**PROJECT REPORT**

**ON**

**ADMISSION PREDICTION**

Submitted By :- (Group-3,Project-1)

|  |  |  |
| --- | --- | --- |
| Name | Email | College |
| Tirth Patel | tirthpatel23t@gmail.com | Ahmedabad Institute Of Technology |
| Aniket Sharad Khatode | [aniketak03@gmail.com](mailto:aniketak03@gmail.com) | Amrutvahini College Of Engineering |
| Visista | Visista17@gmail.com | Annamacharya Institute Of Technology And Sciences |

Under The Guidance Of :-

Gurvansh Singh

M.Tech

Knowledge Solutions India

**Abstract**

Artificial Intelligence has become one of the most emerging field recently. Machine Learning is one of the subset of Artificial Intelligence. There are majorly three types of Machine Learning : Supervised Machine Learning, Unsupervised Machine Learning, Reinforcement Machine Learning. This project is based on Supervised Machine Learning. In Supervised Machine Learning, inputs and outputs both are known and on the basis of them predictions are done. Examples of Supervised Machine Learning are training data or voice or image to predict output. It is also used in forecasting or recommendation system.

**Content**

1. Abstract………………………………………………………………...…2
2. Content…………………………………………………………………...3
3. Chapters :-
4. Introduction…………………………………………………4
5. Libraries…………………………………………………….5
6. Mathematics………………………………………………...6
7. Dataset………………………………………………………7
8. EDA & Data-Preprocessing………………………………...8
9. Models……………………………………………………...10
10. Accuracy Checking Methods……………………………….14
11. Graphs……………………………………………………………...……15
12. Conclusion…...……………………………………………………….....18
13. References………………………………………………………………19

**Chapters**

1. **Introduction :-**

This project is based on Supervised Machine Learning. So,we require both inputs and outputs. Inputs are features (independent variables). They are generally indicated by ‘x’. Outputs are dependent variables. They are generally indicated by ‘y’. In this dataset, features are GRE scrores, TOEFL scores, University Rating, Statement of Purpose (SOP), Recommendation Strength (LOR), CGPA, Research. Independent variable is Chance of Admit. In this project, ML models to predict the ‘Chance of Admit’ with minimum MSE and RMSE and maximum R-Square score are built. Following models are built :

1. Multiple Linear Regression
2. Random Forest Regression
3. Multiple Linear Regression with Principal Component Analysis
4. Random Forest Regression with Principal Component Analysis
5. **Libraries :-**

The most important step is importing required libraries so that every necessary modules can be used in code conveniently. ‘import’ keyword is used to import libraries. Libraries included in this project are :

1. Numpy
2. Pandas
3. Matplotlib
4. Scikit-learn
5. Scipy
6. Mlxtend
7. Statsmodels
8. Xlrd
9. **Mathematics :-**

Concepts of mean, median, mode are used in data-preprocessing. In principal component analysis, concepts of linear algebra like variance, standard deviation, eigen vector etc. are used.

*Mean* : Mean is the average of the values.

*Median* : Median is the middle value of set which is already sorted. If there

two median, then mean of two values is median.

*Mode* : Mode is the most repeated value

*Standard Deviation* : Standard Deviation shows how spread out the values

in Dataset.

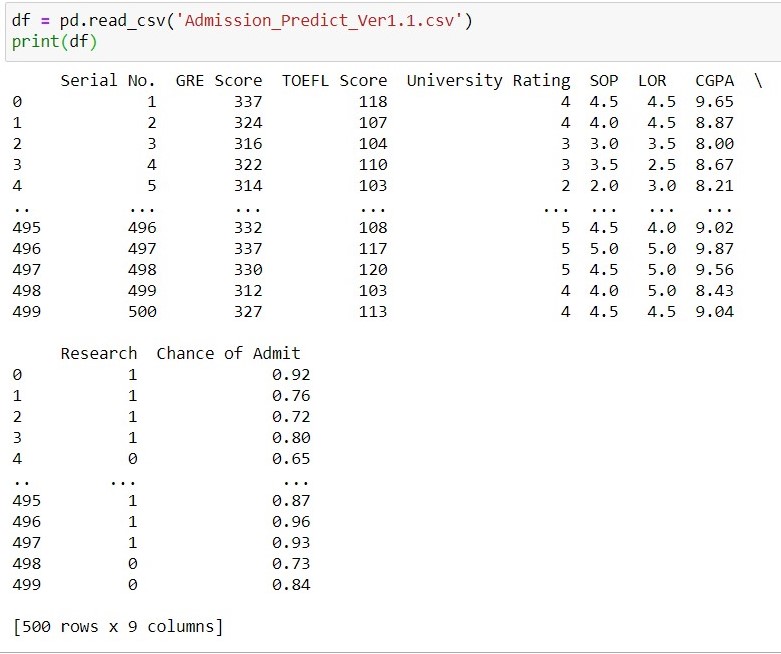
*Variance* : Variance is square of Standard Deviation.

*Eigenvector* : An Eigenvector is vector whose direction remains unchanged

when a linear transformation is applied to it.

1. **Dataset :-**

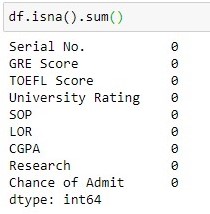
After importing libraries, second step is to read dataset. Pandas library object is used to read dataset. In this dataset, there are GRE scrores, TOEFL scores, University Rating, Statement of Purpose (SOP), Recommendation Strength (LOR), CGPA, Research, Chance of Admit columns. Dimension is 500 rows and 9 columns.

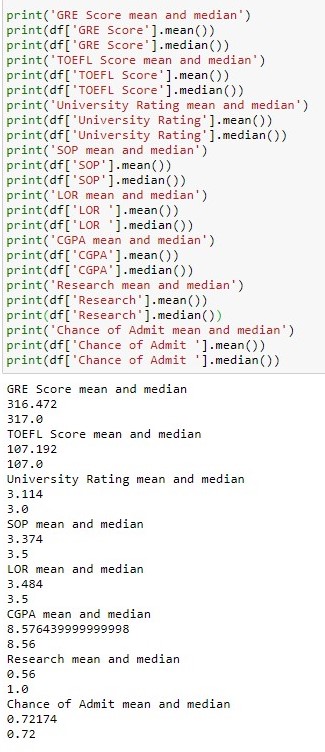


Admission\_Predict\_ver1.1.csv

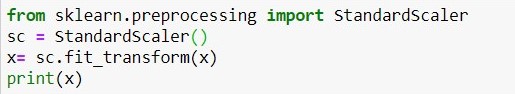
1. **EDA & Data-Preprocessing :-**

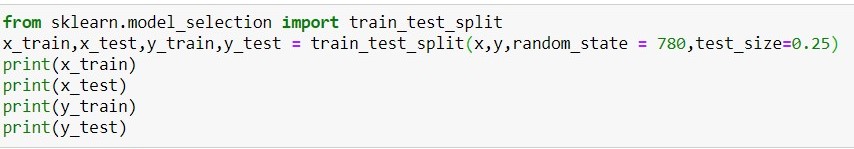
Exploratory Data Analysis & Data-Preprocessing are necessary steps before creating any model. After EDA on this dataset, it is found that there are no null values, outliers, missing value.





In Data-Preprocessing, x (features) and y (output) are extracted from dataset. There is no categorical data so encoding is not required. As there is large difference in values, scaling of features is done. After that using train\_test\_split(), x and y are splitted into train set (x\_train, y\_train) and test set (x\_test, y\_test) for training and testing respectively.

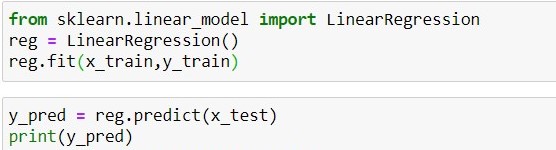




test\_size and train\_size can be given exclusively or bydefault it is 0.25 (25% of dataset). random\_state distributes dataset randomly into train and test according to the seed value.

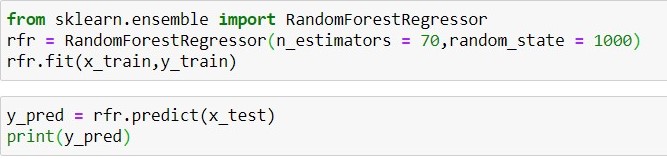
1. **Models :-**
2. **Multiple Linear Regression**

The regression is used to find the relationship between variables. In Machine Learning and in Statistical modelling, that relationship is used to predict the outcome of future events. Linear Regression uses the relationship between the data-points to draw a straight line through all them. Multiple Regression is like Linear Regression, but with more than one independent value or variables. Output will always be one. ‘x\_train’ and ‘y\_train’ are trained by fitting them in model using the object of LinearRegression class. Predict() takes ‘x\_train’ as argument to predict the output (y\_pred). This predicted output (y\_pred) can be compared with actual output (y\_test) to check accuracy. Plotting more than 3 dimension is not possible so Automatic Backward Elimination is used. It removes the least significant features which affects less in predicting output. After that again training and modelling can be done for predicting output. scatter() and plot() are used for plotting graph.



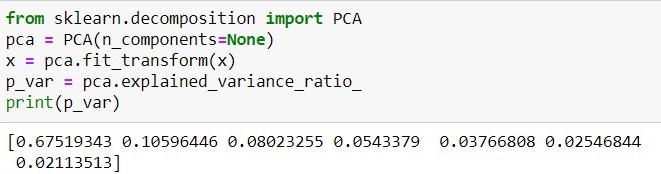
1. **Random Forest Regression**

Random Forest is a popular Machine Learning Algorithm. It can be used for both Classification and Regression problems. Here it is used as Random Forest Regression as it is a regression problem. It is based on the concept of ensemble learning. Multiple algorithms are combined to solve complex problem & to improve the performance of the model. It takes prediction from each tree and based on the average, predicts the final output. ‘x\_train’ and ‘y\_train’ are trained by fitting them in model using the object of RandomForestRegressor class. Predict() takes ‘x\_train’ as argument to predict the output (y\_pred). This predicted output (y\_pred) can be compared with actual output (y\_test) to check accuracy. The advantage of RF is that it takes less training time compared to other algorithm. It predicts output with high accuracy even for large dataset. scatter() and plot() are used for plotting graph.



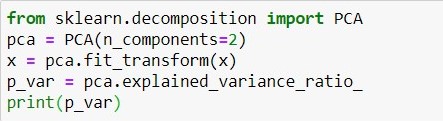
1. **Multiple Linear Regression with Principal Component Analysis**

Plotting more than 3 dimension is not possible and moreover working with multiple features can be inefficient. Principal Component Analysis is most popular dimensionality reduction algorithm. It detect correlation and covariance between features (as it is unsupervised algorithm it considers only input). The goal is to reduce the dimensions of a d-dimensional dataset by projecting it onto a (k)-dimensional subspace (where k<d) in order to increase the computational efficiency while retaining most of the information. Using the object of PCA class, fit\_transform() takes x as argument and fit x in model and transform x into required features. After PCA, only significant features are left and these features are splitted into train set and test set. Now modeliing procedure is similar to Multiple Linear Regression as described above.



Initially, n\_components=none is taken because eigen values are not known and also eigen vectors to create w matrix, are not known.

‘p\_var’ is calculated which describes percentage variance of each column. On the basis of it eigen vectors can be considered. Here they are first two so, n\_components=2 is taken.



1. **Random Forest Regression with Principal Component Analysis**

Principal Compoenet Analysis is done on dataset by performing EDA and Pre-processing. As described above, dimensionality reduction is done similarly. After splitting x (new features) and y into train set and test set, both are trained and fitted in random forest regressor model and ouput is predicted in a similar way as mentioned in 3) RFR.

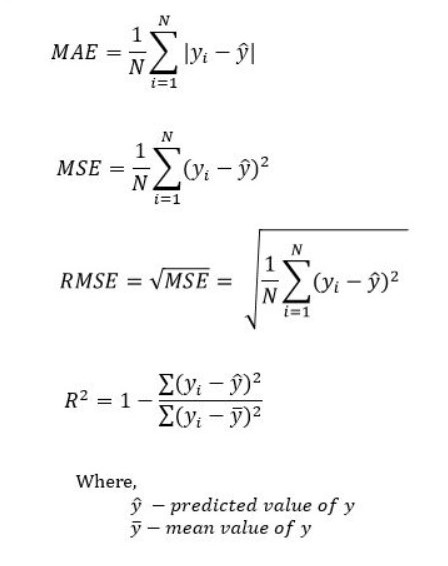
1. **Accuracy Checking Methods :-**

*MAE* : Mean Absolute Error represents the difference between original values & predicted values extracted by averaged difference.

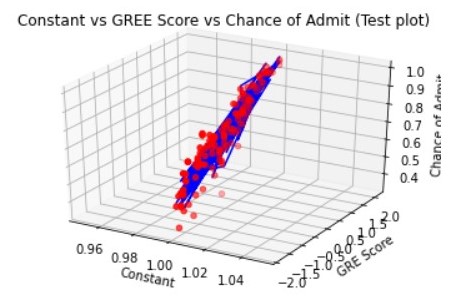
*MSE* : Mean Squared Error represents the difference between original values & predicted values extracted by squared the average distance.

*RMSE* : Root Mean Squared Error is the error rate by the square root of MSE.

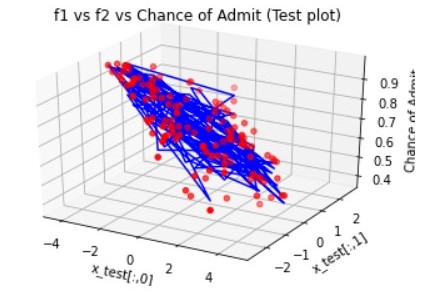
*R-Squared* : It represents the coefficient of how well the values fit compared to the original values. Its value lies between 0 & 1. 1 indicates model fits perfectly to dataset.



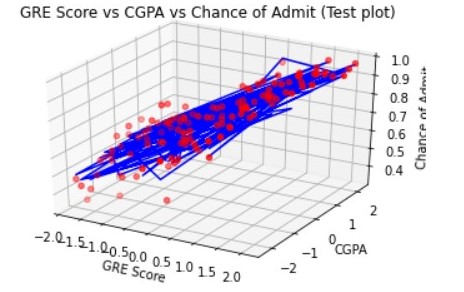
**Graphs**



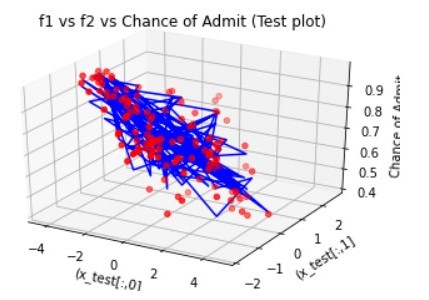
MLR



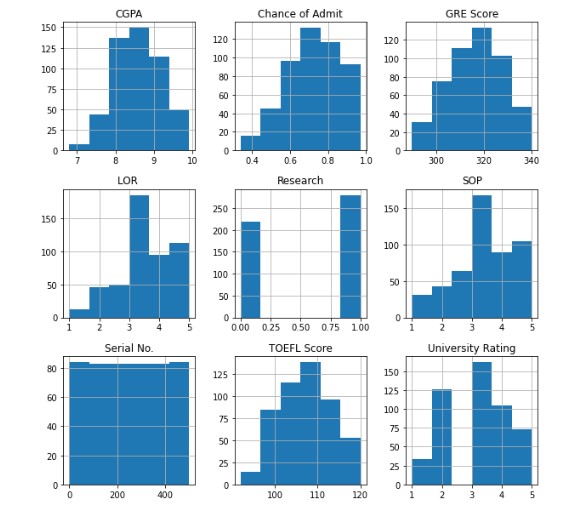
MLR with PCA



RFR



RFR with PCA

****

**Conclusion**

From the below given table, we can conclude that Mean Squared Error of Multiple Linear Regression is least i.e., **0.0025** which can be considered as good among all other models. Root Mean Squared Error of Multiple Linear Regression is least i.e., **0.0500** which can be considered good among all other models. R-Squared Score of Multiple Linear Regression is Highest i.e., **88%** which can be said that model best fits to dataset. Random Forest Regression with Principal Component Analysis has highest MSE and RMSE which is not good for model. Moreover it has least R2\_score i.e., 0.810 among all models, which is shows model does not fit well in dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ACM\Model | MLR | MLR\_PCA | RFR | RFR\_PCA |
| MSE | 0.0025 | 0.0032 | 0.0028 | 0.0040 |
| RMSE | 0.0500 | 0.0569 | 0.0533 | 0.0637 |
| R2\_Score | 0.883 | 0.837 | 0.867 | 0.810 |

**References**

<https://www.mentorrbuddy.com/home>

<https://plotly.com/python/v3/ipython-notebooks/principal-component-analysis/>

<https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html#:~:text=The%20MSE%2C%20MAE%2C%20RMSE%2C,difference%20over%20the%20data%20set.>

<https://www.javatpoint.com/>