INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

PREDICTING AT-RISK STUDENTS AT DIFFERENT PERCENTAGES OF COURSE LENGTH FOR EARLY INTERVENTION USING MACHINE LEARNING

Nannapuraju Nishma[1], Research Scholar, Department of CSE,
J.S. University, Shikohabad,Firozabad (Dist.) , U.P.-283135
Email: nishma.nannapuraju@gmail.com

Dr.Vikram Singh Rathore[2], Professor, Department of CSE,
J.S. University, Shikohabad, Firozabad (Dist.), U.P.-283135
Email: rathorevsmail74@gmail.com

Dr. Vittapu Mani Sarma[3], Professor & HOD, Department of CSE,
Malla Reddy Engineering College and Management Sciences, Medchal, Hyderabad –Telangana - 501401
Email: manisharma.vittapu@gmail.com

----------ABSTRACT----------

With the assistance of online learning technologies such as Massive Open Online Courses (MOOCs), Learning Management Systems (LMS), and Virtual Learning Environments (VLEs), you are able to study at your own speed and in the environment that is most comfortable for you. There are many benefits associated with the use of online learning tools; however, there are also some downsides associated with them. Some of these negatives include students' lack of excitement, high dropout rates, independence in behaviour management, and the need to support students in developing their own learning goals. We provide a prediction model that takes into account the difficulties that students who are considered to be at risk confront. It is possible for teachers to act at the most appropriate periods in order to stimulate the students' interest in learning and so enhance their academic performance. An extensive number of machine learning and deep learning techniques are used in order to create and assess the prediction model, which shows how students learn from their studies. To determine the relative effectiveness of a number of different machine learning algorithms, we make use of the f-score, accuracy, precision, and support. Over the course of the class, a number of different machine learning techniques are assessed in terms of their accuracy, precision, recall, support, and f-score measurements. Certain times, the model that generates the most accurate forecasts is selected as the one to use. A prediction model has the ability to aid educators in preventing students from dropping out of school. This is accomplished by identifying children who are at risk at an early stage and providing them with the required assistance. In the context of online education, our results brought to light the relevance of time-dependent factors, clickstream data, and the levels of achievement achieved by students. When the exam was administered at 0, 20, 40, 60%, 80%, and 100% of the total duration of the course, the findings showed that the average recall was 0.59%, the average precision was 0.84%, the average accuracy was 0.88%, the average 0.90%, and the average 0.91%. An average F-score of 0.59% was obtained. Predictive models, early intervention, children who are at risk, machine learning, feedforward neural networks, random forests, and the creation of the most accurate forecasts as quickly as possible are some of the search terms that are being discussed.

Keywords: - MOOC, VLEs, LMS, ML, DL, RF.

I. INTRODUCTION

I.

There are now 9.6 million students from all around the world who are enrolling in online educational programs. As the quality of virtual learning environments continues to increase, users are able to circumvent challenges such as time and location, which is making education more readily available and less expensive. Learning is considerably facilitated by these technologies, which are especially helpful in times of extreme uncertainty, such as the worldwide pandemic of the year 2020. This learning tool encourages efficient study habits by continuously updating

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

and improving the courses that are being offered. Even while there are a lot of advantages to using online learning materials, there are also some disadvantages connected with them. In contrast to conventional classrooms, virtual learning environments made it hard for teachers or mentors to track the progress of their pupils and modify the speed at which they were learning.

However, in the event that the children were not interested in it, they were unable to determine the extent of their involvement. It's even possible that some children may choose to drop out of school entirely. As a result of the cumulative effect of these problems, students' conduct and academic performance are greatly damaged, which in turn makes learning less relevant. It is possible for teachers to prevent any of these issues from occurring if they are able to monitor the development and participation of their pupils. The continuance of the learning process is contingent upon the instructors being aware of the activities that their pupils are engaged in. Teachers are able to give their students with data that may be utilised to evaluate their development via the usage of online learning tools. Therefore, it is essential to establish this assumption at an early stage in order to guarantee that the students are heading in the right direction. In order to make an accurate prediction about how students will study in the future, the results of the final exam cannot be utilised alone. Instructors may be able to more effectively convince students and hold them responsible when they fall short of performance expectations if they provide them with early and continual success forecasts. In order to make predictions about the learning styles of students, it is possible to develop a model by using data from online platforms and machine learning techniques. The assignment will be finished after the model that is the most appropriate has been chosen. Following a thorough examination of the Open University Learning Analytics dataset, this project implements its use. The student-centered dataset included a variety of different types of information, including demographic data, clickstream data, test scores, the name of the course, and vle interactions. It is possible that kids who are considered to be at risk will be less likely to drop out of school if the OULA data is included. The Support Vector Machine method, which is particularly effective when dealing with multi-dimensional data, is the one that we use to train the prediction model. The SVM-learned prediction model achieves a high level of efficiency in both memory efficiency and guesswork. For the students, there are four different groupings: pass, fail, withdraw, and distinction. It is possible to do this grading whenever it is most convenient for you, provided that the students have been actively engaged up to that point in the course being evaluated. Educators may use these results to determine which students are successful in the class and which students should be removed from the class. A comparison is made between the best prediction model trained using Random Forest and the model trained with Support Vector Machines. In terms of accuracy, the comparison found that Support Vector Machines (SVM) fared better than Random Forest. This was determined after rigorous preparation and review of the datasets. These results, which are anticipated by the model, are essential for ensuring that students continue to participate in online meetings and for boosting the productivity of virtual learning to a level that is equivalent to that of conventional classroom education. There is a possibility that the results of the prediction model will capture the attention of students and direct them towards the most appropriate major. The recommendations included in this article might be beneficial to educators and administrators who are in charge of online education since they will help them better manage their online classrooms and make decisions that are based on accurate information. There is a possibility that these kinds of modifications might lead to improvements in schooling.

If teachers are able to assess their students early on and determine who would struggle and ultimately fail the class, they may be able to keep their students who are considered to be at risk interested and on track academically. Many educational institutions, whether they are located in a more traditional setting or online, fail to discriminate between the requirements of their students and instead apply a same set of laws to all of them. A prediction model that is capable of promptly detecting when and how to help students should be developed by those who are responsible for the creation of virtual learning environments (VLEs). The ability to get individualised feedback and help from the very beginning of the semester would be made possible for them by this. Teaching has been substantially easier and more effective as a result of the enormous advancements that have been made in the area of Educational Data Mining (EDM) in terms of tools, methods, and materials [7]. Due to the fact that these tools do not aid teachers in identifying students who are at risk at an earlier stage in the course, a significant amount of work is required to determine what is wrong and to ensure that students remain on track. In order to reveal latent study habits that account for the benefits and drawbacks of online learners, a multitude of prediction models have been constructed [8, 9]. These models have been built employing improvements in artificial intelligence, machine learning, and deep learning methodologies. In an

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

attempt to bring down the number of students who drop out of school, researchers could apply machine learning techniques to explore specific factors that have a major influence on the rates of student dropout. Educators are able to determine which students are most likely to withdraw their registration before it is too late by using prediction models that are powered by machine learning. Our work intends to swiftly identify students who are likely to drop out of school by using machine learning to discover indicators connected to students' learning and their usage of virtual learning environments (VLEs). Open University Learning Analytics (OULA) conducted a study of the data acquired on students' involvement with course materials and activities throughout the weeks of online teaching. The findings of this research indicate that students' engagement levels are very variable. It is because of this that a significant proportion of students decide to withdraw from the course. Our results were used as a basis for the creation of a prediction tool that might potentially assist teachers in identifying students who are at danger of dropping out of a course before it is finished. If teachers are able to use the prediction model, they may be able to have more convincing interactions with their students, which may in turn reduce the risk that their students would drop out of school.

The following are the results of this endeavour:

The development and testing of prediction models that make use of a range of machine learning and deep learning techniques makes it feasible to anticipate the exam results of pupils.

The identification of vulnerable virtual learning environment (VLE) students who are on the verge of dropping out of school need to be a top focus.

By using a prediction model in conjunction with customised comments, we are able to provide assistance to teachers so that they can support their students at the most appropriate time.

A discussion of a number of different persuasive tactics that might potentially enhance the academic achievement of students

II. LITERATURE SURVEY

A. Testing the efficacy of educational data mining methods for predicting first-semester programming course failure rates

According to the statistics, a significant number of students do badly in beginning computer science classes, which has caused many teachers to express worry. The result of this is that a great deal of severe problems with predictive features have been brought to light. The purpose of this study is to evaluate the effectiveness of several educational data mining methodologies in identifying first-year students in computer science who have difficulties. Previous research has investigated approaches that are analogous to this one; however, this study is distinct in that it applies a novel way to detect the academic issues that students are experiencing: (i) Determine the usefulness of these approaches for identifying students who are at risk in a timely way so that interventions may be taken to minimise the number of students who are at danger; (ii) Determine the influence that activities such as data preparation and algorithm fine-tuning have on the effectiveness of the methods that have been discussed above. In order to assess the effectiveness of four different prediction approaches, we used two datasets that were produced from beginning computer science courses that were taught by a Brazilian public university. One of the ways was derived from classroom teaching, while the other was derived from self-study.

B. Data mining for the purpose of modelling and forecasting student performance in the classroom

Through the use of data mining techniques, the purpose of this study is to predict and examine the academic achievement of students by using their previous performance in class as well as their degree of involvement in group activities. Educational data mining, often known as EDM, is an innovative technique that may be used to improve the academic performance of pupils. Educational institutions have the ability to perform extensive personality surveys of their student population by using EDU. The data for this study comes from two different classes taken at the institution. The categorisation of the content was accomplished by the use of three unique data mining approaches, namely Network, Decision Tree, and Naïve Bayes. On the basis of the accuracy of their predictions, we conduct an analysis and evaluation of the three models. Without any shadow of a doubt, the Naïve Bayes technique consistently beat the other two methods.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

C. Using machine learning approaches to identify pupils at risk and intervene early

There has been a meteoric rise in the number of people participating in Massive Open Online Courses (MOOCs) in recent years. Students will have access to a wide variety of digital materials via these classes. Numerous factors, including the flexibility of the learning environment and the availability of learning assistance, are being credited with contributing to the rapid increase of the user base. According to the findings of a significant amount of study, the high attrition rate and the low completion rate are big problems. This technique makes it possible to identify students who are at danger of failing high school and dropping out of school at an earlier stage. Two models are produced as a consequence of this: one is designed for kids who are successful academically, while the other is designed for children who are at risk of falling behind. The algorithms could be able to determine which students who are enrolled in an online program would have difficulty and will ultimately withdraw from the program.

III. SYSTEM ANALYSIS

D. Current Setup

If teachers are able to assess their students early on and determine who would struggle and ultimately fail the class, they may be able to keep their students who are considered to be at risk interested and on track academically. Many educational institutions, whether they are located in a more traditional setting or online, fail to discriminate between the requirements of their students and instead apply a same set of laws to all of them. A prediction model that is capable of promptly detecting when and how to help students should be developed by those who are responsible for the creation of virtual learning environments (VLEs). The ability to get individualised feedback and help from the very beginning of the semester would be made possible for them by this. Because of the considerable advancements that have been made in educational data mining (EDM) technology, techniques, and solutions, education professionals are now able to enable learning that is both more expedient and less complicated.

1.The downsides

- The current method of predicting the stock market is plagued by a number of problems, and data scientists often come into these problems when they are trying to develop a forecasting model.
- It is clear that the stock market is sensitive to unpredictability in many other areas, including public opinion about corporations and the leadership of national governments, as seen by the fact that this method is useless.

E. Suggested Framework

Using Random Forest and Naive Bayes Classifier, the suggested method investigated a number of factors that have a substantial influence on the percentage of students that drop out of school. This was done with the intention of lowering the failure rate. Through the use of prediction models that are driven by Random Forest and Naive Bayes Classifier, educators have the ability to identify students who are at danger of dropping out of school and develop ways to keep these students enrolled in the classroom. By using machine learning to discover characteristics related with students' learning behaviours and interactions with the virtual learning environment (VLE), the major purpose of our research is to quickly identify students who are at risk of failing to meet the requirements of the classroom. If teachers are able to use the prediction model, they may be able to have more convincing interactions with their students, which may in turn reduce the risk that their students would drop out of school.

During the course of this research, several machine learning and deep learning techniques were used to construct and evaluate prediction models for the purpose of estimating the performance of students on exams.

The identification of vulnerable virtual learning environment (VLE) students who are on the verge of dropping out of school need to be a top focus.

By using a prediction model in conjunction with customised comments, we are able to provide assistance to teachers so that they can support their students at the most appropriate time.

A discussion of a number of different persuasive tactics that might potentially enhance the academic achievement of students

The positive aspects

I am concerned that the prediction was an accurate one.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

At an early stage, it is possible to identify children who are at danger.

in a position to evaluate the performance of students

Help instructors become more aware of when they should act in order to provide the most help to pupils.

For the most part,

IV. METHODOLOGY

F. Information Detail

It was the Open University Learning Analytics Dataset (OULAD) that was used; this dataset is accessible to the general public and was developed by the Open University in the England. The student data is comprised of seven distinct tables, each of which contains information on individuals who are students. The information that is maintained in these tables includes details on the profiles of students, assessments, interactions with the virtual learning environment (VLE), registration for classes, and actual classes that students have attended. If you use key markers, you can join two tables together. By examining the clickstream data (the number of clicks) that is stored in the student database, it is possible to get insight into the daily activities and interactions that students have with the virtual learning environment (VLE). There is a possibility that the test results of the students may be found in the student-modulepresentation dataset triplet. During the 2013–2014 academic year, there were a total of 32,593 students who registered in seven different classes spread out across twenty-two different semesters. The Open Data Institute (http://theodi.org/) is the organisation that provides access to OULAD for the general public. You are able to get access to the dataset by visiting https://analyse.kmi.open.ac.uk/open_dataset.

G. Preparing Data for Use

In an effort to enhance the performance of the prediction models, we either removed any instances of missing variables (such as nulls or noise) or made use of the mean values of those variables that were obtained from the OULAD. An example of this would be the lack of date values in the evaluations table, which would ordinarily specify the times at which the tests were given and returned. The date mean value was used to fill in any and all occurrences of dates that came with values that were either non-existent, null, or missing. One of the reasons for this is that dates are an essential component in the process of identifying children who are at risk at an early stage.

H. Engineering Features

20% of the total time allotted for the course was given to each student, followed by 40%, 60%, 80%, and eventually 100% of the time. Therefore, it is possible to predict the performance of the students in a timely way using this method. On top of that, we reasoned that the demographic information of the pupils could offer some insight into how well they would do on future examinations. For the purpose of predicting the students' future performance, we used demographic data alone, demographic data plus twenty percent of the course completion data, demographic data plus forty percent of the course completion data, and so on. Several additional criteria were built by making use of the data that was already available in order to make predictions about the performance of students at different stages during the course. The Relative Score (RS) variables were built so that it would be possible to show how well students performed at 20%, 40%, 60%, 80%, and 100% of the course block conclusion. It is important to note that the variables are twenty, forty, sixty, eighty, and one hundred roubles.

Variables (LS20, LS40, LS60, LS80, and LS100) were used to display the number of late submissions at the conclusion of the course session at two percent, forty percent, sixty percent, eighty percent, and one hundred percent, respectively. It was decided to define variables with the names AS20%, AS40%, AS60%, AS80%, and AS100% in order to show the real assessment scores at 20%, 40%, 60%, and 100% of the course module conclusion, respectively. The level of involvement that students had with the virtual learning environment (VLE) was evaluated via the use of clickstreams, which were an important factor in determining the durations of the different course blocks. It was decided to construct the sum_clicks and mean_clicks variables first. The total and average clicks are shown by these variables at 20%, 40%, 60%, 80%, and 100% of the way through the course module. These variables are as follows: SC20%, SC40%, SC60%, SC80%, SC100%, AC20%, AC40%, AC60%, AC80%, and AC100%, respectively. The information that was previously contained in the student profiles and assessment tables combined has been consolidated into a single table. This table covers both demographic and evaluation data. Personal data was coupled with information from the virtual learning environment (VLE), such as clickstream data, in order to evaluate how students used the learning tools provided by the VLE during a particular course block.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

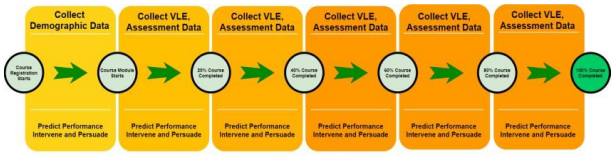

FIGURE 1. Predicting and intervening at-risk students at different percentages of the course length.

V. CONCLUSION

When teachers take the effort to prepare ahead of time and provide early help to students, it is beneficial for both the teachers and the students. Educators have the ability to recognise students who are not contributing to the class, who are losing their concentration, or who may be leaving the class altogether. Since instructors now have access to this new information, they are in a better position to support their students in their academic pursuits by intervening at the proper moments. A model that was developed using Support Vector Machines was suggested by us in order to forecast the academic achievement of students. The model was based on the outcomes of their exams, clickstream data, and personal information. When the two prediction models are compared, the Support Vector Machine (SVM) has a higher success rate than the Random Forest. It is considered that rating elements and clickstream variables had the most impact on the end results of the children because of the difference that they made. It is possible for both instructors and students to make use of this form of prediction tool in order to guarantee that the learning process is proceeding in the appropriate way. The performance of support vector machines (SVMs) is superior than that of other models when it comes to putting students into four categories: pass, fail, withdrew, and distinction. These categories indicate the anticipated levels of performance that students should have. The instructors are able to influence the thoughts of students who are most likely to drop out of school with the assistance of these groups since they are able to act quickly. It's possible that instructors may concentrate more on some students if they are having difficulty. Because this strategy increases the usefulness of online learning, it is beneficial to students in terms of their academic performance. WORK IN THE FUTURE, PART VI Deep learning models and natural language processing techniques have the potential to investigate the activity-wise relevance that has a substantial influence on students' performance. This may be accomplished by replicating the textual components that are related with student input. For the purpose of improving the quality of study, regression methods may also be used, in addition to the utilisation of percentages to display the accomplishments of pupils.

II. REFERENCES

1) A. Mubarak, H. Cao, and S. A. M. Ahmed, ‘‘Predictive learning analytics using deep learning model in MOOCs’ courses videos,’’ Edu. Inf. Technol., vol. 6, pp. 1–22, Jul. 2020.

2) Hernández-Blanco, 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.

3) Sekeroglu, K. Dimililer, and K. Tuncal, ‘‘Student performance prediction and classification using machine learning algorithms,’’ in Proc. 8th Int. Conf. Educ. Inf. Technol., Mar. 2019, pp. 7–11.

4) Romero, S. Ventura, and E. García, ‘‘Data mining in course management systems: Moodle case study and tutorial,’’ Comput. Edu., vol. 51, no. 1, pp. 368–384, Aug. 2008.

5) G. Akçapçnar, M. N. Hasnine, R. Majumdar, B. Flanagan, and H. Ogata, ‘‘Developing an early- warning system for spotting at-risk students by using eBook interaction logs,’’ Smart Learn. Environ., vol. 6, no. 1, p. 4, Dec. 2019.

6) J. Figueroa-Ca nas and T. Sancho-Vinuesa, ‘‘Predicting early dropout student is a matter of checking completed quizzes: The case of an online statistics module,’’ in Proc. LASI-SPAIN, 2019, pp. 100–111.

7) J. Xu, K. H. Moon, and M. van der Schaar, ‘‘A machine learning approach for tracking and predicting student performance in degree programs,’’ IEEE J. Sel. Topics Signal Process., vol. 11, no. 5,pp. 742– 753, Aug. 2017.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

8) L. Cen, D. Ruta, L. Powell, B. Hirsch, and J. Ng, ‘‘Quantitative approach to collaborative learning: Performance prediction, individual assessment, and group composition,’’ Int. J. Comput.-Supported Collaborative Learn., vol. 11, no. 2, pp. 187–225,Jun. 2016.

9) L. P. Macfadyen and S. Dawson, ‘‘Mining LMS data to develop an ‘early warning system’ for educators: A proof of concept,’’ Comput. Edu., vol. 54, no. 2, pp. 588–599, Feb. 2010.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.

10) Muhammad Adnan, Asad Habib, Jawad Ashraf, Shafaq Mussadiq, Arsalan Ali Raza, Muhammad Abid, Maeryam Bashir, AND Sana Ullah Khan, "Predicting At-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models", Digital Object Identifier 10.1109/ACCESS.2021.3049446

11) N. Mduma, K. Kalegele, and D. Machuve, ‘‘Machine learning approach for reducing student’s dropout rates,’’ Int. J. Adv. Comput. Res., vol. 9, no. 42, 2019, doi: 10.19101/IJACR.2018.839045.

12) 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,’’ Procedia Comput. Sci., vol. 148, pp. 87–96, Jan. 2019.

13) R. F. Kizilcec, M. Pérez-Sanagustín, and J. J. Maldonado, ‘‘Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses,’’ Comput. Edu., vol. 104, pp. 18– 33, Jan. 2017.

14) R. S. Baker and P. S. Inventado, ‘‘Educational data mining and learning analytics,’’ in Learning Analytics. New York, NY, USA: Springer, 2014, pp. 61–75.

15) 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