RESEARCH ARTICLE OPEN ACCESS

Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models

Mrs. B.Vijaya MCA M.Tech [1] , Shamkepalli Ameena [2]

[1] Associate Professor, Department of Computer Applications

[2] Student, Department of Computer Applications

[1],[2] Chadalawada Ramanamma Engineering College(Autonomous)

ABSTRACT

Online learning platforms such as Massive Open Online Course (MOOC), Virtual Learning Environments (VLEs), and Learning Management Systems (LMS) facilitate thousands or even millions of students to learn according to their interests without spatial and temporal constraints. Besides many advantages, online learning platforms face several challenges such as students' lack of interest, high dropouts, low engagement, students' self-regulated behavior, and compelling students to take responsibility for settings their own goals. In this study, we propose a predictive model that analyzes the problems faced by at-risk students, subsequently, facilitating instructors for timely intervention to persuade students to increase their study engagements and improve their study performance. The predictive model is trained and tested using various machine learning (ML) and deep learning (DL) algorithms to characterize the learning behavior of students according to their study variables. The performance of various ML algorithms is compared by using accuracy, precision, support, and f-score. The ML algorithm that gives the best result in terms of accuracy, precision, recall, support, and f-score metric is ultimately selected for creating the predictive model at different percentages of course length. The predictive model can help instructors in identifying at-risk students early in the course for timely intervention thus avoiding student dropouts. Our results showed that students' assessment scores, engagement intensity i.e. clickstream data, and time-dependent variables are important factors in online learning. The experimental results revealed that the predictive model trained using Random Forest (RF) gives the best results with averaged precision = 0.60%, 0.79%, 0.84%, 0.88%, 0.90%, 0.92%, averaged recall = 0.59%, 0.79%, 0.84%, 0.88%, 0.90%, 0.91%, averaged F-score = 0.59%, 0.79%, 0.84%, 0.88%, 0.90%, 0.91%, and average accuracy = 0.59%, 0.79%, 0.84%, 0.88%, 0.90%, 0.91% at 0%, 20%, 40%, 60%, 80% and 100% of course length. INDEX TERMS Predictive model, earliest possible prediction, at-risk students, machine learning, feedforward neural network, random forest, early intervention

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

I. INTRODUCTION

Worldwide, there are 9.6 million users enrolled in online education and with the rapid innovations in the development of virtual learning environments, the platforms aid in overcoming the difficulties of space and time, making education easily accessible and affordable. These innovations are so beneficial and convenient as it ensures seamless learning process in case of uncertainties like the pandemic situation the world faced in 2020. These learning platforms enable effective learning behavior providing interesting and interactive classes. Although there are many advantages of this virtual learning platforms, there still occurs some challenges. Unlike physical classes, mentors or instructors in virtual learning environments could not monitor students' learning behaviors and keep them in the right learning pace. They couldn't identify students' interest in case of low engagement. There are even chances of course dropouts. Due to all these challenges, the effectiveness of learning gradually decreases which greatly affects students' performance and behavior. To prevent all these drawbacks, instructors should be able to know the performance of students and their activeness throughout the course. Instructors must understand the activities of students in some way to effectively continue the learning process. Data generated from virtual learning platforms can help instructors to analyze students' performance. This prediction should be earlier to ensure whether students are in the right track. Predicting the students' learning behavior at the end of the course only with the final assessment scores will be of no use. Predicting student performance early and at any stage

of the course will help instructors to persuade and warn students in case of low activeness. A model that predicts the student's learning behavior with the data generated from online platforms can be built using machine learning algorithms and choosing an optimal one will do the needful. Open University Learning Analytics dataset is considered and analyzed for this project. The dataset contained student centered information like demographics data, virtual learning environment interaction, assessment scores, course name, clickstream data. Analyzing OULA dataset can help prevent at-risk students from dropping out of their course. Support Vector Machine algorithm is used to train the predictive model as it works well with high dimensional data. This predictive model trained with SVM consumes less memory and is efficient in prediction. It categorizes students into four classes namely, Pass, Fail, Withdrawn and Distinction. This categorization can be done in any stage of the course taking into account students' participation till that stage. These results can help instructors to evaluate student performance and persuade the students who are about to fail and will withdraw the course. The predictive model trained with SVM is compared against the model trained with Random Forest which is proved to be the best. The results of the comparison showed that, after sufficient preprocessing and carefully evaluating the dataset, SVM has a greater accuracy than Random Forest. These derived results from the predictive model serves as a vital information for preventing students from getting deviated from the online classes and enhance the effectiveness of virtual learning similar to the physical classes. The results of the predictive model can persuade students to the right learning path and improve their interest. Such processes can assist virtual learning administrators and instructors for developing an effective framework for online learning and improve their decision-making process. These types of enhancements can aid in better learning process.

Identifying students at-risk of dropout and failure as early as possible during a course could help the instructors to execute timely and necessary interventions/persuasions to help students to remain steady during their studies [6]. Generally, in traditional classroom settings and online learning settings, a general approach is followed where the same guidelines are defined for all students ignoring individual discontentment. To provide personalized feedback and support right from the start of the semester, VLE designers require the development of a predictive model that makes rapid decisions about how and when to intervene students for support. Educational Data Mining (EDM) tools, techniques, and products have progressed significantly, helping educators to make education easy and effective [7]. However, these techniques lack in identifying at-risk students earlier in the course timeline, compelling instructors to perform significant manual work for students problem identification to keep them on track. The emergence of Artificial Intelligence (AI), machine learning (ML), and deep learning (DL) techniques have facilitated and enabled researchers to develop series of predictive models to reveal hidden study patterns which explain the strength and weakness of online students [8], [9]. To reduce the dropout rates, researchers can use ML techniques to study different variables that significantly affect the students' dropout. Predictive models powers by ML techniques can present the accurate picture of students that are likely to quit their study thus facilitating instructors to come up with preventive measures before dropout behavior occurs. The prime objective of our research study is the earliest possible identification of students who are at-risk of dropouts by leveraging ML techniques to understand variables associated with the learning behavior of students and how they interact with the VLE. By analyzing Open University Learning Analytics (OULA) dataset, it was observed that students are inconsistent in their online learning activities throughout course weeks resulting in high dropouts at the end of the course. Based on our observations, we developed a predictive model that can identify students at-risk of dropout at the very start of the course. The predictive model is capable of facilitating instructors to intervene students through persuasive messages that encourage students to keep themselves on the right track thus avoiding dropouts.

The contribution of this study include:

- Developing and evaluating predictive models using several ML/DL algorithms to predict students' performance scores.
- Earliest possible identification of students in VLE who are at-risk of dropout during the course.

- Integrating personalized feedbacks with a predictive model to help instructors in interventing students at an optimal time.
- Discussing various persuasion techniques that may help students in increasing their study performance

II. LITERATURE SURVEY

Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses

The data about high students' failure rates in introductory programming courses have been alarming many educators, raising a number of important questions regarding prediction aspects. In this project, present a comparative study on the effectiveness of educational data mining techniques to early predict students likely to fail in introductory programming courses. Although several works have analyzed these techniques to identify students' academic failures, our study differs from existing ones as follows: (i) investigate the effectiveness of such techniques to identify students likely to fail at early enough stage for action to be taken to reduce the failure rate; (ii) analyse the impact of data preprocessing and algorithms fine-tuning tasks, on the effectiveness of the mentioned techniques. In our study we evaluated the effectiveness of four prediction techniques on two different and independent data sources on introductory programming courses available from a Brazilian Public University: one comes from distance education and the other from on-campus.

Modeling and predicting students' academic performance using data mining techniques

The main objective of this study is to apply data mining techniques to predict and analyze students' academic performance based on their academic record and forum participation. Educational Data Mining (EDM) is an emerging tool for academic intervention. The educational institutions can use EDU for extensive analysis of students' characteristics. In this study, we have collected students' data from two undergraduate courses. Three different data mining classification algorithms (Naïve Bayes, Neural Network, and Decision Tree) were used on the dataset. The prediction performance of three classifiers are measured and compared. It was observed that Naïve Bayes classifier outperforms other two classifiers

Detecting atrisk students with early interventions using machine learning techniques

Massive Open Online Courses (MOOCs) have shown rapid development in recent years, allowing learners to access high-quality digital material. Because of facilitated learning and the flexibility of the teaching environment, the number of participants is rapidly growing. However, extensive research reports that the high attrition rate and low completion rate are major concerns. In this project, the early identification of students who are at risk of withdrew and failure is provided. Therefore, two models are constructed namely at-risk student model and learning achievement model. The models have the potential to detect the students who are in danger of failing and withdrawal at the early stage of the online course.

III. SYSTEM ANALYSIS

Existing System

Identifying students at-risk of dropout and failure as early as possible during a course could help the instructors to execute timely and necessary interventions/persuasions to help students to remain steady during their studies. Generally, in traditional classroom settings and online learning settings, a general approach is followed where the same guidelines are defined for all students ignoring individual discontentment. To provide personalized feedback and support right from the start of the semester, VLE designers require the development of a predictive model that makes rapid decisions about how and when to intervene students for support. Educational Data Mining (EDM) tools, techniques, and products have progressed significantly, helping educators to make education easy and effective

Disadvantages

➢ In the existing work, the system in which Stock market prediction is full of challenges, and data scientists usually confront some problems when they try to develop a predictive model.

➢ This system is less performance in which it is clear that there are always unpredictable factors such as the public image of companies or political situation of countries, which affect stock markets trend.

Proposed System

In Proposed System, to reduce the dropout rates, used Naive Bayes Classifier and Random Forest techniques to study different variables that significantly affect the students' dropout. Predictive models powers by Naive Bayes Classifier and Random Forest techniques can present the accurate picture of students that are likely to quit their study thus facilitating instructors to come up with preventive measures before dropout behavior occurs. The prime objective of our project is the earliest possible identification of students who are at-risk of dropouts by leveraging ML techniques to understand variables associated with the learning behavior of students and how they interact with the VLE. The predictive model is capable of facilitating instructors to intervene students through persuasive messages that encourage students to keep themselves on the right track thus avoiding dropouts.

The contribution of this project include:

- Developing and evaluating predictive models using several ML/DL algorithms to predict students' performance scores.
- Earliest possible identification of students in VLE who are at-risk of dropout during the course.
- Integrating personalized feedbacks with a predictive model to help instructors in interventing students at an optimal time.
- Discussing various persuasion techniques that may help students in increasing their study performance

Advantages

- Prediction is accurate
- Identifying at-risk students earlier
- Able to predict student scores
- Help instructors in interventing students at an optimal time.
- Efficient

IV. METHODOLOGY

Data Description

We used a freely available Open University Learning Analytics Dataset (OULAD), provided by the Open University, UK. Students' data is spread across 7 tables each containing students centered information such as students' demographics, students' Virtual Learning Environment (VLE) interaction, assessments, course registration, and courses offered. Tables relate to each other through key identifiers. Students' daily activities and VLE interaction are represented as clickstreams data (number of clicks) stored in the student VLE table. Students' assessment scores are stored in a dataset triplet called student-modulepresentation. The OULAD was generated for the year 2013 and 2014 containing 7 courses, 22 module-presentations with 32,593 registered students. OULAD is freely accessible at https://analyse.kmi.open.ac.uk/open_dataset and has been certified by the Open Data Institute http://theodi.org/

Data Preprocessing

To enhance the performance efficiency of the predictive models, all missing variables instances in the form of nulls, or noise were removed or replaced by their mean values from the OULAD. As an example, the date values were missing in the assessments table which represents the date the assessments were taken/submitted. As the date is an important variable in the early prediction of at-risk students, all the date instances having N/A, null, or missing values were replaced by the date mean value.

Feature Engineering

For the earliest possible prediction of students' performance, we divided the course length into 5 parts i.e. 20%, 40%, 60%, 80%, and 100% of course completed. We also assumed that demographic data solely can also be used to predict students' upcoming performance in assessments and final exams. The students' future performance prediction was

determined by modeling the predictive models using only demographics data, using demographics and 20% course completion data, demographics, and 40% course completion data, and so on. To predict students' performance at different times of course module, several new variables were created from the existing variables. Relative Score (RS) variables were created to represent student relative performance at 20%, 40%, 60%, 80%, and 100% of course module completion (RS20, RS40, RS60, RS80, RS100). Variables indicating the number of late submissions were created when 20%, 40%, 60%, 80%, and 100% of the course module was completed (LS20, LS40, LS60, LS80, LS100). Variables representing the raw assessment scores were also created at 20%, 40%, 60%, 80%, and 100% of the course module completion (AS20%, AS40%, AS60%, AS80%, AS100%). Variables representing students' VLE interaction in the form of clickstreams were created for the different percentage of course module length. Two types of variables namely sum_clicks and mean_clicks were created to indicate the sum of clicks and average clicks at 20%, 40%, 60%, 80%, and 100% of the course module completion (SC20%, SC40%, SC60%, SC80%, SC100%, AC20%, AC40%, AC60%, AC80%, AC100%). Students' demographics table was merged with the students' assessment table to get demographics and assessment data in one table. Moreover, the VLE information i.e. students clickstream data was also merged with demographics data to know students' interaction with VLE learning contents during a course module.

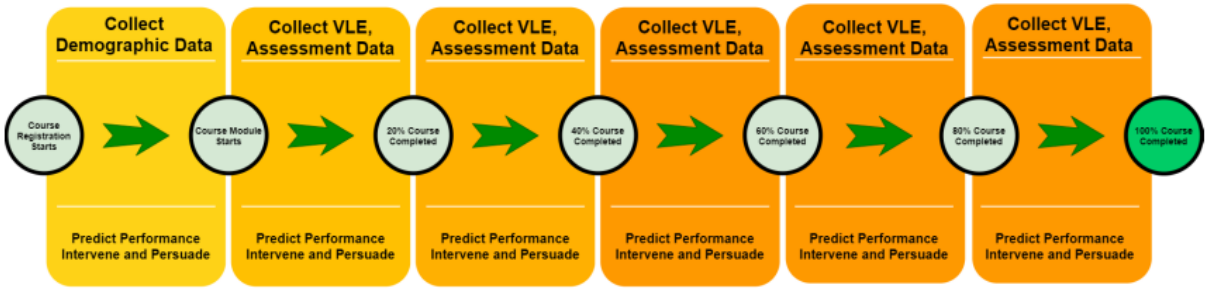

FIGURE 1. Predicting and intervening at-risk students at different percentages of the course length.

V. CONCLUSION

Predicting and intervening students early in the course provides benefits to both students and instructors. It provides an opportunity for instructors to identify students at-risk like students who are likely to dropout from course, low engagement, decline in interest. With this information, instructors can intervene at the optimal time to improve students' study behavior. In this paper, we proposed a predictive model trained on Support Vector Machine for predicting students' performance based on demographics variables, clickstream variables, and assessment variables. It has been compared with the Random Forest predictive model and the comparison proved that SVM has the highest performance scores. Among all the features, clickstream variables and assessment variables are considered having the most significant impact on the final result of the students as their contribution was more. Such a predictive model can enable instructors and students to ensure whether their learning behavior is on the right track. As a result of comparison, it is proved that SVM model predicts students' performance in a more accurate way of classifying them into four classes – Pass, Fail, Withdrawn and Distinction. These classifications help instructors to persuade students who are likely to withdraw the course by making timely interventions. Instructors can also concentrate more on those students whose performance is low. This process improves the study performance of students and increases the efficiency of virtual learning. VI. FUTURE WORKS Activity wise significance with a prominent influence on the students' performance by modeling textual• variables related to students' feedbacks can be examined by utilizing deep learning models and natural language processing techniques. Along with the students'

performance, the percentage of score can also be displayed using regression• techniques for better analysis.

REFERENCES

[1] A. 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] A. 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] B. 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] C. 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.

[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.

[10] 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.

[11] 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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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