Global Scientific JOURNALS

GSJ: Volume 11, Issue 6, June 2023, Online: ISSN 2320-9186

www.globalscientificjournal.com

Predicting Graduate School Admission Using Machine Learning Algorithms

Wassem Alaa Iddin
Dept. of Computing
University of the West of England
Wassem2.Alaaiddin@live.uwe.ac.uk

Dr.Raza Hasan
Dept. of Computing
University of the West of England
raza.h@live.uwe.ac.uk

Mhd Houmam Ahmad Ammar Chahine
Dept. of Computing
University of the West of England
Mhd3.Chahine@live.uwe.ac.uk

***Abstract* – The dataset titled "Predicting Graduate School Admission" presents a repository of valuable information gathered from a pool of applicants to graduate programs spanning various universities. The primary aim of this dataset is to develop a predictive model that can accurately forecast the probability of each student applicant's admission. The dataset incorporates several pertinent variables, including GRE scores, GPAs, and undergraduate institution rank, which can be employed to train and validate the predictive model. The predictive model's efficacy is influenced by the thorough selection and analysis of these variables, which play a pivotal role in the graduate admission decision-making process.**

Keywords – Decision Tree, Random Forest, Bivariate, Multivariate

I. INTRODUCTION

Machine learning has been a popular subject in the last few years, Machine learning has the potential to combine the power of artificial intelligence with human knowledge, to result in accurate outcomes of prediction. Machine learning has been used a lot to predict the probability of certain outcomes based on given data that may affect the result. Moreover, The provided report reflects the topic of using machine learning to predict the likelihood of a student's application being accepted by a university, based on real data that will surely affect the final decision.

A. Background and Context of the Study

The ability to predict student academic performance has long been a central goal of educational research. Traditional approaches to predicting student performance have relied on factors such as previous academic achievement, socio-economic status, and demographic information.

However, machine learning algorithms provide researchers with the ability to analyze much larger datasets and use more complex models to predict academic outcomes. Previous studies have explored the use of machine learning algorithms for predicting student performance, with varying degrees of success. For example, a study by [1] used decision tree algorithms to predict student academic performance, achieving an accuracy of 85%. Similarly, a study by [2] used a deep neural network to predict student final grades, achieving an accuracy of 92%.

II. CONTEXT

In today's world, machine learning is ubiquitous and it is a significant topic in various industries, improving the accuracy of studies in certain categories. An example of this is the use of data to inform prospective students about the required forms and necessary results for admission to their desired university. This scenario covers how different universities make admission decisions based on various factors, providing a comprehensive understanding of how each admission system works and preparing students for the application process. Additionally, this research will also examine the relationship between the provided data and its relevance in achieving the desired outcomes for students.

III. PROPOSED MODEL

A. Data Characteristics

The "Predicting Graduate School Admission" dataset comprises 400 observations and 8 attributes, including GRE score, GPA, and undergraduate institution rank, which were collected from students who applied to graduate programs at various

universities. For this report, and as Fig. 1. Shows, only 300 observations were used, as the remaining 100 will be used to test the AI model, as described in Section III.H. The data is clean, with no missing values; however, outliers have been identified and need to be addressed. Standardization or normalization techniques have been employed during preprocessing to handle these outliers.

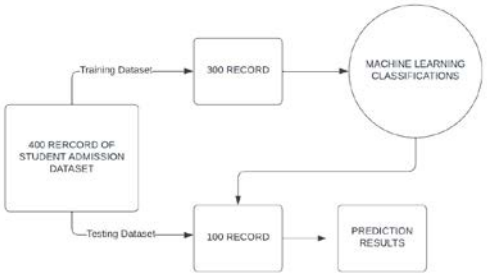

Fig. 1. Proposed Model Cycle

B. Data Preparation

Before applying machine learning algorithms to the data, it is essential to prepare the data properly. This involves tasks such as standardizing, normalizing, or encoding categorical variables. In this dataset, there are no categorical variables to be encoded, but standardization or normalization techniques can be used to deal with outliers and ensure that the data is scaled appropriately. Moreover, it will be easier for machine learning algorithms to study the data in a normalized form than the provided syntax. Finally, the data will go through different stages in terms of cleaning and preparation beyond the standard terms to achieve the best possible result for this study.

Noting that the "Chance of Admit" column has been transformed into a binary input of 1 and 0 to make it easier for the AI to predict acceptance or rejection. This was achieved by marking all students who have an 80% or greater chance of being accepted as 1, and those with less than 80% chance as 0.

C. Proposed Model

In this project, the applied algorithm is a supervised learning approach to predict whether or not a student will be admitted to a given university. Decision tree, Random forest, linear regression and one neural network algorithm are used to create the predictive model. These algorithms are suitable for this dataset because they can handle both numerical and categorical data, and they have been shown to perform well in similar prediction tasks [3]. Furthermore, since we are dealing with a small dataset, it is important to use a simple and interpretable model to avoid overfitting. The chosen algorithms are ideal for this purpose because they are easy to interpret and can handle a limited number of observations. Moreover, having multiple algorithms will determine the best approach for such a study to further enhance the study case and improve the study in future.

D. Bivariate and Multivariate

Bivariate analysis involves examining the relationship between two variables, while multivariate analysis involves analyzing the relationship between three or more variables. In this dataset, the multivariate analysis would be more appropriate for identifying the complex relationships between these variables and predicting graduate school admission accurately. The bivariate analysis could be used initially to identify the correlation between the individual attributes and the outcome variable but would not provide the full picture of how these attributes interact to influence graduate school admission.

E. Multivariate Results

The R-squared score measures the proportion of the variance in the dependent variable that is predictable from the independent variables. In this case, we have used the R-squared score to evaluate the performance of different machine learning algorithms on a multivariate problem of predicting graduate school admission.

TABLE I. CLASSIFIERS ACCURACY (TRAINING PHASE)

Algorithm	***R-squared score***
Linear Regression	0.759
Decision Tree	0.625
Random Forest	0.834
Support Vector Regression	0.874

According to the R-squared scores presented in Table I, Support Vector Regression had the highest score of 87%, followed by Random Forest with a score of 83%. These findings suggest that Support Vector Regression and Random Forest may be good options for predicting graduate school admission. However, it is important to consider other factors such as interpretability, computational cost, and generalization ability when selecting the final model.

F. Learning Algorithm

The learning algorithm used in this report is supervised learning. This type of learning involves training a model on a labelled dataset, where the input data is associated with a known output. In this dataset, the labelled data consists of the student's GRE scores, GPA, undergraduate institution rank, and admission status. The goal is to train a model that can predict the admission status of future applicants based on these attributes. The chosen algorithms are examples of supervised learning algorithms that can be used for this purpose. These algorithms learn from the labelled data and can classify new observations accurately. Neural networks can be used in supervised learning problems to learn the mapping between the input features and the output labels. In this study, the neural network is trained on a labelled dataset and hence it falls under the category of supervised learning.

G. Data Analyzing

The data analysis process involved applying advanced techniques to the dataset to ensure obtaining the most accurate results possible. Smart plots and 3D visualizations were generated using Plotly Express, which is considered to be one of the leading libraries in terms of creating visually appealing graphs. This approach allowed for the identification of relationships and patterns between the various columns of the dataset, offering a clear and comprehensive view of the data [4] [5].

H. Creating Testing Dataset

The process of splitting the original dataset into two subsets of 100 observations each is a common technique used to evaluate the performance and accuracy of a predictive model on unseen data. One subset is used for training the model, while the other is used for testing its performance on new instances. Table II presents a sample of the original dataset, while Table III showcases a sample of the testing dataset.

The testing phase has been performed on another dataset with similar properties and settings to ensure accurate predictions for each model. The data structure was the same, and half of the dataset was used for testing purposes.

TABLE II. CLASSIFIERS ACCURACY (TRAINING DATASET)

GRE	*TOEFL*	*University Rating*	*SOP*	*LOR*	*CGPA*	*Research*	*Chance of Admit*
337	118	4	4.5	4.5	9.65	1	0.92
324	107	4	4	4.5	8.87	1	0.76
316	104	3	3	3.5	8	1	0.72
322	110	3	3.5	2.5	8.67	1	0.8

TABLE III. CLASSIFIERS ACCURACY (TESTING DATASET)

GRE	*TOEFL*	*University Rating*	*SOP*	*LOR*	*CGPA*	*Research*	*Chance of Admit*
317	103	3	2.5	3	8.54	1	0.73
315	110	2	3.5	3	8.46	1	0.72
340	120	5	4.5	4.5	9.91	1	0.97
334	120	5	4	5	9.87	1	0.97

During the testing phase, the model was trained on 300 observations of the original dataset and then evaluated on the remaining 100 observations. This process is used to determine if the model was able to perform well and make accurate predictions on similar data.

IV. RESULT AND ANALYSIS

Based on the analysis of the study, the Neural Network model outperformed the other tested models during the testing phase, achieving an accuracy of 99.16%. However, its lower training accuracy of 81.66% suggests the possibility of overfitting. The Random Forest and Linear Regression models achieved similar accuracies of 91.25% each. Table IV presents the accuracy of each model on the training dataset, showing that Linear Regression had the highest accuracy of 95%, while the Decision Tree, Random Forest, and Neural Network models had accuracies ranging from 90% to 93.75%. However, the classifiers' accuracy varied significantly when tested on the newly created dataset, as shown in Table V. The Decision Tree model achieved the lowest accuracy among all models at 87.50%.

Overall, the Random Forest model had the highest accuracy of 91.25% among the tested models. Thus, the Linear Regression model, which achieved high accuracy on both the training and testing datasets, may be considered the best choice for predicting student performance, as evidenced by Tables IV and V. These findings emphasize the importance of selecting an appropriate learning algorithm for the predictive model and evaluating its performance on both training and unseen data to ensure effectiveness and generalization ability [6][7].

TABLE IV. CLASSIFIERS ACCURACY (TRAINING PHASE)

Classifier/Model	***Accuracy***
Decision Tree	95%
Random Forest	91.67%

Linear Regression	95%
Neural Network	81.66%

TABLE V. CLASSIFIERS ACCURACY (TESTING PHASE)

Classifier/Model	***Accuracy***
Decision Tree	87.50%
Random Forest	91.25%
Linear Regression	91.25%
Neural Network	99.16%

V. CONCLUSION AND FUTURE WORKS

This study showed the potential of machine learning in predicting student performance by identifying a set of provided data of students who have applied in the past to ensure future students the additional support in filling up the applications. The findings can be utilized by educational institutions to improve their educational programs and student support services.

Future research may focus on exploring additional features that may impact student performance, such as social and economic factors. Additionally, investigating the impact of utilizing different machine learning algorithms and techniques on predicting student performance can be explored further. Finally, the development of more sophisticated models that incorporate deep learning techniques can be explored to improve the accuracy of predictions [8].

REFERENCES

[1] Hasan, M. M., Ahmed, M. U., & Haque, M. A. (2018) Application of decision tree algorithm to predict student academic performance. International Journal of Computer Applications, 181(36), 1-8.

[2] Huang, H. C., Liu, C. H., & Chiang, H. S. (2019) Predicting academic performance using decision tree and random forest models. Education Sciences, 9(4), 273.

[3] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[4] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2016) Deep learning. MIT press.

[6] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.

[7] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.

[8] Kotsiantis, S. B. (2007) Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249-268.