Prediction of MS Graduate Admissions using Decision Tree, Linear regression and Random forest algorithms.

A. Sai Prudhvi
B160865CS

P. Rahul
B160607CS

Ch. Jagadeesh
B160547CS

M. Sharath
B160855CS

***Abstract*—The number of students wanting to pursue higher education abroad is increasing rapidly every year, especially in the US-based universities, and they find it difficult to find the best university. And so, this paper helps predict Indian students eligibility to be admitted in the best university based on the attributes such as GRE score,TOEFL score,CGPA,research papers etc. The possibility of their chance of admit is then calculated using their scores . This project is done using machine learning techniques namely decision tree, linear regresion and randomm forest algorithms to predict the output that helps the students to admit to the best university and also helps the student find the possibility to admit to other universities based on their scores. Out of the three algorithms linear regression showed higher accuracy**

I. INTRODUCTION

A main subject for debate is the future of the conventional higher education model. Of course, these concerns are highly dependent on the type of academic institution under consideration, such as public colleges and universities. Most Indian students students face difficulties in enrolling in foreign universities , especially U.S.-based universities, and thus graduate admission is useful in bridging the gap between Indian students and U.S.-based universities where students can find better universities they need without any inconvenience. The number of Indian students (both undergraduate and graduate) enrolling in the US crossed 1million for the first time in 2015-16 as per a report backed by the State department.So there has to be a web application which Could help the students to find their respective universities with minimal admission counselling costs and apply to the Universities they are Eligible, and of interest to them. The Prediction makes use of classification and regression algorithms to predict the chance of a student getting an admission in a university. Supervised machine learning problems are a class of problems which Can generalize and learn from the Training data, i.e. Set of data in which the correct classification is established.

II. LITERATURE REVIEW

Machine learning techniques have many applications for analyzing data and other information in the context of educational settings. This area of study is commonly known as ”educational data mining” (EDM) and is a recent area of Emerging field with its own journals, conferences and community of researchers. A subset of EDM research focusing on data analysis to improve clarity and predictability of higher education institutions on the size of their student bodies is often referred to as enrollment management. Enrolment management is ”a concept of organization and a structured collection of practices aimed at controlling student enrolment.”.A recent study published this year reveals some key factors in the decision-making process and thus allows for the suggestion of a recommendation algorithm that allows applicants to make informed decisions about where to apply. There are many websites that predict admission to college from an aspiring student’s perspective, each website is unique and our website displays top ten universities with their official websites and their score range that they expect from their students.

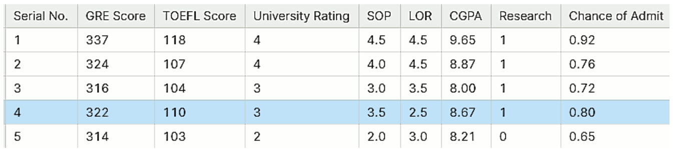
III. SYSTEM ARCHITECTURE

Architecture consists of the Django and User Interface architecture. The user interface is implemented using HTML , CSS and Javascript. And the decision tree classifier algorithm is used to analyze the data and Django framework that consists of collecting models that facilitate development and is also used for both frontend and backend, making it a convenient way to generate dynamic HTML pages using template system. It consists of a view of where business logic is implemented and returns user response. And template is used for building interactive web pages and framework uses HTTP to request backend information.

IV. DATA

We performed analysis on kaggle dataset. The dataset has 400 rows and 9 columns.

GRE is a complex feature consisting of GRE Quant, GRE Verbal and GRE AWA score. Similarly, TOEFL is also a complex feature consisting of TOEFL score and essay score. Chance of admit is calculated based on the inputs.



Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4.0	4.5	8.87	1	0.76
3	316	104	3	3.0	3.5	8.00	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.80
5	314	103	2	2.0	3.0	8.21	0	0.65

Fig. 1. Dataset
1

V. Data Analysis

Plotting between dependant and independant variables(features) to obtain the relationship between the variables.

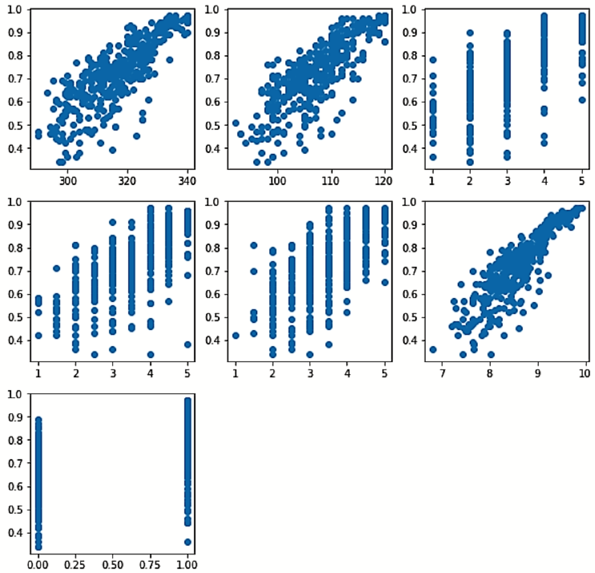
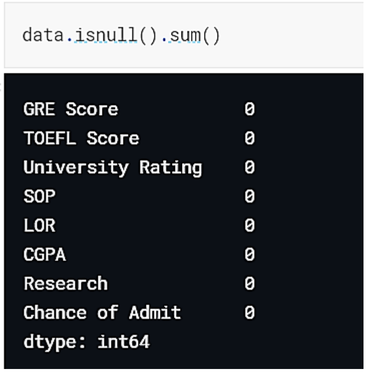


Fig. 2. plot between dependant and independant variables

In machine learning models null values affect the accuracy,so we should find the null values in all the columns and replace them with mean or median value of that particular column.



```
data.isnull().sum()
```

```
GRE Score            0
TOEFL Score          0
University Rating    0
SOP                  0
LOR                  0
CGPA                 0
Research             0
Chance of Admit      0
dtype: int64
```

Fig. 3. Removal of null values

It is observed that no column has null values, hence we can proceed forward. Generally outliers also affect the accuracy, so we need to boxplot to find the outliers.

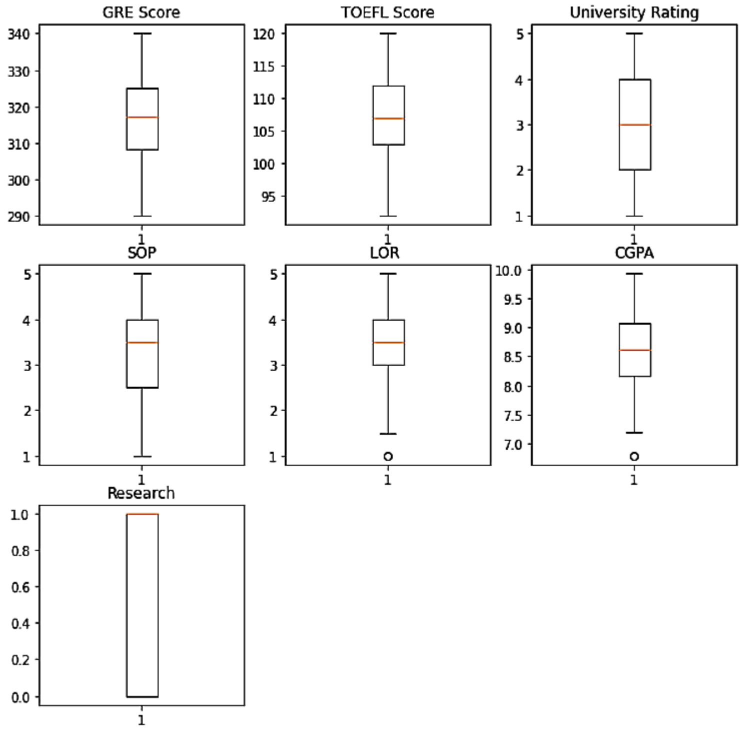
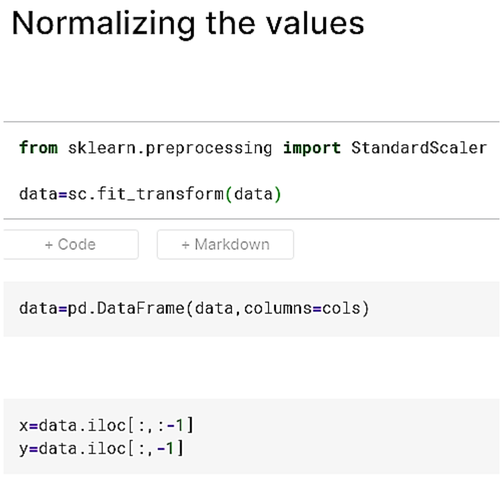


Fig. 4. boxplot

It is observed that there are no outliers below Q1 quartile and above Q4 quartile. Generally if there are any outliers, we remove them or replace with mean or median. To give equal importance to each and every value in the datset, we normalise the values.



Normalizing the values

```
from sklearn.preprocessing import StandardScaler

data=sc.fit_transform(data)
```

```
data=pd.DataFrame(data,columns=cols)
```

```
x=data.iloc[:,:-1]
y=data.iloc[:,-1]
```

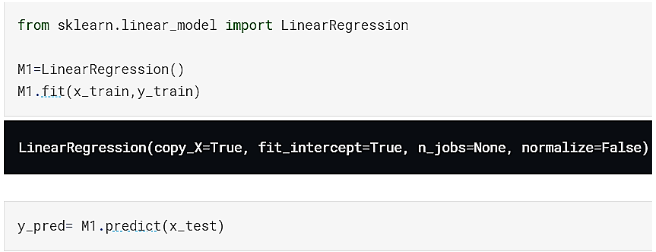
Fig. 5. Normalisation of values

VI. ALGORITHM

In this paper we are going to use three algorithms namely Linear regression, Decision tree and Random Forest algorithms.

A. Linear regression

Linear Regression is an algorithm of machine learning, based on supervised learning. It performs regression. Linear regression executes the task of predicting a dependent variable value (y) based on an independent variable (x) given. So this technique of regression finds a linear relation between x(input) and y(output).



```
from sklearn.linear_model import LinearRegression

M1=LinearRegression()
M1.fit(x_train,y_train)
```

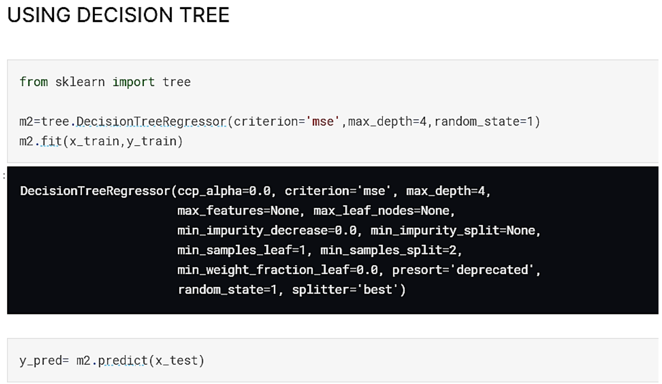
```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

```
y_pred= M1.predict(x_test)
```

Fig. 6. Linear regression code sample

B. Decision Tree

Decision tree algorithm falls under the Supervised Learning category. They can be used to solve regression problems as well as classification problems. Decision tree uses tree representation to solve the problem where each leaf node corresponds to a class label and attributes on the tree's internal node are represented. Using the decision tree we may represent any boolean function on discrete attributes.



USING DECISION TREE

```
from sklearn import tree

m2=tree.DecisionTreeRegressor(criterion='mse',max_depth=4,random_state=1)
m2.fit(x_train,y_train)
```

```
DecisionTreeRegressor(ccp_alpha=0.0, criterion='mse', max_depth=4,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort='deprecated',
                      random_state=1, splitter='best')
```

```
y_pred= m2.predict(x_test)
```

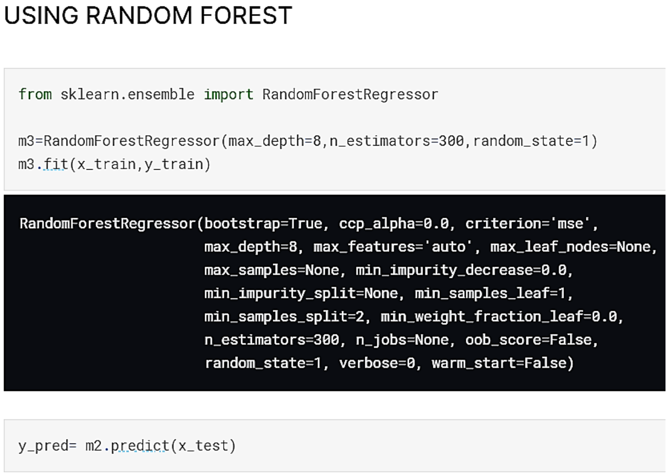
Fig. 7. Decision Tree code sample

C. Random Forest

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

VII. RESULTS

Here we are calculating the chance of admit value.It lies between 0 to 1.From sklearn we imported metrics class which contains different errors which are used to calculate the errors of our model.The different metrics used are Mean absolute error,Mean squared error,Root mean squared error and other metrics used is R2 score.We predicted the chance of admit using different models.The errors obtained are:



USING RANDOM FOREST

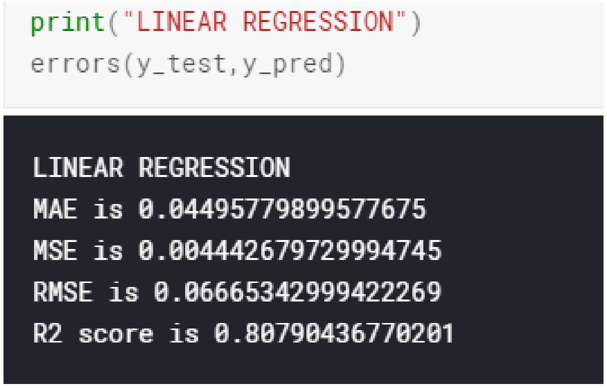
```
from sklearn.ensemble import RandomForestRegressor

m3=RandomForestRegressor(max_depth=8,n_estimators=300,random_state=1)
m3.fit(x_train,y_train)
```

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=8, max_features='auto', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=300, n_jobs=None, oob_score=False,
                      random_state=1, verbose=0, warm_start=False)
```

```
y_pred= m2.predict(x_test)
```

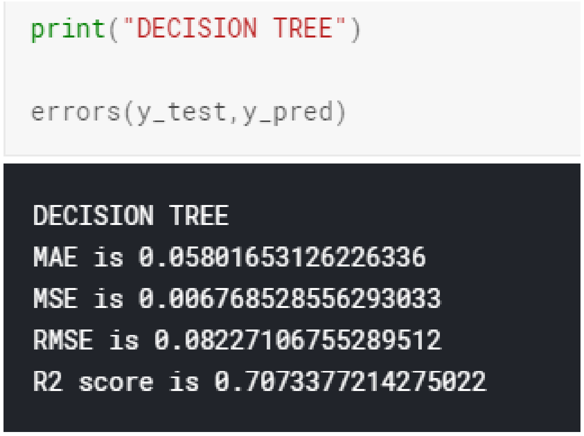
Fig. 8. Random Forest code sample



```
print("LINEAR REGRESSION")
errors(y_test,y_pred)
```

```
LINEAR REGRESSION
MAE is 0.04495779899577675
MSE is 0.004442679729994745
RMSE is 0.06665342999422269
R2 score is 0.80790436770201
```

Fig. 9. Using Linear Regression

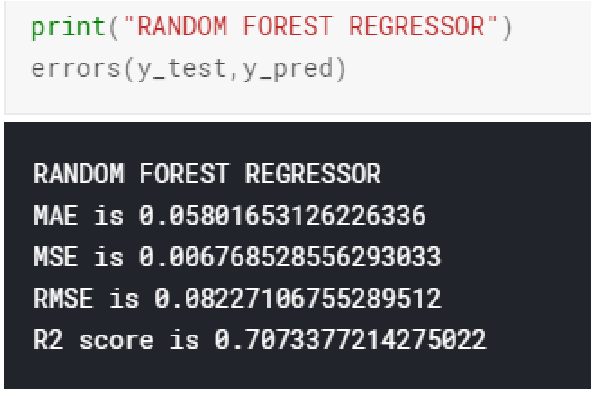


```
print("DECISION TREE")

errors(y_test,y_pred)
```

```
DECISION TREE
MAE is 0.05801653126226336
MSE is 0.006768528556293033
RMSE is 0.08227106755289512
R2 score is 0.7073377214275022
```

Fig. 10. Using Decision Tree



```
print("RANDOM FOREST REGRESSOR")
errors(y_test,y_pred)
```

```
RANDOM FOREST REGRESSOR
MAE is 0.05801653126226336
MSE is 0.006768528556293033
RMSE is 0.08227106755289512
R2 score is 0.7073377214275022
```

Fig. 11. Using Random Forest

It is observed that linear regression gave good R2 score as well as less errors compared to decision tree and random forest.

VIII. CONCLUSION

In this paper we focused more on predicting the MS Graduate admission dataset using Linear regression,Decision tree and Random forest.Applied many pre-processing techniques to improve the accuracy and to reduce the errors.It is observed that principal component analysis is not showing any significant change.Among all, linear regression showed better performance.While calculating feature_importances in Random Forest Regressor, CGPA attribute showed higher value compared to other attributes.

REFERENCES

[1] Jiawei Han And Micheline Kamber,Data Mining Concept and Techniques, Copyright 2006, Second Edition.

[2] Chen Jin, Luo De-lin and mu Fen-xiang An improve ID3 Decision tree algorithm. IEEE 4th International Conference on computer Science Education

[3] Bauer, E. Kohavi, R. (1999). An empirical comparison of voting classification algorithms. Machine Learning, 36(1/2), 105–139.

[4] Ho, T. K. (1998). The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(8), 832–844.

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2992018

[6] Journal of Educational Data Mining. Available online: http://jedm.educationaldatamining.org/index.php/JEDM (accessed on 15 December 2018).

[7] Educational Data Mining Conference 2018. Available online: http://educationaldatamining.org/EDM2018/ (accessed on 15 December 2018).

[8] Romero, C.; Ventura, S. Educational data mining: A survey from 1995 to 2005. Expert Syst. Appl. 2007, 33, 135–146. [Google Scholar] [CrossRef]

[9] Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. Expert Syst. Appl. 2014, 41, 1432–1462. [Google Scholar]

[10] American graduate admissions: both sides of the table http://hdl.handle.net/2142/92866

[11] Hossler, D.; Bean, J.P. The Strategic Management of College Enrollments, 1st ed.; Jossey Bass: San Francisco, CA, USA, 1990. [Google Scholar]

[12] Kuncheva, L.I. Combining Pattern Classifiers: Methods and Algorithms, 2nd ed.; McGraw Hill; John Wiley Sons, Inc.: Hoboken, NJ, USA, 2014. [Google Scholar]

[13] Alpaydin, E. Introduction to Machine Learning, 3rd ed.MIT Press: Cambridge, MA, USA, 2010. [Google Scholar]

[14] Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. Informatica 2007, 4, 249–268. [Google Scholar]

[15] Lapovsky, L. The Changing Business Model For Colleges And Universities. Forbes 2018. Available online:https://www.forbes.com/sites/lucielapovsky/201 8/02/06/the-changing-business-model-for-colleges-anduniversities/bbc03d45ed59 (accessed on 15 December 2018).