International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 09 | Sep 2019 www.irjet.net p-ISSN: 2395-0072

Analysis of Student Performance Using Machine Learning Techniques

G.Umadevi[1]

[1]Research Scholar, School of Computer Science, Engineering and Applications, Bharathidasan University, Trichy

---***---

Abstract - *Educational data has become a vital resource in this modern era, contributing much to the welfare of the society. Educational institutions are becoming more competitive because of the number of institutions growing rapidly. The higher education institutions has potential knowledge such as academic performance of students, administrative accounts, potential knowledge of the faculty, demographic details of the students and many other information in a hidden form. The technique behind the extraction of the hidden knowledge is Knowledge Discovery process. Data mining helps to extract the knowledge from available dataset and should be created as knowledge intelligence for the benefit of the institution. Higher education does categorize the students by their academic performance. In this paper, perform comparative study to various classification algorithms such as J48, Decision table, and K-Nearest neighbor algorithm (K-NN) can be implemented in prediction model for student datasets. The datasets are clustered using K-Means algorithm before classification. Finally compare the results in terms of error rate metrics in data mining and shows that K-NN algorithm can be provide less error rate than the existing algorithms*

*Key Words: (***Prediction analysis, Classification, Nearest Neighbor, Machine learning algorithm, Educational data***)…*

1. INTRODUCTION

Student's academic performance affected by many factors, like personal, socio-economic and other environmental variable. Knowledge about these factors and their effect on student performance can help managing their effect. Recently, much attention has been paid to educational mining research. Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people learning activities in educational environment. Predicting student's performance becomes more. Challenging due to the large volume of data in educational databases. The topic of explanation and prediction of academic performance is widely researched. The ability to predict student performance is very important in educational environments. Increasing student success is a long term goal in all academic institutions. If educational institutions can predict students' academic performance early before their final examination, then extra effort can be taken to arrange proper support for the low performing students to improve their studies and help them to success. On the other hand, identifying attributes that affect course success rate can assist in courses improvement. Newly developed web-based educational technologies and the application of quality standard offer researchers' unique opportunities to study how students learn and what approaches to learning lead to success. The main objective of the paper is to identify both factors that affect courses success rate and student success rate then using these factors as early predictor to expected success rate and handling their weakness. Data Mining (DM) concept is to extract hidden pattern and to discover relationships between parameters in a vast amount of data. There are many achievements of DM techniques in many areas such as engineering, education, marketing, medical, financial, and sport. It shows the DM technique's ability in providing the alternative solution for decision makers in solving problem arise in particular areas. The exploration data in educational field using DM techniques are called as Educational Data Mining (EDM). EDM is concerned with extracting a pattern to discover hidden information from educational data. DM provides various methods for analysis process which include classification, clustering, and association rule. Classification, which is one of the prediction types classifies data (constructs a pattern) based on the training set and uses the pattern to classify a new data (testing set). Clustering is the process of grouping records in classes that are similar, and dissimilar to records in other classes. In relationship mining, the goal is to discover the relationship exist between parameter. In this study, the classification method is selected to be applied on the students' data. Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into separate groups. The basic data mining task is shown in fig 1

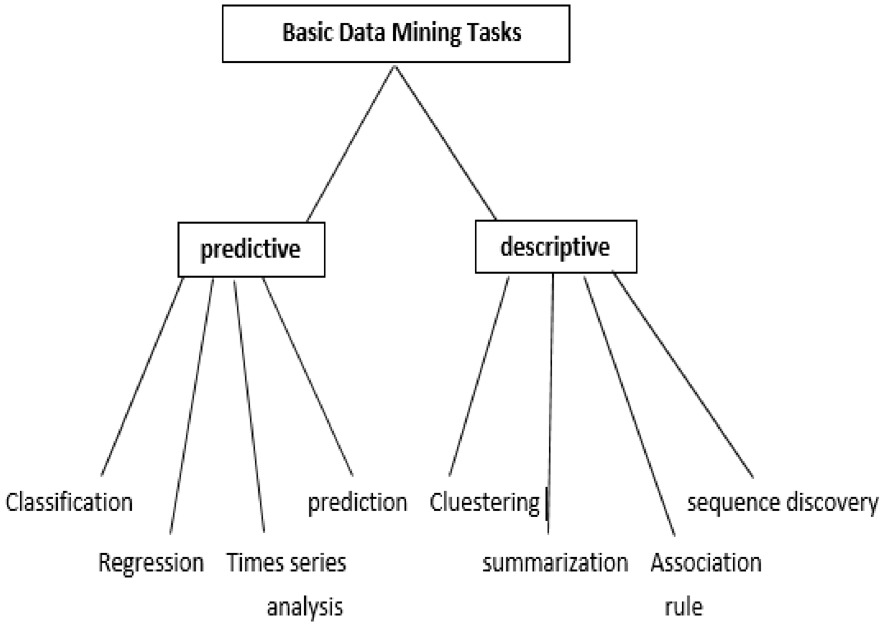

Figure1: Data mining tasks

2. RELATED WORK

Over the last decade ongoing development of statistical modeling tools has led to a growing sophistication in the methods used to analyses relationships between the distributions of species and their environment. We will review the literature of classification analysis and the commonly used techniques in modeling classification problems in this section.

Zacharoula Papamitsiou, et .al [1] provide the reader with a comprehensive background for understanding current knowledge on Learning Analytics (LA) and Educational Data Mining (EDM) and its impact on adaptive learning. It constitutes an overview of empirical evidence behind key objectives of the potential adoption of LA/EDM in generic educational strategic planning. LA and EDM constitute an ecosystem of methods and techniques (in general procedures) that successively gather, process, report and act on machine readable data on an ongoing basis in order to advance the educational environment and reflect on learning processes. In general, these procedures initially emphasize on measurement and data collection and preparation for processing during the learning activities.

Alireza Ahadi, et .al [2] designed Methods for automatically identifying students in need of assistance have been studied for decades. Initially, the work was based on somewhat static factors such as students' educational background and results from various questionnaires, while more recently, constantly accumulating data such as progress with course assignments and behavior in lectures has gained attention. We contribute to this work with results on early detection of students in need of assistance, and provide a starting point for using machine learning techniques on naturally accumulating programming process data. This study is driven by the question of identifying high and low-performing students as early as possible in a programming course to provide better support for them. By high- and low-performing students, we mean students in the upper- and lower half of course scores, and by early, we mean after the very first week of the programming course. This means that instructors could plan and provide additional guidance to specifically selected students already during the second week of the course.

Abeer Badr El Din Ahmed, et .al [3] decision tree method is used on student's database to predict the student's performance on the basis of student's database. We use some attribute were collected from the student's database to predict the final grade of student's. This study will help the student's to improve the student's performance, to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time. Currently the amount huge of data stored in educational database these database contain the useful information for predict of students performance. The most useful data mining techniques in educational database is classification. In this paper, the classification task is used to predict the final grade of students and as there are many approaches that are used for data classification, the decision tree (ID3) method is used here.

Christopher G, et.al [4] e discover video watching behavioral quantities that are correlated with student performance, and show that they can be used to enhance CFA prediction. Additionally, we identify the "early detection" capability of clickstream data, showing that the incremental improvement is higher in the first few course weeks. Moreover, this work is the first to study CFA prediction in the context of MOOC. Each of these are important steps in studying the SLN of MOOC users. Student performance prediction is an intriguing research area, and especially so for MOOC because of its potential benefits, such as the definition of different SLN graph structures that can help an instructor manage her course more effectively. In this paper, using data from one of our own MOOC offerings, we applied some standard algorithms to CFA prediction in this setting, and showed how one type of behavioral data collected about students – video-watching clickstream events can be used as learning features to improve prediction quality. Through evaluation, we saw that our scheme outperformed the standards under each dataset partition and metric considered, and that the improvement was particularly pronounced in the beginning of the course. Also, we saw that it is useful to parse the clickstream data into summary quantities for each user video pair, because in doing so is possible to identify intervals for these quantities that indicate a higher likelihood of a user being CFA or not in answering the corresponding question.

Fadhilah Ahmad, et.al [5] studied the classification method is selected to be applied on the students' data. This research aims to do a comparative analysis among the three selected classification algorithms; Decision Tree (DT), Naïve Bayes (NB), and Rule Based (RB). The comparative analysis is done to discover the best techniques to develop a predictive model for SAP. This pattern will be used to improve the SAP and to overcome the issues of low grades obtained by students. The amount of data stored in an educational database at IHL is increasing rapidly by the times. In order to get the knowledge about student from such large data and to discover the parameter that contributed to the students' success, the classification techniques are applied to the students' data. This study also conducts a comparative analysis of three classification techniques; DT, NB, and RB using WEKA tool. The experimental result shows that the RB has the best classification accuracy compared to NB and DT. The model will allow the lecturers to take early actions to help and assist the poor and average category students to improve their results

3. STUDENT PERFORMANCE PREDICTION USING CLASSIFICATION ALGORITHM

Objective of classification analysis is to explain variability in dependent variable by means of one or more of independent or control variables. The determination of explicit form of classification equation is the ultimate objective of classification analysis. It is finally a good and valid relationship between study variable and explanatory variables. Such classification equation can be used for several purposes. For example, to determine the role of any explanatory variable in the joint relationship in any policy formulation, to forecast the values of response variable for given set of values of explanatory variables. The classification equation helps understands the interrelationships of variables among them. There are various types of classification are implemented in existing framework. The datasets are uploaded in WEKA tool with any WINDOWS OS configuration. Table 1 shows the dataset attributes

S.no	Attributes	Description
1	Gender	Student's gender
2	Nationality	Student's nationality
3	Place of birth	Student's Place of birth
4	Stage ID	Educational level student
5	Grade ID	Grade student
6	Section ID	Classroom student
7	Topic	Course topic
8	Semester	School year semester
9	Relation	Parent responsible for student
10	Raised hands	How many times the student raises his/her hand on classroom
11	Visted resoures	How many times the student visits a course content
12	Announcement view	How many times the student checks the new announcements
13	Discussion	How many times the student participate on discussion groups
14	Parent answering survey	Parent answered the surveys which are provided from school or not
15	Parent school satisfaction	The Degree of parent satisfaction from school
16	Student Absence Days	The Degree of parent for each student
17	Class	Class of the student

Table 1: Dataset Description

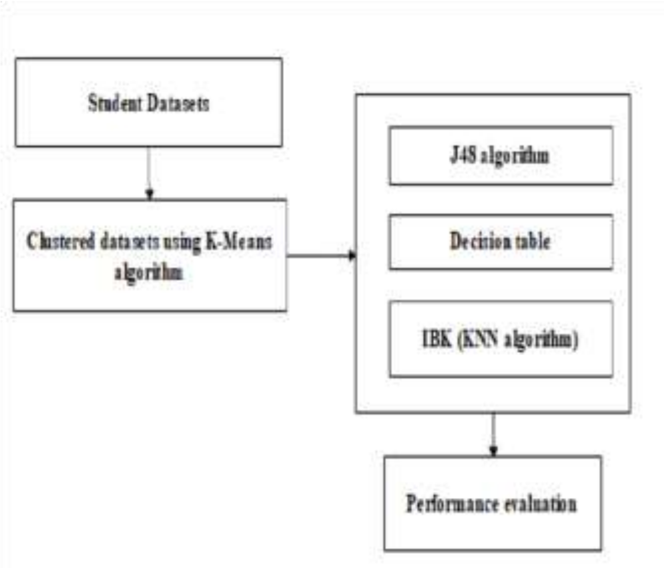

Fig 1: Proposed framework

Fig 1 provide framework for the proposed system include steps such as clustering and classification. K-means clustering can be applied after that applies classification algorithms such as J48 algorithm, Decision table and KNN algorithm. Finally compare the classification algorithm in terms of error rates in performance evaluation

3.1 clustering algorithm

3.1.1 Simple K- Means Clustering algorithm

The student datasets are clustered using simple K-means algorithm. They are provided with a hard and fast of data instances that have to be grouped in keeping with a few notion of correspondence. The algorithm devises access only to the set of features describing each object; it is not given any information as to where each of the instances should be placed within the partition. K-way clustering is a method generally used to mechanically partition a statistics set into okay organizations. It proceeds by selecting k initial cluster centers and then iteratively refining the results. The algorithm converges when there is no further change in assignment of instances to clusters. The student datasets are grouped as two clusters named as normal and abnormal. K-means clustering can be applied after perform the preprocessing for uploaded student datasets. In WEKA tool, Choose cluster options and click drop down list to pick simple K means algorithm and then start cluster to group the classes as 0 and 1.

The basic algorithm pseudo code as follows:

Input: X = {x1, x2, x3,…..,xn} be the set of data points ,

Y= {y1,y2,y3…yn} be the set of data points

V = {v1,v2,v3,….,vn} be the set of centers

Step 1: Select 'c' cluster centers arbitrarily

Step 2: Calculate the distance between each pixels and cluster centers using the Euclidean Distance metric as follows

$$Dist(X,Y) = \sqrt{\sum_{j=1}^{n}(X_{ij} - Y_{ij})^2}$$

X, Y are the set of data points

Step 3: Pixel is assigned to the cluster center whose distance from the cluster center is minimum of all cluster centers

Step 4: New cluster center is calculated using

$$V_i = \frac{1}{C_i}\sum_{1}^{ci} x_i$$

Where Vi denotes the cluster center, ci denotes the number of pixels in the cluster

Step 5: The distance among every pixel and new obtained cluster facilities is recalculated

Step 6: If no pixels were reassigned then stop. Otherwise repeat steps from 3 to 5

3.2 classification algorithm

3.2.1 J48 algorithm:

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. Like CLS and ID3, C4.5 generates classifiers expressed as decision trees, but it can also construct classifiers in more comprehensible rule set form

3.2.1.1 J48 tree construction:

Decision trees are trees that classify instances by sorting them based on feature values Given a set S of cases, C4.5 first grows an initial tree using the divide-and-conquer algorithm as follows: If all the cases in S belong to the same class or S is small, the tree is a leaf labeled with the most frequent class in S. Otherwise, choose a test based on a single attribute with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition S into corresponding subsets S1, S2… according to the outcome for each case, and apply the same procedure to each subset. Decision trees are usually unvaried since they use based on a single feature at each internal node. Most decision tree algorithms cannot perform well with problems that require diagonal partitioning. A decision tree model is a standout amongst the most widely recognized information mining models. This calculation utilizes a recursive apportioning approach. A decision tree is the prototypical information mining instrument, generally utilized for its simplicity of translation. It comprises a root node split by a solitary variable into two segments. Thus, these two new segments turn out to be new nodes that may then each further split on an individual (and unique) variable. This partitioning proceeds until no further part would enhance the execution of the model. The basic algorithm is shown in fig 1

ALGORITHAM 1 C4.5 (D)

Input: an attribute-value dataset D

Tree = {}

if D is "pure" OR other stopping criteria met then

terminate
end if

for all attribute a € D do

compute information theoretic criteria if we split on a
end for

: a_{best}=*Best attribute according to above c***riteria**

tree = create a decision node that tests a_{best} **in the root**

D_v = induced sub − datasets from D based on a_{best}
for all D **do**
$Tree_v$=C4.5 (D_v)

Attach $Tree_v$ **to the corresponding branch of tree**
End for

return Tree

Fig 2 J48 Algorithm

3.2.2 DECISION TABLE

Decision tables are a concise visual representation for specifying which actions to perform depending on given conditions. They are algorithms whose output is a set of actions. The information expressed in decision tables could also be represented as decision trees or in a programming approach in terms of if then else rules if-then-else rules. Each decision corresponds to a variable, relation or predicate whose possible values are listed among the condition alternatives. Each action is a procedure or operation to perform, and the entries specify whether (or in what order) the action is to be performed for the set of condition alternatives the entry corresponds to. The algorithm, decision table, is found in the Weka classifiers under Rules. The simplest way of representing the output from machine learning is to put it in the same form as the input. The use of the classifier rules decision table is described as building and using a simple decision table majority classifier. The output will show a decision on a number of attributes for each instance. The number and specific types of attributes can vary to suit the needs of the task.

a) Entropy using the frequency table of one attribute

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

b) Entropy using the frequency table of two attributes

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

3.2.3 K-Nearest Neighbor algorithm:

Suppose that an object is sampled with a set of different attributes, but the group to which the object belongs is unknown. Assuming its group can be determined from its attributes; different algorithms can be used to automate the classification process. A nearest neighbor classifier is a technique for classifying elements based on the classification of the elements in the training set that are most similar to the test example. With the k-nearest neighbor technique, this is done by evaluating the k number of closest neighbors. The k-nearest neighbors algorithm is one of the simplest machine learning algorithms. It is simply based on the idea that ―objects that are ‗near' each other will also have similar characteristics. Thus if you know the characteristic features of one of the objects, you can also predict it for its nearest neighbor. k-NN is an improvisation over the nearest neighbour technique. It is based on the idea that any new instance can be classified by the majority vote of its ‗k' neighbours, - where k is a positive integer, usually a small number. kNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time. When dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance. A major problem when dealing with the Euclidean distance formula is that the large values frequency swamps the smaller ones. The algorithm steps as follows:

for all the unknown samples Un Sample(i)
for all the known samples Sample(j)
compute the distance between
Un samples(i) and Sample(j)
end for
find the k smallest distances
locate the corresponding samples
Sample(j1),….,Sample(jK)
assign Un Sample(i) to the class which appears more frequently
end for
The basic diagram of KNN is shown in fig 3

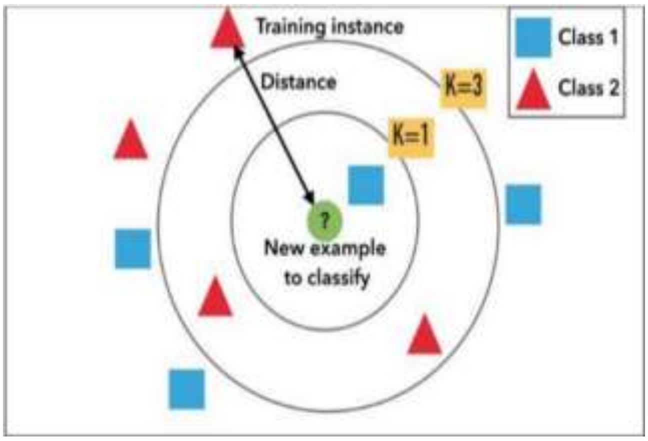

Fig 3: KNN classification

4. EXPERIMENTAL RESULTS

We can upload the datasets for 480 student and collect the samples from KAGGLE student database. And using 17 attributes for predicting student‗s performance. The dataset contains the attributes such as gender, nationality, place of birth, staged, grade id, sectioned, topic, semester, relation, raised hands. Visted resources, announcement view, discussion, parent answering survey, parent school satisfaction, student absence days, class. These attributes can perform classification and clustering using tool named as WEKA for WINDOWS OS with any configuration. We can evaluate the performance of each algorithm and compare the performance based on MSE, RMSE, RAE, RRSE and shown in table and performance graph.

1) MSE:

MEAN SQUARED ERROR (MSE) is by far the most common measure of numerical model performance. It is simply the average of the squares of the differences between the predicted and actual values. It is a reasonably good measure of performance, though it could be argued that it overemphasizes the importance of larger errors. Many modeling procedures directly minimize the MSE.

2) RMSE:

The ROOT MEAN SQUARE ERROR (RMSE) serves to aggregate the magnitudes of the errors in predictions into a single measure of predictive power. RMSE is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent.

3) RAE:

The RELATIVE ABSOLUTE ERROR (RAE) in some data is the discrepancy between an exact value and some approximation to it. An approximation error can occur because:

1. The measurement of the data is not precise due to the instruments.

2. Approximations are used instead of the real data (e.g., 3.14 instead of π).

In the mathematical field of numerical analysis, the numerical stability of an algorithm indicates how the error is propagated by the algorithm.

4) RRSE:

The ROOT RELATIVE SQUARED ERROR (RRSE) is relative to what it would have been if a simple predictor had been used. More specifically, this simple predictor is just the average of the actual values. Thus, the relative squared error takes the total squared error and normalizes it by dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted.

Algorithms	MSE	RMSE	RAE	RRSE
Decision table	0.29	0.37	68	79
J48 algorithm	0.21	0.36	49	78
Ibk	0.18	0.32	42	75

Table 2: Performance table

The overall performance of the results is shown as

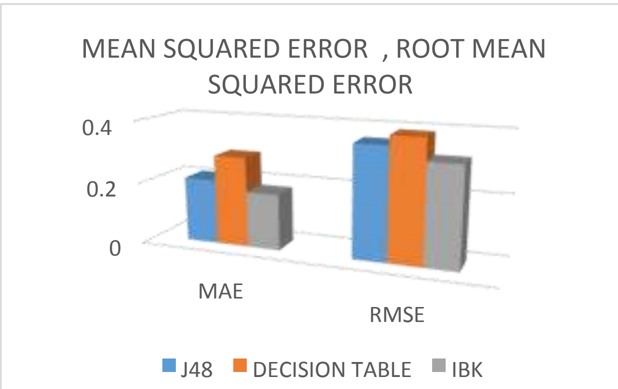

FIG 4 MSE AND RMSE GRAPH

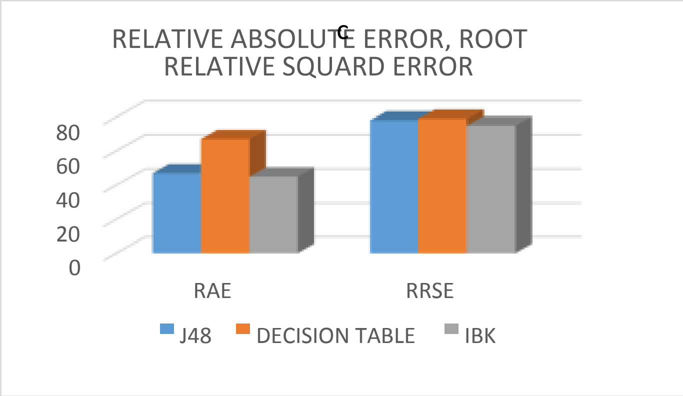

FIG 5 RAE AND RRSE GRAPH

KNN can be outperforms than the existing algorithms and provides reduce number of error rate values. In this paper, a novel approach based on KNN significant academic attributes for performance predictions. The experiment displays good performance of the proposed algorithm and was compared to similar approaches over the same dataset. By analyzing the experimental results, it is observed that the KNN algorithm turned out to be best classifier for student performance prediction because it contains more accuracy and least error rate.

3. CONCLUSION

Using data mining technology for student performance prediction has become the focus of attention in educational data mining. Data mining technology provides an important means for extracting valuable rules hidden in student data and acts as an important role in student performance prediction. In the current study, have demonstrated, using a large sample of student datasets with classification. In this research work, the classification rule algorithms namely J48 algorithm, Decision table and IBK algorithm are used for classifying datasets which are uploaded by user. By analyzing the experimental results it is observed that the IBK algorithm has yields better result than other techniques. In future we tend to improve efficiency of performance by applying other data mining techniques and algorithms.

REFERENCES

[1] Papamitsiou, Zacharoula, and Anastasios A. Economides. "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence." Journal of Educational Technology & Society 17.4 (2014): 49.

[2] Ahadi, Alireza, et al. "Exploring machine learning methods to automatically identify students in need of assistance." Proceedings of the Eleventh Annual International Conference on International Computing Education Research. ACM, 2015.

[3] Ahmed, Abeer Badr El Din, and Ibrahim Sayed Elaraby. "Data Mining: A prediction for Student's Performance Using Classification Method." World Journal of Computer Application and Technology 2.2 (2014): 43-47.

[4]Brinton, Christopher G., and Mung Chiang. "Mooc performance prediction via clickstream data and social learning networks." Computer Communications (INFOCOM), 2015 IEEE Conference on. IEEE, 2015.

[5] Ahmad, Fadhilah, NurHafieza Ismail, and Azwa Abdul Aziz. "The prediction of students' academic performance using classification data mining techniques." Applied Mathematical Sciences 9.129 (2015): 6415-6426