DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING
BACHELORS 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	**SAMRA SALEEM**	**CS-19103**
S2	**MUBINA WAHEED**	**CS-19110**
S3		

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 organization? **[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 performed 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)
= ______________________

Table of Contents

Problem Statement

The objective of the project is to develop a machine learning system to predict the final CGPA of a student at the end of fourth year given GPs of the courses obtained in initial years (up to first, second or third year).

Abstract

The report comprises of the in-depth analysis of the data and performance of various supervised machine learning algorithms. The dataset comprises of the grades of students in different courses from FE to BE and their respective final CGPAs. The data was preprocessed using various preprocessing techniques and then transformed into cleaned form ready for processing. The problem is a classical regression problem as the target column contains a continuous range of values. Several ML models have been applied such as Linear Regression, K-Nearest Neighbours and Random Forest and their corresponding performance, accuracy and error rate is observed . The algorithm with most accurate results, low error rate and improved performance is selected for final prediction.

Libraries and Tools Used

- Programming Language used : Python 3.10
- Tool : Jupyter Notebook
- Libraries
  - Sklearn
  - Numpy
  - Pandas
  - Tabulate
  - Matplotlib
  - Seaborn

Data Preprocessing

Pre-processing is an important step in data mining. The purpose of data pre-processing is to convert the data into a suitable form that can be used by algorithms. Three main pre-processing steps have been applied to the dataset which are data cleaning, features encoding, and features selection. The pre-processing was implemented using Python language.

Encoding

The initial data provided for prediction was in categorical form containing GPS in the form of grades (A,B,A- etc). The data was encoded into numerical form using ordinal encoding. The grades were replaced by their corresponding GPA e.g. grade A+ was mapped to 4.0 gpa and grade F was mapped to 0.0 gpa. The grades WU, W and I are encoded with 0.0 gpa.

Feature Selection

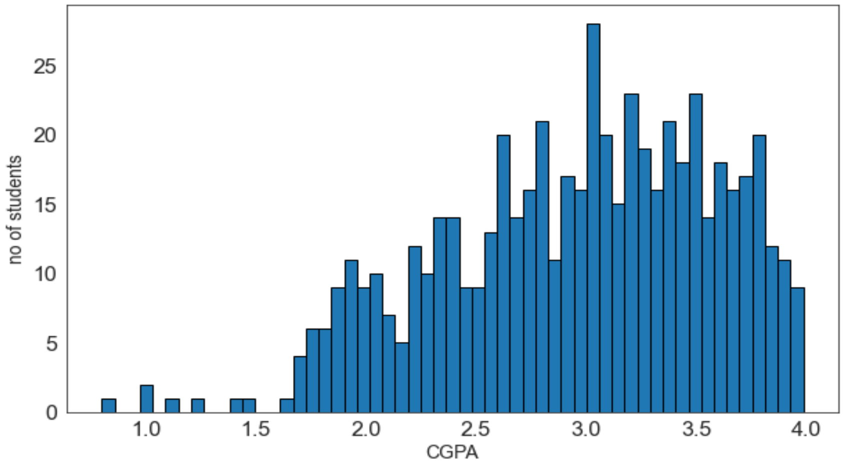
The dataset contained 571 instances and 43 features. Out of these 43 columns, only 34 columns were required for prediction. The first column named Seat No and all the columns corresponding to 4th year courses were simply dropped because they did not have any impact on the model's prediction. After selecting relevant features, the dataset was again subjected to feature selection for individual models. Model-1 contained only FE courses i.e. only 12 columns, Model-2 contained both FE and SE courses i.e. 23 columns and Model-3 contained FE, SE and TE courses i.e. all 34 columns. All the columns corresponding to courses were treated as input and the column of CGPA was treated as target column for all 3 models.

Handling Null Values

The null values were handled individually for each model. Model-1 contained only 5 rows having null values. Model 2 contained only 9 rows having null values. Model 3 contained only 12 rows having null values. As the total number of rows containing null values in each model is very less than the total no of instances, therefore, the rows having null values were simply dropped. If these null values were filled with approximate values, the data would have become noisy and would have impacted the overall performance of the model. Remaining rows in each model were:

- Model-1: 565 rows
- Model-2: 562 rows
- Model-3: 559 rows

Histogram for Data Distribution



The above histogram shows that the range of target variable is in between 0 to 4.0. Moreover, it can be observed that the majority of the students obtained a CGPA between 2.5 and 3.7.

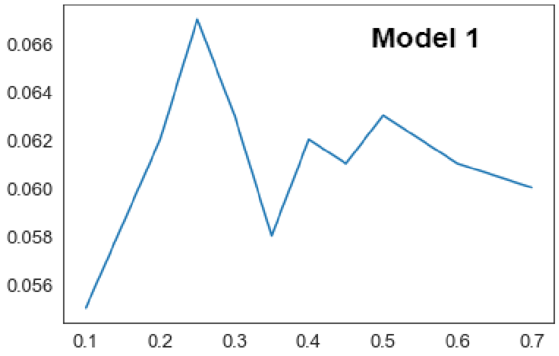
Checking Multicollinearity

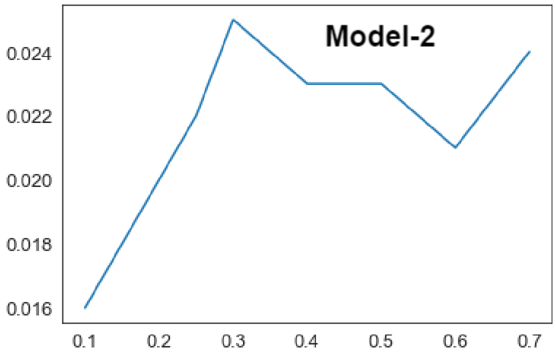
The data was analysed for determining how much the features are correlated with one another. All the features showed a correlation less than 0.7. Therefore, the probability of the occurrence of multicollinearity is very low. Hence, no features were removed.

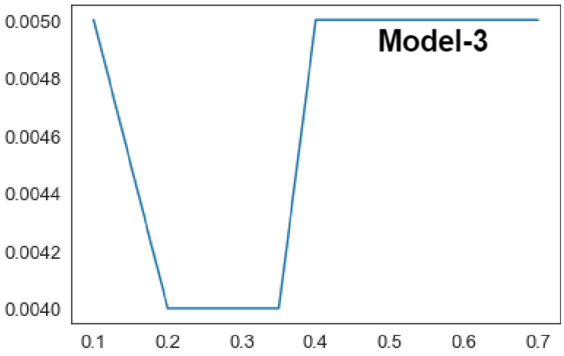
Literature Review of Algorithms Used

Linear Regression

Regression techniques are used to predict continuous outcomes rather than predicting discrete class labels. In order to apply Linear Regression, the data has been split into training and testing sets. The optimal test size was determined by observing the training and testing accuracy of the algorithm on different values of test size ranging from 0.1 to 0.7 for each model separately. Test size resulting in high train-test score and low mean squared error is selected for further processing. The optimal test size used for all 3 models is as follows







- Model-1 : Test size of 0.1 is used
- Model-2 : Test size of 0.1 is used
- Model-3 : Test size of 0.2 is used

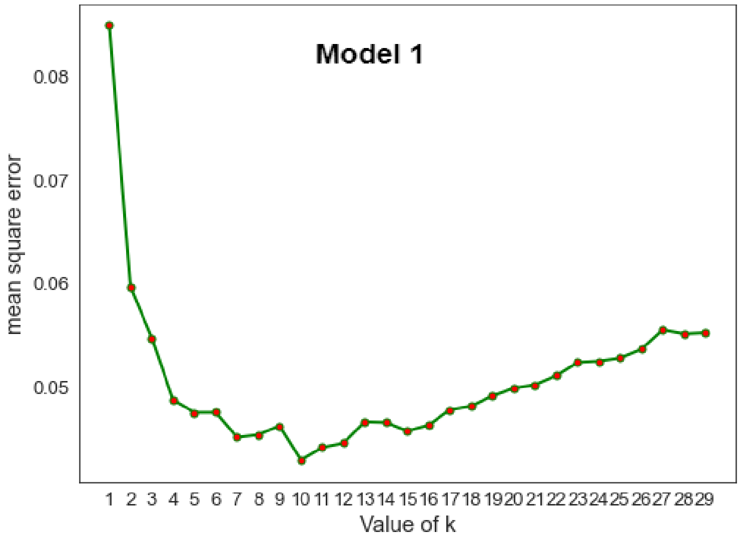
PolyNomial Regression

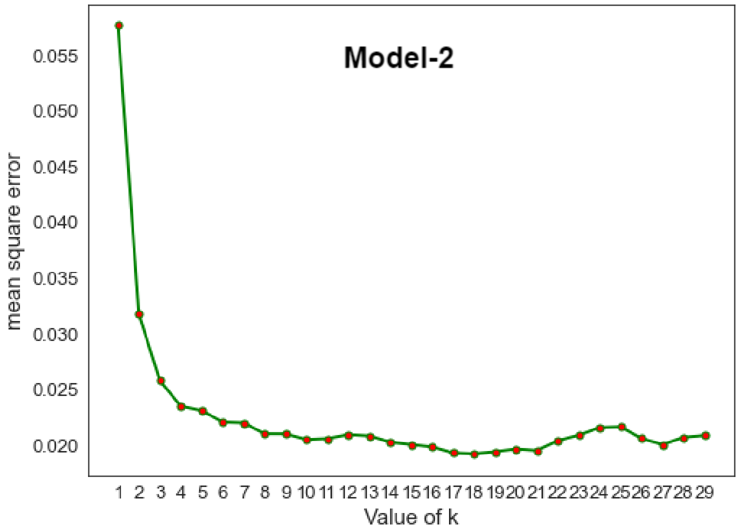
Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial. PolyNomial

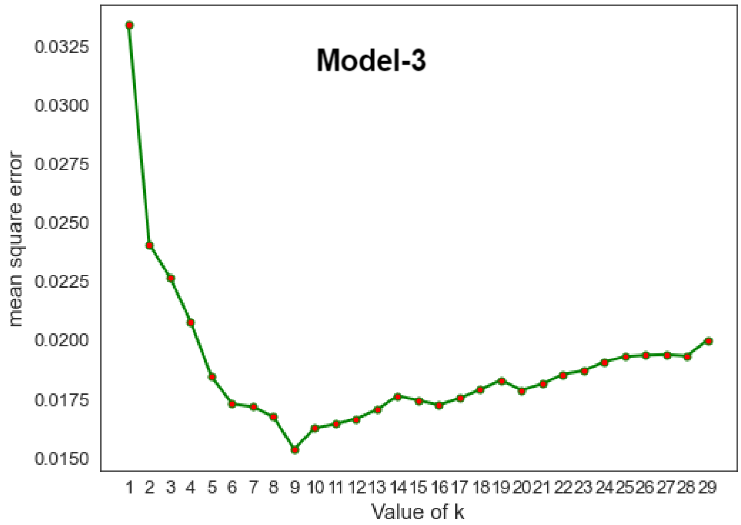
Regression with different degrees was applied on all three models and their corresponding accuracy and performance was observed. This algorithm fitted best on model-1 (with degree 2) as the number of features are low in that model. However, the model started to overfit when subjected to higher degree polynomials in all three models.

K Nearest Neighbours Regressor

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. The optimal value of k is selected by fitting the algorithm on different values of k and observing their train test scores and corresponding mean squared error. For each model, the value of k resulting in low mean squared error and high train-test accuracy was selected for further processing.







Following values of k have been selected for prediction:

- Model-1: K=10
- Model-2: K=21
- Model-3 K=9

Random Forest Regressor

Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. In this project this algorithm is used with default parameters.

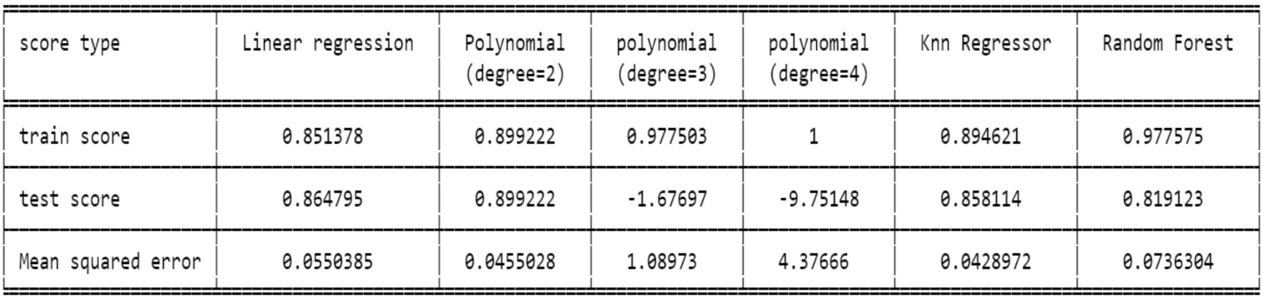
Metrics

As it is a regression problem, the following metrics have been used to evaluate the performance of the model.

- Mean Squared Error
- R-score

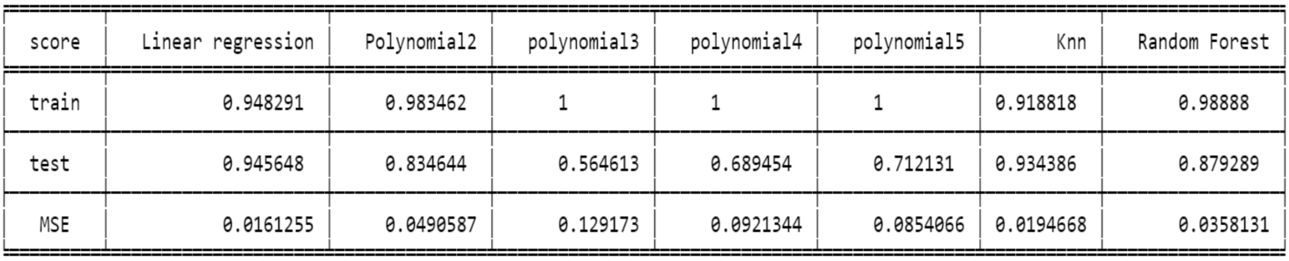
Tabular Representation of Performance of Models

Model 1



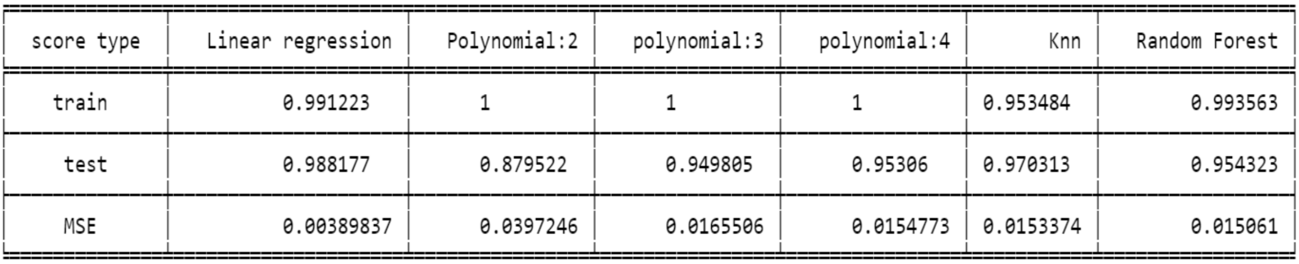
score type	Linear regression	Polynomial (degree=2)	polynomial (degree=3)	polynomial (degree=4)	Knn Regressor	Random Forest
train score	0.851378	0.899222	0.977503	1	0.894621	0.977575
test score	0.864795	0.899222	-1.67697	-9.75148	0.858114	0.819123
Mean squared error	0.0550385	0.0455028	1.08973	4.37666	0.0428972	0.0736304

Model 2



score	Linear regression	Polynomial2	polynomial3	polynomial4	polynomial5	Knn	Random Forest
train	0.948291	0.983462	1	1	1	0.918818	0.98888
test	0.945648	0.834644	0.564613	0.689454	0.712131	0.934386	0.879289
MSE	0.0161255	0.0490587	0.129173	0.0921344	0.0854066	0.0194668	0.0358131

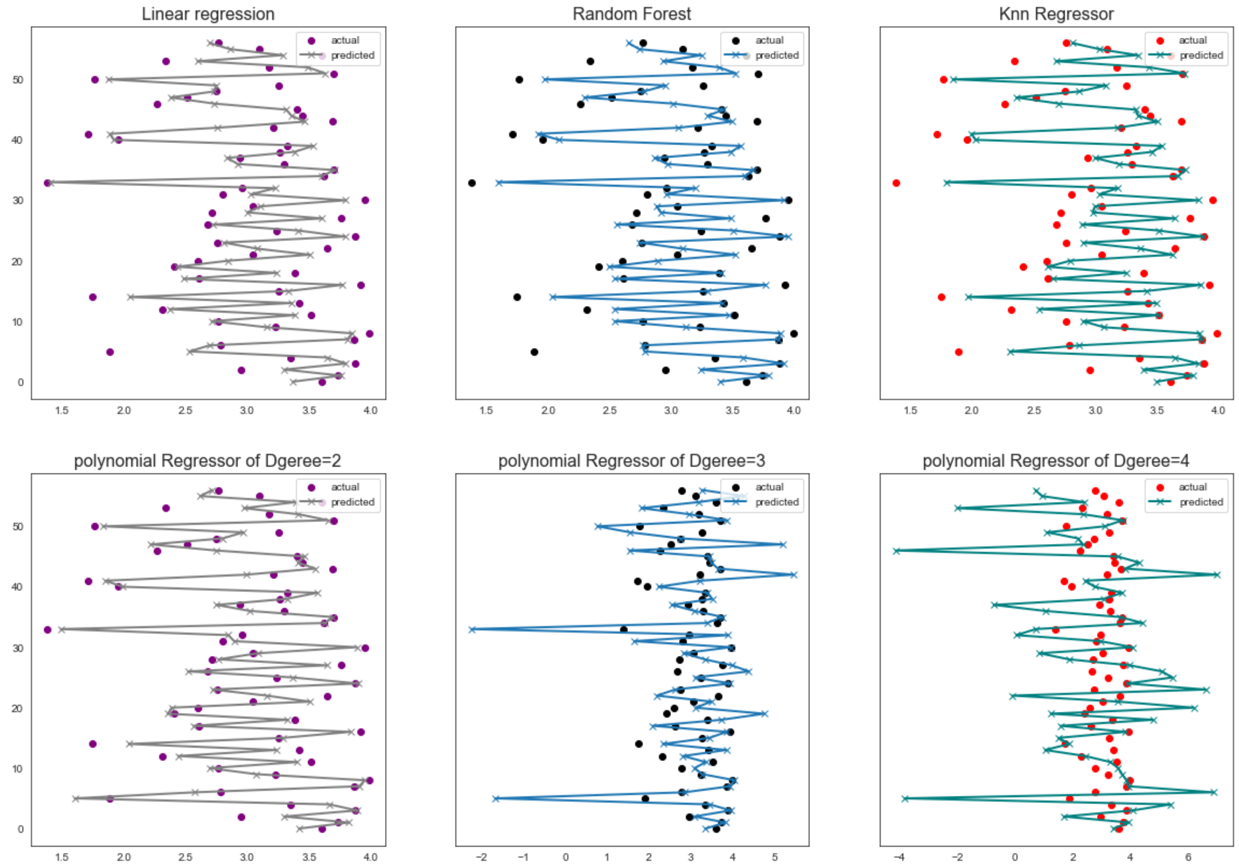
Model 3



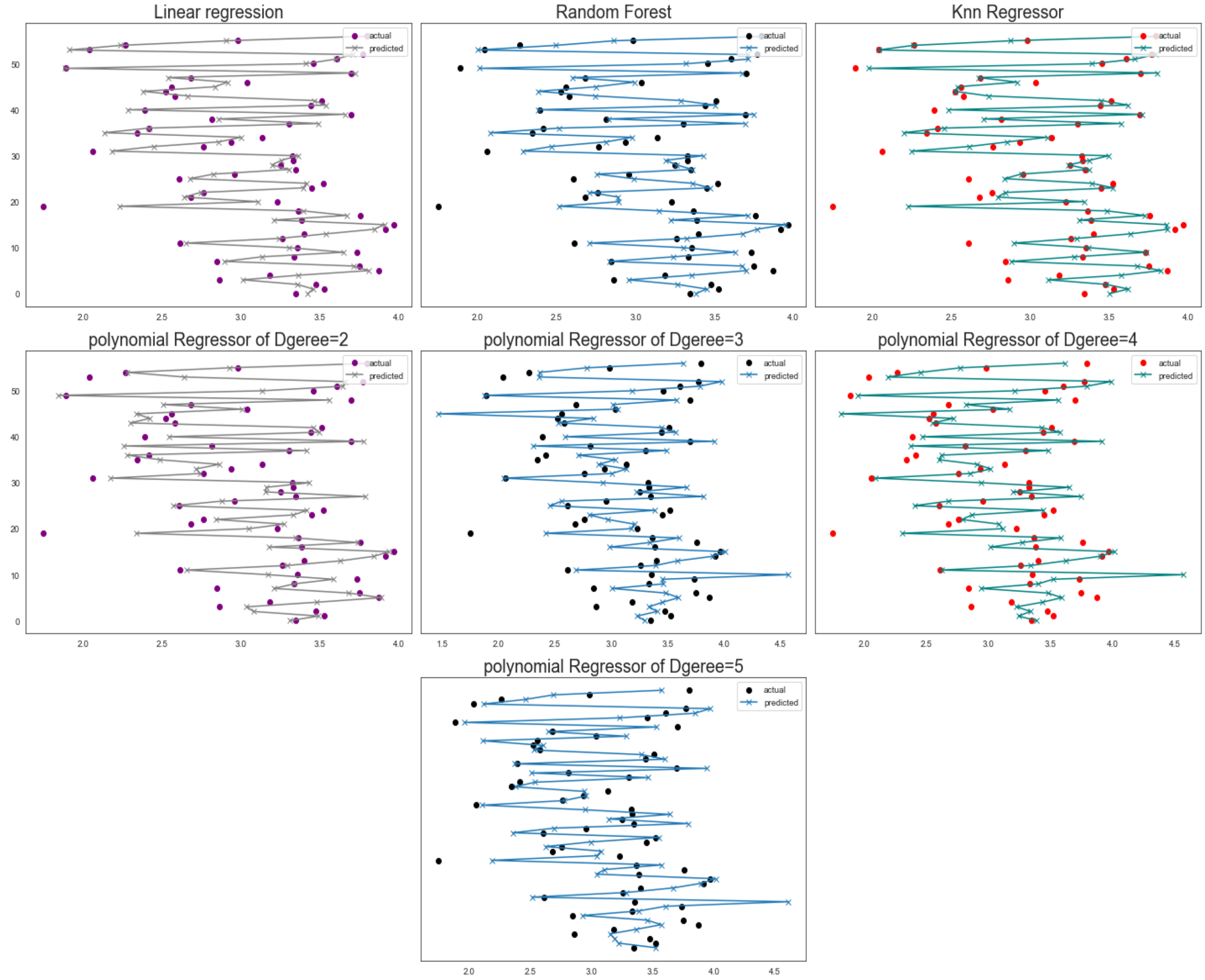
score type	Linear regression	Polynomial:2	polynomial:3	polynomial:4	Knn	Random Forest
train	0.991223	1	1	1	0.953484	0.993563
test	0.988177	0.879522	0.949805	0.95306	0.970313	0.954323
MSE	0.00389837	0.0397246	0.0165506	0.0154773	0.0153374	0.015061

Graphical Comparison Between Algorithms

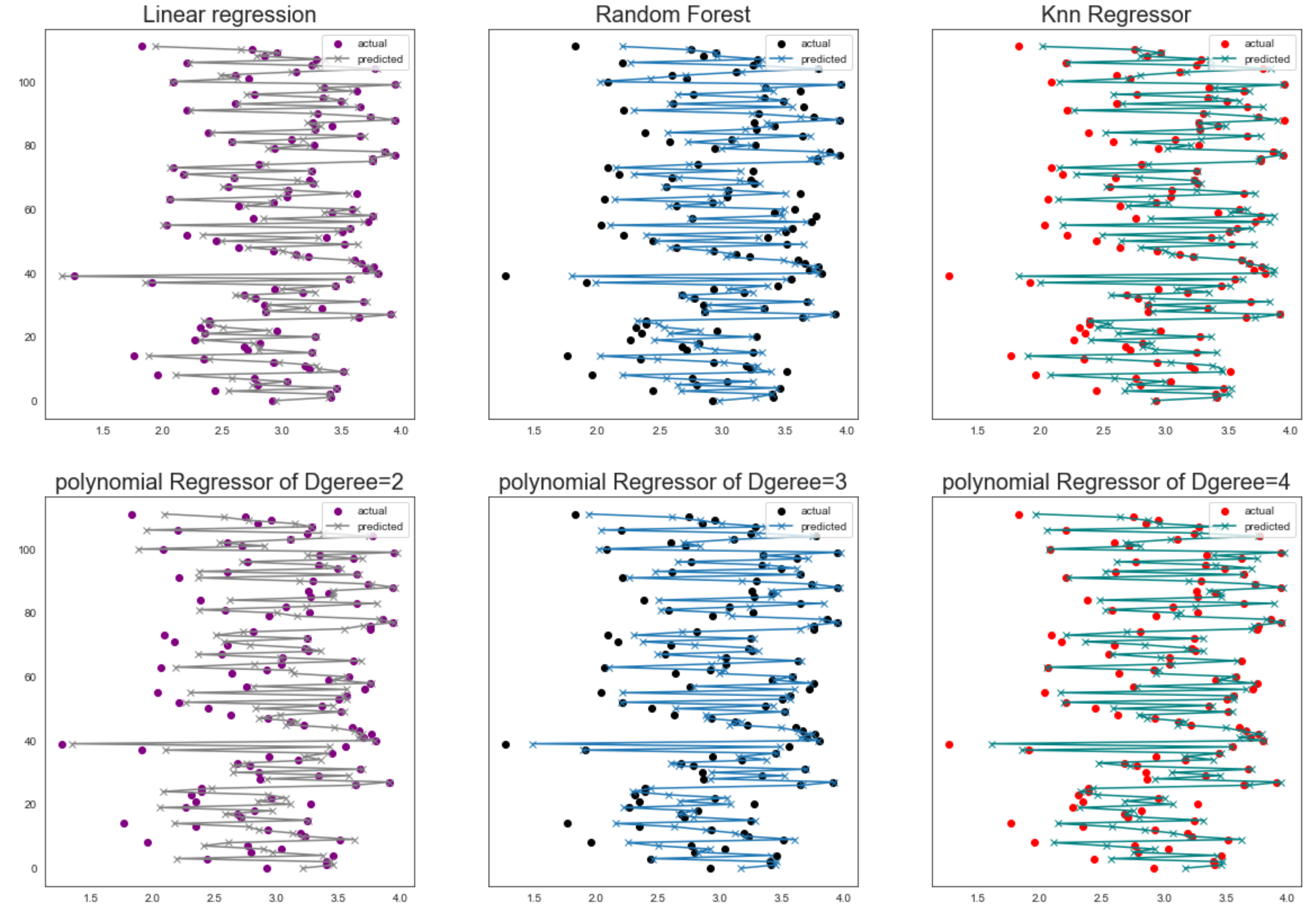
Model 1



Model 2


Linear regression
Random Forest
Knn Regressor
polynomial Regressor of Dgeree=2
polynomial Regressor of Dgeree=3
polynomial Regressor of Dgeree=4
polynomial Regressor of Dgeree=5
actual
predicted
0
10
20
30
40
50
1.5
2.0
2.5
3.0
3.5
4.0
4.5

Model 3



The above graphs demonstrate how well the predicted values of CGPA map to actual values of CGPA on various algorithms.

Comparison Between Algorithms

Model 1

Polynomial regression with degree 2 gives the most better performance on model 1, with lowest mean squared error. However, on higher degrees of polynomials such as 3 or 4 the model starts to overfit resulting in high train accuracy, low test accuracy and high test error. Linear Regression gives a low training and test score when fitted on model 1. KNN also gives a moderate accuracy on model 1. Random forest overfits the model resulting in high training score and low testing score. Model also performs well on KNN algorithm with low mean squared error and high train-test accuracy.

Model 2

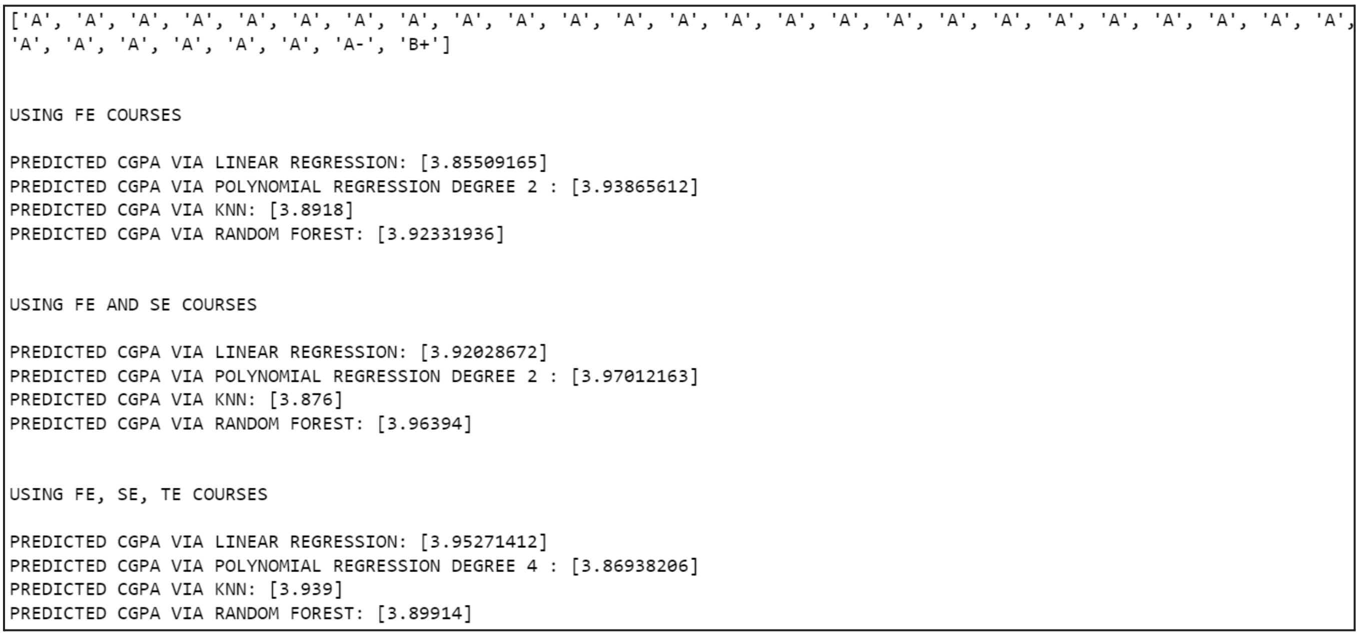
Linear Regression fits best on model 2, resulting in high train-test accuracy and low mean squared error. Polynomial regression leads the model to overfit. Though higher degree polynomials give highest accuracy, this algorithm is computationally expensive. Training time is very high and the number of features are also increased excessively. Random forest also gives the best and improved performance. Model also performs well on KNN algorithm with low mean squared error and high train-test accuracy. However, since the mean squared error of Linear Regression is lower than Random forest and KNN, therefore, Linear Regression algorithm is selected for future predictions.

Model 3

Linear Regression fits best on model 2, resulting in high train-test accuracy and low mean squared error. The model also performs well on polynomial regression. Though higher degree polynomials give highest accuracy, it is computationally expensive. Training time is very high and the number of features are also increased excessively. Random forest and KNN also perform well and do not overfit. However, since the mean squared error of Linear Regression is lower than Random forest and KNN, therefore, Linear Regression algorithm is selected for future predictions.

Execution of User Program

A user program has been implemented to predict the final CGPA of the student using user input. The execution of this program starts with the main() function. This function takes grades as input of all courses from first year to third year. It preprocesses the user input and predicts the final CGPA of the student on the basis of model1, model2 and model3 using different algorithms with their respective optimal parameters generated above. A demonstration of the user program is given below.



```
['A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A', 'A',
'A', 'A', 'A', 'A', 'A', 'A', 'A-', 'B+']


USING FE COURSES

PREDICTED CGPA VIA LINEAR REGRESSION: [3.85509165]
PREDICTED CGPA VIA POLYNOMIAL REGRESSION DEGREE 2 : [3.93865612]
PREDICTED CGPA VIA KNN: [3.8918]
PREDICTED CGPA VIA RANDOM FOREST: [3.92331936]


USING FE AND SE COURSES

PREDICTED CGPA VIA LINEAR REGRESSION: [3.92028672]
PREDICTED CGPA VIA POLYNOMIAL REGRESSION DEGREE 2 : [3.97012163]
PREDICTED CGPA VIA KNN: [3.876]
PREDICTED CGPA VIA RANDOM FOREST: [3.96394]


USING FE, SE, TE COURSES

PREDICTED CGPA VIA LINEAR REGRESSION: [3.95271412]
PREDICTED CGPA VIA POLYNOMIAL REGRESSION DEGREE 4 : [3.86938206]
PREDICTED CGPA VIA KNN: [3.939]
PREDICTED CGPA VIA RANDOM FOREST: [3.89914]
```

Selecting Optimal Result

- For Model-1, as polynomial regression gave the highest accuracy, therefore it can be concluded on the basis of FE grades that the final CGPA of the student would be 3.93.
- For Model-2, as linear regression resulted in highest accuracy, therefore, it can be concluded on the basis of FE and SE grades that the final CGPA of the student would be 3.92.
- For Model-3, as linear regression resulted in highest accuracy, therefore, it can be concluded on the basis of FE, SE and TE grades that the final CGPA of the student would be 3.95.

An alternate approach for selecting the most valid CGPA, is to take average of all the CGPAs of a model predicted using different algorithms. The resulting average CGPA can be regarded as the final predicted CGPA.

Future Extensions

- The accuracy of the models can be improved by applying different algorithms and preprocessing techniques.
- Credit hours can be associated with courses to improve the prediction.
- Increasing the size of data to generate better results.

Conclusion

From the above analysis, it can be concluded that Polynomial regression (degree-2) gives the most accurate result on model 1, whereas, Linear Regression shows a high accuracy and low mean squared error on model 2 and model 3.