**PREDICTION OF UNIVERSITY ADMISSIONS USING REGRESSION ANALYSIS**

**1.Introduction:**

Today, there are many students who travel to foreign countries to pursue higher education. It is necessary for the students to know what are their chances of getting an admit from such universities. Similarly, it is necessary from the university's perspective to know from the total number of applications, what will be the number of applicants who could get an admit based on certain criteria. Currently, students manually perform statistical analysis before applying to universities to find out the probable chance of getting an admit. Also, universities manually check and count the total number of applicants who could get admission into the university. These methods are slow and certainly not very consistent for students and universities to get an actual result. This method is also prone to human error and thus accounts for some inaccuracies. Since the frequency of students studying abroad has increased, there is a need to employ more efficient systems that handle the admission process accurately from both perspectives. Our goal is to apply machine learning on the Admission Data using Regression Analysis.

**1.1 Objectives of research:**

University admission predictor is a boon to many students. Many students from different places find it hard to know whether they get admission to their preferred university or not.

So, this helps them a lot and eases out their fear predicting how much is their chance of admission in their preferred university.

Students normally get the admission based on different aspects like GRE Score, TOEFL Score, CGPA, University rating, etc., So based on these parameters the students chance of admission in their preferred university can be predicted.

**1.2 Problem Statement:**

Nowadays many students are facing problems regarding the correct prediction of their chance of admission in the universities and they don't have any knowledge of what input plays a major role in getting their admission in the university.

**1.3 Scope of the problem:**

The data set provided to us doesn't contain any missing values and if there is any change in the given inputs like GRE Score, TOEFL Score, University rating, CGPA, etc it will affect the Chance of Admit for the student to get admission in their preferred universities.

**2.Review of the literature:**

This section provides the literature review of the work that has previously done on predicting the chances of student enrolments in the universities. There have been several projects and studies performed on topics related to students admission into universities. (Bibodi et al. (n.d.)) used multiple machine learning models to create a system that would help the students to shortlist the universities suitable for them also a second model was created to help the colleges to decide on enrolment of the student. Nave Bayes algorithm was used to predict the likelihood of success of an application, and multiple classification algorithms like Decision Tree, Random Forest, Nave Bayes, and SVM were compared and evaluated based on their accuracy to select the best candidates for the college. Limitation of this research as that it did only relied on the GRE, TOEFL, and CGPA of the student and missed on taking into consideration other important factors like SOP and LOR documents quality, past work experience, technical papers of the students, etc

**3.Data Collection:**

We used the second method of data collection.The data set which we are using for the development of the model is provided by our organization.

**4.Data Analysis & Interpretation:**

**4.1: Exploratory Data Analysis:**

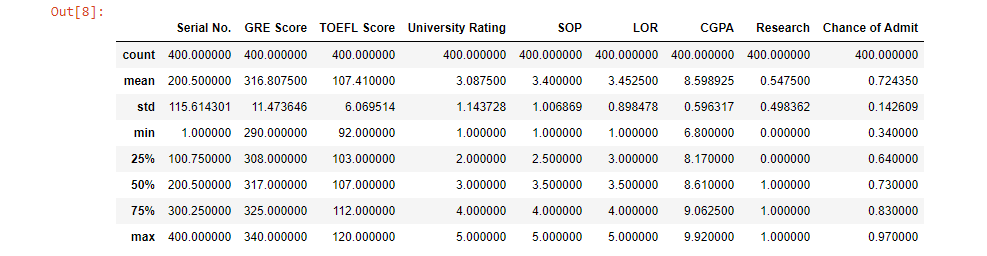
Exploratory Data Analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

It refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

In this we use different visualization techniques like Histograms, Box plots, Heat maps, Scatter plots, etc., present in the libraries matplotlib, Seaborn. And we also get different statistical interpretations like mean, median, max, min, etc., using functions like describe.

Today, there are many students who travel to foreign countries to pursue higher education. It is necessary for the students to know what are their chances of getting an admit from such universities. Similarly, it is necessary from the university’s perspective to know from the total number of applications, what will be the number of applicants who could get an admit based on certain criteria. Currently, students manually perform statistical analysis before applying to universities to find out the probable chance of getting an admit. Also, universities manually check and count the total number of applicants who could get an admit into university. These methods are slow and certainly not very consistent for students and universities to get an actual result. This method is also prone to human error and thus accounts for some inaccuracies. Since the frequency of students studying abroad has increased, there is a need to employ more efficient systems which handle the admission process accurately from both perspectives. Our goal is to apply machine learning on the Admission Data using Regression Analysis.

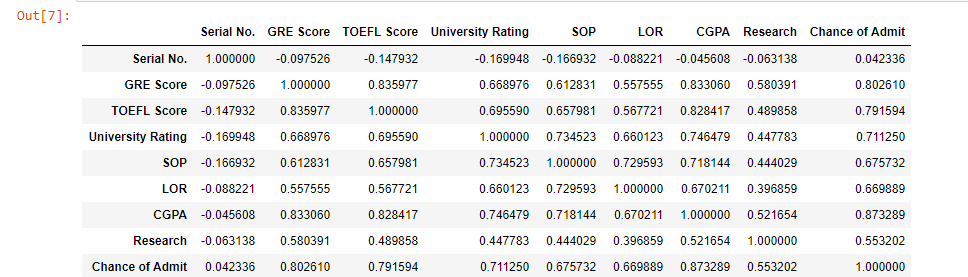
Table4.1.1:Descriptive Analysis



Interpretation: From the above description table 4.1.1 we get mean, std, min,max,count and quartile ranges.

As the mean and median (i.e 50th percentile) values are close to each other we can say that there are no outliers in the data.

Table4.1.2:Correlation



Interpretation: From the above correlation table 4.1.2 the correlation values between Chance of Admit and GRE Score, TOEFL Score, University Rating, SOP, LOR,CGPA are above 0.65 so they are strongly correlated and with Research is 0.55 which is also above 0.5 so it is also strongly correlated and all the above features are positively correlated which says that with increase of value of features, increases the Chance of Admit of Student.

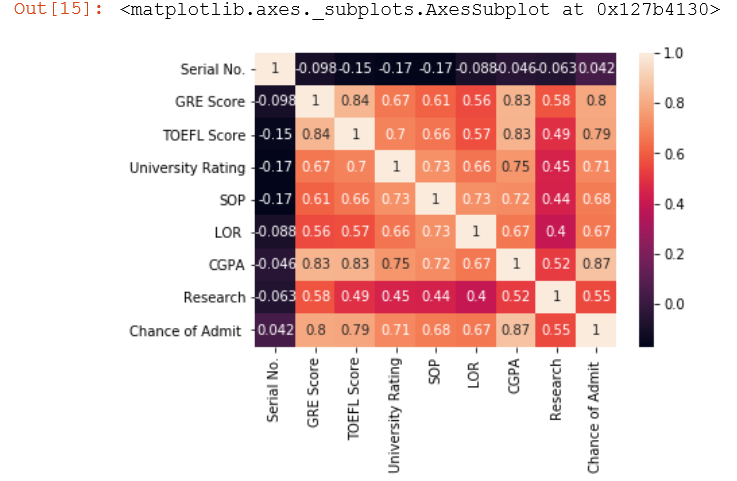
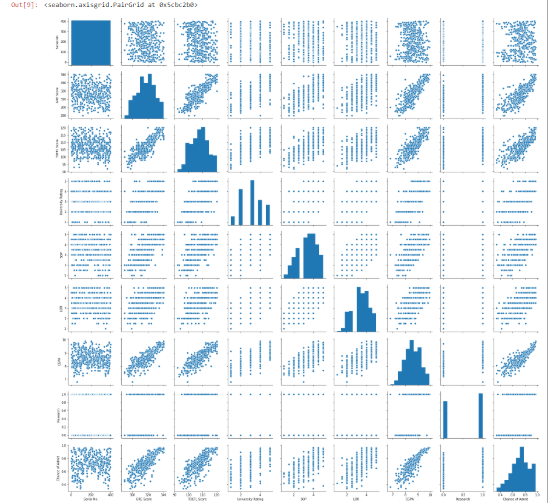


Figure 4.1.3:Heat Maps

Interpretation:From the above figure 4.1.3 we can say that as the colours of GRE Score,TOEFL Score, CGPA with Chance of Admit are brighter which says that they have high correlation.

Next follows the University Rating, SOP,LOR which are little darker when compared to the above three features but are bright so we can say that there is a strong correlation with Chance of Admit.

Table4.1.4: PairPlots(Scatter plot matrix)



Interpretation: From the above scatter plot we can say that as the values of GRE Score, TOEFL Score, CGPA increases Chance of Admit also increases and the features University Rating, SOP, LOR have a considerable affect on Chance of Admit so there is a chance that Chance of Admit also increases.

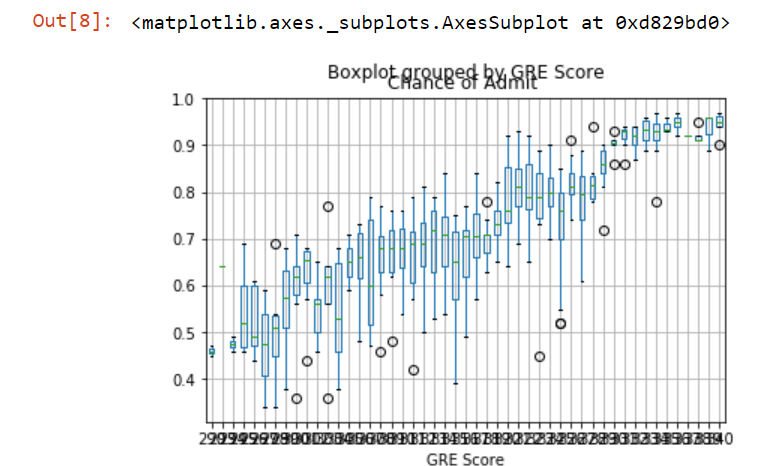


Figure 4.1.4: Boxplot visualization between Chance Of Admit and GRE Score

Interpretation: From the above boxplots, the boxplot for the GRE Score 295 has the minimum and maximum values the least for Chance of Admit and the highest is for 340.As the GRE Score is increasing the boxplots showing minimum and maximum values for the given GRE Score is increasing.

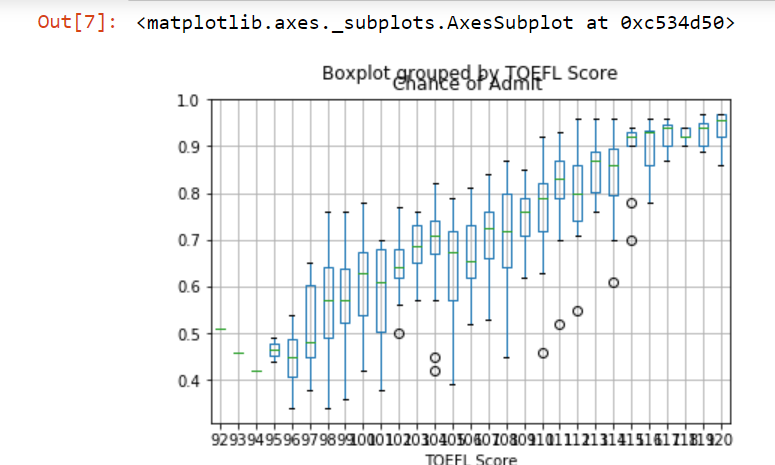


Figure 4.1.5:Boxplot visualization between Chance Of Admit and TOEFL Score

Interpretation:From the above boxplots, the boxplot for the TOEFL Score 96 has the minimum and maximum values the least for Chance of Admit and the highest is for 120.As the TOEFL Score is increasing the boxplots showing minimum and maximum values for the given TOEFL Score is increasing.

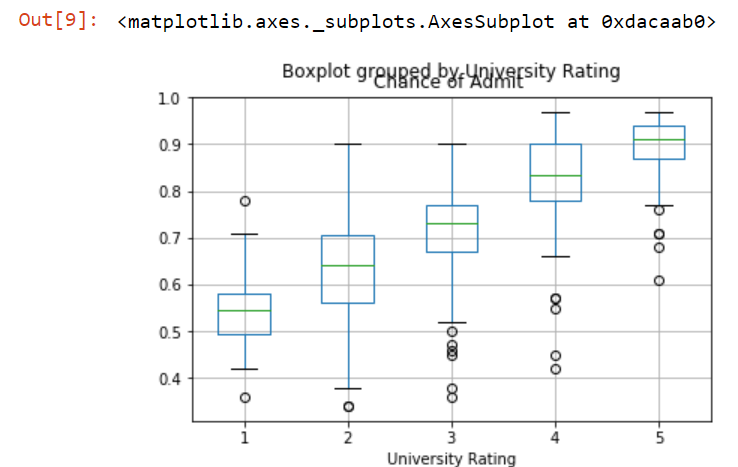


Figure 4.1.6: Boxplot visualization between Chance Of Admit and University Rating

Interpretation: From the above boxplots, the boxplot for the GRE Score 295 has the minimum and maximum values the least for Chance of Admit and the highest is for 340.As the GRE Score is increasing the boxplots showing minimum and maximum values for the given GRE Score is increasing.

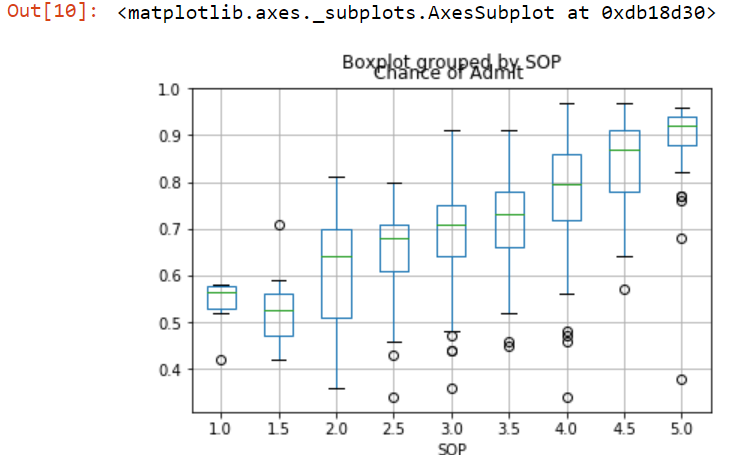


Figure 4.1.7:Boxplot visualization between Chance Of Admit and SOP

Interpretation: From the above boxplots, the boxplot for the SOP 1.5 has the minimum and maximum values the least for Chance of Admit and the highest is for 5.0.As the SOP is increasing the boxplots showing minimum and maximum values for the given SOP is increasing.

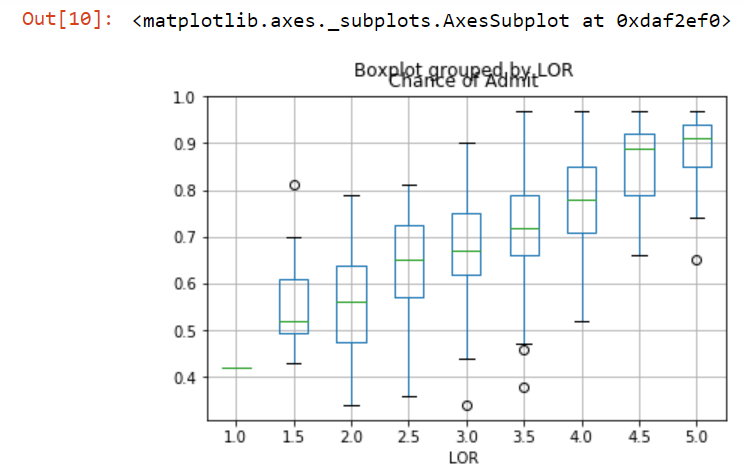


Figure 4.1.8: Boxplot visualization between Chance Of Admit and LOR

Interpretation: From the above boxplots, the boxplot for the LOR 2 has the minimum and maximum values the least for Chance of Admit and the highest is for 5.As the LOR is increasing the boxplots showing minimum and maximum values for the given LOR is increasing.

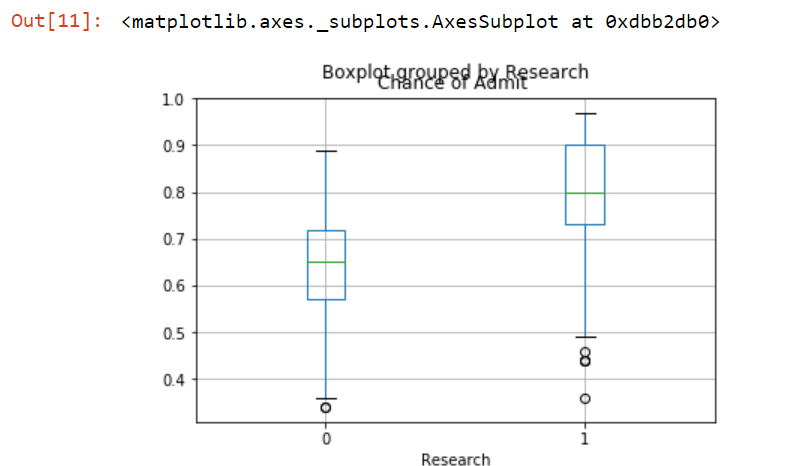


Figure 4.1.10: Boxplot visualization between Chance Of Admit and Research

Interpretation: From the above boxplots, the boxplot for the Research 0 has the minimum and maximum values the least for Chance of Admit and the highest is for 1.As the Research is increasing the boxplots showing minimum and maximum values for the given Research is increasing.

We can say that if a person has a research paper has high chance of admit than who doesn’t have a research paper.

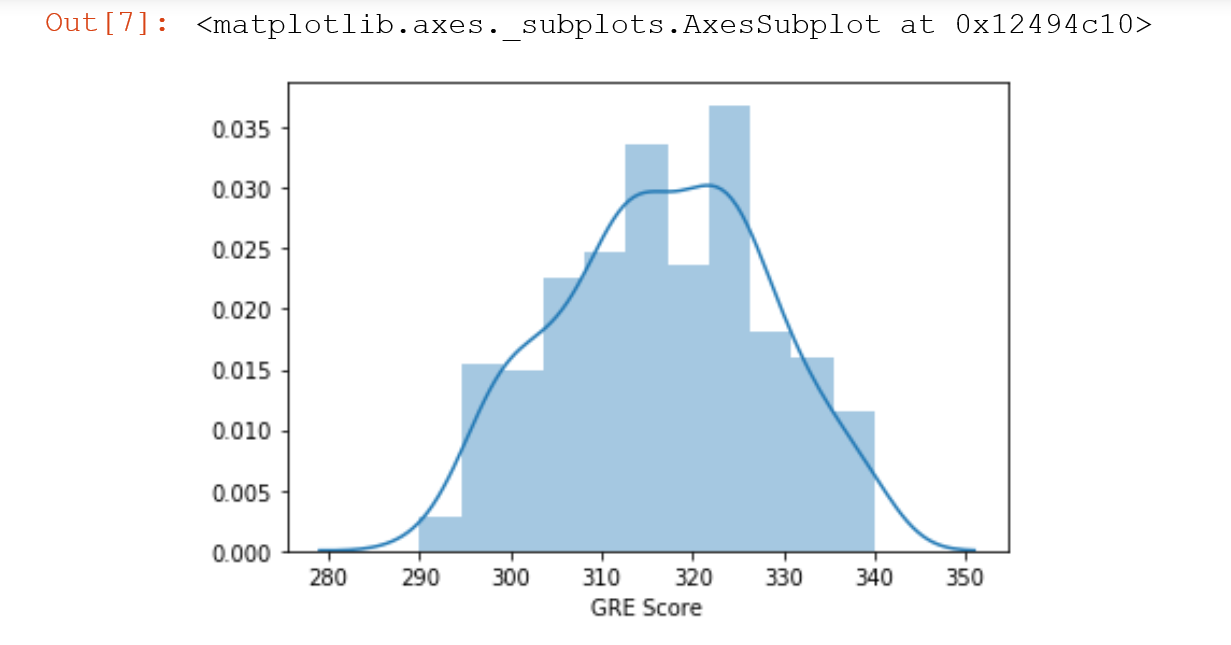


Figure 4.1.11: Distribution of GRE Score

Interpretation: From above figure we can say that the scores are between 290 to 340 and there are more people with GRE Score between 325 to 330.

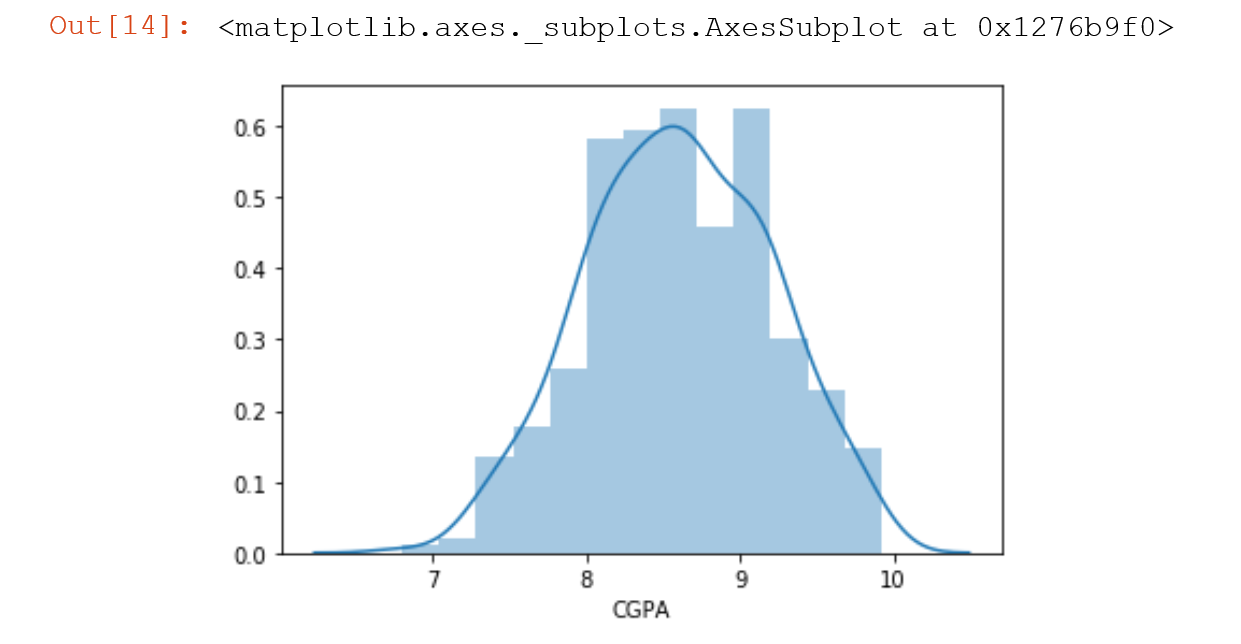


Figure 4.1.12: Distribution of CGPA

Interpretation: From above figure we can say that the scores are between 7 to 10 and there are more people with CGPA between 9 to 9.5.

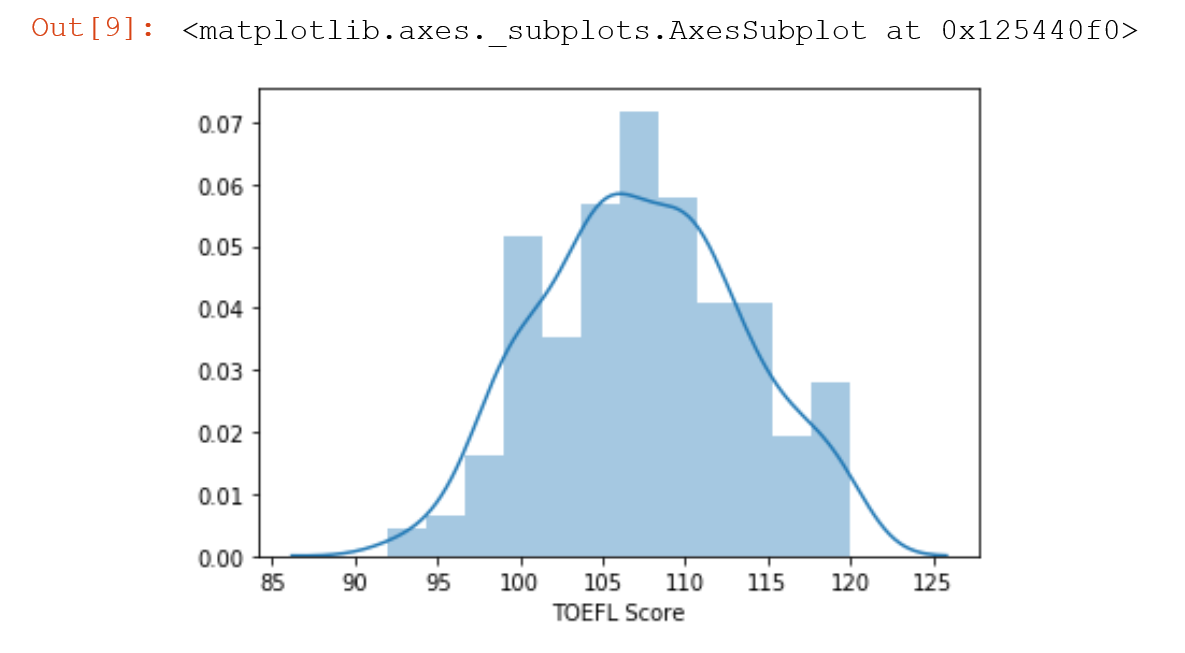
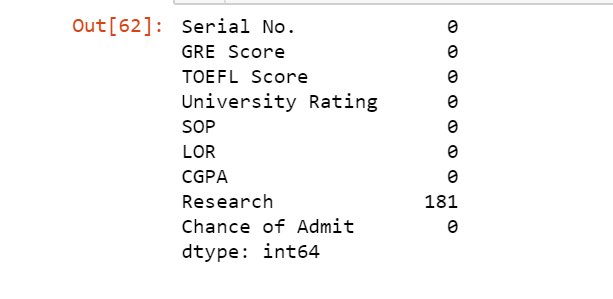


Figure 4.1.13: Distribution of TOEFL Score

Interpretation: From above figure we can say that the scores are between 95 to 120 and there are more people with TOEFL Score between 106 to 109.

**4.2: Data Cleaning:**

The main aim of Data Cleaning is to identify and remove missing values & Null values, Zeros in the data in order to create a reliable dataset. This improves the quality of the training data for analytics and enables accurate decision-making.

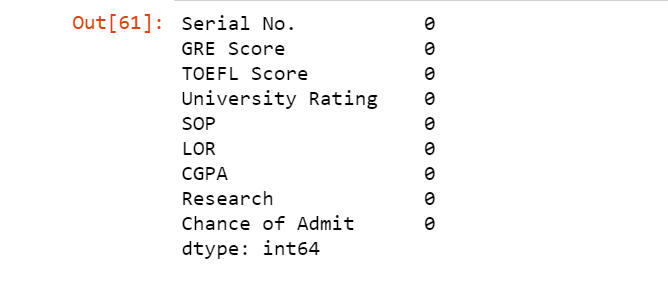


Table4.2.1:Data Cleaning

Interpretation: From the above figure we can say that there are no missing or null values and even zeros except in Research which is categorical data in the given dataset.

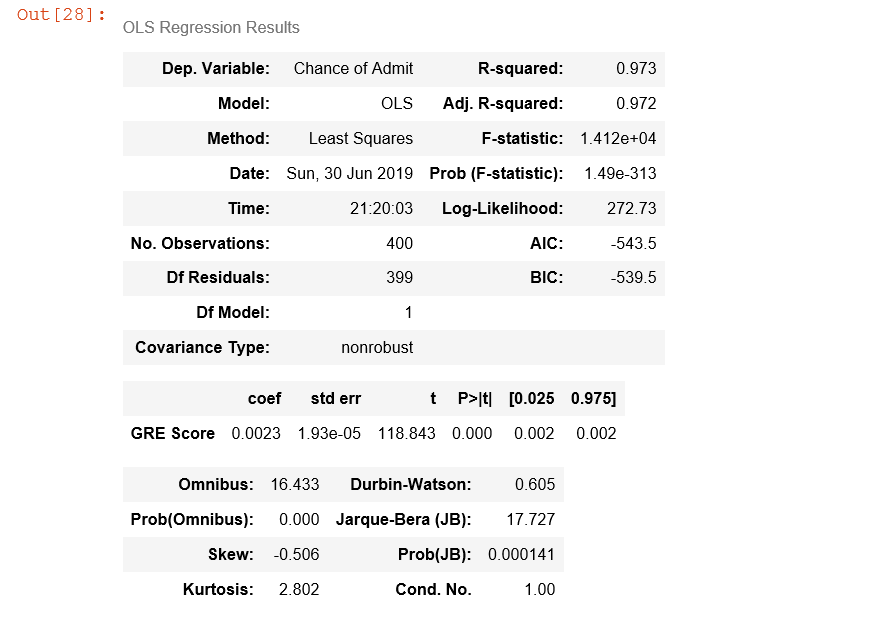
**4.3: Predictive Modeling:**

Predictive modeling is the general concept of building a model that is capable of making predictions. Typically, such a model includes a machine learning algorithm that learns certain properties from a training dataset in order to make those predictions.  
Predictive modeling can be divided further into two sub areas: Regression and pattern classification. Regression models are based on the analysis of relationships between variables and trends in order to make predictions about continuous variables, e.g., the prediction of the maximum temperature for the upcoming days in weather forecasting.  
In contrast to regression models, the task of pattern classification is to assign discrete class labels to particular observations as outcomes of a prediction. To go back to the above example: A pattern classification task in weather forecasting could be the prediction of a sunny, rainy, or snowy day.

**4.3.1:Linear Regression**

The linear regression explains the relationship between**one continuous dependent variable** (y) and **oneindependent variable(x).**

Figure 4.3.1.1:OLS Model between GRE Score and Chance of Admit



Interpretation:From the above OLS model we can conclude that R square and Adj R Square values are high so the model is good model and P value is less than 5% acceptance region i.e., 0.05 alpha then null hypothesis is rejected and alternative hypothesis is accepted i.e., there is a significant relationship among input and output variables.

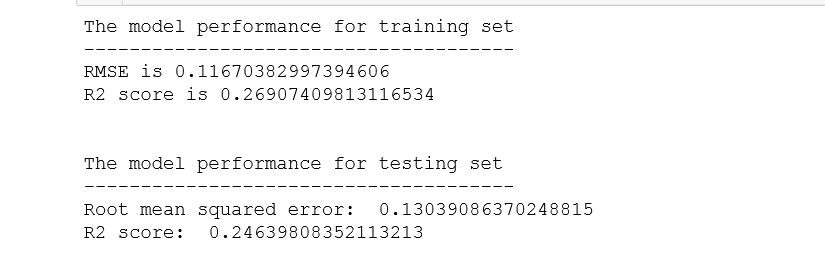
Figure 4.3.1.2:Accuracy between GRE Score and Chance of Admit





Interpretation::From the above we can say that Accuracy for training data is 62% and Accuracy for testing model is 65% and they have less accuracy so it is not a good model

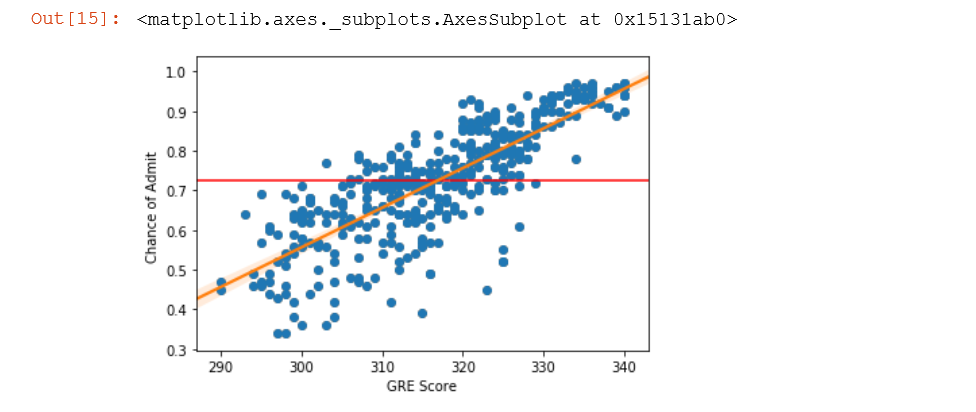
Figure 4.3.1.3:RMSE and R2 Score for GRE Score and Chance of Admit



Interpretation:

From the above model metrics for the regression analysis i.e., Root mean square error and R squared values of the train and test data we can say that the RMSE for the training is little high(0.11) and R squared is little low(0.26). So the model is not good for the training data.

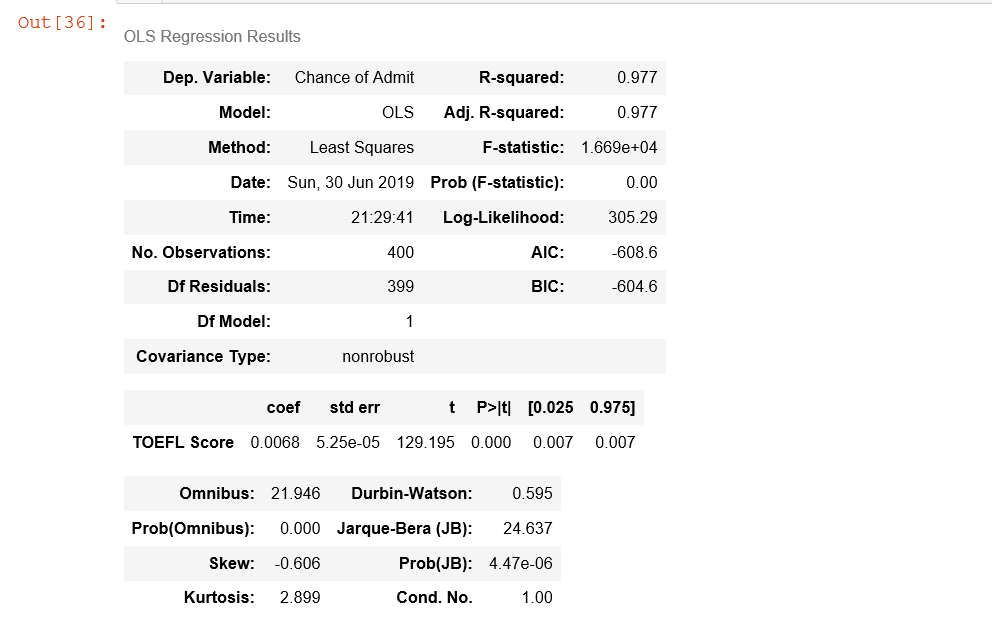
Figure 4.3.1.4:Regression line for GRE Score and Chance of Admit



Interpretation:The above figure shows the regression line for GRE Score and Chance of Admit and based on this regression line we can predict the chance of admit for a new value of GRE Score.

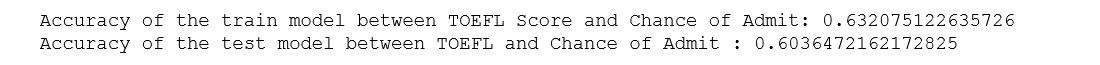
But this regression line is of no use as the accuracy of this model is very low and predicting the value for this model with lower accuracy is useless.

Figure 4.3.1.5:OLS Model between TOEFL Score and Chance of Admit



Interpretation: From the above OLS model we can conclude that R square and Adj R Square values are high so the model is good model and P value is less than 5% acceptance region i.e., 0.05 alpha then null hypothesis is rejected and alternative hypothesis is accepted i.e., there is a significant relationship among input and output variables.

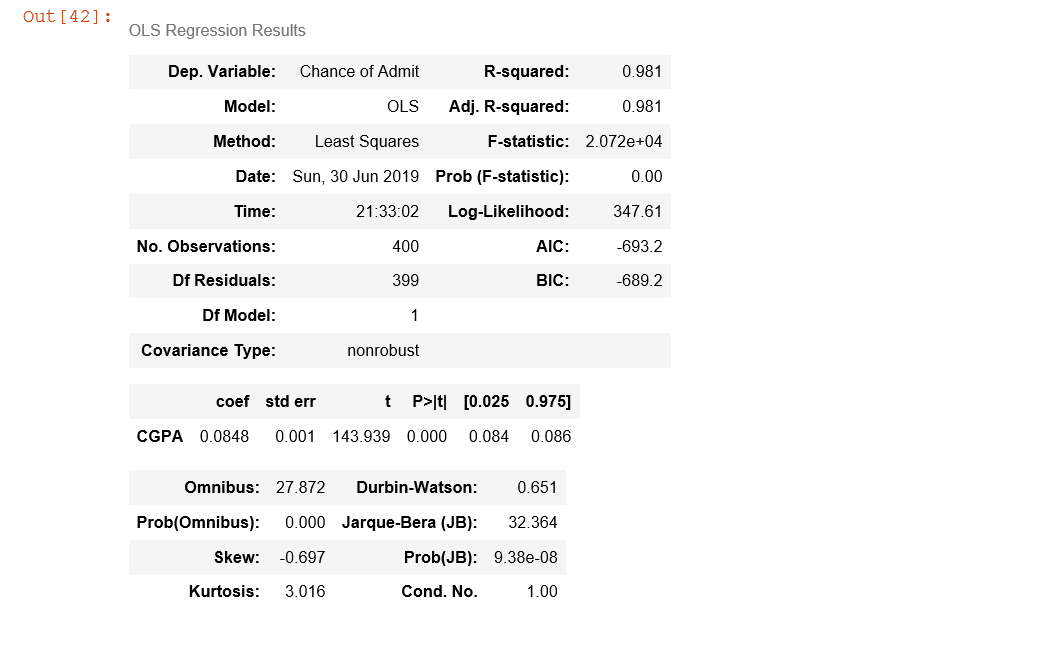
Figure 4.3.1.6:Accuracy between TOEFL Score and Chance of Admit



Interpretation::From the above we can say that Accuracy for training data is 63% and Accuracy for testing model is 60% and they have less accuracy so it is not a good model

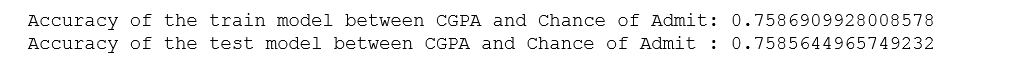
Drawing regression line is of no use as the accuracy of this model is very low and predicting the value for this model with lower accuracy is useless.

Figure 4.3.1.7:OLS Model between CGPA and Chance of Admit



Interpretation: From the above OLS model we can conclude that R square and Adj R Square values are high so the model is good model and P value is less than 5% acceptance region i.e., 0.05 alpha then null hypothesis is rejected and alternative hypothesis is accepted i.e., there is a significant relationship among input and output variables.

Figure 4.3.1.8:Accuracy between CGPA and Chance of Admit



Interpretation::From the above we can say that Accuracy for training data is 75% and Accuracy for testing model is 75% and these both accuracy are same but the accuracy is less so it is not a good model.

Drawing regression line is of no use as the accuracy of this model is very low and predicting the value for this model with lower accuracy is useless.

**4.3.2:Multiple Linear Regression**

The multiple linear regression explains the relationship between**one continuous dependent variable** (y) and **two or more independent variables**(x1, x2, x3… etc)**.**

Features: GRE Score, ToeflScore,CGPA

Figure 4.3.2.1:OLS Model

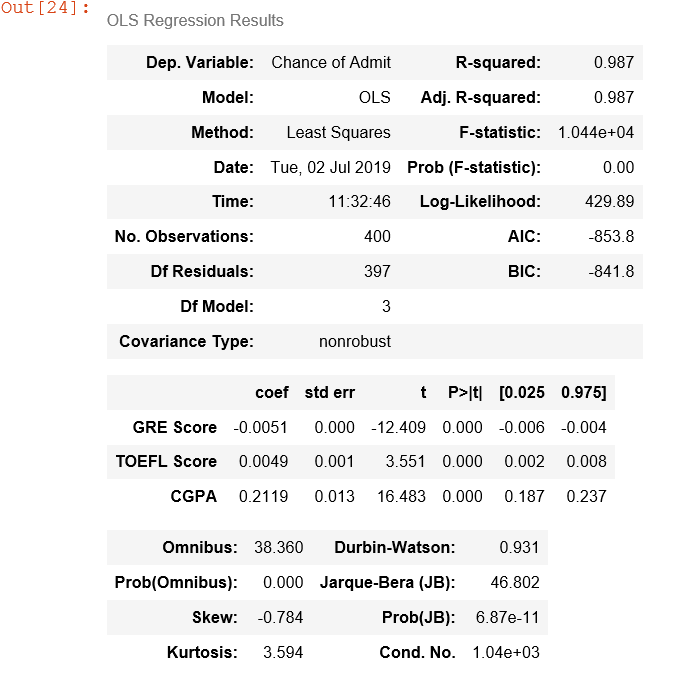


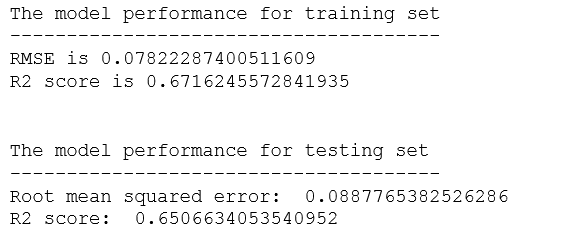
Figure 4.3.2.2: Accuracy for training and testing data.



Interpretation: From the above we can say that Accuracy for training data is 78% and Accuracy for testing model is 77% and these both accuracies are almost near so it is a good model for predicting and is not much biased model.

But is having less values so we cannt use the model based on the values of accuracy values.

Figure 4.3.2.3:RMSE and R2 Score

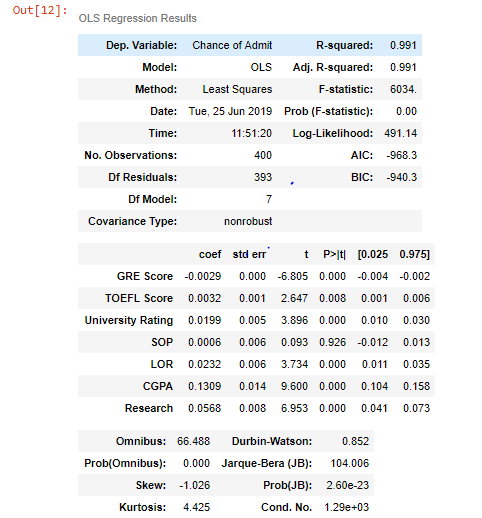


Interpretation: From the above model metrics for the regression analysis i.e., Root mean square error and R squared values of the train and test data we can say that the RMSE for the training is very low(0.07) and R squared is high(0.67). So the model is not so good for the training data i.e.,.For the testing data also the RMSE is low(0.088) and R squared is high(0.65).

So if we take only 3 features the accuracy is low so we need to take all the features that effect the output.

Features:’GRE Score’,’ Toefl Score’,'University Rating', 'SOP',’LOR ', 'CGPA', 'Research',’CGPA’

Figure 4.3.2.4:OLS Model

****

Interpretation: From the above OLS model we can conclude that R square and Adj R Square values are high so the model is good model and P value is less than 5% acceptance region i.e., 0.05 alpha then null hypothesis is rejected and alternative hypothesis is accepted i.e., there is a significant relationship among input and output variables.

Figure 4.3.2.5: Accuracy for training and testing data.

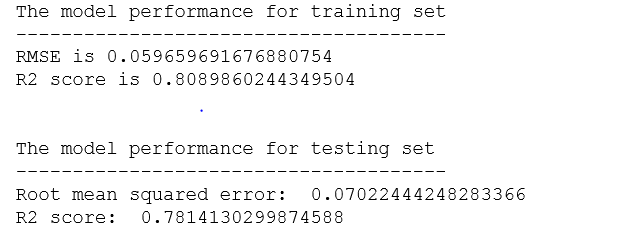
**C:\Users\Yadavalli\Desktop\datasets\train.PNG**

**C:\Users\Yadavalli\Desktop\datasets\test.PNG**

Interpretation: From the above we can say that Accuracy for training data is 80% and Accuracy for testing model is 81% and these both accuracies are almost near so it is a good model.

And by comparing the accuracies we can say that all the features are important and not considering any one feature may lead to the decrease in the accuracy of the model

Figure 4.3.2.6:RMSE and R2 Score



Interpretation: From the above model metrics for the regression analysis i.e., Root mean square error and R squared values of the train and test data we can say that the RMSE for the training is very low(0.059) and R squared is very high(0.080). So the model is very good for the training data i.e., if the data is present in the training data then the result is shown with greater accuracy.For the testing data also the RMSE is low(0.070) and R squared is high(0.78).

Both for the training and testing data the accuracies are almost close and so the model is not much biased and is a good model for predictions.

**4.3.3: Decision Tree Regression**

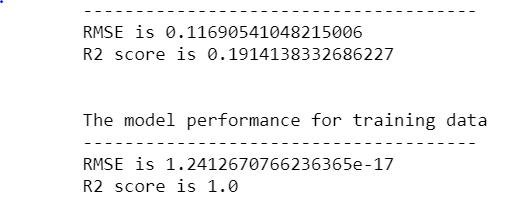
Decision tree builds regression or classification models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

Figure 4.3.3.1:Accuracy for training and testing data

Interpretation: From the above accuracy values of the train and test data we can say that the accuracy values of the train data is 1 and that of the test data is 0.19.

From this we can say that the decision tree model works the best for the training data and the worst for the testing data.

Figure 4.3.3.2:



Interpretation:

From the above model metrics for the regression analysis i.e., Root mean square error and R squared values of the train and test data we can say that the RMSE for the training is very low(1.24e-17) and R squared is very high(1.0). So the model is very good for the training data i.e., if the data is present in the training data then the result is shown with greater accuracy. But for the testing data the RMSE is high and R squared is low.

So for the Decision tree Regression even if the model is showing high metrics for the training but it is showing less for the testing data so this model is highly biased and not a good choice to use for predicting.

**4.3.4: Random Forest Regression**

**Random forests** or **random decision forests** are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics))of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set)

Accuracy of the model for the train data:

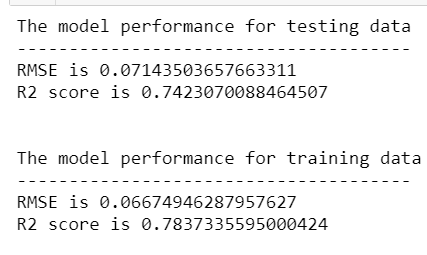


Accuracy for the test data:



Interpretation:

From the above accuracy values of the train and test data we can say that the accuracy values of the train data is 0.790 and that of the test data is 0.783. Where we can say that the model is not biased but the accuracy of the model is less than 80 percent.





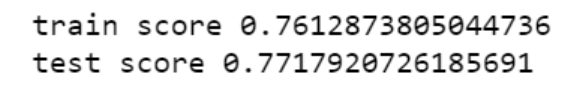
From the above model metrics for the regression analysis i.e., Root mean square error and R squared values of the train and test data we can say that the RMSE for the training is low(0.066) and R squared is high(0.78). So the model is good for the training data i.e., if the data is present in the training data then the result is shown with a good accuracy. For the testing data the RMSE is low and R squared is high.

From this we can that the model is not biased as it is showing almost the same values for both training and testing data with a nearly 80 percent accuracy.

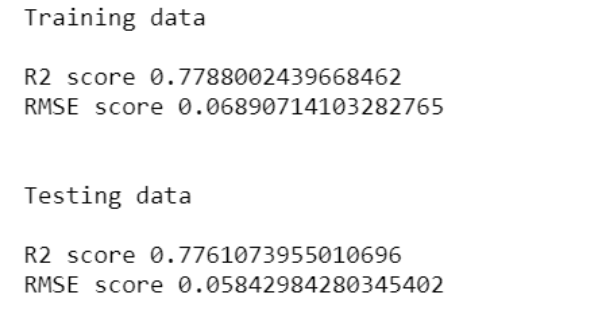
**4.3.5: Support Vector Regression**

Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem. But besides this fact, there is also a more complicated reason, the algorithm is more complicated therefore to be taken in consideration. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin, keeping in mind that part of the error is tolerated.

Accuracy of the model for train and test data :



From the above accuracy values of the train and test data we can say that the accuracy values of the train data is 0.760 and that of the test data is 0.773. Where we can say that the model is not biased but the accuracy of the model is nearly 80 percent.

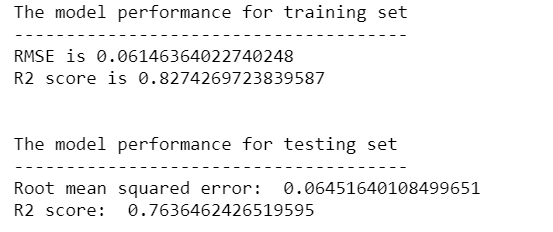


From the above model metrics for the regression analysis i.e., Root mean square error and R squared values of the train and test data we can say that the RMSE for the training is low(0.068) and R squared is high(0.77). So the model is good for the training data i.e., if the data is present in the training data then the result is shown with a good accuracy. For the testing data the RMSE is low and R squared is high.

From this we can that the model is not biased as it is showing almost the same values for both training and testing data with a nearly 80 percent accuracy.

**4.3.6: Neural Networks with Regression**

Neural networks are well known for Classification problems, But they can also be used for regression models.DNN stands for "deep neural network," which is a neural network with multiple layers.A DNN Regression model is similar, but instead of predicting a category, it predicts a numeric value in a continuous range.



From the above model metrics for the regression analysis i.e., Root mean square error and R squared values of the train and test data we can say that the RMSE for the training is low(0.061) and R squared is high(0.82). So the model is good for the training data i.e., if the data is present in the training data then the result is shown with a good accuracy. For the testing data the RMSE is low and R squared is high.

From this we can that the model is not biased as it is showing almost the same values for both training and testing data with a nearly 80 percent accuracy.

**5:Findings and Suggestions:**

First using the regression we found out that by taking single feature and building the regression model it is giving less accuracy so we consider multiple regression and use only three features that are most correlated with the output Chance of admit and have seen that the accuracy has increased but lesser than required . So we considered alls the features and have seen that the accuracy of the model has been increased. So we crated a model considering all the features that effect the chance of admit.

For the decision tree regression for the train data the accuracy is very high but for test data the accuracy is very low and so is highly biased model. So we can say that we cannot use decision tree regression for the prediction of the new data.

For the random forest regression the train and test data accuracies are very close and so there is no much bias in the model. So it can be used for prediction of the new data but with an accuracy of nearly 80 percent.

For the Support Vector Regression the train and test data accuracies are equal (upto two decimal points) and so there is no much bias in the model. But for the Support Vector Machine the data needs to be very large for the higher accuracy of the model.As our data is less Support Vector Regression may not lead to a correct prediction of the new data. But the models accuracy is nearly 78 percent for both test and train data.

**6:Conclusion:**

The main objective of this research was to develop a prototype of the system that can be used by the students aspiring to pursue their education. Multiple machine learning algorithms like Linear Regression,Multiple Linear Regression,Decision Tree Regression,Support Vector Regression etc were developed and used for this research.

**RMSE AND R2 SCORE FOR LINEAR MODEL:**

**TRAINING DATA:**

|  |  |  |  |
| --- | --- | --- | --- |
| **INPUT** | **OUTPUT** | **RMSE** | **R2 SCORE** |
| GRE | CHANCE OF ADMIT | 0.11670382 | 0.269074 |
| TOFEL | CHANCE OF ADMIT | 0.10588406 | 0.3983216 |
| CGPA | CHANCE OF ADMIT | 0.09568096 | 0.5086915 |

**RMSE AND R2 SCORE OF TESTING DATA:**

|  |  |  |  |
| --- | --- | --- | --- |
| **INPUT** | **OUTPUT** | **RMSE** | **R2SCORE** |
| GRE | CHANCE OF ADMIT | 0.1303908 | 0.2463980 |
| TOFEL | CHANCE OF ADMIT | 0.12243583 | 0.33554620 |
| CGPA | CHANCE OF ADMIT | 0.10958873 | 0.46767176 |

**RMSE AND R2 SCORE FOR DIFFERENT MODELS:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODEL** | **RMSE** | | **R2 SCORE** | |
| **Testing** | **Training** | **Testing** | **Training** |
| Multiple Linear Regression | 0.0702244 | 0.0596 | 0.781413 | 0.80898 |
| Decision Tree Regression | 0.116905 | 1.24126 | 0.1914 | 1.0 |
| Random forest Regression | 0.07143 | 0.06674 | 0.74230 | 0.783733 |
| Support vector Regression | 0.05842 | 0.068907 | 0.776107 | 0.778800 |
| Neural Networks with Regression | 0.061463 | 0.064516 | 0.827426 | 0.763646 |

The overall objective of the research was achieved successfully as the system allow the students to save the extra amount of time and money that they would spend on education consultants and application fees for the universities where they have fewer chances of securing admission. Also, it will help the students to make better and faster decision regarding application to the universities.

**7:Bibilography:**

[1]"Skewness,"[Online].Available:https://www.uky.edu/Centers/HIV/cjt765/9.Skewn ess%20and%20Kurtosis.pdf. [Accessed 17 05 2017].

[2]. Livieris, I.E.; Kotsilieris, T.; Tampakas, V.; Pintelas, P. Improving the evaluation process of students’ performance utilizing a decision support software. *Neural Comput. Appl.* **2018**. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Improving%20the%20evaluation%20process%20of%20students%E2%80%99%20performance%20utilizing%20a%20decision%20support%20software&author=Livieris,+I.E.&author=Kotsilieris,+T.&author=Tampakas,+V.&author=Pintelas,+P.&publication_year=2018&journal=Neural+Comput.+Appl.&doi=10.1007/s00521-018-3756-y)] [[CrossRef](https://dx.doi.org/10.1007/s00521-018-3756-y)].

[3].Chang, L. Applying Data Mining to Predict College Admissions Yield: A Case Study. *New Dir. Institutional Res.* **2006**, *131*, 53–68. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Applying%20Data%20Mining%20to%20Predict%20College%20Admissions%20Yield:%20A%20Case%20Study&author=Chang,+L.&publication_year=2006&journal=New+Dir.+Institutional+Res.&volume=131&pages=53%E2%80%9368&doi=10.1002/ir.187)] [[CrossRef](https://dx.doi.org/10.1002/ir.187)].

[4].Powell, F. Universities, Colleges Where Students Are Eager to Enroll. U.S. News and World Report. 2018. Available online: <https://www.usnews.com/education/best-colleges/articles/2018-01-23/universities-colleges-where-students-are-eager-to-enroll> (accessed on 15 December 2018).

[5].Dormann, C.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; García Marquéz, J.R.; Gruber, B.; Lafourcade, B.; Leitão, P.J.; et al. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2012**, *36*, 27–46. [[Google Scholar](https://scholar.google.com/scholar_lookup?title=Collinearity:%20A%20review%20of%20methods%20to%20deal%20with%20it%20and%20a%20simulation%20study%20evaluating%20their%20performance&author=Dormann,+C.&author=Elith,+J.&author=Bacher,+S.&author=Buchmann,+C.&author=Carl,+G.&author=Carr%C3%A9,+G.&author=Garc%C3%ADa+Marqu%C3%A9z,+J.R.&author=Gruber,+B.&author=Lafourcade,+B.&author=Leit%C3%A3o,+P.J.&publication_year=2012&journal=Ecography&volume=36&pages=27%E2%80%9346&doi=10.1111/j.1600-0587.2012.07348.x)] [[CrossRef](https://dx.doi.org/10.1111/j.1600-0587.2012.07348.x)].

1. P. Geurts, D. Ernst., and 1. Wehenkel, Extremely randomized trees, Machine Learning[J], 63(1), 3 -42, 2006.
2. G. Valentini& F. Masulli. Ensembles of Learning Machines[J]. Workshop on Neural Nets, 2002, (2486): 3-20.
3. homas, G, Dietterich. Ensemble learning[J]. Handbook of Brain Theory and Neural Networks, 2002, (2).
4. Tom. Mitchell. Machine Learning[M]. 2003.

[10].1. Breiman. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32