1st International Conference on Multidisciplinary Research (ICMR-2018)
NIILM University, Kaithal, Haryana, (India)
4th-5th August 2018 www.conferenceworld.in ISBN:978-93-87793-38-5

Academic Performance Prediction Using Classification Technique

Aikjot Kaur Narula#, Dr.Raman Maini*

#Student, Department of Computer Engineering, Punjabi university Patiala, India,

** Professor, Department of Computer Engineering, Punjabi university Patiala, India,*

ABSTRACT:

Data mining in current time is becoming important process. Large number of applications that are data centric for both input as well as output. This outputted data need to be processed to extract student performance for its success percentages. Based on performance prediction of student good quality decisions can be taken for student using Boruta as classifier algorithm which classify the attributes based on the z-scores value of the shadow factors. It specifically based on the Random Forest classifier. Each attribute will be given numerical numbering based on the importance. Highest priority values for the attribute is identified. Those attributes which has large variations with the maximum value will be deleted from the list. The performance attributes to predict highest priority features is identified on the basis of numerical scheme adopted. Only those attributes are being considered which are comparative to the maximum ranked attribute. So the Boruta has been applied on the dataset of the students different attributes to identify the performance of the student who is likely to fail. The performance of the Boruta has been compared to the Random forest another classifier algorithm. Based on the selected features the performance of the Boruta is having good performance compared to the Random forest algorithm. The performance of the Boruta is better than the Random forest algorithm in terms of the accuracy for the prediction and execution time.

Keywords: Boruta, Classifier, Random forest.

1. INTRODUCTION

In current time there are large number of applications which are producing the data in bulk amount. This data belongs to multimedia. It can be text, images, videos etc. A better Tool is required to process this data to extract the useful facts. First and foremost step is to mine the data. So that relevant information can be extracted from the large repository of data. Data mining from the large extraction of the data is to extract the findings, correlations, patterns, trends, or relationship. Clustering the useful data later can be used in large organizations for decision making purpose. Decision quality can be improved if the extracted data will be useful in current context. Due to the size of the data it will be very difficult to extract the useful facts. It is harder to retrieve knowledge from the several datasets with growing amount. It will be very difficult to mine through the data

manually. Some automatic machine based on intelligent architecture should be builded. So that it can extract the useful information from the row facts. There are various techniques available which are machine learning based automatically adjust it the complexity of the data to extract pattern and relationships.

1.1. Data Mining Techniques

We can categories the data mining techniques in two classes. One is the descriptive and other is predictive approach. In descriptive approach the whole data will be subdivided into small groups. Developing and identifying the relationship between these groups or variables. In predictive approach prediction is done for one variable based on the value of other variable. Data mining stands for two ways of performing. One is descriptive technique and other is predictive technique. Descriptive technique is based on clustering, Association, Sequential analysis. Predictive technique is based on classification and Regression itself.

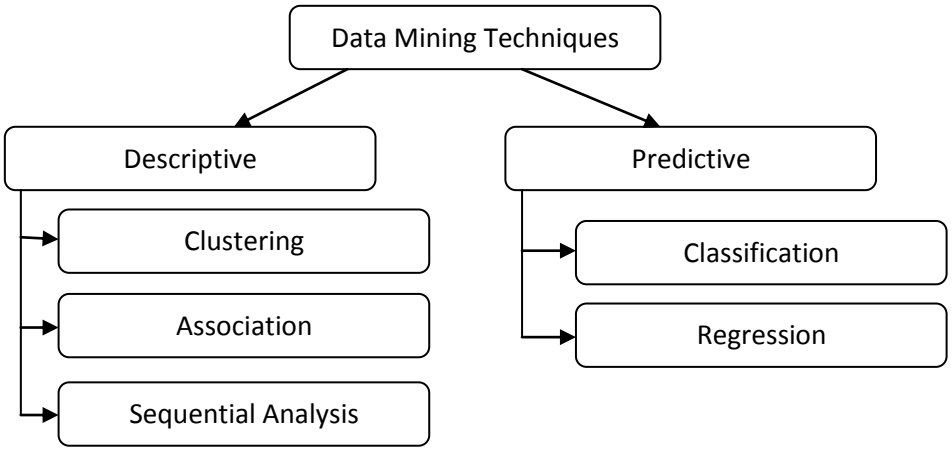

Fig. 1 Data Mining Techniques[1].

a. Classification approach

Classification approach is supervised approach. It involves two steps. In first step the data is distributed based on tuple values in the training set. For the data to be distributed in various classes the data values is known in for each variable. In second data some test data will be supplied to the model of classification to check the accuracy of the system. If the accuracy is to the required level then model can be applied to unknown tuple.

b. Clustering approach

It is the unsupervised learning approach. It will classify the dataset items into different classes. Each class element are similar in nature. One class element are different from other class. This type of technique is useful for fields like business data processing and image processing applications. Following are the classification methods.

i. Tree based classification Algorithm: it is the regression based model. Where whole data will be classified into different small units. While subdivision the tree type structure will be builded. This tree as the final product consists of various leaf and parent elements.

ii. Decision stump: A decision stump is the learning based model. It is machine learning model. It is also know as one level decision tree based model. It considers one level root element and its immediate neighbors that is leaf elements. There are various algorithms like viola-Jones in case of face detection is based in decision stump based approach.

iii. Rule based classification: it is the process of Rule based classification is also called as separate and conquer approach. It will classify the data items in to separate groups and dealing with each group till all the groups are treated.

iv. K-Mean algorithm: it is iteration based approach. Where each items is iteratively moved along the various clusters. This process goes in till the required level is not achieved. It is also be viewed as squared error algorithm.

1.2 Process of data mining

Database: it is the central repository where whole data is being collected. This large repository includes relevant and non relevant data.

Target Data: it is the process of selection of the data. Some filtering process is required to filter the data to extract the useful data.

Processing data: it is to process the data to be transformed to the format appropriate for the process of mining.

Transformed Data: it is the data format where whole data is to be transformed to the format which is suitable for the machine learning system. Used for both classification and clustering techniques.

Patterns and model: it is the framework where large repository of the data using certain model be having specified pattern and models. So that system can be understood there on for drawing conclusions.

Knowledge: it is the last step where the data extracted from the mining process is used for some decision making purpose based on extracted knowledge.

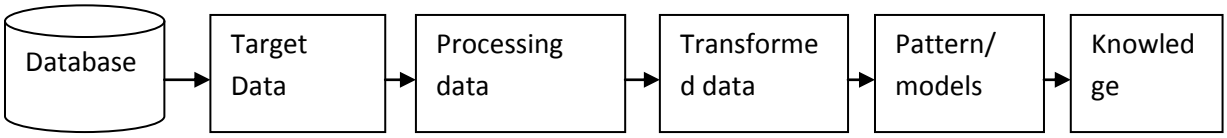

Fig. 2 Process[1]

2. LITERATURE SURVEY

A. Nichat et. al (2017): Analyze and predict the performance of a student with the help of classification technique such as decision tree technique. The author processed the tasks based on various attributes in order to predict the performance of student activities. This research is more concerned about the improvement of classification techniques. These techniques are used to analyze the skilled expertise based on academic performance. The results indicates that data mining techniques are effective in improving the tools for the analysis of student performance[1].

A.V. Kadu et al. (2015): Predicted the performance of students with the help of data mining. In this research, data mining is used to predict the academic outcomes based in early performance and characteristics of other students. The objective of this research was to predict students who are at higher risk of not achieving the good honor degree. The authors proposed the interventions to identify the student performance as early as possible. After that, various strategies are recommended and then measured the improvement in performance[2]

A. Abu Saa et. al (2016): Analyzed the educational data related to the performance of students. Educational data mining is related to mining educational data used to evaluate the interesting patterns and knowledge in educational organizations. This research is concerned with the student's performance. The author explored the several factors in order to affect the performance of the student in higher education. This research also identified the quantitative model which best predicts the classifies the performance of students based on social and personal factors.[3]

E.B. Costa et. al (2017): in this paper author has put the study for the educational system effectiveness. It to mine the education related data for the decision for the student performance at the introductory level. The authors also analyzed the impact of the data preprocessing and fine tuning task algorithms.[4]

C.J.V. Arnedo et. al(2016): author in this paper has works on the prediction system. Used for prediction purpose. It is the black box based model. The input data classifier provides the rich output dataset. The authors also proposed the graphical tools in order to exploit the output performance. It also helps in providing the meaningful guide to students as well as teache

3. ALGORITHM

3.1 Random forest

Algorithm for the construction Random Forest is

- Assume the Training cases be considered as n. And the number of variables considered be M which are being used as classifier.
- The variables considered to build a tree is P. this P is always less than m.
- For Building each tree element the careful selection of the attributes values will be considered. New data can be selected based on the majority votes in the tree.
- Evaluate the best split based on the variables lies in the training set.
- Each tree is spanned fully. It is not filtered tree.
- Best split is one which has least error.

3.2 Boruta Algorithm

It is the classification technique. It is based on Random Forest technique. Boruta is wrapped wrapper around Random forest. Random forest is the quick classification technique. Where the classification is done without tuning up the parameters . it gives numerical estimation to the features based on there importance. It is the ensemble technique which classify the data based in voting of the multiple unbiased weak classifiers like decision tree. It is to consider the whole tree rather than the selected elements. Boruta Uses the existing attributes for the selection based on ranking of the attributes.

In nutshell we can say Boruta is developed on the same idea uses by the random forest classifier. It keeps the numerical numbering to the attributes based on the importance. Higher randomness will give more clear picture about the importance of the attribute.

1. Build a information system by addition of the all the variables. These variables are at least five in number. These attributes are being considered as the show attributes.
2. simply shuffle the attributes to generate the randomness in the attribute values.
3. Run the Random forest technique for the classification to compute the Z- scores.
4. Put the Shadow attributes based on the Z-score value. More Z-Scores value attributes will be kept first.
5. For each attribute with undetermined importance performs two sided test of the equality.
6. Remove those attributes permanently from the list which has Z-Scores value significantly lower than the maximum value.
7. Keep those attributed whose Z-Index value is comparable to the maximum Z-Scores value. Either equal or higher than the maximum Z-Index Value.
8. Once attributes are selected based on importance shadow attributes are deleted.
9. Repeat the process till the all attributes are treated.

Random forest is another classification algorithm. It is the technique to prepare the decision tree for the training data. Later on comparing the test data with the decision tree. Random forest give rank the importance to the variables.

3.3 PSEUDO CODE OF ALGORITHM

1. Collect_attributes(i..n) n>=5
2. Shadow_attributes(i..n)
3. Shuffle_attributes()
4. Evaluate_MZSA()
5. selection_attributes()
6. goto step 4
7. end

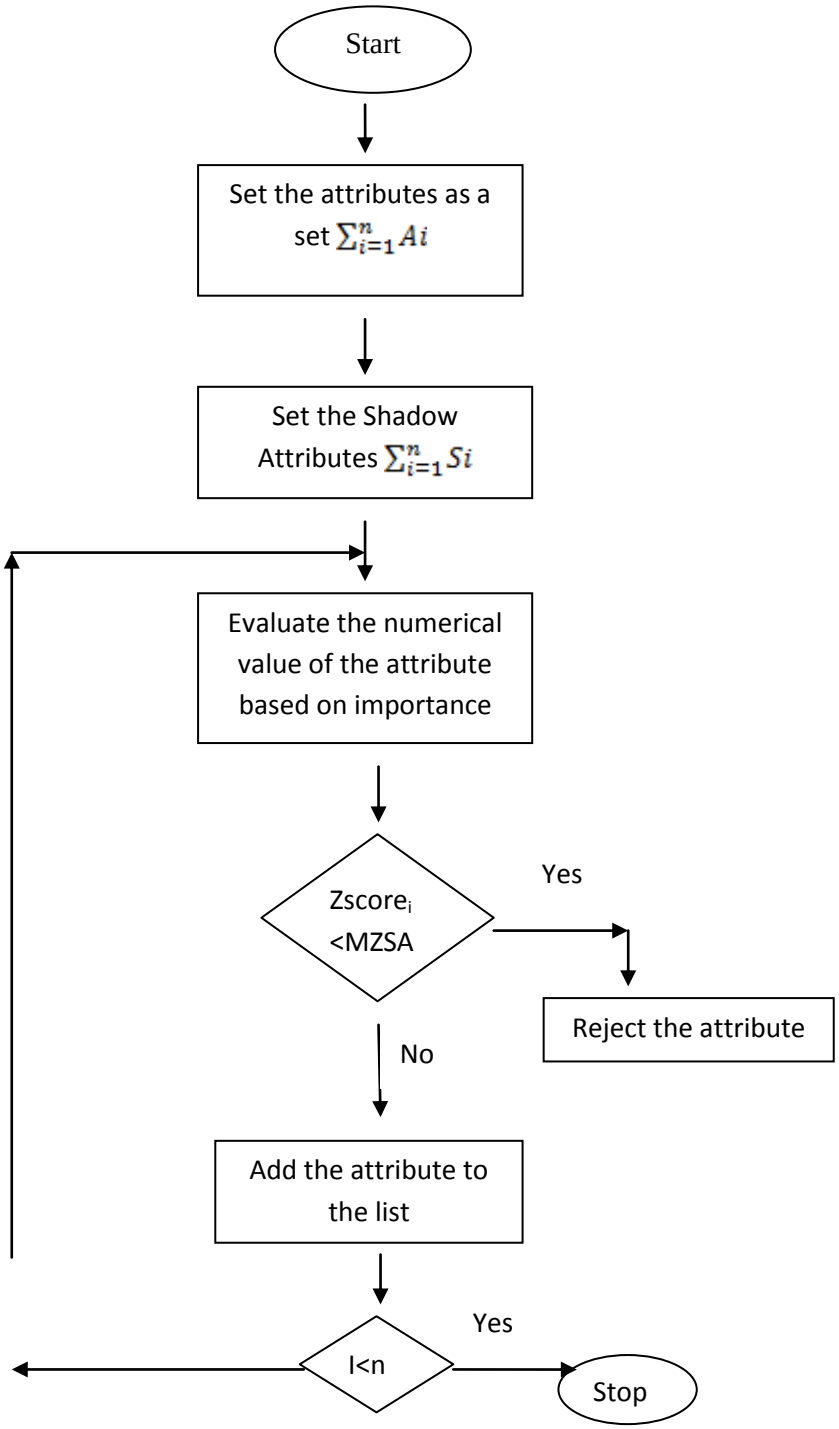

Fig. 3 Flowchart

4. RESULTS AND ANALYSIS

Classification using Boruta over to the dataset for the student to detect the pass percentage. Dataset has 52 attributes for the students. Boruta has rejected 44 attributes based on importance. The importance is calculated based on Max. zscore value of the shadow attributes.

Attributes	Count
Total Attributes	52
Rejected	44
Confirmed	8

Table 1. Attributes

Importance comparison for the Boruta and Random Forest

While applying the Boruta and the Random Forest the selected attributes like failure.x, higher.x, gout.x, G2.x, G3.x, G1.y, G2.y, G3.y Minimum important value and Maximum important value has been indentified. The values are being identified based on z-score value of the shadow attributes.

	minImp for Boruta	maxImp for Boruta	minImp for Random Forest	maxImp Forest
failures.x	2.5230278	7.430368	1.108128	3.063549
higher.x	-0.689697	4.295483	1.733931	-0.82893
goout.x	0.2785523	6.66213	-1.20626	0.063104
G2.x	20.9377576	32.23	1.477549	11.70916
G3.x	25.9353302	47.89364	1.963274	9.656467
G1.y	2.3401021	7.87386	-1.48184	3.161633
G2.y	4.2188705	9.888235	0.298414	5.507902
G3.y	3.0619491	9.010738	0.723159	2.841596

Table 2 Max and Min Important attributes

Area Curve for Min importance value for Boruta

Graph in Fig. 4 is representing area under curve for the minimum importance of the attributes based on shadow attributes. This has been identified by comparing the attributes scores with zscore values. So that only those attributes are being selected which are having value greater than the zscore value of the shadow attributes.

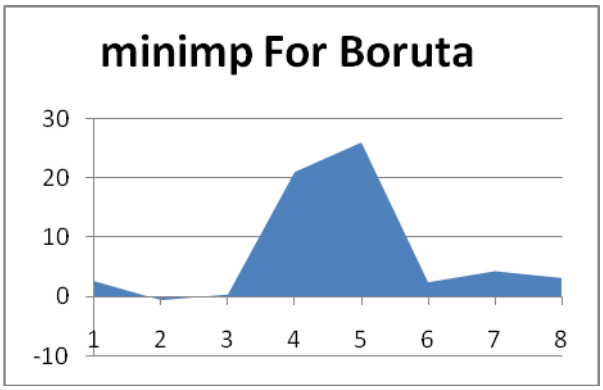

Fig. 4 Min Importance of attributes for Boruta

Area Curve for Max importance value for Boruta

Fig. 5 shows the Area Curve for the Max importance of attributes for the dataset under Boruta Algorithm. This graph shows the area under curve for the max important attributes. So that these attributes are selected by comparing the zscore value of the shadow attributes.

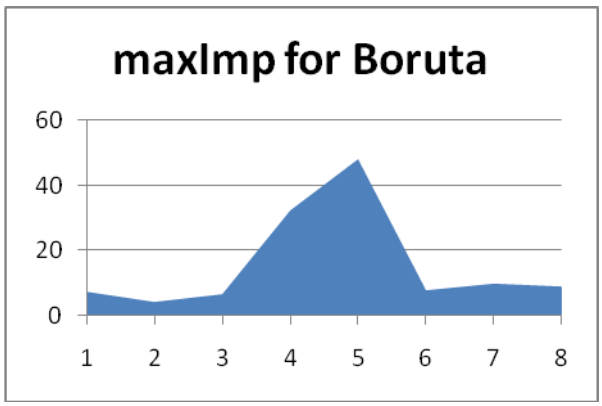

Fig. 5 Max Importance of attributes for Boruta

Area Curve for Min importance value for Random Forest

Fig.6 shows the Area Curve for the Min importance of attributes for the dataset under Random Forest Algorithm.

Area Curve for Max importance value for Random Forest. Various attributes values are negative. Means the random forest algorithm is less efficient in determining the attributes important scores.

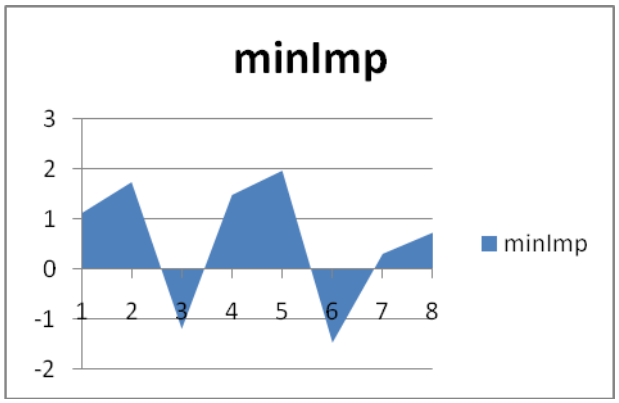

Fig. 6 Max Importance of attributes for Boruta

Area Curve for Max. importance value for Random Forest

Fig. 7 shows the Area Curve for the Max importance of attributes for the dataset under Random Forest Algorithm. Graph represents the area under curve for the maximum importance of the attributes. These attributes are the selected attributes for the identification of the student performance in terms of success. That means probability of the fail.

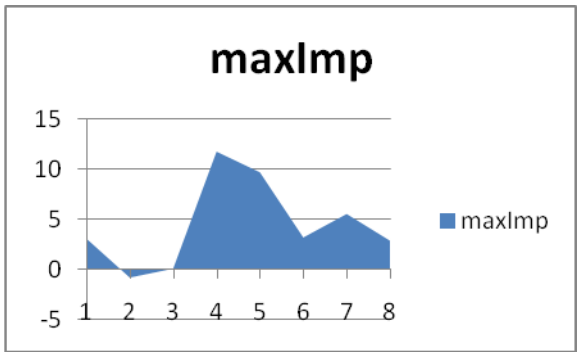

Fig. 7 Max Importance of attributes for Boruta

Comparison of min and Max for the Boruta and Random Forest

Fig. 8 shows the Comparison for Min and Max importance of the attributes for the Boruta and Random Forest. Boruta is performing better than the Random Forest.

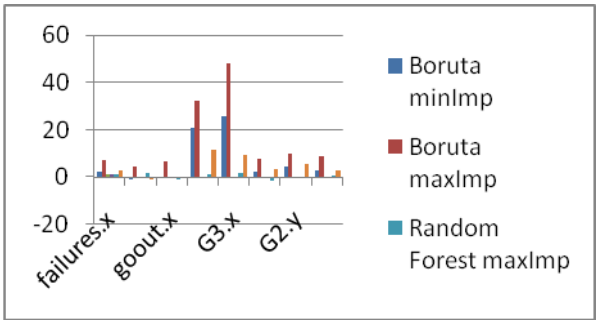

Fig. 8 Comparison for Min and Max for Boruta and Random forest

ACCURACY Comparison for the Boruta and Random Forest

Table 3 shows the Accuracy comparison for the confirmed attributes. Random forest has decrease in the accuracy for the confirmed attributes in comparison to the Boruta Technique. There is substantially decrease in the accuracy for the Random forest. This means the accuracy by Boruta is better than the Random Forest technique.

The Mean Decrease of Accuracy of classification for Boruta and Random Forest		
1	failures.x	6.85350955
2	higher.x	1.76561752
3	goout.x	1.92246387
4	G2.x	36.71049688
5	G3.x	34.14380367
6	G1.y	8.69889191
7	G2.y	9.71254211
8	G3.y	10.32725056

Table 3 Accuracy Comparison for the confirmed Attributes

Range Values for G Attributes

Table 4 shows the values range for the G2.x,G3.x,G1.y,G2.y,G3.y. based on these values range the attributes are defined in tree type structure.

	Min	Max
G2.x	0	17.5
G3.x	0	11.5
G1.y	0	15.5
G2.y	0	13.4
G3.y	0	11.2

Table 4 Values Range

5. CONCLUSION AND FUTURE WORK

Data mining is the most progressive and helpful field for the extraction and manipulation of data to produce useful information. It extracts useful information and identifying it from the records stored in their data warehouses databases and data repository. Boruta uses classification technique to find importance of the attributes The importance of attributes is based on the z-scores values of the shadow attributes. Boruta in comparison to the Random forest technique has shown better accuracy and takes less execution time. In current research objective was to predict student's risk of failure in exams. With the help of Boruta Algorithm prediction of student the performance in exams. In current research large dataset with various attributes for the students for the prediction taken. and gave early warning for the student performance. In future this technique can be applied in various other strategic fields to know the performance of the Boruta Technique.

REFERENCES

1. A. Nichat, P.H. Chu, and P.Y. Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011," Expert Systems with Applications, vol. 39, no. 12, pp. 11303–11311, 2017.

2. A.V. Kadu, S. Kannan, and K. Nagarajan, "Data Mining: Analysis of student database using Classification Techniques," International Journal of Computer Applications, vol. 141, no. 8, pp. 22–27, 2015.

3. A. Abu Saa and A. M. J. Md. Zubair Rahman, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016.

4. E.B. Costa and S. Goel, "Data Mining - Techniques, Methods, and Algorithms: A Review on Tools and their Validity," International Journal of Computer Applications, vol. 113, no. 18, pp. 22–29, 2017.

5. C.J.V. Arnedo , A. Maté, and M. Marco, "Application of Data Mining techniques to identify relevant Key Performance Indicators," Computer Standards & Interfaces, vol. 54, pp. 76–85, 2016.

6. C. C. B. Burgos, M. L. Campanario, D. D. L. Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," Computers & Electrical Engineering, 2017.

7.C. C. B. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," Computers & Education, vol. 51, no. 1, pp. 368–384, 2008.

8.Y. Altujjar, W. Altamimi, I. Asl-Turaiki, and M. Al-Razgan, "Predicting Critical Courses Affecting Students Performance: A Case Study," Procedia Computer Science, vol. 82, pp. 65–71, 2016.

9.G. Badr, A. Algobail, H. Almutairi, and M. Almutery, "Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department," Procedia Computer Science, vol. 82, pp. 80–89, 2016.

10.Harwati, A. P. Alfiani, and F. A. Wulandari, "Mapping Students Performance Based on Data Mining Approach (A Case Study)," Agriculture and Agricultural Science Procedia, vol. 3, pp. 173–177, 2015.

11.M. I. Al-Twijri and A. Y. Noaman, "A New Data Mining Model Adopted for Higher Institutions," Procedia Computer Science, vol. 65, pp. 836–844, 2015.

12.H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," Procedia Technology, vol. 25, pp. 326–332, 2016.

13.R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, "Data mining models for student careers," Expert Systems with Applications, vol. 42, no. 13, pp. 5508–5521, 2015.

14.M. Mayilvaganan and D. Kalpanadevi, "Cognitive Skill Analysis for Students through Problem Solving Based on Data Mining Techniques," Procedia Computer Science, vol. 47, pp. 62–75, 2015.

15.Y.H. Hu, C.L. Lo, and S.P. Shih, "Developing early warning systems to predict students' online learning performance," Computers in Human Behavior, vol. 36, pp. 469–478, 2014.

16.C. Angeli, S. K. Howard, J. Ma, J. Yang, and P. A. Kirschner, "Data mining in educational technology classroom research: Can it make a contribution?" Computers & Education, vol. 113, pp. 226–242, 2017.

17.M. Chalaris, S. Gritzalis, M. Maragoudakis, C. Sgouropoulou, and A. Tsolakidis, "Improving Quality of Educational Processes Providing New Knowledge Using Data Mining Techniques," Procedia - Social and Behavioral Sciences, vol. 147, pp. 390–397, 2014.

18.J. D. Gobert, Y. J. Kim, M. A. S. Pedro, M. Kennedy, and C. G. Betts, "Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld," Thinking Skills and Creativity, vol. 18, pp. 81–90, 2015.

19.Buldu and K. Üçgün, "Data mining application on students' data," Procedia - Social and Behavioral Sciences, vol. 2, no. 2, pp. 5251–5259, 2010.

20.M.Kumar and Rajesh "Predicting Upcoming Students Performance using Mining Technique," International Journal of Modern Trends in Engineering & Research, vol. 4, no. 7, pp. 38–44, 2017.

21.T. Devasia, V. T. P, and V. Hegde, "Prediction of students performance using Educational Data Mining," 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), 2016.

22.C. E. L. Guarin, E. L. Guzman, and F. A. Gonzalez, "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining," IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 10, no. 3, pp. ss119–125, 2015.sss