International Journal of Early Childhood Special Education (INT-JECSE)
DOI:10.9756/INTJECSE/V14I5.160 ISSN: 1308-5581 Vol 14, Issue 05 2022

PREDICTION OF GRADUATE ADMISSION USING MACHINE LEARNING

Dr.A.Anjaiah
Associate Professor
Dept . of. CSE
St.Peter's Engineering College(A)
Hyderabad ,TS, India
anjaiah@stpetershyd.com

NVLN Ramakrishna
IV B.Tech
Dept .of. CSE
St.Peter's Engineering College(A)
Hyderabad, TS, India
ramakrishna7946@gmail.com

Thota Bhavani
IV B.Tech
Dept .of. CSE
St.Peter's Engineering College(A)
Hyderabad, TS, India
bhavanithota562000@gmail.com

K.Krishna Chaitanya
IV B.Tech
Dept .of. CSE
St.Peter's Engineering College(A)
Hyderabad, TS, India
chaitu082001@gmail.com

***Abstract*— Graduate degrees are becoming more popular as a result of the current, extremely competitive work environment. Both applicants and university entrance faculty members have been burdened by this, which has also increased workload. Many students in today's educational environment prefer to continue their education after completing an engineering programme or any graduate degree programme at universities abroad, such as those in the USA, UK, and other countries. The TOFL and GRE exams, which are required for studying abroad, must be taken by students who want to pursue master's degrees at universities overseas. One of the most important things students must think about after taking the examinations is preparing their SOP and LOR.**

There are some consultancies and internet tools that suggest institutions, but many charge exorbitant fees for their services, and the online tools are sometimes are not correct.

Thus, we suggest this initiative. employing machine learning to predict graduate admission, which will inform students of their chances of admission to other colleges. To obtain the prediction, we use various machine learning models. These models ought to be highly accurate and ought to take into account all the important variables that are crucial to the student admissions process. The anticipated results offer students a precise sense of their prospects of admission to a particular university. The suggested model makes use of both classification and regression algorithms.

Keywords— admissions, graduate studies, machine learning Regression, Classification.

I. INTRODUCTION

A form of artificial intelligence (AI) called machine learning enables computers to automatically get better over time. As its applications expand to new fields such as agriculture, finance, electronic commerce, logistics, marketing, and security, it is now playing a larger role in our daily lives. Additionally, machine learning makes a growing number of applications possible that weren't

before. The use of machine learning in education is likewise growing faster. Many researchers and scientists have recently expressed interest in using machine learning in educational settings.

In this regard, early admissions prediction is seen as a crucial subject for both the university and new graduate students. Sadly, recently graduated students frequently do not know what is necessary to get admitted to a university postgraduate programme. Because of this, they waste their limited time and resources concentrating on activities that won't improve their prospects of getting accepted into graduate programmes.

What artificial intelligence algorithm is the most effective in predicting university admission? are the primary and actual difficulties surrounding this topic. What factors are second-most crucial in determining acceptance chances?

The main goal of this research is to offer a machine learning-based approach for early admission prediction to institutions. Linear regression, support vector regression, decision tree regression, and random forest regression are just a few of the supervised machine learning techniques that will be used to analyse our dataset in the first stage. These methods also include Gradient Boosting Classifier, KNN, Random Forest, and Logistic Regression. The second objective is to evaluate and contrast all of the evaluation criteria and algorithms that were used to create a prediction model.

II. LITERATURE SURVEY

The goal of many aspiring graduate students is to finish their coursework and get ready for the master's degree, which is the next step. Many of them might be interested in learning more about the prerequisites for university entrance as well as the universities to which they might be admitted based on those prerequisites[1, 2].

The assessment of student enrollment prospects in universities is covered in this section's review of the relevant literature. Numerous studies and initiatives on the subject of university admission have been conducted, and these have made extensive use of machine learning models to assist students in getting accepted into the colleges of their choice.

The main limitation of earlier research in this field is that it did not take into account all the variables that will affect a student's admission, such as TOEFL/IELTS, SOP, LOR, and undergrad score. An earlier study in this area evaluated the likelihood that a student application to a certain university would be accepted using the Naive Bayes algorithm. However, the current method makes use of several classification and regression techniques to estimate the likelihood that an applicant would be admitted while assessing the applications submitted by university-bound international students.

By comparing prospective students' scores to those of university students currently enrolled, this model was created to predict how well they will do The programme so predicted whether the potential student should be admitted to the university based on a variety of student scores.. In the current approach, comparisons are done between students who were accepted into institutions and those who had their applications denied, ensuring that the method is completely error-free.

III. METHODOLOGY

The whole approach is depicted by the following flowchart:

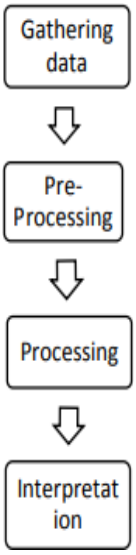


Figure 2. Flow Chart of The Technique

Understanding the problem: Prior to moving ahead, it is important to take some time to consider the issues or worries that students had throughout the time leading up to their admission. The goals of this study should then be to find answers to those issues.

Data Preparation: The data needs to be cleaned, which entails taking out any noise, filling in any missing numbers or outlier values, and deciding on the traits and aspects that will be of the utmost significance in the admissions process for students.

Building Models: For admittance to a specific university, a number of ML models must be created using different machine learning techniques, and a user interface must be created to allow for access to those models [3].

Evaluation: Based on their accuracy ratings, developed models are assessed. When the model is finished, it will be integrated for the final deployment.

cleansing and analysis of data: • Examining feature values to determine [5] until you see the range or distribution of values typical for each feature, what needs to be cleaned or pre-processed • Anomalies such as incorrect data formats, incorrect units for a certain column, or a dearth of cases for a particular class can be found.or you might discover missing or noisy data. • You can determine if the issue is indeed solvable without the use of machine learning. The data cleansing procedure offers the following major advantages:

1. By doing this, significant flaws and inconsistencies that are unavoidable when combining data from many sources into one dataset are removed.
2. Everyone will be more productive thanks to data cleansing software because they will be able to
3. Less errors result in happier customers and happier employees.
4. The capability to visualise the various operations, as well as the purpose and source of your data.

Because we analysed the data, there aren't any outliers or missing numbers, thus there's no need to deal with them in this set of data. We can fill in (or remove) any missing numbers and outliers using any method.

Dataset:

The data in use [6] consisted of 500 instances, each of which represented a candidate. Neither categories nor null value entries were present in the dataset. This dataset was obtained from UCLA's admissions history information. There are eight attributes total in the dataset, and each one is a number:

I. GRE score (General Record Examinations); this score assesses general proficiency in English and mathematics at the undergraduate level. This score has a value between 260 and 340.

II. The TOEFL score, which gauges a student's proficiency in English, is the second factor. This score has a value between 0 and 120.

III. The SOP (Statement of Purpose), which is a statement outlining the applicant's motivation for submitting the application On a scale of one to five, something is scored.

IV. The LOR (Letter of Recommendation) evaluates the validity of the applicant's recommendation. On a scale of one to five, something is scored.

V. Cumulative Grade Point Average (CGPA), which is based on the applicant's success in

undergraduate courses. On a scale of one to ten, something is scored.

VI. University Ranking, which is based on the applicant's prior school's standing. On a scale of one to five, something is scored.

VII. The applicant's familiarity with research is taken into account when grading their research experience. This value can either be one or zero.

VIII. Chance of Admission, or the proportion of applicants admitted to graduate school. This characteristic is the intended value for which the rate from zero to one will be anticipated.

	A	B	C	D	E	F	G	H	I
1	Serial No.	GRE Score	TOEFL Sco	University	SOP	LOR	CGPA	Research	COA
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

Admission_Predict_Ver1.2

Figure 1. Data from Admission Prediction CSV

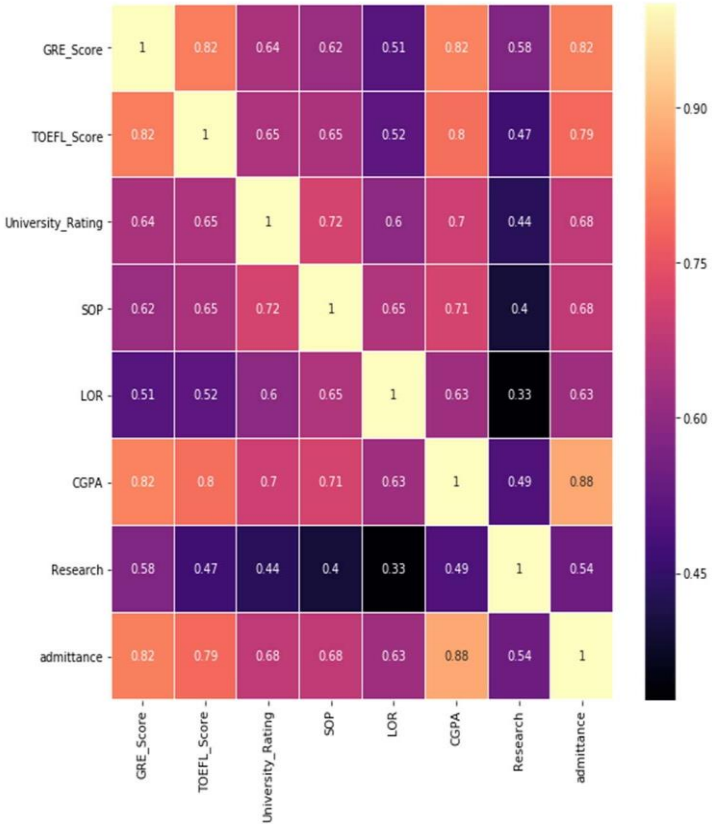
Data Visualization:

Our ability to recognise the characteristics and labels will depend on the results of the data analysis. Based on the aforementioned data, we must first take into account the parameters that significantly affect or have an influence on chance of admission[4].

• We can learn about specific characteristics that are more impacted by the analysis or visualisation. Correlation matrices are one such data visualisation technique.

In order to identify the predictive analysis techniques that would be most helpful in developing a suitable model, correlation matrices can help.

It is evident that the admission rate and all features have a uniformly positive connection. The "CGPA" factor has a significant value of 0.88, making it the most important factor in determining admission rates, while the "Research" factor has



the least significant impact on admission rates, with a value of 0.54.

Fig.1Correlation matrix incorporating all eight qualities

IV. ALGORITHMS

For predictive modelling using machine learning, there are numerous algorithms. The sections below describe the algorithms used in the proposed system to create the regression and classification predictive models.

A. Evaluation models

Linear Regression (LR):

The simple and easiest supervised machine learning algorithm is linear regression LR [10].A relationship between a dependent variable and one or more independent variables can be modelled using this method. Finding a straight regression line through the points is what it entails.

Support Vector Regression (SVR) : Another well-liked Machine Learning technique that is utilised for both classification and regression is Support Vector Regression (SVR) [11]. With only a few small exceptions, it is same as linear regression. SVR enables us to choose the acceptable level of error in our predictive model and will select the best line to suit the data.

Random Forest: [12] One of the most well-known and effective machine learning algorithms that can handle both classification and regression tasks is the random forest technique. In a limited number of decision reads, this method generates forests.

Random forest regressor:

In the Random Forest Regression (RFR) ensemble learning method, a large number of decision trees are built during the training phase, and the average prediction of each tree is used to enhance the prediction.

Random forest classifier:

The class that the majority of trees in the random forest choose as their output is what is used in the random forest classifier task.

Gradient Boosting regressor and classifier:

By integrating the weak learners or weak predictive models, the gradient boosting algorithm creates an ensemble model. Starting with a constant, such as the average of the target values, the model is fitted. The estimators or decision trees are fitted to predict the negative gradients of the samples in following rounds. Each iteration updates the gradients (for every subsequent estimator). The outcome or the contribution from each additional tree or set of estimators is shrunk using a learning rate. When the target column is continuous, we use the gradient boosting regression; however, when the problem is a classification problem, we use the gradient boosting classifier. Between the two, the "Loss function" is the only distinction.

By utilising gradient descent to add weak learners, the major objective of this exercise is to minimise this loss function. We will have different loss functions for regression issues, such Mean Squared Error (MSE), and for classification problems, like log-likelihood, since it is based on a loss function.

Logistic Regression:

This approach uses a threshold and a sigmoid function. The Logistic Regression algorithm assigns weights to each feature using a cost function, much like the Linear Regression algorithm. The approach uses gradient descent to minimise the cost function.

K-Nearest Neighbours:

The K-Nearest Neighbors algorithm is the one that is most frequently used for classification and regression purposes. The term "KNN" stands for "k nearest neighbor," where "k" is an integer value

that indicates how many neighbours should be compared. It is applicable to both classification and regression. To find the mode value, for instance, if the k value is 5, it will compare the five closest neighbours.. If the k value is 6, on the other hand, it will take the six closest values and determine the mean value.

B.Evaluation Methods

A crucial component of creating a powerful machine learning model is model evaluation. There are numerous evaluation techniques available. We will discuss the three main metrics we will use in our evaluation—R-squared (R2), Mean Square Error (MSE), and Root Mean Square Error—in the sections that follow (RMSE).

R-Squared (R2 or the coefficient of determination):

A tool for assessing the effectiveness of simple linear regression is R-Squared (also known as R2 or the coefficient of determination) [8]. It determines how well the point distribution is described by the regression equation or how well the model fits the observed data. In other words, the scatter plot scatters more widely about the regression line the closer the coefficient of determination is near 0.On the other hand, when the R2 gets closer to the regression line, the cloud of dots around it becomes smaller. 1. R2 equals 1 when all points on the regression line are perfectly aligned.

Mean Square Error (MSE):

The square of the prediction between the model and the observations is what is known as the arithmetic mean ([9]). In the context of a single or several regressions, this value must be minimised. The technique is dependent on the residuals' mean being null, or not present. However, their squares' average rarely equals zero.

Root Mean Square Error (RMSE)

It is a typical method for figuring out the inaccuracy in model evaluation studies. The square root of a mistake is its mean across all squares.

B. Comparision of machine learning algorithms for predicting university admission

Regression algorithms and classification techniques are compared in Table 1 for each type of algorithm (logistic regression, KNN, random forest classifier and gradient boosting classifier).

TABLE I. COMPARING MODELS

TABLE II. ALGORITHMS	***Accuracy score***
Linear Regression	0.821208
Support Vector Regressor	0.759781
Random Forest Regressor	0.795852
Gradient Boosting Regressor	0.796148
Logistic Regression	0.9250
K-Nearest Neighbours	0.8875
Random Forest Classifier	0.9500
Gradient Boosting Classifier	0.9750

Due to its high R2 score in comparison to other machine learning models, as shown in table 1, it is evident that the Gradient Boosting Classifier offers higher performance.

V.CONCLUSION

By creating a prototype of the system that students wishing to further their education at other universities can use, the major mission of this work is to compare machine learning algorithms and discover the best suitable algorithms to forecast the chance to get admitted in Foreign Universities. For the suggested work, numerous machine learning algorithms were employed. Comparing Gradient Boosting Classifier to other regression and classification algorithms, it became clear that it was the most appropriate for the system's development. With an average accuracy of 97%, students can use the model to predict their chances of admission to a specific university.

A GUI was created to make the application useable and user-friendly from a non-technical standpoint. The technique enables students to save a sizeable amount of time and money that they would otherwise spend on academic advisors and application costs for universities where they have little possibility of admittance, so successfully achieving the intended purpose of the research. The major limitation of our study is that we only used data from Indian students who were pursuing master's degrees in computer science in the US to build our models, and we only included a small number of universities with varying scores. In the future, the curriculum may be expanded to include more details about fresh colleges and courses. By modifying node-red, the system can also be converted into a web-based application.

Finally, students can utilise machine learning model to predict their odds of admission to a specific university with great accuracy.

ACKNOWLEDGEMENT:

We are thankful to St. Peter's Engineering College for helping us with laboratory and continuing support to prepare this paper in a brighter manner.

REFERENCES

[1] Chen, Y., Pan, C. C., Yang, G. K., & Bai, J. (2014, August). Intelligent decision system for accessing academic performance of candidates for early admission to university. In *2014 10th International Conference on Natural Computation (ICNC)* (pp. 687-692). IEEE

[2] Hasan, M., Ahmed, S., Abdullah, D. M., & Rahman, M. S. (2016, May). Graduate school recommender system: Assisting admission seekers to apply for graduate studies in appropriate graduate schools. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 502-507). IEEE.

[3] L. Breiman, Accuracy Predictors, Machine Learning, 24(2):123-140,1996.

[4] Data Visualizaton, Machine Learning https://www.analyticsvidhya.com/blog/2017/09/common-machin e-learning-algorithms/

[5] Data Cleaning and Analytics, Machine Learning https://archieve.ics.uci.edu/ml/index.php

[6] Mair C, Kadoda G, Lefley M, Phalp K, Schofield C, Shepperd M, Webster S. "An investigation of machine learning based prediction systems". *Journal of systems and software*. 2000 Jul 15;53(1):23-9

[7] https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

[8] Miles, J. (2014). R squared, adjusted R squared. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat06627

[9] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate re search, 30(1), 79-82. https://doi.org/10.3354/cr030079

[10] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

[11] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199-222. https://doi.org/10.1023/b:stco.0000035301.49549.88

[12] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

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

[14] Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collabora tive filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22(1), 5-53. https://doi.org/10.1145/963770.963772

[15] https://www.saedsayad.com/decision_tree.htm

[16] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[17] L. Breiman, Accuracy Predictors, Machine Learning, 24(2):123-140,1996.

[18] D.M Blei, A.Y. Ng, and M. I. Jordan, Latent Dirichlet allocation,Journal of Machine Learning Research,3:993-1022, 2003.

[19] Kotsiantis SB, Zaharakis I, Pintelas P. "Supervised machine learning: A review of classification techniques". *Emerging artificial intelligence applications in computer engineering.* 2007 Jun 10;160(1):3-24.

[20] Data Visualizaton, Machine Learning https://www.analyticsvidhya.com/blog/2017/09/common-machin e-learning-algorithms/