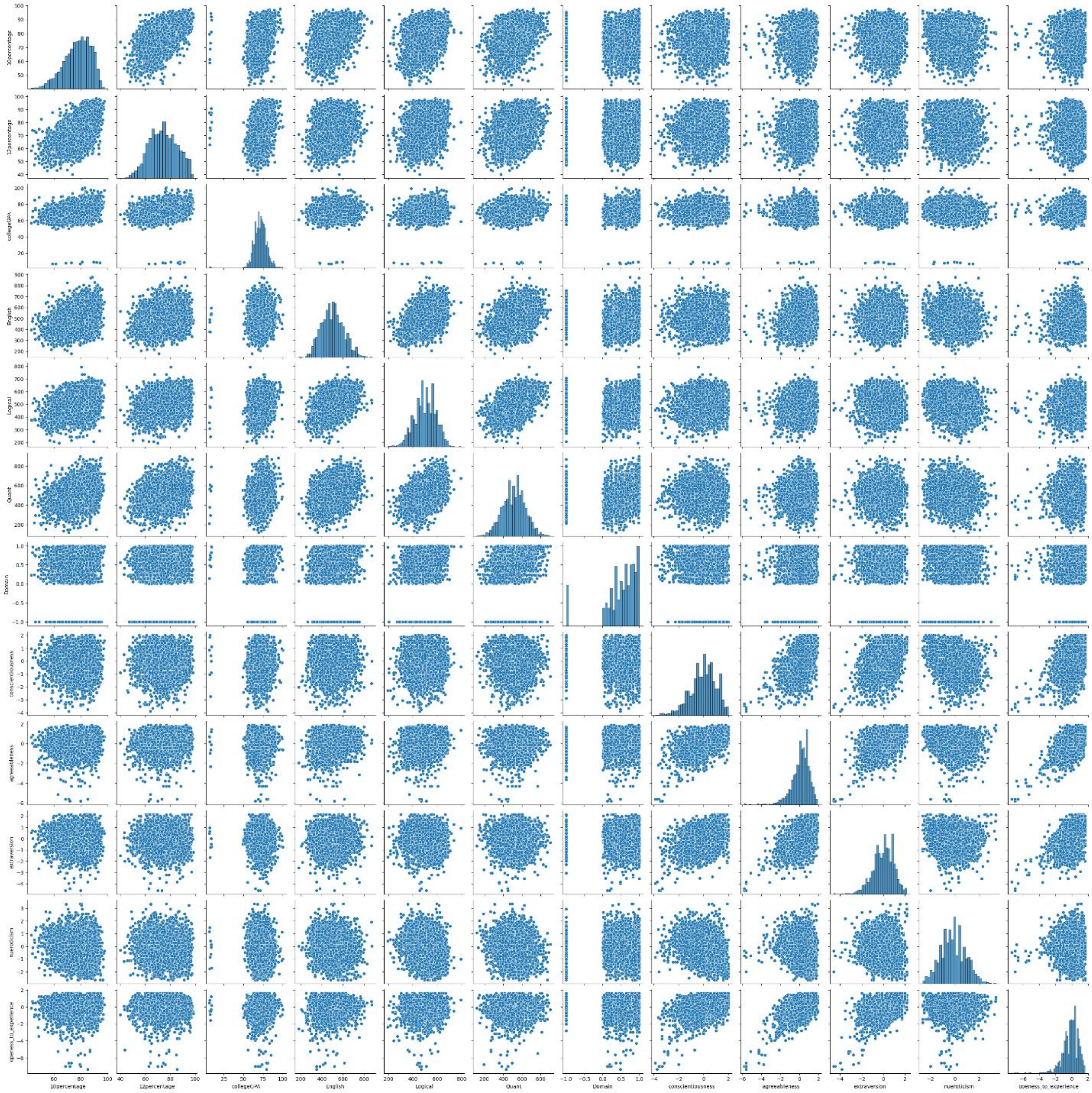
```
+-------+------+-------------------+------------+--------------------+------------+------------+--------------------+---------+-----------+------------+--------------------+----------+-----
|     ID|Gender|                DOB|10percentage|             10board|12graduation|12percentage|             12board|CollegeID|CollegeTier|      Degree|      Specialization|collegeGPA|Colle
+-------+------+-------------------+------------+--------------------+------------+------------+--------------------+---------+-----------+------------+--------------------+----------+-----
| 604399|     f|1990-10-22 00:00:00|        87.8|                cbse|        2009|        84.0|                cbse|     6920|          1| B.Tech/B.E.|instrumentation a...|     73.82|
| 988334|     m|1990-05-15 00:00:00|        57.0|                cbse|        2010|        64.5|                cbse|     6624|          2| B.Tech/B.E.|computer science ...|      65.0|
| 301647|     m|1989-08-21 00:00:00|       77.33|maharashtra state...|        2007|       85.17|amravati division...|     9084|          2| B.Tech/B.E.|electronics & tel...|     61.94|
| 582313|     m|1991-05-04 00:00:00|        84.3|                cbse|        2009|        86.0|                cbse|     8195|          1| B.Tech/B.E.|computer science ...|      80.4|
| 339001|     f|1990-10-30 00:00:00|        82.0|                cbse|        2008|        75.0|                cbse|     4889|          2| B.Tech/B.E.|       biotechnology|      64.3|
| 609356|     f|1989-12-02 00:00:00|       83.16|                icse|        2007|        77.0|                cbse|    10950|          1|M.Tech./M.E.|instrumentation a...|     99.93|
|1081649|     f|1989-04-17 00:00:00|        72.5|         state board|        2007|        53.2|         state board|    14381|          2| B.Tech/B.E.|mechanical engine...|      68.0|
| 610842|     f|1991-04-11 00:00:00|        77.0|         state board|        2009|        88.0|         state board|    13208|          2| B.Tech/B.E.|computer science ...|      71.0|
|1183070|     m|1992-11-25 00:00:00|        76.8|         state board|        2010|        87.7|         state board|     5338|          2| B.Tech/B.E.|information techn...|     73.15|
| 794062|     f|1993-03-15 00:00:00|        57.0|         state board|        2009|        73.0|         state board|     8346|          2| B.Tech/B.E.|computer science ...|     70.08|
|1088206|     m|1990-06-21 00:00:00|        77.0|         state board|        2008|        75.0|         state board|    13424|          2| B.Tech/B.E.|electronics & tel...|      62.0|
|1279958|     m|1992-07-02 00:00:00|        81.2|         state board|        2008|        79.9|         state board|       64|          2| B.Tech/B.E.|instrumentation a...|     67.67|
| 471413|     f|1991-12-24 00:00:00|        85.0|         delhi board|        2009|        88.0|     all india board|       57|          2| B.Tech/B.E.|information techn...|      85.0|
|1088423|     f|1992-08-09 00:00:00|        90.0|         state board|        2009|        82.1|         state board|     2998|          2| B.Tech/B.E.|computer science ...|      85.0|
|1066680|     m|1991-09-12 00:00:00|        86.4|                cbse|        2009|        86.2|                cbse|     1906|          2| B.Tech/B.E.|electronics and c...|      81.4|
| 407672|     m|1990-09-30 00:00:00|       84.13|                   0|        2008|        77.0|                   0|     3801|          2| B.Tech/B.E.|information techn...|      75.2|
| 205633|     m|1988-05-25 00:00:00|        81.7|                 hse|        2005|        75.8|                chse|     5508|          2| B.Tech/B.E.|computer engineering|      78.7|
| 924541|     f|1992-02-25 00:00:00|        86.0|                cbse|        2010|        89.0|                cbse|      429|          2| B.Tech/B.E.|computer science ...|      73.9|
| 512353|     m|1991-10-25 00:00:00|       66.15|         state board|        2009|        54.0|         state board|     3603|          2| B.Tech/B.E.|computer science ...|      66.0|
|1136577|     m|1991-05-22 00:00:00|       79.29|                icse|        2009|       68.67|                cbse|     5298|          2| B.Tech/B.E.|computer science ...|      76.0|
+-------+------+-------------------+------------+--------------------+------------+------------+--------------------+---------+-----------+------------+--------------------+----------+-----
only showing top 20 rows
```



```
summary=df.describe()
summary.show()
```

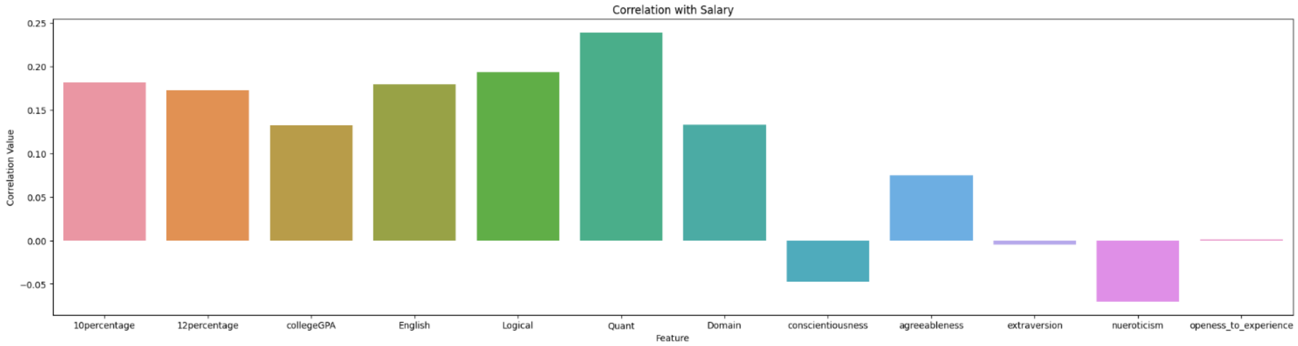
```
+-------+-----------------+------+------------------+--------------------+------------------+-----------------+--------------------+-----------------+------------------+-----------+-
|summary|               ID|Gender|      10percentage|             10board|      12graduation|     12percentage|             12board|        CollegeID|       CollegeTier|     Degree|
+-------+-----------------+------+------------------+--------------------+------------------+-----------------+--------------------+-----------------+------------------+-----------+-
|  count|             2998|  2998|              2998|                2998|              2998|             2998|                2998|             2998|              2998|       2998|
|   mean| 664892.583388926|  null| 77.66626417611741|                 0.0|2008.0807204803202|74.34106070713808|                 0.0|5210.210807204803|1.9246164109406272|       null|
| stddev|364895.0767164758|  null|10.002784713349195|                 0.0| 1.631814337668459| 11.1202990784381|                 0.0|4776.609877326887|0.2640533288018007|       null|
|    min|            11244|     f|              43.0|                   0|              1998|             40.0| board of interme...|                2|                 1|B.Tech/B.E.|a
|    max|          1297877|     m|             97.76|west bengal board...|              2012|             98.7|west bengal state...|            18409|                 2|        MCA|t
+-------+-----------------+------+------------------+--------------------+------------------+-----------------+--------------------+-----------------+------------------+-----------+-
```

The present study employs a histogram and scatter plot matrix to showcase the data distributions and their association with the dependent variable, namely salary. It is noteworthy that the majority of the variables exhibit a normal distribution pattern, while a few are categorical. These visual representations serve as important tools for analyzing and interpreting the data, thereby contributing to the overall rigor and validity of the research findings



## Correlation analysis:

We conducted a correlation analysis and found that quantitative skills, logical reasoning, and domain knowledge are positively correlated with the dependent variable salary. On the other hand, conscientiousness and neuroticism showed negative correlations with salary. Furthermore, extraversion and neuroticism did not show any correlation with salary.



## Analysis Using Big Data Analysis Queries

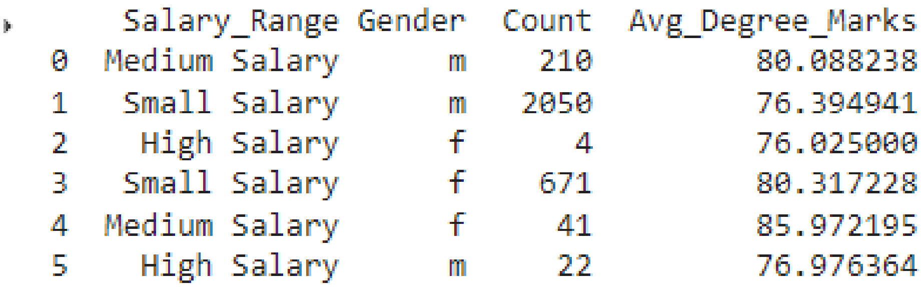
Different type of bigdata queries were utilized that to know the relationships between feature variables.

The PySpark queries are follows:

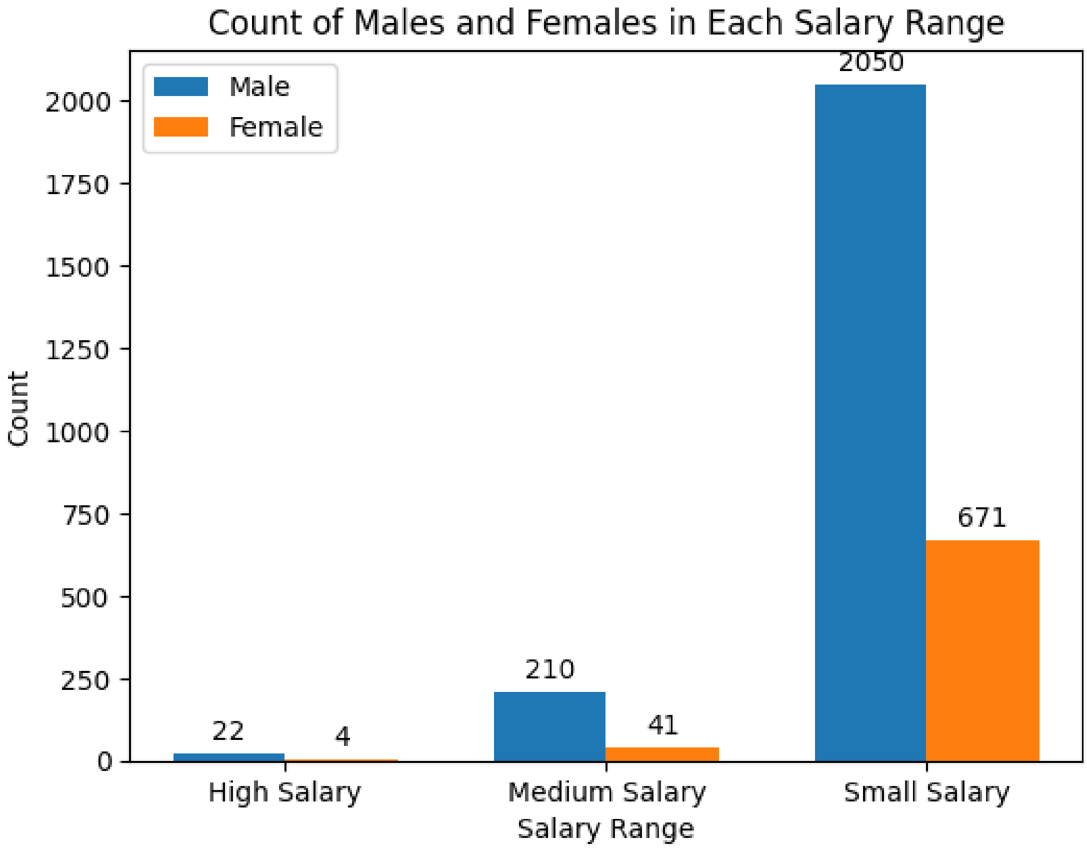
### Query to Group and Calculate Statistics for Salary and Gender in a Pandas Data Frame.

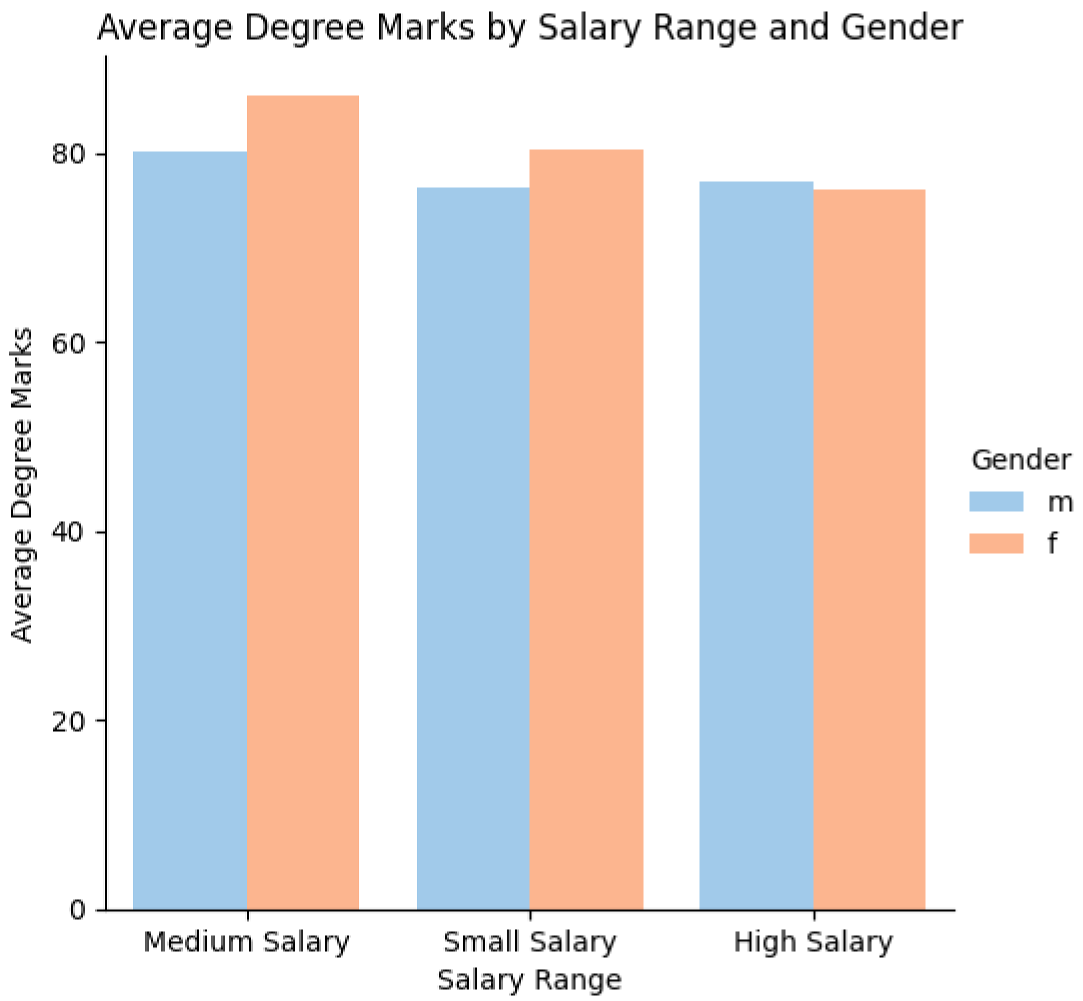
This query is grouping a Data Frame by salary range and gender, and then calculating the count of rows and average degree marks for each group. The resulting Data Frame is then converted to a Pandas Data Frame and returned as the output of the query.

```
df.groupBy(when(df.Salary < 500000, 'Small Salary').when(df.Salary < 1000000, 'Medium Salary').otherwise('High Salary').alias('Salary_Range'),'Gender').agg(count('*').alias('Count'),avg('10percentage').alias('Avg_Degree_Marks')).toPandas()
```



```
     Salary_Range Gender  Count  Avg_Degree_Marks
0   Medium Salary      m    210         80.088238
1    Small Salary      m   2050         76.394941
2     High Salary      f      4         76.025000
3    Small Salary      f    671         80.317228
4   Medium Salary      f     41         85.972195
5     High Salary      m     22         76.976364
```





Interpretations:
Based on the information in the table chart, we can make several observations and analyses:

1. The 'Small Salary' range has the highest count of data points, with 2,050 data points for males and 671 data points for females. This suggests that a larger number of people in the dataset have salaries in the 'Small Salary' range.
2. The 'Medium Salary' range has a comparatively lower count of data points, with 210 data points for males and 41 data points for females. This suggests that fewer people in the dataset have salaries in the 'Medium Salary' range.
3. The 'High Salary' range has the lowest count of data points, with only 22 data points for males and 4 data points for females. This suggests that very few people in the dataset have salaries in the 'High Salary' range.
4. The average degree marks for each salary range and gender combination provide insight into the relationship between education and salary. The 'Medium Salary' range has the highest average degree marks for both males (80.088238) and females (85.972195), suggesting that people with higher degrees are more likely to earn salaries in the 'Medium Salary' range.

5. The 'Small Salary' range has a lower average degree marks compared to the 'Medium Salary' range, indicating that people with lower degrees are more likely to earn salaries in the 'Small Salary' range. However, females in the 'Small Salary' range have a higher average degree marks (80.317228) compared to males (76.394941).
6. The 'High Salary' range has a relatively low average degree marks for both males (76.976364) and females (76.025000), suggesting that higher degrees may not necessarily guarantee a high salary in this range. However, the small sample size for this range could skew the results.

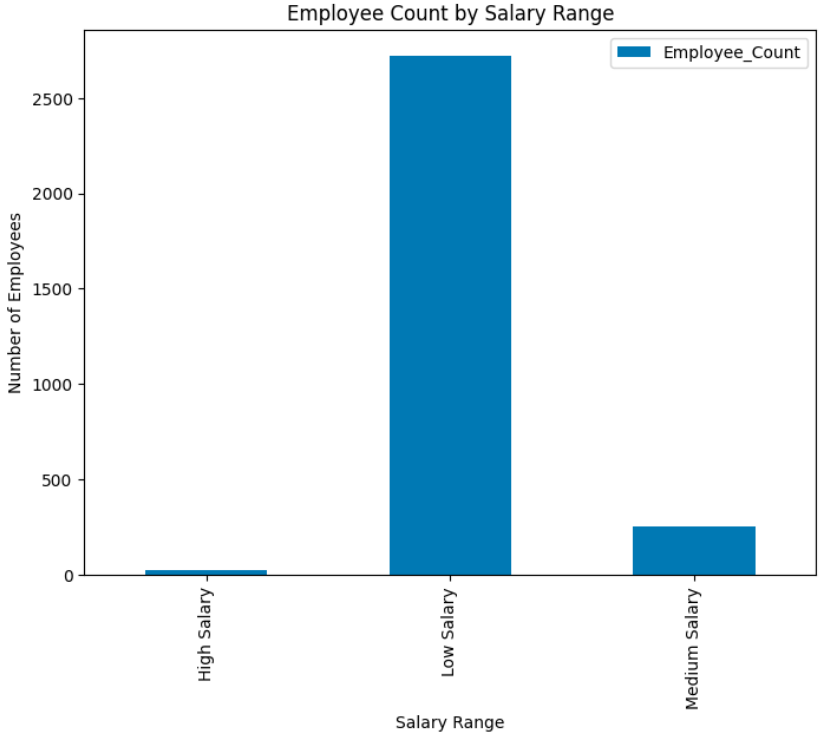
## PySpark BigData query to analyze the dataset for checking the correlations of average college percentage with salary range.

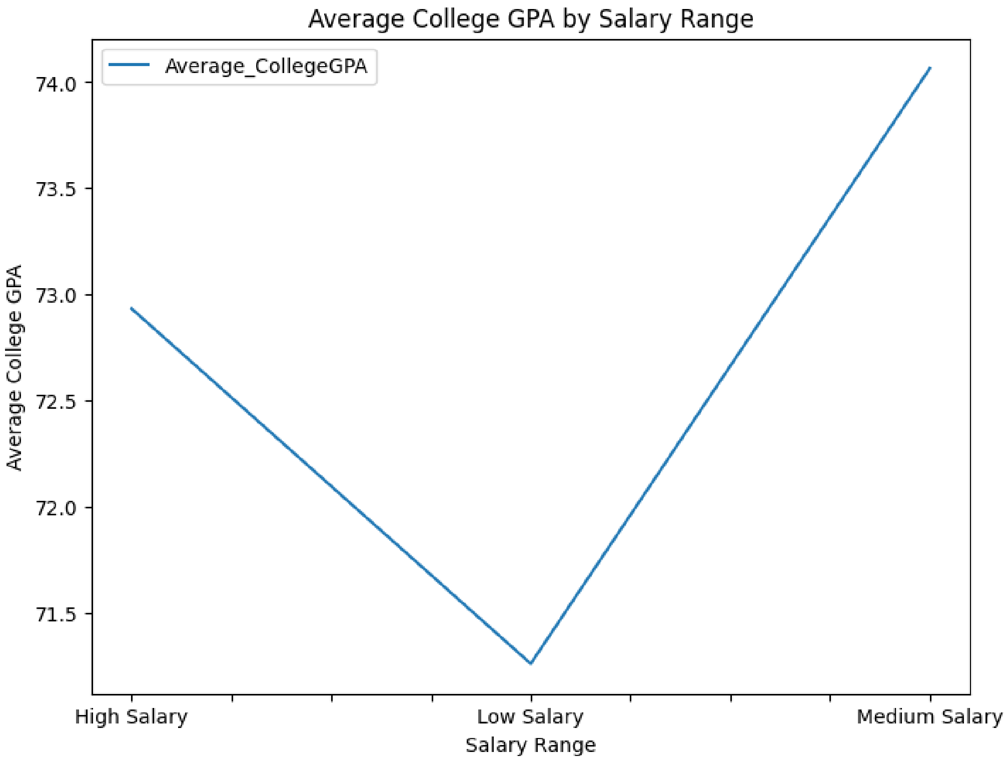
The focus of the analysis was to check the correlation between the average college percentage of the graduate students and their corresponding salary range. This analysis will help in understanding how the academic performance of the students affects their earning potential.

Query:

```
from pyspark.sql.functions import when, count, avg

df.groupBy(when(df.Salary < 500000, 'Low Salary').when((df.Salary >= 500000) & (df.Salary < 1000000),
'Medium Salary'). otherwise('High Salary').alias('Salary_Range')).agg(count('*').alias('Employee_Count'),
avg('collegeGPA').alias('Average_CollegeGPA'))
```





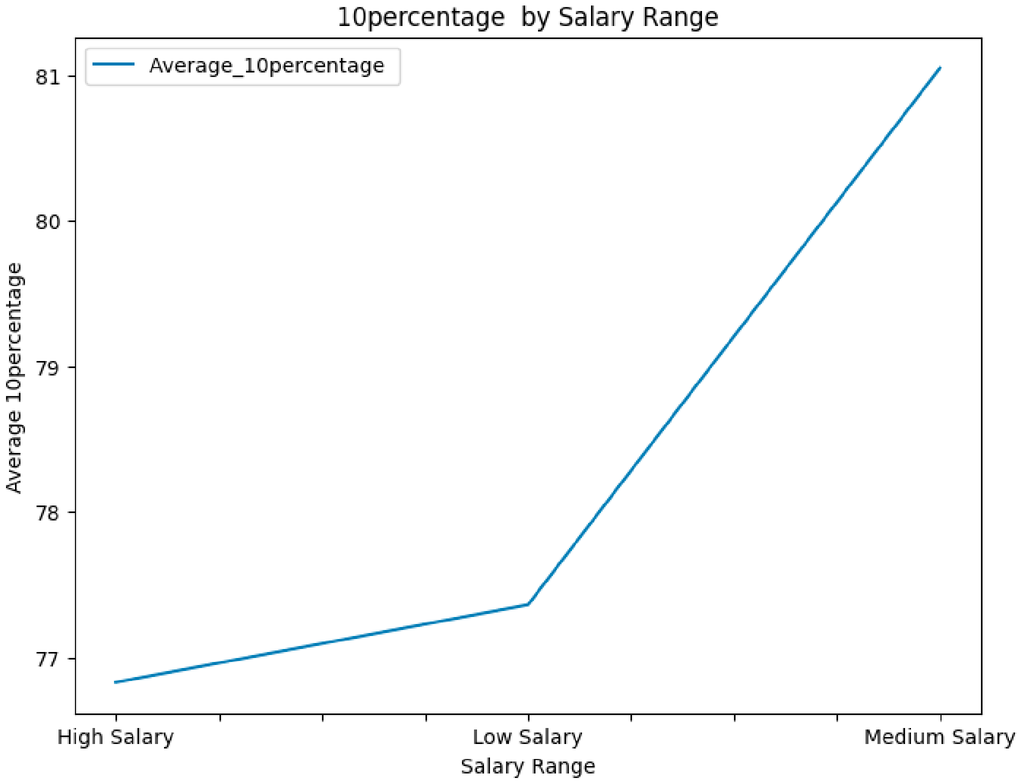
The above charts and graphs result from query represents that we have the highest number of employees in the small salary range pay package. The above line chart show that less average college gpa has the lowest salary, but this not the final results we will analyze it more because there are so many factors that cause the low salary value for the employee.

## Query to analyze the relationships between 10the class percentage with respect to salary range

This PySpark query groups and aggregates employee data based on their salary range. The 'Salary_Range' column is created using the 'when' function and the 'Salary' column. The data is then grouped using the 'groupBy' function and aggregated using the 'agg' function to count the number of employees and calculate the average of the '10percentage' column for each salary range. The resulting columns are renamed using the 'alias' function. The query provides insights into the distribution of employees by salary range and their average educational qualifications.

Query:

```
df.groupBy(
    when(df.Salary < 500000, 'Low Salary')
    .when((df.Salary >= 500000) & (df.Salary < 1000000), 'Medium Salary')
    .otherwise('High Salary')
    .alias('Salary_Range')
).agg(
    count('*').alias('Employee_Count'),
    avg('10percentage').alias('Average_10percentage ')
)
```



Based on the line chart generated from the query results above, it is evident that the average percentage of 10th class holds the highest percentage in the medium salary range, but indicates a lower average percentage in the high salary range.

# Project Management

## Implementation Status Report

### Dataflow Diagram

In the level 1 DFD, the user interacts with the system through the main screen. The data processing unit performs data cleaning and preprocessing on the input data, which is then fed into the prediction algorithm. The prediction algorithm uses various machine learning models to predict graduate salaries based on the available data. The results of the prediction algorithm are then displayed to the user through the report. The user can provide feedback on the accuracy of the results through the feedback module, which can be used to improve the prediction algorithm in the future. The database stores all the relevant data for the system.

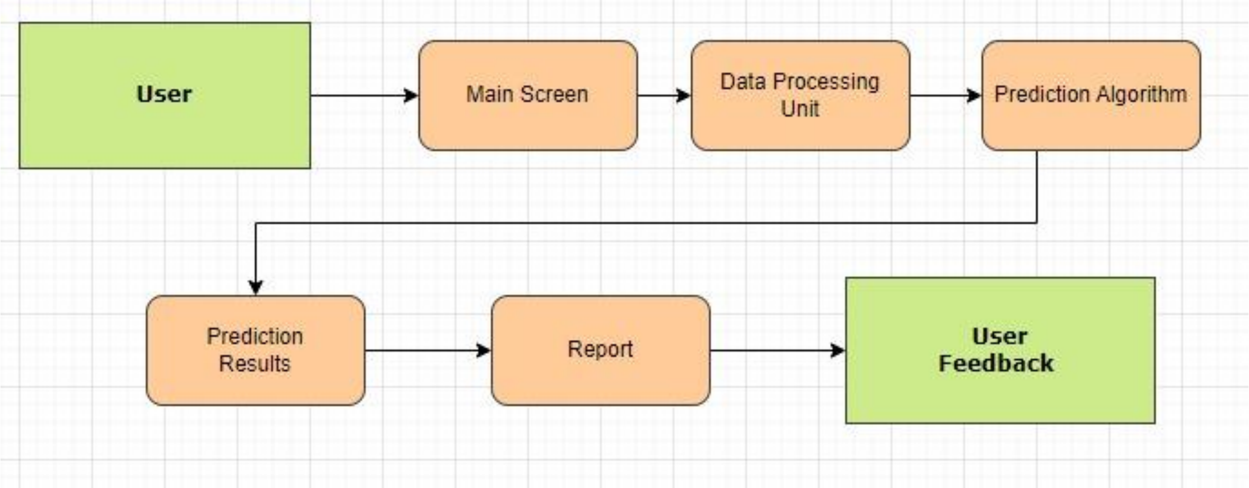


*Figure 1 Data Flow Diagram*

## Algorithm

**Step 1:** Collect the dataset containing information about graduate students' salaries and other relevant factors such as CGPA, board exams, specialization, location, and tier.

In this step the data were collected from Kaggle.

**Step 2:** Preprocess the data by removing null values, dealing with missing data and outliers, and converting categorical variables into numerical ones.

Preprocessing and Data cleansing play an important role in the analysis and prediction of data and machine learning algorithms. We have removed the null values from the data standardized the data by converting the categorical variables into machine learning model understandable form. It avoids biased results and improve the model's prediction accuracy.

**Step 3:** Analyze the data to understand the distribution of features and their relationships with the target variable, graduate salaries. Answer questions such as the effect of CGPA, board exams, specialization, location, and tier on salaries, and which streams have the highest earning potential.

Different type data exploration and analysis done in increment 1, we have used summary statistics, correlation analysis and visualized the data.

**Step 4:** Split the dataset into training and testing sets.

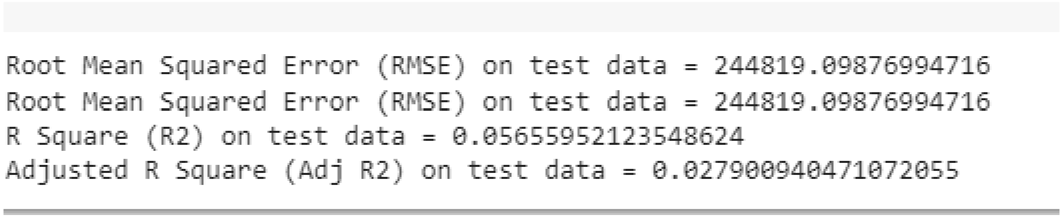
To evaluate the performance of machine learning models the original dataset split into 70/30 train test ratios. We have trained the model with 70% of the dataset.

**Step 5:** Build and evaluate different machine learning models such as linear regression, decision trees, and random forests to predict graduate salaries based on the available data.

In this step the following models were applied on the dataset and evaluated the performance of each models also:

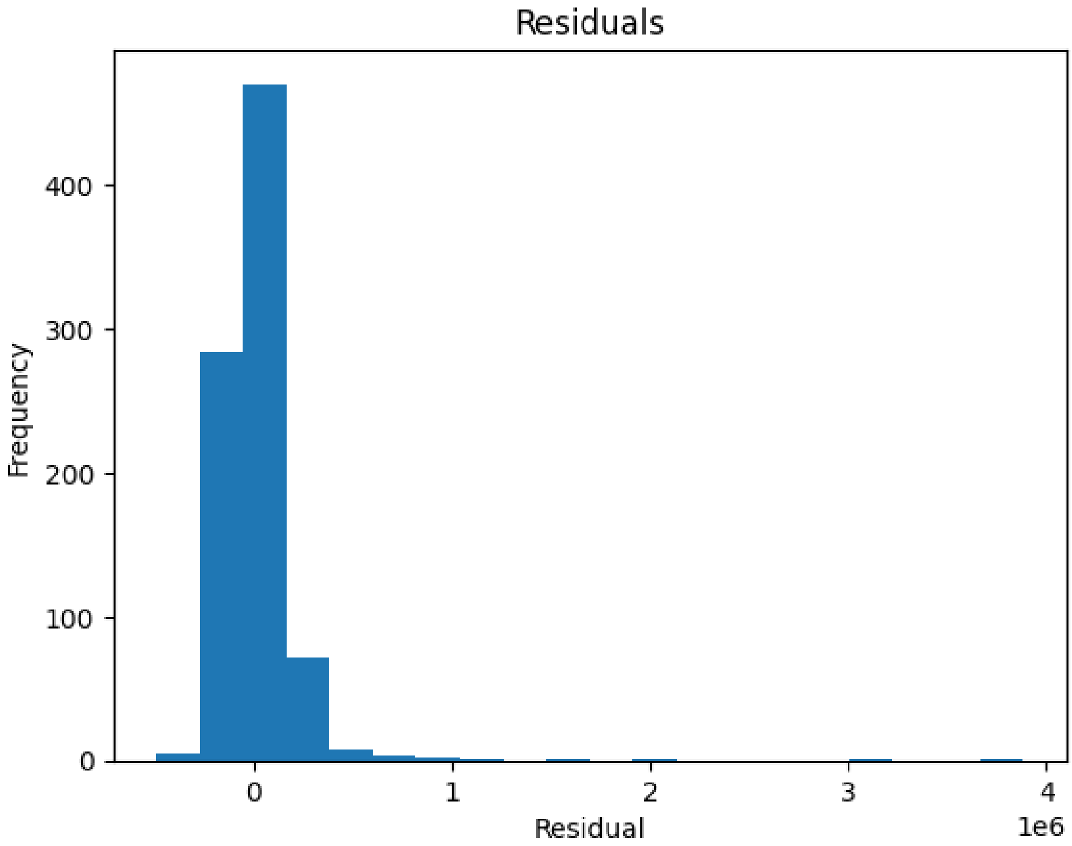
1. Regression Model:

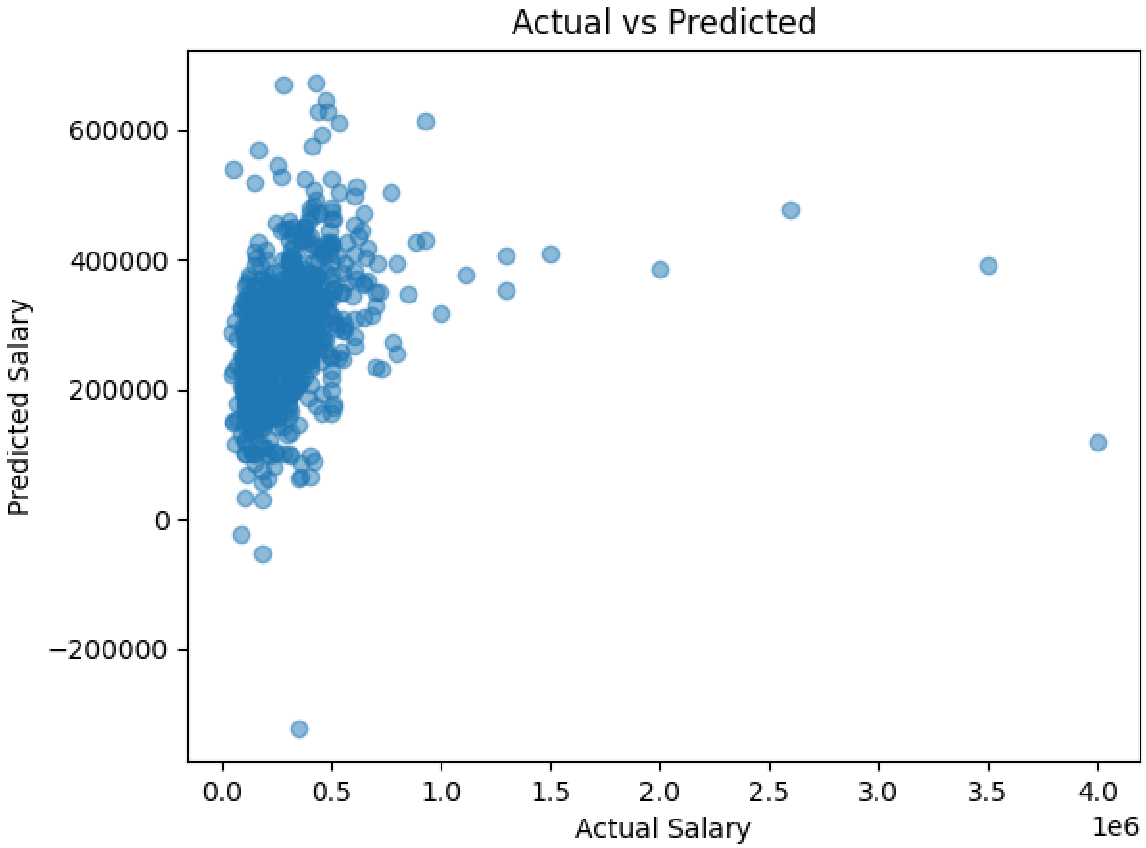
The regression model was applied successfully and got the below score shown in figure.



```
Root Mean Squared Error (RMSE) on test data = 244819.09876994716
Root Mean Squared Error (RMSE) on test data = 244819.09876994716
R Square (R2) on test data = 0.05655952123548624
Adjusted R Square (Adj R2) on test data = 0.027900940471072055
```

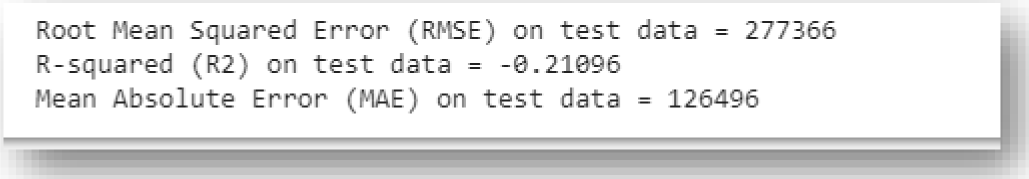
The regression model was successfully applied to test data, and the evaluation metrics were calculated. The Root Mean Squared Error (RMSE) on test data is 244819.09876994716, which indicates the average difference between the actual and predicted values of the target variable. The R Square (R2) on test data is 0.05655952123548624, which indicates the proportion of variance in the target variable that is explained by the regression model. The Adjusted R Square (Adj R2) on test data is 0.027900940471072055, which adjusts the R Square value for the number of predictor variables in the model. These metrics can be used to evaluate the performance of the regression model and make decisions on whether to refine the model or use it for predictions.


Residuals
Frequency
0
100
200
300
400
Residual
0
1
2
3
4
1e6

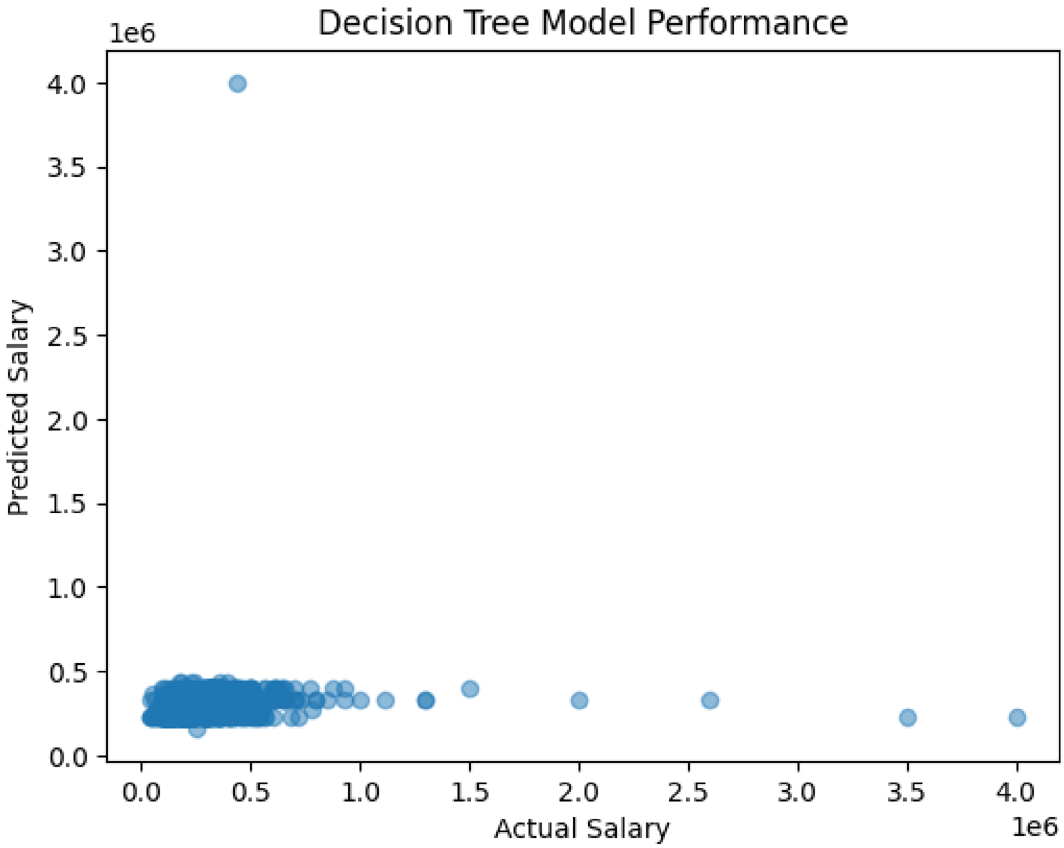


**2. Decision Tree Model:**

The Root Mean Squared Error (RMSE) on the test data is 277366, which means that the average difference between the predicted values and the actual values is 277366. The R-squared value on the test data is -0.21096, which indicates that the model is performing poorly in predicting the salaries of graduate students based on the available data. An R-squared value of 1 would indicate a perfect fit, and a value of 0 would indicate that the model is no better than predicting the mean value of the target variable. A negative value for R-squared suggests that the model is performing worse than the mean prediction. Additionally, the Mean Absolute Error (MAE) on the test data is 126496, which means that on average, the model is off by 126496 in its predictions of graduate salaries. Overall, these results suggest that the model needs to be improved to make more accurate predictions.



```
Root Mean Squared Error (RMSE) on test data = 277366
R-squared (R2) on test data = -0.21096
Mean Absolute Error (MAE) on test data = 126496
```

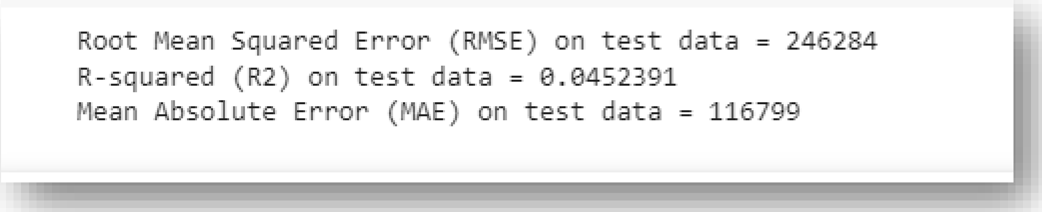


### 3. Random Forest Model

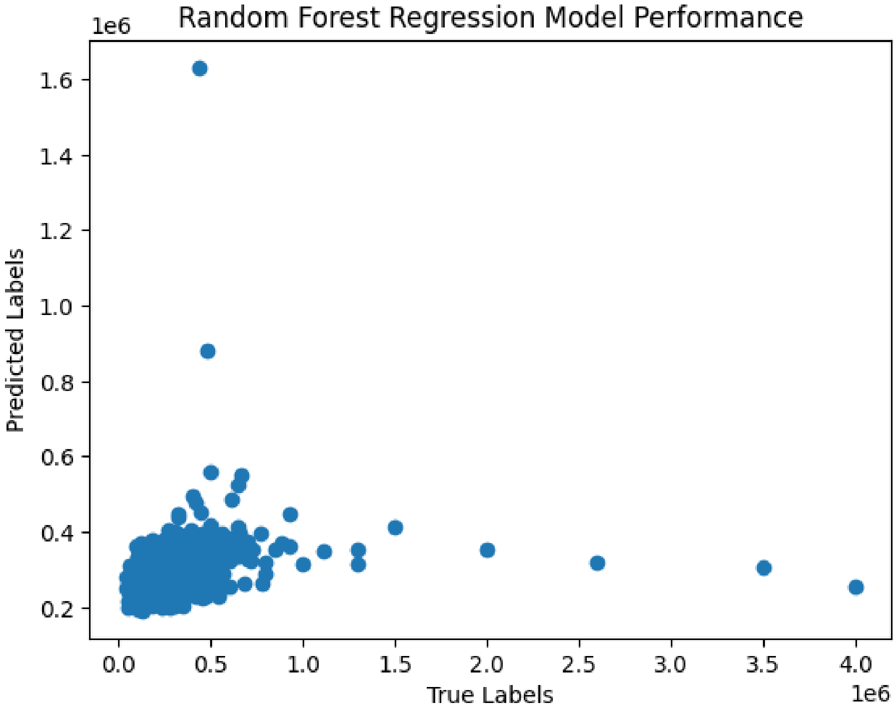
Random forest model were applied successfully, in below figure these are the evaluation metrics for a model on test data. The Root Mean Squared Error (RMSE) is 246284, which indicates the average difference between the predicted and actual values of the target variable (in this case, salary) is around 246284. A lower RMSE value is generally desirable, as it indicates that the model's predictions are more accurate.

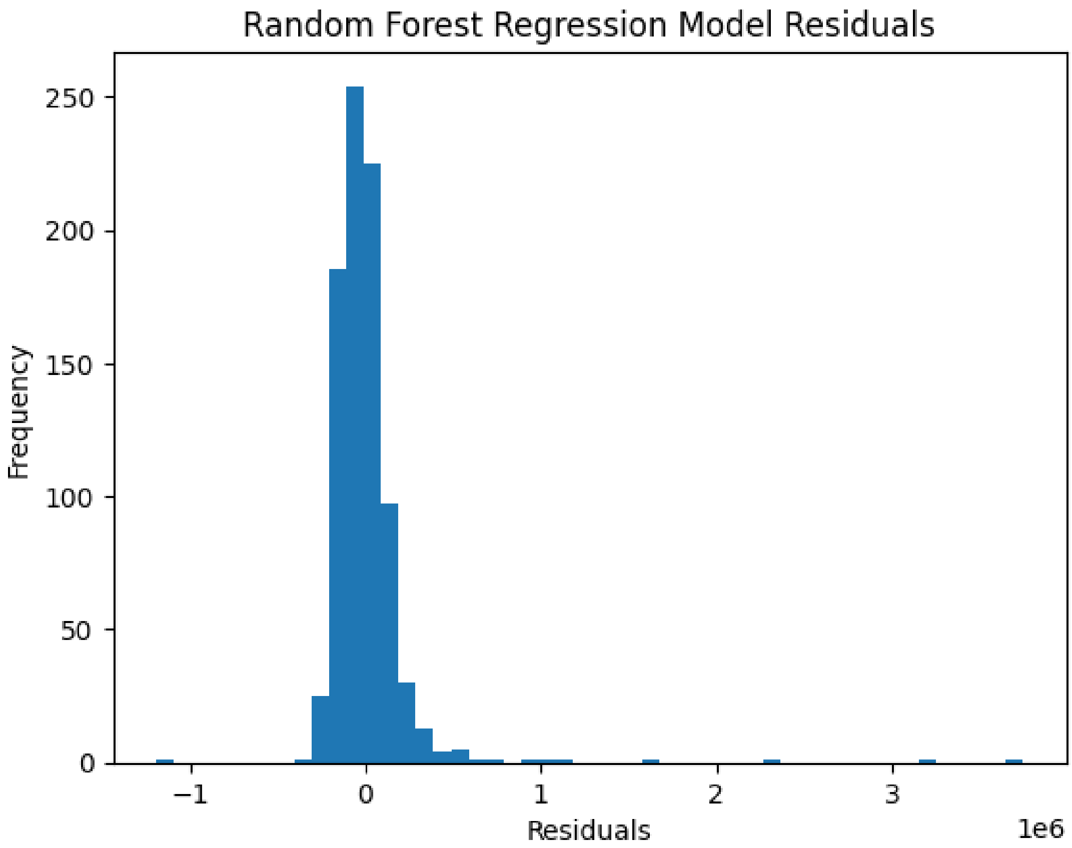
The R-squared (R2) value of 0.0452391 means that only 4.5% of the variance in the target variable (salary) is explained by the model. This value ranges from 0 to 1, with 1 indicating a perfect fit to the data. A higher R2 value is generally desirable, as it indicates that the model is able to explain more of the variation in the target variable.

The Mean Absolute Error (MAE) on test data is 116799, which is the average absolute difference between the predicted and actual values of the target variable. A lower MAE value is generally desirable, as it indicates that the model's predictions are more accurate.



```
Root Mean Squared Error (RMSE) on test data = 246284
R-squared (R2) on test data = 0.0452391
Mean Absolute Error (MAE) on test data = 116799
```



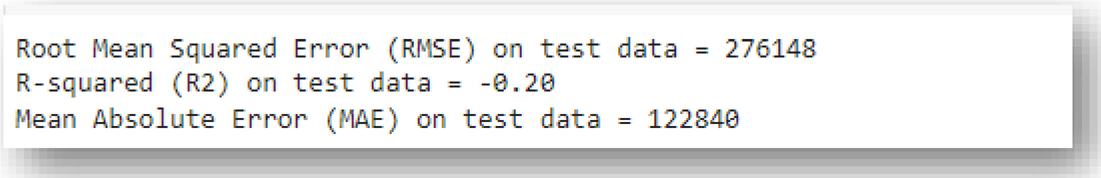


4. **Gradient Boost Model**

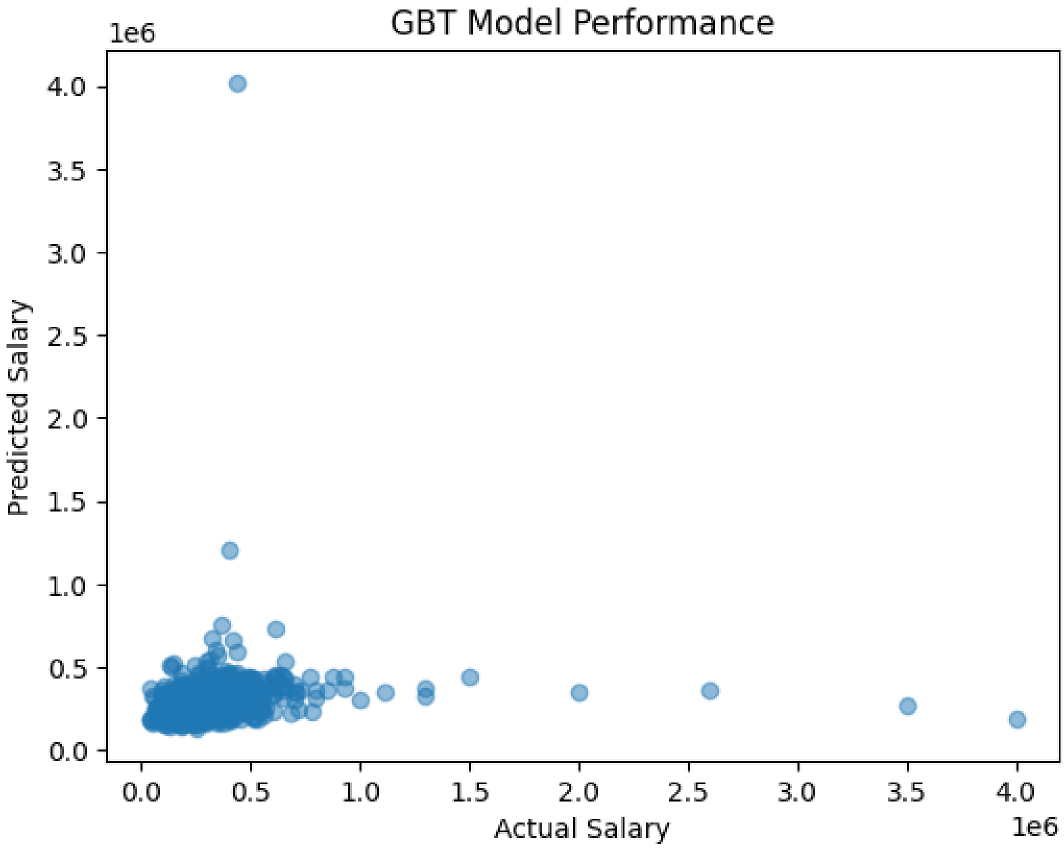
The RMSE is a measure of the difference between predicted and actual values, and in this case, it indicates that the model has an average error of 276148 in its predictions on the test data.

The R-squared (R2) value indicates how well the model fits the data, and in this case, the negative value of -0.20 suggests that the model is a poor fit for the data.

The MAE is another measure of the difference between predicted and actual values, and in this case, it indicates that the model has an average absolute error of 122840 in its predictions on the test data.



```
Root Mean Squared Error (RMSE) on test data = 276148
R-squared (R2) on test data = -0.20
Mean Absolute Error (MAE) on test data = 122840
```



**Step 6:** Choose the most suitable model based on the evaluation metrics such as mean squared error, mean absolute error, and R-squared value.

| Model | RMSE | R-squared | MAE |
|---|---|---|---|
| Linear regression | 244817.27 | 0.0565 | - |
| Decision Tree | 277366 | -0.2109 | 126496 |
| Random Forest | 246284 | 0.0452 | 116799 |
| Gradient Boosting | 276148 | -0.20 | 122840 |

The evaluation metrics in performance evaluation table, the most suitable model would be the Linear Regression model as it has the lowest Root Mean Squared Error (RMSE) and the highest R-squared (R2) value. The Decision Tree and Gradient Boost models have negative R-squared values indicating that they are not suitable for this dataset. The Random Forest model has a lower RMSE compared to Linear Regression, but a lower R-squared value indicating that it may not be as good a fit for the data.

**Step 7:** Use the chosen model to predict the salaries of graduate students and analyze the effect of different factors such as CGPA, board exams, specialization, location, and tier on their earnings.

**Step 8:** Provide insights and recommendations based on the results to help students make informed decisions about their career paths, recruiters identify high-performing students early in their academic careers, and guide the development of academic programs based on the success of certain programs in producing graduates with high salaries.

## Implementation Progress Report

In this increment, we applied a single model on the gathered data and evaluated its performance. The tasks were divided among team members as follows: Divya Anusha Chandrupatla, Mani Venkata sai snehith Nadipally, Sannihitha kolipaka, and Vamsi krishna chittyimadha, with each member contributing 25% to the project.

In 2nd increment we have applied and evaluated multiple models on dataset, due to limited number of dataset and dataset values in the dataset.

## Team Management

Created the WhatsApp group for all the members to easily communicate the projects matters. Different type of models applied by each member as follows:

1. Divya Anusha Chandrupatla : Applied and evaluated Gradient Boost Model
2. Mani Venkata sai snehith Nadipally: applied and evaluated Linear Regression
3. Sannihitha kolipaka : applied and evaluated the Decision tree model
4. Vamsi krishna chittyimadha : applied and evaluated Random Forest model

Dataflow diagram, algorithm, and report writing done combine..

## Issues/Concerns:

In this project, we encountered challenges related to learning and adopting new technologies. Additionally, project management and work distribution proved to be laborious tasks that required significant effort to overcome.

## Summary

In this increment, we have successfully gathered the necessary data for our project and performed data analytics using PySpark and other big data tools. We have preprocessed the data by cleaning and transforming it to prepare for further analysis. We have also analyzed the data to gain insights into the relationship between different factors such as grades, board, location, and tier, and their impact on the salaries of graduate students. Our focus in this increment was on identifying the correlations between the average college percentage and salary range using PySpark queries.

Moving forward, we plan to continue our analysis by selecting a suitable machine learning model to predict the salaries of graduate students based on the available data. We will also explore the significance of the project, including its potential to guide academic program development and help students negotiate fair salaries during the hiring process. Finally, we will conclude by highlighting the major features and objectives of our project and discussing the significance of the findings that we will present in our final report.

## Conclusions

In conclusion, this increment of the project successfully achieved its objectives of gathering relevant data and performing data analytics using Big Data tools. The gathered data was preprocessed and analyzed to identify correlations between academic performance and potential salaries of graduate students. The PySpark BigData query was used to analyze the dataset for checking the correlations of average college percentage with salary range. The results of this analysis will inform the feature design of the machine learning model to be developed in the next increment. Overall, this increment sets a strong foundation for the development of a predictive model for graduate student salaries, which can provide valuable insights for students, recruiters, and universities.

In conclusion, Increment 2 of the project has successfully accomplished its objectives of training and testing four different machine learning models to predict graduate student salaries. The management of the team was also efficiently handled. After careful evaluation, the linear regression model was chosen as the most appropriate model for the project. The trained model was used to predict salaries of graduate students and analyze the impact of different factors such as CGPA, board exams, specialization, location, and tier on their earnings. The results of this analysis provide valuable insights for students, recruiters, and universities. Overall, Increment 2 builds upon the strong foundation laid in the previous increment, and sets the project on a path towards achieving its ultimate goal of developing a predictive model for graduate student salaries using Big Data tools.

## References

[1] A. Wadhwani, Suryawanshi, S., & Padghan. "Salary prediction using machine learning." International Journal of Advanced Research in Computer Science. (accessed.

[2] L. Dong, Wu, Z., Li, Y., & Zhang, L, "Research on the correlation between big data and salary level based on machine learning. ," 2020.

[3] R. M. Oducado and A. Penuela, "Predictors of academic performance in professional nursing courses in a private nursing school in Kalibo, Aklan, Philippines," *Asia Pacific Journal of Education, Arts and Sciences,* vol. 1, no. 5, pp. 21-28, 2014.

[4] W.-H. Ko, "The relationships among professional competence, job satisfaction and career development confidence for chefs in Taiwan," *International Journal of Hospitality Management,* vol. 31, no. 3, pp. 1004-1011, 2012.