International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 10, October-2018

A REVIEW OF DM APPROACHES FOR PREDICTING STUDENT'S PERFORMANCE

Ankita Singh Tomar[1], Rajendra Kumar Gupta[2], Khushboo Agarwal[3]

[1]Research scholar, MITS, Gwalior, India
[2]Faculty of CSE/IT, MITS, Gwalior, India
[3]Faculty of CSE/IT, MITS, Gwalior, India

ABSTRACT- To judge the performance of students in advanced education has now become a great challenge not only in academic but in curriculum activities as well. In this manner, it is important to adequately analyze the data which is utilized for educating and learning processes. Mining data from educational dataset helps to extract useful information for enhancing the teaching and learning abilities. This survey presents the overview of various Data Mining Techniques in terms of advanced education as a new research domain known as educational data mining (EDM) for foreseeing the execution behavior of the students and to find out the reasons behind their failures. The intended review of literature is assessed to briefly evaluate the work done in educational field. Through this task, we extract those factors that describe student's attainments and discover those ones who require special concern by introducing various data mining methods which assimilate classification, clustering, regression, prediction and so forth.

Key words – Data mining, EDM, performance prediction, data mining techniques.

I. INTRODUCTION

Data mining may be introduced as the process of extraction of an enormous amount of information as indicated by various points of view for sorting valuable data, which is collected and assembled in common areas, such as data warehouses and large databases using advanced data mining techniques, tools and algorithms in numerous domains. In other words, DM is the technique of mining valid, previously unknown, intelligible and actionable data from large databases to make it useful information and then using it to make essential commercial and intellectual decisions.

Educational data mining is a standout application amongst the most vital uses of DM which converts raw educational data into relevant data. EDM mainly focuses on presenting new and more proficient strategies for investigating the interesting sorts of information for students that originate from educational environments. Its main objective is to understand how students learn and distinguish the settings in which they figure out how to enhance educational results and to pick up bits of knowledge into and comprehend educational phenomena [1].

The main challenge of Educational data mining was that traditionally, researchers have been utilizing techniques, for example, to gather information related to student's learning experiences through physical methods like interviews, meetings, questionnaires and classrooms undergoing. These strategies are extremely tedious and very little efficient. EDM gathers information from academic databases which is utilized to create different systems and to perceive designs that are unique. The obtained information would then be able to get utilized in advanced education institutions to upgrade their basic leadership process, to enhance student's learning behavior, to limit failure rates, to understand student's behavior in a better way, to help educators and instructors and to enhance teaching skills. It creates computerized techniques for recognizing designs in substantial accumulations of unarranged data which is extremely difficult to investigate or deal with because of the extensive volume of information it exists within.

From a practical point of view, EDM permits to discover new data that is based on the level of data used by the students in order to survey academic systems, to potentially improve the significant aspects in quality of education and to lay the reason for a more effective learning process which depends on the type of users and data used in this process [2]. The type of data used in EDM is:

Offline Data: In this method, Data are generated through various methods like traditional and modern classrooms, interactive instructing, learning conditions, student/teachers data, students' participation, Psychological behavior, course data, information gathered from the academic section of an institution etc [3].

Online Data: This Data are created from the geologically isolated stakeholders of the education; distance educations, electronic training, and computer-supported collaborative learning utilized in social communicating sites and online group convocations [4].

Stakeholders are those clients from whom the business draws its assets like lenders, chiefs, workers, government proprietors (investors), providers, associations and the network. The different stakeholders of EDM can be extensively grouped underneath in Table 1.

TABLE 1. EDM USERS/STAKEHOLDERS.

Types of Users	***Reasons to apply DM techniques.***
Learners/Students	To educate guardians of students about their kid's development in a specific course and furthermore to propose fascinating inclining encounters to the students.[5]
Educators/ Teachers/ Instructors	To recognize the necessity of help for anticipating students' execution. To investigate students' learning and customize behavior. To discover students' normal and unpredictable behavior.
Course Developers /EducationalResearchers	To assess and create data mining methods for viability and exactness among students. To look at data mining systems for suggesting most valuable strategies among all. Attempt to discover new and more productive approaches for enhancing the performance of student.[5]
Organizations/Learning Providers/Universities/ PrivateTrainingCompanies	To increment and smoothen the selection procedures in educational institutions. To propose profitable courses for each class of students. To choose most proficient candidates for graduation and locate the most gainful method for enhancing the methods of evaluations.
System Administrators/ NetworkAdministrator	To examine and create educational data mining applications in educational domain. To improve academic program offers and find the viability of the distance analyzing methods.

II. OBJECTIVES OF EDM

The following four goals of EDM introduced by Baker and Yacef [6] are:

1)Prediction of future learning behavior of students and improvement in their aptitudes like student's personal information, inspiration, meta-perception, etc

2)Discovering or enhancing interested models

3)Studying the impacts of academic support

4)Enhancing logical knowledge about learning and graduators

III. EDM PHASES

Educational data mining mainly aims to extract the valid patterns from the huge quantity of data which is gathered from various academic sources. It usually consists of five phases [7]

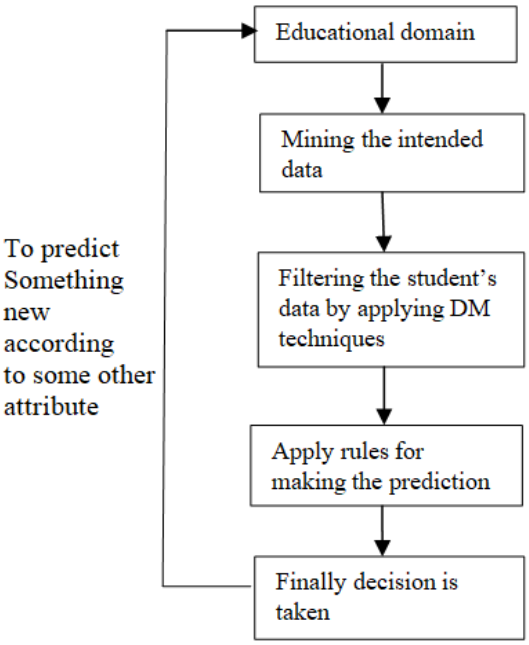

Fig 1. EDM Phases

1. In the first step, EDM establish relationship between data by searching through training set of data.
2. In this step of EDM, whatever relationship was arisen in previous step is cross-checked for validation by applying data set.
3. In the third phase, various data mining methods like clustering, classification, regression, (ARM) association rule mining, neural network etc are used for predicting the learning environment.
4. In the next phase the predicted values are examined.
5. In the final step, decisions are made according to the output gained to make useful strategy for academic institutes with the help of prediction.

IV. MODELS OF EDUCATIONAL DATA MINING

A number of techniques for educational data mining have been introduced which includes clustering regression classification etc, but all techniques lie in the following specified categories discussed in fig 2:

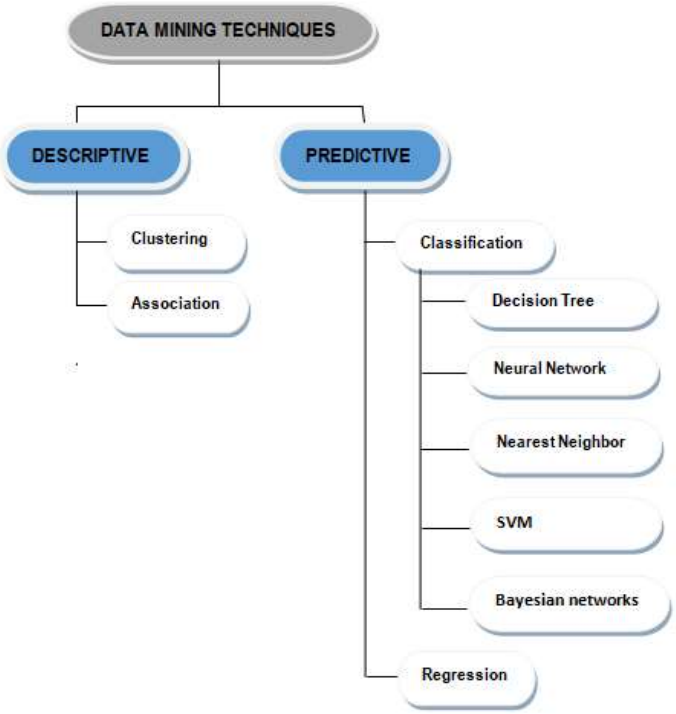

Fig.2 Data Mining Techniques.

1. Descriptive: This analytics summarizes the provided raw material by sorting out the interpretable things to make it understandable by the users.

1.1 Clustering: It is the method of grouping data according to the categories for achieving the best possible rate of similarity and dissimilarity among all groups**.** When data set is specific, then the clustering is more efficient [8].
There are three major clustering methods: K-means clustering, hierarchical method and density method.

1.1.1 K-means Clustering: It was first used by James Mac Queen (1967) is the most frequently used partitioned clustering algorithm because of its feasibility and efficiency and higher accuracy. The major reason behind the usefulness of this clustering technique is that has the ability to generate tighter clusters in contrast to that of hierarchical clustering. But sometimes it seems hard to compare the exact quality of the clusters produced, makes it sometimes less efficient [9].

1.1.2 Hierarchical Clustering: It is an agglomerative (bottom up) clustering technique begins with one factor (singleton) and repeatedly adjoins two or more appropriate clusters. It gets stop when certain (k) number of clusters is attained. In this technique the prior information of the clusters in not required and hence it becomes easy to use.

1.1.3 Density based Clustering: It consists of two different concepts of density i.e, reachability and connectivity. One of the benefits of the DBSCAN is that unlike Hierarchical clustering, it does no longer feels the necessity for the information about the number of clusters in the data set. But it provides excellent result only in the case of well-formed clusters and it cannot form proper cluster of data sets having a large alterations in densities [10].

1.2 Association: ARM is the process of finding frequent patterns, relationships, associations or casual structures from the set of items or data found in varieties of databases like relational databases, transactional databases, and various types of data depositories and dataware houses to discover the guidelines that concludes how or why the specific devices are often bought together in any transaction having more than one items. In Association rule mining, the Market basket analysis is said to be the most recommended application domain preferred by the researchers [11]. The ARM can be briefly described within two factors:

a) Find all generic item sets: Each and every item sets will appear often as a predetermined minimum aid count.

b) Originate strong rules of association from the frequent item sets: The rules should satisfy the property of minimal support and confidence. These rules are known as strong association rules. [12]

2. Predictive: It develops a model that can conclude more than one component of data into a single aspect. It is an approach carried out on a database either to predict response variable based totally on predictor variable.

2.1 Classification: In this technique, a series of data is classified in order to achieve more accurate predictions of the target class for every case of the data

2.1.1 Decision tree: It is a support tool for making decisions which makes use of a tree-like graph of decisions and their visible importance, which consists of chances of event results, resource costs, and utility. ID3 and C4.5 are the two algorithms mainly used for decision making process.

2.1.1.1 ID3 (Iterative Dichotomiser 3): ID3 developed by Ross Quinlan is a simple decision tree learning technique. The fundamental reason behind ID3 algorithm is to develop the decision tre4e by introducing a top-down, greedy search through the provided sets of data for cross-checking every characteristics at each and every tree node for choosing the most useful attribute for classifying a given set. [13].

2.1.1.2 C4.5: This algorithm is the advanced version of ID3 algorithm developed by Ross Quinlan. C4.5 handles both particular and continuous attributes to construct a decision tree so as to cope with continuous attributes. It has some extra features like coping with missing values, categorizing the continuous attributes, and sorting of decision trees, rule derivation and others.

2.1.1.3 Hunt's Algorithm: In this technique, a decision tree is evolved in a loop pattern via means of partitioning the training documents repeatedly into refined subsets. The algorithm recurses till each and every leaf node is found. This technique combines two steps for constructing a decision tree.[14]
Step 1: This step examines whether each file in a node is of the identical class. In that case, the node is named as a leaf node with its classification, which involves the class name of all the archives.
Step 2: If a node is not pure, in that condition choose/creates an attribute test condition to divide the records into two clean data sets. From here a child node is created for each and every subset**.**

2.1.1.4 RndTree (Random Forest): It is a supervised classifier that builds a decision tree that assumes randomly selected components say k at every node of the tree without reducing. The algorithm can work both with classification and regression problems. It's a bagging tree that highlights the ability of multiple varied analyses and ensemble learning to provide deep data understanding [15].

2.1.2 Neural networks: Neural network is the illustration of biological neural system in the form of computational models with adaptive human brain. It executes information at an excessive velocity for fixing complex queries like classification and prediction. Although, ANN has proposed various models, the feedforward neural networks (FNNs) are the most well known and broadly used in a number of applications.

2.1.3 K Nearest neighbor: KNN is an algorithm for classifying data which is useful in knowing the cluster to which a data point belongs to by looking at the data points surrounding it. The k-nearest-neighbor can be named as "late learner" algorithm because of not producing a prototype of the data set in advance. The solely estimation it produce is only when it is requested to choose the data point's neighbors which makes KNN optimized for data mining process. [16]

2.1.4 SVM: Support Vector Machine is a differentiate classification method which is examined by partitioning a hyperplane. SVM produces a hyperplane, for classification and regression techniques. It discovers the closest data vectors called support vectors (SV), to the decision confinement in the training set and a given new test vector can be detached by utilizing only the given closest data vectors [17].

2.1.5 Bayes Theorem's: According to this theorem, the existence of a specific feature in a domain should not relate to any other feature. This theorem is easy to use since it requires just a single scan of the training data, also it effortlessly handles mining data by simply eliminating that probability. The Formula is as per the following:[18]

$$\mathbf{P(c/x)} = \frac{P\left(\frac{x}{c}\right) P(c)}{P(x)} \quad \text{...(1)}$$

2.2 Regression: It is a DM method for predicting the values of a number. Other attributes like benefits, contracts, accommodation price, property rate, region, climate or distance can also be estimated using Regression. As an illustration, the value of a building based on its location, number of rooms, area size and other components can be determined using regression model. Regression analysis originates the values of a configurations for a variable that convert the function to a healthy set of data. The categories of regression functions are:

a) A linear regression approach best fits when the correlation among the predictors and the output can be approximated with a straight line [19].

$$Y = \beta_2 X + \beta_1 + e \quad \text{...(2)}$$

b) Multivariate linear regression refers to linear regression with multiple predictors ($\mathbf{X}_1$, $\mathbf{X}_2$, …, $\mathbf{X}_p$).

$$\mathbf{Yi = \beta 0 + \beta 1 X1i + \beta 2 X2i + \ldots + \beta p Xpi + \varepsilon i} \quad \text{...(3)}$$

V. A COMPREHENSIVE REVIEW OF LITERATURE

Since EDM is a moderately new method of research in the data extraction methodology, it can be defined as the implementation of DM tools and techniques for the analysis of data found in academic areas. It has successfully overcome the limitations faced during traditional methods of mining academic data. After analyzing different literature it has been found that for the prediction of performance placement, scholarship for recruitment process, etc, different factor plays a vital role to measure the accurate results like student's grade in semester exam, parental education, residential area, source of learning, family annual income, student's interest in extra activities, student's family background, student's learning behavior, language he's familiar with, and many other attributes related to student's routine helps to analyze the prediction of different domains [20][21]. Some literature survey areas of different authors are mention below in table 2.

TABLE 2. A COMPREHENSIVE REVIEW OF LITERATURE

S.No.	Year, Author(s)	New Research	Technology used	Algorithm
1.	2015**,** Pooja M Dhekankar, et al.	Students are categorized into grade in order to improve their academic performance [22].	Association rules, clustering, classification and Outlier detection.	Rule Induction and Naïve Bayesian
2.	2014, P.Veeramuthu	Assess and evaluate decision making processes [23].	Clustering	k- means
3.	2017,AnduelaLil e	Improve the learning behavior of students and to increase their profits [24].	Attribute Weighting, Clustering, Classification, Association Mining	Weighting by Information Gain, Relief, Chi-Squared, Uncertainty, K Means, Tree Induction, Apriori, FPGrowth algorithm, Create Association Rule, GSP
4.	2014, Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta,	Identifying attributes that influenced students 'third semester performance prediction [25].	Classification	J48 , Random Tree
5**.**	2014, ShradhaShet and Gayathri	Overcome the problem of low grade of students [26].	Ensemble classification	J48, Naive bayes, decision table
6.	2017,Abdulsala m SulaimanOlaniy, et al.	To improve the level of decision making [27].	Association rule mining	Apriori Algorithm
7.	2015, OmprakashChan drakar	Determine the performance of students in their examinations and anticipate the outcome of the forthcoming examination [28]	Association Rule mining	Apriori algorithm
8.	2013, D. Magdalene Delighta Angeline	Identify the average and below average students and to improve their performance to provide good results [29].	Association rule mining	Apriori algorithm
9**.**	2015**,** Mahendra Sahu, et al.	Obtain information of student's academic progress, degradation in their potentiality, abandonment as well as retention of students [30].	Association rule mining	FP Growth
10.	2010,M.Ramasw ami and R.Bhasakaran	Analyze the connection between variables which helps to predict the outcome of the performance at higher secondary school education [31].	Classification Tree	CHAID
11.	2016, LotfiNajdi, Dr. Brahim ER-RAHA	Enrich the learning characteristics of graduate students and help to adapt teaching strategies according to the identified student profiles [32].	Clustering	K-means
12.	2012, EdinOsmanbego vic, Mirza Suljic	Help students and teachers to enhance student's performance and behavior; reduce failures and improve the quality of learning [33].	Classification	Naive Bayes, multilayer perceptron and C4.5
13.	2012, DorinaKabakchi eva	Develop models of DM to anticipate performance of students based on their personal characteristics [34].	Classification	Rule learner-OneR, C4.5, MLP and K-NN

Table 3 shows the comparison between various techniques of data mining for predicting the accuracy.

Table 3.Comparing Different Methods For Predicting accuracy [35]

S.no	Author	Techniques	Accuracy
1.	Oktariani Nurul Pratiwi 2013 [36]	OneR J48 K Star Naïve Bayes	78.66% 79.61% 74..52% 76.75
2	Ajay Shiv Sharma, S.S 2014 [37]	Logistic regression	83.33%
3	Vikas chirumamilla, B.S. 2014 [38]	Naïve Bayes C4.5	66.18% 77.78%

VI. CONCLUSION

A number of problems are currently faced by academic systems due to various reasons. Through the data mining technology, various techniques are introduced which are considered beneficial for a system to overwhelm the issues and make the advancement in the traits and techniques for academic institutes. In this survey paper, various data mining technologies are elaborated in the literature survey (Table 2) for predicting the accomplishment and demeanor of the students. This can be helpful for an academic system to get improved by enabling better learning process of the students by generating a practical point of view among students to think and to understand. The focused target of Educational Data Mining (EDM) is to provide better academic knowledge to the students, to enhance their behavior and performance and to understand student's detention and contrition by finding new methods of making personalized learning endorsements for each and every student. In this study, a wide sense of the variety of research currently being conducted in EDM was reviewed, by applying data mining methods.

9. REFERENCES

1. Cristobal Romero and Sebastian Ventura, "Data Mining in education" 14 dec 2012.

2. Ritter, S., Harris, T., Nixon, T., Dickison, D., Murray, R., Towle, B. (2009), "Reducing the knowledge tracing space". In International Conference on Educational Data Mining, Cordoba, Spain.

3. Baker, R. S. J. D, "Data mining for education." International encyclopedia of education 7 (2010): 112-118.

4. Jindal and Dutta Borah, "A survey on Educational Data Mining and Research Trends". In International Journal Of Database Management Systems,2013, Vol.5, No.3.

5. Cristóbal Romero and Sebastián Ventura, "Educational Data Mining: A Review of the State-of-the-Art". IEEE transactions on systems, man, and cybernetics—part c.

6. Agathe Merceron, Kalina Yacef(2005), "Educational data mining: A case study" in Proceedings of the 12th international Conference on Artificial Intelligence in Education AIED

7. R. Baker (2010), " Data Mining for Education" In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), vol. 7, pp. 112-118. Oxford, UK: Elsevier.

8. Lotfi NAJDI, Dr. Brahim ER-RAHA, " Use of Unsupervised Clustering to Characterize Graduate Students Profiles based on Educational Outcomes" International Journal of Computer Techniques -– Volume 3 Issue 2, Mar-Apr 2016

9. Tanvir Habib Sardar, Zahid Ansari, "An Analysis of MapReduce Efficiency in Document Clustering using Parallel K-Means Algorithm", Future Computing and Informatics Journal, 2018

10. Density based clustering algorithms-DBSCAN and SNN by Adriano Moreira, Maribel Y.Santos and Sofia Carneiro.

11. Dr. Varun Kumar1, Anupama Chadha, "Mining Association Rules in Student's Assessment Data", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012, ISSN (Online): 1694-0814

12. Han Jiawei and Micheline Kamber, "Data Mining: Concepts and Technique", Morgan Kaufmann Publishers, 2000

13. Romero, c., ventura, s., hervás, c., gonzales, p. (2008), "Data mining algorithms to classify students", International Conference on Educational Data Mining, Montreal, Canada.

14. R.C. Barros et al., "Automatic Design of Decision-Tree Induction Algorithms", SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-14231-9_2, 2015

15. Ajay Kumar Mishra, Bikram Kesari Ratha, "Study of random tree and random Forest data mining algorithms for microarray data analysis", International journal on advance electrical and computer engineering (IJAECE), ISSN (Print): 2349-932X Volume-3, Issue-4, 2016.

16. Sayali D. Jadhav, H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611 Volume 5 Issue 1.

17. Guleria Pratiyush, Sood Manu, "Classifying Educational Data Using Support Vector Machines:A Supervised Data Mining Technique", Indian Journal of Science and Technology, Vol 9(34), DOI: 10.17485/ijst/2016/v9i34/100206, September 2016.

18. Mokhairi makhtar, hasnah nawang, syadiah nor wan shamsuddin," Analysis on students performance using naïve Bayes classifier", Journal of Theoretical and Applied Information Technology 31st August 2017. Vol.95. No.16

19. Agresti, A. (1996), "An introduction to Categorical Data Analysis". Wiley: New York.

20. Pal, B. K. (2011), "Mining Educational Data to Analyze Students Performance". IJACSA .

21. S.Pal, U. a. (2011), "Data Mining :A prediction of performer or under performer using classification"(IJCSIT). IJCSIT .

22. Pooja M .Dhekankar and Dinesh S, "Data Analysis of Student Performance by using Data Mining Concept", International Journal on Recent and Innovation Trends in Computing and Communication Volume 3,Issue: 5, 2015.

23. P.Veeramuthu, Dr.R.Periyasamy, V.Sugasini, "Analysis of Student Result Using Clustering Techniques", International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014.

24. AnduelaLile, "Analyzing E-Learning Systems Using Educational Data Mining Techniques", Mediterranean journal of social science Vol. 2, No. 3, September 2011.

25. Tripti Mishra, Dr. Dharminder Kumar, Dr. Sangeeta Gupta, "Mining Students' Data for Performance Prediction", 2014 Fourth International Conference on Advanced Computing & Communication Technologies

26. ShradhaShet, Gayathri, "Approach for Predicting Student Performance Using Ensemble Model Method", International Journal of Innovative Research in Computer and Communication EngineeringVol.2, Special Issue 5, October 2014.

27. Abdulsalam Sulaiman Olaniyi, Hambali Moshood Abiola, Salau Ibrahim Taofeekat Tosin, Saheed Yakub Kayode, Akinbowale Nathaniel Babatunde, "knowledge discovery from educational database using apriori algorithm", Computer Science and Telecommunications 2017|No.1(51).

28. Omprakash Chandrakar, Jatinder kumar R. Saini, "Predicting Examination Results using Association Rule Mining", International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 1, April 2015.

29. D. Magdalene Delighta Angeline, "Association Rule Generation for Student Performance Analysis using Apriori Algorithm", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 1, March-April 2013.

30. Mahendra Sahu, Smita Bagde, Rubina Sheikh, Narendra Dhawade, "Knowledge Discovery from Student Database using Association Rule Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 10, October 2015.

31. M. Ramaswamiand R. Bhaskaran, “A CHAID Based Performance Prediction Model in Educational Data Mining”, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010.

32. Lotfi NAJDI, Dr. Brahim ER-RAHA, “Use of Unsupervised Clustering to Characterize Graduate Students Profiles based on Educational Outcomes”, International Journal of Computer Techniques -– Volume 3, Issue 2, Mar- Apr 2016.

33. Edin Osmanbegovic, Mirza Suljic, “Data mining approach for predicting student performance”, Journal of Economics and Business, Vol. X, Issue 1, May 2012.

34. Dorina Kabakchieva, “Student performance Prediction by Using Data Mining Classification Algorithms”, International Journal of Computer Science and management Research Vol 1 Issue 4 November 2012.

35. Tripti Dwivedi and Diwakar Singh, ”Analyzing Educational Data through EDM Process: A Survey”, International Journal of Computer Applications (0975 – 8887) Volume 136 – No.5, February 2016.

36. Pratiwi, O. N. (2013), "Predicting student Placement Class using Data mining", 2013 IEEE International Conference on Teaching ,assesment and learning for Engineering(TALE), (p. 618). Bali Dynasty Bali Resort,Kutta Indonesia.

37. Ajay Shiv Sharma, S. P. (2014), "Placement preidction system using Logistic Regression", IEEE

38. Vikas Chirumamilla, B. S. (2014), "A Novel approach to predict student placement Chance with Decision Tree induction", IJST .