International Journal of Neural Networks and Advanced Applications
DOI: 10.46300/91016.2020.7.10 Volume 7, 2020

Modelling Student's performance using data mining techniques in a higher learning environment in the Pacific

Ravneil Nand, Ashneel Chand
The University of the South Pacific
Suva
Fiji

Abstract—The students' performance in higher education has become one of the most widely studied area. Modelling student performance play a pivotal role in forecasting students' performance where the data mining applications are now becoming most widely used techniques in this study. There are various factors, which determine the student performance. Eight attributes are used as input, which is considered most influential in determining students' performance in the Pacific. Statistical analysis is done to see which attribute has the highest influence to student performance. In this research, different algorithms are utilized for building the classification model, each of them using various classification techniques. Some of classification techniques used are Artificial Neural Network, Decision Tree, Decision Table, and Naïve Bayes. The WEKA explorer application and R software are used for correlation test between different variables. The dataset used in this research is an imbalanced set, which is later transformed to balance set through under sampling. Neural Network is one of the classification techniques that has done well on both, imbalanced and balanced dataset. Another technique which has done well is Decision tree. Statistical analysis shows that internal assessment has weak positive relationship with student performance while demographic data is not. Further observations are reported in this research in relation to two types of datasets with application to different classification techniques.

Keywords—Decision Tree, Classifier, Correlation test, Neural Network, Pacific

I. Introduction

THE performance of students are means of ways to determine their excellence in education [1]. It is one the main factors that universities work on to improve on their quality of graduates since universities aims to produce graduates with not just to have an academic transcript but immerse knowledge and skills which is welcomed by employers. Performance of students play a pivotal role in any institute, which can aid in eradicating problems that may exist. It also give a chance to the lecturer and educators to review the course partially or as a whole.

The widespread use of technology and blended learning has facilitated many changes in the education sector including higher education in the Pacific [2]. The virtual learning space allow students to get access to education and resources irrespective of their geographical locations [3]. Data mining techniques are becoming a very common tool in understanding and solving educational and administrative issues in higher education [4] [5] [6]. Amicably, it has been used on modelling student's performance. Authors Ibrahim & Rusli [7], Borkar & Rajeswari [8] and Kalyani, et al. [9] have highlighted that Neural Network plays a vital role in predicting the students' performance.

There are various factors that start affecting student's performance. Some of these can be their home environment or accommodation, which means are the students living in campus, renting or living with their family [8] [10]. The authors further alluded that student's gender, age and marital status might affect student's performance as well. Their learning styles, study habits, language, hobbies and interests. Student tutorial attendance, mid semester marks, assignment marks, and

total coursework is also a very important part of their overall performance. The number of students opts to study particular programme that is of national interest and where scholarship can be easily accessed. Later students realized they have chosen a wrong course after doing few units. This may have also let wrong choice and which annihilates the student performance. Hence the course selection is utmost vital when making decisions [11]. Their programme may also be a factor as some students feel they have enrolled in a programme not suitable for them as there is initial and changing motivation [12]. Number of units a student is doing, its modes, part time or full time student and its level the student is doing can also be a result of student's performance. Few factors that influences the student's selections course characteristics, student's workload, course type and time, etc [11].

The core objective of any university is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students' performance [13]. Students' performance at end semester final examination and grading depicts the success of the student performance. There are many approaches that can be used to o evaluate students' performance. Early intervention to predict students' performance minimise the risk of poor performance [13] [3] [2].

Hu [14] and Sharma [2] stipulated that an early warning system is necessary which can help to identify at-risk students, or predict student-learning performance by analysing learning portfolios recorded in a learning management system (LMS). Early warning system is a process that identifies at –risk students, which converts the information into an intelligent action. This assists academic institutions in improving the retention and success rate of students and to get overview of performance before the examination to reduce the risk of failure.

The academic institutions are most often adjudged by the grades achieved by the students and their overall performance. Educational Data Mining (EDM) is a growing research field which assist academic institutions to improve the performance of their students [15]. The respectable quality of educational dataset can yield healthier results and henceforward the decisions based on such quality dataset can upsurge the quality of education by predicting the performance of students

This research paper examines three various classification techniques namely decision tree, multilayer perceptron and naïve bayes to model students' performance. It is also proposed to see which of the three classification algorithms gives best results that can provide important reference materials for the planning of the future success of the students and faculties. Statistical analysis will be also carried out in order to evaluate and compare the linear dependencies between two variables.

The rest of this paper is organized as follows: Section II is on background while section III delineates on methodology. Section IV gives the insight of the results obtained. Section V presents the discussions, and Section VI concludes the study.

II. Background

Flexibility is one of the hallmarks of The University of the South Pacific (USP) education. The smaller centres are part of the larger campuses spread in remote locations or on the smaller islands in some regional countries in the Pacific [16]. All member country have at least one campus that varies from others in size and student population and centres spread across the outer islands. Laucala campus is the headquarters that coordinates and facilitates most of the courses offered in the region by various modes. The advancement in technology has facilitated many changes in the education sector at higher education. Online and Blended mode is now also used as medium of study offered by universities [17]. USP is a one of its type in the pacific region as it is owned by its 12 member countries in the pacific – Cooks Islands, Fiji, Kiribati, Marshall Islands, Nauru, Niue, Samoa, Solomon Islands, Tokelau, Tonga, Tuvalu, and Vanuatu [2].

III. Methodology

A. Research Framework

Face to Face (F2F) is the traditional mode of study at higher education, where communication that takes place between students and instructors. To compliment this mode, Blended mode has been introduced which typically involves F2F and online activities [17]. Online delivery mode is whereby students do not get any F2F meetings. Both modes is been complimented by well-established instructional design and pedagogical framework which provides a greater understanding of learning in online space. Authors in Kardan, et al. [11], and Baragash & Al-Samarraie [17] states that Learning Management System (LMS) is the most widely used platform in higher education to promote active learning. This platform provides avenues to monitor students in real time to identify students that needs additional learning support in a course. LMS allows educators to get date of student's participation in online activities [3] [16] such as online quiz, discussion forum, lecture capture view, number of times students visited Moodle page, etc. Literature also shows that, logs of online activity and students data captured in an LMS can be used to predict student performance. Forecasting student performance is critical for higher education institution as the student's performance is a way to measure the lectures performance.

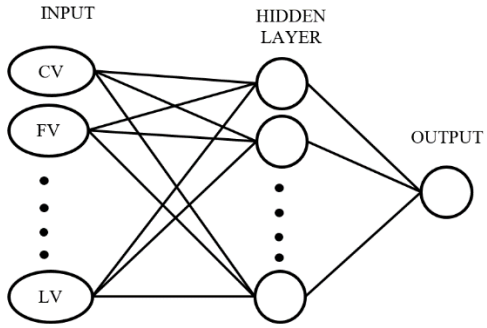

Fig. 1 Artificial neural network diagram. Key: CV, FV and LV are Course View, Forum View and Lecture View respectively.

In this paper we are predicting the student's performance by the time the course has finished within the semester using different attributes. Literature also depicts that various techniques has been used to forecast student's performances using different attributes.

B. Data Collection and Processing

Blended and Online atmosphere allows for a regular tracking of learner involvement in learning activities and monitoring of student learning and the students can be tracked on real time. Centre for flexible learning (CFL) which look after Moodle has all the log activities of the students who are registered on Blended and Online mode. The log activities of 651 students was requested from CFL. The log activities included test marks, assignment marks, time spent on Course view, Forum view and Lecture capture view for 200 and 300 level students. Mode of study, gender and grades was also collected concurrently as given in the Table 1.

Date cleaning has been the first step before it has been normalized. The real time data obtained was normalized as in [18].

In order to meet the objectives of this research, a prediction model was built to determine the pass or fail of students. The research involves qualitative research method where numerical data was collected such as mid-semester marks and coursework marks and which was analysed using statistics method. The research could be considered as a qualitative research since it includes the textual data and the use of the diagrams. The models were used to analyse and to predict student performance. To carry out the research, student's data were collected in csv format. Statistical analysis was done using R software. The analysis was conducted using Weka where decision table, decision tree, Neural Network (NN) and naïve bayes classifiers was used.

Table1: Attributes and Values

Attributes	**Description**	**Values**
Course level	Level of course	200 level 300 level
Gender	Student gender	Male Female
Test	Total Marks of students in test	Total Test Marks
Assignment	Total Marks of students in Assignment	Total Marks
Course view	Student Total Course views per semester	Total Course views
Forum view	Student Total Forum views per semester	Total Forum views
Lecture View	Student Total Lecture Capture views per semester	Total Lecture Capture views
Grade	Final Grade of student at end of semester	Pass Fail
Mode	The mode in which the student is enrolled in the course	B O

WEKA is a free software that is collection of machines learning algorithm for solving the real-world data mining problems [8]. It contains features such as machine learning, data mining, pre-processing, classification, regression, clustering, association rules, attribute selection, and in this research a confusion matrix was used to determine the model accuracy [8] [14]. Three algorithms that were used in this research are decision tree, neural network, and naive bayes algorithms. Missing values was replaced by a question mark "?". The data gathered was an unbalanced data and the model has been tested for both balance and imbalance data. Under-sampling was used to create the balanced dataset. Balance data is where there were equal number of pass and fail, while imbalanced data had more pass (548) then compared to fail (103). The data arrangement was left random for both balanced and imbalance data. Numerous different classification algorithms are applied during the performed research work, selected because they have likely to yield worthy results. Popular WEKA classifiers used in this research includes, Trees (J48), Bayesian classifiers (NaiveBayes) and Function (multilayer perceptron). Different test options namely; use training set, cross validation (10 fold), and percentage split (70%) has been used in all above mentioned cases.

C. Statistically analysis

Calenge [19] states that R is an integrated suite of software facilities for data manipulation, calculation and graphical display. A large, coherent, integrated collection of intermediate tools for data analysis and its operation is given in figure 2. R is an interpreted language, very simple and intuitive. To evaluate and compare the linear dependencies between two variables, statistically analysis is done. Firstly, the variables were classified into either ordinal or nominal datasets. Later, based on the data type, best suited test were conducted on different variables. The two common tests were Spearman Rho and Pearson Test.

Spearman Rho is a nonparametric measure of rank correlation which is suitable for ordinal or nominal datasets and Pearson test is a parametric correlation test where the variables are from normal distribution [20].

D. Performance measurement

To evaluate and compare the performance of the classification models, classification accuracy, error rate and error type is used [14]. Same has been done in this research and confusion matrix used to calculate the above mentioned.

IV. Results

Firstly, the association between variables was carried out. The associations between the Test, Assignment and the Grades were investigated using the Spearman's ranked correlation test as seen in Table 2. This is done based on normalty test, where it was found that the sample data does not follow a normal distribution, therefore, nonparametric test is used. Table 2 shows a moderate correlation between Test and Grades, and Assignment and Grades with correlation coefficient of .315,

and .314 respectively. As for Test and Assignment it shows moderate negative correlation with correlation coefficient of -0.300.

Table 2: Spearman's ranked correlation test

		Test	Assignment	Grade
Test	Correlation Coefficient	1.000	**-0.300****	**0.315****
	Sig. (2-tailed)	.	5.09E-15	1.89E-16
	N	651	651	651
Assignment	Correlation Coefficient	**-0.300****	1.000	**0.314****
	Sig. (2-tailed)	5.09E-15	.	2.37E-16
	N	651	651	651
Grades	Correlation Coefficient	**0.315****	**0.314****	1.000
	Sig. (2-tailed)	1.89E-16	2.37E-16	.
	N	651	651	651

Table 3 shows association between CourseView, ForumView, LCaptureView, Gender, Mode, Levels with Grades. CourseView, ForumView, Gender and Mode have a negative weak correlation with Grades while LCaptureView and Levels has very weak positive correlation since its closer to zero.

Table 3: Spearman's ranked correlation test

		Grade
CourseView	Correlation Coefficient	**-0.1189**
	Sig. (2-tailed)	0.0024
	N	651
ForumView	Correlation Coefficient	**-0.131**
	Sig. (2-tailed)	0.0008
	N	651
LCaptureView	Correlation Coefficient	**0.037**
	Sig. (2-tailed)	0.350
	N	651
Gender	Correlation Coefficient	**-0.1104**
	Sig. (2-tailed)	0.0048
	N	651
Mode	Correlation Coefficient	**-0.1413**
	Sig. (2-tailed)	0.00030
	N	651
Levels	Correlation Coefficient	**0.0763**
	Sig. (2-tailed)	00516
	N	651

A newly proposed Funnel model given in Fig. 2 conceptualizes the results of this research. The output of the model shows the student performance that is the Grade. The factors that have the positive impact on the Grade are internal assessments and external assessments. These are the inputs; Assignment, Test and other factors.

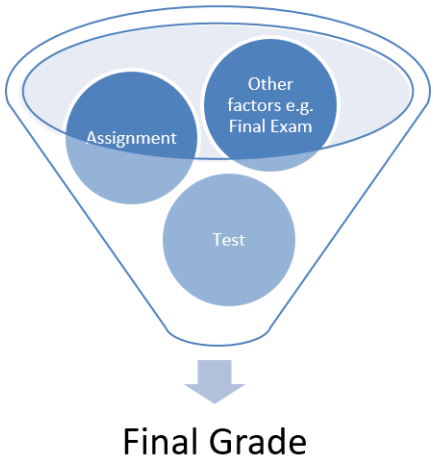

Fig. 2 Funnel model showing Assignment and Test as main contributors to student performance.

Different algorithms are utilized for building the classification model, each of them using various classification techniques. The WEKA Explorer application is used. Tables 4 – 6 represents the accuracy and True Positive Rate (TPR) of various classifiers for balanced data. The best results are made bold in the tables, indicating best accuracy and TPR.

Table 4 shows the results of percentage split of different classifiers where P and F were arranged in order (Balanced Data). Artificial Neural Network (ANN) has produced better results than the rest with 96.8% accuracy. Table 5 shows the results of cross validation for different classifiers where P and F were arranged in order. Again it was seen that ANN had better results. Table 6 shows the results of use of training set of different classifiers where P and F were arranged in order. This time, ANN had better results than other 2 methods.

Table 4: Percentage split of different for Balanced Data

Decision Tree		Neural Network		Naïve Bayes	
Accuracy	TPR	Accuracy	TPR	Accuracy	TPR
95.2	91.4	**96.8**	**94.1**	95.2	91.4

Table 5: Cross validation of different classifiers for Balanced Data

Decision Tree		Neural Network		Naïve Bayes	
Accuracy	TPR	Accuracy	TPR	Accuracy	TPR
92.7	88.6	**93.7**	**90.2**	92.3	87.8

Table 6: Training set of different classifiers for Balanced Data

Decision Tree		Neural Network		Naïve Bayes	
Accuracy	TPR	Accuracy	TPR	Accuracy	TPR
95.6	92	**96.1**	**92.8**	92.3	87.8

Tables 7 – 9 shows the accuracy and True Positive Rate (TPR) of various classifiers for imbalanced data. Bold values indicate the best results obtained. Table 7 shows the results of

percentage split of different classifiers where P and F was not arranged in order. It can be seen that ANN is one of the algorithms with best results. Table 8 shows the results of cross validation for different classifiers where P and F was not arranged in order. This time around, ANN again has the best results in comparison to other two methods. The accuracy is 89.4%. Table 9 shows the results of use of training set of different classifiers where P and F were not arranged in order. It can been seen that ANN has got best results with 94.3% accuracy.

Table 7: Percentage split of different classifiers for Imbalanced Data

Decision Tree		Neural Network		Naïve Bayes	
Accuracy	TPR	Accuracy	TPR	Accuracy	TPR
85.6	95.6	**86.2**	**96.9**	81.5	91.9

Table 8: Cross validation of different classifiers for Imbalanced Data

Decision Tree		Neural Network		Naïve Bayes	
Accuracy	TPR	Accuracy	TPR	Accuracy	TPR
88.8	90.9	**89.4**	**91**	86.5	90.4

Table 9: Training set of different classifiers for Imbalanced Data

Decision Tree		Neural Network		Naïve Bayes	
Accuracy	TPR	Accuracy	TPR	Accuracy	TPR
91.9	92.3	**94.3**	**93.8**	86.2	90.3

Tables 10 shows the ranking of accuracy of various classifiers for balanced and imbalanced dataset. The best results can be seen for ANN.

Table 10: Ranking of different classifiers based different methods.

	Balanced			Imbalanced		
Method	D.Tree	ANN	Naïve Bayes	D.Tree	ANN	Naïve Bayes
% Split Data	2	1	2	2	1	3
10 Folds	2	1	3	2	1	3
Training Set	2	1	3	2	1	3
Total Rank	6	**3**	8	6	**3**	9
Mean Rank	2.0	**1.0**	2.7	2.0	**1.0**	3.0

V. Discussion

This section will discuss the results of the experiment in predicting students' performance. This analysis is based on the ranking of prediction methods and also the main important factors that may influence the students' performance. Table 10 shows the prediction accuracy of different classification method through ranking for predicting students' performance.

In Table 10, one method that has outperformed in imbalanced dataset are Artificial Neural Network (ANN). Neural Network has the best ranks in balanced data and even in unbalanced data. For the balanced data, ANN is able to achieve rank number 1 in 3 out of 3 (100%). Rest of the two methods, Decision Table and Naïve Bayes, are not able to come first in any of the instances. As for the imbalanced data, ANN is able to achieve number 1 rank in 3 out of 3 (100%) cases again as in balanced dataset. The two methods, Decision Tree and Naïve bayes are not able to achieve first rank in any of the cases. It can be seen that in both the data, Naïve Bayes has got the worst result. A classification technique which has produced best results in both datasets is ANN.

Tables 4-9 shows also the highest prediction accuracy of each methods in respect to the dataset. The highest prediction accuracy of ANN is (96.8%) followed by Decision Tree by (96.3%). Next, Naïve Bayes which is (95.2%). All the methods have highest accuracy recorded in balanced dataset. According to literature under sampling can sometimes lose some important attributes but in this case, it has proved to be beneficial.

The result of prediction accuracy is highly dependent on the attributes or features that were used during the initial prediction process. Neural Network method gave the highest prediction accuracy because of the influence from main attributes. Test, Assignment, CourseView, ForumView, LCaptureView, Gender, Mode and Levels have all played a very important role in it. Accordingly to the statistical analysis seen in Tables 2-3, the two most influential attributes are Test and Assignment. One of the advantages of Neural Network is the ability to capture nonlinear relationships easily. It is also referred as adaptive system due to its ability to readily update the historical data like a human brain. So, the model always functions beyond the knowledge base. In addition, the strength of neural network is the ability to learn from a limited set of data such that smaller dataset.

VI. Conclusion

In educational data mining method, predictive modelling is usually used in predicting student performance. In order to build the predictive modelling, there are several tasks used, which are classification, regression and categorization. The most popular technique in predicting student's performance is classification model. There are a number of algorithms under classification model that have been applied to predict student's performance. Among the algorithms used are Decision Tree, Artificial Neural Networks (ANN), Naive Bayes, Decision Table and Support Vector Machine. The classifiers used in this research are ANN, Decision Tree, and Naive Bayes.

Predicting students' performance is mostly useful to the educators and learners as it aims to improve the learning and teaching process. This research has applied different classification techniques in order to predict student performance through use of different attributes beneficial in a High Education Institute. One method that has showed good performance in balanced and imbalanced dataset is ANN where the Test and Assignment mark plays a big role in students'

success. The internal assessments has increased the accuracy of ANN. For prediction techniques, the classification method is frequently used in educational data mining area, where Neural Network and Decision Tree are the two methods highly used by the researchers for predicting student's performance and it has also performed well on the datasets. In conclusion, the predicting of student's performance has motivated us to carry out further research in terms of meta-analysis. It will help the educational system to monitor the students' performance in a better way where important features need special attention.

References

[1] P. M. Arsad and N. Buniyamin, "A neural network students' performance prediction model (NNSPPM)," in *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, 2013.

[2] B. Sharma, R. Nand, M. Naseem and E. V. Reddy, "Effectiveness of online presence in a blended higher learning environment in the Pacific," *Studies in Higher Education,* pp. 1-19, 2019.

[3] A. Jokhan, B. Sharma and S. Singh, "Early warning system as a predictor for student performance in higher education blended courses," *Studies in Higher Education,* pp. 1-12, 2018.

[4] M. Agaoglu, "Predicting instructor performance using data mining techniques in higher education," *IEEE Access,* vol. 4, pp. 2379-2387, 2016.

[5] I. M. Tarun, "Prediction models for licensure examination performance using data mining classifiers for online test and decision support system," *Asia Pacific Journal of Multidisciplinary Research,* vol. 5, no. 3, pp. 10-21, 2017.

[6] N. Venkatesan, "Role of Data Mining Techniques in Educational and E-learning System," *Asia Pacific Journal of Research,* vol. 2, 2013.

[7] Z. Ibrahim and D. Rusli, "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression," in *21st Annual SAS Malaysia Forum, 5th September*, 2007.

[8] S. Borkar and K. Rajeswari, "Attributes selection for predicting students' academic performance using education data mining and artificial neural network," *International Journal of Computer Applications,* vol. 86, no. 10, 2014.

[9] B. S. Kalyani, D. Harisha, V. RamyaKrishna and S. Manne, "Evaluation of Students Performance Using Neural Networks," in *International Conference on Intelligent Computing, Information and Control Systems*, 2019.

[10] W. W. Guo, "Incorporating statistical and neural network approaches for student course satisfaction analysis and prediction," *Expert Systems with Applications,* vol. 37, no. 4, pp. 3358-3365, 2010.

[11] A. A. Kardan, H. Sadeghi, S. S. Ghidary and M. R. F. Sani, "Prediction of student course selection in online higher education institutes using neural network," *Computers & Education,* vol. 65, pp. 1-11, 2013.

[12] C. Sinclair, "Initial and changing student teacher motivation and commitment to teaching," *Asia-Pacific Journal of Teacher Education,* vol. 36, no. 2, pp. 79-104, 2008.

[13] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv preprint arXiv:1201.3417,* 2012.

[14] Y.-H. Hu, C.-L. Lo and S.-P. Shih, "Developing early warning systems to predict students' online learning performance," *Computers in Human Behavior,* vol. 36, pp. 469-478, 2014.

[15] M. Zaffar, S. Iskander and M. A. Hashmani, "A study of feature selection algorithms for predicting students academic performance," *Int. J. Adv. Comput. Sci. Appl,* vol. 9, no. 5, pp. 541-549, 2018.

[16] B. N. Sharma, A. D. Jokhan, R. Kumar, R. W. Finiasi, S. Chand and V. Rao, "Use of short message service for learning and student support in the Pacific region," Springer, 2015.

[17] R. S. Baragash and H. Al-Samarraie, "Blended learning: Investigating the influence of engagement in multiple learning delivery modes on students' performance," *Telematics and Informatics,* vol. 35, no. 7, pp. 2082-2098, 2018.

[18] A. Chand and R. Nand, "Rainfall prediction using Artificial Neural Network in the South Pacific region," in *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2019.

[19] C. Calenge, "The package "adehabitat" for the R software: a tool for the analysis of space and habitat use by animals," *Ecological modelling,* vol. 197, no. 3-4, pp. 516-519, 2006.

[20] S. Arndt, C. Turvey and N. C. Andreasen, "Correlating and predicting psychiatric symptom ratings: Spearmans r versus Kendalls tau correlation," *Journal of psychiatric research,* vol. 33, no. 2, pp. 97-104, 1999.