As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 9(02): 102-107, 2023
Copyright © 2023 International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0902019
Available online at: http://www.ijmtst.com/vol9issue02.html

Detecting At-Risk Students with Early Interventions Using Machine Learning Techniques

Dr.D.Suneetha[1] | K. Tejaswini [2]| M. Sankar [2] | M. Kiran kumar [2]| K. Akhil [2]

[1]Professor & HOD, Department of CSE, NRI Institute of Technology, India
[2]B.Tech Student, Department of CSE, NRI Institute of Technology, India

To Cite this Article
Dr.D.Suneetha, K. Tejaswini, M. Sankar, M. Kiran Kumar, K. Akhil. Research on Detecting At-Risk Students with Early Interventions Using Machine Learning Techniques. International Journal for Modern Trends in Science and Technology 2023, 9(02), pp. 102-107. https://doi.org/10.46501/IJMTST0902019

Article Info
Received: 04 January 2023; Accepted: 01 February 2023; Published: 04 February 2023.

ABSTRACT

In recent years, massive open online courses (MOOCs) have flourished, giving students easy access to informative videos, podcasts, and other online resources. Facilitated learning and adaptability in the classroom are driving a surge in enrollment. Despite this, a large body of evidence indicates that poor completion and high attrition rates are serious issues. This research offers a method for spotting at-risk pupils before they drop out of school entirely. As a result, we build not one but two models: one to identify students at risk and another to measure their academic progress. The models may help identify at an early stage in an online course those students who are likely to fail and drop out. All classifiers perform well across both models, with GBM achieving the maximum performance (with values of 0.894 and 0.952 for the first and second models, respectively) and RF achieving the lowest performance (yielding a value of 0.86) in the at-risk student framework. The suggested frameworks may be utilized to help teachers provide extra help to kids who need it most during times of crisis.

KEY WORDS: Machine learning, massive open online courses, receiver operator characteristics, area under curve.

1. INTRODUCTION

The use of information and communication technology (ICT) has become more common and now plays an essential part in the instructional process. The advancement of information and communications technology has helped to bolster the support of the academic curriculum and enables the construction of a virtual classroom. ICT has the potential to enhance student results and gives teachers the ability to assist pupils in working through problems. As a result, high-quality instruction may be provided via the use of virtual learning [1].

The rapid expansion of information and communications technology (ICT) has stimulated the expansion of massive open online courses (MOOCs) in higher education. MOOCs provide a wide range of multimedia capabilities, making it possible to establish a learning environment that is participatory. Students have access to material from all around the globe because to the availability of MOOCs, which are great digital learning tools [2].

As a result of the elimination of the financial and geographical barriers that were connected with the conventional method of instructing students, a number of the most prestigious educational institutions have begun offering courses online as an alternative to the conventional mode of instruction. Because of the quick ShiruiPan served as the assistant editor who was responsible for managing the evaluation of this manuscript and giving his or her approval for it to be published. The poor completion rates of massive open online courses (MOOCs) is a key concern connected to the proliferation of online courses in higher education [3].

One of the measures that may be utilized to increase completion rates is the identification of students who are considered to be at risk. Finding pupils who are likely to struggle academically and doing so in a timely way may assist teachers in delivering instructional interventions and improving the structure of classes [4]. Because teachers will be able to provide students feedback in real time if they have access to a solution for prompt intervention [5,] retention rates will likely increase. Researchers looked into the factors that led to the cancellation of the class.A variety of causes have been proposed to account for this phenomenon.The lack of interest on the part of students is the primary factor that leads to their withdrawing from their online classes [6]. According to the findings of a recent study [7], researchers believe that students' levels of motivated engagement in online classes may either drop or rise depending on social, cognitive, and environmental aspects. The progression of a student's level of motivation is a key determinant of whether or not they will drop out of school. Investigating how a learner's behavior changes during a series of classes is one way to quantify motivational trajectories [7]. The vast majority of studies, up to this point, have not paid much attention to analyzing the link between motivational trajectories, student learning accomplishment, and at-risk pupils in the context of an online environment. Predicting whether or not students will complete massive open online courses (MOOCs) may give significant information that can assist instructors in identifying students who are at risk early on. Despite the fact that a number of publications have been published in the literature offering strong learning frameworks for online courses, it is still difficult to obtain a high prediction accuracy of student performance in the long-term spanning numerous datasets [8, 9].

In the course of this study, two case studies will be carried out. In the first research, a unique dropout predicting model is proposed. This model is able to provide timely intervention help for students who are at danger of dropping out of school. The use of machine learning allows for the identification of probable patterns of learner attrition based on the activities of the course as well as the previous behavior of the learners themselves. The level of student involvement, in combination with a student's level of motivation in prior courses, was analyzed in order to determine how each factor influences a student's decision to continue participating in the current course. A model for predicting the academic achievement of students is provided in the second case study. The model provides fresh insight into the fundamental aspects of learning activities and may assist instructors in the process of tracking the academic progress of their students. The use of machine learning allows for the tracking of student performance and the provision of vital information to educators, allowing for the development of courses in accordance with the students' levels of educational attainment. In addition to this, it might assist academic advisers in identifying students who are not doing well academically and providing assistance for such students.

The remaining parts of this work are structured in the following manner. In the next section, "Section II," you will get an overview of the most recent studies in the topic. In section III, the methodology of the suggested approach is described. This includes a description of the dataset, several methodologies, and the results of simulations. The findings of this study, as well as potential directions for further investigation, are discussed in Section IV.

2. LITERATURE SURVEY

In massive open online courses (MOOCs), one of the primary areas of concern is student attrition and learning outcomes. In this part, we present an overview of the most recent studies on the identification of kids who are at risk of dropping out of school or not achieving the desired level of academic success.

Using student attitudes and clickstream data as baseline characteristics, feedforward neural networks were constructed in [10] in order to identify students in MOOCs who were at danger of failing the course. In 2014, the data was compiled from three million student click logs and five thousand student forum postings that were gathered via the Coursera platform. Managing a dataset that had inconsistencies was one of the primary challenges posed by this investigation. This problem was solved by using Cohen's Kappa criterion rather than accuracy in the analysis. When both sets of characteristics were included in the analysis, the findings revealed an accuracy of 74%. When we took off the emotional characteristics, this number dropped to 70 percent.

Using a number of different machine learning methods, such as regularized logistic regression, support vector machines, random forest, decision tree, and Naive Bayes, the researchers in [11] were able to identify pupils who were considered to be at risk. A number of characteristics were extracted from the behavioral log data. These characteristics included the amount of time that students spent on the home page as well as the number of times they visited the page overall. According to the findings, models using regularized logistic regression earned the greatest AUC.

Using a number of different machine learning methods, such as regularized logistic regression, support vector machines, random forest, decision tree, and Naive Bayes, the researchers in [11] were able to identify pupils who were considered to be at risk. The behavioral log data was analyzed to extract a set of characteristics, one of which was the number of times students visited the homepage throughout the session as well as how long each visit lasted. According to the findings, models using regularized logistic regression earned the greatest AUC. A model known as the ConRec Network, which is a deep neural network, was presented in [12]. In this study, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) were coupled in order to determine whether or not students enrolled in the online course 'XuetangX' are likely to drop out of the program within the following ten days. The student records were organized in a way that corresponded to a series of time stamps, and they included a variety of data, such as the event time, the event type, and the date the student enrolled. The hybrid neural network model may be broken down into its two component sections, which are the bottom and higher parts, respectively. In the bottom section, the hidden layer of the CNN was used to automatically extract features. In the top section, RNN was used to generate a forecast by accumulating and combining the characteristics that were extracted at each point in time. The model was evaluated in contrast to a number of standard procedures. The findings suggested that all of the models had comparable levels of performance. The results of the F1-score were found to fall somewhere in the range of 90.74 to 92.48. Although there was a comparable level of performance between the ConRec Network model and the baseline methods, the authors argued that the ConRec Network model is more effective than the baseline methods due to its capacity to automatically extract features from student records without the necessity of feature engineering [12].

Researchers have taken into account a variety of factors in order to determine the level of student learner achievement in an online environment. These factors include the length of time that students spend interacting with digital resources, the time that students submit assessments, the total number of attempts that are made, the educational level, the geographical location, and the gender of the student. In the study [13], Genetic Algorithms (GA) were used in order to achieve optimal performance of the feature set. According to the data, the most highly rated characteristics are those that are associated with behavioral qualities rather than demographic characteristics. For the purpose of predicting student performance, four different classifiers—decision tree, neural network, Naive Bayes, and k-nearest neighbor—were taken into consideration. When the GA-optimized feature set was included into the simulation, it was found that accuracy increased by 12% overall. When the decision tree was utilized with all of the features, the accuracy was 83.87%; however, when the GA-optimized feature set was employed, the accuracy increased to 94.09% [13]. Hidden Markov models were used in order to investigate the extent to which latent factors, when considered in combination with observable variables, may have an effect on student performance in online educational settings. In [8], it was suggested that a two-layer hidden Markov model, abbreviated as TL-HMM, may be used to infer latent student behavioral patterns. The ability of TL-HMM, in contrast to that of conventional HMM, to find the micro-behavioral patterns of pupils in more detail and to identify transitions between latent states sets it apart from conventional HMM. For instance, when students engage in quizzes, they often take part in the corresponding forum discussions as well. Additionally, the model is able to learn precise transitions between the date of the quiz evaluation and the date of submission. According to the findings of the study, kids who do well academically have less latent behavioral states because they have adequate information, and as a result, they do not need further assistance [8].

3.PROPOSED SYSTEM

Two different datasets are used in the studies that we conduct. The first group of data comes from online classes offered by Harvard University and the Massachusetts Institute of Technology, while the second group of data is associated with online classes offered by Open University.

In the process of producing online courses, Harvard University and the Massachusetts Institute of Technology (MIT) worked together. The clickstream, which is the number of events that correspond to a user's engagement with courseware, is the most important property of the Harvard dataset. This quantity is represented by the word count. Activities such as clicking on a chapter or on forum postings, as well as browsing the main page of videos, count toward qualification. Before the user may officially enroll in a class, they have to first register for that class [14]. The user must go through five different web sites in order to finish the registration procedure.

Learners are expected to read a certain number of chapters, which is indicated by the "Nchapters" component. The value that is stored in the variable denoted by the name "Nplay video" is the number of times that the learner watched a certain video. The 'Explored' characteristic is a binary distinction between learners who engage in exploratory activities. A student must have accessed more than half of the material in the course in order to be considered an explorer of the material. The 'Viewed' feature is likewise a binary feature, and it is set to 1 whenever a student accesses the home page of assignments and associated videos [15]. This feature is set to 1 when a student views the page.

In addition to the date that each student had their most recent engagement with the courseware, that date is also recorded in the dataset as part of the learner's participation in a particular course. The educational level of the learners is reflected in the "LoE DI" attribute, which is a demographic characteristic. [15] In addition to "race," "ethnicity," and "nationality," additional sorts of demographic data are also reported. The grade on the assignment is an indication property that reflects the percentage of participants that were successful or unsuccessful. The Harvard dataset is broken down into its component parts and summarized in Table 2.

The Open University in the UK provided this research with the second database to use in their investigation [16]. The University of the Open TABLE 2. Harvard dataset summary. provides a variety of online courses for students at both the undergraduate and graduate levels. The Open University disseminated a graphical user interface known as the Open University Learning Analytics Dataset in the academic year 2013–2014. (OULAD) This dataset includes fields for demographic information, information about behavior, and information about time. It incorporates a collection of tables that pertain to student performance, student personal information, in addition to student engagement elements with online courses. The student has the ability to engage with a variety of different kinds of digital content, including PDF files, access to the main page and subpages, and the ability to participate in quizzes [16]. The Tutor Marked Assessment (TMA) and the Computer Marked Assessment are the two distinct kinds of tests that may be taken (CMA). The final grade is determined by computing the weighted total of all of the assessments (50 percent) and the final examinations (50 percent). The "Student Assessment" table contains information relating to the outcomes of student assessments, such as the date the assessment was turned in and the assessment score. The evaluations are required to be included in the dataset. For this reason, students who want to continue enrolled in the class are expected to complete a series of evaluations, one of which is a final examination. If a student earns an overall grade that is higher than 40% [16], then the student will be considered successful in the class. The OULAD dataset is broken down into its component parts and summarized in Table 3.

The data from the students' Virtual Learning Environments (VLEs) were gathered on a daily basis, and feature extraction was carried out on those datasets. The features that were taken from the VLE depend on the clickstream features. Eleven different kinds of VLE activities are included in the OULAD dataset. We compiled an overall total of the number of clicks that each student participated in throughout each activity during the whole of their participation in the course, from the moment they registered for it to the moment they withdrew their enrollment. Following the same methodology as the previous study [5], we extract twenty-two characteristics from the VLE. The OULAD dataset is summarized in Table 1, which can be found here.

Table 1: OULAD Dataset Overview

Features	Description
id_student	Learner identification number
age_band	Learner age
Gender	Learner gender
highest_education	Learner educational level
Region	Learner geographic area
studied_credits	The number of credits for the module that the learner is currently involved
disability	Indicator of student disability
num_of_prev_attempt	Number of times that student undertook the course
imd_band	Socio-economic indicator measure of student economic level
leaerning activity	The type and number of daily activities that the student undertakes
grades	The student's assessment marks
date_registration	The date of learner registration in the course
date_unregistration	The date that the learner quit the course

"Introduction to Computer Science," "Circuits and Electronics," "Health in Numbers: Quantitative Methods in Clinical & Public Health Research," and "Human Health and Global Environmental Change" are the four courses that have been chosen for analysis in this study based on the Harvard dataset. These four courses are: "Introduction to Computer Science," "Circuits and Electronic

[17] The primary objective of the class titled "Introduction to Computer Science" is to educate students on how to use computation while addressing problems. The subject matter covered in the course titled "Circuits and Electronics" includes an introduction to lumped circuit abstraction. The class was developed specifically for first-year students at the Massachusetts Institute of Technology, but it is now open to students all around the globe and offered online. [18]

The objective of the health research class known as "Health in Numbers: Quantitative Methods in Clinical and Public Health Research" is to educate students in the application of quantitative research techniques to the process of monitoring the medical histories of patients. Students learn how to study how changes in the global environment might influence the health of humans in the course titled "Human Health and Global Environmental Change." [Courses] [Course Titles] The selection of these specific four classes was based on the fact that these were the only classes that included information about time periods [19].

In terms of the OULAD dataset, the sole VLE data that was accessible connected to the course known as "Social Science," which was introduced over the span of two semesters during the school year 2013-2014 [16].

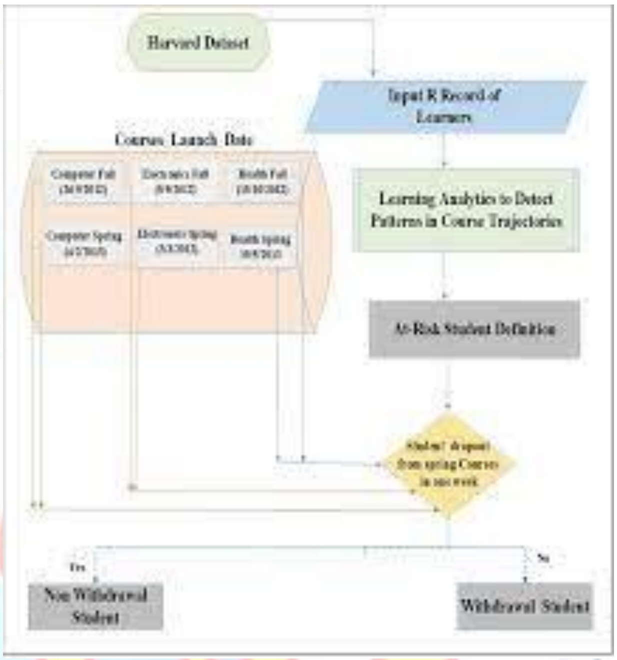

Figure1: Proposed System Architecture

4. RESULTS

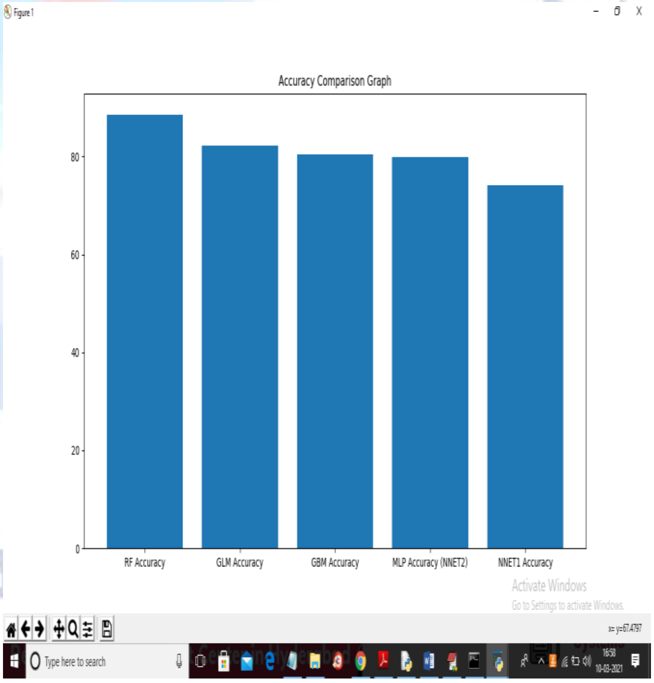

Figure 2: Accuracy of various algorithms

5. CONCLUSION

This study used two case studies to provide decision-makers the chance to intervene early and help students at danger of withdrawal and failure. Engagement, motivation, and withdrawal rates were studied in the first case study. The second case study offered a learning accomplishment model to identify at-risk children and investigate student failure determinants.

The dropout prediction model helps instructors provide early assistance for at-risk pupils. Online course withdrawals are mostly due to student motivation trends. Feature selection improves machine learning model prediction and reduces computing costs. The filter approach for feature selection may also solve overfitting. This research might help instructors identify pupils who need more assistance by tracking student motivational state.

The learning achievement model examined variables affecting at-risk pupils using Harvard and OULAD statistics. Both datasets show that clickstream characteristics strongly predict online course failure.

Future study will validate the suggested methodology with new datasets. Online datasets from multiple providers offering courses on the same subjects would be useful to analyze subject trends. Deep learning can also anticipate course dropouts. Deep learning can extract characteristics from student records by inferring temporal events across MOOC datasets. Deep convolutional neural networks can assess student behavior and motivation and determine their effects on at-risk kids.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] M. R. Ghaznavi, A. Keikha, and N.-M. Yaghoubi, ''The impact of information and communication technology (ICT) on educational improvement,'' Int. Educ. Stud., vol. 4, no. 2, pp. 116–125, 2011.

[2] J. Sinclair and S. Kalvala, ''Student engagement in massive open online courses,'' Int. J. Learn. Technol., vol. 11, no. 3, pp. 218–237, 2016.

[3] H. B. Shapiro, C. H. Lee, N. E. W. Roth, K. Li, M. Çetinkaya-Rundel, and D. A. Canelas, ''Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers,'' Comput. Educ., vol. 110, pp. 35–50, Jul. 2017.

[4] J.-L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, and W. He, ''Identifying at-risk students for early interventions—A time-series clustering approach,'' IEEE Trans. Emerg. Topics Comput., vol. 5, no. 1, pp. 45–55, Jan./Mar. 2017.

[5] R. Alshabandar, A. Hussain, R. Keight, A. Laws, and T. Baker, ''The application of Gaussian mixture models for the identification of at-risk learners in massive open online courses,'' in Proc. IEEE Congr. Evol. Comput. (CEC), Jul. 2018, pp. 1–8.

[6] M. Barak, A. Watted, and H. Haick, ''Motivation to learn in massive open online courses: Examining aspects of language and social engagement,'' Comput. Edu., vol. 94, pp. 49–60, Mar. 2016.

[7] J. C. Turner and H. Patrick, ''How does motivation develop and why does it change? Reframing motivation research,'' Educ. Psycholog., vol. 43, no. 3, pp. 119–131, 2008.

[8] C. Geigle and C. Zhai, ''Modeling MOOC student behavior with twolayer hidden Markov models,'' in Proc. 4th ACM Conf. Learn. Scale, 2017, pp. 205–208.

[9] Altair. (2019). Improve Retail Store Performance Through In-Store Analytics. [Online]. Available: https://www.datawatch.com/in-action/usecases/retail-in-store-analytics/

[10] D. S. Chaplot, E. Rhim, and J. Kim, ''Predicting student attrition in MOOCs using sentiment analysis and neural networks,'' in Proc. 17th Int. Conf. Artif. Intell. Educ., 2015, pp. 7–12.