International Conference of Sciences, Engineering & Environmental Technology (ICONSEET), 6(35): 282 – 289, 2021
ISSN 0794-9650

www.repcomseet.com

ANALYSIS OF STUDENTS' ACADEMIC PERFORMANCE USING TRADITIONAL AND MACHINE LEARNING CLASSIFIERS

Olalude, Gbenga A., Amusan, Ajitoni S., and Adeshina, Ibrahim O.

Department of Statistics, Federal Polytechnic Ede, Osun State, Nigeria

Author's Correspondence E-mail: Olalude.gbenga@federalpolyede.edu.ng

Abstract: Students' academic performance was classified as a dependent factor of variables which include: Age upon admission, years spent at home before admission, program type, high school type, 0'level, and gender. Unlike traditional statistical techniques of analyzing data, machine learning algorithms have made it easier to feed computer with (partitioned) dataset and then slightly program it in order to obtain a model with the most precise classification. This study as well juxtaposed the classification and predictive performance of traditional statistical classifiers as compared with machine learning algorithms. Simultaneously, this study aims to classify and predict students' academic performance in Federal Polytechnic Ede, Osun state. The considered classifiers include Logistic Regression, Decision Tree, Support Vector Machine and Naive Bayes. These classifiers were subjected to varying train-test ratio ranging from 90-10 to 50-50. We observed that some of our classifiers perform below 50% accuracy, however decision tree algorithm yielded the most precise classification as it outperforms both Naïve Bayes and logistic regression largely, but support vector machine slightly. We found that age of students upon admission and the number of years spent at home after completing their secondary school influence their performance at the higher institutions. Although, the older students (> 21 years upon admission) are found to perform less compared to the younger ones, however, the number of years stayed at home further influence the possibility of good performance while in school. The implication of the study herewith is that students who gained admission at age older than 21 years and that have spent above two years at home are likely to experience poor academic performance. Students that are keen on acquiring tertiary education after the secondary education are encouraged to seek education while still at young age and those at home seeking admission should continually study at home while awaiting admission results.

Keyword: Classification; Machine learning; Support vector machine; Decision tree; Naive Bayes, Logistic regression

1.0 Introduction

Predicting students' academic performance is an integral part of an education system, as the overall growth of the education system is directly proportional to the success rate of the students in their examinations. Therefore, there are many situations where the performances of the students' are necessary to be predicted, for example to identify eligible students for participating in placement activities, to identify students eligible for scholarships and to find the weak students so that remedial action can be taken for their betterment. Often times, students' characteristics or experience in a course might determine if they will pass or fail (Kweon, Ellis, Lee & Jacobs, 2017). However, these sets of characteristics are considered personal which work in line with other external (e.g social) factors. Justifying from the point of view of these characteristics, one might model these characteristics to serve as basis of determining the possible outcome of a student after his study activities using statistical classification techniques.

With the advancement of technology in recent days, machine learning which as a branch of data mining is becoming more popular than the traditional techniques of data analysis (e.g. regression analysis, Monte Carlo) for its effectiveness in decision making processes and model formulation. Traditional data technique is a manual statistical procedure being programmed to automate a statistical technique. This indicates that without anyone

programming the logic, one has to manually formulate or code rules. On the other hand, machine learning is an automated process which increases the value of an embedded analytics in many areas, including data prep, natural language interfaces, automatic outlier detection, recommendations, causality and significance detection. All of these features help speed users insight and reduce decision bias.

Not to exaggerate, Xu, et al. (2019) has lauded the predictive performance of machine learning algorithms as compared to the traditional statistical techniques. They established the fact that machine learning gives room for the computer to learn from the available data (training data) until it return an optimum model for predictive actions on the unseen data (test data). Without any form of prejudice, traditional data analysis techniques are quite easy to approach for a lay-man but not really a good fit when it comes to big data analysis. Despite this advantage, traditional data analysis techniques might fail to render the most precise result based on other data characteristics e.g. imbalance dataset.

Considering the complexity in understanding the concept of machine learning and the ease with the use of traditional data analysis technique, one might be tricked to opt for the latter at the expense of the former. However, the likes of Dekker et al., (2009), Wolff et al., (2020) and Asif, et al., (2018), have made it clear the margin of advantage that machine learning algorithms have over tradition data analysis techniques. Similarly, Churpek et al. (2016) who compared the performance of several machine learning algorithms (Random forest, Gradient boosted machine, Bagged trees, Support vector machine, Neural network, Decision tree and K Nearest Neighbor algorithms) relative to logistic regression, using observational cohort dataset from hospitalized ward patients for detecting clinical deterioration on the wards in a large, multicenter database has found that several machine learning methods perfectly predicted clinical deterioration compared to the logistic regression model.

In relation to our study, Cortez and Silva (2008) have previously reviewed and predicts the performance of secondary school students' in two essential courses (Mathematics and Portuguese) using their previous score in the prior sessions and other demographic factors. The authors employed four data mining methods of Decision trees, Random forests, Neural networks and Support vector machines approach. They obtained a result which indicates that the prediction was attainable provided the grades of the previous session were known. This endorses that the prediction of students' performance is premised on previous performance and hence indicates that a student's performance is closely related to the performance in previous course (most likely a prerequisite course).

Another similar study by Sembiring et al. (2011) showed that machine learning capabilities provided effective improving tools for student performance. The study further showed how useful machine learning can be in higher education particularly to predict the final performance of students. The researchers collected data from students by using questionnaire to find the relationships between behavioral attitude of students and their academic performance after which they applied Decision tree and Support vector machine algorithms to predict the students' final grade. Also, the students were clustered into groups using kernel K-means clustering. They pinpointed from their model that there is a strong correlation between mental condition of students and their final academic performance. To combine the theme of all previous efforts, we will apply three machine learning algorithms in addition to logistic regression classifier which will serve as the relative group with which we will compare our other models and the most precise algorithm will serve as basis of our future students' academic performance classifier.

2.0 Data and Methods

We obtained social, demographic and educational records of students from the Department of Statistics, Federal Polytechnic Ede, Osun State. The obtained data comprise Student's ID (a unique key), Age of student before admission, Number of years stayed at home before admission, Program type (FT/PT), High school type (public/private), O'level exam, Gender, and Final grade. Selection of the sample was random across each academic level. Table 1 shows the overview of the students' academic data. A sample of 249 students' records was obtained across the department of statistics (Federal Polytechnic Ede) for Nation Diploma (ND) program. The target variable is the grade variable on 5 ordinal level (Probation, Pass,

Table 1: Overview of data used for the study

SN	Student's ID	Age before admission	Years at home before admission	Program type	High school type	O'level Credits	Gender	Grade
1	0086	19	1	FT	NECO	5	M	Lower
2	1470	20	3	FT	NECO	5	F	Lower
3	1849	18	2	FT	NECO	7	F	Upper
4	0568	18	1	FT	NECO	8	M	Lower
...	...	...	...	...	...	...	...	...
246	2895	18	3	FT	NECO	8	F	Pass
247	3549	18	1	FT	NECO	8	F	Lower
248	2763	19	2	FT	WAEC	8	F	Pass
249	3327	20	2	FT	NECO	8	F	Lower

Source: Department of Statistics, Federal Polytechnic Ede, Osun State (2021)

Lower, Upper and Distinction); other variables (except students' ID and serial number) are the dependent variables used across all model fitted in this study. Table 2 presented the descriptive view into students' characteristics indicated that minimum age of student upon entrance into the Polytechnic is 15 years, average of 20 years and maximum age of 27 years. It took a minimum of 1 year to wait for admission but on the average 2 years and at most 10 years after their secondary school activity.

Table 2: Descriptive Analysis Table

Column	Min	Max	Mean	Std. deviation	Variance
Age upon admission	15	27	19.96	2.1	4.41
Years at home before admission	1	10	2.41	1.27	1.62
O level credit	5	9	7.36	1.12	1.25

Source: Extracted from R Studio result output (2021)

We used three machine learning algorithms which include Decision tree, Naive Bayes, Support vector machine and one traditional classifier (the logistic regression) to classify students' final grade. The Decision Tree classifier is simply a flowchart diagram with the terminal nodes demonstrating classification decisions. Commencing with our dataset, we measured the entropy to find a way to split the set until all the data belong to the same class. There are several approaches to decision trees like ID3, C4.5, CART and many more.

Naive Bayes uses the Bayes' theorem in simple terms with an additional assumption that all predictors are independent. In other words, this classifier assumes that the presence of one particular feature in a class doesn't affect the presence of another one. It is similar to the Bayes' theorem however with more than one independent variable. The Bayes theorem given by [1] becomes [2] if there several independent variables

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})} \quad \textbf{[1]}$$

$$\hat{y} = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}} pv\ c\ (C_k) \prod_{i=1}^{n}\ p(x_i \mid C_k) \quad \textbf{[2]}$$

The Support Vector Machine is a step further away from the support vector classifier which tries to classify a given data point based on two supporting edges of the distinct group (support vector) using either the maximal margin classifier (which is sensitive to outliers and do not leave room for misclassification) or the soft margin classifiers (which allows for bias at an advantage of lower variance). Support vector machine uses either polynomial kernel which uses [3] to determine higher dimensional relationship from which classification is being made or the radial kernel which uses [4]. Where a and b are the support vectors (values at the edge of each clusters/groups), and value of r and d are best selected using cross-validation.

$$(a \times b + r)^d \quad \textbf{[3]}$$

$$e^{-\gamma(a-b)^2} \quad \textbf{[4]}$$

Finally the logistic regression is a logit transformation of the normal regression model which takes the form in [5] below:

$$\log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \textbf{[5]}$$

By simple algebra, we obtain p to be [6]

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2+1}} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}} = S_b(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \quad \textbf{[6]}$$

Where S_b is the sigmoid function which is basis for finding the probability that Y = 1 (that is an event occurring). For the case where there are K levels of Y, the unordered log odd can be obtained using [7] and [8] for ordinal cases.

$$p(C_k \mid \phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \quad \textbf{[7]}$$

where $\phi_k = [\phi_{k1},..,\phi_{kM}]^T, w_k = [w_{k1},..,w_{kM}]^T$

$a_k = w^T\phi + b$ is the "activation" *and* $b = biases$

$$\log \frac{P(Y \le j)}{P(Y > j)} = \text{logit}(P(Y \le j)) = \beta_{j0} - \eta_1 x_1^- \cdots - \eta_p x_p \quad \textbf{[8]}$$

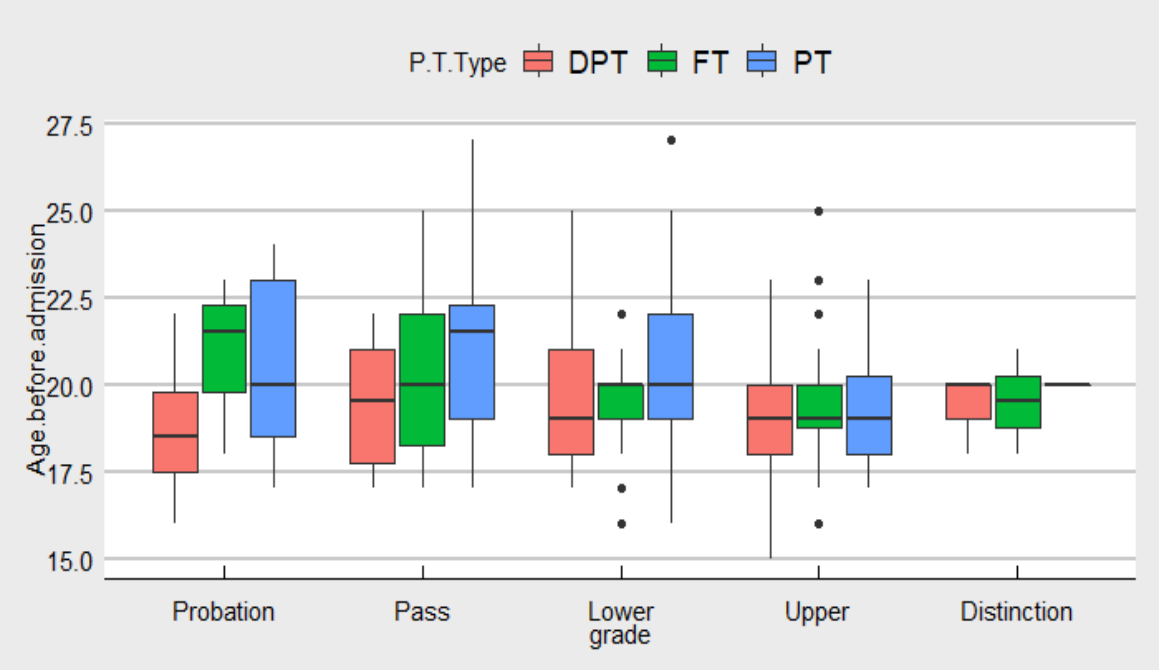
Our model comprises 6 independent variables (Age at admission, Years before admission, Program type, High school type, 0'level, and Gender) against the ordered grade variable (Probation, Pass, Lower, Upper and Distinction). To ensure greater accuracy, we fed the algorithms with varying training data of different ratios starting from 90% training and 10% testing to 50% training and 50% test. We took records of varying performance measure on each occasion and a summarize result is obtained to compare between models. We compared between algorithms and select the algorithm with the highest prediction accuracy. It is important to emphasize that a better alternative to our method is to use cross validation to juxtapose between performance of varying model like the study of Ramezan, Warner and Maxwell (2019) why they applied the population k-fold cross validation to selected the most precise model. However, we prefer the manual method so as to monitor the trend of performance at each iteration.

3.0 Empirical Result

Judging from the grade distribution by age plot presented in Figure 1, it is evident that people who studied full time and end up in probation have the maximum average age (over 21 years) while those that are admitted on daily part time program who ended up in probation are having the minimum average age (below 18 years). We

can suggest that age is a key factor that is been influenced by program type from the former result, however we can oppose this claim also by the latter result. This indicates that a more sophisticated inferential analysis is needed to make a good decision from this dataset. We proceed from the descriptive analysis to fitting four classification algorithms. We consider the use of decision tree and support vector machine from the machine learning models and the ordinal logistic regression and Naive Bayes from the traditional data classifiers. The use of ordinal logistic regression is as a result of the target variable (grade) having five ordered factor levels as previously stated.

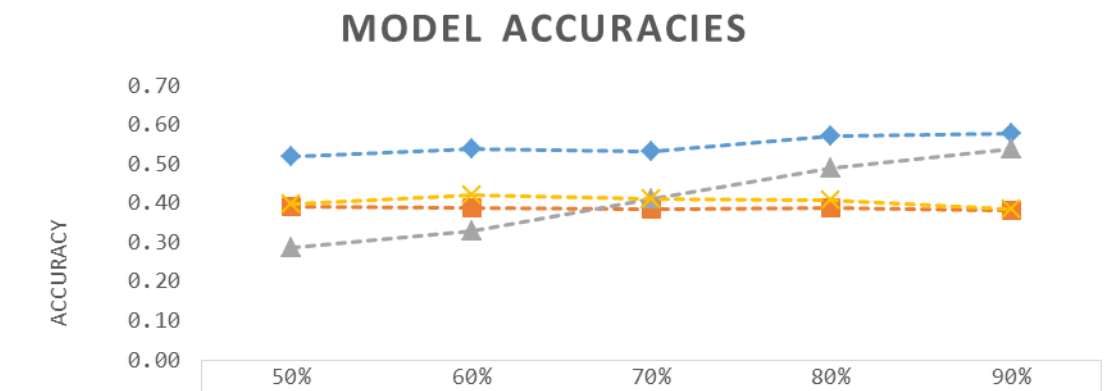
Figure 1: Grade Distribution by age



Source: Extracted from R Studio result output (2021)

R programming language was adopted to perform various training and testing iteration using the Caret package from the program and the accuracy of each model is shown in figure 2. On all occasions, the result is similar to what is obtained in Table 2. It is evident that decision tree outperforms other classification algorithms. Thus, it is considered as the most appropriate algorithm for classifying our academic performance data. Our findings is similar to the findings of Sembiring et al., (2011), that machine learning algorithms are more efficient in classifying dataset than the traditional logistic regression analysis.

Figure 2: Model Accuracies

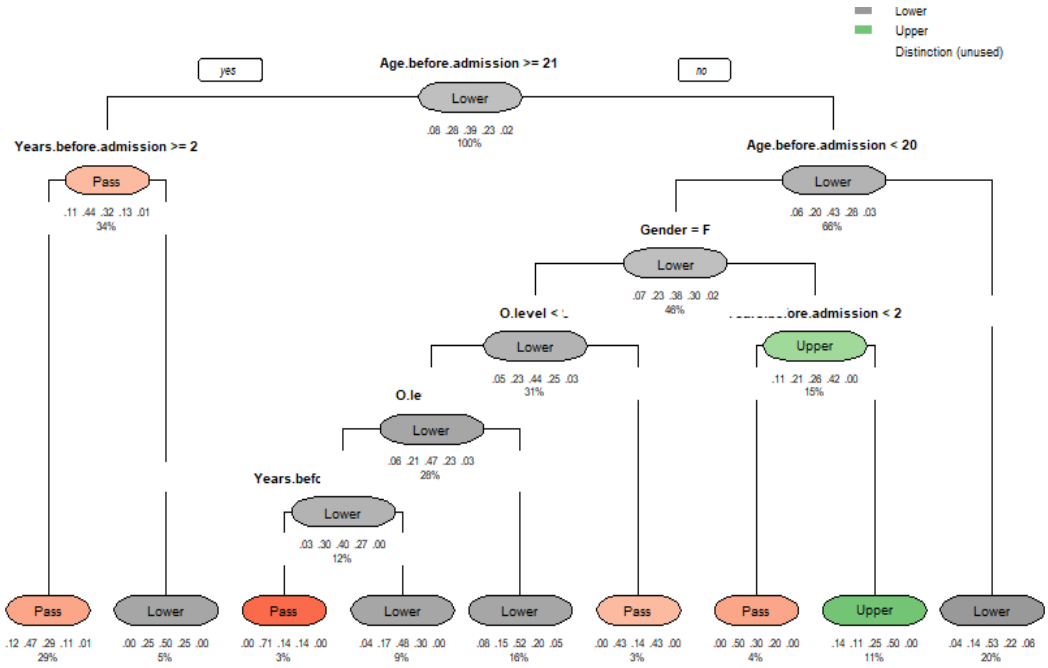


	50%	60%	70%	80%	90%
decision	0.52	0.54	0.53	0.57	0.58
support	0.39	0.39	0.39	0.39	0.38
naive	0.29	0.33	0.41	0.49	0.54
logistic	0.40	0.42	0.41	0.41	0.38

Source: Extracted from R Studio result output (2021)

By observing the decision tree of 90% training data in Figure 3, it is evident that 21% of students, who are older than a threshold of 21 years, lapsed more than 2 years before gaining admission, ending with a "PASS" grade while only 5% of those managed that reach the "LOWER" grade. On the other hand, 11% of male students who were younger than the threshold of 20 years and lapsed lesser than 2 years before gaining admission managed to achieve "UPPER CREDIT" while majority of female students end up either with a "PASS" or "LOWER CREDIT".

Figure 3: Decision Tree Plot for 90% training dataset



Source: ***Source: Extracted from R Studio result output (2021)***

4.0 Conclusion and Recommendations

We classified students' academic performance based on some characteristics that influenced their performance using the traditional and machine learning classifiers. We considered three machine learning algorithms, decision tree, support vector machine, and Naive Bayes algorithms. All the models improve as more training samples are fed into the data used for the classification, except for support vector machine which has a reduction in accuracy beyond 60%. It is evident that decision algorithm performs best (though with a fair accuracy of 55% on average) compared to other models. This indicates that decision tree algorithm is the best model that suits our data. We as well noticed that Naive Bayes algorithm outperforms both support vector machine and logistic regression algorithms, but also perform below 50% accuracy. We conclude by stating that machine learning algorithms are more accurate in prediction making compared to traditional classifiers. With reference to students' academic performance, the age of students upon admission and the number of years spent at home after completing their secondary school have been found to influence the performance of students at the higher institutions. Although, the older students (> 21 years upon admission) are found to perform less compared to the younger ones, the number of years stayed at home further influence the possibility of good performance while in school. The implication of the study herewith is that students who gained admission at age older than 21 years and that have spent above two years at home are likely to experience poor academic performance. Hence, students that are keen on acquiring tertiary education after the secondary education are encouraged to seek education while still at young age. Also, while still at home seeking admission into higher institution of learning, they should keep themselves busy studying at home to avoid poor academic performance when they eventually gained admission into their desired higher institution.

References

Asif, M., Martiniano H. F., Vicente A. M., & Couto F. M. (2018). Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. *PLoS ONE* 13(12): e0208626.

Churpek, M. M., Yuen, T. C., Winslow, C., Meltzer, D. O., Kattan, M. W., & Edelson, D. P. (2016). Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *PMC, 44*(2), 368.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance, *EUROSIS-ETI*, 5-12.

Dekker, A., Cary, D-O., Ruysscher, D. D., Lambin, P., Komati, K., Fung, G., Yu, S., Hope, A. De Neve, W., & Lievens, Y. (2009). Survival Prediction in Lung Cancer Treated with Radiotherapy: Bayesian Networks vs. Support Vector Machines in Handling Missing Data, *ICMLA '09* , 494-497.

Kweon, B.-S., Ellis, C. D., Lee, J., & Jacobs, K. (2017). The link between school environments and student academic performance. *ScienceDirect, Elsevier, 23*, 35-43.

Ramezan, C., A. Warner, T., & Maxwell, A. E. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sens*., 11(2), 185. https://doi.org/10.3390/rs11020185

Sembiring, S., Zarlis, M., Hartama, D., Wani, E., & Ramliana, S. (2011). Prediction of Student Academic Performance by an Application of Data Mining Techniques. *International Conference on Management and Artificial Intelligence IPEDR* (6).

Wolff, S., O'Donncha, F., & Chen, B. (2020). Statistical and machine learning ensemble modelling to forecast sea surface temperature. *Journal of Marine Systems, Elsevier, 208,* https://doi.org/10.1016/j.jmarsys.2020.103347

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computer in Human Behavior, Elsevier*, 98: 166-173.