# Predicting Graduate Admissions: An Analysis of Factors Affecting Admission Success and Model Results

05/06/2024

# Final Project - “Yashwanth Reddy Kovvuri”

## Introduction

Graduate application entails of complex procedures such as gathering all the required data and information that the applicant has to secure a position in the graduate school. Often totalizing, the universities look into numerous particulars, which are usually: standardized test scores like GRE, TOEFL and IELTS, academic performance, recommendation letters, statements of purpose, and research experience and other required documents.

The goal of this measuring is that let a knowledge processing candidates qualifications and their success in an advanced training. While students process their educational plan, they have to deal with a number of conditions and difficulties that incesmentally shape their application strategy. This introduction is meant to cover the main aspects of graduate admission and lay the foundation of understanding the overall studying process which not only includes the assessment and final selection of the applicants.

## Data Source

This data set used for analysis contains 400 records and also contains 9 variables sourced from Kaggle.

The data set is taken from : https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

### Aim

The aim of chance of admission is derived into different goals:

* **Identifying the factors** which plays a significant impact on the chance of admission to graduate programs which involves identifying the various variables in the dataset such as GRE score, TOEFL Score, CGPA, Research which helps in determining the variables influence on chance of admit.
* **Examining the relationships** among different variables is used to understand how the variables interact to admission chances. By analyzing the relations and other patterns, we can have many valuable insights that shows the connections between factors and the chance of admit.
* **Developing predictive models** like linear regression and random forest helps to estimate the probability of admission taking the given variables which involves selecting most relevant results and using them to build a model which can predict the acceptance to difference universities based on the requirements.
* The main aim is to provide the applicants a valuable insights of the chance of admit which helps them to understand which factors help them to improve their chance of admission to their dream university..

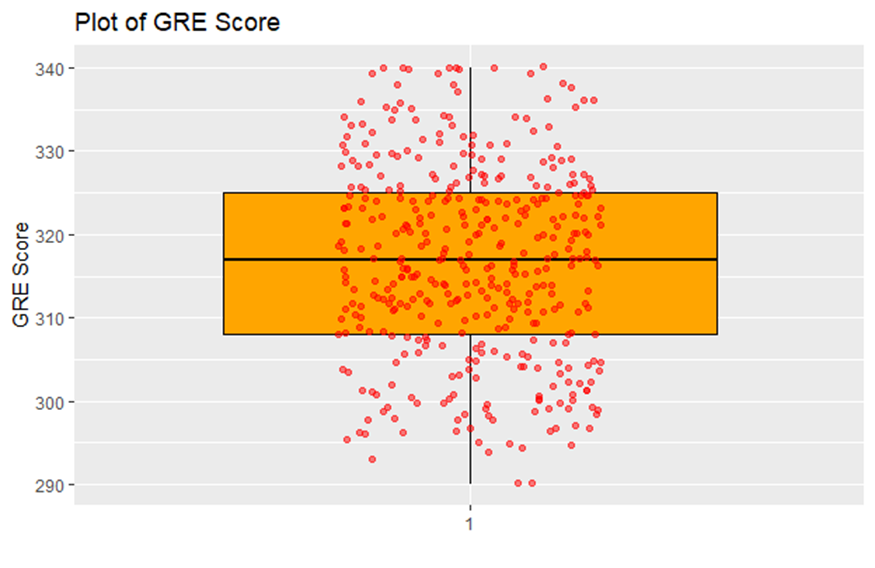
### Variables

The variables are classified as :

1. **GRE** Scores ( out of 340 )
2. **TOEFL** Scores ( out of 120 )
3. **University Rating** ( out of 5 )
4. **Statement of Purpose** and
5. **Letter of Recommendation Strength** ( out of 5 )
6. **CGPA** ( out of 10 )
7. **Research Experience** ( either 0 or 1 )
8. **Chance of Admit** ( ranging from 0 to 1 ) - we are gonna predict this using above variables.

## Visualization

#### Visualization1

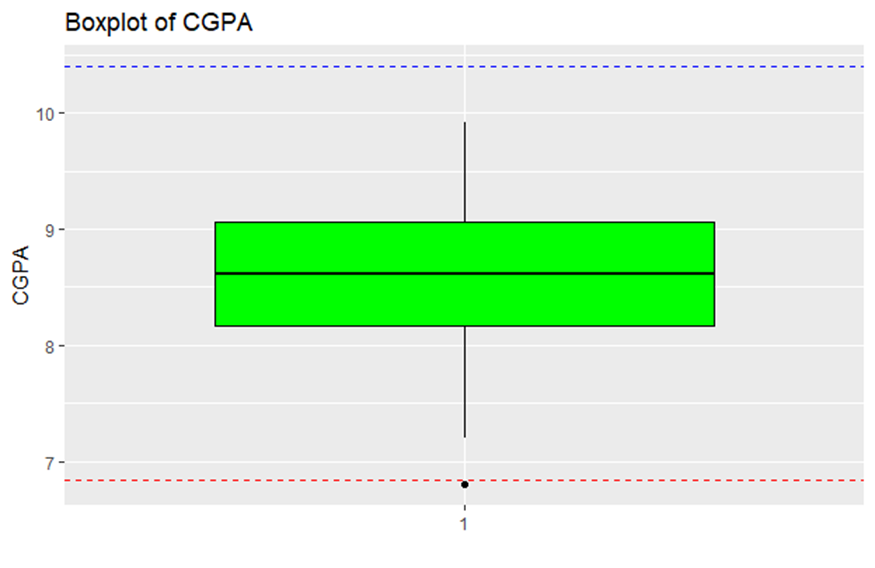


The **upper boundary and lower boundary** is calculated for **finding** the **outliers**. The GRE score higher than the upper range, shows that the outliers on higher side and if the GRE score is lower then the lower range, then the outlier is on lowest side. Also, shows the the number of outliers which is used for data quality in statistical analysis.

The orange box plot shows the interquartile range of GRE scores which is from 310 to 325. The bulk GRE scores are in the middle.The median GRE score is is by the line in the box which is around 320.

The red dots on the box plot which are spreaded horizontally to reduce overlapping and helps in finding the density of data points.

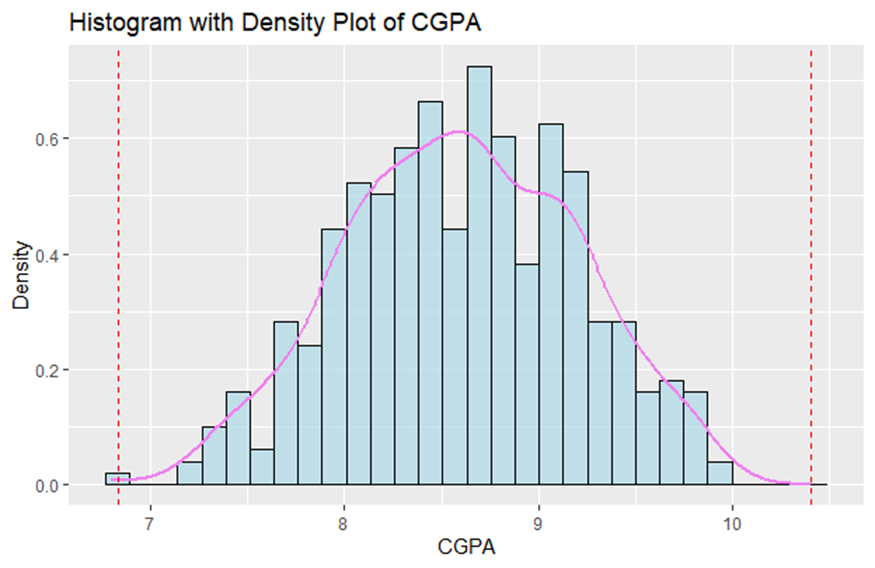
#### Visualization2



**Box plot** The green boxplot consists of the CGPA data which is around 8.0 to 9.0 and the majority of the students in the range of 50% i.e. in the middle of all CGPA values.

The median CGPA is 8.5 which is visualized as a line in the middle of the box plot.

The upper range shows the highest CGPA scores which is 10 and the lower range shows less variation at lower end where it shows only few students scored lower than 8.0.



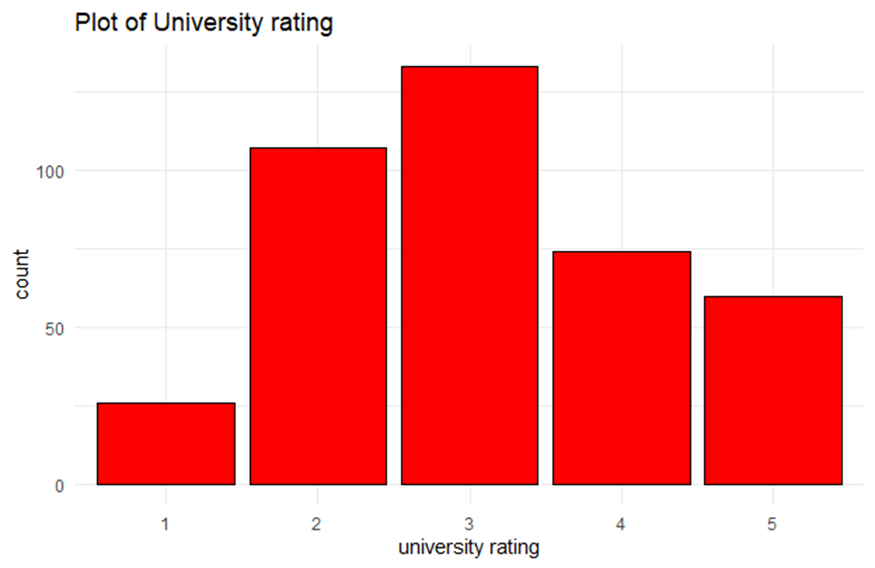
**Histogram** The blue bars is representing the frequency of students with CGPA ranges in different category.Most students land in the range between 8.0 to 9.5.

The density plot is used to estimate the data distribution which contains higher and lower density where it peaks around 9 CGPA which confirms that the most common CGPA is 9.0.

The dashed lines indicates the cutoffs within CGPA distribution and indicates the lower and upper ranges which helps to show the minimum requirement.The left line is used to show the CGPA around 7.0 and the right line shows the CGPA above 9.5.

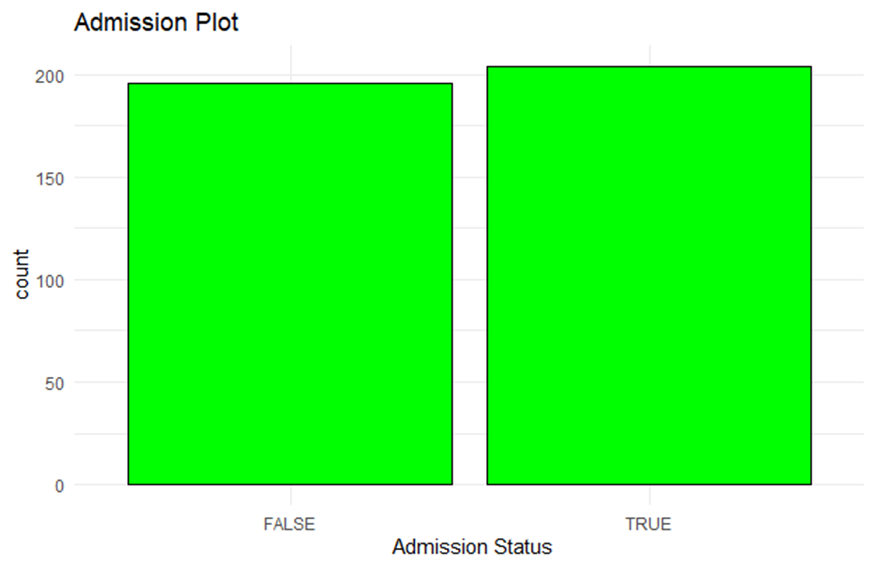
The histogram shows the performance of students based on their CGPA’s.

#### Visualization3



The bar height is used count the universities which are rated around 1 to 5 where; Rating - 1 : indicates the lowest rated university Rating - 5 : indicates the least common university. The most common and largest population are in universities which are rated as 2 and 3.

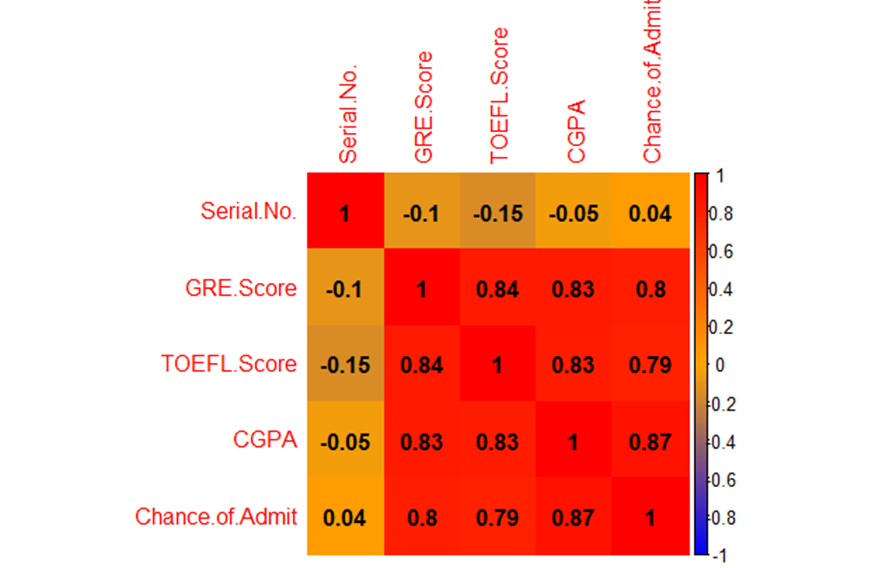
#### Visualization4



Admission Plot shows a bar chart comparing the counts of two admission statuses: The function of memory, therefore, is to either distort or retain accurate information. The bars represented the heights of the number instances for each status. If the bars are evenly thick, it means that roughly the same number of applications to this program were either declined (FALSE) or approved (TRUE). Such a chart is helpful in determining a very quick overview of the balance or distribution of admission outcomes in the data set, suggesting that the given data set most probably has the same amount of positive and negative cases resulting in fair training of predictive models without any hidden biases towards any one outcome.

#### 

#### Visualization5

The above correlation matrix is used to display the relationships between various variables.

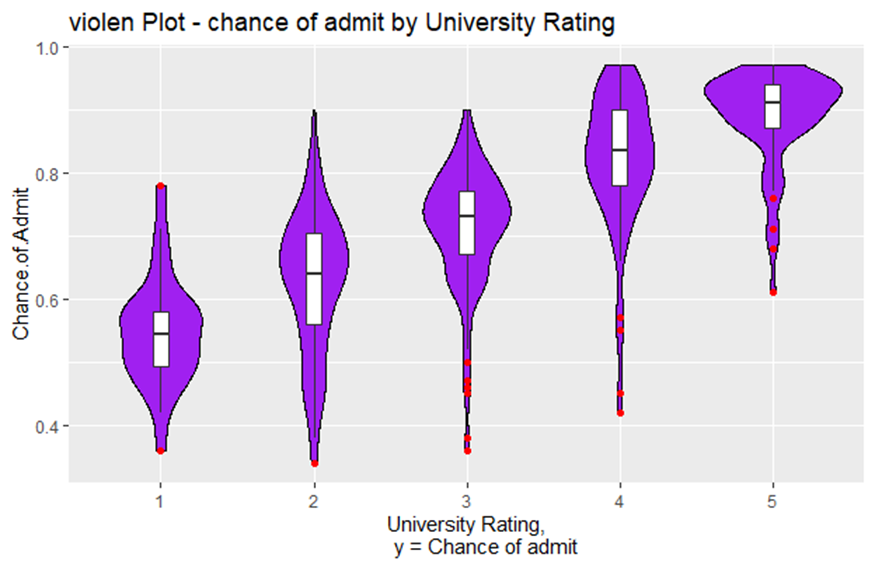
The “serial.no.” has no correlation with any academic variables.

The “GRE scores and TOEFL score” show strong relationship of 0.84 which indicates the candidates who score good on one of the tests tend to score well on other and high chance of admission.

Also, The “CGPA” plays and important role for an admission which also adds a value for the admission including “GRE and TOEFL scores”.

The “Chance.of.Admit” has highest correlation with GRE(0.8), TOEFL(0.79) and CGPA(0.87) which shows that higher values are associated with higher admissions.

#### Visualization6



The university which is rated as 1 is used to display a narrow distribution with low chance of admit which ranges between 0.4 and 0.6

The university which is rated as 2 is used to display wider distribution than rating 1 with chances between 0.4 and 0.7

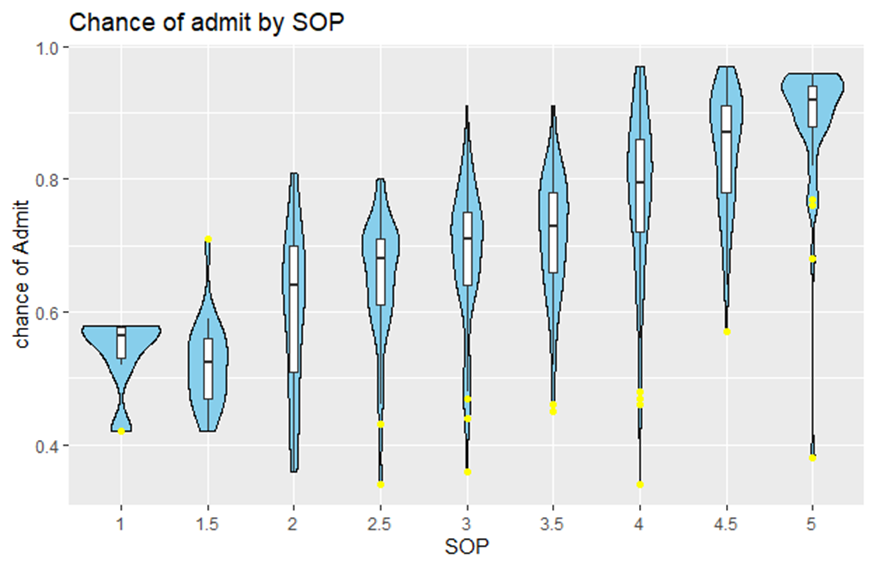
The university which is rated as 3 as a broader range of admit chance which ranged between 0.45 and 0.75.

The university which is rated as 4 shows larger distribution extending from 0.5 to 0.85 with tight concentration shows higher chances of admit.

The university which is rated a 5 is wide and assymetrical distribution which shows the chance of admit from 0.6 to 1.0 which is a high probabilty of admission in high rated universities.

The ‘red points’ are the outliers with each category of university rating.

#### Visualization7



The SOP1 shows very lower chance for getting an admission of a value below 0.6 and notable outlier around 0.4.

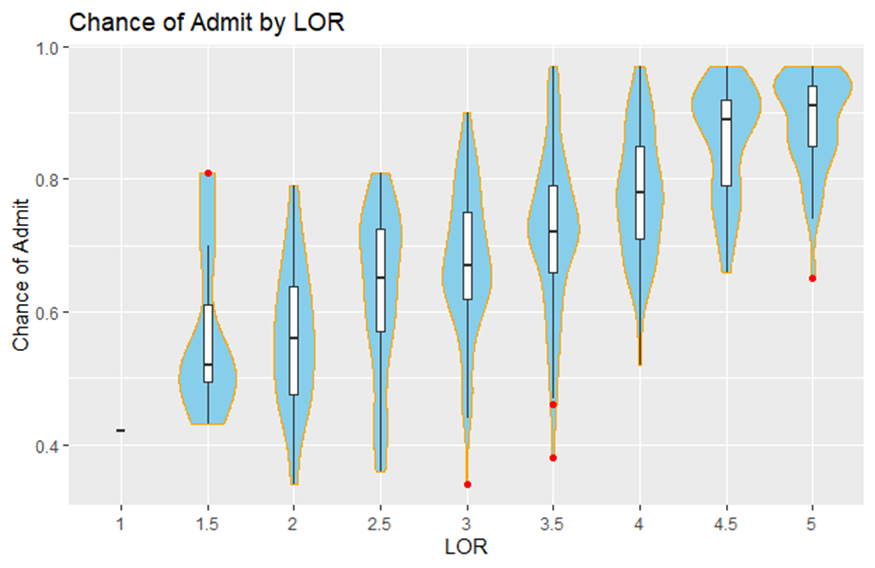
The SOP2 shows a little better chance of getting an admission probabilities compared to SOP1 but low.

The SOP2.5 and SOP2 shows a wide range of admission possibility which extends into 0.7, but the SOP2.5 have high range of distribution towards high chance.

The SOP 3 and SOP 4.5 have increasing median value and range of admission chances and with SOP 4.5 have high chance.

The SOP5 is where the majority of points clustered around 0.8 to 1.0 and have broader distribution.

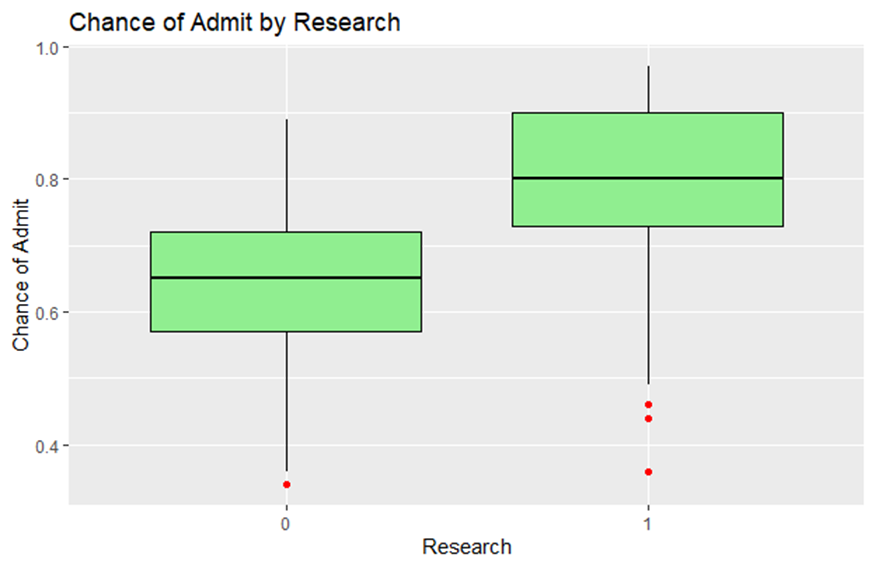
#### Visualization8



The violin plot shows the relationship of “Letters of Recommendation” (LOR) score and the “Admit Chance” into the university’s program, with scores from 1 to 5 being distributed. Prospective students with LOR ratings at the bottom (1 and 1.5) tend to rarely get accepted as their chances are less than 0. 6. LOR ratings between 3 and 4 score relatively higher admissions probabilities compared to the rest. 5 the clusters turns out to be denser with higher median values of admissions possibilities. The LOR rating between 0 and 5 demonstrates the most favorable outcomes, with the biggest standard deviation occurring at the highest rating of 5.

Red dots show outliers, pointing out instances where the candidates either exceeded the typical probabilities of acceptance to a great extent or fell short of such typical probability mistakingly. These figure illustrates strong effect of powerful recommendation letters on the chance of university admission.

#### Visualization9



The box plot represents the “Probability of Admit” depending on whether internship has completed(denoted as 1) or nor(denoted as 0). Subjects with research experience demonstrate a median level of certainty advancing to the next round of admission, and the middle 50% range spans from about 0.

0.65 to 0.9, students who do not have research experience have a low median admissions chance, which reaches zero to the most competitive colleges.

0.55 to 0.75, there are exceptions with respect to both groups especially for the highly experienced applicants who would have admission probability well below the trend line. Such plot emphasizes the key role of undergraduate research experience in providing competitive advantage when it comes to admission into a university program.

## Modeling

For modeling, we need to remove certain values Because the data set contains large number of values with inadequate representation. It would be difficult to create a model with less amount of data. Therefore, by filtering the data set to include accurate representations from each factor.

renaming and reordering the variables is done. so that the data set will be accurate enough for modeling and can give as accurate predicted results. Also, finding the dummy variables helps in creating perfect model by predicting accurately.

**We need to split the data set into training and testing set. I have divided the data as 80% of data as training and the remaining 20% data as testing which will be used for predictions.** The trained data will be used for training the model and the testing data will be used for predictions.

### Linear Regression

A linear regression model in order to predict the value of “Admit Probability” as an outcome variable with the help of the predictor variables of a dataset admitted\_train. The predictors that you have chosen are “GRE” (Graduate Record Examination) which is standardized test in the US for evaluating the qualifications of candidates for graduate study in the US.

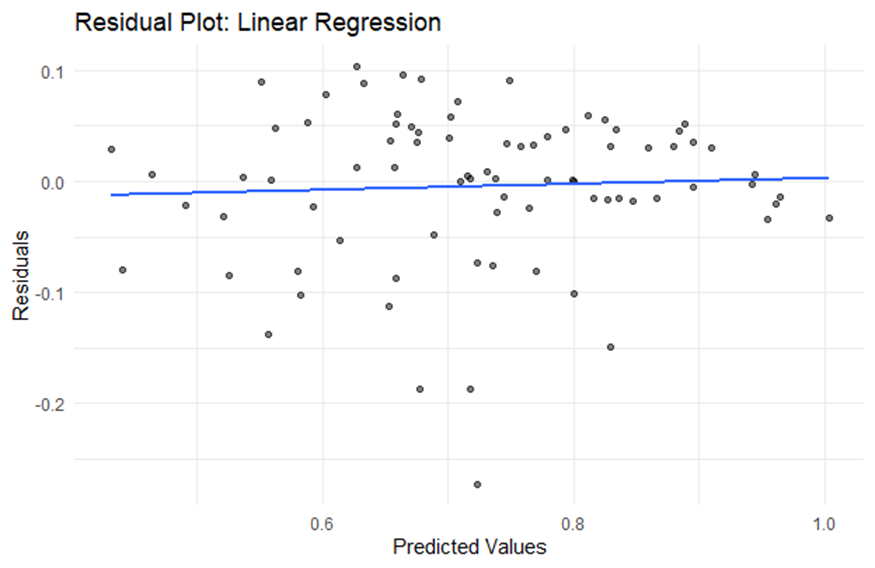
Students provide several pieces of information during the counseling session such as their CGPA, whether they have research experience or not (“Research”). This divergence employs linear\_reg() function from parsnip package which is a tidy modelling framework. The line of code set\_engine(“lm”) indicates that lm function that belongs to r base will be used to execute processing commands for the linear regression model. As the second part of the model setup, it is fitted using the fit function. This strategy, representing a standard structure of a predictive(forward-looking) characteristic modeling observed in statistical analysis, is focused on the explanation of a role of different academic modes and research experience in the probability of admission to some university program. The features selected reflect a theory that both test scores and research potential are key features in selecting a candidate.

**Tidy Model**

The tidy() output gives a compact summary displaying the model’s coefficients and statistics pertaining to lm\_fit, a linear regression model trained on the training data. Coefficients are shown as well as their estimates, standard errors, t-values and probabilities.

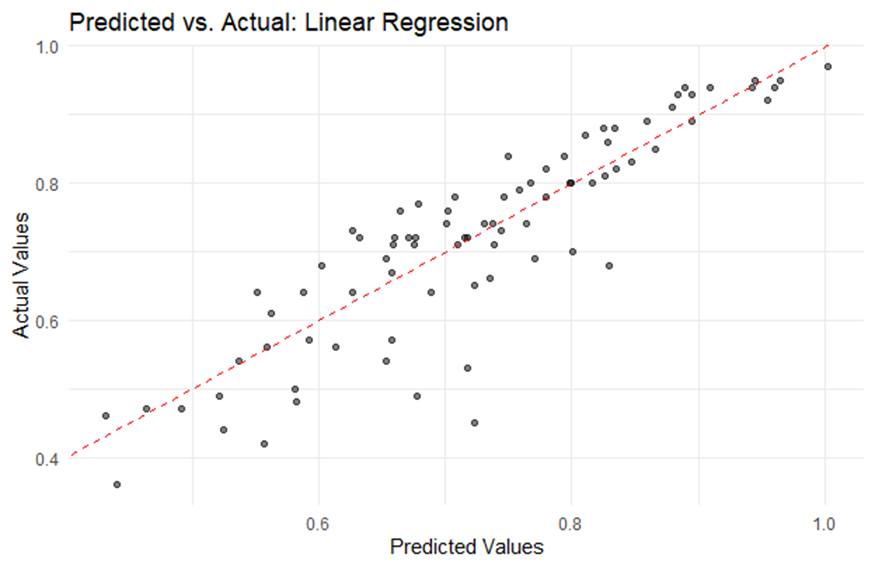
### Linear Regression Graph

#### Residual



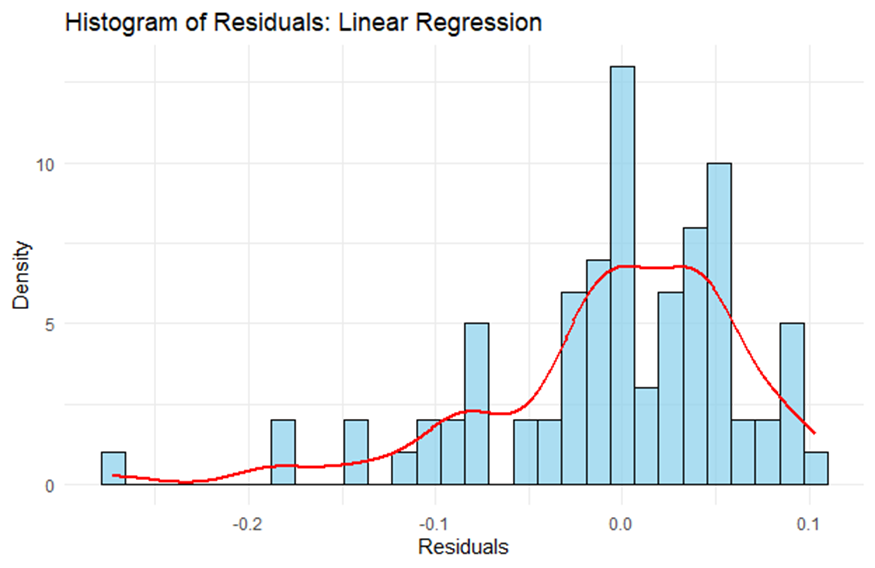
The residual plots a line graph that illustrates the goodness of model fit for a linear regression model. It plots the residuals (the differences between predicted and observed values) on the vertical axis along the predicted values on the horizontal one. Ideally, the residuals should be randomly distribute near the horizontal line of zero which shows no pattern. The residuals scatter around the zero line with no particular pattern that indicates that the given model is inline with the assumptions of linearity, independence, and homoscedasticity(equal variance). Nevertheless, there is a slight deviation that is mainly the case with higher scores of the dependent variable where the residuals are more spread out.

**Actual Vs Predicted**



The scatter plot between the predicted values of linear regression model and observed values as well as the common method of assessing model performance. The squares represent individual data points, with the predicted chance of admit plotted on the x-axis against the actual chance on the y-axis. The dotted red line of which represents 100 % prediction accuracy is used to show how close than the actual data the predictions are. The observations on this line are all part of the prediction. The distribution of points hints that the model arguably looks quite good, especially in the middle ranges of readings. Nevertheless, some points are located a bit farther away from the line which may indicate that different data points diverge from the predicted values, especially at the end of the predication spectrum. Let us draw a conclusion that the model is almost accurate. Nevertheless, it can be an object of improvement as it relatively lacks precision at the ends of the range of predictions.

#### Distribution of residual



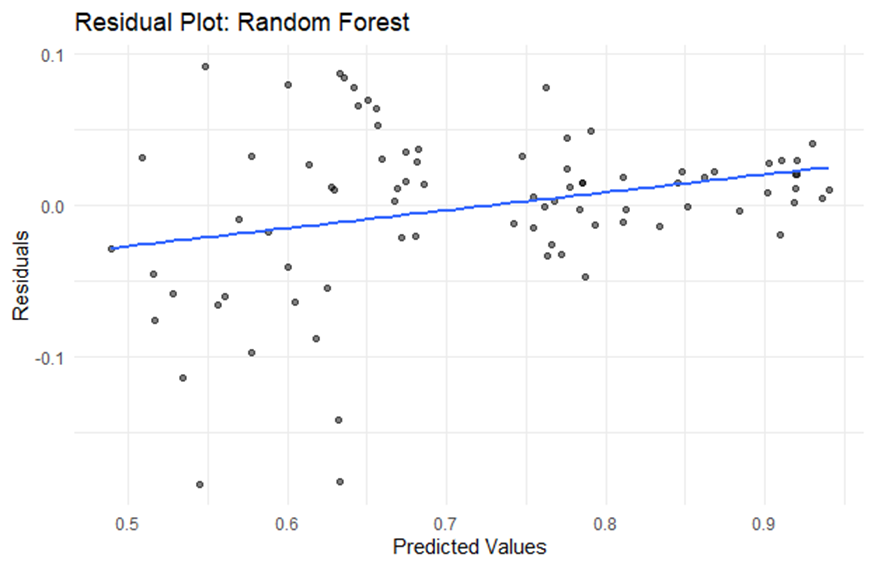
The histogram for the residuals of linear regression model and a red line representing the density estimate. The histogram specifics the distribution of residuals, which are the difference between the observed and the predicted values expected by the model. If residuals are as close to a normal distribution as possible, centered around zero, the model may be deemed to present a very good fit. For this plot, the residuals are genuinely clustered around zero, with a tinge of skewness, having the tail extending towards the negative side. On the other hand, a line can be seen half of the way out in both directions, suggesting the presence of outliers or extreme residual values. With red density curves, we can easily see the overall shape of the distribution and this confirms that there is no perfect symmetry but roughly symmetric. The pattern, therefore, alludes to the fact that the model, though a decent fit for the data, may also have systemic errors or phenomena not captured by the model, as portrayed by the skewed values and the distribution tails.

### Random Forest

The Random Forest Model which underlies this is a more enlightened machine learning technique that predicts the probability of being admitted to graduate studies. This is possible thanks to the fact that decision trees are trained as a set of ensemble methods. The latter outperform linear algorithms such as linear regression both for the accuracy and stability. Random Forest’s capability for handling nonlinear relationships among multiple predictors which consist of GRE scores, being able to represent them even better, as well as TOEFL scores, CGPA, and research experience is significant for solving this problem. It often leads to the smaller RMSE and MAE, therefore, what in turn means the higher accuracy in forecast, and the higher R-squared, which symbolizes the greater power of the explanatory. On the other side, it allows to pull some of the factors out of the equation, reducing the risk of overfitting and leaving reliable factors which are the admit key engines. Universities might, therefore, as an application of those qualities, embrace Random Forest as one of the instruments to sharpen their admissions by discovering insights through data.

### Random Forest Graph

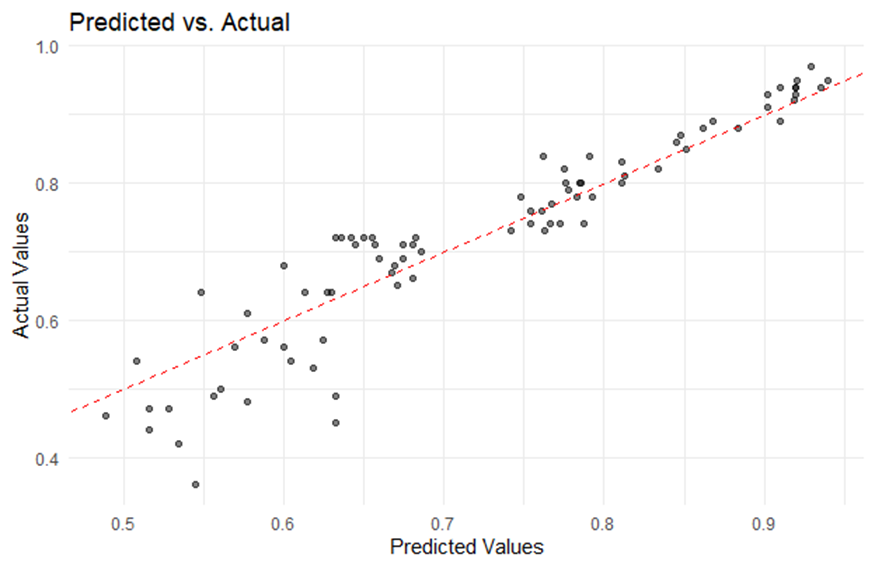
#### Residual



Residual plot for Random Forest regression model, and it helps in visualizing the distance between known actual values and predicted values in a dataset (residual errors)The horizontal consolidates for responded values, and the vertical imagining for the residuals. Ideally, residuals should be uniformly distributed around the horizontal line at zero point, this way, we know that the prediction of the model doesn’t show any biases at different levels.

In the course of conception of your plot it comes across that there is a slight upswing of residuals in terms of which they increase as predicted values increase listening to the assumption that the model most likely underpredict the results at higher values. Such composition often meaniled the presence of the non-linearity in the data in which there might exist no totally deferred capturing by the Random Forest model. In addition, the presence of the several points that drastically move off zero could be the instances of outliers or model has the significantly higher error typically. The plot is very effective in terms of diagnostics of model’s performance and pinpointing the part of model generated predictions that need improvements.

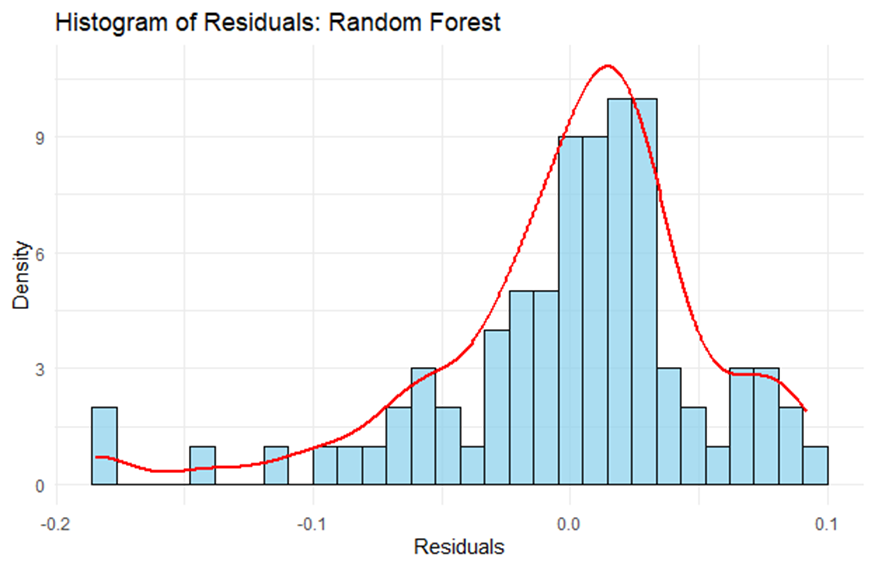
#### Actual vs Predicted



It shows a chart with the predicted values and the actual values, usually to judge the accurateness of the model. The x-axis draws the predicted values while the y-axis presents the observed values. The bottom dotted line illustrates a perfect forecast and refers to the state where the forecasted values and the actual ones match together.

From the plot in this case we can see that most points are rather close to the line suggesting the model used performs adequately in predicting the outcomes. But we have breaks, which specifically appear for the lower and higher predicted cognitive function where the model underestimates or overestimates, respectively. These output plots serve to visually scrutinize the model’s capability within different value range and detect the areas in which the model may require additional shaping or the data shows a high degree of variability.

#### Distribution of Residual



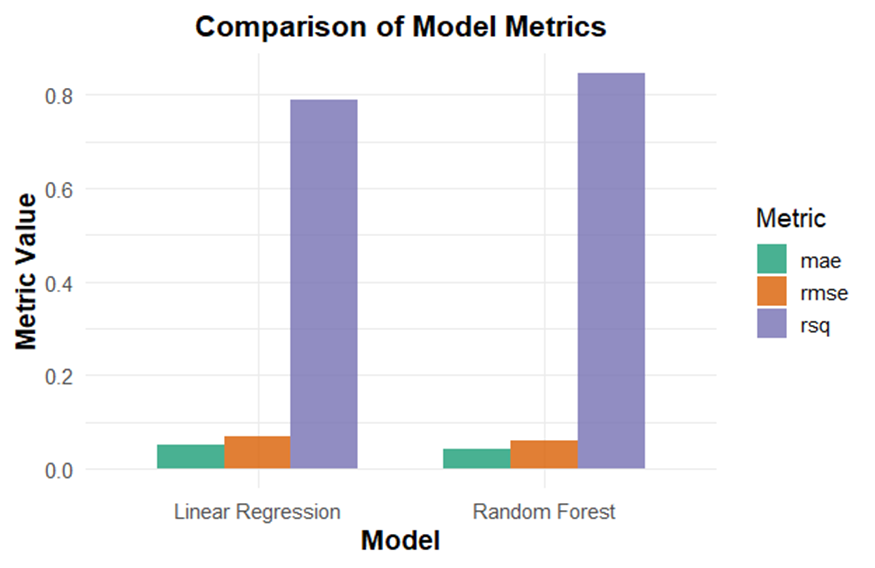
The histogram of residuals from a Random Forest model. These are the deviation from the predicted values to the actual values. Optimally, the residuals should become randomly dispersed around zero meaning that model is biased in no way.

In your plot, the residuals are on average centered on zero, and the general shape closely matches the blue line, which represents the normal distribution. Thus, the model errors are relatively unbiased and normally distributed, which illustrates and reflects model performance. Nevertheless, there is some behavior at the tails especially towards the negative side which could cause prediction of some data points to be less accurate. This histogram assists in model diagnostics, checking the error distribution and refining the model training or data preprocessing procedures.

### Comparison

The metrics data presents a comparison between two models: Linear Regression and Random Forest are the models which are very popular. Here random forest model gives the the smallest measured value (0). Cgroup (0. 069) corresponds to LINreg (0. 059) indicating that RF may be precise. Very closely, Random Forest also portrays smaller values in MAE. While both approaches showed some correlation (0. 446 versus Linear Regression (0. 0503)), it was apparent that the latter was more precise (p= 0. 0503). Moreover, R-squared (RSQ) metric, which is responsible for measuring vairiance, is exceeded for the Random Forest model (0). (853) about 78. 9% for Linear Regression compared with Random Forest (0.884) implying that Random Forest accounted for more variability in the data.

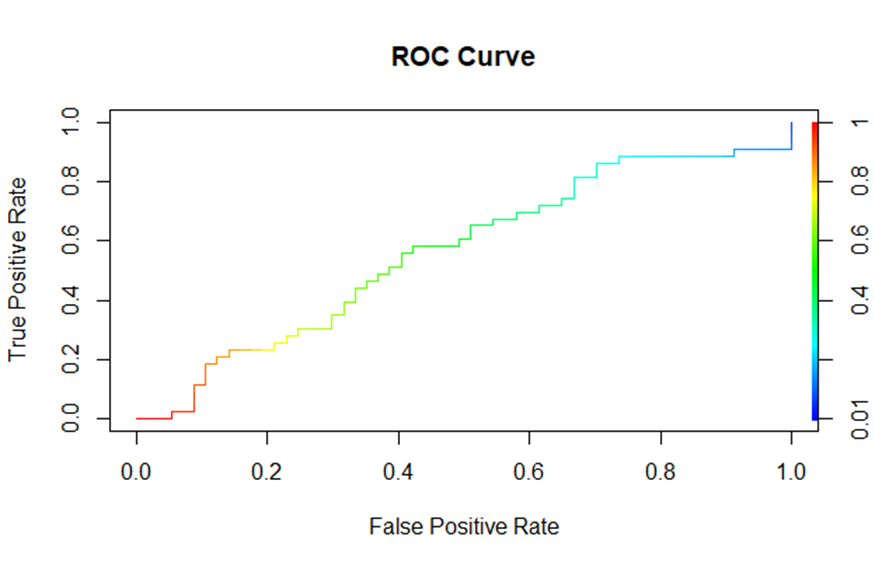
The **Random Forest model is more precise compared to the Linear Regression model.**



Plot shows comparison of Model metrics : Linear Regress vs Random Forest. Metrics assessed are MAE(mean absolute error), RMSE(root mean squared error) and R-squared (R^2), respectively. In case of both Linear Regression and Random Forest models, we can clearly see the **Random Forest model achieves significant improvement with respect to all three evaluated metrics.**

Otherwise, the RSQ value, implying the percentage of the dependent variable variation that is covered by the independent variable(s), is much higher for the Random Forest, proving a much better fit on the data. In addition, lower MAE and RMSE values achieved by Random Forest algorithm indicate a model which shows less average error per prediction and less variability in its error as compared to the Linear Regression model. With this bar chart one merges the performance capabilities of Random Forest which was chief in this specific comparison and thus make it a most probable model of the dataset palatable.

**ROC CURVE of training dataset**

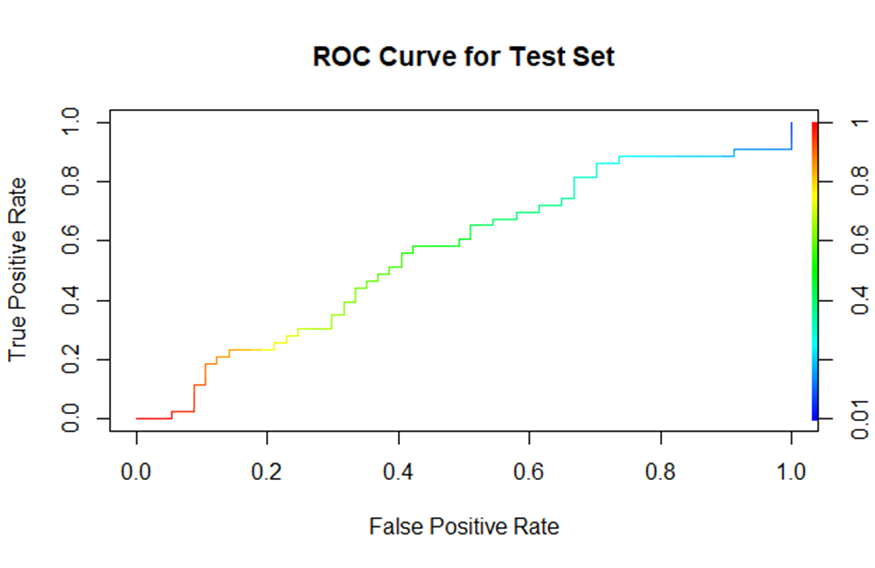


ROC curve is used for measuring the reliability of binary classification models by plotting the True Positive Rate (TPR) on the Y-axis and False Positive Rate (FPR) on the X-axis at different threshold levels. The different parts are colored differently—red, yellow, and green—to indicate distinctive performance phases.

The range to the right demonstrates a sudden increase in TPR and a relatively insignificant increase in FPR, indicating good performance. On the way to yellow and then green, the rate of positive TPR-change is slowing down signaling that the model is no longer as efficient as before. And the last blue line might determine the point of saturation where changes of the threshold have no effect on true positive and false positive values.

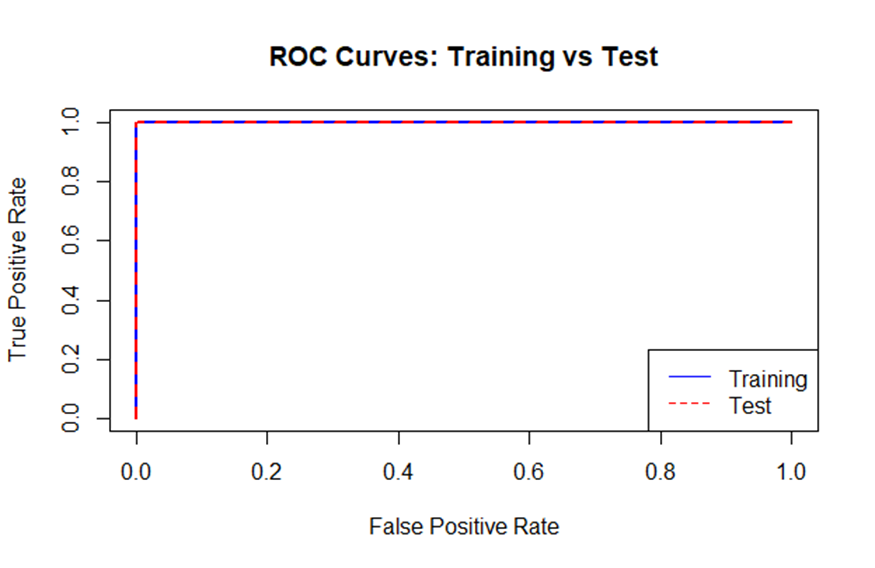
An appropriate ROC curve shows a large area under the curve (AUC), whose maximal value must be 1. The nearer the AUC to 1 means, the better the model is at discriminating between the classes. The plot demonstrates the relationship between sensitivity and specificity and the optimal threshold to achieve this balance.

**ROC CURVE of testing dataset**



The line chart charts a model’s performance in the area of binary classification by drawing genuine positive rates (TPRs) against false positive rates (FPRs) for different threshold specifications. Improvement in output performance is indicated with the red to green gradient whereas the better the threshold, the greener the gradient. A vertical initial steeper rise over the red section reminds of a high value sensitivity at the low threshold, and many positive test results with very few false-positive ones are depicted. On green area of the curve false positive increase in association with the height of the curve, which leads to a loss of sensitivity. While each model will produce a specific plot, the best model will push the curve further toward the top-left corner, creating a low FPR and high TPR. The ROC curve demonstrates that the model is, in general, accurate, with a pronounce performance at lower threshold levels, having a good way to differentiate between real and fake positives.

**ROC CURVE Comparison**



The plot shows the ROC curves of both training as well as testing dataset of a classifier model and illustrates the model’s class difference potential. The ROC curves show the TPR versus FPR below which the threshold levels are plotted. Here the blue line is the train set and the red dashed line is the test data. The lines are almost situated right over each other, almost far the upper corner of the plot what indicates excellent model performance with a high true positive rate and very low false positive rates for both of the datasets. Such a satisfactory alignment serves as an indication that the model is free of overfitting on data that is used to train the model. It shows such strong accordance between training and tests success usually is a sign of powerful model.

**AUC**

The **AUC value was “1” for both training and testing validation sets** which indicates that the model can properly discriminate between positive cases and negative cases. In this case, the model could be good enough, but we cannot rule out the possibility of overfitting where the model learns good on the training data but without generalizing.

## Conclusion

The withholding of Graduate Admissions dataset revealed **that Random Forest outshined Linear Regression respectively in the precision, accuracy and the explained data variance.** High AUC values both in the process of images training and test sets implicate that the models have a perfect classification between benign and malignant lesions which may indicate overfitting. Through exploratory data analysis (EDA) and correlation studies it was revealed that there were most probably very strong ties between the GRE, TOEFL and CGPA scores with the “Chance of Admit,” therefore indicating that academic results plays a major role when making admission decisions. Besides, in university selection, research involvement and university rankings also have a considerable weight. When looking at the outstanding model performance, model overfitting with signs of its imperfection, happens and this requires undertaking additional actions for model robustness. Ultimately, this study has the power to reveal predictive modeling’s benefit in graduate admission, as further trials may help identify other indicators that influence decision-making.

Overall, The project is useful for universities, students to provide them with data-driven perspective on admission process which can be used to improve a quick decision making.