SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.04 NO. 3 2025
DOI: 10.51984/SUCP.V4I3.4299

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Confrence Proceeeding homepage: http://www.sebhau.edu.ly/journal/CAS

# Classification Model to Predict the University Major Using Decision Tree Algorithm

*Yahyia BenYahmed[a], Almahdi Alshareef[b], Salah Amar[b], Mohammed Zadana[b]

[a] Department of Computer Science, Faculty of Education, University of Fezzan, Libya,
[b] Department of Computer Science, Faculty of Information Technology, Sebha University, Libya

**Keywords:**
Data analysis
Decision Tree
Intelligent model
Predicting achievement
University Major

**A B S T R A C T**

Data mining education has emerged as a powerful method for predicting students' academic success and uncovering hidden links in educational data. The goal of this paper is to create an intelligent prediction system to assist in guiding and counseling students who were preparing to start college. The main issue addresses in the study is that many students end up specializing in areas they do not desire, often influenced by the advice of friends or family. The performances of Decision Tree (ID3), Naïve Bayes (NB), and k-Nearest Neighbor (kNN) algorithms, among the data mining algorithms, are calculated and compared to predict the final appropriate university major for students. This involves analyzing the data and evaluating the training performance of classifiers in terms of accuracy, precision, sensitivity (recall), and F-measure. The study utilizes actual, private data to construct the system. After undergoing various processing procedures, the student data from Sebha Universities are organized into different categories and prepared for the algorithms to be trained on, with an accuracy rate of 64.28%, the Decision Tree method proves to be the most effective among the three classification algorithms that are tested.

# نموذج تصنيف للتنبؤ بالتخصص الجامعي باستخدام خوارزمية شجرة القرار

يحيى بن يحمد[a]، المهدي الشريف[b]، صالح عمر[b]، محمد زادنا[b]

[a] كلية التربية، جامعة فزان، ليبيا
[b] كلية تقنية المعلومات، جامعة سبها، سبها، ليبيا

**الكلمات المفتاحية:**
تحليل البيانات
شجرة القرار
النموذج الذكي
التنبؤ بالإنجاز
التخصص الجامعي

**الملخص**

برزت تقنية التنقيب في البيانات التعليمية كطريقة فعّالة للتنبؤ بالنجاح الأكاديمي للطلاب وكشف الروابط الخفية في البيانات التعليمية. هدفت هذه الورقة إلى إنشاء نظام تنبؤ ذكي للمساعدة في توجيه وإرشاد الطلاب الذين يستعدون لبدء الدراسة الجامعية .تناولت الدراسة القضية الرئيسية المتمثلة في أن العديد من الطلاب ينتهي بهم الأمر بالتخصص في مجالات لا يرغبون فيها، متأثرين غالبًا بنصائح الأصدقاء أو العائلة. تم حساب أداء خوارزميات شجرة القرار، وخوارزمية بايز الساذجة، وخوارزمية "k" لأقرب جار، من بين خوارزميات التنقيب في البيانات، ومقارنتها للتنبؤ بالتخصص الجامعي المناسب للطلاب. تضمن ذلك تحليل البيانات وتقييم أداء تدريب المصنفات من حيث الدقة والإحكام والحساسية (التذكر) ومقياس-F. استخدمت الدراسة بيانات فعلية خاصة لبناء النظام. بعد الخضوع لإجراءات معالجة مختلفة، تم تنظيم بيانات الطلاب من جامعات سبها في فئات مختلفة وإعدادها لتدريب الخوارزميات عليها، وبمعدل دقة بلغ 64.28%، أثبتت طريقة شجرة القرار أنها الأكثر فعالية بين خوارزميات التصنيف الثلاث التي تم اختبارها.

## 1. Introduction

Data mining is the process of collecting meaningful information from massive databases in order to uncover patterns, trends, and insights for higher education [1]. Numerous facets of students' academic performance and university results have been predicted via the application of machine learning algorithms and educational data mining. Numerous studies have shown how well data mining works to predict students' academic success, degree completion rates, and suitable university majors. A machine learning algorithm-based model was presented by [2] to predict undergraduate students' final exam results. Educational data mining has been used to create information management systems at colleges and universities, offering effective ways to evaluate student performance and highlight student achievements [3]. The DM method and association rule application research are introduced in [4] for the purpose of developing IMS in colleges and universities. Additionally, the data indicated that students who select one course also typically select the other course. In [5] developed a hybrid model to predict a good fit in major for students to decrease dropout risk, utilizing machine learning algorithms. However, from prediction students' academic performance and degree completion to predicting the right university major, these studies show

*Corresponding author:
E-mail addresses: yah.benyehmed@fezzanu.edu.ly, (A. Alshareef) alm.alshareef@sebhau.edu.ly
*Article History : Received 20 February 2025 - Received in revised form 01 September 2025 - Accepted 07 October 2025*

the promise of data mining and machine learning algorithms.

Assisting students in choosing suitable university major and forecasting academic achievement, intelligent models especially those grounded in machine learning and artificial intelligence are being employed more often [6]. For prediction purposes, these models are intended to give instructors and students insightful information that will assist them in making well-informed decisions on possible majors and academic pathways. It's crucial to remember, though, that the use of these models has sparked questions regarding bias and the possibility of inaccurate predictions[7]. The ethical and just application of the predictions these intelligent models produce must thus be carefully considered, even if they may be very useful tools. However, intelligent models that draw from artificial intelligence and machine learning have demonstrated potential for both academic success prediction and helping students choose the right university major. If these models are to direct students' educational pathways, it is imperative that potential biases and ethical issues be taken into account[8, 9].

The Decision Tree algorithm (DT) is a flexible supervised machine-learning technique that may be used to solve classification and regression issues. It is a common data mining approach for creating classification models with a tree-like topology [10]. One clever approach to predict the right university major is to build one using a decision tree algorithm. Simple to comprehend, analyze, and visually represent, decision tree algorithms provide categorization and prediction models [11]. Through the creation of models based on training data, the algorithm may be used to predict academic performance and accomplishment, including finished GPA [12] . For the purpose of solving regression and classification issues, decision trees are among the most widely used and straightforward classification methods [13]. DT algorithms are being utilized by researchers to forecast and identify the elements that influence the learning performance of first-year college students [14].

University major means specifying a path for a specific subject in any life science, in which students can specialize or resort to while looking to obtain a university degree. Choosing a field of study is a difficult decision, as most university graduates change their major if given the opportunity to return to the stage of choosing a university major. Due to the availability of many specializations and fields, it has become difficult for students to choose the appropriate specialization for them, so the idea of the need to build a specific system that predicts the appropriate specialization came up [15, 16]. In light of this, many universities or higher education institutions have updated their acceptance criteria for students and started introducing prediction techniques for the field of specialization [17].

Given these challenges facing the student, the need has led to the use of different technologies or systems, meaning comprehensive theories, to provide a comprehensive understanding of the student’s choice of the appropriate major for him. This study will rely on the use of the DT algorithm to direct students to the appropriate university field, compare them, and rely on the best. DT has shown promise in predicting academic performance and assisting students in selecting appropriate university majors. However, it's crucial to address potential biases and ethical considerations when using these models to guide students' educational paths.

The paper is organized as follows. The literature-related work is reviewed in the section that follows. The study methodology and dataset description are presented in the third part, which also looks at the algorithms employed in the tests. Next, we'll go over the findings, analysis, and research contributions. We wrap up the analysis by outlining the limits of the research and suggesting further lines of inquiry.

## 2. Related works

Various data mining approaches have been used to predict a student's appropriate university major and predict student achievement. Algorithms used to predict final grades by searching for patterns in students' study-related data and social behavior traits. The findings revealed that students' social behavior features improved prediction for one-quarter of the courses. To predict student performance, the method employed collaborative filtering techniques. Because there is no prior knowledge about students' knowledge, abilities, or excitement for certain courses, early grade prediction is more challenging. Still, evidence has shown that the prediction is enhanced by information regarding students' activities during the course of the semester [18].

### 2.1 The Learning Performance of the Proposed Recommendation System

Numerous mechanisms and case studies have been proposed to advise and guide students in choosing their college majors, which can be a crucial and confusing decision in their lives. The mechanisms aim to bridge the gap between academia and industry by providing different perspectives [19].

According to [20], a variety of explainable machine learning approaches were tested to predict students' right undergraduate major (field of specialization) before admission at the undergraduate level, including DT, Extra tree classifiers ETC, Random forest RF classifiers, Gradient boosting classifiers GBC, and Support Vector Machine (SVM). The findings indicate that grades in high school and university, as well as admission examinations, are excellent criteria for recommending the best undergraduate major since these input qualities most reliably predict the student's field of expertise. In addition, the ineffective educational system and parental decision-making leave students unsure about their major or line of career. Classifiers can help solve this issue in part [21]. Classifiers can be used to assess the correctness of the decisions that the students will make. These classifiers, which include the DT and Random Forest classifiers, aid in determining the prediction's accuracy. The DT classifier predicts the right answer with greater accuracy. Students will be able to make informed decisions about their studies and career paths with the aid of this classifier. A DNN-based career track recommender system was presented in a different research by [22] to help guidance counselors help their students choose a career path that suits them and includes 1500 students. The academic strand in senior high school was predicted with an accuracy rate of 83.11%, the DNN algorithm's prediction of students' academic strand works rather well. However, merely by utilizing the suggested approach, guidance counselors were able to handle kids' issues more effectively.

Recently, the study of [21] recommended the best college major to provide the best programs for students, Fuzzy Logic is utilized to create a recommendation system that compares university academic offerings with the set of student requirements. The end-of-work experimental test, which gauges the degree of correctness of the task, yields encouraging findings, showing a respectable number of participants in the right main selection. In [23], the researchers explored using Hebron University as a case study to construct an expert system to aid high school students in selecting the proper university major. The system analyzes the student's qualities and talents and displays the finest university programs that suit his personality using both rules and machine learning algorithms. According to the statistics, 85% of the students were happy with the system and would suggest it to others. Furthermore, the work published in [24] offered a data mining technique based on rough set theory for students' college major selection. Using data mining techniques, this study tries to identify the main elements that impact high school students' major decision. A questionnaire was created to collect information from students at several institutions. The study's findings indicated that estimated reductions had a considerable effect on students' choice of institution and college major. This research assists students in not constantly changing their major due to a poor choice of major, which leads to unhappiness with their major.

Additional comparison study is being undertaken to evaluate five approaches of recommender systems for university study field and career domain recommendations [25]. The proposed methodologies took into account user-based and item-based collaborative filtering, demographic-based recommendation, case-based reasoning, and ontology. To assess the efficacy and efficiency of the deployed procedures, a case study of Lebanese high school pupils is examined. The results show the suggested hybrid strategy received an average of 95% usability feedback and 92.5% student satisfaction, indicating that it may be a viable solution for similar issues in any application domain.

### 2.2 Intelligent Techniques for the Proposed Appropriate University Major Students

Numerous studies have shown how well data mining works to predict students' academic success, degree completion rates, and suitable

university majors [25-27]. The following are a few similar studies using data mining techniques to predict suitable university majors.

**2.2.1 Decision Trees (DT)**

Decision trees (DT) are a sort of machine learning algorithm that has been utilized in numerous research to predict academic results and aid in university key decision-making. For example, one study [12] offered decision tree-based prediction models for academic success utilizing college students' support networks. During the COVID-19 epidemic, 484 students enrolled at a big public institution in the United States provided data. A research used the Chi-Square Automatic Interaction Detection (CHAID) and Cforest algorithms to predict students' academic success, as evaluated by their self-reported GPA. The study compares single-tree and random forest models for predicting academic achievement. Each algorithm discovered distinct characteristics significant for different student demographics, with some overlap depending on accuracy and variance. Decision trees algorithms that classified successful and failed students were created in [28, 29] using the CHAID and CART algorithms applied to the student enrollment data of information system students at the Open Polytechnic of New Zealand. With CHAID, the accuracy was 60.5%, while with CART, it was 59.4%. In one study, the learning activities of students were predicted using the k-means clustering method. Students and teachers may find use for the information produced by applying data mining techniques. According to the authors in [30], DT algorithms can represent the grade prediction problem by taking into account the courses that students must complete in order to earn their degree. The research carried out tests using both public and institute data sets. A typical Chinese higher education system provided the educational dataset of 1,325 students and 832 courses. DT, such as ID3, CART, and C4.5, have demonstrated improved precision in forecasting student achievement.

DT were used in another study [31] to predict STEM community college students' degree completion within three years. A total of 283 students' data, comprising 14 variables such as age, gender, degree, and college GPA, were used to create the model. The findings provide crucial information on how to create a student support system that is more effective and responsive. Moreover, recommendation algorithms that choose the best university major based on student data have been developed using decision trees. The research employed the Decision Trees method to forecast the suitable major by utilizing input factors associated with data and MBA students' experiences. K-fold cross-validation was used for the system's validation. The outcomes demonstrated that the random forest outperformed the other categorization methods, with an accuracy of 97.70% as opposed to the published research's 75.00%. The degree percentage, MBA percentage, and entrance exam result were identified to be the most important elements that contributed to the model [26]. In a different study, learning performance prediction models of freshmen are constructed using family background variables that can be acquired before the semester begins. Random Forests, Decision Trees (C5.0), CART, and multilayer perceptrons (MLP) algorithms are used to predict and identify the factors influencing the learning performance of first-year university students. An actual sample of 2407 freshman from 12 departments at a Taiwan vocational institution will be used. CART beats other algorithms, according on the results [32]. Furthermore, utilizing Random Forests, the Decision Tree method has been published for predicting university students' academic achievement and major. The classification tree approach is applied over a ten-year period, data from every undergraduate course taken by a student at a major institution in Canada is analyzed. Two reliable classifiers and a variable significance analysis give important information to university administrators [33].

**2.2.2 Naive Bayes (NB)**

Bayesian learning is a sort of machine learning algorithm that uses Bayes' theorem to update the probability of a hypothesis as new evidence becomes available. It is employed in educational data mining for categorization and prediction purposes [3].

Bayesian Networks can be used to predict academic success [34]. The authors did a systematic evaluation and comparative study of educational data mining strategies for predicting student performance. The study attempted to reformulate the problem into clustering, classification, and regression problems, as well as examine alternative techniques, such as DT and Bayesian networks.

Based on a student's characteristics and performance, the NB technique is a probabilistic classifier that may be used to predict the right university major. NB has been used successfully in several research to predict academic achievement and student graduation [35-37]. To ascertain whether a student would graduate on time or be a late pass, the technique examines a number of variables, including gender, high school origin, entrance exam results, groups, scholarships, and GPA [38]. In predicting the graduation of Information Systems Study Programs at the Faculty of Computer Science, Sriwijaya University, the NB classifier approach obtained an accuracy of 97.6378%, an error rate of 2.3622%, a precision of 90%, and a recall of 100%. In [39] focused on using the NB algorithm to predict student graduation by creating a model that can identify and predict students who are likely to graduate late, allowing institutions to develop and implement appropriate retention and remediation plans. For both late pass and on-time graduating students, the results yielded accurate predictions.

Other study suggested a framework for utilizing NB and additional classification methods like Decision Tree and Rule-Based classification to predict first-year bachelor students' academic achievement in a Computer Science course. The information used in this study was acquired for the academic year 2012–2016 from Bulacan University in the Philippines through the IT program. This study will be very important and helpful, especially to university officials, as it will provide them a way to determine which students, depending on model-included characteristics, will finish college [40, 41].

**2.2.3 k-Nearest Neighbors (kNN)**

k-Nearest Neighbors (kNN) technique is one of the machine learning algorithms used in a study to predict students' academic achievement. It has a very high accuracy rate and can be used to predict students' final test marks [3]. The kNN technique was one of the appropriate classifiers for predicting student undergraduate major in a study that employed supervised learning approaches to discover high school students and university students regarding their grades, extra-curriculum activity and university CGPA [42]. Another research employed the KNN algorithm to predict students' academic achievement and suggested two similarity measures to cope with category factors without turning them into numerical data. It demonstrates that the uggested approach works better than the conventional one, which computes the distance between the nominal variables by requiring the encoding of nominal variables. According to the results, the suggested method performs 14% more accurately than the baseline and is not affected by outliers [43]. In a related research, the KNN technique demonstrated precision values ranging from 76 to 82% when analyzing the students' year of study, and it was shown to be the best model for predicting academic success for each semester when compared with four other algorithms [44]. Based on input data pertaining to the student's academic background and the labor market, the study employed kNN to predict the relevant major. K-fold cross-validation was used for the system's validation.

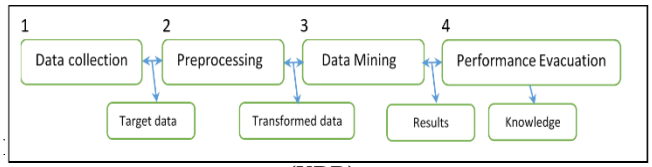
Additional similar studies that use a non-extensive data collection and only include characteristics that are simple to gather at the start of a study program to predict the academic achievement of master's students. The study gathered over 700 student records from the University of Mannheim's Faculty of Business Informatics and Mathematics in Germany between 2010 and 2018 Alshammari et al. [45]. The study assessed how well kNN with other five algorithms predicted academic achievement of students. It also investigated the use of SMOTE to enhance predictions and address unbalanced datasets. The prediction models' results show that a limited number of characteristics may be used to accurately predict academic success. Moreover, comparisons of k-Nearest Neighbor with other algorithms performances were made in research [46]. The College of IT at the University of Basra released datasets from its bachelor's degree programs in 2017–2019 to forecast students' success on their final exams. This study shows that machine learning models can accurately predict academic outcomes, which gives schools useful information on how to find at-risk pupils and customize treatments.

These studies show how data mining techniques may be used to predict student performance and the appropriateness of a university

major. In order to assist students in making educated decisions, these strategies are used to assess educational data, forecast student performance, pinpoint the elements that influence academic achievement, and suggest suitable courses, college majors, or future occupations.

## 3. Methodology

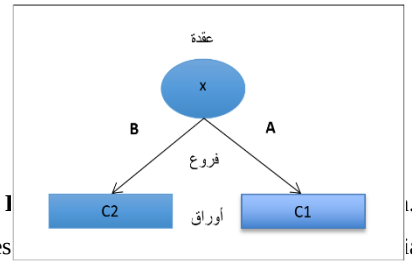
The methodology of this study has been conducted using Knowledge Discovery for Databases (KDD) approach. KDD process is defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [47]. The steps of KDD process are illustrated in Fig. 1.




(KDD).

### 3.1 Proposed mining approach

It is one of the classification approaches based on the construction of a tree structure to represent the rules collected from the classification process. It is made up of a set of nodes that represent test points, which results in branches that indicate the course of the test result, which leads to leaves that represent the conclusion at the end (Karitey et al, 2020). The components of the decision tree are depicted in Fig.2.




The test nodes ia the branches. Rather, the ensuing branches are linked to additional test nodes and finally finish in leaves. The top node is stated as a parent node of the lower node, whereas the lower nodes are child nodes of the higher node. The decision tree is generated from the data table that comprises Because there are several independent variables Xi and one variable is the objective function [48], each independent variable can have a test node, as illustrated in Fig. 3.

‘

**Fig. 3:** How the decision tree works**.**

This means that the test nodes in the decision tree represent the fields in the database, that is, the independent variables, and the branches that emerge from these nodes represent the value contained in these fields, and this value is the test node so that the data set is divided according to one of the categories in the objective function. into several groups, and at the end of the tree at the edges of the branches there are leaves, that is, the categories in the objective function. The tree classifies by arranging from top to bottom and starts from a single node called the wall node. It forms paths to other nodes by dividing The data is divided into parts iteratively until reaching the leaves, thus forming several paths, and each path is interpreted as a rule to reach a specific class in the objective function, and thus reach the desired decision, and this means the decision tree [48]. It consists of the following:

**a. Root node:** It is a single node that represents a parent node for all the nodes in the complete decision tree, and it does not have any branches entering and leaving branches that are linked either to an input node to form child nodes for the root node or directly linked to a leaf, and it expresses the most important independent variables that divide the data set by categories.

**b. Input nodes:** It is a group of other nodes formed in the tree, and each node has a branch inside either from the root node or from other input nodes that form a parent node and branches outside that are linked to other income nodes that form child nodes for this node or are linked to a leaf, and at any node from these nodes, a subtree is formed with child nodes and leaves at lower levels branching off from it, and this node expresses the independent variables in the data set.

**c. Leaves:** These leaves express the end of the path in the tree and have a branch entering from an input node or directly from the root node, and no branches emerge from it. They express the classes in the objective function and thus the final decision.

***Bayesian Classifier Algorithm***

It is a method that requires less amount of data preparation to calculate variables. It is well structured to deal with both real and distinct data. Stratified conditional independence is evident between subsets of variables in the Bayesian classifier. It provides a graphical model of the causal relationship to the performance of the learning process. The Bayesian classifier relies on the probability rule. Bayesian conditional, which is a technique for estimating the probability of the existence of a property given a data set as evidence or input, and is called Bayes' rule, theory, or theorem [49].

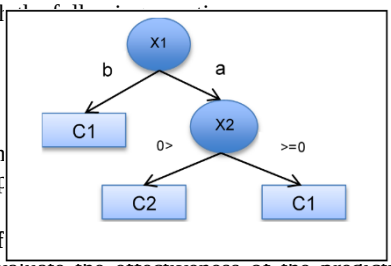
$$p(A\backslash B) = \frac{P(B\backslash A)P(A)}{P(B)} \quad (1)$$

It is a text of what is usually known as Bayes' theorem.

It reads: The conditional probability of event A knowing event B is equal to the probability of B knowing A multiplied by the probability of A divided by the probability of B. For example, when students in a college are classified into different divisions internally, on the condition of obtaining certain grades in some courses.

***kNN algorithm***

It is the most effective directed (supervised) machine learning algorithm. The kNN algorithm is considered one of the predictive and descriptive classification algorithms. This algorithm classifies the data to find the nearest neighbor easily and effectively. The algorithm relies in its work on measuring the Euclidean distance between each point and the point closest to it, and when the data is close to In some cases, the Euclidean distance is very small between each point and its neighboring point, but as the data values diverge, the distance between the points becomes very large, hence the title of the algorithm, as the letter k indicates the cases that will be classified based on the distances between them, that is, between their neighbor [50]. The distance is calculated with




(2)

whereas:

d(i,j): represen

$x_i$, $x_j$: They rep

### 3.2 Methods f

In order to evaluate the effectiveness of the prediction model and algorithms, the predicted values must be calculated compared with the actual values. If the values are equal by a large percentage, the mean of a good prediction model is that there are multiple criteria for the effectiveness of the prediction. The table shows the possible outcomes of predicting binary values [51] . Table (1) shows the evaluation of forecast effectiveness.

**Table 1:** The evaluation of forecast effectiveness.

| | Predicted as True | Predicted as False |
|---|---|---|
| Actually True | True Positive | False Negative |
| Actually False | False Positive | True Negative |

A table that evaluates the effectiveness of prediction. The matrix that displays the possible prediction results is called the confusion matrix [51]. There are different evaluation criteria that can be obtained from these values, one of which is accuracy.

In this experiment, the appropriate university major data is fed into a predictor that uses DT3, NB, and kNN algorithms to predict the data.

The Confusion Matrix is used to assess the performance of the predictors. Researchers utilize a range of performance criteria to evaluate the effectiveness of machine learning algorithms. The Confusion matrix is used to assess classifier performance. In the confusion matrix, the total of the diagonal elements is referred to as correctly categorized cases, while the remainder are referred to as wrongly classified instances. We employed the Accuracy (Acc), Precision (Prec), Recall (Sens), F-measure, and training duration performance indicators to assess and compare the performance of presented prediction models.

Accuracy is calculated by dividing the number of test records by the number of successfully classified records. The percentage of True Positive (TP) records to the total number of True Positive (TP) records in a certain class is called precision. There are two types of recall: true positives and false negatives. The total number of records properly categorized to the total number of records in a class is known as the recall ratio (FN). The most important and commonly used factor to measure the performance of the classifier is accuracy. Accuracy (Acc) is calculated by the ratio of correct prediction samples to the total samples in the dataset. Equation (1) defined Acc as:

By dividing the total number of test records by the total number of successfully categorized records, accuracy is calculated. Precision is defined as the ratio of TP records to all TP records in a given class. Recall comes in two flavors: true positives and false negatives. The recall ratio is the proportion of records that were correctly classified to all the records in a class (FN). Accuracy is the most significant and frequently used factor to assess the performance of the classifier. The ratio of valid prediction samples to all samples in the dataset is used to compute Accuracy (Acc). According to Equation (3), Acc is:

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

The following formulas were used to compute the precision:

$$Precision\ (Prec) = \frac{TP}{TP + FP} \quad (4)$$

The ratio of real projected positive samples to all positive samples is known as Sensitivity (Sens) or recall. However, the ratio of genuine projected negative samples to all negative samples is known as selectivity. Equation (5), respectively, stand for Sens.

$$Sensitivity\ or\ Recall\ (Sens) = \frac{TP}{TP + FN} \quad (5)$$

The cyclical mean between recall and precision was defined by the F-measure. A model is deemed successful if its value is one, while a value of 0 indicates ineffective performance. The following is the F-measure equation.

$$F - measure = \frac{2TP}{2TP + FP + FN} \quad (6)$$

### 3.3 Data Collection

In this study, we relied on ready-made data stored in the database of the Information Development Center for students who enrolled at Sebha University from the year 2017 to the year 2021. The database contains information for 2693 students, and contains 22 characteristics in the various colleges of the university, in addition to that it was used. With a questionnaire specifically for students and the purpose for which it was conducted, to know the reasons that led the student to choose this specialization. This questionnaire was conducted by (Amel Albohali, Almahdi Alshareef, Hassan Al Gaddafi, 2021), from which we chose some questions to strengthen the characteristics of databases.

### 3.4 Data Preprocessing

***Missing Values Processing***

Loss of data is a major problem for the researcher, and not treating it appropriately may cause the researcher some problems, such as reducing the sample size to an inappropriate size, or obtaining poor results. There are several methods used to treat missing values. We have used methods that are compatible with the types of data in the study's database, which are a most frequent substitution method and filling missing values with arithmetic mean method have applied into the data to solve missing values.

***Data Merging***

Merging data from several sources helps in condensing information and avoiding excessive repetitions or contradictions, which contributes to increasing the efficiency, speed, and accuracy of analysis or mining procedures. Heterogeneity in the data structure poses a major challenge to data merging operations. In the end, a complete database with a unified structure and organized data is accurately distributed in its appropriate places.

Accordingly, the process of merging student data for the College of Medicine, the College of Arts, and the College of Education in all its specializations was carried out, so that the data is in one class dedicated to the College of Medicine, one class dedicated to the College of Arts, and one class dedicated to the College of Education, meaning that the class contains the data of students in these colleges, i.e. college Nursing, the College of Pharmacy, the College of Dentistry, and the College of Human Medicine are included in a class designated for the College of Medicine or the College of Medical Sciences with all its branches and specializations, and the College of Arts with all its specialties, which also includes the College of Law. The records of students from those colleges have been merged into one class designated for the College of Arts, and the College of Education with all its existing branches. At Sebha University, its records were combined into one class dedicated to the College of Education.

***Features selection***

In this step, the features of interest and influence that affect the dependent variable were selected, which are gender, secondary specialization, year of enrollment, residential address, marital status, and secondary school. Eleven questions were selected from the total questionnaire questions, which the researchers considered to have a greater impact than the total number of questions. Some characteristics that have no effect on the student's choice of major were excluded and removed, such as: place of birth, academic number, religion, and several questions from the questionnaire used to prepare the database, which do not have a significant impact on the choice of major.

***Data Changing***

In this process, converting and changing data is one of the important stages in preparing data for analysis and mining. This is because algorithms understand numbers, not texts. We need to convert each text category into numbers so that the algorithms can process it using mathematical equations. There are several ways to change categorical data. In this study, we have applied the Label Encoding method. In this method, a unique integer is assigned to each category in the column, such as: In the Gender column, the value of male is replaced with the number (1), and the female is replaced with the number (2), and this is how it is dealt with. With the rest of the columns contained in the database

The Fig. 4 describes the data before processing and Fig. 5 describes the data after processing.

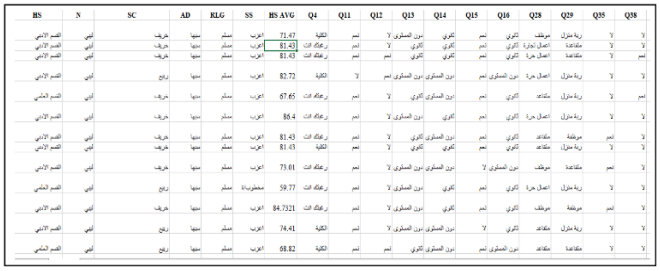


**Fig. 4:** The data before processing.

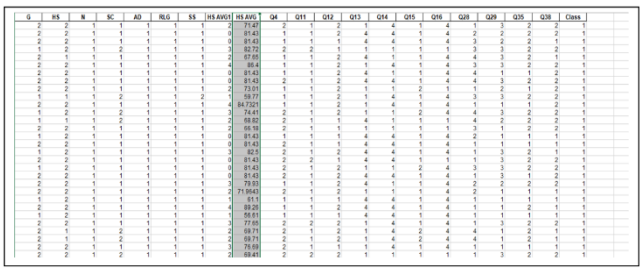


**Fig. 5:** The data after processing.

## 4. Training and Test data

After completing the data processing steps, the data is ready to be entered into the classification algorithms. This data consists of 2632 records distributed across the seven colleges. Fig. 6 shows the number of students in each college. With regard to building the model, the percentage split method was followed. The training and testing process, which divides the data according to a ratio determined by the researcher during the research process, where it is trained and then tested and the final model is given. We have divided the data into 80% training data and the remaining 20% of the data to test the model's performance.


**Fig. 6:** The number of students from each college stored in the database.

We have noticed from the figure that there is a large discrepancy in the number of students in each college, as the number of students is as follows: (Sciences 923 students, Information Technology 429 students, Engineering 333 students, Education 299 students, Arts 266 students, Medicine 244 students, Economics (Commerce and Science). Politics (138 students), this discrepancy may cause a defect in the training of the model and the accuracy of its results, as it is processed to produce good and accurate results.

## 5. Experiment and results

In the study, we have relied on the use of a decision tree to predict the choice of the major closest to a student based on his grades and some personal information, and this algorithm was compared to some classification algorithms, such as the NB algorithm and the nearest neighbor algorithm, in order to come up with the best results and prove the validity of some previous studies (Karitey, et al. ,al,2020), (Mohamed Hegazy & Hoda, 2020), in which the DT3 algorithm outperformed the rest of the classification algorithms. Prediction models were created using the most common learning tool Anaconda and relying on the Jupyter Notebook terminal interface found in this tool, because it is easy to use. It provides graphical interfaces that are easy to handle.

Given that the nature of the data is unbalanced, as we indicated in the previous paragraph 4, two experiments were conducted on the data. The first experiment in which the previous algorithms were applied to the data with its initial distribution, and in the second experiment the data was balanced and made equal in all (Class) colleges, through Under sampling method: This method stipulates equality between all categories by reducing the large categories and equating them to the minority so that they become balanced in all categories. They have equalized and a random sample of 100 records from each category was taken from each category, in order to enhance the accuracy of the model.

**Table 2:** The results of the three algorithms in the first experiment.

| Algorithms | Accuracy | Precision | Recall | f-Measure |
|---|---|---|---|---|
| ID3 | %61.85 | %65 | %66 | %65 |
| NB | %10.05 | %15 | %16 | %3 |
| kNN | %55.97 | %57 | %60 | %58 |

We have noted from Table (2) that the DT3 algorithm is the best algorithm compared to the rest of the algorithms. By all measures used, the accuracy rate of the DT3 algorithm is 61.85% and the kNN algorithm is 55.97%. The Naive Bayes algorithm showed a weak result compared to the two algorithms. This decline in this algorithm may be due to a natural result. Data, or its number and dimensions, or it may be due to the fact that the data was unbalanced, and the result of the current study agreed with the study (Karitey, et, al, 2020), in terms of the superiority of the decision tree algorithm over the rest of the other algorithms. This is a result that we cannot rely on from Where the accuracy measure, due to the unbalanced nature of the data, as a result we have improved the performance of the results by equalizing all categories, to extract a reliable accuracy result, as equality between characteristics was done by Under Sampling, this technique aims to reduce the large categories and equate them with the small ones. This is done by taking a random sample from all categories equally. In this study, equality was achieved by taking 100 student records for each of all categories at random. Fig. 6 shows the categories after the balancing process.

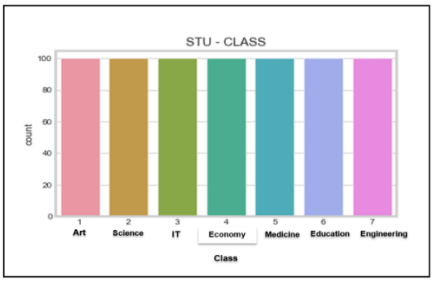



**Fig. 6:** The number of students from each college in the database after balancing.

Table (3) shows the results of the algorithms after the process of balancing the categories. It was shown that even in this experiment, even after the process of balancing the categories, the best algorithm was the DT3 algorithm with an accuracy rate of 64.28%, which increased by three percentages compared to the previous experiment in which the accuracy percentage in the decision tree algorithm was 61.85. %, we notice that there is not a big difference between the two experiments in this algorithm, but the result that this experiment showed in terms of the accuracy measure can be relied upon, because the categories in it were equal, unlike the first experiment, and we also notice a very large increase in the NB algorithm from In the first experiment, the accuracy rate was 10.05%, and in this experiment the accuracy rate was 60.71%, a big difference between the two experiments, and this shows that the NB algorithm does not work correctly when the classes are multiple and unbalanced.

**Table 3:** Algorithms results after the category balancing process.

| Algorithms | Accuracy | Precision | Recall | f-measure |
|---|---|---|---|---|
| ID3 | 64.28% | 36% | 66% | 63% |
| NB | 60.71% | 58% | 67% | 53% |
| KNN | 49.28% | 49% | 47% | 46% |

### *Confusion Matrix*

When the Confusion Matrix algorithm has applied, the classified probabilities of the DT3 algorithm were determined. The probabilities predicted correctly by the classifier were: in the first category (Faculty of Arts), the test sample reached 26 records. The model correctly predicted 22 records, and in the second category (Faculty of Arts), the test sample reached 26 records. The test sample contained 23 records. The model correctly predicted 16 records. In the third category (total), the test sample included 21 records, of which the model correctly predicted 10 records. In the fourth category (total), the test sample contained 31 records, the model correctly predicted 15 records. In the fifth category (college), the test sample contained 15 records. The model correctly predicted only 7 records. In the sixth category (college), the test sample contained 13 records. The model correctly predicted 6 records. In the seventh category (college), the test sample contained 11. Record The model predicted 10 records correctly. Fig. 7 depicts the confusion matrix of the decision tree algorithm prediction model for second experiment.

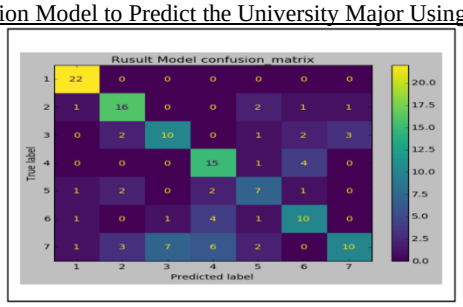



**Fig. 7:** The confusion matrix of the DT3 algorithm result.

***Testing the system on a real sample of current Sebha University students and comparing it to the system's outputs***

The data of a number of students have taken, and they numbered twenty students from random colleges. These students' real desire was to enter this major. Most of them the system predicted their major correctly, some of them incorrectly, and some of them were an approximation, meaning that the student's current major is, for example, information technology. The system has shown it as Sciences. The following table shows the test results.

**Table 4:** The results of the test samples.

| Sample | Insert | DT classification result |
|---|---|---|
| 1 | Female, scientific, Libyan, Spring, Sabha, single, 78%, my desire, no, no, yes, secondary, middle school, yes, university graduate, employee, housewife, yes | Economy (positive) |
| 2 | Male, academic, Libyan, Spring, Sabha, single, 86, my desire, no, no, university graduate, high school, yes, graduate, employee, female employee, yes | Engineering (negative) |
| 3 | Male, academic, Libyan, spring, Sabha, suitor, 80%, my desire, no, yes, no, graduate, female graduate, no, graduate, employee, housewife, no | Economy (positive) |
| 4 | Male, academic, Libyan, fall, Sabha, single, very good, my desire, no, yes, yes, high, university graduate, no, university graduate, employee, female employee, yes | Economy (positive) |
| 5 | Male, scientific, Libyan, Spring, Sebha, single, 73.7%, my desire, no, no, yes, university graduate, middle school, yes, university graduate, self-employed, housewife, yes | Economy (positive) |
| 6 | Male, literary, Libyan, fall, Sabha, single, 80.35, my desire, yes, yes, no, graduate, university, high school, no, none, employee, female employee, yes | Arts (positive) |
| 7 | Male, literary, Libyan, fall, Sabha, single, 80.35, my desire, yes, yes, no, graduate, university, high school, no, none, employee, female employee, yes | IT (positive) |
| 8 | Male, academic, Libyan, Spring, Sabha, single, 89, my desire, yes, no, yes, graduate, graduate, yes, graduates, employee, female employee, yes | Science (negative) |
| 9 | Male, academic, other, spring, Sabha, single, 91, my desire, yes, yes, no, university graduate, primary school, yes, graduate, employee, housewife, no | IT (positive) |
| 10 | Male, academic, Libyan, fall, Sabha, 84, my desire, yes, yes, no, university graduate, high school, yes, university graduate, employee, female employee, yes | IT (positive) |
| 11 | Male, academic, Libyan, Spring, Sabha, 82, my desire, yes, no, no, university graduate, middle school, no, none, employee, housewife, yes | Science (negative) |
| 12 | Female, academic, Libyan, fall, Sabha, 80, my desire, yes, no, no, university graduate, high school, yes, university graduate, self-employed, employee, yes | IT (positive) |
| 13 | Male, academic, Libyan, Spring, Sabha, 78, my desire, yes, no, no, secondary, primary, yes, university graduate, employee, housewife, no | IT (positive) |
| 14 | Male, academic, Libyan, Spring, Sabha, 83, friend, yes, no, no, university graduate, high school, no, none, employee, female employee, yes | IT (positive) |
| 15 | Male, literary, Libyan, spring, other, 77, my desire, yes, no, no, university graduate, university graduate, no, nothing, employee, female employee, yes | Arts (positive) |
| 16 | Male, literary, Libyan, fall, Sabha, 73, my desire, yes, yes, no, university graduate, middle school, yes, university graduate, employee, housewife, yes | Arts (positive) |
| 17 | Female, literary, Libyan, fall, Sabha, 81, my desire, yes, yes, yes, university graduate, high school, yes, university graduate, employee, housewife, yes | Arts (positive) |
| 18 | Female, literary, Libyan, Spring, Sabha, 74, family, yes, yes, no, university graduate, middle school, no, none, employee, housewife, no | Arts (positive) |
| 19 | Male, literary, Libyan, Spring, Sabha, 71, my desire, no, yes, no, university graduate, high school, yes, high school, employee, female employee, yes | Arts (positive) |
| 20 | Male, literary, Libyan, fall, Sabha, 76, friend, yes, no, yes, university graduate, university graduate, no, none, employee, employee, yes | Arts (positive) |

For further clarification, the samples in which positive is written in the DT3 classification result field in the table above mean that the system predicted the correct college in which the student is currently studying, while negative means the opposite or an approximation as we explained previously, including the error rate was calculated. Based on these 20 samples, it turns out that the error rate is 15%.

## 6. Conclusion

In this study, a predictive model is presented that helps students choose the appropriate major based on previous data. Real data was collected for students at Sebha University, and several steps were performed on it to segment and prepare it. Then, the procedure was done by entering the data into artificial intelligence algorithms, which are DT3 and NB. And the kNN, to compare them and rely on the best algorithm that shows the best result in terms of accuracy measure, by creating a model to choose the appropriate major for students, was able in this study to obtain several positive and effective results in building a method to help future students enter university based on personal data. With an accuracy rate of 64.28%, the DT3 method proved to be the most effective among the three classification algorithms for tested. Contributing to directing students to the appropriate major, building a system based on data mining techniques that helps Sebha University improve the process of guiding students to the appropriate major and reduce student distraction during the process of choosing a university major. In the future, algorithms such as NN, SVM or Logistic Regression may be used to help better understand the problem under study. Also, delve more into specialization or prediction, such as general prediction and specific specialization. A mobile application can also be designed to help students make the appropriate decision smoothly or quickly.

## 7. Reference

[1] Aggarwal CC. Data mining: the textbook. Springer; 2015.
[2] Yağcı M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments 2022; 9:11.
[3] Batool S, Rashid J, Nisar MW et al. Educational data mining to predict students' academic performance: A survey study. Education and Information Technologies 2023; 28:905-971.
[4] Kukkar A, Sharma A, Fan J, Zhang M. Data mining applications in university information management system development. 2022.
[5] Jacob D, Henriques R. Educational data mining to predict bachelors students' success. Emerging Science Journal 2023; 7:159-171.
[6] Shoaib M, Sayed N, Singh J et al. AI student success predictor: Enhancing personalized learning in campus management systems. Computers in Human Behavior 2024; 158:108301.
[7] Angeioplastis A, Aliprantis J, Konstantakis M, Tsimpiris A. Predicting student performance and enhancing learning outcomes: a data-driven approach using educational data mining techniques. Computers 2025; 14:83.
[8] Alyahyan E, Düştegör D. Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education 2020; 17:1-21.
[9] Ciolacu M, Tehrani AF, Binder L, Svasta PM. Education 4.0-Artificial Intelligence assisted higher education: early recognition system with machine learning to support students' success. In: 2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging(SIITME). IEEE; 2018. pp. 23-30.

[10] Xie Y. Efficiency of Hybrid Decision Tree Algorithms in Evaluating the Academic Performance of Students. International Journal of Advanced Computer Science & Applications 2024; 15.
[11] Matzavela V, Alepis E. Decision tree learning through a predictive model for student academic performance in intelligent m-learning environments. Computers and Education: Artificial Intelligence 2021; 2:100035.
[12] Frazier A, Silva J, Meilak R et al. Decision Tree-Based Predictive Models for Academic Achievement Using College Students' Support Networks. arXiv preprint arXiv:2108.13947 2021.
[13] Izza Y, Ignatiev A, Marques-Silva J. On explaining decision trees. arXiv preprint arXiv:2010.11034 2020.
[14] Ibrahim IH, Garba EJ, Adejumo A. Predictive Model for Identification and Analysis of Factors Impacting Students Academic Performance Using Machine Learning Algorithms. Kasu Journal of Computer Science 2024; 1.
[15] Bridet L, Leighton MA. The major decision: Labor market implications of the timing of specialization in college. 2015.
[16] Singh LB, Chaturvedi RK, Mehdi SA, Srivastava S. Student's Preference towards Specialization Selection: An Exploratory Perspective. Adhyayan: A Journal of Management Sciences 2020; 10:10-16.
[17] Obias DC, Matitu CAB, Almodovar BJ et al. Exploring the Influence of Out-of-Specialized-Field Teaching on Educators Development and Motivation. International Journal of Multidisciplinary: Applied Business and Education Research 2025; 6:787-804.
[18] Villar A, de Andrade CRV. Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. Discover Artificial Intelligence 2024; 4:2.
[19] Bansal PK, Ahmed M. Fuzzy logic-based analysis of student behavior patterns in educational environments. In: AIP Conference Proceedings. AIP Publishing LLC; 2025. p. 020047.
[20] Alsayed AO, Rahim MSM, AlBidewi I et al. Selection of the right undergraduate major by students using supervised learning techniques. Applied Sciences 2021; 11:10639.
[21] Roman M, Ullah A, Ullah MA et al. PREDICTING ACADEMIC SUCCESS: A MACHINE LEARNING APPROACH USING DECISION TABLES AND RANDOM FORESTS ALGORITHMS. Spectrum of Engineering Sciences 2025; 3:205-213.
[22] Hernandez R, Atienza R. Career track prediction using deep learning model based on discrete series of quantitative classification. Applied Computer Science 2021; 17.
[23] Iwadi I, Ali D, Jabari M, Sukic E. The Relation of Artificial Intelligence Technology Application with Administrative Performance: A Case Study of Staff in Directorates of Education in the Hebron Governorate in Palestine. TEM Journal 2024; 13.
[24] Kuczera K, Dziembek D. Application of Rough Set Theory to Improve the Efficiency of Higher Education Systems. In: European Conference on Artificial Intelligence. Springer; 2024. pp. 237-249.
[25] Lahoud C, Moussa S, Obeid C et al. A comparative analysis of different recommender systems for university major and career domain guidance. Education and Information Technologies 2023; 28:8733-8759.
[26] Zayed Y, Salman Y, Hasasneh A. A Recommendation System for Selecting the Appropriate Undergraduate Program at Higher Education Institutions Using Graduate Student Data. Applied Sciences 2022; 12:12525.
[27] Al-Shalabi L. A Data Mining Model for Students' Choice of College Major Based on Rough Set Theory. J. Comput. Sci 2019; 15:1150-1160.
[28] Beseiso M. Enhancing student success prediction: A comparative analysis of machine learning technique. TechTrends 2025; 69:372-384.
[29] Salloum SA, Salloum A, Shaalan K et al. Comparative Analysis of Classical Machine Learning Techniques for Predicting Students' Exam Performance. In: International Conference on Breaking Barriers with Generative Intelligence. Springer; 2024. pp. 219-227.
[30] Zhang Y, Yun Y, An R et al. Educational data mining techniques for student performance prediction: method review and comparison analysis. Frontiers in psychology 2021; 12:698490.
[31] Richards Z, Kelly AM. STEM enrollment decision trees as graduation predictors for community college students enrolled in remedial mathematics. Community College Review 2025; 53:85-104.
[32] Huynh-Cam T-T, Chen L-S, Le H. Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance. Algorithms 2021; 14:318.
[33] Latif G, Abdelhamid SE, Fawagreh KS et al. machine learning in higher education: students' performance assessment considering online activity logs. IEEE access 2023; 11:69586-69600.
[34] Saeedi S, Božanić D, Safa R. Strategic analytics for predicting students' academic performance using cluster analysis and Bayesian networks. Educ. Sci. Manag 2024; 2:197-214.
[35] Hussen AAME, Saikhu A. Modeling Of Student Graduation Prediction Using the Naive Bayes Classifier Algorithm. In: 2024 3rd International Conference on Creative Communication and Innovative Technology (ICCIT). IEEE; 2024. pp. 1-8.
[36] Zheng X, Li C. Predicting students' academic performance through machine learning classifiers: a study employing the Naive Bayes classifier (NBC). International Journal of Advanced Computer Science and Applications 2024; 15.
[37] Akanbi OB. Application of naive bayes to students' performance classification. Asian journal of probability and statistics 2023; 25:35-47.
[38] Meiriza A, Lestari E, Putra P et al. Prediction graduate student use naive bayes classifier. In: Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019). Atlantis Press; 2020. pp. 370-375.
[39] Rawal A, Lal B. Predictive model for admission uncertainty in high education using Naïve Bayes classifier. Journal of Indian Business Research 2023; 15:262-277.
[40] Perez JG, Perez ES. Predicting student program completion using Naïve Bayes classification algorithm. International Journal of Modern Education and Computer Science 2021; 13:57-67.
[41] Rimal Y, Sharma N, Alsadoon A. The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. Multimedia Tools and Applications 2024; 83:74349-74364.
[42] Thamilselvan R, Rajalaxmi R, Gothai E et al. Forecasting Students Academic Achievement Using Machine Learning Techniques. In: 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA). IEEE; 2024. pp. 1-6.
[43] Jawthari M, Stoffová V. Predicting students' academic performance using a modified kNN algorithm. Pollack Periodica 2021; 16:20-26.
[44] Contreas-Bravo LE, Nieves-Pimiento N, González Guerrero K. Prediction of University-Level Academic Performance through Machine Learning Mechanisms and Supervised Methods. Ingeniería 2023; 28.
[45] Alturki S, Cohausz L, Stuckenschmidt H. Predicting Master's students' academic performance: an empirical study in Germany. Smart Learning Environments 2022; 9:1-22.
[46] Gupta V, Singhal P, Khattri V. Enhancing Predictive Accuracy in Education: A Detailed Analysis of Student Performance Using Machine Learning Models. Tuijin Jishu/Journal of Propulsion Technology 2024; 45:2024.
[47] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI magazine 1996; 17:37-37.
[48] Dehghani AA, Movahedi N, Ghorbani K, Eslamian S. Decision tree algorithms. In: Handbook of hydroinformatics. Elsevier; 2023. pp. 171-187.
[49] Sabzevari Y, Eslamian S. Bayesian theory: Methods and applications. In: Handbook of Hydroinformatics. Elsevier; 2023. pp. 57-68.
[50] Li H. K-nearest neighbor. In: Machine Learning Methods. Springer; 2023. pp. 55-66.
[51] Reeta R, Pavithra G, Priyanka V, Raghul J. Predicting autism using naive Bayesian classification approach. In: 2018 International Conference on Communication and Signal Processing (ICCSP). IEEE; 2018. pp. 0109-0113.