Analysis of the Performance of the Main Algorithms for Educational Data Mining: A Review

Mustafa Abdalrassual jassim
University of Monastir, Tunis, Al Muthanna University, Muthanna, Iraq
mustafa944@mu.edu.iq

Abstract
Data mining is the process of extracting useful and valuable information from vast amounts of data. Algorithms and various tools to use are some of the most popular data mining applications to estimate future events based on past experiences. Many researchers use techniques and tools to extract useful data to address and solve higher education problems in this context. EDM analyzes educational data using methods and algorithms to develop and apply DM data extraction to the information stored in academic data repositories. As a result, it provides essential knowledge of the teaching and learning process for successful and effective educational planning. This paper focuses on comparing the algorithms' performance and applied to the same educational data set. Educational Data Mining (EDM) uses these algorithms to explore patterns, educational statistics, and predictions in the data. Thus, statistics are generated based on all classification algorithms. A comparison of all eight classifiers was made to predict results and find the best performance classification algorithm accurately. This paper aims to use a literary survey to determine the most appropriate algorithm according to EDM's needs.

Keywords: Educational Data Mining (EDM), Classification, Prediction, Academic performance.

1. Introduction
The primary goal of any educational institution is a student's academic success. EDM is a new area of research that helps education institutions enhance their students' performance [1]. EDM analyzes educational data using methods and algorithms to develop and apply data mining (DM) to the information stored in academic data repositories [2]. EDM uses these algorithms to explore patterns, predictions, and educational statistics in data that illustrate learner performance [3]. And this data must always be processed and analyzed accordingly to produce information and knowledge. Organizational information creation and distribution are a strategic priority supporting higher education in the roadmap to prepare, update, and develop academic and study indicators. For a fair data analysis process, the creation of technical infrastructure is important. One of the necessary technologies for this infrastructure is a data warehouse. A data warehouse is a multi-dimensional data warehouse which is primarily used for analysis [4]. Data mining analysis covers, among other things, marketing, health, finance, and insurance. EDM is known as the DM app in educational contexts. EDM is a significant advancement that focuses on modelling to enhance learning experiences and organizational skills. To improve HEI decision-making, EDM uses software tools to identify trends and patterns for education data [5]. Researchers have done a lot of research to extract 'learning outcomes' from students and factors which affect academic performance [6]. EDM manages a growing new strategy for investigating instruction and information and using data mining techniques on raw data emerging from prescriptive frameworks to useful information that may significantly influence research and prescriptive practices. Data mining uses a great role in enhancing the educational field's various aspects, from anticipating student performance and teacher excellence or even administrative improvements and appropriate resources [7]. There are many educational fields that data mining is used as analysis and perception, providing criticism to support teachers, recommendations to students. Predicting substitute implementation, student presentation, detection of alternative practices, student assembly, failure,
development of idea maps, curriculum building, planning, and reservation [8]. The current study also compares these algorithms and the methods used and suggests the best accuracy and efficiency in some scenarios. This paper's findings will help data scientists in further studies to choose the correct techniques and algorithms. The research question is: What are the best techniques and algorithms for EDM for student analysis of dropout, degree, failure rate, student progress, proper decision-making, etc.? This question's answer can be essential for education researchers to limit experiment time and analyze student dropout and graduation rates. We hope you use this paper to gain a clear insight into which methods and algorithms are most adequate for studying these academic indicators.

2. **Background**

This work's research objective is to study current methods and algorithms for EDM; the most appropriate educational institutions' data analysis experience. In this, the latest developments in the EDM subject and current methods and algorithms of the EDM needed to design the experiments are described.

![Diagram](image)

**Figure 1.** Overview of how EDM methods are applied

There are increasing numbers of EDM applications. they can be grouped into the following four categories:

1- Student Modeling: Student data (including knowledge, grades, etc.) and EDM techniques can be used to design a customized learning process by modelling differences between students.
2- Modelling of the domain's knowledge structure: Methods are created that combine modelling frameworks and search algorithms to discover data-based domain models.
3- Educational support: Effective educational support can be identified.
4- Scientific Research: Applications may help develop and test educational, scientific theories and the formulation of new hypotheses.

3. **Classification Algorithms in Educational Data Mining**

Classification is one of the most commonly used methods for classifying large data sets. This approach involves algorithms for supervised learning and data adaptation. The aim is to establish the relationship between the variable of interest of the qualitative form and other variables observed [9]. Classification analysis can be used to query, determine, or forecast conduct through an algorithm. It works by generating a collection of training data containing a variety of essential characteristics and potential outcomes. The classification algorithm's task is to figure out how the collection of traits reaches its end, which determines and assigns categories to a group of data to be evaluated more precisely. This approach uses mathematical
methods such as decision trees, J48, Random Forest, SVM, linear programming, neural grids, and statistics. Figure (1) demonstrates EDM grading algorithms.

3.1. Classification
Classification is a data mining technology that assigns classes to different data to support predictions and analyses. Classification is one of the ways of efficiently processing large data sets. The classification objective is to correctly predict the target population in the data for each event. The classification technique involves learning and classification from the data. It is the most widely used DM method, and it is used for...
developing classes and assigning datasets to each. In short, if the target variable is discrete, then this is a classification problem, and if the target variable is continuous, it is a regression task.

3.2. Clustering
It is a specific group of organisms based on their properties and grouped according to their similarities. Concerning data extraction, this methodology divides the data by implementing a particular linking algorithm, which is most suitable for analyzing the required information. Clustering helps to split the data into several subsets. Each of these subsets contains similar data to each other, and these subsets are called groups. Now that our customer base’s data is broken down into groups, we can make an informed decision about who we think is best suited for the product.

3.3. Association Rule Mining
It is a method that seeks to recognize recurring patterns, correlations, correlations, or causal structures of data sets found in various databases, such as link databases, transactional databases, and other types of the data warehouse. Looking at a group of transactions, digging into the rules of correlation aims to find the rules that enable us to predict a particular item's occurrence based on other elements' events in the transaction. Correlation is a data mining function that detects the likelihood of co-occurrence of things in the set. Relationships between simultaneous elements are expressed as association rules.

3.1.1. J48 Decision Tree
J48 algorithm is used to classify various applications and to achieve correct classification results. J48 is an ID3 extension [10]. Additional features of J48 are its missing traits and choice of tree pruning; it is an algorithm for creating a decision tree generated by C4.5 (an extension of ID3) [11]. It is also known as a statistical classifier. To classify the decision tree, we need a database.

3.1.2. Neural Network
A neural network is a collection of algorithms that aim to identify fundamental relationships in several data through a framework that imitates the human brain's workings [12]. In this context, neural networks refer to neuronal systems organically or artificially. Neural networks can be adapted to evolving inputs. The structure thus achieves the best outcome without re-designing output parameters [13]. In developing trading systems, the idea of neural networks, which has its origins in artificial intelligence, has become very common.

3.1.3. Support Vector Machine
Support vector machine (SVM) is a supervised machine learning algorithm for classification, regression and variation detection. The SVM algorithm aims to find a break between two groups of objects, assuming that the greater the breakup, the robust the classification [14]. The Support Vector Machine (SVM) operates similarly to the C4.5 algorithm, except that no decision trees are used by the SVM at all. To classify knowledge into two groups, SVM learns data sets and determines the hyperlink.

3.1.4. Random Forest
It is one of the theories of statistical learning, and the method is used with multiple decision trees to make predictions and use voting to obtain the final results of the prediction [15]. The size of the sub-sample is always the same as the sample of the original input but replacement is used to draw examples.

3.1.5. C4.5 Algorithm
The C4.5 algorithm is used for data mining as a decision tree classification to make a decision based on a given data set (univariate or multivariate predictions) [16]. Classifier here refers to a data mining method that takes the data we need to identify the new data category and predict it.

3.1.6. K-Nearest Neighbors Algorithm
K-Nearest Neighbors (KNN) is a simple algorithm that stores all existing states and classifies new cases based on a similar scale. In the early 1970s, KNN had already been used as a non-parametric method in statistical estimation and pattern recognition [17]. KNN, also known as K-Nearest Neighbor, is a moderated learning and pattern classification algorithm that helps us find which class the new entry belongs to (test value) when selecting k-Nearest neighbors and calculating the distance between them.

3.1.7. Naive Bayes Algorithm
A group of classification algorithms based on Bayes' theorem are Naive Bayes classifiers. It's not a single algorithm; it's an algorithm group. They all share a common principle: each pair of features is labeled as an independent. Naive Bayes is provided with a standardized table training dataset [18]. So, it is treated as a moderated learning algorithm. It helps us predict the probability of an occurrence based on the circumstances we are familiar with for these cases.

3.1.8. ID3 Algorithm
ID3 is an algorithm developed by Ross Quinlan [19] and used to generate a decision tree from a data set while learning the decision tree. ID3 is an introduction to C4.5 and is commonly used to process machine learning and natural languages. The resulting tree is used to identify possible samples.

4. Literature Search Procedure and Criteria
The main goal of all educational institutions is to improve the standard of education and improve academic performance. Educational data mining is a rising area of research in which universities improve their students' understanding [20]. There are many advantages of data mining technology in the education sector. Leveraging DM technologies in the education sector is a growing and new area of research. Educational data exploration. Participates in developing methods that aid in the specific search of data sets from the educational setting [21]. (EDM) is a sophisticated system that allows exploration of knowledge from academic settings by developing and applying data extraction methods and algorithms to the information stored in data warehouses in higher education institutions [22]. One of the main applications of higher education in predicting student performance. This will help the stakeholders to understand the impact of various factors on academic performance, thus enabling them to take prompt and adequate remedial action [23]. Many techniques and algorithms have been applied to higher education data, including student performance, marital status, failure, and education to make correct decisions and develop academic performance to obtain accurate results. Among these algorithms are classifiers and the J48 Decision Tree, Random Forest Algorithm, SVM, C4.5 Algorithm, ID3 Algorithm, Naïve Bayes Algorithm. Accurate results and various statistics have been shown using techniques and algorithms to develop academic data mining, as shown in Table 1, and used by many researchers.

| Table 1. Comparison of EDM classification algorithms |  |  |
| Authors                  | Problem / Objective                                                                 | Algorithm / Methods                                                                 | Best Algorithm   |
|-------------------------|---------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|------------------|
| Hussain, et al [9]      | Techniques of data mining for improving student academic performance and preventing   | J48, PART, Random Forest and Bayes Network.                                          | Random Forest    |
|                         | dropouts.                                                                             |                                                                                      |                  |
| Algur, et al [25]       | To extract student performance data and data.                                         | Random Forest and J48                                                               | Random Forest    |
| Anoopkumar, et al [26]  | Enhance students 'classification and prediction accuracy.                              | Naïve Bayes, J48, Bayes Net, SVM, REP Tree and Random Forest                        | J48              |
| Salal, et al [27]       | Student Performance Dataset (to predict student results)                               | Naïve Bayes, J48, Random Forest, REP Tree, JRip, OneR, SimpleLogistic and ZeroR.    | J48              |
| GUNASUNDARI, et al [28] | Predicting student results, identification of inappropriate student behaviors, group  | Naïve Bayes KNN, Random Forest, Neural Network                                       | Random Forest    |
|                         | ing of students and modeling of students.                                              |                                                                                      |                  |
| Kapur, et al [29]       | Measuring student potential and how to improve it                                     | J48, Random Forest, Naïve Bayes, Naïve Bayes and K-star.                            | Random Forest    |
| Rahayu, et al [30]      | The objective of the analysis is to forecast student success using rating algorithms  | J48, KNN, Naïve Bayes and KNN.                                                      | Naïve Bayes      |
|                         | on the basis of course assessments.                                                   |                                                                                      |                  |
| Saa, et al [2]          | Prediction of academic performance and an early warning mechanism to forecast         | Naïve Bayes, KNN and Random Forest                                                  | Random Forest    |
|                         | failure of students and lower academic performance.                                    |                                                                                      |                  |
| Al-Shehri, et al [32]   | Predicting student performance                                                        | SVM algorithm and KNN algorithm                                                       | SVM              |
| Mueen, et al [33]       | Teachers' Assistant to boost the academic performance of students.                    | Naïve Bayes, Decision Tree, and Neural Network                                       | Naïve Bayes      |
| Asanbe, et al [34]      | Predicting teacher performance                                                         | ID3, MLP, C4.5 decision tree                                                         | C4.5             |
| Mishra, et al [22]      | apply different classifiers to find students' employability and to improve employability | C4.5, J48, Random Forest                                                             | J48              |
| Kumar, et al [20]       | Student academic performance                                                           | Naïve Bayes, Decision Tree, PART and Random Forest                                  | Random Forest    |
| Authors            | Description                                                                 | Algorithms                               | Algorithm |
|--------------------|------------------------------------------------------------------------------|------------------------------------------|-----------|
| Sunday, et al [4]. | The analysis of student performance in educational programming with classification techniques | J48, ID3 algorithm.                      | J48       |
| Ahmed, et al [33]. | Computation of student academic performance using a predictive data mining approach | KNN, J48, Bagging, Random forest, and Naïve Bayes | KNN       |
| Lee, et al [10].   | Predict if there is a real possibility of academic leakage                   | J48, KNN, and SVM                        | J48       |
| Delima, et al [8]. | Accuracy of students' responses in the actual assessment made using the LMS | KNN, C4.5 and Naïve Bayes algorithms      | Naïve Bayes |
| Jawthari, et al [7]. | Use student data to improve student performance, as well as teacher performance. | SVM, Random Forest                       | Random Forest |

There are other EDM methodologies, which have not been used widely in education, so I have not covered them.

5. Results
EDM in classification uses algorithms to predict future results in student performance and improve the institution's quality. Several researchers have conducted research to predict student performance. In this research, different classification algorithms were analyzed based on results, accuracy, and performance. Different classifier algorithms used in EDM by various authors are KNN, Naïve Bayes, SVM, Random Forest, Neural Networks, C4.5, J48, ID3. An assessment of EDM methods and algorithms has been conducted. This analysis shows that the Random Forest and J48 algorithms have better accuracy in grading students in analyzing student dropout and graduation rates, improving teaching and learning, and improving academic indicators, which leads to making the right decisions to develop academic indicators. Table (1) mentioned the algorithms most often used by researchers to explore educational data. Figure 1 shows the algorithms mostly used in EDM.
Figure 3. Classification Algorithms of EDM

6. Conclusion and Future Work
EDM is a relatively new area with the scope of many potentials to help the community if used in the right way. In this paper, we compared eight algorithms. Its primary focus is on exploring and creating valuable knowledge from educational information systems from schools to colleges and universities. To improve decision-making and enhance student performance, predicting student performance is one of EDM's top research areas. The importance of this review paper is to provide consistent evaluation criteria for comparing different technologies for each application. Forecast accuracy is used as a performance forecast file for the techniques considered. This paper demonstrates that a practical approach does not inherently imply that it is useful in another application. In future work, further surveys should be done for each EDM application to identify the most effective technology or explore new technologies more precisely. Moreover, the Extended Evaluation should use benchmarks in the evaluation.

References
[1] Maryam Zaffar, Manzoor Ahmed Hashmani, and KS Savita. Performance analysis of feature selection algorithm for educational data mining. In 2017 IEEE Conference on Big Data and Analytics (ICBDA), pages 7–12. IEEE, 2017.
[2] Oswaldo Moscoso-Zea, Pablo Saa, and Sergio Lujan-Mora. Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. Australasian Journal of Engineering Education, 24(1):4–13, 2019.
[3] Ravinder Ahuja, Animesh Jha, Rahul Maurya, and Rishabh Srivastava. Analysis of educational data mining. In Harmony Search and Nature Inspired Optimization Algorithms, pages 897–907. Springer, 2019.
[4] Kissinger Sunday, Patrick Ocheja, Sadiq Hussain, Solomon Sunday Oyelere, Oluwafemi Samson Balogun, and Friday Joseph Agbo. Analyzing student performance in programming education using classification techniques. 2020.
[5] Dongjo Shin and Jaekwoun Shim. A systematic review on data mining for mathematics and science education. INTERNATIONAL JOURNAL OF SCIENCE AND MATHEMATICS EDUCATION, 2020. (ITHET), pages 1–6. IEEE, 2016.

[6] Zengzhen Shao, Hongxu Sun, Xiao Wang, and Zhongzhi Sun. An optimized mining algorithm for analyzing students’ learning degree based on dynamic data. IEEE Access, 8:113543–113556, 2020.

[7] Moohanad Jawthari and Veronika Stoffova. Effect of encoding categorical data on student’s academic performance using data mining methods. In The International Scientific Conference eLearning and Software for Education, volume 1, pages 521–526. " Carol I" National Defence University, 2020.

[8] Allemar Jhone P Delima, Jan Carlo T Arroyo, Michelle C Elape, and Markdy Y Orong. The use of schoology as learning management system in the college of computing education: A response assessment using data mining techniques. International Journal, 9(3), 2020.

[9] Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwib, and Najoua Ribata. Educational data mining and analysis of students’ academic performance using weka. Indonesian Journal of Electrical Engineering and Computer Science, 9(2):447–459, 2018.

[10] Luis Earving Lee, Salvador Ibarra Martínez, Jos’e Antonio Cast’an Rocha, Jes’us David Ter’an Villanueva, Julio Laria Menchaca, Mayra Guadalupe Trevi’no Berrones, Emilio Cast’an Rocha, et al. Evaluation of prediction algorithms in the student dropout problem. Journal of Computer and Communications, 8(03):20, 2020.

[11] Amritpreet Kaur, E Roop, and L Sharma. Classification and prediction based enhanced j48 and reetree algorithms to predict corona virus pandemic. no. July, 2020.

[12] Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1150–1160, 2020.

[13] Qian Hu and Huzefa Rangwala. Towards fair educational data mining: A case study on detecting at-risk students. In Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020), pages 431–437, 2020.

[14] Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Salem Alelyani, and Raheel Nawaz. Predicting academic performance of students from vle big data using deep learning models. Computers in Human Behavior, 104:106189, 2020.

[15] Hui-Chun Hung, I-Fan Liu, Che-Tien Liang, and Yu-Sheng Su. Applying educational data mining to explore students’ learning patterns in the flipped learning approach for coding education. Symmetry, 12(2):213, 2020.

[16] Ginika Mahajan and Bhavna Saini. Educational data mining: A state-of-the-art survey on tools and techniques used in edm. International Journal of Computer Applications & Information Technology, 12(1):310–316, 2020.

[17] Lili Dwi Yulianto, Agung Triayudi, and Ira Diana Sholihati. Implementation educational data mining for analysis of student performance prediction with comparison of k-nearest neighbor data mining method and decision tree c4. 5. Jurnal Mantik, 4(1, May):441–451, 2020.

[18] Ekansh Maheshwari, Chandrima Roy, Manjusha Pandey, and Siddharth Swarp Rautray. Prediction of factors associated with the dropout rates of primary to high school students in india using data mining tools. In Frontiers in Intelligent Computing: Theory and Applications, pages 242– 251. Springer, 2020.

[19] Saja Taha Ahmed, Rafah Al-Hamdani, and Muayad Sadik Crock. Developed third iterative dichotomizer based on feature decisive values for educational data mining. Indonesian Journal of Electrical Engineering and Computer Science, 18(1):209–217, 2020.

[20] Mukesh Kumar and AJ Singh. Evaluation of data mining techniques for predicting student’s performance. International Journal of Modern Education & Computer Science, 9(8), 2017.
[21] Md Rakib Ahmed, Sheikh Tariqul Islam Tahid, Nazneen Akter Mitu, Pipasha Kundu, and Soniya Yeasmin. A comprehensive analysis on undergraduate student academic performance using feature selection techniques on classification algorithms. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pages 1–6. IEEE, 2020.

[22] Tripti Mishra, Dharminder Kumar, and Sangeeta Gupta. Students’ employability prediction model through data mining. International Journal of Applied Engineering Research, 11(4):2275–2282, 2016.

[23] GG Gokilam and K Shanthi. Performance analysis of various data mining classification algorithms on diabetes heart dataset. COMPUSOFT, An international journal of advanced computer technology, 5(3):2320–0790, 2016.

[24] Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwib, and Najoua Ribata. Educational data mining and analysis of students’ academic performance using weka. Indonesian Journal of Electrical Engineering and Computer Science, 9(2):447–459, 2018.

[25] Siddhu P Algur, Prashant Bhat, and Nitin Kulkarni. Educational data mining: Classification techniques for recruitment analysis. International Journal of Modern Education & Computer Science, 8(2), 2016.

[26] M Anoopkumar and AMJZ Rahman. Model of tuned J48 classification and analysis of performance prediction in educational data mining. Int. J. Appl. Eng. Res, 13(20):14717–14727, 2018.

[27] YK Salal, SM Abdullaev, and Mukesh Kumar. Educational data mining: Student performance prediction in academic. IJ of Engineering and Advanced Tech, 8(4C):54–59, 2019.

[28] P Gunasundari. A survey on data mining: Trends, techniques, tools and domain specific research issues, applic. Tathapi with ISSN 2320- 0693 is an UGC CARE Journal, 19(19):575–587, 2020.

[29] Bhrigu Kapur, Nakin Ahluwalia, and R Sathyaraj. Comparative study on marks prediction using data mining and classification algorithms. International Journal of Advanced Research in Computer Science, 8(3), 2017.

[30] Syarifah B Rahayu, Nur D Kamarudin, and Zuraini Zainol. Case study of upnm students performance classification algorithms. J Eng Technol, 7(4.31):285–289, 2018.

[31] Amjad Abu Saa, Mostafa Al-Emran, and Khaled Shaalan. Mining student information system records to predict students’ academic performance. In International conference on advanced machine learning technologies and applications, pages 229–239. Springer, 2019.

[32] Huda Al-Shehri, Amani Al-Qarni, Leena AlSaati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, and Sunday O Olatunji. Student performance prediction using support vector machine and k-nearest neighbor. In 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), pages 1–4. IEEE, 2017.

[33] Ahmed Mueen, Bassam Zafar, and Umar Manzoor. Modeling and predicting students’ academic performance using data mining techniques. International Journal of Modern Education and Computer Science, 8(11):36, 2016.

[34] MO Asanbe, AO Ososifan, and WF William. Teachers’ performance evaluation in higher educational institution using data mining technique. International Journals of Applied Information System (IJJAIS) Volume, 10:10–15, 2016.