" Enhancing Online Learning Engagement And Performance Through Predictive Analysis "

Lakshmi Naga Priya
Computer Science And Engineering
(of Affiliation)
Narasaraopeta Engineering College
(JNTUK)
Narasaraopeta,India
aduripriya@gmail.com

Anjani Thanmayee Chandaluri
Computer Science And Engineering
(of Affiliation)
Narasaraopeta Engineering College
(JNTUK)
Narasaraopet,India
anjanithanmaye7@gmail.com

Leela Pavani Vemula
Computer Science And Engineering
(of Affiliation)
Narasaraopeta Engineering College
(JNTUK)
Narasaraopet,India
vleela800@gmail.com

***Abstract*— This research focuses on predictive analysis, a machine learning technique aimed at reliable performance prediction in online learning systems, encompassing various courses and platforms. Our goal is to register new users and mitigate dropout rates by providing personalized test results and enabling users to access their performance data at anytime. Through this research, we aim to elucidate students learning behaviors in relation to their study factors. Leveraging random forest and diverse models, our predictive analysis trains and tests data to develop robust prediction models. Our findings demonstrate that random forest yields superior accuracy, offering valuable insights into student learning dynamics. The experimental results revealed that the predictive model trained using Random Forest (RF) gives the best results with averaged average accuracy = 71%, 74%, 79%, 80%, 80%, 69%.**

Keywords—co Supervised learning, Predictive analysis, Performance prediction, Machine learning, and feature selection

I. Introduction

There is rapid evolution in online learning systems, including MOOC and Virtual Learning Environments - VLE has revolutionized the access to the education. Effectively evaluating and analyzing student performance and data generated within these platforms is very important for instructors to monitor and comprehend students' learning progress. Early detection of students' performance in VLEs enables timely intervention, guiding them on the right path. Previous studies have shown that variables stored in database records related to students' learning can aid instructors in predicting future performance L. P. Macfadyen et al [1]-[3]. Detecting students who are about to dropout or failure at the outset of the course allows instructors to implement timely interventions, fostering stability in students' studies O.E. Aissaoui et al[6]. Traditional approaches, both in physical classrooms and online settings, often employ a one-size-fits-all approach, overlooking individual variations. To offer personalized reports and support from the beginning of the course, there is a need for a predictive model that rapidly decides when and how to provide support to students through intervention. Despite advancements in Educational Data Mining (EDM) tools and techniques, they still do not possess the ability to early identify students having risk in the course. timeline, necessitating significant manual effort from tutors or instructors for the identification of problem and provide support for students who registered for their courses.

II. BACKGROUND AND RELATED WORK

A. Educational Data Mining

The emerging field of "educational data mining" is devoted to analyzing large-scale data sets.There has been an increase in research using statistical and predictive models in recent years to analyze data from formal and the informal educational Environments. For instance, several studies E. B. Costa, S.M. Jayapraksh, A.Cano,S. Palmer et al[10]-[13] have explored The impact of demographic factors on fostering successful learning outcomes and student retention. Additionally, efforts have been made to facilitate early intervention and provide timely feedback to guide at-risk students A. Economides, J.J.Maldonado et al [14]-[16].
Studies conducted at Open University, UK A. Wolff, M.Hlosta et al [17], [18], have tried to use a variety of predictor characteristics to identify pupils who are having risk, emphasizing the importance of incorporating demographic variables alongside student behavior variables for improved predictive models. Y.Cui and Choi et al [19] categorized variables contributing to student dropouts into three categories: demographic variables, course structure and requirement variables, and environmental/context factors. Time-series clustering approaches - Y. Lee and J. Choi et al [20] have been utilized for early identification of students having risk demonstrating higher accuracy compared to traditional aggregation methods. Moreover, clickstream data have been found to provide more accurate and objective measures of students' online engagement compared to self-reported data at P. H. Winne et al [21].

III. Data Description

We utilized the Open University Learning Analytics Dataset (OULAD), containing student-centric information like demographics and VLE interactions. The student VLE table records all of the activities that students do within the VLE, including daily interactions and clickstream data. The triplet of datasets known as student-module presentation contains the assessment scores. Data from seven courses and 22 module presentations, totaling 32,593 enrolled students, are included in the OULAD for the years 2013 and 2014. Because of its accreditation by the Open Data Institute, the OULAD is readily available to the public, guaranteeing its dependability and openness.OULAD is publicly accessible through the Open Data Institute's certification. For access, visit https://analyse.kmi.open.ac.uk/open_dataset.

A. Abbreviations and Acronyms

Support Vector Machine – SVM

Random Forest – RF

K-Nearest Neighbors – KNN

Extra Tree Classifier – ETC

Ada Boost Classifier – ABC

Gradient Boost Classifier - GBC

B. Data Preprocessing

In order for our prediction models to work more efficiently, we implemented a preprocessing step to handle missing variables effectively. Instances of missing variables, such as nulls or noise, were addressed by either removing them or substituting them with their respective mean values taken from the Learning Analytics Dataset from Open University.

For example, in the assessments dataset where date values indicating the dates assessments were taken or submitted were missing, which is a crucial variable for early prediction of risk of students.

we took the following steps:

- **Identification of Missing Values**: We identified instances with missing date values, denoted as N/A, null, or any other form of missing representation.
- **Replacement with Mean Values**: All instances with missing date values were substituted with the mean date value derived from the OULAD. This approach ensures that the predictive models have complete and accurate data to work with, thereby enhancing their performance.

C. *Feature Engineering*

In order to provide the earliest feasible projection of pupils' performance, we adopted a dynamic approach, providing performance forecasts at any percentage of course completion rather than dividing it into fixed intervals. In this method, we utilized demographic data alone as well as in combination with varying percentages of course completion data to develop predictive models.
Integration of students' demographic data with assessment and VLE information was achieved by merging the demographics table with the assessment table and the clickstream data, respectively. This allowed us to analyze students' interactions with VLE learning contents alongside their demographic and assessment data throughout a course module.

D. *Data Analysis*

We have gathered datasets from esteemed sources, including online learning platforms and educational institutions. These datasets constitute the cornerstone for our predictive model development and analysis. Our data collection encompasses the following datasets: “assessments.csv”,”vle.csv”, “StudentRegistration.csv”,”courses.csv” "studentinfo.csv" and "studentAssessment.csv".

The two main datasets are displayed below:

StudentInfo dataset:

The StudentInfo dataset contains 12 columns aare studied_credits, num_of_prev_attempts,code_presentation, code_module, id_student, region, gender, and highest_education and finally disability. This dataset tells about student Result.
Here code_module and code_presentation describes details about course user opted. imd_band is a percentage which describes about the student education details.
This is labeled data as this contains final_result which is the key factor in prediction.

IV. Data Visualization

The below graph specifies about the comparison of code modules. Based on this graph we can predict the highly popular module. Based on the graph we predict the students are very much interested towards the BBB,FFFmodules.

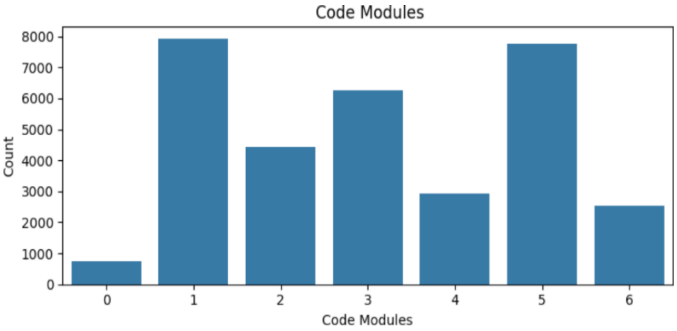


Fig 1 – Comparision of Code Modules

Evalution Criteria

ACCURACY:

To calculate accuracy, divide the number of classes that were correctly predicted by the total number of classes. It illustrates how accurate the model's predictions are on average.

PRECISION:

The precision measures the percentage of genuine positives among all the positive predictions that the model produces. It is useful for assessing the reliability of positive predictions, especially when the positive class is rare.

V. Experimental Results

A. *Constructing predictive models solely with demographic data*

Six predictive models that were exclusively trained on demographic data using K-fold cross-validationThe outcome was the goal variable in these models. while all other demographic variables served as input. recall, accuracy, Precision and F-score values for different positions of students' final results indicate notably poor performance across all predictive models. Particularly concerning is the significant under performance observed for "Fail" as result. Within proactive support systems, timely Identification of students having risk is critical. Thus, the predictive models' effectiveness in predicting outcomes for students at risk of failure holds heightened importance, enabling early interventions to improve student performance.

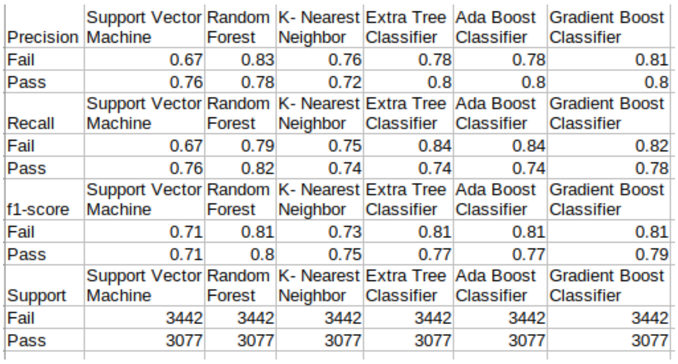
B. *Constructing predictive models using Demographics and click-stream data*

In an effort to improve the predictive models performance, We incorporated both clickstream data and demographic information into our predictive model training and testing process andwe incorporated clickstream data (Students' engagement with the VLE, as indicated by the number of clicks throughout the course duration.) alongside demographics for training purposes.clickstream datais helpful in identifying about the submission of the student and Upon examining the heatmap it became evident that the relation between the final result and other variables remained consistent, with no notable positive and negative correlations observed Between demographics, clickstream data, and the end result. Although sum_clicks100 and mean_clicks100 have a slight connection, it is statistically insignificant. As a result, we incorporated all demographic and clickstream information to train and evaluate the prediction models.

C. *Constructing predictive models using Demographics ,Clickstream, Assessment Scores by Feature Engineering*

A merge operation was carried out to combine Distinction and Pass classes into a Pass class and Withdrawn and Fail classes into a Fail class in order to improve and enhance the results of the models. These combined courses share similar characteristics and information, warranting a feature engineering technique to bolster predictive model performance, particularly for the Fail class, where pupils need the instructor's help and are in danger.

The following figure illustrates a notable increase in predictive model performances post feature engineering. Every prediction model, on average, had accuracy, F-score, precision, and recall values higher than 79%. All baseline models were regularly outperformed by RF, with SVM showing the lowest performance. Gradient Boost, AdaBoost, and ExtraTree classifiers displayed similar performance levels, closely trailing Random Forest. Consequently, the RF classifier was chosen to train and test predictive models across different course module durations.



Precision	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0.67	0.83	0.76	0.78	0.78	0.81
Pass	0.76	0.78	0.72	0.8	0.8	0.8
Recall	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0.67	0.79	0.75	0.84	0.84	0.82
Pass	0.76	0.82	0.74	0.74	0.74	0.78
f1-score	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0.71	0.81	0.73	0.81	0.81	0.81
Pass	0.71	0.8	0.75	0.77	0.77	0.79
Support	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	3442	3442	3442	3442	3442	3442
Pass	3077	3077	3077	3077	3077	3077

Fig 2 - performance of models after performing feature engineering.

VI. Using Pervasive Techniques to Intervene with Students

To enhance students' study behavior through intervention and persuasion, timing is crucial. Utilizing the RF predictive model, which demonstrated satisfactory results (80% accuracy, precision, recall, f-score), interventions can be initiated after any percentage of the course length. Moreover, with detailed demographic data, interventions can commence at the course outset. Figure 6 outlines trigger types for students that is delicate, improving, and consistent, emphasize praise, reward, appreciation, and social acceptance.. Triggers for at-risk students incorporate fear, hope, and suggestion, while those To improve student consistency and improvement, highlight praise, reward, appreciation, and social acceptance.. The timing of trigger delivery depends on the predictive model's stage-specific insights. For instance, a hope-based trigger for at-risk students might state."Our predictive model describes about your performance either pass or fail and shows many graphs representing Code Module Distribution, Gender Distribution, Age Band Distribution, Region Distrubution"

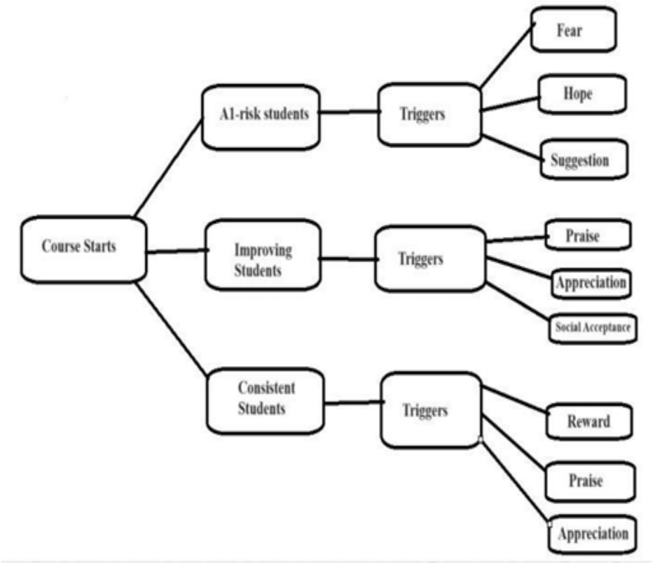


Fig 3 - Different triggers for students of different performance

VII. Conclusion,Limitation And Future Work

The primary goal of the undertaking is to use the Machine learning algorithm to estimate student risk at any stage of course length. Four classification metrics were used in the study and for evaluations. The research revealed in addition to demonstrated data using clickstream data and assignment ratings significantly improved the models performance. The best performing algorithm is Random Forest which resulted 79% accuracy was chosen to forecast student performance. Clickstream data and assessment scores have the biggest influence on the outcome of all the variables.

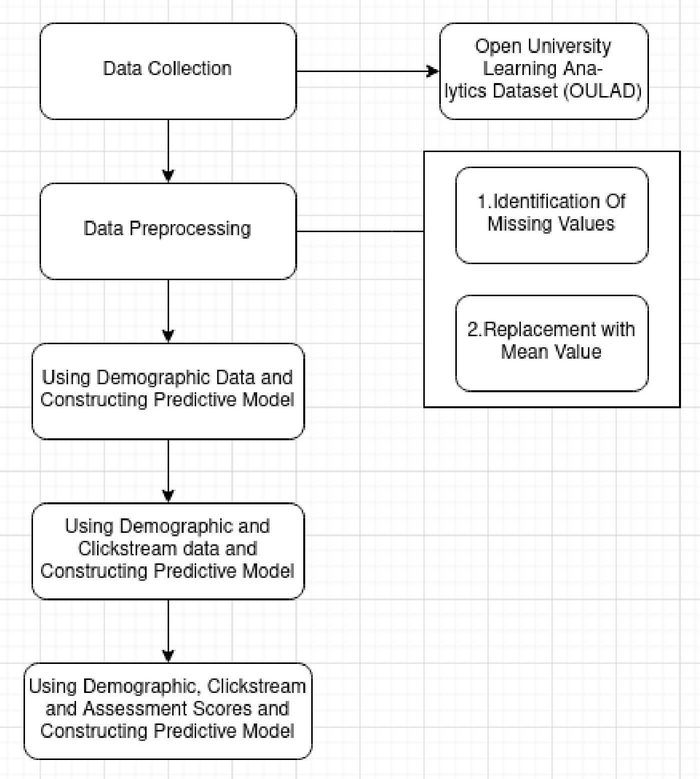
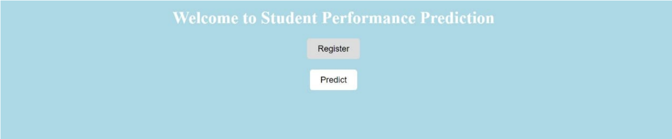

Data Collection
Open University Learning Ana-lytics Dataset (OULAD)
Data Preprocessing
1.Identification Of Missing Values
2.Replacement with Mean Value
Using Demographic Data and Constructing Predictive Model
Using Demographic and Clickstream data and Constructing Predictive Model
Using Demographic, Clickstream and Assessment Scores and Constructing Predictive Model

Fig 4 – Project Flow Chart

In our project, firstly users can either register for a new course or to enquire about the results using predict option. If user wants to register then the user will provide their details and an id is issued to them when they were successfully registered. If the user want to predict the result they have to provide details like code of the courses and some mandatory fields about the user.

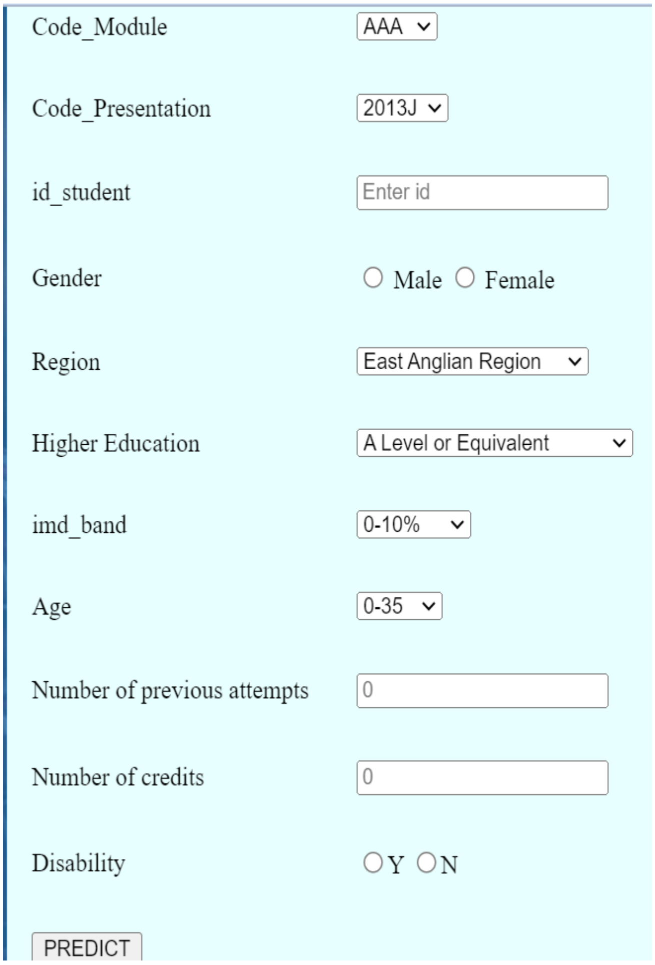
The Random Forest predictive model's outcomes underscore its efficacy in promptly identifying students at risk by performance. Such data based investigations can aid VLE administrators and tutors in shaping online learning frameworks, enriching the decision-making process. However, we acknowledge the need for more extensive analyses to assess diverse online activities within the OULAD. Specifically, Further research is warranted to explore how different early intervention tactics might be smoothly integrated into the online learning environment to encourage students to stay on track. Our future endeavors will focus on scrutinizing the activity-specific significance that significantly influences student performance through the dissemination of textual messages and reminders.

Output Screens



Welcome to Student Performance Prediction

Fig 5 – Home Screen



Code_Module AAA

Code_Presentation 2013J

id_student Enter id

Gender ○ Male ○ Female

Region East Anglian Region

Higher Education A Level or Equivalent

imd_band 0-10%

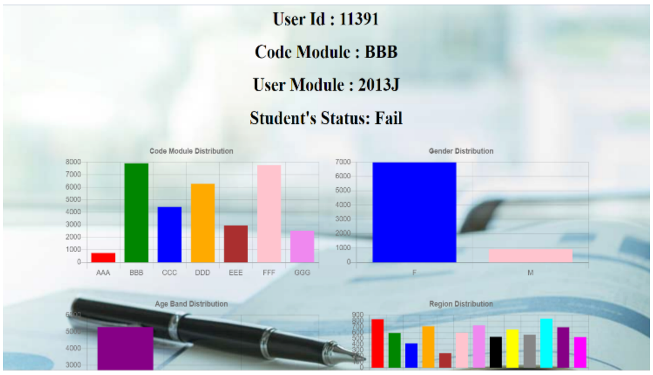
Age 0-35

Number of previous attempts 0

Number of credits 0

Disability ○ Y ○ N

Fig 6 - Registration Form And Prediction Form



User Id : 11391

Code Module : BBB

User Module : 2013J

Student's Status: Fail

Fig 7 – Prediction Success

VIII. *References*

[1] L. P. Macfadyen and S. Dawson , "developing 'early warn system' for educators: A proof of concept," Comput. Edu., vol. 54, no. 2, pp. 588– 599, Feb. 2010.

[2] 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.

[3] S. Valsamidis, S. Kontogiannis, I. Kazanidis, T. Theodosiou, and A. Karakos, "A clustering methodology of Web log data for learning management system," J. Educ.Technol. Soc., vol. 15, no. 2, pp. 154–167, 2012.

[4] 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.

[5] O. E. Aissaoui, Y. E. A. El Madani, L. Oughdir, and Y. E. Allioui, "Combining machine learning algorithms to predict the learners' learning styles," Procedia Comput. Sci., vol. 148, pp. 87– 96, Jan. 2019.

[6] 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.

[7] 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.

[8] K. S. Rawat and I. Malhan, "A hybrid 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.

[9] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araájo, and J. Rego, "Evaluating the effectiveness of data mining for early prediction of students' academic failure in introductory programming courses," Comput. Hum. Behav., vol. 73, pp. 247–

256, Aug. 2017.

[10] 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.

[11] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, 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.

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

[13] 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.