e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

PREDICTING AT-RISK STUDENTS IN VIRTUAL LEARNING ENVIRONMENT USING SUPPORT VECTOR MACHINE

S Keerthika[*1], V P Janani[*2], T Karthika[*3]

[*1]Assistant Professor, Department Of Computer Science And Engineering, Velalar College Of Engineering And Technology, Erode, Tamil Nadu, India.

[*2,3]Student, Department Of Computer Science And Engineering, Velalar College Of Engineering And Technology, Erode, Tamil Nadu, India.

ABSTRACT

With the increasing uncertainty in today's world, education pattern is gradually transforming from physical classes to virtual classes. Virtual learning environments are the source of interaction between students and the mentors. It facilitates millions of students to learn according to their interest without any constraints like time and place. Besides many advantages, there still occurs some challenges like low engagement, course dropouts due to lack of continuous monitoring and so on. In this paper, we have considered OULA dataset for training a predictive model built with Support Vector Machine. The data included demographics data, clickstream data and the assessment scores among which clickstream data and assessment scores had a significant impact on the final prediction. The performance of the proposed model is compared with the model built using Random Forest algorithm. The experimental results revealed that the predictive model built using Support Vector Machine has a greater accuracy in prediction. This early prediction enables instructors to monitor students' performance and intervene at any stage of the course. Students and instructors can keep track of the learning pace and ensure whether students are in the right track.

Keywords: OULA Dataset, Random Forest, Support Vector Machine, Virtual Learning Environment.

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

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

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.

II. RELATED WORKS

Choosing the suitable algorithms and techniques for different purposes is a vital part in the domain of Machine Learning. Various works related to machine learning prediction are carried forward as follows. In [3], two datasets – Student Performance Dataset and Students Academic Performance Dataset are considered and analyzed using three machine learning algorithms. The results showed that Support Vector Regression had a better prediction among Back Propagation and Long-Short Term Memory. These results were based on eighteen experiments and proved that accuracy can be further improved by necessary preprocessing steps. In [5], early prediction system is built in the setting of eBook-based teaching and learning. It used students' eBook data to build the predictive model. The model is built on thirteen machine learning algorithms with 10-fold cross validation for all the models. Accuracy and kappa metrics were used for comparison of models among which Random Forest had a greater accuracy.

In [6], early prediction of at-risk students is carried out to identify dropout-prone students. The model is built on decision tree considering mainly the quiz completion, participation in forum and access to bulletin board. Furthermore in [7], a novel machine learning method is proposed that address challenges like – different students from different backgrounds and courses selected, lack of information about courses, students' evolving progress. Based on students' evolving performance states, a bi-layered structure with multiple base predictors and a group of ensemble predictors is built to make predictions. Then, it is necessary to discover course relevance for constructing base predictors. To achieve this, a data-driven approach based on probabilistic matrix factorization and latent factor models were proposed.

Finally, in [8], the focus is towards the collaborative learning based on quantitative evaluation and prediction of group performance. Initially, machine learning techniques were used to predict group performance by interaction among members exploring the application of Extreme Learning Machine and Classification and Regression Trees using live interaction data. Then a comparative model to restore individual student performance is proposed which is used in generative mixture model. The individual grade expectations from student are compared to actual group performance to define collaboration synergy. This evaluation indicates a great level of predictability of group performance solely based on style and mechanics of collaboration.

III. PROPOSED WORK

The predictive model is designed in such a way that the model predicts the students' performance and provides classification results like Pass, Fail, Withdrawn or Distinction. The above-mentioned results can be obtained in any stage of the course and instructors of Virtual learning environments can intervene and persuade students for preventing course dropouts. This will improve the students' performance and help keep them on the right track.

The model is built using Support Vector Machine. This algorithm works well with the high dimensional data as there are greater number of features in the dataset. SVM works well as there is a clear separation between classes. It is more memory efficient and gives an efficient accuracy for the model. The OULA dataset is loaded which is freely provided by the Open University, UK. Students' data is spread across tables each containing

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

students centered information such as students' demographics, students' virtual learning environment interaction, assessments, course registration, and courses offered. Activities of students and their interaction with VLEs are represented in terms of clickstream data. Figure 1 and 2 shows an instance of the OULA dataset.

module	presentation	id_student	gender	Region	highest_education
AAA	2013J	11391	M	East Anglian Region	HE Qualification
AAA	2013J	28400	F	Scotland	HE Qualification
AAA	2013J	30268	F	North Western Region	A Level or Equivalent
AAA	2013J	147756	M	North Region	Lower Than A Level
AAA	2013J	146188	F	West Midlands Region	A Level or Equivalent

Figure 1: Instance of OULA dataset

imd_band	age_band	sum_click	assessment_score	credits	disability	final_result
90-100%	55<=	10	84	240	N	Distinction
60-70%	35-55	8	70	60	N	Pass
30-40%	35-55	1	40	60	Y	Withdrawn
50-60%	35-55	11	69	60	N	Pass
20-30%	0-35	2	30	60	Y	Fail

Figure 2: Instance of OULA dataset (continued)

The dataset is loaded initially and it is split into training and testing data using K-fold cross-validation technique where the value of k was set to 10. The libraries are imported from weka tool. It contains various machine learning algorithms that can be used for data mining tasks. These are used for availing the functionalities within the java code. In the pre-processing step, the data is handled for missing data. The missing data is replaced with the average mean values. This is done to enhance the performance of the model.

These training and testing datasets with and without target variables are loaded into the environment of the proposed framework respectively. A model with Random Forest is built for comparison. The proposed predictive model built using SVM is compared against the model built with RF. The models are compared in terms of recall, precision, f1 score and accuracy. The comparison proved that the model built using SVM performs better than RF. Thus, SVM outperforms RF and can be used in the model to produce an accurate outcome.

IV. EVALUATION

Evaluating the model is the most important part of building a machine learning model. It works on a constructive feedback model. After building a model, feedback from performance metrics is considered to make improvements. This process is continued until a desirable accuracy is achieved. The performance of the model is measured by evaluating it. It is vital to check the accuracy of the model prior to computing predicted values.

Accuracy is a metric that gives the fraction of results that the model predicted right. Accuracies of Support Vector Machine and Random Forest are compared and is proved that SVM had a greater accuracy. Recall, precision and f1-score are the most important classification metrics and are plotted for SVM and RF for comparison. Recall is the true positive rate and gives the fraction of positive values we identified correctly with the total values (True Positive + False negative). Precision gives the fraction of correctly identified true positive values with the all predicted positive values (True positive + False positive). The average value of recall and precision is the fi-score.

Thus, the proposed model proves to be best at giving better accuracy. The values of all these performance metrics had a greater value for SVM than RF. The table 1 shows the comparison of evaluation metrics of SVM and RF. It is apparent that the accuracy score of SVM is far higher than RF and this difference is mainly due to the preprocessing step and the features considered.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

Table 1: Comparison of evaluation metrics of SVM and RF

Algorithms	Accuracy	Precision	Recall	F1-Score
Random Forest	0.74	0.74	0.72	0.73
Support Vector Machine	0.94	0.94	0.96	0.95

From the below visualization, it is evident that the model built using Support Vector Machine has a greater accuracy with fair values of performance metrics.

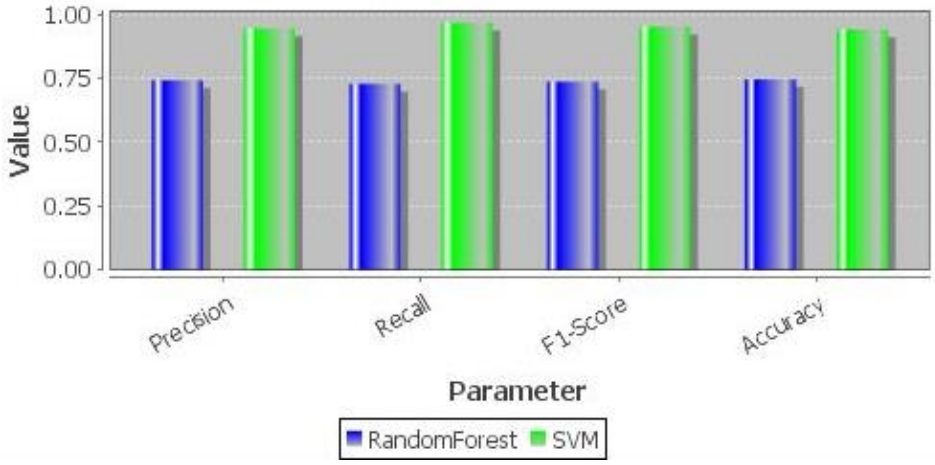

Figure 3: Comparison chart of SVM and RF

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.

VII. 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. 1306039.

[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.