

# Analysis on College Admission Data

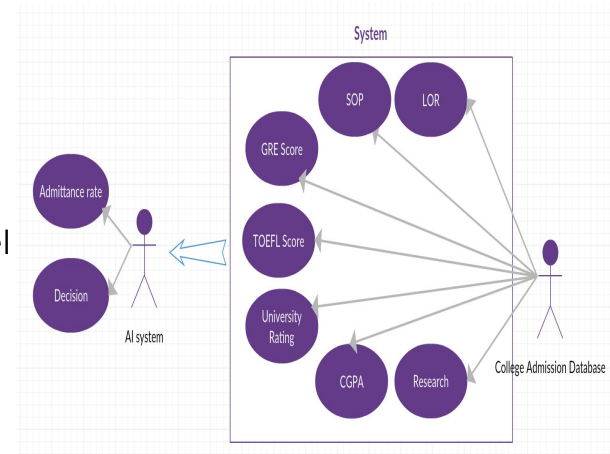
Himani Gupta-95529  
Pragya Moharatha-95474  
Priya Phapale - 94662





# Introduction

- College application process can be quite intimidating.
- Objective here is to analyze student information and predict admit/acceptance chances for the student.
- Decision tree/ Linear Regression for the predictive model analysis
- Optimize admission process by helping make a decision for student selection.





# Data Source

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	4.5	8.87	1	0.76
3	316	104	3	3	3.5	8	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.8
5	314	103	2	2	3	8.21	0	0.65
6	330	115	5	4.5	3	9.34	1	0.9
7	321	109	3	3	4	8.2	1	0.75
8	308	101	2	3	4	7.9	0	0.68
9	302	102	1	2	1.5	8	0	0.5
10	323	108	3	3.5	3	8.6	0	0.45

Input Dataset is taken from **Admission\_Predict.csv**

- ❖ Here, **Categorical Variables** are University Ranking, SOP, LOR
- ❖ They are **Ordinal Categorical Variable**.
- ❖ **Continuous Variables** are GRE Score, TOEFL Score, CGPA, Chance Of Admit
- ❖ **Discrete Variables** are Serial No and Research



## Data - Check for missing values

```
self.data.notnull().sum() / len(self.data)).sort_values(ascending=False)
```

Chance of Admit	1.0
Research	1.0
CGPA	1.0
LOR	1.0
SOP	1.0
University Rating	1.0
TOEFL Score	1.0
GRE Score	1.0
Serial No.	1.0

Data had no  
missing values in  
any columns



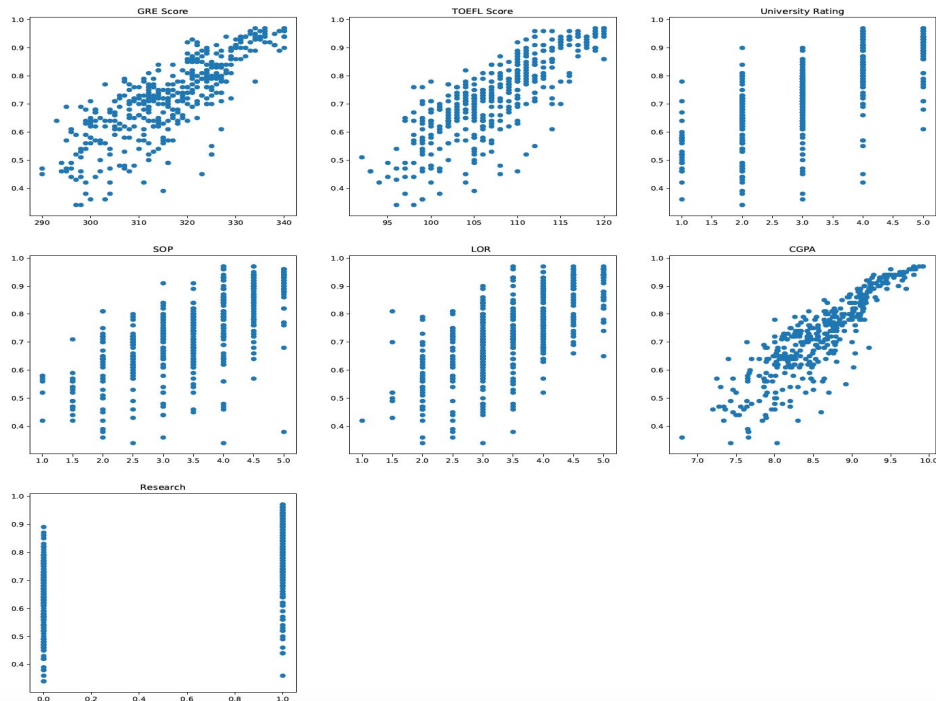
## Data summary statistics

```
self.data = pd.read_csv("Admission_Predict.csv")  
describe = self.data.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
Serial No.	400.0	200.500000	115.614301	1.00	100.75	200.50	300.2500	400.00
GRE Score	400.0	316.807500	11.473646	290.00	308.00	317.00	325.0000	340.00
TOEFL Score	400.0	107.410000	6.069514	92.00	103.00	107.00	112.0000	120.00
University Rating	400.0	3.087500	1.143728	1.00	2.00	3.00	4.0000	5.00
SOP	400.0	3.400000	1.006869	1.00	2.50	3.50	4.0000	5.00
LOR	400.0	3.452500	0.898478	1.00	3.00	3.50	4.0000	5.00
CGPA	400.0	8.598925	0.596317	6.80	8.17	8.61	9.0625	9.92
Research	400.0	0.547500	0.498362	0.00	0.00	1.00	1.0000	1.00
Chance of Admit	400.0	0.724350	0.142609	0.34	0.64	0.73	0.8300	0.97



# Data Visualization



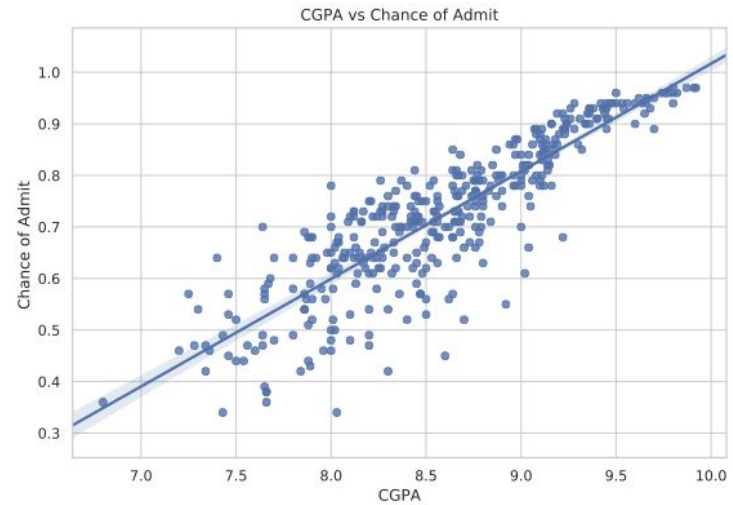
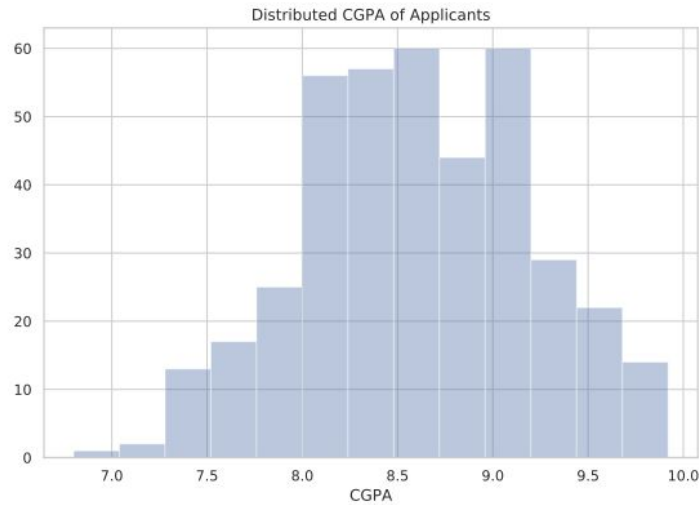
Features Vs. Target

## Features:

- GRE Score
- TOEFL Score
- University Rating
- SOP
- LOR
- CGPA
- Research

## Target:

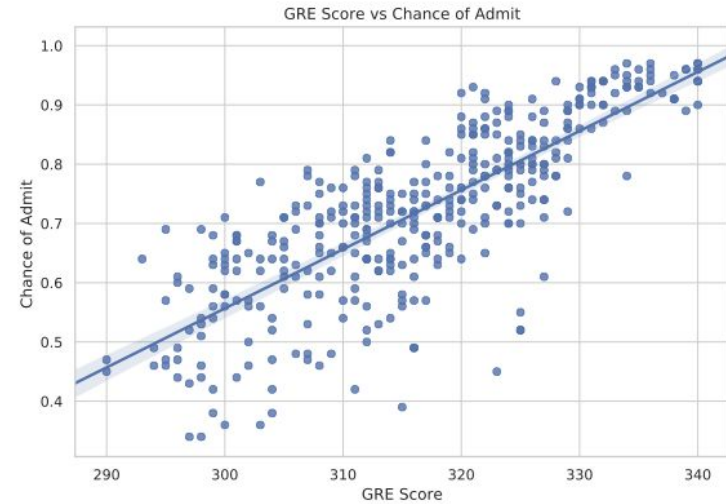
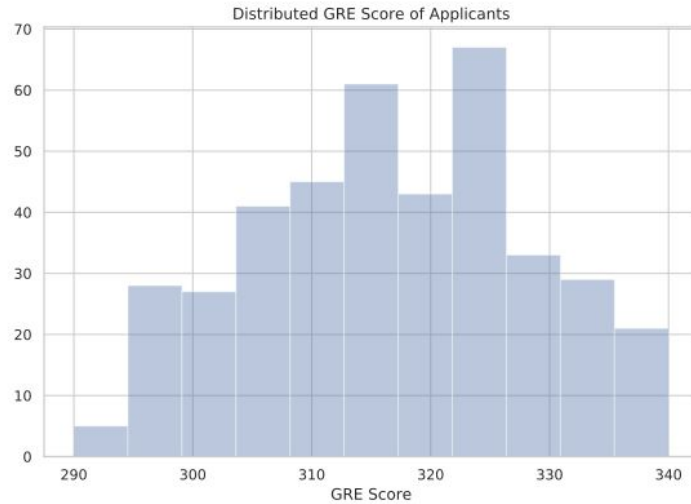
- Chance of Admit



## CGPA - Analysis

slope=0.20884722950069118,  
Rvalue = 0.8732890993553,  
stderr=0.0058403454390059265

Intercept=-1.0715116629342307,  
Pvalue =2.3365140004985676e-126,

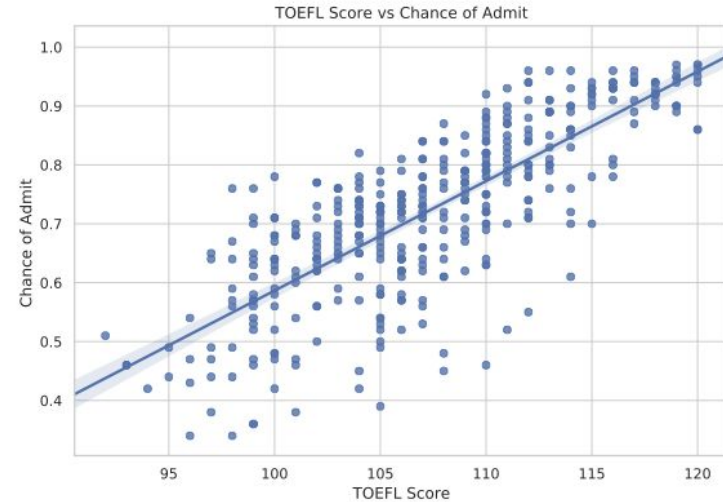
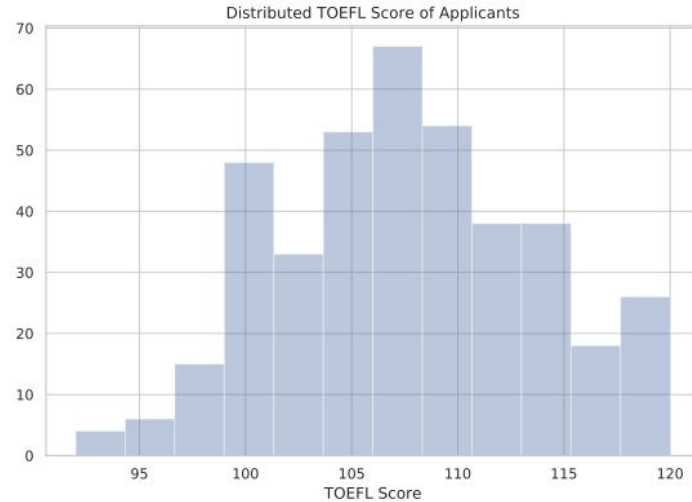


## GRE score - Analysis

Slope = 0.009975882025681376,  
Rvalue = 0.8026104595903499,  
Stderr = 0.00037163616056649364

Intercept = -2.4360842448510525,  
Pvalue = 2.458112414179777e-91,





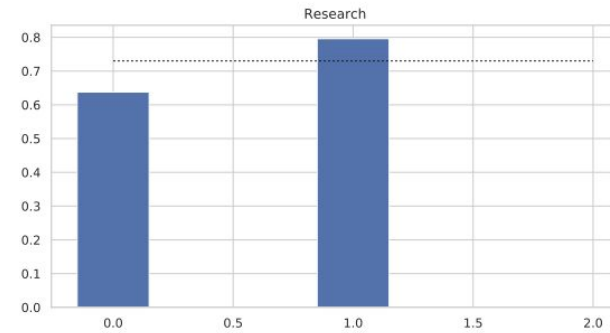
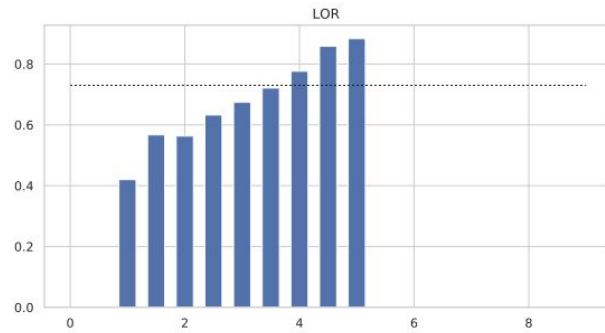
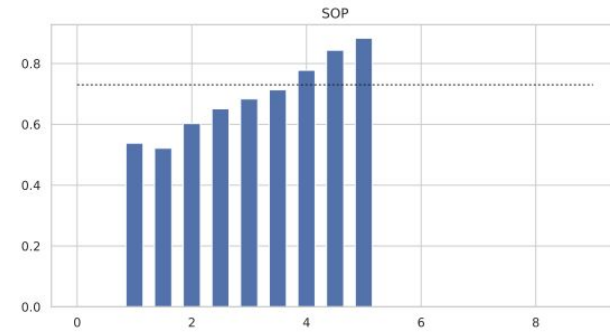
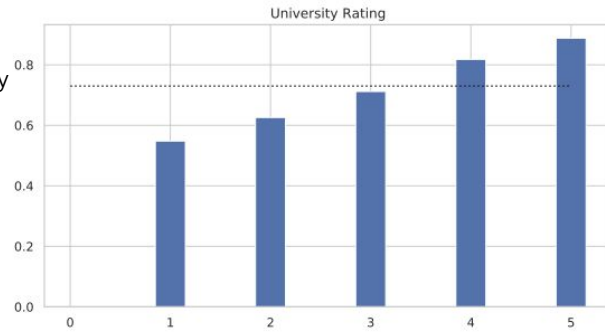
### TOEFL score - Analysis

Slope=0.018599296811431715,  
Rvalue =0.7915939869351045,  
stderr=0.0007196600972616436

Intercept=-1.2734004705158803,  
Pvalue =3.6341021759972773e-87,



Median of y  
= 0.73



Discrete/Categorical features



# Prediction - Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \dots$$

	x1	x2	x3	x4	x5	x6	x7
features	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research
Coefficients ( $\beta$ )	0.0017676	0.0029032	0.00771492	-0.0053149	0.0282844	0.11736896	0.019229

## Chance of Admit (Y)

$= x_1 * 0.00176755 + x_2 * 0.0029032 + x_3 * 0.00771492 + x_4 * (-0.00531496) + x_5 * 0.02828439 + x_6 * 0.11736896 + x_7 * 0.01922889 - 1.2748826685520998$

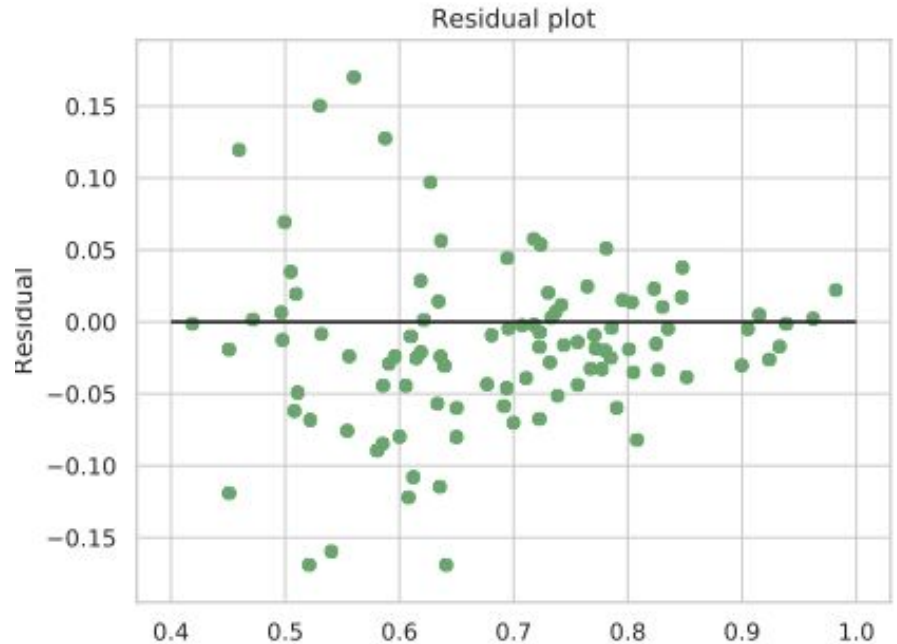
Intercept ( $\beta_0$ ) = - 1.2748826685520998

GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit	predictions_LR
324	110	3	3.5	3.5	9.04	1	0.82	0.800940
325	107	3	3.0	3.5	9.11	1	0.84	0.804872
330	116	4	5.0	4.5	9.45	1	0.91	0.905113
312	103	3	3.5	4.0	8.78	0	0.67	0.723805
333	117	4	5.0	4.0	9.66	1	0.95	0.923824



# Prediction - Linear Regression

- ❖ Multiple Linear Regression model since  $n > 1$
- ❖ Model summary:
  - $R^2 = 0.73$ ,**
  - Mean Squared error = 0.0035,**
  - Mean Absolute error = 0.0433**
- ❖ Since  $R \text{ square} = 1 - (SSE_{\text{Error}} \div SSE_{\text{Total}}) > 0.7$ , the relationship can be used.



# Prediction - Decision Tree Regressor



- Regression- Creates model to predict value of target (numerical variable) using one or more predictors (numerical and categorical variables)
- Decision tree regressor is a tree kind of model which has decision nodes and leaf nodes
- Standard deviation is use to calculate homogeneity of numerical samples
- Constructing decision tree is all about finding an attribute that returns highest standard deviation reduction
- Use that node as root node and split further
- Coefficient of Deviation is used to decide when to stop branching

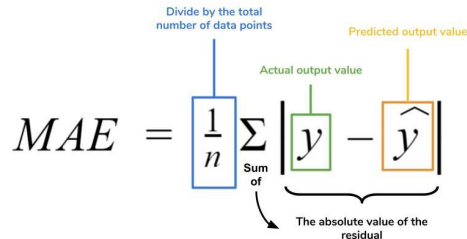
# How it works?

- Step1 -Calculate standard deviation of the target attribute
- Step 2- After splitting further calculate standard deviation for each branch.
- Step 3- (Standard deviation before the split) - (Resulting standard deviation) = standard deviation reduction.
- Step 4- The attribute with the largest standard deviation reduction is chosen for the decision node
- Step 5-The dataset is divided based on the values of the selected attribute and process continues
- Step 6- Model evaluation- MAE, MSE, R2

$$SD = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

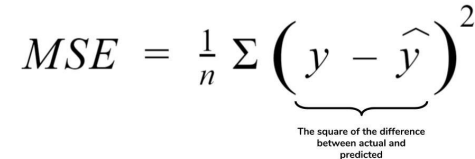
$$S(T, X) = \sum_{c \in X} P(c) S(c)$$

$$SDR(T, X) = S(T) - S(T, X)$$



The diagram shows the Mean Absolute Error (MAE) formula with several annotations. A blue box around  $\frac{1}{n}$  is labeled "Divide by the total number of data points". A green box around  $y$  is labeled "Actual output value". An orange box around  $\hat{y}$  is labeled "Predicted output value". A bracket under the absolute value term  $|y - \hat{y}|$  is labeled "The absolute value of the residual". The word "Sum of" is written below the summation symbol  $\Sigma$ .

$$MAE = \frac{1}{n} \sum |y - \hat{y}|$$



The diagram shows the Mean Squared Error (MSE) formula. A bracket under the difference term  $y - \hat{y}$  is labeled "The square of the difference between actual and predicted".

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$

# User interface

My Application

GRE Score

310

TOEFL Score

108

University Rating

4

SOP

4.5

LOR

4.5

CGPA

8.8

Research

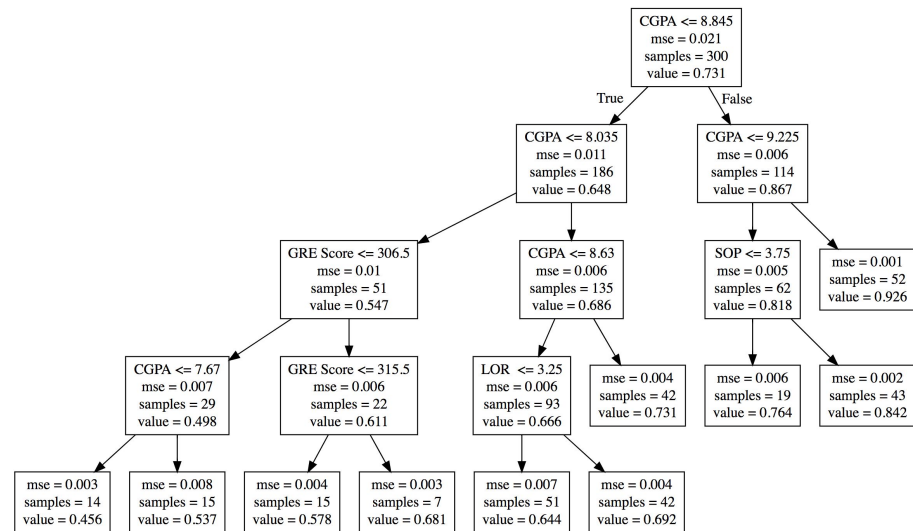
0

Predict

Quit

Predicted Acceptance rate :73.0%

Chances are: Low



## Sample Output

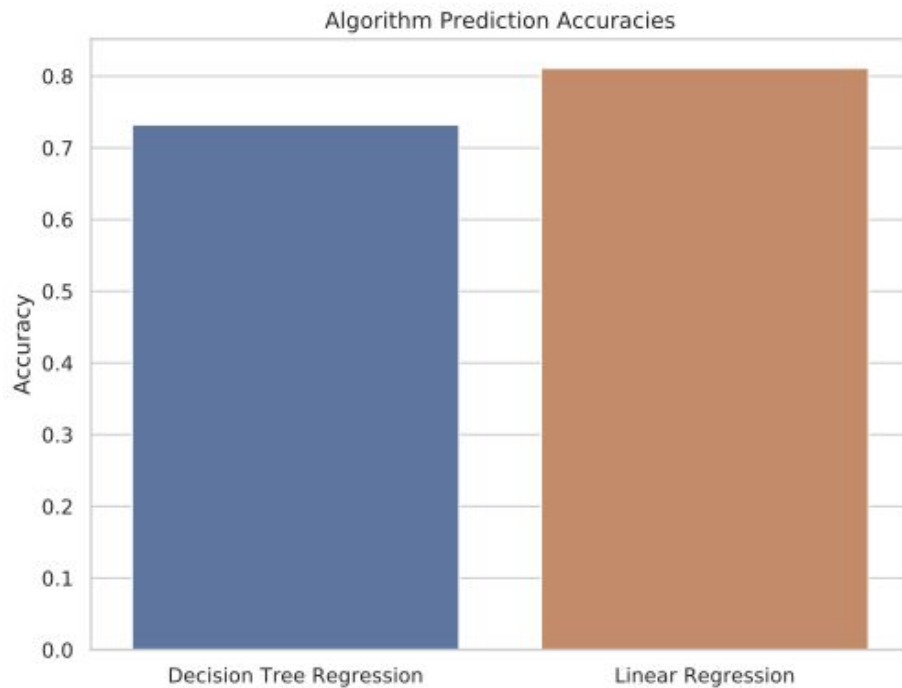
GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit	predictions_DR
324	110	3	3.5	3.5	9.04	1	0.82	0.763684
325	107	3	3.0	3.5	9.11	1	0.84	0.763684
330	116	4	5.0	4.5	9.45	1	0.91	0.925769
312	103	3	3.5	4.0	8.78	0	0.67	0.730714
333	117	4	5.0	4.0	9.66	1	0.95	0.925769

## Model Summary:

- **R2 = 0.735**
- **MAE = 0.055**
- **MSE = 0.0050**



# Comparison analysis





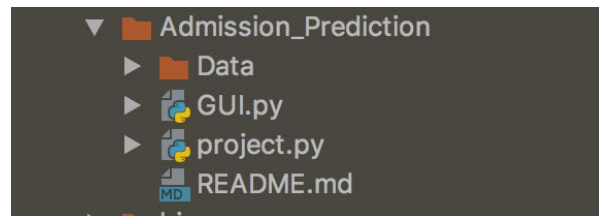


# Code - [Github](#)

Python libraries:

- Numpy
- Pandas
- Matplotlib
- Sklearn
- Os
- Seaborn
- scipy

Code Structure



# Thank You