Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education

Sarfraz Nawaz Brohi[1](✉), Thulasyammal Ramiah Pillai[1],
Sukhminder Kaur[1], Harsimren Kaur[2], Sanath Sukumaran[1],
and David Asirvatham[1]

[1] Taylor's University, Subang Jaya, Selangor, Malaysia
{SarfrazNawaz.Brohi, Thulasyammal.RamiahPillai, Sukhminder.Kaur, Sanath, David.Asirvatham}@taylors.edu.my
[2] Hilti Asia IT Services, Petaling Jaya, Malaysia
simyaulekh10@gmail.com

Abstract. In this research, we compared the accuracy of machine learning algorithms that could be used for predictive analytics in higher education. The proposed experiment is based on a combination of classic machine learning algorithms such as Naive Bayes and Random Forest with various ensemble methods such as Stochastic, Linear Discriminant Analysis (LDA), Tree model (C5.0), Bagged CART (treebag) and K Nearest Neighbors (KNN). We applied traditional classification methods to classify the students' performance and to determine the independent variables that offer the highest accuracy. Our results depict that the data with the 11 features using random forest generated the best accuracy value of 0.7333. However, we revised the experiment with ensemble algorithms to reduce the variance (bagging), bias (boosting) and to improve the prediction accuracy (stacking). Consequently, the bagging random forest outperformed other methods with the accuracy value of 0.7959.

Keywords: Predictive analytics · Machine learning · Higher education

1 Introduction

Likewise, most industries, data also plays an important role in higher education. Apart from traditional data, nowadays the education organizations are collecting data from social media and location-based streams. The collected data can be used to mine and construct predictive analytical models to enhance students' success rate. The data mining and analytics software can be utilized to provide immediate outcomes to the instructors about the learner's academic performance. Such tools can analyze patterns and predict outcomes such as potential cases of dropping out, requiring extra assistance or demanding challenging assignments [1]. Each learner has different behavior in different modules due to varying strengths and weaknesses. Predictive analytics in education can assist in differentiated learning [2]. Nowadays, higher education institutions are investing in building predictive analysis tools. A prediction model was built in the Nottingham Trent University (NTU) [3]. The application pointed out the four

M. H. Miraz et al. (Eds.): iCETiC 2019, LNICST 285, pp. 254–261, 2019.
https://doi.org/10.1007/978-3-030-23943-5_19

important factors such as library usage, attendance on campus, attendance in the tutorial and the use of the study portal. The software triggers the NTU instructor when the student's rate of engagement is decreasing. The use of predictive analysis can be useful for institutions to raise their profit, revenue, and financial planning. Education organizations are adopting data science practices to predict enrollment trends and operational needs [4]. Moreover, the number of dropouts has become a serious issue for the education organization. Completion and persistence rates are important because they measure how well an institution is serving its students.

According to the National Student Clearinghouse Research Center (2016), the persistence rate from year one to year two is 72.1%. Persistence rates continue to decline during and after the second year of colleges. The total completion rate of students who start college and completed within six years is 54.8% nationally. The persistence and completion rates are lower for part-time students. Rates of completion and persistence vary between two and four-year colleges, public and private, full-time versus part-time students [5]. In order to contribute to the domain of predictive analytics in higher education, in this research, we carried out an in-depth review of the research in the area and compared machine learning algorithms to evaluate their accuracy rate. The rest of this paper is structured as follows. Section 2 elaborates the predictive analytics models and other approaches developed using machine learning algorithms to achieve a wide variety of goals in the education sector. Section 3 briefly describes the methodology of this research. Section 4 contains a detailed discussion of our experiment results. Finally, we have discussed the future direction of this research in Sect. 5.

2 Related Work

A predictive analysis system was built to measure student satisfaction level towards an online course program based on data collected from students who enrolled in the summer-session at western university [6]. Each of the learners was given a survey that included questions on demographics, student satisfaction, and five predictor variables. Correlation, regression, and anova analysis were used to build the prediction model. The study found that the interaction framework with the inclusion of two predictors is internet self-efficacy and self-regulation. Instructors are encouraged to design more collaborative activities to enhance learners' interaction, and student satisfaction level can be improved by performing internet related-tasks. Furthermore, student retention is one of the major issues in higher institutions, especially in the online course program. Several researchers found that the rate of dropouts is increasing [7–9]. Shimin et al. [10] built prediction models using RapidMiner 5.3 that can predict whether a student will engage further after registering to a specific course. The prediction variables are built using J-48 and J-Rip decision tree [10]. The two algorithms tested resulted in producing a high-performance model that provides indicators for predicting the future of a student who has registered in a specific course program. A study was conducted in Delhi Technological University on building predictive analytics model based on the data collected from the National Informatics Center Delhi with the goals of predicting which students will enroll in the particular course and what is the current demand, which programs are trending, and which are becoming obsolete? [11]. The prediction models were built using the decision tree and neural network techniques. The outcomes show that the highest precision indicators were found in decision tree modeling.

The Social Networks Adapting Pedagogical Practice (SNAPP) is a predictive analytics software that is used by University of Wollongong to generate data visualization of user interaction, activity patterns of behavior on the forum [12]. SNAPP is used to map learners' level on engagement and activity to identify learners who are at risk of failing a subject due to lower participation. The tools are used to generate data reports, which include monitoring login frequency, dwell time and number of downloads. Connect for Success is another example of predictive analysis software used by Edith Cowan University (ECU), the system works based on enrolment data, and it is an early warning tool that is used to improve learner success and to improve graduation rates [13]. Automated Wellness Engine (AWE) is an alert system that is designed and built to improve learner engagement and retention rates at the University of New England. AWE is a software-based prediction model that is using emoticons to identify activity embedded in the student portal (myUNE) and another system that identify learners' interaction with the university and instructor. Based on the data collected by AWE, it is smart enough to predict learners who are at high-risk or struggling or may be experiencing disengagement from their course [14]. Open University Australia (OUA) developed Personalized Adaptive Study Success (PASS). PASS is a software tool that is used for predictive analysis to enhance student engagement and retention in an online environment. The model was built based on the individual characteristic, social web, curriculum and physical data collected from some systems [15]. The software assists learners to be aware of their upcoming academic performance and suggests what the students could do better to improve their performance.

3 Methodology

The data was obtained from the Kaggle [16, 17]. It contains 480 student records in rows and 16 features in the columns. The features are classified into three major groups such as Demographic Features (DF), Academic Background Features (AF), and Behavioral Features (BF). The demographic features are nationality, gender, place of birth, and parent responsible for the student. The academic background features consist of the educational stage, grade level, section ID, semester, topics and student absence days. The behavioral features are discussion groups, raised a hand in class, opening resources, viewing announcements, answering the survey by parents and parent school satisfaction. We have utilized these features using traditional classification methods such as random forest and Naive Bayes. Moreover, we applied ensemble algorithms to choose the correct features and to predict the students' performance with high accuracy. The data provider already pre-processed the data. There were no missing fields. The data was pre-processed again before the analysis using R software. The data was split into 75% for training and 25% for testing.

4 Results and Discussion

We used a combination of ensemble techniques that will improve the accuracy of machine learning algorithms results by reducing the variance (bagging), bias (boosting) and improve the accuracy of the prediction (stacking). However, we have compared the results using traditional methods such as random forest and naive Bayes. We tested the behavioral features such as group discussion and resources visited to determine whether these features are contributing factors in students' performance. Firstly, we checked the importance of the features of the dataset. The variables are shown in Fig. 1 circular bar chart and valued by importance in Table 1.

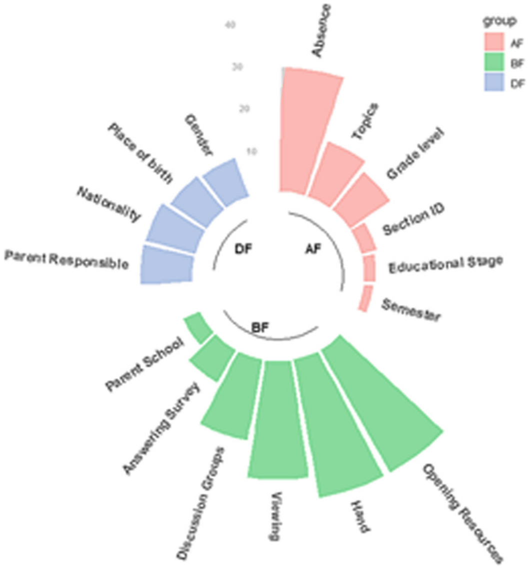

Fig. 1. The categories of the features.

We conclude that behavioral features are the most important features among the three groups of the features from the circular bar chart as shown in Fig. 1. Secondly, academic background features and finally demographic features. We used traditional classification methods to classify the students' performance and to determine the independent variables that offer the highest accuracy. The students' performance was classified using the random forest and Naive Bayes using all the independent characters. The most important 10 features are highlighted with bold font in Table 1. The best (10, 11 and 12) features were selected to classify the performance of the students. The data with the 11 features using random forest gave the best accuracy value (0.7333) as shown in Table 2. The more features taken does not promise higher accuracy. However, we must choose the correct features to improve the accuracy of the algorithms.

Table 1. The importance of features.

Features category	Feature	Importance	Place
Demographic Features	**Parent responsible for student**	**11.95**	**7**
	Nationality	**11.84**	**8**
	Place of birth	**10.10**	**10**
	Gender	9.42	11
Academic Background Features	**Student absence days**	**29.02**	**3**
	Topics	**13.64**	**6**
	Grade level	**11.38**	**9**
	Section ID	3.55	14
	Educational stage	2.73	15
	Semester	2.02	16
Behavioral Features	**Opening Resources**	**32.96**	**1**
	Raised hand in class	**32.94**	**2**
	Viewing Announcements	**27.77**	**4**
	Discussion groups	**19.16**	**5**
	Answering survey by parents	8.69	12
	Parent school satisfaction	4.20	13

All six behavioral variables were not selected as features to classify the performance in this case. However, we tried using one behavioral feature either group discussion or visited resources. The accuracy increased to 0.6917 when the feature visited resources was included in the classification algorithms. This shows that the behavioral variable visited resources is an important feature, and it should be included in the feature selection. This can be seen in Tables 1 and 3. It has established that the better importance of the features improves accuracy. Finally, we have utilized three ensemble methods namely boosting, bagging and stacking to improve the accuracy using all the sixteen features. The boosting machine learning algorithm namely basic tree model (C5.0) and stochastic gradient Boosting model (gbm) were used. The bagging algorithms such as bagged CART (treebag) and random forest (rf) were used. The results are given in Table 4. The ensemble stacking sub-models namely random forest, gbm, K Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) were used and the results can be seen in Table 5. The accuracy of bagging random forest had outperformed the other methods with the accuracy of 0.7959. In this study, the ensemble bagging random forest had given the best result. We need to take more combination of features to increase the accuracy of the performance of the student. We can also improve the result by including more features such as hours spent in the module and the students' interest in the particular module to classify the students' performance in the future study.

Table 2. Accuracy using Random forest and Naive Bayes based on the importance of the features.

			95% Confidence interval	
		Accuracy	LB	UB
All 16 Independent variables	Random Forest	0.725	0.636	0.8025
	Naive Bayes	0.625	0.532	0.7117
10 Best Independent Variables	Random Forest	0.7	0.6096	0.7802
	Naive Bayes	0.6583	0.5662	0.7424
11 Best Independent Variables	Random Forest	**0.7333**	0.6449	0.8099
	Naive Bayes	0.6667	0.5748	0.7501
12 Best Independent Variables	Random Forest	0.7083	0.6184	0.7888
	Naive Bayes	0.5917	0.4982	0.6805

LB = Lower boundary UB = Upper boundary

Table 3. Accuracy with and without behavioral features using Random forest and Naive Bayes.

			95% Confidence interval	
		Accuracy	LB	UB
Without all behavioral variables	Random Forest	0.6583	0.5662	0.7424
	Naive Bayes	0.6	0.5066	0.6883
Behavioural variables *group discussion*	Random Forest	0.6583	0.5662	0.7424
	Naive Bayes	0.5417	0.4483	0.6329
Behavioural variables *visited resources*	Random Forest	**0.6917**	0.6009	0.7727
	Naive Bayes	**0.6917**	0.6009	0.7727

Table 4. Accuracy using boosting and bagging ensemble algorithm.

		Min	1st Qu	Median	Mean	3rd Qu	Max	NA's
Boosting	**C5.0**	0.6327	0.7262	0.7732	0.7771	0.8084	0.9375	0
	gbm	0.6531	0.7500	0.7708	0.7626	0.7917	0.8367	0
Bagging	**treebag**	0.6596	0.7372	0.7836	0.7749	0.8154	0.9167	0
	rf	0.6809	0.7672	0.8125	**0.7959**	0.8333	0.8750	0

Table 5. Accuracy using an ensemble stacking algorithm.

		95% Confidence interval	
	Accuracy	LB	UB
Random Forest	0.7396	0.64	0.8238
GBM	0.7188	0.6178	0.8058
LDA	0.7500	0.6512	0.8328
KNN	0.6354	0.5309	0.7313

5 Future Direction

Likewise, most of the recent research trends of technology such as Big Data, Cloud Computing, and Edge Computing, predictive analytics has several barriers to its implementation in education organization and various other sectors where security, privacy, quality and transparency of data play a critical role. Both instructors and students have raised data privacy and ownership concerns. The management must undertake the initiative to ensure ethics and permissions are maintained at all levels. As an example the University of California has addressed the concern of how students data privacy is violated, as the institution use a third party service platforms to gather data, the management realized that the vendors were using the student data to commercialize or sell their product [18]. There is an utmost need to formulate secure and privacy preserved data collection as well as analysis techniques. Most importantly, users of the data must be aware of the fair usage and treatment of their records. In education organizations, predictive analytics must be carried out in compliance with legal standards such as FERPA. Another concern raised on predictive analytics is the ethics, transparency and legal compliance on the use of data. Furthermore, data must be gathered from trusted sources, and systems must be in place to ensure that data is reliable, and analyzing such data would lead to suitable insights and appropriate actions. Predictive analytics also requires domain experts as well as data scientists [15]. The human brain has to make the decision, not analytic tools, hence the number of experts, background and wisdom do matter in making a decision. Data is just a help to produce a result. Lack of vision and familiarity is considered as the major barriers to predictive analytics.

Formulating predictive analysis modeling requires a lot of thought process to address the problem and the goal of predictive analytics modeling. The lack of data warehousing among institution makes the process of predictive analysis as a challenge. An institution cannot get the right type of data if there is no central repository of data that is accessible and transparent. The lack of data warehousing is also affecting the complexity of analytical tools. One of the major barriers to implementing data analytics in higher education is cost. Management tends to view analytics as an investment as the tools are expensive and tend to think it will not return the investment. To come up with predictive analysis, the first spending would be in human resources such as hiring a data analyst. The second comes to the data warehousing which includes the infrastructure, platform and services that are being purchased to build predictive analytics [19]. Apart from these barriers, the accuracy of the predictive model is also very important to make better decisions. Depending on the features of the dataset, a predictive model should be generated by trying a combination of various machine learning algorithms, and the model should be validated to obtain optimum accuracy. For example, in this research, based on our experience of data analytics, first, we found that random forest and naive bayes gave the accurate results, and then we enhanced the accuracy by using ensemble techniques with random forest. In the future direction of this research, we plan to use advanced deep learning algorithms to build a predictive analytics model for higher education.

References

1. Harel, E., Sitko, T.: Digital Dashboards: Driving Higher Education Decisions. Educause Center for Applied Research, Boulder (2003)
2. Johnson, L., Levine, A., Smith, R., Stone, S.: The 2010 Horizon report. The New Media Consortium, Austin, TX (2010). http://wp.nmc.org/horizon2010
3. Watson, H.J.: Business analytics insight: hype or here to stay? Bus. Intell. J. **16**(1), 4–8 (2011)
4. Burke, M., Parnell, A., Wesaw, A., Kruger, K.: Predictive analysis of student data (2017). https://www.naspa.org/images/uploads/main/PREDICTIVE_FULL_4-7-17_DOWNLOAD.pdf
5. Shapiro, D., et al.: Completing College: A National View of Student Completion Rates – Fall 2011 Cohort (Signature Report No. 14), December 2017. National Student Clearinghouse Research Center, Herndon, VA (2017)
6. Long, P., Siemens, G.: Penetrating the fog: analytics in learning and education. EDUCAUSE Rev. **46**, 30 (2011). http://net.educause.edu/ir/library/pdf/ELI7079.pdf
7. Willging, P.A., Johnson, S.D.: Factors that influence students' decision to dropout of online courses. J. Asynchronous Learn. Netw. **13**(3), 115–127 (2009)
8. Boston, W.E. et al.: Comprehensive Assessment of Student Retention in Online Learning Environments. School of Arts and Humanities, APUS. Paper 1 (2011)
9. Hoskins, S.L., Van Hooff, J.C.: Motivation and ability: which students use online learning and what influence does it have on their achievement? Communications **36**(2), 177–192 (2005)
10. Kai, S., et al.: Predicting student retention from behavior in an online orientation course
11. Hawkins, B.L.: Accountability, demand for information, and the role of the campus IT organization. In: Katz, R.N. (ed.) The Tower and the Cloud, pp. 98–104. Educause, Boulder (2008). www.educause.edu/thetowerandthecloud/PUB7202j
12. Bakharia, A., Dawson, S.: SNAPP: a bird's-eye view of temporal participant interaction. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge, pp. 168–173 (2011)
13. Jackson, G., Read, M.: Connect 4 success: a proactive student identification and support program, pp. 1–5. ECU, Australia (2012). fyhe.com.au/past_papers/papers12/Papers/9B.pdf
14. Leece, R., Hale, R.: Student engagement and retention through e-Motional intelligence. UNE, Australia (2009). http://www.educationalpolicy.org/events/R09/PDF/Leece_E-Motion.pdf
15. Atif, A., Richards, D., Bilgin, A., Marrone, M.: A panorama of learning analytics featuring the technologies for the learning and teaching domain. In: Carter, H., Gosper, M., Hedberg, J. (Eds.) Electric Dreams. Proceedings ascilite 2013, Sydney, pp. 68–72 (2013)
16. Amrieh, E.A., Hamtini, T., Aljarah, I.: Mining educational data to predict student's academic performance using ensemble methods. Int. J. Database Theor. Appl. **9**(8), 119–136 (2016)
17. Amrieh, E.A., Hamtini, T., Aljarah, I.: Preprocessing and analyzing educational data set using X-API for improving student's performance. In: 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), November 2015, pp. 1–5. IEEE (2015)
18. Nissenbaum, H.N.: Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford Law Books, Stanford (2010)
19. Denley, T.: How predictive analytics and choice architecture can improve student success. Res. Pract. Assess. **9**(2), 61–69 (2014)