

# **A Comparison of Classification Models in Predicting Graduate Admission Decision**

**Adrita Iman**  
**Purdue University Fort Wayne**

**Xiaoguang Tian**  
**Purdue University Fort Wayne**

*While the decision process of graduate admissions remains elusive, specific criteria are decidedly significant in determining admission outcome. Prospective students applying to graduate programs experience a real predicament of selecting the right schools to invest limited resources for the application. This paper presents comparisons of various machine learning classification models, including Naïve Bayes, Logistic Regression, Multilayer Perceptron and Decisions Tree models, in predicting the admission outcome of candidates with a set of known parameters using a dataset of 400 applicant records. By comparing the performance metrics of these methods, the study finds Naïve Bayes to be the most accurate model for this type of dataset. Predictive models such as the ones discussed in this paper can be a valuable tool for prospective students in shortlisting universities in their application process. The study also proposes a framework that incorporates machine learning-based classification into the admissions decision process. Implementing such methods may help support graduate admissions committees in streamlining large pools of applications or observing and understanding trends in their past admission decisions.*

*Keywords: graduate admissions, data mining, predictive analysis, machine learning, supervised learning, cross-validation*

## **INTRODUCTION**

In recent years, there has been increased consideration of using Artificial Intelligence and Machine Learning-based predictive models for gauging student performance and evaluation. Universities are being urged to adopt data-driven, evidence-based approaches to improve student retention, reduce drop-out rates, and improve the quality of education by removing bias from grading and evaluation of students (Nuutila et al., 2018; Gabriele et al., 2016). Many researchers have addressed this call for implementing predictive models capable of identifying key performance determinants of students (Barik et al., 2020; Deo et al., 2020; Fernández-García et al., 2020). There seems to be a need to improve overall transparency in student evaluation to benefit both students and educational institutions.

Applying to graduate programs is a crucial yet tedious process for prospective students. There is immense pressure to build a robust and impressive profile that appeals to admission committees, which involves taking expensive standardized tests and writing a solid personal statement, among various other factors. Students struggle to select universities that are the best fit for them based on their career goals while

also maximizing their chances of admission using limited financial resources. Similarly, the admissions decision process demands much time and effort from admissions committees. Screening vast numbers of applications to assess the candidates' readiness for the program and potential success is not a simple task. With ever-increasing competition for graduate school admissions, the process needs to be meticulous, efficient, and adaptive all at the same time.

The main objective of this study is to compare the performance of various supervised machine learning classification models in predicting admission decision based on several predictors, such as test scores, GPA, research and quality of admission essays, commonly considered by universities. This research paper also proposes a machine learning-based decision framework for higher education institutes using machine learning approaches.

## RELATED WORKS

Existing research applying data mining and machine learning predictive capabilities to explore university admissions and student performance suggest that most universities base their admission decisions largely on the academic profiles of applicants. Adekitan and Noma-Osaghae (2019) found this to be true for a university in Nigeria. They argued that the results of their study indicate a need for re-evaluation of this focus on academics and pursue other factors that may be better indicators of student success in higher education, such as sports, leadership experiences, extracurricular projects, etc. On the other hand, Patterson et al. (2012) found that the GPA of first-year college students could be predicted using their SAT scores, corroborating the validity of using standardized test scores in gauging the potential for academic success. Further, Fu (2012) found high school GPA and test scores to be correlated with the academic success of first-year college students. Eberly et al. (2013) showed that university first-years with less than 2.55 high school GPA consistently experienced low academic success. These studies highlight the importance of the academic focus in admissions criteria since universities want to admit students with higher likelihood of success in their educational journey and achieve higher graduation rates.

Few studies have addressed the admission criteria and academic success of graduate applicants from a predictive analysis standpoint to authors' knowledge. An exploration of internet blogs and forum discussions on this topic suggest that admitted graduate students consistently tend to have high CGPA and high standardized test scores. Acharya et al. (2019) compared regression models for predicting graduate admission using an earlier version of the same dataset. They found that high test scores and CGPA resulted in greater chances of admission. Muratov et al. (2017) looked at multivariate predictive models to forecast academic performance in the Doctor of Pharmacy schools. They found that undergraduate GPA and PCAT (standardized test required for pharmacy school) scores were strong determinants of selecting candidates with high success potential in the rigorous PharmD programs.

Arguably, other factors also play significant roles in the admission decision. Universities frequently emphasize the importance of the statement of purpose (SOP) for undergraduate admission and graduate admission. In particular, the SOP is often mentioned as the essential piece in a graduate school application packet because it gives the reviewers an understanding of the candidate's goals and interests in their own words. The SOP also allows them to determine how well the candidate matches their unique program, gauge the candidate's motivation to pursue higher education and their passion for success (Cornell, 2020).

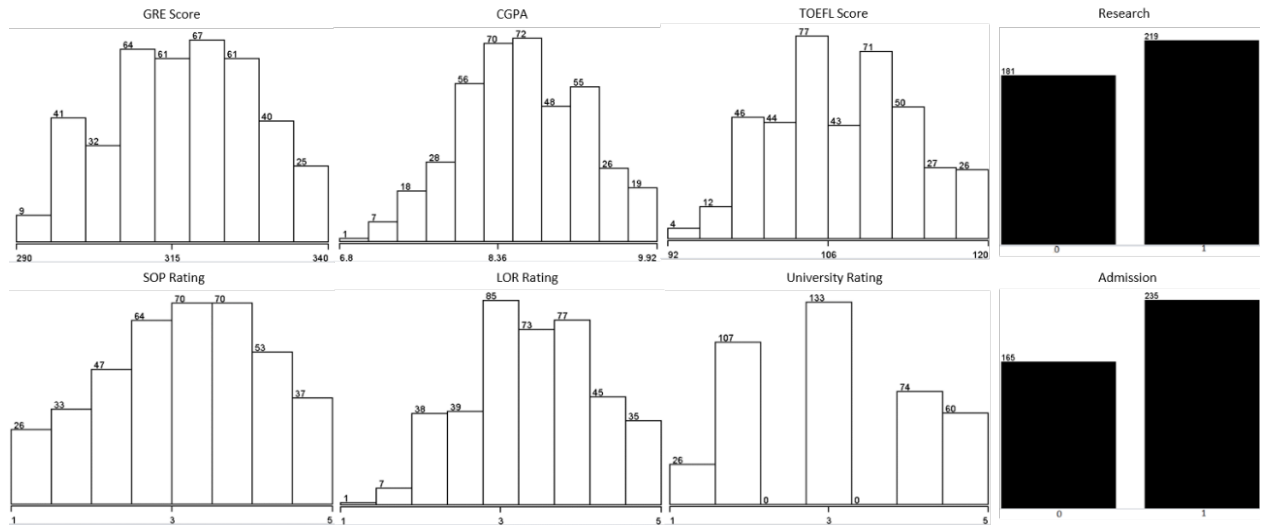
Researchers and institutions are increasingly considering using predictive capabilities of machine learning and artificial intelligence in education. Such methods have been suggested for creating student grading models that remove human bias, predicting student performance for early intervention and assistance, improving student retention in programs, predicting college completion or graduation and maximizing enrollment in the admissions process (Kucak et al., 2018; Zhu, 2015; Slim et al., 2018; Eberly et al. 2013; Chang, 2006). This paper explores the predictive power of seven supervised machine learning models in using parameters available in a dataset of 400 records of applicant profile information to predict admission outcome. The objective of the study is to determine which classification model performs the best as a predictive model. The study also addresses the gap in the literature of exploring admissions criteria in

graduate applicants by further examining the data to provide insights into what attributes carry more weight in forming the admission decision.

## DATASET

The study uses an open-source dataset from Kaggle, an online community of data scientists and machine learning practitioners (Acharya et al., 2019). This dataset contains 400 records of candidate information, including several parameters that graduate admissions committees give consideration. The parameters included are: GRE and TOEFL scores, ungraduated CGPA (scale of 10), a rating for the strength of Statement of Purpose (scale of 5), a rating for the strength of Letters of Recommendation (scale of 5), a binary measure of whether the student has research experience (0 = no, 1 = yes), rating of University applied to (scale of 5) and a binary measure of the admission outcome (0 = not admitted, 1 = admitted). Note that this dataset was collected mainly from Indian students. International applicants to US universities are required to provide language proficiency tests such as the TOEFL exam. Descriptive analyses of the dataset are shown in Figure 1, and Table 1, and Table 2.

**FIGURE 1  
OVERALL DISTRIBUTION OF EACH ATTRIBUTE IN THE DATASET**



**TABLE 1  
GENERAL STATISTICS OF CONTINUOUS VARIABLES**

|                    | GRE Score | CGPA  | TOEFL Score | SOP   | LOR   | University Rating |
|--------------------|-----------|-------|-------------|-------|-------|-------------------|
| Maximum            | 290       | 6.8   | 92          | 1     | 1     | 1                 |
| Minimum            | 340       | 9.92  | 120         | 5     | 5     | 5                 |
| Mean               | 316.808   | 8.599 | 107.41      | 3.4   | 3.453 | 3.087             |
| Standard Deviation | 11.474    | 0.596 | 6.07        | 1.007 | 0.898 | 1.144             |

**TABLE 2**  
**COUNTS OF CATEGORICAL DATA**

|   | Research | Admission |
|---|----------|-----------|
| 0 | 181      | 165       |
| 1 | 219      | 235       |

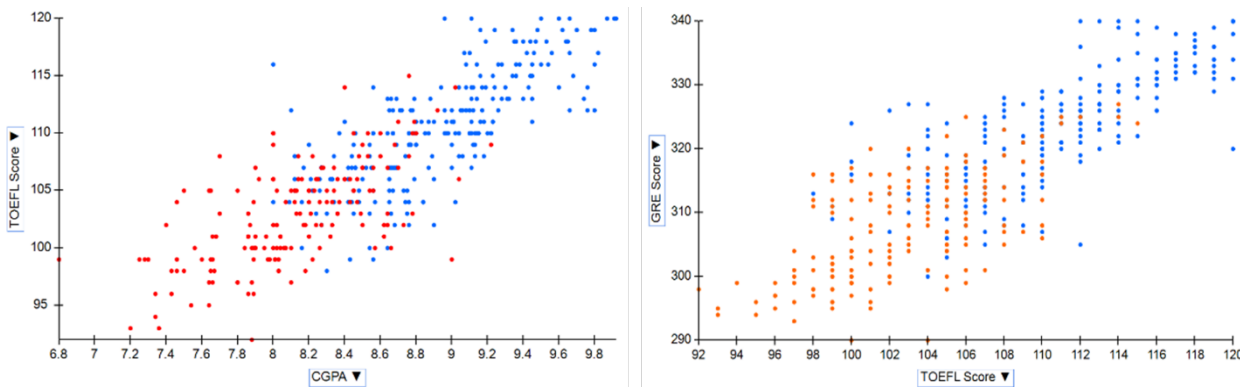
Exploring the dataset further using data mining tools on Microsoft Excel, the research found high levels of correlation between specific parameters. This underscores the importance of particular parameters prevalently accepted as strong determinants of admission decisions.

**TABLE 3**  
**CORRELATION BETWEEN PARAMETERS IN THE DATASET**

|                   | GRE Score | TOEFL Score | University Rating | SOP   | LOR   | CGPA  | Research | Admission |
|-------------------|-----------|-------------|-------------------|-------|-------|-------|----------|-----------|
| GRE Score         | 1.000     |             |                   |       |       |       |          |           |
| TOEFL Score       | 0.836     | 1.000       |                   |       |       |       |          |           |
| University Rating | 0.669     | 0.696       | 1.000             |       |       |       |          |           |
| SOP               | 0.613     | 0.658       | 0.735             | 1.000 |       |       |          |           |
| LOR               | 0.558     | 0.568       | 0.660             | 0.730 | 1.000 |       |          |           |
| CGPA              | 0.833     | 0.828       | 0.746             | 0.718 | 0.670 | 1.000 |          |           |
| Research          | 0.580     | 0.490       | 0.448             | 0.444 | 0.397 | 0.522 | 1.000    |           |
| Admission         | 0.666     | 0.625       | 0.575             | 0.543 | 0.533 | 0.683 | 0.503    | 1.000     |

(Correlation > 0.7 are highlighted)

**FIGURE 2**  
**SCATTERPLOTS COMPARING ADMISSION OUTCOMES BASED ON (LEFT) GRE SCORE V. TOEFL SCORE, AND (RIGHT) TOEFL SCORE V. CGPA. BLUE = ADMISSION CLASS 1 (ADMITTED); RED = ADMISSION CLASS 0 (NOT ADMITTED)**



The scatterplots in Figure 2 depict a relationship between GRE scores, TOEFL Scores and CGPA with admission outcome. It is clear from these figures that higher values for these parameters are strongly related to positive admission outcome. Additionally, a PCA (principal components analysis) of the dataset reveals that the CGPA parameter alone explains 69.55% of the dataset. The study also ran a feature selection of best subsets using XLminer to determine the most important predictors of the output variable (admission)

through Welch's Test and F-Test of the top-ranked features. These results indicate that CGPA and TOEFL Scores (for international applicants) may be the most critical consideration in a graduate program's admissions decision.

**TABLE 4**  
**FEATURE SELECTION USING WELCH'S TEST AND F-TEST ON THE DATASET**

| <b>RANK</b> | <b>VARIABLE</b> | <b>WELCH: P-VALUE</b> | <b>F: P-VALUE</b> |
|-------------|-----------------|-----------------------|-------------------|
| <b>1</b>    | CGPA            | 1.85952E-55           | 3.6E-56           |
| <b>2</b>    | TOEFL Score     | 1.58814E-45           | 1.14E-44          |

While such exploratory analysis helps explore the dataset to visualize patterns and gauge the importance of specific parameters, they cannot provide an informed prediction of the admission outcome. Machine learning models can be employed for predictive purposes.

## **METHODOLOGY**

The dataset is analyzed using seven classification models, namely Naïve Bayes, Multilayer Perceptron, Logistic Regression, Random Forest, REP Tree, Random Tree, and J48 on the Weka software. Each method is tested using five sampling methods, and metrics evaluating the model's performance in each case are summarized.

### **Naïve Bayes**

Naïve Bayes is a simple classification technique based on Bayes' theorem. The model converts the dataset into a frequency table and calculates probabilities (Hand et al., 2001). Then, it calculates the posterior probability for each class using the Bayesian equation, and the class with the highest posterior probability is the outcome of the prediction. Naïve Bayes is considered one of the most efficient inductive machine learning algorithm (López-Úbeda et al., 2020).

**TABLE 5**  
**EVALUATION METRICS OF NAÏVE BAYES CLASSIFICATION MODEL**

| <b>Sampling Method</b>             | <b>Accuracy</b> | <b>Precision</b> | <b>Recall</b> | <b>F-measure</b> | <b>ROC Area</b> |
|------------------------------------|-----------------|------------------|---------------|------------------|-----------------|
| <b>10-fold cross-validation</b>    | 84.5            | 0.851            | 0.845         | 0.846            | 0.926           |
| <b>5-fold cross-validation</b>     | 84.75           | 0.854            | 0.848         | 0.848            | 0.926           |
| <b>66% Training/34% Validation</b> | 86.0294         | 0.863            | 0.86          | 0.861            | 0.934           |
| <b>80% Training/20% Validation</b> | 88.75           | 0.888            | 0.888         | 0.888            | 0.946           |
| <b>50% Training/50% Validation</b> | 84              | 0.844            | 0.84          | 0.84             | 0.916           |

### **Multilayer Perceptron**

Multilayer Perception (MLP) is a type of feedforward artificial neural network (ANN) that refers to networks composed of multiple layers of the perceptron (Simon, 1999). It consists of at least three layers: input, hidden, output. Apart from the input nodes, each of the other nodes is a neuron that uses a nonlinear activation function. MLP can distinguish data that is not linearly separable. The model passes data through the input nodes, multiplies with weights and computes a sum. Then, it adds a bias factor and feeds the sum through the activation function.

**TABLE 6**  
**EVALUATION METRICS OF MLP CLASSIFICATION MODEL**

| Sampling Method                    | Accuracy | Precision | Recall | F-measure | ROC Area |
|------------------------------------|----------|-----------|--------|-----------|----------|
| <b>10-fold cross-validation</b>    | 81.75    | 0.818     | 0.818  | 0.818     | 0.889    |
| <b>5-fold cross-validation</b>     | 83.5     | 0.836     | 0.835  | 0.835     | 0.899    |
| <b>66% Training/34% Validation</b> | 84.5588  | 0.847     | 0.846  | 0.846     | 0.92     |
| <b>80% Training/20% Validation</b> | 85       | 0.852     | 0.85   | 0.85      | 0.929    |
| <b>50% Training/50% Validation</b> | 83       | 0.83      | 0.83   | 0.83      | 0.905    |

### Logistic Regression

Logistic regression, or logit model, extends the idea of linear regression to predict categorical output variables (Strano and Colosimo, 2006). This statistical model uses a logistic function to model the binary dependent variable, derived from the function of probability and odds. [Logit =  $\log(\text{odds}_{\text{class}})$ ;  $\text{odds}_{\text{class}} = \text{probability}_{\text{class}} / (1 - \text{probability}_{\text{class}})$ ].

**TABLE 7**  
**EVALUATION METRICS OF LOGISTIC REGRESSION CLASSIFICATION MODEL**

| Sampling Method                    | Accuracy | Precision | Recall | F-measure | ROC Area |
|------------------------------------|----------|-----------|--------|-----------|----------|
| <b>10-fold cross-validation</b>    | 84.25    | 0.843     | 0.843  | 0.843     | 0.926    |
| <b>5-fold cross-validation</b>     | 84       | 0.841     | 0.84   | 0.84      | 0.924    |
| <b>66% Training/34% Validation</b> | 86.0294  | 0.861     | 0.86   | 0.86      | 0.937    |
| <b>80% Training/20% Validation</b> | 88.75    | 0.888     | 0.888  | 0.888     | 0.954    |
| <b>50% Training/50% Validation</b> | 82.5     | 0.825     | 0.825  | 0.825     | 0.919    |

### Random Forest

Random forest is an ensemble learning method for classification (Ho, 2002). It constructs multiple decision trees with the training data and outputs the class that occurs most frequently. The decision at each node is selected at random instead of utilizing a systematic approach (Raghu and Sriraam, 2018). Random forest usually outperforms the random tree because it uses averages to improve predictive accuracy and control over-fitting.

**TABLE 8**  
**EVALUATION METRICS OF RANDOM FOREST CLASSIFICATION MODEL**

| Sampling Method                    | Accuracy | Precision | Recall | F-measure | ROC Area |
|------------------------------------|----------|-----------|--------|-----------|----------|
| <b>10-fold cross-validation</b>    | 81       | 0.809     | 0.81   | 0.809     | 0.905    |
| <b>5-fold cross-validation</b>     | 81.5     | 0.814     | 0.815  | 0.814     | 0.899    |
| <b>66% Training/34% Validation</b> | 83.0882  | 0.831     | 0.831  | 0.831     | 0.919    |
| <b>80% Training/20% Validation</b> | 85       | 0.85      | 0.85   | 0.85      | 0.93     |
| <b>50% Training/50% Validation</b> | 80       | 0.8       | 0.8    | 0.799     | 0.905    |

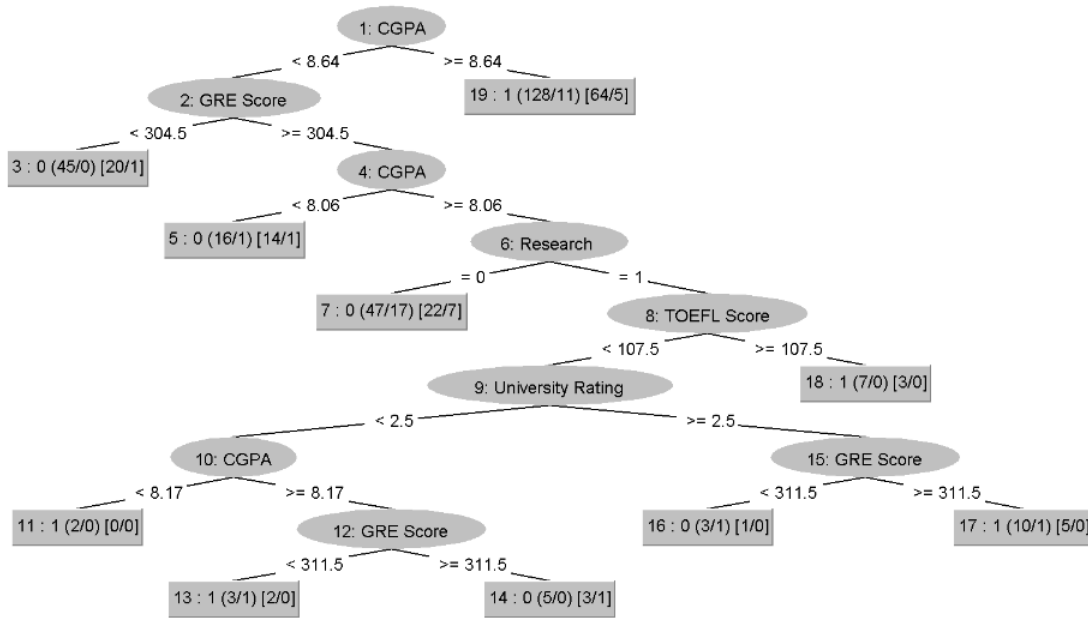
### REP Tree

The REP tree algorithm is a fast decision tree learner programmed into Weka (Raviya and Gajjar, 2012). It is based on the C4.5 algorithm and can produce discrete outcome trees. The model builds a decision tree using information gain or variance and prunes it using reduced-error pruning with back-fitting.

**TABLE 9**  
**EVALUATION METRICS OF REP TREE CLASSIFICATION MODEL**

| Sampling Method                    | Accuracy | Precision | Recall | F-measure | ROC Area |
|------------------------------------|----------|-----------|--------|-----------|----------|
| <b>10-fold cross-validation</b>    | 79.25    | 0.793     | 0.793  | 0.793     | 0.851    |
| <b>5-fold cross-validation</b>     | 80       | 0.799     | 0.8    | 0.799     | 0.863    |
| <b>66% Training/34% Validation</b> | 80.8824  | 0.828     | 0.809  | 0.803     | 0.883    |
| <b>80% Training/20% Validation</b> | 83.75    | 0.851     | 0.838  | 0.838     | 0.903    |
| <b>50% Training/50% Validation</b> | 81       | 0.812     | 0.81   | 0.808     | 0.864    |

**FIGURE 3**  
**REP TREE IN 66% TRAINING SPLIT**



Looking at the top-most nodes in the REP Tree, it can be deduced that the first and most important determinant of the admission outcome. 128 of 139 records with CGPA greater than 8.64 was classified as admitted. This highlights the importance of CGPA in the admission decision. The next parameter considered, based on node 2, is the GRE Score.

**Random Tree**

Random Tree is a supervised classifier in Weka. It also uses an ensemble learning algorithm that generates many individual learners or trees. The model then employs a bagging idea to produce a random set of data for constructing a decision tree. Each node is split using the best split among all variables.

**TABLE 10**  
**EVALUATION METRICS OF RANDOM TREE CLASSIFICATION MODEL**

| Sampling Method                    | Accuracy | Precision | Recall | F-measure | ROC Area |
|------------------------------------|----------|-----------|--------|-----------|----------|
| <b>10-fold cross-validation</b>    | 79       | 0.793     | 0.79   | 0.791     | 0.788    |
| <b>5-fold cross-validation</b>     | 78.25    | 0.781     | 0.783  | 0.781     | 0.769    |
| <b>66% Training/34% Validation</b> | 75.7353  | 0.759     | 0.757  | 0.758     | 0.757    |
| <b>80% Training/20% Validation</b> | 83.75    | 0.837     | 0.838  | 0.837     | 0.833    |
| <b>50% Training/50% Validation</b> | 79       | 0.79      | 0.79   | 0.789     | 0.784    |

**J48**

J48 classifier is the C4.5 algorithm for building decision trees is implemented in Weka (Sawavanan and Gayathry, 2018). The J48 algorithm has high-performance accuracy in classification and is often called one of the best machine learning algorithms to examine the data categorically and continuously.

**TABLE 11**  
**EVALUATION METRICS OF J48 CLASSIFICATION MODEL**

| Sampling Method                    | Accuracy | Precision | Recall | F-measure | ROC Area |
|------------------------------------|----------|-----------|--------|-----------|----------|
| <b>10-fold cross-validation</b>    | 81.25    | 0.812     | 0.813  | 0.812     | 0.849    |
| <b>5-fold cross-validation</b>     | 79.5     | 0.795     | 0.795  | 0.795     | 0.809    |
| <b>66% Training/34% Validation</b> | 82.3529  | 0.824     | 0.824  | 0.822     | 0.794    |
| <b>80% Training/20% Validation</b> | 85       | 0.852     | 0.85   | 0.85      | 0.911    |
| <b>50% Training/50% Validation</b> | 80.5     | 0.805     | 0.805  | 0.804     | 0.801    |

**RESULTS AND EVALUATION**

After evaluating all models, their performance in predicting the outcome of the validation set based on learning from the training set was cross-compared.

**TABLE 12**  
**COMPARING TOP 10 MODELS BASED ON ACCURACY**

| Classifier                   | Sampling | Accuracy | Precision |
|------------------------------|----------|----------|-----------|
| <b>Naïve Bayes</b>           | 80/20    | 88.75    | 0.888     |
| <b>Logistic Regression</b>   | 80/20    | 88.75    | 0.888     |
| <b>Naïve Bayes</b>           | 66/34    | 86.0294  | 0.863     |
| <b>Logistic Regression</b>   | 66/34    | 86.0294  | 0.861     |
| <b>Multilayer Perceptron</b> | 80/20    | 85       | 0.852     |
| <b>Random Forest</b>         | 80/20    | 85       | 0.85      |
| <b>J48</b>                   | 80/20    | 85       | 0.852     |
| <b>Naïve Bayes</b>           | 5-fold   | 84.75    | 0.854     |
| <b>Multilayer Perceptron</b> | 66/34    | 84.5588  | 0.847     |

Naïve Bayes and Logistic Regression have the highest overall accuracy in predicting admission outcome, i.e. admitted. The 80% training/20% validation sets naturally have a higher accuracy due to the large training set. However, the accuracy for 66%/34% split is also relatively high for these two models at ~86% accuracy in correctly classifying admitted candidate. Comparing the cross-validation also indicates how accurately the models performed. For both 10- and 5-fold validation with Naïve Bayes and Logistic Regression, accuracy is higher than with the other models. This further supports that these two models'



better fits for predicting admissions. Naïve Bayes has slightly higher precision than logistic classification for the cross-validation samplings.

**TABLE 13**  
**COMPARING PERFORMANCE METRICS OF CROSS-VALIDATION SAMPLING**

| <b>Classifier</b>            | <b>Sampling</b> | <b>Accuracy</b> | <b>Precision</b> |
|------------------------------|-----------------|-----------------|------------------|
| <b>Naivebayes</b>            | 10-fold         | 84.5            | 0.851            |
| <b>Naivebayes</b>            | 5-fold          | 84.75           | 0.854            |
| <b>Multilayer Perceptron</b> | 10-fold         | 81.75           | 0.818            |
| <b>Multilayer Perceptron</b> | 5-fold          | 83.5            | 0.836            |
| <b>Logistic</b>              | 10-fold         | 84.25           | 0.843            |
| <b>Logistic</b>              | 5-fold          | 84              | 0.841            |
| <b>Random Forest</b>         | 10-fold         | 81              | 0.809            |
| <b>Random Forest</b>         | 5-fold          | 81.5            | 0.814            |
| <b>Reptree</b>               | 10-fold         | 79.25           | 0.793            |
| <b>Reptree</b>               | 5-fold          | 80              | 0.799            |
| <b>Random Tree</b>           | 10-fold         | 79              | 0.793            |
| <b>Random Tree</b>           | 5-fold          | 78.25           | 0.781            |
| <b>J48</b>                   | 10-fold         | 81.25           | 0.812            |
| <b>J48</b>                   | 5-fold          | 79.5            | 0.795            |

## CONCLUSION AND DISCUSSION

The study identified the Naïve Bayes and Logistic Regression classification model as good predictive models for graduate admission based on parameters of candidates long-established as crucial in informing the decision. Naïve Bayes outperforms Logistic Regression by a very small margin. Naïve Bayes has a few advantages that make it a widely accepted classification model, favored by analysts in many domains. Some of its benefits include that it is a simple model, making it easy and quick to implement; does not require large training data; can process both continuous and discrete variables. Furthermore, it is not too sensitive, which allows the ability to ignore irrelevant data features.

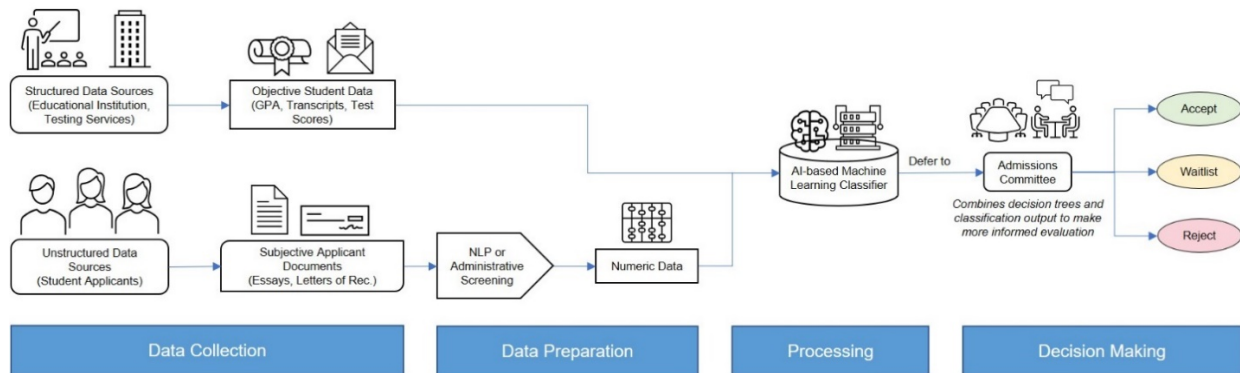
The study results show that classification models, trained on existing records of previous students, can be a valuable tool in predicting the possible outcome of prospective graduate school candidates by inputting variables such as test scores, grade point average, quality of essay and letters, etc. In practice, the admission decision is ultimately subjective to a great extent. Graduate programs especially tend towards a more thoughtful, holistic review of a broad range of qualities in a candidate, including personal attributes to complement the cognitive parameters to increase the likelihood that a student admitted to a master's or doctoral program will be among those most likely to succeed. However, the proposed approach can be a helpful tool for both students and admissions committees to consider some of the quantitatively definable attributes that still play significant roles in the decision process.

The dataset used contained numeric ratings for the statement of purpose essay and letters of recommendation. While the classification models are to be used by institutions, their predictive validity can be enhanced by transforming subjective admission requirements, such as application essays and recommendation letters, to numerical data through Natural Language Process, or assigning numeric ranks based on the quality of writing and embedding the new variables as predictors for machine learning modelling.

The decisions trees generated using the tree classifiers evaluated in this paper mirror the process of graduate admission decision quite well (see Figure 3). Relying on the decision tree based flow chart, admissions committees can review a parameter such as CGPA or GPA score first and use a rough cut-off

number to filter out less competitive applications. Then, they may check another parameter and apply a cut-off on that. Occasionally, they need to reconsider those that do not meet the original cut-off by cross-checking multiple parameters.

**FIGURE 4**  
**WORKFLOW PROCESS DEPICTING POSSIBLE APPLICATION OF MACHINE LEARNING MODELS IN ADMISSIONS DECISION**



This comparative study serves as a proof of concept for incorporating machine learning-based classification models into the admission decision process. The research does not suggest that artificial intelligence can or should take the place of an admission committee, but instead that it may assist with decision-making. A proposal for how machine learning may enhance the admission decision and support education management is described in Figure 4.

## FUTURE RESEARCH

The use of Machine Learning methods has been extensively studied in various fields, including medical diagnosis, financial trading and education performance, and they have shown strong potential to improve predictive capabilities in numerous ways (López-Martínez et al., 2018; Gerlein et al., 2016; Houssein et al., 2020). The study is one of few that has compared the ability of machine learning classification models in predicting graduate student admission outcome. Further research is needed to understand better the importance of the various determinants in this decision outcome. A similar data set may be collected from a different group of applicants to evaluate the predictive power of the classification models tested in this study and explore the significance of the various candidate attributes in informing admission decision in those groups. For example, US domestic students do not need to provide TOEFL scores and have CGPA recorded on a scale of 4.0. Such differences may affect the outcome of the predictive analysis. While prospective students have much to gain from such predictive models, graduate institutions have also called for more data-driven research demonstrating the link between admissions criteria and student success (Eberle et al., 2013; Kent and McCarthy, 2016). Thus, this area of research has a multitude of implications in enrollment decisions.

## REFERENCES

- Acharya, M., Asfia, A., & Antony, A. (2019). A comparison of regression models for prediction of graduate admissions. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. IEEE.
- Adekitan, A.I., & Noma-Osaghae, E. (2018). Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Education and Information Technologies*, 24(2), 1527–1543.
- Barik, L., Barukab, O., & Ahmed, A. (2020). Employing artificial intelligence techniques for student performance evaluation and teaching strategy enrichment: An innovative approach. *International Journal of Advanced and Applied Sciences*, 7(11), 10–24.
- Chang, L. (2006). Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research*, (131), 53–68.
- Cornell University. (2020). *Writing Your Statement of Purpose*. Graduate School. Retrieved from <https://gradschool.cornell.edu/diversity-inclusion/recruitment-2/prospective-students/writing-your-statement-of-purpose/>
- Deo, R.C., Yaseen, Z.M., Al-Ansari, N., Nguyen-Huy, T., Langlands T.A.M., & Galligan, L. (2020). Modern artificial intelligence model development for undergraduate student performance prediction: An investigation on engineering mathematics courses. *IEEE Access*, 8, 136697–136724.
- Eberle, W., Talbert, D., Simpson, E., Roberts, L., & Pope, A. (2013). Using machine learning and predictive modeling to assess admission policies and standards. *Proceedings of the 9th Annual National Symposium*. The University of Oklahoma.
- Fernández-García, A.J., Rodríguez-Echeverría, R., Preciado, J.C., Manzano, J.M.C., & Sánchez-Figueroa, F. (2020). Creating a recommender system to support higher education students in the subject enrollment decision. *IEEE Access*, 8, 189069–189088.
- Fu, Y. (2012). *The effectiveness of traditional admissions criteria in predicting college and graduate success for American and international students*. The University of Arizona.
- Gabriele, A.J., Joram, E., & Park, K.H. (2016). Elementary mathematics teachers' judgment accuracy and calibration accuracy: Do they predict students' mathematics achievement outcomes? *Learning and Instruction*, 45, 49–60.
- Gerlein, E.A., McGinnity, M., Belatreche, A., & Coleman, S. (2016). Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications*, 54, 193–207.
- Hand, D.J., & Yu, K. (2001). Idiot's Bayes — not so stupid after all? *International Statistical Review*, 69(3), 385–399. doi:10.2307/1403452
- Ho, T.K. (2002). A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis and Applications*, 5(2), 102–112. doi:10.1007/s100440200009
- Houssein, E.H., Emam, M.M., Ali, A.A., & Suganthan, P.N. (2020). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. *Expert Systems With Applications*, 114161.
- Kent, J.D., & McCarthy, M.T. (2016). *Holistic Review in Graduate Admissions: A Report from the Council of Graduate Schools*. Council of Graduate Schools.
- Kucak, D., Juricic, V., & Dambic, G. (2018). Machine learning in education - A survey of current research trends. *Proceedings of the 29th DAAAM International Symposium*, pp. 406–410.
- López-Martínez, F., Schwarcz, A., Núñez-Valdez, E.R., & García-Díaz, V. (2018). Machine learning classification analysis for a hypertensive population as a function of several risk factors. *Expert Systems With Applications*, 110, 206–215.
- López-Úbeda, P., Díaz-Galiano, M.C., Martín-Noguerol, T., Ureña-López, A., Martín-Valdivia, M., & Luna, A. (2020). Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches. *Expert Systems With Applications*, 160, 113647.

- Muratov, E., Lewis, M., Fourches, D., Tropsha, A., & Cox, W. (2017). Computer-assisted decision support for student admissions based on their predicted academic performance. *American Journal of Pharmaceutical Education*, 81(3), 46. doi:10.5688/ajpe81346
- Nuutila, K., Tuominen, H., Tapola, A., Vainikainen M.P., & Niemivirta, M. (2018). Consistency longitudinal stability and predictions of elementary school students' task interest success expectancy and performance in mathematics. *Learning and Instruction*, 56, 73–83.
- Patterson, B.F., Mattern, K.D., & Swerdzewski, P. (2012). Are the best scores the best scores for predicting college success? *Journal of College Admission*, 217, 34–45.
- Raghu, S., & Sriraam, N. (2018). Classification of focal and non-focal EEG signals using neighborhood component analysis and machine learning algorithms. *Expert Systems With Applications*, 113, 18–32.
- Raviya, K., & Gajjar, B. (2012). Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA. *Paripex - Indian Journal of Research*, 2(1), 19–21. doi:10.15373/22501991
- Sawavanan, N., & Gayathry, V. (2018). Performance and classification evaluation of J48 algorithm and kendall's based J48 algorithm (KNJ48). *International Journal of Computer Trends and Technology*, 59(2), 73–80.
- Simon, H. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.
- Slim, A., Hush, D., Ojah, T., & Babbitt, T. (2018). Predicting student enrollment based on student and college characteristics. *International Educational Data Mining Society*.
- Strano, M., & Colosimo, B.M. (2006). Logistic regression analysis for experimental determination of forming limit diagrams. *International Journal of Machine Tools and Manufacture*, 46(6), 673–682.
- Zhu, X. (2015, March). Machine teaching: An inverse problem to machine learning and an approach toward optimal education. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).