Classification, Clustering and Association Rule Mining in Educational Datasets Using Data Mining Tools: A Case Study

Sadiq Hussain[1](✉), Rasha Atallah[2], Amirrudin Kamsin[3], and Jiten Hazarika[4]

[1] Dibrugarh University, Assam, India
sadiqdu@rediffmail.com
[2] Faculty of Computer Science and IT,
University of Malaya, Kuala Lumpur, Malaysia
[3] Department of Computer System and Technology,
University of Malaya, Kuala Lumpur, Malaysia
[4] Department of Statistics, Dibrugarh University, Assam, India

Abstract. Educational Data Mining is an emerging field in the data mining domain. In this competitive world scenario, the quality of education needs to improve. Unfortunately most of the students' data are becoming data tombs for not analyzing the hidden knowledge. The educational data mining tries to uncover the hidden knowledge by discovering relationships between student learning characteristics and behavior. With this educational data modeling, the educators may plan for future learning pedagogy to support the student's learning style. This knowledge may be applied by the academic planners to improve the quality of education and decrease the failure rate. In this paper, we had collected real dataset containing 666 instances with 11 attributes. The data is from the Common Entrance Examination (CEE) data of a particular year for admission to medical colleges of Assam, India conducted by Dibrugarh University. We tried to find out the association rules using the data. Various clustering and classification methods were also used to compare the suitable one for the dataset. The data mining tools applied in the educational data were Orange, Weka and R Studio.

Keywords: Classification · Clustering · Association rule mining
Educational data mining · Data mining tools

1 Introduction

Educational Data Mining is a sub area of Data Mining Domain. This new area has immense potential to mine various aspects for betterment of students as well as in decision making by the authorities of educational institutes. There is a collection of student data year after year without much utility. These data may be classified or clustered to distinguish between excellent, good and academically poor students. The excellent and good students may be encouraged to perform better, while the academically poor students may be given remedial classes and extra attention to prevent

R. Silhavy (Ed.): CSOC 2018, AISC 765, pp. 196–211, 2019.
https://doi.org/10.1007/978-3-319-91192-2_21

dropouts and to enrich the quality of education across higher educational institutes across the globe. So, the aim of Educational Data Mining is to extract some hidden yet useful information from the large educational institutes' datasets which varies from the schools to university levels [1]. In case of unsupervised learning or descriptive analysis, the clustering may be applied. For supervised learning where the class labels are known, classification techniques may be used. Using the student datasets, various aspects may be predicted. The authorities may use these predictions for quality enhancements of the institutes. The association rule mining may be applied to discover some of the interesting relation among the attributes of the datasets [2].

There are various clustering techniques used in the field of knowledge extraction. The clustering techniques used by in this paper are K-means, Hierarchical and Partitioning Around Medoids. For classification, neural networks, naïve bayes and decision tree methods were used. For association rule mining, Apriori algorithm was used. A comparison was also made among these techniques. The data mining tools used for this work are Weka, Orange and R. The reasons for selecting these tools are that all are open source, easy to use and platform independent data mining tools. All have scripting interface. They are good at data visualization and analysis [3]. The Government of Assam authorized Dibrugarh University which is the easternmost University of India, conducted common entrance examinations (CEE) for a particular period for admission of students to medical colleges of Assam. The collected data were of students who came for counseling cum admission into medical colleges of Assam in the year 2013. The data were collected as per our requirement by one of the authors of this paper through direct personal interview at the time of admission. Altogether, data with 12 attributes were collected. The CEE rank and CEE percentage is ambiguous data, as if the CEE percentage is high so is the rank. So, only the CEE percentage termed as performance of the candidate is used as response variable whereas the rank of the candidate was not used. The performance is converted to categorical data from the real data. Not only performance but all the explanatory variables data were converted to categorical data in case of classification and association. For Clustering, the data was not converted and only selected numerical attributes were used to find some meaningful clusters.

The rest of the paper is organized as follows: Sect. 2 present Literature Review, Sect. 3 describes Methodology, Sect. 4 presents Experiments and Results and the Sect. 5 describes the Conclusion and Future Work.

2 Literature Review

This section explains various works done by the researchers on educational datasets using clustering and classification techniques.

2.1 Clustering Techniques Used by the Researchers on Educational Datasets

DeFreitas et al. [4] made a comparative analysis of clustering algorithms. The dataset used was Learning Management System log data. They compared K-means, DBSCAN and BIRCH methods to select the most suited clustering methods for educational data sets.

Based on the silhouette coefficient score, it was found that K-means performs better than the other two algorithms. It is also better from the point of view of distribution of clusters.

Dutt et al. [5] reviews how the large data that is generated by the educational institutes may be utilized properly. Most of such data remains unused and the useful information is not extracted. This application of Clustering, Classification, prediction modeling is very low in case of educational setup. The authors surveys different clustering methods applied in the field of Educational Data Mining (EDM).

Nagy et al. [6] proposed a student advisory framework which offers consultation to first year students of Cairo Higher Institute for Engineering. This intelligent system is based on clustering and classification techniques. This framework aims for decreasing the rate of failures which was very high otherwise.

Oyelade et al. [7] proposed a deterministic educational model to evaluate the performance of student academic achievement. The data was collected from a private institute based at Nigeria. The students' results were clustered by using K-means algorithm. The authors claimed that these mining results would be helpful in effective decision of higher authorities of the institutes.

2.2 Classification Techniques Used by the Researchers on Educational Datasets

Almarabeh [8] used Weka tool to compare five different classifier available in the toolkit for the University student dataset. The classifiers were Bayes Network, J48, ID3, Neural Network and Naïve Bayes. It was found that Bayes Network classifier performs better than the others by using different evaluation measures.

Bhardwaj et al. [1] used predictive model for classification of good learners and slow learners. The dataset used from different affiliating institutes under Dr. R.M.L. Awadh University, Faizabad, India and it contains 300 instances. They had used 17attributes for the classification task. The high attributes were obtained from the dataset to predict the academic performance of the students.

Yadav et al. [2] applied decision tree algorithms for prediction of engineering students' results. The algorithms used were C4.5, ID3 and CART. The results of the analysis helped the academically weaker students to perform better. The data was collected from VBS Purbanchal University with 90 records for the session 2010. C4.5 decision tree algorithm performs better than other two with 67.78% accuracy.

3 Methodology

This section describes the about various techniques used by the data miners for educational datasets, e.g. classification, clustering, association rule mining, classification errors, evaluation methods, and data mining open source tools.

3.1 Unsupervised Learning and Clustering Methods

Clustering is grouping or partitioning data into some subsets. The clustering is made in such a way that the object belong to one cluster similar to other objects in that cluster and the dissimilar clusters belong to other clusters. Clustering, an unsupervised learning technique may be classified depending on the data types used, similarity measures and the theory involved for defining the cluster [9]. There are two types of clustering [10], hard and soft. In the hard clustering, each data point belongs to one cluster or not, but in case of soft clustering each data point may belong to one or more clusters.

In this paper, we used K-means clustering, Hierarchical Clustering, Partitioning Around Medoids. K-means algorithm tries to group n items into some user defined subgroups k where k should be less than or equal to n. The grouping is done in iterative manner by minimizing the sum of squared distances and centroids of the items until there are no longer any changes in the structure or a threshold is reached. So, Partitioning Around Medoids (PAM) technique needs to define the number of clusters in advance like k-means algorithm. PAM is based on medoids that are centrally located in clusters [11]. PAM is another partitioning algorithm like k-means whereas the later performs better in wide variety of datasets [12]. Hierarchical Clustering technique is used to group items into a tree of clusters. This method has two types; divisive and agglomerative. In divisive clustering, the single cluster is iteratively splitted to make more clusters where as the clusters are merged into a single cluster in case of agglomerative. The merging or splitting is stopped until a certain condition is reached [13].

We used one neural reduction unsupervised technique called Self Organizing Map (SOM) for visualizing its data. SOM is useful for visualizing high dimensional data for low dimensional view and is also useful for dimensionality reduction [14]. The clustering methods were performed by Orange and R Studio.

3.2 Supervised Learning and Classification Methods

In this paper, we used three classification techniques; decision tree (J48 in Weka), neural network (Multilayer Perceptron in Weka) and Naïve Bayes. Various researchers have used these classification techniques to classify educational datasets [1, 15, 16]. All the classifications were performed by the WEKA software. The decision tree is a predictive supervised modeling technique where all the class labels are known. It is used for predicting and analyzing the data. This technique builds a tree like top down model. The C4.5 algorithm (J48 in case of WEKA) implements decision tree algorithm by using pruning which means that the node may be deleted if it does not add any significance. In WEKA, the neural network classifier is implemented by using Multilayer Perceptron algorithm. Neural network is emulation of human neuron system. In pattern recognition scenario, neural network is a very popular classifier [17]. It maps the input data to acceptable output, and feeds forward in nature. It has many layers including hidden layers which generally use sigmoid activation functions. Naïve Bayes is predictive classification technique which is based on bayes' probability theory with strong assumptions among the fields. Naïve Bayes classifier is suitable for small datasets and it is simple and easy to interpret. The model can represent only linear class boundaries with categorical data, so the representation is difficult compared to decision tree classification algorithm in such cases [18].

3.3 Association Rule Mining

The occurrence of an item may be predicted by using the occurrence of other items in the transactions. The rules that define such transactions in the form X → Y are called association rules [19]. Support is termed as frequency of occurrence of set of items or itemset, while confidence is fraction of transaction that contains the itemset. The frequent itemset is the itemset whose support is greater than the minimum support threshold. While generating association rule, minimum support threshold, size, dimension and average transaction width are the factors that affects complexity in mining. Agrawal et al. [20] proposed the Apriori Algorithm which is based on the principle that any subset of a frequent itemset is frequent, and any superset of infrequent itemset should not tested or generated.

3.4 Classification Errors

Mean Absolute Error (MAE) is the average of absolute differences of the predicted and observed samples [21]. Root Mean Squared Error (RMSE) is square root of the average of the difference between the forecast and observed values. Both MAE and RMSE vary from zero to infinity. The lower the value of both, the better is the results. Since MAE do not use the square, so it is more robust to outliers. RMSE is more useful when the large errors are not expected.

$$\mathrm{MAE} = \frac{1}{n}\sum\nolimits_{i=1}^{\mathrm{n}} |\mathrm{X_i} - \mathrm{X}| \tag{1}$$

where n = the number of errors, $|\mathrm{x_i} - \mathrm{x}|$ = the absolute errors.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{obs,i} - X_{model,i}\right)^2}{n}} \tag{2}$$

where $\mathrm{X_{obs}}$ is observed values and $\mathrm{X_{model}}$ is modeled values at time/place i.

3.5 Accuracy, Confusion Matrix, Silhouette and Multidimensional Scaling

Accuracy is the proportion of the true positives and true negatives to the total number of cases [8]

$$\mathrm{Accuracy} = \frac{(\text{True Positives} + \text{True Negatives})}{(\text{True Positives} + \text{True Negatives} + \text{False Negatives} + \text{False Positives})} \tag{3}$$

To view the performance of the machine learning algorithms especially for classifier, a confusion matrix or error matrix is obtained where each row represents the instances of predicted values and column the actual ones [22]. It is a special contingency table with predicted and actual classes.

Silhouette value is a measure to find out how close each point to its own cluster than to its neighboring clusters [23]. This value ranges from −1 to +1. The value close +1 means that the clustering has some impact and the samples are close to its own cluster where as the value close to −1 indicates the cluster has no meaning or objects are in wrong clusters. Silhouette Plot may be used to visualize the clusters.

Multidimensional Scaling (MDS) visualizes the level of similarity among the objects in the datasets [24]. This visualization is made with the help of distance matrix.

3.6 Data Mining Tools

The following data mining tools had been used for the analysis.

3.6.1 Orange

Orange is python based open source machine learning data mining tool [25]. It uses various visualizations for the mined data. It is interactive in nature. So, orange is visual programming environment for data analysis and mining. It runs on Linux, Windows and Mac. The current version is 3.7. Data fusion, text mining and bioinformatics packages may be installed as add-ons.

3.6.2 Weka

Waikato Environment for Knowledge Analysis (Weka) is another machine learning open source under GNU Genera Public License for classification [26], regression, association, visualization. Weka is written in Java. It was developed at University of Waikato, New Zealand. This tool runs on all the computing platforms and may be used easily because of its GUI.

3.6.3 R Studio

R Studio is a freeware for integrated development environment of R programming language [27]. It is available for two editions. One is R Studio Desktop and another is R Studio Server. R Studio is written in C++ and Qt framework is for its graphical programming. R packages are implemented in the R Studio. R programming is machine learning programming language used extensively in the field of data mining.

4 Experiments and Results

4.1 Data Cleaning and Feature Selection

The collected data has 666 instances. Some of the attributes are ambiguous. The rank of the candidate is removed from the dataset in the data cleaning phase of the data pre-processing. Performance is the response variable and others are explanatory variables. Performance attribute is taken as numerical attribute in case of k-means clustering whereas in other cases it is termed as categorical attribute. The following is the data table used for the data analysis (Table 1).

The attributes were ranked using InfogainAttribute Eval ranker search method of Weka. InfoGainAttributeEval method is used for feature selection. Entropy measures the degree of impurity. It can be assumed that the dataset is less impure if the degree is

Table 1. Attribute Description with their values

Attribute	Description	Values
Performance	Performance in Common Entrance Examination (CEE)	{'Excellent','Vg','Good','Average'} If the percentage is top 100, then Excellent If the percentage is next 200, then Very Good (Vg) If the percentage is next 200, then Good The rest is termed as Average Here the percentage means the percentage in the CEE Examination
Gender	Sex of the Candidate	{'male','female'}
Caste	Caste of the Candidate	{'General','OBC','SC','ST'} OBC – Other Backward Caste SC – Schedule Caste ST – Schedule Tribes
coaching	Whether the candidate attended any coaching classes within Assam, outside Assam or not	{'NO','WA','OA'} No – No Coaching WA – Within Assam OA – Outside Assam
Class_ten_education	Name of the board where the candidate studied at Class X level	'SEBA','OTHERS','CBSE'
twelve_education	Name of the board where the candidate studied at Class XII level	'AHSEC','CBSE','OTHERS'
medium	Medium of instructions for the study at Class XII level	'ENGLISH','OTHERS','ASSAMESE'
Class_X_Percentage	The percentage secured by the candidate at Class X standard	'Excellent','Vg','Good','Average' If the percentage is above 80%, then Excellent If the percentage is less than 80% but more than or equal to 70%, then Very Good (Vg) If the percentage is less than 70% but more than or equal to 60%, then Good The rest are termed as Average
Class_XII_Percentage	The percentage secured by the candidate at Class XII standard	'Excellent','Vg','Good', 'Average' If the percentage is above 80%, then Excellent If the percentage is less than 80% but more than or equal to 70%, then Very Good (Vg) If the percentage is less than 70% but more than or equal to 60%, then Good The rest are termed as Average
Father_occupation	The occupation of the father of the candidate	'DOCTOR','SCHOOL_TEACHER','BUSINESS','COLLEGE_TEACHER', 'OTHERS','BANK_OFFICIAL','ENGINEER','CULTIVATOR'
Mother_occupation	The occupation of the mother of the candidate	'OTHERS','HOUSE_WIFE','SCHOOL_TEACHER','DOCTOR', 'COLLEGE_TEACHER','BANK_OFFICIAL','BUSINESS', 'CULTIVATOR','ENGINEER'

close to zero. A good attribute is an attribute that reduces the most entropy and are highly ranked [28]. The caste was the highest rank attribute followed by Class_XII_Percentage, Father_Occupation, Mother_Occupation, Class_X_Percentage and Medium. The other high ranked attributes are obvious, but the Caste attribute rank is quite surprising with the ranked value 0.51393.

4.2 Association Rule Mining

The authors tried to find the association rules on the datasets using Orange. The orange canvas for the associations with widgets looks as in the Fig. 1.

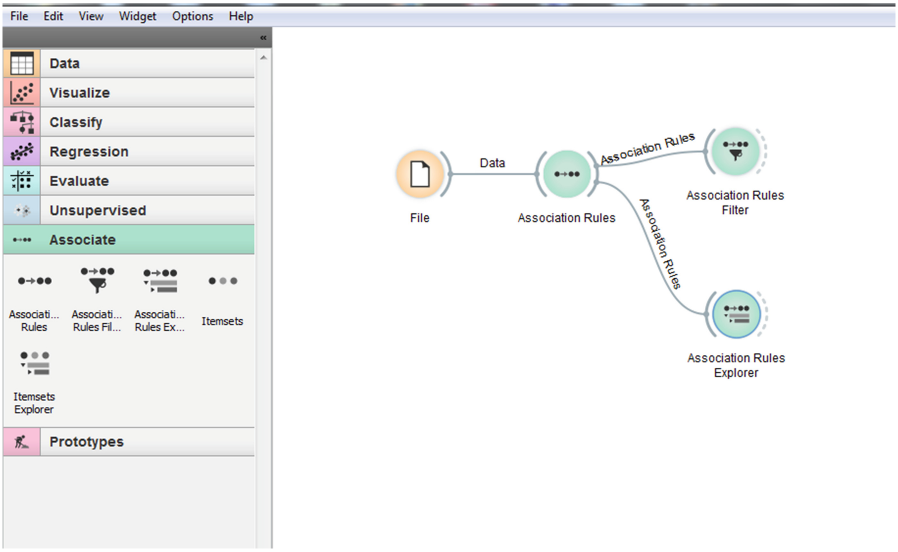

Fig. 1. Orange canvas for the association rules

With the threshold for support and confidence set as 50% and 60% respectively, the followings are the rule statistics, rule matrix and the extracted rules (Tables 2 and 3).

Table 2. Rule statistics for the datasets

Association Rules Filter Fri Nov 03 17, 15:51:26
Rules statistics
Total number of rules: 14
Support: 51%–65%
Confidence: 63%–95%
Number of rules in the graph: 14
Support: 51%–65%
Confidence: 63%–95%
Selected rules: 14
Support: 52%–65%
Confidence: 64%–96%

Table 3. Selected rules obtained by Orange

Supp	Conf	Antecedent →	Consequent
0.527	0.794	Mother_occupation=HOUSE_WIFE	medium=ENGLISH
71	0.744	Class_ X_Percentage=Excellent	Class_XII_Percentage=Excellent
0.527	0.655	medium=ENGLISH	Mother_occupation=HOUSE_WIFE
0.652	0.849	Class_ X_Percentage=Excellent	medium=ENGLISH
0.517	0.673	Class_ X_Percentage=Excellent	coaching=WA
0.515	0.862	Class_XII_Percentage=Excellent	medium=ENGLISH
0.517	0.766	coaching=WA	Class_ X_Percentage=Excellent
0.529	0.889	Class_ten_education=SEBA	twelve_education=AHSEC
0.524	0.651	medium=ENGLISH	coaching=WA
0.652	0.810	medium=ENGLISH	Class_ X_Percentage=Excellent
0.515	0.640	medium=ENGLISH	Class_XII_Percentage=Excellent
0.529	0.957	twelve_education=AHSEC	Class_ten_education=SEBA
0.524	0.777	coaching=WA	medium=ENGLISH
0.571	0.955	Class_XII_Percentage=Excellent	Class_ X_Percentage=Excellent

4.3 Classification

Three Classification Algorithms were used in the datasets viz. Naïve Bayes, Neural Network and Decision Tree using Weka. The Table 4 shows the classification summary of the three classifiers.

Table 4. The comparison of classifiers on the educational datasets

Classifier	Correctly classified instances	Incorrectly classified instances	Accuracy	MAE	RMSE
Decision Tree (J48)	431	235	64.71%	0.2296	0.3388
Neural Network (MLP)	605	61	90.84%	0.0658	0.1861
Naïve Bayes (NB)	385	281	57.81%	0.2567	0.3625

The following figures compare the correctly classified instances with incorrectly classified instances and MAE and RMSE errors (Figs. 2 and 3).

The Table 5 shows the confusion matrix of the Neural Network (Multilayer Perceptron).

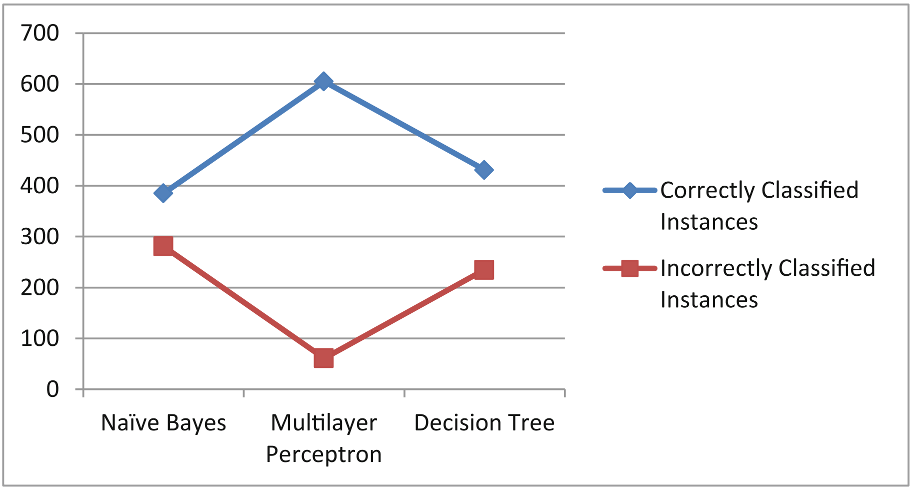

Fig. 2. Comparison of three classifiers

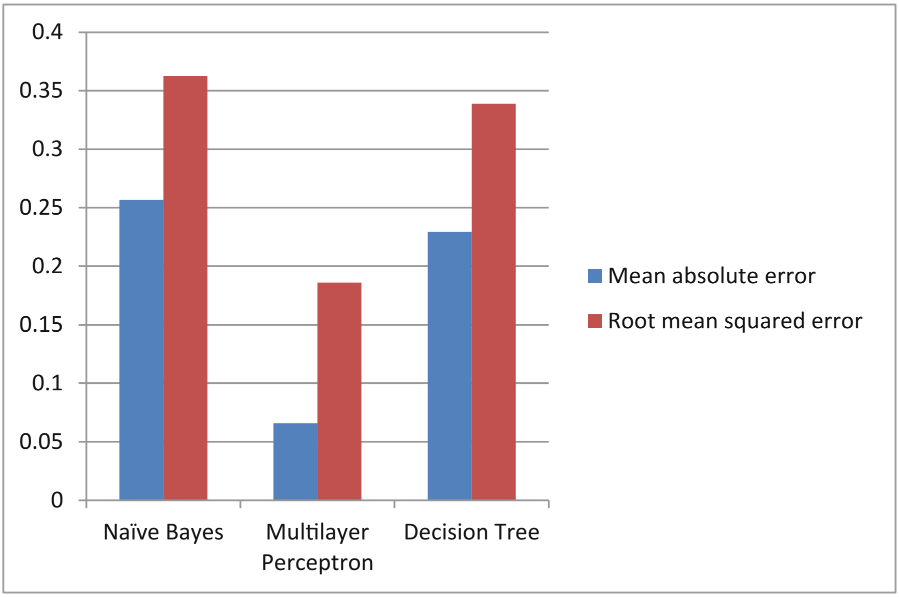

Fig. 3. Comparison of three classifiers using MAE and RMSE

Table 5. Confusion matrix for multilayer perceptron classification

a	b	c	d	Classified as
84	12	4	1	a = Excellent
4	**182**	11	1	b = Vg
5	12	**185**	8	c = Good
0	1	2	**154**	d = Average

4.4 Clustering

The clustering was applied in the datasets using Orange and R Studio. The Clustering methods applied was K-means clustering, Hierarchical Clustering and Partitioning Around Medoids (PAM). K-means and PAM performs better than the Hierarchical Clustering in one considers the Silhouette score. The following figure shows the dendrogram of Hierarchical Clustering using R Studio. For PAM and K-means, the silhouette score is 0.54 which means that a moderate score is achieved. The experiment is done with varying number of clusters. But it was found that as the number of clusters increase the silhouette score decreases. So, the performance is good with number of clusters i.e. k = 3. The following figures depict the cluster analysis results (Figs. 4, 5, 6, 7, 8, 9 and 10).

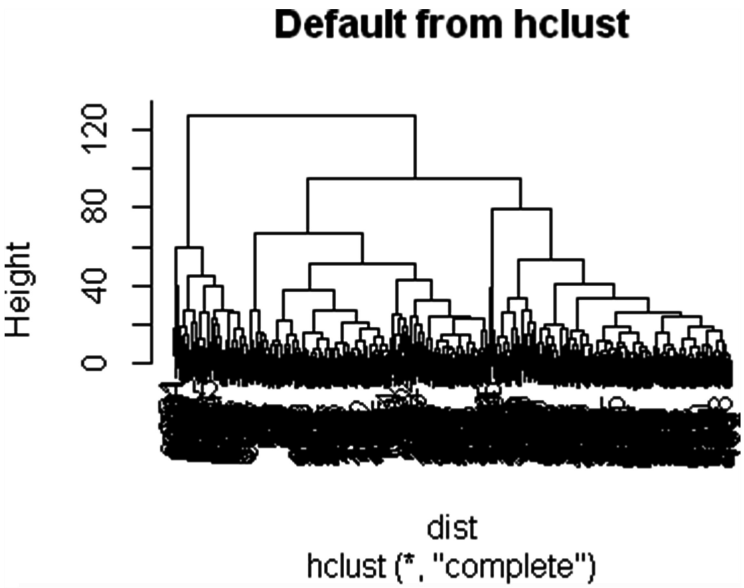

Fig. 4. Dendrogram of Hierarchical Clustering using R Studio

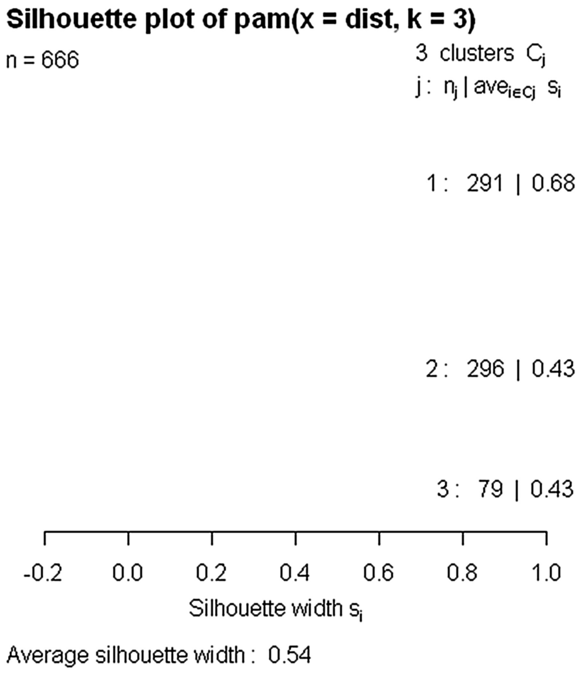

Fig. 5. Silhouette Plot of Pam with 3 Clusters using R Studio

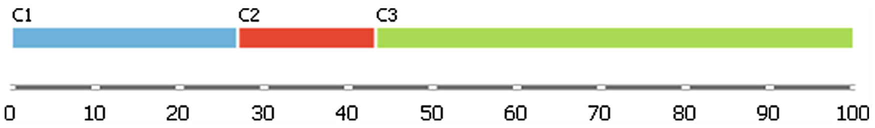

Fig. 6. Box Plot of 3 Clusters using K-means Algorithm using Orange

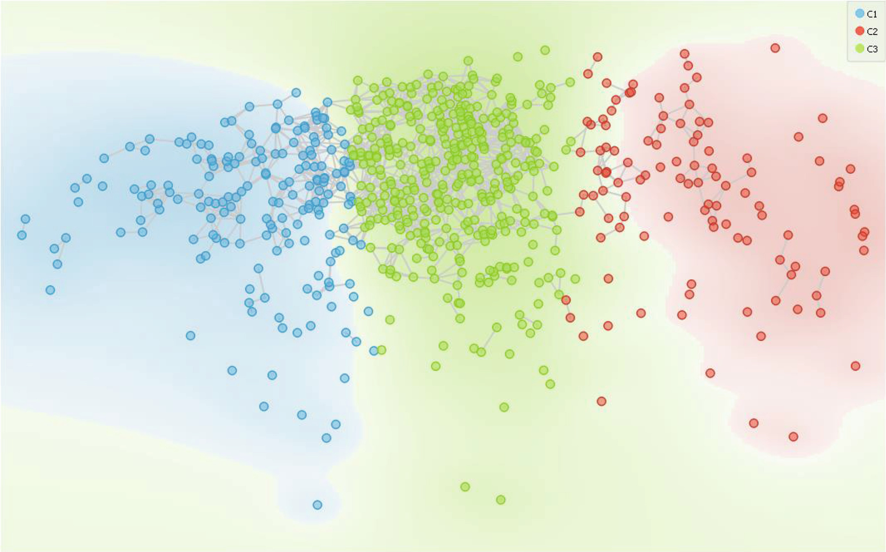

Fig. 7. MDS of K-means Clustering using Orange

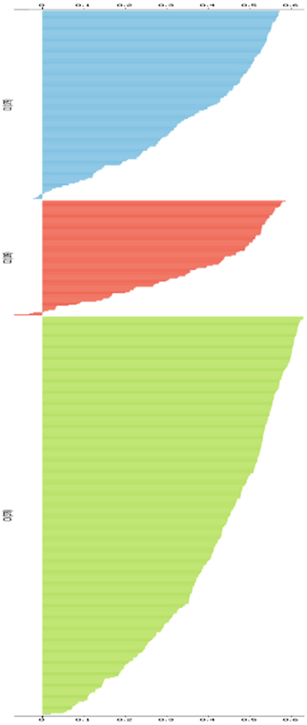

Fig. 8. Silhouette Plot of K-means Clustering using Orange

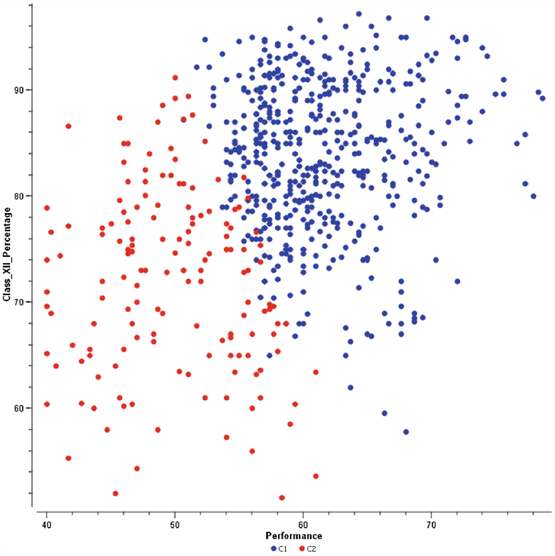

Fig. 9. Scatter Plot of two Clusters using K-means Clustering (Performance against Class XII Percentage)

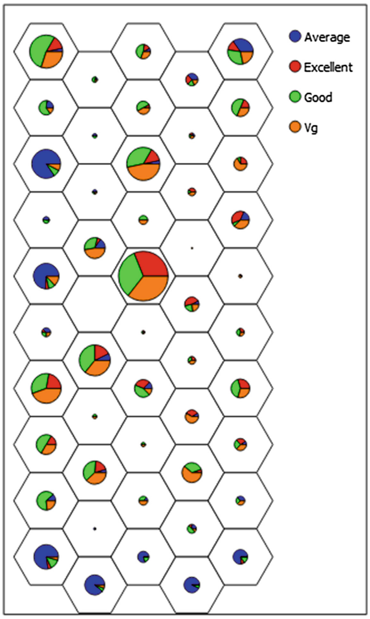

Fig. 10. SOM Visualization using Orange

5 Conclusion and Future Work

The authors tried different data mining techniques using different visualization tools. In case of association rule mining, apriori algorithm was used to mine some of the interesting rules. In case of Classification, three classifiers were used and based on its accuracy and classification errors; it was found that the neural network was the best classifier. Although neural network performs well on big dataset and was believed that it is not the best classifier for educational datasets, but in our case, the neural network classifier outperforms the other classifier with 90.84% accuracy. At the end, the authors tried out for any meaningful clustering structure. It was observed that PAM and K-means clustering performs better than hierarchical clustering with silhouette with 0.54 and number of clusters is three.

We may try to find out the trends of CEE data using time series methods as future work. The classification and clustering may be applied to the students with different sets of relevant attributes. The highly influence factors for cracking the competitive examination like CEE may also explored as future work.

References

1. Bhardwaj, B.K., Pal, S.: Data mining: a prediction for performance improvement using classification. Int. J. Comput. Sci. Inf. Secur. (IJCSIS) **9**(4), 136–140 (2012)
2. Yadav, S.K., Pal, S.: Data mining: a prediction for performance improvement of engineering students using classification. World Comput. Sci. Inf. Technol. J. **2**(2), 51–56 (2012)
3. Kukasvadiya, M.S., Divecha, N.H.: Analysis of Data Using Data Mining tool Orange. Int. J. Eng. Develop. Res. **5**(2), 1836–1840 (2017)
4. DeFreitas, K., Bernard, M.: Comparative performance analysis of clustering techniques in educational data mining. IADIS Int. J. Comput. Sci. Inf. Syst. **10**(2), 65–78 (2015)
5. Dutt, A., Aghabozrgi, S., Ismail, M.A.B., Mahroein, H.: Clustering algorithms applied in educational data mining. Int. J. Inf. Electron. Eng. **5**(2), 112–116 (2015)
6. Nagy, H.M., Aly, W.M., Hegazy, O.F.: An educational data mining system for advising higher education students. Int. J. Comput. Inf. Eng. **7**(10), 1226–1270 (2013)
7. Oyelade, O.J., Oladipupo, O.O., Obagbuwa, I.C.: Application of K-means clustering algorithm for prediction of students' academic performance. Int. J. Comput. Sci. Inf. Secur. **7** (1), 292–295 (2010)
8. Almarabeh, H.: Analysis of students' performance by using different data mining classifiers. Int. J. Mod. Educ. Comput. Sci. **9**(8), 9–15 (2017)
9. Sivogolovko, E., Novikov, B.: Validating cluster structures in data mining tasks. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT 2012, p. 245. ACM, New York (2012)
10. Everitt, B.: Cluster Analysis. Wiley, Chichester (2011). ISBN 9780470749913
11. Park, H.S., Jun, C.H.: A simple and fast algorithm for K-medoids clustering. Exp. Syst. Appl. **36**(2), 3336–3341 (2009)
12. Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithms and validity indices. IEEE Trans. Patt. Anal. Mach. Intel. **24**(12), 1650–1654 (2002)
13. Berkhin, P.P.: A Survey of Clustering Data Mining Techniques. Springer, Heidelberg (2006)

14. Chang, W.L., Pang, L.M., Tay, K.M.: Application of self-organizing map to failure modes and effects analysis methodology. Neurocomputing (2017). https://doi.org/10.1016/j.neucom.2016.04.073
15. Ahmed, A.B.E.D., Elaraby, I.S.: Data mining: a prediction for student's performance using classification method. World J. Comput. Appl. Technol. **2**(2), 43–47 (2014)
16. Pandey, U.K., Pal, S.: Data mining: a prediction of performer or underperformer using classification. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) **2**(2), 686–690 (2011)
17. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley-Interscience Publication, New York (2000)
18. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. **29**, 103–130 (1997)
19. Jiawei, H., Micheline, K.: Data Mining: Concepts and Techniques. Elsevier Book Series (2000)
20. Rakesh, A., Ramakrishnan, S.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pp. 487–499 (1994)
21. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**, 79–82 (2005)
22. Powers, D.M.W.: Evaluation: from precision, recall and f-measure to roc., informedness, markedness & correlation. J. Mach. Learn. Technol. **2**(1), 37–63 (2011)
23. de Amorim, R.C., Hennig, C.: Recovering the number of clusters in data sets with noise features using feature rescaling factors. Inf. Sci. **324**, 126–145 (2015). https://doi.org/10.1016/j.ins.2015.06.039
24. Borg, I., Groenen, P.: Modern Multidimensional Scaling: Theory and Applications, pp. 207–212, 2nd edn. Springer, New York (2005). ISBN 0-387-94845-7
25. Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Stajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: data mining toolbox in Python. JMLR. **14**(1), 2349–2353 (2013)
26. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, San Francisco (2011)
27. Verzani, J.: Getting Started with RStudio, p. 4. O'Reilly Media, Inc. (2011). ISBN 9781449309039
28. Sharma, A., Dey, S.: Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. IJCA **3**, 15–20 (2012). Special Issue on Advanced Computing and Communication Technologies for HPC Applications ACCTHPCA