A Prediction Model of Newly Admitted Students in the Level Exam Using Data Mining

توقع لأداء الطلاب المقبولين الجدد في امتحان المستوى باستخدام التنقيب في البيانات

Yousef Saleh Abuzir
Professor/ Al Quds Open University/ Palestine
yabuzir@qou.edu

يوسف صالح أبو زر
أستاذ دكتور / جامعة القدس المفتوحة/ فلسطين

Isam Younis Amro
Associate Professor/ Al-Quds Open University/ Palestine
iamro@qou.edu

إسلام يونس عمرو
أستاذ مشارك / جامعة القدس المفتوحة/ فلسطين

Bassam Tork
Assistant Professor/ Al-Quds Open University/ Palestine
btork@qou.edu

بسام ترك
أستاذ مساعد / جامعة القدس المفتوحة/ فلسطين

Maher Issa
Lecturer / Al-Quds Open University/ Palestine
missa@qou.edu

ماهر عيسى
مدرس / جامعة القدس المفتوحة/ فلسطين

Received: 29/09/2020, **Accepted**: 22/11/2020
DOI: 10.33977/2106-000-004-003
http://journals.qou.edu/index.php/PJTAS

تاريخ الاستلام: 2020/09/29، تاريخ القبول: 2020/11/22
E-ISSN: 2521-411X
P-ISSN: 2520-7431

ABSTRACT

In this research, we will use the Data Mining technique as a prediction model to predict the student’s grade in the level exam. At the same time, we are interested in finding the main factors that affect the grade. In order to analyze and predict what will happen during the various stages of the enrollment process at the University, data mining models will be used. This will help the university determine the interventions and measures needed and take the required action accordingly at the right time. To perform the analytics and the predictions, we used Waikato’s Knowledge Analysis Environment (WEKA) tool and different algorithms such as K-Means, logistic regression, Kohonen’s Self Organizing Map (KSOM), as well as EM to identify the most influential factors that predict student’s grade in the level exam. The results of this research showed that EM offers great value to determine the main parameters that affect the student’s final grade in the level exam. The other three algorithms, logistic regression, K-Means, and KSOM are advanced predictive models for the student’s grade in the level exam.

Keywords: Data Mining, logistic regression, neural networks, level exam grade, K-means, EM Algorithm, Kohonen’s Self-Organizing Map (KSOM) and Clustering.

الملخص

هذا البحث، سنستخدم تقنية التنقيب في البيانات كنموذج تنبؤ لتوقع علامة الطالب في اختبار المستوى. في الوقت نفسه، نحن مهتمون بإيجاد العوامل الرئيسة التي تؤثر على هذه العلامة أو النتيجة. من أجل التحليل والتنبؤ بما سيحدث خلال المراحل المختلفة لعملية التسجيل في الجامعة، يمكن استخدام نماذج التنقيب عن البيانات التي ستساعد الجامعة في تحديد التدخلات، والتدابير، واتخاذ الإجراءات اللازمة، وفقًا لذلك في الوقت المناسب. لإجراء التحليلات والتنبؤات، استخدمنا أداة Waikato’s Knowledge Analysis Environment (WEKA) وخوارزميات مثل(K-Means)، والانحدار (اللوجستي)، وخريطة (Kohonen) ذاتية التنظيم (KSOM) و(EM) لتحديد العوامل الأكثر تأثيرًا على تنبؤ علامة للطالب في اختبار المستوى. أظهرت نتائج هذا البحث أن (EM) تظهر أداء جيد لتحديد العوامل الرئيسة التي تؤثر على العلامة النهائية للطالب في اختبار المستوى. تعد خوارزميات الثلاثة الأخرى المستخدمة الانحدار (اللوجستي، K-Means، KSOM) نموذجًا تنبئيا لعلامة الطالب في امتحان المستوى.

الكلمات المفتاحية: تنقيب في البيانات، الانحدار اللوجستي، الشبكات العصبية، علامة امتحان المستوى،K-means ، EM خوارزمية، خريطة كوهن ذاتية التنظيم (KSOM) والتكتل.

INTRODUCTION

The amount of structured and unstructured educational data is increasing rapidly in the last decade. The data hides valuable information about students like major, high school background, admission type, grade point average (GPA), attendance, assignments, quizzes, lab work, tests, final exams, and extracurricular activities. Furthermore, social interaction networks, psychometric factors, and students’ demographics like address, age, gender, and family background are very interesting and are very interesting and manageably assessed.

Educational data mining is an emerging discipline that is used to extract information from educational data, which is important for various stakeholders like students, academic advisors, teachers, administration, and educational systems.

Nowadays, many data and information related to academic and education field are available on academic portals, online repositories, and research centers. Researchers can use this information and apply different data analyses to obtain important information to support different educational system stakeholders. They can use data mining techniques in many areas in education and university admission.

This study focuses on identifying the main parameters that affect the prediction of the student’s grade in the level exam using data mining. Different algorithms are used to assist in predicting and deciding the main parameters that affect the analysis and prediction.

In this study, we obtained the data from one of the educational branches of al-Quds Open University and the Ministry of Education in Palestine. Different algorithms have been used to analyze the data like K-Means, logistic regression, Kohonen’s Self Organizing Map (KSOM), and EM on the dataset. According to the analysis of the data and the results, the most accurate results are achieved by EM to predict the key parameters that affect the prediction of the student’s grade in the level exam. While logistic regression, K-Means, and KSOM could be used as effective tools in predicting concrete compressive strength.

The remainder of this paper is organized as follows. In Section 2, we review existing literature of the researches in data mining in education. An overview of data mining techniques and WEKA is presented in section 3. Section 4; describes our approach, framework model, and methods used in this study, followed by a discussion regarding the findings from this research in Section 5, and a summary and conclusions in Section 6.

LITERATURE REVIEW

The main idea of data mining predictive models is to use historical data to predict the new and future values of an outcome based on one or more input parameters. Data mining has been used to assist decision making in different fields, including healthcare (Abuzir Y. et al. 2020), civil engineering (Abuzir Y and Abuzir S. 2020), the stock market (Abuzir Y. et al., 2019), Agriculture (Abuzir, 2017), manufacturing, service, and academia. Several studies have discussed the development of data mining techniques for decision-making in higher education.

Ahmed and Elaraby (2014) applied educational data mining to predict the student's performance using the decision tree (ID3) classification method. The data set consisted of 1548 records obtained from an educational institution in the years 2005-2010 from the Information System Department. The study could identify students who needed special attention to reduce the failing ratio and take the required action accordingly at the right time.

Mohammad M. Abu Tair and Alaa M. El-Halees (2012) used educational data mining to discover knowledge. The collected data covered fifteen years from the College of Science and Technology in the Islamic University of Gaza in Khan Younis. The authors used two classification methods to predict the grade of graduate students. In addition, they discovered association, clustering, and outlier detection rules where they described the extracted knowledge for each of them. The study could predict low grades on time and, consequently, helps college management predict those students from the beginning and enhance their performance before graduation.

Brijesh Kumar Baradwaj and Saurbh Pal (2011) conducted a study on student performance based on information like attendance, seminar, assignment, and class test. The study was performed on a data set of 50 students from VBS Purvanchal University, Jaunpaur (Uttah Pradesh) Computer Applications Department of Master of Computer Applications course (MCA) in the years 2007-2010. The study could help identify the students who needed special attention to reduce the failing ratio and take the needed action for the next semester's examination.

Sonali Agarwal, G.N. Pandey, and M. D. Tiwari (2012) conducted a comparative analysis on community college student database among various classification approaches. Support Vector Machine was established as the best classifier with minimum root mean square error (RMSE) and maximum accuracy. The Radial Basis Kernel was identified as the best choice for the Support Vector Machine. The study showed the importance of data availability and the use of different parameters to evaluate students' admission academic performance and, finally, the placement test.

Alaa El-Halees (2009) used educational data mining to analyze students' behavior in a database course. The author preprocessed the data then applied various data mining approaches to discover classification, association, clustering, and outlier detection rules. He extracted knowledge from each of them that describes students' behavior. The study showed how useful data mining could be to improve students' performance.

Wati et al. (2017) , compares the efficiency of data mining techniques using Naïve Bayes Classifier and Tree C4.5 algorithms. Algorithms are used to predict student-learning outcomes. The result shows lower average accuracy for Naïve Bayes Classifier and Tree C4.5

Shahiri et al. (2015), Algarni, (2016), Dahiya V. (2018), Purani et al. (2019) and AlHakami (2020) presented a survey on various components of educational data mining along with its objectives. The main objectives of these researches are to present different data mining techniques used in education. They also strive to compare and evaluate the performance of the different data mining techniques in predicting, advising, dropout, and analysis of students' learning environment. They also want to provide appropriate recommendations and meet the main goals of data mining for education.

Data mining in the educational systems is one of the most recent research topics. In the field of education, many research applied different approaches to data mining and Artificial Neural Network technologies. Data mining techniques are applied for education in multiple case studies (Villanueva, 2018). Table 1 summarizes different data mining techniques widely applied in different domains linked to education and its objectives.

Table 1 Classification of Data Mining Researches in the Domains of Education

Domain	Description	DM Techniques	References
Dropping out or Retention Analysis	Analysis of factors related to dropout and student retention.	Decision Trees, Classification, Neural Networks.	(Bayer et al., 2012), (Thomas, 2015), (Yukselturk &Education, 2014).
VLO or VLE Analysis	Analysis of VLO virtual learning objects or Virtual Learning Environment (VLE).	Correlation Analysis, Regression Trees, Classification, Clustering, Sequential Patterns, Bayesian Networks, Neural Networks, Association rules, Linear regression.	(Ali Yahya et al., 2013), (He, 2013), (Rabbany et al., 2014), (Dutt et al., 2015).
Performance and students evaluation Analysis.	Analysis of the performance of students or their assessment during face-to-face or virtual courses.	Decision Trees, Regression Trees, Classification, Clustering, Sequential Patterns, Bayesian Networks, Neural Networks, Association rules.	(Badr et al., 2014), (Hu et al., 2014), (Shahiri & Husain, 2015).
Generation of Educational Recommendations.	Generate recommendations for the educational process.	Decision Trees, Markov Chains Clustering, Sequential Patterns, Association rules.	(Hung et al., 2012), (Chalaris et al., 2014).
Learning pattern Identification	Analysis of the ways in which virtual students develop in the learning environment and try to establish the way in which they learn.	Decision Trees, Classification, Clustering, Sequential Patterns, Bayesian Networks, Neural Networks, Association rules, Linear regression.	(Chalaris et al., 2014), (Belsis et al., 2014), (Mayilvaganan & Kalpanadevi, 2015).
Students patterns Identification	Data analysis of educational environments, which identified patterns among students.	Correlation Analysis, Decision Trees, Classification, Clustering, Differential Sequence Mining, Sequential Patterns, Bayesian Networks, Association rules.	(Mugla, 2014), (Campagni et al., 2015).
Students related Prediction:	Predictions relating to students, predictions in the final grades, performance, behavior in specific courses, etc.	Decision Trees, Classification, Clustering, Sequential Patterns, Bayesian Networks, Neural Networks, Association rules, Linear regression.	(Barracosa & Antunes, 2011), (López et al., 2012), (Oladokun et al., 2008), (Şen et al., 2012), (Kaur et al., 2015), (Trivedi et al., 2016).

The study of Mengash (2020) focuses on ways to support universities in admissions decision-making using data mining techniques to predict applicants' academic performance at the university. The results demonstrate that applicants' early university performance can be predicted before admission based on certain pre-admission criteria (high school grade average, Scholastic Achievement Admission Test score, and General Aptitude Test score). The study used the Artificial Neural Network technique with an accuracy rate above 79%.

The work of Sani and Babandi (2020) aims to analyze and evaluate student performance in the Department of Computer Science, Jigawa State Polytechnic. A decision tree model is applied during the experiment. The results indicate that it is possible to predict graduation performance; in addition, a procedure for evaluating the performance for each course has been identified.

The work of Mayreen A. & Alexander H. (2019) presents the outcomes of linking an educational data mining approach to the model of students' academic performance. Three data mining classification models (Naïve Bayes, Decision Tree, and Deep Learning in Neural Network) were defined to analyze the data set and predict students' performance. Results show that the Deep Learning classifier beats the other two classifiers by gaining an overall forecast accuracy of 95%. Their analysis and information about prediction help college administration and faculty members improve education and make changes if necessary.

The paper of Patel (2020) presents a literature research on data mining methods used to predict student's performance from 2002 to 2020. This paper reviews work done by different researchers to predict student's performance from all perspectives. The paper also discusses commonly used attributes in different research for the student performance analysis.

The study of Alyahyan and Düştegör (2020), aims to provide a step-by-step set of guidelines for educators willing to apply data mining techniques to predict student success. This study provided educators with an easier access to data mining techniques, enabling all the potential of their application to the field of education.

AN OVERVIEW OF DATA MINING AND WEKA

Data mining is a process or a technique of applying different algorithms on a huge dataset for extracting beneficial information or knowledge. Its intelligent tools are required to apply data mining techniques to manipulate datasets.

Data mining is often used as a combination of intelligent and non-traditional sciences like business analytics, mathematics, logic, statistics, artificial intelligence, machine learning, and artificial neural networks Dweib and Abuzir (2018) and Abuzir and Baraka (2019). In data mining, researchers can use different algorithms in the analysis of the data. These algorithms can be used for Classification, Clustering, Prediction, Decision Trees, Association, and Sequential Patterns (Brown, 2012).

There are different steps in data mining as follows (as shown in Figure 1):

- Data Selection,
- Preprocessing - Data Cleaning,
- Data Transformation,
- Data Mining - Pattern Evaluation and Knowledge Presentation
- Evaluation - Decisions/Use of Discovered Knowledge

There are different useful tools of Educational Data Mining:

- WEKA (Waikato Environment for Knowledge Analysis)
- KEEL (Knowledge Extraction Based on Evolutionary Learning)
- RapidMiner
- R language
- KNIME (Konstanz Information Miner)
- ORANGE

WEKA is an abbreviation for Waikato's Knowledge Analysis Environment. It is an open-source tool developed at the University of Waikato in New Zealand. WEKA is a Java-based tool that involves many open-source data mining and machine learning algorithms. WEKA has the following features (Abuzir Y and Abuzir S., 2020):

- Data processing tools.
- Classification, clustering algorithms and relationship mining (association rules, correlation and sequential patterns).
- User graphical interface.
- WEKA data mining and Machine learning tools

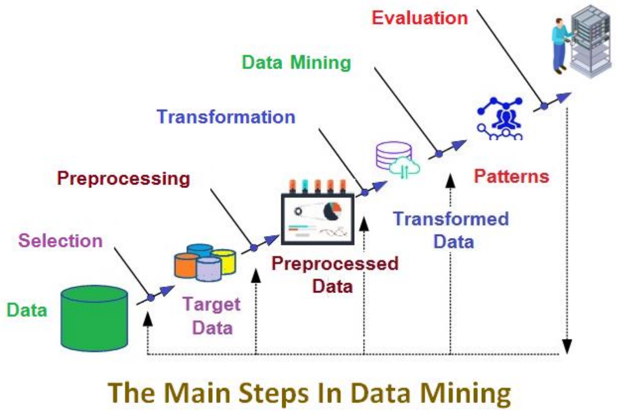


Figure 1 The Main Steps in Data Mining

MATERIALS AND METHODS

A brief overview of different data mining algorithms is needed to predict the output of students. This section will discuss the basis of data mining algorithms with their performance on our case study and the effect of the students' different attributes on the prediction model. Figure 2 illustrates the process of prediction model flow.

Huge volumes of data are now stored in educational information systems, and it comes from various sources, different formats, and different granularity levels. The problems of educational data mining must be explored and analyzed carefully. This will help us achieve the basic goals of data mining in education.

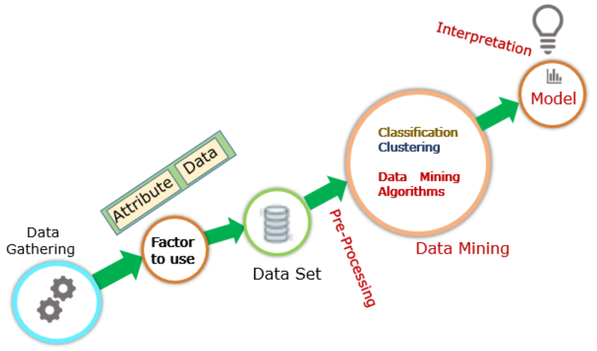


Figure 2 Process for Predication Model

This paper investigates different data mining techniques and algorithms that can be adequately utilized to note the issues of predictions of students' performance in the level exam and predict the main factors that affect the level exam results. WEKA workbench as a data mining tool has many tools, algorithms, and graphics techniques that can be used in our study to predict the performance of newly admitted students in the level exam. Most of the algorithms are built in this tool. The features of the datasets and algorithms employed in this analysis are addressed in the following paragraphs and subsections. It addresses in depth the methodological approach used to establish the prediction model of the key factors affecting the prediction of the level exam performance for the newly admitted students.

Datasets

This research aims to use data mining techniques to predict the student's performance in the level exam. It helps the admission officers at the University decide how to allocate different university recourses for the newly admitted students. The idea behind this use is to help enhance the quality of managing student admission.

As a first step, we collect the data. The data used for this purpose were collected from the Admissions Office in a local branch of al-Quds Open University. As input to the prediction model, 17 variables are selected. We collected two types of data: Personal Attributes and Academic Attributes. The selection of attributes is based on their capacity to provide appropriate predictability. The collected data relates to the following personal attributes of students: Tawjihi seat number, student ID number, intended majors, academic plans (joined academic term), gender, job position, marital status, and residency. In addition to that, data regarding the students' grades in the following level exams, Arabic, English, and Computer, were collected. Other data were collected from the Ministry of Higher Education, which contained grades from the Tawjihi General Exam, such as Arabic Language, English Language, and Computer. As an Output variable for our model, we selected students' grades for the level exam. However, data regarding the students' financial issues and financial aid was not collected, due to privacy concerns. After eliminating incomplete data, the sample comprised 5621 student records.

In the second step, we removed some of the parameters and preprocessed the data. Later, we analyzed these data using the WEKA tool to identify any existing patterns and develop predictive analytics models. Using this model, we predict the main parameters that affect the prediction of the grade for level exams and estimate the level exams' grade for the newly admitted students. This model can be adopted and customized by institutions to predict the grade for the different level exams.

We obtained the statistical analysis using WEKA to complete Table 2. WEKA supports users with two ways to split data:

- The first method is training and supplied test set
- the second method is a percentage split

These groups are not included in each other during the training phase. To do the statistical analysis of the datasets, we divided the datasets (5621 records) into two groups: Training set (3710 students records) 66% and testing set (1911 students records) 34%. After splitting the data into training and testing data, the statistical analysis and data mining algorithms are accomplished to present the results.

Table 2 Numeric Datasets for Student Ranges (WEKA)

Name of Parametert	Maximum	Minimum)	Mean	SDV
CUM_AVG	87.62	0	64.282	13.247
Tarabic	90	0	50.579	11.724
Tenlish	100	0	44.212	13.261
Tcomputer	97	0	49.118	10.843
UArabic1	90	0	50.782	5.11.547
UEnglish	100	0	44.08	13.539
UComputer	97	0	49.611	10.94

DATA MINING ALGORITHMS

Data mining techniques are used to create a model according to which hidden data can discover new knowledge. The essential tasks of data mining techniques are predictions based on the automatic discovery of new relationships and attribute dependencies in the data observed.

There are many different machine-learning algorithms used in data mining models. In this study, we test four different Machine Learning (ML) algorithms to find the main parameters that affect the student's grade of the level exam and predict the grade level exam. The following paragraphs show a brief description of these ML algorithms.

Logistic Regression: It is a statistical tool used for estimation and prediction by using the logic function.

EM (Expectation Maximization) is a clustering algorithm used in data mining. It is based on two iterative steps. The first one is the centroid, where each object is assigned to the most likely cluster. In the second step, we recomputed (Least Squares Optimization) for the centroid.

Another algorithm is the Kohenon Self Organizing Map (KSOM), also called vector quantization. KSOM is an artificial neural network used in unsupervised learning and is considered one of the most known clustering algorithms.

Simple K-Means Clustering is used as an unsupervised learning algorithm and uses Euclidean distance measure to compute distances between instances and clusters.

Given the data <x1, x2,…,xn> and K, assign each xi to one K clusters, C1…Ck, the following algorithm is used to apply K-Means:

```
1-Set $\mu_1$ ... $\mu_k$  randomly
2-Repeat the following steps until convergence:
    2.1-Assign each point xi to the cluster with closest mean $\mu_j$
    2.2- Calculate the new mean for each cluster (equation 1)
```

$$\mu j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \text{ …….(1)}$$

In our study, we used the four following data mining techniques: Logistic Regression, EM, KSOM, and K-means algorithms to analyze data regarding the student's grade in the level exam. This type of analysis will then be used to help in decision making at the University and make a plan to facilitate other general decision-making regarding the allocation of staff load and other facilities and administration issues.

We utilized these different data mining algorithms with different configurations to identify useful data patterns and errors and predict approximate and effective results.

We analyzed and evaluated the results. The results will guide us towards identifying the ideal profile of parameters that minimizes scores of the different parameters like Standard Deviation (SD) and Root Mean Squared Error (RMSE).

RESULTS AND DISCUSSION

In this study research, we used the Data Mining Tool WEKA. It is one of the most professional and extensive packages for machine learning algorithms.

In order to understand the importance of the input variables in our model, it is very important to analyze the impact of input variables on the output variable during the prediction using the ML algorithms. The used algorithms provide very different results based on the relevance of features in a different way. Some features are not applicable to determine the output. The following seven features were removed in the analysis of our dataset: TWA_SEAT, BRANCH_NAME, STUD_ID, JOIN_TERM_NO, JOB_A_NAME, and MARITAL_STATUS.

In this section, we utilized data mining techniques in predicting and finding the main parameters that affect the students' grades in the level exam. We discuss, compare, and evaluate the ML algorithms using our dataset.

We used a linear regression algorithm because it is a simple regression algorithm; it is fast to train and showed a good performance since our output variable (Ucomputer grade in computer level exam) for our data is a linear combination of the inputs.

We used WEKA to evaluate the linear regression on our problem before moving onto more complex algorithms in case it performs well. By using WEKA we find that the following formula can predict the Linear regression for Ucomputer:

0.9 * Tcomputer + 5.48

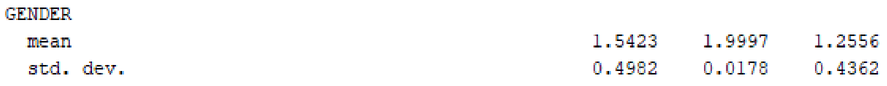
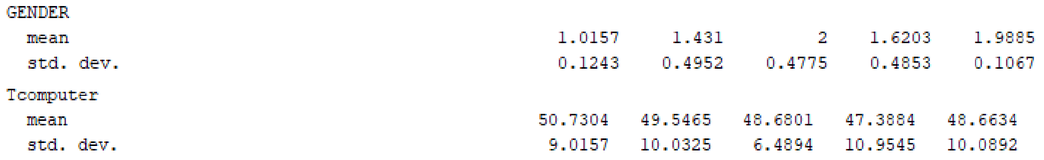
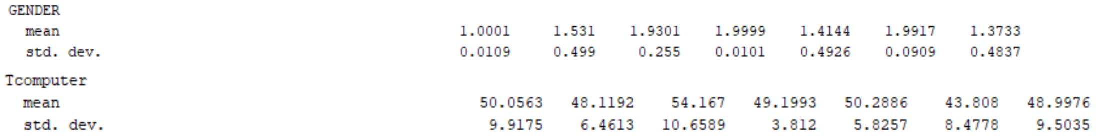
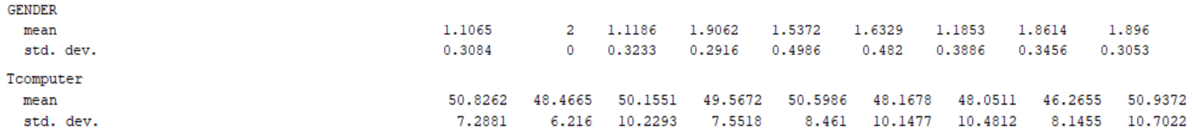
With
Degrees of freedom = 3560
R^2 value = 0.76328
Adjusted R^2 = 0.76322
SE of Coef = 0.00837.

The result obtained by WEKA minimizes the square of the absolute sum of the learned coefficients, which is equal to 0.00837, which gives us a good performance of the prediction for the student's grade in the level exam.

We used WEKA Tool to compute and visualize the results. Table 3 represents the results of EM algorithm. We used a different number of clusters (K=3,5,7, and 9), as shown in Table 3. During the first time, we tested our system with a number of clusters equal to 3; then we ran with 5 clusters, and so on. The performance of EM algorithm was evaluated based on the different number of clusters, as illustrated in Figure 3 and Table 3.

Figure 3 illustrates the relationship between the main parameters that affect the student's grade in the level exam using EM Algorithm. As shown in these figures, the values of the student's grade in the level exam are based on gender and computer grade in the Tawjihi.

Table 3 Results For EM Algorithms Using WEKA

Number of Clusters	Results For EM
EM (with K= 3)	 GENDER mean 1.5423 1.9997 1.2556 std. dev. 0.4982 0.0178 0.4362
EM (with K= 5)	 GENDER mean 1.0157 1.431 2 1.6203 1.9885 std. dev. 0.1243 0.4952 0.4775 0.4853 0.1067 Tcomputer mean 50.7304 49.5465 48.6801 47.3884 48.6634 std. dev. 9.0157 10.0325 6.4894 10.9545 10.0892
EM (with K= 7)	 GENDER mean 1.0001 1.531 1.9301 1.9999 1.4144 1.9917 1.3733 std. dev. 0.0109 0.499 0.255 0.0101 0.4926 0.0909 0.4837 Tcomputer mean 50.0563 48.1192 54.167 49.1993 50.2886 43.808 48.9976 std. dev. 9.9175 6.4613 10.6589 3.812 5.8257 8.4778 9.5035
EM (with K= 9)	 GENDER mean 1.1065 2 1.1186 1.9062 1.5372 1.6329 1.1853 1.8614 1.896 std. dev. 0.3084 0 0.3233 0.2916 0.4986 0.482 0.3886 0.3456 0.3053 Tcomputer mean 50.8262 48.4665 50.1551 49.5672 50.5986 48.1678 48.0511 46.2655 50.9372 std. dev. 7.2881 6.216 10.2293 7.5518 8.461 10.1477 10.4812 8.1455 10.7022

In EM algorithm, the best parameters are selected based on their Standard Deviation Values. Table 4 shows the list of the main factors that affect the student's grade in the level exam with their standard deviations.

The second model uses the KSOM algorithm. This algorithm is employed to illustrate the components that affect the student's grade in the level exam. For the KSOM algorithm, the main components that affect the student's grade in the level exam are gender and Tcomputer grades. Figure 4 shows the results.

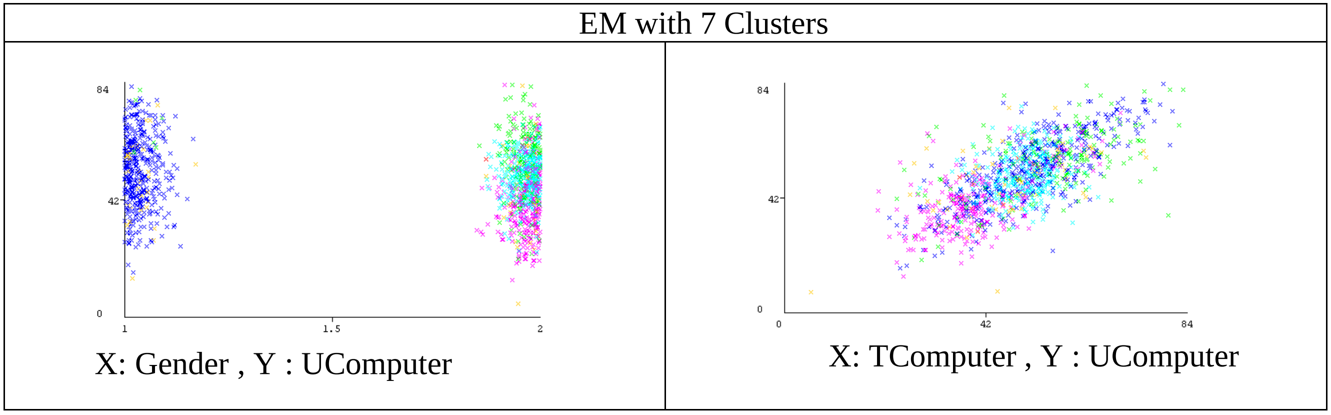
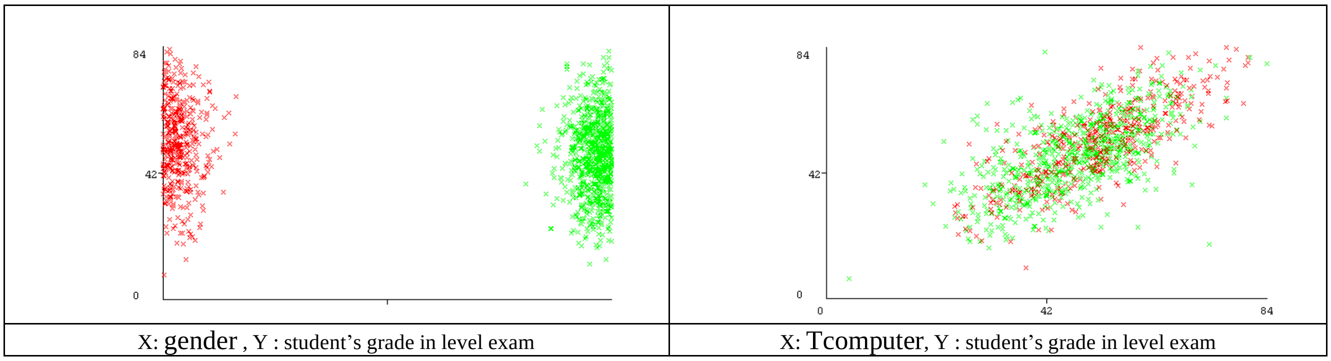

EM with 7 Clusters

Figure 3 Plotting of the main components that affect the student's grade in the level exam using

Figure 5 illustrates a comparison between the EM and KSOM algorithm. As the figure shows, the predicted model for the two components is highly similar. The performance of gender and Tcomputer on the student's grade in the level exam has the same significant effect. The analysis of the two graphs shows that the two algorithms have the same effect between the potentially used two input parameters gender and Tcomputer.

Table 4 List of the Main Component that Affects Student's Grade in the Level Exam (EM)

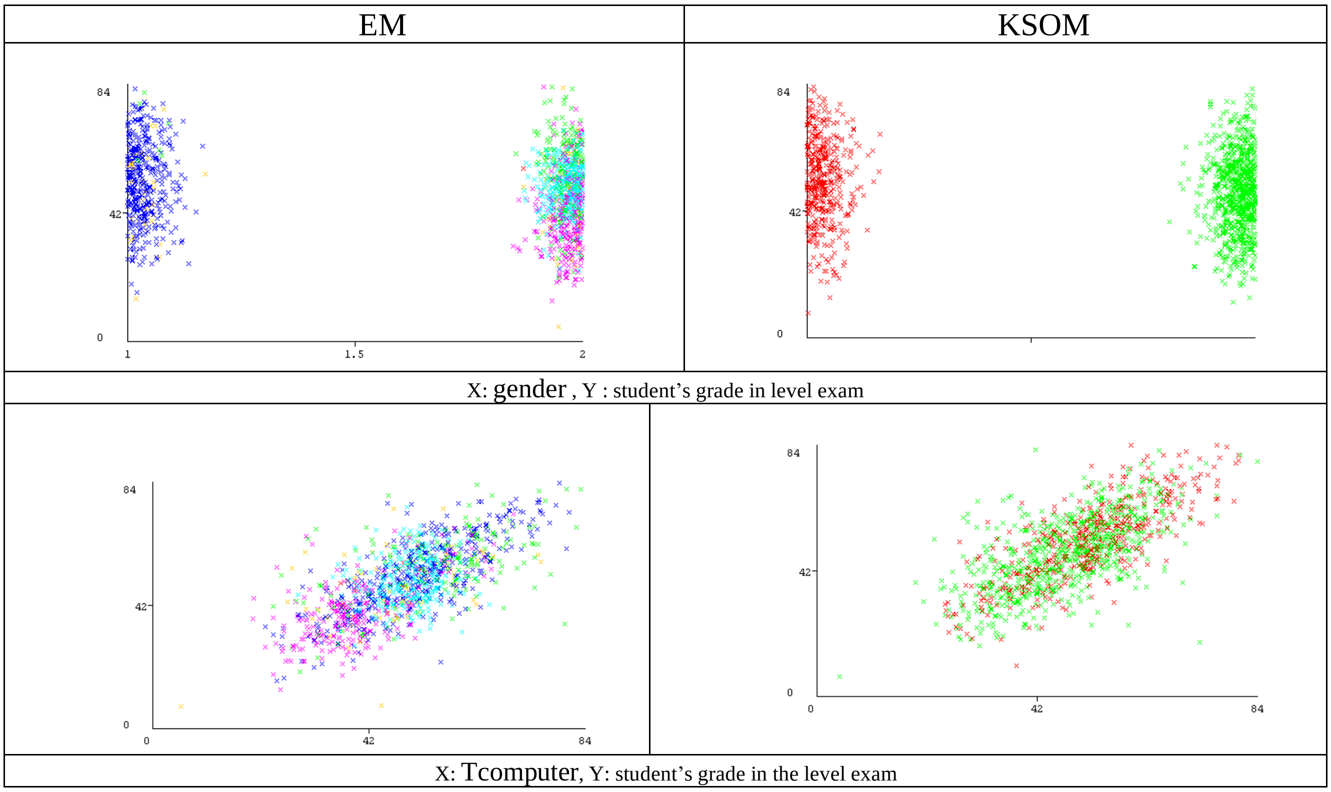
Number of Clusters	Standard. Deviation	St Dev. Without subject parameter	Predict Components
3	0.0178	0.0647	Gender
5	0. 0.1067	0.3542	Gender
	6.4894	5.7734	Tcompputer
7	0.0101	0.0007	Gender
	3.812	6.1907	Tcompputer
9	0.2916	0.0001	Gender
	6.216	5.7946	Tcompputer



X: gender , Y : student's grade in level exam

X: Tcomputer, Y : student's grade in level exam

Figure 4: Gender and Tcomputer grade versus student's grade in the level exam (Ucomputer) (KSOM)



EM

KSOM

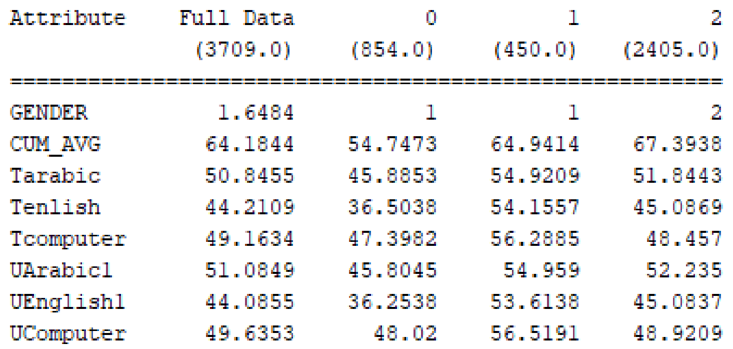
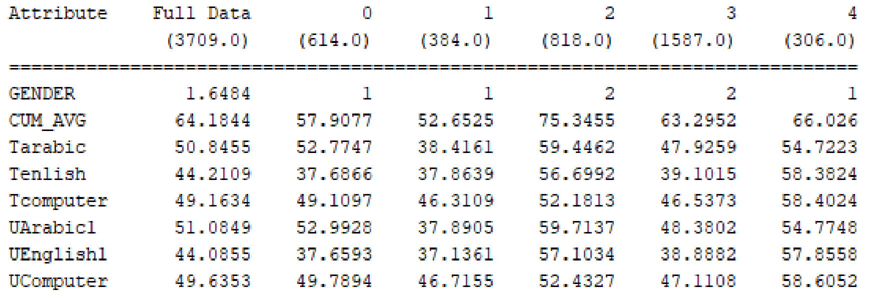
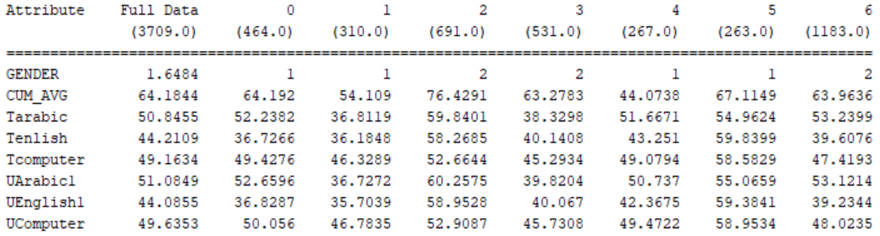
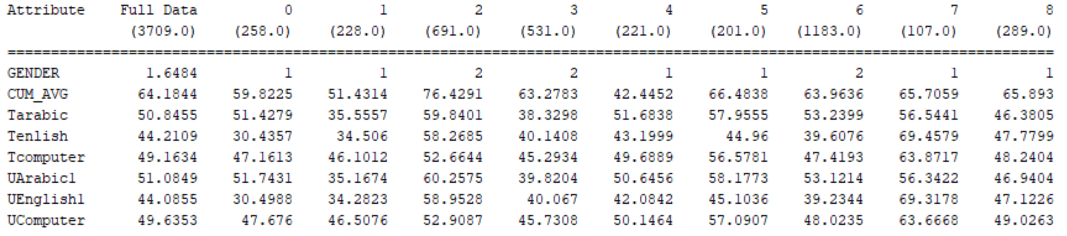
X: gender , Y : student's grade in level exam

X: Tcomputer, Y: student's grade in the level exam

Figure 5: Comparing EM and KSOM Algorithms

The K-means algorithm is applied to the datasets, using different values for k = 3, 5, 7, and 9. Table 5 shows the results of clustering with the different values for K= 3, 5, 7, and 9.

Table 5 Results for K-Means (with K= 3, 5, 7, and 9)

Number of Clusters	Results For K-Means (with K= 3,5, 7 and 9)
K-Means (with K= 3)	
K-Means (with K= 5)	
K-Means (with K= 7)	
K-Means (with K= 9)	

```
Attribute     Full Data         0         1         2
               (3709.0)   (854.0)   (450.0)  (2405.0)
=======================================================
GENDER           1.6484         1         1         2
CUM_AVG         64.1844   54.7473   64.9414   67.3938
Tarabic         50.8455   45.8853   54.9209   51.8443
Tenlish         44.2109   36.5038   54.1557   45.0869
Tcomputer       49.1634   47.3982   56.2885    48.457
UArabicl        51.0849   45.8045    54.959    52.235
UEnglishl       44.0855   36.2538   53.6138   45.0837
UComputer       49.6353     48.02   56.5191   48.9209
```

```
Attribute    Full Data         0         1         2         3         4
              (3709.0)   (614.0)   (384.0)   (818.0)  (1587.0)   (306.0)
==========================================================================
GENDER          1.6484         1         1         2         2         1
CUM_AVG        64.1844   57.9077   52.6525   75.3455   63.2952    66.026
Tarabic        50.8455   52.7747   38.4161   59.4462   47.9259   54.7223
Tenlish        44.2109   37.6866   37.8639   56.6992   39.1015   58.3824
Tcomputer      49.1634   49.1097   46.3109   52.1813   46.5373   58.4024
UArabicl       51.0849   52.9928   37.8905   59.7137   48.3802   54.7748
UEnglishl      44.0855   37.6593   37.1361   57.1034   38.8882   57.8558
UComputer      49.6353   49.7894   46.7155   52.4327   47.1108   58.6052
```

```
Attribute    Full Data         0         1         2         3         4         5         6
              (3709.0)   (464.0)   (310.0)   (691.0)   (531.0)   (267.0)   (263.0)  (1183.0)
============================================================================================
GENDER          1.6484         1         1         2         2         1         1         2
CUM_AVG        64.1844    64.192    54.109   76.4291   63.2783   44.0738   67.1149   63.9636
Tarabic        50.8455   52.2382   36.8119   59.8401   38.3298   51.6671   54.9624   53.2399
Tenlish        44.2109   36.7266   36.1848   58.2685   40.1408    43.251   59.8399   39.6076
Tcomputer      49.1634   49.4276   46.3289   52.6644   45.2934   49.0794   58.5829   47.4193
UArabicl       51.0849   52.6596   36.7272   60.2575   39.8204    50.737   55.0659   53.1214
UEnglishl      44.0855   36.8287   35.7039   58.9528    40.067   42.3675   59.3841   39.2344
UComputer      49.6353    50.056   46.7835   52.9087   45.7308   49.4722   58.9534   48.0235
```

```
Attribute    Full Data         0         1         2         3         4         5         6         7         8
              (3709.0)   (258.0)   (228.0)   (691.0)   (531.0)   (221.0)   (201.0)  (1183.0)   (107.0)   (289.0)
================================================================================================================
GENDER          1.6484         1         1         2         2         1         1         2         1         1
CUM_AVG        64.1844   59.8225   51.4314   76.4291   63.2783   42.4452   66.4838   63.9636   65.7059    65.893
Tarabic        50.8455   51.4279   35.5557   59.8401   38.3298   51.6838   57.9555   53.2399   56.5441   46.3805
Tenlish        44.2109   30.4357    34.506   58.2685   40.1408   43.1999     44.96   39.6076   69.4579   47.7799
Tcomputer      49.1634   47.1613   46.1012   52.6644   45.2934   49.6889   56.5781   47.4193   63.8717   48.2404
UArabicl       51.0849   51.7431   35.1674   60.2575   39.8204   50.6456   58.1773   53.1214   56.3422   46.9404
UEnglishl      44.0855   30.4988   34.2823   58.9528    40.067   42.0842   45.1036   39.2344   69.3178   47.1226
UComputer      49.6353    47.676   46.5076   52.9087   45.7308   50.1464   57.0907   48.0235   63.6668   49.0263
```

Based on the analysis of the result of K-Means, we find that the most factors that affect the student’s grade in the level exam are the result of the computer course in Tawjihi (Tcomputer) (Table 5). Referring to the achieved results, Table 5 presents a summary of the key attributes that affect the student’s grade in the level exam using the three different algorithms.

Referring to the results shown in Table 6, the K-Means algorithm shows that CUM_AVG, Tenglish, and Tcompute are the most common parameters that affect the student’s grade in the level exam. When we use both EM and KSOM algorithms, two parameters are considered, which are: Gender and Tcomputer (Table 7). At the same time, the K-Means algorithm includes a distinguished component, the CUM_AVG. It is clear that all three algorithms show intersection and provide different information. In general, the analysis concludes that Tcomputer is a common parameter that affects the student’s grade in the level exam.

By using WEKA, we find that the actual average of the student’s grade in the level exam is equal to 49.6353. The results show that the K-Means can be successfully used to give a more accurate prediction for increasing the student’s grade in the level exam compared to the average of the actual data, which is 49.6353 (Table 8).

Table 6 K-Means - The Most Factors that Affect the Computer Level Result

Number of Cluster	Parmeters
3	Tcomputer
5	CUM_AVG, Tenglish and Tcomputer
7	CUM_AVG, Tenglish and Tcomputer
9	Tenglish and Tcomputer

Table 7 Summary of the Main Components that Affect the Computer Level Result Using the 3 Algorithms

K-Means	EM	KSOM
CUM_AVG, Tenglish and Tcomputer	Gender and Tcomputer	Gender and Tcomputer

Table 8 A Relation Between No. of Cluster, Sum of Squared Errors and the Student's Grade in the Computer Level Exam Using K-Means

No. of Clusters	Sum Of Squared Errors (SSE)	Number Of Iterations	student's grade in level exam for computer is (49.6353)
3	319.8002246319065	16	56.5191
4	353.5744326617288	6	50.953
5	240.38699317691064	17	58.6052
6	253.00159228731195	16	56.5191
7	215.58737786916973	41	58.9534
8	208.46352694465233	41	62.5657
9	203.7570220482533	41	63.6668
12	174.06781954211337	96	63.4604
15	156.2830245315655	54	55.0647
20	136.53349982348624	67	62.1059
25	124.69899067623594	36	63.7429
30	83.77	17	63.3709
50	62.3	16	67.23

In summary, this study reviewed previous researches on educational data mining and predicting students' performance. These different studies delve into analyzing the data mining techniques and parameters that will help predict student performance in the level exams.

We tested three different algorithms for comparative evaluation of finding the main parameters that affect the student's grade in the level exam using the WEKA tool. Based on the result found, it is clear that the results of the performance of EM algorithm are the most accurate and effective for finding various parameters affecting the student's grade in the level exam. The achieved results exhibit Logistic linear regression, K-Means, and KSOM are the most adequate algorithms for predicting and improving the student's grade in the level exam.

Based on the results from the testing and evaluation, the researchers found out that applying the different data mining algorithms to our datasets effectively predicts and improves the student's grade in the level exam increases the performance of the student's grade in the level exam from 49.6353 to 67.23.

Our analysis shows that there is a slight difference between the data mining algorithms. The output of each data mining algorithm is similar, and the performance of each of them is suitable for the prediction and improving the grade of the student in the level exam.

In addition, our prediction model will help the University take the required actions at the right time and improve the students' performance. The records of newly admitted students will be fed to the model. Those who have less chance to pass the level exam will be advised by their supervisors to pay more attention to the preparation of the exam or take supporting courses to improve their chances to pass the exams.

Based on the analysis of the data and interpretation of the results, we can improve the computer level exam's performance by accepting more female students and students with high computer results in Tawjihi.

Interpreting the result of the data mining techniques tells us that we can increase the average student's grade in the level exam from 49.6353 to 67.23 by working on the two factors that affect Computer-level exams. These factors are Computer results in Tawjihi and Gender.

The results can be used and utilized by the Student Admission, Academic Affairs, and Enrolment Office at the University for planning purposes:

1-Admit policy for the students:

- Accepting more females than male
- Accepting students with high results in Computer Exam.

2- Based on the results, the University can exclude some students from the level exams. This will make a profit from the use of resources, labs, technicians, staff, and financial and human resources.

3- Effective management of staff load.

CONCLUSIONS

This research aims to find out the most useful attributes, which are used for predicting the performance of students in the level exam, and determine which data mining techniques and parameters are best to improve the accuracy of the prediction mechanism in the educational information system.

This paper discussed developing a prediction model for finding the main parameters that affect the students' grades in the level exam.

We used seven input parameters to predict the student's grade in the level exam using different data mining algorithms (Logistic Regression, EM, KSOM, and K-Means). The actual input parameters consist of 7 parameters and one output student's grade in level exam (Ucomputer) with 5621 student records.

Results showed that using data mining techniques effectively predicts the main parameters that affect students' grades in the level exam. The analysis shows that Logistic linear regression, K-means, and KSOM algorithms are the most accurate algorithms to predict the student's grade in the level exam. At the same time, EM is useful for predicting the main parameters that affect the student's grade in the level exam.

In general, data mining techniques are very effective tools in predicting the student's grade in the level exam as well as the main factors that affect and improve the performance of student's grade in the level exam. The results can be used by the student admission, academic affairs, and enrolment office for planning purposes. Our study may be expanded to include an additional range of parameters to improve the prediction of the students' grades in the level exam.

REFERENCES

- *Abuzir Y., Abuzir M. and Abuzir A., Using Artificial Neural Networks (ANN) To Detect the Diabetes, accepted in COMMUNICATION & COGNITION (C&C) Journal, V53, N2-1 (2020). Ghent, Belgium.*
- *Abuzir Y., Abuzir S., Data Mining Techniques for Prediction of Concrete Compressive Strength (CCS), Palestinian Journal of Technology and Applied Sciences (PJTAS), No 3 (2020).*
- *Abuzir Y. and Baraka A.M. , Financial Stock Market Forecast Using Data Mining in Palestine, Palestinian Journal of Technology and Applied Sciences, No 2 (2019).*
- *Abuzir Y., Predict the Main Factors that Affect the Vegetable Production in Palestine Using WEKA Data Mining Tool, Palestinian Journal of Technology and Applied Sciences, pp 58-71, No 1 (2018).*
- *Abeer B. A., Ibrahim S. E., Data Maning: A Prediction for Student's Performance Using Classification Method, 2014.*
- *Mohammed M. Abu Tair, A. M. E, Mining Educational Data to Improve Students' Performance: A Case Study, International Journal of Information 2 (2), 2012.*
- *Brijesh K. B., Saurabh P., Mining Educational Data to Analyze Students' Performance, 2011.*
- *Sonali A., Pandey G. N., and Tiwari M. D., Data Mining in Education: Data Classification and Decision Tree Approach, 2012.*
- *El-Halees A., Mining Students Data to Analyze Learning Behavior: A Case Study, 2009.*
- *Wati, M., Indrawan, W., Widians, J.A., & Puspitasari, N. (2017). Data mining for predicting students' learning result. 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), 1-4.*
- *Shahiri A. M. , Husain W. , and Rashid N. A. , "A review on predicting student's performance using data mining techniques," Procedia Computer Science, Vol. 72, pp. 414–422, 2015.*
- *Algarni A., "Data Mining in Education" International Journal of Advanced Computer Science and Applications, Vol. 7, Issue. 6, pp. 456-461, 2016.*
- *Dahiya V., "A Survey on Educational Data Mining", International Journal of Research in Humanities, Arts and Literature, Vol. 6, Issue.5, pp. 23-30, 2018.*
- *Purani K.D. and Chaudhary M.B.(2019), Educational Data Mining: A Survey of Analyzing Student Academic Performance Methods, International Journal of Computer Sciences and Engineering, Survey Paper Vol.-7, Issue-2, Feb 2019.*
- *AlHakami, H., Alsubait, T., & Al-Jarallah, A.S. (2020). Data Mining for Student Advising. International Journal of Advanced Computer Science and Applications, 11.*
- *Dweib M., Abuzir Y. (2018), Optimization of the Neural Networks Parameters, Palestinian Journal of Technology and Applied Sciences, pp 34 -47, No 1 (2018).*
- *Brown M., Data mining techniques, IBM DeveloperWork, December 2012, online https://www.ibm.com/developerworks/library/ba-data-mining-techniques/*
- *Villanueva A., Moreno L.G. & Salinas M.J.(2018), Data mining techniques applied in educational environments: Literature review, Digital Education Review - Number 33, June 2018.*
- *Bayer, J., Bydzovská, H., & Géryk, G. (2012). Predicting dropout from social behaviour of students. International Educational Data Mining Societ, (Dm), 103–109.*
- *Thomas, J. (2015). Predicting College Students Dropout using EDM Techniques, 123(5), 26–34.*
- *Yukselturk, E., & Education, C. (2014). Predicting Dropout Student□: An Application of Data Mining Methods In An Online Education Program, 17(1).*
- *https://doi.org/10.2478/eurodl-2014-0008*
- *Yahya A., A., Osman, A., & Abdu Alattab, A. (2013). Educational Data Mining□: A Case Study of Teacher's Classroom Questions. IEEE 13th International Conference On, (February), 92–97.*
- *He, W. (2013). Examining Students ' Online Interaction in a Live Video Streaming Environment Using Data Mining and Text Mining Computers in Human Behavior. Computers in Human Behavior, (February), 90–102. https://doi.org/10.1016/j.chb.2012.07.020*
- *Rabbany, R., Elatia, S., Takaffoli, M., & Zaïane, O. R. (2014). Collaborative Learning of Students in Online Discussion Forums: A Social Network Analysis Perspective. EDM 2014, 1–25.*
- *Dutt, A., Aghabozrgi, S., Akmal, M., Ismail, B., & Mahroeian, H. (2015). Clustering Algorithms Applied in Educational Data Mining, 5(2), 112–116. https://doi.org/10.7763/IJIEE.2015.V5.513*
- *Badr, A., Din, E., & Elaraby, I. S. (2014). Data Mining□: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and*

Technology, 2(2), 43– 47. https://doi.org/10.13189/wjcat.2014.020203

- *Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. Computers in Human Behavior, 469–478.*
- *Shahiri, A. M., & Husain, W. (2015). A Review on Predicting Student's Performance using Data Mining Techniques. Procedia Computer Science, 72, 414–422. https://doi.org/10.1016/j.procs.2015.12.157*
- *Hung, J., Hsu, Y., & Rice, K. (2012). Integrating Data Mining in Program Evaluation of K-12 Online Education, 15, 27–41.*
- *Chalaris, M., Gritzalis, S., Maragoudakis, M., & Sgouropoulou, C. (2014). Improving Quality of Educational Processes Providing New Knowledge using Data Mining Techniques. Procedia - Social and Behavioral Sciences, 147, 390–397. https://doi.org/10.1016/j.sbspro.2014.07.117*
- *Belsis, P., Chalaris, I., Chalaris, M., & Skourlas, C. (2014). The Analysis of the Length of Studies in Higher Education based on Clustering and the Extraction of Association Rules. Procedia - Social and Behavioral Sciences, 147, 567–575. https://doi.org/10.1016/j.sbspro.2014.07.159*
- *Mayilvaganan, M., & Kalpanadevi, D. (2015). Cognitive Skill Analysis for Students through Problem Solving Based on Data Mining Techniques. Procedia - Procedia Computer Science, 47, 62–75. https://doi.org/10.1016/j.procs.2015.03.184*
- *Mugla, H. G. (2014). Modeling Student Performance in Higher Education Using Data Mining Modeling Student Performance in Higher Education Using Data Mining, (February 2016). https://doi.org/10.1007/978-3-319-02738-8*
- *Campagni, R., Merlini, D., Sprugnoli, R., & Verri, M. C. (2015). Data mining models for student careers. Expert Systems with Applications, 42, 5508–5521.*
- *Barracosa, J., & Antunes, C. (2011). Anticipating Teachers' Performance. KDD 2011 Workshop: Knowledge Discovery in Educational Data, 77–82.*
- *López, M. I., Luna, J. M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. International Educational Data Mining Society, 148–151.*
- *Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores□: A data mining approach. Expert Systems with Applications, 39, 9468–9476. https://doi.org/10.1016/j.eswa.2012.02.112*
- *Kaur, P., Singh, M., & Singh, G. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia - Procedia Computer Science, 57, 500– 508. https://doi.org/10.1016/j.procs.2015.07.372*
- *Trivedi, S., Pardos, Z. A., Sárközy, G. N., & Heffernan, N. T. (2016). Spectral Clustering in Educational Data Mining. EDM 2011, (February), 129–138.*
- *Mengash H. A., "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," in IEEE Access, vol. 8, pp. 55462-55470, 2020, doi: 10.1109/ACCESS.2020.2981905.*
- *Babandi U. et al. (2020), Data Mining: Predicting of Student Performance Using Classification Technique. in International Journal of Information Processing and Communication (IJIPC) Vol. 8 No. 1 pp. 92-101.*
- *Amazona, M. V., & Hernandez, A. A. (2019). Modelling Student Performance Using Data Mining Techniques. Proceedings of the 2019 5th International Conference on Computing and Data Engineering - ICCDE' 19. doi:10.1145/3330530.3330544.*
- *Ganorkar S.S., Tiwari N., Namdeo V. (2020) Analysis and Prediction of Student Data Using Data Science: A Review. In: Zhang YD., Senjyu T., SO–IN C., Joshi A. (eds) Smart Trends in Computing and Communications: Proceedings of SmartCom 2020. Smart Innovation, Systems and Technologies, vol. 182, pp 443-448, Springer, Singapore.*
- *Patel V. K., Pawar M., Goyal S. (2020), Predicting Student's Performance Using Data Mining Techniques: A Survey From 2002 To 2020, International Journal of Scientific & Technology Research-IJSTR,Volume 9 - Issue 6, June 2020 Edition.*
- *Alyahyan E. and Düştegör D. (2020), Predicting academic success in higher education: literature review and best practices, International Journal of Educational Technology in Higher Education (2020) 17:3.*