UNIVERSITY ADMIT ELIGIBILITY PREDICTOR

M.Rajeswari
Department of Electronics and Communication Engineering
Panimalar Engineering College

A.Sweatha
Department of Electronics and Communication Engineering
Panimalar Engineering College

K.Sarika
Department of Electronics and Communication Engineering
Panimalar Engineering College

R.Niranjani
Department of Electronics and Communication Engineering
Panimalar Engineering College

Abstract— Student admission problem is very important in educational institutions. This paper addresses machine learning models to predict the chance of a student to be admitted to a master's program. This will assist students to know in advance if they have a chance to get accepted. Newly graduate students usually are not knowledgeable of the requirements and the procedures of the postgraduate admission and might spent a considerable amount of money to get advice from consultancy organizations to help them identify their admission chances. Human consultant and calculations might be bias and inaccurate. The machine learning models are multiple linear regression, k-nearest neighbor, random forest, and Multilayer Perceptron. Experiments show that the Multilayer Perceptron model surpasses other models.

Keywords— K- Nearest neighbors,Multilayer Perceptron, Multiplelinear regression, Random forest, Student admission

I. INTRODUCTION

The aim of this project is to help students in shortlisting universities with their profiles. The predicted output gives them a fair idea about their admission chances in a particular university. This analysis should also help students who are currently preparing or will be preparing to get a better idea. The world markets are developing rapidly and continuously looking for the best knowledge and experience among people. Young workers who want to stand out in their jobs are always looking for higher degrees that can help them in improving their skills and knowledge.

This fact has motivated us to study the grades of students and the possibility of admission for master's programs that can help universities in predicting the possibility of accepting master's students submitting each year and provide the needed resources.

- Graduate Record Exam[1] (GRE) score. The score will be out of 340 points.
- Test of English as a Foreigner Language[2] (TOEFL) score,which will be out of 120 points.
- University Rating (Uni.Rating) that indicates the Bachelor University ranking among the other universities. The scorewill be out of 5.
- Statement of purpose (SOP) which is a document written to show the candidate's life, ambitious and the motivations for the chosen degree/ university. The score will be out of 5 points.
- Letter of Recommendation Strength (LOR) which verifies the candidate professional experience, builds credibility,boosts confidence and ensures your competency. The score is out of 5 points.
- Undergraduate GPA (CGPA) out of 10
- Research Experience that can support the application, such as publishing research papers in conferences, working as research assistantwith university professor (either 0 or 1).

One dependent variable can be predicted which is chance of admission, that is according to the input given will be ranging from 0 to 1.

II. TECHNICAL BACKGROUND

A. Shapiro-Wilk Normality Test

The Shapiro-Wilks test is a test performed to detect whether a variable is normally distributed or not depending on the p- value. In case the p-value was less than or equal 0.05, the test will reject the null hypothes . Otherwise, the variable is normally distributed. It is good to mention that Shapiro test does has limitations. Moreover, it is biased toward large samples. The larger the sample, the more possibility to get a statistically significant results[1].

B. Multiple Linear Regression

Multiple linear regression is a statistical technique used to predict a dependent variable according to two or more independent variables. As well as, present a linear relationship between them and fit them in a linear equation. The format of the linear equationis as following [2]:

$$y_i=\beta_0+\beta_1 xi_1+...+\beta_n x_n+\epsilon \quad (1)$$

where, for i=n observations:

y_i=dependent variable

x_i= independent variables

β_0=y-intercept

β_n=slope coefficients for each independent variable

ϵ=the model's error term or residuals

C. K-Nearest Neighbor

K-nearest neighbor (KNN)is a supervised machine learning algorithm used for classification and regression problems [3], [4]. It is based on the theory of similarity measuring. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for regression as well as for classification but mostly it is used for the classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

Therefore, to predict a new value, neighbors should be put into consideration. KNN uses some mathematical equations to calculate the distance between points to find neighbors. In a regression problem, KNN is used to find the mean of the klabels. While in classification problems, the mode of k labels will be returned[5].

D. Random Forest

The random forest algorithm is one of the most popular and powerful machine learning algorithms that is capable of performing both regression and classification tasks [6]. This algorithm creates forests within number of decision reads. Therefore, the more data is available the more accurate and robust results will be provided[7].

Random Forest method can handle large data sets with higher dimensionality without overfitting the model. In addition, it can handle the missing values and maintains accuracy of missing data [7].

E. Multilayer Perceptron

Multilayer perceptron is a supervised deep artificial neural network machine-learning algorithm used to predict value of a dependent variable of a dataset according to weights and bias [6]. Weights are updated continuously when finding any error in classification. The first layer is the input layer, more than one layer are presented next as hidden layers where each layer will contain a linear relationship between the previous layer, and the final layer is output layer that makes decisions and predicts [8].

Forward pass and backward pass can be performed. Forward pass is where information flows from left to right, in other words, the flow will be input,hidden layers, and output in order. On the other hand, backward weights will adjust according to the gradient flow in that direction [8].

F. Logistic Regression

Logistic regression algorithm is used to identify the probability of occurrence of an event based on single predictor variable. Multivariate Logistic regression can be used to determine the probability of the occurrence of an event based on multiple predictor variables. The class variable that has to be predicted has to be binary or dichotomous. Logistic Regression is also a supervised machine learning algorithm which used data with predetermined classes to create a model and perform predictive analysis on unseen data.

G. Decision Tree

It is a supervised machine learning algorithm. Due to its simple logic, effectiveness and interpretability it the most widely used classification algorithm. The model works by creating a tree-like structure by dividing the data-set into several smaller subsets based on different conditional logic. The main components of the decision tree are the decision nodes, leaf nodes and the branches. Nodes with multiple branches are the decision nodes, nodes with no branches are called the leaf nodes, and the top node is called the root node of the decision tree. The nodes are connected to each other via branches based which are different conditions. The root and decision nodes are created by computing the entropy and information gain for the data-set.

III. RELATED WORK

A great number of researches and studies have been done on graduation admission datasets using different types of machinelearning algorithms.

One impressive work by Acharya et al. [9] has compared between 4 different regression algorithms, which are: Linear Regression, Support Vector Regression, Decision Trees and Random Forest, to predict the chance of admit based on the best model that showed the least MSE which was multi- linear regression. Then compute error functions for the developed models and compare their performances to select the best performing model out of these developed models the linear regression is the best performing model.

In addition, Chakrabarty et al. [10] compared between both linear regression and gradient boosting regression in predicting chance of admit; point out that gradient boosting regression showed better results.

Gupta et al. [11] developed a modelthat studies the graduate admission process in American universities using machine learning techniques. The purpose of this study was to guide students in finding the best educational institution to apply for. Five machine learning models were built in this paper including SVM (Linear Kernel),AdaBoost, and Logistic classifiers.

Waters and Miikkulainen [12] proposed a remarkable article that helps in ranking graduation admission application according to the level of acceptance and enhances the performance of reviewing applications using statistical machine learning.

Sujay [13] applied line arregression to predict the chance of admitting graduate students in master's programs as a percentage. However,no more models were performed.

Bayesian Networks were used by (Thi et al. (2007)) to create a decision support system for evaluating the application submitted by international students in the university. This model was designed to predict the performance of the aspiring students by comparing them with the performance of students currently studying in the university and had similar profile during their application. In this way based on the current students profile the model predicted whether the aspiring student should be granted admission to the university. Since the comparisons were made only with the students who were already admitted in the university and the data of the students who were denied admission were not included in the research this model proved to be less efficient due to the problem of class imbalance.

(Abdul Fatah S; M (2012)) developed a model that can provide the list of universities/colleges where the which best suitable for a student based on their academic records and college admission criteria. The model was developed by applying data mining techniques and knowledge discovery rules to the already existing in-house admission prediction system of the university.

(Mane (2016)) conducted a similar research that predicted the chance of a student getting admission in college based on their Senior Secondary School, Higher Secondary School and Common Entrance Examination scores using the pattern growth approach to association rule mining. The performance of both the models was good the only drawback was the problem statement was single university-centric.

Janani Pet al. [14] proposed a developed project uses machine learning technique specifically a decision tree algorithm based on the test attributes like GRE, TOEFL, CGPA, research papers etc. According to their scores the possibilities of chance of admit is calculated. The developed model has 93% accuracy.

Navoneel Chakrabartyet al. [10] proposed a comparison of different regression models. The developed models are gradient boosting regress or and linear regression model. Gradient boosting regression or have to score of 0.84. That surpassing the performance of linear regression model. They

computed different other performance error metrics like mean absolute error, mean square error, and root mean square error.

Chithra Apoorva et al. [15] proposed different machine learning algorithms for predicting the chances of admission. The models are K- Nearest Neighbor and Linear Regression, Ridge Regression, Random Forest. These are trained by features have a high impact on the probability of admission. Out of the generated models the linear regression model have 79% accuracy.

Mishra and Sahoo (2016) conducted a research from a university point of view to predict the likelihood of a student enrolling in the university after the have enquired about of courses in the university. They used K-Means algorithm for clustering the students based on different factors like feedback, family income, family occupation, parents qualification, motivation etc. to predict if the student will enroll at the university or not. Depending upon the similarity of the attributes among the students they were grouped into clusters and decisions were made. The objective of the model was to increase the enrollment of the students in the university.

In research conducted by (Jamison (2017)) the yield of college admission was predicted using machine learning techniques. Yield rate can be defined as the rate at which the students who have been granted admission by the university actually enroll for the course. Multiple machine learning algorithms like Random Forest, Logistic Regression and SVMwere used to create the model.

IV. MODEL DESIGN

A. *Independent Variable Importance*

To find the importance range of the independent variables, Random Forest classifier can be used. The higher the value, the more important it is.

Table 2 Independent Variable Importance

Independent Variable	***Rank***
GRE	1.6818745
TOEFL	1.3432065
Uni. Rating	0.4589954
SOP	0.7489156
LOR	0.3707980
CGPA	2.6919350
Research	0.2661699

The results in table 2 show that the most important variable is CGPA as it has the highest ranking among all other variables. And the second highest variable is GRE.

B. *Histogram*

A histogram plot is used to present the frequencies of continuous numbers and to show the distribution of the data selected. Outliers and skewness can be predicted from a histogram along with some other features. Skewness measures how much a graph is asymmetric.

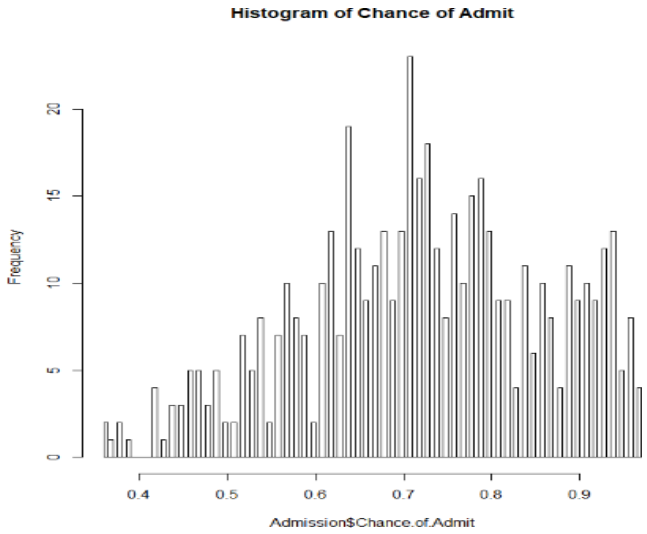

Figure 2 Histogram of Chance of Admit

Figure 2 shows the histogram graph of the dependent variable. Chance of Admit with skewness−0.25 to the left.

The histogram graphs of the most important independent variables are presented also according to the importance test.

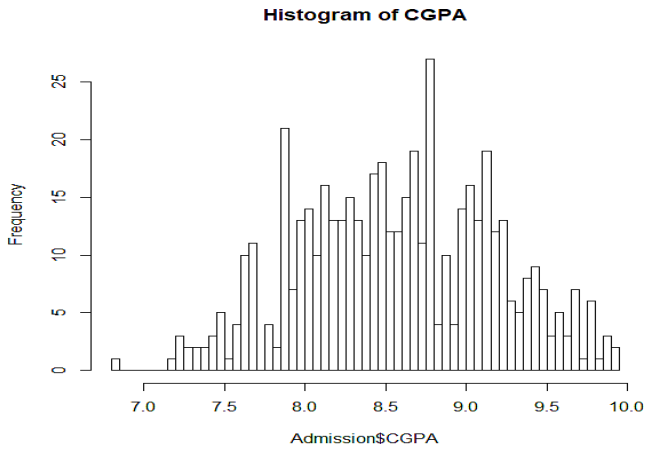

Figure 3 Histogram of CGPA

Figure 3 shows the histogram of CGPA, the most important independent variable, with skewness −0.0283553 to the left.

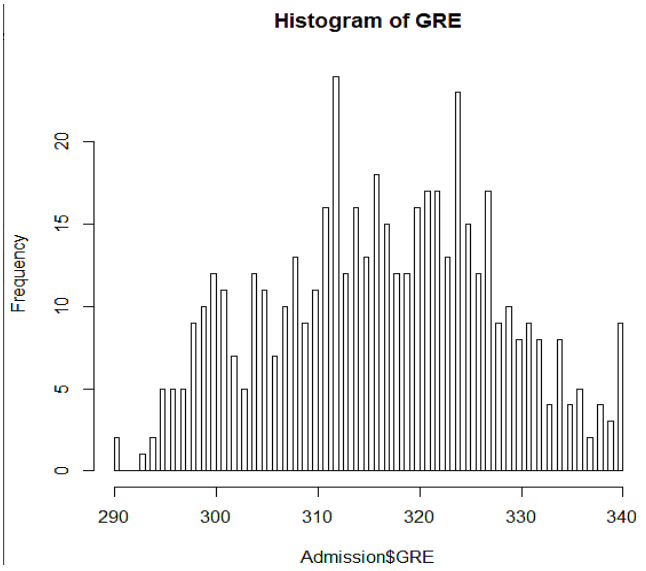

Figure 4 Histogram of GRE

Figure 4 shows the histogram of GRE with skewness−0.04 to the left.

C. Normality Test

Normality test shows whether a parameter is normally distributed or not. Shapiro test is used to perform normality test. If the p-value is greater than 0.05, this means it is normally distributed. Otherwise, the graph is not normally distributed.

Table 3 Shapiro-Wilk NormalityTest

Parameter	***p-value***
Chance of Admit	3.239e-6
CGPA	0.01171
GRE	0.0001245

Table 3 displays the p-value of each parameter obtained from Shapiro-Wilk test; all three parameters' p-value are less than 0.05. Therefore, the null hypothesis is rejected, and variables are not normally distributed.

D. Multicollinearity Issue

Multicollinearity is a huge issue that exists whenever an independent variable is highly correlated with one or more independent variables in a multiple regression equation. If VIF is > 10, high multicollinearity is found. This problem can lead to unstable regression model. In other words, any slight change in the data will lead to a huge change in the coefficients of the multiple linear regression model [16], [17].

In conclusion, there is no multicollinearity problem since all the values are less than 10. This also leads to the fact that our regression model is stable.

E. Linear Regression

According to the linear regression model applied, the equation that represents regression model is:

Regression model= -1.33 +)0.002 * GRE + 0.0026 * TOEFL +0.005* Uni.Rating + 0.004 * SOP +0.013 * LOR + 0.118 * CGPA + 0.023 * Research)

According to Pr(>|t|) value from the linear regression test, all variables have a statistically significant role except for columns 3, 4, which are Uni.Rating and SOP. Also, the R- squared value = 0.83. which means that 83% of variation in our dataset can be explained with our model. The p-value is 2.2e-16, which is way less than 0.05 so we reject the null hypothesis and the model is statistically significant.

V. RESULTS AND DISCUSSIONS

A. *Statistical Test*

According to the normality test, the dependent variable is not normally distributed. Therefore, non parametric test will be performed using PHStat. The test is one-way ANOVA which is performed to determine whether three samples or more have any statistically significant differences between their means or not [18].

The test shows that p-value equals 0.97, which is greater than 0.05, thus, the null hypothesis cannot be rejected, and the tests are not statistically different.

B. *Mean absolute error*

The different regression models are performed on Admission dataset through Weka in order to decide which model performs the best based on mean absolute error (MAE) value. The results are shown in Table 4:

Table 4 Performance Analysis

Regression model	***MAE value***
Multi line arregression	0.0343
Random Forest	0.0363
KNN	0.0544
Multilayer perceptron	0.0337

According to table 4, multilayer perceptron has the smallest MAE equivalent to 3.37% which means that it is the best model.

VI. CONCLUSION

Student admission problem is very important in educational institutions. In this project addresses machine learning models to predict the chance of a student to be admitted. This will assist students to know in advance if they have a chance to get accepted. In this paper, machine learning models were performed to predict the opportunity of a student to get admitted to a master's program. The machine learning models included are multiple linear regression, k-nearest neighbor, random forest, and Multilayer Perceptron. Experiments show that the Multilayer Perceptron model surpasses other models.

As for the future work, more models can be conducted on more datasets to learn the model that gives the best performance.

VII. REFERENCES

1. S. S. Shapiro, M. B. Wilk, and B. T. Laboratories, "An analysis ofvariance test for normality," 1965.
2. G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," *Procedia - Soc. Behav. Sci.*, vol. 106, pp. 234–240, 2013.
3. C. Lopez-Martín, Y. Villuendas-Rey, M. Azzeh, A. Bou Nassif,and S. Banitaan, "Transformed k-nearest neighborhood output distance minimization for predicting the defect density of software projects,"*J. Syst. Softw.*, vol. 167, p. 110592,Sep. 2020.
4. A. B. Nassif, O. Mahdi, Q. Nasir,M. A. Talib, and M. Azzeh,

   "Machine Learning Classifications of Coronary Artery Disease," in*2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2018, pp. 1–6.
5. N. S. Altman, "An introduction to kernel and nearest-neighbor non parametric regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185,1992.
6. A. B. Nassif, M. Azzeh, L. F. Capretz, and D. Ho, "A comparison between decision trees and decision tree forest models for software development effort estimation," in *2013 3rd International Conference on Communications and Information Technology, ICCIT 2013*,2013, pp. 220–224.
7. T.K. Ho, *Random Decision Forests*. USA: IEEE Computer Society,1995. A. B. Nassif, "Software Size and Effort Estimation from Use Case Diagrams Using Regression and Soft Computing Models,"University of Western Ontario, 2012.
8. D. E. Rumelhart, G. E. Hinton, and R.

J. Williams, “Learning internal representations by error propagation,” *MIT Press. Cambridge, MA*,vol. 1, no. V, pp. 318–362, 1986.

9. M. S. Acharya, A. Armaan, and A. S. Antony, “A comparison of regression models for prediction of graduate admissions,” *ICCIDS 2019 - 2nd Int. Conf. Comput.Intell. Data Sci. Proc.*, pp. 1–5, 2019.
10. N. Chakrabarty, S. Chowdhury, and S. Rana, “A Statistical Approach to Graduate Admissions’ Chance Prediction,” no. March,pp. 145–154, 2020.
11. N. Gupta, A. Sawhney, and D. Roth, “Will i Get in? Modeling the Graduate Admission Process for American Universities,” *IEEE Int.Conf.Data Min. Work. ICDMW*, vol. 0, pp. 631–638,2016.
12. A. Waters and R. Miikkulainen, “GRADE : Graduate Admissions,”pp. 64–75,2014.
13. S.Sujay, “Supervised MachineLearning Modelling & Analysis for Graduate Admission Prediction,” vol. 7, no. 4, pp. 5–7, 2020.
14. Janani P, Hema Priya V, Monisha Priya S, Prediction of MS Graduate Admissions using Decision Tree Algorithm ,International Journal of Science and Research (IJSR) ISSN: 2319-7064 ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426.
15. Chithra Apoorva D A, Malepati Chandu Nath, Peta Rohith, Bindu Shree.S, Swaroop.S modelling the Prediction for University Admission using Machine Learning. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-6 March 2020.
16. D.E. Farrar and R. R. Glauber, “Multicollinearity in regression analysis; the problem revisited,” no. 1, pp. 5–7, 2003.
17. R.M. O’Brien, “A caution regarding rules of thumb for variance inflation factors,” *Qual. Quant.*, vol. 41, no. 5, 2007.
18. E.Ostertagová and O.Ostertag, “Methodology and Application of Oneway ANOVA,” *Am. J. Mech. Eng.*, vol. 1, no. 7, pp. 256–261,2013.