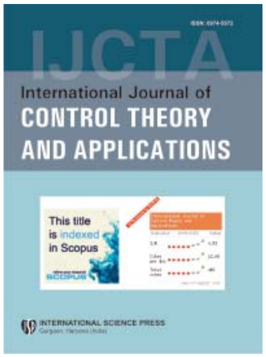

International Journal of Control Theory and Applications

ISSN : 0974–5572

© International Science Press

Volume 10 • Number 23 • 2017

A Study on Classification and Clustering Data Mining Algorithms based on Students Academic Performance Prediction

K. Govindasamy[a] and T. Velmurugan[b]

[a]*Research Scholar, VELS University, Chennai, India.*

E-Mail: mphilgovind@gmail.com

[b]*Associate Professor, PG and Research Department of Computer Science, D. G. Vaishnav College, Chennai, India.*

E-Mail: velmurugan_dgvc@yahoo.co.in

Abstract: Academic Data mining is a recent and quickly growing a very important technique in the investigation of data generated in Educational domain. In this article, an analysis of student results of UG and PG degree was carried out using some of the classification and clustering algorithms in data mining. The data set about student details are collected from four private colleges in Tamil Nadu state of India. The primary objective of this research is to evaluate some of the algorithms in the prediction of student's academic performance in their end semester examinations. The frequently used clustering algorithms such as Expectation Maximization (EM) and *k*-Means algorithm, and classification algorithms such as C4.5 algorithm, *k*-Nearest Neighbor algorithm and Naïve Bayes algorithm are utilized to carry out for the prediction of students' performance. The performance of these algorithms is analyzed based on their accuracy of results. Also, the performances of these algorithms were compared with one another by means of classification accuracy.

Keywords: *Education Data Mining, Classification Algorithms, Students performance, Clustering Algorithms.*

1. INTRODUCTION

Data mining in simple terms can be told as a method for extracting meaningful set of patterns in huge/bulk quantities of data sets. Recently there is an increasing awareness in data mining, where academic data mining is being investigated widely along with the help of learning systems. Academic performance prediction of students is actually a challenging task in current scenario. The growth in information and statement technologies has changed the way in which large quantities of information are accessed, such that the work of academic leaders is reduced or made easy. Important decisions can be made by the academic leaders with the help of the huge data available to them using various algorithms.

Usually educational organizations used to collect huge amount of data which would be relevant to faculty members, students, etc. But the importance of data that is collected is unknown. The data that are used in generating simple queries or traditional reports may be insignificant, which will not contribute to the process of inference/decision making in the educational organizations. The collected data may also contain such insignificant data. Also the volume and complexity of the collected data may be very high such that it is not easy to handle. If that is the case then the collected data may not be used and memory is occupied unnecessarily. The available data can be made usable if and only if it is converted into useful information by exploiting potentiality of the collected data. A wide range of data mining algorithms such as J48, AD Tree, C4.5, Random Tree, etc [2] are used to extract useful information from potential data gathered in various educational organizations. When the data mining algorithms are used effectively to predict students' performance it leads to development of the students which in turn leads to development of the nation [1].

The rest of the article is organized as follows. Section 2 discusses about various research articles related to data mining techniques used in predicting students' performance were discussed. Section 3 explores about the dataset and data mining algorithms such as C4.5 algorithm, *k*-nearest neighbor algorithm, Naïve Bayes algorithm, EM algorithm and k-Means algorithm used for predicting academic performance of students in detail. The prediction results of each algorithm were examined in detail and compared with each other to evaluate the performance of the algorithms is given in section 4. Finally, section 5 concludes and given inference from the Experiments and also provides suggestion for further research.

2. RELATED WORK

In educational data mining various research have been done in predicting students' performance using different data mining techniques such as clustering, classification, neural networks, etc. Some of the methodologies from different research articles were discussed in this section. Xindong Wu et al. have surveyed various data mining algorithms of different categories such as association analysis, clustering, classification, statistical analysis and link mining in their research. Based on their survey they provided top 10 data mining algorithms which come under the above categories. The top 10 algorithms are C4.5, *k*-Means, Support Vector Machine (SVM), Apriori, EM, PageRank, AdaBoost, K nearest neighbor algorithm, Naive Bayes, and CART [18]. In this article five of the ten algorithms under the category classification and clustering are used for predicting student's academic performance with real time data sets from educational organizations.

Archana T and Usha Devi Gandhi have analyzed various research articles related to educational data mining from the year 2000 to 2016 and provided inferences about the same. They have considered prediction of student performance in traditional classroom environment and online tutoring system, and also early prediction of student dropout and retention. They concluded that effective usage of educational data mining leads to development of nation [1]. Agrawal Bhawana D and Gurav Bharti Bhave reviewed various Data Mining Techniques such as association rule mining and classification in Education Domain. They mainly focused on predicting the low performance of students in academics at school using decision tree algorithms such as J48, AD Tree, C4.5, Random Tree, etc. They concluded that studentfailure prediction at school level can be a difficult task as it is amultifactor problem and also the available data are usually imbalanced [2].

Ajith P et al. used Association Rules Instead of tree based classification to perform student performance analysis since tree based classification is complicated to understand and depends on the technical competency of the decision maker. They analyzed that Association Rules aims at discovering implicative tendencies that can be valuable information for the decision-maker which is absent in tree based classifications. So they used a new interactive approach top rune and filter discovered rules. They integrated user knowledge in the post processing task and created a Rule Schema formalism extending the specifications to obtain association rules from knowledgebase. Based on their research they concluded that the results obtained using Association Rules are better to understand and can be applied to real time use when compared to tree based classifications [4].

Baker RSJD has done a research on data mining in education, in which he analyzed various approaches such as clustering, relationship mining, prediction, discovery with models, and distillation of data for human judgment. In the illustrative example provided in his article, an analysis was made on junior school students who benefitted from re-reading and students who did not benefit from re-reading. This analysis done using data mining determined that students with overall low reading speed who were receiving special needs learning support actually benefitted from re-reading [7]. Anuradha C and Velmurugan T have evaluated various classification algorithms in predicting student's performance in their research. They have analyzed the performance of classification algorithms such as J48, OneRip, JRip, Naïve Bayes classifiers and Bayesian Net classifiers using data set of students from three private colleges in Tamilnadu state of India. They illustrated that the prediction rates of the above said algorithms are not uniform, which varies from 61-75%. Also they have found that the data attributes (first and second classes)have significantly influenced the classification process. They have also suggested to use larger data sets for more accurate results [6].

Dinesh Kumar A and Radhika V have done a survey on student performance prediction in which they analyzed the predictive model in data mining. According to their survey the factors such as student family income, learning behavior, student family size and mother's qualification affected the student performance. Using the predictive model in data mining academic performance (success and failure) of student can be predicted which helps the teachers to concentrate more on the students who might tend to fail in future [8]. Dorina Kabakchieva has predicted Bulgarian university student's performance by using data mining classifier methods such as KNN, J48, OneRip, JRip, Naïve Bayes classifiers and Bayesian Net classifiers in her research. But the predication rates were not up to the mark, ranging between 52-67%. However the conclusions obtained from the research were used for providing recommendations to the university management, concerning the sufficiency and availability of university data, and also helped in improving the data collection process of university [9].

Ogunde AO and Ajibade DA have used a data mining technique called ID3 Decision tree algorithm to predict Graduation grades of University students. They used data such as entrance examination score and grade in secondary school as input for prediction. A model for predicting students' graduation grades was generated by training the gathered data. They categorized five classes such as Pass, Third Class, Second Class Lower, Second Class Upper and First Class for which the accuracy/true positive rateis obtained as 0%, 37.5%, 65%, 66.7%, and 30% respectively using the ID3 algorithm [12]. Ajay Kumar Pal and Saurabh Pal have analyzed educational data in different degree colleges and institutions affiliated with VBS Purvanchal University, Jaunpur, India for predicting performance of students using classifiers such as ID3, ADT and Bagging. They used WEKA tool for this process. In their research according to the data set used, ID3 Classification is chosen as the best algorithm since it has the highest accuracy of 78% and least time taken to generate the model. Also they inferred from all the classifiers results that the students who are likely to fail may be successfully identified [3].

Anju Rathee and Robin Prakash Mathurhave surveyed various decision tree classification algorithms such as ID3, C4.5 and CART for evaluating student academic performance of students in their research. These algorithms were used on internal exam data of students to predict their performance in the university end semester exam. The prediction results provided by these algorithms enabled the tutors to know the slow learners and improve their performance. They concluded that the C4.5 is the best algorithm among all the three because it provides better accuracy and efficiency than the other algorithms[5]. Hashmia Hamsa et al. have used two classification methods, decision tree and fuzzy genetic algorithms for predicting academic performance of students in their research. They considered parameters such as admission scores, sessional marks and internal marks in their dataset for their research. They developed prediction model for Bachelor and Master degree student in Electronics and Communication and Computer Science. They concluded that the prediction from decision tree categorized more students in risk class and prediction from fuzzy genetic algorithm categorized more students in safe class as students between safe and risk class are only considered for fuzzy genetic algorithm [10].

Mojisola G. Asogbon et al. performed academic performance prediction of students using Multiclass support vector machine in their research. They used student dataset from the University of Logos, Nigeria for evaluating the performance of Multiclass support vector machine predictor. They used 7-fold cross validation technique to enhance the performance of Multiclass support vector machine, which effectively predicted academic results/performance for all categories of students. They suggested that with the help of prediction results the managements of the institutions can place the students into appropriate faculty programs [11]. Romero C and Ventura S have done a detailed survey of educational data mining from the year 1995 to 2005. In their article they discussed about various data mining tools available for effective analysis and applications of data mining in various educational areas. They also suggest the usage of e-learning recommendation agents which analyses what a student is doing and recommends relevant actions they think would be beneficial to the student. Also they suggested that integrating recommenders with domain knowledge and onto logies will yield more benefits[13]. Shanmuga Priya K and Senthil Kumar AV did research on improving students' performance using classification technique (ID3 Algorithm) in data mining. They used the post graduate internal exam student data of the department of Information Technology, Hindustan College of Arts and Science, Coimbatore for their research. They included a special attribute, extra-curricular activities in addition to the regular attributes in their data set which created an impact such that it contributes to the gain of the prediction and help improve students' performance [14].

Suman and Pooja Mittal Pdid a comparative analysis on role of data mining in education. In their article they compared classification techniques such as naive net, Bayes net and decision tree, and clustering techniques such as hierarchal, k-mean, DBSCAN and OPTICS. They also discussed on which algorithm is suitable for which domain, especially which algorithms are suitable for prediction in education sector[15]. Trivedi A evaluated students classification based on decision tree in his research methodology. In his research he developed a prediction model for student results in any educational organization, where this model is based on five subject marks of all students. This model based on decision tree predicts the classes accurately as the collected test data is from a valid source[17].

Surjeet KumarYadav et al. have studied various data mining applications for predicting the student's performance in their research. In their article they discuss the usage of decision trees in educational data mining to predict students' performance. Decision tree algorithms such as ID3, C4.5 and CART were applied on past performance of students' data to generate the model, where this model was used to predict the performance of students. They used WEKA tool for analysis of the above mentioned algorithms. Their results illustrates that the best algorithm for data classification is CART [16]. Some of the data mining techniques discussed in this section were used in the proposed method for predicting student academic performance. Reviewing these research articles gave an insight for applying and evaluating different classification and clustering data mining algorithms.

3. PROPOSED METHODOLOGY

In this article, various clustering and classification algorithms were examined and compared using MATLAB (MATrix LABoratory) software and results are discussed. For preprocessing of initial data WEKA software is used. When performing data mining, a large part of the work is to manipulate data as the data may be available in any form without a proper structure. MATLAB has a lot of toolboxes for data mining so the part of coding the algorithm can be quite short. And when manipulating data, MATLAB is definitely better. Since it is developed to work with matrices, deleting a row, a column, transposing a matrix or calculating the determinant, all these can be done in one line of code. The MATLAB software is one of the effective tools which can be used for analysis in data mining. The algorithms used in this article are C4.5, Naive Bayes, Nearest Neighbor algorithm (IBk), *k*-Means and EM algorithm.

Classification is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis. It is one of the several methods which intend to make the analysis of very large data sets effective. A well-planned data classification system makes essential data easy to find and retrieve. Data classification procedures and guidelines define what categories and criteria the organization will use to classify data. Classification models predict categorical class label, such classification model will be built to categorize student performance as first class, second class, third class, etc. The classification algorithms used in the proposed methodology are C4.5 algorithm, k-Nearest Neighbor algorithm and Naïve Bayes algorithm.

Clustering is the grouping of a particular set of objects based on their characteristics and aggregating them according to their similarities. With respect to data mining this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis. It allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships. The clustering algorithms used in the proposed methodology are EM and k-Means algorithm.

3.1. C4.5 Algorithm

C4.5 Algorithm generates Decision Trees which can be used for classification problems. It improves the ID3 algorithm by dealing with both continuous and discrete attributes, missing values and pruning trees after construction. It is a supervised learning algorithm which requires a set of training examples, where each example can be seen as a pair of input object and desired output value (class). The algorithm analyzes the training set and builds a classifier that must be able to correctly classify both training and test examples. A test example is an input object and the algorithm must predict an output value (the example must be assigned to a class). The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The technique uses Gain Ratio instead of Information Gain for Splitting purpose [6].

$$\begin{aligned} \text{Gain Ratio (D, S)} &= \text{Gain (D, S)/Split INFO} \\ \text{Where Split INFO} &= \left[\sum_{i-1}^{S} \frac{Di}{D} \log_2 \frac{Di}{D}\right] \end{aligned} \tag{1}$$

The algorithm, summarized as follows.

Step 1: Create a node N;

Step 2: If samples are all of the same class, C then

Step 3: Return N as a leaf node labeled with the class C;

Step 4: If attribute-list is empty then

Step 5: Return N as a leaf node labeled with the most common class in samples;

Step 6: Select test-attribute, the attribute among attribute-list with the highest information gain;

Step 7: Label node N with test-attribute;

Step 8: For each known value aiof test-attribute

Step 9: Grow a branch from node N for the condition test-attribute = *ai*;

Step 10: Let sibe the set of samples for which test-attribute = *ai*;

Step 11: If siis empty then

Step 12: Attach a leaf labeled with the most common class in samples;

Step 13: Else attach the node returned by generate decision tree (*si*, attribute-list, and test-attribute)

3.2. Naive Bayes Classifier

A Naive Bayes classifier assigns a new observation to the most probable class, assuming the features are conditionally independent when the class value is given.Naive Bayes classifier is used for analysis in this method, which is important for several reasons. It is very easy to construct, not needing any complicated iterative parameter estimation schemes. This means it may be readily applied to huge datasets. It is easy to interpret, so users unskilled in classifier technology can understand why itis making the classification it makes. And finally, it often does surprisingly well.

A Bayesian classifier is based on the idea that the role of a (natural) class is to predict the values of features for members of that class. Bayesian classifiers are based on Bayes theorem, which says

$$P\,(cj \mid d) \;=\; p\,(d \mid cj)\;P\,(cj\,)\;p\,(d) \tag{2}$$

$P\,(cj \mid d)$ = Probability of instance d being in class cj,

$p\,(d \mid cj)$ = Probability of generating instance d given class cj,

$P\,(cj)$ = Probability of occurrence of class cj,

$p\,(d)$ = Probability of instance d occurring

3.3. *k*-Nearest Neighbor Classifier (*k*-NN)

k-NN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. *k*-NN has been used in statistical estimation and pattern recognition already in the earlier days as a non-parametric technique. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor. KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry The K-nearest-neighbor (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set.

Distances: For the computation of distance between two scenarios using some distance function $d(xy)$, where xy are scenarios composed of N features, such that

$$x = \{x_1 \ldots x_N\}$$

$$y = \{y_1 \ldots y_N\}$$

Two distance functions are discussed in this summary:

Absolute distance measuring: $$d_A(x, y) = \sum_{i=1}^{N} | x_i - y_i | \tag{3}$$

Euclidean distance measuring: $$d_E(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \tag{4}$$

Because the distance between two scenarios is dependant of the intervals, it is recommended that resulting distances be scaled such that the arithmetic mean across the dataset is 0 and the standard deviation 1. This can be accomplished by replacing the scalars xy with $x'y'$ according to the following function:

$$x' = \frac{x - \overline{x}}{\sigma(x)} \tag{5}$$

Where x is the unscaled value, $\overline{x}$ is the arithmetic mean of feature Equation 4), $\sigma(x)$ is its standard deviation (see Equation 5), and x'

The arithmetic mean is defined as: $$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6}$$

x across the data set (see is the resulting scaled value.

We can then compute the standard deviation as follows:

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{7}$$

3.4. EM Algorithm

EM algorithm handles data summaries more effectively by taking each clustering feature as a data object. It explicitly processes features such as cardinality, mean, and the second-order statistics of a sub cluster. It describes a sub cluster of data items more accurate, and so less sensitive to the data summarization procedure. The generated clusters are very close to the original ones. The general EM algorithm is a common iterative scheme to maximize the likelihood. Thus a maximization likelihood estimate can be obtained. The general EM algorithm is profitably applied on incomplete-data problems, one typical example of which is cluster analysis if class indicator is regarded as missing values. Its basic idea is to associate with the given incomplete data problem, a complete-data problem for which the maximum likelihood estimate is computationally tractable.

$$\Lambda(p, \lambda) \;=\; -\textstyle\sum_{i=1}^{n} p_i \log_2 p_i - \lambda\,(\textstyle\sum_{i=1}^{n} p_i - 1) \tag{8}$$

Where, *p* is an open set of attributes subject to constraints.

3.5. *k*-Means Algorithm

k-Means algorithm, is used to solve the *k*-Means clustering problem. The first step in this algorithm is to decide the number of clusters. It is mandatory that the number of clusters decided should match the data. An incorrect choice of the number of clusters will invalidate the whole process. An empirical way to find the best number of clusters is to try *k*-Means clustering with different number of clusters and measure the resulting sum of squares. Then the center of the clusters should be initialized. The closest cluster should be attributed to each data point an dthe position of each cluster is set to the mean of all data points belonging to that cluster. This process should be repeated until convergence. If there are n data points *xi*, *i* = 1...*n* that have to be partitioned in *k* clusters. The goal is to assign a cluster to each data point. *k*-Means is a clustering method that aims to find the positions μ*i*, *i* = 1... *k* of the clusters that minimize the distance from the data points to the cluster.

Implicit objective function in *k*-Means measures sum of distances of observations from their cluster centroids, called Within-Cluster-Sum-of-Squares (WCSS). This is computed as

$$\mathrm{K} \;=\; \sum_{j=1}^{k}\sum_{i=1}^{n} || x_i^{(j)} - c_j ||^2 \tag{9}$$

where $|| x_i^{(j)} - c_j ||^2$ is a chosen distance measure between a data point $x_{(i)}^{(j)}$ and the cluster center c_j, is an indicator of the distance of the *n* data points from their respective cluster centers. The algorithm is composed of the following steps:

Step 1: Place *k* points into the space represented by the objects that are being clustered. These points represent initial group centroids.

Step 2: Assign each object to the group that has the closest centroid.

Step 3: When all objects have been assigned, recalculate the positions of the *k* centroids.

Step 4: Repeat steps 2 and 3 until the centroids no longer move.

This produces a separation of the objects into groups from which the metric to be minimized can be calculated. The *k*-means is simple clustering algorithm that has been improved to several problem domains.

3.6. Problem Statement

Educational organizations have huge amount of data relevant to students. Such data can be effectively used to predict the academic performance of students by using various data mining algorithms. Based on the research on various articles it can be inferred that the same algorithm provides different results under different circumstances, which is mainly due to the amount of data being used and the attributes selected in the data set. So, care must be taken in selecting the data set and choosing the attributes/variables in the data set. Also, comparing and evaluating various data mining algorithms is a challenging task as the parameters used. Forpredicting the performance of the data mining techniques may vary depending upon the category of the mining technique. Hence,utmost care should be taken in comparing the data mining algorithms when predicting academic performance of students.

3.7. Description of Dataset

Database of certain students were collected from four private Arts and Science Colleges in Chennai city of Tamilnadu, India. A total of 108 students' details were available in the database. The database contains details of the students along with their performance in internal exams and end semester exams. This database is mainly used for evaluating the performance of various classifications and clustering algorithms to predict the academic performance of the students in theirend semester examinations. In this research, classification algorithms such as C4.5 algorithm, Naive Bayes Classifiers and *k*-NN algorithm, and clustering algorithms such as EM algorithm and k-Means clustering algorithm are compared.

Initially the database is created by obtaining basic data of students from admission department in the College. Then, the database is strengthened by adding mark details of students which were obtained from corresponding subject departments. Some of the sensitive information, which may add more value to the data set, was directly collected from the concerned students through questionnaire. All the above information consolidated as a whole form the complete dataset for the proposed methodology. In the dataset, the output attribute is the Student End Semester Examination Marks (ESM) which will be usually in numerical form (percentage). Values of the output attribute can be categorized into four classes such as Distinction Class (output attribute value >60%), First Class (output attribute value from 45 to 60%), Second Class (output attribute value from 36 to 45%), Fail (output attribute value <36%). Description for the attributes defined for current research is given below:

S.N: Serial Number

Name: Student Name

Sex: Gender Male/Female

Branch: B.A., Eng, B.A., Tam, BBA, BCA, B.Com,

SSG: 10TH Marks in State Board

HSG: 12TH Marks in State Board

Medium: Studies in school - English/Tamil

LOC: Student Staying (Rural, Town, Urban)

HOS: Student staying in Hostel or not

FSIZE: Family Size

TFA: Type of Family (Individual/Joint/Orphan)

FINY: Family Income - Yearly

FQUAL: Father and Mother Qualification

MQUAL: Mother Qualification

PSM: Previous Semester Mark

CTG: Class Test Grade

SEM_P: Seminar Performance

ASS : Assignment

ATT: Class attendance for student

ESM: End Semester Examination

4. EXPERIMENTAL RESULTS AND OBSERVATION

Student academic performance is predicted based on multiple input attributes. Algorithms such as C4.5, Naive Bayes and KNN were used on the input attributes to generate a classification model in-order to predict academic performance of students. In this research, WEKA application was used for preprocessing input data and MATLAB is used for evaluating the performance of various data mining algorithms. For all classification algorithms two testing were performed, which are cross validation and percentage split in-order to ensure exact comparison of the classification algorithms. For clustering algorithms testing is done based on the cluster size. Student marks are categorized into various classes such as Distinction, First, Second and Fail, which are considered as clusters. Cluster size is formed with respect to the clusters by the clustering algorithm, which represents the number of students in that particular category. Before performing the test, data must be preprocessed initially. Distribution of data in the stage of preprocessing is illustrated in Figure1 and is given below.

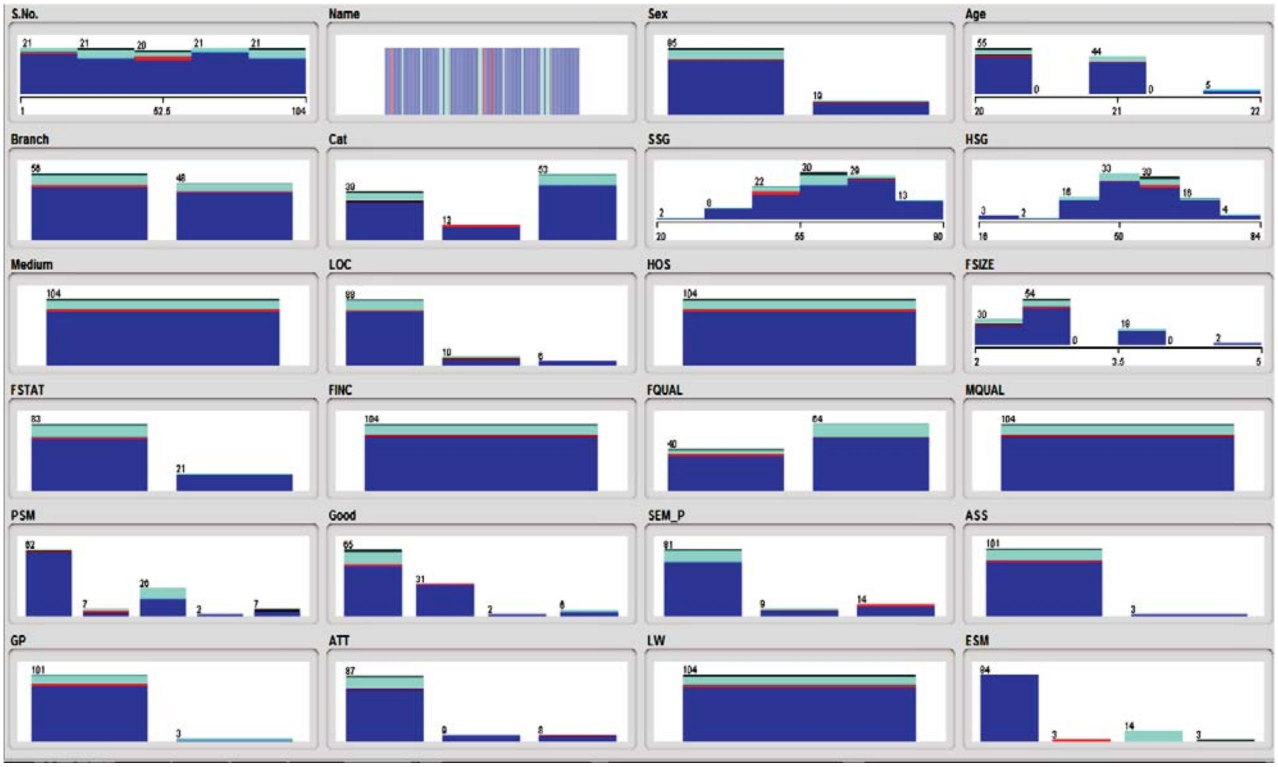

Figure 1: Data Preprocessing using WEKA

4.1. Results of C4.5 Algorithm

C4.5 algorithm was analyzed based on the collected dataset and the results of the analysis are illustrated clearly in the Table 1. It can be vividly seen from the Table 1, that when cross-validation testing is done for C4.5it correctly classified about 83.7% and when percentage split testing is done for C4.5it correctly classified about 80.1%. The results given in the Table 1illustrates that for three of the classes Distinction (90-100%), First (42.9%-100%)and Second (34.5-92.9%) the True Positive (TP) Rate is high, whereas the Class Fail (0%) has very low TP rate. The Distinction class (100%) and Second class (88.6-100%) have high Precision, and First class (13-54.5%) has medium precision, whereas the class Fail (0%) has very low Precision.

Table 1
Results of C4.5 algorithm

Class	*C4.5-Cross validation*		*C4.5-Percentage split*	
	Precision	*TP Rate*	*Precision*	*TP Rate*
Distinction	1	1	1	0.900
First	0.545	0.429	0.130	1
Second	0.886	0.929	1	0.345
Fail	0	0	0	0
Average	0.818	0.837	0.897	0.801

4.2. Results of Naive Bayes Algorithm

Naive Bayes was analyzed based on the collected dataset and the results of the analysis are illustrated clearly in the Table 2. It is evident from the Table 2 that when 10-fold cross-validation testing is done on Naive Bayes algorithm it correctly classified about 76.9% and when percentage split testing is done on Naive Bayes algorithm it correctly classified about 85.7 %.The results given in the Table 2illustrates that for the class Second the TP rate (91.7-96.6%) are high, and for all the other classes Distinction, First and Fail the TP rate is very low. For the Second class the precision (83.7-90.3%) is also high, for First class the precision is medium, whereas for Distinction and Fail classes the precision is very low.

Table 2
Results of Naive Bayes Algorithm

Class	*Naïve Bayes Cross validation*		*Naïve Bayes Percentage split*	
	Precision	*TP Rate*	*Precision*	*TP Rate*
Distinction	0	0	0	0
First	0.273	0.214	0.667	0.667
Second	0.837	0.917	0.903	0.966
Fail	0	0	0	0
Average	0.713	0.769	0.806	0.857

4.3. Results of *k*-NN Algorithm

Table 3
Results of *k*-NN Algorithm

Class	*k-NN-Cross validation*		*k-NN -Percentage split*	
	Precision	*TP Rate*	*Precision*	*TP Rate*
Distinction	0	0	0	0
First	0.400	0.429	0.333	0.333
Second	0.841	0.881	0.839	0.897
Fail	0	0	0	0
Average	0.733	0.769	0.724	0.771

k-NNalgorithm was analyzed based on the collected dataset and the results of the analysis are illustrated clearly in the Table 3. It is obvious from the Table 2 that when 10-fold cross-validation testing is done on *k*-NN algorithm it correctly classifies about 76.9% and when percentage split testing is done on *k*-NN algorithm it correctly classifies about 77.1%.

The results given in Table 3illustrates that for the Second class the TP Rate (88.1-89.7%) are high, for First class the TP rate is medium and for the classes Distinction and Fail the TP rate is very low. For the Second class the precision is high, for First class the precision is medium and for the classes Distinction and Fail the precision is very low.

4.4. Results of EM and *k*-Means clustering

Table 4 shows the clustering results for EM and *k*-Means algorithms. The EM algorithm correctly clusters about 74% and *k*-Means accurately clusters about 81%.

The results from Table 4 show that the accuracy is high for the First cluster in *k*-Means, whereas the third cluster has high accuracy rate in EM. It can be seen that the size of clusters are not same for both the algorithms. In *k*-Means clustering the third and fourth clusters have minimum accuracy. In EM algorithm the first and second clusters have less accuracy. The time taken by EM algorithm to build model is 0.3 second whereas the time taken by *k*-Means algorithm to build model is 0.1 second.

Table 4
Clustering results for the EM Algorithm and *k*-Means clustering

	Cluster #	*k-Means Cluster size*	*k-Means Accuracy in %*	*EM Cluster size*	*EM Accuracy in %*
1.	(Distinction)	34	92	27	79
2.	(First)	21	81	36	75
3.	(Second)	44	79	39	83
4.	(Fail)	5	76	3	81

4.5. Performance Comparison of Algorithms

In classification algorithms it is found that C4.5 provided good and consistent prediction results for Distinction Class, whereas Naïve Bayes and *k*-NN algorithms provided good prediction results for Second Class, and poor prediction results for Distinction and Fail classes.Overall considering average of all classes, C4.5 algorithm outperforms all other classification algorithms such as Naïve Bayes and *k*-NN algorithms with average highest accuracy of 62.7%.

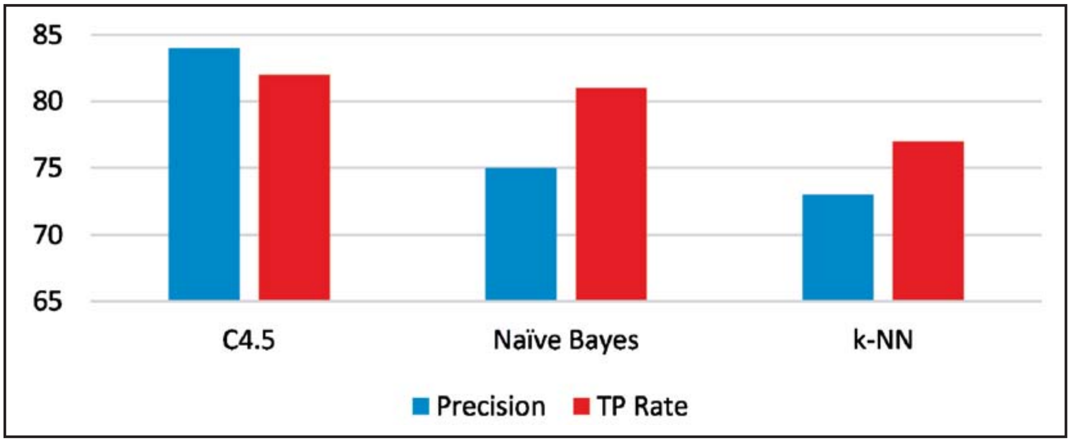

Figure 2: Classification Algorithms Comparison

In clustering technique, *k*-Means algorithm provides better accuracy for the student category Distinction and First Class, whereas EM algorithm provides better accuracy for the category Fail and Second Class. Overallconsidering average of all classes, *k*-Means algorithm performs better than EM algorithm with respect to clustering algorithms used in this work.

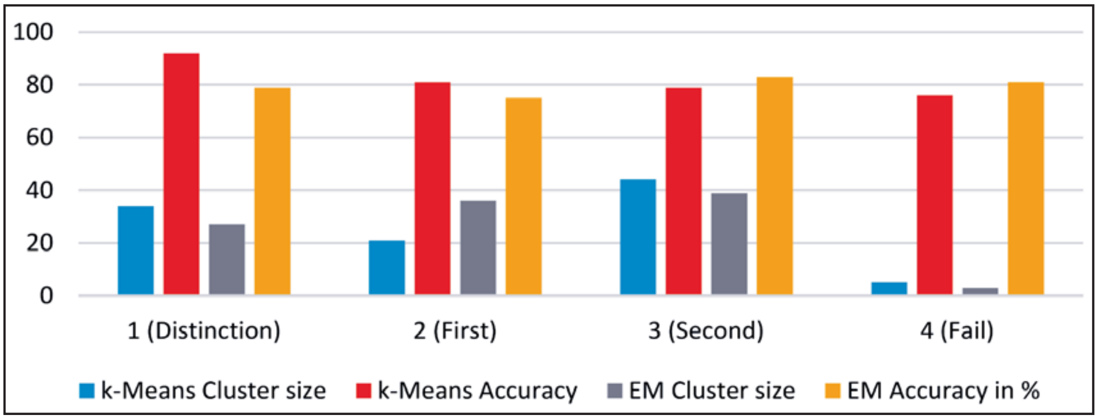

Figure 3: Clustering Algorithms Comparison

It can be observed from the table 5 that the time taken to provide results by both clustering and classification algorithms are almost similar with negligible difference. The overall accuracy presented in the Table 5 is calculated by taking average of all the precision fields of classes (Distinction, First, Second and Fail) in both cross validation and percentage split of each classification algorithm. For clustering algorithms the overall accuracy is calculated by taking average of accuracy of all the classes. It can be clearly seen from the table that the clustering algorithms outperforms classification algorithms in terms of accuracy. In particular k-Means algorithm has the highest accuracy of 82% and k-NN algorithm has the lowest accuracy of 38.7%.

Table 5
Performance Comparison of Algorithms

S.No	*Algorithm*	*OverallAccuracy in %*	*Time Taken in seconds*
1.	C4.5	62.7	0.28
2.	Naïve Bayes	41.9	0.24
3.	*k*-NN	38.7	0.22
4.	EM	79.5	0.29
5.	*k*-Means	82	0.30

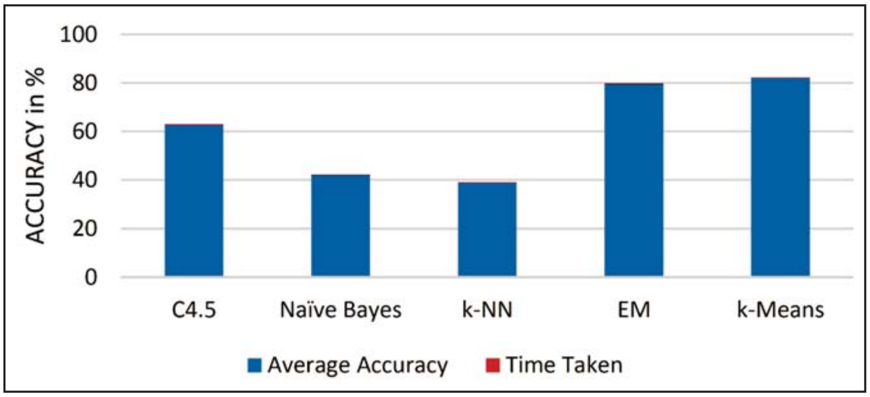

Figure 4: Performance Comparison of Classification and Clustering Algorithms

5. CONCLUSION

In this research work, classification and clustering algorithms were examined and compared based on the students' data set via its attribute values. The results show that C4.5 algorithm outperforms from all other classification algorithms used in this work with average highest accuracy of 62.7%. The performance of *k*-Means algorithm is well compared with EM clustering algorithm which have overall accuracy of 82%. Comparatively the performance of clustering algorithms is well in the prediction of student performance than the classification algorithms. The accuracy rate also indicates that the same results obtained by this experimental work for the chosen data set. This research can be prolonged further by adding different students' attributes that have impact on academic performance, which are not used in this data set. Also, the size of data set may be increased such that the student data from all colleges in a whole district is used instead of data from few colleges in a particular region to increase the accuracy of results.

REFERENCES

[1] Archana T and Usha Devi Gandhi, *"Prediction of Student Performance in Educational Data Mining - A Survey"*, International Journal of Pharmacy & Technology, Vol. 8, No. 3, pp. 17757-17763, 2016.

[2] Agrawal Bhavana D and Gurav Bharti B,*"Review on Data Mining Techniques used For Educational System"*, International Journal of Emerging Technology and Advanced Engineering. Vol. 4, No.11, pp. 325–329, 2014.

[3] Ajay Kumar Pal and Saurabh Pal,*"Analysis and Mining of Educational Data for Predicting the performance of Students"*, International Journal of Electronics Communication and Computer Engineering. Vol. 4, No.5, pp. 1560–1565, 2013.

[4] Ajith P, Tejaswi B, Sai MSS,*"Rule Mining Framework for Students Performance Evaluation"*,International Journal of Soft Computing and Engineering. Vol. 2, No. 6, pp. 201–206, 2013.

[5] Anju Rathee and Robin Prakash Mathur,*"Survey on Decision Tree Classification algorithm for the Evaluation of Student Performance"*, International Journal of computers & Technology. Vol. 4, No. 2, pp. 244–247, 2013.

[6] Anuradha C and Velmurugan T, *"A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance"*, Indian Journal of Science and Technology, Vol 8, pp. 1-12, 2015.

[7] Baker RSJD,*"Data mining for education"*, International encyclopedia of education. Vol.7, pp. 112–118, 2010.

[8] Dinesh Kumar A, Radhika V, *"A Survey on Predicting Student Performance"*,International Journal of Computer Science and Information Technologies. Vol. 5, No. 5, pp. 6147–6149, 2014.

[9] Dorina Kabakchieva,*"Predicting Student Performance by using Data mining Methods for Classification"*, Bulgarian Academy of Science, Cybernetics and Information Technologies. Vol. 13, No. 1, pp. 61–72, 2013.

[10] Hashmia Hamsa, Simi Indiradevi and Jubilant J Kizhakketthottam, *"Student Academic Performance prediction model using decision tree and fuzzy genetic algorithm"*, Elseiver, Procedia Technology, Vol. 25, pp. 326-332, 2016.

[11] Mojisola G. Asogbon, Oluwarotimi W. Samuel, Mumini O. Omisore, and Bolanle A. Ojokoh, *"A Multi-class Support Vector Machine Approach for Students Academic Performance Prediction"*, International Journal of Multidisciplinary and Current Research, Vol. 4, pp. 210-215, 2016.

[12] Ogunde AO, Ajibade DA,*"A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm"*, Journal of Computer Science and Information Technology. Vol. 2, No. 1, pp. 21–46, 2014.

[13] Romero C and Ventura S,*"Educational data mining: A survey from 1995 to 2005"*, Expert systems with applications. Vol. 33, No. 1, pp. 135–146, 2007.

[14] Shanmuga Priya K and Senthil Kumar AV, *"Improving the student's performance using Educational data mining"*, International Journal of Advanced Networking and Application. Vol. 4, No. 4, pp. 1680–1685, 2013.

[15] Suman and Pooja Mittal,*"A Comparative Study on Role of Data Mining Techniques in Education: A Review"*, International Journal of Emerging Trends & Technology in Computer Science. Vol. 3, No. 3, pp. 65-69, 2014.

[16] Surjeet KumarYadav, Brijesh Bharadwaj, and Saurabh Pal. *"Data mining applications: A comparative study for predicting student's performance."*, International Journal of Innovative Technology & Creative Engineering, Vol. 1, No. 12, pp. 13-19, 2012.

[17] Trivedi A,*"Evaluation of Student Classification Based On Decision Tree"*, Int Journal of Advanced Research in Computer Science and Software Engineering. Vol. 4, No. 2, pp. 111–112, 2014.

[18] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand and Dan Steinberg. *"Top 10 Data mining Algorithms"*, Springer, Knowledge Information System, Vol. 14, pp. 1–37, 2013.