International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Preprocessing and Feature Extraction Process in Predicting Students Performance using Clustering Technique

K. Govindasamy, T. Velmurugan

Abstract: ***Analyzing students' performance patterns using various statistical techniques always remain a critical task for many researches. This research paper focuses on basic preprocessing and extraction techniques followed before analyzing the student's performance in academics. The collected information from various colleges has to be cleaning process for removing irrelevancy in data. The repeated records and unfilled data are removed in cleaning stage. The features are to be extracted from the preprocessed record. The research mainly concentrated on analyzing students' performance from collected information. The extraction process implemented in this research work carefully examined in extracting the necessary features for analyzing process.***

Index Terms: Educational Data Mining, Feature extraction, Clustering, E-Learning.

I. INTRODUCTION

Identifying students' performance and creating necessary methodology for teaching based on students grasping mechanism remains a toughest job for many academicians. The area of educational data mining plays an important role in examining many important researches in the field of education. The Educational data mining helps the researches and academicians to find out the behavioral patterns followed by students in learning process and examines the teaching methodology implemented in teaching process.

The educational data mining process followed in examining the performance of students and mechanism followed in implementing the teaching aides can be clearly studied in this research paper. The basic strategy followed in educational data mining plays a unique responsibility in examining student's behavioral patterns followed in learning process. The exact measurement criteria for factors influencing the academic performance of students remains a toughest job for many research area. The records collected from academic colleges may contain redundancy and irrelevancy, which may effects the performance of analyzing. The initial step in educational data mining process removes such irrelevancies and improves the accuracy of analyzing stage. The second stage of every mining process continues from feature extraction process, which is mostly avoided in many data mining process. Few researchers mainly focus on feature extraction process while dealing with educational data mining for analyzing students' academic performance, which may eventually improves the efficiency of proposed method in analyzing stage. The basic idea behind the feature extraction process is data mining, pattern reorganization and machine learning concepts, which remains a major area for researches to focus for analyzing. The unpredictable data from the collected record set have to be removed in feature extraction process. The feature extraction process also helps in improving the accuracy in proposed model and improves the better understandability in reading the outcome. The feature extraction process also plays a major role in improving the efficiency in predicting academic performance of the students. The process is broadly divided into three major models such as embedded, filter and wrapper. Basic idea behind the embedded model is to propose some learning algorithms for training and associating the record set. Filtering techniques usually carried out in the preprocessing stage of every data mining stages. Finally wrapper approaches are used in educational data mining for evaluating the necessary features using learning algorithms.

The feature selection process followed in many researches uses many classification algorithms for analyzing the academic performance of college students. The different patterns are identified and taken for the analyzing stage. The accuracy based on Recall, Precision, F- Measure and correctly classified values are identified for students' academic performance prediction.

II. LITERATURE REVIEW

Educational data mining is a major area mainly focus for predicting the students' academic performance in learning aspects. The extraction of necessary information from collected educational record set and analyzing the information are known as educational data mining. The evaluation of research field and recent improvement in educational filed leads to produce a challenging task in evaluating students' performance in academics. The necessary steps for educational data mining starts from preprocessing following feature extraction process and ends with analyzing stage with necessary clustering and classification algorithms. Students information collected from various institutions are useful in examining the performance based on living location and basic educational background.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Mr. K. Govindasamy*, Research Scholar, Department of Computer Science, School of Computing Science, VISTAS, Chennai, Tamil Nadu, India.
Dr. T. Velmurugan, PG and Research Department of Computer Science, D.G.Vaishanav College, Chennai, Tamil Nadu, India.

The strategy followed in educational data mining starts from questionaries' setting and collecting information from various colleges. The process also useful to the academicians in studying the behavioral patterns followed by the students from various locations, either rural of urban.

Classification algorithm K Means calculates initial centroids values randomly, which makes the iteration count reduction and significantly improves the time spent on evaluation process [1]. When common record set of data are to be trained, different attributes can involve in clustering the vertically presented data [2]. The classification algorithms and clustering algorithms are examined for fixing many problems in data mining techniques [3]. The record set collected from industries are analyzed with various classification algorithmic techniques such as Fuzzy C- Means and K- Means, which also examines the working of various clustering algorithms and tested with accuracy. The K Mean clustering concept of evaluating the students' academic performance were involved in research work and produced necessary outcome based on accuracy [4]. The research work also examines the factors that influencing academic work such as midterm exams, final exams, class quizzes and assignments. The work also suggests that all information's sharing between class advisor and students should be in proper manner to avoid unnecessary fallout in information lose.

The study focus the instructors to turn their attention towards the dropouts from institutions and eventually increase the qualification of the students [5]. Smooth Support Vector Machine is an application introduced, which works with the principles of classification algorithms and kernel K means algorithmic techniques [6]. The resultant information produces a suggestion to academicians to form a psychometric factors analyzing committee for students, as well as instructors. The measures taken for predicting students' academic performance improves the overall performance of the students in using K Mean clustering algorithm [7]. The monitoring system for examining the higher education students are also improved and tested with academic performance. The focus of the research work also extended to systematically review the different clustering techniques involved in educational data mining in predicting the performance of the students.

The classification techniques involved in the research work finds relevancy in testing data collected from record set [8]. The research work also focus in examining many research problems in mining tools and algorithms in classification, clustering, association, neural networks etc., and uses java script coding in producing results[9]. The paper discusses briefly about communications between different analyzing tools presented in educational data mining. Association rule based clustering and classification based clustering techniques are discussed with various problems solutions [10]. The simulation of the results are carried out with the usage of WEKA tool, which makes feature extraction process very easy and carrying the dataset to next analyzing process. The training data are used for the research work and resultant are showed with the mean and standard division. The proper working and strategy followed in WEKA analyzing tools are clearly shown with examples [11][20]. The classification facility of data are also examined with necessary reports. The discussion about Bulgarian educational sector and implementing the prediction process of pre university students characteristics are examined and classified with the necessary results [11, 12, 13].

III. RESEARCH METHODOLOGY

The process of preprocessing and feature extraction process followed in this research work can be explained with the architecture diagram figure 1.

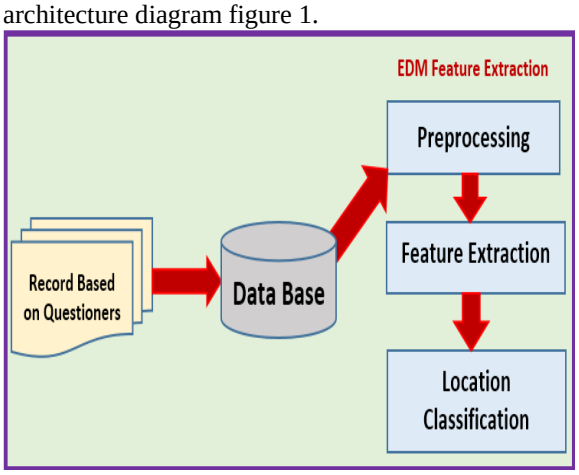

Figure 1. Architecture Diagram for Feature Extraction

The research paper discuss about the feature selection process and students location classification based on certain rules. This research papers ultimate goal is to bring a clear idea about the different classification algorithms used for feature extraction or selection process deployed in educational data mining process. The process followed in this research work starts from collecting questioners from various colleges. The questioners are framed not only for evaluating the performance of the students and supervisor, it also examines the location based students understandability. The collected information contains various information of students such as family size, family type, family annual income, parent's qualification, living locations, test marks based on E learning tools, Average among various tests conducted based on E Learning methodology and finally understandability about E Learning method [14,15,16]. The information's collected from various colleges are stores in database for further inspection. The information collected are used in preprocessing stage for removing irrelevancy in record, then carried out to feature extraction process as in figure 1. The necessary features are collected for final evaluation process and classification are carried based on the location of the students.

A. *DATA DISCRETION*

The information about students from various arts colleges from different locations are collected. The necessary questioner are set for evaluating the students learning capability and understandability. The questioners consists of different questions for evaluating the students' performance and usage of learning methodology. The process also added some questions related to supervisor needed for learning methodology.

The information consist of 4391 records of students who learn through E Learning and attended tests based on academic performance. The students record also examines whether the students living locality plays a major role in learning understandability and usability of learning methodology [17,18,19].

B. *Experimental Tools and setup*

The collected information can be used for many educational based research. The initial stage used for preprocessing and feature selection are implementing the collected dataset into the WEKA tool for classification process. The irrelevancy in collected record are removed through certain algorithmic steps implemented. Though the WEKA tool is open source and easy to handle classification process, beginning process initiated with WEKA. The process after feature selection and comparison of various classification algorithm may starts for analyzing stage in MATLAB.

C. *Steps in pre-processing and feature selection*

The necessary steps followed in preprocessing and feature selection process are explained bellow with steps.

Read Entries in SR
For each Entry in SR
Read fields (SP, LL)
If SP = 'good' and LL= 'Village'
Then
Get SR1= SR, R1 and CSCS
If Get SR1= {ST Name, ST Sex, FQ, MQ, FT, B and S}
Then
Remove SR1
Save R1 and SR1
End if
Else
Next Entry
End if.

The Steps deployed above helps in removing the irrelevant record from collected record. The name of the student and student's genre details are not need for this research work. The qualification of father and mother also not determines the learning method followed by the student in E Learning tool, so the father and mother qualifications are also removed from collected database. Some of the information's collected from the students contains duplicates and counts about the blood relations, which is also not needed for this research work. So the steps deployed before feature extraction process removes the irrelevant data from collected record.

D. *Feature extraction process using algorithms and classification process*

The categorized record are arranged accordingly and compared with various classification algorithms and results for accuracy is produces. The research work carried out uses different evaluating process such as Principle components, Chi Squared, Filtered attribute Evaluation, Correlation Based Feature Selection subset Evaluation, Gain attribute Evaluation and Relief attribute Evaluation. The clustering algorithms used for classification process are K- means, K-Mediods, Fuzzy C Mean, Expectation- Maximization and Random Swap Expectation Maximization.

IV. RESULTS AND DISCUSSIONS

The research work carried out in this paper demonstrates the feature extraction process along with some clustering techniques for classification process. The efficiency of the clustering algorithm and effectiveness of the algorithm are calculated from Recall, Precision, F- Measure and correctly classified values accuracy. The extracted features are evaluated through some of existing algorithms and results are shown. The features are trained separately with each and every clustering algorithm for classification purpose.

The formulation used for calculating the Recall, Precision and F – measures are calculated using the true positive and false positive error rate. The accuracy calculation is also performed using the true negative and false negative error rate in evaluation process.

$$Recall = \frac{TP}{TP + FN} \quad (1)$$

Were TP determines True positive error rate and FN determines false negative error rate.

The instants correctly classified in given dataset can be shown with TP and un-correctly classified item set can be determined with FN as in equation 1.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Were TP determines True positive error rate and FP determines false positive error rate. The results carries the information about positive error rate which is unnecessary placed in visualization process as in the equation 2.

F- Measures actually determines the harmonic average of recall and precision. Usually F- Measures are calculated with the following equation 3.

$$F - Measures = \frac{2(Recall * Precision)}{Recall + Recision} \quad (3)$$

A. *Principle component analysis*

The analyzing techniques used for examining the clustering algorithms Principle component analysis is one of the statistical outcome, which uses an quadratic transformation methodology for converting a set of annotations correlated variables into a set of linearly uncorrelated values, basically known as Principle component.

Table 1. Principle Component Analysis

Clustering Algorithms	Recall	Precision	F-Measure
K- means	0.550	0.58	0.546
K- Mediods	0.445	0.446	0.446
Fuzzy C Mean	0.695	0.702	0.696

Expectation Maximization	0.577	0.586	0.577
Random Swap Expectation Maximization	0.67	0.660	0.660

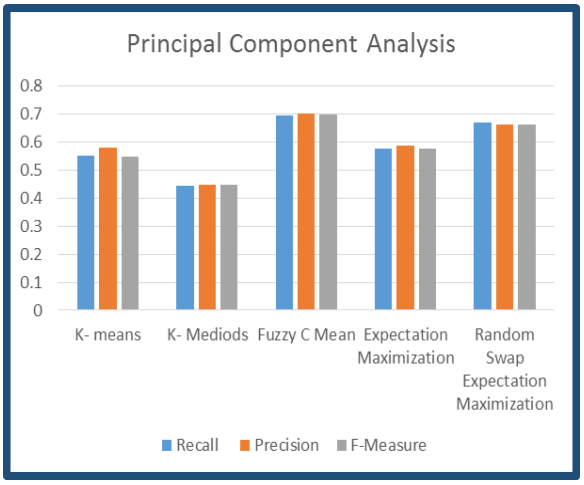

Figure 2. Principle Component Analysis

B. *Chi Squared Test*

The distribution of sampling remains a statistical theory test, chi-squared distribution hold the position of true in null theory. Basically chi-squared testing is used for frequencies checking among two or more categories of significantly different variants. The analyzing tool classifies observations into various classes, which is known to be null hypothesis in some theory. The ultimate goal of this test is to go through whether the observations are correctly correlated to null hypothesis or not. This test often known as sum of squared error test for sample variance classification process.

Table 2. Chi Squared Test

Clustering Algorithms	Recall	Precision	F-Measure
K- means	0.695	0.699	0.693
K- Mediods	0.584	0.612	0.572
Fuzzy C Mean	0.716	0.719	0.717
Expectation Maximization	0.675	0.675	0.675
Random Swap Expectation Maximization	0.654	0.652	0.652

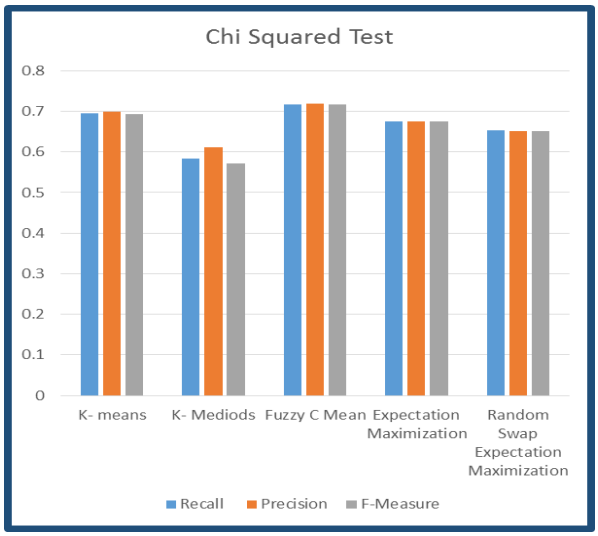

Figure 3. Chi Squared Test

C. *Filtered attribute Evaluation*

The features gathered from the result of preprocessing method should undergone the process of filtering attributed evolutionary process for inspecting the basic accuracy of each and every classified variables.

Table 3. Filtered Attribute Evaluation

Clustering Algorithms	Recall	Precision	F-Measure
K- means	0.689	0.692	0.689
K- Mediods	0.584	0.612	0.572
Fuzzy C Mean	0.737	0.742	0.738
Expectation Maximization	0.730	0.739	0.74
Random Swap Expectation Maximization	0.654	0.652	0.652

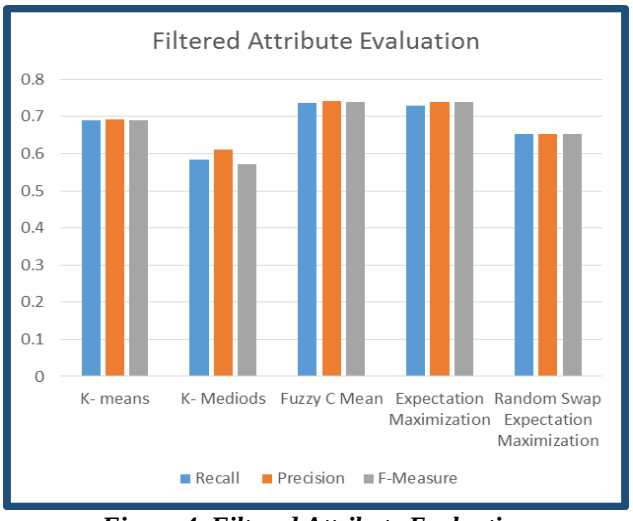

Figure 4. Filtered Attribute Evaluation

D. *Correlation-Based Feature Selection subset Evaluation*

The truly classified features can be evaluated perfectly with the usage of correlation based selection functionality. If the evaluated class as a unique identification compared with other evaluated classes are known to be correctly correlated features. The adaption of this theory leads to a production of highly correlated features. This technique for classification algorithms finds a best solution to training data for suitable measures for improving the training data.

Table 4. Correlation-Based Feature Selection

Clustering Algorithms	Recall	Precision	F-Measure
K- means	0.67	0.660	0.659
K- Mediods	0.584	0.612	0.572
Fuzzy C Mean	0.633	0.65	0.634
Expectation Maximization	0.619	0.628	0.622
Random Swap Expectation Maximization	0.67	0.668	0.656

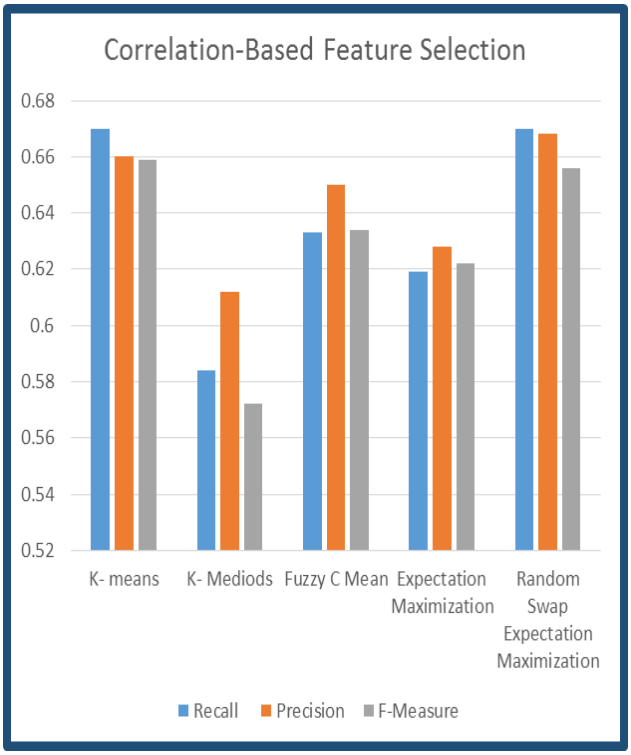

Figure 5 Correlation-Based Feature Selection

E. *Gain attribute Evaluation*

The Gain attribute selection method works with the information worthiness of attributes with respect to that of the observed class. The attribute evaluation process takes the correctly classified variant and fits the best attributes into the observed class. The examination process for calculating the perfectly classified data item is completed for identifying the mismatched data record set.

Table 5. Gain attribute Evaluation

Clustering Algorithms	Recall	Precision	F-Measure
K- means	0.675	0.690	0.676
K- Mediods	0.584	0.612	0.572
Fuzzy C Mean	0.709	0.72	0.710
Expectation Maximization	0.689	0.716	0.688
Random Swap Expectation Maximization	0.654	0.652	0.652

F. *Relief attribute Evaluation*

Basically Relief is an algorithm used to take a filter method approach for feature selection process that is very sensitive for interacting the features of record. The binary problems in discrete method and numerical features can be solved with the usage of Relief attribute evaluation process. Relief also measures the top ranked features and arranges accordingly for future references. The weightage produced with the effort of ranking is mostly used for down streaming modeling. The identification of best scoring feature techniques is carried out with the means of nearest neighbor algorithmic technique.

Table 6. Relief attribute Evaluation

Clustering Algorithms	Recall	Precision	F-Measure
K- means	0.709	0.714	0.709
K- Mediods	0.584	0.612	0.572
Fuzzy C Mean	0.76	0.757	0.874
Expectation Maximization	0.67	0.666	0.658
Random Swap Expectation Maximization	0.654	0.652	0.652

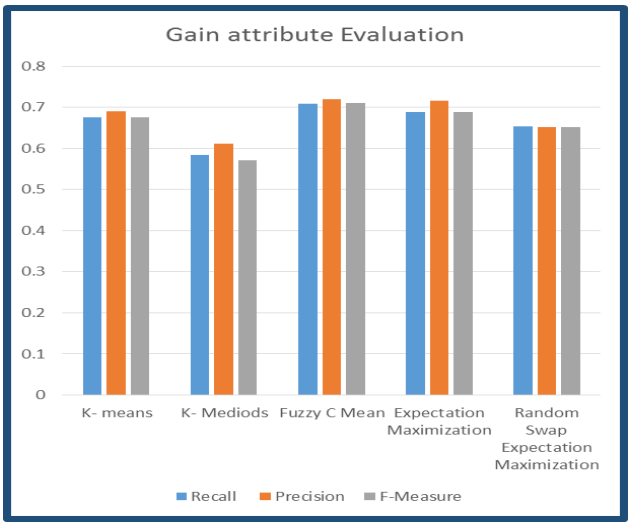

Figure 6. Gain attribute Evaluation

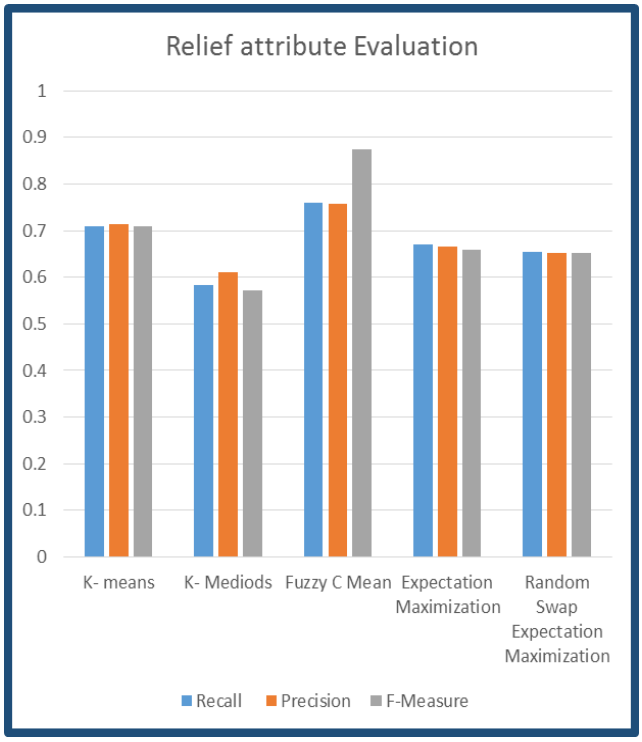

Figure 7. Relief attribute Evaluation

G. *Performance Evaluation*

The performance evolutionary process in every research work carries values to proposed algorithmic techniques. The performance of each and every algorithm is measured with constrains of mean and variance between the algorithms. Table 7 shows the difference between each and every algorithms in correctly classified instance. The process is taken to the next level for testing the accuracy of the algorithms in evaluation process.

Table 7. Correctly Classified Instance in Percentage

Feature Selection Algorithms	Correctly classified instance in %					Mean	Variance
	K-means	K-Mediods	Fuzzy C Mean	EM	RSEM		
Principle component	55.87	45.44	70.44	58.64	66.97	59.47	.00802 9786
Chi Squared Test	70.44	59.33	72.54	68.36	66.28	67.40	.00370 7004
Filtered attribute Evaluation	69.76	59.33	74.61	73.92	66.28	68.78	.00390 2547
Correlation-Based Feature	66.97	59.33	64.19	62.81	66.97	64.06	.00384 4472
Gain attribute Evaluation	68.37	59.33	71.83	69.75	66.28	67.13	.00346 0798
Relief attribute Evaluation	70.83	59.33	76.00	66.97	66.28	67.88	.00401 6358

The accuracy of each and every tested algorithm is shown in below table 8. The adapted technique random swap expectation maximization algorithm produces more accurate result compared with other algorithms.

Table 8. Accuracy of Classification Algorithms

Feature Clustering Algorithms	Accuracy
K- means	59.7130
K- Mediods	60.0547
Fuzzy C Mean	57.9595
Expectation Maximization	81.0977
Random Swap Expectation Maximization	82.1087

V. CONCLUSION

The research work pays a unique attention in evaluating the student's behavioral patterns in following the methodology of using the learning aides like E Learning tools. The collected records from various colleges are used for examining the efficiency of E Learning tools and efficiency of the students in understanding the base of the learning tool. The research work pays more care in preprocessing technique in removing the irrelevant data from the collected record and feature selection process carried out.

This research work also compares various classification algorithms and finds the best solution in evaluation process.

REFERENCES

1. Azhar Rauf, Sheeba,"Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, Vol. 12 (7), Pp. 959-963, 2012.
2. Jaideep Vaidya, "Privacy Preserving K-Means Clustering over Vertically Partitioned Data", In proceeding of SIGKDD '03, Washington, DC, USA, August 24-27, 2003.
3. N. Sivaram, "Applicability of Clustering and Classification Algorithms for Recruitment Data Mining", International Journal of Computer Applications, Vol. 4(5), July 2010.
4. Md. Hedayetul Islam Shovon, "Prediction of Student Academic Performance by an Application of K-Means Clustering Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2(7), July 2012.
5. A.S. . Arunachalam., and T. Velmurugan, "A Survey on Educational Data Mining Tools and Techniques", International Journal of Data Mining Techniques and Applications, Vol. 5 (2), PP. 167-171, 2016.
6. Sajadin Sembiring, "Prediction of Student Academic Performance by an Application of Data Mining Techniques", International Conference on Management and Artificial Intelligence IPEDR, IACSIT Press, Vol. 6, 2011.
7. Oyelade, O. J, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", (IJCSIS) International Journal of Computer Science and Information Security, Vol.7, 2010.
8. Trilok Chand Sharma," WEKA Approach for Comparative Study of Classification Algorithm", (IJARCCE) International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2(4), April 2013.
9. A.S.Arunachalam and K.Rajeswari , "An Inclusive Survey of Student Performance With Various Data Mining Methods "International Journal of Engineering and Technology (IJET) vol.7,No.2.33, pp.522-525,2018.
10. Shilpa Dhanjibhai Serasiya," Simulation of Various Classifications Results using WEKA", International Journal of Recent Technology and Engineering (IJRTE), Vol. 1(3), August 2012.
11. Swasti Singhal," A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE),Vol. 2(6), May 2013.
12. D.Kabakchieva, "Analyzing University Data for Determining Student Profiles and Predicting Performance", Cybernetics and Information Technologies, Vol.1(3), March 2013.
13. A.S. Arunachalam., and T. Velmurugan., "Measures for predicting success factors of ELearning in Educational institutions", International Journal of Pure and Applied Mathematics, Vol. 118 (18), PP: 3673-3679, 2018.
14. T. Velmurugan and C. Anuradha, "Performance Evaluation of Feature Selection Algorithms in Educational Data Mining," Performance Evaluation, vol. 5, 2016.
15. Aha, David W., and Richard L. Bankert. "Feature selection for case-based classification of cloud types: An empirical comparison." In Proceedings of the AAAI-94 workshop on Case-Based Reasoning, vol. 106, p. 112. 1994.
16. A. S. Arunachalam., and T. Velmurugan., "Analyzing Student Performance using Evolutionary Artificial Neural Network Algorithm", International Journal of Engineering and Technology, Vol. 7(2.26), PP: 67-73, 2018.
17. Kabakchieva Dorina, "Predicting student performance by using data mining methods for classification", Cybernetics and Information Technologies, Vol. 13, No. 1, pp. 61-72, 2013.
18. Kotsiantis .S. , C. Pierrakeas, and P. Pintelas, "Predicting students' performance in distance learning using machine learning techniques", Applied Artificial Intelligence , Vol. 18, No. 5, pp. 411-426, 2004.
19. Pandey.U. K and S. Pal, "Data mining: A prediction of performer or underperformer using classification", International Journal of Computer Science and Information Technology, Vol. 2, No. 2, pp. 686-690, 2011.
20. Dharmarajan, K., and M. A. Dorairangaswamy. "Web usage mining: improve the user navigation pattern using fp-growth algorithm." Elysium journal of engineering research and management (EJERM) 3.4 (2016).