IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11580

# A Systematic Analysis on Role of Data mining algorithms in the field of Educational Data mining

**Karthick S [1], Kanimozhi V A[2], Malathi V A[3] Vibinchandar S[4]**

Assistant Professor, Department of English, Sri Krishna Adithya College of Arts and Science, Coimbatore[1]

Assistant Professor, PG & Research Department of Mathematics, Sri Ramakrishna College of Arts and Science, Coimbatore[2]

Assistant Professor, Department of Computer Science Education, RVS College of Education, Coimbatore[3]

Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science, Coimbatore[4]

**Abstract:** Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. EDM refers to the techniques, tools, and research designs utilized to obtain information from educational records, typically online logs, and examination results, and then analyses this information to formulate conclusions. The problem with educational data is it is a data rich and information poor collection. Data Mining is a process of finding potentially useful patterns from huge data sets. It is a multi-disciplinary skill that uses machine learning, statistics, and AI. Data mining process have the ability to discover the hidden knowledge present within the collection of educational data and then identify students' performance with great accuracy. Also, it is playing a vital role in the process of diagnosis and prediction of problems in students' education. This proposed paper presented a detailed systematic study of the role of various data mining techniques and algorithms in the field of educational data mining.

**Keywords:** Educational Data Mining, Data Mining algorithms, Student Performance, Naïve Bayes, Neural Network

## I. INTRODUCTION

A. What is Educational Data mining?

EDM stands for Educational Data Mining. It can be defined as the technique for finding the specific types of data that come from the education system and implementing those techniques to understand students and the system better.

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in. [2]

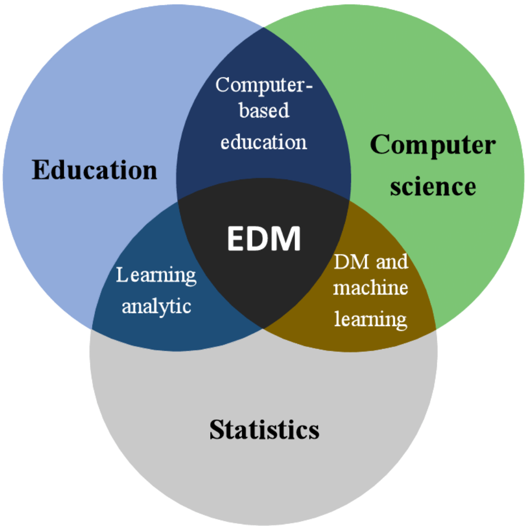

Fig.1 Educational Data Mining

DOI: 10.17148/IJARCCE.2022.11580

## II. ROLE OF DATA MINING ALGORITHMS IN DATA MINING

A. What is Data mining?

Data Mining is a non-trivial methodology for locating valid, novel, possible, helpful and ultimately comprehendible patterns in knowledge. It puts along a spread of tools and techniques which will be applied to the processed data so as to find hidden patterns.

B. What is Machine Learning?

This is the algorithm part of the data mining process. It provides computers with the ability to learn without being explicitly programmed. This taxonomy or way of organizing machine learning algorithms is useful because it forces us to think about the the roles of the input data and the model preparation process and select one that is the most appropriate for our problem in order to get the best result. [4]

C. Supervised Learning:

Input data is called training data and has a known label or result. A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. [4]

D. Unsupervised Learning:

Input data is not labelled and does not have a known result. A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity. [4]

E. Semi-Supervised Learning:

Input data is a mixture of labelled and unlabelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions. [4]

## III. TYPES OF ALGORITHMS

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases. [2]

1. Classification**:** Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large.

2. Clustering: Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes.

3. Prediction: Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables.

4. Association rule: Association and correlation is usually to find frequent item set findings among large.

5. Neural networks: Neural network is a set of connected input/output units and each connection has a weight present with it. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. [4]

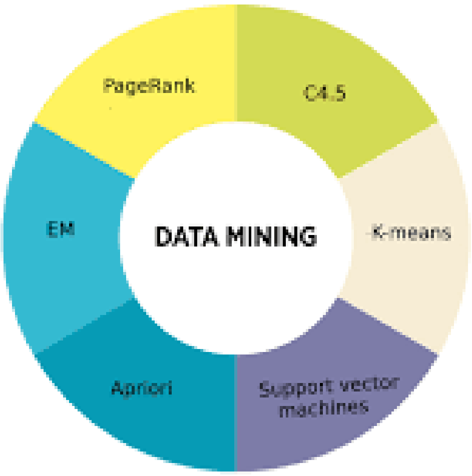

Fig 2. Types of Data Mining algorithms

DOI: 10.17148/IJARCCE.2022.11580

## IV. LITERATURE REVIEW

Table 1. Shows various Data Mining algorithms & techniques used in Educational Data mining:

| S.No | Author | Techniques Used | Accuracy | Year |
|---|---|---|---|---|
| 1 | Raheela Asif, Agathe Merceron, Syed Abbas Ali, Najmi Ghani Haider [5] | Naive Bayes | 83.65% | 2017 |
| | | Random Forest Trees with Gini Index | 71.15% | |
| | | Neural Networks | 62.50% | |
| 2 | Amjad Abu Saa [6] | Classification and Regression Tree (CART) | 40% | 2016 |
| 3 | Eduardo Fernandes [7] | Classification model – Gradient Boosting Machine | Gradient Boosting technique outperforms | 2018 |
| 4 | Bo Guo [8] | Deep Learning | 78.4 | 2015 |
| 5 | Ms.Tismy Devasia, Ms.Vinushree T P, Mr.Vinayak Hegde [9] | Naïve Bayes | Naïve Bayes outperforms | 2016 |
| 6 | Prof. Priya Chandran Ms. Sanakhatun Shakirali Shaikh [10] | K-Means | K-Means outperforms | 2018 |
| 6 | Sohan Kamble [11] | Naïve Bayes | Naïve Bayes outperforms | 2019 |
| 7 | Ram Dayal Tanwar | K-Means | K-Means outperforms | 2019 |
| 8 | Mudasir Ashrafa, | Naïve Bayes | 97.5% | 2020 |
| 9 | Annisa Uswatun Khasanah | Bayesian Network | 98.08% | 2017 |

## V. DETAILED EXPLANATION OF DATA MINING METHODOLOGY IN EDUCATIONAL DATA MINING FIELD

A. Raheela Asif et al. [5]

In this proposed work, author used data mining methods to study the performance of undergraduate students. Two aspects of students' performance have been focused upon. First, predicting students' academic achievement at the end of a four year study programme. Second, studying typical progressions and combining them with prediction results. Two important groups of students have been identified: the low and high achieving students. The results indicated that by focusing on a small number of courses that are indicators of particularly good or poor performance, it is possible to provide timely warning and support to low achieving students, and advice and opportunities to high performing students. [5]

B. Amjad Abu Saa et al. [6]

This proposed work is equally concerned with this subject, specifically, the students' performance. This study explored multiple factors theoretically assumed to affect students' performance in higher education, and finds a qualitative model which best classifies and predicts the students' performance based on related personal and social factors. [6]

C. Eduardo Fernandes et al. [7]

Author presented a predictive analysis of the academic performance of students in public schools of the Federal District of Brazil during the school terms of 2015 and 2016. Initially, proposed work concentrated on a descriptive statistical analysis to gain insight from data. Subsequently, two datasets were obtained. The first dataset contains variables obtained prior to the start of the school year, and the second included academic variables collected two months after the semester began. Classification models based on the Gradient Boosting Machine (GBM) were created to predict academic outcomes of student performance at the end of the school year for each dataset. Results showed that, though the attributes 'grades' and 'absences' were the most relevant for predicting the end of the year academic outcomes of student performance.[7]

DOI: 10.17148/IJARCCE.2022.11580

D. Bo Guo et al. [8]

In this proposed work author developed a classification model to predict student performance using Deep Learning which automatically learns multiple levels of representation. It pre-train hidden layers of features layer wisely using an unsupervised learning algorithm sparse auto-encoder from unlabelled data, and then it utilized supervised training for fine-tuning the parameters. Author trained the model on a relatively large real world students dataset, and the experimental results showed the effectiveness of the proposed method which can be applied into academic pre-warning mechanism. [8]

E. Tismy Devasia et al. [9]

This proposed system was a web based application which makes use of the Naive Bayesian mining technique for the extraction of useful information. The experiment was conducted on 700 students' with 19 attributes in Amrita Vishwa Vidyapeetham, Mysuru. Result proved that Naïve Bayesian algorithm provides more accuracy over other methods like Regression, Decision Tree, Neural networks etc., for comparison and prediction. This system aimed at increasing the success graph of students using Naïve Bayesian and the system which maintains all student admission details, course details, subject details, student marks details, attendance details, etc. It took student's academic history as input and gives students' upcoming performances on the basis of semester. [9]

F. Priya Chandran et al. [10]

In this research paper, author developed a model to predict the student's performance by using clustering model with the help of k-means algorithm. Author focused on social media perspective data, which is strongest source of EDM and nowadays, it is more widely used to be as an important factor in the field of EDM. It is observed that, the model built using EDM techniques incorporate students overall behavior and pedagogy to analyze the knowledge and teaching and learning outcomes. [10]

G. Sohan Kamble et al. [11]

The proposed work described the overall performance of the student ,in academic, sports ,practicals, culture, social as well as extra curricular activities. This work helped to select the field in future as well as the student weak in which area how to improve his knowledge in that area. Prediction of student behavior is made by the classifier. Here Naïve Bayes algorithm was utilized to predict the student behaviour. Naïve Bayes classified outperformed compared to other classifiers. [11]

H. D Ram Dayal Tanwar et al [12]

Author examined many papers for prediction of students' performance .Here on comparing decision tree and k means it is seen that k means is more efficient as compare to decision tree. Students performance was so important for their future it not only help student but also help teachers institute parents. Elbow method utilized in the optimal solution. Many big institutes used the concept of AI for prediction. With the help of machine learning concept, it is easy to improve the result and future of students. It is not only useful for students but also for teacher and institute to improve their result. [12]

I. Mudasir Ashrafa et al. [13]

In this proposed work several ensemble techniques have been discussed to get a comprehensive knowledge of key methods. Among various ensemble approaches, researchers have practiced boosting mechanism to predict the performance of students. As application of ensemble methods is contemplated to be significant phenomenon in classification and prediction procedures, therefore the researchers exploited boosting technique to develop an accurate prediction pedagogical model, in view of the pronounced nature and novelty of the proposed method in educational data mining.

The base classifiers including random tree, j48, knn and naïve bayes have been evaluated on 10- fold cross validation system. Moreover, filtering procedures such as oversampling (SMOTE) and under-sampling (Spread subsampling), have been exploited to further inspect any significant change in results among meta and base classifiers. Both ensemble and filtering approaches have demonstrated substantial improvement in predicting the performance of students than the application of conventional classifiers. Furthermore, based on the improvement in results two novel prediction models have been propounded after conducting performance analysis on each approach [13]

J. Annisa Uswatun Khasanah et al. [14]

This proposed work used Feature Selection to select high influence attributes with student performance in Department of Industrial Engineering Universitas Islam Indonesia. Then, two popular classification algorithm, Bayesian Network and Decision Tree, were implemented and compared to know the best prediction result. The outcome showed

DOI: 10.17148/IJARCCE.2022.11580

that student’s attendance and GPA in the first semester were in the top rank from all Feature Selection methods, and Bayesian Network is outperforming Decision Tree since it has higher accuracy rate. [14]

## VI. Conclusion

Educational data mining refers to techniques and tools designed for automatically extracting meaning from large repositories of data generated by peoples learning activities in educational settings. It is the application of data mining techniques and methods to educational data for management and research purposes. Aims at understanding the learners’ learning behavior better to set and customize trainings. It requests algorithms to aggregate data collected from online learning systems to bring out patterns. The topic covers prediction, usage and results for pedagogical process improvement. Data mining algorithms like CART, Random Forest-ID3, Decision Tree, Support Vector Machine, and Naive Bayes are widely used for the prediction of students’ performance. This paper has presented a detailed systematic review of various data mining techniques and algorithms which are utilized in the process of diagnosis and prediction of students’ behaviour and performance.

## VII. References

[1] Kanimozhi .V .A, Vibinchandar .S, “Impact of Artificial Intelligence in Feature Selection Methods of Naïve Bayes Prediction Model”, Solid State Technology Volume: 63 Issue: 5, 2020.

[2] https://www.javatpoint.com/educational-data-mining

[3] Vibinchandar .S, Dr. Krishnapriya .V, “Survey On Various Prediction Models For Survival Of Breast Cancer Patients Using Warm Boot Random Forest Classifier”, European Journal of Molecular & Clinical Medicine, ISSN 2515-8260, Volume 07, Issue 09, 2020.

[4] V.A. Kanimozhi, Dr. T. Karthikeyan , “A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease”, International Journal of Advanced Research in Computer and Communication Engineering, ISSN (Online) 2278-1021, Vol. 5, Issue 4, April 2016.

[5] Raheela Asif a, * , Agathe Merceron b , Syed Abbas Ali c , Najmi Ghani Haider, “Analyzing undergraduate students' performance using educational data mining”, Computers & Education, 0360-1315, 2017 Elsevier Ltd.

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

[7] Eduardo Fernandes Maristela Holanda Marcio Victorino Vinicius Borges Rommel Carvalho Gustavo VanErven , “Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil”, Journal of Business Research, 2018

[8] Bo Guo, Rui Zhang, Guang Xu, Chuangming Shi and Li Yang, “Predicting Students Performance in Educational Data Mining”, International Symposium on Educational Technology, 2015

[9] Ms.Tismy Devasia, Ms.Vinushree T P, Mr.Vinayak Hegde, “Prediction of Students Performance using Educational Data Mining, International Conference on Data Mining and Advanced Computing, 2016

[10] Prof. Priya Chandran Ms. Sanakhatun Shakirali Shaikh, “Educational Data Mining: Predicting student's performance using clustering, International Journal of Management, IT & Engineering , Vol. 8 Issue 6, June 2018.

[11] Sohan Kamble , Kodam Saurabh, Maske Adesh, “Student Performance Prediction System”, Mukt Shabd Journal Issn No : 2347-3150, Volume VIII, Issue VIII, AUGUST/2019

[12] Ram Dayal Tanwar, Dr. Rajeev Kumar Gupta, Predicting Students Performance and Review on EDM: Machine Learning Theory, International Journal for Scientific Research & Development| Vol. 7, Issue 05, 2019 | ISSN (online): 2321-0613

[13] Mudasir Ashrafa, Majid Zamanb , Muheet Ahmed, “An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches”, Procedia Computer Science, 2020

[14] Annisa Uswatun Khasanah, Harwati, “A Comparative Study to Predict Student’s Performance Using Educational Data Mining Techniques”, Materials Science and Engineering, 2017

## VIII. Biography

**Mr. Karthick S** is a faculty in English. He has qualified in SLET and NET examinations. His areas of specialization are Fiction and Comparative Literature.

DOI: 10.17148/IJARCCE.2022.11580

**Ms. Kanimozhi.V.A** is a faculty in Computer Science. She has qualified in SLET and NET examinations. Currently she is pursuing Ph.D in computer science. Her areas of specialization are Data mining and Web Development.

**Mrs. Malathi V.A** is a faculty in Computer Science Education. She has qualified in SLET and NET examinations. Her areas of specialization are Data mining and Web Development.

**Mr. Vibinchandar S** is pursuing Ph.D in computer science. He is an International Certified Software Tester by ISTQB. Also he is a corporate trainer and he trained more than 8000 students in Technical & Soft skills. His areas of specialization are Data mining, Software Engineering and Web Development.