www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 03, March 2023, pp : 329-333

e-ISSN : 2583-1062

Impact Factor : 5.725

PREDICTING RISK AT EARLY STAGE DURING STUDENTS ONLINE COURSES USING MACHINE LEARNING MODELS

Krishnakaarthik T[1], Ananthan A[2], Mytheshwaran G[3], Babu R[4], Arun M[5]

[1]Assistant Professor, Nandha College of Technology, Perundurai 638 052, Tamilnadu, India

[2,3,4,5]UG Students - Final Year, Department of Information Technology, Nandha College of Technology, Perundurai 638 052, Tamilnadu, India

ABSTRACT

Virtual Learning Environments (VLEs), Learning Management Systems (LMS), and Massive Open Online Courses (MOOCs) are just a few of the online learning platforms that make it possible for thousands or even millions of students to learn according to their interests and without being restricted by time or space. Online learning platforms have many advantages, but they also face a number of disadvantages, such as students' lack of interest, high dropout rates, low engagement, self-regulation, and being forced to set their own goals. IN this study, we propose a predictive model that looks at the issues that at-risk students face and then makes it easier for teachers to intervene quickly to get students more engaged in their studies and better performing. Various machine learning (ML) and deep learning (DL) algorithms are used to train and test the predictive model to determine how students learn based on their study variables. The accuracy, precision, support, and f-score are used to compare the performance of various ML algorithms. In the end, the ML algorithm with the highest f-score metric, accuracy, precision, recall, and support is chosen to create the predictive model at various percentages of course length. Instructors can use the predictive model to identify students who are at risk early in the course, allowing for prompt intervention and avoiding student dropout. Our findings demonstrated the significance of time-dependent variables, students' assessment scores, engagement intensity, or clickstream data, and online learning.

Keywords: Predictive model, earliest possible prediction, at-risk students, machine learning, feed forward network, random forest, support vector machine, early intervention.

1. INTRODUCTION

Statistics are used in predictive modeling to predict outcomes. Predictive modeling can be applied to any kind of unknown event, regardless of when it occurred, despite the fact that the event one wishes to predict is typically in the future. For instance, crimes are frequently detected and suspects are identified using predictive models after they have occurred. A lot of the time, the model is chosen based on detection theory to try to figure out the likelihood of an outcome given a certain amount of input data. For instance, if an email is given, how likely is it to be spam? Classifiers can be used by models to try to figure out how likely a set of data is to belong to another set. For instance, a model may be utilized to decide if an email is spam or "ham" (non-spam). Predictive modeling is more commonly referred to in academic or research and development contexts as either synonymous with or largely overlapping with the field of machine learning, depending on definitional boundaries. Predictive modeling is frequently referred to as predictive analytics when it is used for business purposes.

1.1 At-Risk Students

In the United States, an "at-risk student" is a student who needs temporary or ongoing intervention to succeed academically. [1] Other names for at-risk students include "at-promise youth" and "at-risk youth."[2] At-risk students are also adolescents who are less likely to successfully transition into adulthood and achieve economic self-sufficiency. [3] At-risk students exhibit emotional or behavioral issues, truancy, low academic performance, a lack of interest in academics, and For instance, a study demonstrated that linear modeling can predict between 80 and 87 percent of the variables that influence a school's retention rate. [4]In the California Penal Codes, Governor Newsom changed all references to "at-risk" to "at promise" in January 2020. After the National Commission on Excellence in Education published the article "A Nation at Risk" in 1983, the term "at-risk" was used. The article said that society in the United States was in danger economically and socially [6]. At-risk students are students who have been told, either officially or unofficially, that they are at risk of failing in school. It is difficult to compare the various state policies on the topic in the United States because different states define "at-risk" in different ways. Compared to other students, "at-risk" students face a variety of obstacles. Students with low socioeconomic status, particularly boys, exhibit feelings of isolation and estrangement in their schools, according to research done by Becky Smerdon for the American Institutes for Research. These students are more likely to be classified as "at-risk" because of their socioeconomic status. In a 2006 speech, educational philosopher Gloria Ladson-Billings claimed that the label itself actually exacerbates the difficulties. "We cannot saddle these babies at kindergarten with this label and expect them to

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 03, March 2023, pp : 329-333

e-ISSN : 2583-1062

Impact Factor : 5.725

proudly wear it for the next 13 years, and then think, well, gee, I don't know why they aren't doing well," is her point of view.

1.2 Machine Learning

The study of computer algorithms that can automatically improve through experience and the use of data are known as machine learning (ML). It is thought to be a component of artificial intelligence. In order to make predictions or decisions without being explicitly programmed to do so, machine learning algorithms construct a model from sample data, or training data. When it is difficult or unfeasible to develop conventional algorithms that can carry out the required tasks, machine learning algorithms are utilized in a wide range of applications, including computer vision, speech recognition, email filtering, and medicine. Other applications include computer vision. Computational statistics, which focuses on making predictions with computers, is closely related to a subset of machine learning; but statistical learning is only one type of machine learning.

1.3 Feed Forward Neural Network

An artificial neural network known as a feed forward neural network does not have cyclical connections between its nodes. As a result, it is distinct from its ancestor: networks of recurrent neurons the first and simplest artificial neural network was the feed forward neural network. The information in this network only moves forward from the input nodes, past any hidden nodes, and on to the output nodes. The network does not contain any cycles or loops. A single-layer perceptron network, which only has one layer of output nodes, is the simplest type of neural network. Through a series of weights, the inputs are fed directly to the outputs. In each node, the sum of the products of the weights and the inputs is calculated, and the neuron fires and takes the activated value (typically one) if the value is higher than a threshold, usually 0; Otherwise, it uses the value that has been disabled (typically -1).

1.4 Random Forest

Random forests, also known as random decision forests, are a type of ensemble learning for classification, regression, and other tasks. During the training phase, a large number of decision trees are created. The class that was chosen by the majority of the trees is what the random forest produces for classification tasks. The individual trees' mean or average prediction is returned for regression tasks. Random decision forests eliminate the tendency of decision trees to over fit their training set. Though their accuracy is lower than that of gradient-boosted trees, random forests generally perform better than decision trees. However, their performance can be affected by data characteristics. Tin Kam Ho developed the first random decision forest algorithm in 1995 by employing the random subspace method. According to Ho, this method is a means of putting Eugene Kleinberg's "stochastic discrimination" approach to classification into practice.

2. LITERATURE SURVEY

Xyang et al., Mushtaq Hussain, [2] has suggested. The student's performance prediction is an important research topic in this system because it can help teachers identify students who require additional assistance and prevent students from dropping out before final exams. Predicting the challenges that students will face during a subsequent digital design class session is the goal of this study. Ouafae EL AISSAOUI and coworkers, [3] has suggested. The construction of an effective student model that represents the characteristics of the student is necessary for the implementation of an effective adaptive e-learning system in this system. One of these characteristics is the learning style, which refers to the manner in which a student prefers to learn. Sunbok Lee and others, [4] has suggested. Schools can use a dropout early warning system to identify students who are at risk of dropping out of school, respond quickly to them, and ultimately assist students who might drop out to continue their education for a better future. However, accurate predictive modeling for a dropout early warning system may be challenging due to the inherent class imbalance between dropout and non-dropout students. Et al., Antonio Hernández-Blanco, [6] has suggested. In this system, the study of Educational Data Mining (EDM) focuses on how to use statistical, data mining, and machine learning techniques to find patterns in large amounts of educational data. While a variety of machine learning methods have been used in this area over the years, Deep Learning has recently received more attention in the educational sector. Abdelkader, Hanan E., et al., [12] has suggested. In this system, every educational organization primarily aims to raise students' academic performance to raise education quality as a whole. In this regard, the rapidly expanding field of research known as Educational Data Mining (EDM) makes use of the fundamental ideas of Data Mining (DM) to assist educational establishments in locating useful information regarding the Student Satisfaction Level (SSL) associated with Online Learning (OL) during the COVID-19 lockdown.

3. EXISTING SYSTEM

People who have strong self-regulated learning (SRL) skills, which are characterized by their capacity to plan, manage, and control their learning process, are able to learn more quickly and perform better than people whose SRL

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 03, March 2023, pp : 329-333

e-ISSN : 2583-1062

Impact Factor : 5.725

skills are weaker. SRL is basic in learning conditions that give low degrees of help and direction, as is normally the situation in Enormous Open Web-based Courses (MOOCs). With prompts and activities, learners can be actively supported and trained to participate in SRL. Nonetheless, successful execution of student emotionally supportive networks in MOOCs requires a comprehension of which SRL methodologies are best and the way that these systems manifest in web-based conduct. There are two new insights into SRL provided by this article. For a diverse adult learner population, we first provide insight into SRL and its behavioral manifestations in MOOCs. Second, by making use of the diversity of the current sample, we demonstrate a number of individual variations in SRL that may serve as a basis for specific interventions like adaptive scaffolding.

4. PROPOSED SYSTEM

Finding and categorizing learning issues faced by Research Scholars is the central focus of this thesis. Follow the positive and negative circumstances of engineering students to complete their Research level experiences. Mining social media data like engineering students' study problems will allow for the classification of a group of engineering students based on their experiences and the identification of their issues that need to be addressed in order to enhance the quality of education. The dataset for load testing the proposed system Testing and training utilize the dataset. The classification and prediction are taken from the random forest. The outcomes of the prediction are then displayed by the support vector machine algorithm. The predictions are then displayed using the KNN, SVM, and RANDOM FOREST algorithms. The implementation of a deep feed forward neural network. The aforementioned algorithms are used to create the classification report. These machine learning algorithms can be used to classify the attributes and values in the dataset that determine whether a student excels, passes, fails, or withdraws.

4.1 Input Dataset

All instances of missing variables in the form of nulls or noise were replaced by the OULAD's mean values to improve the predictive models' performance. For instance, the assessments table lacked the date values, which indicate when the assessments were taken and submitted. All date instances with N/A, null, or missing values were replaced by the date mean value because the date is an important variable in the early prediction of at-risk students.

4.2 Preprocessing

In machine learning and data mining, data preprocessing is used to make input data easier to work with in the preprocessing technique training set, which is a subset used to train a model. Test set—to divide the prediction of the trained method, a subset is used to test the trained model. Is sufficiently huge to yield genuinely significant outcomes. Is a good representation of the entire data set? To put it another way, you shouldn't select a test set that differs from the training set. Your objective is to develop a model that effectively adapts to new data, assuming that your test set satisfies the previous two conditions. New data can be compared to our test set.

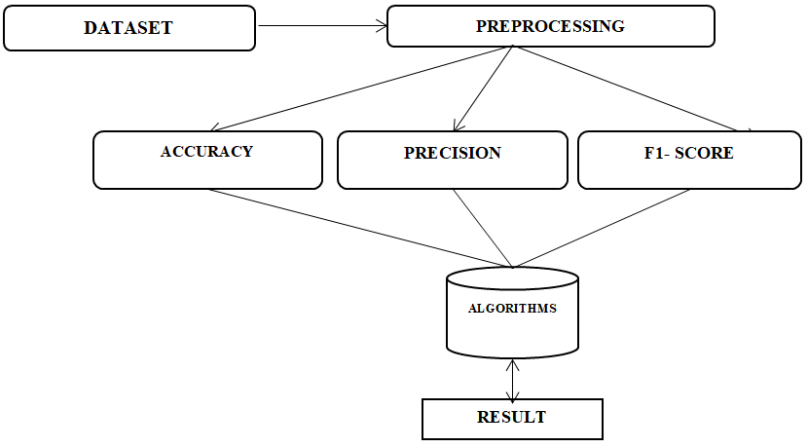


Figure 1. System Flow Diagram

4.3 Classification of ML Algorithms

Classification is a term used in machine learning to describe a predictive modeling problem in which a class label is predicted for a specific piece of input data. Problems with classification include: Classify an example to determine whether or not it is spam. Classify a handwritten character as one of the known characters given the situation. A model will use the training dataset to figure out the best way to map input data examples to particular class labels. In our case, the models for the dataset attributes provide the results for the dataset, and the classification result will be either the candidate is passing, fail, or withdrawn, etc., so the training dataset must be sufficiently representative of the problem.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

Vol. 03, Issue 03, March 2023, pp : 329-333

e-ISSN : 2583-1062

Impact Factor : 5.725

4.4 Comparison Chart

Examination or looking at is the demonstration of assessing at least two things by deciding the important, equivalent qualities of everything, and afterward figuring out which qualities of each are like the other, which are unique, and how much. When two things have different characteristics, the differences can be looked at to see which one is best for a particular job. A comparison is also the description of the similarities and differences between two things. Comparison can take many different forms, with the attributes and parameters used to make each algorithm produce the desired character result varying by field. The obtained result is promoted and represented when the various field-specific machine learning models are compared.

5. EXPERIMENTAL SETUP

The most straightforward performance metric is accuracy, which is simply the ratio of correctly predicted observations to total observations. One might believe that our model is superior if it has high accuracy. Yes, accuracy is an excellent metric, but only when working with symmetric datasets where the values of false positives and false negatives are nearly identical. In order to evaluate your model's performance, you must therefore examine additional parameters.

Accuracy = TP+TN/TP+FP+FN+TN

Precision is the ratio of correctly predicted positive observations to total predicted positive observations; precision is measured in this way. The question that this metric address is how many of the passengers who reported having survived actually did so. The low number of false positives is related to the high precision. We have a pretty good precision of 0.788.

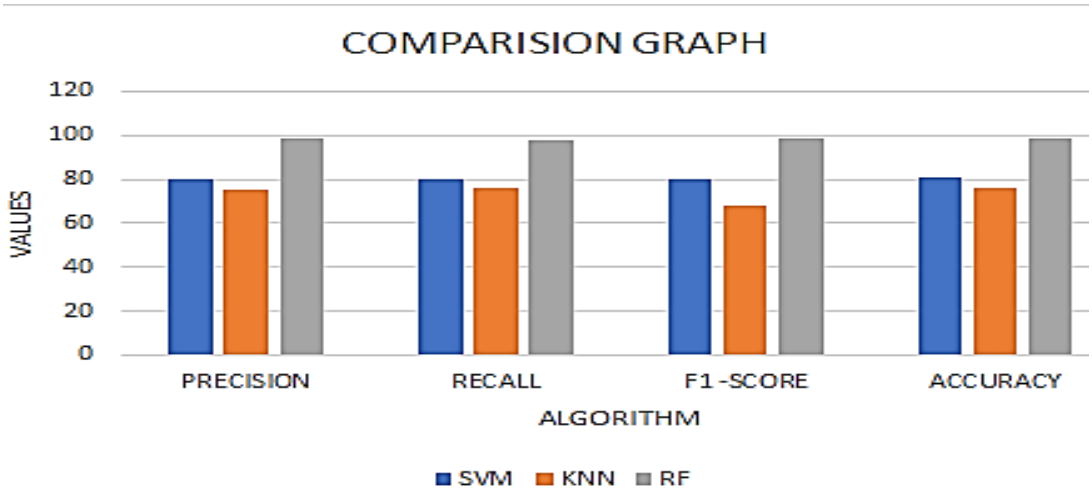
Precision = TP/TP+FP

The ratio of correctly predicted positive observations to all actual class observations is called recall (sensitivity). The responses to the question are: How many of the actual survivors among the passengers did we label? We have a recall of 0.631, which is above 0.5 for this model.

Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, F1 score is 0.701.

Algorithm	Precision	Recall	F1 Score	Accuracy
SVM	80	80	80	81
KNN	75	76	68	76
RF	99	98	99	99



The RANDOM FOREST algorithm outperforms the majority of SVM, KNN, and RANDOM FOREST algorithms in terms of precision, recall, and accuracy, as shown in the aforementioned chart and table.

6. CONCLUSIONS

Both students and instructors benefit from anticipating and intervening with students at various stages of the length of the course. It gives teachers a chance to help students who are at risk of dropping out and make an intervention at the right time to help them study better. For predicting students' performance based on demographics, demographics + clickstream, and demographics + clickstream + assessment variables, we developed a number of predictive models that were trained on a variety of ML and DL algorithms in this study. In the end, the RF predictive model with the highest performance scores was chosen to predict students' performance across course lengths. A predictive model of this kind can make it easier for teachers to make timely interventions and convince at-risk students to do better in school. The clickstream and assessment variables had the greatest impact on the students' final score out of all the variables. According to the findings of this study, predictive models can noticeably benefit from methods like feature engineering. Students' performance was predicted at the very beginning of the course module timeline using only demographic variables.

7. REFERENCES

[1] S. Valsamidis, S. Kontogiannis, I. Kazanidis, T. Theodosiou, and A. Karakos, ‘‘A clustering methodology of Web log data for learning management systems,’’ J. Educ. Technol. Soc., vol. 15, no. 2, pp. 154–167, 2012.

[2] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, ‘‘Using machine learning to predict student difficulties from learning session data,’’ Artif.Intell. Rev., vol. 52, no. 1, pp. 381–407, Jun. 2019.

[3] O. E. Aissaoui, Y. E. A. El Madani, L. Oughdir, and Y. E. Allioui, ‘‘Combining supervised and unsupervised machine learning algorithms to predict the learners’ learning styles,’’ ProcediaComput. Sci., vol. 148, pp. 87–96, Jan. 2019.

[4] J. Y. Chung and S. Lee, ‘‘Dropout early warning systems for high school students using machine learning,’’ Children Youth Services Rev., vol. 96, pp. 346–353, Jan. 2019.

[5] S. A. Salloum, M. Alshurideh, A. Elnagar, and K. Shaalan, ‘‘Mining in educational data: Review and future directions,’’ in Proc. Joint EuropeanUS Workshop Appl. Invariance Comput. Vis. Cairo, Egypt: Springer, 2020, pp. 92–102.

[6] Hernández-Blanco A, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, ‘‘A systematic review of deep learning approaches to educational data mining,’’ Complexity, vol. 2019, May 2019, Art. No. 1306039.

[7] K. S. Rawat and I. Malhan, ‘‘A hybrid classification method based on machine learning classifiers to predict performance in educational data mining,’’ in Proc. 2nd Int. Conf. Commun., Comput. Netw. Chandigarh, India: National Institute of Technical Teachers Training and Research, Department of Computer Science and Engineering, 2019, pp. 677–684.

[8] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauría, J. R. Regan, and J. D. Baron, ‘‘Early alert of academically at-risk students: An open source analytics initiative,’’ J. Learn. Analytics, vol. 1, no. 1, pp. 6–47, May 2014.

[9] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. MousaFardoun, and S. Ventura, ‘‘Early dropout prediction using data mining: A case study with high school students,’’ Expert Syst., vol. 33, no. 1, pp. 107–124, Feb. 2016.

[10] S. Palmer, ‘‘Modelling engineering student academic performance using academic analytics,’’ Int. J. Eng. Edu., vol. 29, no. 1, pp. 132–138, 2013.

[11] Z. Papamitsiou and A. Economides, ‘‘Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence,’’ Edu. Technol. Soc., vol. 17, no. 4, pp. 49–64, 2014.

[12] Peña-Ayala A, ‘‘Educational data mining: A survey and a data mining based analysis of recent works,’’ Expert Syst. Appl., vol. 41, no. 4, pp. 1432–1462, Mar. 2014.

[13] N. Z. Zacharis, ‘‘A multivariate approach to predicting student outcomes in Web-enabled blended learning courses,’’ Internet Higher Edu., vol. 27, pp. 44–53, Oct. 2015.