International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 10 (2020), pp. 2895-2908
 https://dx.doi.org/10.37624/IJERT/13.10.2020.2895-2908

Students Performance: From Detection of Failures and Anomaly Cases to the Solutions-Based Mining Algorithms

Ebtehal Ibrahim Al-Fairouz[1], Mohammed Abdullah Al-Hagery[2]

[1,2]Department of Computer Science, College of Computer, Qassim University, Buraydah, Saudi Arabia.
[1]ORCID: 0000-0002-2667-5925, [2]ORCID: 0000-0001-6939-013X

Abstract

Educational Data Mining (EDM) helps to recognise the performance of students and predict their academic achievements that include the successes aspects and failures, negative aspects and challenges. In the educational systems, a massive amount of students' data has been collected, which has become difficult for officials to search through and obtain the knowledge required to discover challenges facing students and universities by traditional methods. Therefore, the rooted problem is how to dive into these data and discover real challenges that are facing both the students and the universities. The main aim of this research is to extract hidden, significant patterns, new insights from students' historical data, which can solve the current problems, help to enhance the educational process and to improve academic performance. The data mining tools used for this task are classification, regression, and association rules for frequent patterns generation. The research data sets gathered from the College of Business and Economics (CBE). The finding of this research can help to make appropriate decisions for certain circumstances and provide better suggestions for overcoming students' weaknesses and failures. Through the findings, numerous problems related to a students' performance discovered at different levels and in various courses. The research findings indicated that there are many important problems. Consequently, a suggestion of suitable solutions, which can be presented to the relevant authorities for the benefit and improving student performance and activating academic advising.

Keywords: Educational Data Mining, Students Failures, Student Performance, Academic Advising, Association Rules, Anomaly Detection.

1. INTRODUCTION

Educational institutions have information systems designed to provide the information necessary for the management and educational development process. The Educational Information System (EIS) is a means of collecting, analysing, maintaining and distributing information and data, which supports decision making [1]. The data mining processes and tools can extract useful knowledge from these systems, which have accumulated educational data over several years.

Data mining is an essential step in what is referred to as Knowledge Discovery In Databases (KDD) [2]. It can be briefly defined as extracting useful features or unseen patterns from a large data set [3], [4]. The KDD process consists of several steps; the first one involves gathering appropriate data from different sources. The second, data selection, to determine which data is to be used. Third, data pre-processing, which involves filling in missing values, removing outliers and resolving inconsistencies in the data. Fourth, data transformation, by converting data into a format that is appropriate for the mining process. Fifth, data mining algorithms by applying intelligent techniques to extract useful patterns. Finally, the evaluation of results, seen in the patterns that represent knowledge discovered [2], [5].

Predicting student performance is a significant concern for educational institutions [6]. To do this, the field of education has adopted data mining techniques as a way to detect and analyse student performance and predict their learning achievement. The techniques have shown themselves to be capable of preventing failure and focusing on poor performance to guide and help overcome difficulties. Student performance depends on many factors, such as the social, economic and personal; knowledge could be derived from these factors to assess the academic performance of students [7]. Other benefits include better evaluating the institution, helping improve the education process, identifying future requirements, and improving decision making [8]. The importance of this research comes from the promise this field offers, in serving the educational process, the university and the student.

This study aims to discover new patterns and features in the students' academic records. It contributes to predict and improve academic performance using regression and classification techniques on that data for the last five years. Moreover, it identifies the student's weaknesses and failures and explores the knowledge that helps to improve the educational process. Furthermore, it tries to find the reasons for the student's repeated failure in a particular course by use association rules and to activate academic advising for students to overcome or minimise their problems and failures. Also, this research contributes to discovering anomalous values that may provide great benefits in achieving the requirements to raise the level of education quality.

The motivation for doing this research is to help the college of CBE to find useful solutions that help in achieving quality in the educational process. In addition to searching for the reasons that led to the level weakness of some students and their low academic achievement, or searching for outstanding students in its various departments to benefit from their experiences in achieving high academic performance.

2. RELATED WORK

Research in EDM is an interesting domain for academics and researchers, especially in educational institutions. The research in this area generates useful knowledge related to students, instructors, courses and the educational management system, as a whole. Since the knowledge from data collected in educational systems is a veritable gold mine, it is important to make accurate decisions in achieving the requirements for this work, as it helps raise the educational process, in addition to increasing the quality of the educational institution and reducing failure.

Data mining can be used in the area of education for a better understanding of the learning process and acquiring practical knowledge. This, in turn, helps identify problems facing students and reduce failure in academic performance [9]. Data mining in the educational area is called Educational Data Mining (EDM). It has contributed significantly to the measurement of student academic performance and preventing dropouts, and to better understanding failure [7]. The EDM is a research field that assists in discovering ways to enhance the quality of education [10], [11]. It is a computer-based learning method that helps discover new patterns of data sets in educational institutions and represents one particular field of data mining [8].

EDM includes various sets of users or members, including the educational institution's administrators, teaching staff, students, curriculum developers, and planners [10], [12]. Since 1993, many research works have employed EDM, with the number of these studies growing significantly since then [13], [14]. A research works focused on extracting knowledge from student data, predicting performance, evaluating student performance in specific courses or finding an association between courses using various data mining techniques [15].

Some related works have obtained their data from the learning management system (LMS) known as kalboard 360 [16]–[18]. Whereas many studies relied on the analysis of real data from different environments of institutions, such as colleges, universities, or schools using common classification methods, like collected data sets from the College of Computer Applications in India, also, from the National Defense University in Malaysia. Some of the datasets were not enough [19]–[23].

Additionally, some of the previous works utilised limited methods such as the classification and regression methods in their study [24], [25].

One of the common methods that can be employed in this field is the decision tree, using the decision tree model as a classifier or predictor for students' academic data can help to analyse the data and to study student performance and the discovery of their achievements [8], [26].

Besides, applying the Data Mining Tools can constitute a practical guide for decision-makers and teachers in higher education institutions, to identify hidden problems related to student success and failure [27]. Furthermore, the classification techniques are useful to predict a student's career [28].

The use of association rules algorithms can be extensively used in studies related to EDM alongside the other algorithms. The benefit of association rules extraction is to find frequent patterns in databases and to explore the relationship between the various attributes that affect the academic achievement of students [29], [30]. Furthermore, revelation the useful information from behavioral data for students by using association rules. Additionally, by the association rules, we can obtain frequent patterns of behaviors that have a significant impact on student performance and students' Failures cases can be identified. This may help educational institutions understand and improve students' behavior and also make the appropriate decisions, besides, the use of the association rule method that offers insight into improving admissions planning [14], [28], [31].

In this paper, we selected the most significant tools to analysis students' historical dataset from the CBE to identify aspects of student failure, success and predict their academic performance using these technologies, which include classification and regression, Outlier Analysis. Where Outlier Analysis are representing the anomalies cases. Also, the use of technologies that help discover students' achievements and find out the reasons behind some students' failure by using association rules. Also, this paper contributes to the search for anomalies detection that may be distinct cases of the college that help in making the appropriate decisions in the interest of those students.

3. METHODOLOGY

The proposed method uses several various techniques to focus on student performance analysis of the CBE. The overall architecture of the proposed method is shown in Figure 1. In this study, we used the Orange data mining platform as open-source software for data mining and machine learning [32]. The data mining techniques include Linear Regression, Association Rules, Decision Tree, Naive Bayes, and Random Forest. The classification and regression techniques were used to predict students' performance. Whereas, the association rules technique was used for detecting frequents items among students' records; to understand the reasons for their failure.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 10 (2020), pp. 2895-2908
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.10.2020.2895-2908

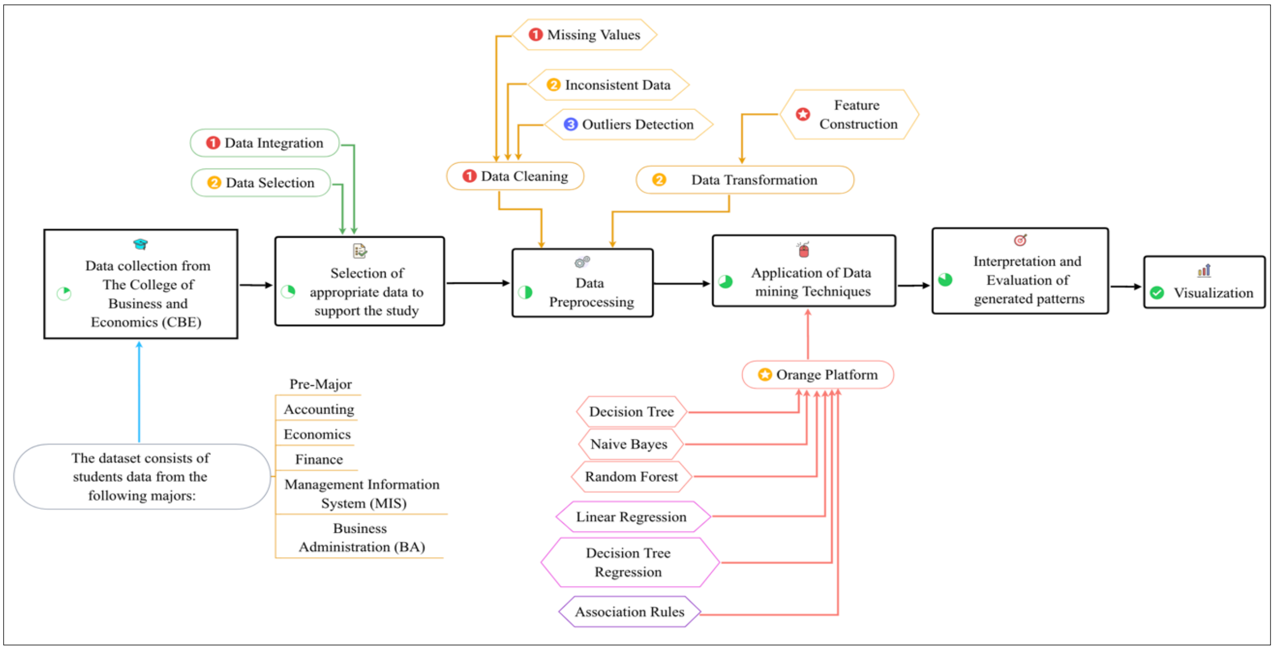

Fig 1. Methodology framework

3.1 Data Collection

This study was conducted on the five-year (2014–2018) data set of undergraduate students enrolled in the CBE. The data set contains male and female students' data from the different departments, namely Management Information Systems (MIS), Finance, Accounting, Economics, Business Administration (BA), and Pre-Major. The total number of records is 72,259 and 14 attributes. The attributes used in this study are described in Table 1.

Table 1: Data set information

#	Attribute Name	Description
1	SEMESTER	This attribute contains the semester such as 382, 391, etc. The meaning of 382: (38) is the year 1438 in the Hijri and (2) is the second semester of this year.
2	COURSE_CODE	The code of the course.
3	COURSE_NAME	The full name of the course.
4	CRD_HRS	The credit hours per semester.
5	STUDENT_ID	A student number is a unique number for each student.
6	GENDER_NAME	Female, Male.
7	ENTRY_DATE	Date of adding the course to the student schedule.
8	CONFIRMED_MARK	Student points of 100 in every course.
9	GRADE_DESC	A+, A, B+, B, etc.
10	CUM_GPA	Cumulative Grade Point Average (CGPA) out of 5.0.
11	SEMESTER_GPA	Semester Grade Point Average (SGPA) out of 5.0.
12	STSTATUS_DESC	Active, graduate, dropped out, etc.
13	MAJOR_NAME	MIS, Accounting, Finance, Economics, BA, or Pre-Major
14	STUDENT_LEVEL	Actually level of the student such as First level, second level, etc.

3.2 Data Pre-processing

Real-world data tends to be noisy, incomplete, and inconsistent. For this reason, the best practice to use before data mining techniques is the application of data pre-processing, which will ensure error-free and high-quality data. The data pre-processing steps are shown in Figure 2.

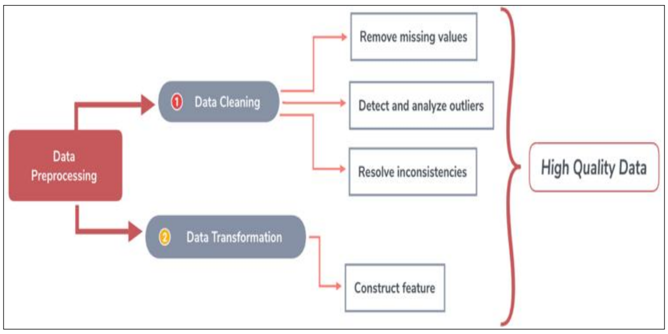

Fig 2. Data Pre-processing

3.2.1 Data Cleaning

(1) Remove missing values

In the first step, we used the Orange platform to clean the data and remove missing values, where the records containing empty values were completely deleted. After the records of missing values were deleted, the data reduced to 52,430 records.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 10 (2020), pp. 2895-2908
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.10.2020.2895-2908

(2) Resolve inconsistencies
Inconsistent data is that contain discrepancies in names or values. It was done through used the Microsoft Excel, involved checking the data set, and used this step as a means of avoiding future errors and conflicts.

(3) Detect outliers and anomalies
As we know, outliers can present as an incorrect values entry, sampling error or exceptional true value. We checked outliers' values to identify them and make sure they are not incorrect values. In the third step of data pre-processing, we reveal the outliers by using the outliers' widget in the Orange platform. The widget revealed 525 outlier cases. Figure 3 displays the outliers' detection by scatter plot.

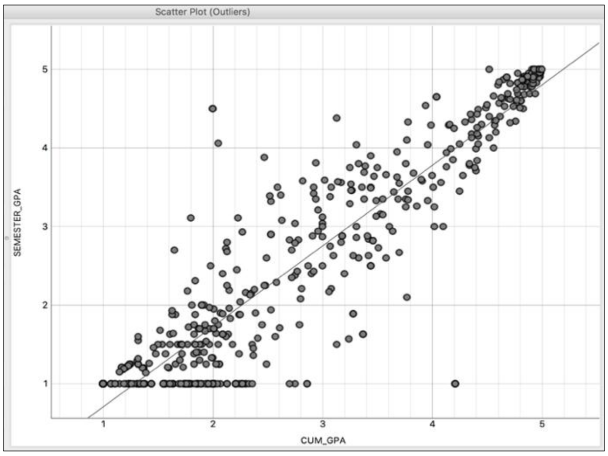

Fig 3. Scatter plot detecting outliers/ anomalies cases

3.2.2 Data Transformation

Following data cleaning, we used data transformation to provide more effective results. It should be noted that some of the proposed algorithms require GPA classification due to it not being able to handle continuous numerical values. Our study classified GPA into five categories.

The first new feature was called "Class_GPA" and was assembled by using CUM_GPA attribute, normally used to split students by their CGPA into multiclass classifications. The CGPA was classified into five categories, as shown in Table 2.

Table 2: CUM_GPA classification

#	CGPA	Class
1	>=4.5	Excellent
2	>=3.75	Very Good
3	>=2.75	Good
4	>=2.0	Acceptable
5	<2.0	Fail

We created a second new feature, called "Class_Semesters," to group the semesters into years, using SEMESTER attribute, as shown in Table 3.

Table 3: SEMESTER classification

#	Semester	Class
1	342-351 (2014)	First Year
2	352-361 (2015)	Second Year
3	362-371 (2016)	Third Year
4	372-381 (2017)	Fourth Year
5	382-391 (2018)	Fifth Year

We created the third new feature, called "Class_Marks," to group students' grades into two sub-groups by using the CONFIRMED_MARK attribute, as shown in Table 4.

Table 4: CONFIRMED_MARK classification

#	CONFIRMED_MARK	Class
1	>=60	P (Pass)
2	<60	F (Fail)

3.3 Application of Data Mining Techniques

(1) Classification Methods

Decision Tree (DT) is a tool that helps support decisions and uses a flow chart in the form of a tree that contains a set of rules can be represented in this form "IF-THEN" [2], [4], [33].

Random Forest (RF), an ensemble method, is normally used to improve accuracy [34]. The principle of the ensemble method is that weak classifiers can be combined to form a strong ensemble model or strong classification method. The RF is a collection of DTs (weak classifiers), with all the outcomes of these DTs collected to produce the RF, which is a strong classifier, then "the average" or "the majority voting" is used to predict the final result [16], [35], [36].

Naïve Bayes (NB) a simple technique for probability classification based on Bayes' theorem. It is called Naïve because it assumes that all attributes are independent of each other, which means the attributes are not correlated with each other [24], [37]. This algorithm is faster because this classifier requires small amounts of training data and less computing than other algorithms [2].

Model performance is measured using the Confusion Matrix. It is a table that contains columns and rows, where the number of columns and rows depends on the number of classes. It displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Several measures can be derived from the confusion matrix to evaluate the performance of models. In this study, our focus is on

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 10 (2020), pp. 2895-2908
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.10.2020.2895-2908

Classification Accuracy (CA), Precision, F1-score and Recall, as seen in Equations (1) to (4) [2].

$$CA = \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$F1\text{-}score = \frac{2\times Precision \times Recall}{Precision+Recall} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

(2) Regression Methods

Linear Regression (LR) is a predictive model used to predict the value of the dependent variable (y) based on the value of the independent variable (x) [10], [38]. LR can produce accurate predictions and is considered one of the easiest techniques to apply. In the LR model, the two-dimensional data is represented as dots falling into a straight line, where the X-axis is the predictor and the Y-axis is the target [39]. The performance of the regression model is evaluated based on four of the most popular metrics: Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R-squared) [40]. The MSE, RMSE, MAE, and R-squared are presented below, from Equation (5) to Equation (8). Where n is the total number of observation/ rows, y_i represents the actual values, $\hat{y}_i$ represents predicted/ estimated values, $\bar{y}_i$ is the mean of the actual y_i values and the i value ranging from 1 to n.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{8}$$

(3) Association Rules

Finding meaningful rules among student data requires the use of Association Rules, which helps to extract frequent patterns between data. Confidence and support measures are used to identify the relationships between transactions. Support refers to the probability that the transaction contains A and B of itemsets A and B. In contrast to this, one can confidently evaluate, to a degree of certainty, the discovered correlation, which is the probability that a transaction containing A also contains B [2]. The user identifies the initial values of minimum support and confidence to produce association rules so that when the generated rules with values of confidence and support for itemsets is lower than the predefined minimum value, these itemsets are not accepted as a frequent itemset; consequently, the generated rules will be rejected [30], [41], [42]. The Equation of support and confidence measures are given in Equation (9) and Equation (10), respectively. A and B are frequent itemsets, P is the probability [2].

$$Support(A \Longrightarrow B) = P(A \cup B) = \frac{freq(A,B)}{N} \tag{9}$$

$$Confidence\ (A \Longrightarrow B) = P(A|B) = \frac{freq(A,B)}{freq(A)} \tag{10}$$

4. EXPERIMENTS AND ANALYSIS

4.1 The General Analysis of Student Performance

Students' performance was analysed through the Orange platform. We used the Distribution widget to shows the values for Class_GPA based on five years of study. We compared the performance of students over five years to determine the possibility of failure and excellence. Table 5 shows the probability of failure and excellence for five years and the total number of student records in each semester. It also displays the percentage of students who excel and fail.

In the next part, we compared the failure and excellence rates, where the records were divided into ten semesters. The goal was to search for the semester that comprised large numbers of students failing and excelling. Table 6 focusses on students who excel and fail and compares this to the rates of excellence and failure throughout the ten semesters.

Students' GPA was analysed based on Major_Name, to know which majors include the most significant number of excellent and failed students. Table 7 shows data on excellent and failed students based on Majors. Since the number of students influences the failure or excellence rate, the total number of students in each major is calculated, as shown in the following Table 7.

In the next part, the data will be analysed based on gender, to identify which gender more often fails to achieve a high CGPA. Table 8 presents data on failed and excellent students based on gender, where the number of failed male students' records was 2,289, whereas the number of failed female students' records was 798. The table also shows the probability of failure and excellence in student records. From this table, it became clear that female students earned higher percentages of distinction.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 10 (2020), pp. 2895-2908
 https://dx.doi.org/10.37624/IJERT/13.10.2020.2895-2908

Table 5: Students who excel and fail, by year

Year	Total No. of Student Records	No. of Excellent Students' Records	No. of Failed Students' Records	% of Excellent Students' Records	% of Failed Students' Records	Probability of Excellent Students	Probability of Failed Students
First Year	9648	682	673	17.23%	21.80%	0.071 ± 0.005	0.070 ± 0.005
Second Year	10697	731	509	18.47%	16.49%	0.068 ± 0.005	0.048 ± 0.004
Third Year	8104	478	425	12.08%	13.77%	0.059 ± 0.005	0.052 ± 0.005
Fourth Year	13737	1105	783	27.92%	25.36%	0.080 ± 0.005	0.057 ± 0.004
Fifth Year	10221	962	697	24.31%	22.58%	0.094 ± 0.006	0.068 ± 0.005

Table 6: Students who excel and fail, by semester

Semesters	Total No. of Students' Records	No. of Excellent Students' Records	No. of Failed Students' Records	% of Excellent Students' Records	% of Failed Students' Records
342	4845	291	384	7.35%	12.44%
351	4803	391	289	9.88%	9.36%
352	5242	391	252	9.88%	8.16%
361	5455	340	257	8.59%	8.33%
362	5288	295	271	7.45%	8.78%
371	2816	183	154	4.62%	4.99%
372	6630	551	243	13.92%	7.87%
381	7107	554	540	14.00%	17.49%
382	7363	774	496	19.56%	16.07%
391	2858	188	201	4.75%	6.51%

Table 7: Students who excel and fail, by majors

Major Name	Total No. of Students' Records	No. of Failed Students' Records	No. of Excellent Students' Records	% of Failed Students' Records	% of Excellent Students' Records	Probability of Failed Students' Records	Probability of Excellent Students' Records
MIS	6731	60	640	1.94%	16.17%	0.009 ± 0.002	0.095 ± 0.007
Finance	8031	76	498	2.46%	12.58%	0.009 ± 0.002	0.062 ± 0.005
Pre-Major	5691	2313	235	74.93%	5.94%	0.406 ± 0.013	0.041 ± 0.005
Accounting	11081	159	1202	5.15%	30.37%	0.014 ± 0.002	0.108 ± 0.006
Economics	3440	0	695	0%	17.56%	-	0.202 ± 0.013
BA	17433	479	688	15.52%	17.38%	0.027 ± 0.002	0.039 ± 0.003

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 10 (2020), pp. 2895-2908
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.10.2020.2895-2908

Table 8: Students who excel and fail, by gender

Gender Name	Total No. of Students' Records	No. of Excellent Students' Records	No. of Failed Students' Records	% of Excellent Students' Records	% of Failed Students' Records	Probability of Excellent Students' Records	Probability of Failed Students' Records
Female	27177	3380	798	85.40%	25.85%	0.124 ± 0.004	0.029 ± 0.002
Male	25230	578	2289	14.60%	74.15%	0.023 ± 0.002	0.091 ± 0.004

4.2 The Experimental Results of Data Mining Method

4.2.1 Experimental Results of Classification

The data set was divided, with 75% training data and 25% test data, with Class_GPA as the target variable. Table 9 presents the evaluation results of DT, RF, and NB. The table contains the results of CA, F1-score, Precision, and Recall. Figure 4 shows the results of the evaluation of the three models.

Table 9: The evaluation results of the prediction

Model	CA	F1-score	Precision	Recall
RF	0.713	0.712	0.715	0.713
DT	0.698	0.697	0.699	0.698
NB	0.594	0.595	0.605	0.594

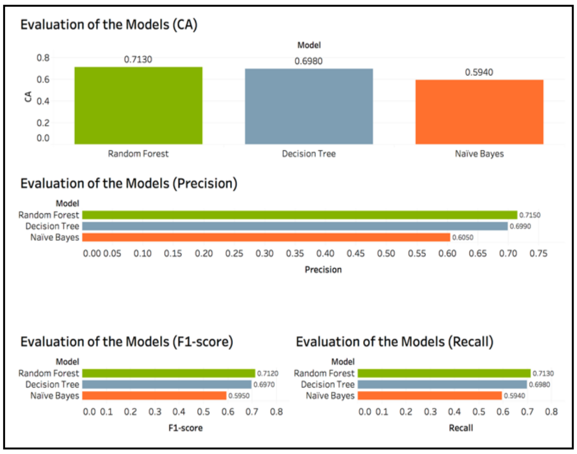

Fig. 4. Results evaluation of the three models

4.2.2 Experimental Results of Regression

Initially, the correlation between the CUM_GPA and the SEMESTER_GPA was examined and drawn on a Scatter Plot to establish a relationship between the two variables. Figure 5 shows the dependent variable being the CUM_GPA, and the independent variable the SEMESTER_GPA. As we can see from the figure, the coefficient of correlation r=0.88, this value indicates that the relationship between the two variables was a strong positive relationship since both runs in a straight line and increase in parallel.

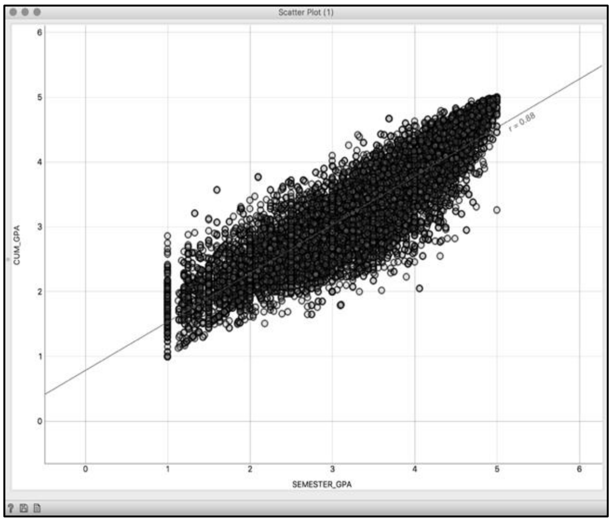

Fig. 5. Scatter plot of CUM_GPA and SEMESTER_GPA

Two regression models were used to predict student performance, those being the LR and the DT. The target variable was the CUM_GPA attribute, with the predictor being the SEMESTER_GPA and an attribute that has a meaningful impact on CGPA. Figure 6 visualises the Error rate of the models using various measures. Whereas, Table 10 presents the evaluation results values of the regression models.

Table 10: Regression models results' evaluation

Model	MSE	RMSE	MAE	R-squared
LR	0.154	0.393	0.315	0.773
DT	0.135	0.368	0.288	0.801

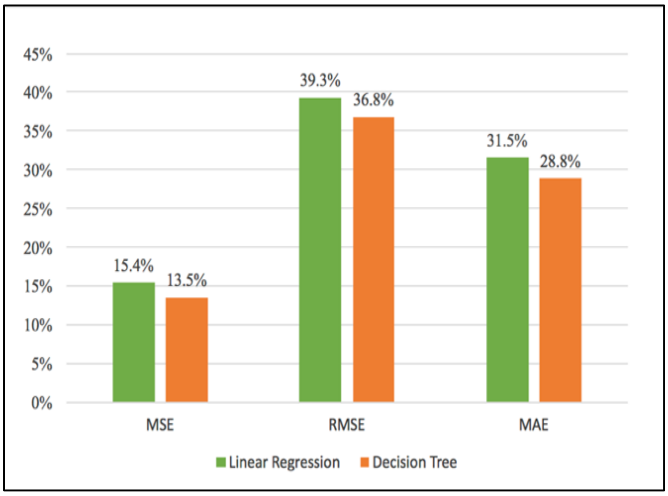

Fig. 6. Error rate of the models.

4.2.3 Experimental Results of Association Rules

Finding a strong association between items in the multidimensional data set is not always easy, due to the variance of data. For this reason, the association between data will be examined in three approaches by selecting two attributes in each method. All the rules generated will need to be higher than the minimum support value and also higher than the minimum confidence value [41]. For this reason, we determined the minimum support value to be 2%. Additionally, we identified the minimum confidence value to be 30% for all approaches.

The first approach is the impact of a subject major on the CGPA, with thirteen rules extracted, and based on two attributes MAJOR_NAME and Class_GPA, as observed in Table 11.

Table 11: The first approach of Association Rules

Rule #	If (Antecedent)	Then (Consequent)	Support	Confidence
1	MAJOR_NAME=BA	Class_GPA= Good	14.4%	43.4%
2	Class_GPA= Good	MAJOR_NAME= BA	14.4%	36.1%
3	MAJOR_NAME= BA	Class_GPA= Acceptable	12.8%	38.5%
4	Class_GPA= Acceptable	MAJOR_NAME= BA	12.8%	46.6%
5	MAJOR_NAME= Accounting	Class_GPA= Good	8.5%	40.2%
6	MAJOR_NAME= Finance	Class_GPA= Good	7.1%	46.1%
7	MAJOR_NAME= MIS	Class_GPA= Good	5.9%	46.1%
8	MAJOR_NAME= Pre-Major	Class_GPA= Fail	4.4%	40.6%
9	Class_GPA= Fail	MAJOR_NAME= Pre-Major	4.4%	74.9%
10	MAJOR_NAME= Pre-Major	Class_GPA= Acceptable	3.4%	31.5%
11	MAJOR_NAME= Economics	Class_GPA= Very_Good	2.4%	36.3%
12	Class_GPA= Excellent	MAJOR_NAME= Accounting	2.3%	30.4%
13	MAJOR_NAME= Economics	Class_GPA= Good	2.2%	32.8%

The second approach is the impact of subject major on the students' marks, with ten rules extracted, and based on two attributes MAJOR_NAME and Class_Marks, as shown in Table 12.

Table 12: The second approach of Association Rules

Rule#	If (Antecedent)	Then (Consequent)	Support	Confidence
1	MAJOR_NAME= BA	Class_Marks =P	29.5%	88.8%
2	Class_Marks= P	MAJOR_NAME= BA	29.5%	33.4%
3	MAJOR_NAME= Accounting	Class_Marks =P	19.3%	91.2%
4	MAJOR_NAME= Finance	Class_Marks =P	14.3%	93%
5	MAJOR_NAME= MIS	Class_Marks =P	11.8%	92%
6	MAJOR_NAME= Pre-Major	Class_Marks =P	7.1%	65.7%
7	MAJOR_NAME= Economics	Class_Marks =P	6.4%	97.8%
8	Class_Marks= F	MAJOR_NAME= BA	3.7%	32.3%
9	MAJOR_NAME= Pre-Major	Class_Marks =F	3.7%	34.3%
10	Class_Marks= F	MAJOR_NAME= Pre-Major	3.7%	32.2%

The third approach is the impact of courses on the students' marks; seven of the rules have been extracted based on two attributes COURSE_NAME and Class_Marks, as presented in Table 13.

Table 13: The third approach of Association Rules

Rule #	If (Antecedent)	Then (Consequent)	Support	Confidence
1	COURSE_NAME= Feasibility analysis of projects	Class_Marks=P	3.3%	99.4%
2	COURSE_NAME= Operations Management	Class_Marks=P	3%	79.6%
3	COURSE_NAME= Introduction to management information systems	Class_Marks=P	2.9%	92.6%
4	COURSE_NAME= Organizational Behavior	Class_Marks=P	2.8%	95.2%
5	COURSE_NAME= Strategic Management	Class_Marks=P	2.8%	98%
6	COURSE_NAME= Saudi Commercial Law	Class_Marks=P	2.5%	98.2%
7	COURSE_NAME= Principles of Management Accounting	Class_Marks=P	2.1%	72%

4.2.4 Experimental Results of Anomalies' Analysis

In the EDM, anomaly detection is not only used to find students with academic problems and poor performance but also to discover students who excel in academic performance. Also, the detection of outliers helps the educational institution make effective decisions that help the student avoid making wrong decisions. In our work, the outliers were discovered in the CBE students' data through the use of outliers' analysis. About 525 anomalies and about 51882 inliers were obtained after applying the outliers' detection method. This analysis helps to detect anomalies that may be distinctive and useful to the CBE. It also helps in discovering cases, which may turn into problems to be avoided. It will also help decisions to be solved, such as trying to find solutions that would help improve the performance of students with low GPAs.

5. FINDING AND DISCUSSION

The following are the results obtained in the pursuit of the study's goals. The results of the analysis of student academic records showed that by comparing five years, that the highest excellence and failure rates occurred in the students' fourth year. In contrast, the lowest excellence and failure rates occurred in the students' third year. It's important to note that the number of student records in the fourth year is higher than in other years; the reason may be due to an increase in the admission rate in that year (2017). Furthermore, by comparing the fourth and fifth year for excellent students, we found that the probability of excelling in the fifth year is higher than the fourth year by between 0.1 and 0.088, which is slightly higher by 0.015 and 0.013, compared to the fourth year. Whereas comparing the fourth and fifth year for failed students, we noted that the probability of failure in the fifth year is between 0.073 and 0.063, which is a slight increase of 0.012 and 0.010, compared to the fourth year. By comparing the ten semesters, the highest failure rate was observed in the first semester of 2017 by 17.49%, and the lowest failure rate was in the first semester of 2016 by 4.99%. As for students who excelled, we noted that the highest rate of excellence in a class was in the first semester of 2018 by 19.56% and the lowest rate was in the second semester of 2016 by 4.62%.

Consequently, these results helped to make an important observation, which is that the number of students in the fourth year has a strong impact on increasing the percentage of excellence and failure in that year, which reached 27.92% and 25.36%, respectively.

Further, the students' GPA was analysed based on the major and, through this analysis, we noted that Accounting students outperformed all students in the excellence classes. In contrast, we found that Pre-Major students were higher in the failure classes. On the other hand, the highest failure rate was in a Pre-Major by 74.93% and the probability of increased failure is estimated to be between 0.419 and 0.396. Since the Pre-Major is the major that contains general study courses from various majors, we found that most students fail in some courses, especially in the first three semesters of study at the college. In contrast, the probability of increasing excellence in the Economics department is the highest among other majors, where the probability value is estimated to be between 21.5% and 18.9%. This is due to one of the following reasons. First, in our opinion, the failure rate in the Economics major is 0%, so it is likely to increase excellence. Second, we think this major may be easy, as it depends on theoretical more than practical courses.

Therefore, these results help the decision-makers to find alternative methodological plans for the difficult specialisations that have a high rate of failure and develop studied plans that contribute to raising students' academic achievement. Moreover, the performance of students was analysed based on gender, and we observed that female students outperformed male students, where the analysis showed that the records of failed male students exceeded the records of female students by 1,491 records. Also, the probability of failure in male students' records was between

0.095 and 0.086. Whereas, the probability of failure in the female students' records was between 0.031 and 0.027. Furthermore, it became clear that the highest percentage of distinction was in the records of female students, where the percentage of excellence was 85.4%. In contrast, the percentage of excellence for male students was 14.6%. We also noted that the probability of a high GPA in female students' records was between 0.128 and 0.120, whereas the probability of a high GPA in male students' records was between 0.025 and 0.021. Overall, the results indicated that female students' outperformed male students and that they are less likely to fail than male students. Besides, the probability of male students obtaining a failed GPA is 7% higher than the probability of excelling. As for female students, the probability of superiority is 9.7% higher than the probability of failing.

Accordingly, these results lead us to the fact that female students are more diligent in obtaining high rates and avoiding failure in their academic performance. These results help the college to try to search for the reasons that led to the failure of male students in their academic performance, educate students by setting up seminars that support them in raising their academic performance, the search behind the reasons that led to their failure and take the crucial decisions to reduce this failure in the coming years

On the other hand, the evaluation results of classification methods showed that RFs achieved the highest scoring 71.3% on CA and Recall, 71.5% on Precision and 71.2% on F1-score. The next algorithm was the DT with 69.7% on F1-score, 69.8% on CA and Recall, and showed slight increases on Precision by 0.1%, which means it scored 69.9%. Meanwhile, the NB appeared to be the worst algorithm, obtaining 59.4% on CA and Recall, 59.5% on F1-score, and 60.5% on Precision. We can conclude from these findings that the performance of the RF algorithm on this type of data set is excellent. Therefore, one of the points to be taken into account is that the principle of RF and the ensemble learning method is proportional to our data set, which is structured data. Where the basic principle of RF is that a group of weak learners can be combined to form a strong collective learner, this principle helped to obtain an adequate evaluation in the classification of student performance. Furthermore, we found that the DT was lower by 1.5% on CA than RF; this indicates that the RF is more accurate with results than the DT, and the DT built according to IF-THEN rules [2]. Accordingly, we concluded from this assessment that a rule-based classifier is proportional to the data set used in this study.

Finally, according to the results, RFs have outperformed the other algorithms in all evaluation measures. This can be used to meet the requirements of the university in achieving quality and discovering weak students, as well as finding students who show excellent and exceptional capabilities

As for the results of the evaluation of regression models, the value of the average of the squared of the errors (MSE) was estimated at 15.4% in LR, whereas in DT it was estimated at 13.5%. The value of the differences between the actual values and the values predicted by the LR (RMSE) was 39.3%, whereas the DT was 36.8%. Also, the value of the average of the absolute values decided, calculating the differences among predicted and actual values (MAE), was estimated at 31.6% in LR, whereas in DT it was 28.8%. Moreover, the value of the proportion of variance of the dependent variable explained by the independent variable (R-squared) was 77.3% in LR, which indicates that the model shows 77.3 % of the variability in the CUM_GPA (the target variable). Whereas, in DT, the R-squared was 80.1%, which indicates that the model explains 80.1 % of the variability in the CUM_GPA. The result of evaluating the models' performance has shown us that the DT model is good and is better than LR, as the error rate in the DT is less than LR.

The association's rules were analysed based on three approaches. The first approach is the impact of a major on the CGPA; we observed from the first and third rule that students in the BA category are most likely to obtain a good GPA, with 43.4% confidence. They are also most likely to obtain an acceptable GPA, with 38.5% confidence. Furthermore, we noted from the second and fourth rule that the vast majority who obtain a good GPA, with 36.1% confidence and an acceptable GPA, with 46.6% confidence are BA students.

As in the fifth, sixth and seventh rule, Accounting, Finance, and MIS students are more likely to get a good GPA with 40.2%, 46.1%, and 46.1% confidence, respectively. It was noted in the eighth and ninth rules, the Pre-Major students, often get a GPA to fail with 40.6% confidence. Also, that the failed students most probably belong to the Pre-Major with 74.9% confidence. As the 10^{th} rule states, students of Pre-Major may obtain an acceptable GPA, with 31.5% confidence.

As for Economics students, the 11^{th} and 13^{th} rules show that they are more likely to have a very good GPA, with 36.3% confidence, and a good GPA with 32.8% confidence. As for the 12^{th} rule, they are the lowest in confidence value, 30.4%; this rule says that if the GPA class belongs to the excellent group, then the major will be Accounting. This rule indicates that most students who excel the most belong to the Accounting major.

Through these thirteen rules, it is clear to us that the highest confidence obtained was 74.9%, which shows that failure rates often occur in the Pre-Major. As we mentioned previously, the Pre-Major is a major that is taken before specialisation and comprises courses from all majors. We surmise that its students often fail because some of their courses are from disciplines they do not like.

We noticed the next rule that scores the highest confidence, at 46.6%, states that if the GPA class belongs to "acceptable," then the major is a BA. The BA is dominated by an acceptable GPA, and it is the most popular specialisation in the CBE with 17,433 records. This discovery may indicate that most students tend to belong to this specialisation due to the belief among many that courses tend to be easy. This may also be due to the popularity of this major, which provides jobs for graduates at many companies and organisations.

The second approach is the impact of a major on the students' marks; we noticed that the rule with the highest confidence, 97.8%, is the seventh rule. This rule shows that if a major is in Economics, it is more likely that it will obtain a mark of "P", which means that Economics students will likely pass all courses. This is followed by the fourth rule, with a 93% confidence. This rule clarifies that if the major is Finance, they are likely to pass the courses. The fifth rule, with 92% confidence, indicated that if the major is Management Information Systems, the marks will constitute a pass. The next rule is the third rule, with 91.2% confidence, denoting that if the major is Accounting, then they are likely to pass the courses too. The last rule with high confidence is the first rule, at 88.8%, which shows the student who belongs to the BA group is most likely to obtain a pass mark. On the other hand, if the major in Pre-Major, then they will pass the courses with a confidence of 65.7%, as in the sixth rule. Whereas the ninth rule states that if the major in Pre-Major, then it is likely that the failure of a course will be obtained with a confidence of 34.3%. The second rule says that if the class mark is "P," it is likely that the major is a BA, with a 33.4% confidence. Whereas the eighth and 10th rules state that if the class of marks is "F" this indicates that the major is a BA or Pre-Major, with little confidence, 32.3% and 32.2%, respectively. Finally, we have concluded that these majors, Economics, Finance, MIS and Accounting, are more likely to get the pass in courses, with high confidence, over 90%. This demonstrates that excellent and interested students always belong to these majors. Furthermore, when students belong to the fields they prefer, they give their best.

The third approach is the impact of courses on the students' marks; the resulting rules show that a student who registered in the course "Feasibility analysis of projects" is most likely to obtain a pass mark, with a 99.4% confidence, as in the first rule. We also noted in the fifth and sixth rules, that with a 98.2% and 98% confidence, and if the course is "Saudi Commercial Law" and the course is "Strategic Management," then students will most likely pass this course. The fourth rule states that if the course is "Organisational Behavior," then students will pass this course, with a 95.2% confidence, as the third rule, with a 92.6% confidence. If the course is "Introduction to Management Information Systems," students are more likely to pass this course. Moreover, we have two rules where we see less confidence than 92.6% by almost 13%, which are the second and seventh rules, with 79.6% and 72%, respectively. As the second rule states, that if the course is "Operations Management," students will succeed in this course. As for the seventh rule, it appears that if the course is "Principles of Management Accounting," students will also pass this course. Finally, according to our experience in the CBE courses, "Feasibility analysis of projects," "Saudi Commercial Law" and "Organisational Behavior" are general education courses in the five departments: Management Information Systems, Accounting, Finance, Economics and Business Administration, the "Strategic Management" course is a general education course in Management Information Systems, Accounting, Economics and Business Administration. These general education courses aim to expand the scope of students' understanding by adding courses from different specialisations, for the student to graduate with knowledge of majors different than the one they primarily studied.

The research findings suggest that the knowledge obtained from the third approach means that students often pass general education courses. Many students intentionally add these courses, either to raise their GPA due to the course being easy or because of the cooperation they are felt with the lecturer. Also, these courses may be added to fill the gap of the academic schedule because some students prefer to not have too much free time in their schedules. Furthermore, one of the conditions of the college is that students must complete the courses of the first three levels before specialisation, with a second condition being the obligation to obtain a GPA higher than 2, conditions that led some students to be shut out of specialisations. So, they have to add these general courses to finish previous courses or raise their GPA. We did note that the rules with the F mark did not appear in this analysis under the measures' selected values. We conclude from this that there were more instances of success than failure.

Furthermore, after the outliers' analysis, we noticed that significant anomalous data appeared in the records of students of Pre-Major. The anomaly was due to the weak SGPA and CGPA. In addition to their course failures. Student failure at the first level was often due to several reasons, such as the difficulty of the courses, the difference in the methods of lecturers teaching the courses or the standardisation of questions (standardised test) between the female and male students department. Also, there may be personal reasons related to the student's social life. Consequently, a strategic plan must be designed to understand difficulties and problems experienced by the students of the first level, and then practical decisions could be made that are appropriate to these problems, to avoid students failing in future years. This brings up the necessity of the academic advisor, especially for Pre-Major students, to guide them in the continuation of their studies and to overcome difficulties. We have observed the problem of "academic separation" in the academic cases of most Pre-Major students. The terms "dropout," "discontinuation of study" and "termination" were also read. We also discovered a group of observed anomalies that serve the college in many respects, especially in obtaining high-quality standards in the education process. Where a group of students was found who have a high CGPA at all levels of study, they nevertheless graduated with an excellent CGPA. The college should, in turn, realize that the excellent students' experience leads to organised volunteer courses. These could be offered by the students who excel, and that can assist students of the same major. Those students' experiences may be used to provide advice to those who wish to join this major and could be achieved through social media.

6. CONCLUSION AND FUTURE WORK

The purpose of this study was to analyse student data in the CBE by extracting new patterns and features from their academic data. It additionally sought to detect anomaly cases. It did this by predicting the academic performance of students over the last five years, from 2014 to 2018, using data mining techniques. Moreover, it identified the students' weaknesses and failures and explored the knowledge that helps to improve the educational process. Furthermore, it tried to find the reasons for the students' repeated failure in a particular course.

This study explored, through the application of data analysis, first, that the probability of excellence and failure was in the fifth year more than in the fourth year (in the first and second semesters of 2018). We found through these results that the rate of excellence in the last year exceeded the failure rate by 2.7%. Second, the probability of increasing excellence among students of the department of Economics was the highest among other majors by more than 18.9%. On the other hand, the probability of increased failure in a Pre-Major was estimated to be more than 39.3%. Third, the probability of excellence in the records of female students was estimated between 12.8% and 12%, whereas the probability of excellence among the records of male students was estimated to be between 2.5% and 2.1%. Therefore, the analysis leads us to the following conclusion: male students and Pre-Major students are more likely to fail and therefore need, in this period, to follow up with academic advisors.

Additionally, according to the results of classification, RF has outperformed the other algorithms in all evaluation measures, with 71.3% of CA and Recall, F1-score 71.2%, and Precision 71.5%. Furthermore, as a result of evaluating the performance of the regression models, we have noticed that the DT model is not only good but is better than LR, as the error rate in the DT is less than LR. On this basis, we conclude from this study that the best classification model is RF, and the best regression model is DT.

Moreover, the results that we reached through the association's rules indicate that the knowledge obtained from the first approach was that failure rates often appeared in the Pre-Major with a 74.9% estimated confidence. The results also showed that if the GPA class belongs to "acceptable," then the major is BA with an estimated 46.6% confidence. Students from the following majors, Economics, Finance, MIS and Accounting, are more likely to get the pass marks in the courses, with an over-90% confidence. Our findings from the third approach suggest that the knowledge obtained shows that students often pass general education courses. It was clear through research and submitted questions to the officials in the CBE that many students intentionally add these courses either to raise their GPA due to the course being easy or because of the cooperation they felt with the lecturer. There is also passion and curiosity felt by some students, who enroll in these courses and obtain valuable information that will benefit them in future.

In future, researchers would need additional data for the analysis, to increase the accuracy of the prediction. They may also want to focus on features that have a substantial impact on student performance, such as high school rate, absences, the number of notifications and the number of failures in a course. Additional models, such as traditional Neural Networks and deep learning, could be employed.

ACKNOWLEDGMENT

The authors would like to thank the College of Business and Economics at Qassim University that provided the data required for this research.

REFERENCES

[1] R. M. Damin, M. A. Kadry, and E. M. Hamed, “An investigation into the use of the education Management Information System (EMIS) in Iraq: Case study,” in *2014 International Conference on Engineering and Technology (ICET)*, 2014, pp. 1–6.

[2] M. K. Jiawei Han Jian Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier, 2012.

[3] R. Lawrance and V. Shanmugarajeshwari, “An assay of teachers’ attainmentusing decision tree based classification techniques,” in *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies, ICCPCT 2017*, 2017.

[4] K. N. Shah, M. R. Patel, N. V Trivedi, P. N. Gadariya, R. H. Shah, and N. Adhvaryu, “Study of Data Mining in Higher Education-A Review,” *International Journal of Computer Science and Information Technologies*, vol. 6, no. 1, pp. 455–458, 2015.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine*, vol. 17, no. 3, pp. 37–54, Mar. 1996.

[6] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, “Predicting Students Performance in Educational Data Mining,” in *2015 International Symposium on Educational Technology (ISET)*, 2015, pp. 125–128.

[7] B. Kumar and S. Pal, “Mining Educational Data to Analyze Students Performance,” *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, pp. 63–69, 2011.

[8] A. I. Adekitan and E. Noma-Osaghae, “Data mining approach to predicting the performance of first year student in a university using the admission requirements,” *Education and Information Technologies*, vol. 24, no. 2, pp. 1527–1543, Mar. 2018.

[9] L. A. Buschetto Macarini, C. Cechinel, M. F. Batista Machado, V. Faria Culmant Ramos, and R. Munoz, “Predicting Students Success in Blended Learning—Evaluating Different Interactions Inside Learning Management Systems,” *Applied Sciences*, vol. 9, no. 24, p. 5523, Dec. 2019.

[10] S. Angra and S. Ahuja, “Implementation of data mining algorithms on student’s data using rapid miner,” in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 2017, pp. 387–391.

[11] R. S. J. Baker, “Data Mining for Education,” 3rd Editio., vol. 7, 2010, pp. 112–118.

[12] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[13] S. Roy and A. Garg, "Predicting academic performance of student using classification techniques," *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, UPCON 2017*, vol. 2018-Janua, pp. 568–572, 2018.

[14] S. Ahmed, R. Paul, and A. S. M. L. Hoque, "Knowledge discovery from academic data using Association Rule Mining," in *2014 17th International Conference on Computer and Information Technology (ICCIT)*, 2014, pp. 314–319.

[15] M. Hasibur Rahman and M. Rabiul Islam, "Predict Student's Academic Performance and Evaluate the Impact of Different Attributes on the Performance Using Data Mining Techniques," *2nd International Conference on Electrical and Electronic Engineering, ICEEE 2017*, no. December, pp. 1–4, 2018.

[16] B. Kapur, N. Ahluwalia, and S. R, "Comparative Study on Marks Prediction using Data Mining and Classification Algorithms," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 3, pp. 632–636, Apr. 2017.

[17] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," in *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*, 2019, pp. 243–247.

[18] J. H. Sharp and L. A. Sharp, "A comparison of student academic performance with traditional, online, and flipped instructional approaches in a C# programming course," *Journal of Information Technology Education: Innovations in Practice*, vol. 16, no. 1, pp. 215–231, 2017.

[19] V. Shanmugarajeshwari and R. Lawrance, "Analysis of students' performance evaluation using classification techniques," *2016 International Conference on Computing Technologies and Intelligent Data Engineering, ICCTIDE 2016*, pp. 1–7, 2016.

[20] S. B. Rahayu, N. D. Kamarudin, and Z. Zainol, "Case Study of UPNM Students Performance Classification Algorithms," *International Journal of Engineering and Technology*, vol. 7, no. December 2018, pp. 285–289, 2018.

[21] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood, K. U. Sarker, and A. Rafi, "Student Academic Performance Prediction by using Decision Tree Algorithm," in *2018 4th International Conference on Computer and Information Sciences (ICCOINS)*, 2018, pp. 1–5.

[22] A. Marwaha and A. Singla, "A study of factors to predict at-risk students based on machine learning techniques," in *Advances in Intelligent Systems and Computing*, 2020, vol. 989, pp. 133–141.

[23] A. I. Adekitan and O. Salau, "Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance," *SN Applied Sciences*, vol. 2, no. 1, pp. 1–15, Jan. 2020.

[24] S. S. Al-Nadabi and C. Jayakumari, "Predict the selection of mathematics subject for 11 th grade students using Data Mining technique," in *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, 2019, pp. 1–4.

[25] H. Mousa and A. Maghari, "School Students ' Performance Predication Using Data Mining Classification," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 6, no. 8, pp. 136–141, 2017.

[26] A. Al Mazidi and E. Abusham, "Study of general education diploma students' performance and prediction in Sultanate of Oman, based on data mining approaches," *International Journal of Engineering Business Management*, vol. 10, pp. 1–11, 2018.

[27] K. Sunday, P. Ocheja, S. Hussain, S. S. Oyelere, B. O. Samson, and F. J. Agbo, "Analyzing Student Performance in Programming Education Using Classification Techniques," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 02, p. 127, Jan. 2020.

[28] P. Rojanavasu, "Educational data analytics using association rule mining and classification," in *ECTI DAMT-NCON 2019 - 4th International Conference on Digital Arts, Media and Technology and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering*, 2019, pp. 142–145.

[29] S. Kotsiantis and D. Kanellopoulos, "Association Rules Mining: A Recent Overview," *GESTS International Transactions on Computer Science and Engineering*, vol. 32(1), pp. 71–82, 2006.

[30] V. Nida Uzel, S. Sevgi Turgut, and S. Ayse Ozel, "Prediction of Students' Academic Success Using Data Mining Methods," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2018, pp. 1–5.

[31] A. F. Meghji, N. Ahmed Mahoto, M. A. Unar, and M. Akram Shaikh, "Analysis of Student Performance using EDM Methods," in *2018 5th International Multi-Topic ICT Conference (IMTIC)*, 2018, pp. 1–7.

[32] A. Naik and L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," *Procedia Computer Science*, vol. 85, pp. 662–668, Jan. 2016.

[33] M. A. Al-Hagery, "Classifiers' Accuracy Based on Breast Cancer Medical Data and Data Mining Techniques," *International Journal of Advanced Biotechnology and Research*, vol. 7, no. 2, pp. 760–772, 2016.

[34] K. Limsathitwong, K. Tiwatthanont, and T.

Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," in *2018 5th International Conference on Business and Industrial Research (ICBIR)*, 2018, pp. 110–114.

[35] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119–136, 2016.

[36] M. A. Al-Hagery, E. I. Al-Fairouz, and N. A. Al-Humaidan, "Improvement of Alzheimer disease diagnosis accuracy using ensemble methods," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 8, no. 1, pp. 132–139, 2020.

[37] A. Abu Saa, "Educational Data Mining & Students' Performance Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, pp. 212–220, 2016.

[38] P. M. Arsad, N. Buniyamin, and J. Ab Manan, "Neural Network and Linear Regression methods for prediction of students' academic achievement," in *2014 IEEE Global Engineering Education Conference (EDUCON)*, 2014, no. April, pp. 916–921.

[39] A. Ahlemeyer-Stubbe and S. Coleman, *A Practical Guide to Data Mining for Business and Industry*, 1st ed. Chichester, UK: John Wiley & Sons, Ltd, 2014.

[40] A. Sandoval, C. Gonzalez, R. Alarcon, K. Pichara, and M. Montenegro, "Centralized student performance prediction in large courses based on low-cost variables in an institutional context," *Internet and Higher Education*, vol. 37, no. June 2017, pp. 76–89, Apr. 2018.

[41] M. A. Al-Hagery, "Extracting hidden patterns from dates' product data using a machine learning technique," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 3, pp. 205–214, Dec. 2019.

[42] S. Hussain *et al.*, "Educational data mining and analysis of students' academic performance using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, p. 447, Feb. 2018.

ABOUT THE AUTHORS:

Ebtehal Ibrahim Al-Fairouz received her BSc in Computer Science from the Qassim University, Buraydah, KSA. She is a teaching assistant in the Department of Management Information System (MIS) at the College of Business and Economics (CBE) and a Master's student in Computer Science Department, Qassim University, KSA. Her research interests include data mining, data analytics, data visualisation and machine learning.

Mohammed Abdullah Al-Hagery received his BSc in Computer Science from the University of Technology in Baghdad Iraq-1994. He got his MSc. in Computer Science from the University of Science and Technology Yemen-1998. Al-Hagery finished his Ph.D. in Computer Science and Information Technology, (Software Engineering) from the Faculty of Computer Science and IT, University of Putra Malaysia (UPM), November 2004. He was the head of the Computer Science Department at the College of Science and Engineering, USTY, Sana'a from 2004 to 2007. From 2007 to this date, he is a staff member at the College of Computer, Department of Computer Science, Qassim University, Buraydah, KSA. He published more than 31 papers in various international journals. Dr. Al-Hagery was appointed the head of the Research Centre at the Computer College, and a council member of the Scientific Research Deanship Qassim University, KSA from September 2012 to October 2018. Currently, he is teaching the master degree students and a supervisor of four master thesis. He is a jury member of several PhD and master thesis, as an internal and external examiner in his field of his specialist.