Developing A Predictive Model for Selecting Academic Track Via GPA by using Classification Algorithms: Saudi Universities as Case Study

Thamer Althubiti, Tarig M. Ahmed, Madini O. Alassafi
Department of Information Technology-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah 21589, Saudi Arabia

Abstract—**The main motivation of any educational institution is to provide quality education. Therefore, choosing an academic track can be clearly seen as an obstacle, for students and universities, which in turn led to imposing a mandatory preparatory year program in Saudi Arabia. One of the main objectives of the preparatory year is to help students discover the right academic track. Nevertheless, some students choose the wrong academic track which can be a stumbling block that may prevent their progress. According to the tremendous growth of using information technology, educational data mining technology (EDM) can be applied to discover useful patterns, unlike traditional data analysis methods. Most of the previous research focused on predicting the GPA after the students choose an academic track. On the contrary, our research focuses on using classification algorithms to develop a predictive model for advising students to select academic tracks via prediction of the GPA based on the preparatory year data at Saudi Universities. Then, compare classification algorithms to provide the most accurate prediction. The dataset was extracted from a Saudi university containing preparatory year data for 2363 students. This work was carried out using five classification algorithms: Gradient Boosting(GB), K-Nearest Neighbors (kNN), Logistic Regression (LG), Neural Network(NN) and Random Forest(RF). The results showed the superiority of the Logistic Regression algorithm in terms of accuracy over the other algorithms. Future work could add behavioral characteristics of students and use other algorithms to provide better accuracy.**

Keywords—Data mining; educational data mining; classification algorithms; logistic regression; neural networks; gradient boosting; k-nearest neighbors; predicting students' performance

I. Introduction

In light of scientific progress and the development of communication and information technologies, there is a huge amount of data stored in database management systems (DBMS)[1]. It does not end with the ability to store this data, but it is more important how to use it in the production of knowledge [1] [2]. Recently, there is an increasing interest in science of data mining (DM). The concept of DM is simply a combination of artificial intelligence, statistics, machine learning, and databases [3] [4]. DM techniques can be used to discover unique patterns and hidden relationships. Data mining outcomes contribute to problem-solving, decision-making, and planning for organizations and companies [3]. It also plays a key role in various fields such as economy, healthcare, and education [4].

Educational data mining (EDM) is interested in discovering hidden relationships in data obtained from educational institutions or learning management systems (LMS). This area of research is used to take advantage of the data to better understand the students and what they learn. EDM is mainly used to predict students' academic performance to help them choose their study track [5]. This helps in making the right decision at the right time.

All countries seek to increase the quality of education. The increasing concern with the quality of education can be clearly seen in the Ministry of Education of Saudi Arabia as it acquires the highest share of the country's budget. According to the budget report, the Kingdom of Saudi Arabia's education budget amounted to 189 billion riyals, representing 18% of the total general budget in 2023 [6]. Additionally, applied of the preparatory year program was started in Saudi universities in 2009 [7]. The preparatory year, the first year of a student's university journey is considered to be the most important in a student's academic study. One of its aims is to prepare the students to choose an academic track based on their results [7] [8].

Due to the increasing number of academic tracks in universities, sometimes students choose a track that is not suitable for them, even if the results of the preparatory year qualify them for this track. This causes the failure of students or graduating with an unsatisfactory GPA. Furthermore, some students have to change their academic tracks after studying for several years, causing wasted effort. A research study conducted by M.J. Foraker at Western Kentucky University (WKU) in 2012 found that 25% of the students changed an academic track once, and 5% more than once [9].

This research aims to predict the academic via GPA of students in Saudi universities. It can help teachers and academic advisors modify students' study plans and improve academic performance. For the purposes of this study, reliance was made on a data set extracted from a Saudi university. It contains data for the preparatory year, by employing five classification algorithms: Gradient Boosting(GB), kNN, Logistic Regression (LG), Neural Network(NN) and Random Forest(RF). In addition, the model is evaluated by comparing the algorithms in terms of accuracy and area under the curve(AUC). To predict students' academic performance, we must get answers to these research questions.

1) How can we predict the right academic track via the GPA and preparatory year data of students in Saudi universities?

2) Which classification algorithm has the highest accuracy in predicting the right academic track for Saudi students?

Our paper is arranged as follows: Section II provides an overview of previous academic work in the field of predicting the academic performance of students in general and Saudi universities in particular and reviews the selected algorithms. Section III explains the methodology and materials used in preparing this paper to predict the academic track based on the GPA. It also describes the contents of the dataset and the tool used to extract the results and methods for evaluating the results of the proposed model. Section IV presents the results of predicting the GPA and the factors affecting students' academic performance in Saudi universities. We conclude our paper with Section V, in which we discuss the experimental results and answers to the research questions and offer the conclusion and future work.

II. Background

A. Predicting Academic Performance for Students

In Nigerian universities, the duration of the study is five years in engineering colleges. Adekitan and Salau [10] questioned about the possibility of predicting the last cumulative GPA based on the results of the first three years. They developed a model by using the KNIME application that experiments with six data mining algorithms (PNN, Random Forest, The Decision Tree, Naive Bayes, Tree Ensemble, and Logistic Regression). The dataset contains a record of 1842 students from different engineering departments. The GPA of the first three years, the first year of academic study, and the department were considered as input. The results showed that the GPA of the third year was the most effective in the prediction of the last cumulative GPA. They also showed that the Logistic Regression algorithm provides higher accuracy than the other algorithms, with an accuracy of 89.15%.

Ginting and Rahman [11] presented a prediction system for the GPA of university students. The proposed system uses an Artificial Neural Network and combines it with a supervised Backpropagation algorithm. The system consists of 18 nodes at the input layer with 24 hidden nodes to produce one node in the output layer. The dataset contains 591 records of students who graduated from an Indonesian university. The system was tested using four different methods, each method changes the number of test and training data. The accuracy of the proposed system was 97.2%.

Zollanvari et al. [12] applied the maximum-weight dependence tree to propose a GPA prediction model. The proposed model is based on the behavioral characteristics of the students. A questionnaire containing 20 questions was distributed in order to find out the behaviors that affect GPA prediction. These questions are based on the educational objectives. The number of students in the dataset is 82 students. The accuracy of the proposed model was 65.85%. Better results can be achieved by increasing the number of students in the dataset and incorporating academic performance with behavioral characteristics.

Putpuek et al. [13] compared the decision tree algorithm and data mining techniques to predict the students' GPA based on personal factors. The selected algorithms were (C4.5 and ID3) and the techniques were (Naïve Bayes and K-NN) and personal factors such as (gender, skills, type of acceptance, etc.). The dataset contains data of 2,281 students graduating from the same college in different years. The results showed the superiority of the Naïve Bayes techniques, as it achieved an accuracy of 43.18%. While ID3, C4.5, and K-NN achieved 41.65%, 42.88%, and 43.05% accuracy results, respectively.

In order to explore factors affecting the students' academic performance, Hamoud et al. [14] proposed a model that compares the algorithms of the Decision Tree (J48, Random Tree, and REPTree). The study was conducted on the students at Computer Science College at the Basra University in Iraq. Data was collected from 161 students' answers to a questionnaire containing 60 questions in different fields. The results showed the superiority of J48. The results also showed that the factors such as the GPA, the father's job, and the quality of food have a high effect on the students' performance. On the contrary, factors such as gender and age have a weak effect.

Pallathadka et al. [15] analyzed four machine learning algorithms to find out the most accurate one. The used algorithms are SVM, C4.5, ID3, and Naive Bayes. Examinations were conducted on the UCI machinery student performance dataset available online. The dataset contains 649 records and 33 factors. The results showed that the SVM algorithm is the most accurate.

B. Predicting Academic Performance for Students in Saudi Universities

In order to solve the problem of students graduating with a low GPA in Saudi Arabia and help through early intervention Alyahyan and Düşteaör[16] developed a model predicting the final GPA based on the results of the first year after the preparatory year. This model is based on decision tree algorithms (Rep Tree, Random Tree, and J48). The dataset contains the record of 339 students and 15 factors such as gender, nationality, subjects' grades, and final GPA. According to the results, the J48 algorithm achieves the highest predictive ability of up to 69.3%.

Al-Barrak and Al-Razgan [17] applied the J48 algorithm on a dataset of 239 female students majoring in computer science at a Saudi university. In order to find which courses, have the most impact on the final cumulative GPA. The dataset contains 16 compulsory computer science courses. Based on the results of the experiment, it was found the two courses' Software Engineering-1 and JAVA-2' have the greatest effect on the final grade.

In order to measure the ability of classification algorithms to predict the GPA Mueen et al. [18] proposed a predictive model based on a student's record in only two courses. They used three classification techniques (Naive Bayes, C4.5, and MLP). This study was conducted on King Abdelaziz

University students in two courses (Programming and Operating Systems). Everything related to the subject, including assignments, tasks, tests, etc., were collected through the Learning Management System LMS. The results showed the superiority of the Naive Bayes classifier over the classifiers, and it achieved a prediction ability to 86%.

Altujjar et al. [19] presented a predictive model to the performance of undergraduate students in the College of Computers at King Saud University using classification algorithms. The model aims to identify important courses that have a significant impact on academic achievement. The ID3 algorithm was used to build the model for each academic year. The dataset consists of 100 student records. The dataset was split into 75% for training and 25% for testing. The results showed that the courses (IT 221), (CSC111), and (CSC113) have a significant and clear impact on the students' academic performance.

Hilal Al-Murabaha [20] analyzed the data of Saudi university students by using classification techniques. The objective is to predict student performance during the undergraduate semester. The dataset contains the record of 225 students and 10 features such as (midterm exams, attendance, final exam score, previous exams score, science experiments, projects. etc.). Five classifiers are applied to analyze student data (Naive Bayes, Bayesian Network, ID3, J48, and Neural Network) using the WEKA tool. The results showed the superiority of Bayesian Network over other classifiers, with an accuracy of 92% also, the amount of data affects the accuracy of the results.

C. Classification Algorithms

1) Neural Networks (NN): Additionally referred to as artificial neural networks (ANNs) or simulated neural networks (SNNs), are at the top of deep learning algorithms. Their call and shape are inspired by the way of the human brain, mimicking the manner that biological neurons signal to each other [21] [22].

(ANNs) re-constructed from node layers, containing an enter layer, one or more hidden layers, and an output layer. each node, or artificial neuron, connects to some other and has an associated weight and threshold [21] [22]. If the output of any individual node is above the specified threshold fee, that node is activated, sending information to the following layer of the group, otherwise no information is surpassed.

2) Gradient Boosting (GB): The Gradient Boosting algorithm is commonly used in the field of machine learning and is often used to build prediction and classification models. It aims to build a strong model using a succession of weak models. At each stage, a new model is built by improving the mistakes of the previous model, achieved by training the new model on the errors of the previous model. This procedure helps reduce bias errors in the final model [23] [24].

3) K-Nearest Neighbors (kNN): In k-NN classification, the output is a class membership. An object is classed via a plurality vote of its pals, with the object being assigned to the class most commonplace amongst its okay nearest neighbors. If k = 1, then the item is without a doubt assigned to the class of that single nearest neighbor [24][25].

4) Logistic Regression (LR): This type of statistical version is regularly used for type and predictive analytics. Logistic regression estimates the possibility of an occasion taking place, together with voting or did not vote, based on a given dataset of impartial variables [26] [27] [28]. Because the final result is a possibility, the structured variable is bounded between 0 and 1. In logistic regression, a logit transformation is implemented on the odds this is the chance of fulfillment divided by way of the probability of failure. that is also typically referred to as the log odds or the natural logarithm of odds [26] [27] [28].

5) Random Forest (RF): This classifier is the most popular. The primary dataset is used to construct a subset of random trees. Each tree contains a different set of features and data to predict a decision. In the end, the most common and frequent decision is chosen [24] [26] [29].

III. Research Method and Material

In the method section, we present the data mining phases that we went go through to develop a predictive model for the academic tracks via GPA based on the preparatory year data in Saudi universities. Our research methodology consists of six phases (Fig. 1) as follows:

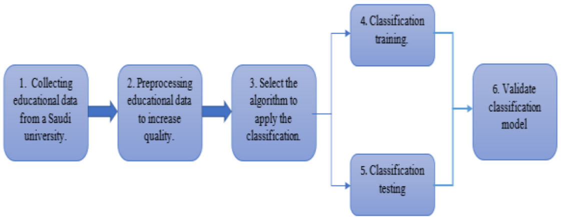
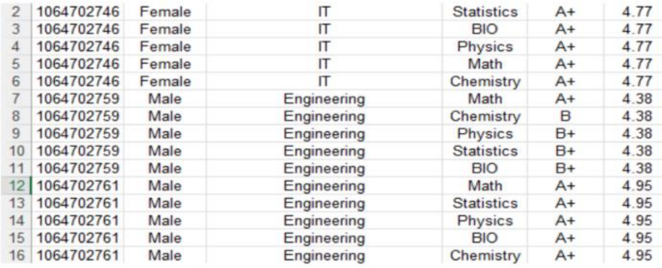

1. Collecting educational data from a Saudi university.
2. Preprocessing educational data to increase quality.
3. Select the algorithm to apply the classification.
4. Classification training.
5. Classification testing
6. Validate classification model

Fig. 1. Research methodology.

A. Data Collection

We obtained the dataset from a Saudi university, which preferred not to be named with taking care of the privacy of students and concealing any data indicating their personality. The dataset contains records of the preparatory year for the scientific subjects (Chemistry, Statistics, Math, Physics and BIO), the final GPA upon graduation from the university and the college to which the student is registered. All these records belong to students graduating from a university in the same year. The structure of the data set was not suitable for the data mining process. The number of records is 12393, and each student had five records in the data set, with each record representing a subject, as shown in the following Fig. 2:



2	1064702746	Female	IT	Statistics	A+	4.77
3	1064702746	Female	IT	BIO	A+	4.77
4	1064702746	Female	IT	Physics	A+	4.77
5	1064702746	Female	IT	Math	A+	4.77
6	1064702746	Female	IT	Chemistry	A+	4.77
7	1064702759	Male	Engineering	Math	A+	4.38
8	1064702759	Male	Engineering	Chemistry	B	4.38
9	1064702759	Male	Engineering	Physics	B+	4.38
10	1064702759	Male	Engineering	Statistics	B+	4.38
11	1064702759	Male	Engineering	BIO	B+	4.38
12	1064702761	Male	Engineering	Math	A+	4.95
13	1064702761	Male	Engineering	Statistics	A+	4.95
14	1064702761	Male	Engineering	Physics	A+	4.95
15	1064702761	Male	Engineering	BIO	A+	4.95
16	1064702761	Male	Engineering	Chemistry	A+	4.95

Fig. 2. Pure dataset.

Students are evaluated in each subject out of one hundred marks distributed as in the following Table I.

TABLE I. STUDENT EVALUATION

Mark	Grade symbol
From 95 to 100	A+
From 90 to less than 95	A
From 85 to less than 90	B+
From 80 to less than 85	B
From 75 to less than 80	C+
From 70 to less than 75	C
From 65 to less than 70	D+
From 60 to less than 65	D

B. Data Pre-Processing

1) Data cleaning and data reduction: During this phase, we made sure that all students took the same courses, so we deleted the data of students who transferred from other universities.

To increase the balance of the dataset and because of the large difference between the number of students graduating from some colleges we have deleted the data of students graduating from the following colleges (Table II):

TABLE II. DELETED COLLEGES

Deleted colleges	Number of students
Arts	23
Law and Political Science	7
Media	4

2) Data transformation: At this stage, the data has been transformed into a format that accepts modeling. The dataset structure for each student was five records. Each record represents a subject. Also, we changed the GPA formula from a numeric to a categorical as follows (Table III):

TABLE III. GPA SYMBOL

GPA	GPA symbol
From 5.00 to 4.50	Excellent
From 4.49 to 3.75	Very_ Good
From 3.74 to 2.75	Good
From 2.74 to 2.00	Pass

After that, we added a new column named (OUTPUT). The students were divided into two values. Any student who achieved a GPA greater than or equal to four will be given in OUTPUT feature a value (RIGHT), and a student with a cumulative GPA of less than four will be given a (WRONG) value in the OUTPUT feature as Table IV.

This procedure helps us to form our hypothesis "When a student achieves a GPA higher than 4.00, then the academic track is correct".

TABLE IV. STUDENT OUTPUT SYMBOL

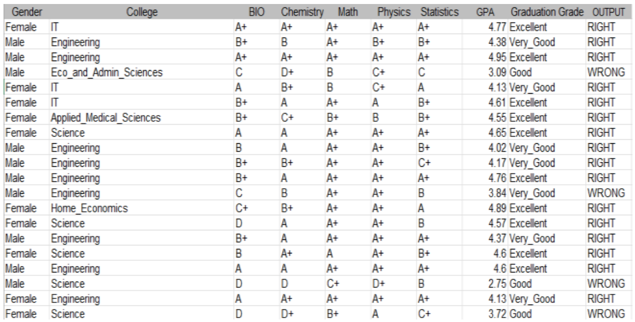
GPA	OUTPUT
From 4.00 to 5	RIGHT
Less than 4.00	WRONG

After completing the data pre-processing stage, we extracted a data set containing 2363 records in Excel format for this study. The features are the following (Table V):

TABLE V. FEATURES ON A DATASET

Features	No. of types	Type
Gender	2	Male, Female
College	12	Applied_Medical_Sciences, Dentistry, Design_and_Built_Environment, Eco_and_Admin_Sciences, Engineering, Home_Economics, Geology, IT, Medicine, Nursing, Pharmacy, Science
BIO	6	A+, A, B+, B, C+, C
Math	6	A+, A, B+, B, C+, C
Chemistry	6	A+, A, B+, B, C+, C
Physics	6	A+, A, B+, B, C+, C
statistics	6	A+, A, B+, B, C+, C
GPA		
Graduation Grade	4	Excellent, Very_Good, Good, Pass
OUTPUT	2	RIGHT, WRONG

Fig. 3 shows the structure of the final dataset.



Gender	College	BIO	Chemistry	Math	Physics	Statistics	GPA	Graduation Grade	OUTPUT
Female	IT	A+	A+	A+	A+	A+	4.77	Excellent	RIGHT
Male	Engineering	B+	B	A+	B+	B+	4.38	Very_Good	RIGHT
Male	Engineering	A+	A+	A+	A+	A+	4.95	Excellent	RIGHT
Male	Eco_and_Admin_Sciences	C	D+	B	C+	C	3.09	Good	WRONG
Female	IT	A	B+	B	C+	A	4.13	Very_Good	RIGHT
Female	IT	B+	A	A+	A	B+	4.61	Excellent	RIGHT
Female	Applied_Medical_Sciences	B+	C+	B+	B	B+	4.55	Excellent	RIGHT
Female	Science	A	A	A+	A+	A+	4.65	Excellent	RIGHT
Male	Engineering	B	A	A+	A+	B+	4.02	Very_Good	RIGHT
Male	Engineering	B+	B+	A+	A+	C+	4.17	Very_Good	RIGHT
Male	Engineering	B+	A	A+	A+	A+	4.76	Excellent	RIGHT
Male	Engineering	C	B	A+	A+	B	3.84	Very_Good	WRONG
Female	Home_Economics	C+	B+	A+	A+	A	4.89	Excellent	RIGHT
Female	Science	D	A	A+	A+	B	4.57	Excellent	RIGHT
Male	Engineering	B+	A	A+	A+	A+	4.37	Very_Good	RIGHT
Female	Science	B	A+	A	A+	B+	4.6	Excellent	RIGHT
Male	Engineering	A	A	A+	A+	A+	4.6	Excellent	RIGHT
Male	Science	D	D	C+	D+	B	2.75	Good	WRONG
Female	Engineering	A	A+	A+	A+	A+	4.13	Very_Good	RIGHT
Female	Science	D	D+	B+	A	C+	3.72	Good	WRONG

Fig. 3. Dataset after pre-processing.

C. Classification Algorithms Selection

After several experiments and an understanding of the characteristics of the classification algorithms, to achieve the best possible results from the data set, the following classification throws were used with default parameters values:

- Neural Network (ANN)
- Gradient Boosting (GB)
- K-Nearest Neighbors (kNN)
- Logistic Regression (LR)

- Random Forest (RF)

D. *Experiments (Training and Testing)*

We chose the Orange Data Mining software to conduct the experiments. Orange data mining is written in Python and is open-source. It was developed at the University of Ljubljana. The program's graphic interface offers an easy experience in handling and ease of learning [30] [31]. It supports several operating systems such as Windows and Linux. It provides the possibility to test algorithms, validation, and prediction.

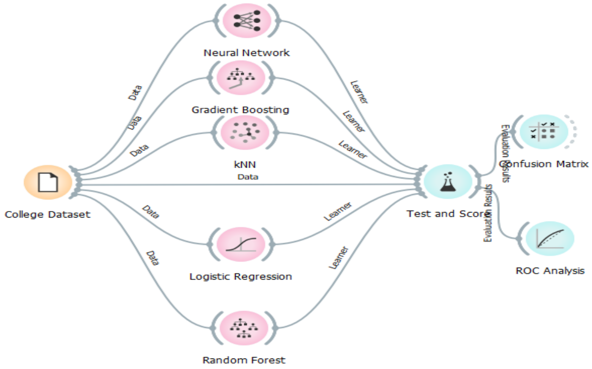
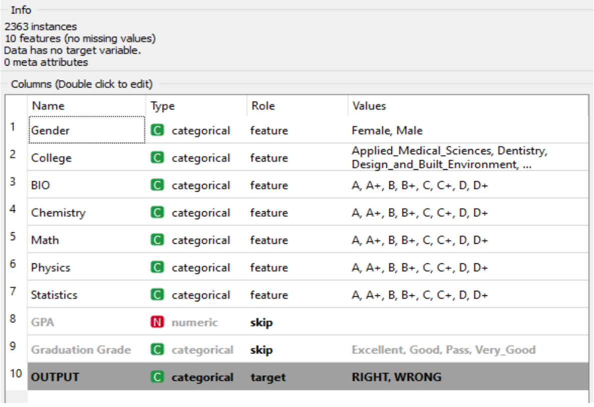


Fig. 4. Orange data mining model.

The data set was used for each student record with nine characteristics. Fig. 4 shows the model created in the orange data mining tool. To explain how the model works, the data set has been loaded and the feature to which each class of data belongs, and which feature of this data is the target as in Fig. 5.



Info

2363 instances
10 features (no missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	Gender	C categorical	feature	Female, Male
2	College	C categorical	feature	Applied_Medical_Sciences, Dentistry, Design_and_Built_Environment, ...
3	BIO	C categorical	feature	A, A+, B, B+, C, C+, D, D+
4	Chemistry	C categorical	feature	A, A+, B, B+, C, C+, D, D+
5	Math	C categorical	feature	A, A+, B, B+, C, C+, D, D+
6	Physics	C categorical	feature	A, A+, B, B+, C, C+, D, D+
7	Statistics	C categorical	feature	A, A+, B, B+, C, C+, D, D+
8	GPA	N numeric	skip	
9	Graduation Grade	C categorical	skip	Excellent, Good, Pass, Very_Good
10	OUTPUT	C categorical	target	RIGHT, WRONG

Fig. 5. Model specifications.

In this model, the OUTPUT was determined as the target and the rest were as features and skipped GPA and Graduation Grade features as in Fig. 3 to 5. Then the data set was linked to the previously selected algorithms widget, as well as to the "Test and Score" widget to display the results. This procedure provides training and testing of all algorithms at the same time which saves a lot of time and effort rather than testing the algorithms individually. CROSS VALIDATION was used to split the data into test and training data. Cross-validation divides the data into several groups called FOLDS. This method splits the dataset randomly into 10 subsets [32]. The model training phase uses nine subsets, while the testing phase uses the final subset. This process is repeated 10 consecutive times each time a different subset is selected in the testing phase [32].

E. *Hypothesis*

When the student graduates with a GPA greater than or equal to (4.00), this means that the academic track chosen by the student is correct and commensurate with the characteristics chosen in the dataset. On the contrary, when a student achieves a GPA less than (4.00), the academic track chosen by the student is wrong.

F. *Validation*

To evaluate the performance of the model, we will rely on the Accuracy, Area under the curve (AUC-ROC), and confusion matrix.

- Area under the curve (AUC-ROC):

Gives an idea of the effectiveness of the model and the AUC score is used to compare the different algorithms. Each classifier will predict either a true or false result. Whenever the AUC value was greater than 0.5 the classifier was able to separate the two results and give a correct result and vice versa if the AUC value were less than 0.5 the classifier would have predicted an opposite outcome. That is, the actual positive is expected to be negative. The use of AUC is used when the data set is unbalanced [33].

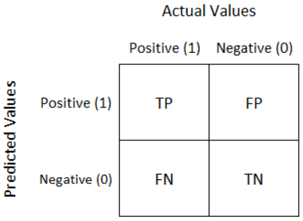
- Accuracy:

In machine learning and data technology, the term accuracy is inevitable almost in every category assignment. this is the most popular measurement or metric used to assess models [31]. We calculate the accuracy by using the equation:

$$\textbf{Accuracy} = TP + TN / TP + TN + FP + FN$$

- Confusion Matrix :

A confusion Matrix is one way to measure the performance of classification models. It can be used when the outputs are two or more classes. The result is an extracted table with four areas. Each area represents the expected and actual value [27] [34] [35] . as shown in Fig.6 are prescribed.



Actual Values

Predicted Values		Positive (1)	Negative (0)
	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 6. Confusion matrix.

IV. Experiments and Results

In this section, we will do the experiments based on the database described in the previous section and using Orange Data Mining Tool. The classification algorithms will be used with default parameters. The algorithm will be evaluated based on the following criteria: accuracy, an area under the curve (AUC-ROC) and a confusion matrix. The goal of the experiments will be to predict the student's GPA based on five subjects studied in the preparatory year by using classification algorithms. To reach the answer to the question, "Is the academic track chosen by the student right or wrong?" The results were as follows in Table VI:

TABLE VI. RESULTS OF THE GPA PREDICTION MODEL

Model	***AUC***	***CA***	***F1***	***precision***	***Recall***	***Specificity***
GP	0.850	0.789	0.789	0.789	0.789	0.788
kNN	0.804	0.749	0.749	0.749	0.749	0.745
LR	0.854	0.791	0.791	0.791	0.791	0.790
ANN	0.846	0.786	0.786	0.786	0.786	0.784
RF	0.817	0.755	0.755	0.755	0.755	0.749

The results of the experiments showed that the LG algorithm provides the best performance with a slight distinction from the GP algorithm. The accuracy and AUC-ROC values for the LG algorithm were 79.1% and 85.4%, respectively. While the GB algorithm attained up to 78.9% accuracy and an AUC-ROC value of 85%. In the same context, the kNN algorithm performed the weakest with an accuracy of up to 75% and an AUC-ROC value of 80.4%. The performance of all algorithms used in the experiment is compared in Fig. 7.

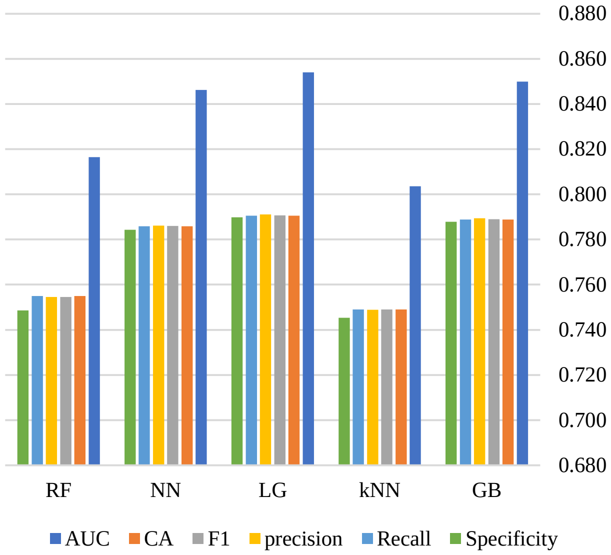


Fig. 7. Comparison performance metrics of algorithms.

The confusion matrix is constructed as shown in Table VII. Where diagonal entries reflect successfully categorized samples, and the remaining entries represent misclassified ones. The results demonstrate that, according to the LG algorithm, actually 1015 predicted the right academic track and 853 the wrong academic track.

TABLE VII. CONFUSION MATRIX FOR ALL ALGORITHMS

LG Algorithm

		Prediction	
		Right	Wrong
Actual	Right	1015	262
	Wrong	233	853

GB Algorithm

		Prediction	
		Right	Wrong
Actual	Right	1014	263
	Wrong	236	850

NN Algorithm

		Prediction	
		Right	Wrong
Actual	Right	1015	262
	Wrong	244	842

kNN Algorithm

		Prediction	
		Right	Wrong
Actual	Right	984	293
	Wrong	300	786

RF Algorithm

		Prediction	
		Right	Wrong
Actual	Right	1001	276
	Wrong	327	759

V. DECISION AND CONCLUSION

The main objective of this research was to develop a model to predict the right academic track Via GPA by using classification algorithms for Saudi university students. Therefore, we used a dataset of Saudi university students containing five scientific subjects studied in the preparatory year in addition to the student's gender, college, and final GPA. We made sure that all students studied the same subjects. Five classification algorithms serve as the basis for the proposed model: gradient boost, kNN, logistic regression, neural network, and random forest. We assumed that when the student achieves a GPA greater than or equal to 4.00 which means the academic track is correct. But if the student achieved a GPA less than 4.00, the student chose the wrong academic track. The results show that the logistic regression algorithm is the most accurate and able to predict correctly. It achieved an accuracy of 79.1% and an AUC of 85.4%. It can be seen that the accuracy of the model is somewhat low. This is due to the small number of features and their confinement to the academic subjects and the gender of the student. Other features can affect the accuracy of the model, such as behavioral characteristics, high school results, and Aptitude and achievement tests. The results justify the validity of the hypothesis that it is possible to predict the academic track based on the GPA where the proposed model was able to predict the final GPA that the student will achieve if he joins a specific academic track. The results of this study are expected to help educational institutions in early intervention to guide students who are struggling to choose the right academic track. Future research can improve accuracy by relying on additional variables including behavioral characteristics, results from aptitude tests, and grades from high school.

REFERENCES

[1] M. Fakhimuddin, U. Khasanah, and R. Trimiyati, "Database Management System in Accounting: Assessing the Role of Internet Service Communication of Accounting System Information," Research Horizon, vol. 1, no. 3, Art. no. 3, Jun. 2021, doi: 10.54518/rh.1.3.2021.100-105.

[2] M. O. Igbinovia and I. J. Ikenwe, "Knowledge management: processes and systems," Information Impact: Journal of Information and Knowledge Management, vol. 8, no. 3, Art. no. 3, 2017, doi: 10.4314/iijikm.v8i3.3.

[3] J. Han, M. Kamber, and J. Pei, "1 - Introduction," in Data Mining (Third Edition), J. Han, M. Kamber, and J. Pei, Eds., in The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 1–38. doi: 10.1016/B978-0-12-381479-1.00001-0.

[4] T. Hastie, R. Tibshirani, and J. Friedman, "Introduction," in The Elements of Statistical Learning: Data Mining, Inference, and Prediction, T. Hastie, R. Tibshirani, and J. Friedman, Eds., in Springer Series in Statistics. New York, NY: Springer, 2009, pp. 1–8. doi: 10.1007/978-0-387-84858-7_1.

[5] C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. d. Baker, Eds., Handbook of Educational Data Mining, 0 ed. CRC Press, 2010. doi: 10.1201/b10274.

[6] "Budget Statement 2023." https://www.mof.gov.sa/en/budget/2023/Pages/default.aspx (accessed Mar. 11, 2023).

[7] D. of A. & Registration, "Preparatory Year." https://admission.kau.edu.sa/Content-210-EN-260921 (accessed Mar. 11, 2023).

[8] H. Brdesee and W. Alsaggaf, "Is There a Real Need for the Preparatory Years in Higher Education? An Educational Data Analysis for College and Future Career Readiness," Social Sciences, vol. 10, no. 10, pp. 1–16, 2021.

[9] M. Foraker, "Does Changing Majors Really Affect the Time to Graduate? The Impact of Changing Majors on Student Retention, Graduation, and Time to Graduate," undefined, 2012, Accessed: Nov. 12, 2021. [Online]. Available: https://www.semanticscholar.org/paper/Does-Changing-Majors-Really-Affect-the-Time-to-The-Foraker/cb8df7853c6937092ec842fdc9f674b5a4767f68

[10] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," Heliyon, vol. 5, no. 2, p. e01250, Feb. 2019, doi: 10.1016/j.heliyon.2019.e01250.

[11] S. L. B. Ginting and M. A. F. Rahman, "DATA MINING, NEURAL NETWORK ALGORITHM TO PREDICT STUDENT'S GRADE POINT AVERAGE: BACKPROPAGATION ALGORITHM," vol. 16, p. 10, 2021.

[12] A. Zollanvari, R. C. Kizilirmak, Y. H. Kho, and D. Hernandez-Torrano, "Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors," IEEE Access, vol. 5, pp. 23792–23802, 2017, doi: 10.1109/ACCESS.2017.2740980.

[13] N. Putpuek, N. Rojanaprasert, K. Atchariyachanvanich, and T. Thamrongthanyawong, "Comparative Study of Prediction Models for Final GPA Score: A Case Study of Rajabhat Rajanagarindra University," in 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), Singapore: IEEE, Jun. 2018, pp. 92–97. doi: 10.1109/ICIS.2018.8466475.

[14] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," IJIMAI, vol. 5, no. 2, p. 26, 2018, doi: 10.9781/ijimai.2018.02.004.

[15] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," Materials Today: Proceedings, p. S221478532105241X, Jul. 2021, doi: 10.1016/j.matpr.2021.07.382.

[16] E. Alyahyan and D. Dusteaor, "Decision Trees for Very Early Prediction of Student's Achievement," in 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia: IEEE, Oct. 2020, pp. 1–7. doi: 10.1109/ICCIS49240.2020.9257646.

[17] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," IJIET, vol. 6, no. 7, pp. 528–533, 2016, doi: 10.7763/IJIET.2016.V6.745.

[18] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques," IJMECS, vol. 8, no. 11, pp. 36–42, Nov. 2016, doi: 10.5815/ijmecs.2016.11.05.

[19] Y. Altujjar, W. Altamimi, I. Al-Turaiki, and M. Al-Razgan, "Predicting Critical Courses Affecting Students Performance: A Case Study," Procedia Computer Science, vol. 82, pp. 65–71, 2016, doi: 10.1016/j.procs.2016.04.010.

[20] H. Almarabeh, "Analysis of Students' Performance by Using Different Data Mining Classifiers," IJMECS, vol. 9, no. 8, pp. 9–15, Aug. 2017, doi: 10.5815/ijmecs.2017.08.02.

[21] K. Mehrotra, C. K. Mohan, and S. Ranka, Elements of artificial neural networks. MIT press, 1997.

[22] L. V. Fausett, Fundamentals of neural networks: architectures, algorithms and applications. Pearson Education India, 2006.

[23] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," Artif Intell Rev, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[24] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," Emerging artificial intelligence applications in computer engineering, vol. 160, no. 1, pp. 3–24, 2007.

[25] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," Applied Computing and Informatics, vol. 12, no. 1, pp. 90–108, Jan. 2016, doi: 10.1016/j.aci.2014.10.001.

[26] C. Zhenhai and L. Wei, "Logistic regression model and its application," Journal of Yanbian University (natural science edition), vol. 38, no. 01, pp. 28–32, 2012.

[27] M.-Y. Yuan, Data mining and machine learning: WEKA application technology and practice. Tsinghua University Press, Beijing, 2014.

[28] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, "Using machine learning to predict physics course outcomes," Phys. Rev. Phys. Educ. Res., vol. 15, no. 2, p. 020120, Aug. 2019, doi: 10.1103/PhysRevPhysEducRes.15.020120.

[29] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Jul. 2013, pp. 1–7. doi: 10.1109/ICCCNT.2013.6726842.

[30] A. Jovic, K. Brkic, and N. Bogunovic, "An overview of free software tools for general data mining," in 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2014, pp. 1112–1117. doi: 10.1109/MIPRO.2014.6859735.

[31] 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, doi: 10.1016/j.procs.2016.05.251.

[32] M. W. Browne, "Cross-Validation Methods," Journal of Mathematical Psychology, vol. 44, no. 1, pp. 108–132, Mar. 2000, doi: 10.1006/jmps.1999.1279.

[33] A. J. Bowers and X. Zhou, "Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes," Journal of Education for Students Placed at Risk (JESPAR), vol. 24, no. 1, pp. 20–46, Jan. 2019, doi: 10.1080/10824669.2018.1523734.

[34] E. Frank and M. A. Hall, Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2011.

[35] O. Caelen, "A Bayesian interpretation of the confusion matrix," Ann Math Artif Intell, vol. 81, no. 3, pp. 429–450, Dec. 2017, doi: 10.1007/s10472-017-9564-8.