A Hybrid Approach for Dropout Prediction of MOOC Students using Machine Learning

Fawaz J. Alsolami
falsolami1@kau.edu.sa
Computer Science Department Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia.

Summary
Massive Open Online Courses (MOOC) is an extensive way of providing online education to the students all over the world. Based on the statistics, this education system have millions of students attending hundreds of courses in different offered programs. Since, MOOC started, it has been facing a challenging concerns, which is also a major difference between the traditional teaching and MOOC, known as "student dropout ratio". With this fact, the overall performance of MOOC is negatively impacted the real purpose of distance learning. Whereas the difference between course registration and course completion ratio in MOOC is quite large. On the better side, the emerging technologies has created several opportunities for the students to get education online, but due to multiple factors the dropout ratio of online students is high as compare to traditional school learning process. This research is focusing on the issue to understand and predict the MOOC dropout ratio. The multiple models and evaluation metrics generating variety of results as extracted from literature review. To tackle this problem, the experiment conducted in this study using KDD MOOC dataset by implementing hybrid approach of machine learning algorithms. The results suggested the appropriate improvements in the dropout accuracy ratio. Based on the final results, the maximum accuracy recorded as 90% that measured through random forest model. Finally, the model can help and assist the online education system to understand the early dropout prediction and to do necessary arrangements.

Key words:
MOOC dropout prediction; MOOC data; KDD dataset; machine learning algorithm.

1. Introduction

MOOC is one of the solutions of providing online education around the world. This type of education widely applied from different universities in the form of offering distance learning courses in multiple discipline [1], [2]. The number of MOOC students has increased rapidly as reached to 110 million excluding China, while offering around 2500 courses and 11 online programs [3]. This facts and figures, highlights the importance and feasibility of online courses and gives a chance for the students to meet with the professionals. Furthermore, it provides the facility at your place, students do not require to visit any place, get the visa, and ticket, where the education reached to your home. Figure 1 is showing the growth of MOOC student in numbers in specific years shown in the graph.

Manuscript received May 5, 2020
Manuscript revised May 20, 2020

In this era, the latest education trends has shown remarkable impact on the society year by year. With its popularity, it became a good choice for the students, which is not only cheap, but affordable in many ways. Reaching to the ideal and most advanced courses by single click was never been that easy in real life [4]. Another research emphasized that, it is not only reasonable and viable for the students, although this system offers the space for the professional, educators, and practitioners to improve their skills and to be well-known in educational world [5]. The main idea of MOOC is to develop constructive relationship between the universities, students, and teachers by connecting and uniting them for single cause and that is education.

Apart from the big success of online education, the system has been criticized by different scholars for some valid reasons encountered in this system. First of all, the large number of registration in each course is one of the challenges in online courses. It is complex, to deal with large group of student at the same time [6]. Another issue highlighted in the previous work is even after the big number of enrolment, the course completion rate is quite low [7]–[9]. In addition, from the instructor's vision, handling number of assessments, and capturing the participant's attention are some other issues raised in this educational scenario [10]. On top of that, in recent researches the issue underlined by the educators and instructors is to predict dropout ratio in online courses. Indeed, the high percentage of dropping the course put the institutes in a complex situation.

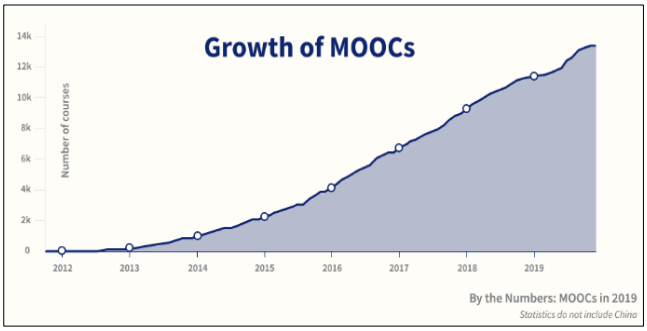

Fig. 1 By the Numbers: MOOCs in 2019 [3]

The factors behind the dropout ratio has been discussed and assessed several times in previous work. A research tried to extract multiple features and behavior of student to understand the reasons behind high dropout ratio and call it as "climbing over the cliff" [11]. For this, the dataset provided by XuetangX (one of the largest online platforms) [9] that has been used in KDD Cup [12] as well for analysis and predicting the behavior of the students. Further details and description of dataset are discussed in the next section "Related Work".

The major contribution of this research are as follows. Firstly, the paper aims to highlights the real challenges of online education especially from the context of MOOC. However, the main purpose of this research is to understand the behavior of the student, build the model, and predict the dropout ratio to support online education system. Therefore, the hybrid machine learning approach used to predict and compare the accuracy for solving dropout problem. For this, the research employed MOOC student's dataset organized and presented in KDD Cup [12] for experiment.

To sum up this section, this paper is presented as follows. The next section highlights the most related research work presented by different scholars. Data description and preprocessing discussed in Section-III. Afterwards, the experiment details and framework design presented in Section-IV that shows the description of selected machine learning model. Next, the implementation of the model and generated results with discussion explained in Section-V and Section-VI respectively. Finally, the last section concludes the paper with purpose and findings of this research and pointing out the factors that can be applied in future research

2. Related Work

Educational data mining extensively applied and used for predicting the student's performance and behavior. Commonly, this approach used to discussed several problems such as to predict graduate performance [13], learning analytics [14], and data mining course for undergraduate students [15]. The purpose is to classify the data, build and train the model, generate rules and patterns, and then use it for future purpose. Recently, the problem related to student's dropout ratio in online courses pointed out and criticized several times.

The problem of student's dropout ratio got much attention from the scholars to assist and help out the online educational institutes, instructors, and other professionals. To understand the student's behavior a research build the model and suggested some solutions [11]. In addition, illustrates the problem by predicting the dropout ratio using weekly performance [16]. Some researches, worked on the log files generated by the institutes to track the student's involvement and participation in the course [17]. From the previous researches, it has been evident that the actual time of the dropout is critical factor to understand. Therefore, a model is required to understand from the experienced data, and how accurately it can predict the student's dropout ratio [18].

Certainly, multiple machine learning algorithms applied to solve this issue such as logistic regression (LR) [19], Naïve Bayes (NB) [20], Decision Tree (DT) [21], Random Forest (RF) [19], Support Vector Machine (SVM) [21], and Gradient Boosting (GB) [22]. Most of the time, researchers applied multiple models to measure the results and accuracy by comparing performance of different models [4], [9], [23]. The best fit model can be used for better prediction of student's dropout, which can help the organization to forecast the actual number of students who will appear in the final exam. Another research conducted which applied the model and measured the accuracy on weekly data [24]. But, the overall data including the log files, participation in assessments, learning methodologies can be used for more understanding.

To understand more about dropout prediction, scholars have presented multiple scenario based on their understanding from the data. For example, whether the student will participate until end of the course or the week in which students are active is the final week [24]–[27]. On the other side, the next methodology is based on the different idea, which is to understand the chances of student's connectivity in future weeks [28], [29]. It revealed that there is no clear definition for measuring the accuracy of the model to early predict the dropout ratio. Another research suggested that, to understand the chances of student's dropout can be measured by looking at the performance and student's attention in current time. Therefore, this problem can be considered as time-series problem, rather to predict on static data [17].

Accordingly, there are several factors highlighted in different researches to understand the reasons behind student's dropout in MOOC. The factors collected by merging the idea and concept from various researches [5]. The author categorized the reasons of dropout in two perspective as mentioned in table 1. The factors mentioned in the below table are considered to be some reasons that force the students to drop the course. In the end, this research output may elaborate some other features that can be added to this list. Whereas, the online educational institutes may assume these factors before introducing new online program and courses. Keep in mind, the factors are not standard, but extracted based on the previous work and suggestions from the scholars. Motivation for the students and time management are some of the issues related to student's concerns.

Table 1: Forcing Factors Behind Student's Dropout [5]

Student Focused Factors	Course Focused Factors
No Motivation	Curriculum Design Issues
Time Management Issues	Lack of Interaction
Lack of Knowledge	Unidentified Fee Collection
Inadequate Background	Isolation

Since, this research idea has been undertaken by different scholars. KDD data set used and applied using machine learning algorithm in the previous work. The main contribution of this paper is to understand and improve the prediction accuracy ratio measured earlier. Student dropout prediction in MOOC using machine learning algorithm presented by [30], illustrated and compared the results using accuracy measured across the sections and then overall average of all courses. The optimal accuracy evaluated through CNN recorded as 86%, whereas the least accuracy measured through DT as 75%. Moreover, another idea presented by [10] where the prediction was measured using weekly data, and then final accuracy assesses for complete understanding.

To summarize this section, here we presented the comparative overview with the help of literature review. The idea was to understand the use and importance of these algorithm for measuring the dropout accuracy. As found in literature review, the research idea has been implemented earlier, where different results generated using KDD data. Therefore, in this research we have extracted the most common methods used for MOOC dropout prediction those are; LR, NB, DT, DT, RF, and SVM. The details of each algorithm is presented in the experiment design section. For now, there is a comparison which extracted from literature review and presented in table 2. It illustrates the type of the models applied on the same KDD data to measure the prediction ratio of dropout student using different scenario. Whereas, the purpose of this study is how to improve the performance of the prediction ratio by applying the implementation of the model simultaneously. In addition, as can be seen from the table, not all algorithms has implemented in one approach. Some of the experiments conducted earlier used alternatives algorithm, which highlights the variety of approaches can be used for MOOC dropout prediction. Therefore, in this research we will use the hybrid approach using all six models in single approach. For this, the experiment design section describes the step by step methodology undertaken in this study. In the end, the results generated in this research compared with the previous work as mentioned in this table. It will provide the comparative analysis on the performance in this study with previous work.

Table 2: The Accuracy of Classifiers in Previous Work – Using KDD Data Set

Algorithm	[31]	[10]	[30]
Decision Tree	X	85%	75%
Logistic Regression	86.78%	84%	81%
Support Vector Machine	88.56%	86%	81%
Random Forest	X	X	82%
Gradient Boosting	89.12%	X	82%
Naïve Bayes	X	X	79%

3. Data Description and Preprocessing

The dataset used in this research is issued by well-known organization and publicly available called KDD Cup-2015 [12]. The original form of the data was not fully integrated with the tool used in this research, is Rapid Miner [32]. Therefore, the necessary steps has been taken to preprocess the data before framework implementation.

Overall, the data file has to log information of thirty days for all the students enrolled in any online course. Altogether, the data belongs to 39 courses and more than 100 thousands of users connected with the system. The log file showing the participation of the users in different ways such viewing course videos, working on assessments, navigating on different course objects or closing the website. According to the original file, there are almost 80 million log information is recorded in the file. In this log information, almost 80 thousands of student who enrolled in one of the course offered by the institution. As the data was huge, for this experiment 10,000 samples of the data selected and implemented through six different models using rapid miner. Furthermore, the data was distributed in different files as described below. The first file belongs to the "enrollment" records of each student. It was divided into two different files; "training" and "testing". The second major file is associated with the history of "log" information. It was the biggest file in this dataset, which kept all the information about student and their communication with the educator's website and server. This file provides the real information and reasons behind the dropping of the course. Based on this information the model can be validated, whether the student will drop or continue the course. Moreover, the next important data file is "truth_train" to know about the particular student drop the course or not. This file is used for training purposes connected with log history. During the

training phase the two files "log" and "truth_train" combined to build the model. The next file in this scenario provided by the KDD is "object", which consist of the information related to courses and modules. Finally, the last file is "date" that describes the beginning and finishing time of the course. Altogether, there is nothing to show about other files except the "log" file, which consist of some interesting facts and figures. These facts are very useful and can help in training of the model and further lead to the decision. Therefore, some statistical and descriptive information about the log data is showing in the following table 3.

Table 3: Some Important List of Attributes of Log File

Log Attribute	Minimum	Average	Deviation
browser_problem	0	16.40	45.66
browser_access	0	11.39	35.91
browser_video	0	8.06	15.10
class_size	385	3423	2245
server_problem	0	2.35	7.44
server_access	0	32.37	51.55
navigate	0	12.79	18.87

The above table illustrated some important attributes mentioned in the log file, which are associated with number of students. The minimum value is denoting that sometime no student faced this problem. Whereas the "class_size" minimum values is expressing the minimum class size for all courses. The table illustrating the information related to each student as compiled with full history. The minimum value is highlighting the minimum number of entries associated with the particular attributes. Then, the average is corresponding to the most likely number of students associated with that attributes. In the last, deviation is telling the possible amount of variation in the single entity. All logs attributes has been selected in this experiment as an independent variables, which are associated with class variable. The class variable is denoting two different types of values that is "Drop" and "Non-Drop". The idea is to train the model using experienced data, and based on the values assigned for log variables. Finally, the model will be tested to understand about the prediction ratio for "Drop" and "Non-Drop" students.

To develop the final version of the data file, possible preprocessing steps has been applied on the files. For example, there were some attributes, which were mentioned using sequence of alphanumeric numbers such as username and course_id. Those attributes were transformed into numerical values to give more understanding and useful for framework implementation, through "transformation" operator available in the rapid miner. Moreover, the log file was managed to be in particular requirements based on machine learning algorithms. Some of the attributes were removed, which were not required for this experiment such as "class_size" and "access". Whereas "bowser_access" were kept in replacement of "access" variable. After taking some other steps, the final version of the single data file created for model implementation. In the modified version of the file, two new attributes were added as "course_No", which indicates number of courses enrolled by any student, and "non_drop", to enter the information about in how many courses students still participating.

4. Experimental Design

MOOC dropout prediction has applied by different scholars using multiple methods. The purpose of prediction is to assist those organizations that offering online education. The rate of prediction and accuracy can provide them to think in an efficient way and reasons behind the dropouts. Therefore, this experiment designed to make best use of data and apply using multiple algorithms to highlight the optimal results and accuracy achievement. The following step-wise approach is showing the overall structure of the experiment conducted in this study.

Stepwise Approach: Predicting Student's Dropout Ratio in MOOC

Begin
{
Phase-I: Data Selection
{
Data Understanding
Data Cleaning
Missing Data Imputation
Selection of Attributes and Data Values
}

Phase-II: Variable Identification
{
Selection of Independent Variables
Selection of Class Variable (Label)
}

Phase-III: Algorithms Implementation using Rapid Miner
{
Data Transformation
Connecting all Model (NB, DT, RF, SVM, GB, LR) with Data
}

Phase-IV: Result and Analysis and Discussion
{
Measuring Accuracy using "Performance" Operator in Rapid Miner
Generating Confusion Matrix for Accuracy
}

Phase-V: Results Comparison and Final Discussion
{
Analysis and Comparison with other models
Analysis on Accuracy Increment or Decrement
}
End
}

To implement the above framework, the following algorithms has been chosen based on the findings of the previous work. The explanation and use of each model is presented below.

4.1 Logistic Regression (LR)

LR is the kind of statistical and classification model, which work on the bases of binary values such as "0" and "1" [33]. The model is developed for predicting and approximating the result values based on binary numbers. The model can be trained using categorical values such as drop (1) and non-drop (0) student to predict for the potential students [12]. Therefore, the model is commonly known and depending on binary values. The results of this model can be measured using probability which occur between "0" and "1". Although, there are multiple extensions presented in previous researches by modifying the LR model [26]. Following is the basic equation (see Eq. 1) use for calculating the probability of given data [34].

$$\text{Logit(p)} = ln\left(\frac{p}{1-p}\right) = \frac{\text{probability of presence of characteristics}}{\text{probability of absence of characteristics}} \quad (1)$$

where *P* is refer to calculate the probability.

4.2 Random Forest (RF)

This is the second model selected in this study to measure the performance of the MOOC dropout prediction. This algorithm falls under the category of classification, data mining and machine learning [35]. The model is famous for ensemble learning approach, which represents the searching of the optimal results by generating multiple trees in the single run [19]. It is a supervised learning model of classification that creates forest of multiple tree. The tree kind methodology is useful for associating multiple attributes of the data called nodes till reach to the result. RF is work by creating sets of trees and then repeating the steps for integrating and train the trees to get the optimal one [36].

4.3 SVM

It is a common approach used for regression and classification to analyze the data and train the model. This method is known as supervised classification method, which means the label and class attributed is already defined in the data set. Using the training dataset, the steps of this approach is to assign number of values to one or multiple class attributes. Therefore, this technique uses non-linear classification approach for predicting the category of the dataset. Finally, the values will be joined by separating and measuring the gap between different groups [24]. Earlier, the techniques is already applied on MOOC dataset, which further analyze the result by comparing it with RF [37].

4.4 Gradient Boosting (GB)

As this research is mainly focused on predicting the dropout ratio of MOOC student's dataset. From the literature it has been evident that the dataset is best use for apply, train, and test the model using classification method [10], [30], [38]. Therefore, the next model selected in this research is GB, which is another kind of learning method use for classification. This method is start the process by generating weak model after analyzing the dataset. Initially, in order to generate the weak models, it helps to approach to the optimal results by reducing error and improving the accuracy [39]. Therefore, it is known for boosting method, as unlike decision tree, combining the weak models this method give boost in generating the strong model [40].

4.5 Naïve Bayes (NB)

The technique is from the family of generating model using probabilistic classifier. NB is one of the common approaches using for prediction based on the selected features. It is also a supervised learning approach and used for educational data mining several times [20]. The approach of this model is to calculate the probability of occurrence of predictors using different other independent attributes. Basically, in the model building the target classes assigned earlier, whereas the remaining attributes consider as an independent [41]. The common equation using for NB is as follows, see Eq. 2 [42].

$$P\left(\frac{A}{B}\right)\frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \quad (2)$$

4.6 Decision Tree (DT)

Finally, the DT is the last classification model chosen in this study. It's another type of algorithm use for forecasting after training of the model. It is a common approach of statistics, data mining, and machine learning as well [43]. The approach of this model is develop a tree based mechanism, therefore, it known as decision tree. The tree, where multiple nodes and branches connected to each other, till the last node,

and the model can provide the decision label for the particular data values [22]. Each node in the decision tree is denoting the class identification of the selected data. Whereas, the branches used for building association between the attributes [21]. Finally, the decision can be taken based on validating the dataset on each node until it reach to the decision [44].

5. Framework Implementation

After appropriate steps taken earlier, the model is implemented using six chosen machine learning algorithms. The approach in this framework is to use the previous data file generated by KDD and to develop the model for prediction. The data file preprocessed using necessary action. Altogether, there was 17 attributes selected as independent and one attributes was chosen as label column. The label column has two different types of values "0" and "1" associating with "non-drop" and "drop" respectively. The overall, experiment conducted using Rapid Miner, as this is tool feasible for and used several times in previous research [45], [46].

The implementations started by importing the dataset using the operator provided by rapid miner called "Read_CSV". Then, as there were altogether, six model selected in this study, therefore, another operator applied to make multiple copies of dataset. For this, "Multiply" operator executed in the process, which further connected with the models. For building the model with enhanced capability, the validation part was performed using cross selection of the dataset. There are two types of validations normally using by the researchers known as; split validation and cross validation. The split validation is supposed to divide the data into two sets by defining the percentage of each. Whereas, the standard value of division is use as 70% and 30% for training and testing purposes respectively [47]. This research used the cross validation using 10-fold method. This method will allow the process to divide data into 10 different subsets of random data. The main purpose of using this validation scheme is to use best use of data. As it allows and consider whole data to be a part of training and testing phase. Therefore, the main data file is directly connected to six separate cross validation operator, which named as per the selected mode. The first part of implementation is shown in figure 2.

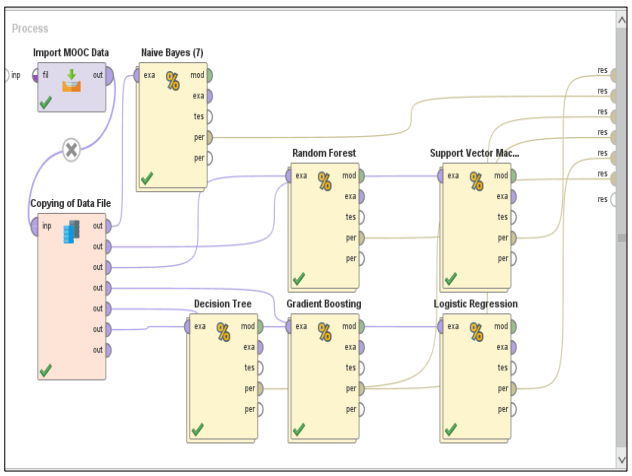

Fig. 2 The First Phase of Implementation

The second phase of implementation is illustrating the process designed under each validation operator as shown in figure 3. In this phase there are three operators used; (i) the machine learning model, (ii) Apply Model, and (iii) Performance. The same strategy applied under each cross validation operation (as shown in figure 2). For each cross validation the particular machine learning algorithm is replaced. This model operator use for train the model using training dataset. The training phase is used to train the model according to the particular algorithm. At the next stage, the developed model needs to be applied. Therefore, the operator "Apply Model" is connected, which receiving the trained model and testing data for implementation as shown in figure 3. This operator is useful for model implementation and generating number of predictions. Finally, the last operator in this figure is "Performance" operator. The operator is applied here to measure the dropout prediction accuracy based on the result generated by testing model. The result and accuracy discussed in the next section.

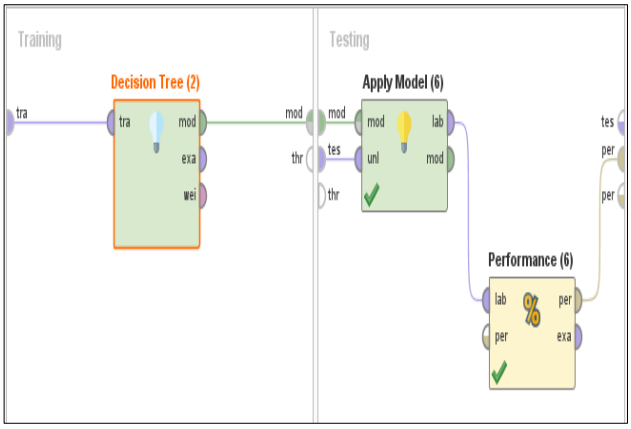

Fig. 3 The Second Phase of Implementation

6. Results and Discussion

The model has been executed and generated prediction accuracy successfully. The performance of the each model is calculated and presented using confusion matrix as shown in table 4. The confusion matrix is a common strategy use for classification model to understand the prediction ratio for each class. The confusion matrix is showing the percentage of model based on the label class in the dataset. As defined earlier the dataset contains two types of classes; "0" and "1", representing the "non-drop" and "drop". Therefore, the result table is illustrating the prediction ratio for each class. Furthermore, the three elements of each confusion matrix such as precision, recall, and accuracy used as evaluation criteria shown in the table. The discussion of the results is defined according to these three criteria. The details of each criteria corresponding to the selected model is presented in subsequent section.

6.1 Precision

In the actual definition of precision, is to understand the accuracy in any condition. For the classification model, the precision is highlighting the ratio of correct prediction divided by total prediction [48]. The percentage is showing the correct prediction for each class. For example, in the result table, it is showing that the through Naïve Bayes, the minimum precision calculated for class "0" that is 61%. It denotes that for "non-drop" students the accuracy of the model is quite low as compare to other model. In addition, in almost every model the performance for the same class is low. Decision tree and support vector machine measured the precision value for non-drop student is 78%. Overall, for class "0" the precision values assessed from 61% to 91%.
As the research conducted to understand the dropout ratio using selected students dataset, it can be clearly witnessed that for class "1" the precision values are quite impressive between most of the models except Naïve Bayes. Among all, the random forest performance is the best for predicting correctly and calculated as 92%. While, four of the models predicted for dropout student as 91%. For the same class the lowest precision is recorded for Naïve Bayes as 84%.

6.2 Recall

Recall is the second validation criteria used in this research to define the performance of the classification models. The recall value calculates the randomly selected class values divided by the total number of item's existence in the dataset [48]. In this situation, as evident from the result table the performance of Naïve Bayes is reasonably low from other's algorithms, which is 48% for class "0". It is the minimum value assessed between all. The maximum recall value for any class is measured as 92% by logistic regression and random forest, where both values are related to dropout student's prediction. Till now, for both of the classes, it is visible that the performance of Naïve Bayes is lowest in all results, while random forest recall values for both class is recorded as highest. Altogether, the decision tree prediction for dropout and non-dropout students selected randomly divided by the actual values of dropout and non-dropout student is measured as 81% for class "0", and 92% for class "1".

6.3 Accuracy

Finally, the last criteria used in this research is known as accuracy. The accuracy used for understanding the overall performance of the classifiers. It can be calculated using the complete values of true predictions, divided by the total number of values for all classes [48]. According to the result discussed in this section and performance of the other criteria, it is strikingly identified that the overall accuracy of random forest is the highest as 90%. It can be considered the best method in this research, which can help the organization to support their online courses by predicting the right number of dropout students. On the other side, based on the results in this research, the performance of the Naïve Bayes is the lowest (80%). There is another good performance recorded in this experiment where gradient boosting (89%) is very near to the random forest (90%), which can be considered as the second best option for MOOC dropout prediction, according to the result generated in this study.

Table 4: The Measured Performances for All Models

Classifier	Class	Precision	Recall	Overall Accuracy
Naïve Bayes	0	61%	48%	80%
	1	84%	88%	
Logistic Regression	0	81%	77%	88%
	1	91%	92%	
Random Forest	0	83%	81%	90%
	1	92%	92%	
Decision Tree	0	78%	79%	88%
	1	91%	90%	
Support Vector Machine	0	78%	79%	88%
	1	91%	91%	
Gradient Boosting	0	86%	77%	89%
	1	91%	88%	

Finally, the last part of this section is to compare the accuracy measured in this research with previous work. table

5 is representing the overall accuracy measured in this research in comparison with related work. Altogether, the experiment conducted in this study using six models, whereas in related work some of the model matched but some are not. Overall, most of the model's accuracy assessed better than related work. The main purpose of this research is to conduct the experiment using MOOC data available online to predict the dropout ratio of the students. As discussed earlier that this problem is quite challenging for the institutes offering online education, where the completion rate of online students is lower than the registration.

To look over the comparison report, for clear understanding and similarity in the experiment, the related work selected in this study used the same dataset of MOOC as applied in this research. Therefore, the comparative report can make a good sense of clarification regarding the accuracy generated in this study. It has been evident that decision tree prediction score in this study is 88%, which is better than related work. Although, in related work-1, the algorithm was not considered, whereas in the related work-2 and 3, it is showing the accuracy as 85% and 75% respectively. Moreover, the second algorithm used is logistic regression, the measured accuracy recorded as 88%, again it enhanced the result as compare to other studies. The reasons, behind the improvement may be the 10-fold cross validation method suggest in this study, to train the model using 10 subsets of data.

Table 5: The Accuracy of Classifiers' Comparison in Previous Work – Using KDD Data Set

Algorithm	This Study	Related Work-1 [31]	Related Work-2 [10]	Related Work-3 [30]
Decision Tree (DT)	88%	X	85%	75%
Logistic Regression (LR)	88%	86.78%	84%	81%
Support Vector Machine (SVM)	88%	88.56%	86%	81%
Random Forest (RF	90%	X	X	82%
Gradient Boosting (GB)	89%	89.12%	X	82%
Naïve Bayes (NB)	80%	X	X	79%

In the same way, support vector machine's performance is very similar to related work-1, but much better than related work 2 and 3. The only algorithm performed low is gradient boosting as compare to related work-1 the performance is low as 0.12%. In the last, according to the experiment conducted in this study, the optimal performance generated by random forest that is 90%. It is better than the performance measured in related work-3, the other work did not considered and applied this algorithm. On this statement, based on the current experiment's results, the random forest can be considered the best model for MOOC student's dropout prediction, and can be act as an optimal model for the online and distance learning education system to use for student's dropouts prediction.

7. Conclusion and Future Work

High dropout ratio in any educational institutes can create a high risk for the learners and teachers as well. The issue has been undertaken in this study, to evaluate this issue and try to provide better solution in predicting the dropout ratio. It can help the organizations to estimate the use of resources, providing online material, and specifically offering the number of seats to the eligible students. Therefore, the hybrid machine learning model applied in this study, which provided the satisfactory and improved accuracy ratio using the selected MOOC dataset. Compared to the other studies, the random forest model selected as optimal model for predicting dropout ratio. The benchmark set in this study can provide the better understating for the organization and researchers on this issue. In future, the framework can be integrated using other features and algorithms for enhancing the accuracy result and capability of the machine learning models.

References

[1] T. L. Friedman, "Come the Revolution - NYTimes.com," New York Times, vol. 12, pp. 5–12, 2012.

[2] H. M. Dai, T. Teo, N. A. Rappa, and F. Huang, "Explaining Chinese university students' continuance learning intention in the MOOC setting: A modified expectation confirmation model perspective," Comput. Educ., vol. 150, 2020.

[3] D. Shah, "MOOC Students Statistics," 2019. [Online]. Available: https://www.classcentral.com/report/mooc-stats-2019/.

[4] J. Chen, J. Feng, X. Sun, N. Wu, Z. Yang, and S. Chen, "MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine," Math. Probl. Eng., 2019.

[5] F. Dalipi, A. S. Imran, and Z. Kastrati3, "MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges," in EDUCON 2018, 2018.

[6] W. Xing and D. Du, "Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention," J. Educ. Comput. Res., 2018.

[7] H. Khalil and M. Ebner, "Moocs completion rates and possible methods to improve retention - a literature review," in World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2014.

[8] D. F. O. Onah, J. E. Sinclair, and R. Boyatt, "Dropout rates of massive open online courses: Behavioural patterns," in International Conference on Education and New Learning Technologies, 2014.
[9] F. Wenzheng, T. Jie, L. Tracy Xiao, Z. Shuhuai, and G. Jian, "Understanding Dropouts in MOOCs," in Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI'19), 2019, pp. 1–8.
[10] J. Chen, J. Feng, X. Sun, N. Wu, Z. Yang, and S. Chen, "MOOC Dropout Prediction Using a Hybrid Algorithm Based on Decision Tree and Extreme Learning Machine," Math. Probl. Eng., vol. 2019, pp. 1–12, 2019.
[11] C. Chen, G. Sonnert, P. M. Sadler, D. D. Sasselov, C. Fredericks, and D. J. Malan, "Going over the cliff: MOOC dropout behavior at chapter transition," Distance Educ., vol. 41, no. 1, pp. 6–25, 2020.
[12] Editor, "KDD Cup 2015," 2015.
[13] J. M. Zimmermann, J., Brodersen, K. H., Heinimann, H. R., &Buhmann, "A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance," J. Educ. Data Min., vol. 7, no. 3, 2015.
[14] G. Siemens and R. S. d Baker, "Learning analytics and educational data mining: towards communication and collaboration," in Proceedings of the 2nd international conference on learning analytics and knowledge, 2012.
[15] F. Saleem and A. Malibari, "DATA MINING COURSE IN INFORMATION SYSTEM DEPARTMENT–CASE STUDY OF KING ABDULAZIZ UNIVERSITY," in 3rd International Congress on Engineering Education, 2011.
[16] B. Jeon and N. Park, "Dropout Prediction over Weeks in MOOCs by Learning Representations of Clicks and Videos," arXiv Prepr. arXiv2002.01955, 2020.
[17] A. A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment," Interact. Learn. Environ., pp. 1–20, 2020.
[18] P. M. Moreno-Marcos, P. J. Muñoz-Merino, J. Maldonado-Mahauad, M. Pérez-Sanagustín, C. Alario-Hoyos, and C. D. Kloos, "Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs," Comput. Educ., vol. 145, p. 103728, 2020.
[19] D. Peng and G. Aggarwal, "Modeling MOOC Dropouts," Entropy, vol. 10, no. 114, p. 49944, 2015.
[20] A. Dangi and S. Srivastava, "Educational data Classification using Selective Naïve Bayes for Quota categorization," in 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE), 2014, pp. 118–121.
[21] J. Liang, J. Yang, Y. Wu, C. Li, and L. Zheng, "Big data application in education: Dropout prediction in edx MOOCs," in Proceedings - 2016 IEEE 2nd International Conference on Multimedia Big Data, BigMM 2016, 2016, pp. 440–443.
[22] J. Liang, C. Li, and L. Zheng, "Machine learning application in MOOCs: Dropout prediction," in ICCSE 2016 - 11th International Conference on Computer Science and Education, 2016, pp. 52–57.
[23] J. Gardner, Y. Yang, R. S. Baker, and C. Brooks, "Modeling and Experimental Design for MOOC Dropout Prediction: A Replication Perspective.," Int. Educ. Data Min. Soc., 2019.
[24] K. M, S. F, Z. Z, and P. N, "Predicting MOOC dropout over weeks using machine learning methods," in Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs, aclweb.org, 2014, pp. 60–65.
[25] S. Nagrecha, J. Z. Dillon, and N. V Chawla, "MOOC Dropout Prediction: Lessons Learned from Making Pipelines Interpretable," in 26th International Conference on World Wide Web Companion, 2017, pp. 351–359.
[26] J. He, J. Bailey, B. I. P. Rubinstein, and R. Zhang, "Identifying at-risk students in massive open online courses," in Proc. of the 29th AAAI Conference on Artificial Intelligence, 2015, pp. 1749–1755.
[27] C. Taylor, K. Veeramachaneni, and U. M. O'Reilly, "Likely to stop? Predicting Stopout in Massive Open Online Courses," arXiv Prepr. arXiv, vol. 1408.3382, 2014.
[28] J. W. You, "Identifying significant indicators using LMS data to predict course achievement in online learning.," Internet High. Educ., vol. 29, 2016.
[29] F. Marbouti, Diefes-Dux, H. A., and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," Comput. Educ., 2016.
[30] L. Qiu, Y. Liu, Q. Hu, and Y. Liu, "Student dropout prediction in massive open online courses by convolutional neural networks," Soft Comput., 2018.
[31] W. Feng, J. Tang, and T. X. Liu, "Understanding dropouts in MOOCs," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, vol. 33, pp. 517–524.
[32] R. M. Team, "Rapid Miner Documentation." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/associations/create_association_rules.html. [Accessed: 10-Mar-2019].
[33] S. Solutions, "Logistic Regression." [Online]. Available: https://www.statisticssolutions.com/what-is-logistic-regression/. [Accessed: 01-Apr-2019].
[34] H. DW Jr, L. S, and S. RX, Applied Logistic Regression, 3rd ed. New Jersey: John Wiley & Sons, 2013.
[35] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE J. Biomed. Heal. Informatics, vol. 18, no. 6, pp. 1750–1756, 2014.
[36] N. Donges, "The Random Forest Algorithm," Towards Data Science, 2018.
[37] S. S. Ayse, "Prediction of Course Completion based on Participants ' Social Engagement on a Social-Constructivist MOOC Platform," UNIVERSITY OF SOUTHAMPTON, 2017.
[38] L. Qiu, Y. Liu, and Y. Liu, "An Integrated Framework with Feature Selection for Dropout Prediction in Massive Open Online Courses," IEEE Access, vol. 6, pp. 71474–71484, 2018.
[39] C. Sheppard, Tree-based Machine Learning Algorithms: Decision Trees, Random Forests, and Boosting. CreateSpace Independent Publishing Platform, 2017.
[40] R. Miner, "Gradient Boosting." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/gradient_boosted_trees.html.
[41] G. S. Abu-Oda and A. M. El-Halees, "Data Mining in Higher Education: University student dropout case study," Int. J. Data Min. Knowl. Manag. Process (IJDKP, vol. 5, no. 1, pp. 97–106, 2015.
[42] R. Gandhi, "Naive Bayes Classifier," Towards Data Science2, 2018.
[43] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques :

a survey," Int. J. Eng. Technol., vol. 7, no. 2.8, pp. 684–687, 2018.
[44] P. Yadav, "Decision Tree in Machine Learning," Towards Data Science, 2018.
[45] S. Angra and S. Ahuja, "Implementation of Data Mining Algorithms on Student's Data using Rapid Miner," in International Conference On Big Data Analytics and computational Intelligence (ICBDACI), 2017, pp. 387–391.
[46] P. Tripathi, S. K. Vishwakarma, and A. Lala, "Sentiment analysis of english tweets using rapid miner," in Computational Intelligence and Communication Networks (CICN), 2015 International Conference on, 2015, pp. 668–672.
[47] Rapid Miner, "Cross Validation Operator." [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/validation/cross_validation.html. [Accessed: 01-Apr-2020].
[48] D. L. Olson and D. Delen, "Performance evaluation for predictive modeling. Advanced data mining techniques," in Advanced Data Mining Techniques, 2008, pp. 137–147.

Fawaz J. Alsolami received the Ph.D. degree in Artificial Intelligence from King Abdullah University of Science and Technology, Saudi Arabia. He is currently working as an Assistant Professor and having the responsibility as Chairman of Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. His research interests include AI, Machine Learning, and Data Mining.