Survey on Educational Data Mining using Various Approaches of Data Mining

Abstract— Educational Data Mining shortly known as (EDM) is emerging field to explore the data from various educational contexts. It provides inherent knowledge about imparting the education, which is used to enhance the quality of teaching and learning. Effective planning can provide personalized education. Educational sector produces data in large amount that is too voluminous and complex to understand. Data mining searches through the large amount of dynamically generated data to present users with the useful and understandable patterns and trends. It has the power to use the raw data effectively which has been produced by universities, to draw the hidden patterns and the relationships among the attributes that are used in predicting the student performance, their behavior effectively. This research compares various techniques of DM and tries to find the most efficient technique among them.

Keywords—Wireless sensor network, clustering, style, styling, insert (*key words****)***

1. INTRODUCTION

In a country's life education plays a vital role to ensure the survival of the state and the nation [1]. In today's scenario educational technologies aide the process of learning and teaching (TL) as they are being used in educational domains including the traditional form of classrooms where it's all about face to face and even the learning platforms available online. Educational actors (students, teachers and administrators) have been benefitted as they are provided with the relevant information in which they have to act upon and thereby end up in promoting the quality based innovations in this domain [2]. These days universities are run in a very powerful and dynamically viable manner. A large amount of data is gathered in the form of marks, records, documents, files, performance all related to student performance [3].

Educational data mining act as a bridge between the two one is the education and the other is computer science. The subfields of computer science, Data mining and Machine learning are used. Data mining is used to uncover the hidden patterns in the unstructured data. It is devoted to discover the knowledge and then generate the relevant information. Due to the data mining advancements it has become possible to mine educational data to get the useful data. This relevant information serves to benefit its handlers [4].

The fast growth in the educational domain brought the fact of the distilling the massive data which led to educational data mining (EDM) emergence [5]. Educational data mining is a specific field of data mining that helps in discovering invisible patterns in the data to take decisions for students, teachers and administrators. EDM make predictions which are further characterize the learner behavior, domain content knowledge, educational functionalities, assessment outcomes, applications and the applications. The outcomes can guide the students in learning process, tutors in teaching process to enhance the educational practices and the administrators in managing process [6].

1.1 Phases of Educational Data Mining

The advancements in the EDM are progressively evolving the data mining techniques to inform the educational domains. Ultimately, the objective is to gather relevant information about the learning and teaching for the pedagogical enhancements. Accordingly, EDM has been divided into four phases.

The first phase, discovers the relationships between data using data mining techniques such as classification, clustering, regression, sequential pattern mining and association rule mining.

The second phase, validates whether the selected relationships are theoretically validated or not. If the relations are validate further processing is done otherwise not.

The third phase, predictions are made according to the relationships that are theoretically validated for the future aspects in educational learning and teaching context.

The fourth phase, supports the predictions made in third phase and make policy level decisions which would help in pedagogical improvement in the educational era [7].

Educational touches of data mining and affects several other aspects of education industry. The key components of EDM are - stakeholders of education, various data mining tools and techniques, educational data,

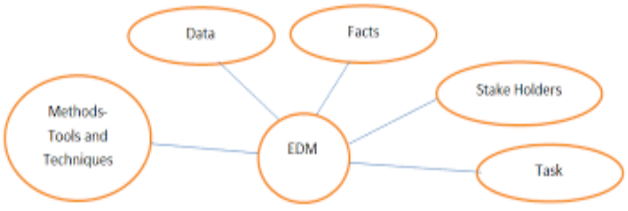

Figure 1: Educational Data Mining Components

Stakeholders

Keeping in mind all the following aspects of the education, that is primary education to the higher education, education stakeholders can be categorized majorly in following:

- Learners/Students: The most essential & component of impacted form is learners. As students are involved directly in the process of the learning, they fall in the primary group of stakeholders. EDM can help them with the personalized education based on

various recommendations and can increase the interestingness of education for students towards learning. Different learning tasks can be formulated in the different group of students based on their needs.

- Faculties: Educators, Teachers, & Instructors are benefitted which can know which user require extra hold up. The student performance prediction becomes simple. Another impact is to help out in classifying of the learners into the groups. It also offer an insight into patterns in to which the students can then learn Regular & irregular. Teachers who can analyze data as well as determine most made errors were at common basis. Further than just the academics, analysis for learning of student's & behavior which can also be able to be done sense if they can need any further support at the time of process of the learning. Teachers can also main stakeholders.
- Parents: Parents are the part of the secondary group. They are liable for helping their kids to get them enroll in the most suitable courses for them.
- Course Researchers and Educational Developers: They are the people who design and modify the course. They are responsible for the growth of education. Developers fall in the group of stakeholders of secondary form .
- Administrators: They could also be able to called as users of hybrid form . EDM is essential for the effective type of utilization of the resources; it help in the determination of what offers which can capture much pupils into different programs as well as courses. They are responsible for various administrative decisions such as infrastructure development and employing the expert faculty.

Data for EDM

Analysis of educational data in large amount is being involved in the decision-making as well as planning for future [8]. The educational data is mix of the structure which has it as well as data of unstructured form being collected from simple as well as sources of complicated form so that this will generate big data. This data could be of the form of the responses shown by huge no: of students to different type of questions or be of large pool of the texts being received by student's from the collectives form eg essays on online as well as other descriptive data of generic form . Such data is from a digital process on a whole those can really be sudivided into 2 types of the info. These 2 categories are then 1st is structured data 2nd is unstructured data.

Structured Data: This data is organized by now & leave possibility at a lesser amount of being too huge and of being too vague. As a o/c, the data is itself the explanatory as well as much more regulated as the unstructured data which is being compared to other. At the given time, this creates data free of charge since the intervention of human as well as brings in the evidence of crystal obvious being away from the prejudices as well as judgments.

Many collecting structured of sources data which could be termed by incorporating the capable form which is embedded & assessment being formative are as follows-

- tutors of Intelligent form
- Simulations
- Semantic tools of mapping
- Learning management systems

Unstructured Data: This form of data doesn't come in from 1 source which is specific & there is no further data model which is predefined. This may contain learning of information which comes from web. This is being included the useful data like a IP address learner's / user's name & may be related to different texts as of the sources eg Internet forums, or audio files ,video clips,.

Here are even sources of potential form where data of unstructured form can be collected from-

- Learning of games
- analyses of Social interaction
- Affect meters
- Body sensors

As can be seen, that educational data are in abundance, and this can makes data mining on educational form an essential exercise which can than make better learning development crosswise all the verticals of education form of industry [9]. All this gathered, structured data & unstructured data when is being analyzed will than be of huge help when it is coming to meeting real goal and to establish specific education goals. The data be both assorted as well as its being classified too. The data is generated from the online sources and offline sources-

- Offline Data: As the name suggests, data which is offline are generated by situations of actual -time and settings situations . Setups eg tests for traditional classroom , contemporary classes based on interaction , teacher interactions and student interactions, actual-time data being derived from various courses as well as different departments of such institute eg schools, colleges, also universities. Some factors are participation levels from students, attendance of students', behavior as well as scores related to attitude.
- Online Data: other than offline data, data which is online is not being dependent upon the kind of location which is geographical . The data are being derived from the weblogs, E-mails, transcripts telephonic conversations, spreadsheets of email medical records, Legal form, and publication databases and many other.

Data Mining Methods

Few of the popular as well as efficient form of methods for data mining is classified as follows [10]:

- Classification: Classification is training technique as well as in testing technique, which categorizes to data which is collected into few of the preset groups. It is a valuable form of technique those for student performances prediction, analysis of risk, systems for student monitoring, as well as for the detection of the errors and many more.
- Clustering: Like the classification, this method puts the same data all together into the clusters, but not for the under the categories of preset form. This method is useful in different learners preferences. Analysis of the students' character of comprehensive

form & method is suitable for collaborative form of learning those are done through the clustering help.

- Statistics: Statistics are being useful for the management system of course and assists in extreme deviations determination from mean. It records, functions of statistical form like the mean, mode & helps in managing of response system of the student.
- Prediction: Extremely useful in passing education industry trends for the future. It is a method that is being utilized to forecast the success rate, rate for dropout & designing ways of the retention.
- Association Rule Mining: This is an important data mining technique in finding various relations among the attributes of data, such as admission, migration, parents-faculty-students relation etc. various patterns for reasons of student's failure can be found out.

The other effective form of methods data mining used in EDM industry neural networks, regression, SVM etc [11].

Functional Tools of the Data Mining on Educational form:

- WEKA named as (Waikato Environment for Knowledge Analysis):The workbench of Weka contains many tools, algorithms method & graphics methods which lead to analysis as well as predictions. Most of the algorithms are inbuilt in this tool.
- KEEL named as (Knowledge Extraction Based on Evolutionary Learning): KEEL application is machine learning software set of which is being designed meant for resolution providing to data mining of numerous problems. It have software techniques collection those are in the manipulation of data & analysis before as well as after process. It is being applied for methods of soft computing in the extracting of useful data about learning of data mining & knowledge.
- KNIME (Konstanz Information Miner):This stage is used in a hole by open source for the analytics of data , reporting, & integration. Generally it is used for the research of pharmaceutical form , this tool of business analysis is presently used in wide form for Educational Data Mining.
- ORANGE: component-based data mining software is what about Orange indicates which suite that is appropriate for the data analysis of explorative form , visualization, & predictions. It must operate different exploration methods perfectly and also aids in the scoring & data filtering as a part of operation for post-processing.

2. LITERATURE SURVEY

Pandey and Pal [12] conducted a study on new comer students will performer or not on the basis of student performance choosing students of 600 no: from various colleges of the Dr.R.M.L. Awadh University, in Faizabad, India. They have been applied in the Bayes Classification on the Category, Language & qualification of background, in conclusion they found of that whether newcomers will perform or are not.

Romero, Cristóbal, et al. [13] compared in the various methods of data mining & techniques for the classifying of students based on the final marks received into their relevant courses. They have used with the real data of the 7 model courses with students of Cordoba University . In their applied method they found that a classifier model is sufficient for educational form of use which has to be accurate and the comprehensible meant for teachers for making decisions.

Galit [14] perform a case study that uses student's data to predict and analyze their learning behavior and to warn students at risk before their final exams.

Osmanbegović, Edin, and Mirza Suljić [15] from University of Tuzla perform data mining techniques and methods for comparing the prediction of students' success, data collected applying from surveys which is conducted at the time of summer semester by Faculty of Economics, academic year 2010-2011, among 1st year students and investigated the result of students' achieved from the high school form and from entrance exam, & effect on success.

Pandey and Pal [16] perform a study on student performance by choosing 60 no: of students from degree college of the Dr. R. M. L. Awadh University, in Faizabad, India. They applied association rule for find interest of the student in opting class for teaching the language.

Cortez, Paulo, and Alice Maria G. Silva [17] used data mining techniques to forecast student's achievement of secondary by school using the real-world data. The 2 core classes were modeled. 4 Data Mining models (i.e. Neural Networks, Decision Trees, SVM and Random Forest) and 3 input selections (e.g. without previous grades and with) were tested. The results shows predictive accuracy can be achieved, although student achievement also depend other relevant.

Han and Kamber [18] describes the software of data mining which allow users so as to analyze the data from various dimensions, various form to categorize it & summarize relationships which then identified at times of the process of mining .

Kumar, S. Anupama, and M. N. Vijayalakshmi [19] perform for the study on the internal assessment for student's data so as to predict their presentation in last exam. They used C4.5 decision tree algorithm. The algorithm accuracy is being compared through algorithm of ID3 also need to be much more effective in form of predicting to the point where time is taken to derive the tree and outcome of the student.

Bhardwaj and Pal [20] perform a study on BCA known as (Bachelor of Computer Application) course of the Dr. R. M. L. Awadh University, in Faizabad, from India. The performance of student is based by just selecting 300 no: of students from the 5 diverse college degree. They used classification method on 17 attribute, and found that the factors like living location, medium of teaching, students_ grade in senior secondary exam, students other habit, mother's qualification, student's family status were highly correlated with the student academic performance and family annual income.

Khan [21] described a performance on the study which include no: of 400 students in which their are 200 boys & 200 girls those are selected from school of senior secondary of the place Aligarh Muslim University, Aligarh, India aim of this research to analyze the rate of success of students in higher secondary in science group. Cluster sampling technique was used to analyze this study. He found that girls who had good income, education, occupation, wealth and place of residence, got greater achievement on the other hand the boys with low living status had relatively greater academic gain also.

3. COMPARISON OF VARIOUS DATA MINING TECHNIQUES

While comparing various data mining techniques, there are some specific steps upon which we have to concentrate and follow the steps accordingly. The methodology consists of collection of dataset, pre-processing the dataset and applying data mining over the datasets to get the result. This research analyses various classification techniques of data mining over the students' performance dataset and tries to find the estimate of the accuracy depending on various parameters.

Naïve Bayes: classifiers of Naive Bayes are then be a collection of algorithms of classification those based on the **Bayes' Theorem**. It's not single algo although a algorithms family where several share general principle, that is each pair of the features is being classified is being independent of every other.

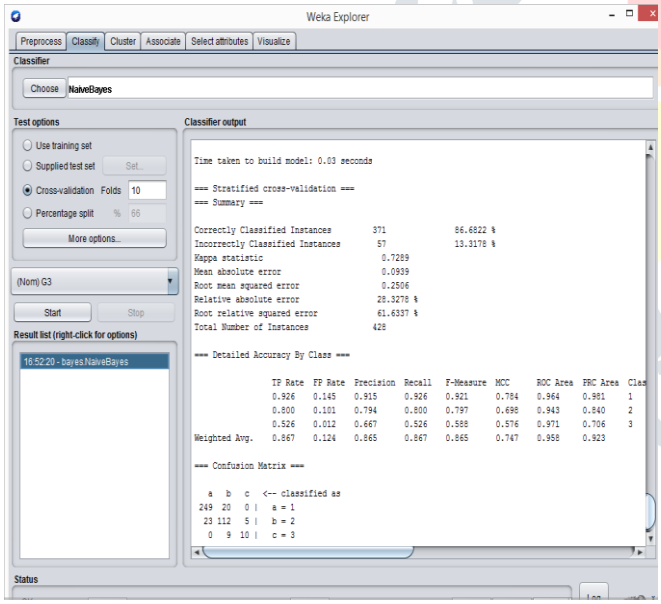

Figure 2 Naïve Bayes algorithm

Decision Table: here Decision table testing is technique for software testing which is being used to test behavior of system for various combinations of i/p. This is a approach of systematic form where the various i/p combinations & their related form of system behavior (Output) are being captured in the tabular form.

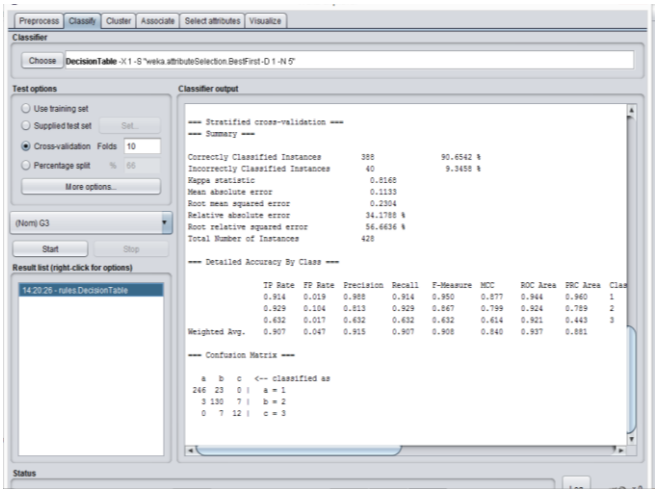

Figure 3 Decision table algorithm

Decision Stump: Here a decision stump is the model of machine learning those consisting of decision tree for one-level . That is, it's decision tree with a internal node (root) which is connected immediately to nodes which is terminal (its leaves). A decision stump forms a prediction which is fully based on value of the feature of single input. at times they are being called as 1-rules. Those Depending on input feature type , there are possibility of several variations .

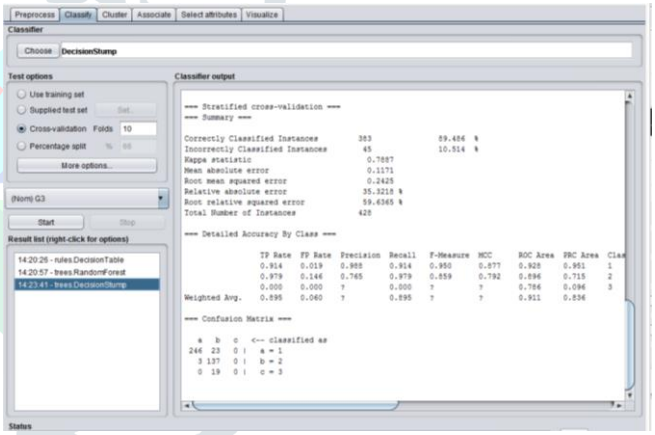

Figure 4 Decision stump algorithm

BayesNet: BayesNet B = < N, A, 0 > is shown as a directed acyclic graph shortly called (DAG) with a distribution of conditional probability (CP table) for each node, collectively represented by α. Each node n ∈ N represents a domain variable, and each arc a ∈ A between the nodes those being represents a dependency of probabilistic form (see Pearl 1988). Normally, a BN can be able to be used to compute probability in a conditional way of 1 node, shown values being assigned to different nodes; for the same reason, a BN could be used as classifier which shows the distribution of the posterior probability of classification node by giving the other attributes values .

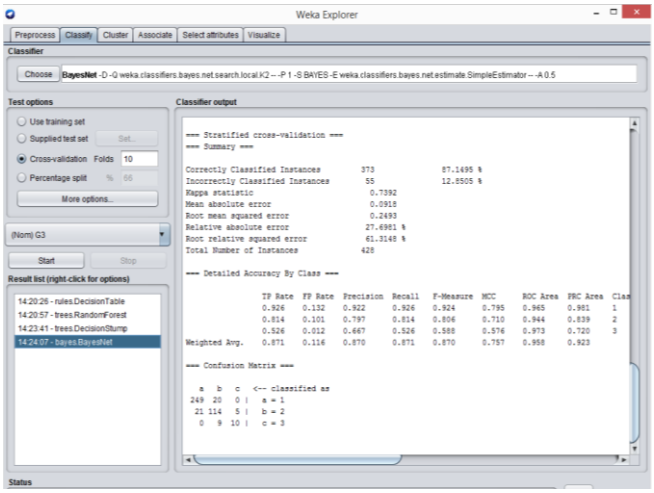

Figure 5 BayesNet algorithm

As we have seen the visualization of various DM techniques we will now compare the accuracy and time taken by each algorithm which will help to create a rough estimation of the accuracy of classification techniques (see table1).

Table 1 Classified Instances

Algorithms	**Correctly classified instances**	**Incorrectly classified instances**
Naïve Bayes	371	57
Decision table	388	40
Decision Stump	383	48
BayesNet	373	55

Performance is being computed via asking classifier to also offer their best guess with the reference to the classification which is for each occurrence in set of test . Then the classifications is predicted which are being compared to real classifications to determine the accuracy. Therefore, if you flip values of 'correct' form that you have given it, the results will be flipped also. Other than classified instances there are various other measures on which our results relies on. Table 2 shows the evaluation criteria for each and every algorithm.

Table 2 Various parameters for classification

Evaluation Criteria	**NaïveBayes**	**Decision table**	**Decision stump**	**Bayesnet**
Kappa statistics	0.7289	0.8168	0.7887	0.7392
Mean absolute error	0.0939	0.1133	0.1171	0.0918
Root mean squared error	0.2506	0.2304	0.2425	0.2492
Relative absolute error	28.3278%	34.1788%	35.3218%	29.6981%

4. CONCLUSION

EDM is upcoming field which is for exploring data in the educational context by applying dissimilar techniques of DM. It offers essential knowledge of the teaching as well as process of learning for education planning which is effective. In this paper, different classified techniques are applied on database of student's to study the students performance who really require a special attention during their study and will able to take appropriate steps at right time. This empirical o/c presents that we be able to produce short however the accurate attributes prediction that help the education professionals as well as institutions to predict such type of students who have the worst performance; they can develop the special techniques to improve them with special attention with confidence.

In future, we will try to find more efficient results of this database by any other technique of classification, clustering or association rule mining.

REFERENCES

[1] Sugiyarti, E., Jasmi, K. A., Basiron, B., Huda, M., Shankar, K., & Maseleno, A. (2018). Decision support system of scholarship grantee selection using data mining. International Journal of Pure and Applied Mathematics, 119(15), 2239- 2249.

[2] Rodrigues, M. W., Zárate, L. E., & Isotani, S. (2018). Educational Data Mining: A review of evaluation process in the e-learning. Telematics and Informatics.

[3] Tegegne, A. K., & Alemu, T. A. (2018). Educational data mining for students' academic performance analysis in selected Ethiopian universities. Information Impact: Journal of Information and Knowledge Management, 9(2), 1-15.

[4] Sheshasaayee, A., & Bee, M. N. (2018). E-learning: Mode to Improve the Quality of Educational System. In Smart Computing and Informatics (pp. 559-566). Springer, Singapore.

[5] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. IEEE Access, 5, 15991- 16005.

[6] Alom, B. M., & Courtney, M. (2018). Educational Data Mining: A Case Study Perspectives from Primary to University Education in Australia.

[7] Tavares, R., Vieira, R., & Pedro, L. (2017, November). A preliminary proposal of a conceptual educational data mining framework for science education: Scientific competences development and self-regulated learning. In Computers in Education (SIIE), 2017 International Symposium on (pp. 1-6). IEEE.

[8] Zhang, W., & Qin, S. (2018, March). A brief analysis of the key technologies and applications of educational data mining on online learning platform. In Big Data Analysis (ICBDA), 2018 IEEE 3rd International Conference on (pp. 83-86). IEEE.

[9] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177- 194.

[10] Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies, 23(1), 537-553.

[11] Manjula, M. (2018). A Systematic Review on Educational Data Mining. International Journal of Scientific Research in Science, Engineering and Technology, 164 -170.Chong, M., A. A., et al. (2005). "Traffic Accident AnalysisUsing Machine learning Paradigms." Informatica 29(1).

[12] U . K. Pandey, and S. Pal, .Data Mining: A prediction of performer or underperformer using classification., (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), pp.686-690, ISSN:0975-9646, 2011.

[13] Romero, Cristóbal, et al. "Data mining algorithms to classify students." Educational Data Mining 2008. 2008.

[14] Galit.et.al, .Examining online learning processes based on log files analysis: a case study. Research, Reflection and Innovations in Integrating ICT in Education 2007.

[15] Osmanbegović, Edin, and Mirza Suljić. "Data mining approach for predicting student performance." Economic Review 10.1 (2012).

[16] U. K. Pandey, and S. Pal, .A Data mining view on class room teaching language., (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[17] Cortez, Paulo, and Alice Maria Gonçalves Silva. "Using data mining to predict secondary school student performance." (2008).

[18] J. Han and M. Kamber, .Data Mining: Concepts and Techniques,. Morgan Kaufmann, 2000.

[19] Kumar, S. Anupama, and M. N. Vijayalakshmi. "Efficiency of decision trees in predicting student's academic performance." First International Conference on Computer Science, Engineering and Applications, CS and IT. Vol. 2. 2011.

[20] B.K. Bharadwaj and S. Pal. Data Mining: A prediction for performance improvement using classification., International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[21] Z. N. Khan, .Scholastic achievement of higher secondary students in science stream., Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.