IDENTIFYING STUDENTS AT RISK FROM ONLINE CLICKSTREAM DATA USING MACHINE LEARNING

Hadeel Alhabdan[1] and Ala Alluhaidan[2]

[1]College of Computing and Information Sciences, Princess Nourah bint Abdulrahman University,Riyadh, Saudi Arabia
[2]Department of Information Systems, College of Computing and Information Sciences, Princess Nourah bint Abdulrahman University,Riyadh, Saudi Arabia

ABSTRACT

This study examines the use of four machine learning methods to identify students at risk from online clickstream data for 60 courses and the students' grades in these courses. To identify students at risk of failing, the study classified students with grades of "F" or "D" as at-risk, while students with grades of "A," "B," or "C" were classified as safe. Logistic regression, decision tree, neural networks and random forest models were used, with each model subjected to eight folds cross-validation. The decision tree model had the lowest performance across all four metrics, followed by the logistic regression model, while the neural network model showed marginally superior accuracy, sensitivity, and F1 score compared to the random forest model. The four machine learning models were found to be reliable in identifying at-risk students based on the provided online clickstream data.

KEYWORDS

Decision tree, Logistic regression, Neural networks, Online clickstream data, Random Forest.

1. INTRODUCTION

The increase in online learning in higher education is primarily due to the rapid development of Internet technology. The COVID-19 pandemic has further impacted the education system, resulting in a shift from offline to online courses [1].

Machine learning techniques have been used to predict student performance and predict both good and bad outcomes. Early prediction of academic performance is important to improve learning outcomes and increase graduation rates. It also serves as a basis for university policies, teaching practices, evaluation of learning effectiveness, feedback from teachers and students, and adaptation of learning environments [1, 2].

By enrolling in online classes, students can acquire knowledge from any location in the globe, at their preferred speed. Nevertheless, like any pedagogical approach, it possesses both benefits and limitations.

David C. Wyld et al. (Eds): NLAI, IBCOM, MLDS, GridCom, SCAI, ICCSEA, SPPR – 2024
pp. 17-24, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142202

Online courses offer exceptional convenience. Individuals can manage their academic pursuits alongside employment or familial obligations, as they can access course materials at their convenience from any location with an internet connection. The interactive elements, such as online forums and virtual classes, cultivate a feeling of cooperation among students [3].

On the other hand, online learning might provide difficulties. The absence of direct interpersonal engagement with educators and classmates may delay the development of social and cooperative learning. Self-discipline is essential for students to successfully manage their time and prevent procrastination. Problems with technology and challenges with internet connectivity have the potential to impair the process of acquiring knowledge [3, 4].

Clickstream data provides higher education institutions with an effective tool for identifying students at risk of failing online courses. This data, gathered from students' interactions with learning management systems, provides essential insights into their learning habits, levels of engagement, and potential difficulties.

Universities can identify disengaged students by studying clickstream data, which includes lower login frequency, limited interaction with course materials, and avoidance of specific modules. Early detection of these trends enables early interventions such as designed academic advice, peer tutoring, and additional support services. Furthermore, clickstream data can assist identify specific areas in which students are struggling, allowing for focused interventions and resource allocation [5].

The objective of this paper is to compare the performances of four machine learning methods to predict at-risk students using clickstream data.

A clickstream dataset that contains information on student interactions, participation in forums and assessment results will be considered for constructing machine learning models and measuring their performances. Cross-validation techniques will be used to assess the accuracy and performance of the developed machine learning models.

The rest of this paper is organized as follows. In Section 2 is the related work. The methodology is in Section 3. Section 4 discusses the results. Finally, in Section 5 are the conclusions.

2. RELATED WORK

This section presents some recent literature on machine learning techniques to analyze and predict student performance in online courses based on study habits.

In [6], Holicza and Kiss used a machine learning algorithm to predict and test student performance decline. The study compared online and offline learning data and found that success in school was related to habits such as sleep, study time, and screen time.

McIntyre [7] aimed to identify key features for accessing online learning in low- and middle-income countries, especially for girls, due to the coronavirus disease (COVID-19) pandemic. This study used data mining and machine learning models to analyze 54,842,787 data points from online learning platforms. Country differences, gender, and COVID-19 have been identified as important characteristics in access to online learning. The data-driven model also provided additional insights into factors such as math skills, year of birth, session difficulty, and time taken to complete the session.

A study by Zhang et al. al [8] proposed an e-learning performance prediction framework based on behavior classification, which uses feature fusion to identify e-learning behaviors. Furthermore, a process behavior classification model that considers the learning process was introduced. Experimental results showed that the BCEP prediction framework showed a good prediction effect and the PBC model outperformed traditional classification methods. This new approach provided a quantitative evaluation solution for e-learning classification methods.

Gao et al. al [9] proposed a deep cognitive diagnosis framework to improve traditional cognitive diagnosis methods through deep learning. It modeled student competency skills based on students' responses to objective and subjective problems and considered attention mechanisms and neural networks. The model predicted student performance based on careless choices and guesses.

In Liu et al., [10] proposed a model to predict student performance based on evolutionary spiking neural networks. The model analyzed the relationship between course and student attributes and used an evolutionary membrane algorithm to improve accuracy. The model was tested on two benchmark datasets and compared with other experimental algorithms. The results showed that this model effectively improves the prediction accuracy of student grades and provides early warning and timely correction for students and teachers.

In Xu et al [11], a model was proposed to predict student performance based on evolutionary spiking neural networks. The model analyzed the relationship between course and student attributes and used an evolutionary membrane algorithm to improve accuracy. The model was tested on two benchmark datasets and compared with other experimental algorithms. The results showed that this model effectively improves the prediction accuracy of student grades and provides early warning and timely correction for students and teachers.

A study by Ramaswamy et al. al [12] present the development of a general predictive model to identify at-risk students in different courses. The CatBoost algorithm performs best when dealing with categorical and missing data, making it a good candidate for solutions on educational datasets.

Zhang et al. al [13] used a tree-based machine learning algorithm to predict the academic performance of undergraduate students in a Chinese university. His three models were created: decision trees, gradient boosted decision trees, and random forests. The results showed that the RF model can identify more than 80% of underperforming and at-risk students, improving the quality of teaching and learning.

3. METHODOLOGY

The following formatting rules must be followed strictly. This (.doc) document may be used as a template for papers prepared using Microsoft Word. Papers not conforming to these requirements may not be published in the conference proceedings.

3.1. Dataset Description

The e-learning data consists of two datasets in excel format, for the Hijri year 1444. The first dataset has 7 attributes with 48480 entries. Three attributes are categorical, while the other four are numerical. The dataset contains statistics on students' interactions with online courses offered during the Hijri year 1444.

The dataset includes the following features: Course code, Reference No, Student ID, Average number of visits to the course, Average entry time for exams, Average Interactions inside course, and Average Time spent in the course.

There are 62 courses listed under the course code feature in the dataset. The dataset reference number field has 595 distinct values. Finally, the dataset student ID field includes 17131 values. The second dataset has 7 attributes. These attributes are named Course code, Course Number, Semester, Reference number, Student ID, Score and Description.

3.2. Data Preprocessing

At the beginning we removed the duplicated records from both given datasets. The two databases share the fields StudID, Course, and RefNo. We assigned the variables StudID and Course as the index keys in both datasets. The two datasets were combined into a single dataset.

The number of records that contain missing values is 1977 records. This refers to the fact that there are students in the first dataset without results in the second dataset. By eliminating all the records containing missing values from the new dataset, 45313 students with results will remain. Some students having online clickstream data withdrew from certain courses, and their scores in these subjects were designated as 'W' or ‘DN’ to indicate the course withdrawal or denial of student from final exam. As a result, their online data proved ineffective for classification purposes. Records with 'W' and ‘DN’ Grade are likewise removed from the dataset.

The dataset contains 4 numerical fields (CourseEntryAvg, ExamEnteryAvg, InteractionAvg, CourseTimeSpent) and three categorical fields (StudID, Course and Grade).

The failed students and students whose scores are D were considered at-risk students and are represented by class 0, whereas the classes C, B and A are considered safe and represented by number 1. Using the MinMax normalization on all the 7 fields of the PNU dataset, a normalized dataset is obtained.

We calculated the correlation matrix for the whole dataset to find whether there are any relationships between the features of Average number of visits to the course, Average entrance time for tests, Average interactions within the course, and Average time spent in the course. The correlation matrix is explained in Figure 3.1 below.

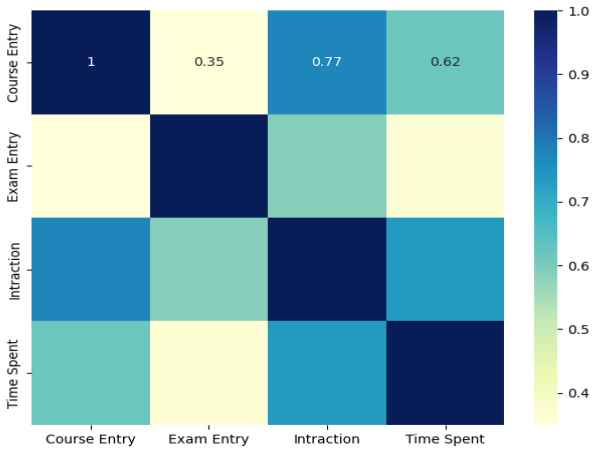

Figure 0.1 Correlation matrix of the PNU dataset features

The correlation matrix shows that the average interactions within the course feature has the highest correlation with the other features, followed by the average time spent in the course feature. Conversely, the average entrance time for tests has the weakest correlation with the other features.

3.3. Data Classification

To develop machine learning binary classifiers for identifying students at risk, we consider both the Students with F and D grades to be at risk, whereas the students with grades C, B and A safe. Hence, we would have a binary classification task.

We will consider four different kinds of machine learning classifiers:

a) Logistic Regression model
b) Decision Tree model
c) Neural network model based on Python keras.
d) Random forest model

where all the models will employ cross validation techniques to avoid the overfitting problem. The logistic regression, decision tree and random forest models are implemented by using the Python's scikit-learn package, while the neural network models are implemented based on Tensorflow Keras package. We will use the accuracy, precision, recall and F1 score metrics as the main evaluation measures for any of the machine learning models under consideration.

4. RESULTS

Figure 4.1 shows that the neural network, random forest and logistic regression models have very close accuracy average percentages. The neural network shows slightly better accuracy over the random forest model and the random forest model shows better accuracy percentages over the logistic regression model. The least accuracy performance is obtained by the decision tree model.

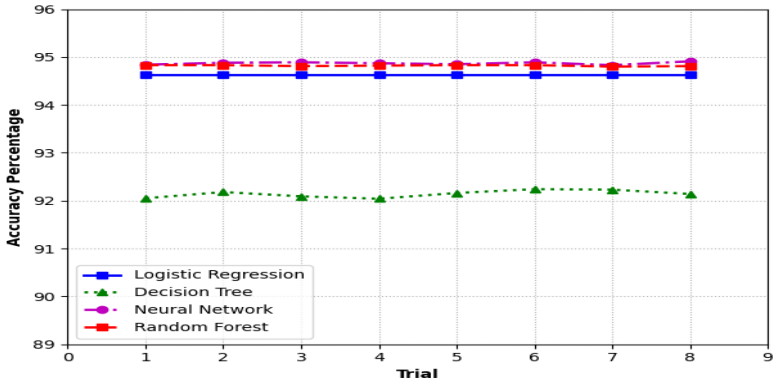

Figure 0.1 Average accuracy percentages of logistic regression, decision tree, neural network and random forest models.

Figure 4.2 shows that the precision obtained by the random forest model is better than the neural network model and the neural network model is better than the decision tree model. The logistic regression model gives the least precision among the four models.

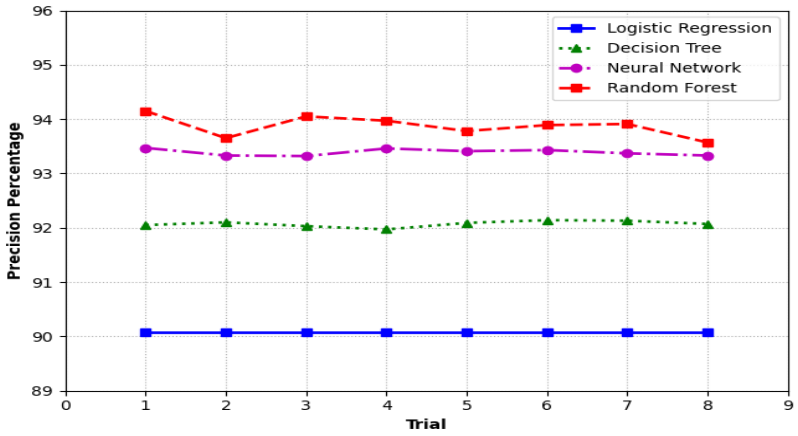

Figure 0.2 Average precision percentages of logistic regression, decision tree, neural network and random forest models.

The recall values illustrated in Figure 4.3, shows close behavior as accuracy. The neural network, random forest and logistic regression models have very close recall average percentages. The neural network shows slightly better recall over the random forest model and the random forest model shows better recall percentages over the logistic regression model. The least recall performance is obtained by the decision tree model.

Finally, in Figure 4.4, we see that the neural network gives slightly better f1 score over the random forest model and the random forest model gives slightly better score over the logistic regression model. The decision tree model gives the least f1 score among the four machine learning models.

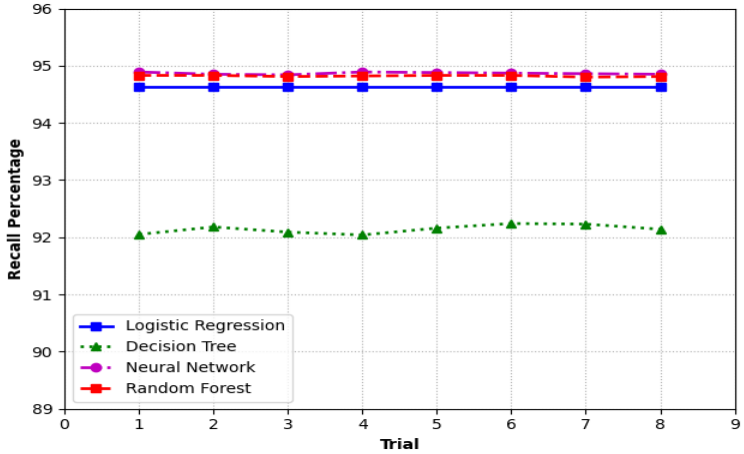

Figure 0.3 Average recall percentages of logistic regression, decision tree, neural network and random forest models.

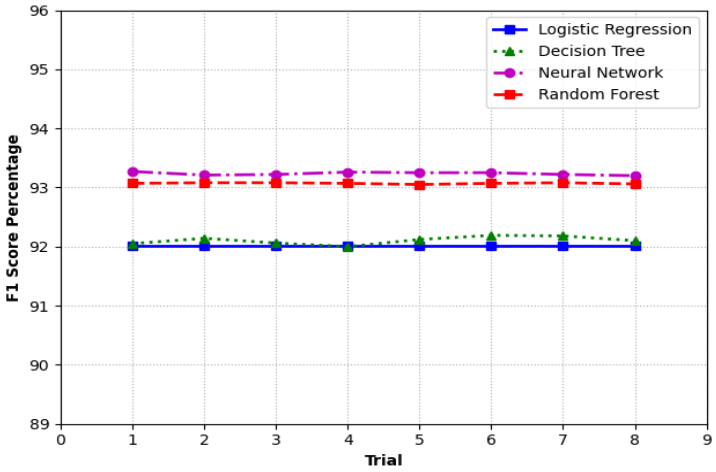

Figure 0 Average F1 score percentages of logistic regression, decision tree, neural network, and random forest models.

5. CONCLUSIONS

This study analyzes internet clickstream data to identify at-risk students. Given two datasets for this study: one including online clickstream data for 60 courses and the other having student outcomes in these courses.

The data preprocessing included linking two datasets, eliminating records with missing values, excluding records with grades 'W' and 'DN', normalizing both categorical and numerical variables (by converting the categorical values to numerical values and then using the minmax normalization), and finally merging certain pairwise grades into single grades to reduce the target variable.

Students with grades of 'F' or 'D' were classed as at-risk students, whereas students with grades of 'A', 'B', or 'C' were classified as safe. As a result, the problem is turned into a binary classification problem.

The logistic regression, decision tree, neural network, and random forest models were tested for their ability to identify students at risk. Each machine learning model went through eight trials, each consisting of a 10-fold cross-validation. Figures 4.1, 4.2, 4.3, and 4.4 show that the decision tree model performed the least well on all four metrics when compared to the other three models. The neural network model exhibited slightly higher accuracy, recall, and F1 score than the random forest model (figures 4.1, 4.3, and 4.4). In comparison to the logistic regression model, both models performed better in terms of accuracy and recall. Figures 4.2 indicate that the random forest model has more accuracy than the other three models, followed by the neural network model.

The four machine learning models used to detect students at risk produced respectable results, demonstrating that these models are dependable in identifying students at risk based on the given online clickstream data.

ACKNOWLEDGEMENTS

The authors would like to thank everyone, just everyone!

REFERENCES

[1] Steven C.H. Hoi, Doyen Sahoo, Jing Lu, Peilin Zhao, Online learning: A comprehensive survey, Neurocomputing, Volume 459, 2021, Pages 249-289

[2] Yan, N. and Au, O.T.-S. (2019), "Online learning behavior analysis based on machine learning", Asian Association of Open Universities Journal, Vol. 14 No. 2, pp. 97-106.

[3] Hai, L.; Sang, G.; Wang, H.; Li, W.; Bao, X. An Empirical Investigation of University Students & Behavioural Intention to Adopt Online Learning: Evidence from China. *Behav. Sci.* 2022, *12*, 403.

[4] Ofori, F.; Maina, E.; Gitonga, R. Using machine learning algorithms to predict students's performance and improve learning outcome: A literature-based review. *J. Inf. Technol.* 2020, *4*, 33–55.

[5] Arcinas, M.M. Design of Machine Learning Based Model to Predict Students Academic Performance. *ECS Trans.* 2022, *107*, 3207.

[6] Holicza B, Kiss A. Predicting and Comparing Students' Online and Offline Academic Performance Using Machine Learning Algorithms. *Behavioral Sciences*. 2023; 13(4):289.

[7] McIntyre, N.A. Access to online learning: Machine learning analysis from a social justice perspective. Educ Inf Technol 28, 3787–3832 (2023)

[8] Qiu, F.; Zhang, G.; Sheng, X.; Jiang, L.; Zhu, L.; Xiang, Q.; Jiang, B.; Chen, P.k. Predicting students' performance in e-learning using learning process and behaviour data. *Sci. Rep.* 2022, *12*, 453.

[9] Gao, L.; Zhao, Z.; Li, C.; Zhao, J.; Zeng, Q. Deep cognitive diagnosis model for predicting students' performance. *Future Gener. Comput. Syst.* 2022, *126*, 252–262.

[10] Liu, C.; Wang, H.; Du, Y.; Yuan, Z. A Predictive Model for Student Achievement Using Spiking Neural Networks Based on Educational Data. *Appl. Sci.* 2022, *12*, 3841.

[11] Xu, J.; Moon, K.H.; van der Schaar, M. A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE J. Sel. Top. Signal Process.* 2017, *11*, 742–753.

[12] Ramaswami, G.; Susnjak, T.; Mathrani, A. On Developing Generic Models for Predicting Student Outcomes in Educational Data Mining. *Big Data Cogn. Comput.* 2022, *6*, 6.

[13] Zhang, W.; Wang, Y.; Wang, S. Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in China. *Educ. Inf. Technol.* 2022, *27*, 13051–13066.