A Comparison of EDM Tools and Techniques

Eman Alshehri[1], Hosam Alhakami[2], Abdullah Baz[3], Tahani Alsubait[4]
College of Computer and Information Systems
Umm Al-Qura University
Makkah, Saudi Arabia

***Abstract*—Several higher educational institutions are adapting the strategy of predicting the student's performance throughout the academic years. Such a practice ensures not only better academic outcomes but also helps the institutions to reorient their curriculums and teaching pedagogies so as to add to the students' learning curve. Educational Data Mining (EDM) has risen as a useful technology in this league. EDM techniques are now being used for predicting the enrolment of students in a specific course, detection of any irregular grades, prediction about students' performance, analyzing and visualizing of data, and providing feedback for overall improvement in the academic spheres. This paper reviews the studies related to EDM, including the approaches, data sets, tools, and techniques that have been used in those studies, and points out the most efficient techniques. This review paper uses true prediction accuracy as a standard for the comparison of different techniques for each EDM applications of the surveyed literature. The results show that the J48 and K-means are the most effective techniques for predicting the students' performance. Furthermore, the results also cite that Bayesian and Decision Tree Classifiers are the most widely used techniques for predicting the students' performance. In addition, this paper highlights that the most widely used tool was WEKA, with approximately 75% frequency. The present study's empirical assessments would be a significant contribution in the domain of EDM. The comparison of different tools and techniques presented in this study are corroborative and conclusive, thus the results will prove to be an effective reference point for the practitioners in this field. As a much needed technological asset in the present day educational context, the study also suggests that additional surveys are recommended to be driven for each of the EDM applications by taking into account more standards to set the best techniques more accurately.**

***Keywords*—Educational Data Mining (EDM); students' performance; prediction; higher education; WEKA**

I. Introduction

Data mining is the most effective process for analyzing big data warehouses to derive valid and useful information, to extract hidden data, and to detect relationships between factors in massive data [1]. The educational data mining (EDM) process uses computational methods to convert raw data from educational systems into useful information to help educational issues [2]. Education nowadays contains several enhancement methods used to supervise and identify the students' academic performance in their studies. Data mining has been considered as one of the most useful processes used to identify students' performance. Presently, the scope of Data Mining has not limited to education only as it covers almost all those domains where data is used. There are many examples of applications using data mining. Retail management is one of such applications. Other examples include applications within banking sector, telecommunication, marketing, hospitality, production management, and so on. These organizations take the benefits from data mining to increase their income and future growth [3]. The data extraction in the field of education through using data mining methods is typically known as Educational Data Mining (EDM). Nowadays, EDM is a new discipline concerned with a different approach [1][4]. In the current context, education provides various ways and systems of learning that students can access. To quote a few, these include: Learning Management System (LMS) which is so popular and needed these days along with conventional classroom learning and Learning Object (LO). Social networking and online forums are other also needed in the E-learning process. Adaptive Hypermedia systems, educational games, concept maps and online exams are other points of educational contact needed by students worldwide. Each of these platforms brings several types of data, which EDM has to handle. [5]. Many educational institutions assess the students' performance depending on the course content and knowing the objectives to fulfill an effective learning process [4]. One of the biggest goals of higher education institutions is to improve and enhance the quality education process for its students. One way to improve the quality level in the education system is by discovering and applying data mining techniques [6]. Data mining is used to predict students' registration in a certain course and to detect any abnormal values of grades, prediction about students' performance, analysis and visualization of data, providing feedback to support instructors, recommendations for students, and so on [7][8]. Data mining techniques also assist in advising students in choosing the appropriate subjects for their undergraduate or postgraduate courses in the university. Data mining discovery has become an area of growing importance, especially in education, as it assists in students' data analysis by using several factors and interpreting it to deliver a useful information [3].

This paper surveys literature review regarding EDM with data sets size and techniques used in such studies. The present study also aims at identifying the most effective technique for Educational Data Mining. This paper is divided into five sections. Section 1 shows an introduction for the paper purpose and structure. Section 2 examines the background of Educational Data mining (EDM) methods. Details about phases of data mining are shown within Section 3. In Section 4, we have discussed some applications of educational data mining, while Section 5 discusses the results. Section 6 concludes this research and posits suggestions for future work in the same domain.

II. Background

EDM certainly helps in reaching the needed goals for the educational process. By applying EDM methods, we can build prediction models for enhancing students' performance [9].

A. Data Mining Techniques

There are many research papers and studies regarding the use of data mining techniques in education. The most common and widely used techniques for predicting students' performance are regression and classification, but other techniques have also been used, such as Clustering [10]. The EDM is useful for improving the process of studying, advising students, finding the reasons leading to dropouts, predicting students' performance, detection of undesirable behaviour. EDM can also help the educators to track academic progress to improve the teaching process. These algorithms in data mining require a quick mention to be familiar with [11]. A list of techniques explanation is stated below:

- **Classification** is one of the data mining applications that divides data into target classes [12]. The classifier algorithm uses a pre-classified prototype for identifying the set of parameters required in classification, for allocating a category to a record. The classification aims to predict the target class for each status of data accurately [13]. In EDM, this technique is used for classifying students based on the characteristics such as age, gender, grades, behavior, etc. The major classification algorithms are BayesNet, Naïve Bayes, C4.5 (J48), ID3 and Neural Network (NN).
  - **J48** classifier is a kind of decision trees algorithms. It consists of many nodes starting from the root till ending with the leaf. This algorithm can fix the issue of overfitting data and un- pruning. It is also able to specify the attributes are relevant or irrelevant at classification. In each node of the tree, J48 chooses the best effective feature to divide its sample into subsets at different classes. J48 can also deal with continuous and discrete data. [14]. Also, J48 algorithm repeatedly classifies data until it reaches the optimum level of categorization.
  - **Naive Bayes** classification is a simple classification algorithm that can calculate the probability by calculating the combination of entries and frequency in the data set [15].
- **Regression** technique is mainly used when there is a need to predict how one or more independent variables are related to dependent ones. Dependent variables are the ones to be predicted, while independent variables are known in prior [13] [16]. In EDM, it is used for the prediction of students' academic performance, prediction of the final grade, etc. Support Vector Machine (SVM), linear regression and neural networks are some common methods for regression within educational data mining [17]. Moreover, Classification and Regression Trees can be used together at the same model like CART technique.
- **Clustering** splits the data set into different groups, known as clusters. Clustering is needed mainly in cases where the ultimate usual data set categories within the data set are not known previously. The data point within a cluster should be same as other data points within the same cluster and different from data points of different cluster [18] [19]. In EDM, the clustering technique is used for grouping according to similarities and differences between students, courses, behavior, etc. [13]. The most popular clustering methods are K-mean and X-mean [20].
- **Decision Trees** are used for applying the classification model in the form of a tree structure. Each inner tree node represents examination for attribute, while a branch is a symbol for the result of that examination, and each leaf node acts as a classification. The classification rule is a path from the root to the leaf. Stability and easy interpretation is the biggest advantage of this technique. It is also suitable for solving different problems in various sectors, such as finance, business, education, etc.

B. Data Mining Tools

There are several tools that help in data mining and most of them are open source. An exhaustive perusal of the literature in this context helped us to list out the tools that have been used frequently in different research studies. The tools are listed below:

- **WEKA** is a java based tool used to process big data sets. It includes different algorithms, which may be applicable within data mining techniques [21]. It can be easily applied to algorithms to obtain quick results [22].
- **RapidMiner** is a tool developed by the rapid miner company. It provides features for machine learning, data mining, etc. It is mainly used for research, rapid prototyping, training, and of course education. It assists all the phases of data mining, including results such as validation, optimization and visualization [23].

III. Data Mining Process

The data mining process consists of phases that allow us to convert unknown information (raw data) into valid and meaningful information (knowledge) [24]. The following Fig. 1 shows knowledge extraction in sequential steps that are part of data mining process.

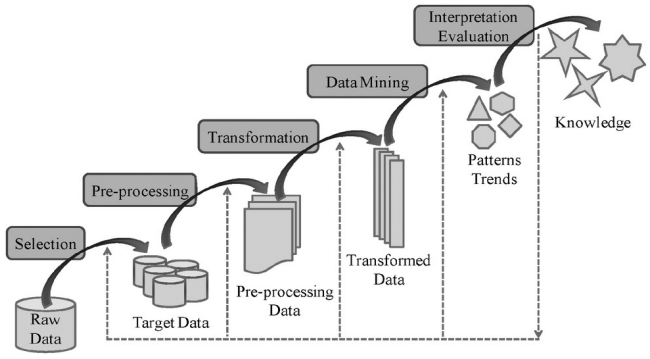

Fig. 1. Data Mining Process[25].

1) Data Selection: This step involves selecting or retrieving a data set in which the process of discovery has to be performed as a pre-processed data [26].
2) Data Pre-Processing: This step involves making the data more reliable in this stage, i.e., remove irrelevant data from the data set, and find the missing values and handle it [27].
3) Data Transformation: In this step the data is transformed and is categorised into appropriate formats for mining, i.e., performing some algorithms as classification and clustering [27].
4) Data Mining: It is one of the most significant steps of the process in which techniques and tools are applied to extract useful patterns. Data mining algorithms include classification, clustering, regression, etc. [28].
5) Pattern Evaluation: In this stage, we can identify specific patterns and evaluate to come up with the desired goals.
6) Knowledge Representation: This is the last phase where the knowledge obtained previously is visually represented to the user. This stage makes use of visualization techniques to help the users to have full images of the results and interpret the outcomes [26].

IV. Literature Review

The literature review undertaken for the present research context specifically includes research studies published in the last five years. These studies have assisted in predicting students' performance through EDM techniques. Moreover, the analysis also revealed different educational data mining techniques that are being used, tools, data set size, the best algorithm used with the highest prediction accuracy. Table I illustrates the summary of the reviewed literature.

A. Prediction of Students' Performance

In 2020, Authors in [29] used Naïve Bayes and J48 techniques for students' academic performance prediction and guided the students by using WEKA tool. They used over 3867 students' records upon of 5 years of Umm Al-Qura University. The results concluded that J48 algorithm achieved the best accuracy of 84.38% while NaiveBayes algorithm gave an accuracy of 46.68%.

Authors in [30] stated that they used educational data mining to predict Semester Grade Point Average (SGPA) of Bachelor of Technology (B. Tech) third-semester computer engineering students. They used a classification based on previous academic performance and student's social conditions. The authors used two classification algorithms were REP Tree and J48 on the data set using 10- fold cross validation to find the relationship between social parameters and students' performance. This was also used for predicting students' performances in the third semester. They applied these algorithms on the data set of 236 at computer engineering students of Punjabi University. Data was collected through a structured questionnaire from students pursuing B. Tech Computer Engineering from the Computer Engineering Department, Punjabi University. They conducted their study on a sample of 260 students having 17 attributes which included social parameters and previous academic performance such as (fathers' and mothers' education, living place during B. Tech, marks obtained in 10th English, marks obtained in 12th English and marks obtained in 10th Math). The study revealed that parent's education affected student's performance. Moreover, second semester performance played an important role for third semester performance. Finally, they found that a J48 Algorithm gave higher accuracy than REP Tress. The accuracy of J48 was 67.37%, while REP Tree was 56.78%.

According to the authors in [31], they worked on predicting the student's final grade based on the information collected in the early stage. Prediction was based on two different training data sets. Each data set contained data of different students in the last four semesters in the period from 2013 to 2015. In their study, the authors used data based on the following parameters (grade, test1, test2, lecture presence, and lab-presence). The research was executed on the basis of two separate experiments. Both experiments had the same goal, as we mention, which was to predict students' final grades. Data mining classification algorithms were applied on both data sets separately, and the goal was to predict students' final marks based on those two data sets. The data was collected from one particular course. The first training data set contained information about a number of students' visits to the lectures and laboratory exercises. The second training data set contained more students' data besides lecture visits which were added as two more parameters. These two parameters were students' results on two tests performed during the semester. The authors concluded that students must be present in one-third of the total number of lectures and laboratory classes to pass a particular exam. The study cited that when the total number of students present in the lecture and laboratory classes was greater than two-thirds of the total number of classes, the students would get a high grade. Also, The authors also used an IBK algorithm, which is the implementation of K-nearest neighbor classifier for the first data set. This is besides J48 classification algorithm, which is an implementation decision tree classifier for another data set. The authors concluded that they found in the Second training that IBK algorithms provided the best performance of 98.58%. Besides, the J48 technique also provided a good performance of 86.40%.

According to [32], the authors used data mining methods for students' academic prediction. They used data based on different parameters, such as teaching material access duration, academic performance for students, including assignments and tests, and discussion forums. In their study, they collected information about students who taken Programming Fundamental and Advanced Operating System courses from August 2014 to May 2015. Then, their study applies three classifiers on the compared, tested, and analyzed data set. The classifiers were Naïve Bayes, Multi-layer perception, and C4.5 (J48). The three classifiers were tested on 38 attributes. They applied ten-fold cross-validation, which means that the data set was randomly divided into ten subsets of the same size. The authors concluded that the Naïve Bayes classifier gives the best overall prediction accuracy than the other two classifiers with 86%.

The study in 2020 [33] by Zainab Mohammed et al. used WEKA tool to predict the academic instructors' performance by Using K-means Clustering and Naive Bayes classifications. They used data set at the UCI website that contained a total of 5820 evaluation scores provided by the university students for the evaluation of the academic instructors' performance

and attributes such as instructor's name, course code values, and the course attendance rate. The result found that Naive Bayes classification had an accuracy of 98.86%, while 98.9% for K-means Clustering.

Al Breiki et al. [34] used data mining algorithms to predict the performance of students by using WEKA tool. They analyzed 145 Students data of two academic years (2009- 2010 and 2014-2015) and attributes including student ID, secondary education GPA (/100), and cumulative GPA (/4.00) at the United Arab Emirates University. They applied eight techniques such as Decision Table (DT), propositional rule learning (JRip), Simple Log Regress (SLR), Gaussian Processes Random Tress (GPRT), K-nearest Neighbors (IBK) and Random Forest (RF). The results conclude that Regression gives the best prediction accuracy with 96.98%, followed by Random Forest 96.4%.

In addition, the study [35] by A. Tekin used three prediction techniques of data mining: SVM, ELM, NN and applied to data taken from students who studies computer science and information technology at the end of their 1st, 2nd and 3rd year courses to predict the GPA at graduation. He collected 127 students' records of the Collage of Computer and Instructional Technology enrolled in the Fırat University in Turkey either from 2006 to 2010 or from 2007 to 2011 with the attributes such as grades of students, cultural courses, and student's GPAs. The result of their study highlight that SVM had higher prediction accuracy with a rate of 97.98%, then ELM comes with a 94.8% accuracy rate, and finally NN with the least accuracy rate of 93.76%.

Y. K. Saheed et al. in their study of 2018 [6] applied a data mining technique to predict students' performance based on ID3, J48, and CART algorithms using WEKA tool. They collected 234 student records from the Faculty of Natural Science and Department of Computer Sciences for the years 2013 and 2014 at a private University in Northern part of Nigeria. The result of their study conclude that J48 and CART resulted in the same accuracy of 98.3%, while an ID3 gave 95.9% accuracy in prediction.

Fadhilah Ahmad et al. in their study [36], applied Decision Tree, Naive Bayes, and Rule-Based techniques for predicting the academic performance of first-year bachelor students at the Collage of Informatics and Computing, University Sultan Zainal Abidin, Malaysia. They collected 497 students' records across the span of eight years from 2006/2007 till 2013/2014. The records included data about students such as previous academic records, the background of family and demographics, etc. The results from the research found that the Rule-Based was the best accuracy comparing to the other classification with accuracy rate of 71.3%, while Naive Bayes and Decision Tree found the accuracy of 67.0% and 68.8%, respectively.

In 2018, the study [4] by Alaa Hamoud et al. presented a model based on decision tree algorithms: J48, Random Tree, and REP Tree to predict students' performance. They surveyed students of the Computer Science and Information Technology College in Basrah University. They collected 161 questionnaires and attributes, including academic information, Social Information, Demographic Data, etc. Finally, their result found that the J48 algorithm had the highest accuracy of 62.1% compared to Random Tree and RepTree algorithms of 61.4%, 60.1%, respectively.

In 2020, the study [37] by Abdullah Baz et al. used Naive Byes classifier for predicting students' academic performance at Umm Al-Qura University, based on the final GPA using WEKA tool. The authors collected a dataset consisting of 138 students with 13 attributes. Finally, the results highlight that the Naive Bayes classification had an accuracy of 72.46%.

In 2019, Ramaswami et al. [38] used four data mining techniques that included Logistic Regression, Random Forest, k-Nearest Neighbour and Naïve Bayes. The authors used different techniques in order to improve the prediction accuracy of students' performance by using Python. They collected 240 students of Xorro-Q (Web-based audience interaction tool) from 2016 to 2017 with the attributes such as activity name, activity, test1, test2, and final exam score. The results found that Random Forest had the best accuracy of 74%.

The study [39] analyzed students' data in higher education to predict students' grades and to enhance the students' performance by finding the connection among three main dimensions. The first one was students' activities through e-Learning. The second dimension was teaching manner. And the last dimension is students' results. Their data was collected from the e-Learning system log file and the database of the British University in Egypt. The study's result appeared that the Naive Bayes network had rate of 87.07% prediction accuracy.

In 2017, Uddin, Humam and Nafis, Md Tabrez [40] applied data mining technology for acquiring student performance during their entire semester by using Rapid Miner Studio tool. They used three different clustering techniques: K-Means, K-Medoids, and X-Means for categorizing students. Data set of 94 students of Bachelor of Technology was collected. The batch included 24 attributes such as Aggregate percentage, industry internships, and projects completed. The results of the study showed that X-Means clustering technique gave the best result for students' performance of 86.17% and the accuracy of 81.91% for K-Means, and 84.04% for K-Medoids.

In 2020, Nemomsa et al. in their study [41] used six different classifiers J48, RandomForest, NaiveBayes, BayesNet, JRip, and PART to predict students' academic performance by categorizing student status into dropout/fail, poor, good, excellent, or average performer through predictive modelling by using WEKA tool. They applied the same classifications on two data sets. They collected 6,573 students from AMU student repositories and some data was collected by using questionnaire-based survey. The data collection covered four departments of Computer Science (CS), Water Resource and Irrigation Engineering (WRIE), Water Supply and Environmental Engineering (WSEE) and Hydraulics and Water Resource Engineering (HWRE) of the second-year first semester. For each department, one major course was selected, with several attributes including gender, department, course, course credit hours, grade, semester average GPA, cumulative GPA, and students' status. The result showed that J48 and JRip classifiers produced the highest classification accuracy of J48 is 99.4%, and for JRip, it was 99.3%.

Another study [42] by S. Senthil and W. LIN applied fourteen classification methods to enhance and evolve models that target students' performance prediction and identify the

impact attributes. The data set was taken from UCI Machine Learning Repository while contained 33 attributes of 649 students. The study's result highlighted that BayesNet, Multilayer-Perceptron, Simple Logistics Pegasos, KStar, JRip, and Random Forest were the best algorithms for accurate prediction. The algorithm with the highest accuracy was the Simple Logistic of 93.2% followed by Random Forest of 93.1%, and the worst algorithm was IBk with 82.1%.

The study [43] by Alhakami et al., analyzed students' performance and achievement regarding ABET files learning by using Naive Bayes and J48 algorithms. They used a data set consisting of 126 students and some attributes such as attendance, quiz, midterm, and grade at Umm Al-Qura University. The results appeared that the J48 classifier had the best accuracy of 100%.

B. Detecting Students Behavior

The study [44] by Pujianto et al. in 2017 aimed to assist the students at the faculty of literature and the likelihood of their success in adapting to new environments. In Indonesia, it has always been an issue for elected students to join the literature faculty, especially those who don't have linguistic qualifications in high school. Their study applied the Naive Bayesian classifier algorithm to predict those students' level of achievements in Literature Faculty who came from the non-linguistics major in senior high school. The authors used data set based on a survey published online for students from literature faculty all around East Java Province in Indonesia. Some attributes used in this study were the English national exam score and the number of books reads per month. Based on the data analysis by using the NBC algorithm, the authors found that the high school senior students who don't have linguistic traits were also able to join the literature faculty. For the analysis of talent and the national exam score, the results showed that the accuracy of NBC was 70%.

On the other hand, to predict the probability of student's graduation at JIIT University of India. R. Ahuja, and Y. Kankane [14] applied Seven techniques for education data mining. The techniques used include KNN, Naïve Bayes, Random Forest, Logistic Regression and Ctree by using R language. The data set used for this research consisted of 35 attributes of students that included both the academic and non-academic data such as students' grades, student's age, size of family, residential, travelling time from home to school. The data was complied by using college reports. The result of the study concluded that the Ctree and Random Forest algorithm performed much better than other algorithms. The prediction accuracy of the two techniques was 90.37% and 89.47%, respectively.

Another study in 2020 [45] by Hooshyar et al. proposed a novel algorithm called PPP to predict students' performance with learning obstacles by way of procrastination behavior. This was attempted by using eight different classifications such as L-SVM, R-SVM, DT, RF, and NN. The authors collected 242 students and 16 attributes such as open date of an assignment, date of first view of the assignment, and date of assignment submission from the University of Tartu in Estonia. The result showed that PPP algorithm had the maximum accuracy rate of 96%.

C. Enrolment Decision for Students

In 2020, Nurhachita et al., in their study [46], presented a comparison between K-Means and Naïve Bayes clustering methods to use data mining on new students' admission at the Universities Islam in Palembang by using Rapid Miner tool. The authors collected data from 2016 to 2019 of 18930 students with attributes such as students' name, school origin, secondary national examination score, and study programs. The result of their research conclude that the Naive Bayes classifier gave an accuracy of 9.08%.

In 2020, the study [47] by Hanan Mengash aimed to focus on helping universities make acceptance decisions by applying data mining techniques for applicants' academic performance prediction. The study used four classification techniques, including Decision Trees, Support Vector Machines, Artificial Neural Network and Naive Bayes, to predict the students' performance at the end of their school years. The data set contained 2039 students' records of Computer and Information Sciences collage at Princess Nourah Bint Abdulrahman University. The results found that the Artificial Neural Network (ANN) had more than 79% accuracy rate, which makes it better than other classification methods. The Naive Bayes had the worst results.

D. Miscellaneous Studies

In 2020, the study [48] by Fatima Alshareef et al. reviewed the related researches in EDM, including applications and techniques, and identified the best algorithm for each of the EDM applications. The authors had relied on the right prediction accuracy and use it as a guide for identifying effective techniques. Thus, their result conclude that Random Forest and Bayesian were the most algorithms performed effectively for predicting the performance of students. Furthermore, Social Network Analysis gave the best functionality for identifying student behaviors. Both Social Network Analysis and Clustering were the most effective algorithms for student modeling and students' grouping, respectively.

In 2019, Francesco Agrusti et al. [49] tried to collect studies that used EDM methods to predict dropouts students. They selected 73 studies related to this topic to analyze. Their study found six classification techniques that were used in that field. These were: Decision Tree, K-NN, SVM, Bayesian Classification, NN, and Logistic regression. Their study highlight that frequency of the use of Decision tree was 67%, followed by Bayesian Classification at 49 49%, Neural Networks 40%, and Logistic regression 34%.

V. Discussion

A thorough review of the existing research studies shows that there are several algorithms for EDM applications in the context of analyzing students' data to support the educational process. Table II shows a comparison among different research studies based on the highest prediction accuracy depending on the use of the techniques. The results found that the J48 and K-mean are the best effective algorithms in predicting students' performance. The EDM techniques that have achieved the highest usage are Bayesian Classification, followed by Decision tree classifiers, then Logistic regression, Neural Networks, and K-Nearest Neighbour. The study paper

TABLE I. SUMMARY OF RESULTS OF RESEARCH SEEKING STUDENTS PERFORMANCE PREDICTION

Ref.	Objective	Techniques used	Sample size	Best algorithm	Prediction Accuracy	Tool
[29]	Prediction Students Performance	J48 ;NB	38671	J48	84.38%	WEKA
[30]		RT ; J48	260	J48	67.37%	WEKA
[31]		IBK; J48; ZeroR; Part	-	IBK	98.58%	WEKA
[32]		NB; MP ; J48	60	NB	86%	WEKA
[41]		J48; RF ; NB; BN; JRip; PART	6573	J48	99.4%	WEKA
[42]		BN ; MP; SL; SPegasos; KStar; RF;JRip	649	SL	93.2%	WEKA
[39]		NB	3040	NB	87.07%	STATA
[40]		K-Means; K-Medoids ; X-Means	94	X-Means	86.17%	RapidMiner
[35]		NN ; SVM; ELM	127	SVM	97.98%	-
[49]		DT; K-NN; SVM;Bayesian; NN; LR	73	DT	67%	-
[38]		NB; LR; K-NN,; RF	240	RF	74%	Python programming
[34]		SLR; DT; GPRT; IBK; RF;MP;SMOReg;LA	145	LA	96.98%	WEKA
[33]		K-means; NB	5820	K-means	98.9%	WEKA
[36]		DT; NB; Rule-Based	497	Rule-Based	71.3%	WEKA
[6]		ID3; J48; CART	234	J48 and CART	98.3%	WEKA
[4]		J48, RT; REP Tree	161	J48	62.1%	WEKA
[37]		NB	138	NB	72.46%	WEKA
[43]		NB;J48	126	J48	100%	WEKA
[45]	Detecting Students Behavior	PPP; R-SVM; L-SVM; DT; NN; NB; GP; ADB; RF;	242	PPP	96%	-
[14]		NB ; KNN ; RF ; Ctree; LR; Rpart; J48	-	Ctree	90.37%	R programming
[44]		NB	50	NB	70%	WEKA
[46]	Enrollment Decision for Students	NB; K-means	18930	NB	9.08%	RapidMiner
[47]		DT; SVM; NB; ANN	2039	ANN	79%	WEKA

TABLE II. COMPARISON OF THE EDM TECHNIQUES REGARDING ON PREDICTION ACCURACY.

Ref.	Techniques	Highest accuracy appeared
[43]	J48	100%
[46]	NavieBayes(NB)	98.86%
[40]	X-Means	86.17%
[35]	Support Vector Machine (SVM)	97.98%
[14]	Ctree	90.37%
[49]	Decision Tree(DT)	67%
[38]	Random Forest(RF)	96.4%
[34]	Logistic Regression(LA)	96.98%
[45]	Neural Network(NN)	96%
[33]	K-means	98.9%
[36]	Rule-Based	71.3%
[6]	CART	98.3%
[4]	RepTree	61.4%
[6]	Iterative Dichotomiser 3(ID3)	95.9%
[39]	IBK	82.1%
[39]	Simple Logistic	93.27%
[44]	JRip	83.46%
[35]	K-Medoids	84.04%

found that most researchers used two algorithms, Naive Bayes and J48. In addition to its easy implementation and high prediction accuracy, Naive Bayes algorithm deals well with missing data. J48 is another easy to implement algorithm, yet it provides high accuracy results, which explains its frequent use. Besides this, J48 can use both discrete continuous values and has the capability of updating and reasoning. However, it is hard to deal with the absence of data through this technique. However, one should note some of the limitations of these techniques such as their need of large data sets for attaining good accuracy. For further discussion, authors in [50][51] report some advantages and disadvantages of both techniques. In addition, the present paper has identified each tool used in the different research studies and a comparison of those techniques has been tabulated below. Table I shows that 20 out of 23 selected research studies mention the software used; therefore, the results highlight that the most widely used tool was WEKA that attracted 15 out of 20 research, which is approximately 75%. Moreover, using many algorithms in software to identify the prediction accuracy is essential for comparing algorithms and to determine the suitable technique that can be used in the given application.

VI. CONCLUSION AND FUTURE WORK

The EDM would help all the educational stakeholders in several ways. For instance, such tools and techniques could support to improve the students' performance and success in academics, leverage teachers' performance, and support decision-making in institutions. Thus, data mining in higher education would help institutions and educators enhance the educational process effectively. The strength of this review paper lies in using the true prediction accuracy as an indicator to determine the highest effective techniques for each EDM applications of the surveyed studies.

The results of this review paper would be an effective reference for researchers, education providers, educational decision-makers, and others so that they can implement and promote educational data mining more efficaciously. This review paper focuses on further developments in the field of education data mining to support academic advising. Moreover, additional surveys have to be considered for every EDM application by including other standards to specify the best algorithms more accurately.

REFERENCES

[1] J. Jacob, K. Jha, P. Kotak, and S. Puthran, "Educational data mining techniques and their applications," in *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 1344–1348, IEEE, 2015.

[2] A. Nguyen, L. Gardner, and D. Sheridan, "Data analytics in higher education: An integrated view," *Journal of Information Systems Education*, vol. 31, no. 1, p. 61, 2020.

[3] N. Walia, M. Kumar, N. Nayar, and G. Mehta, "Student's academic performance prediction in academic using data mining techniques," *Available at SSRN 3565874*, 2020.

[4] A. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting student performance in higher education institutions using decision tree analysis," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, pp. 26–31, 2018.

[5] B. Al Breiki, N. Zaki, and E. A. Mohamed, "Using educational data mining techniques to predict student performance," in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pp. 1–5, IEEE, 2019.

[6] Y. Saheed, T. Oladele, A. Akanni, and W. Ibrahim, "Student performance prediction based on data mining classification techniques," *Nigerian Journal of Technology*, vol. 37, no. 4, pp. 1087–1091, 2018.

[7] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv preprint arXiv:1201.3417*, 2012.

[8] R. Sumitha, E. Vinothkumar, and P. Scholar, "Prediction of students outcome using data mining techniques," *International Journal of Scientific Engineering and Applied Science (IJSEAS)–Volume-2, Issue-6*, 2016.

[9] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices.," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1 – 21, 2020.

[10] B. Bakhshinategh, O. R. Zaiane, S. ElAtia, and D. Ipperciel, "Educational data mining applications and tasks: A survey of the last 10 years," *Education and Information Technologies*, vol. 23, no. 1, pp. 537–553, 2018.

[11] A. B. Zoric, "Benefits of educational data mining," Sep 2019. Copyright - Copyright Varazdin Development and Entrepreneurship Agency (VADEA) Sep 19/Sep 20, 2019; Last updated - 2020-01-22.

[12] H. Almarabeh, "Analysis of students' performance by using different data mining classifiers," *International Journal of Modern Education and Computer Science*, vol. 9, no. 8, p. 9, 2017.

[13] A. B. Zoric, "Benefits of educational data mining," Sep 2019. Copyright - Copyright Varazdin Development and Entrepreneurship Agency (VADEA) Sep 19/Sep 20, 2019; Last updated - 2020-01-22.

[14] R. Ahuja and Y. Kankane, "Predicting the probability of student's degree completion by using different data mining techniques," in *2017 Fourth International Conference on Image Information Processing (ICIIP)*, pp. 1–4, IEEE, 2017.

[15] E. N. Azizah, U. Pujianto, E. Nugraha, *et al.*, "Comparative performance between c4. 5 and naive bayes classifiers in predicting student academic performance in a virtual learning environment," in *2018 4th International Conference on Education and Technology (ICET)*, pp. 18–22, IEEE, 2018.

[16] S. Roy and A. Garg, "Analyzing performance of students by using data mining techniques a literature survey," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, pp. 130–133, IEEE, 2017.

[17] R. B. Sachin and M. S. Vijay, "A survey and future vision of data mining in educational field," in *2012 Second International Conference on Advanced Computing & Communication Technologies*, pp. 96–100, IEEE, 2012.

[18] R. Baker *et al.*, "Data mining for education," *International encyclopedia of education*, vol. 7, no. 3, pp. 112–118, 2010.

[19] S. M. Thakrar, N. Jadeja, and N. Vadher, "Educational data mining: A review.," *IUP Journal of Information Technology*, vol. 14, no. 1, 2018.

[20] C. Anuradha, T. Velmurugan, and R. Anandavally, "Clustering algorithms in educational data mining: a review," *International Journal of Power Control and Computation*, vol. 7, no. 1, pp. 47–52, 2015.

[21] S. Srivastava, "Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining," *International Journal of Computer Applications*, vol. 88, no. 10, 2014.

[22] M. Chunqiao, "Student performance early warning based on data mining.," *International Journal of Performability Engineering*, vol. 15, no. 3, pp. 822 – 833, 2019.

[23] K. Saravanapriya, "A study on free open source data mining tools," *International Journal of Engineering and Computer Science*, vol. 3, no. 12, pp. 9450–9452, 2014.

[24] F. Ali, D. Bhatt, T. Choudhury, and A. Thakral, "A brief analysis of data mining techniques," in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp. 752–758, IEEE, 2019.

[25] M. Figueiredo, L. Esteves, J. Neves, and H. Vicente, "A data mining approach to study the impact of the methodology followed in chemistry lab classes on the weight attributed by the students to the lab work on learning and motivation," *Chem. Educ. Res. Pract.*, vol. 17, pp. 156–171, 2016.

[26] P. Shruthi and B. Chaitra, "Student performance prediction in education sector using data mining," 2016.

[27] F. A. Ibrahim and O. A. Shiba, "Data mining: Weka software (an overview)," *Journal of Pure and Applied Sciences*, vol. 18, no. 3, 2019.

[28] S. Križanić, "Educational data mining using cluster analysis and decision tree technique: A case study," *International Journal of Engineering Business Management*, vol. 12, p. 1847979020908675, 2020.

[29] H. Alhakami, T. Alsubait, and A. Aljarallah, "Data mining for student advising," *International Journal of Advanced Computer Science and Applications Science (IJSEAS)–Volume-11, Issue-3*, 2020.

[30] W. Singh and P. Kaur, "Comparative analysis of classification techniques for predicting computer engineering students' academic performance," *International Journal of Advanced Research in Computer Science*, vol. 7, no. 6, 2016.

[31] M. Ilic, P. Spalevic, M. Veinovic, and W. S. Alatresh, "Students' success prediction using weka tool," *Infoteh-Jahorina*, vol. 15, pp. 684–688, 2016.

[32] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," *International Journal of Modern Education and Computer Science*, vol. 8, no. 11, p. 36, 2016.

[33] Z. M. Ali, N. H. Hassoon, W. S. Ahmed, and H. N. Abed, "The application of data mining for predicting academic performance using k-means clustering and naïve bayes classification.," *International Journal of Psychosocial Rehabilitation*, vol. 24, no. 3, pp. 2143 – 2151, 2020.

[34] B. Al Breiki, N. Zaki, and E. A. Mohamed, "Using educational data mining techniques to predict student performance," in *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pp. 1–5, IEEE, 2019.

[35] A. Tekin, "Early prediction of students' grade point averages at graduation: A data mining approach," *Eurasian Journal of Educational Research*, vol. 54, pp. 207–226, 2014.

[36] F. Ahmad, N. H. Ismail, and A. A. Aziz, "The prediction of students' academic performance using classification data mining techniques," *Applied Mathematical Sciences*, vol. 9, no. 129, pp. 6415–6426, 2015.

[37] A. Baz, F. Alshareef, E. Alshareef, H. Alhakami, and T. Alsubait, "Predicting students' academic performance using naïve bayes," *International Journal of Computer Science and Network Security*, vol. 20, no. 4, 2020.

[38] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, "Using educational data mining techniques to increase the prediction accuracy of student academic performance," *Information and Learning Sciences*, 2019.

[39] E. Abou Gamie, S. Abou El-Seoud, M. Salama, and W. Hussein, "Multidimensional analysis to predict students' grades in higher education," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 14, no. 02, pp. 4–15, 2019.

[40] H. Uddin and M. T. Nafis, "Students academic performance using partitioning clustering algorithms," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, 2017.

[41] G. Nemomsa, D. P. Sharma, and A. Mulugeta, "Predictive modeling for student performance analytics through data mining techniques," *IUP Journal of Computer Sciences*, vol. 14, no. 1, 2020.

[42] S. Senthil and W. M. Lin, "Applying classification techniques to predict students' academic results," in *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pp. 1–6, IEEE, 2017.

[43] H. H. Alhakami, B. A. Al-Masabi, and T. M. Alsubait, "Data analytics of student learning outcomes using abet course files," in *Science and Information Conference*, pp. 309–325, Springer, 2020.

[44] U. Pujianto, E. N. Azizah, and A. S. Damayanti, "Naive bayes using to predict students' academic performance at faculty of literature," in *2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, pp. 163–169, IEEE, 2017.

[45] D. Hooshyar, M. Pedaste, and Y. Yang, "Mining educational data to predict students' performance through procrastination behavior," *Entropy*, vol. 22, no. 1, p. 12, 2020.

[46] N. Nurhachita and E. S. Negara, "A comparison between naïve bayes and the k-means clustering algorithm for the application of data mining on the admission of new students," *Jurnal Intelektualita: Keislaman, Sosial dan Sains*, vol. 9, no. 1, pp. 51–62, 2020.

[47] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020.

[48] F. Alshareef, H. Alhakami, T. Alsubait, and A. Baz, "Educational data mining applications and techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, 2020.

[49] F. Agrusti, G. Bonavolontà, and M. Mezzini, "University dropout prediction through educational data mining techniques: A systematic review," *Journal of e-Learning and Knowledge Society*, vol. 15, no. 3, pp. 161–182, 2019.

[50] S. Roy and A. Garg, "Analyzing performance of students by using data mining techniques a literature survey," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, pp. 130–133, 2017.

[51] J. Charoenpong, B. Pimpunchat, S. Amornsamankul, W. Triampo, and N. Nuttavut, "A comparison of machine learning algorithms and their applications.," *International Journal of Simulation–Systems, Science & Technology*, vol. 20, no. 4, 2019.