DOI: 10.46300/91013.2020.14.13 Volume 14, 2020

Graduate Admission Prediction Using Machine Learning

Sara Aljasmi
Department of Computer Engineering
University of Sharjah
Sharjah, UAE
u16103862@sharjah.ac.ae

Ali Bou Nassif
Department of Computer Engineering
University of Sharjah
Sharjah, UAE
anassif@sharjah.ac.ae

Ismail Shahin
Department of Electrical Engineering
University of Sharjah
Sharjah, UAE
ismail@sharjah.ac.ae

Ashraf Elnagar
Department of Computer Science
University of Sharjah
Sharjah, UAE
ashraf@sharjah.ac.ae

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. 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, Multiple linear regression, Random forest, Student admission

I. INTRODUCTION

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. As a result, the number of students applying for graduate studies has increased in the last decade [1]–[4]. 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.

The dataset [5] presented in this paper is related to educational domain. Admission is a dataset with 500 rows that contains 7 different independent variables which are:

- 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 score will 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 assistant with 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.

It is worth to mention that all tests will be done using R language. Models will be created using Weka, and statistical test will be performed using PHStat.

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 hypothesis. 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 [6].

B. Multiple Linear Regression

Multiple linear regression is a statistical technique used to predict a dependent variable according to two or more

[1] Graduate Record Examination www.ets.org/gre

[2] Test Of English as Foreign Language www.ets.org/toefl

independent variables. As well as, present a linear relationship between them and fit them in a linear equation. The format of the linear equation is as following [7]:

$$y_i=\beta_0+\beta_1 x i_1+...+\beta_n x_n+\epsilon \qquad (1)$$

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 [8], [9]. It is based on the theory of similarity measuring. 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 k labels. While in classification problems, the mode of k labels will be returned [10].

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 [11]. This algorithm creates forests within number of decision reads. Therefore, the more data is available the more accurate and robust results will be provided [12].

Random Forest method can handle large datasets with higher dimensionality without overfitting the model. In addition, it can handle the missing values and maintains accuracy of missing data [12].

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 [13]. 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 [14].

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 [14].

III. RELATED WORK

A great number of researches and studies have been done on graduation admission datasets using different types of machine learning algorithms. One impressive work by Acharya et al. [15] 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.

In addition, Chakrabarty et al. [16] 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. [17] developed a model that 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 [18] 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 [19] applied linear regression to predict the chance of admitting graduate students in master's programs as a percentage. However, no more models were performed.

IV. DATASET PROCESSING AND FEATURE SELECTION

A. Correlated variables

The dataset contained an independent variable to present the serial number of the requests. According to my expertise, it does not correlate to the dependent variable; hence, it was removed from the dataset. It is good to mention that tests showed that there are no missing values in any row of the database.

B. Outliers

Outliers are data values that differ greatly from the majority of a set of data. To find the outliers, there are many methods that can be used, such as: scatterplot and boxplot. In this paper outliers will be investigated using boxplot method.

As for the boxplot, the middle part of the plot represents the first and third quartiles. The line near the middle of the box represents the median. The whiskers on either side of the IQR represent the lowest and highest quartiles of the data. The ends of the whiskers represent the maximum and minimum of the data, and the individual circles beyond the whiskers represent outliers in the dataset [20].

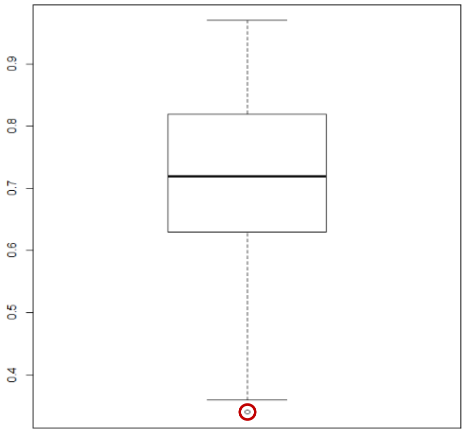

Figure 1 Boxplot of Chance of Admit

As noticed from Figure 1, there are outliers beneath the boundaries of the boxplot.

There are many different methods to deal with outliers, such as:

- Remove the case
- Assign the next value nearer to the median in place of the outlier value
- Calculate the mean of the remaining values without the outlier and assign that to the outlier case

In our case, there are few outliers; therefore, the best option will be removing the rows that contain outliers.

C. *Dataset Devision*

After cleaning the dataset, it was divided randomly into two parts using holdout method. The first part contains 80% of the dataset to present training with 498 observations. The second part contains 20% of the dataset to present testing with 98 observations.

D. *Feature Selection*

In order to perform feature selection, ols_step_best_subset() function has been applied; which will display all possible subsets. Then according to certain criteria, the best subset will be selected. Since there are 7 independent variables, the number of subsets to be tested is 2^7, which equals 128 subsets. To apply feature selection, linear regression equation should be applied. Note that linear regression can be performed only with numeric independent variables.

It is observed that all variables are numeric. It is good to mention that in case of having categorical variable, as.numeric() function can be used to convert data variables to numeric. Next, regression model is created, and feature selection is performed.

The best model in feature selection is presented by the model with either highest R-square or smallest MSEP, which belongs to model 6 which includes all columns except column 4 that presents SOP.

E. *Descriptive Summary*

Descriptive summary provides numerical measures of some important features which describe a dataset. Some features are presented in table 1:

Table 1 Descriptive Summary

Feature	***Value***
Minimum Value	0.3600
First Quartile (Q1)	0.6325
Median	0.7200
Mean	0.7233
Third Quartile (Q3)	0.8200
Maximum Value	0.9700

Note that the first and third quartile are found using the following equations where *N* is equal number of data points.

First Quartile (Q1)=(N+1) × 0.25 (2)

Third Quartile (Q3)=(N+1) × 0.75 (3)

V. 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.

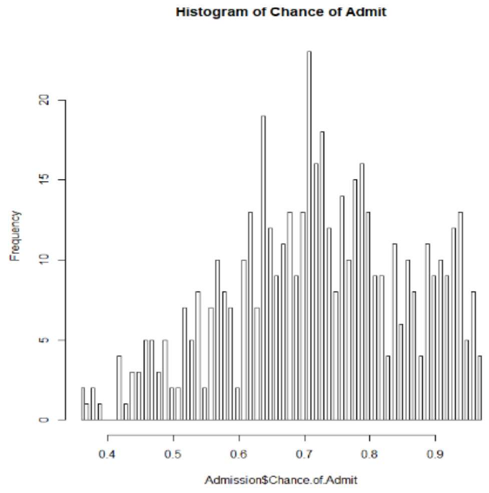

Figure 2 Histogram of Chance of Admit

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.

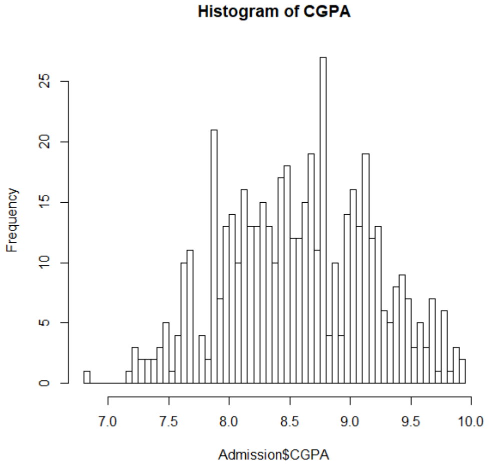

Figure 3 Histogram of CGPA

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

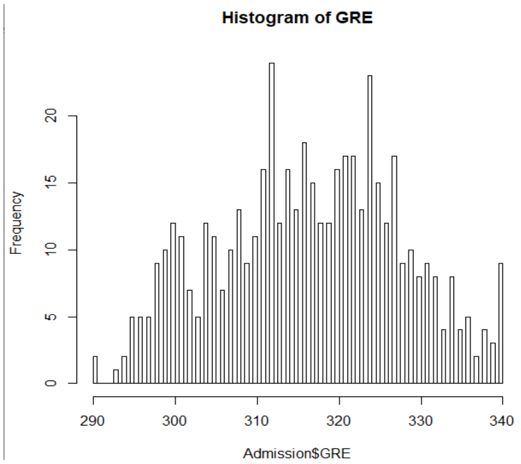

Figure 4 Histogram of GRE

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 Normality Test

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 [21], [22].

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.

VI. Results and Discussions

A. Statistical Test

According to the normality test, the dependent variable is not normally distributed. Therefore, nonparametric 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 [23].

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 linear regression	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.

CONCLUSION

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.

REFERENCES

[1] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-split Optimized Bagging Ensemble Model Selection for Multi-class Educational Data Mining," *Appl. Intell.*, vol. 50, pp. 4506–4528, 2020.

[2] F. Salo, M. Injadat, A. Moubayed, A. B. Nassif, and A. Essex, "Clustering Enabled Classification using Ensemble Feature Selection for Intrusion Detection," in *2019 International Conference on Computing, Networking and Communications (ICNC)*, 2019, pp. 276–281.

[3] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic ensemble model selection approach for educational data mining," *Knowledge-Based Syst.*, vol. 200, p. 105992, Jul. 2020.

[4] A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, "E-Learning: Challenges and Research Opportunities Using Machine Learning Data Analytics," *IEEE Access*, 2018.

[5] M. S. Acharya, A. Armaan, and A. S. Antony, "A Comparison of Regression Models for Prediction of Graduate Admissions," *Kaggle*, 2018. .

[6] S. S. Shapiro, M. B. Wilk, and B. T. Laboratories, "An analysis of variance test for normality," 1965.

[7] G. K. Uyanık and N. Güler, "A Study on Multiple Linear Regression Analysis," *Procedia - Soc. Behav. Sci.*, vol. 106, pp. 234–240, 2013.

[8] C. López-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.

[9] 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.

[10] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992.

[11] 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.

[12] T. K. Ho, *Random Decision Forests*. USA: IEEE Computer Society, 1995.

[13] A. B. Nassif, "Software Size and Effort Estimation from Use Case Diagrams Using Regression and Soft Computing Models," University of Western Ontario, 2012.

[14] 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.

[15] 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.

[16] N. Chakrabarty, S. Chowdhury, and S. Rana, "A Statistical Approach to Graduate Admissions' Chance Prediction," no. March, pp. 145–154, 2020.

[17] 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.

[18] A. Waters and R. Miikkulainen, "GRADE : Graduate Admissions," pp. 64–75, 2014.

[19] S. Sujay, "Supervised Machine Learning Modelling & Analysis for Graduate Admission Prediction," vol. 7, no. 4, pp. 5–7, 2020.

[20] G. Singler, "Statistics Reference Series Part 1: Descriptive Statistics," 2018.

[21] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis; the problem revisited," no. 1, pp. 5–7, 2003.

[22] R. M. O'Brien, "A caution regarding rules of thumb for variance inflation factors," *Qual. Quant.*, vol. 41, no. 5, 2007.

[23] E. Ostertagová and O. Ostertag, "Methodology and Application of Oneway ANOVA," *Am. J. Mech. Eng.*, vol. 1, no. 7, pp. 256–261, 2013.