# PROJECT REPORT ON

# "Graduate Admission Prediction"

**2020-2021**

**SUBMITTED BY**

**Sakshi Deore(33217)**

**Ellika Mishra(33221)**

**Prajakta Keer(33231)**

**Under the guidance of**

**Mrs D.D. Londhe**

**Department of Information Technology**

**Pune Institute of Computer Technology College of Engineering**

**Sr. No 27, Pune-Satara Road, Dhankawadi, Pune - 411 043.**

**2020-2021**

# ABSTRACT

This project tries to understand the Graduate Admissions process by specifically analyzing different features of around 500 students like GRE score, Toefl score etc. Chance of admission of students for postgraduate programs has been given a data set. After extensive data pre-processing, we have tried to build an admission prediction model. The key to analyzing Graduate Admissions data is to analyze data in buckets rather than considering all in one bucket. The project aims to help students choose the right Universities by predicting whether a student will be admitted to a specific University. Similarly, this model can be used by the Graduate Admission Committee to predict the admission of particular students in the University.

# ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us the possibility to complete this report. A special gratitude I give to our guide **Prof. Deepali Londhe**, whose contribution in stimulating suggestions and encouragement, helped us to coordinate this project especially in writing this report.

Furthermore We would also like to acknowledge with much appreciation the crucial role of the staff of the IT Department , who gave the permission to use all required equipment and the necessary materials to complete the project "***Graduate Admission Prediction "***. A special thanks goes to all the teammates, who helped to assemble the parts and gave suggestions about this project. We have to appreciate the guidance given by our **H.O.D. and all other professors**.

# List of Contents

# List of Figures

# CHAPTER 1

# INTRODUCTION

## 1.1 Problem Statement

This project analyzes various features such as GRE Score, TOEFL Score, CGPA, University Rating, Research, Rating of Universities of various students who had applied to get admission in foreign universities and predicts the chance of student getting the admission.

The dataset provided was pretty small. We shuffled the data to avoid patterns. By performing exploratory analysis on the data we found that data was quite normal as it did not contain any null values or outliers. The next step was data pre-processing. Feature scaling helped to normalize the features and to bring them at the same scale. Then we extracted the features. We then split the dataset into a test set and train set at a ratio of 25 : 75.

We developed 2 models for this dataset:

1) Multiple Linear Regression(MLR)

2) Random Forest Regressor(RFR)

## 1.2 Purpose

Students are often worried about their chances of admission in graduate school. The purpose of this project is to help students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their admission chances in a particular university. This analysis should also help students who are currently preparing or will be preparing to get a better idea.

## 1.3 Objectives

1. Read the data and clean it
2. Data Exploration
3. Data Validation
4. Train the ML model
5. Test the accuracy

6. Visualize the data model
7. Prediction of Result

## 1.4 References

[1] Ali Bou Nassif, Ismail Shahin, Ashraf M Elnagar, "Graduate Admission Prediction Using Machine Learning", December 2020, DOI: 10.46300/91013.2020.14.13.

[2]Online Source: https://towardsdatascience.com/graduate-admission-prediction-using-machine-learning-8e09ba1af359

[3]A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, "E-Learning: Challenges and Research Opportunities Using Machine Learning Data Analytics," IEEE Access, 2018.

# CHAPTER 2

# Literature Survey

## 2.1 Introduction

The world markets are developing rapidly and continuously looking for the best knowledge and experience among people. Young workers who want to stand out in their jobs are always looking for higher degrees that can help them in improving their skills and knowledge. As a result, the number of students applying for graduate studies has increased in the last decade. This fact has motivated us to study the grades of students and the possibility of admission for master's programs that can help universities in predicting the possibility of accepting master's students submitting each year and providing the needed resources. The dataset used in this project is related to the educational domain. Admission is a dataset with 500 rows that contains 7 different independent variables which are:

• Graduate Record Exam1 (GRE) score. The score will be out of 340 points.

• Test of English as a Foreigner Language2 (TOEFL) score, which will be out of 120 points.

• University Rating (Uni.Rating) that indicates the Bachelor University ranking among the other universities. The score will be out of 5.

• Statement of purpose (SOP) which is a document written to show the candidate's life, ambition and the motivations for the chosen degree/ university. The score will be out of 5 points.

• Letter of Recommendation Strength (LOR) which verifies the candidate's professional experience, builds credibility, boosts confidence and ensures your competency. The score is out of 5 points

• Undergraduate GPA (CGPA) out of 10

• Research Experience that can support the application, such as publishing research papers in conferences, working as research assistant with university professors (either 0 or 1). One dependent variable can be predicted which is the chance of admission, that is according to the input given will be ranging from 0 to 1.

## 2.2 Detailed Literature Survey

### 1) Multiple Linear Regression

Multiple linear regression is a statistical technique used to predict a dependent variable according to two or more independent variables. As well as, present a linear relationship between them and fit them in a linear equation. The format of the linear equation is as following:

$yi=\beta 0+\beta 1xi1+...+\beta nxn+\epsilon$ (1)

where, for i=n observations:

yi=dependent variable

xi= independent variables

β0=y-intercept

βn=slope coefficients for each independent variable

ϵ=the model's error term or residuals.

### 2) Random Forest

The random forest algorithm is one of the most popular and powerful machine learning algorithms that is capable of performing both regression and classification tasks. This algorithm creates forests within a number of decision reads. Therefore, the more data is available the more accurate and robust results will be provided. Random Forest method can handle large datasets with higher dimensionality without overfitting the model. In addition, it can handle the missing values and maintains accuracy of missing data.

## 2.3 Findings of Literature survey

A great number of researches and studies have been done on graduation admission datasets using different types of machine learning algorithms. One impressive work by Acharya et al. has compared between 4 different regression algorithms, which are: Linear Regression, Support Vector Regression, Decision Trees and Random Forest, to predict the chance of admit based on the best model that showed the least MSE which was multilinear regression. In addition, Chakrabarty et al. compared between both linear regression and gradient boosting regression in predicting chance of admit; point out that gradient boosting regression showed better results. Gupta et al. developed a model that studies the graduate admission process in American universities using machine learning techniques. The purpose of this study was to guide students in finding the best educational institution to apply for. Five machine learning models were built in this paper including SVM (Linear Kernel), AdaBoost, and Logistic classifiers. Waters and Miikkulainen proposed a remarkable article that helps in ranking graduation admission applications according to the level of acceptance and enhances the performance of reviewing applications using statistical machine learning. Sujay applied linear regression to predict the chance of admitting graduate students in master's programs as a percentage. However, no more models were performed.

# CHAPTER 3

# 3. System Architecture and Design

## 3.1 Dataset Description

### A. Correlated variables

The dataset contained an independent variable to present the serial number of the requests. According to my expertise, it does not correlate to the dependent variable; hence, it was removed from the dataset. It is good to mention that tests showed that there are no missing values in any row of the database.

### B. Outliers

Outliers are data values that differ greatly from the majority of a set of data. To find the outliers, there are many methods that can be used, such as: scatter plot and boxplot. In this paper outliers will be investigated using a box plot method. As for the boxplot, the middle part of the plot represents the first and third quartiles. The line near the middle of the box represents the median. The whiskers on either side of the IQR represent the lowest and highest quartiles of the data. The ends of the whiskers represent the maximum and minimum of the data, and the individual circles beyond the whiskers represent outliers in the dataset.

There are many different methods to deal with outliers, such as:

• Remove the case

• Assign the next value nearer to the median in place of the outlier value

• Calculate the mean of the remaining values without the outlier and assign that to the outlier case.

In our case, there are few outliers; therefore, the best option will be removing the rows that contain outliers.

### C. Dataset Division

After cleaning the dataset, it was divided randomly into two parts using the holdout method. The first part contains 80% of the dataset to present training with

498 observations. The second part contains 20% of the dataset to present testing with 98 observations.

### D. Feature Selection

In order to perform feature selection, ols_step_best_subset() function has been applied; which will display all possible subsets. Then according to certain criteria, the best subset will be selected. Since there are 7 independent variables, the number of subsets to be tested is 27, which equals 128 subsets. To apply feature selection, linear regression equations should be applied. Note that linear regression can be performed only with numeric independent variables. It is observed that all variables are numeric. It is good to mention that in case of having categorical variable, as.numeric() function can be used to convert data variables to numeric. Next, a regression model is created, and feature selection is performed. The best model in feature selection is presented by the model with either highest R-square or smallest MSEP, which belongs to model 6 which includes all columns except column 4 that presents SOP.  E. Descriptive Summary Descriptive summary provides numerical measures of some important features which describe  a dataset.

### E. Descriptive Summary

Descriptive summary provides numerical measures of some important features which describe a dataset. Some features are presented in Fig 1:

| Feature | Value |
|---|---|
| Minimum Value | 0.3600 |
| First Quartile (Q1) | 0.6325 |
| Median | 0.7200 |
| Mean | 0.7233 |
| Third Quartile (Q3) | 0.8200 |
| Maximum Value | 0.9700 |

Fig 1

### 3.2 Algorithms

#### 1)Multiple Linear Regression

In Multiple Linear Regression, the target variable(Y) is a linear combination of multiple predictor variables $x_1$, $x_2$, $x_3$, ...,$x_n$. Since it is an enhancement of Simple Linear Regression, the same is applied for the multiple linear regression equation. Below are the main steps of deploying the MLR model:

1. Data Pre-processing Steps
2. Fitting the MLR model to the training set
3. Predicting the result of the test set

**Step-1: Data Preprocessing Step:**

The very first step is data preprocessing which we have already discussed in this tutorial. This process contains the below steps:

- Importing libraries: Firstly we will import the library which will help in building the model. Below is the code for it:

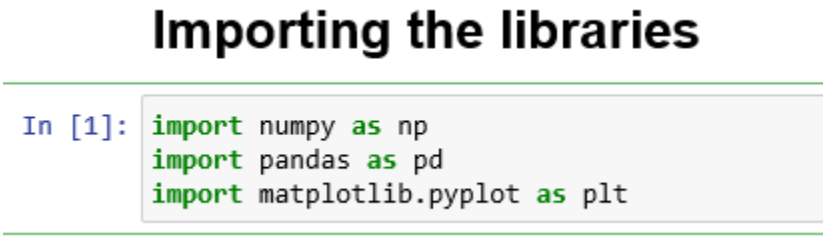


Fig 2

- **Importing dataset:** Now we will import the dataset(‘Admission_Predict_Ver1.1.csv’), which contains all the variables. Below is the code for it:

## Reading the dataset

```
In [2]: dataset = pd.read_csv('Admission_Predict_Ver1.1.csv')
        #print(dataset)
        dataset.columns

Out[2]: Index(['Serial No.', 'GRE Score', 'TOEFL Score', 'University Rating', 'SOP',
               'LOR ', 'CGPA', 'Research', 'Chance of Admit '],
              dtype='object')
```

Fig 3

- **Extracting dependent and independent Variables:**

## Data Preprocessing

```
In [11]: # extracting features
         x = dataset.iloc[:, 1:8].values
         y = dataset.iloc[:, -1].values
         #print(x)
         #print(y)
```

Fig 4

**Step: 2- Fitting our MLR model to the Training set:**

Now, we have well prepared our dataset in order to provide training, which means we will fit our regression model to the training set. It will be similar to as in the Simple Linear Regression model. The code for this will be:

## Multiple Linear Regerssion(MLR)

```
In [14]: # training the model
         from sklearn.linear_model import LinearRegression as LR
         lr = LR()
         lr.fit(x_train, y_train)

Out[14]: LinearRegression()
```

Fig 5

**Step: 3- Prediction of Test set results:**

The last step for our model is checking the performance of the model. We will do it by predicting the test set result. For prediction, we will create a **y_pred** vector. Below is the code for it:

```
In [15]: # predicting the output
         y_pred = lr.predict(x_test)
         #print(np.concatenate((y_pred.reshape(len(y_pred), 1), y_test.reshape(len(y_test), 1)), 1))

In [16]: # checking for errors
         import sklearn.metrics as met
         print('MSE = ', met.mean_squared_error(y_test, y_pred))
         print('RMSE = ', np.sqrt(met.mean_squared_error(y_test, y_pred)))
         print('R2 = ', met.r2_score(y_test, y_pred))

         MSE =  0.0035468543843696306
         RMSE =  0.05955547316888374
         R2 =  0.837218729569142
```

Fig 6

**2)Random Forest**

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

```
In [15]: from sklearn.ensemble import RandomForestRegressor
         rfr=RandomForestRegressor(n_estimators=65, random_state=42)
         rfr.fit(x_train,cy_train)

Out[15]: RandomForestRegressor(n_estimators=65, random_state=42)


In [17]: y_pred=rfr.predict(x_test)
```

Fig 7

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

# CHAPTER 4

# Experimentation and Results

## 4.1 Tools Used

We created 2 models using Python.

**Python -** Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

**Python libraries used -**

**Numpy**:NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

**Pandas**: In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

- read_csv() function to read the dataset

**Matplotlib**: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

- to create scatter plot
- to create 2D plot

**Sklearn** :Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines.

- from pre-processing module import StandardScaler class to perform feature scaling
- from decomposition module imported PCA to reduce the number of features
- from model_selection module imported train_test_split() function to split the dataset
- from linear_model imported LinearRegression class to implement MLR
- imported the metrics module to calculate errors

## 4.2 Accuracy

### 1) Multiple Linear Regression

```
In [24]: # checking for errors
         import sklearn.metrics as met
         print('MSE = ', met.mean_squared_error(y_test, y_pred))
         print('RMSE = ', np.sqrt(met.mean_squared_error(y_test, y_pred)))
         print('R2 = ', met.r2_score(y_test, y_pred))

         MSE =  0.0035008062048216236
         RMSE =  0.05916761111302047
         R2 =  0.8393320898471635
```

Fig 8

### 2) Random Forest

```
In [20]: import sklearn.metrics as met
mse=met.mean_squared_error(cy_test,y_pred)
rmse=np.sqrt(mse)
r2_score=met.r2_score(cy_test,y_pred)
print(mse)
print(rmse)
print(r2_score)
```

```
0.025435266272189352
0.1594843762635994
0.8536722991520771
```

Fig 9

## 4.3 Visualization

### 1) Multiple Linear regression

#### a) Predicted vs Actual output Plot

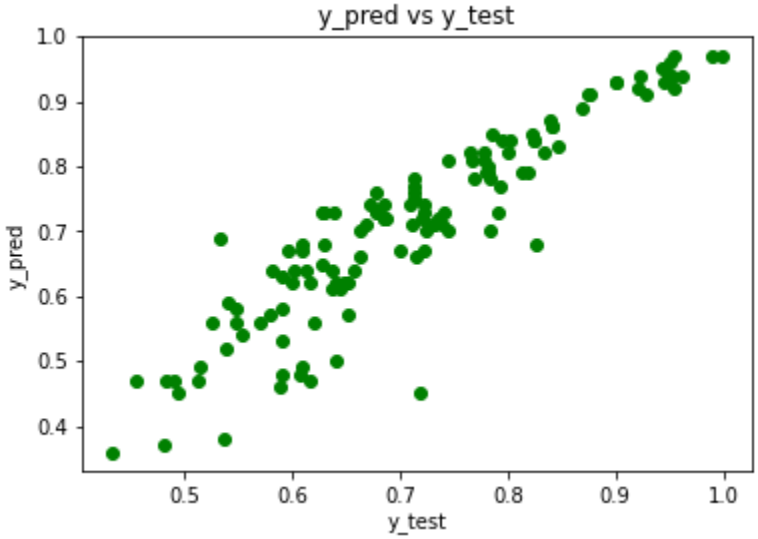


Fig 10

### 2) Random Forest

#### a) Chance of Admit Correlation Coefficients

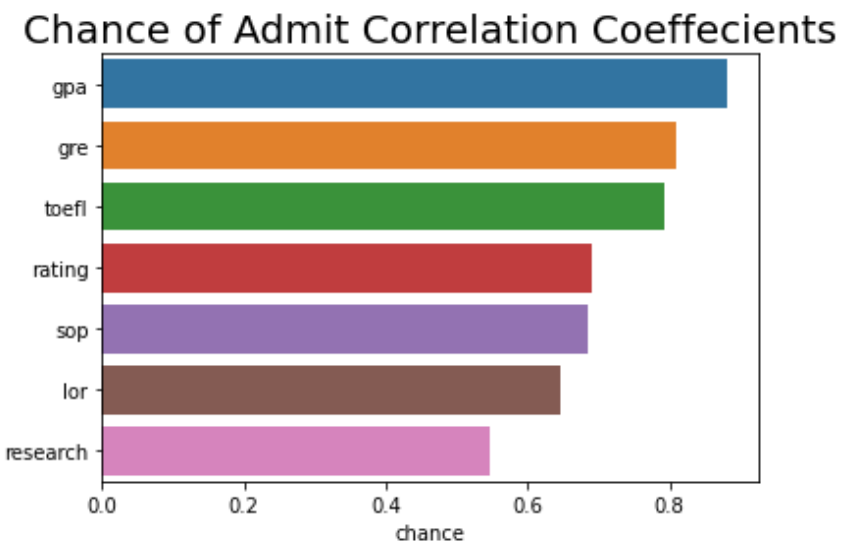
Chance of Admit Correlation Coeffecients
gpa
gre
toefl
rating
sop
lor
research
0.0
0.2
0.4
0.6
0.8
chance

Fig 11

# CHAPTER 5

# Conclusion and Future scope

## 5.1 Conclusion

In this project, we have predicted the chance of admission of a student in universities for a master's program. From the study, it has been observed that CGPA and GRE scores have more impact on chances of Admission.

Accuracy of the 4 models is as follows:

1)MLR - 83.93%

2) RFR – 85.3%

From the above figures it is clear that Random Forest Regressor is the best model with an accuracy of 85.3%

## 5.2 Future Work

As for the future work, more models can be conducted on more datasets to learn the model that gives the best performance.

# REFERENCES


- https://www.researchgate.net/publication/348433004_Graduate_Admission_Prediction_Using_Machine_Learning/link/5ffee71745851553a03dd015/download
- https://towardsdatascience.com/graduate-admission-prediction-using-machine-learning-8e09ba1af359?gi=5afd44d1e83a
- https://www.kaggle.com/adepvenugopal/predicting-graduate-admissions-using-ml