



## A STUDY ON EDUCATIONAL DATA MINING TO IMPROVE STUDENTS PERFORMANCE USING CLASSIFICATION ALGORITHM

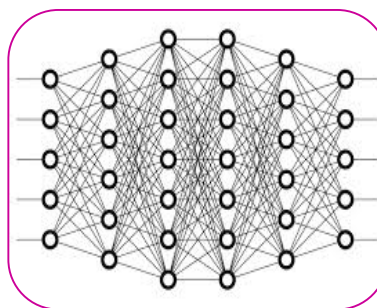
Praveen Kumar A. P.<sup>1</sup> and Dr. M. Sundararajan<sup>2</sup>

<sup>1</sup>Ph.D Research Scholar, Thiruthangal Nadar College, Chennai.

<sup>2</sup>Assistant Professor, Department of Computer Science, Nandanam College of Arts & Science, Chennai.

### ABSTRACT

*Educational data mining is large, complex and various platforms for finding out missing values as well as to improve students' performance with the use of data mining techniques. KDD is a part of data mining which can be a challenging issue is among worldwide. Knowledge discovery data process in the data mining to improve students' performance in self financing institutions can be possible by discovering necessary information from various students' dataset. The KDD process is affected zone patterns can be identified with the proper implementation of data mining technique. This paper focuses about educational data mining and research work carried out on data mining technique to improve academic student result and activities. This paper also focuses on taking a various students activities and based on result points and techniques adopted by different researches, and discusses about the recent and effective algorithm to short out diabetes mellitus.*



**KEYWORDS :** KDD, Classification Algorithms, Cluster, Outlier Detection, Data mining.

### I. INTRODUCTION

#### I. Introduction to Educational Data Mining with Data Mining

Educational Data Mining organizations generally adopt information technology to reduce cost as well as improve efficiency and quality, and with it, research is known to be effective. It is especially used when it draws information from multiple sources of posing special problems and prosperity of data available within the data mining system. Data mining is the application of algorithm for extracting patterns from large volume of data. Educational data mining is the increasing research area in data mining skill. Data mining holds great shows potential for educational organization to allow various education systems to scientifically use data and investigation to get better the universities quality of education and decrease the result education system.

Analysing student's data and information to classify students, or to create decision trees or association rules, to make better decision or to enhance student's performance is an interesting filed of research , which mainly focuses on analyzing and understanding student's educational data that indicates their educational performance , and generates specific rules, classification, and prediction help students in their future educational performance.

Classification is the most familiar and most effective data mining technique used to classify and predict values. Educational Data Mining (EDM) is no exception of this fact, it will use in this research paper to analyze collect students information through a survey, and provide classification based on the collected data to predict and classify students performance in their upcoming semester. The objective of this study is to identify relations between student's personal and social factors, and their academic performance.

This newly discovered knowledge can help students as well as instructors in carrying out better enhanced educational quality, by identifying possible underperformance at the beginning of the semester/year, and apply more patterns to them in order to help them in their education process and get better marks. There are multiple different classification methods and techniques used in knowledge discovery and data mining. It is mostly prejudiced by the social, economic conditions and health policies of the place and varies across different countries, individuals, groups etc. This paper focuses on survey, statistics and analysis on calculate diabetes. Decision Tree and Naïve Bayes algorithm are used to be relevant on a pre-existential data set of forecast diabetes and it consequences in statistics accurateness and presentation improvements.

## II. Goals for Educational Data Mining

It describes the following four goals of EDM:

- ❖ **Predicting student's future learning behavior**
- ❖ **Discovering or Improving domain models**
- ❖ **Studying the effects of educational support**
- ❖ **Advancing scientific knowledge about learning and learners.**

### Predicting student's future learning behavior:

With the use of student modeling, this goal can be achieved by creating student models that incorporate the learners characteristics, including detailed information such as their knowledge, behaviors and motivation to learn.

### Discovering or improving domain models:

Through the various methods and applications of EDM, discovery of new and improvements to existing model is possible.

### Studying the effects of educational support:

It can be achieved through learning systems.

### Advancing scientific knowledge about learning and learners:

By building and incorporate students models, the field of EDM research and technology and software used.

## III. Methods of Educational Data Mining

There are so many methods of educational data mining but all kind of methods lie in one of following specified categories:

**Prediction:** Ryan S.J.D Baker has given a detail explanation of prediction in his paper. He mentioned that "In prediction, the goal is to develop a model which can infer a single aspect of data from some combination of other aspects of data. If we study prediction extensively then we get three types of prediction: Classification, regression and density estimation. In any category of prediction the input variables will be either categorical or continuous. In case of classification, the categorical or binary variables are used, but in regression continuous input variables are used. Density estimation can be done with the help of various kernel functions.

**Clustering:** In clustering technique, the data set is divided in various groups, known as clusters. When data set already specified, then the clustering is more useful. As per clustering phenomenon, the data point of one cluster and should be more similar to other data points of same cluster and more dissimilar to data points of another cluster. There are two ways of initiation of clustering algorithm. Firstly start the

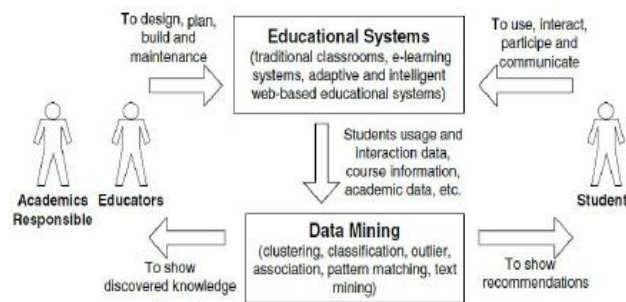
clustering algorithm with no prior assumptions and second is to start clustering algorithm with a prior postulate.

**Relationship Mining :** Relationship mining generally refers to contrive new relationship between variables. It can be done on a large data set, having a number of variables. Relationships mining is an attempt to discover the variables which is most closely associated with the specified variables. There are four types of relationship mining: association rule mining, correlation mining, sequential pattern mining and causal data mining . Association data mining is based on if-then rule that is if some particular set of variable value appears then it generally have a specified value. In correlation mining, the linear correlations are discovered between variables.

**Discovery with Models:** It includes the designing of model based on some concepts like prediction, clustering and knowledge engineering etc.. This new created model’s predictions are used to discover a new predicted variable.

**IV. DATA MINING PROCESS**

The objective of this study is to discover relations between student’s personal and social factors, and their educational performance in the previous semester using data mining tasks. Henceforth, their performance could be predicted in the upcoming semesters. Correspondingly a survey has constructed with multiple personal, social and academic questions which will later be preprocessed and transformed into nominal data which will use in data mining process to find out the relations between the mentioned factors and student’s performance. The student’s performance is measure and indicates by Grade Point Average (GPA) which is real number out of 4.0. This study has conduct on a group of students enrolled in different colleges in various Universities and Autonomous colleges in tamilnadu state. Fig.1 showing educational data mining cyclic process.



**Table1. Attributes Description and Possible Values**

ATTRIBUTE	DESCRIPTION	POSSIBLE VALUES
GENDER	Student’s gender	{Male; Female}
FIRST LANG	First Language	{Tamil,English,Hindi,Sanskrit,Malayalam}
COMM	Category	{OC,BC,MBC,DNC,SC,ST}
TEACHLANG	Teaching Method Medium	{Tamil, English}
HSP	High school Percentage	{Excellent,Very Good, Good, Poor}
LOC	Living Location	{Urban, Suburban, Rural}
SPON	Does the students have sponsor	{Yes, No}
DISCOUNT	Students discounts	{Yes, No}
TRANSPORT	How the students comes to the University or Colleges	{Private Car, University Bus, Public Transport, Own Vehicle }
INCOME	Parents Annual Income	{Low, Medium, High}
GPA	Previous semester GPA	{>3.60(Excellent),3.00-3.59(Very Good),2.50-2.99(Good),<2.5 (Pass)}

**Table 2 ALGORITHMS USED IN EXISTING RESEARCH WORKS**

ALGORITHM	OPERATION	STATUS	ALTERNATIVE	RESULTS
Fuzzy Logic Set, DT[13]	Accuracy, Sensitivity	Efficiency	C4.5	100/78.79
J48[6]	Accuracy	Detection	KNN	73.8/76.3
Naïve Bayes, J48[7]	Performance	Detection	DT	76.30/73.69
Particle Swarm optimization[5]	Accuracy, Sensitivity	Detection	ANFIS	99.8
Naïve Bayes,J48[2]	Accuracy	Avoid	SVM	79.56/76.95
Naïve Bayes[4]	Accuracy	Avoid	SVM	99.2/99.6
Naïve Bayes, Logistic[3]	Accuracy	Detection	KNN	74.4/56.7
TPNN, eTPNN, sTPNN[8]	Accuracy, Sensitivity	Detection	ANN	88/90/90
FCM[1]	Accuracy	Detection	SVM	59.5/95.7
ANN-K-Means[15]	Accuracy	Avoid	K-Means ANN	99.20/97.5
Naïve bayes, C4.5[12]	Accuracy	Detection	K-Means	90.9/100

Various algorithms and applications which are used by many research experts are précised in the above table 2 to get best algorithm for finding the accuracy, specificity, and error rate in prediction in educational data mining student's results and the best are J48, Naïve Bayes, and AGKNN algorithms.

## V. DISCUSSION

It is obvious from the literature that the Accuracy, sensitivity, specificity, error rate, was decided by the dataset taken from various students. Efficient algorithms, Tools or Simulation models improve the Accuracy of the classification and prediction. Majority of the research works in Educational system had been used the 'WEKA Tools' for their implementation; subsequently they obtained the better results followed by the implemented using MATLAB. Naïve bayes was frequently used and popular algorithm to classify and predict in the ground of health care. The Next most used algorithm is J48. The high level of accuracy has been produced using Naïve bayes algorithm.

## VI. CONCLUSION

In order to scrutinize the better classification algorithm to predict students performance in various universities, related research papers have been gathered. This study shows that the best part of using Tools and algorithms. In this respect we concluded that the WEKA Tools and Naïve Bayes algorithm gives the better accuracy in the countryside of educational system. A further study is under progress to implement with actual dataset of researchers research work. Appearance of comprehensive algorithms and tools will be presented in upcoming paper.

## REFERENCES

- [1]. Baradwaj, B.K. and Pal S., 2011. Mining Educational data to analysis students performance. (IJACSA) International Journal of Advanced Computer Science and Applications,vol.2,No.6,2011.
- [2] Ahmed,A,B,E,D.and Elaraby,I.S., 2014. Data Mining : A prediction for student's performance using classification method. World journal of computer Application and technology,2(2),pp.43-47.
- [3]Pandey,U.K and Pal,S.,2011.Data mining : A prediction performer or underperformer using classification.(IJCSIT) International Journal of Computer Science and Information technologies ,Vol.2(2),2011,686-690.
- [4] Bhardwaj,B.K. and Pal,S.,2012. Data Mining: A prediction for performance improvement using classification. (IJCSIS)International Journal of Computer Science and Information Security,Vol.9,No.4,April 2011.

- 
- [5] Yadav,S.K., Bharadwaj,B, and Pal, S.,2012. Data Mining Applications: A Comparative Study for predicting student's performance. International Journal of Innovative Technology& Creative Engineering(ISSN:2045-711), Vol.1, Nov.12,December.
- [6] Yadav,S.K, and Pal, S.,2012. Data Mining: A prediction for performance improvement of engineering students using classification. World of Computer Science and Information Technology Journal (WCSIT).(ISSN:2221-0741), Vol.2, No.2.51-56,2012.
- [7] Naeimeh Delavari, Somnuk Phon-Amnuaisuk," DataMining Application in Higher Learning institutions", Informatics in Education, Vol.7,No.1,2008.
- [8] Dina Abdulaziz Alhammadi," Data Mining in Education-An Experimental Study", International Journal of Computer Applications Volume 62,No.15, January 2013.
- [9] Dr.Mohd Maqsood Ali,"Role of data mining in education sector", International Journal of Computer Science and Mobile Computing Vol.2,Issue.4,April 2013.
- [10] S.Lakshmi Prabha, Dr.A.R Mohamed Shanavas," An Analysis of students performance using classification algorithms", IOSR Journal of Computer Engineering, Volume 16, Issue 1, January 2014.