Predicting Student Assessment Marks to Prevent Student Attrition in MOOCs

Angelin Ann Jacob
PES2UG19CS041
DEPARTMENT OF CSE,
PES UNIVERSITY,
BENGALURU,KARNATAKA
pes2ug19cs041@pesu.pes.edu

B Vivek Sai Chinna
PES2UG19CS077
DEPARTMENT OF CSE,
PES UNIVERSITY,
BENGALURU,KARNATAKA
PES2UG19CS077@pesu.pes.edu

Deepa Shree C V
PES2UG19CS105
DEPARTMENT OF CSE,
PES UNIVERSITY,
BENGALURU,KARNATAKA
pes2ug19cs105@pesu.pes.edu

Rongali Lalith Vardhan
PES2UG19CS337
DEPARTMENT OF CSE
PES UNIVERSITY
BENGALURU, KARNATAKA
PES2UG19CS337@pesu.pes.edu

***Abstract* — This paper presents the approach of machine learning techniques to help instructors understand students better with the help of assessment scores. Ensemble learning was mainly used to perform the predictions. We have omitted the demographic features of the student and have focussed only on the academic side of the dataset. The workflow of the current project includes EDA, model building and hyperparameter tuning. We have performed a few EDA techniques like standard scaling, one-hot-encoding as a part of pre-processing. Moving on to model-building, we have worked on Random Forest Classifiers, AdaBoost and XGBoost classifiers. As the initial accuracies and other performance metrics were low, hyperparameter tuning has been carried out using cross validation techniques like Grid-Search CV. Our work mainly aims at assisting course instructors and designers to improve student experience in online courses and increase student attention rates and participation levels.**

***Keywords* — assessment mark prediction, vle, parquet files, ensemble learning, one-hot-encoding, hyperparameter tuning**

I. INTRODUCTION

Growth of technology has led to the introduction of digitalization and virtualisation in various domains. Education is one amongst these. A lot of MOOC platforms have been introduced with the goal of providing quality education to everyone. Students have been enrolling in these courses and the outbreak of the pandemic has seen a sharp increase in these numbers. Though people enrol in courses, only a handful of them manage to complete them and earn a certificate. A published study of 17 Harvard and MIT online courses on diverse topics[1], offered on the provided. The formatter will need to create these components, incorporating the applicable criteria that follow. edX platform between fall 2012 and summer 2013 reveals that course completion rates remain abysmally low (only 43,196 of the 841,687 registrants).

The current problem at hand has been an area of huge interest in the field of education mining. Researchers have been working for a while to leverage machine learning to improve the MOOC experience. Given the current situation of the world, there is a high chance that most of the learning will continue being online for a long time to come. Hence, it is crucial to assist students at the right time, to ensure they progress well in the course.

Our aim is to predict student assessment marks beforehand, so that attention can be paid to the students who are at a risk of dropping out. It is not only important to predict if the student is at a risk of attrition, but also important to help improve his/her performance. So, our aim is to predict the score of the student, rather than just predicting if he/she will pass or fail the course. By doing this, the instructor can know which students are lagging behind well in advance. This would help take suitable measures to assist students in their betterment and also complete the course with a better grade.

II. LITERATURE REVIEW

Before moving on with the problem at hand, we explored a few other works and infer insights from their approaches.

[2]The paper focuses on investigating the performance of the student in the course. This paper answers (1) to predict whether a student will drop out from the course (2) to predict whether a student that does not drop out will pass or fail the course. The paper uses Regressions, ANNs and Expert Learning. Students were also categorized on the

basis of their daily use of VLE. The paper also discusses how recent approaches include a hybrid approach involving Convolution Neural Network and Recurrent Neural Network, K-means clustering combined with Apriori Algorithms to find Association Rules. But the paper finalises on ANN, Distributed Random Forest, Gradient Boosting Machine, Deep Learning and Generalised Linear Model were used.

[3]This paper focuses on the comparison of Artificial Neural Network (ANN) and Random Forest (RF) models to predict performance of the students based on their demographic and assessment. Feature engineering is also put to use for better enhancement of the model. This paper also historicizes the primitive models which were put to use like.All different types of activities of the VLE environment were combined into a single metric(number of days).

[4]This paper focuses on a newer, more novel approach called Generalised additive models of location, scale and shape (GAMLSS). GAMLSS is implemented in R through the gamlss package. The goal of the following modelling is to illustrate how GAMLSS can be used in practice and has no attempt at making theoretical LA/EDM-related claims based on the OULAD data set. A likelihood ratio aimed at determining whether the GAMLSS scale and shape parameters were constant for all observations suggested these parameters were not constant.

[5]Student grades are predicted and classification is carried out. Decision Trees, K Nearest Neighbors and Logistic Regression are the models used. A few dimensionality reduction techniques like PCA, LDA have also been used. The dataset used is OULAD. Kappa and Accuracy were used to measure the performance of the models.

[6]This paper uses time series classification techniques by exploiting VLE(Virtual Learning Environment) data. LSTM is used to build the model. The model consists of many non-linear layers and the complexity increases with the depth. The best performance was obtained with a 3-layer LSTM. The model works well with sequential data but fails to capture patterns when introduced to aggregate data.

[7]The authors aim to predict the graded performance of the students on MOOC assessments. Personalised Linear Regression Model(PLMR) is used to perform the predictions. The EdX MOOC dataset has been used here. Performance metrics used are the F1-Score and the R2 score.

[8] This paper aims for classifying the students into Withdrawn, Fail, Pass, and Distinction, rather than only Completers and Non-completers (two categories). The author predicted the final outcome of the student at the very beginning of the course by considering the first assignment marks. Pearson correlation test was done to measure strength association between variables. The predictive model was implemented through machine learning algorithms Decision Tree, Random Forest and Bayesian Additive Regression Trees (BART). Finally, the author concluded that the BART algorithm gave good values compared to the other models.

[9]Here,XG boost machine learning algorithm was implemented on a model for predicting withdrawal of students on different levels of OULAD dataset by author.Here precision was equal to accuracy and the recall was recorded as 1. This means the implemented algorithm has shown every single instance as ‘not withdrawn’. This would lead to incorrect predictions.

[10] This paper showed the results of a classification study where the author had developed a model on predicting the success or failure of student's exams.Based on students interaction with the virtual learning platforms(VLE), the author has applied adaptive stream data analysis techniques on OULAD dataset for predicting the student’s final outcome.RF and ARF were implemented here.At last the author concluded Random forest machine learning algorithm has given the the highest accuracy to the model.

[11]. This paper focuses on the Business intelligence framework for AE-LS to monitor and manage the performance of the learner more effectively. It is a set of approaches that combine to give learners an adequate content meeting their requirements. To achieve all the four major cycles, Learning Analytics uses different methods and techniques such as: Statistics, Information Visualization, Data Mining,Business Intelligence, and Social Network analysis. A model of DWH has been proposed.

[12]In this paper we analyse the interaction between the students and electronic learning systems. This type of analysis serves in predicting the student scores, in alerting students-at-risk, and in managing the degree of student engagement to the educational system. The approaches in this work implements the divide and conquer algorithm on the feature set of an educational data set to enhance the analysis and prediction accuracy. SVM, Decision trees and neural networks are used here. A comparison among these algorithms shows that the Random Forest algorithm has the best classification/prediction accuracy 84.99.

[13] The paper uses SVM to predict the final result from the learner’s activity. In the experiment focusing on the presence or absence of behavioral features, in the case of ANN with the best result, it is classified with an accuracy of 79.1% .

Our Plan: Most of the papers we surveyed mostly focus on the final pass or fail result but not the assessment marks. We aim to predict student assessment marks to prevent student attrition using ensemble learning classification techniques. This would help the instructor

and the course organizers take early action to improve student experience.

III.DATASET AND PREPROCESSING

The dataset used is OULAD[2].This dataset contains information about students who have enrolled in MOOCs of the famous Open University of UK. The acronym OULAD is read as Open University Learning Analytics Dataset. The dataset was collected from Kaggle. It contains seven csv files. [14] The schema of the dataset and the relation between the seven files, has been provided in figure 1. All the seven csv files were used to pre-process the data and get more insights by plotting various graphs between various attributes. The attributes presented in Table 1 have been used to carry out the prediction process for the current project.

The sum of the clicks was added up to give one count per day. The mentioned columns are looked up for null values and the appropriate values were filled in by analysing all the csv files of the dataset. None of the outliers were removed, as they were all valid and contribute towards training the models. The categorical columns were one-hot encoded. Hence, three columns were obtained from the assessment_type variable. The values are 1 and 0, which correspond to a binary value of yes or no in the corresponding column.The final dataset was compressed and stored in the form of a parquet file. This format is almost one-fourth times smaller than the usual csv files. The score is divided into ten different classes. The ranges and the respective classes are as given in table 2.

COLUMN	DESCRIPTION	Type
Num_of_prev_attempts	Number of times the student has already taken up the course	Integer
Studied_credits	The total number of credits offered by the course	Integer
Date_assessment	The deadline of the assessment in days from the start of the module	Integer
Sum_click	The number of times a student has accessed the learning material in a day	Integer
Date_registration	The date of registration of the student in days from the start of the module	Integer
Weight	The weightage of each assessment towards the final score	Integer
Score	The target variable, the score of student in each assessment	Integer
Assessment_type	The type of assessment. Mainly three types, (tutor marked, computer graded, exam)	object

Table 1 – List of attributes used for the prediction

Also, after the dataset is loaded, a few columns are downsized to a lower bit version of the same datatype. The score and number are downsized to 8-bit from 16-bit integers for faster training.After obtaining the required variables and removing the duplicates, the dataset had around 3 million rows, which is a good amount of data for training a model.

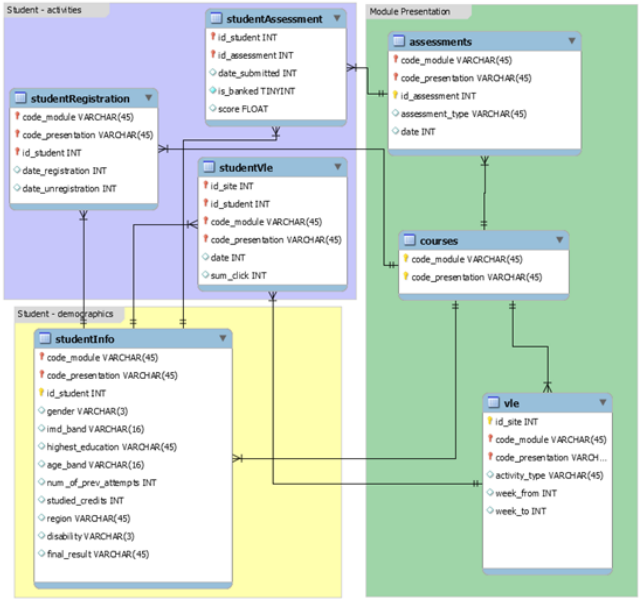


Figure 1: Schema of the dataset

Class	Range of Score
0	0-10
1	11-20
2	21-30
3	31-40
4	41-50
5	51-60
6	61-70
7	71-80
8	81-90
9	91-100

Table 2 - Range of Scores and the corresponding target classes

IV. MODEL BUILDING

The current project focuses on the use of ensemble learning to achieve the set goal. Ensemble Learning is the process which involves several weak learners that are combined to give one complex model. A few models might perform well in a few scenarios. By using ensemble learning, we try to leverage all the possible advantages of each of the models and combine them to build a good, complex model, capable of handling all the kinds of test cases.This paper mainly focuses on two kinds of ensemble learning techniques:

1. Bagging: Also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once. This is a type of parallel learning.
2. Boosting: In boosting, a random sample of data is selected, fitted with a model and then trained sequentially—that is, each model tries to compensate for the weaknesses of its predecessor.

A total of three models have been used to perform the predictions.

I. Random Forest

Random Forest is a bagging technique. As the name implies, they are a collection of decision trees. For classification the output of the random forest is selected by the method of majority voting.

For the current problem, the numerical values, except the score, are standardised before passing for training.

Grid Search Cross Validation is also performed to get a better set of hyperparameters. After running Grid Search Cross validation for 3 epochs, the optimal depth of the random forest model was obtained as 36. The log loss was obtained as 1.77. The confusion matrix and the classification report are shown in Figure 2 and 3 respectively.

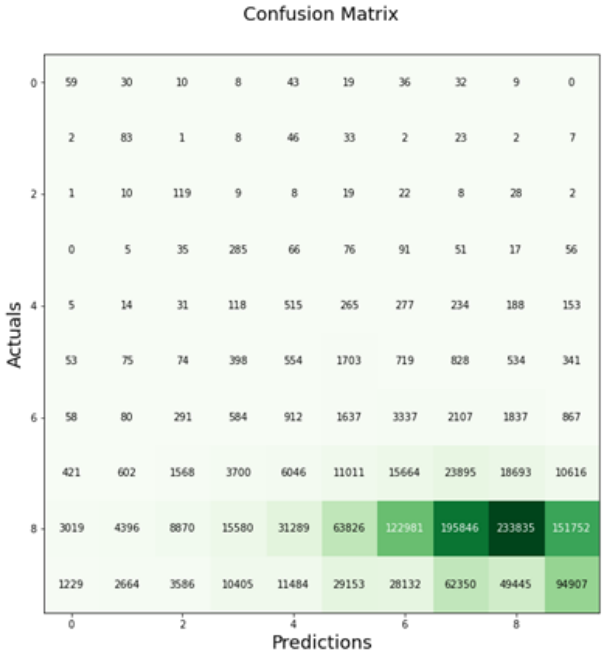


Figure 2- Confusion matrix of random forest classifier

	precision	recall	f1-score	support
0	0.41	0.29	0.34	4847
1	0.44	0.30	0.36	7959
2	0.44	0.30	0.36	14585
3	0.41	0.30	0.34	31095
4	0.41	0.28	0.33	50963
5	0.37	0.27	0.32	107742
6	0.37	0.31	0.34	171261
7	0.38	0.40	0.39	285374
8	0.40	0.49	0.44	304588
9	0.46	0.47	0.47	258701
accuracy			0.41	1237115
macro avg	0.41	0.34	0.37	1237115
weighted avg	0.40	0.41	0.40	1237115

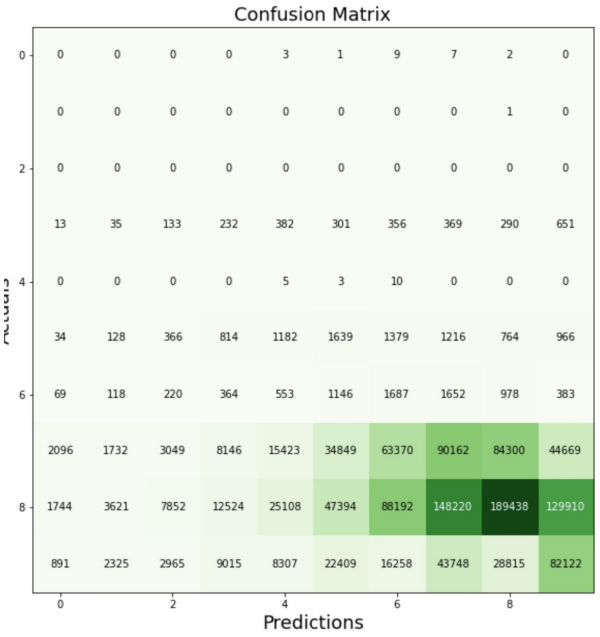
Figure 3- Classification report of random forest classifier

II. AdaBoost

AdaBoost, also called Adaptive Boosting, is a technique in Machine Learning used as an Ensemble Method.

The most common algorithm used with AdaBoost is decision trees with one level that means decision trees with only 1 split.

These trees are also called Decision Stumps. This model gives equal weights to all the data points and then assigns higher weights to the points that are wrongly classified and reduces the weights to the points which are wrongly classified. It keeps training the model unless a lower error rate is achieved.



	precision	recall	f1-score	support
0	0.00	0.00	0.00	4847
1	0.00	0.00	0.00	7959
2	0.00	0.00	0.00	14585
3	0.08	0.01	0.01	31095
4	0.28	0.00	0.00	50963
5	0.19	0.02	0.03	107742
6	0.24	0.01	0.02	171261
7	0.26	0.32	0.28	285374
8	0.29	0.62	0.40	304588
9	0.38	0.32	0.35	258701
accuracy			0.30	1237115
macro avg	0.17	0.13	0.11	1237115
weighted avg	0.27	0.30	0.24	1237115

Figure 4 and 5 - Confusion matrix Classification report of Adaboost

III. XGBoost

XGBoost is an optimized distributed gradient boosting library. It is a part of the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way. Based on Decision-tree ensemble learning algorithm, it is used when the data is unstructured and dealing with highly tabular data.XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

XGBoost provides built-in Regularization for helping us to reduce the highly dimensional dataset as OULAD at hand. Weights are produced appropriately and this is the key factor for our project, as features rely heavily on a few attributes like (credit score, date_registration and

date_unregistration) while other attributes do not contribute a lot for predictions.

The metrics return a value of 1.71694 for the XGBoost model.

As we see that the confusion matrix shows the model performing well on students who receive higher marks (7,8,9 classes). The model classifies 8 very well, but has a chance of misclassifying 8 as 9/7. But this classification error is acceptable as we can consider 7 to 9 as a distinction, so students are assured distinction either way. Since the difference between 7 and 8, or 8 and 9 does not prove much, this error is acceptable.

We also use the confusion matrix to give us some insights about the predictability of the model and what we can improve in order to help us increase the same. We also see that the model does not predict classes as efficiently as higher classes. This can be attributed to the reason that the model sees higher classes than lower classes/failures.

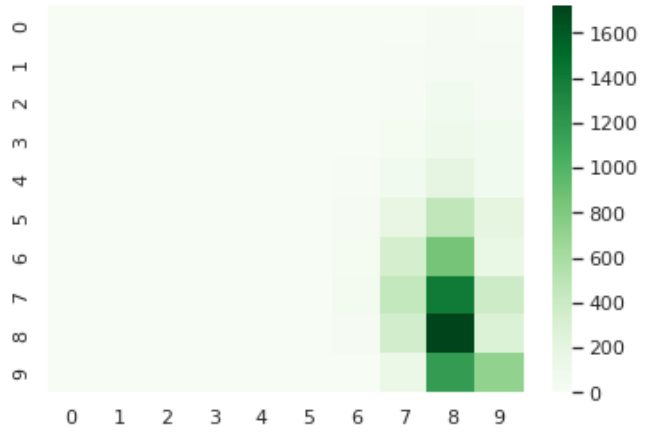


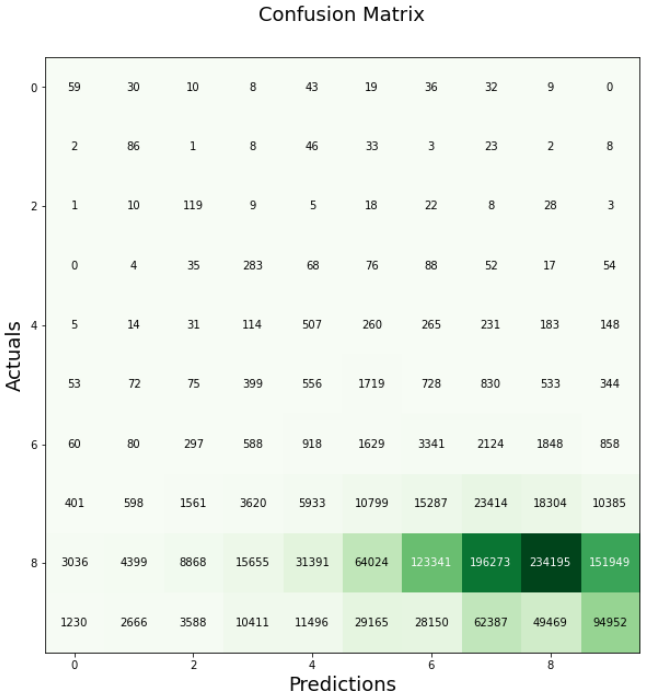
Figure 6 - Confusion matrix of XGBoost classifier

	precision	recall	f1-score	support
0	0.51	0.12	0.19	1323
1	0.00	0.00	0.00	2107
2	0.00	0.00	0.00	3911
3	0.00	0.00	0.00	8241
4	0.20	0.00	0.00	13836
5	0.24	0.02	0.04	28754
6	0.29	0.02	0.03	45492
7	0.28	0.24	0.26	76017
8	0.30	0.68	0.41	81078
9	0.38	0.40	0.39	69241
accuracy			0.31	330000
macro avg	0.22	0.15	0.13	330000
weighted avg	0.29	0.31	0.25	330000

Figure 7 - Classification report of random forest classifier

IV. LightGBM

Light Gradient Boosting Machine (LightGBM) is another classifier, which as the name implies, works on the principle of boosting.It is a fast, gradient based algorithm which can be used for both regression and classification. The preprocessing carried out was the same as random forest.Grid Search was performed. A log loss of around 1.76 was obtained. The classification report and the confusion matrix are given below.



	precision	recall	f1-score	support
0	0.24	0.01	0.02	4847
1	0.41	0.01	0.02	7959
2	0.53	0.01	0.02	14585
3	0.42	0.01	0.02	31095
4	0.29	0.01	0.02	50963
5	0.32	0.02	0.03	107742
6	0.28	0.02	0.04	171261
7	0.26	0.08	0.12	285374
8	0.28	0.77	0.41	304588
9	0.32	0.37	0.34	258701
accuracy			0.29	1237115
macro avg	0.34	0.13	0.10	1237115
weighted avg	0.30	0.29	0.21	1237115

Figure 8 and 9 - Confusion matrix and Classification report of LightGBM

IV. PERFORMANCE METRICS

We use Classification report, Confusion Matrix and most importantly log-less functions as our performance metrics.

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Actuals and Predicted values are present across the diagonals. All metrics return promising results with some error in predictions of class 8 and 9, which is fairly acceptable.

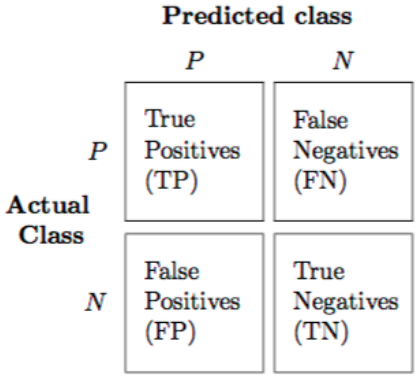


Figure 8 - Confusion Matrix by outcomes and observations

Log Loss is the most important classification metric based on probabilities. It's hard to interpret raw log-loss values, but log-loss is still a good metric for comparing models. For any given problem, a lower log loss value means better predictions.We have used the log loss function in the scikit learn package as our metrics for measuring the predictability for the classification of our model.

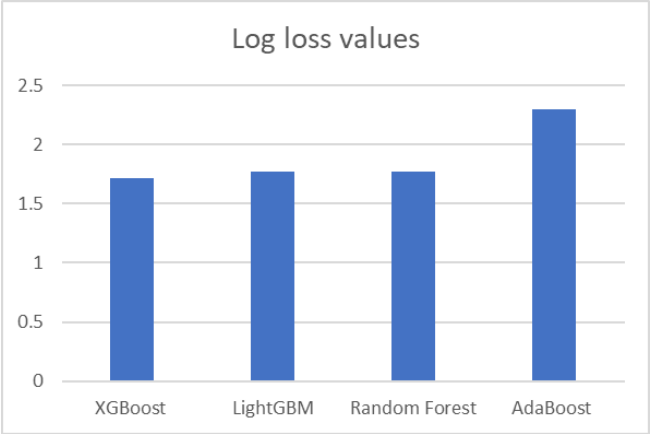
$$logloss = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{M} y_{ij}\log(p_{ij})$$

- N is the number of rows
- M is the number of classes

Figure 9 - Cross Entropy/Log Loss for MultiClass Classification

Out of all the models, XGBoost gives the lowest log loss value(i.e 1.71), followed by LightGBM, Random Forest and then finally AdaBoost.

Model Name	Log loss values
XGBoost	1.71694
LightGBM	1.76862
Random Forest	1.77233
AdaBoost	2.29396



V. CONCLUSION

We leveraged Machine learning techniques, to assist the course designers and instructors to identify at-risk students and help improve their performance. While the other papers talked about predicting at-risk students, we aimed at not only detection but also improvement of the student performance by predicting assessment marks. Since this is an area of interest and a lot of research is going on, we wish to use big data techniques and a few other sophisticated techniques to obtain better predictions.

VI. REFERENCES

[1] Ho, Andrew & Reich, Justin & Nesterko, Sergiy & Seaton, Daniel & Mullaney, Tommy & Waldo, Jim & Chuang, Isaak. (2014). HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013. SSRN Electronic Journal. 10.2139/ssrn.2381263

[2] OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques Nikhil Indrashekhar Jha1, Ioana Ghergulescu and Arghir-Nicolae Moldovan School of Computing, National

College of Ireland, Dublin, Ireland Adaptemy, Dublin, Ireland

[3] A Machine Learning Based Approach for Student Performance Evaluation in Educational Data Mining Muhammad Sammy Ahmad

[4] Distributional regression analysis of learning analytics and educational data Fernando Marmolejo-Ramos · Mauricio Tejo · Marek Brabec · Jakub Kuzilek · Srecko Joksimovic · Vitomir Kovanovic · Jorge Gonzalez ´· Raydonal Ospina

[5] Poudyal Sujan, Nagahi Morteza, Nagahisarchoghaei Mohammad and Ghanbari Ghodsieh, "Machine Learning Techniques for Determining Students' Academic Performance: A Sustainable Development Case for Engineering Education," 2020

[6] Aljohani, N.R.; Fayoumi, A.; Hassan, S.-U. "Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment," *Sustainability* **2019**, *11*, 7238. https://doi.org/10.3390/su11247238

[7] Zhiyun Ren;Huzefa Rangwala;Aditya Johri, "Predicting Performance on MOOC Assessments using Multi-Regression Models," 2016,arXiv,cs.CY.

[8]"Early Predictor for Student Success Based on Behavioural and Demographic Indicators", Efthyvoulos Drousiotis1, Lei Shi2, Simon Maskell1,2021

[9] "Identifying at-risk students across different stages of distance learning courses and identifying their most relevant predictors at each stage",Paul Duncker, 2020

[10] "Educational Stream Data Analysis: A Case Study" Gabriella Casalino, Giovanna Castellano, Gennaro Vessio,2020

[11] El Janati, Salma & Maach, Abdelilah & El Ghanami, Driss. (2019). Learning Analytics Framework for Adaptive E-learning System to Monitor the Learner's Activities. International Journal of Advanced Computer Science and Applications. 10. 275-284. 10.14569/IJACSA.2019.0100835.

[12] Eslam Abou Gamie, M. Samir Abou El-Seoud,Mostafa A. Salama,"A layered-analysis of the features in higher education data set",(2019)

[13] Kitanaka, Yuki & Takeuchi, Kazuhiro & Hirokawa, Sachio. (2017). Predicting Learning Result of Learner in E-learning Course with Feature Selection Using SVM. 122-125. 10.1145/3175536.3175567.

[14]Kuzilek, J., Hlosta, M. & Zdrahal, Z. Open University Learning Analytics dataset. *Sci Data* **4,** 170171 (2017). https://doi.org/10.1038/sdata.2017.171