IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 31-35

A Comparative Study on Educational Data Mining Using Classification Techniques

N. Mohamed Farook Ali[1,] Dr. N. Sasirekha[2]

[1]*Ph.D Research Scholar, PG and Research Department of CS Vidyasagar College of arts and science, Udumalpet.*

[2]*Assistant Professor, PG and Research Department of CS Vidyasagar College of arts and science, Udumalpet.*

Abstract: *Most of the institutions are looking forward for the solution for student's better performance in their Education. Every institution is ready to implement different methods to understand students thinking and their needs. Educational data mining (EDM) is a most important and focused research area. In EDM different type of student data were collected and implemented through various data mining techniques. Classification Algorithms like Decision trees, K- Nearest Neighbor, Neural Networks, Naïve Bayes and so on were used. This paper on the EDM compares the classification algorithms used in EDM..*

Keywords: *Educational Data Mining, Prediction, Classification Algorithms.*

I. INTRODUCTION

Educational Data Mining is consider as an emerging data mining research area, and used to develop methods for exploring hidden and unique large-scale data that come from educational institutions and through those methods the better understanding of students is possible.

Educational data is collected from students through the interactive learning environments, administrative of colleges, schools or universities and computer based collaborative and sometime it has multiple levels of meaningful hierarchy, the properties of the data itself determines those things. The following factors are plays an important role in study of educational data such as time, sequence and context [9].

Data mining is very popular for its decision making process. It discovers novel and potential useful information from vast amount of data, so it is also known as Knowledge Discovery in Database (KDD). In recent years the researchers have increasing interest in educational research. In Educational Data Mining there are many techniques are used such as Naïve Bayes, Neural Network, Fuzzy Logic, Genetic Algorithm and so on.

It is a big challenge to measure academic performance of students because it is based on different factors like psychological factors, understanding levels, capacity to learn, performance in exams and so on. So the research's main scope is to find out the factors that affect students performance.

This study focuses on results of different Educational Data Mining research and its accuracy based on selection of different algorithms. The researchers collected data about students with different aspects. The collected data is analyzed with various data mining classification algorithms to predict the future results based on student performance. The algorithm which gives the significant result is known as better algorithm for a particular group of data and this is achieved through the data mining methods [5].

II. LITERATURE SURVEY

There are many papers published in the of Educational Data Mining research. Much research is done in predicting academic performance of a student. So in this literature survey some of the research results regarding student performance and future predictions are analyzed and discussed.

Amjad Abu Saa (2016) conducted a research work and they used classification techniques to predict results. There are multiple classification techniques available in data mining such as, Decision trees, K-Nearest Neighbor (K-NN), Neural Networks and Naïve Bayes. In this research 270 students records were collected. The possible values are retrieved from each record and also it describes attributes of each data. From the result Naïve Bayes model produces the high potential results also different data patterns are found through Naïve Bayes model. Finally found that the performance of the student is not totally based on their academic efforts, in spite, some other related factors that have equal to high influences as well[1].

Ahmed Mueen, Bassam Zafar and Umar Manzoor (2016) conducted a similar research on Educational Data Mining; Student's performance was predicted based on student's academic records and their forum participation. For this research two undergraduate course data were collected. To predict student's performance they used three classification models like Naïve Bayes, Neural Networks and Decision Trees. The results show that Naïve Bayes model gave better result comparing to other two models by obtaining the overall prediction accuracy of 86% [2].

Ashwin Satyanarayana, and Mariusz Nuckowski (2016) conducted a survey over 788 school student dataset which includes 37 questions each. In this research multiple classifiers were used like Decision Trees, Naïve Bayes and Random Forest to get accuracy over student's data by eliminating noisy instances. From this study association rules were identified that affect student's results using a set of rule based techniques like Apriori, Filtered Associator and Tertius. In this regard the result was found that prior work there was no filtering on student data has been performed and focused on using single classifiers. So in this study comparison of single filters and ensemble filters was done and it is concluded that ensemble filters works better for identifying and removing noisy instances [3].

Another comparative study was done by Bhrigu kapur, Nakin Ahluwalia and Sathyaraj (2017). They compared six algorithms like J48, Random Forest, Naïve Bayes, Naïve Bayes Multinomial, K-Star and IBK. They used 480 entry of data set and implemented through Weka tool. The Survey conducted based on seven attributes and found Random Forest algorithm provides more accuracy compared to other algorithms [4].

K. Prasada Rao et. al (2016) [5] conducted a survey over 200 college students. In this research classification techniques were used on student database to predict the learning behavior of student's. From this research, the researcher identified the slow learners, and effectively the action taken to rectify the failures and take appropriate action to qualify the weaker students in perfect manner. In this study the performance of J48, Naïve Bayes and Random forest algorithms were compared. Finally the researcher got accuracy using Random forest algorithm when the data set is in massive size.

A research carried out by the team [6] (2016), and the performance of the student's were predicted. Classification techniques were used to create prediction module of the system to predict the future values. Various parameters like previous academic performance were considered to predict student's academic results and placement. The dashboard is the module which describes the whole overview of the institution in a graphical representation of data. Decision tree algorithms ID3 and C4.5 were implemented to generate reports based on structured database. From this research ID3 algorithm provided the best accuracy of 95.33%.

U. K. Pandey, and S. Pal (2011) [3] [7] conducted a survey with 200 student's records. To find student's division based on previous year database they used Bayesian classification method. Through this research the dropout ratio of student's were reduced. This study also used to find the student's who those are all need special attention to reduce failing ratio and necessary action to be taken at right time.

Another survey conducted (2017) [8] to predict academic performance of final year engineering student's. They collected last 2 to 5 years passed out student's data of various departments. The algorithm used for the system was the ID3 algorithm. There are two formats of input. First type of input was data collected through survey and the second type was data entered manually by each student into a forum. Thus the classification and analysis of the system was performed. Based on various attributes, overall student's performance is provided.

III. PREDICTIVE ANALYTICS

When there is a situation to predict the future events, there is a need to perform predictive analytics process. Predictive analytics is the branch of the advanced analytics. To make prediction about future results, the management, business process modeling and information technology should bring together and it uses a number of data mining predictive modeling and analytical techniques. The patterns which are found in from implemented data can be used to identify the risk and opportunity for future. To assign a score or value to a particular set of conditions, the predictive analytics models should identify relationships among different factors to access risk.

Predictive analytics performs the process of uncovering the patterns and relationships in both the structured and unstructured data. The organization is creating it through the data mining and text analytics along with statistics. For the analysis process the structured data can be used like name, age, gender, address and course of a student. The textual data retrieved from social media or from a note from any enquiry centre or any type of open text which used to be extracted from the plain text are used in the model building process is known as unstructured data.

Predictive analytics is used to suggest actions or to perform actions already decided for the future benefit of a organization. It also provides not only single option but also provides many different decision options to get benefits from prediction and its implications. So, an organization become proactive, forward looking, anticipating outcomes and act well based on the data not on the assumption [15].

IV. DATA MINING PREDICTION USING CLASSIFIERS

4.1 Classification

In the field of Data Mining Classification is a main and most important technique. Class and category of a value is identified using classification based on the previously categorized values. Some of the important classification techniques are discussed.

4.2 Classification Algorithms used for Prediction

Naive Bayes Algorithm: Naive Bayes Algorithm is considered as very efficient and easy algorithm. The classification rate is considerably very high and most of the cases it predicts accurate results. This algorithm produces good results only when the data set is very large [10].

Random Forest Algorithm: To improve the predictive accuracy, Random Forest Algorithm uses average values of the model. It is a meta-estimator, so it fits a number of decision trees on various sub samples of datasets. It also controls over fittings. The original sample size is always matches the sub-sample size but the samples are drawn with replacement [11].

Neural Network Algorithm: Neural Network Algorithm endeavors to recognize hidden relationships in a set of data through many processes like the way the human brain operates. A Neural Network Algorithm is a series of algorithms and it produces the biggest result without requiring redesigning the output criteria when it adapt to changing input [12].

K-Nearest Neighbor Algorithm: KNN doesn’t allow the classes to be separated linearly; it is a main and distinct advantage of KNN algorithm. It handles noisy data well enough. When the dataset is very large KNN consumes more time to process, it is considered as main drawback [10].

J48 Algorithm: The ID3 algorithm’s features are extended into J48 Algorithm. J48 has the following additional features such as, it accounts the missing values, decision tree pruning, continuous attribute value ranges, derivation of rules, etc. it is an open source java implementation of the C4.5 Algorithm in WEKA Data Mining tool [13].

ID3 Algorithm: ID3is a decision tree algorithm. Compared to C4.5 decision tree algorithm, the ID3 Algorithm gives accurate results. Space consumption of ID3 Algorithm is very small and its detection rate is very high. The rule which is produced by ID3 Algorithm is very hard to prove. The searching time is comparatively very high. ID3 Algorithm needs a large memory to store a tree structure [10].

C4.5 Algorithm: It is also a decision tree algorithm. Some of the features of C4.5 algorithm is as follows, it can be easily interpreted and very easy to implement. It accepts both continuous and discrete values. Some of the limitations are, when a small deviation in data can lead a completely different decision tree. It cannot work with small dataset [10].

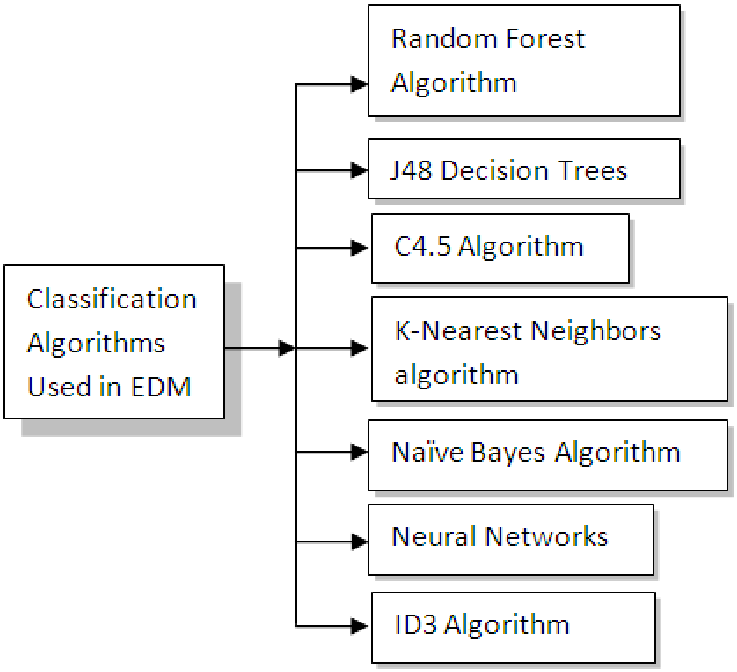

Figure 1: Classification Algorithms used in EDM [14].

The pros and cons of different classification algorithms are discussed in the following table. Through this table there is a possible to know the limitations of the algorithms.

Table 1: Pros and Cons of Classifiers.

Classifiers	**Pros**	**Cons**
Naive Bayes	Easy to interpret and has the capability of updating and reasoning	Needs a large training data
Random Forest Algorithm	Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases.	Slow real time prediction, difficult to implement, and complex algorithm.
Neural Network Algorithm	Can work with incomplete data and has the capability of both updating and reasoning	Difficult to deal with missing data and does not support mixed variable. Needs a lot of data to train.
K-Nearest Neighbor Algorithm	Can work with incomplete data and has the capability of updating	Needs a lot of data and it can’t deal with missing data
Decision Tree Algorithms	Easy to interpret and has capability of reasoning	Needs lot of data, difficult to deal with missing data

The following table describes objective of the research and algorithms which are used by different researchers and shows the best algorithm from the set of algorithms based on their result accuracy.

Table-2: Classification Algorithms compared in EDM

Reference	Problem / Objective	Algorithms / Methods	Best Algorithm
1	To create a qualitative model which best classifies and predicts the students performance based on related personal and social factors.	ID3 Decision trees, K-Nearest Neighbor (K-NN), Neural Networks, Naive Bayes	Naive Bayes
2	To apply data mining techniques to predict and analyze student academic performance.	Naive Bayes, Neural Networks, Decision Trees	Naive Bayes
3.	Using filters to identifying and eliminating noisy instances.	J48, Random Forest, Naïve Bayes	Naive Bayes
4	Predicting student performance using classification techniques. Compare and predict well performed Algorithm	J48, Random Forest, Naïve Bayes, Naive Bayes Multinomial.	Random Forest
5	Compares the performance of J48, Naïve Bayes and Random forest algorithm to predict better accuracy.	J48, Naive Bayes, Random forest	Random forest
6	To generate query specific reports of the academic performance of a group of students.	ID3, C4.5	ID3
7	Reduce the failing ratio and reduce the dropout ratio through classification algorithm	Naive Bayesian classification	Naive Bayes
8	To create a predictive model for academic performance, and the model whose construction achieved through ID3 algorithm.	ID3	ID3

V. CONCLUSION

Student performance prediction has become very popular in Educational Data Mining. It is used to improve the performance of students and also improves the quality of the institution. In EDM, Classification Algorithms are used to predict the future results. Many researchers have been done research to predict student performance. In this paper, different classification algorithms were analyzed based on the results, accuracy and performance, best algorithm is reported. Different classifier algorithms used for EDM by different authors are Naïve Bayes, K-Nearest Neighbor, Random Forest, Neural Networks, C4.5, J48, ID3. Form these algorithms, most of the cases Naïve Bayes algorithm produced better accuracy than other algorithms. This study paper will be a base for the further research in the field of Educational Data Mining.

REFERENCES

[1]. Amjad Abu Saa. (2016) “Educational Data Mining & Students’ Performance Prediction” International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.

[2]. Ahmed Mueen, Bassam Zafar and Umar Manzoor. (2016) “Modeling and Predicting Students’ Academic Performance Using Data Mining Techniques” I.J. Modern Education and Computer Science, 2016, 11, 36-42.

[3]. Ashwin Satyanarayana, Mariusz Nuckowski, “Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance” Spring '2016' Mid. Atlantic 'ASEE' Conference, 'April' 8.9,'2016' GWU.

[4]. Bhrigu Kapur, Nakin Ahluwalia and Sathyaraj R, “Comparative Study on Marks Prediction using Data Mining and Classification Algorithms”, International Journal of Advanced Research in Computer Science, 8 (3), March-April 2017,632-636.

[5]. Prasada Rao, K. , M. V.P. Chandra Sekhara, and B. Ramesh. "Predicting Learning Behavior of Students using Classification Techniques." International Journal of Computer Applications (0975 – 8887) Volume 139 – No.7, April 2016.

[6]. Siddhi Parekh, Ameya Nadkarni, and Riya Mehta (2016) "Results and Placement Analysis and Prediction using Data Mining and Dashboard." International Journal of Computer Applications (0975 – 8887) Volume 137 – No.13, March 2016 In Proceedings of the 22nd International Conference on World Wide Web, pp. 413-418. ACM.

[7]. 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.

[8]. Chandini Lulla, Yash Agarwal, Snehal Kankariya, Prateek Sakaray, Pankaja Alappanavar,” Student Academic Performance Prediction using Machine Learning and Data Mining Techniques”, IJCSMC, Vol. 6, Issue. 5, May 2017, pg.301 – 307.

[9]. http://educationaldatamining.org.

[10]. Sagardeep Roy, Anchal Garg “Analyzing Performance of Students by Using Data Mining Techniques- A Literature Survey” 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)GLA University, Mathura, Oct 26-28, 2017.

[11]. https://www.analyticsindiamag.com/7-types-classification-algorithms.

[12]. https://www.investopedia.com/terms/n/neuralnetwork.asp.

[13]. Gaganjot Kaur, Amit Chhabra “ Improved J48 Classification Algorithm for the Prediction of Diabetes” International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014.

[14]. https://data-flair.training/blogs/classification-algorithms.

[15]. https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/