ENHANCING ONLINE LEARNING ENGAGEMENT AND PERFORMANCE THROUGH PREDICTIVE MODELING

Shaik Khaja Mohiddin Basha[1], Aaduri Priya [2], Chanduluri Thanamayee [3],
Vemula Pavani[4] [1] Professor, [2, 3 & 4] Student
[1]adurilakshminagapriya@gmail.com, [2] anjanithanmayee7@gmail.com, [4] leelapavani123@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

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 findings showed that the Random Forest (RF) predictive model yielded the highest averaged precision scores = 72%, 74%, 79%, 80%, 81%, 77%, averaged recall = 71%, 74%, 79%, 80%, 81%, 69%, averaged F-score = 71%, 74%, 79%, 80%, 81%, 69%, and average accuracy = 71%, 74%, 79%, 80%, 80%, 69%.

KEYWORDS: Supervised learning, Predictive analysis, Performance prediction, Machine learning, and feature selection.

I. INTRODUCTION

There is rapid evolution in online learning systems, including Massive Open Online Courses - 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].

However, Forecasting students' progress at the outset of the course, rather than after completion, poses a greater advantage. Incorporating a predictive model that accurately forecasts students' learning behavior at the outset of a course through behavior data analysis presents a significant challenge. In the realm of online learning, where vast amounts of data are generated daily, machine learning (ML) techniques can play a pivotal role When analyzing defining variables of students, offering insights advantageous for both tutors and students. S. Valsamidis,M Hussain et al[4], [5].

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.

Machine learning advancements have enabled researchers to create predictive models revealing concealed study patterns, elucidating both the strengths and weaknesses of students who are pursuing course online. J.Y Chung,D.Thomas et al[7], [8]. ML techniques can be employed to study variables significantly affecting student dropout, offering accurate insights into students likely to discontinue their studies. The core aim of this project is the identification or the early identification of students having risk by providing their performance results Utilizing ML techniques to comprehend variables linked to their learning behavior and interaction with the VLE.

The study's contributions encompass:

- Creating and assessing predictive models employing diverse ML/DL algorithms for forecasting students' performance.
- Determining a student's risk in VLEs before the course even begins.

Through the evaluation of the Open University Learning Analytic datasets, our study noted irregularities in online learning engagement among students across course weeks, resulting in elevated dropout rates by the course conclusion. Utilizing these findings, we constructed a predictive model capable of early identification of dropout-prone students. This model aids students in monitoring their performance and maintaining progress. This model facilitates students to know about their performance and stay on track.

II. BACKGROUND AND RELATED WORK

A. EDUCATIONAL DATA MINING

The emerging field of "educational data mining" is devoted to analyzing large-scale data sets using statistical and machine learning (ML) methods in order to obtain insights on the behavior patterns and learning environments of students. Machine learning approaches have been used in a number of EDM studies to identify factors that have a substantial impact on students' performance, dropout rates, engagement levels, and interactions in online learning environments. While a great deal of research focuses on examining factors that are obtained from students' online activities, several studies also include demographic data in order to evaluate their influence on study habits. The length, content type, class time, and other interaction activities were the first important variables taken into account for study. But as online learning systems developed, more factors were added to the study, including location, click-stream data, assignment and assessment scores, and online interactions. Determine

While most research concentrates on gathering information and forecasting students' performance at the conclusion of the course, there is a dearth of solutions aimed at preventing dropout and failure. With online learning platforms generating huge amount of data from the course's onset, there is potential to develop comprehensive predictive models analyzing variables from the beginning of the course, facilitating early intervention to prevent failures and dropouts.

A study by K.S. Rawat and I.Malhan [9] implemented four machine learning algorithms to identify students having high risk of failure and dropout at early in the course, with the most accuracy being shown by Support Vector Machine (SVM).

Predictive model adaptation to particular learning platforms is still difficult, though, because online platforms, instructional methods, and course structures differ.

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].

Recent studies have focused on analyzing clickstream data to measure students' online engagement, with limited research aiming to facilitate timely interventions for students. Gupta and Sabitha utilized Decision Tree algorithms to determine significant features contributing to student retention in MOOCs, while Akçapnar Gökhan created an early warning system to forecast the likelihood of academic failure based on student eBook reading statistics.

It's still difficult to predict how well pupils will perform early in the course. due to the diverse nature of online learning environments. In any case, machine learning strategies like Irregular Woodland (RF), Choice Trees (DT), Bolster Vector Machine (SVM), and K-Nearest Neighbors (KNN) calculations are progressively being utilized to distinguish learning designs and encourage convenient interventions. For instance, Krösi and Farkas et al [22] utilized Recurrent Neural Network (RNN) algorithm to predict students performance and engagement, while Alberto C. and John D. L. et al [23] employed a multi-view genetic programming approach for classifying students and triggering timely alerts for at-risk students.

By using logistic regression to predict student dropouts, Lara et al. were able to significantly lower dropout rates through the deployment of tutoring action plans. Knowledge discovery in databases techniques has been introduced to extract information about students' interaction with e-learning systems, aiding instructors in improving students' study performance.

While existing research addresses various aspects related to dropout prediction, early intervention, and classification of student performance groups, there is a gap in predicting students having risk at different course lengths. This proposed predictive model aims to fill this gap by enabling educational institutions and instructors to identify students having low score early in the course and intervene through persuasive techniques to improve their study performance.

We must take into account demographic information, a crucial component for early intervention, in order to anticipate children at an earlier stage. Assessment results and Clickstream data are significant time-dependent variables in addition to Demographic data.With an accuracy rate of 79%, it uses a random forest model to identify students who are having risk in the online courses and provides them with intense early assistance.

Our project concentrates on leveraging machine learning methods to enhance online learning outcomes. We explore the application of Support vector machines and artificial neural networks are used to predict online course performance and student participation. We gather and analyze data from diverse online learning platforms to train and validate our predictive models.

This study employs predictive modeling techniques to improve online learning engagement and performance. Models for machine learning, such as logistic regression and decision trees,, random forest, and gradient boost classifiers, are utilized to forecast student outcomes in online courses. The authors utilize features such as student interaction data, course material engagement, past performance, and other relevant metrics to train the

predictive models.

In our project, we aim to enhance online learning engagement and performance through the application of

predictive modeling techniques. We utilize a blend of machine learning models specifically tailored to the online learning environment. These models are designed to predict student outcomes and levels of engagement by incorporating features such as student interaction data, course material engagement, past performance, and other pertinent metrics.

Key features such as course progress, interaction frequency, and assessment scores are utilized to train our models.

The study employs techniques such as decision trees, k-nearest neighbors(KNN), and logistic regression to predict students performance and also the engagement. Data from various online learning platforms and courses are utilized to train and test the models, aiming to provide personalized recommendations and interventions for optimizing learning experiences.

These studies exemplify the diverse array of techniques and methodologies employed in predective model. They underscore the significance of factors such as course performance, learner statistics, and environmental conditions in accurately predicting learning outcomes and engagement levels.

III. DATA DESCRIPTION

For our research, we used the Open University Learning Analytics Dataset (OULAD), which was provided by the Open University in the United Kingdom. This dataset comprises seven tables, each containing student-centric information such as demographics, Virtual Learning Environment (VLE) interactions, assessments, course registrations, and offered courses. These tables are interconnected via key identifiers, facilitating relational data analysis.

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.

For access to the dataset, interested parties can visit https://analyse.kmi.open.ac.uk/open_dataset.

OULAD has the potential to be a perfect platform for early forecasting of at-risk pupils with careful investigation and efficient modeling.

A. 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 table 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:

1. **Identification of Missing Values**: We identified instances with missing date values, denoted as N/A, null, or any other form of missing representation.

2. **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.

By addressing missing variables in this manner, we aimed to optimize The efficiency of our predictive

models' performance. enabling more accurate and reliable predictions of students risk at any stage of the course length.

This preprocessing step is very essential to ensure that these predictive models are trained on high-quality data, minimizing the impact of missing values on the model's performance and ensuring robust predictions.

B. 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.

C. 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.

Column Name	Column Description
code_module	This Column describes about Course opted by student
code_presentation	Describes about the current module
id_student	Tells Student Id
gender	Tells Student Gender
region	Tells Student Region
highest_education	Tells Info about Education
imd_band	Describes about the imd band
age_band	Tells Student Age
num_of_prev_attempts	Tells No of attempts given before
studied_credits	Tells Credit points
disability	Tells aboutMental Disability
final_result	Result of student

Fig 1 - StudentInfo Dataset

This studentInfo Dataset is the main Dataset which contains all the details about the student and in main prediction these have more influence on greatest education, disability, number of prior tries, age band, imd band, and student credits.

StudentAssessments dataset:

The dataset student Assessment dataset contains various attributes like

Assessment Id, Student Id, Date of Assessment Submitted,Score secured by student.

Column Name	Column Description
id_assessment	Unique Id of Assessment
id_student	Unique Id of Student
date_submitted	Assignment Submitted Date
score	Score of student

Fig 2 – studentAssessments.csv

D. 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.

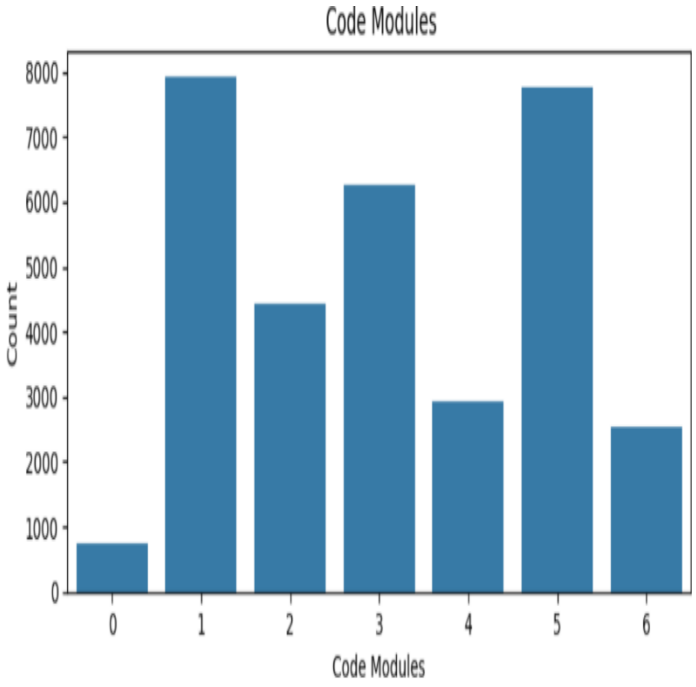

Fig 3 - Comparision of code Modules

The data contains many differennt courses and many different code modules in it we found the late submissions by student in different courses.

The below graph “0” represents late submission of assessment by students and ”1” represents the submission of assessment by students at any time for different course modules .

The below graph specifies about the late submissions of students in different courses.

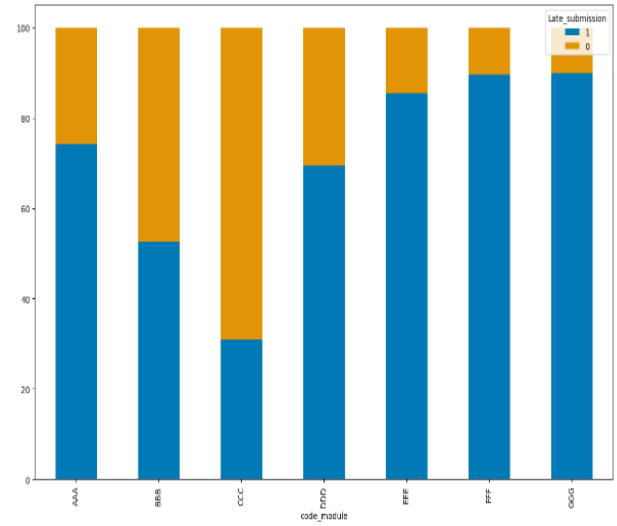

Fig 4 - Late Submissions

In the below graph “0” represents late submission of assessment by students and ”1” represents the submission in time for different Assessments.

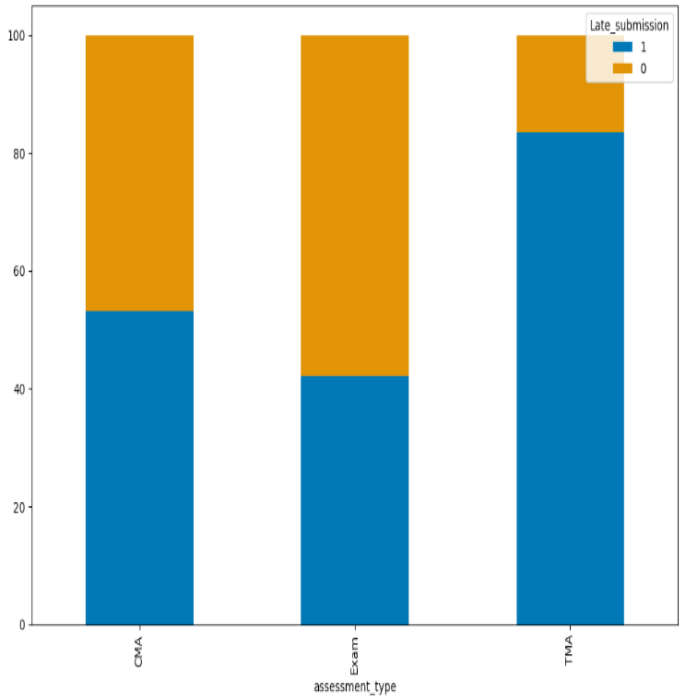

Fig 5 – Assessment Types

IV. EXPERIMENTAL SETUP FOR PREDICTIVE MODELING

In order to model the Learning Analytics Dataset (OULAD) at Open University, we selected six machine learning algorithms to train and test predictive models across different course stages. These algorithms were chosen to divide students performance into two categories : Fail (students completing the course without passing marks or students submitting after the time), Pass (students completing the course with a passing score). Python 3.11.2 scripts were used to aid of Python packages like scikit-learn, NumPy, and seaborn, for building these prediction models.

A. EVALUATION CRITERIA

Prior to advancing with Predictive models are trained and tested at several points during the course session, using the dataset underwent partitioning into training and testing sets employing the method of K-fold cross-validation, where k is set to 10. The dataset is divided into k subsets using this procedure. If k-1 subsets are used to train the model, while 1 subset is kept for testing the model, making it easier to assess the model's performance on untested data.

The following measures have been chosen to assess the models' performance:

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.

RECALL:
Recall makes ensuring that no examples of the positive class are missed by the predictive model. It assesses how well the model can distinguish, from all real positive cases, all pertinent examples of the positive class.

F-SCORE:
The harmonic mean of recall and precision is represented by the F-score, sometimes referred to as the F1 measure. It offers a fair evaluation of the model's performance, especially in situations where there is a class disparity.
When dealing with classification issues involving unbalanced class distributions, the F1 measure is especially helpful.

When taken as a whole, these evaluation criteria offer valuable information on the strength and efficiency of the predictive models created to identify student risk at any point during the course module.

V. EXPERIMENTAL RESULTS

A. PHASE 1: USING ONLY DEMOGRAPHICS DATA TO CONSTRUCT PREDICTIVE MODELS

Precision	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0	0.29	0.26	0.27	0.37	0.38
Pass	0.41	0.43	0.42	0.43	0.46	0.43
Recall	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0	0.22	0.3	0.26	0.07	0.09
Pass	0.7	0.52	0.48	0.5	0.72	0.72
f1-score	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0	0.25	0.28	0.26	0.12	0.15
Pass	0.5	0.47	0.45	0.46	0.54	0.54
Support	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	1411	1411	1411	1411	1411	1411
Pass	2472	2472	2472	2472	2472	2472

Fig 6 Performance scores of prediction models based on demographic data are shown in Fig. 6.

Six predictive models that were exclusively trained on demographic data using K-fold cross-validation with a value of k set to 10 are shown in Figure 6. The 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 at-risk students 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. PHASE II: USING CLICKSTREAM DATA AND DEMOGRAPHICS FOR CONSTRUCTING PREDICTIVE MODELS

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.

Precision	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0.34	0.44	0.31	0.39	0.45	0.49
Pass	0.49	0.59	0.56	0.58	0.58	0.5
Recall	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0.02	0.26	0.3	0.27	0.18	0.25
Pass	0.87	0.8	0.65	0.72	0.79	0.85
f1-score	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	0.04	0.32	0.6	0.32	0.26	0.33
Pass	0.63	0.68	0.58	0.64	0.67	0.7
Support	Support Vector Machine	Random Forest	K- Nearest Neighbor	Extra Tree Classifier	Ada Boost Classifier	Gradient Boost Classifier
Fail	1411	1411	1411	1411	1411	1411
Pass	2472	2472	2472	2472	2472	2472

Fig 7 - Predictive models' performance score when trained with clickstream data and demographics.

Figure 7 displays the performance outcomes of the predictive models developed using demographics, clickstream data. Notably, classifiers such as Random forest, ET, and Gradient Boost demonstrate satisfactory results for the Pass class.

However, Performance scores for Fail and Distinction classes.remain unsatisfactory. Despite the improvements observed when compared to models trained solely on demographics data, the predictive models' performance is still far from acceptable standards.

C. PHASE III: CONSIDERING ASSESSMENT SCORES DEMOGRAPHICS AND CLICKSTREAM, FOR DEVELOPING PREDICTIVE MODELS

To improve prediction model performance, evaluation scores were combined with demographic and clickstream data. The following figure displays the performance scores of predictive models trained on demographics, clickstream, and assessment data. Notably, the average score variable showed a moderate negative connection with the final result, indicating that final result scores increased while average assessment scores decreased. Additionally, both mean clicks and sum clicks variables revealed significant positive relationships. Using assessment score factors. Furthermore, a little correlation was discovered between the late submission variable and the final result.

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	0.79	0.78	0.81
Pass	0.76	0.78	0.72	0.77	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	0.79	0.84	0.82
Pass	0.76	0.82	0.74	0.77	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	0.79	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 8 - The prediction model's performance score when trained with assessment scores, demographics, and clickstream data.

Figure 8 showcases the prediction models' performance

values that were trained using assessment, demographic, and clickstream data. Remarkable enhancements were observed in predictive model performance across Pass, and Fail classes upon integrating assessment data. Specifically, SVM and K-NN models need improvements in performance.

D. 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.

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 9- performance of models after performing feature engineering.

Fig 9 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.

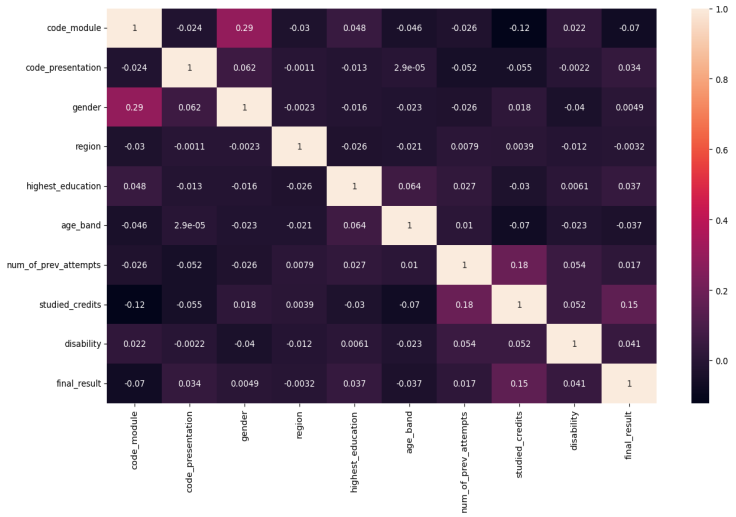

Fig 10 – HeatMap representing Correlation Matrix

Fig 10 represents the correlation matrix it is used to identify the best relation and according to the graph we can know that the best relation is found between gender and code_module but as the value is 0.29 which is very small we can neglect it.

VI. INTERVENING STUDENTS THROUGH PERSUASIVE TECHNIQUES

The Behavior Model of Fogg emphasizes the simultaneous presence of ability, motivation, and triggers to effectively influence positive attitudes. 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"

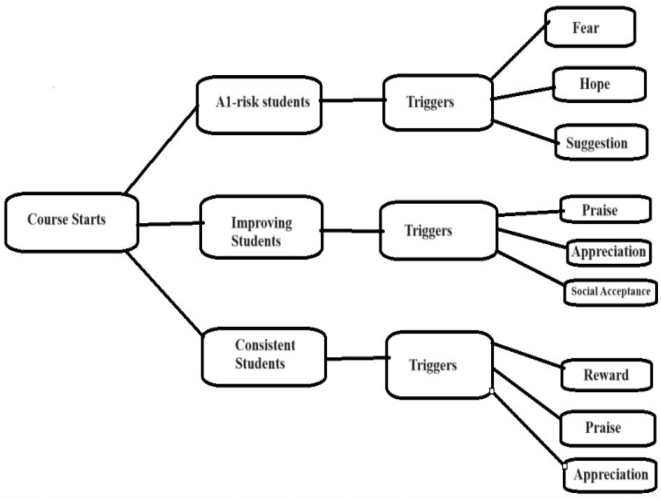

Fig 11 – Different triggers for students of different performances.

VII. CONCLUSION, LIMITATION, FUTURE SCOPE

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 80% accuracy was chosen to forecast student performance. Clickstream data and assessment scores have the biggest influence on the outcome of all the variables.

ALGORITHM	ACCURACY
Support Vector Machine	71.61
K-Nearest Neighbor	80.01
Extra Tree Classifier	78.4
Gradient Boost Algorithm	80.43
Random Forest Classifier	80.75
Ada Boost Classifier	79.45
MLP Classifier	69.55

Table 1 – accuracy of different algorithms

Flow Chart:

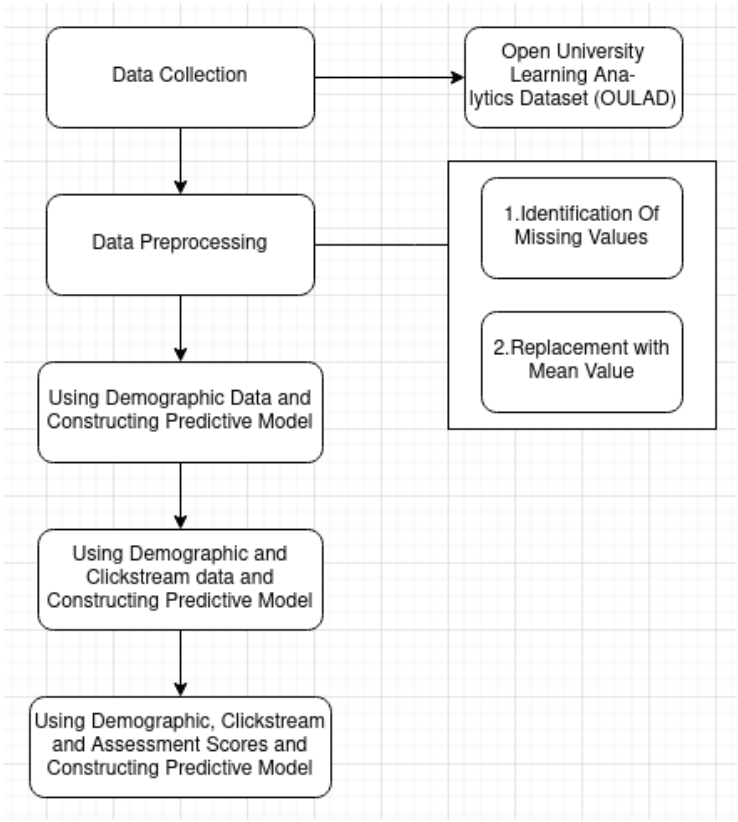

Fig 12 – Flow Chart for project

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 onlineactivities 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.

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] J. Kuzilek, M. Hlosta, D. Herrmannova, Z. Zdrahal, and A. Wolff, "Ou analyse: Analysing at-risk students at the open universities," Learn. Analytics Rev., vol. 8, pp. 1–16, Mar. 2015.

[17] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment," in Proc. 3rd Int. Conf. Learn. Analytics Knowl., 2013, pp. 145–149.

[18] M. Hlosta, D. Herrmannova, L. Vachova, J. Kuzilek, Z. Zdrahal, and A. Wolff, "Model student online behaviour in a virtual learning environment," 2018, arXiv:1811.06369. [Online]. Available: http://arxiv.org/abs/1811.06369

[19] Y. Cui, F. Chen, and A. Shiri, "Scale up predictive models for early detection of at-risk students: A feasibility study," Inf. Learn. Sci., vol. 121, nos. 3–4, pp. 97–116, Feb. 2020.

[20] Y. Lee and J. Choi, "A review of online course dropout research: Implications for practice and future

research," Educ. Technol. Res. Develop., vol. 59, no. 5, pp. 593–618, Oct. 2011.

[21] P. H. Winne, "Improving measurements of self-regulated learning," Educ. Psychologist, vol. 45, no. 4, pp. 267–276, Oct. 2010.

[22] R. Baker, B. Evans, Q. Li, and B. Cung, "Does inducing students to schedule lecture watching online classes improve their academic performance? An experimental analysis of a time management intervention," Res. Higher Edu., vol. 60, no. 4, pp. 521–552, 2019.

[23] J. M. Lim, "Predicting successful completion using student delay indicators in self-paced online courses," Distance Edu., vol. 37, no. 3, pp. 317–332, Sep. 2016.

[24] J. Park, K. Denaro, F. Rodriguez, P. Smyth, and M. Warschauer, "Detecting changes in student behavior