applied sciences

MDPI

Article

Enhanced Student Admission Procedures at Universities Using Data Mining and Machine Learning Techniques

Basem Assiri [1,*], Mohammed Bashraheel [2] and Ala Alsuri [2]

[1] Computer Science Department, College of Computer Science and Information Technology, Jazan University, Jazan 82817, Saudi Arabia
[2] Department of Information Technology and Security, College of Computer Science and Information Technology, Jazan University, Jazan 82817, Saudi Arabia; msbashraheel@jazanu.edu.sa (M.B.); ayusef@jazanu.edu.sa (A.A.)
* Correspondence: babumussmar@jazanu.edu.sa

Abstract: The progress of technology has played a crucial role in enhancing various fields such as education. Universities in Saudi Arabia offer free education to students and follow specific admission policies. These policies usually focus on features and scores such as the high school grade point average, general aptitude test, and achievement test. The main issue with current admission policies is that they do not fit with all majors, which results in high rates of failure, dropouts, and transfer. Another issue is that all mentioned features and scores are cumulatively calculated, which obscures some details. Therefore, this study aims to explore admission criteria used in Saudi Arabian universities and the factors that influence students' choice of major. First, using data mining techniques, the research analyzes the relationships and similarities between the university's grade point average and the other student admission features. The study proposes a new Jaccard model that includes modified Jaccard and approximated modified Jaccard techniques to match the specifications of students' data records. It also uses data distribution analysis and correlation coefficient analysis to understand the relationships between admission features and student performance. The investigation shows that relationships vary from one major to another. Such variations emphasize the weakness of the generalization of the current procedures since they are not applicable to all majors. Additionally, the analysis highlights the importance of hidden details such as high school course grades. Second, this study employs machine learning models to incorporate additional features, such as high school course grades, to find suitable majors for students. The K-nearest neighbor, decision tree, and support vector machine algorithms were used to classify students into appropriate majors. This process significantly improves the enrolment of students in majors that align with their skills and interests. The results of the experimental simulation indicate that the K-nearest neighbor algorithm achieves the highest accuracy rate of 100%, while the decision tree algorithm's accuracy rate is 81% and the support vector machine algorithm's accuracy rate is 75%. This encourages the idea of using machine learning models to find a suitable major for applicants.

Keywords: students; university admission; major selection; data mining analysis; machine learning models

Citation: Assiri, B.; Bashraheel, M.; Alsuri, A. Enhanced Student Admission Procedures at Universities Using Data Mining and Machine Learning Techniques. *Appl. Sci.* **2024**, *14*, 1109. https://doi.org/10.3390/app14031109

Academic Editors: Yu Liang, Wenjun Wu and Ying Li

Received: 17 December 2023
Revised: 18 January 2024
Accepted: 22 January 2024
Published: 29 January 2024

1. Introduction

This paper is an extension of a work originally presented at the 7th International Conference on Data Science and Machine Learning Applications (CDMA) [1]. In today's world, technology plays a crucial role in the development of various fields, including medicine, education, industry, economy, and securities. Consequently, technological advancement has become essential as it strengthens the quality, facilities, and overall improvement of these fields. Education is one of the most significant fields, so developed nations typically place emphasis on their educational systems. This contributes to a capable, efficient, and thriving society. Countries provide different kinds of educational systems and teaching methods

suiting all fields of science. Guiding people correctly to suitable educational systems and fields will improve people's achievements, leading to success and satisfaction [2,3].

Studying at university is a significant stage for students, as it enables them to secure employment opportunities. Every year, a large number of high school graduates pursue job opportunities or enroll in universities. Students who enroll in universities must avoid favoring prestigious majors since they may not always align with their interests and capabilities. According to research, several factors can affect a student's decision when selecting a major at a university. These may include their background, economic status, time and financial constraints, psychological factors, gender, skills, job market conditions, and the impact of family and peers [4,5]. The factors mentioned above can have an impact on student choices and the processes of decision-making.

Free higher education is provided to students in public universities in Saudi Arabia, and earning a degree enhances the chances of acquiring appropriate employment opportunities with a decent income. As a result, a considerable number of high school graduates seek admission to universities. As an illustration, the Admission and Registration Deanship at Jazan University reported an annual enrollment of approximately 13,000 students from 2017 to 2020. However, unfortunately, there are many cases where the academic major that students favor may not be the optimal selection, leading to potential failure or the need to transfer to a different major after several years, where the failure rates reach 28–33% [6–8]. As a result, universities permit students to choose their preferred areas of study while assessing their aptitude through evaluations such as the high-school grade point average (GPAH), general aptitude test (GAT), and achievement test (AT). The GAT is a widely used test around the world [9,10] that assesses students' cognitive abilities, e.g., their comprehension and analytical skills. In addition, it evaluates students' performance in mathematics and Arabic. The AT evaluates students' success in mathematics, physics, chemistry, and biology during their high school education. In fact, it is not mandatory for every candidate, e.g., those who seek admission to Arabic Language or Islamic Studies programs, to take the test [9,10].

Following this, a student's weighted score (WS) is calculated by combining their GPAH (50%) and GAT score (50%). Furthermore, a student's qualifying score (QS) is determined by combining their GPAH (30%), GAT score (30%), and AT score (40%). These percentages may be modified in certain situations, such as during the COVID-19 pandemic.

The primary concern with these scores is that they conceal certain crucial details that influence a student's choice of major at university. For instance, some students may have the same WS, but among them, some may have an average GPAH and a high GAT score, while others may have a high GPAH and an average GAT score. Additionally, individuals with an average GPAH and a high AT score might not be suitable for an English language major, but they could still be admitted to this major based on their overall score. Moreover, this study attempts to mitigate the influence of human factors by utilizing data mining and machine learning methods.

Furthermore, machine learning is applied in various fields, including education, healthcare services, marketing, and finance [11]. Machine learning models are heavily dependent on data and statistics. As a subset of artificial intelligence, machine learning can learn from data, recognize patterns, and make decisions with minimal human involvement. Indeed, there are two categories of machine learning models: supervised and unsupervised learning. Supervised learning involves providing the model with training examples to teach the algorithms how to generate the correct output. In unsupervised learning, algorithms can learn and discover hidden patterns on their own [12].

This research examines the admission standards for universities in Saudi Arabia, focusing on the impact of concealed information (which underlies students' GPAH, WS, and QS scores) on their selection of majors. First, a questionnaire is presented to experts to comprehend the ideal process of selecting a student's major. Second, we gather actual data about Jazan University students to examine the current state of the process of selecting a major. Here, we use data mining techniques to analyze the collected data and establish

connections between student performance in the university, indicated by GPAU, and admission features such as GPAH, GAT, AT, WS, and QS. Actually, four types of relationship analysis are employed: modified Jaccard analysis, approximated modified Jaccard analysis, data distribution analysis, and correlation coefficient analysis. The research modifies Jaccard index [13] to suit the characteristics of the students' data records. Indeed, Jaccard analysis measures the similarity between two sets of data. The objective is to find the percentage of similarity between the elements in the sets. Our data sets include numerous attributes, including GPAU, GPAH, GAT, AT, WS, and QS. Each attribute is considered as a set, and the degree of similarity is calculated. However, since our data set includes students' records, we initially propose the notion of modified Jaccard analysis, which measures the similarity between the elements of two sets and of the same record. For instance, it examines the similarity between GPAU and GPAH for the same student. Next, we use approximated modified Jaccard analysis to measure similarity in a more lenient manner, treating values that are near each other as similar. Moving forward, we conduct an analysis of the data distribution, followed by using correlation coefficient analysis. Accordingly, the investigation shows that current admission procedure has a generalization issue, where the relationships tend to differ significantly from one major to another.

Thirdly, this article suggests incorporating additional features, like the grades earned in high school courses, into supervised machine learning models to accurately find suitable majors. The paper specifically applies K-nearest neighbor (KNN), decision tree (DT), and support vector machine (SVM) algorithms to classify students into their respective major categories [14–16]. The experimental simulation revealed that KNN provides the most accurate results, with a 100% success rate. In comparison, the DT has an accuracy rate of 81% and the SVM has an accuracy rate of 75%. Employing this process has the potential to increase enrolment rates for students in majors that are best suited to their individual strengths and abilities. This enables students to graduate on time with a wealth of knowledge, practical experience, and overall satisfaction. Additionally, this technique can substantially reduce the frequency of students transferring, withdrawing, or failing courses.

1.1. K-Nearest Neighbor

The initial application of the KNN model was in statistical settings in the early 1970s [14]. The model operates by identifying a collection of K instances that are the closest in distance to a given point. The optimal K-value is used to cluster the given data. The variable K represents the number of neighboring data points that are taken into account during the classification process.

1.2. Decision Tree

A decision tree is a type of supervised learning method that is commonly used for both classification and regression [15]. The DT algorithm finds the value of an unknown data point by analyzing its attributes. Each attribute is evaluated as true or false using an if statement. In this study, the DT model is utilized specifically for classification purposes.

1.3. Support Vector Machine

SVM is a commonly used supervised learning algorithm that is applicable for various tasks such as classification, regression, and outlier detection. In SVM, the objective is to construct a hyperplane, which is basically a line that separates data points into different classes, and this hyperplane is used to predict the classes of new data points. The SVM algorithm is able to handle data that have multiple features represented in high-dimensional space, as mentioned in reference [16].

2. Related Work

Many researchers have examined the challenge of selecting a suitable university major. Some studies have specifically emphasized the difficulty of making this decision due to external factors such as familial or cultural pressure, as referenced in studies [4,5].

Social and community pressures, including the influence of friends, instructors, and family members, can significantly impact a student's choice of major, particularly in the fields of science, technology, engineering, and mathematics. Research has indicated that mothers, even those who have not attended college themselves, have the most significant influence in guiding their sons' and daughters' decisions regarding their choice of major [6]. According to research, gender plays a role in selecting a major. Malgwi et al. performed a survey and discovered that females tend to consider their abilities when selecting a major, whereas males are more inclined to choose majors that have better job prospects [11]. Many factors influence this issue, including economic factors, familial considerations, and personal preferences, as well as gender. Montmarquette et al. found that male students tend to be more careful in their choice of major compared to their female counterparts [12]. Overall, these studies demonstrate that the process of choosing a major is complex and multifaceted.

In addition, universities have various plans in place to help students select an appropriate major. While the specific criteria may vary among institutions, many universities make major selection a part of the admissions process [3,4]. Conversely, some universities adopt a general curriculum in the first year, allowing students to explore various fields before deciding on a major. However, some students may still struggle to make a decision after the first year, and as a result, their GPA may suffer if they take courses in diverse areas like mathematics or physics [4]. Furthermore, a different study indicates that certain universities utilize various methods to assist students in choosing their majors. One approach involves selecting majors based on enrollment essays, which are evaluated using automated computer systems to analyze their content, logic, and psychological information [17]. Moreover, the availability of jobs and the demand for certain majors rely heavily on countries' policies and marketplaces. While various countries employ different education strategies, a student's decision to select a suitable major is crucial, as it helps a country advance and enables students to succeed in their careers [18–20]. Thus, enrolling in a major without considering its suitability can result in failure, which can have detrimental effects on students, families, educational systems, and even countries. Therefore, various studies provide guidance to students on selecting a suitable major [21,22].

Many countries utilize pre-college examinations to assess students' academic abilities and accomplishments. In certain countries, assessments like the American College Test (ACT) and Scholastic Assessment Test (SAT) are employed to authenticate the accuracy of student GPAH. According to research, it is not sufficient to solely rely on either tests or GPAH. However, for a more precise evaluation of students' abilities and academic achievements, we can consider both pre-college examinations and GPAH together [23–26].

In the Kingdom of Saudi Arabia, universities provide free education, and to maintain the large number of applicants, they have special admission criteria. They allow applicants to rank their preferable majors. Then, enrollment is based on GPAH and two pre-university tests, which are the GAT and AT. A study was conducted on students at Damam University, and it was found that there is a strong relation between students' GPAU and their GAT, AT, and GPAH scores. The study shows that the influence of the GAT and AT is important for students pursuing medical and science majors but has less significance for students pursuing humanities majors [9]. The influence of the GAT and AT on students' performance at King Saud University was also studied. It was found that GPAH affects the college GPAU more than the GAT and AT [10].

Furthermore, many works use artificial intelligence techniques such as data mining, machine learning, and deep leaning models to analyze the issue of choosing majors at university and to guide students. A proposed study will focus on a comparison between first-year students and other students. It will analyze some factors such as a student's general ability and personality characteristics to help students select suitable majors. It will use artificial intelligence and an expert system rather than traditional methods [5,27]. Another work categorized student performance and learning aspects using an intelligent system, which contained adaptive neuro fuzzy and learning vector quantization network methods, to help institutions classify students based on their ability to progress [18].

Another work used game theory for admission polices [28]. Xu et al. considered the influence of pedagogical process of students on their background, academic performance, and subject choices. This research applied a machine learning model to make predictions related to an acquired program relying on students' progress during the college period [29]. Some others used parallelism to enhance machine learning models and their distribution processes [30,31]. Ahmad et al. used data mining to predict students' ability to progress while studying at university. Their research applied decision tree, naïve Bayes, and rule-based models, and they found that the rule-based model was the best in terms of prediction accuracy, which reached 71% [32].

The purpose of this work is to examine admission procedures for universities. It measures the relation between the main features of current students' admission procedures, such as GPAU and other scores such as GPAH, GAT, AT, WS, and QS. A Jaccard analysis is used to measure the similarities among datasets [33–35]. Some works modify or relax the Jaccard similarity index to fit it to different kinds of problems. This also helps in finding the similarity and relations between structured and unstructured data [35–37].

Additionally, this study employs data mining techniques to compile and analyze the provided data. Finally, this paper presents machine learning models that help to find the most fitting majors for university students during the admission process. These models suggest a suitable major based on GPAH, individual scores of high school courses, and some other qualitative and quantitative standardized tests.

3. Methodology

The process of this investigation involved six stages, namely, collecting data, preparing data, data mining, creating a relevant dataset, developing machine learning models, and making assumptions. In this research, data collection and analysis took two paths to investigate the major selection process, which assessed how it should be and how it is currently. First, a questionnaire was given to domain experts to investigate the major selection process, which assisted in determining the correct process and procedure. Second, we gathered real-word information about students and used data mining to examine the present state of procedures related to the selection of majors.

3.1. Data Collection

This section looks into the viewpoints of professionals and experts in 19 different majors at Jazan University, as outlined in Table 1. This study was carried out using a survey that is available in both Arabic and English and is presented below:

- Goal: The objective of the survey was to enable specialists to identify the key requirements for admission in every major. Specialists in the field ensured data accuracy and the comprehension of all aspects of the issue. This facilitated the creation of a precise dataset and provided proper guidance to machine learning models.
- Participants: The respondents, who were the primary stakeholders, consisted of 61 male and female faculty members from various majors at Jazan University, as demonstrated in the first and second columns of Table 1.
- Design: Our proposal involved an internet-based survey that was administered to 61 faculty members from 19 different majors at Jazan University, as depicted in the first column of Table 1. The survey was conducted between April and June of 2020 and comprised seven straightforward questions to identify the primary prerequisites that students must meet to be admitted into majors. The main inquiry sought faculty members' input about the significant high school subjects that concern each university major (based on their respective fields). The second question concerned the minimum scores that students must achieve in these subjects (determined by the first question). The third question concerned the minimum GAT and AT scores that students must obtain to be eligible for these majors. The fourth question inquired about the pre-admission exams that are mandatory for specific college majors. The answers show that pre-admission exams are not required for about 95% of the university's majors.

The fifth question focused on students' success factors. Lastly, the survey concluded the input with an open question.

- Results: Table 1 provides a summary of the questionnaire results. The first column indicates the colleges and majors, while the second column displays the number of faculty members participating from various majors. The third column outlines the significant courses associated with each major. This information was used to create our dataset (which provided training examples for machine learning models) to help students choose an appropriate major at university. The fourth column of the table displays the minimum scores required for each major, including GPAH, GAT, and AT. However, it should be noted that the responses to this question were not entirely accurate, indicating that the faculty members may not have a complete understanding of the structure of the GAT and AT. For instance, a Mathematics faculty member indicated that a GAT score of 60 and an AT score of 60 are sufficient, even though this falls below the acceptance threshold at Jazan University. The last column in the table outlines the essential skills and qualities that students need to possess to excel in each major. The findings of the questionnaire suggest that each major has specific courses that are of utmost importance, while other courses may not be as necessary.
- Practical aspects: A comprehensive dataset needed to be designed, including specific training examples based on the survey's results and further analyzed data. After the initial questionnaire results and real-world data examples were combined, the accuracy of the training process within the machine learning models could be improved. By analyzing the overlap between majors, 60 training examples were used and categorized into four labels or classes, as shown in Tables 2 and 3.

Table 1. The questionnaire results.

Colleges and Majors		Participants	Important Courses	GPAH, GAT, and AT	Characteristics and Skills
Majors of Medicine		5	Biology and Chemistry	90, 90, 90	- Analysis - Memorizing - Discipline - Being up to date - Time management - Patience - Hard working
Majors of Dentistry		3	Biology and Chemistry	90, 90, 85	
Majors of Public Health and Tropical Medicine		4	Biology	90, 90, 85	
Majors of Nursing		3	Biology	90, 85, 85	
Majors of Applied Medical Science		3	Biology	90, 85, 85	
Majors of Pharmacy		5	Chemistry and Biology	87.5, 85, 85	
Majors of Engineering		3	Math, Physics, Chemistry, and English Lang	85, 80, 85	- Critical thinking - Analysis - Discipline - Being up to date - Time management - Hard working - Creativity
Majors of Computer Science and IT		4	Math and English Lang	80, 85, 80	
Majors of Design and Architecture		3	Art, History, Geography, and Science	80, 80, 80	
Science	Math	8	Math and English Lang	80, 80, 80	- Analysis - Memorizing - Discipline - Time management - Hard working
	Physics		Math, Physics, and English Lang		
	Chemistry		Math, Chemistry, and English Lang		
	Biology		Biology		
Majors of Business Administration		3	Math and English Lang	80, 75, 70	
Majors of Education		3	No Preference	75, 75	
Majors of Sharia and Law		3	Arabic Lang and Religious Studies	75, 75	
Arts and Humanities	English	11	English Lang	75, 75	- Analysis - Memorizing - Discipline - Time management - Hard working - Creativity
	Arabic		Arabic Lang		
	Others		Arabic Lang, History, and Geography		

Table 2. Dataset sample that includes features and labels (the first row includes the number of rows and columns as well as the classes).

# of Examples	# of Columns	Classes																	
60	20	1	2	3	4														
Quran	Tafseer	Hadith	Toheed	Fegh	Arabic	Math	Physics	Chemistry	Biology	Geology	English	Computer Science	Sociology	Research Skills	Applied Skills	Physical Education	GAT	AT	Class
80	80	80	80	80	80	75	75	75	75	75	80	80	75	80	75	80	75	75	4
95	99	98	99	99	100	96	98	100	100	99	97	96	99	94	96	100	91	87	1
80	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	85	80	80	3

Table 3. The four classes in the dataset.

#	Labels (Classes)	Colleges and Majors
1	Medicine and Healthcare	Majors of Medicine, Dentistry, Public Health and Tropical Medicine, Nursing, Pharmacy, and Applied Medical Science
2	Engineering and Computation	Majors of Engineering, Computer Science and IT, Design and Architecture
3	Science and Management	Majors of Science and Business Administration
4	Theoretical Science	Majors of Education, Sharia (and Law), and Arts (and Humanities)

3.2. Real Data Preparation

To support the survey, student information was also required to analyze and measure the correlation between GPAU and various other factors such as GPAH, GAT, AT, WS, and QS. This investigation helped in dataset (training examples and labels) preparation. This process was conducted as illustrated below:

- Data specification: This involved identifying the necessary data for this study, which included high school records, GPAU, GPAH, GAT, AT, WS, and QS. The data covered both male and female students across all undergraduate majors at Jazan University.
- Requesting data: This involved seeking permission from the Jazan University administration to access the requested data, ensuring that this study used real-world data.
- Data cleaning: Once the data were obtained, data cleaning was conducted to address any missing values and to eliminate noisy data. This study considered data from 2018, 2019, and 2020.
- Data ordering: This re-arranged the data based on the domain experts' answers (in the survey). We divided students' data into two categories, scientific majors and theoretical (literature) majors, since they have different properties. For example, students pursuing scientific majors use QS, while theoretical major students use WS.

3.3. Data Mining Process

After receiving the students' information from Jazan University, we investigated the patterns and relationships among the data as follows:

(1) The correlation between the GPAU and GPAH for students pursuing all majors.
(2) The correlation between the GPAU and GAT for students pursuing all majors.

(3) The correlation between the GPAU and WS for students pursuing theoretical majors.
(4) The correlation between the GPAU and AT for students pursuing scientific majors.
(5) The correlation between the GPAU and QS for students pursuing scientific majors.

In this part, we applied data mining to analyze the given data and to find the relations between student performance at university, represented by GPAU, and other scores such as GPAH, GAT, AT, WS, and QS. We used four kinds of relation analysis, which were modified Jaccard analysis, approximated modified Jaccard analysis, data distribution analysis, and correlation coefficient analysis. As mentioned earlier, this study proposed modified Jaccard and approximated modified Jaccard analyses to suit the structure of our data. Our data consisted of students' records in different majors, and each record had many attributes such as GPAU, GPAH, GAT, AT, WS, and QS.

This investigation transformed the data and information into knowledge, providing deep insights into how these individual scores influence students' major selections and how these scores are reflected in students' success in their majors. This helped to diagnose the presence and the size of the problem and to adjust the measurements while devising a solution, which will be explained in detail later.

3.4. Dataset Preparation

We followed the guidance of the questionnaire and correlation results to prepare our dataset in a way that was suitable for use with a machine learning model. This involved building training examples in which all data were numerically represented and divided into features and labels.

Our dataset was made up of 20 columns and 60 rows. Of these columns, 19 represented features, including scores of students in various high school courses such as Quran, Tafseer, Hadith, Toheed, Feqh, Arabic, Math, Physics, Chemistry, Biology, Geography, English, Computer Science, Research Skills, Applied Skills, and Physical Education. Additionally, GAT and AT scores were also included among the features. The last column represented the label, which was the suitable class, as illustrated in Table 2. The dataset contained 60 rows. The first row provided a summary of the numbers of rows, columns, and classes. The second row displayed the features, while the remaining rows presented training examples. As previously mentioned, there are four classes represented in Table 3.

3.5. Machine Learning Procedures

To assist students in selecting a suitable major, this paper suggests using supervised machine learning models. These models aim to enhance the precision of finding students' university majors. This study used KNN, DT, and SVM and compares the accuracy of these three classifiers in relation to our problem. This study analyzed GPAH features by evaluating individual scores in all high school courses, as well as GAT and AT scores, to anticipate the correct major. Additionally, certain courses (features) that were significant for each major (as demonstrated in Table 1) were assigned greater importance for each major.

3.6. Assumption

This study assumes that using individual scores of high school courses (which are hidden behind GPAH) has a notable influence on finding the appropriate major for university applicants.

4. Data Mining Analysis and Experimental Simulation

The analysis and the experimental simulation used two methods, which were data mining and machine learning processes. The first method used data mining to determine the correlation among the dataset's main features, focusing on the correlations between GPAU and other scores such as GPAH, GAT, AT, WS, and QS. We emphasize the positive and negative influences of some features and illustrate the importance of conducting more investigations into current university student acceptance procedures.

The second approach utilized machine learning models to demonstrate the recommended process for admitting university students. The experimental simulation investigated the consideration of additional features such as high school course grades. The proposed admission procedures were analyzed using three machine learning models.

4.1. Data Mining (Correlation Analysis)

This section involves analyzing the connections and similarities among the principal features of students' admission procedures. The analysis explored the relationships between students' GPAU and GPAH, GAT, AT, WS, and QS. To facilitate this investigation, all GPAU scores were converted to a 100-point scale. Then, we conducted modified Jaccard, approximated modified Jaccard, data distribution, and correlation coefficient analyses. With the use of Microsoft Visual C++17 software, the code was written in C++ to find the mentioned relations.

4.1.1. Modified Jaccard Analysis (MJA)

As mentioned before, Jaccard analysis measures the similarity between two sets of data. It determines the similarity percentage of the elements of the sets. A and B are two sets, and Jaccard analysis measures the similarity between A and B using the following formula [13]:

$$J(A,B) = |A \cap B| / |AUB|$$

Since our datasets consisted of many features, such GPAU, GPAH, GAT, AT, WS, and QS, we considered each feature as a set, and we measured the similarity. We measured the similarities between GPAU from one side and GPAH, GAT, AT, WS, and QS from the other side. However, our dataset consisted of students' records, so we proposed the concept of MJA, which measures the similarities between the elements of two sets and the same records. For example, it checks the similarity of GPAU and GPAH for the same student. Figure 1 shows the difference between Jaccard analysis and the proposed MJA. For four students, GPAU = {99, 91, 86, 66} and GPAH = {91, 86, 85, 80}. It is clear that the first elements of the two sets belong to one student (one record), the second elements of the two sets belong to another student (one record), and so on.

GPAU	GPAH
99	91
91	86
86	85
66	80
|GPAU ∩ GPAH| = {91, 86}	
|GPAU U GPAH| = {99, 91, 86, 66,85, 80}	
Jaccard = 33.34%	

GPAU	GPAH
99	91
91	86
86	85
66	80
Records similarity = {}	
MJA = 0%	

Figure 1. Illustrates a comparison between Jaccard analysis and MJA.

Therefore, the result of the traditional Jaccard analysis was approximately 33.34%, as it found two similarities {91, 86} out of six elements {99, 91, 86, 66, 85, 80}. It compared the elements of the two sets regardless of their records. The result of the MJA was 0%, as it checked the similarity of the first elements in the two sets, then the second elements, and so on. In other words, it considered the similarity only if the student's GPAU matched his/her GPAH. Then, we separated the investigation results into two groups, which were scientific colleges, as shown in Table 4, and theoretical colleges, as shown in Table 5. Clearly, Table 4 shows the number of students and the MJA calculation between the GPAU and other scores (which are GPAH, GAT, AT, and QS), and Table 5 also shows the MJA calculation between the GPAU and GPAH, GAT, and WS. The general results of the MJA for scientific colleges

in Table 4 show very low similarities, as they do not exceed 9%, and the most similar percentages lay between 3% and 5%. In addition, the MJA results differ from one college to another. The MJA results between GPAU and QS are the highest for 45% of scientific majors such as dentistry, public health and tropical medicine, pharmacy, design and architecture, and business administration. The table also shows that for 27% of majors, such as medicine, nursing, and applied medical sciences, GPAU to GPAH is the best indicator, while GPAU to AT is the best indicator for engineering, computer science and information technology, and science colleges and GPAU to GAT is the best indicator for some other colleges. However, there are some colleges that have the same MJA results for two or more indicators. For instance, the nursing college has the same result for GPAU to GPAH and GPAU to GAT. Moreover, most of the findings also apply to the MJA results for theoretical colleges that appear in Table 5.

Table 4. Illustrates MJA between GPAU and GPAH, GAT, AT, and QS for scientific colleges.

College	No. of Students	GPAU—GPAH	GPAU—GAT	GPAU—AT	GPAU—QS
Medicine	458	0.09	0.04	0.03	0.04
Dentistry	185	0.02	0.04	0.04	0.05
Public Health Tropical Medi	620	0.02	0.02	0.02	0.03
Nursing	521	0.03	0.03	0.02	0.02
Applied Medical Sciences	1363	0.03	0.02	0.02	0.02
Pharmacy	454	0.04	0.03	0.03	0.05
Engineering	945	0.01	0.02	0.03	0.02
Computer Science and IT	1515	0.01	0.03	0.04	0.03
Design and Architecture	409	0.02	0.02	0.03	0.05
Science	2056	0.01	0.02	0.03	0.02
Business Administration	2637	0.03	0.03	0.02	0.03

Table 5. Illustrates MJA between GPAU and GPAH, GAT, and WS for theoretical colleges.

College	No. of Students	GPAU—GPAH	GPAU—GAT	GPAU—WS
Education	603	0.05	0.01	0.03
Sharia and Law	1216	0.04	0.02	0.02
Arts and Humanities	2957	0.01	0.03	0.02

To summarize, the MJA does not show strong relations between student performance at university (GPAU) and the other features. Thus, this study proposes further investigating feature relations using scores behind these features.

4.1.2. Approximated Modified Jaccard Analysis (AMJA)

The low MJA results encouraged us to relax this analysis and instead propose AMJA. Since MJA considers a similarity only if the value of the element in set A matches the exact value of the corresponding element in set B, this paper proposes AMJA, which measures the approximated similarities in a range (+10/−10) of values. We relaxed the similarity measure to (+10/−10) under the guidance of the university's grading system since each grade represents about 10 points. For example, grade A lies between 90 and 100, grade B lies between 80 and 89, and so on. This relaxation allowed us to find more relations among our features.

AMJA was applied to the example in Figure 1, in which GPAU = {99, 91, 86, 66} and GPAH = {91, 86, 85, 80}. The AMJA resulted in 75% approximated similarity because the difference between the first elements of the two sets was less than 10. This was also true for the second and the third elements. However, it did not count the fourth elements {66–80}, as the difference exceeded 10 points. As usual, we separated the AMJA results into two groups, which were scientific colleges, as shown in Table 6, and theoretical colleges, as shown in Table 7.

Table 6 shows the AMJA calculations between the GPAU and the other scores (GPAH, GAT, AT, and QS), and Table 7 shows the MJA calculations between the GPAU and GPAH, GAT, and WS. The numbers of students are the same as those in Tables 4 and 5.

Table 6. Illustrates AMJA between GPAU and GPAH, GAT, AT, and QS for scientific colleges.

College	GPAU—GPAH	GPAU—GAT	GPAU—AT	GPAU—QS
Medicine	38.56	35.73	38.56	40.52
Dentistry	30.65	31.18	44.09	47.31
Public Health and Tropical Medi	12.56	21.09	25.44	28.34
Nursing	24.14	23.37	29.17	27.97
Applied Medical Sciences	22.36	22.07	27.27	29.18
Pharmacy	31.37	23.75	31.81	40.74
Engineering	12.37	27.17	28.75	28.96
Computer Science and IT	11.15	26.25	28.96	28.96
Design and Architecture	19.51	17.07	25.85	42.2
Science	6.17	24.99	29.7	28.44
Business Administration	19.98	20.02	18.99	27.29

Table 7. Illustrates AMJA between GPAU and GPAH, GAT, and WS for theoretical colleges.

College	GPAU—GPAH	GPAU—GAT	GPAU—WS
Education	50.67	3.31	9.77
Sharia and Law	25.72	2.71	8.46
Arts and Humanities	12.95	1.69	6.22

Although the AMJA is very relaxed, the relations between GPAU and other scores are not very tight. The results of the AMJA show that the relation between GPAU and QS is the best indicator, even if it does not suit some majors such as nursing and science. Also, we cannot ignore the huge percentages of non-similarities that range between 52 and 72% for all colleges.

For theoretical colleges, the AMJA of GPAU and GPAH is the best indicator, even if non-similarities lay between 49 and 87% for all colleges. The relation is very low for GPAU and WS, which is the main measure in the current acceptance procedures at Saudi universities.

4.1.3. Non-Similarity Distribution Using Approximated Modified Jaccard Analysis (DAMJA)

Before proceeding, we investigated the non-similarity percentages to see the data distribution of these non-similar values. In Tables 8 and 9, non-similar values are classified as high or low. If the difference between a student's GPAU and the corresponding score (GPAH, GAT, AT, WS, or QS) was greater than +10 points, it was classified as high, and if the difference between a student's GPAU and the corresponding score was smaller than −10 points, it was classified as low; otherwise, it was classified as an approximated similarity (the approximated similarities already appear in Tables 6 and 7). The numbers of students are the same as those in Tables 4 and 5.

Table 8 shows a classification of similar and non-similar AMJA results. For the major of medicine, first, the approximated similarity of GPAU and GPAH was 38.56%. No student had +10 GPAU in comparison with GPAH, and for 61.4% of students the GPAU was more than 10 points less than the GPAH. Second, for GPAU and GAT, the approximated similarity was 35.73%, where 29.19% was high and 35.1% was low. Third, for GPAU and AT the approximated similarity was 38.56%, where 18.95% was high and 44.5% was low. Fourth, for GPAU and QS the approximated similarity was 40.52%, where 11.98% was high and 47.49% was low. According to this investigation, a student who is classified as high performs better at university, which is a good indicator for major selection, even if they are not classified as an approximated similarity. This means that even if GPAU and QS have

the highest approximated similarity, GPAU and AT are better since the total number of both the approximated similarity and high classification is the highest. This is also applicable for the majors of dentistry, public health, engineering, computer science and information technology, and business administration. In addition, the approximated similarity of GPAU and GAT is the best indicator for the majors of nursing, applied medical sciences, pharmacy, design and architecture, and science.

Table 8. Illustrates AMJA and non-similar distribution that lays between GPAU and GPAH, GAT, AT, and QS for scientific colleges.

College	GPAU—GPAH			GPAU—GAT			GPAU—AT			GPAU—QS		
	Compatible	High	Low	Compatible	High	Low	Compatible	High	Low	Compatible	High	Low
Medicine	38.56	0.00	61.41	35.73	29.16	35.16	38.56	18.95	42.58	40.52	11.98	47.49
Dentistry	30.65	0.54	68.83	31.18	46.22	22.65	44.09	26.88	29.00	47.31	18.82	33.87
Public Health and Tropical Medicine	12.56	0.48	87.00	21.09	42.84	36.16	25.44	39.77	34.83	28.34	21.74	49.92
Nursing	24.14	0.00	75.91	23.37	54.25	22.44	29.17	42.15	28.76	27.97	33.14	38.89
Applied Medical Sciences	22.36	0.22	77.45	22.07	48.56	29.42	27.27	40.84	31.92	29.18	27.86	42.96
Pharmacy	31.37	0.22	68.43	23.75	57.32	19.00	31.81	45.75	22.41	40.74	32.03	27.23
Engineering	12.37	0.11	87.52	27.17	29.81	43.00	28.75	34.46	36.88	28.96	15.75	55.29
Computer	11.15	0.99	87.96	26.25	46.80	26.90	28.96	44.53	26.57	28.96	25	46.04
Design and Architecture	19.51	0.49	80.00	17.07	75.17	7.80	25.85	64.15	10.00	42.2	41.22	16.59
Science	6.17	0.78	93.11	24.99	45.19	29.91	29.7	40.25	30.00	28.44	21.29	50.27
Business Administration	19.98	2.19	77.80	20.02	61.45	18.66	18.99	69.47	11.53	27.29	44.84	27.87

On the other hand, for the majors of, Sharia and law, and arts and humanities, the approximated similarity of GPAU and GPAH is the highest, but by considering the advantages of those students who have high performances at university, the approximated similarity of GPAU and GAT becomes the best indicator for education majors and Sharia and law majors.

In short, such variations demonstrate the weakness and inconsistency of current admission procedures and features.

Table 9. Illustrates AMJA and non-similar distribution that lays between GPAU and GPAH, GAT, AT, and QS for theoretical colleges.

College	GPAU—GPAH			GPAU—GAT			GPAU—WS		
	Compatible	High	Low	Compatible	High	Low	Compatible	High	Low
Education	50.67	44.87	4.47	3.31	96.69	0.00	9.77	9.23	0.00
Sharia and Law	25.72	72.14	2.14	2.71	97.12	0.16	8.46	91.45	0.08
Arts and Humanities	12.95	85.16	1.89	1.69	98.11	0.20	6.22	93.71	0.07

4.1.4. Correlation Analysis

In this part, we analyze the relations among the main features of student acceptance procedures using the correlation coefficient. The correlation coefficient is used to measure statistical values, which range between 1 and −1, and to create a relationship between them using the following formula [38]:

$$P(X,Y) = E\,(X - \mu x)(Y - \mu y)/\sigma X \cdot \sigma Y$$

The Pearson product–moment correlation coefficient (ρ) is a measure of the correlation between two variables, X and Y. Here, E is the expectation; μx and μy are the mean values of X and Y, respectively; σX is the standard deviation of X; and σY is the standard deviation of Y. This formula considers the means and the expectations of X and Y.

We separated the investigation results into two groups, which are scientific and theoretical colleges. Tables 10 and 11 show the correlation coefficient calculation between GPAU and the other scores (GPAH, GAT, and WS).

Table 10. Illustrates the correlation between GPAU and GPAH, GAT, AT, and QS for scientific colleges.

College	GPAU—GPAH	GPAU—GAT	GPAU—AT	GPAU—QS
Medicine	0.33	0.16	0.20	0.26
Dentistry	0.28	−0.06	0.21	0.16
Nursing	0.45	0.13	0.23	0.31
Pharmacy	0.40	0.10	0.20	0.30
Applied Medical Sciences	0.42	0.21	0.41	0.45
Public Health	0.38	0.28	0.39	0.46
CSIT	0.50	0.47	0.52	0.61
Design	0.53	0.33	0.44	0.53
Engineering	0.50	0.43	0.50	0.57
Science	0.52	0.36	0.53	0.58
Business	0.36	0.21	0.43	0.41

Table 11. Illustrates the correlation between GPAU and GPAH, GAT, AT, and QS for theoretical colleges.

College	GPAU—GPAH	GPAU—GAT	GPAU—WS
Education	0.34	0.01	0.26
Sharia	0.38	0.31	0.43

The correlation coefficient between GPAU and GPAH is the highest for the medicine, dentistry, nursing, and pharmacy majors, as well as some theoretical majors such as education and Sharia. Meanwhile, for majors such as applied medical sciences, public health, computer science and information technology, design, engineering, and science the correlation coefficient between GPAU and QS is the best indicator, while for the art major the correlation coefficient between GPAU and WS is the best indicator. In addition, for the business major the correlation coefficient between GPAU and AT is the best indicator.

In short, the correlation between GPAU and GPAH is suitable for some colleges, and considering other scores has a negative impact on the correlation. The qualifying score is an appropriate indicator for other scientific colleges, while the AT is a suitable indicator for the business college. These results show that the scores currently considered for students' acceptance are not suitable for many majors. Thus, this study proposes further investigating the detailed scores behind GPAH, AT, and QS.

4.2. Machine Learning Models

This study focused on the admission procedures of universities, considering additional features such as high school course grades. This research employed three machine learning models to analyze the suggested acceptance procedures. This study was conducted using Windows 10, and Python was the programming language. It utilized a 2.90 GHz Intel Core (TM) i7 CPU and 4 GB of RAM for processing. Each test was conducted five times, and the average results are reported.

This section of the study aimed to enhance the university admission procedures by testing the dataset and features. This research evaluated the dataset using three machine learning models, which were KNN, DT, and SVM. The parameters for these models were set to their default values. The dataset used for this study comprised 19 features, which included high school courses, GAT, and AT. The classifier generated four classes that corresponded to specific domains, as listed in Table 3.

During the KNN processing, an accuracy rate of 100% was achieved when K = 1. However, increasing the value of K resulted in a reduction in accuracy. For instance, when

K = 3, the accuracy of the KNN model was 91%, as demonstrated in Figure 2. In Figure 2, the data points represent examples from the dataset, while the background areas indicate the four classes, with class 1 represented by white, class 2 represented by yellow, class 3 represented by pink, and class 4 represented by gray. The figure demonstrates that the majority of the points were classified accurately in the intended classes.

Figure 3 shows the result of the DT, where the accuracy rate reached 81%. It uses the same class backgrounds as Figure 2. Obviously, some points were classified on the borders between the classes, and one of the points was misclassified. Furthermore, when the SVM classifier was applied to the dataset, an accuracy rate of 75% was achieved, as illustrated in Figure 4, which also uses the same class backgrounds as Figure 2. The figure reveals that the SVM struggled with classification due to the narrow margins between classes. Additionally, a point was misclassified, and some points were located on the borders between the classes. Based on the results obtained by running the three classifiers, it was proven that the KNN machine learning model provided the highest accuracy rate and was the most suitable classifier for the proposed method.

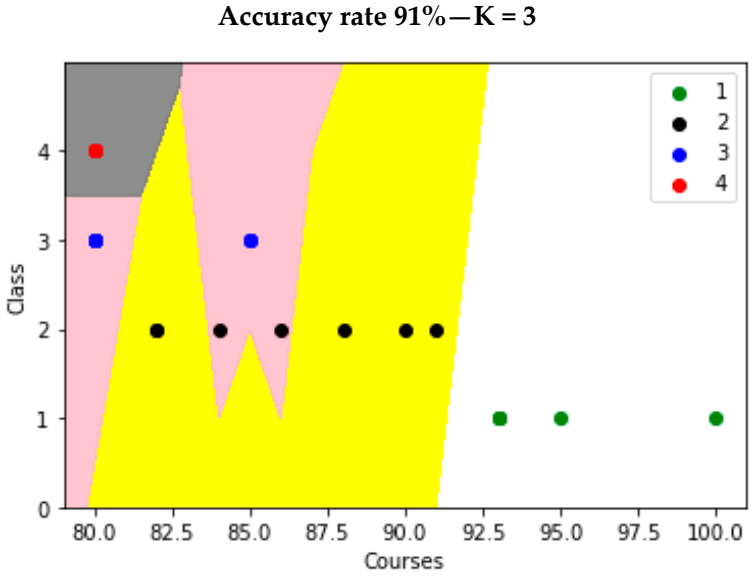

Figure 2. Illustrates KNN with an accuracy rate of 91% with K = 3.

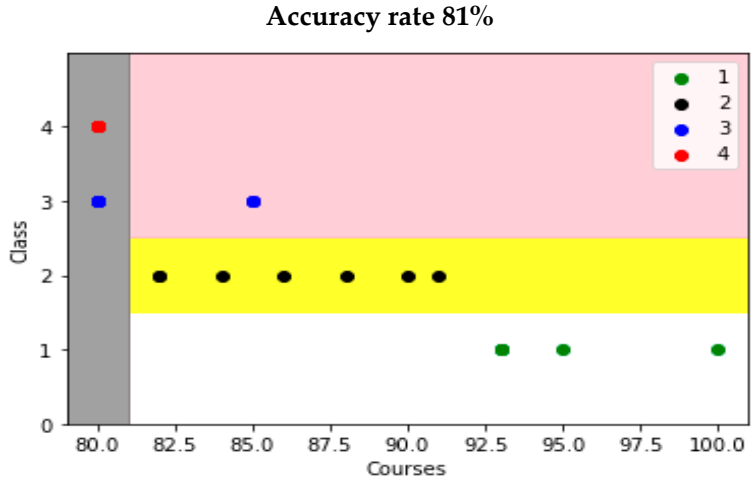

Figure 3. Illustrates DT with an accuracy rate of 81%.

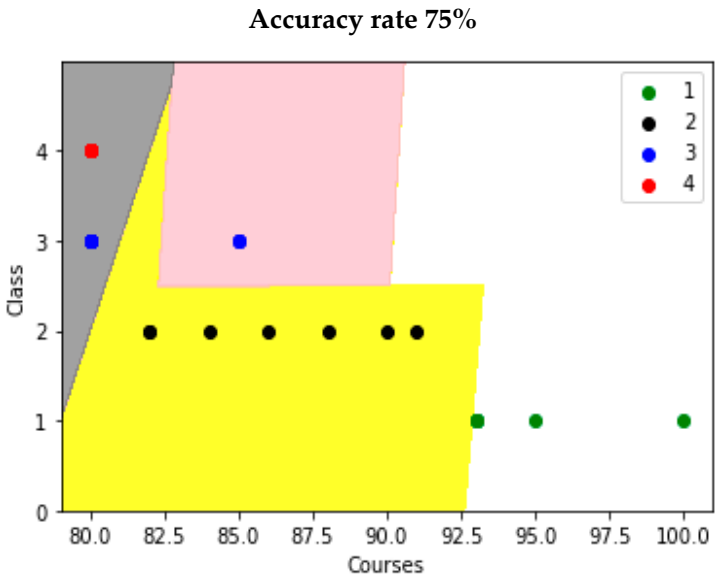

Figure 4. Illustrates SVM with an accuracy rate of 75%.

5. Discussion and Recommendation

According to the findings of this research, the current admission policies at universities are not accurate. These polices rely on some composed numerical scores, which are WS and QS. Such composed scores are calculated according to scores such as GPAH, GAT, and AT. These scores are calculated as a projection of a set of data, so they hide important information. Indeed, GPAH represents the scores of about 17 courses. GAT is a number that represents students' aptitude in mathematics and Arabic. AT is another number representing mathematics, physics, chemistry, and biology. However, the medical major gives more importance to English, chemistry, and biology. The scores of the three courses are hidden behind GPAH and AT.

The main issue with accumulated scores is that they do not accurately represent the values they are based on. For example, there are two stores (Store 1 and Store 2). The average sales figure for Store 1 for the last two years (y1 and y2) is 100,000, and the average sales figure for Store 2 for the last two years is also 100,000. However, the sales for Store 1 in the first year were Store1.y1 = 20,000 and the sales for Store 1 in the second year were Store1.y2 = 180,000, while the sales for Store 2 were Store2.y1 = 99,000 and Store2.y2 = 10,100. Obviously, the sales of Store 1 increased incredibly last year, while the sales of Store 2 were stable for the last two years. In this way, the average does not show some important details behind the value. The same issue is applicable to student GPAH and other scores, as shown in Figure 5. Figure 5 shows the course scores and GPAH for two students (GPAH = 85 for both students, represented by red line). Figure 5 highlights the distances between GPAH and individual course scores. It is clear that the GPAH for student 2 is more representative (represented by the green line), while the course scores for student 1 (represented by the blue line) have more distance from the GPAH. In fact, standardized measures can be used, such as variance and standard deviation. Clearly, the variance and standard deviation for student 2 are smaller than those for student 1, which means that the GPAH for student 2 is more representative of the course scores.

Another example demonstrating the issues of WS and QS is shown in Table 12. Table 12 illustrates that two students applied to medical college. Student x has GPAH = 95, GAT = 91, and AT = 89, while student y has GPAH = 94, GAT = 89, and AT = 89. Using the current enrollment policy, student x has a better chance to be accepted as a result of having higher scores for GPAH and GAT. In Table 6, we can find the important courses for each major. Obviously, student x has higher GPAH and GAT scores because of their high scores in mathematics, physics, Arabic, and others. However, considering the hidden (detailed) scores, student y is better in chemistry and biology, which are more important when joining

medical college. In addition, student y has a lower GAT score because of their performance in mathematics, while y's English score is better than x's. English scores are more important than mathematics scores for students who apply to medical college. Indeed, the same concept is applicable to engineering, computer science and information technology, and business colleges.

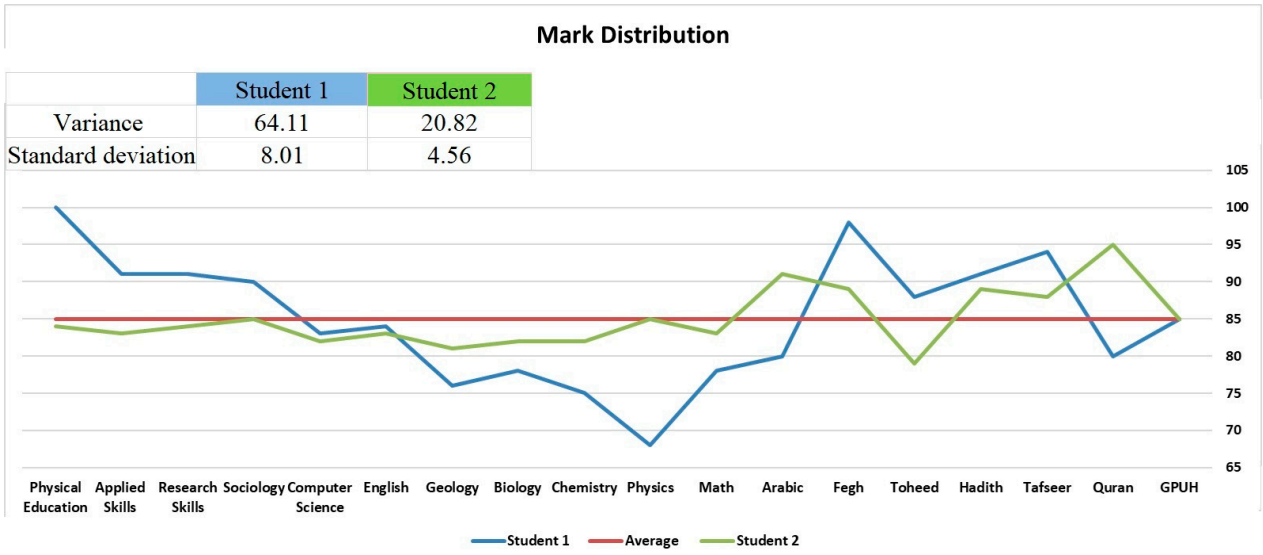

Figure 5. Illustrates distributions of course scores for two students and compares them to the GPAH, which is 85 for both students. It also calculates the variance and standard deviation.

On the other hand, the majors of education, Sharia, and arts and humanities do not consider AT. GPAH and GAT affect the enrollment in these three majors negatively. GPAH has the highest correlation with GPAU for education and Sharia majors, while WS has the highest correlation for arts and humanities majors (as shown previously in Table 5). However, GPAH and QS scores give more credit to mathematics, physics, chemistry, and biology, which are not relevant to theoretical majors. In addition, an English major requires good performance in English courses, while the scores of mathematics, physics, chemistry, and biology are not important. Moreover, AT tests students' aptitude in mathematics and Arabic, which are not important for an English major.

In practice, the proposed machine learning model can be deployed in the form of an application programming interface, where a student's scores can be entered as inputs and the output is a suitable class (major). In addition, the current admission exams and scores should be changed to show more hidden details. Therefore, we end this paper with the following recommendations:

1. To enhance the accuracy of the classification, the GAT score should be divided into two separate scores: GAT-Math and GAT-Language.
2. For students applying to an English major, a specific English test should be incorporated or the high school English course score should be used for evaluation.
3. The AT score should take into account the proficiency in the English language.
4. To provide a more detailed evaluation, the AT score should be divided into five separate scores: AT-Math, AT-Physics, AT-Chemistry, AT-Biology, and AT-English.
5. The WS and QS should be replaced by a relative course-based model, as depicted in Table 1, rather than using them for evaluation purposes.

Table 12. Example of scores of two students showing the detail scores.

Student x			**Student y**		
Mathematics		99	Mathematics		90
Physics		100	Physics		91
Chemistry		90	Chemistry		100
Biology		91	Biology		100
English		92	English		97
Arabic		100	Arabic		94
GPAH		95	**GPAH**		94
GAT = 91	Math	91	**GAT = 89**	Math	87
	Arabic	91		Arabic	93
AT = 89	Math	92	**AT = 89**	Math	88
	Physics	90		Physics	86
	Chemist	86		Chemist	90
	Biology	88		Biology	92

6. Conclusions

In conclusion, this research enhances the process of student admission at universities. This research investigated the current acceptance procedures using GPAH, QS, and WS. This paper highlighted the issue of free education, a large number of applicants, and high failure rates. We first considered experts' points of view using a survey. Then, we used data mining analysis to investigate the relations between student performance and the admission features. Four data mining techniques were applied to real-world data. We created suitable methods such as modified Jaccard and approximated modified Jaccard analyses that fit our problem. The results show that the current admission procedures are not adjusted for all majors, as the relations vary from one major to another. Next, we determined the importance of considering the individual grades of high school courses and linked them to each major. Finally, we suggested using successful students' records as training examples for each major and we applied machine learning models. In the future, this model can be the basis of a smart application to guide students and universities in the admission process.

Author Contributions: Conceptualization, B.A., M.B. and A.A.; methodology, B.A., M.B. and A.A.; software, B.A., M.B. and A.A.; validation, B.A., M.B. and A.A.; formal analysis, B.A., M.B. and A.A.; investigation, B.A., M.B. and A.A.; resources, B.A., M.B. and A.A.; data curation, B.A.; writing—original draft preparation, M.B. and A.A.; writing—review and editing, B.A.; visualization, B.A., M.B. and A.A.; supervision, B.A.; project administration, B.A.; funding acquisition, B.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend the appreciation to the Deputyship for Research Innovation, Ministry of Education in Saudi Arabia for funding this research work and covering the APC through project number ISP-2024.

Data Availability Statement: Restrictions apply to the availability of these data. Data was obtained from Jazan University with the permission of Jazan University HABO-10-Z-001, with Reference No: REC-43/05/094.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Assiri, B.; Bashraheel, M.; Alsuri, A. Improve the Accuracy of Students Admission at Universities Using Machine Learning Techniques. In Proceedings of the 7th International Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 1–3 March 2022; pp. 127–132.
2. Kumar, M.; Singh, A.J.; Handa, D. Literature survey on student's performance prediction in education using data mining techniques. *Int. J. Educ. Manag. Eng.* **2017**, *7*, 40–49. [CrossRef]
3. Kumar, M.; Singh, A.J.; Handa, D. Literature survey on educational dropout prediction. *Int. J. Educ. Manag. Eng.* **2017**, *7*, 8. [CrossRef]
4. Miller, G.A. Undergraduates' Decision-making Processes in College Major Selection. Ph.D. Thesis, Hofstra University, Hempstead, NY, USA, 2018.
5. Albakri, B.; Abuhamdeiyeh, S.; Mousa, A. Rule-Based Expert System to Lead Freshmen Students in Choosing a Suitable College Major. In Proceedings of the 10th IADIS International Conference on Information Systems, Budapest, Hungary, 10–12 April 2017.
6. Whitehead, A. Examining influence of family, friends, and educators on first-year college student selection STEM major selection. *J. Mason Grad. Res.* **2018**, *5*, 58–84.
7. Casinillo, L. Factors affecting the failure rate in mathematics: The case of Visayas State University (VSU). *Rev. Socio-Econ. Res. Dev. Stud.* **2019**, *3*, 1–18.
8. Bennedsen, J.; Caspersen, M.E. Failure rates in introductory programming. *AcM SIGcSE Bull.* **2007**, *39*, 32–36. [CrossRef]
9. Hassan, S.M.; Al-Razgan, M. Pre-university exams effect on students GPA: A case study in IT department. *Procedia Comput. Sci.* **2016**, *82*, 127–131. [CrossRef]
10. Alghamdi, A.; Hamdan, K.; Al-Hattami, A.A. The Accuracy of Predicting University Students' Academic Success. *J. Saudi Educ. Psychol. Assoc.* **2016**, *186*, 1–10.
11. Malgwi, C.A.; Howe, M.A.; Burnaby, P.A. Influences on students' choice of college major. *J. Educ. Bus.* **2005**, *80*, 275–282. [CrossRef]
12. Montmarquette, C.; Cannings, K.; Mahseredjian, S. How do young people choose college majors? *Econ. Educ. Rev.* **2002**, *21*, 543–556. [CrossRef]
13. Niwattanakul, S.; Singthongchai, J.; Naenudorn, E.; Wanapu, S. Using of Jaccard coefficient for keywords similarity. In Proceedings of the International Multiconference of Engineers and Computer Scientists, Hong Kong, China, 13–15 March 2013; Volume 1.
14. Thanh Noi, P.; Kappas, M. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors* **2018**, *18*, 18. [CrossRef]
15. Decision Trees. Scikit. (n.d.). Available online: https://scikit-learn.org/stable/modules/tree.html (accessed on 3 November 2021).
16. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
17. Boyd, R.L.; Pennebaker, J.W. Building a Personalized College Major Selection Web Page. *PsyArXiv* **2018**. [CrossRef]
18. Iraji, M.S.; Aboutalebi, M.; Seyedaghaee, N.R.; Tosinia, A. Students classification with adaptive neuro fuzzy. *Int. J. Mod. Educ. Comput. Sci.* **2012**, *4*, 42. [CrossRef]
19. Adu, E.O.; Oshati, T. Psychological Variables as Correlate of Students' Academic Achievement in Secondary School Economics in Oyo State Nigeria. *J. Psychol.* **2014**, *5*, 125–132. [CrossRef]
20. Kiaghadi, M.; Hoseinpour, P. University admission process: A prescriptive analytics approach. *Artif. Intell. Rev.* **2023**, *56*, 233–256. [CrossRef]
21. Abu Saa, A.; Al-Emran, M.; Shaalan, K. Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technol. Knowl. Learn.* **2019**, *24*, 567–598. [CrossRef]
22. Ribera, A.K.; Miller, A.L.; Dumford, A.D. Sense of peer belonging and institutional acceptance in the first year: The role of high-impact practices. *J. Coll. Stud. Dev.* **2017**, *58*, 545–563. [CrossRef]
23. Kutscher, M.; Nath, S.; Urzúa, S. Centralized admission systems and school segregation: Evidence from a national reform. *J. Public Econ.* **2023**, *221*, 104863. [CrossRef]
24. Noble, J.; Sawyer, R. *Predicting Grades in Specific College Freshman Courses from ACT Test Scores and Self-Reported High School Grades*; ACT Research Report Series; American College Testing Program: Iowa City, IA, USA, 1999.
25. Noble, J.; Sawyer, R. *Predicting College Grades from ACT Assessment Scores and High School Course Work and Grade Information*; American College Testing Program: Iowa City, IA, USA, 1991.
26. Betts, J.R.; Morell, D. The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *J. Hum. Resour.* **1999**, *34*, 268–293. [CrossRef]
27. Sadiq, F.A.; Mitlif, R.J.; Abbas, J. A computational mechanism for making admission decisions in the centralized admission system. In *AIP Conference Proceedings*; AIP Publishing: Melville, NY, USA, 2023; Volume 2414.
28. Lee, F.X.; Suen, W. Gaming a Selective Admissions System. *Int. Econ. Rev.* **2023**, *64*, 413–443. [CrossRef]
29. Xu, J.; Moon, K.H.; Van Der Schaar, M. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 742–753. [CrossRef]
30. Hossain, M.A.; Assiri, B. Emotion specific human face authentication based on infrared thermal image. In Proceedings of the 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 13–15 October 2020.

31. Numan, M.; Subhan, F.; Khan, W.Z.; Assiri, B.; Armi, N. Well-organized bully leader election algorithm for distributed system. In Proceedings of the 2018 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), Serpong, Indonesia, 1–2 November 2018; pp. 5–10.
32. Fadhilah, A.; Ismail, N.H.; Abdul Aziz, A. The prediction of students' academic performance using classification data mining techniques. *Appl. Math. Sci.* **2015**, *9*, 6415–6426.
33. Jaccard, P. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* **1908**, *44*, 223–270.
34. Jaccard, P. The distribution of the flora of the alpine zone. *New Phytol.* **1918**, *11*, 37–50. [CrossRef]
35. Albatineh, A.N.; Niewiadomska-Bugaj, M.; Mihalko, D.P. On similarity indices and correction for chance agreement. *J. Classif.* **2016**, *23*, 301–313. [CrossRef]
36. Albatineh, A.N.; Niewiadomska-Bugaj, M. Correcting Jaccard and other similarity indices for chance agreement in cluster analysis. *Adv. Data Anal. Classif.* **2011**, *5*, 179–200. [CrossRef]
37. Eelbode, T.; Bertels, J.; Berman, M.; Vandermeulen, D.; Maes, F.; Bisschops, R.; Blaschko, M.B. Optimization for medical image segmentation: Theory and practice when evaluating with dice score or jaccard index. *IEEE Trans. Med. Imaging* **2020**, *39*, 3679–3690. [CrossRef]
38. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4.