ISSN No:-2456-2165

Anticipating College Admissions: An Algorithmic Approach

[1]Eturu Harshith
[1]Department of Computer Science and Engineering,
Vellore Institute of Technology, AP

[2]Jemunigani Jishnu
[2]Department of Computer Science and Engineering,
Indian Institute of Technology, Palakkad

Abstract:- "In the pursuit of one's desires, Abraham Lincoln depicts the essence of discipline as the capacity to prioritize what is genuinely desired over immediate gratification. Numerous students face obstacles when tasked with choosing the college that best meets their individual needs. This is commonly attributed to a number of factors, such as defective college assessments, a lack of awareness, and dubious predictions, resulting in misgivings after the admissions process has been completed. In order to resolve this issue within the student population, our paper presents a solution in the form of an innovative web application. The primary objective of this application is to aid students in making well-informed decisions before being assigned to an institution, which is accomplished through the development of an automated prediction model customized for a particular college admissions system.

Keywords:- *Analysis, Prediction, College, Allotment, Accuracy.*

I. INTRODUCTION

In a global market that is becoming increasingly interconnected, businesses consistently pursue the expertise and experience of individuals. Young professionals aspiring to excel in their professions frequently pursue advanced degrees to expand their knowledge and abilities. However, this can be a time-consuming and dubious endeavor, as the likelihood of admission is unknown. As a result, the investment in labor may prove futile.

This initiative seeks to provide students with a reliable estimate of their likelihood of admission to specific institutions, enabling them to make informed decisions prior to the allocation process. This project's application is readily deployable for use at a variety of universities.

This study employs a dataset pertaining to the field of education that contains 500 rows and seven distinct independent variables. Using data from "**Kaggle**" and advanced machine learning algorithms, the study predicts prospective university admission rankings.

II. BASIC CONCEPTS USED AND LITERATURE REVIEW

➢ *Literature Review:*

Numerous ardent graduate students aspire to conclude their coursework and become eligible to pursue a master's degree. There is a natural curiosity among them regarding the prerequisites for university admission and the institutions that might admit them based on their qualifications.

A number of machine learning models have been utilized in several programs and research on themes linked to university admission to help students get admitted into the institutions of their choosing. There are several data models that might help students with this, and historical data demonstrates that the Naive Bayes method has been used to estimate the likelihood of admission.

Comparing this model with previous models will help us draw out better results from shady outcomes.

Discussion of other drawbacks and the scope of this project shall be discussed in the following topics:

- *Tools and Techniques used:*

This section describes the fundamental concepts underlying the tools and techniques utilized for this project.

We have utilized Jupyter Notebook to execute our code and retrieve results.

Project Jupyter's objective is to provide open-source software, open standards, and interactive computing services for a variety of programming languages, including Python. Utilizing regressions and machine learning models:

➢ *Linear Regression:*

Linear regression is one of the most fundamental and widely used Machine Learning techniques. As a statistical method for predictive analysis, it provides a straightforward and effective method for predicting outcomes. As a widely adopted technique, linear regression demonstrates its value in a variety of Machine Learning applications.

➢ *Lasso Regression:*

With less absolute shrinkage and a selection operator,a LASSO model increases predictability and comprehension through variable regularization and selection, a methodology used in statistics and machine learning.

➢ *Super Vector Model:*

Support vector machine (SVM) is a supervised machine learning model that focuses specifically on two-classification problems. In the field of supervised machine learning, SVMs excel at utilizing classification techniques to manage such scenarios.

➢ *Decision Trees:*

The decision tree is a supervised learning technique that is frequently used for both classification and regression tasks, though it is favored for classification. As a tree-structured classifier, it employs leaf nodes to represent classification outcomes and inner nodes to encapsulate the unique characteristics of the dataset.

➢ *Random Forest:*

Random Forests are an effective machine learning algorithm for supervised learning. This adaptable method is well-suited for addressing a wide variety of machine learning problems, including regression and classification tasks. It is based on the concept of ensemble learning and employs multiple classifiers to address complex problems and improve overall model performance. Random Forests' widespread adoption can be attributed to their adaptability and efficacy in addressing diverse classification and regression problems.

➢ *KNN:*

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a supervised learning classifier that makes predictions or classifications for individual data points based on proximity by identifying their nearest neighbors within the dataset.

III. PROBLEM STATEMENT / REQUIREMENT SPECIFICATIONS

In this section, you should compose the Problem Statement. Your clients here are students who aspire to gain university admission. GRE, TOEFL, SOP, LOR, and the number of studies are significant factors in determining admission. On the basis of a variety of student scores, the model predicts whether or not an applicant should be admitted to university. The objective is to predict a student's likelihood of admission to a particular university, given the data provided. We have compiled the "admission_ predict.csv" data set for 500 students.

➢ *Data Description:*

Table 1 Data Description

VARIABLE	DEFINITION
GRE	GRE score achieved
TOEFL	Toefl score achieved
LOR	Letter of recommendations given/issued
SOP	Statement of purpose
University Ranking	
Research	Number of research papers issued
CGPA	Cumulative score of a student in all semesters.

Variable Name	Type	Predictor/Response
Serial No.	Continuous	Predictor
GRE Score	Continuous	Predictor
TOEFL Score	Continuous	Predictor
University Rating	Continuous	Predictor
SOP	Continuous	Predictor
LOR	Continuous	Predictor
CGPA	Continuous	Predictor
Research	Binary	Predictor
Chance of Admit	Binary/Categorical	Response

Fig 1 Variables and their Classifications

➢ *Project Planning*

The project will be divided into several stages, starting with data acquisition and preparation.

- *Define the Problem:*

The issue statement and project scope must be established in the first stage, which also includes selecting the schools and organizations that will be studied.

- *Data Collection:*

The collecting of data necessary for the analysis, such as GPA, TOEFL, and GRE scores, is the second stage. This data may be acquired via questionnaires and surveys as well as freely available resources like university websites.

- *Data Preprocessing:*

In the third phase, the data are preprocessed and cleaned in order to eliminate outliers, fill in blanks, and normalize the data.

- *Exploratory Data Analysis:*

Exploratory data analysis is done in the fourth step to uncover further information about the data, such as connections between admission decisions and GRE and TOEFL scores.

- *Feature Engineering:*

The fifth stage involves extracting valuable elements from the data, such as merging the GRE and TOEFL scores or including other factors like LOR and extracurricular activities.

- *Model Selection:*

Selecting an appropriate machine learning model for the study, such as decision trees or logistic regression, is the sixth stage.

- *Model Training and Evaluation:*

In the seventh phase, the selected model is trained on the preprocessed data, and its performance is evaluated using metrics such as recall, precision, and accuracy.

- *Deployment and Maintenance:*

When this model is created, it will need to be continuously monitored and updated in a production environment, like a web application or a mobile app, in order to maintain accuracy and effectiveness over time.

In general, a project that plans to forecast college admission based on GRE and TOEFL results should make sure that the data is preprocessed and analyzed using the proper machine learning techniques and that the model is deployed in a user-friendly and maintainable way.

➢ *Project Analysis*

In this project, we aim to develop a forecasting model that accurately predicts a student's likelihood of gaining admission to UCLA. The project had multiple phases, including data collection and planning, design, and performance evaluation. Numerous factors affect the success of a project, including the quality of the data, the validity of the model, and the precision of the prediction. To ensure the success of our plans, we require a thorough analysis of data, models, and forecasts. The identification of relevant factors that can influence the price of a product, such as trading volume, market sentiment, and company news, is particularly dependent on data analysis. We will conduct data analysis in order to comprehend the data, including patterns, relationships, and patterns.

Model analysis will include the evaluation of model parameters and hyperparameters. To ensure that the model does not fit the training data, we will perform cross-validation and evaluate its performance using various metrics such as mean squared error and mean error. An estimate review will involve comparing the student's scores to evaluate their accuracy. We will also use visualization techniques to analyze forecast events and compare them with past trends. In addition, we will analyze the limitations of the project, such as the available data, the complexity of the model, and the impact of unforeseen events on the product. We will consider the ethics of using the forecasting model to raise funds and ensure that the project complies with ethical and legal requirements.

After the requirements are collected or the problem statements is conceptualized, there is no room for ambiguity in this project.

➢ *System Design*

- *Design Constraints:*

Design restrictions are significant elements that must be taken into account during the creation of any project. We have used a Jupyter notebook, which is extracted from Anaconda. The goal of Project Jupyter is to provide open-source software, open standards, and interactive computing services for a variety of programming languages.

- *Data Availability:*

When creating a prediction model, a significant constraint is the quantity and quality of the data. The data must be reliable, accurate, consistent, and devoid of mistakes and missing numbers. Additionally, there must be enough data to both train and test the machine learning model's accuracy.

- *Model Interpretability:*

The model used to forecast college admission should be interpretable, which implies that the end users should have no trouble understanding the characteristics and decision-making process employed by the model.

- *Privacy and Security:*

When creating a system to predict college admission, the privacy and security of the student's data are essential. Making sure the system complies with applicable privacy rules and regulations and that the data is maintained securely is crucial.

- *Scalability:*

The system must be built to manage massive volumes of data and be scalable to support an expanding population of students and universities.

- *Performance:*

The system needs to be able to deliver precise forecasts in a timely manner. To enable the system to process vast volumes of data, its performance has to be optimized.

User Experience: The system needs to be accessible and user-friendly for end users. It should be simple to use and give consumers results that are clear and succinct.

In this section, we'll go through the system's design for the Student Admission Predictor. The system's flow is depicted in the diagram below By using the user interface developed in Shiny, the student will enter the information for his or her profile.

✓ To provide the users with the desired outcome, the user interface code will communicate with the KNN and Decision Tree models.
✓ The outcome of the models' execution will be shown to the learner as the output collection method on the user interface.

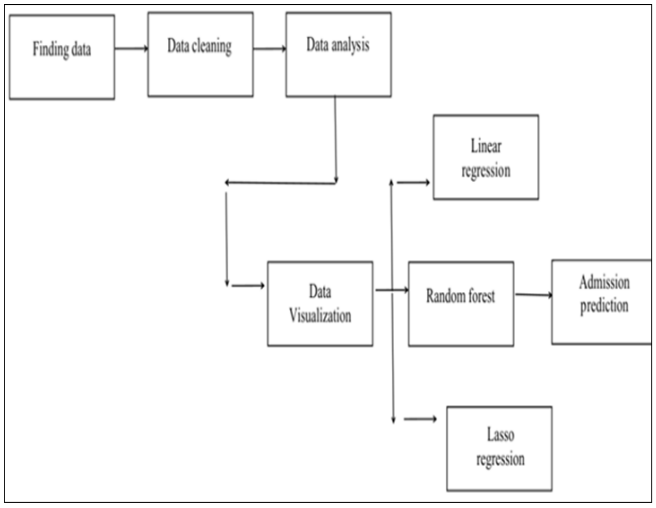


Fig 2 System Architecture or Block Diagram

ISSN No:-2456-2165

IV. IMPLEMENTATION

The first crucial step in creating a model for our use case is choosing the appropriate data collection method. For our forecasts, we have chosen a dataset that includes all the crucial elements that influence the likelihood of adoption. Deal with fields that have missing values; data cleaning comes next. Once The Data is appropriate for analysis, we use a variety of tools and libraries to visualize it.

A. Methodology or Proposal:

➢ *Problem Understanding:*

In the beginning, we must take some time to consider the issues or worries that students have prior to admission, and we must set the eradication of those issues as the goal of this study.

➢ *Data Preparation:*

Data should be cleaned, which entails taking out any noise from the data, filling in any missing values or wildly out-of-range values, and finishing any qualities or features that will be extremely important in the student admissions process.

➢ *Building Models:*

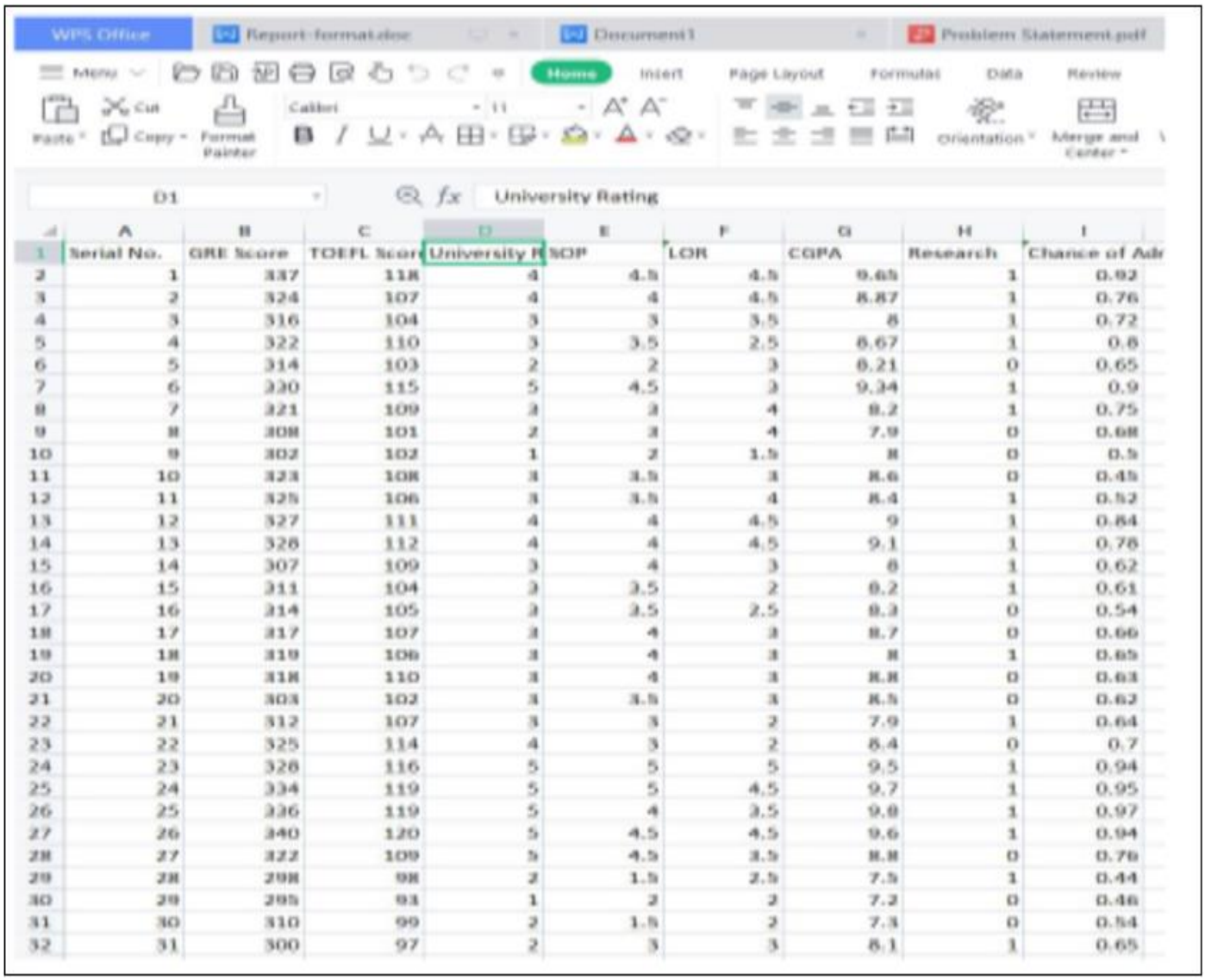
For admittance to a certain institution, a number of ML models must be created using various machine learning methods, and a user interface must be created to access those models.

➢ *Evaluation:*

The accuracy scores of developed models are examined. The model will be combined for final deployment after it has been finished.

➢ *Data Cleaning:*

You may determine what needs to be cleaned up or processed by looking at feature values until you find the range or distribution of values that is normal for each characteristic. Based on our examination of the data,No missing values exist, and There is no need to account for anomalies. manage this set of information.



	A	B	C	D	E	F	G	H	I
1	Serial No.	GRE Score	TOEFL Score	University R	SOP	LOR	CGPA	Research	Chance of Adr
2	1	337	118	4	4.5	4.5	9.65	1	0.92
3	2	324	107	4	4	4.5	8.87	1	0.76
4	3	316	104	3	3	3.5	8	1	0.72
5	4	322	110	3	3.5	2.5	8.67	1	0.8
6	5	314	103	2	2	3	8.21	0	0.65
7	6	330	115	5	4.5	3	9.34	1	0.9
8	7	321	109	3	3	4	8.2	1	0.75
9	8	308	101	2	3	4	7.9	0	0.68
10	9	302	102	1	2	1.5	8	0	0.5
11	10	323	108	3	3.5	3	8.6	0	0.45
12	11	325	106	3	3.5	4	8.4	1	0.52
13	12	327	111	4	4	4.5	9	1	0.84
14	13	328	112	4	4	4.5	9.1	1	0.78
15	14	307	109	3	4	3	8	1	0.62
16	15	311	104	3	3.5	2	8.2	1	0.61
17	16	314	105	3	3.5	2.5	8.3	0	0.54
18	17	317	107	3	4	3	8.7	0	0.66
19	18	319	106	3	4	3	8	1	0.65
20	19	318	110	3	4	3	8.8	0	0.63
21	20	303	102	3	3.5	3	8.5	0	0.62
22	21	312	107	3	3	2	7.9	1	0.64
23	22	325	114	4	3	2	8.4	0	0.7
24	23	328	116	5	5	5	9.5	1	0.94
25	24	334	119	5	5	4.5	9.7	1	0.95
26	25	336	119	5	4	3.5	9.8	1	0.97
27	26	340	120	5	4.5	4.5	9.6	1	0.94
28	27	322	109	5	4.5	3.5	8.8	0	0.76
29	28	298	98	2	1.5	2.5	7.5	1	0.44
30	29	295	93	1	2	2	7.2	0	0.46
31	30	310	99	2	1.5	2	7.3	0	0.54
32	31	300	97	2	3	3	8.1	1	0.65

Fig 3 Data Cleaning and Preparation Process for Student Admissions Prediction Model

ISSN No:-2456-2165

➢ *Data Visualization:*

The label we must take into consideration from the aforementioned data is "Chance of Admission", and we must then take into account the characteristics that impact or significantly affect Chance of Admission. Once the data has been analyzed, we will be able to identify features and labels.

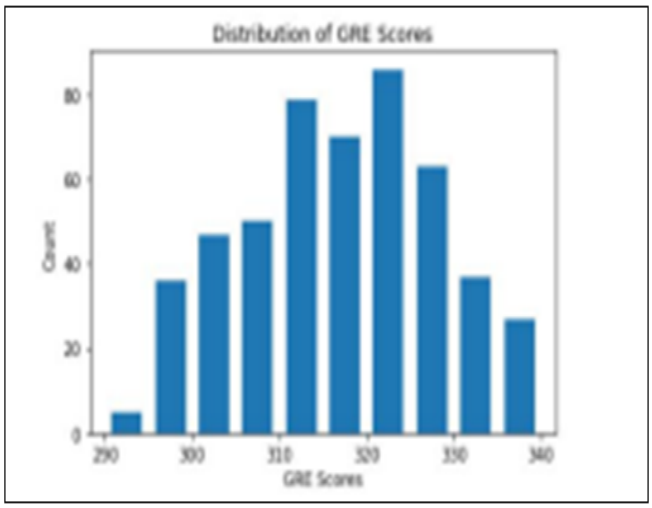


Fig 4 Distribution of GRE Scores

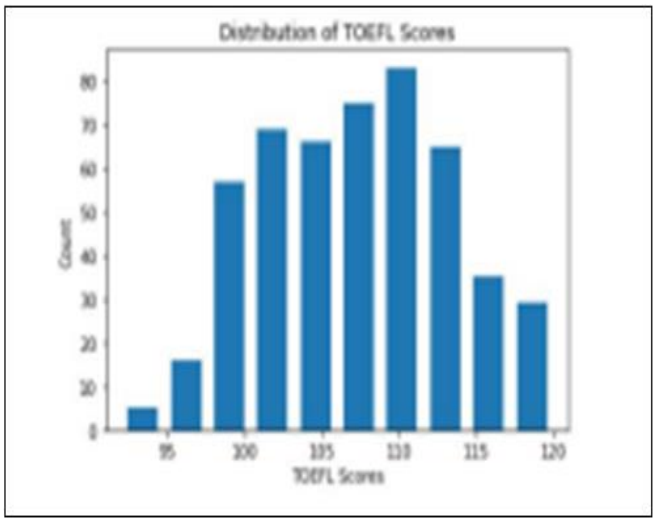


Fig 5 Distribution of TOEFL Scores

B. Testing or Verification Plan

In this study, the original dataset was divided into training and test portions (80% and 20%, respectively). Machine learning models—Logistic Regression, Linear regression, decision tree model, K-NN, and random—were subsequently trained to correspond with the training data. The odds of training data predicted by admitting the trained models All models were performed in the Anaconda-specific Jupyter environment for coding and model training.

➢ *Testing Analysis is Made in the following Table:*

Table 2 ML Model Performance on Admission Prediction (Test Data)

Test ID	Test Case	Title: Test Condition	System	Behavior Expected Result
T01	Checking for missing values	df_copy.isnull()	Working fine/compatible	No null values are present.
T02	Evaluating eachwith cross values	using grid search	Working fine/compatible search	Return the train score.
T03	Cross-validation validation linear regression	cross_val_score(Linea r Regression (no)) with rmalize=True)	Warning encountered, but. Got got the accuracy	The accuracy.

C. Result Analysis or Screenshots

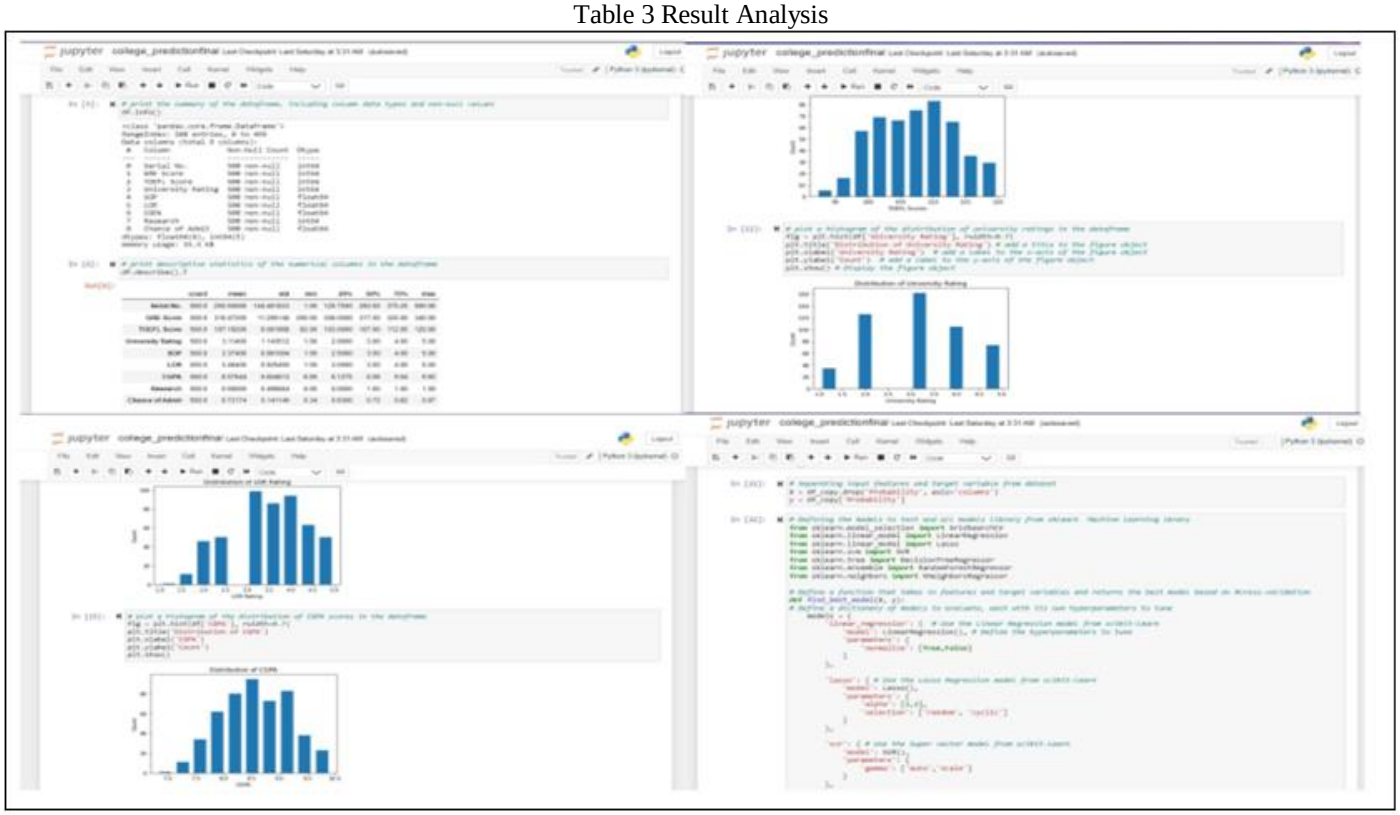
The findings of this investigation seem to suggest that it makes a significant contribution to the response variable "Chance of Admit." The likelihood of admission increases with higher GRE and TOEFL scores. Based on the aforementioned parameters, the model may be used to predict the likelihood of admission with an accuracy of 82.5%. This approach will help institutions forecast admission decisions and streamline their selection and scheduling processes.

The model supported the premise that GRE, TOEFL, and other test results are required for admission to Master's degree programs. These measures can be used to assess the outcomes of the GRE and TOEFL-based college entrance prediction systems.

The system can be deemed effective in predicting college admissions based on GRE and TOEFL results if the accuracy, precision, recall, and F1 scores are high and the confusion matrix displays a decent balance between true positives and false positives.

ISSN No:-2456-2165

Table 3 Result Analysis



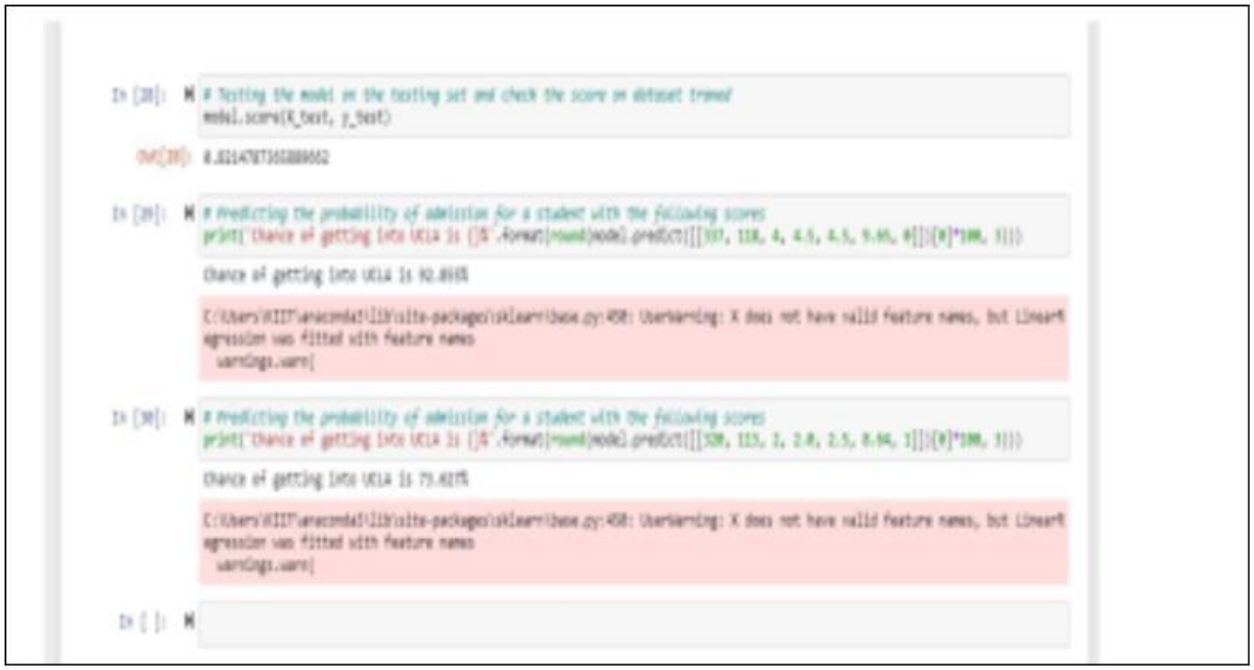


Fig 6 Coefficient Determination

D. Quality Assurance:

Quality assurance processes ensure that standards are reliable, accurate, valid, and meet end-user expectations.

- Good Data: Good data used to train and evaluate the model is important. Be mindful of the accuracy and reliability of the forecast. Data must be clean, consistent, and error-free. Data security processes may include data validation, data maintenance, and data normalization processes.
- Model Validation: The accuracy and reliability of the model must be checked through rigorous testing and validation.
- Maintenance: The model should be maintained and updated to ensure it remains accurate and viable over time. This will include monitoring performance standards, updating weak standards, and reintroducing new data standards.

ISSN No:-2456-2165

V. STANDARDS ADOPTED

➢ *Design Standards*

- *Project Goal:*

To achieve the highest accuracy of possibility for a student to get admitted to UCLA. This project assists students in getting an inerrant chance of admission to a particular college and helps them make better choices before allocation. Using this model, we can deploy a whole application for the college.

Outcome: Students will receive a percentage of possibility based on their given inputs. Model analysis will include the evaluation of model parameters and hyperparameters. To ensure that the model does not fit the training data, we will perform cross-validation and evaluate its performance using various metrics, such as mean squared error and mean error. With the training of the model, we have arrived at the conclusion that **Linear Regression** gives the most accurate chances.

Risks: Since the model is not completely accurate, there might be a slight risk, which, as seen in most of the cases, is negligible. Another risk we might encounter is if the college providing the admission is not in accordance with the given data or has changed the admission procedure.

- *UML Diagram:*

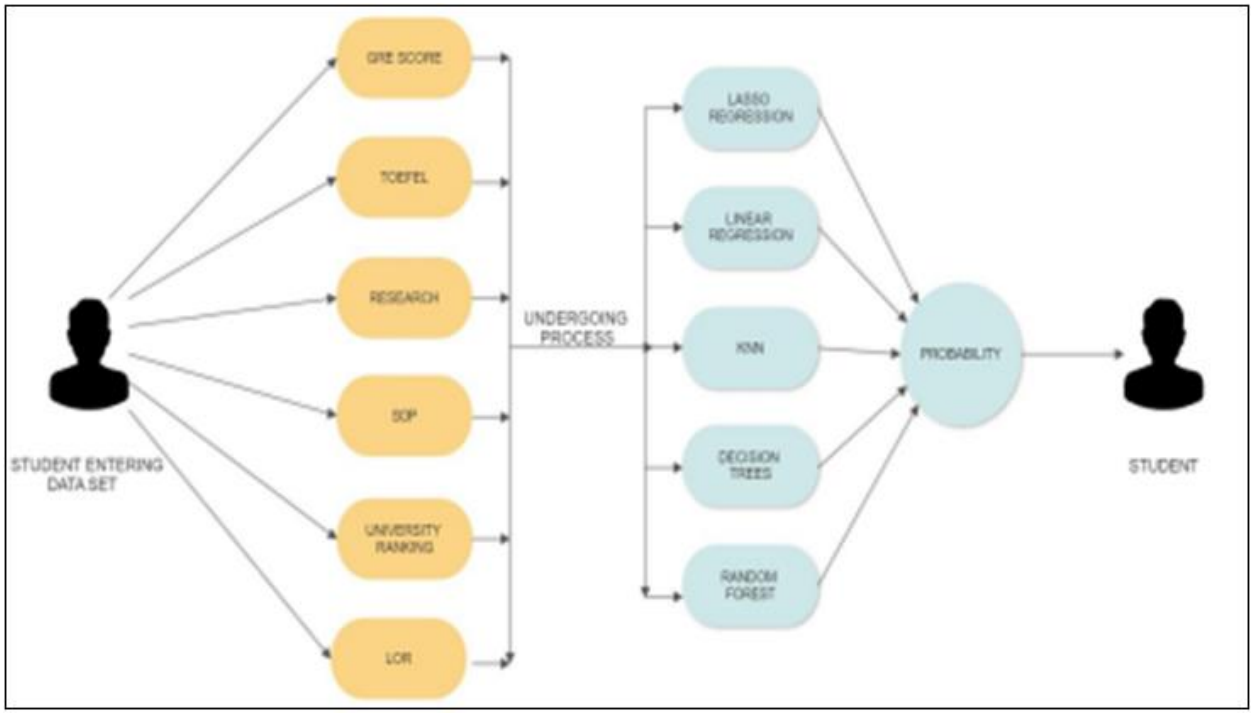


Fig 7 UML Diagram

VI. CONCLUSION AND FUTURE SCOPE

➢ *Conclusion*

Each year, a large number of students begin their academic careers by enrolling at colleges and universities. Nonetheless, a significant proportion of these pupils confront difficulties due to limited resources, a lack of prior knowledge, and a lack of decision-making skills. As a result, issues arise, such as applying to unsuitable institutions, resulting in time, money, and effort being squandered.

To address this issue, our initiative focuses on assisting students in selecting the university that best meets their needs. We stress the importance of applying to colleges where there is a reasonable possibility of acceptance, as opposed to squandering time on applications to institutions where acceptance is unlikely. Students can save both time and money by limiting their college applications to those with the highest likelihood of acceptance.

In this study, machine learning models were used to forecast a student's chance of admission to a master's program. K-nearest neighbor, random forest, lasso regression, super vector model, and linear regression are the machine learning models that are featured, out of which linear regression proved to come up with the highest accuracy.

➢ *Future Scope*

- *Community Learning:*

Community learning techniques can be used to combine multiple models to improve the precision and robustness of the prediction.

- *Feature Engineering:*

Advanced design techniques can be used to capture complex patterns and patterns in data, such as in media analysis or related social media products.

- *Defined Models:*

The use of defined models helps to clarify and better understand the factors affecting student results and the predictions made by the model.

- *Integration with Trading Platforms:*

The integration of forecasting models with web applications can allow users to check with more universities, as this model provides the chances of admission for a single university. A web application can host machine learning for numerous universities, and students can enroll themselves in the model's most accurate predicted college.

Overall, the future of this work is broad and has great potential for further research and development in the field of prediction. With the conducted experiments, we have drawn the conclusion that **Linear Regression** gives the highest accuracy.

This model's superior performance in comparison to other existing models is primarily attributable to a crucial distinction in its approach. While previous studies in this field relied on the Naive Bayes algorithm to predict a student's likelihood of admission to a particular university, they failed to account for all the essential variables that influence admission decisions, such as TOEFL or IELTS scores, a statement of Purpose (SOP), a letter of Recommendation (LOR), and an undergraduate GPA.

Our model, on the other hand, has been meticulously designed to account for each of these critical factors, resulting in more accurate and trustworthy results. By incorporating these variables, we expect our model to produce more accurate predictions and more accurate results when assessing the likelihood of student acceptance.

REFERENCES

[1]. College Admission Prediction using Ensemble Machine Learning Models Vandit Manish Jain1, Rihaan Satia2
[2]. University Admission Prediction using Machine Learning Kruthika CS, Apeksha B, Chinmaya GR, Madhumathi JB, and Veena MR
[3]. Predicting Graduate Admissions using ML From Kaggle, Venugopal Adep