# REVIEW ON EDUCATIONAL DATA MINING

[1]Mrs. Khushbu Gupta, [2]Esheeta Saha
[1]Assistant Professor, [2]Student
[1,2]Department of Computer Science and Application
St. Aloysius' College (Autonomous), Jabalpur (M.P.), India

***Abstract :***In this recent era, educational institutes and government are concerned to improve the students' performance so as to achieve good academic results. To make this thought work, a lot of manual efforts is being applied from past which were very tedious to tackle. As in the present day, increase in data and difficulty in tackling the data is creating a problem in getting out the useful information. So this problem is overcome with the help of educational data mining (EDM). EDM is playing a vital role to discover the hidden information out of the data which remains untouched. Many researchers and educational scholars have presented different attributes which are affecting the students' performance. This paper reviews different classifiers which help to predict the root cause of affecting students' academic performance. Further some areas where EDM could be applied to help students/instructors/teachers are mentioned here.

***IndexTerms* - DM, EDM, KDD, GPA, CGPA, Weka, RapidMiner, R Tool, Python.**

## I.INTRODUCTION

Data Mining (DM), is a broad area of work in computer science which helps to discover interesting facts and figures that remains previously unknown or untouched many a times but tends to be potentially useful for different purposes. DM is also popularly known as Knowledge Discovery in databases (KDD). DM has come in existence due to massive growth in data, lack of knowledge form the massive data and as a need for analysis.

Educational Data Mining (EDM) is sub-area of work in DM. EDM is a transpiring discipline, concerned with developing methods for exploring unique and increasing large scale data that come from educational sector. Analysis of educational data is not a new process but with the increase in computing power and DM techniques, EDM has come into existence after a series of EDM workshops in many International research conferences held from 2000 to 2007.

In 2008, group of researchers initiated an annual international research conference on EDM firstly held in Montreal, Quebec, Canada. In 2009, EDM researchers initiated Journal of Educational Data Mining (academic journal) to share and widely spreading the research results. In 2011, EDM researchers initiated the International Educational Data Mining Society to associate EDM researchers and expand the field. [10]

EDM has a vivid utilization like analyzing the similar and different categorical students, analyzing the course to be provided, analyzing the implementation of best methods to teach different categorical students, bringing improvement in students' academic performance, bringing improvementin teaching, assisting educators, improving the morals in students and building different models to help predict the nature of learning and students' performance in terms ofterm-end results.

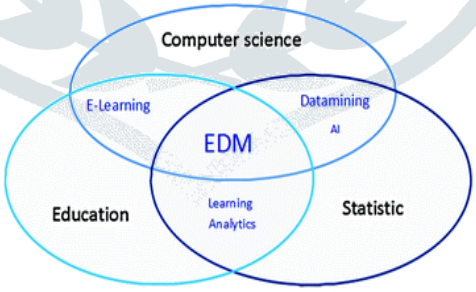


fig. EDM [11]

## II.TOOLS

### 2.1 Weka

Weka (Waikato Environment for Knowledge Analysis) is Java based machine learning software for data mining task. Weka has environment for both pre-written codes as well as for developing new machine learning algorithms. Weka is popularly known free data mining tool with advanced text mining features [12].

### 2.2 RapidMiner

RapidMiner is open-source data science software for data preparation, machine learning, deep learning, data mining, text mining and predictive analytics. It is graphical user interface which makes users easily to work on it.

### 2.3 R Tool

R is both a free software environment and programming language for statistical computing and graphics making it popularly used by researchers and professionals for their work. R is helping data miners, scholars and researchers for data analysis across different sectors and turning out as powerful analytical tool.

### 2.4 Python

Python is general purpose and high level programming language. Features of Python like high readability, platform-independence, interactive program, etc. has made it widely used among scholars and researchers for data mining. Inclusion of packages for scientific use and data analysis made it a powerful tool for data mining.

## III. LITERATURE REVIEW

Identifying, predicting and classifying the students into different categories on the basis of Cumulative Grade Point Average (CGPA) grade which may help the faculty to provide remedial for weak students as well as to upgrade the grades of all students. Dataset of 391 students' academic achievement was collected from Electrical Engineering faculty of a Malaysian public university for the year 2005-2007 and Neuro-Fuzzy classification was applied to it. The main objective of the developed tool is to provide early attention to students predicted as weak category in the form of tutorials and motivation. [1]

Prediction of the factors that is responsible for undergraduate retention in ERAU and similar institutes. Dataset of 972 students enrolled in 2018 at ERAU is collected. On this dataset algorithms like Logistic Regression, Naïve Bayes, KNN, Random Forest, MLP and Decision Tree were applied out of which Logistic Regression gave the best result using Weka. The aim is to provide financial help, motivation and appropriate aide to those students who quit their studies by timely meddling to sustain a defined GPA. [2]

Determining the strategies for recruitment of effective professors and granting policies to motivate existing professors for boosting the standard of research and teaching-learning process. Dataset of 1992 teachers' data is gathered from AICTE mandatory disclosure documents of South Indian engineering colleges. Association rule mining using Apriori algorithm was applied on dataset with the help of R tool. The aim is to build the mind set of students for self-learning and encouraging professors for enhancing their work on research, patents, awards, book publications and R & D grants. [3]

Predicting instructors' performance based on questionnaires evaluation from students opinion about course and instructors' performance. Data is gathered from departments of Marmara University, Istanbul, Turkey. Out of this data, 1995 observations are used for training and 855 observations are used for testing. Decision tree, Support vector machines, Artificial neural networks and discriminant analysis classification techniques were applied for building seven classifier models. C5.0 is the best classifier showing accuracy, precision and specificity. Aim is to make instructors to improve their performance, make evaluation of course effective and expressive, and improvement in analysing measurement criteria. [4]

Main objective is to analyse and improve the performance of first year Computer Application (BCA) studentsinprogramming skills using prediction and classification algorithms. Dataset of 300 records were collected from colleges affiliated to Madras University and were applied on Multilayer Perception, Naïve Bayes, SMO, J48, REPTree classification algorithms using WEKA. The result of experiments based on prediction accuracy is MLP with highest 93% , J48 with 92%, REPTree with 91% , SMO with 90% and Naïve Bayes with lowest of 84%. This work is fruitful for the institution, teachers and students to understand the scope of improvement in the programming performance. [5]

Prediction of factors which determines the student's academic performance is done on dataset from UCI machine learning repository. Naïve Bayes, J48 decision tree and MLP classification algorithms are applied on dataset with the help of WEKA. Classification resulted out to be Naïve Bayes as 68.60% accuracy, J48 as 73.92% accuracy and MLP as 51.13% accuracy. Factors which vitally affect final grades of students are previous grades, consumption of alcohol on workday and weekends, mother's education. The purpose of this classification was to recognize the factors influencing student's academic performance. [6]

Discovering the level of course knowledge from students' performance and accordingly providing apt remedial measures so as to upgrade performance in courses. Dataset is gathered from MS Ramaiah Institute of Technology from the year 2014 to 2016. Students' performance was measured according to bloom taxonomy in continuous internal evaluations. R programming and Python is used as prediction techniques on the collected dataset. Linear regression algorithm and association rule mining is applied to decide the cluster of course knowledge level in which a student fits to. [7]

Student performance depends on many different factors affecting educational process. A student performance prediction model was proposed with the application of KNN and Naïve Bayes classification algorithms on rapid miner IDE. The performance of any algorithm changes upon changing the data set and IDE. Two experiments were performed on data set of secondary schools containing 500 records with 8 attributes (gender, DOB, specialization, city, school name, status, father's job and student status) out of 2000 records gathered from ministry of education in Gaza Strip for 2015 year. As a result Naïve Bayesian showed higher accuracy (93.17%) as compared to KNN. The purpose of this classification is to upgrade the performance of the student by making early predictions in their performance which may encourage the students to perform better and excel in their education process. [8]

Academic year end performance of public school students of Federal District of Brazil is predicted to make the teachers, guidance counsellors for providing the students with appropriate aide to lessen the number of failures. Dataset was collected from database of State Department of Education of the Federal District of Brazil for year 2015 and 2016. Two dataset were used, first

classified students by variables before the school year commenced and second classified students by variables including academic variable after commencing of two months of school. Descriptive statics, CRISP-DM methodology and GBM (Gradient Boosting machine) algorithm were used on the dataset. Classification models were measured using ROC(Receiver Operating Characteristics) curve. Result indicated that the attributes *grades* and *absence* are most important to predict the year end results, still *neighborhood*, *school* and *age* even affects the result of a student. [9]

The above literature review is summarized as:

Table: Summary of review

| S. No. | Year | Title | Dataset | Tool(s) and Technique(s) | Results | Future work |
|---|---|---|---|---|---|---|
| 1 | 2015 | Educational Data Mining for Prediction and Classification of Engineering Students Achievement | Dataset of 391 students' academic achievement is collected from Electrical Engineering faculty of a Malaysian public university for the year 2005-2007 | Neural network, Decision tree, KNN, Bayesian Networks, Neuro-Fuzzy, SVM were studied and compared | Neuro-Fuzzy classification is proposed | Subject grades would be selected using criteria like is it a core subject, subject that have prerequisite, will be analyzed to correlate with final grade and then fed into the classifier |
| 2 | 2016 | Use Educational Data Mining to Predict Undergraduate Retention | Dataset of 972 students enrolled in 2018 at ERAU is collected | Logistic Regression, Naïve Bayse, KNN, Random Forest, MLP and Decision Tree using WEKA | Logistic Regression gave the best result using WEKA | Not specified |
| 3 | 2016 | Teacher Recruitment Data Analytics using Association Rule Mining in Indian Context | Dataset of 1992 teachers' data is gathered from AICTE mandatory disclosure documents of South Indian engineering colleges | Association rule mining using Apriori algorithm with the help of R tool | Administration will be helped to provide better education in educational institutions | Research work could be extended to help management of other colleges, schools and universities |
| 4 | 2016 | Predicting Instructor Performance Using Data Mining Techniques in Higher Education | Dataset is collected from departments of Marmara University, Istanbul, Turkey. 1995 observations are used for training and 855 observations are used for testing | Decision tree, Support vector machines, Artificial neural networks and discriminant analysis classification techniques | C5.0 is the best classifier showing accuracy, precision and specificity | Not specified |
| 5 | 2017 | Classification and Prediction based Data Mining Algorithms to Predict Students' Introductory programming Performance | Dataset of 300 records is collected from colleges affiliated to Madras University | Multilayer Perception, Naïve Bayes, SMO, J48, REPTree classification algorithms using WEKA | Accuracy shown is MLP with highest 93% , J48 with 92%, REPTree with 91% , SMO with 90% and Naïve Bayes with lowest of 84% | Not specified |
| 6 | 2017 | Predicting Academic Performance of Student Using Classification Techniques | Dataset is collected from UCI machine learning repository | Naïve Bayes, J48 decision tree and MLP classification algorithms using WEKA | Accuracy shown is Naïve Bayes as 68.60%, J48 as 73.92% and MLP as 51.13% | Not specified |
| 7 | 2017 | Predicting the Course Knowledge Level of Students using Data Mining Techniques | Dataset of BE and MTech course students from MS Ramaiah Institute of Technology for the year 2014 to 2016 is collected | R programming and Python | Linear regression algorithm and association rule mining helps to decide the cluster of course knowledge level in which a student fits to | Not specified |
| 8 | 2017 | Students | Data set of | Naïve Bayesian | Naïve Bayesian | More classification |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Performance Prediction Using KNN and Naïve Bayesian | secondary schools containing 500 records with 8 attributes out of 2000 records is gathered from ministry of education in Gaza Strip for the year 2015 | and KNN using Rapid miner IDE | showed higher accuracy (93.17%) as compared to KNN | algorithms can be applied on different educational datasets |
| 9 | 2018 | Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil | Dataset is collected from State Department of Education of the Federal District of Brazil with 238575 records of 2015 and 247297 records of 2016 | Descriptive statics, CRISP-DM, GBM and ROC | *grade* and *absence* are most important to predict the year end results, still *neighborhood*, *school* and *age* even affects the result of a student | Knowledge of research can help education specialist to take apt decisions for providing aide to the students and can be extended for UG students. Further, more attributes can be searched which affect students performance as well as other well-known data mining techniques |

## IV. CONCLUSION AND FUTURE SCOPE

In this paper, introduction of educational data mining and different techniques of DM used in educational data are briefly described. EDM is flourishing a lot in order to help the educational sector to work for the at-risk students anddifferent unknown attributes affecting students' performance are coming forth which are making EDM researchers to work vigorously. In the future EDM can be applying in the following areas:

1. Prediction of stream/branch/course to be taken by the matriculation students for their further studies.
2. Prediction of stream to be pursued by the Undergraduate student for their Post Graduation.
3. Comparison in the pattern and tricks of managing their studies by the hosteller and localite students.
4. Predicting the stability of an instructor of any school/college/institute/university from their educational profile.
5. Predicting the profession to be opted by students by collecting data related to their academic, personal, family background and demographic.

## V. REFERENCE


**[1]** Buniyamin, N., Mat U. b. and Arshad, P. Md. 2015. Educational Data Mining for Prediction and Classification of Engineering Students Achievement. 2015 IEEE 7th International Conference on Engineering Education (ICEED '15), 49-53.

**[2]** Lehr, S., Liu, H., Klinglesmith, S., Konyha, A., Robaszewska, N. and Medinilla, N. 2016. Use Educational Data Mining to Predict Undergraduate Retention. 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT '16), 428-430.

**[3]** Kadappa, V., Guggari, S. and Negi, A. 2016. Teacher Recruitment Data Analytics using Association Rule Mining in Indian Context. 2016 IEEE International Conference on Data Science and Engineering (ICDSE '16).

**[4]** Agaoglu, M. 2016. Predicting Instructor Performance Using Data Mining Techniques in Higher Education., Manuscript ID Access-2016-00572, submitted for publication to IEEE.

**[5]** Sivasakthi, M. 2017. Classification and Prediction based Data Mining Algorithms to Predict Students' Introductory programming Performance. Proc. International Conference on Inventive Computing and Informatics (ICICI '17), 346-350.

**[6]** Roy, S. and Garg, A. 2017. Predicting Academic Performance of Student Using Classification Techniques. 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON '17) GLA University, Mathura.

**[7]** Parkavi, A. and Lakshmi, K. 2017. Predicting the Course Knowledge Level of Students using Data Mining Techniques. 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM '17), 128-133.

**[8]** Amra, I. A. A. and Maghari, A. Y. A. 2017. Students Performance Prediction Using KNN and Naïve Bayesian. 2017 8th International Conference on Information Technology (ICIT '17).

**[9]** Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R. and Erven, G.V. 2018. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research (2018) available at https://doi.org/10.1016/j.jbusres.2018.02.012.

**[10]** https://en.wikipedia.org/wiki/Educational_data_mining

**[11]** https://media.springernature.com/original/springer-static/image/chp%3A10.1007%2F978-3-319-02738-8_1/MediaObjects/314483_1_En_1_Fig1_HTML.gif

**[12]** https://en.wikipedia.org/wiki/Weka_(machine_learning)