



# Computational Prediction Algorithms and Tools Used in Educational Data Mining: A Review

Ameer K. AL-Mashanji <sup>1\*</sup>, Aseel Hamoud Hamza <sup>2</sup> and Laith H. Alhasnawy <sup>3</sup>

<sup>1</sup>Presidency of University, University of Babylon, amir.mashanji@uobabylon.edu.iq, Babylon, Iraq.

<sup>2</sup>College of Law, University of Babylon, aseel.hamod@uobabylon.edu.iq, Babylon, Iraq.

<sup>3</sup>Presidency of University, University of Babylon, laith.alhasnawi4@uobabylon.edu.iq, Babylon, Iraq.

\*Corresponding author email: amir.mashanji@uobabylon.edu.iq; mobile: 07704319686

## خوارزميات التنبؤ الحاسوبية والأدوات المستخدمة في التنقيب عن البيانات التعليمية: مراجعة

امير علي كاظم<sup>1\*</sup>، اسيل حمود حمزة<sup>2</sup>، ليث حامد حمزة<sup>3</sup>

1 رئاسة الجامعة، جامعة بابل، amir.mashanji@uobabylon.edu.iq، بابل، العراق

2 كلية القانون، جامعة بابل، aseel.hamod@uobabylon.edu.iq، بابل، العراق

3 رئاسة الجامعة، جامعة بابل، laith.alhasnawi4@uobabylon.edu.iq، بابل، العراق

Received: 31/1/2023 Accepted: 13/3/2023 Published: 31/3/2023

### ABSTRACT

Abstract In recent days, a wide variety of tools have appeared for performing educational data mining (EDM). The current education systems show that there are several factors affecting students' performances. First and foremost, students need motivation in order to learn and this motivation results into their success. The prediction of student performances is an important field of research in Educational Data Mining, particularly through the application of different data mining techniques. The majority of EDM research focuses on prediction algorithms. The current work presents a review of the data mining predicting algorithms and tools that have been adopted in EDM. It also provides insight into the algorithms and powerful data mining tools that most widely used in student performance prediction. This will mainly be of use for educators, instructors and institutions, increasing the students' levels of study.

### Keywords:

Classification Algorithms, Data Mining Tools, Educational Data Mining, and Regression Algorithms

### الخلاصة:

في الأيام الأخيرة ، ظهرت مجموعة متنوعة من الأدوات لأغراض أداء التنقيب عن البيانات التعليمية (EDM). تظهر أنظمة التعليم الحالية أن هناك عدة عوامل تؤثر على أداء الطلاب. أولاً وقبل كل شيء ، يحتاج الطلاب إلى الدافع من أجل التعلم وهذا الدافع يؤدي إلى نجاحهم. يعد التنبؤ بأداء الطلاب مجالاً مهماً للبحث في استخراج البيانات التعليمية ، لا سيما من خلال تطبيق تقنيات التنقيب عن البيانات المختلفة. تركز غالبية أبحاث EDM على خوارزميات التنبؤ. يقدم العمل الحالي مراجعة لخوارزميات التنقيب عن البيانات والأدوات التي تم تبنيها في EDM. كما يوفر نظرة ثاقبة للخوارزميات وأدوات التنقيب عن البيانات القوية الأكثر استخداماً في التنبؤ بأداء الطلاب. سيكون هذا مفيداً بشكل أساسي للمعلمين والمرشدين والمؤسسات ، مما يزيد من مستويات الطلاب الدراسية.

### الكلمات المفتاحية:

خوارزميات التصنيف ، أدوات التنقيب عن البيانات ، التنقيب عن البيانات التعليمية وخوارزميات الانحدار .



## Introduction

Introduction Knowledge Discovery in Databases (KDD) is a process that involves analyzing and modeling large databases automatically explorative [1]. It is a controlled procedure whereby new and understandable patterns of validity and use are identified out of a large, complex data set. Data mining is an essential process in KDD, as it infers the algorithm for exploring data, developing models, and identifying patterns that were not known before. This model is necessary for the analysis and prediction of phenomena in data [1]. Educational Data Mining (EDM) is a field of research that deals with the data of educational institutions. This could include predicting student performances and learning analytics to classify students according to their learning performances [2]. EDM presents an analysis of data for information systems that support learning in a variety of educational institutions, such as schools, colleges, and universities, as long as they provide a conventional teaching model. EDM mainly aims to predict students' results and their modeling, which represent two essential issues within the educational context [2].

The work conducted in the present paper can be outlined in the following way. Section (2) describes the predicting techniques adopted in EDM. Section (3) draws a comparison between the research works that make use of predicting techniques in EDM. Section (4) states the conclusions of the present work.

## Materials and Methods

In EDM, prediction techniques are adopted for predicting the performance of students. This act requires several tasks, including classification and regression.

### 1. Classification Techniques

Classification is considered to be one of the supervised learning techniques which aim towards creating a classified model for classifying class labels for unknown data. In other words, a classifier is formed using the training set, and it is applied in the classification of unknown data into predetermined classes that already exist [3]. The classifier evaluation involves taking an already classed input, through which its accuracy is obtained by the rate of correctly classified data items [3]. The following section describes some of the algorithms used in the classification process.

#### 1.1 K-Nearest Neighbor

K- Nearest neighbor (KNN) classifier is a completely different approach to classifying data. No obvious universal model is built, as it has only a local and implicit estimation[4]. The essential concept here is the classification of objects utilizing the examination of K-class values in similar data points. The class that is selected could be the class or class distribution with the highest frequency in the neighborhood. There are only two tasks to be learned in the KNN classifier: selecting the number of (k) neighbors and the (d) representing distance metric [4].



## 1.2 Naive Bayes

Naive Bayes (NB) classifier is a statistical classifier that assumes that features are statistically independent of each other. NB classifier is based on the assumption that the impact of attributes does not depend on the value of other attributes in a certain class [4]. As a result of this independence, the NB classifier is highly scalable and can learn quickly when using a large dataset with high dimensional attributes. This is useful for many real-world datasets, such as speech data, text image data, spam filtering, and medical image processing [4].

## 1.3 Decision Tree

Decision Tree (DT) is a commonly used classification model. It consists of a tree whereby branch nodes represent a choice between several options. The leaf nodes represent decisions. DT is often applied in gathering information to help in decision-making. DT begins at a root node for users for taking action. Starting from this node, the DT learning algorithm splits into other nodes. As a result, the branches in the DT represent possible scenarios of decisions and their results [5].

## 1.4 Bayes Network

Bayes Network (BN) (also known as a network decision, Bayes net, or belief network), this statistical classifier has a visual representation in form of a graph structure through a directed no-period graph. It is used in predicting class memberships through the probability of a certain sample belonging to a given class. For example, BN represents the probabilistic relationships between symptoms and diseases. Given the symptoms, where the network is used to calculate the probabilities of having different diseases [6].

## 1.5 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) consists of several input and output units that are connected with a certain weight. Throughout the learning stage, the weights are adjusted to help the network in predicting the correct labels for the input tuples. MLP is well-studied for continuous-valued inputs and outputs. MLP is considered to be better than the other networks in terms of pattern or trend identification in data. It is therefore suitable for predicting student performances [7].

## 1.6 Support Vector Machine

Support Vector Machine (SVM) is a successful method to be used with non-linear class boundaries. However, a drawback is that there is often a lack of data to learn composite non-linear models. This method is based on the idea that in the mapping of data into higher dimensions, a linear appears for the classes. Practically, only absolute mapping occurs utilizing the kernel functions [8].



## 1.7 Random Forest

Random Forest (RF) is an ensemble predictor technique that depends on the decision tree model. RF algorithm can be used for both regression and classification tasks. It works well with large datasets and achieves accurate results. With regression, the result is returned based on the average output of all trees while with classification, the result is returned based on the votes of all trees [9].

## 2. Regression Techniques

Algorithms in this type predict continuous values determined by the input variable. This technique is often used when the target variables represent quantities, like incomes, scores, heights, or weights, whereby the output is a continuous value, such as an integer or floating-point value [10].

### 2.1 Linear regression

Linear regression is a supervised learning algorithm that forms a special case of regression analysis. The main idea of this model is to explain the relationship between one dependent variable (usually denoted by Y and indicating the target class) and one or more independent variables (usually denoted by x and indicating the features), using a straight line [10].

### 2.2 Support Vector Regression

In problem regression, SVM is called (Support Vector Regression) SVR. It is a supervised learning method characterized by the use of kernels as it can handle nonlinear prediction very efficiently through a nonlinear kernel function. One of its most important strengths is that it is used to build classification models or regression methods as well as to achieve significant results with large datasets. SVR works well with high-dimensional datasets and thus avoids the curse of the dimensionality problem [11].

## An Overview of Important EDM Tools

The tools included in the next section offer a variety of algorithms that can be used to predict and find relationships in educational data.

- **WEKA:** It is a free and open-source software package developed at the University of Waikato in New Zealand that assembles a set of data mining and model-building algorithms. WEKA contains a free set of classification, regression, clustering, and feature selection algorithms and a set of algorithms for data analysis, predictive modeling, and visualization tools. The logo of the Weka tool and the Main Interface is shown in Figure (1) and Figure (2) respectively [12].



WEKA

Figure. 1 The Logo of Weka Tool

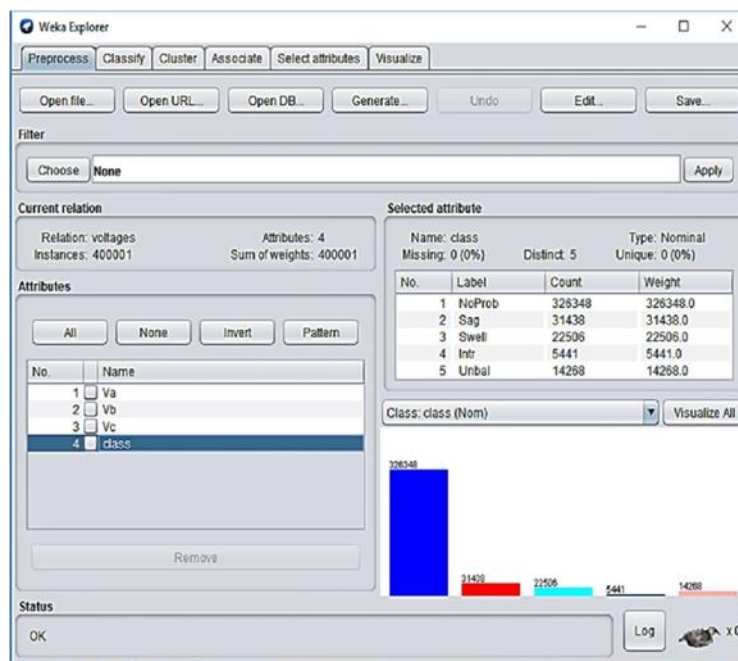
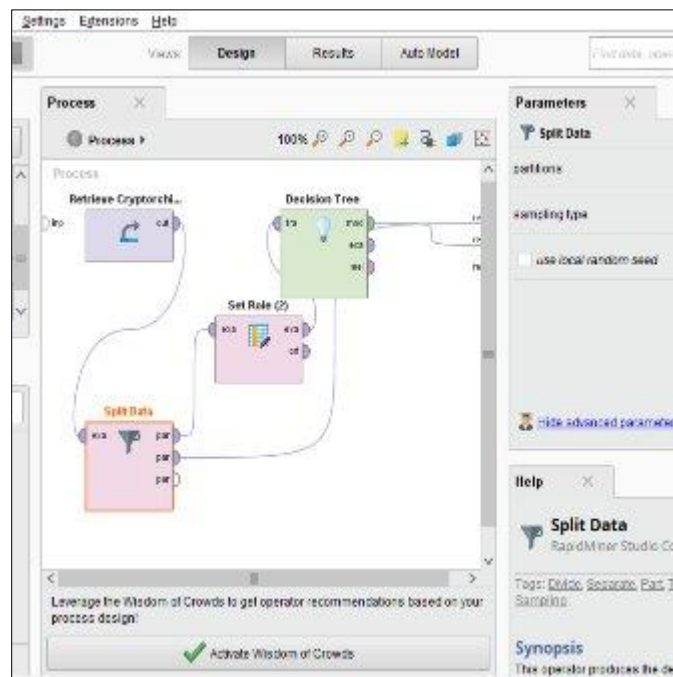


Figure 2 The Main Interface of the Weka Tool

- Rapid Miner:** Rapid Miner contains a very wide range of regression and classification algorithms as well as clustering algorithms, association rule mining, and other applications. It is a package for building models and performing data mining analysis developed by a company of the same name. It provides a varied environment for preparing data, deep learning, machine learning, and performing predictive analysis. It supports all steps of machine learning processes including data preparation, results visualization, model validation, and optimization. The Rapid Miner tool logo and main interface are depicted in Figure (3) and Figure (4) respectively [12].



**Figure. 3 The Logo of Rapid Miner Tool**



**Figure. 4 The Main Interface of the Rapid Miner Tool**

### Prediction Methods Used In EDM

To present a comprehensive view of the predicting algorithms used in Educational Data Mining, there are (18) related works selected that are of relevance to the subject of this study. The student performances were predicted using a varying range of classification and regression algorithms. These works have been analyzed for improving the student's performances in light of the predicted results. Through such predictions, instructors and institutions can identify weaknesses at an early stage and treat or assess them appropriately, to improve their overall performances.

The scope of studies involved in this review are published between 2017 and 2022. Table (1) below draws a comparison between the predicting algorithms used in Educational Data Mining.





Table 1: Prediction algorithm as Applied in EDM

Ref.	Algorithm and Task Type	Tools Used and	Year	Aim of Paper
[13]	DT, NB and RF	Weka	2017	Prediction of Student Performances
	Classification			
[14]	NB and KNN	Rapid Miner	2017	Prediction of Students' Academic Performances
	Classification			
[15]	LR and MLP	Weka	2017	Predicting Student Performance in Final Examination
	Regression			
[16]	RF and LR	Weka	2017	For Predicting the Alumni Earning
	Regression			
[17]	RF and DT	Weka	2018	For Predicting The Grades Students in Module
	Classification			
[18]	KNN	Weka	2018	For Predicting the Performance of Students to Prevent Non Active
	Classification			
[19]	BN, NB, MLP DT and RF	Weka	2018	To Predict the Student's performance
	Classification			
[20]	RF, DT and	Weka	2018	



	BN			Prediction of the Performance of Students and to Prevent Drop Out
	Classification			
[21]	SVM and KNN	Weka	2018	Prediction of the alumni Income
	Classification			
[22]	RF and DT	Weka	2019	Prediction of the grades of students in the Research Project.
	Classification			
[23]	KNN, NB, and DT	Rapid Miner	2019	Prediction of the Student's Performance into non Excellent or Excellent
	Classification			
[24]	MLP ,RF, SVM, DT and NB.	Rapid Miner	2019	For Predicting the Final Student Grade
	Classification			
[25]	RF and SVM	Weka	2019	Predict the Alumni Completion
	Regression			
[26]	KNN and DT	Rapid Miner	2020	Analysis of the Student Performance
	Classification			
[27]	RF, SVR, and LR	Weka	2020	For Predication The Graduate's Earnings

مجلة جامعة بابل للعلوم التطبيقية وعلوم الحاسوب  
 Journal of Babylon University for Applied Sciences and Computer Science

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com ISSN: 2312-8135 | Print ISSN: 1992-0652





	Regression			
[28]	DT and NB	Weka	2021	For predicting the Student's Performance Depending on their Past Academic Records
	Classification			
[29]	DT, NB, MLP, and RF	Weka	2022	To predict students' final grades depending on past research
	Classification			
[30]	NB, RF, MLP, and DT	Weka	2022	For Prediction Students Academic Performance
	Regression			

In light of the comparison made in Table (1), the prediction algorithms are illustrated in Figure (5) below based on the frequency of use in Educational Data Mining.

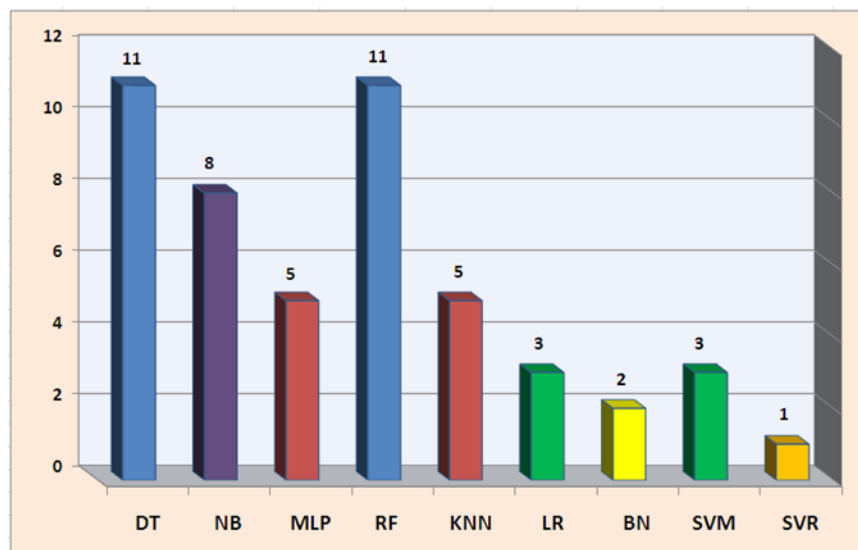


Figure.5 Prediction Algorithm of EDM

Figure (6) show that the majority of researchers in EDM made use of the Weka tool in predicting the students' performances, whereas Rapid Miner is hardly used in the compared works.

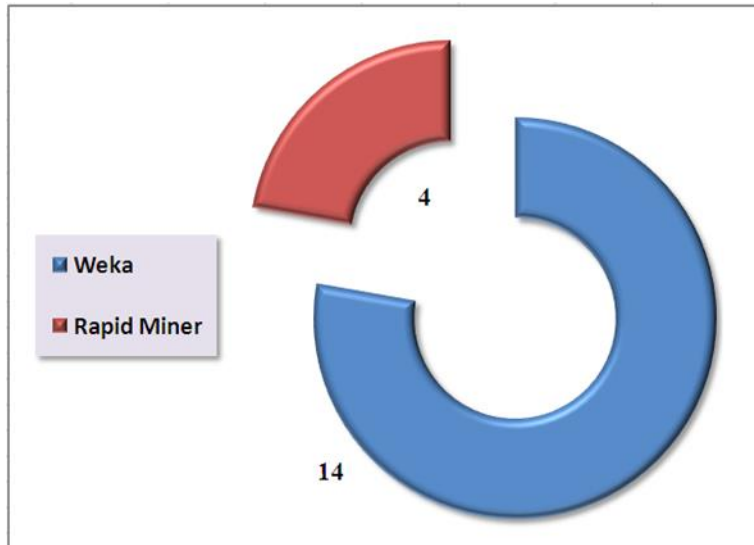


Figure. 6 Tools used in EDM

Figure (7) below shows that the majority of researchers used the classification task in their research and rarely used the regression task for predicting student performances.

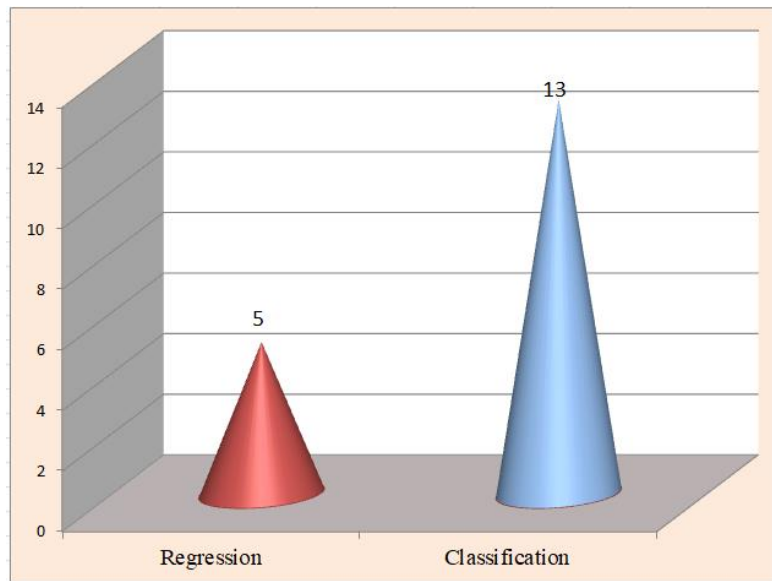


Figure. 7 Prediction task used in EDM



## Conclusion

The present review study discussed several predicting tools and techniques that have been used in the educational context for predicting the performances of students. Such predictions are useful for parents, teachers, and scientific institutions in the future. The conclusion can be drawn that the majority of works address the same predicting algorithms and it has been found the RF and DT algorithms are most frequently used by researchers in the prediction tasks in their research. Therefore, It is suggested that these algorithms can be used in EDM prediction, using different student variables and data mining techniques for obtaining better results. It was also found that most of the researchers used the classification task in their research and rarely used the regression task to predict students' performance. In addition, most EDM prediction researchers used the Weka tool to predict student performance, while Rapid Miner was rarely used. Finally, we hope this review is useful for researchers to learn about the emerging algorithms and tools in the prediction of EDM.

## Conflict of interests.

There are non-conflicts of interest.

## References

- [1] A. A. Saa, "Educational data mining & students' performance prediction", *International Journal of Advanced Computer Science and Applications*, vol.7,no.5, pp.212-220, 2016.
- [2] M. Zaffar, M. A. Hashmani and K. S. Savita, "Performance analysis of feature selection algorithm for educational data mining", In *2017 IEEE Conference on Big Data and Analytics (ICBDA)*,2017, pp. 7-12. DOI: [10.1109/ICBDAA.2017.8284099](https://doi.org/10.1109/ICBDAA.2017.8284099)
- [3] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: methods, metrics, and applications", *IEEE Access*, 5, 2017, pp. 10562-10582. DOI: [10.1109/ACCESS.2017.2706947](https://doi.org/10.1109/ACCESS.2017.2706947)
- [4] B. Kapur, N. Ahluwalia and R. Sathyraj, "Comparative Study on Marks Prediction using Data Mining and Classification Algorithms", *International Journal of Advanced Research in Computer Science*, vol.8,no.3, 2017.
- [5] S. Sivakumar, S. Venkataraman and R. Selvaraj, "Predictive modeling of student dropout indicators in educational data mining using improved decision tree", *Indian Journal of Science and Technology*, vol.9,no 4, pp.1-5,2016.
- [6] D. Kabakchieva, "Predicting student performance by using data mining methods for classification", *Cybernetics and information technologies*, vol.13 ,no.1, pp.61-72, 2013.
- [7] R. M. Ahmed, N. F. Omran and A. A. Ali, "Predicting and Analysis of Students' Academic Performance using Data Mining Techniques", *International Journal of Computer Applications*, vol. 182 ,no.32, pp.0975 - 8887,2018.
- [8] Y. Al Amrani, M. Lazaar and K. E. El Kadiri , "Random forest and support vector machine based hybrid approach to sentiment analysis", *Procedia Computer Science*, vol.127, pp.511-520,2018. DOI: [doi.org/10.1016/j.procs.2018.01.150](https://doi.org/10.1016/j.procs.2018.01.150)
- [9] A. L. Boulesteix, S. Janitza, J. Kruppa and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*,vol. 2,no.6, pp.493-507,2012. DOI: [doi.org/10.1002/widm.1072](https://doi.org/10.1002/widm.1072)



- [10] I. Kawashima and H. Kumano, "Prediction of mind-wandering with electroencephalogram and non-linear regression modeling", *Frontiers in Human Neuroscience*, vol.11, pp.1-10,2017 .DOI: [doi.org/10.3389/fnhum.2017.00365](https://doi.org/10.3389/fnhum.2017.00365)
- [11] Li, Y. Bontcheva and H. Cunningham , "Adapting SVM for data sparseness and imbalance: a case study in information extraction", *Natural Language Engineering*, 15(2), pp. 241-271, 2009. DOI: [doi.org/10.1017/S1351324908004968](https://doi.org/10.1017/S1351324908004968)
- [12] A. Jovic, K. Brkic and N. Bogunovic , "An overview of free software tools for general data mining", In *2014 37th International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 1112-1117, 2013. DOI: [10.1109/MIPRO.2014.6859735](https://doi.org/10.1109/MIPRO.2014.6859735)
- [13] M. Kumar and A. J. Singh, "Evaluation of Data Mining Techniques for Predicting Student's Performance", *International Journal of Modern Education & Computer Science*, vol.9,no.8,2017. DOI: [10.5815/ijmeecs.2017.08.04](https://doi.org/10.5815/ijmeecs.2017.08.04)
- [14] I. A. A. Amra and A. Y. Maghari, "Students performance prediction using KNN and Naïve Bayesian", In *2017 8th International Conference on Information Technology (ICIT)*, pp. 909-913,2017. DOI: [10.1109/ICITECH.2017.8079967](https://doi.org/10.1109/ICITECH.2017.8079967)
- [15] F. Widyahastuti and V. U. Tjhin, "Predicting students performance in final examination using linear regression and multilayer perceptron", In *2017 10th International Conference on Human System Interactions (HSI)*, pp. 188-192, 2017. DOI: [10.1109/HSI.2017.8005026](https://doi.org/10.1109/HSI.2017.8005026)
- [16] T. A. Wotaifi and E. S. Al-Shamery, "Fuzzy-Filter Feature Selection for Envisioning the Earnings of Higher Education Graduates", *Compusoft*, vol.7,no.12,pp. 2969-2975, 2018.
- [17] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood and K. U. Sarker, "Student academic performance prediction by using decision tree algorithm", In *2018 4th international conference on computer and information sciences (ICCOINS)*, pp. 1-5, 2018.DOI: [10.1109/ICCOINS.2018.8510600](https://doi.org/10.1109/ICCOINS.2018.8510600)
- [18] S. Wiyono and T. Abidin, "Implementation of K-Nearest Neighbour (Knn) Algorithm To Predict Student'S Performance", *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, vol.9,no.2, pp.873-878, 2018. DOI: [doi.org/10.24176/simet.v9i2.2424](https://doi.org/10.24176/simet.v9i2.2424)
- [19] M. Zaffar, M. A. Hashmani, K. S. Savita and S. S. H. Rizvi, "A study of feature selection algorithms for predicting students academic performance", *International Journal of Advanced Computer Science and Applications*, vol.9,no.5 , pp.441-440, 2018.
- [20] S. Hussain, N. A. Dahan, F. M. Ba-Alwib and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA", *Indonesian Journal of Electrical Engineering and Computer Science*, vol.9,no.2, pp.447-459,2018. DOI: [10.11591/ijeecs.v9.i2.pp447-459](https://doi.org/10.11591/ijeecs.v9.i2.pp447-459)
- [21] Strand, Miranda and Tommy Truong, "Predicting Student Earnings After College".
- [22] E. C. Abana, " A decision tree approach for predicting student grades in Research Project using Weka", *Int. J. Adv. Comput. Sci. Appl*, vol.10,no.7, pp.285-289,2019.
- [23] W. F. W. Yaacob, " Supervised data mining approach for predicting student performance", *Indones. J. Electr. Eng. Comput. Sci*,vol.16,no.3,pp. 1584-1592, 2019. DOI: [10.11591/ijeecs.v16.i3.pp1584-1592](https://doi.org/10.11591/ijeecs.v16.i3.pp1584-1592)
- [24] J. Sultana, M. U. Rani and M. A. H. Farquad, " Student's performance prediction using deep learning and data mining methods", *Int. J. Recent Technol. Eng*,vol. 8,no.1S4, pp.1018-1021,2019.
- [25] T. A. Wotaifi, " Mining of Completion Rate of Higher Education Based on Fuzzy Feature Selection Model and Machine Learning Techniques", *Int. J. Recent Technol. Eng*, vol.8,no.2S10, pp. 393-400,2019.DOI: [10.35940/ijrte.B1067.0982S1019](https://doi.org/10.35940/ijrte.B1067.0982S1019)
- [26] L. D. Yulianto, A. Triayudi and I. D. 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: 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*, vol.4,no1, pp.441-451,2020.
- [27] T. A. Wotaifi and E. S. Al-Shamery, "Modified Random Forest based Graduates Earning of Higher Education Mining", *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.*,vol. 12, pp. 56-65,2020.



- [28] A. Ranjan and R. Raj, "Deep, A., and Senapati, K. K. Student Performance Prediction Using Classification Algorithms", In *Proceedings of the Fourth International Conference on Microelectronics, Computing and Communication Systems*, Springer, Singapore, pp. 469-477, 2021.
- [29] M. Kumar, C. Sharma, S. Sharma, N. Nidhi and N. Islam, "Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance", In *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, pp. 1013-1017, 2022. DOI: [10.1109/DASA54658.2022.9765236](https://doi.org/10.1109/DASA54658.2022.9765236)
- [30] M. P. R. I. R. Silva, R. A. H. M. Rupasingha and B. T. G. S. Kumara, "A Comparative Study of Predicting Students' Academic Performance Using Classification Algorithms", In *2022 2nd International Conference on Advanced Research in Computing (ICARC)*, pp. 385-390, 2022. DOI: [10.1109/ICARC54489.2022.9753729](https://doi.org/10.1109/ICARC54489.2022.9753729)