Data Analytics and Artificial Intelligence
Vol: 2(5), 2022
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/data-analytics-and-artificial-intelligence/

University Admit Eligibility Predictor

Archana Janani S,*Jeeva Ganesan T
Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India.
*Corresponding author Email:jeevaganesan3737@gmail.com

Abstract. Students are often worried about their chances of admission to University. The aim of our 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.

1. 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. 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 (GRE) score. The score will be out of 340 points.
- Test of English as a Foreigner Language (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 1
- 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).

2. Technical Background

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 equation is as following

$$Yi=\beta0+\beta1xi1+...+\betanxn +\epsilon$$

Where, for i=n observations: Yi=dependent variable xi= independent variables β0=y-intercept βn=slope coefficients for each independent variable ϵ=the model's error term or residuals

K-Nearest Neighbor

K-nearest neighbor (KNN) is a supervised machine learning algorithm used for classification and regression problems. 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.

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

3. Related Work

A great number of researches and studies have been done on graduation admission datasets using different types of machine learning algorithms.We has compared between different regression algorithms, which are: K-Nearest Neighbor , Multiple linear regression and Random Forest, to predict the chance of admit based on the best model that showed the least MSE which was multilinear regression.

4. Dataset Processing and Feature Selection

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.

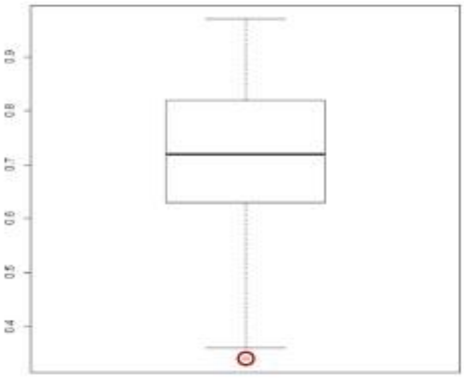

FIGURE 1. Outliers

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.

5. Model Design

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 1. Independent variable

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

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 2. P value

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

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 Rsquared 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.

6. Result

Mean absolute error: The different regression models are performed on Admission dataset through Weka in order to decide whichmodel performs the best based on mean absolute error.

TABLE 3. MAE value

Regression model	***MAE value***
Multi linear regression	0.0343
Random Forest	0.0343
KNN	0.0544

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

7. 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 neighbour, random forest Experiments show that the multiple regression 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.