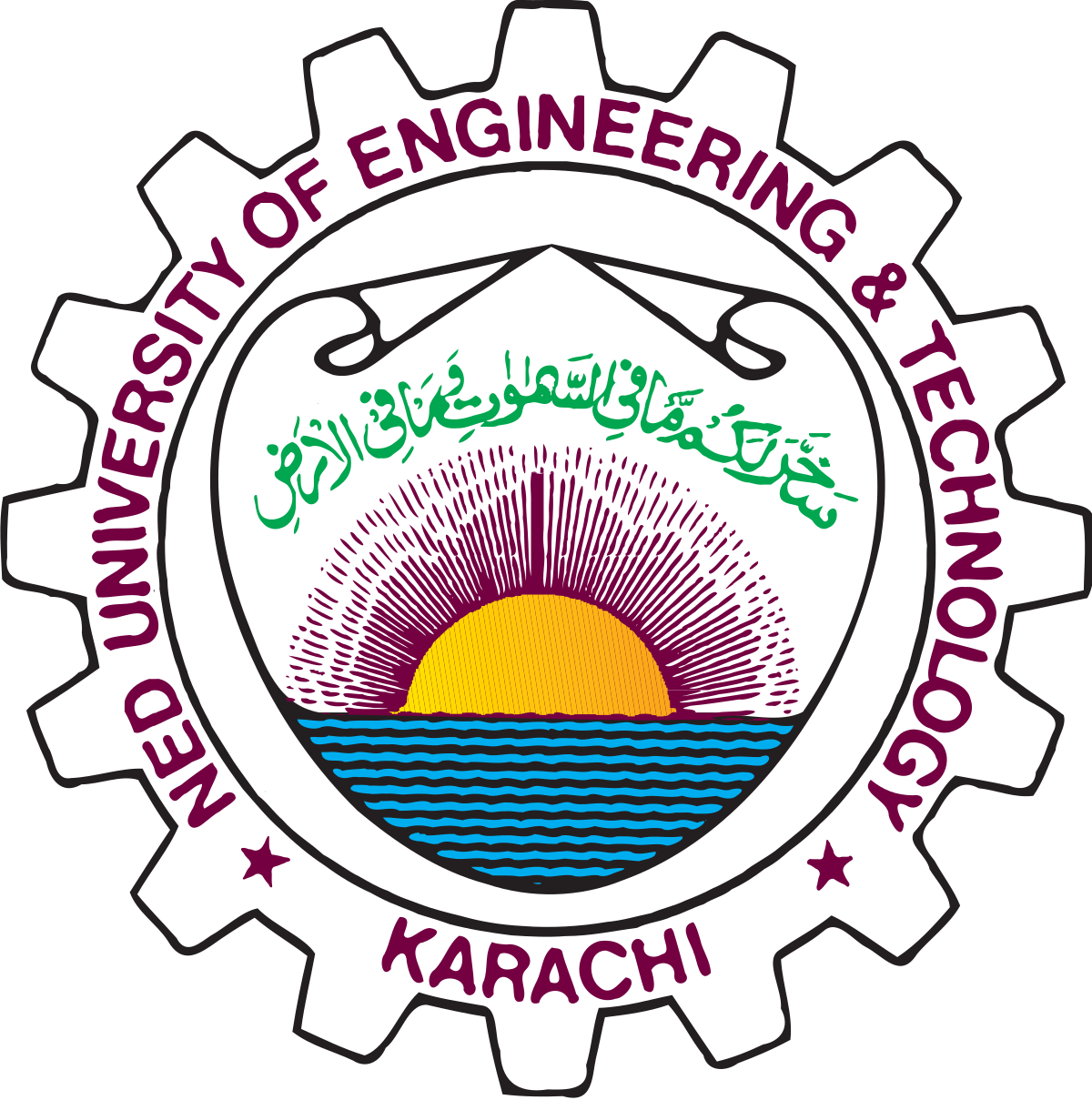
TERM PROJECT

Prediction of Cumulative Grade Point Average



**Team Members**

**UMME HANI (CS-19054)**

**SYEDA RIDA FATIMA(CS-19055)**

SUBMISSION DATE: 24.07.2022

1. DATA PREPROCESSING

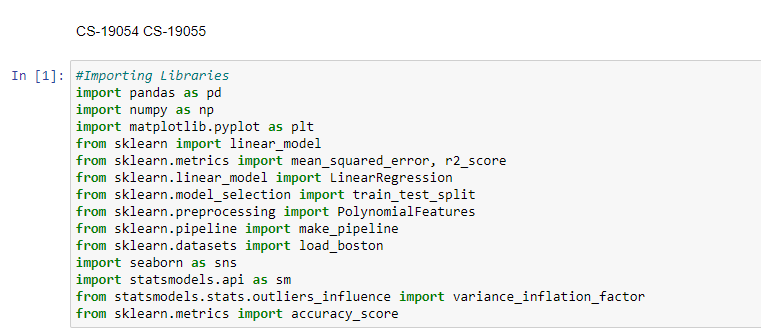
When it comes to creating a Machine Learning model, data preprocessing is the first step marking the initiation of the process. Typically, real-world data is incomplete, inconsistent, inaccurate (contains errors or outliers), and often lacks specific attribute values/trends. This is where data preprocessing enters the scenario – it helps to clean, format, and organize the raw data, thereby making it ready-to-go for Machine Learning models.

1.1 IMPORT ALL THE CRUCIAL LIBRARIES

Since Python is the most extensively used and also the most preferred library by Data Scientists around the world, we’ll show you how to import Python libraries for data preprocessing in Machine Learning.

The predefined Python libraries can perform specific data preprocessing jobs. Importing all the crucial libraries is the second step in data preprocessing in machine learning. The three core Python libraries used for this data preprocessing in Machine Learning are:

* NumPy – For scientific calculation in Python and inserting any type of mathematical operation in the code, we have imported this library.
* Pandas – For data manipulation and analysis we have imported this library.



1.2 IMPORTING DATA SET



df= pd.read\_csv(‘Dataset.csv’)

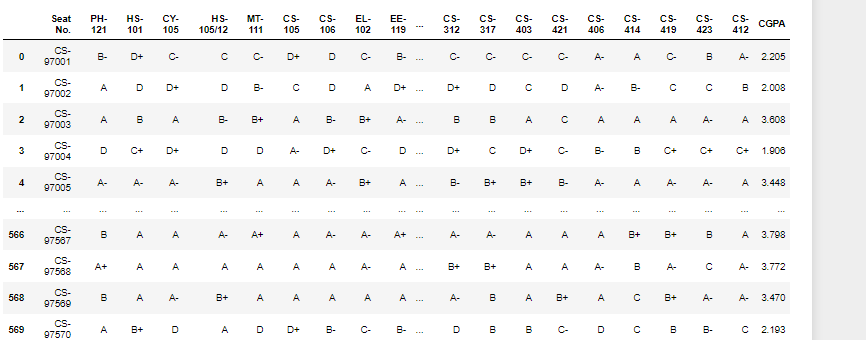
In this line of code, “df” denotes the name of the variable wherein we have stored the dataset. The function contains the name of the dataset as well. Once we execute this code, the dataset will be successfully imported.

During the dataset importing process, there’s another essential thing we have done– extracting dependent and independent variables. For every Machine Learning model, it is necessary to separate the independent variables (matrix of features) and dependent variables in a dataset.

In our data set:

dependent variables= CGPA

Independent variables=GPA of courses start with course code 1,2,3 and 4

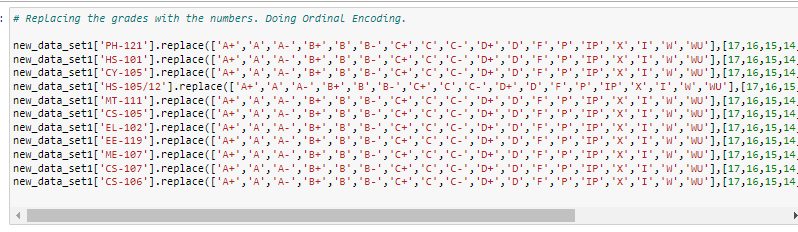


1.3 ENCODING THE CATEGORICAL DATA

We had used **.replace()** function and we have replaced the grades as per following convention:

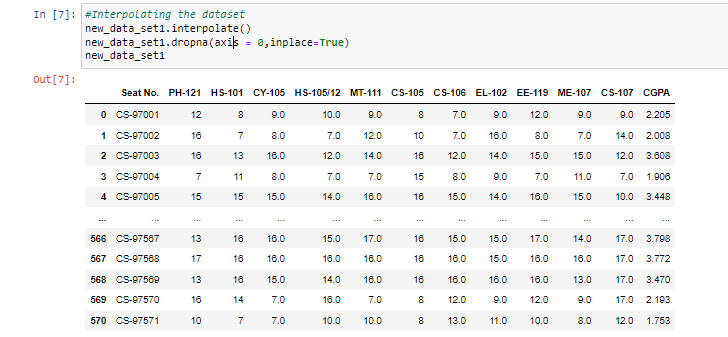
['A+','A','A-','B+','B','B-','C+','C','C-','D+','D','F','P','IP','X','I','W','WU']

[17, 16, 15, 14, 13,12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0]



### 1.4 IDENTIFYING AND HANDLING THE MISSING VALUES

We have deleted the particular row which entirely contains null values, called as interpolation.



1.5 SPLITTING THE DATASET

The dataset was splitted into 2 parts. One is the training part and the other one is the test part. The size of the training set is 80% and the test size is 20%. We used the test\_size function to split the training and testing set.The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. The random state hyperparameter in the train\_test\_split() function controls the shuffling process. With random\_state=None , we get different train and test sets across different executions and the shuffling process is out of control. With random\_state=0 , we get the same train and test sets across different executions.

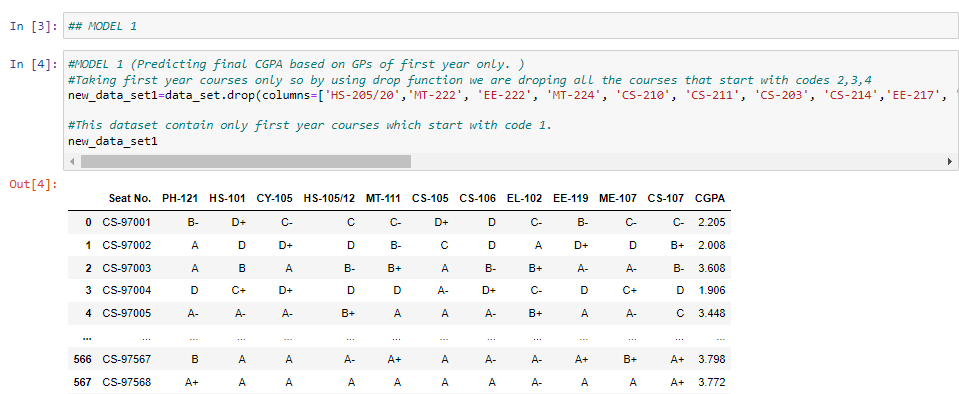


1.6 MODEL SELECTION

We have applied the Machine Learning Algorithm on the courses of First year to predict first year GPA for model 1 and for model 2 we had chosen first year and second year courses to predict second year GPA. These 2 model selections were done randomly.

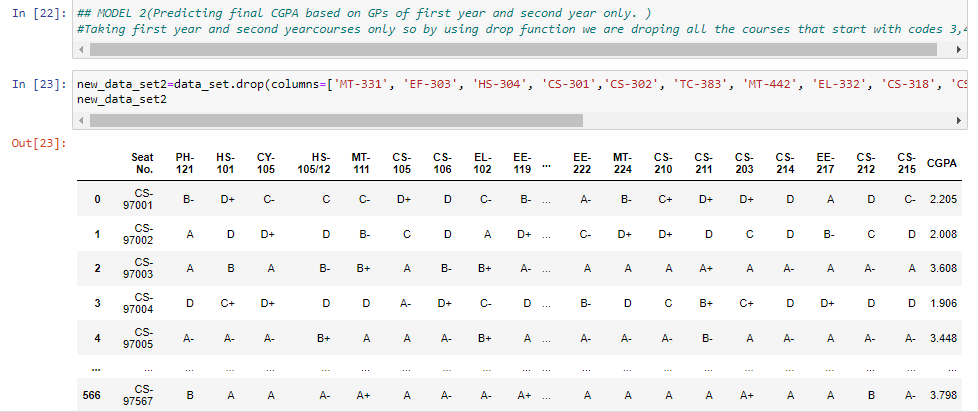
FIRST YEAR MODEL

Model 1 contains courses of the first year only. In the dataset, the courses where course code starts with 1 are FE courses (e.g.: CS-106). In general there are 11 FE courses in the dataset for which we have to predict GPA. The courses are:



FIRST AND SECOND YEAR MODEL

Model 1 contains courses of the first year and second year. In the dataset, the courses where course code starts with 1 are FE courses (e.g.: CS-106) and the courses where course code starts with 2 are SE courses (e.g.: MT-222). There are 11 courses of FE and 11 courses of SE in the dataset. In general this model contains 22 columns (input column) for which we have to predict GPA.



1.7 SELECTION OF MACHINE LEARNING ALGORITHM

We had implemented our models on 2 Algorithms.

* Linear Regression
* Polynomial Regression

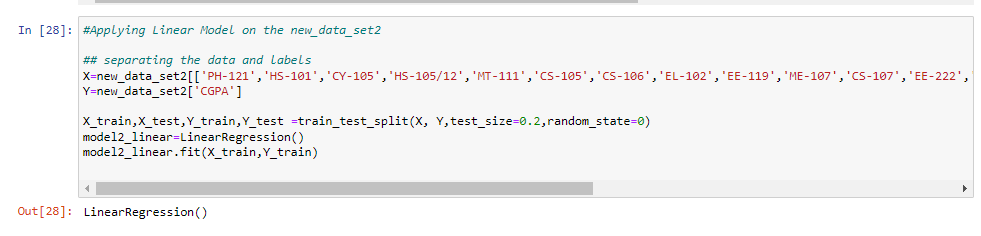
REGRESSION

**Regression analysis** is a statistical technique for determining the relationship between a single dependent (criterion) variable and one or more independent (predictor) variables. The analysis yields a predicted value for the criterion resulting from a linear combination of the predictors.

LINEAR REGRESSION

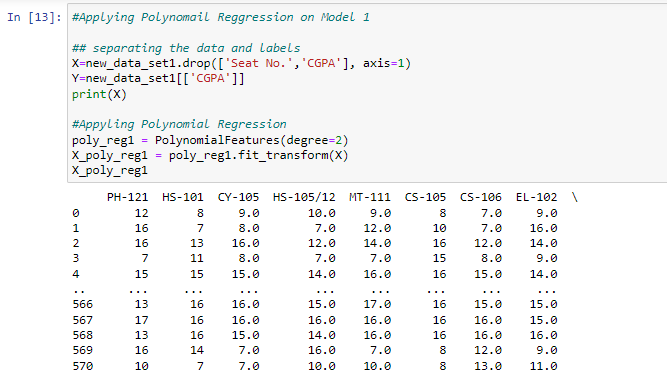
We had chosen a Linear Regression Algorithm to predict the GPA of both models. **Linear regression** is easier to use, simpler to interpret, and we obtain more statistics that help us to assess the model. It is one of the most commonly used predictive modeling techniques.It is represented by an equation 𝑌 = 𝑎 + 𝑏𝑋 + 𝑒, where a is the intercept, b is the slope of the line and e is the error term. This equation can be used to predict the value of a target variable based on a given predictor variable(s).

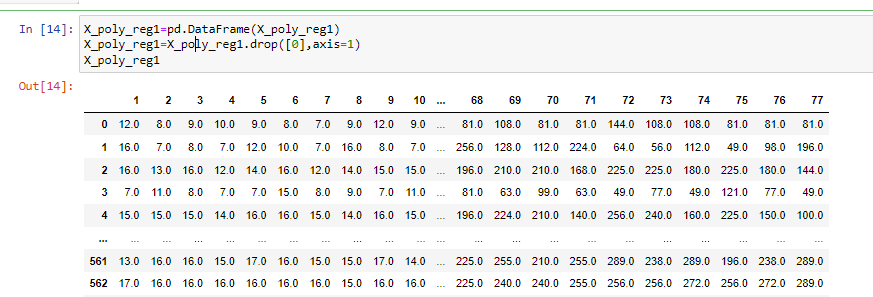


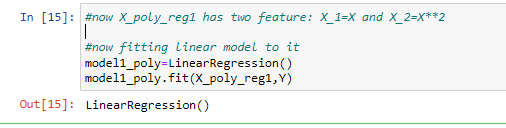


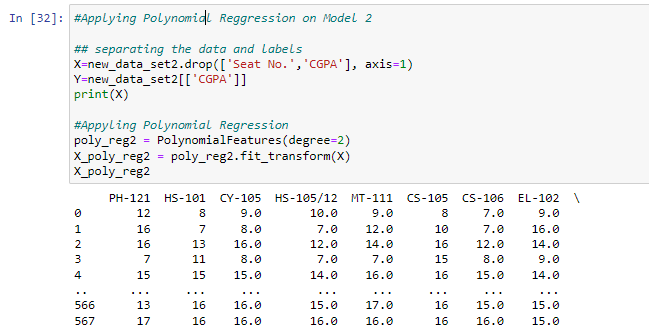
POLYNOMIAL REGRESSION

Polynomial provides the best approximation of the relationship between the dependent and independent variable. That’s why we had chosen a Polynomial Algorithm to predict the GPA of both models. It is used to study the isotopes of the sediments, study the rise of different diseases within any population, study the generation of any synthesis etc. Polynomial Regression is generally used when the points in the data are not captured by the Linear Regression Model and the Linear Regression fails in describing the best result clearly. A Broad range of functions can be fit under it. Polynomial basically fits a wide range of curvature







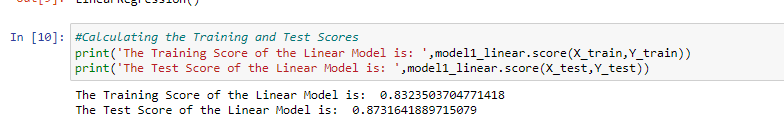


1.7 COMPARISON OF SCORES

FIRST YEAR MODEL

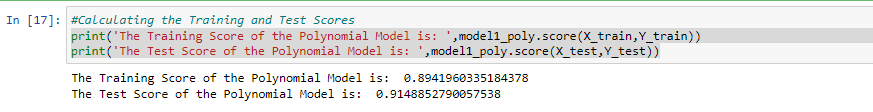
TRAINING SCORE TEST SCORES OF LINEAR REGRESSION ALGORITHM

The training score and testing score of of the Linear Regression model 1 is:



TRAINING SCORE TEST SCORES OF POLYNOMIAL REGRESSION ALGORITHM

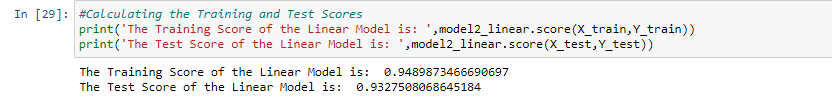
The training score and test score of the Polynomial Regression model is



FIRST YEAR AND SECOND YEAR COURSES MODEL 2

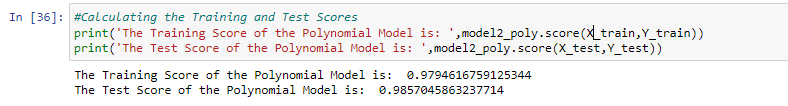
TRAINING SCORE TEST SCORES OF LINEAR REGRESSION ALGORITHM

The training score and testing score of of the Linear Regression model 2 is



TRAINING SCORE TEST SCORES OF POLYNOMIAL REGRESSION ALGORITHM

The training score and test score of the Polynomial Regression model 2 is



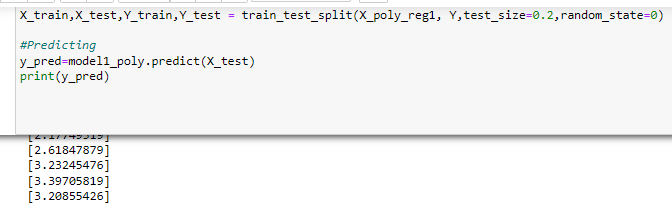
1.8 PREDICTION ON TESTING DATA

MODEL 1

LINEAR REGRESSION

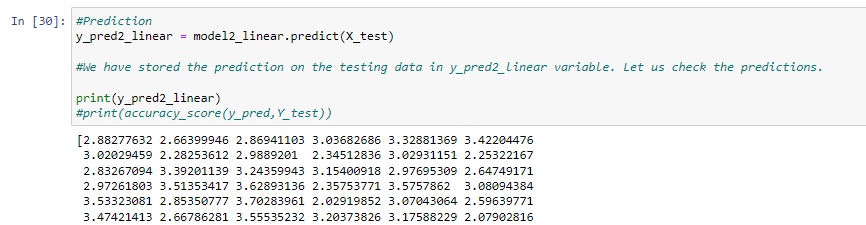


POLYNOMIAL REGRESSION

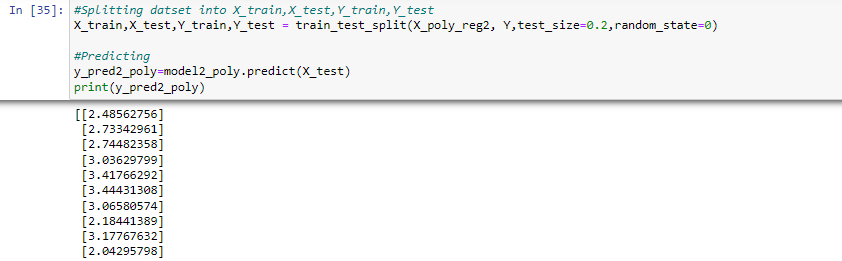


MODEL 2

LINEAR REGRESSION



POLYNOMIAL REGRESSION



1.9 DISTINGUISHING FEATURES

The distinguishing features of our model is that we had calculated the value of R2 Squared and Mean Squared Error. These two are the important features in the statistical analysis of Regression models. We calculated the values of R2 Squared and Mean Squared Error in both our model 1 and model 2.

R SQUARED VALUE

R-Squared (R² or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit). A higher R-squared value will indicate a more useful beta figure. For example, if a stock or fund has an R-squared value of close to 100%, but has a beta below 1, it is most likely offering higher risk-adjusted returns.

MEAN SQUARED ERROR

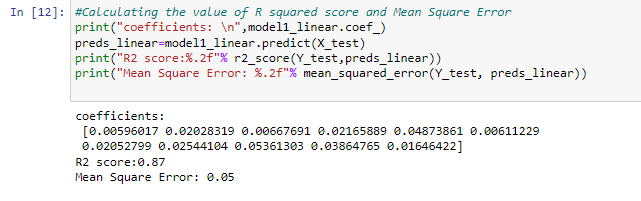
The mean squared error (MSE) tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the “errors”) and squaring them. The squaring is necessary to remove any negative signs. The smaller the mean squared error, the closer you are to finding the line of best fit. Depending on your data, it may be impossible to get a very small value for the mean squared error.

1.10 COMPARISON OF SCORES OF R2 SQUARE AND MEAN SQUARE ERROR OF BOTH MODELS

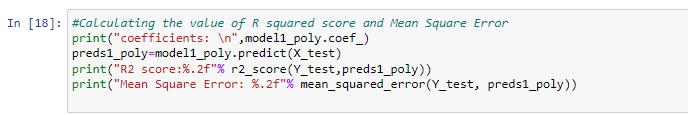
We had done a comparison on the value of R2 squared and MSE on both the Algorithms of model 1 and Model 2. This comparison tell us which model and algorithm was fitting best.

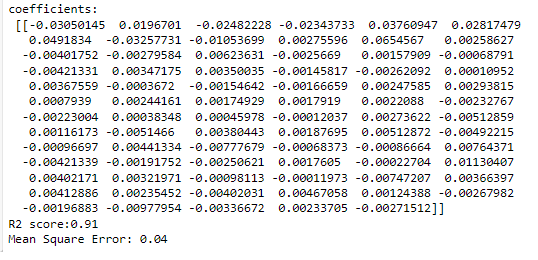
COMPARISON OF MODEL 1 ALGORITHMS

LINEAR REGRESSION



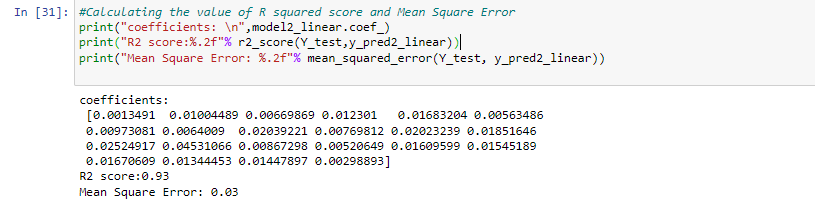
POLYNOMIAL REGRESSION



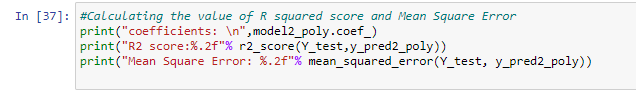


COMPARISON OF MODEL 2 ALGORITHMS

LINEAR REGRESSION



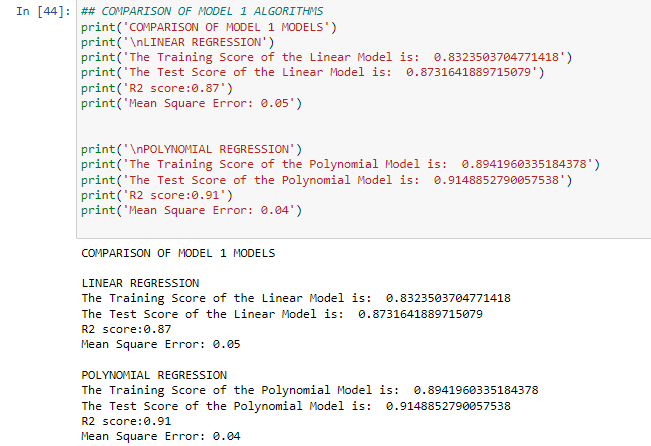
POLYNOMIAL REGRESSION



1.11 COMMENT ON THE IMPLEMENTED ALGORITHMS

MODEL 1

The training and testing score of Polynomial Regression is greater than Linear Regression. It means that Polynomial Regression best fits the model. The value of R2 square and MSE of Polynomial is also greater than Linear Regression. The higher the value of R2 Square and the lesser the value of MSE the better the model it is. Both the models are predicting good. If the training score is better than the testing score then the model will overfit and if the training score is bad and good generalization on testing data then the model is underfit.

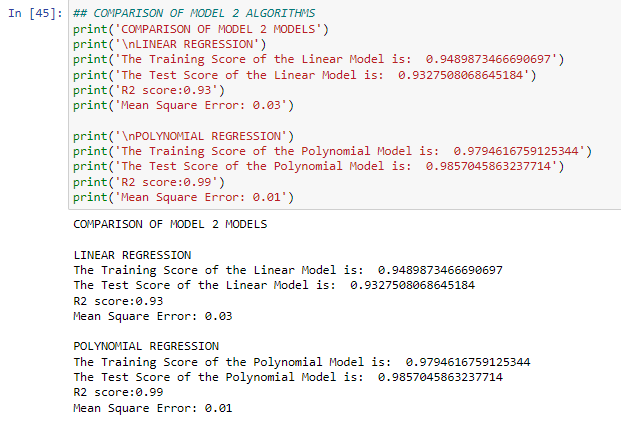


SOLUTION OF UNDERFITTING

Our model gives a good accuracy score on both training and testing data but the score of training data is less than the score on the generalization of testing data then our model is underfit. To resolve underfitting we will add more data points in our data sets.

MODEL 2

The training and testing score of Polynomial Regression is greater than Linear Regression. It means that Polynomial Regression best fits the model. The value of R2 square and MSE of Polynomial is also greater than Linear Regression. The higher the value of R2 Square and the lesser the value of MSE the better the model it is. Both the models are predicting good. If the training score is better than the testing score then the model will overfit and if the training score is bad and good generalization on testing data then the model is underfit.



SOLUTION OF OVERFITTING

Our model gives a good accuracy score on both training and testing data but the score of training data is less than the score on the generalization of testing data in polynomial regression and the score of training dataset is higher than the testing dataset then our model is overfit .To resolve overfitting we will regularize the features.To regularize we will use ridge regularization techniques in which we will reduce the weight of the feature rather than eliminate them.

1.12 USER INTERFACE

We had deployed our trained model on a simple interface. We use simple coding for taking inputs , store them in a list/array and then pass it to our trained model and with the help of .predict() function we predict the CGPA

1.13 FUTURE EXTENSION

For the future extension of this project we will deploy our trained models on heroku and also use different python libraries for the interactive interface. We will extend our model by giving constructive feedback and convert the CGPA into grades. We also decided to try new machine algorithms.