Uzma Aijaz : CS_19052
Khadija : CS_19075
Musfira Fayyaz : CS_19303

DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING
BACHELOR IN COMPUTER SYSTEMS ENGINEERING
Course Code: CS-324
Course Title: Machine Learning
Complex Engineering Problem
TE Batch 2019, Spring Semester 2022
Grading Rubric
TERM PROJECT

Group Members:

Student No.	Name	Roll No.
S1	**Uzma Aijaz**	**CS-19052**
S2	**Khadija**	**CS-19075**
S3	**Musfira Fayyaz**	**CS-19303**

CRITERIA AND SCALES				**Marks Obtained**		
				S1	**S2**	**S3**
Criterion 1: Does the application meet the desired specifications and produce the desired outputs? (CPA-1, CPA-2, CPA-3) **[8 marks]**						
1	2	3	4			
The application does not meet the desired specifications and is producing incorrect outputs.	The application partially meets the desired specifications and is producing incorrect or partially correct outputs.	The application meets the desired specifications but is producing incorrect or partially correct outputs.	The application meets all the desired specifications and is producing correct outputs.			
Criterion 2: How well is the code organized? **[2 marks]**						
1	2	3	4			
The code is poorly organized and very difficult to read.	The code is readable only to someone who knows what it is supposed to be doing.	Some part of the code is well organized, while some part is difficult to follow.	The code is well organized and very easy to follow.			
Criterion 3: Does the report adhere to the given format and requirements? **[6 marks]**						
1	2	3	4			
The report does not contain the required information and is formatted poorly.	The report contains the required information only partially but is formatted well.	The report contains all the required information but is formatted poorly.	The report contains all the required information and completely adheres to the given format.			
Criterion 4: How does the student perform individually and as a team member? (CPA-1, CPA-2, CPA-3) **[4 marks]**						
1	2	3	4			
The student did not work on the assigned task.	The student worked on the assigned task and accomplished goals partially.	The student worked on the assigned task and accomplished goals satisfactorily.	The student worked on the assigned task and accomplished goals beyond expectations.			

Final Score = (Criterial_1_score x 2) + (Criteria_2_score / 2) + (Criteria_3_score x (3/2)) + (Criteria_4_score)
= _____

Uzma Aijaz : CS_19052
Khadija : CS_19075
Musfira Fayyaz : CS_19303

PREDICTION OF CUMULATIVE GRADE POINT AVERAGE

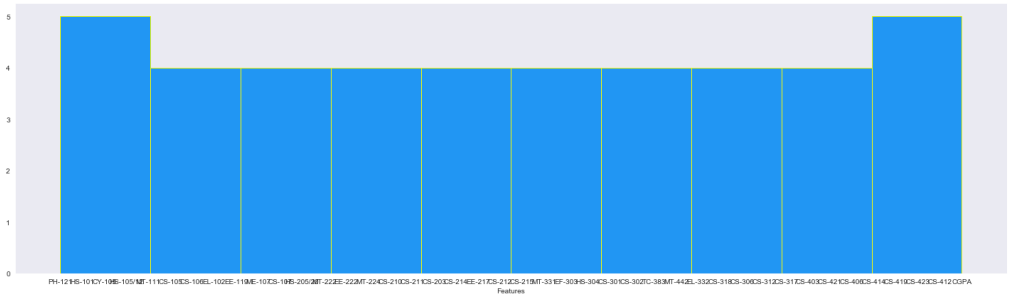
1. Pre-Processing Step Details:

Data preprocessing is a crucial step in Machine Learning which helps to enhance the dataset and extract useful insights from the data. The following steps are followed for the preprocessing of The Grades Dataset.

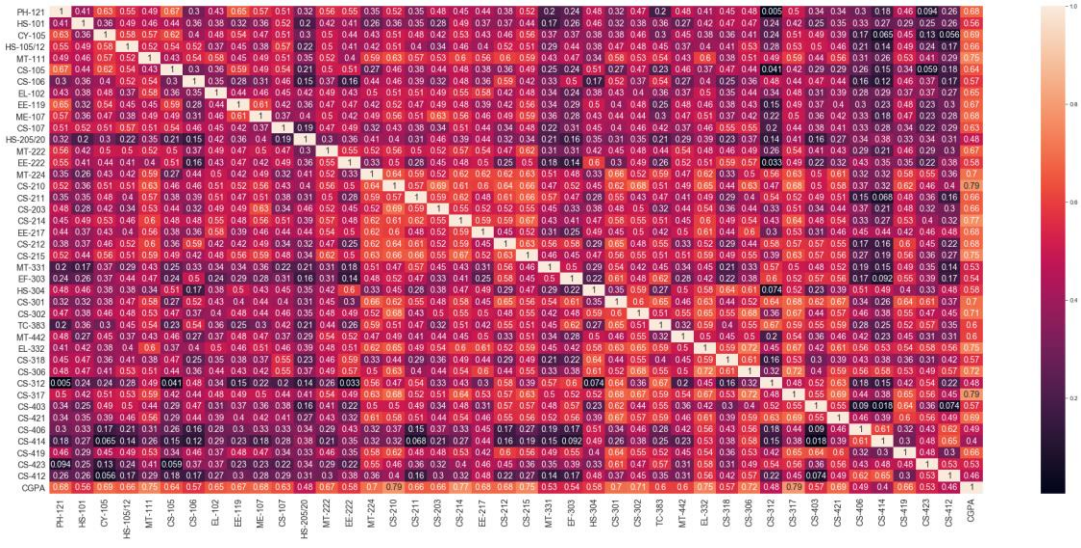
- **Importing Libraries:** Some crucial Python libraries are imported for the preprocessing. They include; **NumPy** for all kinds of scientific and mathematical calculations. **Pandas** is used for data manipulation, analysis, and management of the dataset. **Matplotlib** is a 2D Python library that is used to plot graphs and see the correlation. **Scikit-learn** is also used which provides a selection of efficient tools for machine learning and statistical modeling.
- **Acquiring Dataset:** The given dataset namely "The_Grades_Dataset" is placed in the working directory and imported into the notebook
- **Identification and Removal of Missing values:** The dataset contained many null values such as in the entries of columns: CY-105, EE-222, and many more. Such cells are filled up with the mode value of their respective column. The mode is taken as the entries were in non-numeric form and their mean or median couldn't be found
- **Removing Extra Features:** Unnecessary features such as "Seat No" is dropped as it doesn't play part in the prediction of CGPA.
- **Categorical Data Encoding:** String values are converted to numeric values by taking A+ as 11.0, A as 10.0, A- as 9.0, B+ as 8.0, B as 7.0, B- as 6.0, C+ as 5.0, C as 4.0, C- as 3.0, D+ as 2.0, D as 1.0, F as 0. Meanwhile, I, W, and WU are also taken as 0 as they result in incompletion and withdrawal and no CGPA calculation is required for that.

Uzma Aijaz : CS_19052
Khadija : CS_19075
Musfira Fayyaz : CS_19303

- **Graphical Analysis:** Feature histogram is made to check out the distribution of each feature which gives the following result.



- **Correlation Between Features:** A correlation graph is also made to visualize the relationships among all the features.



- **Scaling the Data:** The data values are scaled using the Standardization technique as the dataset contains outliers and standardization brings down all the features to a common scale without distorting the differences in the range of the values.
- **Splitting the Dataset:** The cleaned dataset is then divided into training and testing parts for the prediction in the ratio of 80:20.

2. Models and Machine Learning Algorithms used:

- **MODEL# 1:** The first model contains GPs of only first-year courses in the input set and is used for training and to predict the final CGPA.

  The Machine Learning algorithms used for the training of Model # 1 are as follows:

- ★ **KNN Regressor Algorithm:** KNN regression is a non-parametric method that, intuitively, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.
- ★ **Decision Tree Regressor:** Decision Tree Regressor builds regression in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

- **MODEL# 2:** The second model predicts the final CGPA based on the GPs of the first year and second.

  The Machine Learning algorithms used for the training of Model # 2 are as follows:

  - ★ **SVM:** SVMs are used to find the best line in two dimensions or the best hyperplane in more than two dimensions to help us separate our space into classes.
  - ★ **Random Forest Regressor:** Random Forest Regression is a supervised learning algorithm that uses an ensemble learning method for regression. The ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

- **MODEL# 3:** The third model predicts the final CGPA based on the GPs of the first, second, and third years.

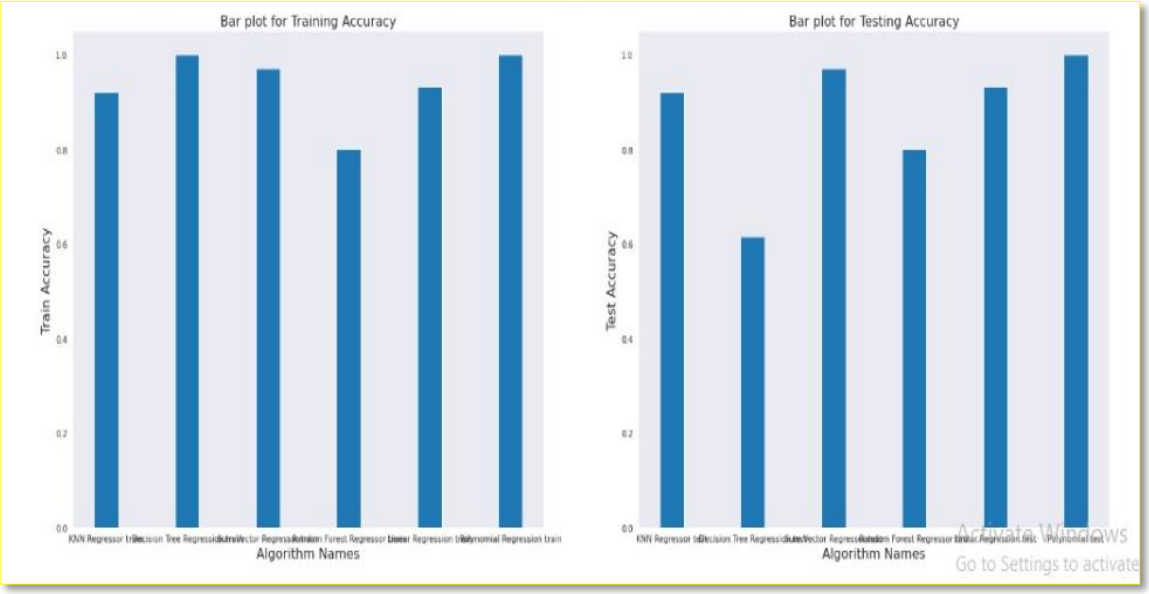
  The Machine Learning algorithms used for the training of Model # 3 are as follows:

  - ★ **Linear Regression:** Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.
  - ★ **Polynomial Regression:** Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial. Polynomial Regression is sensitive to outliers so the presence of one or two outliers can also badly affect the performance.

Uzma Aijaz : CS_19052
Khadija : CS_19075
Musfira Fayyaz : CS_19303

3. Graphical Analysis of all Models:

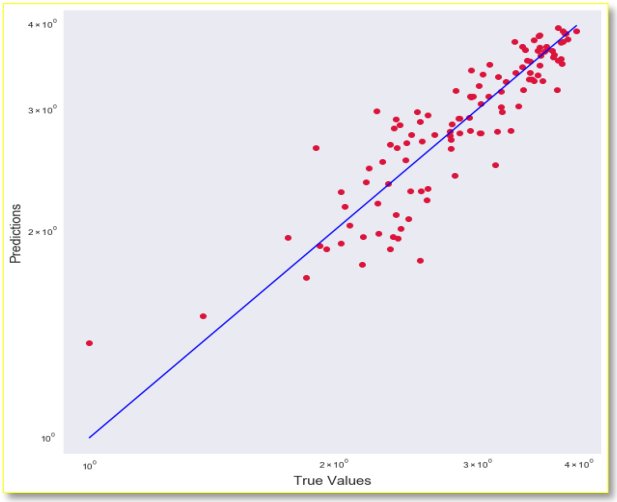
The graphical of all three models are given as follows:



4. Performance of Machine Learning System, Issues, and Improvements:

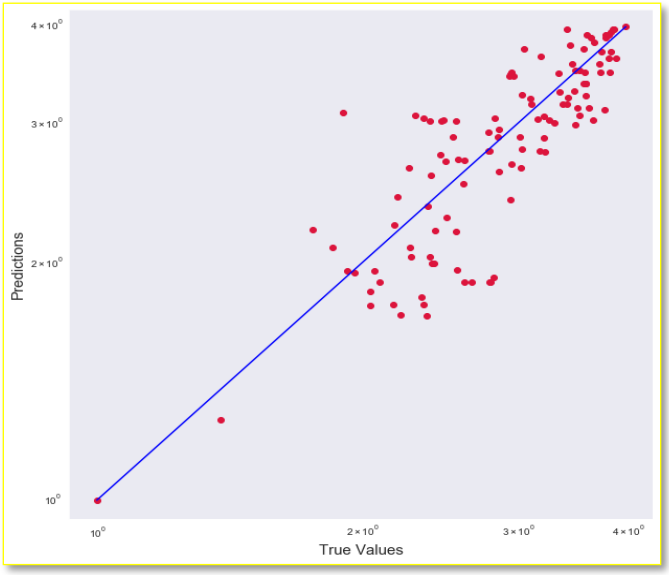
- **Model # 1:** The first algorithm applied to Model # 1 is **KNN Regressor Algorithm** which gives an Accuracy: of 80.53% KNN works well on Model # 1.

  The graph between true and predicted values gives the following result:

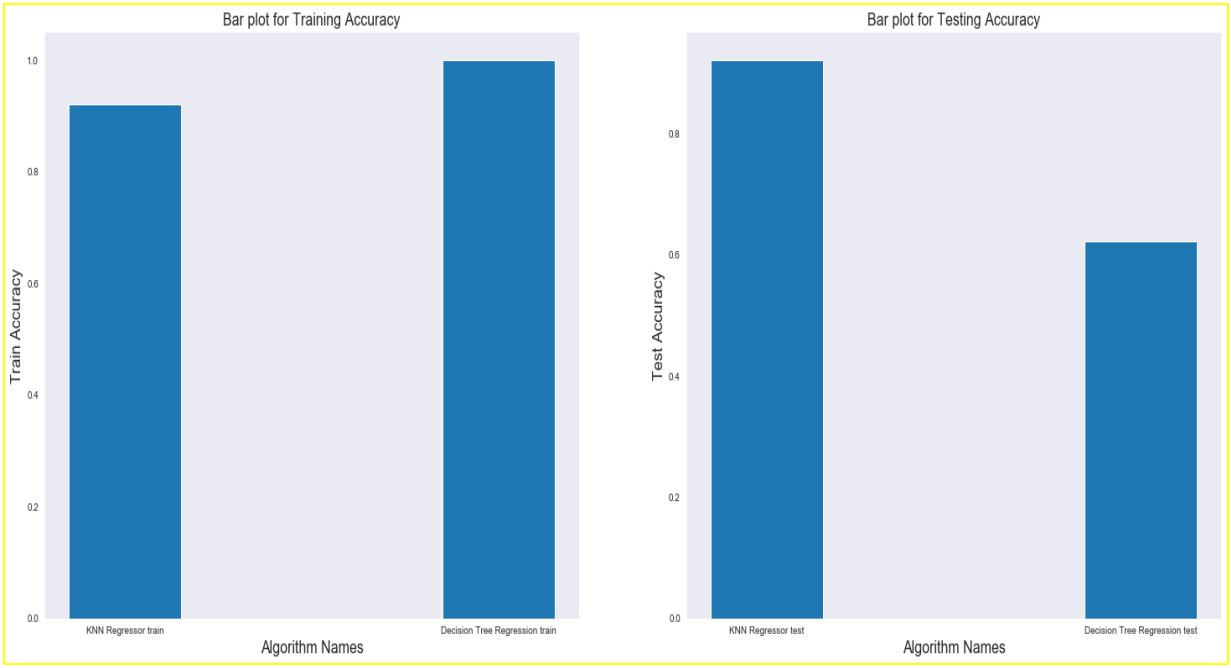


Uzma Aijaz : CS_19052
Khadija : CS_19075
Musfira Fayyaz : CS_19303

The second algorithm applied to Model # 1 is the **Decision Tree Regressor Algorithm** which gives an Accuracy: of 62.32%. Decision Tree Regressor is causing an issue of **Overfitting** as it has a training accuracy of 99.99% while testing accuracy is just 62.3%. The graph between true and predicted values gives the following result:



The combined graph for **KNN Regressor** and **Decision Tree Regressor** gives the following output.



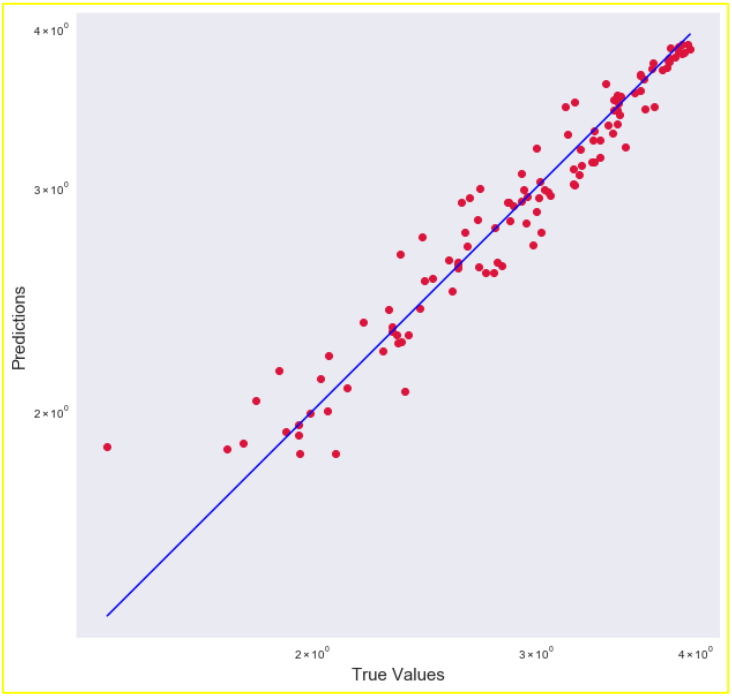
The KNN regression predicted nearest to the actual output as compared to the Decision Tree regression model.

```
Score for subject PH-121 is: a
Score for subject HS-101 is: a+
Score for subject CY-105 is: b
Score for subject HS-105/12 is: b+
Score for subject MT-111 is: c+
Score for subject CS-105 is: d
Score for subject CS-106 is: d+
Score for subject EL-102 is: a
Score for subject EE-119 is: a+
Score for subject ME-107 is: b
Score for subject CS-107 is: b+
CGPA using KNN Regressor algorithm: [3.005]
CGPA using Decision Tree Regression: [2.612]
```

- **Model # 2:**

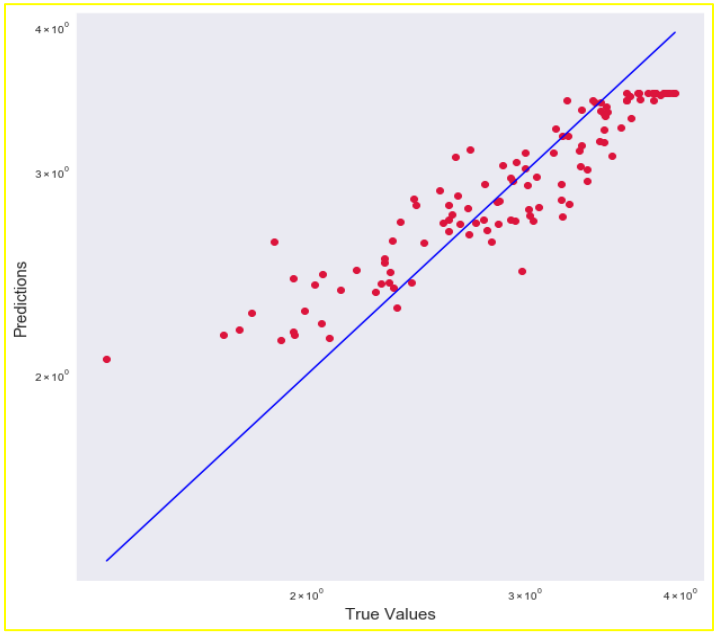
The first algorithm applied to Model # 2 is the **SVM Regressor Algorithm** which has ves Accuracy: of 94.61%. SVM perfectly fits model # 2 as it gives 96.88% training accuracy and 94.6% testing accuracy.

The graph between true and predicted values shows the following result:

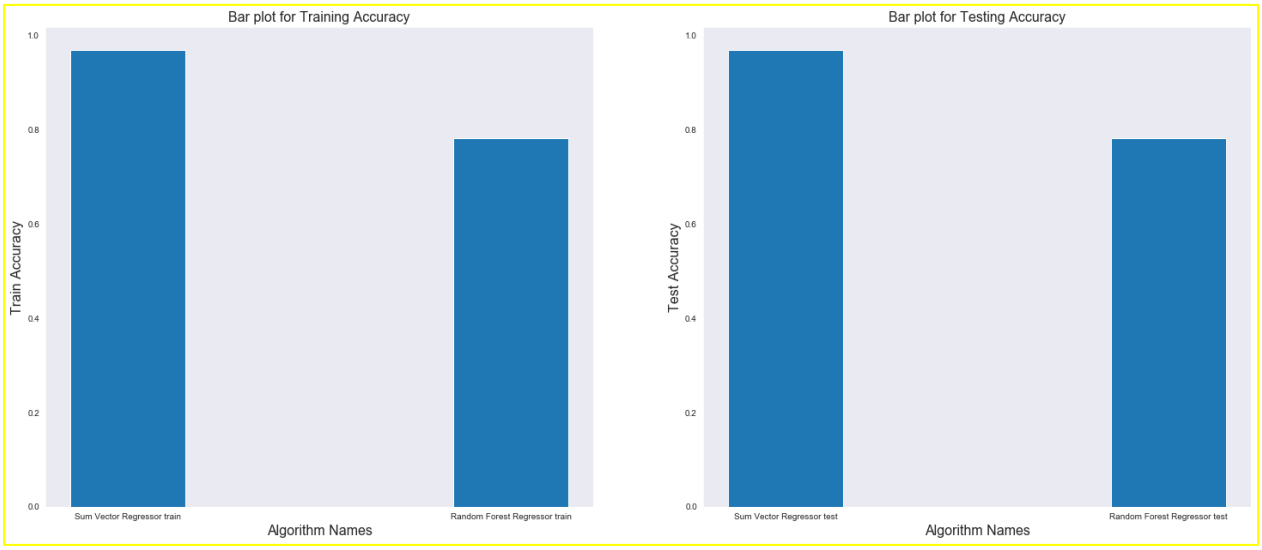


The second algorithm applied to Model # 2 is the **Random Forest Regressor Algorithm** which gives an Accuracy: of 80.96%. Random Forest Regressor performs well as it gives 78.15% training accuracy and 80.962% testing accuracy.

The graph between true and predicted values shows the following result:



The combined graph for **SVM Regressor Algorithm** and **Random Forest Regressor Algorithm** gives the following output.

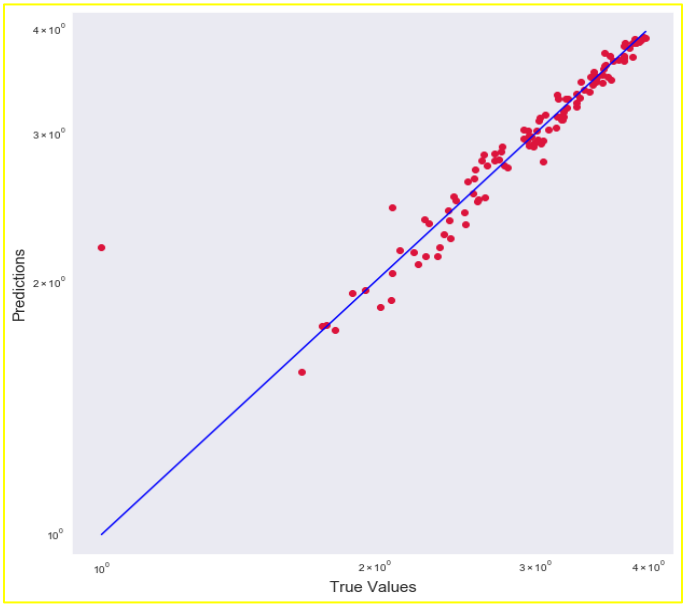


Vector Machine Algorithms cannot predict the nearest value of output, and Random Forest too, but Random Forest behaves better than SVM.

```
Score for subject PH-121 is: a
Score for subject HS-101 is: a
Score for subject CY-105 is: a+
Score for subject HS-105/12 is: a+
Score for subject MT-111 is: a
Score for subject CS-105 is: a
Score for subject CS-106 is: a
Score for subject EL-102 is: a
Score for subject EE-119 is: a
Score for subject ME-107 is: a
Score for subject CS-107 is: a+
Score for subject HS-205/20 is: a+
Score for subject MT-222 is: a+
Score for subject EE-222 is: a
Score for subject MT-224 is: a
Score for subject CS-210 is: a
Score for subject CS-211 is: a
Score for subject CS-203 is: a
Score for subject CS-214 is: a
Score for subject EE-217 is: a
Score for subject CS-212 is: a
Score for subject CS-215 is: a
CGPA using Sum Vector Regressor algorithm: [2.52051168]
CGPA using Random Forest Regressor: [3.51832053]
```

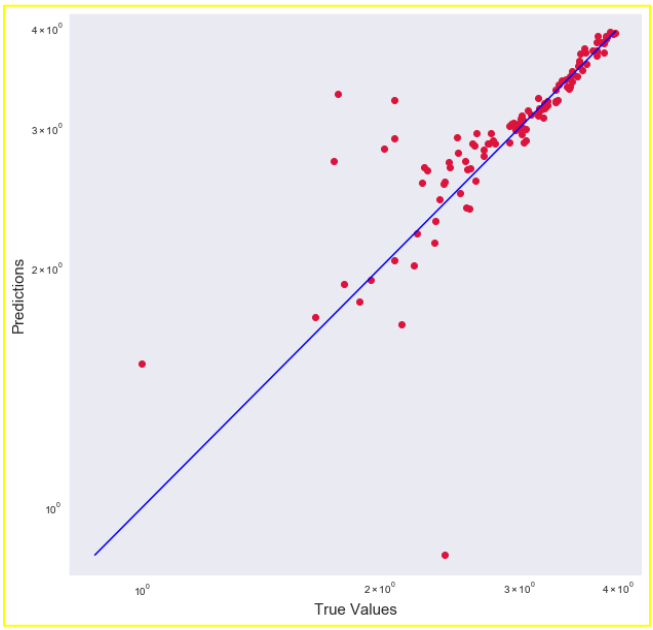
- **Model # 3:**

The first algorithm applied to Model # 3 is the **Linear Regression Algorithm** which gives an Accuracy: of 94.1%. For better fitting of the model, Polynomial Regression is used which starts to make the model overfit. The graph between true and predicted values shows the following result:

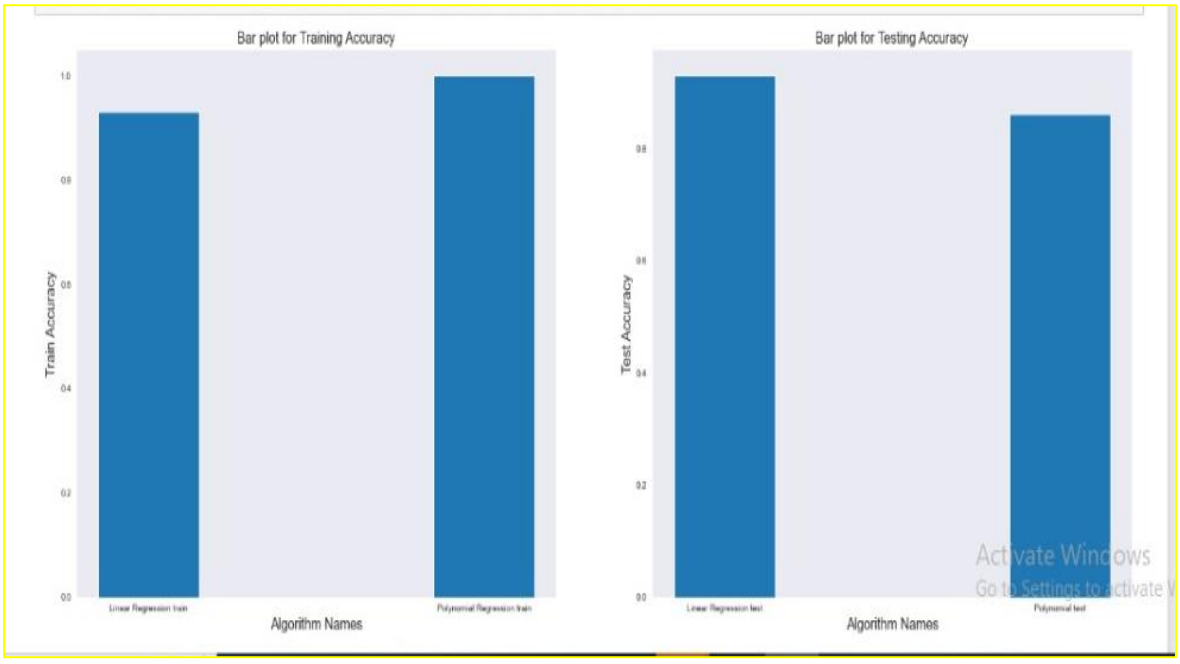


Uzma Aijaz : CS_19052
Khadija : CS_19075
Musfira Fayyaz : CS_19303

The second algorithm applied to Model # 3 is the **Polynomial Regression Algorithm** which gives an Accuracy: of 75.3%. The graph between true and predicted values shows the following result:



The combined graph of the **Linear Regression Algorithm** and **Polynomial Algorithm** gives the following output.



The Linear Regression generates a linear line and may miss the data of training features so it may cause underfitting. Then we treat the model with Polynomial Regression, but Polynomial Regression overfits the model, that's why the predicted value is very much unpredictable for both of the algorithms.

```
Score for subject PH-121 is: d
Score for subject HS-101 is: d
Score for subject CY-105 is: d
Score for subject HS-105/12 is: d
Score for subject MT-111 is: d
Score for subject CS-105 is: d
Score for subject CS-106 is: d
Score for subject EL-102 is: d
Score for subject EE-119 is: d
Score for subject ME-107 is: d
Score for subject CS-107 is: d
Score for subject HS-205/20 is: a
Score for subject MT-222 is: a
Score for subject EE-222 is: a
Score for subject MT-224 is: c
Score for subject CS-210 is: c
Score for subject CS-211 is: c
Score for subject CS-203 is: c
Score for subject CS-214 is: c
Score for subject EE-217 is: b
Score for subject CS-212 is: b
Score for subject CS-215 is: b
Score for subject MT-331 is: b
Score for subject EF-303 is: b
Score for subject HS-304 is: b
Score for subject CS-301 is: b
Score for subject CS-302 is: b
Score for subject TC-383 is: b
Score for subject EL-332 is: b
Score for subject CS-318 is: b
Score for subject CS-306 is: b
Score for subject CS-312 is: b
Score for subject CS-317 is: b
CGPA Linear Regression: [4.74370991]
CGPA Polynomial Regression: [11.64927888]
```

Improvement of Overfitting:

Overfitting can be removed by using:

- Cross-validation
- Regularization techniques are given below:
- Ridge Regression
- Lasso Regression
- Elastic net