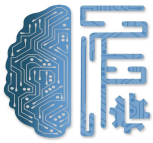
Machine Learning Intern
December 2022



College Admission Prediction Using Machine Learning

Sachin Pokharel

Contents

Chapter 1

Project Report

.

1.1 Problem Statement

Many students apply to colleges and universities every year, but only a portion is accepted. The college admission process is very competitive as there are many more qualified applicants than spots available at top colleges and universities. The admission process can be stressful and uncertain for students and their families. Many students are unaware of the standards and procedures for college admissions, so they seek consultants and spend a significant amount of money on guidance and the admission process. Furthermore, it is very stressful for colleges and universities to review and evaluate all the applications as it requires more time and resources.

1.2 Market/Customer/Business Needs Assessment

As mentioned in the problem statement, there is a market need for this AI system which targets students and educational institutions. The application can provide valuable information and proper guidance to the students planning for their higher education and help them understand their chances of being accepted to different colleges based on their academic qualifications and other factors.

Education institutions, colleges and universities can make informed decisions about the eligibility of the applicants, whether to take as a student or reject them. With the help of this application, the institutions can better predict the applicants who are most likely to succeed academically and graduate. Moreover, the admission process will be less stressful as most of it will be automated, which saves a lot of time and resources and optimize their plan for future growth.

1.3 Target Specifications and Characterization

This application focuses on students, education institutions, education consultants and other companies that deal with education. The application can also be optimized by adding new features so that more potential groups can be interested in the product.

- **Students:** Students who are planning for their further education and who require seeking guidance about the application process can take leverage of this application. It helps to evaluate themselves, understand their chances of being accepted to colleges, and make an informed decision about colleges to apply to.
- **Education Institutions:** Education Institutions can be benefited from using this system as it helps them to predict the likelihood that an applicant will be accepted to their college. This allows them to make more informed decisions about which applicants to accept and can also help them manage their enrollment targets.
- **Education Consultants:** Education Consultants can use this system to help their clients make informed decisions about where to apply to college, which can help the consultants provide more accurate and personalized guidance.

1.4 External Search

These research journals and papers provided valuable insights and ideas that were relevant and helpful for developing the College Admission Prediction prototype. The articles are mentioned below:

- Centralized admissions for engineering colleges in India [1].
- Adaptive admissions process for effective and fair graduate admission [2].
- A machine learning approach for graduate admission prediction [3].
- Student Academic Performance Prediction using Supervised Learning Techniques [4].
- Predicting graduate student success: A comparison of neural networks and traditional techniques [5].

1.5 Bench Marking Alternate Products

There are several alternatives to a college admission prediction system that consumers may consider when looking for products to help navigate the college admission process. Some of the alternatives are as follows:

- **College Search Portals:** Several web portals allow students to search for colleges based on their preferences and interests. Most of these portals provide information about admission requirements, acceptance rates and other relevant details.
- **Admission Consultants:** Students can seek a private admission consultant to help with the college admission process, help with personal essays and provide guidance and support throughout the application process.

- **College Admission Fairs:** Many colleges and universities conduct college admissions fairs every year, where students can meet with the admission committee from different institutions. Attending these fairs is an excellent opportunity for students to learn more about colleges and seek guidance about admission.

However, these alternatives have their weaknesses. For example, college search websites can be a valuable resource for students looking for information on colleges and universities, but there is no personal/customized assessment of a student's chance of being accepted. Similarly, many students cannot afford to seek an admission consultant as it is expensive and takes a lot of resources and time.
College Admission Prediction System has many advantages over these alternatives. The application is based on data and statistical models, which can better assess applicants' chances of being accepted to a college. Similarly, by using this application, consumers can save a lot of their time as it is a quick and easy way to know the student's chances for admission. As the system is based on the data, a wider range of data sources can provide a more comprehensive assessment of students' profiles than other alternatives.

1.6 Applicable Patents

There are some patents which relate to the concept of prediction of an applicant's acceptance into universities and colleges. These patents are listed below:

- System and Method for Probabilistic Prediction of an Applicant's Acceptance (U.S. Provisional Patent Application No. 61/792,342).
- Method and System for College Matching (U.S. Provisional Application No. 61/908,098).
- College admission optimizer for an individualized education consulting system and method (U.S. application Ser. No. 13,931,232).

1.7 Applicable Regulations

A number of laws and regulations need to be considered while developing this application. These regulations can be applied based on the context and use of this application. Some of the rules that can be applied are mentioned below:

- A privacy law could be applied, which involves collecting, storing or using personal data about the students. So, there may be a requirement for permission to collect and use personal data and the security and confidentiality of this data.
- A number of education laws should be considered which apply to the colleges and universities during the admission process.
- A number of anti-discrimination laws should be considered as the application should be free from any gender bias and discrimination on the basis of certain characteristics, race or religion.

- Environmental regulations should also be considered if the development and scalability of the application require big hardware, which can have environmental impacts.

1.8 Applicable Constraints

There are several constraints that can restrict the development and use of this AI system. The applicable constraints are discussed below:

- **Space:** We may need to consider the space requirements for big data servers for storing and processing data depending on the size and complexity of the application. Moreover, if any physical infrastructure is used to support the system, that needs to be handled.
- **Budget:** The budget requirement also needs to be considered, as the development and maintenance of this application can be expensive if we need to purchase or rent data from various sources or hire staff to develop and maintain the application. Moreover, it is essential to consider the budget and allocate resources properly if we need to do marketing.
- **Expertise:** We may need to hire staffs with expertise in machine learning/deep learning, data analysis and software development depending on the complexity of the application. There could be potential groups of people that needs to be hired who expertise in relevant fields such as education/admission counselors.
- **Data Availability:** Data has an important part in the training and evaluating the system. The system may be limited by the availability and quality of data that is used for training and testing. The prediction may be inaccurate if the data is biased or incomplete.
- **Consumer Acceptance:** The approval of this system may be limited for further use if the system does not meets the expectations of the consumers on the quality, reliability and adaptability.

1.9 Business Model

There are several business opportunities with this application which mainly targets students, higher institutions and consultants. Some of the business ideas that could be applicable are discussed below:

- **Subscription:** This application could offer monthly or yearly subscription-based access to potential customers with pricing tiers for different access levels. Example - Tier 1, Tier 2 and Tier 3.
- **Pay-per-use:** It can also have a pay-per-use feature which allows the customers to pay a fee each time they use this application.
- **Advertising:** This application could also be monetized by allowing universities and colleges to advertise their business in exchange for a fee.

- **Tutoring:** This application can also offer coaching or tutoring services to help the student prepare for college admission exams to improve their chances of getting accepted to their desired universities based on the evaluation.

1.10 Concept Generation

The advancement of AI has been increasing significantly in these recent years in the education domain. However, there is still a need for automated systems in this domain to make more informed decisions. There are still lots of issues applying to college as a student because it is a complex and stressful process which needs a lot of resources and time. The students may feel overwhelmed by the many factors that can impact their chances of getting accepted, such as grades, test scores, recommendations and personal letters. Some are unaware of the admission process or may not know how to proceed with the application process effectively. There are also additional challenges, such as risk bias and discrimination against the students.

With universities and colleges, every year, they receive a massive number of applications and evaluating each application effectively takes a lot of resources and time. There are cases where an applicant who is a good fit for the institution but still discarded as the process is handled manually. So, the institutions need to identify and target students who are more likely to be a good fit which this application provides them. So, to tackle all these significant problems, the concept of this AI application was developed.

1.11 Concept Development

The development of a College Admission Prediction System involves several steps, which are discussed below:

- Identifying the feature attributes and variables important for predicting college admissions.
- Collecting the data on identified variables for many students who have applied to institutions.
- Performing exploratory data analysis on the data to understand patterns and trends. The study can be done using visualizations.
- Cleaning and preprocessing the data includes handling missing values, correcting errors, scaling and standardizing so that the data is in a consistent format.
- Selecting different machine learning models to use for prediction. There are various models, such as Random Forest, KNN, SVM or even Neural Networks.
- Training the model on the data and adjusting the hyper-parameters to minimize the prediction error.
- Performing validation on the model by testing the model on unseen data.

- Fine-tuning of the models if there is still room for improvement.
- Deploying the model into production and using it for predicting college applicants.

1.12 Final Product Prototype

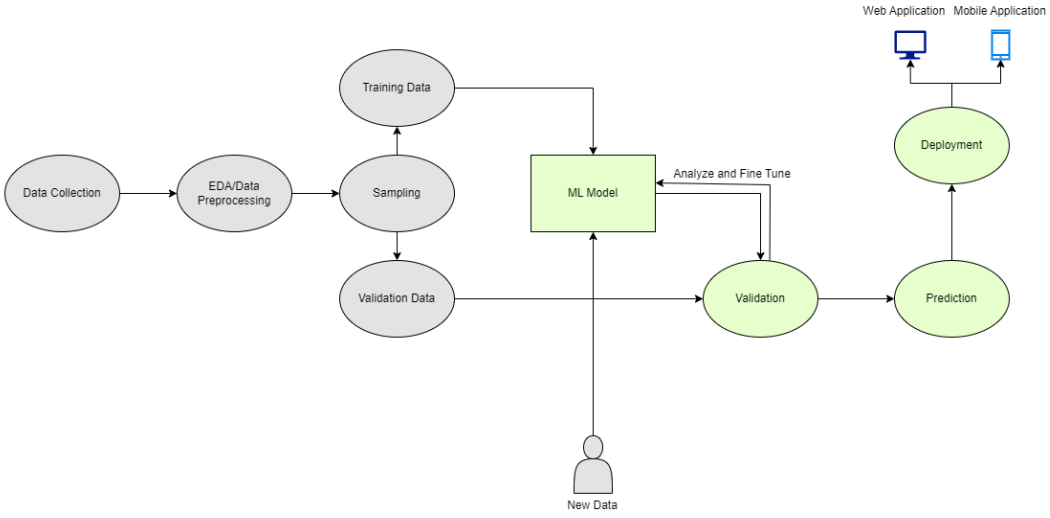


Figure 1.1: Schema for College Admission Prediction System.

After the application is developed and deployed, students will need to provide the system with information about their academic backgrounds, test scores, grades, and extracurricular activities and achievements. These are all crucial variables which have a lot of impact on the admission process. The students will need to enter this information in the application as user input. The platform can be web or mobile based, with a front-end layout for entering inputs. The application will then use the information/data provided by the student and generates the output, which is the predicted probability of admission chance.

Chapter 2

Product details

2.1 Data Sources

There are different data sources that can be used to develop this application. The databases from college/university admission offices, which have detailed records of the students who apply for admissions, can be one of the sources. The records contain information about their academic record, test scores, extracurricular activities, recommendations and personal essays. Moreover, students are also a data source as they directly provide the required information on the application, which will be in the database as training data. Furthermore, the data also can be obtained from various other sources, such as educational databases and academic organizations, which helps the application to be more comprehensive and accurate.

2.2 Project Requirements

Many algorithms, frameworks and software tools can be used to develop this application. The tools and techniques that will be used depending on the business goals, the size of the project, the available resources and the team's expertise.

Some of the algorithms, frameworks and techniques that will be used in developing this application are as follows:

- Statistical techniques like regression analysis and correlation can be used to identify relationships between different variables in the data to improve prediction models based on the relationships using Pandas and Numpy.
- Data visualization tools such as Matplotlib, Seaborn, Tableau, and Power BI can be used to explore and understand the data and to create visualizations for generating meaningful insights and patterns of the data.
- Machine Learning Algorithms can be used to build prediction models based on these patterns, which include Random Forests, SVM, KNN, Gradient Boosting and neural networks.
- Flask, Django or other front-end development frameworks can be used to create web applications for this project.
- For deployment, AWS or Google Cloud Platform can be used.

The team members developing this application depend on the specific skills and resources available to the team. These are some of the roles that will be needed for development:

- **Data Scientist:** Responsible for data analysis, building and training the model and also responsible for the prediction model.
- **Software Engineer:** The software engineer will be responsible for building the software system that will be used to deploy the prediction model. They will develop the client user interface and integrate the prediction model into the system.
- **Project Manager:** Responsible for organizing and coordinating the team's work and communicating with the clients.
- **Domain Expert:** The team may also need a domain expert who has expertise in the college admission process. This person could help the team understand the admission process and can ensure whether the developed system reflects reality accurately or not.

2.3 Budget Allocation

Roles	**Budget (Per Annum)**
Data Scientist	8 LPA
Software Engineer	5.5 LPA
Project Manager	7 LPA
Domain Expert	5 LPA
Total	**25.5 LPA**

Table 2.1: Budget Allocation for College Admission Prediction

Chapter 3

Code Implementation

This chapter implements basic visualization, exploratory data analysis, ML modelling and validation. A sample dataset relevant to this application was found on the Kaggle website. The dataset is created by **Mohan S Acharya** to predict the applicants' chances of getting graduate admission. Although the dataset does not have many records, this dataset somewhat reflects the actual real-world data for college admission prediction.

3.1 Dataset

The dataset contains seven variables that need to be considered important during the college admission process. The variables are as follows:

1. GRE Score (out of 340)
2. TOEFL Scores (out of 120)
3. University Rating (out of 5)
4. Statement of Purpose and Letter of Recommendation Strength (out of 5)
5. Undergraduate GPA (out of 10)
6. Research Experience (either 0 or 1)
7. Chance of Admit (ranging from 0 to 1)

	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
0	337	118	4	4.5	4.5	9.65	1	0.92
1	324	107	4	4.0	4.5	8.87	1	0.76
2	316	104	3	3.0	3.5	8.00	1	0.72
3	322	110	3	3.5	2.5	8.67	1	0.80
4	314	103	2	2.0	3.0	8.21	0	0.65

Figure 3.1: First five records of the dataset

3.2 EDA for College Admission Prediction

3.2.1 Statistical Analysis

	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000

Figure 3.2: Statistical summary of the dataset

Observations:

- The average score of admission to the graduate college was 0.72, which lies in the 50th percentile.
- Applicants' chances of admission range from 34% to 97%.
- There are no null/missing values in the dataset as each of the features have 500 records.

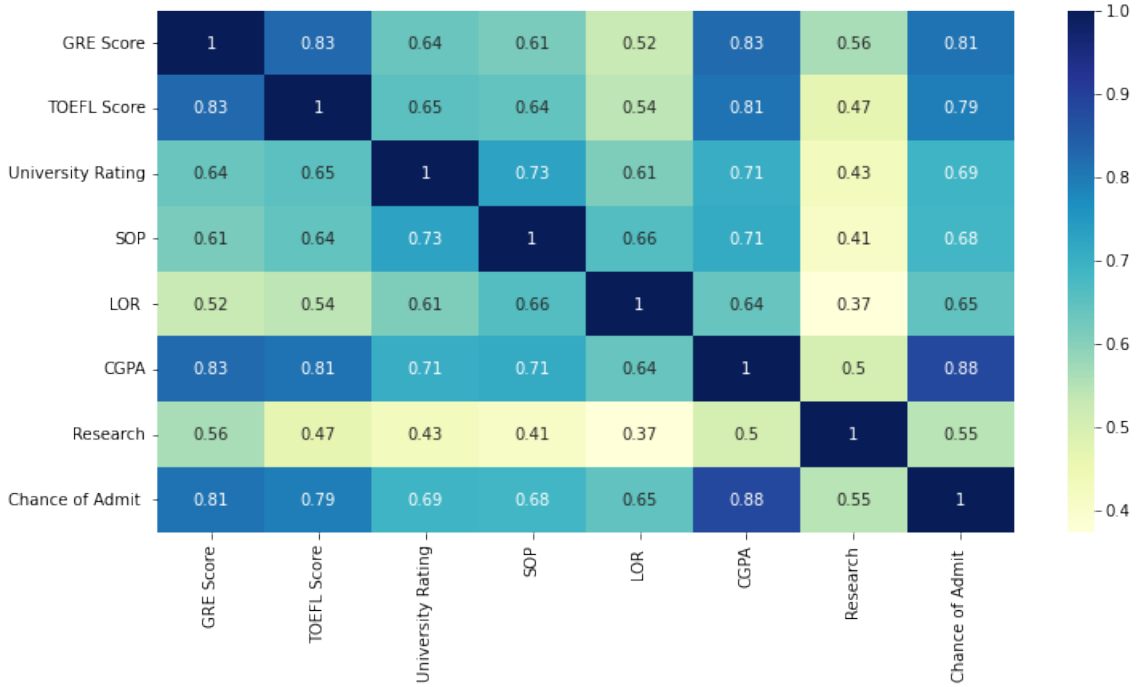


Figure 3.3: Checking the correlations between the features using seaborn heatmap

Observations:

- CGPA and GRE Score has the highest positive correlation with the chance of admit, which means the applicants with higher CGPA and GRE score were most likely to be accepted to the universities.
- Those who have a higher CGPA have a high GRE Score.

3.2.2 Bar Plots (Categorial Features)

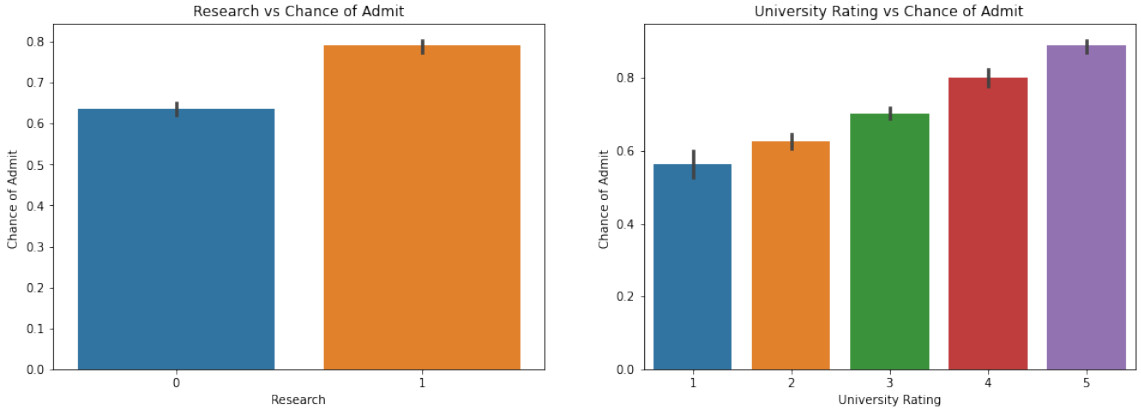


Figure 3.4: Checking the correlations between the features using seaborn heatmap

3.2.3 Distribution Plots (Numerical Features)

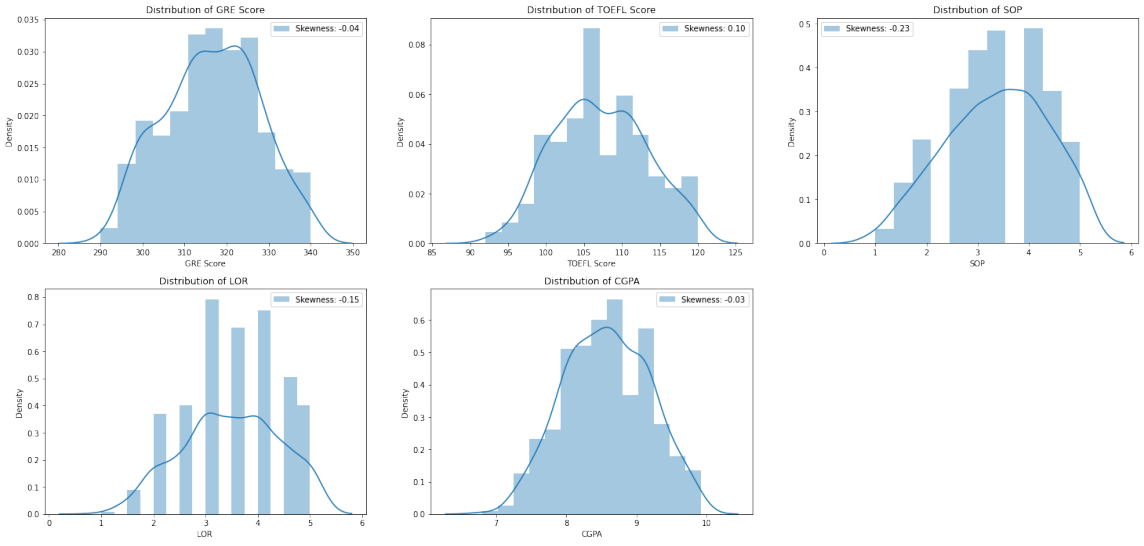


Figure 3.5: Distribution plots of different features

Observations (Categorical Features):

- The applicants who have published research papers before having a higher chance of being accepted. However, the applicants without research were also accepted with an acceptance rate of approx 62%.
- Applicants from top-rated universities had a higher chance of acceptance.

Observations (Numerical Features):

- From the distribution plots we can see that, GRE Score, SOP, CGPA, and LOR are negatively skewed, which means the distribution is slightly shifted to the left, and there are slightly more values on the right side of the distribution. It is also worth noting that the degree of skewness is relatively small so it is not that significantly skewed.
- The academic performance and test scores of the applicants indicate that the applicant's profiles had variations.

3.2.4 Regression Plots

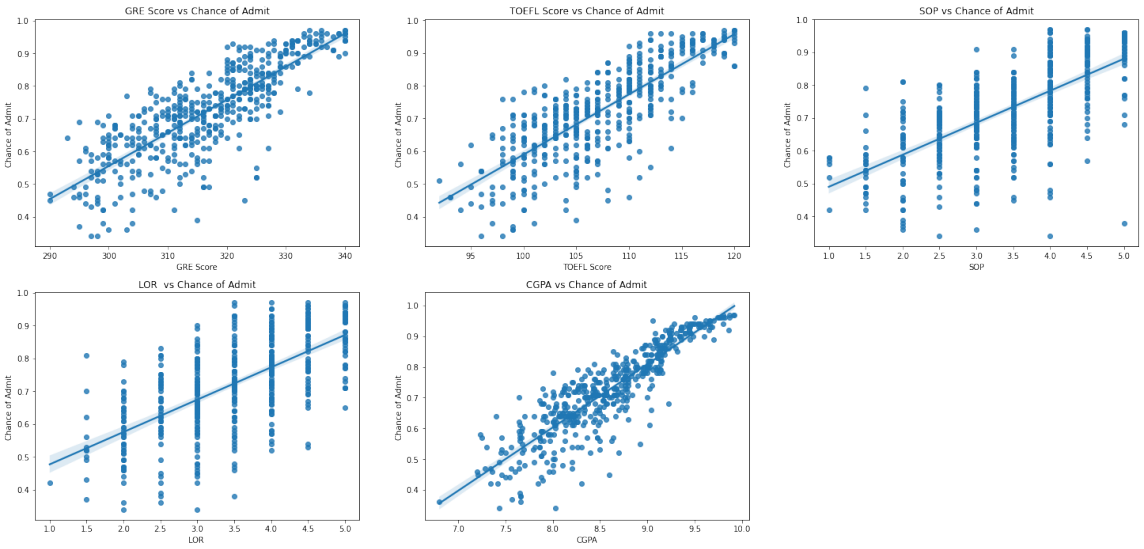


Figure 3.6: Regression plot between chance of admit and other features

Observations:

- Applicants with higher CGPA, GRE, and TOEFL scores were most likely to be accepted.
- LOR and SOP have some impact on the admission chances but there were applicants having low LOR and SOP ratings but still got admitted.

3.3 Machine Learning Modeling

Four machine learning models were trained on this dataset. The models were Linear Regression, Decision Tree Regressor, Random Forest Regressor and Gradient Boosting Regressor. 6-Fold Cross-validation was performed on the data, and the models were evaluated using mean squared error (MSE) metrics. The trained models can improve by fine-tuning the selection of hyper-parameters and considering other feature engineering techniques. To get a general idea, simple models have been used.

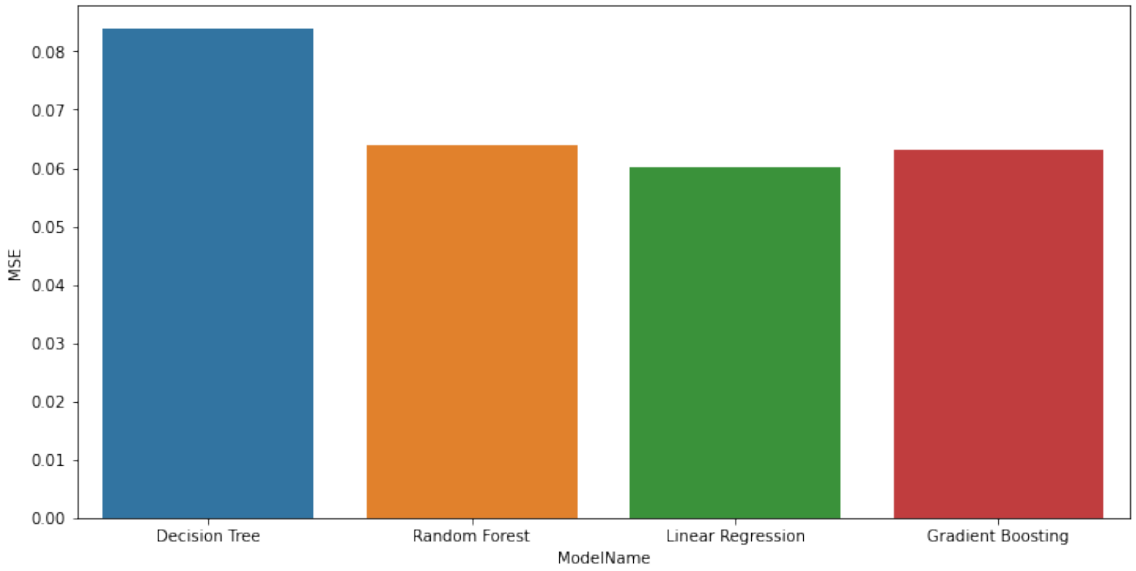


Figure 3.7: Mean Squared Error obtained from different ML Models

3.4 Validation on Test Data with Linear Regression

From the above graph, Linear Regression had the lowest mean squared error out of all four algorithms. So, validation of test data was performed on Linear Regression. R2 score metric was used to perform validation on the test data, which resulted that 81% of the variance in the dependent variable can be predicted from the independent variables.

Source Code: `https://github.com/p-sachin/college_admission_prediction`

Chapter 4

Conclusion

A college admission prediction application is a helpful application for students, education institutions and education consultants. It can help students make an informed decision on where to apply and also allow educators to understand better and support their student's academic journey. While no prediction system can guarantee admission to a particular college or university, it can provide valuable guidance and support to those navigating the complex world of higher education.

Bibliography

[1] S. Baswana, P. P. Chakrabarti, S. Chandran, Y. Kanoria, and U. Patange, "Centralized admissions for engineering colleges in india," *INFORMS Journal on Applied Analytics*, vol. 49, no. 5, pp. 338–354, 2019.

[2] J. Zimmermann, A. von Davier, and H. R. Heinimann, "Adaptive admissions process for effective and fair graduate admission," *International Journal of Educational Management*, 2017.

[3] A. AlGhamdi, A. Barsheed, H. AlMshjary, and H. AlGhamdi, "A machine learning approach for graduate admission prediction," in *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, pp. 155–158, 2020.

[4] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques.," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, 2019.

[5] B. C. Hardgrave, R. L. Wilson, and K. A. Walstrom, "Predicting graduate student success: A comparison of neural networks and traditional techniques," *Computers & Operations Research*, vol. 21, no. 3, pp. 249–263, 1994.