Junran Cao, Ricardo Diaz, Jeremiah Ogunla, and Osemekhian Ehilen
Professor Amir Hossein Jafari
Intro to Data Mining – DATS 6103
12/6/2021

Final Group Project Report – Group 4

Background

This report provides an overview of the Introduction to Data Mining (DATS 6103) Final Project, prepared by Junran Cao, Ricardo Diaz, Jeremiah Ogunla, and Osemekhian Ehilen. The project examined the factors that affected admission to the master's programs in the U.S. to see how well they predicted the admission. The project aimed to answer whether schools evaluated applicants based on academic criteria, such as college GPA and standardized test scores, or on non-academic criteria, such as the strength of personal statement/recommendation letter and research background. The project also examined if these criteria were equally important or one outweighed the other.

This report first describes the dataset and the data mining algorithms used. It then explains the setup of the project, which includes data cleaning, preprocessing, utilizing algorithms, and coding. Finally, the report describes the results and provides a conclusion. The list of references is displayed, as well as documented computer listings.

The Dataset

The dataset was obtained from Kaggle. Inspired by the UCLA Graduate dataset, it was created by several Indian students in 2019.[1] The data on 500 applicants were collected. The dataset contained seven predictor variables and one outcome variable. The predictor variables were GRE score (out of 340), which was expressed as an interval, integer variable; TOEFL score (out of 120), which was expressed as an interval, integer variable; University rating (out of 5), which was expressed as an ordinal, float variable; Statement of purpose, which was expressed as an ordinal variable and was named 'SOP'; strength of letters of recommendation (out of 5), which was expressed as an ordinal, float variable and was named 'LOR'; Undergraduate GPA (out of 10), which was expressed as an interval, float variable and was named 'CGPA'; Research experience (either 0 or 1), which was expressed as a categorical variable and was named 'Research', and 'Chance of admit' (from 0-1), which was expressed as an interval, float variable. The dataset contained ten columns, with one of the columns named 'Serial No.' The dataset contained no missing values.

Data Mining Algorithms

The data mining algorithms used for the project were linear regression, Decision Tree, and Random Forest. Linear regression was used to make strong assumptions about the relationship between the predictor variables (x) and the outcome variable (y). With our dataset being mostly quantitative, the outcome variable being continuous, and the predictor

---

[1] Mohan S Acharya, Asfia Armaan, Aneeta S Antony: A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019.

variables being categorical or quantitative, linear regression would increase the inferential power well.

The linear regression equation is:

$\hat{y} = b0 + b1\ x1 + b2\ x2 + \ldots$ where
b0: intercept
bi: slope/rate of change

a linear regression is shown below:

.

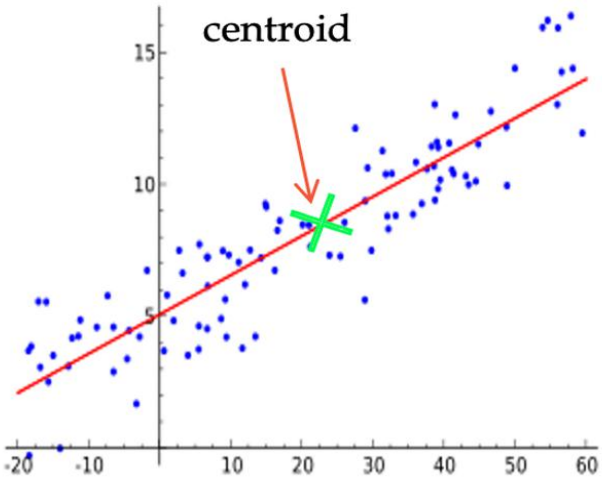

Figure 1[2]

Decision Tree was used for the ease of visualization and interpretation. A decision tree is a flowchart-like tree structure where an internal node represents a feature, the branch represents a decision rule, and each leaf node represents an outcome. The node on the top is the root node, which learns to separate based on the attribute value. The root node also recursively separates the tree.[3] Accordingly, a decision tree can be interpreted as if-this-then-that.

The illustrations of a decision tree and how it works are shown below.

---

[2] Tin, Martin. 'Intro to Linear Regression - Machine Learning 101.' *Medium*, DataDrivenInvestor, 6 Oct. 2020, https://medium.datadriveninvestor.com/machine-learning-101-part-1-24835333d38a.

[3] Python Decision Tree Classification with Scikit-Learn Decisiontreeclassifier.' *DataCamp Community*, https://www.datacamp.com/community/tutorials/decision-tree-classification-python.

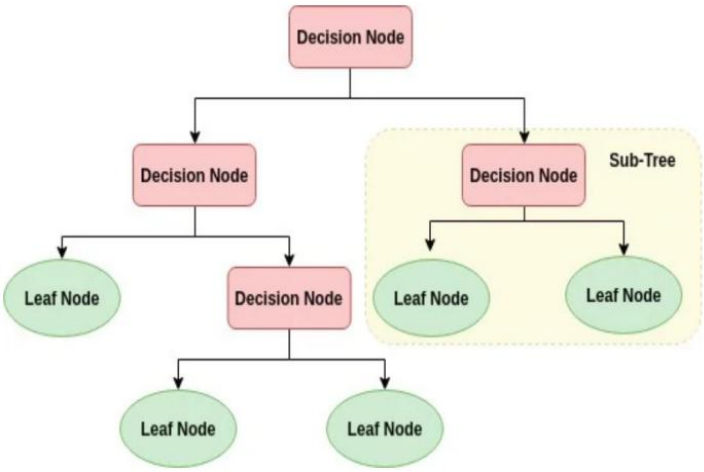

Figure 2[4]

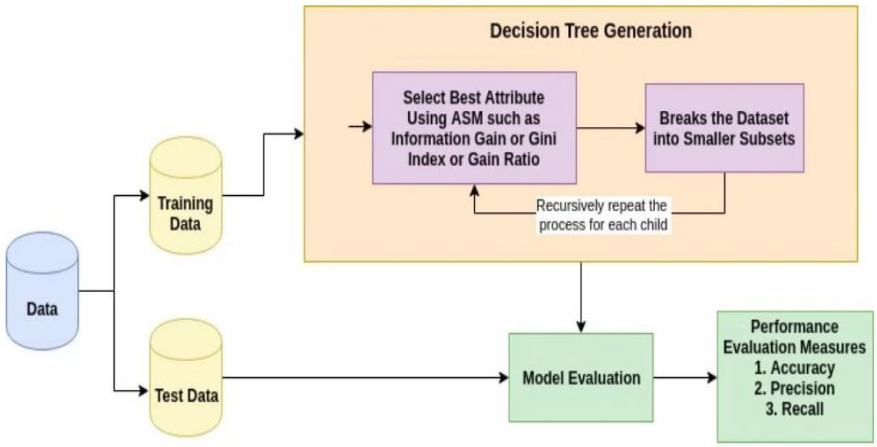

Figure 3[5]

An ensemble of decision trees, Random Forest, was also used. In a random forest, there is no correlation between different decision trees. When classifying, the new input sample enters, and each decision tree in the forest is evaluated and classified individually. Each decision tree, therefore, has its own classification result, and the most appeared classification result among the decision trees becomes the final result of the random forest.

An illustration of Random Forest is shown below:

[4] 'Python Decision Tree Classification with Scikit-Learn Decisiontreeclassifier.'
[5] Ibid.

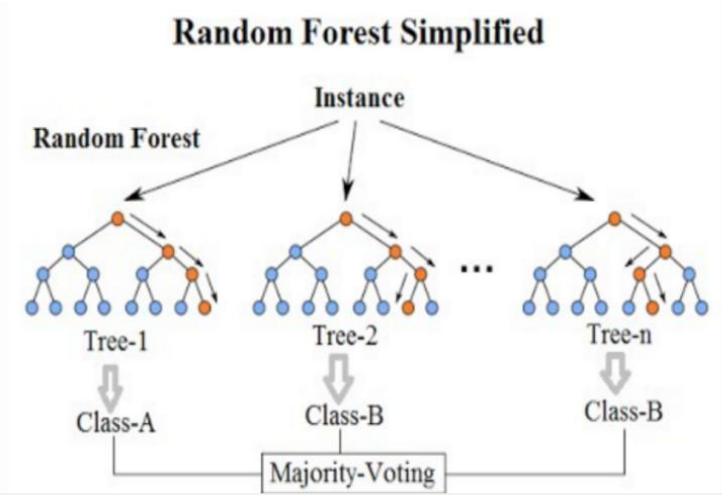

Figure 4[6]

The algorithm used in data preprocessing is binning. Data binning is a way of assigning original data values into bins which they fit according to their size. The original values are then replaced by values representing the corresponding intervals.

An illustration of data binning is shown below:

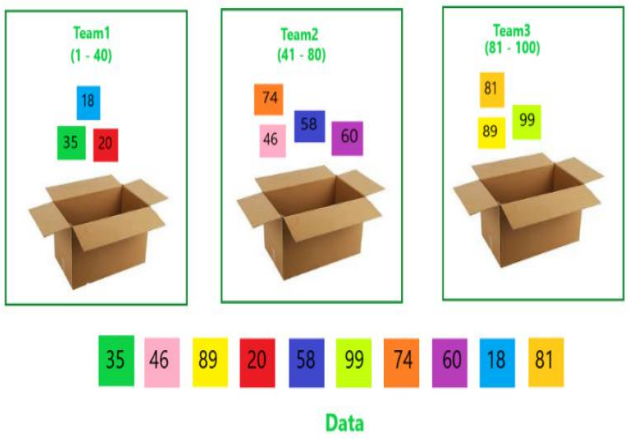

Figure 5[7]

Experimental Setup

Cleaning and Preprocessing

We started cleaning and preprocessing the data with eliminating the whitespaces in 'Chance of admit ' and 'LOR ' by renaming the variables as 'Chance of admit' and 'LOR'. Then, we made 'University Rating' a categorical variable. We also set 'Serial No.' as the index. Moreover, we binned 'GRE Score' and 'TOEFL Score' into four groups (1, 2, 3, 4) to better reflect the data. After that, we checked for 'NA' values and found no missing values. To "restrict" the data, we selected 'Chance of admission' with the values that were greater than or

---

[6] *Random Forest – Random Forest - EasyAI*. https://easyai.tech/en/ai-definition/random-forest/.

[7] Ibid.

equal to 0.01 and less than or equal to 1; 'GRE score' with the values that were greater than or equal to 1 and less than or equal to 340; 'TOEFL score' with the values that were greater than or equal to 1 and less than or equal to 120; 'University rating' with the values that were greater than or equal to 1 and less or equal to 5; 'SOP' with the values greater than or equal to 1 and less than or equal to 5; 'LOR' with the values that were greater than or equal to 1 and less than or equal to 5; 'CGPA' with the values that were greater than or equal to 1 and less than or equal to 10; and 'Research' with the values greater than or equal to 0 and less than or equal to 1. We also checked outliers with inter-quartile range detection for the integer and float variables, which are 'GRE Score', 'TOEFL Score', 'SOP', 'LOR', 'CGPA', and 'Chance of admit', showing in which array the outliers were. Inter-quartile range detection would reduce how the data is skewed, making it more normally distributed.[8] Finally, in case the authors accidentally entered wrong values for categorical variables when preparing the dataset, we also checked outliers for the categorical variables, which are 'University Rating' and 'Research'.

Visualization

Data visualization was an integral part of our data analysis; it helped us to better understand the data before conducting further analysis.

Heatmap

The heat map below sheds light on the correlations between the admission features.

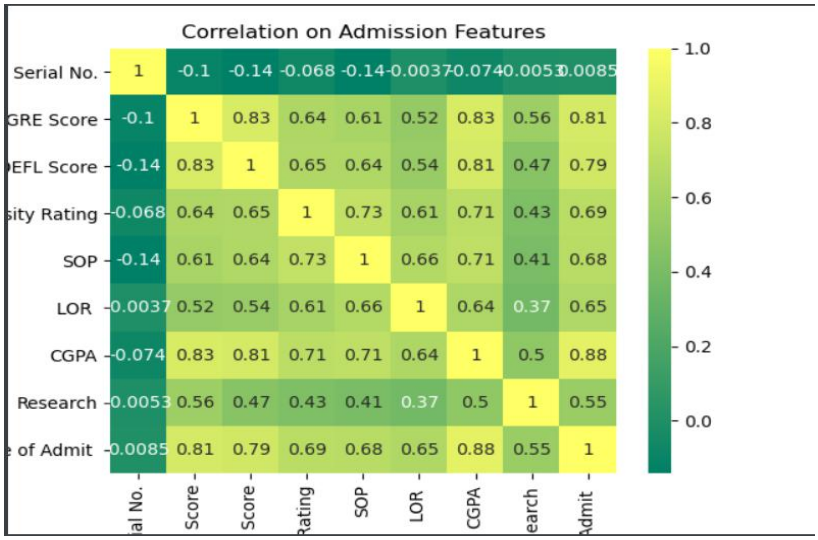

Figure 6

In a heat map, correlation ranges from -1 to +1. Values closer to zero mean there is no linear trend between the two variables. The closer to 1 the correlation is, the more positively correlated they are. A correlation closer to -1 is similar, but instead of both increasing, one variable will decrease as the other increases.[9]

---

[8] Zach. (2021, March 12). *Interquartile Range vs. standard deviation: What's the difference?* Statology. Retrieved December 6, 2021, from https://www.statology.org/interquartile-range-vs-standard-deviation/.

[9] *How can one interpret a heat map plot*. Cross Validated. Retrieved December 1, 2021, from

Figure 6 showed that 'CGPA', 'GRE Score', and 'TOEFL Score' were the three predictor variables that were more correlated with 'Chance of admit'.

Joint plots

Based on the preliminary results generated by the heat map, we then used joint plots to further understand the relationship between 'CGPA', 'GRE Score', 'TOEFL Score', 'Research', 'University rating', and 'Chance of admit'.

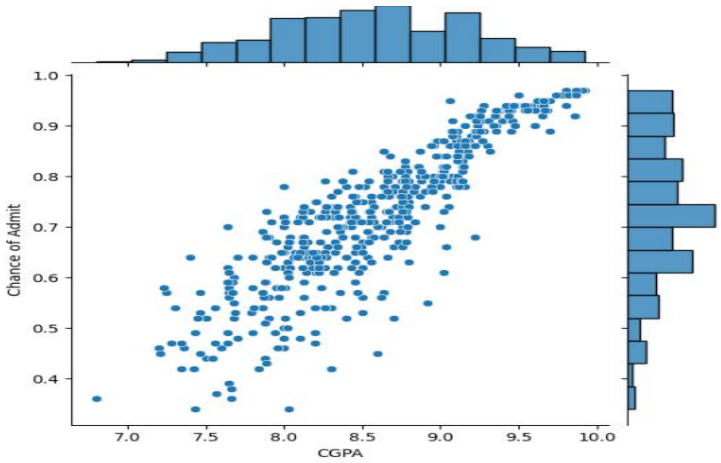

Figure 7

The central graph of Figure 7 showed a bivariate graph showing how 'Chance of admit' varied with 'CGPA'. From the graph, we observed a linear relationship between 'CGPA' and 'Chance of admit'. The plot that was placed horizontally at the top of the bivariate graph showed the distribution of 'chance of admit'. The third plot that was placed on the right margin of the bivariate graph with the orientation set to vertical showed the distribution of 'chance of admit'.[10] From Figure 7 showed that the applicants with a higher GPA tend to have a higher chance of admission.

---

https://stats.stackexchange.com/questions/392517/how-can-one-interpret-a-heat-map-plot.

[10] Team, P. (n.d.). *How to plot a jointplot with 'hue' parameter in Seaborn - Pretag*. Pretag development team. Retrieved December 1, 2021, from https://pretagteam.com/question/how-to-plot-a-jointplot-with-hue-parameter-in-seaborn.

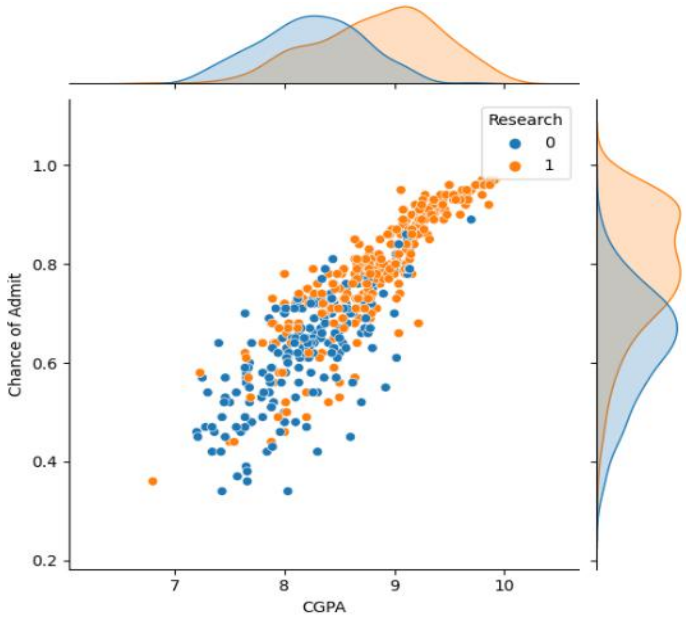

Figure 8

We then added 'Research' into the visualization. As shown by Figure 8, the applicants who had higher college GPA also tend to have research experience and a higher chance of admission.
This is particularly interesting since our correlation heatmap did not show a significant relationship between chance of admission and research.

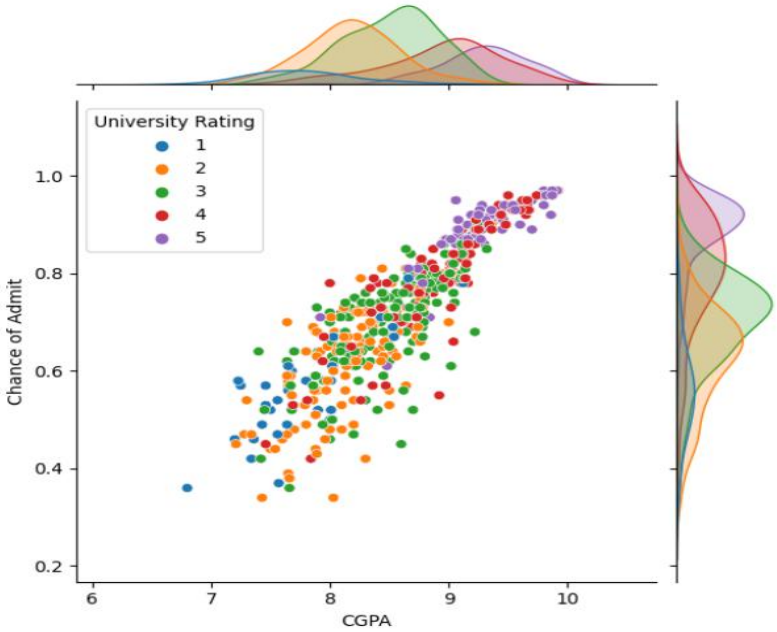

Figure 9

Plotting the relationship between 'University Rating', 'CGPA' and 'Chance of admit', we saw that the applicants with higher college GPA tend to come from higher-rated schools and have a higher chance of admission.

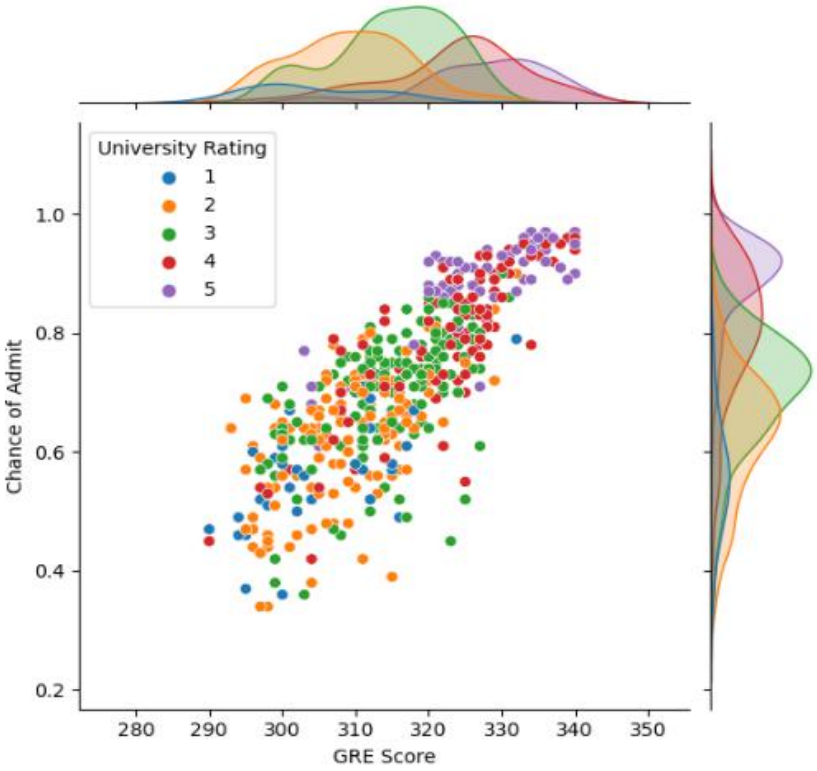

Figure 10

From Figure 10, we saw that the applicants with higher scores tend to come from higher-rated schools and have a higher chance of admission.

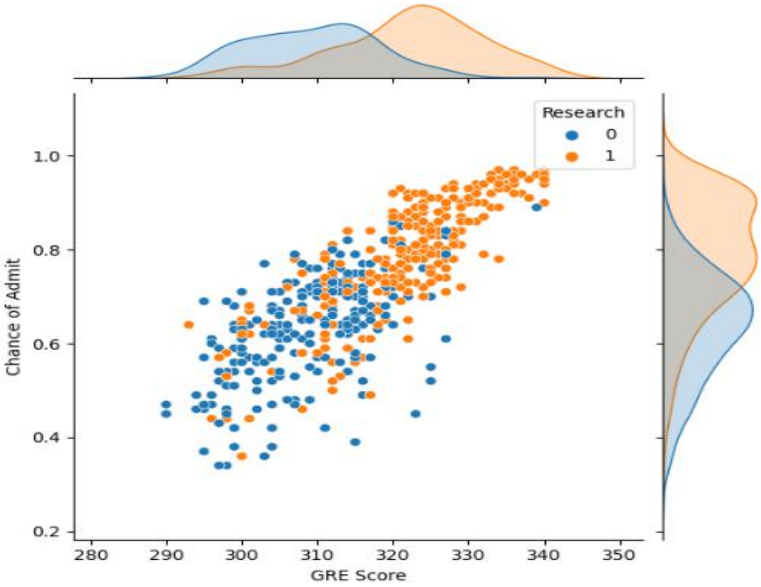

Figure 11

Figure 11 showed that the applicants with higher GRE scores also tend to have research experience and a higher chance of admission.

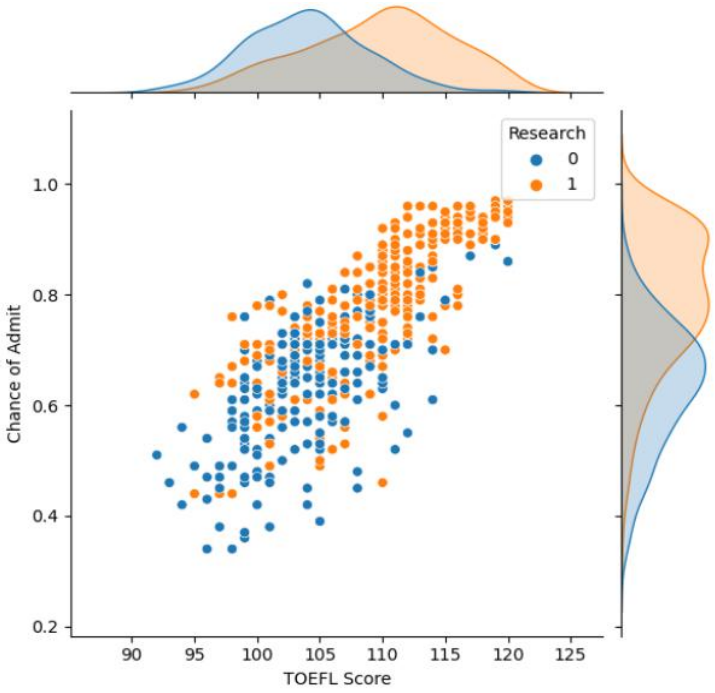

Figure 12

Plotting the relationship between 'TOEFL Score', 'Research' and 'Chance of admit', we again saw that the applicants with higher TOEFL scores also tend to also have research experience and a higher chance of admission.

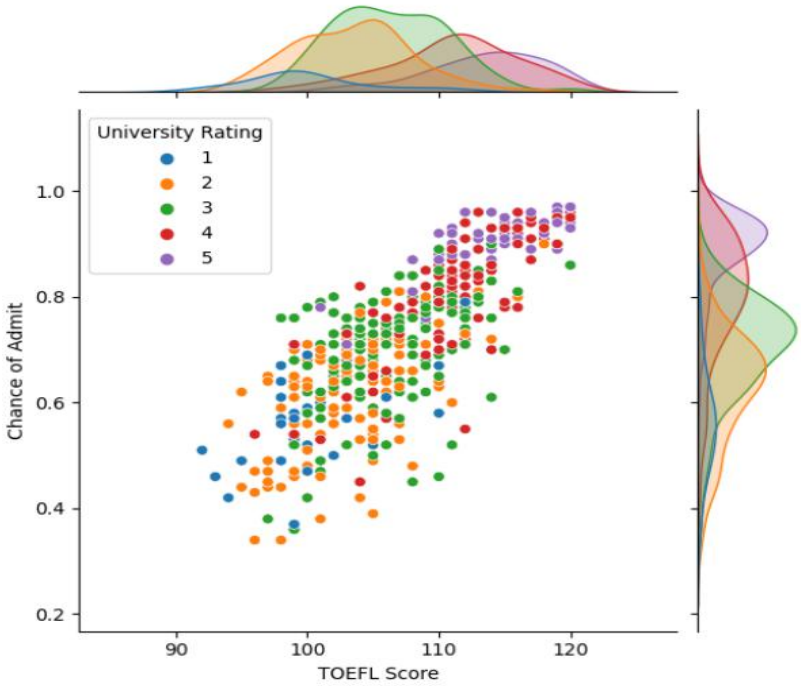

Figure 11

Plotting the relationship between 'TOEFL Score', 'University Rating', and 'Chance of admit', we saw the applicants with higher TOEFL scores tend to come from higher-rated schools and have a higher chance of admission.

Box plots

Boxplots showed how the values in the data are spread out by depicting groups of numerical data through their quartiles.[11]

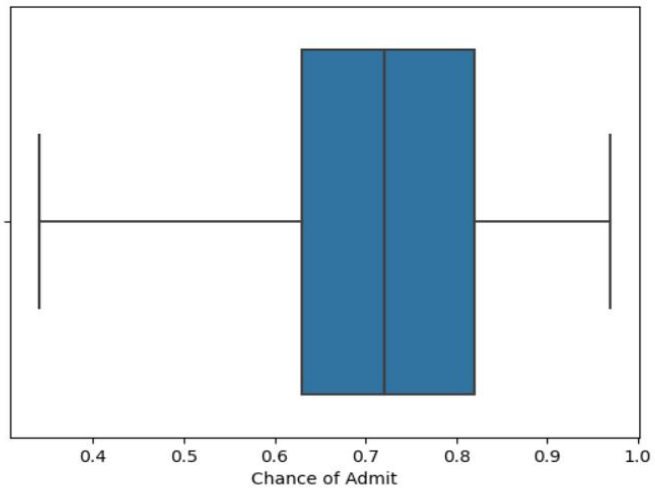

Figure 12

Figure 12 showed how 'Chance of admit' was spread out among the applicants in our dataset.

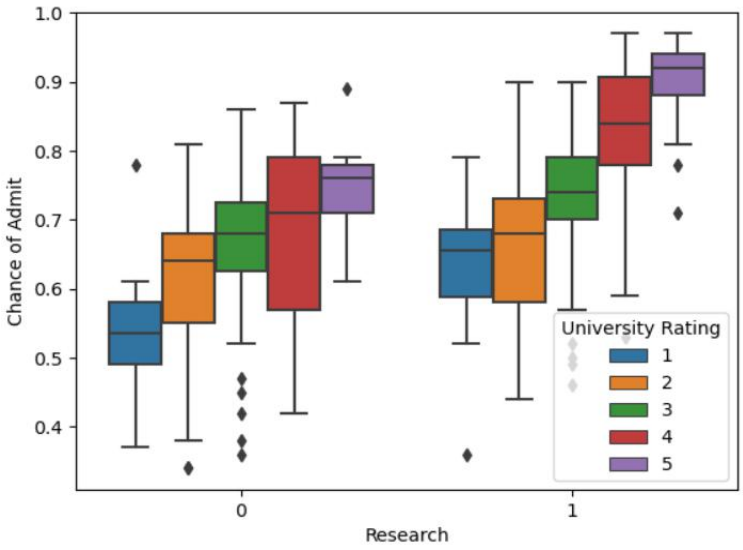

Figure 13

We also used boxplots to explore the relationship between 'Research', 'University Rating', and 'Chance of admit'. Figure 13 showed that the applicants with research experience tend to have a higher chance of admission, with their university rating taken into account.

[11] Galarnyk, M. (2020, July 6). *Understanding Boxplots*. Medium. Retrieved December 1, 2021, from https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51.

Implementation of Data Mining Techniques in Python

We implemented Decision Tree, Random Forest, and Linear regression in our data analysis using Python.

A cross-validation procedure was also implemented to check for the best accuracy score among several regression models.

An illustration of cross-validation is shown below.

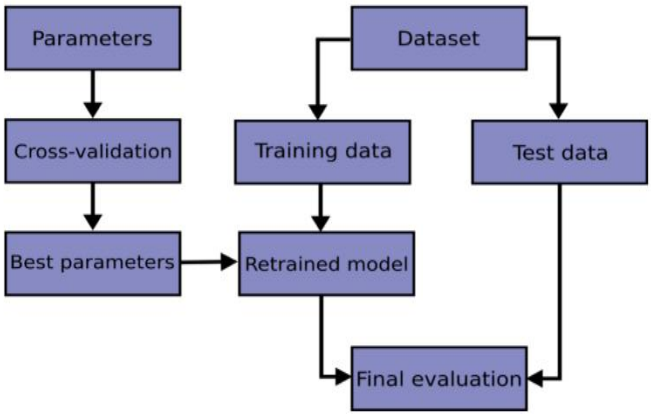

Figure 14

The results from cross-validation (Figure 15) showed that linear regression and Random Forest are the two models with higher accuracy scores.

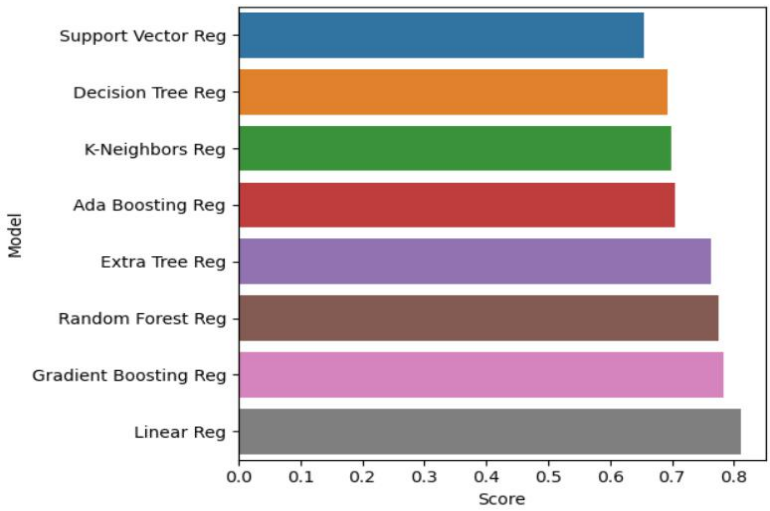

Figure 15

Functions

Our MAIN code started with the following functions:

· Data_admit() function downloaded the dataset and preprocessed it to prepare it for other functions and classes to call.

· The main() function created the pyqt5 application.

· The App() class created the application's features, such as Menu buttons with methods that calls CorrelationPlot(), box(), DecisionTree(), Regression(), cross(), RandomForest() classes for data visualization, machine learning models, cross validation.

Results

Linear Regression

Our linear regression analysis started with a density plot showing how well the model is predicting 'Chance of admit'. As shown in Figure 16, the predicted 'Chance of admit' was very close to the actual 'Chance of admit', telling us that the model was satisfactory. In this model, we tried scaling and not scaling the features, and the performance scores were the same.

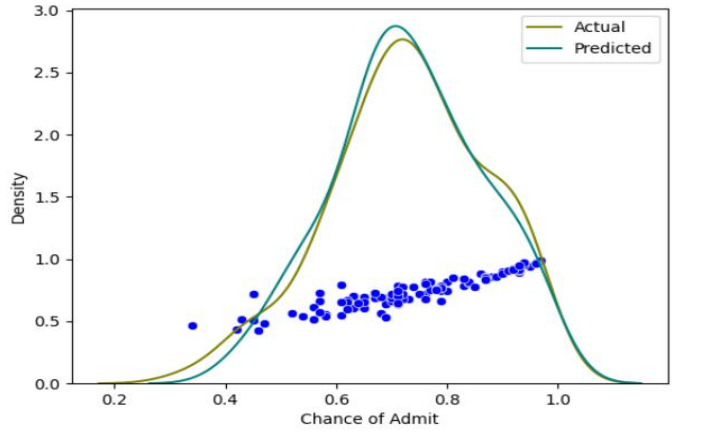

Figure 16

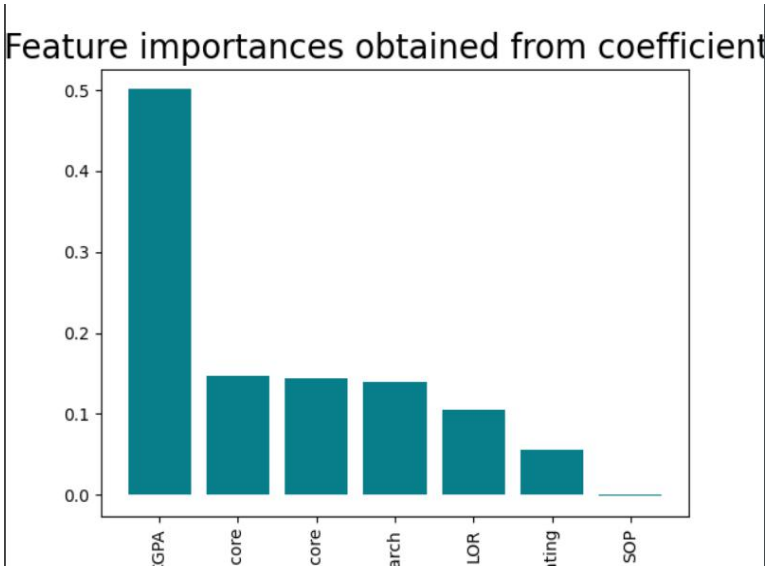

Figure 17

As illustrated by Figure 17, the linear regression analysis showed that college GPA was the most important feature in predicting 'Chance of admission'. In addition, 'GRE score', 'TOEFL score', and 'Research' weighed almost equally. We also saw that 'LOR' and 'University

Rating' weighed less. Finally, we saw that 'SOP' weighed the least. We figured that the results were mostly in line with the conventional theories of predicting graduate school admission.

Decision Tree

A Decision Tree Regressor Analysis is a graphic representation of various alternative solutions available to solve a problem. With a decision tree, the alternative solutions and possible choices are illustrated graphically, making it easier to make well-informed choices.[12]

The goal of our Decision tree algorithm was to predict the dependent variable ('chance of admission') from our independent variables ('GRE score', 'TOEFL score', 'University rating', 'SOP', 'LOR', 'CGPA', and 'Research').

The approach was to build a tree structure through a series of binary splits (true/false) from the root node by having branches passing several decision nodes (internal nodes), until reaching the leaf nodes.

Attribute Selection

1. We selected all the independent variables for the purpose of analysis. Note that users of GUI would be able to select desired independent variables in building the algorithm.

2. The root node was selected by recursively splitting the training samples, using the features with the highest Mean Squared Error (MSE).

3. The nodes were iterated until reaching a maximum depth of 5 on their way to the leaf node, helping us reaching a reasonable accuracy. The maximum depth of the tree was also specified to prevent the tree from becoming too deep - a scenario that would lead to overfitting.[13]

---

[12] Soner Yildirim (2020 February 11). Decision Trees and Random Forests — Explained Detailed explanation with theory and examples with code. Retrieved December 1, 2021, from https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd

[13] Victoria Rios (2021 February 23. Regression Model Selection (step-by-step) on Covid-19 Vaccination Progress Worldwide dataset (python) Retrieved December 1, 2021, from https://medium.com/data-science-ai-and-machine-learning-for-dummies/regression-model-selection-step-by-step-on-covid-19-vaccination-progress-dataset-python-22fe8e6f37ca.

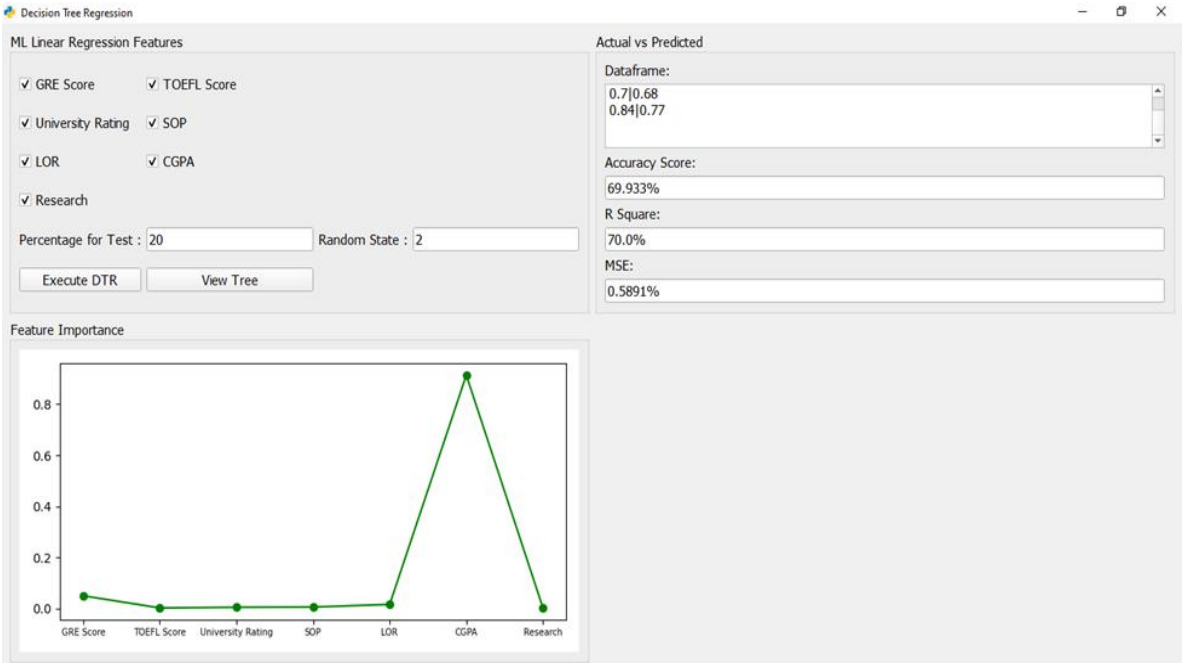

Figure 18

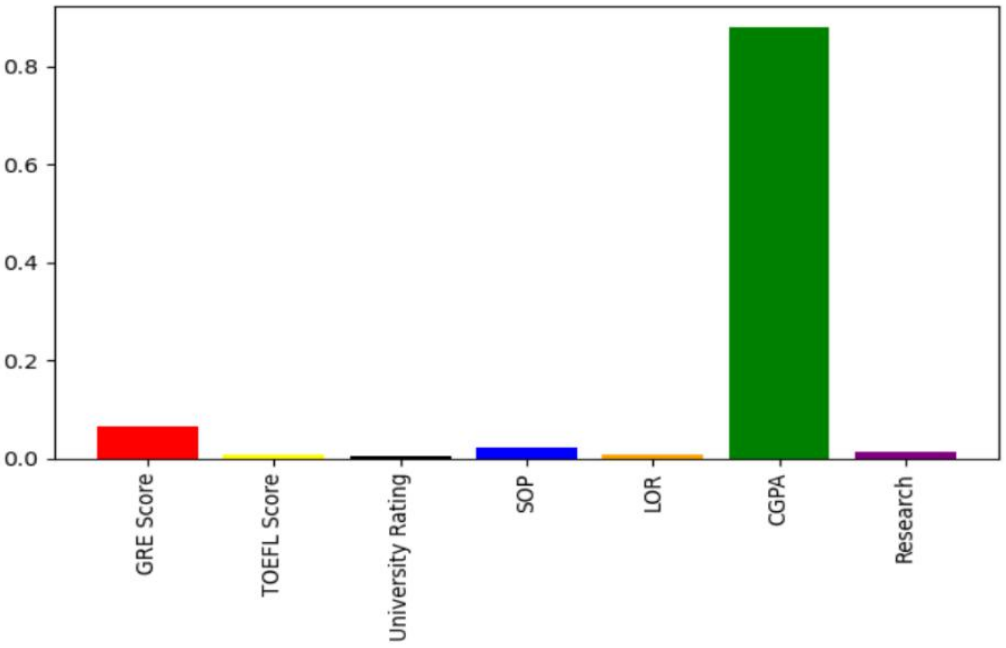

Figure 19

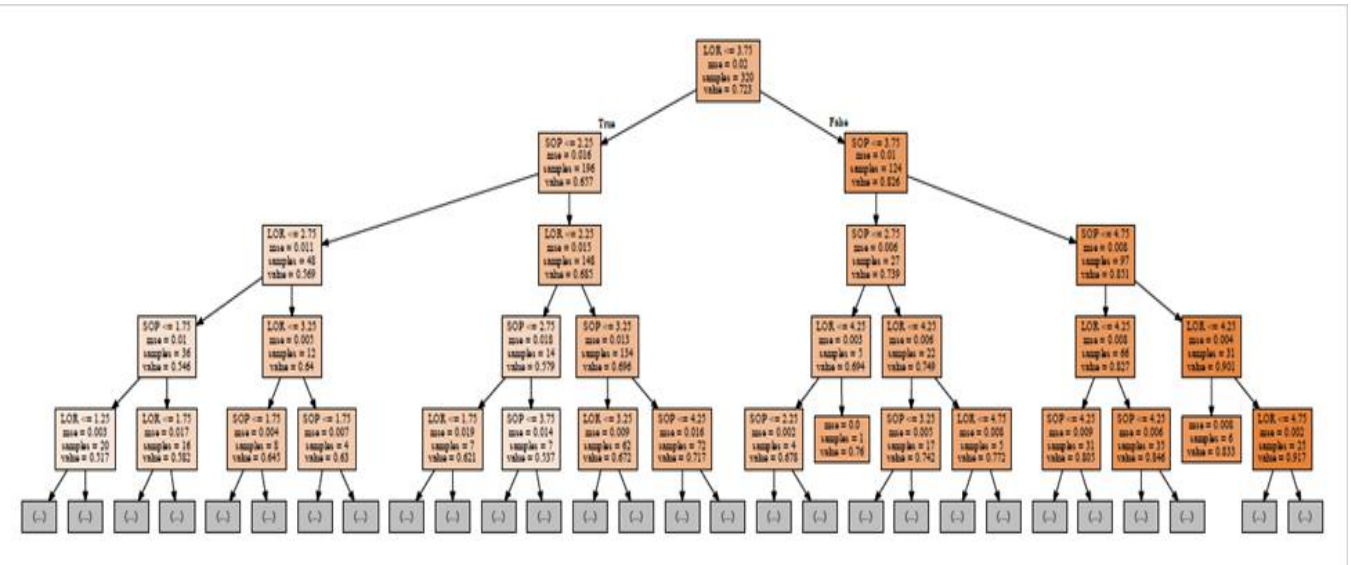

Figure 20

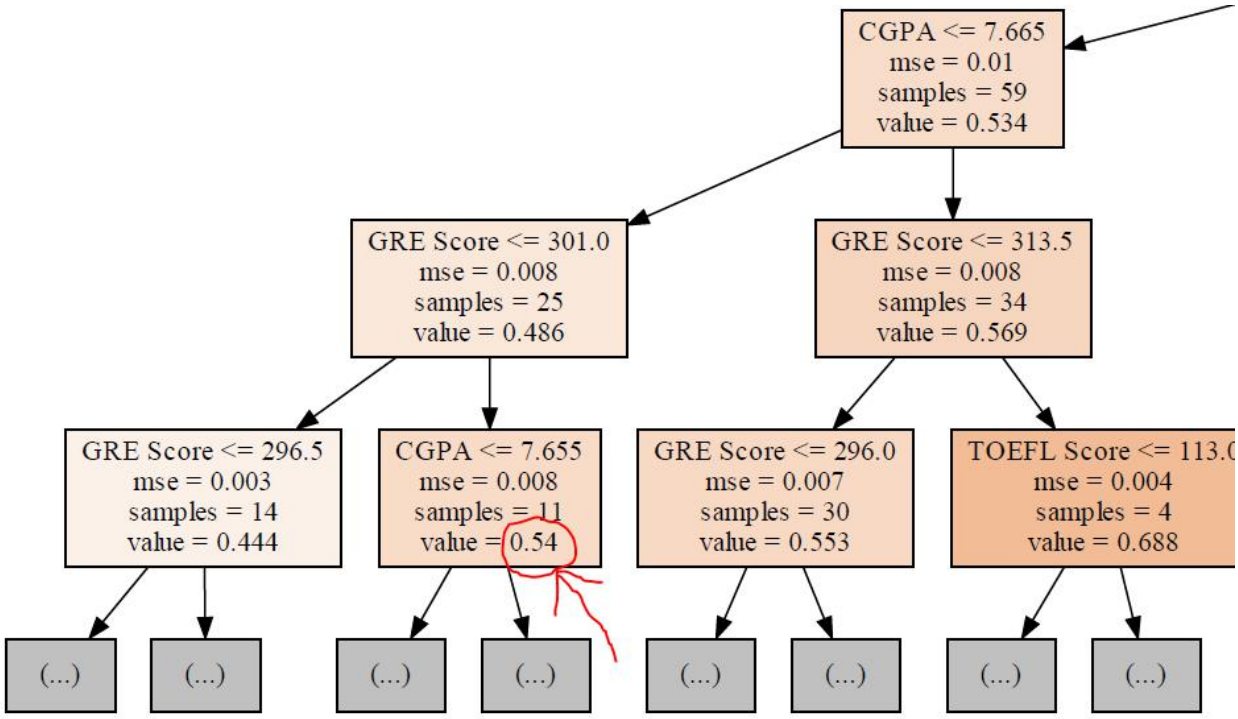

Figure 21

CGPA<=7.665 and GRE Score<=301 , predicted values=0.54

As we can see, the Decision Tree analysis showed that college GPA and GRE score were the two features with higher importance. However, this was not totally in line with our EDA and the feature importance results yielded by other models. Moreover, the regressor score of the Decision Tree model was around 69%. Considering also how the model is predicting the data, as shown by Figure 22, we decided to discard the Decision Tree model.

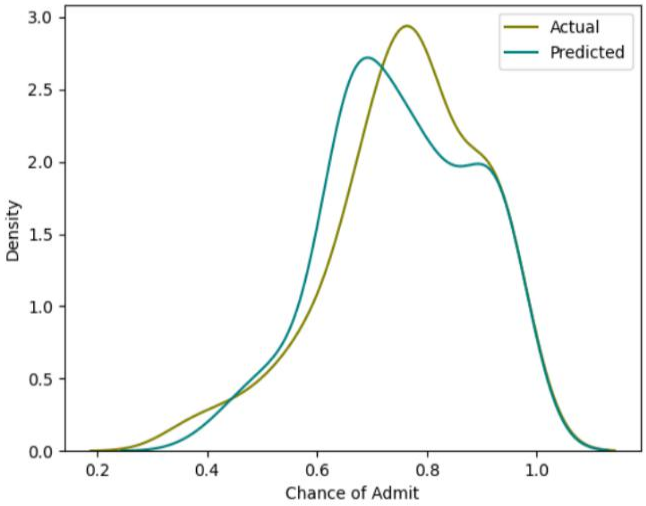

Figure 22

Random Forest

Random forest regression is great for obtaining non-linear relationships between input features and the target variable.

Our random forest started at the very top with one node. This node was then split into decision nodes. These nodes were then split into their respective right and left nodes.

At the end of the leaf node, the average of the observations that occurred within that area was computed. The nodes at the bottom are referred to as leaves or terminal nodes.

The value in the leaves was usually the mean of the observations occurring within that specific region. Looking at the leftmost leaf node below in Figure 24, the predicted value (0.444), was the average of the 14 samples. Generally, Random Forest produces better results and works well on large datasets. Its bootstrapping aggregation makes it a better model over decision trees as its predictions are based on the averages of a decision tree.

Figure 23 shows the feature importance. As we can see, college GPA, GRE score, and TOEFL score were the three comparatively important features, followed by "University rating", "SOP", "LOR", and "Research".

The Random Forest model yielded a regressor score of a minimum of 76% and a maximum of 79%, meaning that it was a better model.

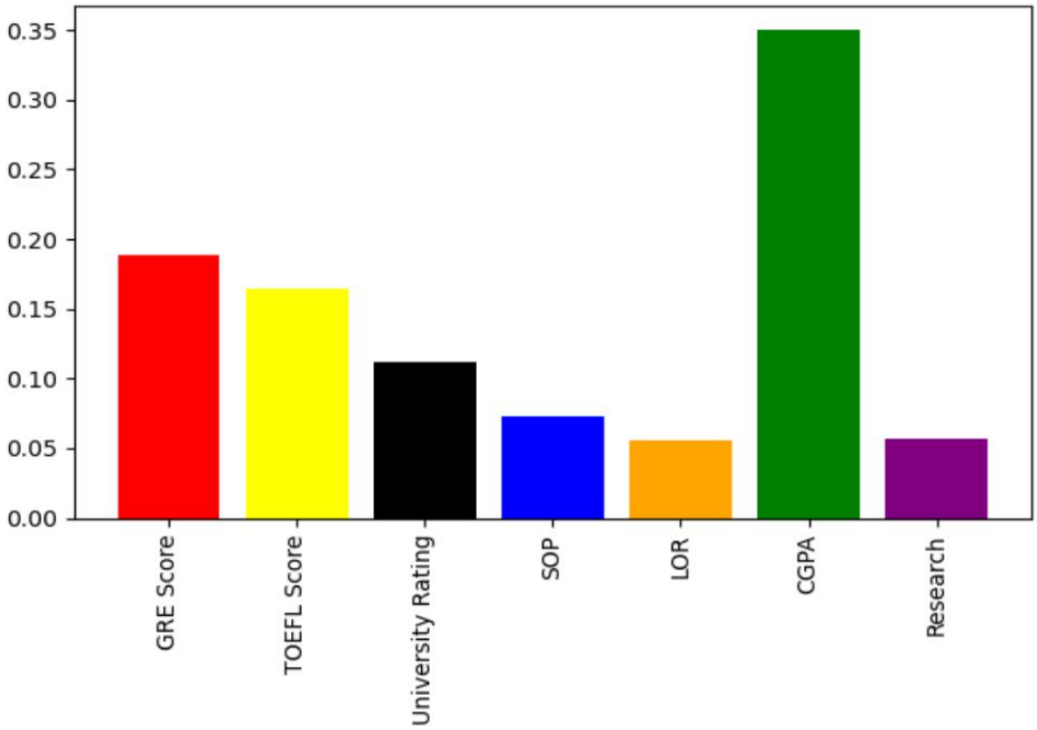

Figure 23

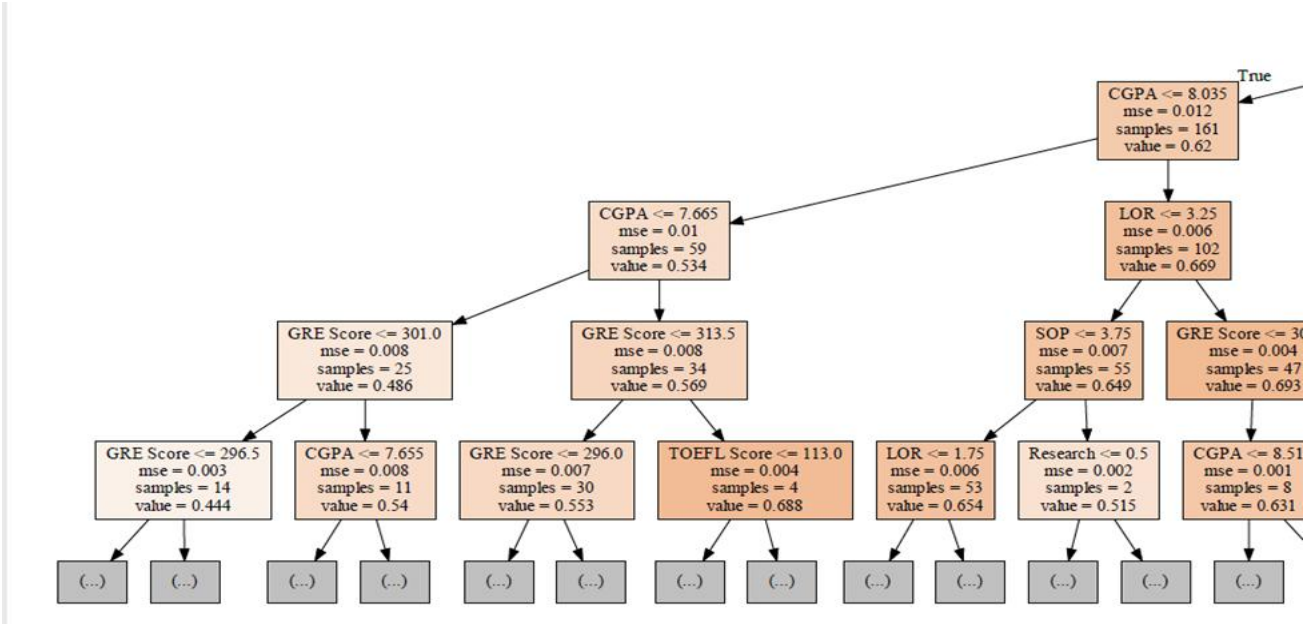

Figure 24

Model Evaluation

Models	MSE	R-SQUARED
LINEAR REGRESSION	0.0037	0.784
RANDOM FOREST REGRESSION	0.0042	0.783
DECISION TREE REGRESSION	0.0059	0.699

Figure 25

Figure 25 showed that the Linear Regression model has the lowest mean square error (MSE) with a value of 0.0037 and with a R-squared value of 0.784, meaning that the model describes 78.4 % of the dataset.

Summary

In sum, we concluded that college GPA was the most important feature in predicting admission to the master's programs in the U.S. Standardized test scores (i.e., GRE score and TOEFL score) could be considered as the feature with secondary importance. We also concluded that letters of recommendation, university rating, personal statements, and research experience weighted less, suggesting that non-academic criteria were less valued in the admission process.

From this project, we learned that it was important to have a variety of models in conducting data analysis to select the best model and, therefore, acquire the most accurate results. We also learned that it was always useful to visualize the data prior to building the models to get a better understanding of the dataset and have a benchmark for the analysis followed.

To make our results even more accurate, we would love to include more predictor variables and have a larger sample size.

References

By: IBM Cloud Education. (n.d.). *What is overfitting?* IBM. Retrieved December 1, 2021, from https://www.ibm.com/cloud/learn/overfitting#:~:text=Overfitting%20is%20a%20concept%20in,unseen%20data%2C%20defeating%20its%20purpose.

Galarnyk, M. (2020, July 6). *Understanding Boxplots*. Medium. Retrieved December 1, 2021, from https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51.

Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019.

Python Decision Tree Classification with Scikit-Learn Decisiontreeclassifier." *DataCamp Community*, https://www.datacamp.com/community/tutorials/decision-tree-classification-python.

Random Forest – Random Forest - EasyAI. https://easyai.tech/en/ai-definition/random-forest/.

Soner Yildirim (2020 February 11). Decision Trees and Random Forests — Explained Detailed explanation with theory and examples with code. Retrieved December 1, 2021, from https://towardsdatascience.com/decision-tree-and-random-forest-explained-8d20ddabc9dd

Team, P. (n.d.). *How to plot a jointplot with 'hue' parameter in Seaborn - Pretag*. Pretag development team. Retrieved December 1, 2021, from https://pretagteam.com/question/how-to-plot-a-jointplot-with-hue-parameter-in-seaborn.

Tin, Martin. "Intro to Linear Regression - Machine Learning 101." *Medium*, DataDrivenInvestor, 6 Oct. 2020, https://medium.datadriveninvestor.com/machine-learning-101-part-1-24835333d38a.

Victoria Rios (2021 February 23. Regression Model Selection (step-by-step) on Covid-19 Vaccination Progress Worldwide dataset (python) Retrieved December 1, 2021, from https://medium.com/data-science-ai-and-machine-learning-for-dummies/regression-model-selection-step-by-step-on-covid-19-vaccination-progress-dataset-python-22fe8e6f37ca

Zach. (2021, March 12). *Interquartile Range vs. standard deviation: What's the difference?* Statology. Retrieved December 6, 2021, from https://www.statology.org/interquartile-range-vs-standard-deviation/.

Appendix

We imported our dataset from kaggle api, as shown by the codes below.

```
os.environ['KAGGLE_USERNAME'] = 'koyanjo'
os.environ['KAGGLE_KEY'] = '33bfba07e0815efc297a1a4488dbe6a3'

from kaggle.api.kaggle_api_extended import KaggleApi

dataset = 'mohansacharya/graduate-admissions'
path = 'datasets/graduate-admissions'

api = KaggleApi()
api.authenticate()

api.dataset_download_files(dataset, path)

api.dataset_download_file(dataset, 'Admission_Predict.csv', path)
```