

Analyzing Student Behavior to Predict Student Result in VLE (Virtual Learning Environment)

Umair Bin Ahmad¹ | Abdullah Riaz² | Muhammad Abubakar³

Mukarram Ahmad⁴ | Muhammad Mubashar⁵

Department of Computer
Science, Information
Technology University, Lahore
Pakistan.^{1,2,3,4,5}

¹ MSDS 19036

² MSDS 19054

³ MSDS 19086

⁴ MSDS 19090

⁵ MSDS 19095

Abstract

Today, distance learning has become a very popular trend and has had a positive impact on learning, especially with the development of technology, online virtual learning environments play a vital role in providing education. These virtual learning environments generate a huge amount of data. By collecting and analyzing learner's data, their learning experience can be improved by providing them informed guidance and improving course materials. By integrating data streams into progressive decision making, the current development of interdisciplinary learning analysis research presents important challenges for the use of behavioral analysis. Identifying students at risk of failure is related to many educational strategies to improve their capabilities and skills through timely intervention by academic institutions. Foreseeing student performance is a crucial decision-making problem, including data from several modules that can be fused into standardized vector to make a decision. This research exploits a sequential classification problem to predict success and failure of students, by analyzing actionable data in the form of learners' interactional actions with the online virtual learning system, using openly available OULA (Open University Learning Analytics) dataset. Various machine learning algorithms were deployed to train model and tested and achieved accuracy of about 87.8 %, 83.20 % precision, and 87.8 % recall.

Keywords

Machine Learning, VLE (Virtual Learning Environment), OULA (Open University Learning Analytics), EDM, LSTM (Long Short Term Memory), Student assessments.

1. Introduction:

With the rapid development of information technology, the amount of student data collected by the education sector has increased significantly. In addition, Virtual Learning Environments emerged and moved courses to the Internet. Such online learning systems also contribute in generating adequate data as a by-product, to discover and mine appropriate student behaviors to address issues in online learning environments and encompassing the early prediction of student's success or failure. Distance learning has become an important part of the education system. This is very beneficial for full-time staff, military personnel, and non-residents or individuals in remote areas that cannot attend classroom lectures. Even ordinary students can benefit from it to gain professional knowledge. However, because there is very little interaction between students and organizations, it is difficult to analyze student needs and overcome shortcomings. It is important to monitor student participation and evaluate their progress over time. With the recent rise of the learning analytics community, educator's collaboration not only examines student behavior to gain a comprehensive understanding; But they also have the best policies for maintaining the best teaching methods that lead to a successful educational environment, which helps resume behavior of the Institute. All aspects of learning analysis in higher education defined by integrating academic analysis to provide financial support to their resource allocation procedures explain learners demonstrate corrective action to demonstrate behavior and exercise stability teacher and lecturer strategies. It helps a college to reduce student learning the attrition rate ultimately improves the completion rate of the university.

For our research we used (OULA) Open University Learning Analytics dataset. The Open University is one of the world's largest distance learning university. Currently, approximately 170,000 students have registered for different courses. The data collected has enough student information to predict their final result. The main idea is to involve students in the module, make them perform better, and improve the content or style of the module when needed. One of the key features that directly affects the end result is the online clicks. The number of clicks each student has on the module is an important factor as it shows that they have been actively researching and learning from the module. Teaching materials, lecture notes and other content are delivered to students via the VLE. Students' interactions with the educational materials are logged and stored. Dataset consists of both student demographic data and interaction data with the university's VLE (i.e. clicks on the various modules, such as viewing and downloading course materials and, submitting assessments, visiting discussion forums etc.)

In our proposed research work we analyzed student behavior by evaluating the clickstream behavior of students and try to predict student final result as pass or fail. Raw data is

processed into executable form, which characteristically contributes in the compilation of interactive student activities in a click-stream format.

After preprocessing various machine learning algorithms were applied (i.e. Logistic Regression, Random Forest, GridSearch). Our system correctly predicts with accuracy 87% of the success or failure of students on the basis of whole semester sum clicks. Some preliminary analysis of the identification attributes for better understanding of data. After selecting important features, we use three machine learning algorithms (Logistic Regression, Random Forest, GridSearch) to predict the final result of students. For evaluation, accuracy map is drawn for each model. Confusion matrices were also drawn that gave the true positive rate (TPR) and false positive rate (FPR).

2. Literature Review:

Now a days, distance learning has become very popular trend and plays a positive impact on learning, especially with the advancement of technology online virtual learning environments had played a vital role in providing education. According to the National Center for Education Statistics, more than 5M students are enrolled in distance education courses at present. It is highly advantages for full time workers, students, nonresidents and individuals specially living in remote areas who are unable to attend classroom lectures.

In online virtual learning environments, the interaction between students and organization is very less, [1] It is hard to analyze the needs and requirements of students to overcome their deficiencies. So, it is important to monitor student behavior and engagement and assess their progress over a period.

Many online virtual learning environments are now a days available and accessible to students which offer a variety of short courses, nano degree programs. These online learning systems also generate immense amount of data, that can be used to explore and mine appropriate student behaviors to cope with problems arises in online learning environments.

Increasing trend and fame of MOOCs (Massive Open Online Courses) and remote learning makes it an interesting area of research. Most of the research work has been done on virtual learning environment. Recently data mining techniques are widely used on educational dataset. EDM (Educational Data Mining) has been very hot research area in recent years.

[2] This paper introduces the use of EDM (Education Data Mining) and discuss data mining techniques for the prediction of student learning outcomes using data sets of senior students. J48 algorithm and WEKA was used to analyze student's data. Main aim of this research is to focus on discovering several indicators that may effect on hypothetical performance of students.

[3] An Examination of Online Learning Effectiveness Using Data Mining; this paper discusses the techniques of data mining, where students online experience is assessed based on their log files. But there is not predicted outcome of student's success & failure.

[4] In this article, a method was proposed to improve the student's performance by using K-mean clustering algorithm to map the student's data and group the data set into clusters, but there is no future performance prediction.

[5] In this research paper, a comprehensive description of the dataset along with the data collection technique is described, the whole process of data preparation is given. Attributes of each table is described. Authors have performed data visualization to ensure that all features or classes have good distribution.

[6] The reference paper describes, the authors had done analysis on OULA dataset to understand how different features such as region, module, assessment and sum of clicks affects the final result. Moreover, decision tree is used to recognize important features and top ten features or attributes were used for classification. For model evaluation Pearson chi-squared is used.

[7] In this research article, deep neural networks (LSTM model) was trained and deployed to predict early withdrawal of students by analyzing their 25 weeks data. For this research work OULA (Open University Learning Analytics) dataset was used. Furthermore, data was merged into single table and converted into week wise activities, then various techniques such as ANN (Artificial Neural Networks) and LSTM (Long Short-Term Memory) was applied to envisage forthcoming withdrawal of students.

[8] In this reference paper, the authors conducted a case study, to reveal online learning patterns and behavior in online learning environments by applying data mining techniques on log files generated by Virtual Learning environments and Learning management systems. This paper demonstrated how data mining techniques can be applied to benefit online coaching and education with suggestions for online instructors, and courseware developers.

Overall, we are working on prediction of student's success and failure by studying their demographic information and analyzing their behavior on virtual learning environment (VLE). Our study will help organization to know student deficiencies and make strategy to overcome and enhance participants' learning experience.

3. Data

3.1 Data Description

The data used for this research is Open University learning analytics dataset [5]. It contains 7 tables which are interlinked through unique identifiers. These tables contain information about courses, assessment, student's info and their interactions with virtual learning

environment for the courses which they select. These tables contain seven courses named as modules which are started twice a year. There are courses presentation represents the year and the month in which course module is started. Course module consists of year, the “B” represents the presentation starts in February and “J” represents the presentation starts in October. Every course length is represented in days. Student Info table represents the demographical data of a student. This dataset contains the information of 32593 students over a course of 9 months from 2013 to 2014 in which 7 courses are offered twice a year. Data is classified into four sections on the base of student final result. Data is classified as 9% of distinction student, 38% pass student, 31% withdrawn students and 22% fail students. This dataset also contains the information of student interactions which have almost 20 different categories of each student. These categories contain the click of pdf files, video links, quizzes, assignment submission etc. These activities correspond to 20 different kinds, namely dual_pane, data_plus, external_quiz, folder, forums, glossary, homepage, htmlactivity, ouelluminate, ou_collaborate, ou_content, ou_wiki, page, questionnaire, quiz, repeat_activity, resource, shared_subpage, subpage, and url, with each activity indicating a particular behavior in the learning environment. Activity types along their description is given below in Appendix Table 5. There are total number of 15382 pass instances with total number of 32603275 clicks and total number of 6680 pass instances with total number of 4765668 clicks.

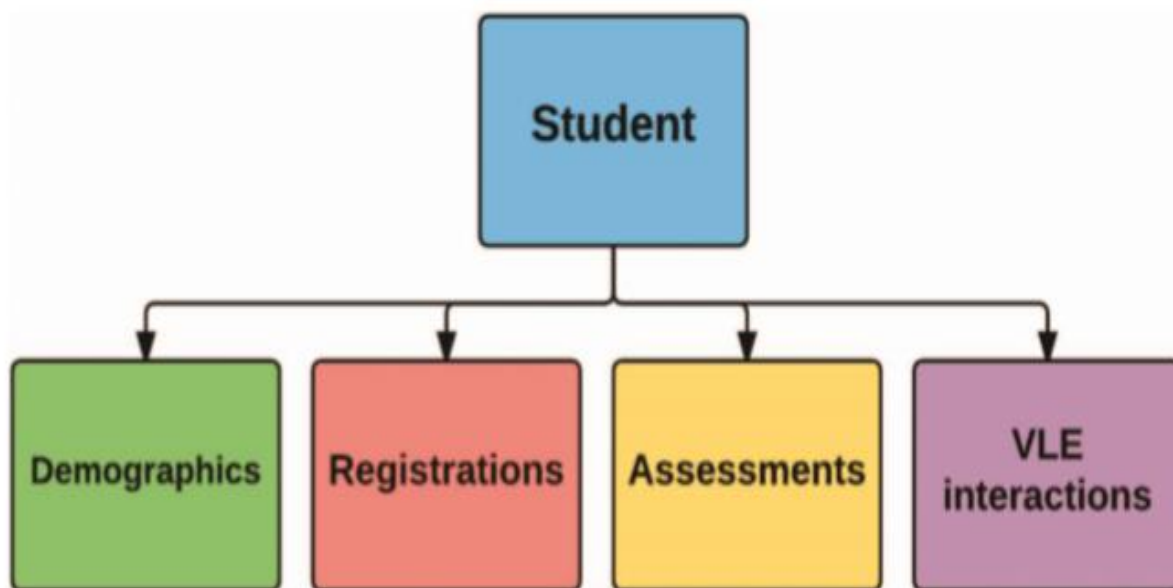


Fig 1 Overall dataset structure

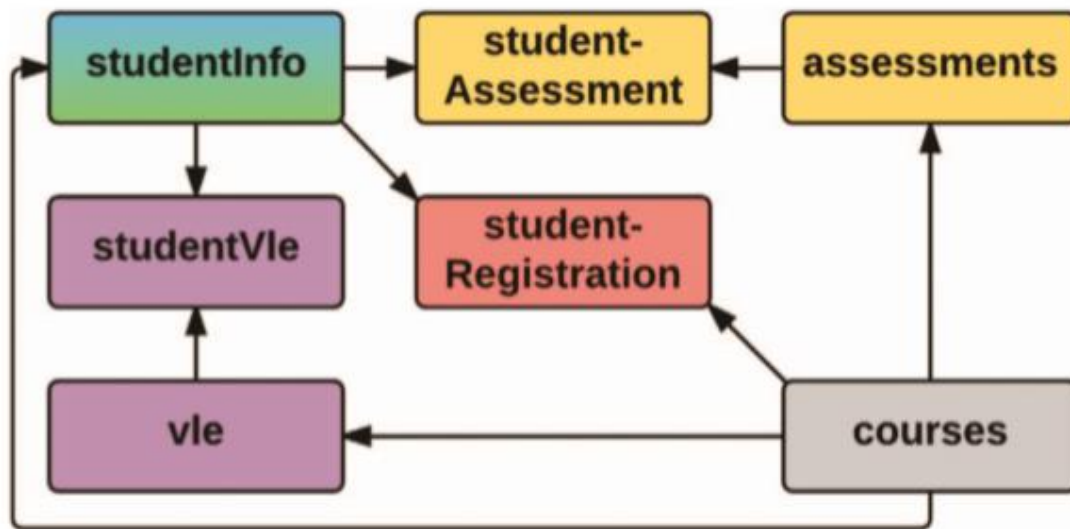


Fig 2 Tables structure

3.2 Data Preprocessing

As we discussed that the data set of OULA is spread into different tables. Our goal is to predict the student is pass or fail by its demographical and clickstream behavior. For that in studentInfo table all the entries of students who had withdrawn the course are discarded, and a single class was created by merging both entries of “Pass” and “Distinction”. All the null values were found in “imd_band” were replaced by the mode of that column and then analysis is done on data by plotting different graphs which helps in understanding better understanding on data.

In Fig 3 we plot the graph of age according to the target class which is in our case is the final result. It shows that as students of higher age are passing as compared to the lower age students. And the fail ratio is more in younger students. So, this can be a feature to predict the final result of a student. And in Fig 4 plot the graph of disability according to the final result. And it shows that passing ratio is higher in case of those students who are not disable and failure rate is higher in case of disability so this is also a feature. Fig 5 shows the relationship between higher education and final result. And in this case, it shows that the

students with higher education have higher passing ratio and the failure rate is higher in case of no formal qualification.

So, this is also a factor which effect the final result. Based on the analysis, it can be seen in Fig 6 that gender does not show any key role in final result as it is almost the same in both cases for male and female.

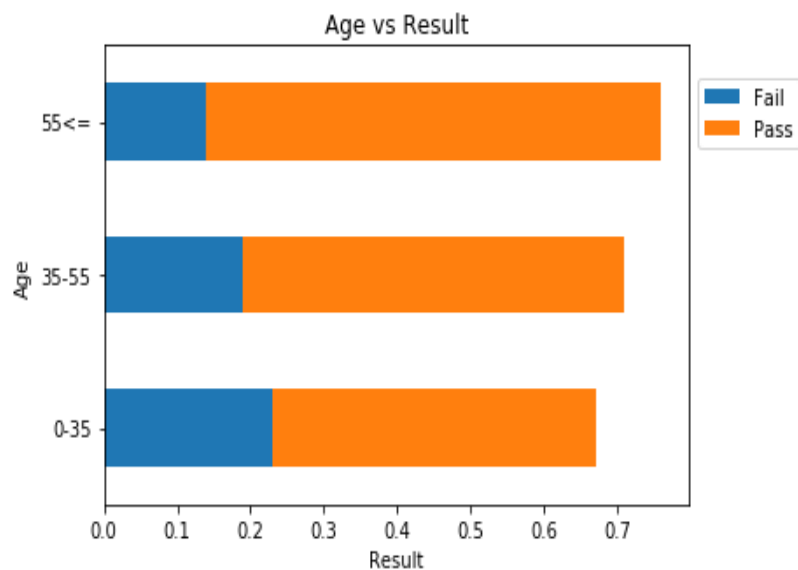


Fig 3 Gender VS Final Result

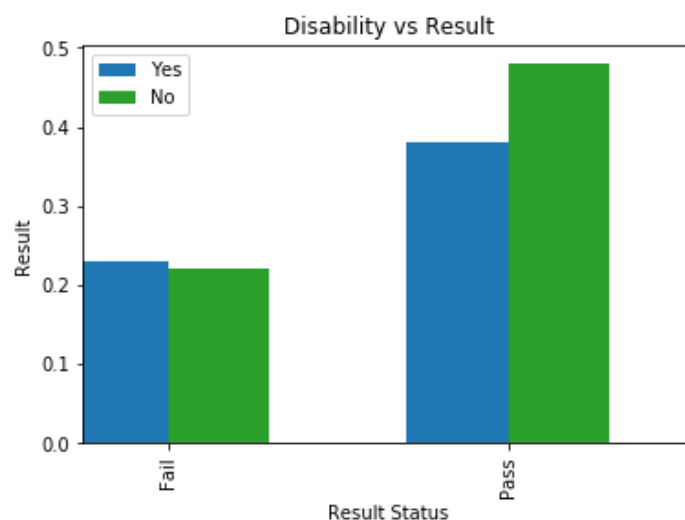


Fig 4 Disability vs Result

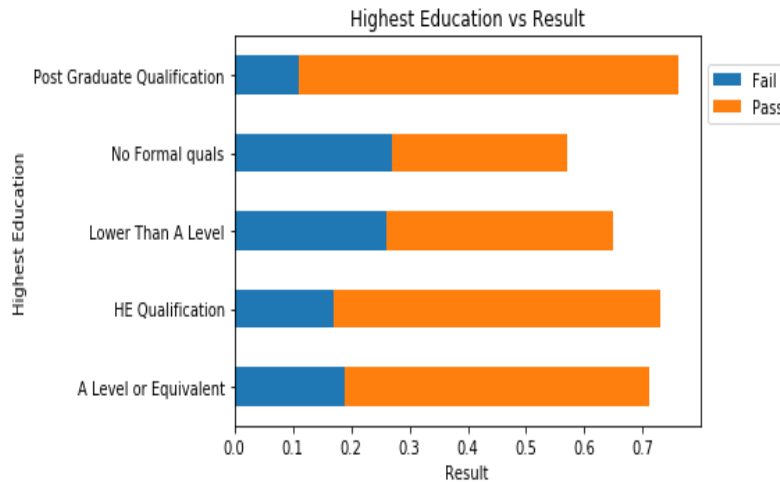


Fig 5 Highest education vs Result

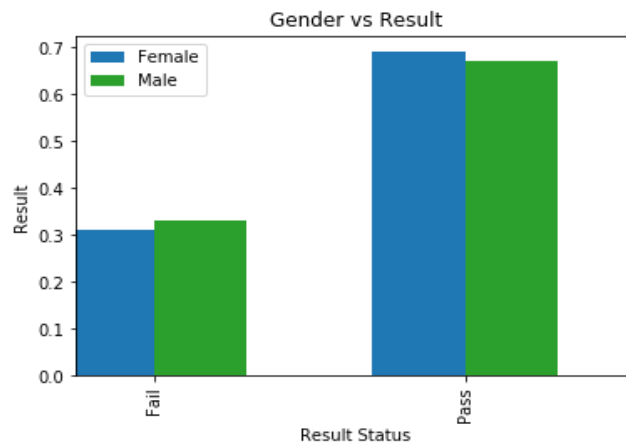


Fig 6 Gender vs Result

As our goal is to predict the final result on the basis of student's demographic behavior as well as on the base of clickstream behavior which is given in the studentVle table. In that table click stream of every student is given in the form of days which we sum on the base of student id and code presentation which is actually represent that when course starts because different students take different courses and if a student withdrawn in a representation take that course again next presentation. So, we sum 'click ids' by using both student id and code presentation to make sure that every student's clickstream should be summed up for every code presentation if students take admission in different code presentation. and then we merged the studentVle table with the student Info table again on the base of student id and code presentation so that the final result of every student id should be placed respectively to its code presentation.

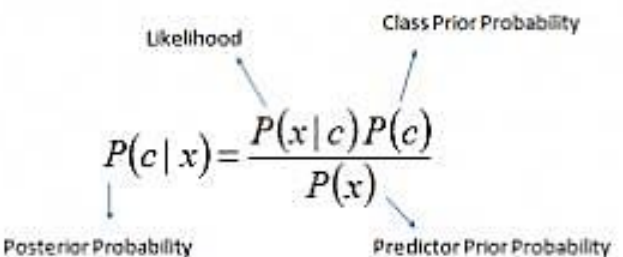
And we are also working on the early prediction of student final result that will they pass or fail by using clickstream behavior. For that instead of sum the whole clicks we divide the clicks in months. And by the total days we have the records of 9 months.

4. Models

4.1 Naïve Bayes Classifier

It is a classification algorithm based on Bayes theorem. It is fast, reliable and have high speed on large dataset. Naïve Bayes assumes that the effect of features in dataset are independent of each other and the presence of one feature is unrelated to other features. This is why it is known as Naïve.

It calculates class (Pass or Fail) for each attribute using Bayes formula given below:



The diagram shows the formula $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ with arrows pointing from labels to the terms: 'Likelihood' points to $P(x|c)$, 'Class Prior Probability' points to $P(c)$, 'Posterior Probability' points to $P(c|x)$, and 'Predictor Prior Probability' points to $P(x)$.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$
$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig 7 Naïve Bayes

- $P(c|x)$ is the posterior probability of class ($c \rightarrow$ Pass or Fail) given predictor ($x \rightarrow$ Age, Highest Education, Sum of Clicks, IMD Band, Disability, Gender etc.).
- $P(c)$ is the previous (Pass or Fail) probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class (Pass or Fail).
- $P(x)$ is the prior probability of predictor (Age, Highest Education, Sum of Clicks, IMD Band, Disability, Gender).

It classifies student as Pass or Fail based on student online activities and demographic behavior. It calculates the probability of Pass or Fail in following steps:

- It calculates the prior probability of class (Pass/Fail).
- Find Likelihood probabilities of each feature (sum of clicks, age etc.) for each class (Pass/Fail).
- Find posterior probability using Bayes formula.
- Assign class Pass or Fail based on which one has the highest probability.

Naïve Bayes classification works for single class prediction as well multi class prediction.

4.2 Logistic Regression

Logistic regression is a supervised classification algorithm and it use probability to predict target value in range of $[0,1]$ i.e. 0 as pass, 1 as fail. Logistic Regression is also known Sigmoid function given as follow:

$$g(z) = \frac{1}{1+e^{-z}}$$

Sigmoid function is S-shaped curved that can take any real-valued number and map between 0 and 1. See Fig. 8

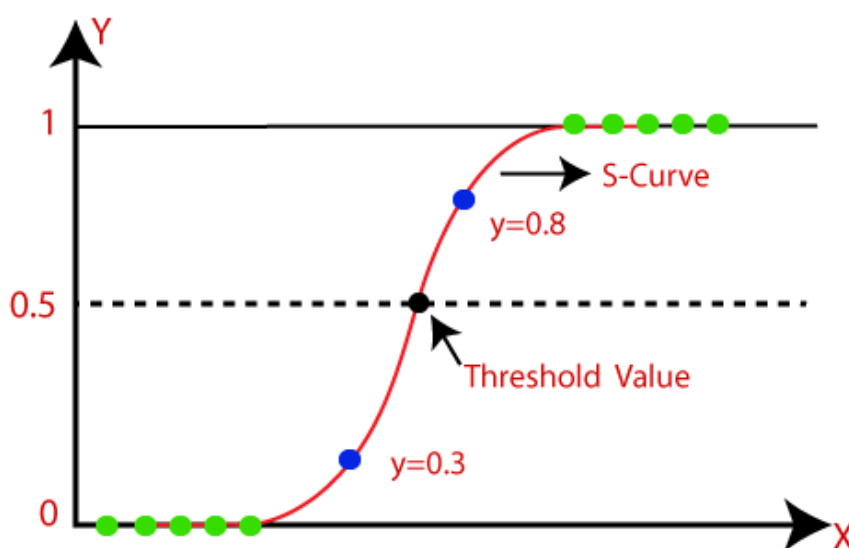


Fig 8 Sigmoid Curve

The value of threshold is important and is totally dependent on classification problem. The threshold value effect the Precision and Recall of our model. We want both Precision and Recall to be 1 in idea case scenario but we have to decide on the tradeoff between Precision and Recall. Low Precision/High Recall where we want to decrease the number of false negatives without decreasing the number of false positive or High Precision/Low Recall where we want to reduce the number of false positives without reducing the number of false negatives.

Logistic Regression is classified as:

- **Binomial:** Target class has two possible outcomes 0 or 1.
- **Multinomial:** Target class have three or more possible outcomes like class 0, class 1, class 2 etc.
- **Ordinal:** Target class have ordinal outcomes like Good, Very Good, Poor, Very Poor etc.

We are using binomial logistic regression as we are predicting student final result as Pass or Fail.

4.3 Random Forest

Random forests are ensemble methods for solving supervised machine learning tasks. Basically, it's a bunch of decision trees to solve the problem. Decision tree is supervised machine learning tasks for classification and regression. A very common example of decision trees is used to predict the temperature. For the prediction of temperature, we need to ask gradually different questions for more accurate prediction of temperature. Gradually, a relevant set of questions helps us to reduce dimensions of dataset and make accurate predictions. [9] [10]

There are basic terminologies used with Decision trees as shown in Fig 9:

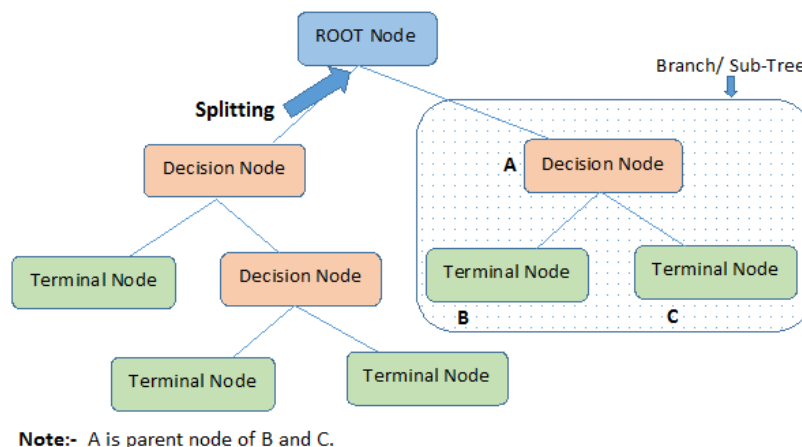


Fig 9 Decision Trees

- **Root Node:** It's the root node represents the most important feature of dataset. It characterizes the entire population or sample.
- **Splitting:** In this process we will divide a node into two (left node or right node) or more sub-nodes.
- **Decision Node:** Decision node is when a sub-node is splits into more sub-nodes.
- **Leaf/ Terminal Node:** No children in a Leaf Node is a last node.
- **Pruning:** Decreasing the size of nodes is called Pruning.
- **Branch / Sub-Tree:** A subsection of decision tree is called branch or sub-tree.
- **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

Algorithms used in decision trees are:

- ID3
- Gini Index

- *Chi-Square*
- *Variance Reduction*

In random forest decision trees uses Gini Index. By using the Gini index, we select the root node and divide the dataset into left tree and the right tree. We split the data into best split of data using random forest. For simplicity, we're just going to try every possible split and use the best one. One with the highest Gini Gain value is set root node.

First of all, we calculate the Gini Impurity of dataset for the selection of root node using formula:

$$G_{initial} = \sum_{i=1}^3 p(i) * (1 - p(i))$$

In our proposed research, we have code presentation, sum click, id student, code module, highest education, imd band, age band, studied credits and disability features. Random forest select sum click as root node making it as the most important feature among them.

Then, we calculate the Gini Impurities of the two branches G left and G right.

Gini Gain:

Gain=G_initial - G_Left - G_Right

Through this process, we build decision tree. Random forest is bunch of Decision trees bundled together like ensemble methods.

Decision tree individually used for classification then why we use random forest. Because random Forest when we can solve the same problems using Decision trees.

- Decision trees are easily implemented but they lack in accuracy. Decision trees works very effectively with the training data that was used to build them but they give poor results when try to prediction with test dataset.
- Decision tree shows Overfitting property.

Bagging:

Random forest uses bagging technique for classification or regression. [10] It is an algorithm that train a bunch of decision trees of a given dataset with n points:

- Its samples with replacement given n training instances from the dataset.
- It trains a decision tree on the n samples.
- For some t, repeat t times.

For prediction using this model with the trees, we aggregate the predictions from all the decision trees and adopts the majority votes generated by decision trees for class labels (Pass or Fail).

5. Experiments and Evaluation

In purposed system, we use three classification methods Random Forest, Logistic Regression and Naive Bayes for classification of students. We trained the dataset on these models with different parameters and compare accuracy, among all Random forest gives highest accuracy.

5.1 Random Forest

As discussed above Random Forest uses ensemble technique of bagging for classification. In Bagging, random forest use decision trees and on the base of each vote by the decision tree, classify the student as Pass or Fail. We provide two parameters max_depth and estimators to the Random Forest algorithm. We feed these parameters with different values for example n_estimators=[1,2,3,4,5,6,7,8,9] and max_depth=[2,4,6,8,10,12,14,16] and we use(cv=5) 5 fold cross validation data for training and testing data. Before feeding this data to the model, we standardize the data using 'Min Max Scaler'. Min Max Scaler standardizes the data between min and max value. Using random forest, best parameters which are selected by random forest is n_estimator=9 and max_depth=16 and we get 86% accuracy on validation data and 87% on testing data. Accuracy scores are given in table 1.

Table 1

Accuracy	87.8%
Precision	83.20%
Recall	87.80%

Confusion matrix for Random Forest is given in Fig. 10 below.

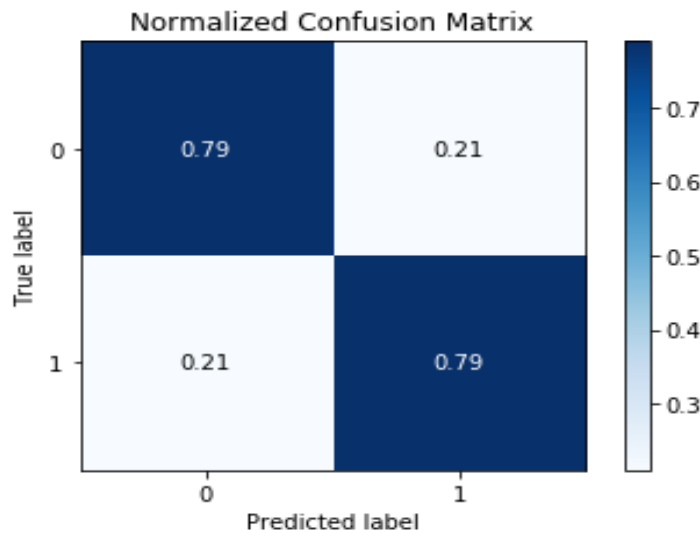


Fig 10 Confusion Matrix (Random Forest)

5.2 Logistic Regression

Before data splitting, we have done resampling. Resampling technique is user when data is highly imbalanced. There are two techniques in resampling “oversampling” and “under sampling”. In oversampling technique adding more samples in minority class and under sampling technique consist of removing samples from majority class. In our case we have adopt oversampling technique because our sample “Pass” class have 15382 entries and in case of “Fail” it has 6680 entries. So, we have done oversampling and our pass and fail both have 15382 entries. After that we standardize the training data using ‘Min Max Scaler’ technique. And we have set the parameters of logistic regression $c=100$ and $tol=.01$ (tol is tolerance) and we get 73% accuracy. Scores table is shown in table 2, along with confusion matrix in fig. 11

Table 2

Accuracy	78.3%
Precision	79.48%
Recall	70.14%

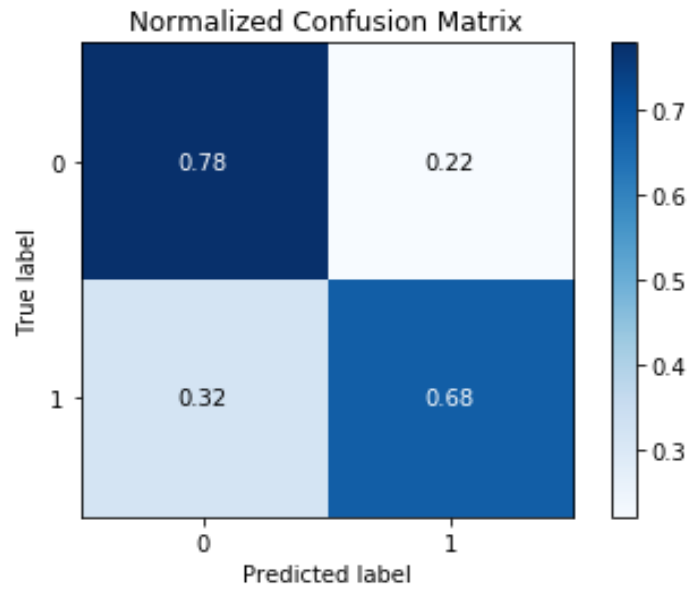


Fig 11 Confusion Matrix for Logistic Regression (with sampling)

and without resampling using the original data when we applied logistic regression, we get the 78.4% accuracy. (See table 3 and fig 12.)

Table 3

Accuracy	78.3%
Precision	79.48%
Recall	70.14%

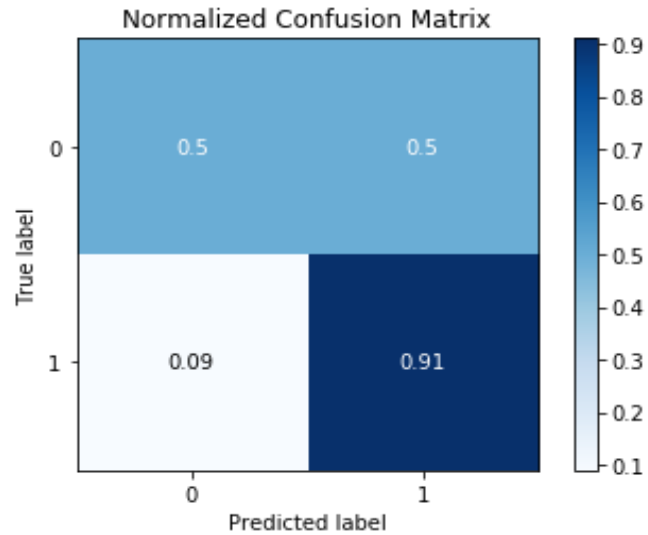


Fig 12 Confusion Matrix for Logistic Regression (without sampling)

5.3 Naïve Bayes

Working of naïve bayes classifier is discussed above, by applying naïve bayes model with oversampling we get the 74.2% accuracy. as shown in table 4. Confusion matrix for Naïve Bayes is shown in fig. 13.

Table 4

Accuracy	73.0%
Precision	77.99%
Recall	67.23%

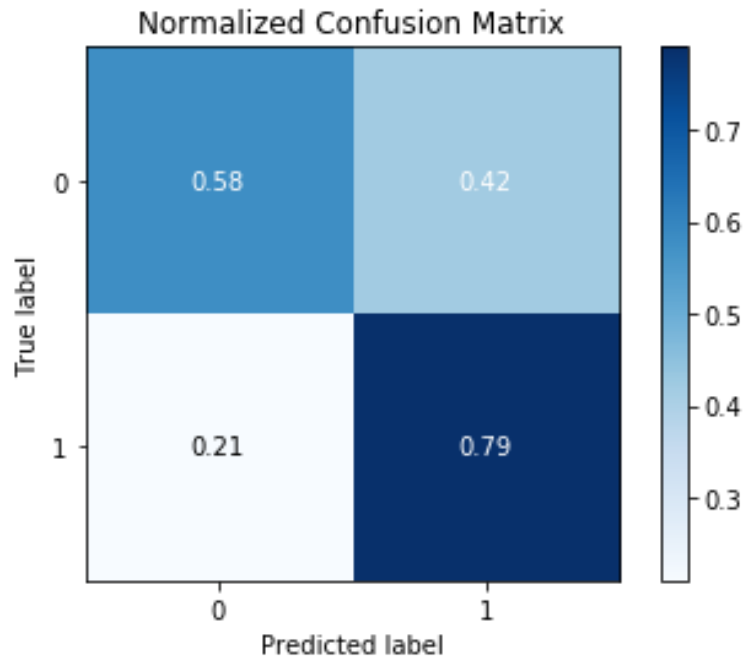


Fig 13 Confusion Matrix Naïve Bayes (with oversampling)

6. Conclusion

Our research paper presents the usefulness of Machine Learning models in predicting success and failure of students from the Open University Learning Analytics [1] dataset, by monitoring their behavior i.e. clickstream in an online Virtual Learning Environment. We achieve accuracy of about 87% in predicting success and failure of students. Our current method only utilizes student's engagement pattern in VLE to determine their performance. Moreover, other significant aspects i.e. their scores and assessment submission can also be combined to gain more accurate results.

Prediction of student success or failure will help organizations initiate the development of student support cell and steering boards in the online education community, in the system of providing students with the best results by providing appropriate advice and through a one-to-one platform discussion with student consultation.

7. Future Work

The forthcoming perspective in this research area is early prediction of student success and failure by analyzing weekly behavior of student in online learning environment. Regarding the early prediction of student result, we have started work and working on it. We have preprocessed the data and working on applying the time series models. We have processed the data into 20 different online activity types according to week ranging 0-38 for every student. We will predict the student final result in week 25 and inform student about his/her progress. We will also apply some deep learning models i.e. Recurrent Neural Networks, LSTM models for better accurate results. Further, scope of this research will be extended in better decision making in online VLE's, and recommendations will be given to students based on their interest. All relevant courses will be offered to students based on their interest.

Furthermore, scope of this research can be extended to online learning websites i.e. coursea, Udemy, Udacity etc.

References

- [1] S. U. Aishwarya Ramachandrappa, "Open University Learning Analysis".
- [2] E. V. R. Sumitha, "Prediction of Students Outcome Using Data Mining Techniques," *International Journal of Scientific Engineering and Applied Science (IJSEAS)* – Volume-2, 2016.
- [3] Z. T. H. V. d. M. Nurbiha A. Shukor, "An Examination of Online Learning Effectiveness Using Data Mining," *Procedia - Social and Behavioral Sciences*, 2015.
- [4] M. S. ., S. J. Parneet Kaura, "Classification and Prediction based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Science Direct Procedia Computer Science*, 2015.
- [5] M. H. Z. Z. Jakub Kuzilek, "Open University Learning Analytics dataset," *Scientific Data*, 2017.
- [6] N. S. Vivek Doijode, "Predicting student success based on interactions with Virtual Learning Environment," *Analytics Experience*, 2016.
- [7] H. W. R. A. A. Saeed-Ul Hassan, "Virtual Learning Environment to Predict Withdrawal by Leveraging Deep Learning," *International Journal of Intelligent Systems*, 2019.
- [8] K. Z. Jui-long Hung, "Revealing Online Learning Behaviors and Activity Patterns and Making Predictions with Data Mining Techniques in Online Teaching," *MERLOT Journal of Online Learning and Teaching*, 2008.
- [9] "Decision Trees," 26 October 2018. [Online]. Available: <https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb>.
- [10] V. Zhou, "Random Forests for Complete Beginners," 12 April 2019. [Online]. Available: <https://towardsdatascience.com/random-forests-for-complete-beginners-2014d9ed91c0>.
- [11] F. Y. T. S. A. K. Okubo, "Students' performance prediction using data of multiple courses by recurrent neural network," in *In 25th International Conference on Computers in Education, ICCE. Asia-Pacific Society for Computers in Education*, 2017.
- [12] [Online].

Appendix

Table 5 Activity and Description

Act-Code	activity_type	description
A-1	resource	Typically comprises of PDF resources i.e. lecture notes, books etc.
A-2	oucontent	signifies contents of assignments, which learners' have to pass during presentation
A-3	url	Comprises of links to internal or external resources or for sample video/audio contents
A-4	glossary	comprise of basic glossary related to content of course
A-5	ouelluminate	Consists of online tutorial/live session
A-6	Shared_subpage	Consists of materials shared among several courses and/or faculty
A-7	dualpane	the site is divided into two parts; one containing information and second for activity related to the information
A-8	ou_wiki	Open University wikipedia content
A-9	homepage	course homepage
A-10	forum_ng	discussion forum
A-11	questionnaire	questionnaires related to course
A-12	repeat_activity	usually indicates to content from preceding weeks of course
A-13	subpage	indicates to other sites in the course together with basic instructions
A-14	ou_collaborate	online video discussion rooms (students -tutors)
A-15	page	contains informations and directions related to course
A-16	folder	contations files related to course
A-17	data_plus	additional information audios/videos/pdf
A-18	external_quiz	points to external quiz relevant to course (depending on course content is usually similar to quiz)
A-19	htmlactivity	interactive html page with educational learning content
A-20	quiz	course quiz