Journal of Information Systems Engineering and Management
2025, 10(15s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Employing Supervised Learning Techniques for College Major Prediction: Empowering Decision-Making in University Admission Systems

Monira Aloud [1], Nora Alkhamees [1], Nora Almezeini [1], and May AlYahya[2]
[1]Department of Management Information System, College of Business Administration, King Saud University, Saudi Arabia
[2] Department of Management, College of Business Administration, King Saud University, Saudi Arabia

ARTICLE INFO

Received: 02 Dec 2024

Revised: 25 Jan 2025

Accepted: 02 Feb 2025

ABSTRACT

Introduction: Choosing an appropriate study major is a significant challenge for prospective university students. Many students enroll in programs misaligned with their interests or abilities, leading to major changes, extended study duration, financial losses, and the displacement of other potential students. Addressing this issue requires a systematic approach to assist students in making informed decisions.

Objectives: This study aims to develop an intelligent decision support model to predict the most suitable undergraduate major based on students' academic performance, interpersonal influences, and external factors. In additiona, it seeks to identify the key variables that influence major selection.

Methods: A dataset was collected from 430 participants through a structured questionnaire comprising 18 questions. Various supervised learning techniques, including Support Vector Machine (SVM), Random Forest, and Naïve Bayes, were employed to predict the most suitable major. Furthermore, Random Forest and XGBoost were used to analyze relationships between study majors and input variables to determine the most influential factors.

Results: The models demonstrated that SVM, Random Forest, and Naïve Bayes provided the most accurate predictions. The analysis identified creativity, technical skills, and GPA as the top three factors influencing students' study major selection.

Conclusions: The findings emphasize the importance of data-driven decision-making in academic advising. By leveraging machine learning techniques, institutions and career counselors can guide students toward suitable majors, reducing the likelihood of major changes and optimizing educational resources.

Keywords: data-mining techniques; supervised learning techniques; admission prediction; preadmission criteria; student major.

INTRODUCTION

Today, high school students face challenges when selecting the most appropriate university major. Universities should thus strive to maintain dynamic and efficient admission criteria based on valid and reliable evidence that can help potential students enrol in majors based on their capabilities and interests. For college graduates, choosing a study major impacts their career options, especially in professions such as engineering, which require significant investments in human capital [1].

Choosing the right undergraduate major has become a global educational and economic problem that needs further study. For instance, although the United States spends over $9 billion on first-year undergraduate students, 30% do not return for their second year [2]. According to several studies, 75% of dropouts occur during the first few weeks of a student's first year [3]. Of the 98% of students who started a bachelor's program in 2011–2012, 33% switched their study program by 2014, in their third year, according to the US Department of Education (NCES) [4]. The statistics also indicate that roughly 1 in 10 students switched majors more than once [4]. These numbers necessitate making every effort to support students in selecting the right major and universities in reviewing their admission criteria.

Social scientists have been studying for a long time how college major selection affects results in the job market [1]. A considerable number of studies have been done to understand the forces behind the students' major choice [5,6]. The decision to enroll in a study major is significant for several reasons, including emotional well-being [1]. Understanding how major decisions drive changes in the market skill demands is indeed crucial to being studied in the macroeconomy as it influences the workforce skill composition.

Research interests in applying machine learning (ML) and data mining (DM) techniques in education have grown recently, especially among higher education institutions. According to a recent study, rather than being based on knowledge-rich facts, education decisions are typically based primarily on the perceptions and experiences of educational management stakeholders and students [7]. Over the last decade, research on ML, and education DM has played an important part in investigating issues, such as insights on student performance. In addition, various research papers have used ML and DM techniques to predict student enrollment, acceptance to universities, and admission into their selected colleges/majors [7-10]. Despite the large number of social studies on student study major choice and the factors influencing it [1, 5,6], few [7-11] employ ML and DM techniques using a dataset of student data.

Thus, this study investigated the ability to use supervised learning techniques to develop an intelligent decision support model that predicts the student's major based on their preadmission profile, interpersonal effects, and external influences. Predicting a student's study major is a classification problem in DM; therefore, this study identified the input variables behind the right selection of the student's major. We trained and tested seven supervised learning techniques on a student dataset. The dataset was gathered from an online survey of King Saud University's (KSU) College of Business Administration (CBA) undergraduate students in October 2022. The dataset contains student admission scores, including overall higher school grades and entry test scores. It also includes information about the family's educational background, working experience, interpersonal effects, and additional external influences.

This study contributes to literature in two ways. First, using supervised learning classification techniques (seven supervised learning techniques), we develop and evaluate a model that can predict the most applicable student major based on the student’s preadmission profiles, interpersonal effects, and external influences. It is important to note that variables such as interpersonal effects and extracurricular activities are not considered in the admissions process for undergraduate programs in Saudi universities. These underused input variables are the motivation of this research. In addition, we identify the best performance of the seven supervised learning techniques to forecast the appropriate student study major using accuracy, precision, recall, and F1-Measure metrics.

The study's second contribution is examining the importance of the factors used in predicting their major. Therefore, this study identifies the student preadmission data, interpersonal effects, and the external influences that most precisely predict student majors, so that the admission office in universities can update the admission criteria accordingly. The majority of the few published research conducted in Saudi Arabia on this topic have been limited to predicting students' performance concerning their undergraduate course grades and admission test scores [8].

The rest of this paper is structured as follows. Section 2 offers a background of the study domain and presents the related work in the literature. Section 3 defines the research methodology and the collected dataset. Section 4 presents the experimental results. Finally, the conclusion and possible future research extensions are presented in Section 5.

BACKGROUND AND RELATED WORKS

Admission Process in Saudi Arabia Universities

Saudi universities have central rules and systems for undergraduate admissions regulated by the Saudi Ministry of Education [12]. To apply to an undergraduate degree, an applicant must submit the scores from two standardized exams administered by the National Assessment Center for Higher Education1. These two standardized exams are the General Aptitude Test (GAT) and the Scholastic Achievement Admission Test (SAAT). The GAT measures students' knowledge, rational thinking, problem-solving, and inductive/deductive abilities through mathematical and verbal skills. In contrast, the SAAT measures ability, knowledge, and reasoning in biology, chemistry, physics, and mathematics [12].

The weighted average result of two or three factors, including the High School Grade Average (HSGA), the SAAT, and the GAT, determines admissions to a specific undergraduate program. Each university gives these factors different weights based on the number of seats available each year. In addition, each undergraduate program establishes a cut-off point for weighted averages. For instance, in the academic year 2022–2023, the weights of these three admission criteria for college of business administration in the four most famous Saudi public universities were, for three criteria (HSGA, SAAT, and GAT), respectively, 50%, 0%, and 50% at KSU [13]; 40%, 30%, and 30% at King Abdulaziz University (KAU) [14]; 10%, 40%, and 50% at King Fahd University of Petroleum & Minerals (KFUPM) [15]; and 30%, 40%, and 30% at Princess Nourah Bint Abdulrahman University (PNU) [16]. Thus, it is clear that each university has different requirements for admission.

However, determining appropriate weights for these three criteria is difficult due to an absence of current findings and research papers in Saudi Arabia analyzing the association between these criteria and student university performance. The work by [8] determined the appropriate weights for these three criteria for admission to the college of Computer Science at PNU. This has been done using DM techniques to identify the relationship on a dataset of 2000 students. This study will support the selection of students major using a more suitable weighting scheme that gives proper criteria priority.

Student Major Specialization

Social scientists have long studied how major college selection affects job market outcomes [1]. The works by [1,5,6] offer extensive reviews of the studies in this area of research. These studies have examined the impact of several factors such as sex, employment, earnings, early work, family background, peers, and role models, on choosing a study major.

The significance of the high school curriculum and its connection to major program choice is emphasized by [5]. Student ability established prior to college is likely to be a key factor in the college major choice [1]. Previous studies have shown significant differences in students' ability to meet the requirements of college preadmission across several majors (as measured by preadmission exam results such as the SAT math and verbal scores) [17].

Several studies in literature have extensively examined students' intentions to pursue science, technology, engineering, and mathematics (STEM) college majors [45-47]. Factors such as students' demographic and family backgrounds play a significant role, particularly among female students [47]. Additionally, parents' occupations and engagement have been found to influence students' STEM major choice [48, 49]. Other influential factors include students' academic grades and interest in STEM subjects [45-47], along with their high school learning experiences.

In selecting a college major, students are found to be affected by the potential job salaries instead of merely the initial salaries [18]. According to [19], the likelihood of a student majoring in a field associated with the major of a close relative is clearly associated with the relative's earnings at the time the student chooses the major. The impact of confidence and background variables on the selection of STEM majors are found to be very critical [46]. The study conducted by [46] emphasizes the significance of students' confidence levels in their academic and mathematical abilities when it comes to making initial choices for STEM majors. These findings provide valuable insights for educators, counselors, and policymakers who are interested in promoting and supporting STEM-related majors and careers. By understanding the impact of confidence on major selection, stakeholders can develop strategies and interventions that enhance students' confidence in their abilities, ultimately encouraging their pursuit of STEM fields.

The study by [20] used a survey approach to examine the crucial factors that impact a student's choice for a study program in business college. There were 37 factors in total, including interpersonal effects (e.g., interest in the study program) and external influences (e.g., job availabilities). The results indicate that previous studies may have

[1] https://etec.gov.sa/en/About/Centers/Pages/qiyas.aspx

underestimated the overall impact of interpersonal factors. Students pointed out that abilities usually used in conjunction with a certain study program influence their choice of a college major. For instance, the marketing major pointed out that communication and creativity skills are important factors, and the Management Information System (MIS) major indicated technical skills as necessary.

By exploring how undergraduate students' evaluation criteria, social influences, and instructional methods link to undergraduate students' choices for a business major, the study by [21] built on previous works in the literature considering students' choice of a study major. The findings indicate that while attitudes did not significantly influence students' intentions to choose a business major, both of the subjective norm and perceived behavioral are important factors. There were no changes found in any of the teaching methods that affected students' attitudes toward pursuing a business major.

Supervised Learning Techniques

In this section, we will provide an overview of the seven models utilized in this study.

Naïve Bayes (NB): is one of the simplest and most effective classifiers in terms of predictive performance. It is based on the Bayes Theorem and assumes that the value of any attribute on a particular class is conditionally separate from any other attributes' values [34]. This assumption is called class conditional independence. This implies that a NB model maintains a record of how frequently a value from the target field occurs alongside a value from the input field.

Logistic Regression (LR): This classifier is based on the algorithm for kernel logistic regression proposed by Keerthi et al. [35]. It uses the myKLR tool implemented by Stefan Rueping [36]. It provides fast and good results for many tasks and can be used for regression and classification purposes.

Deep Learning (DL): DL is an advanced ML technique and is based on multi-layer artificial neural network. It employs sophisticated deep neural network algorithms inspired by the way the human brain functions. This model performs significantly better than the ML techniques when analyzing large datasets with numerous features when dealing with unstructured data [37].

Decision Trees (DT): Simple and widely used predictive model. It classifies data to produce a collection of nodes in a tree-shaped model. Each node in the DT symbolizes a splitting rule for a specific attribute, and its branches represent the possible values. A DT uses a series of if–else criteria to represent and classify data. This model performs well especially for categorical and numerical attributes [8].

Random Forest (RF): A classifier developed by Breiman in 2001 [38], which constructs a large number of DTs at training time and outputs the majority votes over the forecasts provided by the trees. RF operates rapidly and effectively across an extensive collection of DTs with different subsets of a dataset to increase its forecast performance. It is an effective learning model employing prediction accuracy [39].

XGBoost: Gradient boosted trees is a feature-based classification model proposed by Friedman in 2001 [40]. It returns a prediction model that consists of an ensemble of weak prediction models, usually DTs. It is often employed with fixed-size DTs.

SVM: This classifier was developed in 1995 [41]. It is based on finding the optimal distance hyperplane between two classes. SVM identifies a hyperplane that clearly divides data points into several classes in an N-dimensional space (where N represents the number of features). Decision boundaries, known as hyperplanes, assist in categorizing data points [42].

Supervised Learning Techniques Used in Education Predictive Model

Classification is a DM technique that is also known as supervised learning. Several studies have used supervised learning techniques to examine the impact of different input variables in choosing the right students' majors. These studies used Decision Tree (DT) [7, 22-24, 47], k-nearest neighbor algorithm [24], associate rules [22], Support Vector Machine (SVM) algorithm [7, 25], rule-based classification [26], and Random Forest (RF) [7, 24, 25, 27, 45] among others. The majority of the research papers in the literature have focused on students in their preparatory year (first year) to use the Grade Point Average (GPA) [25] in predicting students' majors. Other studies have focused on sub-majors in a particular field, such as first-year engineering students [24, 25] and information systems students [23]. In addition, several studies [45-47] have explored students' intentions to select STEM majors in college.

According to the used set of input variables, previous studies have used the following factors: (i) final scores of high school and preadmission tests [7, 26, 45, 47], (ii) first-year GPA [25], final grades of courses at first year [22, 23], (iii) student interest in the filed [7], (iv) student's interpersonal skills [24, 26, 27], (v) work experience [7, 23], socio-economic background [22], and gender [45].

The study by [22] used a DT and an association rule to predict the best learning institution for specific students. The information was collected from 1,109 students from three Thailand universities. The findings indicated that four factors were important. These four factors are the reputation of the university, confidence in the university, learner abilities, and family financial income.

The study by [25] applied different ML techniques to recommend a suitable major for students in their first year. In particular, the study focused on a small engineering college sample. Moreover, it aimed to select for each student an engineering major from seven majors based on their first-year performance. The recommendation model is fed with data from each engineering major. Afterwards, an ML algorithm with the best performance was selected based on the relevant set of attributes for that major. The results demonstrated that the proposed system accurately and efficiently recommends the student to the appropriate engineering major(s).

The study by [27] analyzed the differences among students of nine majors in Bangladesh. It proposed the suitable major among these nine majors for future students by evaluating their grades, personality characteristics, and other factors. The data was gathered from 103 respondents. They used their college grades and Big-5 behaviour attributes to build supervised learning models using hierarchical classification using RF Classifier and a one-level RF Classifier. Furthermore, the RF classifiers have been used to analyze factors of students' personality and intelligence over several majors to predict the right study majors based on their academic performance, interpersonal skills, and intelligence. The accuracy of the results at level one 96.1% and 94.72% at the second level, correspondingly. In addition, the study examined whether the model could suggest a student for future postgraduates.

Based on current job markets and the applicant's work experience, the study by [7] evaluated several ML approaches (DT, Extra tree classifiers, RF, XGBoosting, and SVM) to predict students' right undergraduate major before admission at the undergraduate level. The relationship between a student's major and other input variables is also evaluated using statistical analysis. The results show that higher student marks in higher secondary, student interest in the field, and preadmission test scores play an important role in the student's major.

The study discussed in [45] aimed to identify the primary factors that impact high school students' decision to pursue an engineering major. Through an analysis of comprehensive data from the High School Longitudinal Study of 2009 using the RF method, the researchers ranked the predictive power of various factors related to high school experiences. The findings revealed that student gender emerged as the most influential factor, followed by high school math achievement and student beliefs and interests in math and science. These results indicate the need for further investigation and systemic involvements to address gender disparities in engineering major selection. Additionally, the study emphasizes the significance of promoting high school math achievement and interest in math and science to encourage more students to consider and pursue engineering as a career path.

Therefore, this paper addressed the major selection problem by recommending the most appropriate study major for students based on their preadmission profile, interpersonal effects, and external influences. Consequently, it was desired to develop an intelligent form of decision support system to support the student in enrolling in a suitable study major. Hence, the current study examined the potential of using supervised learning techniques to build an intelligent decision support model to predict the student study major. In doing so, this study seeks to identify the set of input variables (factors) that are most strongly linked and tied to selecting the appropriate study major for students.

DATA AND METHODOLOGY

The survey approach was selected as the preferred data collection method due to its ability to capture comprehensive information about the factors influencing students' choice of study majors. The survey participants were specifically targeted from the CBA at KSU, ensuring that the collected data would be representative of the student population under investigation. This approach allowed for a systematic exploration of the various factors that influence students' decision-making processes when selecting a business major.

At KSU, students admitted to the CBA undergo a comprehensive two-year program focused on general business studies. This curriculum is designed to provide students with a solid foundation in fundamental business principles and concepts. As they progress through this initial phase of their academic journey, students are faced with an important milestone at the end of their second year—making a definitive decision regarding their choice of study major. Within the CBA, students have the opportunity to specialize in one of six distinct majors: accounting, economics, finance, management, marketing, and Management Information Systems (MIS).

The timing of this decision-making process is considered critical, as it offers a unique opportunity to capture the factors that influence students' choices during this pivotal period. Therefore, our survey specifically targets CBA students who have completed the first two years of their studies and have already specialized in one of the available majors. By focusing on this subset of students, we aim to gain valuable insights into the factors that impact their decision-making processes, considering they have already made a commitment to a particular area of business study. This targeted approach allows us to delve deeper into the factors influencing students' choices within specific majors and explore potential variations across different academic paths.

In order to develop a reliable predictive model for determining the most suitable study major for individual students, this study conducted an evaluation of seven different supervised learning models: Naïve Bayes (NB), Logistic Regression (LR), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosting Tree (XGBoost), and Support Vector Machine (SVM). Each of these models possesses unique characteristics and capabilities, enabling a comprehensive analysis of the collected data. The selection of these models was based on their established effectiveness in handling categorical data and uncovering hidden patterns [7, 22-27, 45-47]. Specifically, Naïve Bayes, DT, RF, XGBoost, and LR were chosen for their suitability in datasets containing categorical variables such as pre-admission tests, family educational background, and job security. These models incorporate specialized mechanisms to handle categorical data efficiently, such as leveraging conditional probabilities, partitioning data based on discrete attribute values, exploring random feature subsets, employing carefully crafted splitting strategies, and encoding categorical variables as binary dummy variables. Furthermore, these advanced supervised learning techniques excel at uncovering complicated patterns and relationships within the data. The DL models, including neural networks, are proficient at capturing hidden patterns, while tree-based models like DT, RF, and XGBoost excel at identifying non-linear relationships and feature interactions. LR models are particularly valuable for identifying the significant factors that influence prediction outcomes.

Data Collection

We began developing our survey with an extensive review of existing literature on the determinants influencing students' selection of study majors. We specifically focused on the field of business majors, considering it as the focal point of our study. Several factors were collected from previous well-conducted studies, notably those conducted by [28-32]. Drawing from this body of literature, we created a list of potential factors that should be included in our survey questionnaire [28-32].

To ensure the questionnaire's validity and effectiveness, we collaborated with a diverse group of esteemed faculty members specializing in various business disciplines. Their expertise played a pivotal role in meticulously examining and evaluating the survey questionnaire. As a result of this rigorous examination, selected questions were considered redundant and subsequently eliminated, while other questions underwent revisions to enhance clarity and precision.

The finalized version of the survey encompassed a total of 18 factors. These factors were classified into three distinct categories: (a) four preadmission scores and tests, specifically the High School Grade Average (HSGA), General Aptitude Test (GAT), Subject Area Achievement Test (SAAT), and Grade Point Average (GPA); (b) eight interpersonal effects; and (c) six external influences. Participants were requested to evaluate the relevance or impact of each factor on a 5-point scale, ranging from "totally irrelevant" (1) to "very relevant" (5). Table 1 provides a comprehensive overview of the data fields employed in this study, shedding light on the specific variables considered.

Following the necessary minor adjustments and refinements, the survey was thoroughly disseminated to the target population of CBA students at KSU. To ensure maximum participation, multiple channels were utilized for survey distribution. Firstly, the survey was distributed directly to CBA students via their KSU student email accounts, enabling widespread access and encouraging engagement. Secondly, the survey was shared through CBA students' clubs, leveraging these communities to reach a diverse range of students and collect a good representation of the student body. Lastly, the instructors themselves played a crucial role in facilitating survey participation by actively encouraging their students to complete the survey.

A total of 435 students actively participated in the survey. These students were drawn from several patches and represented a diverse grouping of majors within the CBA, as shown in Table 2. The inclusion of students from different patches ensured a broad cross-section of the student population, allowing for a robust analysis of the factors influencing study major selection across different contexts. Additionally, the participation of students from various majors enabled a comprehensive exploration of the decision-making processes across different fields of business study.

Table 1. Variables in dataset

Factor	Data Filed	Category
Admission Test Scores	High School Grade Average (HSGA)	Numeric
	General Aptitude Test (GAT)	Categorical
	Scholastic Achievement Admission Test (SAAT)	Categorical
	GPA	Categorical
Interpersonal Effects	Interest in the field	Categorical
	Opportunity to use creativity	Categorical
	Opportunity to use technical skills	Categorical
	Opportunity to use communication skills	Categorical
	Opportunity to own a business	Categorical
	Easy Major	Categorical
External Influences	Job security	Categorical
	Job availability	Categorical
	Prestige associated with major	Categorical
	Previous work or course experience in a major	Categorical
	Department reputation	Categorical
	Influence of parents	Categorical
	Influence of relatives (brother, sister, uncle, cousin)	Categorical
	Influence of friends or graduates	Categorical

Table 2. Demographic Data

Major	n	%
Accounting	61	14.02%
Economic	59	13.56%
Finance	89	20.46%
Marketing	73	16.78%

Management	69	15.86%
MIS	84	19.31%

Reliability

Reliability is a crucial aspect when assessing the stability and consistency of test results. It determines the extent to which a measurement consistently produces similar outcomes. A high level of reliability indicates that the measurement is dependable. To evaluate the internal uniformity and reliability of the study's dependent variable, the coefficient alpha measure, widely utilized in statistics, is employed.

According to previous research [33], a coefficient alpha value of 0.7 suggests satisfactory internal consistency. In this study, Table 3 presents the coefficient alpha values for the scaled variables, which encompass 18 questions and were responded to by 435 participants. Notably, the coefficient alpha for the scaled variables is calculated to be 0.85. This demonstrates a high level of internal consistency and reliability, indicating that the measurement of the study's dependent variable yields dependable and consistent results.

Table 3. Questionnaire Reliability

Cronbach's Alpha	No. of Items	No. of Respondents	% of Respondent
0.85	20	435	100%

RESULTS AND DISCUSSION

Prediction Model Results

We aim to develop a student major prediction model that is capable of accurately predicting student major in business schools based on student scores and other external features. We employed different ML techniques (supervised learning methods to be more particular), namely, we used NB, RL, DL, DT, RF, XGBoost, and SVM.

To ensure the reliability and accuracy of our model, we utilized a stratified sampling approach to divide the dataset into training and testing sets. This technique enables us to obtain representative samples that effectively capture the characteristics of the entire dataset. By dividing the input records into distinct subsets known as strata, we are able to minimize bias. Subsequently, we randomly sampled data from each stratum to ensure a comprehensive representation of the dataset [42, 50 -52].

Table 4 presents the results using different metrics precision, recall, F-measure, and accuracy. We think precision is the most critical and important metric, as it gives an impression regarding the quality of major prediction rather than quantity. This model aims to help predict the student major (precision) rather than the number of students who will specialize in a certain major (recall).

Table 4. Student Major Prediction Model Results by Different ML Techniques

	Precision	**Recall**	**F-measure**	**Accuracy**
NB	68.8	64	66.3	66.5
LR	52.16	46.5	49.16	48.2
DL	64	62.7	63.34	62.6
DT	52.7	50.17	51.4	51.4
RF	69.7	66.7	68.16	65.8
GBOOST	64.8	65.3	65.	65.8
SVM	70.5	70.17	70.33	70.6

Table 4 shows that SVM, RF, and NB achieved the best performance. The SVM achieved more than 70%, whereas RF and NB scored nearly 70%. Table 5 gives a closer look at the results, and we wanted to further explain the performance of NB, RF, and SVM by showing the confusion matrix. It can be found that the performance of evaluation metrics (especially the precision) is high for all majors except accounting and finance in some cases. This is due to the close relationship between accounting and finance with regard to their shared characteristics, making the prediction mission even more challenging. Features such as job availability and high salaries correlate with finance and accounting majors. To illustrate more, by the NB algorithm, almost half of the accounting major students were predicted as finance majors (7 out of 16 accounting major students were predicted as finance). The same applies to other ML techniques in accounting and finance. In contrast, the high performance in majors that has no conflicts with other majors (i.e., has no shared features). For example, using the SVM technique, the MIS major performance score was >82%, the economic score was 77%, and the marketing score was 83%. This clearly indicates the effect of features and their relation to each other in predicting student majors. The major management performance in some cases was also low (refer to SVM in Table 5), the prediction of management score was 48% (out of 25 predicted as a management major, only 12 were true management, i.e., almost half of the instances), while six were accounting, two were finance, two were marketing, two were MIS, and finally, one was economic majored). Shared characteristics between the management major and all other majors can cause this. The management was the only major that had false predictions associated with all other majors, which also explains the low prediction performance in the management major. Additionally, the percentage of management courses in CBA BSc curriculum courses in KSU (which was the data source in this study) constitutes 40% of all business courses in all CBA majors. This is another explanation for the relationship between management and all other majors.

Furthermore, Table 6 compares the precision results of the student major prediction model for all six majors with and without closely related majors. Thus, it presents the precision score for all six majors, for all majors without considering accounting and finance majors, and finally, all majors without finance, accounting, and management to present an increase in the performance percentage without considering close related majors. Obviously, the precision score using the SVM jumped from 71% to almost 75% and then to 82%.

Table 5. Confusion Matrix for NB, RF, and SVM

ML Technique	**Confusion Matrix**								
NB			**Actual**						
			MIS	Eco	Mkt	Fin	Mgt	Acc	Precision
	Predicted	MIS	22	1	2	3	1	0	75.86
		Eco	0	6	0	0	3	0	66.67
		Mkt	0	3	16	0	2	0	76.19
		Fin	0	2	2	13	1	7	52.00
		Mgt	1	3	4	4	18	2	56.25
		Acc	0	0	0	2	0	7	77.78
		Recall	95.65	40.00	66.67	59.09	72.00	43.75	
RF			**Actual**						
	Predicted		MIS	Eco	Mkt	Fin	Mgt	Acc	Precision
		MIS	22	0	0	8	0	0	73.33
		Eco	0	14	0	1	1	1	82.35
		Mkt	0	0	15	2	2	0	78.95
		Fin	3	2	0	11	1	7	45.83
		Mgt	1	4	1	3	15	6	50.00
		Acc	0	0	0	1	0	4	80.00
		Recall	84.62	70.00	66.67	93.75	78.95	22.22	

SVM			Actual						
	Predicted		MIS	Eco	Mkt	Fin	Mgt	Acc	Precision
		MIS	24	0	0	5	0	0	82.67
		Eco	0	14	0	0	2	2	77.78
		Mkt	0	0	15	1	1	1	83.33
		Fin	0	3	2	16	1	3	64.00
		Mgt	2	1	2	2	12	6	48.00
		Acc	0	0	0	3	1	6	60.00
		Recall	92.31	77.87	78.95	59.26	70.59	33.33	

Table 6. Precision with and without Close Related Study Majors

	Precision for All Six Majors	**Precision without Finance and Accounting**	**Precision without Finance, Accounting, and Management**
NB	68.8	70.25	74
LR	51.16	54	58
DL	64	62.25	66.6
DT	52.17	60.5	66
RF	69.7	72.5	79.6
XGBOOST	64.8	67.8	71
SVM	70.5	74.25	82

Factor Importance Ranking

The main goal of the factor importance ranking is to rank a subset of input variables (factors) that can accurately characterize the input data while mitigating the impacts of irrelevant variables (factors) and still produce accurate prediction results. Therefore, to compare and further validate the effectiveness of the used factors, we have examined the importance through (a) the values of the average factors affecting the forecast of the target variable (study major) as generated by the XGBoost algorithm and (b) the weights of factors generated using the RF algorithm. We have selected both the XGBoost and RF algorithms to determine the factor importance scores, given their good performance in predicting students' study major. According to Chen et al., XGBoost maintains the important factors required for prediction [43], while the RF feature selection method is very effective and valuable for selecting important factors [44].

Figure 1 shows the most important factors for each of the six majors as generated by the XGBoost algorithm. The green bar in this context implies that the factor positively correlates with predicting the student study major. In contrast, the red bars indicate the irrelevance to the prediction of study major. Evidently, the study major is sensitive to the higher salary factor for four majors (Figure 1). The findings align with the results presented by [18], which confirm that when choosing a college major, students are primarily influenced by the potential job salaries rather than just the initial salaries. Furthermore, the results indicate a significant influence of parents, close relatives, friends, and graduates on the selection of a management study major. These findings are consistent with the research conducted by [19], which highlights that the probability of a student choosing a major related to that of a close relative is strongly correlated with the relative's earnings at the time of the student's major selection.

Furthermore, each study major is associated with specific factors that play a significant role in predicting its selection. For example, accounting majors are influenced by job availability, finance majors consider preadmission tests and scores, marketing majors value creativity, management majors are influenced by others, MIS majors prioritize technical skills, and economics majors tend to favor an easy major. The findings of a survey-based study conducted by [20] partially support these observations. The study aimed to identify the crucial factors that impact students' choices of study programs in business colleges. According to the survey responses, students indicated that their choice of a college major is influenced by the abilities typically associated with that particular field of study. For instance, marketing majors highlighted the importance of communication and creativity skills, while MIS majors emphasized the necessity of technical skills.

We have conducted an extensive analysis of factor importance for all six majors using the RF algorithm, and the results are presented in Table 7. According to the generated weights, creativity, technical skills, and GPA emerge as the top three factors that significantly influence the choice of study major. It is worth noting that the top five highly weighted factors obtained through the RF algorithm align with those obtained from the XGBoost algorithm. These findings support and reinforce the results of previous studies in the literature. Specifically, the impact of the following factors has been well-documented: final scores of high school and preadmission tests [7, 26], first-year GPA [25], student interest in the field [7], student's interpersonal skills [24, 26, 27], work experience [7, 23], and socio-economic background [22]. By confirming the significance of these factors, our study adds to the existing body of research and strengthens the understanding of the determinants behind students' choice of study major.

Table 7. average generated weight of the Factors (independent variables) by the RF algorithm

Factor	Weights
Creativity	0.241
Technical Skills	0.101
GPA	0.077
Easy Major	0.073
Communication Skills	0.066
Higher Salary	0.065
Parents Educational Background	0.053
Own a business	0.053
HSGA	0.052
Interest in Major	0.046
Close Relatives Educational Background	0.044
Prestige	0.044
SAAT	0.036
GAT	0.033
Previous Work/Course Experience	0.030
Reputation	0.030
Job Availability	0.023
Influence of Friend/Graduate	0.016

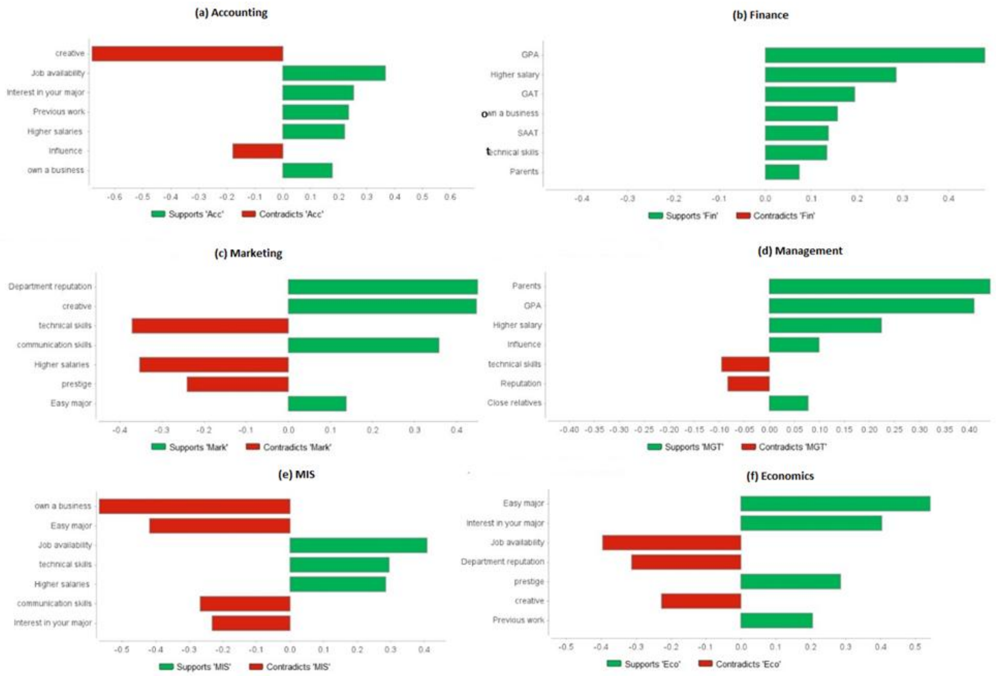

Figure 1: Feature Importance by XGBoost Algorithm.

CONCLUSION

This study investigated the importance of employing supervised learning techniques to predict students' study major based on their preadmission profile, interpersonal effects, and external influences. Furthermore, we identified a number of factors that should be considered when developing an intelligent decision support system for university admission. Our main goal and intention is that this study would serve as a prototype for research in developing a decision support system for university admission.

Using a student dataset, we trained and evaluated seven supervised learning models (i.e., NB, RL, DL, DT, RF, XGBoost, and SVM). This dataset includes entry test results and overall students' high school grades. In addition, the dataset contains details about the family's educational history, interpersonal data, and job market expectations.

The findings demonstrate that a good-performance prediction model might be developed using preadmission and interpersonal factors to predict a study major. For instance, the SVM model precision performance was >70%. Thus, the findings support the feasibility of prediction modeling in finding the potential and appropriate student study major, where decision-makers in the university can utilize these prediction models to review the preadmission criteria. Furthermore, the experimental results show that considering a variety of factors yields better results, but it is crucial to investigate the relevance of each, as not all factors are equally important.

According to our findings, a student's competencies and personality greatly influence their decision in selecting their business study major. Our findings demonstrate the importance of assigning weight to the SAAT score in the preadmission criteria. This is because it is one of the preadmission criteria that most reliably predicts the student study major. Based on this finding, KSU's admission decision-makers are advised to evaluate the admissions criteria and include a weight to the SAAT score. Generally, Saudi universities are advised to consider more preadmission variables such as interpersonal effects and extracurricular activities that may potentially affect the selection of student's major and hence future student performance.

It is important to acknowledge the limitation of sample size in this study. While the survey approach was utilized to collect data for the development of the student study major prediction model, it is worth noting that the sample size consisted of CBA students at KSU, who have the opportunity to major in six different fields. Although the sample size was considered appropriate for the purposes of this study and represented the population of interest, it is essential to consider the potential impact on the generalizability of the findings. Future research could explore larger sample sizes, encompassing a more diverse range of students from various educational institutions, to further validate and expand upon the predictive models in the context of multi-class study major selection.

Implementing a predictive model for undergraduate student study major in real university settings presents challenges that require careful consideration. Data privacy concerns must be addressed by implementing robust attribute selection and ensuring compliance with privacy regulations. Continuous model updates are crucial to maintaining accuracy and relevance over time, necessitating regular monitoring, stakeholder feedback incorporation, and staying informed about advancements in the field. Accurately capturing student interpersonal skills and references poses a challenge, which can be overcome through structured assessments and standardized evaluation methods. University administrators should establish clear data governance policies, enhance data protection measures, foster transparency, and gain stakeholder acceptance. Involving students and faculty in the model development process is essential for successful implementation. By overcoming these challenges, universities can leverage predictive models to facilitate informed decision-making and improve undergraduate student study major selection outcomes.

There are several research directions to expand and build on the work conducted in this study. Firstly, future research can explore the impact of cultural factors on study major selection by extending the analysis to non-business majors such as medicine, computer science, engineering, and sciences. This investigation would provide valuable insights into the essential factors that shape students' choices in these fields. Additionally, alternative ML and DM techniques can be employed to develop an intelligent decision support system for university admission, improving the knowledge model and enhancing its effectiveness. Furthermore, incorporating dynamic data sources, such as labor market data, into the predictive model can offer real-time information on job market trends, enabling students to align their study major choices with emerging career opportunities. A future study could enrich the analysis by exploring non-quantitative factors, including personal interests, motivation, extracurricular activities, prior work experiences, and career aspirations. Qualitative research methods like interviews or focus groups could complement the quantitative approach, providing a more holistic view of significant selection in undergraduate study majors. These research directions would contribute to the ongoing development and refinement of the model, expanding its utility and impact on educational decision-making.

REFRENCES

[1] A. Patnaik, M. Wiswall, and Basit Zafar, 2020, "College majors", NBER Working Papers 27645 SSRN Electronic Journal, National Bureau of Economic Research, Inc, 2020.

[2] M. Schneider, “Finishing the First Lap: The Cost of First Year Student Attrition in America’s Four Year Colleges and Universities”. American Institutes for Research, 2010.

[3] P.M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, “Prediction in MOOCs: A review and future research directions”. IEEE Transactions on Learning Technologies, 12 (3) (2018), 384-401.

[4] K. Leu, “Beginning College Students Who Change Their Majors within 3 Years of Enrollment”. NCES: Washington, DC, USA, 2017.

[5] J. G. Altonji, E. Blom, and C. Meghir. “Heterogeneity in human capital investments: High school curriculum, college major, and careers”. Annual Review of Economics, 4(1),185–223, 2012.

[6] J.G. Altonji, P. Arcidiacono, and A. Maurel. “The analysis of field choice in college and graduate school. In Handbook of the Economics of Education, 5, 305-396, 2016.

[7] A. O. Alsayed, M. S. M. Rahim, I. AlBidewi, M. Hussain, S. H. Jabeen, N. Alromema, N., S. Hussain, and M. Jibril, “Selection of the right undergraduate major by students using supervised learning techniques”. Applied Sciences. 11(22), 10639, 2021.

[8] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," in IEEE Access, vol. 8, pp. 55462-55470, 2020.

[9] L. Tan, J. B. Main, and R. Darolia, “Using random forest analysis to identify student demographic and high school-level factors that predict college engineering major choice”. Journal of Engineering Education. 110(3), 572-593, 2021.

[10] M.M. Awaliyah, A. Kurniawati, and A. Rizana, “Profile matching for students specialization in industrial engineering major. IOP Conference Series: Materials Science and Engineering. 830(3), 032063, 2020.

[11] F. Atuahene, “An analysis of major and career decision-making difficulties of exploratory college students in a Mid Atlantic University”. SN Social Sciences. 1(4), 80, 2021.

[12] National Center for Assessment (Qiyas). (September 18, 2022). Establishment of national center for assessment in higher education. [Online]. Available: https://etec.gov.sa/en/productsandservices/Qiyas/Education/Pages/default.aspx

[13] King Saud University. (September 18, 2022). Deanship of admission and Registration. [Online]. Available: https://dar.ksu.edu.sa/sites/dar.ksu.edu.sa/files/imce_images/admission2022.pdf

[14] King Abdul Aziz University. (September 18, 2022). Admission. [Online]. Available: https://admission.kau.edu.sa/Files/210/Files/162695_Admission_Guide.pdf

[15] King Fahd University of Petroleum and Minerals. (September 18, 2022). Undergraduate admission (bachelor degree). [Online]. Available: http://www.kfupm.edu.sa/departments/admissions/default.aspx

[16] Princess Nourah Bint Abdulrahman University. (September 18, 2022). Deanship of admission and registration. [Online]. Available: https://www.pnu.edu.sa/en/Deanship/Registration/Pages/AdmissionGuide.aspx

[17] S. E. Turner and W. G. Bowen. “Choice of major: The changing (unchanging) gender gap”. ILR Review, 52(2):289–313, 1999.

[18] M. C. Berger. “Predicted future earnings and choice of college major”. ILR Review, 41 (3):418–429, April 1988. ISSN 0019-7939418-429.

[19] X. Xia. “Forming wage expectations through learning: Evidence from college major choices”. Journal of Economic Behavior and Organization, 132(PA):176-196, 2016.

[20] D. Roach, R. McGaughey, and J. Downey, "Selecting a business major within the College of Business”, Administrative Issues Journal,2(1), article 112012.

[21] M. S. Hiatt, J. A. Swaim, and M. J. Maloni, “Choosing an undergraduate major in business administration: Student evaluative criteria, behavioral influences, and instructional modalities”, The International Journal of Management Education, 16(3), 2018, 524-540, ISSN 1472-8117.

[22] K. Pupara, W. Nuankaew and P. Nuankaew, "An institution recommender system based on student context and educational institution in a mobile environment," 2016 1-6, doi: 10.1109/ICSEC.2016.7859877.

[23] C. Fiarni, E. M. Sipayung, and P. B. T. Tumundo, “Academic decision support system for choosing information systems sub majors programs using decision tree algorithm”. Journal of Information Systems Engineering and Business Intelligence, 2019, 5, 57-66.

[24] S. Iyer, and C. Variawa, “Using machine learning as a tool to help guide undeclared/undecided first-year engineering students towards a discipline”. In Proceedings of the Canadian Engineering Education Association (CEEA), Ottawa, 8-12, 2019, 1-17.

[25] M. Ezz, and A. Elshenawy, “Adaptive recommendation system using machine learning algorithms for predicting student’s best academic program”. Education and Information Technologies, 25(4), 2020, pp. 2733-2746.

[26] M. Ayman AlAhmar. “A prototype rule-based expert system with an object-oriented database for university undergraduate major selection”. International Journal of Applied Information Systems, 4(8), 2012,38-42.

[27] N. Kamal, F. Sarker and K. A. Mamun, "A comparative study of machine learning approaches for recommending university faculty," 2nd International Conference on Sustainable Technologies for Industry 4.0, 2020, pp. 1-6, doi: 10.1109/STI50764.2020.9350461.

[28] W. Zhang, “Why is: Understanding undergraduate students’ intentions to choose an information systems major”. Journal of Information Systems Education, 18(4), 2007,447-458.

[29] J. Downey, R. McGaughey, and D. Roach, “MIS versus computer science: An empirical study of the influences on the student's choice of major”. Journal of Information Systems Education, 20(3), 2009,357- 368.

[30] D. Kim, F. S. Markham, and J. D. Cangelosi, “Why students pursue the business degree: A comparison of business majors across universities”. Journal of Education for Business, 78(1), 2002,28-32.

[31] J. Downey, R. McGaughey, and D. Roach, “Attitudes and influences toward choosing a business major: The case of information systems”. Journal of Information Systems Education, 10, 2011,231-251.

[32] D. Roach, R. McGaughey, and J. Downey, James, " Selecting a Business Major within the College of Business" Administrative Issues Journal. 2(1), 2012, pp. 107-121.

[33] U. Sekaran, and R. Bougie. “Research Methods for Business: A Skill Building Approach.” John Wiley & Sons, 2016.

[34] T.M. Mitchell, “Machine Learning”, McGraw-Hill, 1997

[35] S. S. Keerthi,K. B. Duan, S. K. Shevade, and A. N. Poo Sathiya, et al. "A fast dual algorithm for kernel logistic regression." Machine learning 61.1 (2005): 151-165.

[36] S. R¨uping. myKLR - Kernel logistic regression. Technical report, University of Dortmund, Department of Computer Science, 2003.

[37] A. Mathew, Amitha, P. Amudha, and S. Sivakumari. "Deep learning techniques: An overview." International conference on advanced machine learning technologies and applications. Springer, Singapore, 2021.

[38] L. Breiman, Leo. "Random forests." Machine Learning 45.1 (2001): 5-32.

[39] C. Crisci, Carolina, Badih Ghattas, and Ghattas Perera. "A review of supervised machine learning algorithms and their applications to ecological data." Ecological Modelling 240 (2012): 113-122.

[40] J. H. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of Statistics 29(5) (2001): 1189-1232.

[41] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[42] V. H. Nhu, A. Shirzadi, H. Shahabi (2020). Shallow landslide susceptibility mapping: A comparison between logistic model tree, logistic regression, naïve bayes tree, artificial neural network, and support vector machine algorithms. International journal of Environmental Research and Public Health, 17(8), 2020, 2749.

[43] C.Chen, Q.Zhang, B.Yu, Z.Yu, P.J. Lawrence, Q.Ma, Y.Zhang, Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier, Computers in Biology and Medicine, 123, 2020, 103899, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2020.103899.

[44] R. Chen, S.W., CDewi, S.W., Huang. Selecting critical features for data classification based on machine learning methods. Journal of Big Data 7, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4

[45] L. Tan, J., Main, J. B., and R. Darolia, R. Using random forest analysis to identify student demographic and high school-level factors that predict college engineering major choice. Journal of Engineering Education, 2021, 110(3), 572–593. https://doi.org/10.1002/jee.20393

[46] M. Moakler, and M. Kim, College Major Choice in STEM: Revisiting Confidence and Demographic Factors. The Career Development Quarterly, 2014, 62: 128-142. https://doi.org/10.1002/j.2161-0045.2014.00075.x

[47] CN., Chang, S., Lin, OM., Kwok. Predicting STEM Major Choice: a Machine Learning Classification and Regression Tree Approach. Journal for STEM Educ Res 6, 358–374 (2023). https://doi.org/10.1007/s41979-023-00099-5

[48] N., Howard, K., Howard, R., Busse, & C., Hunt. Let's talk: An examination of parental involvement as a predictor of STEM achievement in math for high school girls. 2019. Urban Education, 58(4), 586–613. https://doi.org/10.1177/0042085919877933

[49] M., Moakler, & M., Kim. College major choice in STEM: Revisiting confidence and demographic factors. 2014, The Career Development Quarterly, 62(2), 128–142. https://doi.org/10.1002/j.2161-0045.2014.00075.x

[50] S.K. Thompson, (2012). Sampling (Vol. 755). John Wiley & Sons.

[51] X. Meng, (2013, May). Scalable simple random sampling and stratified sampling. In International conference on machine learning (pp. 531-539). PMLR.

[52] E. Liberty, K. Lang, & K. Shmakov, (2016, June). Stratified sampling meets machine learning. In International conference on machine learning (pp. 2320-2329). PMLR.