applied sciences

MDPI

Article

Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques

Alhuseen Omar Alsayed [1,2,*], Mohd Shafry Mohd Rahim [1], Ibrahim AlBidewi [3], Mushtaq Hussain [4], Syeda Huma Jabeen [5], Nashwan Alromema [6], Sadiq Hussain [7] and Muhammad Lawan Jibril [8]

[1] Department of Computer Science, School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, Johor Bahru 81310, Johor, Malaysia; shafry@utm.my
[2] Deanship of Scientific Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[3] Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; ialbidewi@kau.edu.sa
[4] Department of Computer Science and Information Technology, Virtual University of Pakistan, Lahore 54000, Pakistan; mushtaq.hussain@vu.edu.pk
[5] Brainnettom Center, Institute of Automation, University of Chinese Academy of Sciences, Beijing 100190, China; hjabeen2@gmail.com
[6] Department of Computer Science, Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdulaziz University, Jeddah 21589, Saudi Arabia; nalromema@kau.edu.sa
[7] Examination Branch, Dibrugarh University, Dibrugarh 786004, India; sadiq@dibru.ac.in
[8] Department of Computer Science, Federal University of Kashere, Barri 771103, Gombe State, Nigeria; lawan.jibril@fukashere.edu.ng
* Correspondence: nuriy3@graduate.utm.my; Tel.: +966-543-169-128

check for updates

Citation: Alsayed, A.O.; Rahim, M.S.M.; AlBidewi, I.; Hussain, M.; Jabeen, S.H.; Alromema, N.; Hussain, S.; Jibril, M.L. Selection of the Right Undergraduate Major by Students Using Supervised Learning Techniques. *Appl. Sci.* **2021**, *11*, 10639. https://doi.org/10.3390/app112210639

Academic Editors: Suhuai Luo and Kamran Shaukat

Received: 25 September 2021
Accepted: 8 November 2021
Published: 11 November 2021

Abstract: University education has become an integral and basic part of most people preparing for working life. However, placement of students into the appropriate university, college, or discipline is of paramount importance for university education to perform its role. In this study, various explainable machine learning approaches (Decision Tree [DT], Extra tree classifiers [ETC], Random forest [RF] classifiers, Gradient boosting classifiers [GBC], and Support Vector Machine [SVM]) were tested to predict students' right undergraduate major (field of specialization) before admission at the undergraduate level based on the current job markets and experience. The DT classifier predicts the target class based on simple decision rules. ETC is an ensemble learning technique that builds prediction models by using unpruned decision trees. RF is also an ensemble technique that uses many individual DTs to solve complex problems. GBC classifiers and produce strong prediction models. SVM predicts the target class with a high margin, as compared to other classifiers. The imbalanced dataset includes secondary school marks, higher secondary school marks, experience, and salary to select specialization for students in undergraduate programs. The results showed that the performances of RF and GBC predict the student field of specialization (undergraduate major) before admission, as well as the fact that these measures are as good as DT and ETC. Statistical analysis (Spearman correlation) is also applied to evaluate the relationship between a student's major and other input variables. The statistical results show that higher student marks in higher secondary (hsc_p), university degree (Degree_p), and entry test (etest_p) play an important role in the student's area of specialization, and we can recommend study fields according to these features. Based on these results, RF and GBC can easily be integrated into intelligent recommender systems to suggest a good field of specialization to university students, according to the current job market. This study also demonstrates that marks in higher secondary and university and entry tests are useful criteria to suggest the right undergraduate major because these input features most accurately predict the student field of specialization.

Keywords: machine learning; learning analytics; student field forecasting; imbalanced datasets; explainable machine learning; intelligent tutoring system

1. Introduction

Today, higher education institutions face considerable difficulties, such as the absence of government funding, competitive job markets, admission processes, student strength, and selections of student specializations [1,2]. Student specialization selection is an area of educational research that has received little attention, although it is critical in recognizing students' interests and preparing them for a future career [3]. Student specialization is a worldwide educational problem that needs to be investigated. For example, in the USA, approximately 30% of year-one students do not return for their second year, and more than $9 billion is spent on these students [4]. Furthermore, the completion rates of 4-year degrees in the US are approximately 50% [5]. These alarming figures require every possible effort to support students and higher education institutions in this critical issue. According to a study conducted by the United States Departments of Education (NCES), of the 98% of students that declared a bachelor's degree major in 2011–2012, 33% changed their major by 2014 during their third year of study [6]. Moreover, approximately 35% of college students who declared their majors to be STEM programs and 29% of students who declared their majors to be STEM-related programs eventually changed their majors after 2 years of study [7].

Student specialization selection can indicate the choice of an appropriate specialization/major that leads to a high level of satisfaction, success in allotment, graduation within a time frame, or other more specific milestones [8]. In an educational institution, the selection of the right undergraduate major by students is a major challenge when progressing to an academic level because students do not know about the job market and the demand for the required skills.

1.1. Student Field Specialization (Undergraduate Major Course)

Field specialization selection means selecting the right undergraduate major for students, for example, engineering, computer science, and management [9,10]. Universities are required to fulfill students' academic disciplines. One essential goal of universities is to aid student admission into their desired college specialization. What student admission means varies depending on the context of the university requirements, students' academic results, and other related factors [11,12]. Universities provide student admission centers and student counselors or advisors to help students meet their educational needs. Recommending suitable colleges and fields (suitable undergraduate majors) based on students' attributes and preferences is one service that could be provided by admission departments. However, due to the growing numbers of fields, students, and available skills, these advisors sometimes fail to help students with their selections [3,13,14]. Due to the substantial amount of work required by these advisors, who are not able to handle this situation, students have insufficient knowledge about how to select an appropriate field (major) in their undergraduate program that fits their preferences, personality, subjects of interest, and career type that he or she likes [3,13,14].

1.2. Significance of Predicting Student Field Specialization

The undergraduate major (field specialization) is an important research topic because an incorrect undergraduate major selection affects students' academic lives, learning, and careers [15]. Students in every country face challenges in selecting the right undergraduate major. From the time students decide to continue their higher education, they are confronted with decisions concerning their education, many of which can be challenging. When students attend college, they choose a major based on several factors, such as their friends, parents, future opportunities, and, most importantly, the student advisory center. Some students are not fully aware of the importance of their academic abilities and job market demands [16]. They may depend on others' opinions, which may lead to the incorrect and unsuitable major selection. Incorrect academic decisions have a considerable and direct impact on students' success and future lives [17]. If a student chooses an unsuitable academic major and continues to have low grades and fails to raise his or her CGPA within a year, the student will be dismissed from college.

Thus, the choice of a field (student major) for a new student can be a difficult decision; therefore, universities need to use a student intelligent counseling system because the correct student feedback has been shown to decrease the course dropout rate and increase graduation rate [18]. Long et al. [19] indicated that improving and enhancing the matching of students with their university specialization could substantially assist in decreasing the level of study discontinuation among younger students, which would also contribute to opening up spots for other potential students and in decreasing the inefficient utilization of public resources and funds for higher education. Therefore, this paper addressed the field of specialization suggestion problem by suggesting appropriate study fields for students at early stages, according to the current job market, education history, and career goals for the students. Hence, it is desirable to develop sophisticated forms of intelligent recommendation tools to help students in selecting an appropriate field of specialization

1.3. Machine LearningTechniques Used in an Education Predictive Model

Artificial intelligence (AI) is an important concept in the field of science and is currently a promising technological revolution. It uses machines to develop a concept of intelligence that is more like the human brain. There have been various fields that have taken advantage of implementing AI in their day-to-day business processes. In the area of computer sciences, the concept of artificial intelligence is widely utilized, and it is considerably related to the concept of machine learning (ML). ML is a sub-field of AI that identifies complex and hidden patterns or knowledge from large amounts of data and then makes smart decisions on unseen data [20]. One of the key features of ML is a training model utilizing different dependent and independent variables, which further depends on different utilized learning algorithms types (supervised, semisupervised, or unsupervised). ML is mostly used for predictions in many field to provide solutions to questions such as global solar radiation [21], weather predictions [22], flight time deviation [23], mortality rates in COVID-19 patients [24], predict bank failures [25], credit default prediction in bank [26], cyber security [27], bankruptcy prediction [28], filter e-learning contents [29] and efficient processes for manufacturing industries [8]. It can also be used for predicting rates in student dropouts in any course [17] and for understanding unique student learning styles [30]. Moreover, ML algorithms can assist the educational sector by constantly evaluating student academic performance. Due to the vast and dynamic implementation of ML, as well as its capability to learn from any dataset and to predict and classify future transactions, we have selected multiple ML algorithms for use in this study.

In recent years, research interest in the application of ML in education has increased, particularly among higher-education institutions. A recent study discovered that educational-related decisions are frequently made based on educational management stakeholders'/ students' impressions and experience, rather than based on knowledge-rich data. Unfortunately, it is a challenging task to make a suitable choice of the subject matter at an early stage due to the convoluted interaction of a variety of factors [31–33]. ML approaches are designed to make necessary educational information readily available to knowledge consumers. ML techniques have also been shown to be beneficial in improving outcomes at several educational institutions and student management centers by making necessary educational information readily available to students and other individuals [34,35]. In the past 10 years, investigations on ML and education data mining [EDM] have played a significant role in exploring educational problems [31,36], such as understanding student performance [37,38] and educational institution performance [39]. These techniques have also been used to predict student engagement and difficulty in online education [40,41] and in recommending suitable colleges and courses [9,42–45]. ML is increasingly prevalent and vital in educational contexts, in terms of predicting and identifying quality educational-related problems for students and decision-makers, as well as in enhancing other managerial services pertaining to streamlining students' needs. Furthermore, numerous research studies on education have predicted admission to universities, student allotment, and admission into their desired colleges/majors by using ML techniques [33,46,47].

The current study investigated the best ML classifier that is suitable for building student field specialization intelligent systems, which can predict student study fields based on student academic history and the job market. Predicting student study fields is a classification problem; therefore, we verified the performance of common ML classification algorithms, such as Decision Tree (DT), extra tree classifiers [ETC], random forest (RF) classifiers, and gradient boosting classifiers [GBC], and Support Vector Machine [SVM], on the current study dataset. The performances of these algorithms are good, based on categorical data [48]. Additionally, these algorithms are easily described, understandable, and implemented [48]. The current study also verified the performance of the SVM on the current study dataset because it can examine both linear and nonlinear data [49].

1.4. Innovation of the Current Study

The current study investigated the student field of specialization by using students' previous histories and job market information utilizing different ML techniques. ML approaches have been employed to accurately forecast college selection and select the best fit student major by means of common classification algorithms with diverse feature sets [1]. To achieve our current study goal, we trained different common ML models (extra tree classifiers [ETC], random forest [RF] classifiers, decision tree [DT] classifiers, gradient boosting classifiers [GBC], and support vector machine [SVM]) based on the current job market and students' previous histories. The results showed that RF and GBC predicted student majors with higher accuracy, as compared to DT, SVM, and ETC. Based on the Spearman correlation method, the study concluded that higher marks in higher secondary levels, entry tests, and universities, are good criteria for suggesting student field specialization. Furthermore, student work experiences and job placements are additional factors that are strongly related to student field specializations. In addition, these ML models and features could be of high value in developing an intelligent system to easily recommend a specialization to potential applicants who are often unsure of their desired fields of specialization. Finally, this paper differs from other research in this area of predicting student field specializations because it is based on the job market and student histories and experiences.

The current study investigated the following research questions.

Question 1: Can we model the student undergraduate major path choice according to the job market and student academic history by applying different ML algorithms, and if so, which ML classifier offers optimal performance in predicting student undergraduate major selection?

Question 2: How is a student's undergraduate major path choice associated with that student's previous academic performance and the job market?

Contributions: This study possesses contributions as enlisted below in the domain of selection of majors by students.

1. The research utilized Kaggle repositories to devise a ML approach in selecting the field of specialization by students for future endeavors.
2. Several supervised learning techniques with 10-k fold cross-validation were utilized and yielded that RF, SVM, and GBC were the suitable classifiers for predicting student undergraduate major.
3. The influential factors related to selecting the right undergraduate major were also showcased. According to my knowledge, no work has been done to find student influential factors.
4. The findings may be integrated into the intelligent field recommender system for predicting suitable fields for students according to the job market.

The rest of this paper is organized as follows. The literature review of the student study field is discussed in Section 2. The research materials and methods are discussed in Section 3, which contains all of the details of our proposed framework of the student study field selection system. Section 4 describes the experiment and discusses the results of the current study, where the performances of different ML algorithms are tested on the current study dataset. Finally, Section 5 presents the conclusions and future work of the current study.

2. Literature Review

Several studies have been conducted to investigate student field specialization (student major) using ML. Past research has used different ML techniques and input features to study the relationship between student data and student majors. Alshaikh et al. [3] built a recommendation system to suggest suitable colleges for KAU students based on the students' grades, college specializations, and enrollment requirements. They applied this system to a dataset of 960 KAU preparatory students in 2017. Two methods were used to evaluate the accuracy of the k-nearest-neighbor algorithm. In the first method, the dataset was split into two datasets, 20% of the dataset for testing and the remaining 80% for training, which generated 70.83% accuracy. The second approach applied k-fold cross-validation, where the dataset was split into K smaller sets and the test was applied K times. Pupara et al. [50] have proposed an accurate institutional recommender system (RS) that was developed by combining decision tree and association rule methodologies. The RS is intended to assist students in selecting acceptable colleges based on their context and educational institution information using a mobile device. Ezz and Elshenawy [9] presented an adaptive recommendation system for predicting a suitable engineering department for students enrolled in an engineering preparatory year college using classification methods such as SVM, k-nearest neighbor (KNN), linear regression (LR), quadratic discriminant analysis (QDA), and RF. The system recommends a suitable engineering department among seven engineering departments for each student based on his academic performance and the proposed system has an average accuracy of 82.57%. In the study of Salaki et al. [51], 3 ML algorithms, namely, naïve Bayes (NB), RF, and sequential minimal optimization (SMO), were trained on a dataset collected from three different educational colleges in Bangladesh to identify and select the best groups of educational majors to streamline the selection of a suitable direction for new students. The results showed that RF had the best performance, with 84.9% accuracy, 84.9% precision, 84.6% sensitivity, and an F-measure of 84.3%. In a study conducted by Fiarni et al. [52], an academic decision support system was built in the IS department to classify and recommend IS sub-majors for students using a C4.5 decision tree classifier and a rule-based approach. Bautista et al. [10] adopted 8 methods (namely, the J48 tree classifier, logistic function classifier, naïve Bayes (NB), nominal regression, decision tree CHAID, neural network multilayer perceptron, neural network radial basis function, and nearest neighbor) to recommend a suitable specialization for engineering students. The first three methods were tested in Weka and achieved accuracies of 80.5%, 64.30%, and 60.11%, respectively. The last 5 experiments conducted in SPSS, yielded accuracies of 64.00%, 68.20%, 71.30%, 61.00%, and 71.20%. Moreover, the J48 tree classifier performed the best, with the highest accuracy of 80.5%. A study conducted by Kularbphettong and Tongsiri [53] aimed to develop a decision support system for student major selection using two ML methods, J48 and Bayesian network algorithms (BNs). Their results showed that BNs performed the best, with 92.13%, 0.93, and 0.91 accuracy, precision, and F-measure, respectively. Meng and Fu [35] applied 8 classification methods, namely, SVM, decision tree (DT), naïve Bayes (GNB), RF, gradient boosting decision tree (GBDT), convolutional neural network (CNN), collaborative filtering (CF), and recurrent neural network (RNN), to recommend appropriate college majors. RF performed best, with an accuracy of 97.87% and an f-score of 96.60%. Wei et al. [54] proposed an improved SVM-based prediction system model for predicting second major selection. Their experimental results indicated that the proposed method performed best, with an accuracy of 87.36%, AUC of 0.8735, the sensitivity of 85.37%, and specificity of 89.33%, Sethi et al. [55] conducted a study to predict the appropriate study stream for students in higher secondary education. They found that the neural network (NN) outperformed the other approaches with a classification accuracy of 86.72%, the sensitivity of 0.92, specificity of 0.82, and MCC of 0.72. Abosamra et al. [56] examined various types of ML predictions models on a dataset, which gave the best choice as a (NN) architecture that provides 6.26 an average root mean squared error, and a mean absolute error of 5.74 based on a scale of 0 to 100.

The artificial neural network (ANN) method was adopted by Latifah et al. [57] to predict suitable student specialization in a dataset of 314 students based on student records from the iGracias Integrated Academic Information System at Telkom University in 2016, and they achieved an accuracy of 94.81%. The NB method and analytic hierarchy process (AHP) techniques were adopted by Zubaedah et al. [58] to build a decision support system to predict suitable specialization in technical faculty in Indonesia. A rule-based classification algorithm (PART) was adopted by Tamiza et al. [59] to propose an intelligent model for selecting and predicting suitable university specialization. The model achieved an accuracy of 73.7%. Iyer and Variawa [60] built a system model to guide first-year undeclared/undecided engineering students to predict suitable engineering majors. They found that the RF approach outperformed the other classification algorithms, with the highest accuracy of 57%. AlAhmar [61] developed a rule-based expert system that suggested majors for students at the undergraduate level. Kamal et al. [62] has used RF classifiers to analyze students' personality and intelligence across various majors and academic programs and predict suitable college majors for students based on academic results, personality, and level of intelligence with an accuracy of level one at 96.1% and 94.72% at second level respectively, moreover, they have investigated that their framework has potential to recommend a student towards future higher degree options.

Although the attractiveness of higher education institutions in many areas of student field of study selection has been extensively researched, there is a paucity of evidence available for modeling the relationship between these factors and intelligent recommendation of student fields based on the job market and student history and experience. To our knowledge, no studies have been conducted on the use of any ML algorithm specifically designed for the purpose of predicting student specialization and identifying the extent to which various parameters contribute to the determination of the specialization of students. As a result of this discovery, we were inspired to conduct our current study. As a result, the current research has implications for higher education, students, and the labor market.

3. Materials and Methods

In the current study, we developed an intelligent system for the field of specialization selection. We trained and tested various ML models on a student dataset because such techniques are suitable for categorical data. The proposed framework has the following stages: data preprocessing, visualization model selection modules, and model deployment. Figure 1 shows all the steps of the proposed framework. Overall, this section provides all the implementations of the student study field intelligent system in the form of the below sub-section.

3.1. Data Description

The data collection consists of several steps. Before the implementation process, consistent and appropriate educational data are required to achieve acceptable results. In this experiment, the dataset was published in Kaggle [63]. The details are shown in Table 1. This dataset was collected from MBA students of CMS Business School in January 2020 and was published on the Kaggle website [63]. This dataset contains placement data of students, including secondary school, higher secondary school, and entry test scores. The dataset also includes work experience, degree percentage, and salary offered by the organization. The salary information represents the importance of the field in the job market. The degree percentage shows the student's interest in the field. The current database contains 216 student records and 19 input features. The first experimental dataset is shown in Table 1, and the target variable is specialization.

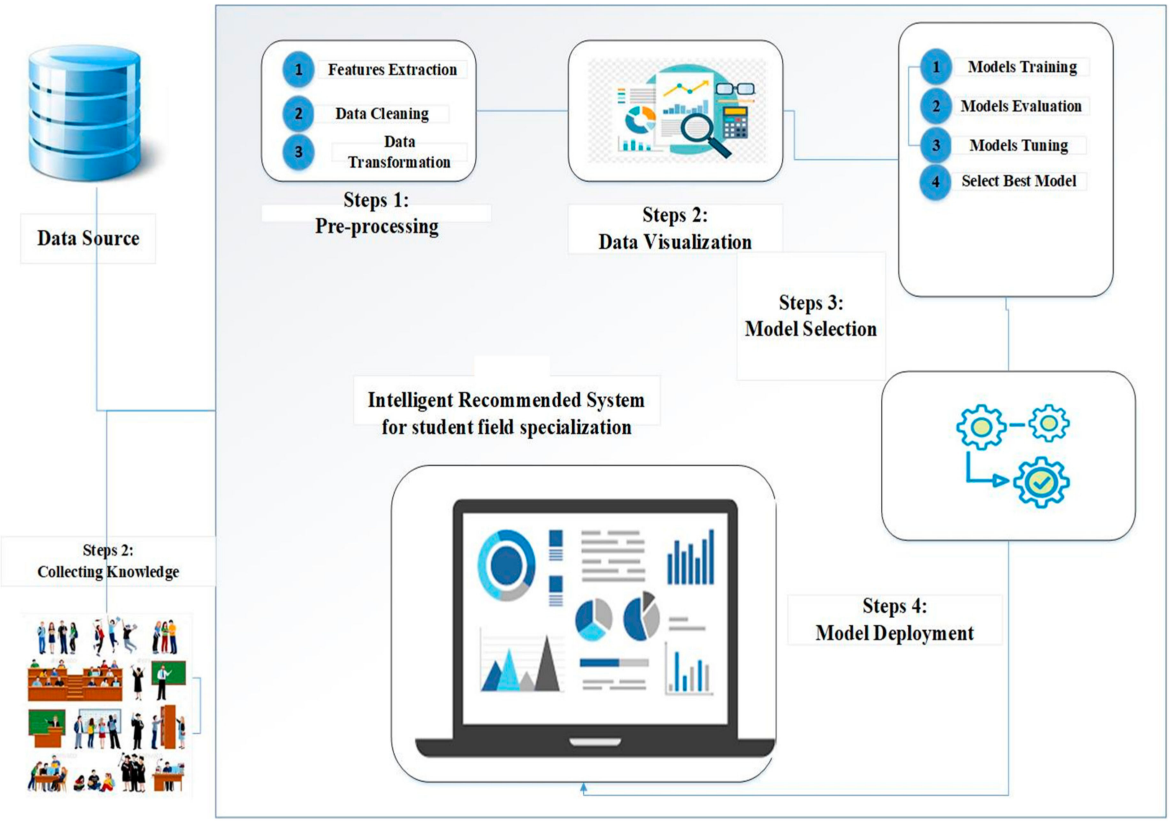

Figure 1. Intelligent recommender system for student field recommendation.

Table 1. First experiment dataset description.

Features Name	Description	Data Type
Gender	Student Gender (Male/Female)	Categorical
SSC_P	Student secondary school percentage	Numeric
SSC_b	Student secondary school board studied (Class 10)	Categorical
HSC_P	Higher secondary school percentage (Class 12)	Numeric
hsc_b	Higher secondary school board studied (Class 12)	Categorical
hsc_s	High secondary school (Class 12) specialization	Categorical
degree_p	Degree percentage	Numeric
degree_t	Degree type	Categorical
workex	Work experience	Categorical
etest_p	Score on entrance test	Numeric
specialization	Degree specialization	Categorical
mba_p	Student percentage in MBA	Numeric
status	Student placement status	Categorical
salary	Student salary	Numeric
ssc_p_catg	Student secondary school percentage in 3 categories (85%+, 60–85%, <60%)	Categorical
hsc_p_catg	Higher secondary school percentage in three categories (85%+, 60–85%, <60%)	Categorical
mba_p_catg	Student percentage in MBA in 3 categories ((85%+, 60–85%, <60%)	Categorical
degree_p_catg	Degree percentage in 3 categories (85%+, 60–85%, <60%)	Categorical
etest_p_catg	Percentage on entrance test in 3 categories (85%+, 60–85%, <60%)	Categorical

3.2. Proposed Framework

Undergraduate major selection is a crucial and challenging decision for the university and the student during the process of admission to fulfill their future success. Due to bad counseling by the admission office center in the university, students go into the wrong study field; as a result, student learning will be affected [15]. Every country faces problems with students selecting the right undergraduate major course, and past studies have indicated that institutions have seen a significant increase in the number of students enrolling as undeclared. In reality, it is estimated that over 50% of students enter college undecided and that approximately 75% of students change their majors at least once before they complete their degree [64,65].

This study builds a student major intelligent system that provides feedback to students and universities about study field selection based on the data extracted from the university database. Finding and building an intelligent system model of students' field of specialization could be of high value in developing an intelligent system to easily recommend a specialization to potential applicants who are often unsure of their desired field of specialization. In the current study, we verified the performance of different supervised ML techniques to predict students' fields of specialization based on student history and the job market. High-performance ML algorithms can provide a high degree of support to student major intelligent systems. Based on performance, the student major intelligent system can provide various support to educational institutions on many issues, such as (1) helping the university administration staff in making quick decisions about students' fields of study; (2) recommending personalized study fields according to students preferences and the job market; (3) decreasing the workload for the admission office; (4) enrolling students according to their preferences and job markets; (5) identifying at-risk students and chances of success of students at early stages; and (6) intervening with the student at early stages so that course dropout rates will decrease, as well as to ensure that students go into the right study fields. Figure 1 shows the basic architecture of the intelligent system for student field of specialization recommendation. The proposed student field intelligent system framework has four major phases, which are shown below.

Step 1 Pre-processing: In the first phase known as Phase-1 (preprocessing), raw information (216 student records) was collected from a university database, as shown in Table 1. Subsequently, we applied different preprocessing techniques by using the Python module, such as removing missing records, deleting irrelevant student records, normalization, outlier detection, and hot encoding. To increase the proposed system performance, we also created new features by creating different categories at different education levels (ssc_p_catg, hsc_p_catg, mba_p_catg, degree_p_catg and etest_p_catg). To remove the missing records, we used different missing record techniques. Sometimes, ML techniques do not process the categorical technique; therefore, we applied the hot encoding technique. Additionally, the cleaned data were normalized because the ML model does not work correctly on non-normalized data. Finally, the preprocessing module converted the data into an acceptable form for the ML models, and the data were ready for the next phase. The current study dataset contained 19 input features (previous exam history, salary, and experiences of students), and the explanation of our current study dataset is shown in Table 1. Specialization is the target variable (Y) that finds the class of independent variables. When students belong to Management and Human resource management (HR) study fields (Mk and HR) then target variable (Y) is set "1" and if students belong to Management and Finance study field then target variable (Y) is set to "0".

Step 2 Data Visualization: In the second phase (data visualization), the clean data were visualized by using a different Python library, which shows how important the input feature is in predicting student study fields. This visualization is used to better understand the current study data.

Steps 3 Model Selection: In the third Phase (model selection modules), different supervised ML techniques were trained and tested using a 10-fold cross-validation method on the clean data by using the Scikit-learn Python library. It is a free ML library of python. In this library, we can easily implement ML algorithms. The current study ML algorithms were trained and tuned on training data and tested on test data using 10-fold cross-validation. After training, the model will find a pattern between input features and out variables or find the best model on the current study dataset. In addition, to select high-performance models, different performance metrics, such as accuracy and a true positive rate, were used. Finally, ML models with high accuracy are selected for intelligent field systems:

Step 4 Model Deployment: In Phase four, an intelligent field of specialization system was developed with the help of a high-performance ML model (RF, GBC, and SVM) because RF, GBC, and SVM predict field specialization with high accuracy. Consequently, intelligent detection of the student field of specialization provides decision-makers, academic advisors, students, and other individuals with knowledge and person-specific information, which is intelligently filtered or presented at the appropriate time, to improve education and the student's best-fit specialization field [3,14,35,66,67]. Additionally, the proposed intelligent field of specialization system will help university admission offices in daily activities.

4. Results and Discussion

Consequently, the selection of an appropriate and suitable field of study is a paramount issue for both students and educational institutes. Therefore, in this section, we investigated the student field (student major) of specialization selection using different ML techniques based on student academic history and the current job market. We also visualized the current dataset to further understand the input variables and performed several experiments using Python to answer the research questions.

We performed data visualization using Python to further understand the experimental dataset. These visualization results show the importance of input attributes in predicting field specialization. Figure 2 shows that higher secondary students who obtained jobs mostly majored in commerce and science. This result further indicates that commerce and science fields are currently in the greatest demand.

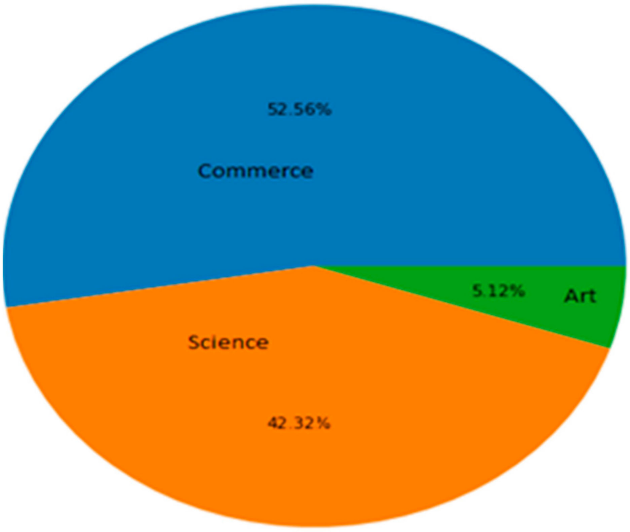

Figure 2. The proportion of different student fields of specialization in higher secondary school.

Figure 3 shows that commerce and management students in higher secondary schools mostly take science and technology fields as postgraduates.

In Figure 4, the blue portion (1) represents Mkt and finance, and the yellow portion (0) represents the Mkt and HR. Figure 4 indicates that 55.81% of students take the MKT& Finance program in postgraduate education, and others take the MKT & HR field.

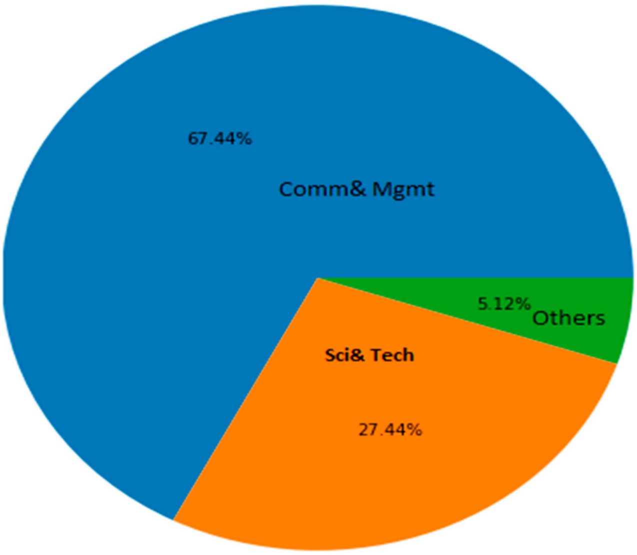

Figure 3. The proportion of different student fields of specialization in degree.

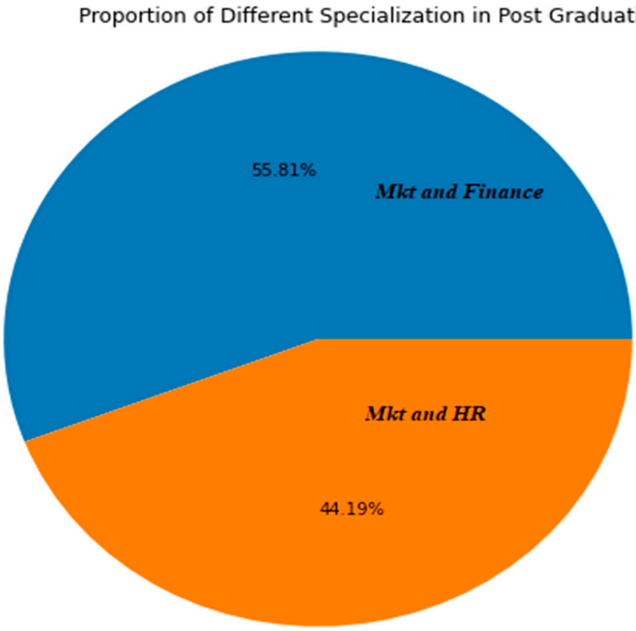

Figure 4. The proportion of different student fields of specialization in postgraduate education.

At this step, we have obtained considerable knowledge about the data and can easily build models. Although there are lots of studies related to the selection of majors by students and its influential factors [1–7,68], studies with a machine learning approach in this domain are limited [7]. In this study, we investigate the student study field based on student job markets, student academic history, and job experience. As the number of factors, size of the datasets, methods applied varied in different studies, it is impractical to compare these studies. Some of these studies were listed in Table 2. To find an appropriate ML model and student factors for our proposed recommendation system, this study conducts two experiments on datasets using the Python Sklearn library. Sklearn is a Python library that is used to train and deploy ML models. In this section, we investigate the following research question:

Table 2. Related studies are based on major influence factors for the recommendation of the future domain.

Reference	Year of Publication	Country	No of Participants	Some of the Most Influential Factors in Selecting Major	Methods Used
[69]	2006	USA	89	personal interest in the subject matter, long-term salary prospects, probability of working in the field after graduation, starting salary, and prestige of the profession	Average Importance
[70]	2017	Korea	816	collaborative learning; technology-based learning; self-regulated learning; hands-on activities; belief in major benefits	Cronbach's Alpha, exploratory factor analysis (EFA), and confirmatory factor analysis (CFA)
[71]	2016	Korea	195	Gender, Grade, Acquaintance's recommendation, Daily hours of study, Place of residence	Fisher's exact test, *t*-test, one way ANOVA, Mann Whitney test, and ANCOVA
[14]	2019	Saudi Arabia	239 prospective participants and 392 university participants	the outcome of student's qualification exams and overall high school grades	Fuzzy Expert System
[72]	2015	Indonesia	40	value of national examination, the value of the placement test, and value of School Exams	Fuzzy Multiple Attribute Decision Making (FMADM)
[73]	2011	Iran	465	Students' interest and decision	Structural Equation Modeling
[74]	2004	USA	114	Financial aid, previous education, potential career/degree characteristics, and information sources.	chi-square and analysis of variance for mean differences
[55]	2020	India	550	Marks in Board exam, Family income, Scholarship, etc.	SVM, k-nearest neighbor and Neural Networks

Question 1: Can we model the student undergraduate major path choice according to the job market and student academic history by applying different ML algorithms, and if so, which ML classifier offers optimal performance in predicting student undergraduate major selection?

The first experiment was conducted to explore this question. In the first experiment, we applied several Tree based ML models (DT algorithm, RF algorithm, extra tree classifier, and XGBoost) and SVM to our dataset by using their default parameters. Decision-tree has been widely implemented in various domains, such as in medical fields [75], marketing prediction tasks [76], and education [77,78], due to its various well-known attractive features [79]. Features such as simplicity, comprehensive calculations, no required parameters, and the capability of handling mixed types of data, encouraged us to select DT in this study. Random forests are used in this study due to being easy and stable with many interesting properties. One of these interesting properties is that they provide a powerful computation of variables [80]. The extra tree classifier is one of the learning algorithms that can aggregate the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. It has been applied in this study because it is similar to RF; however, it is faster, and its method in the construction of the decision tree in the forest is optimal. The tree-based algorithm is simple and requires less data; additionally, it is easy to understand and easily implemented [81,82]. Conversely, deep learning (artificial neural network) is complex, computationally expensive, and requires more data [83]. Additionally, we did not use the naive Bayes algorithm, which is a very commonly used algorithm to solve real-life problems because it overlooked how to calculate probabilities [84]. We used SVM because the performance is good using small datasets [85,86]. Moreover, it does not apply a strict requirement on the number of samples and sample points; additionally, it can process error distributions and can be easily promoted. XGBoost was used in this study because it has higher predictive accuracy than other ML algorithms, such as SVM and DT [87,88]. Our dataset contains both numeric and categorical attributes. Therefore, the selected model must perform well on categorical data. For the first experimental dataset, we used 19 input features, which are shown in Table 1, and the target variable was *specialization*. First, we converted the target variable (specialization) into a binary form (0,1) by using python, wherein "0" denotes marketing and finance (Mk&Fin) and "1" represents the Market and Human resources Field (Mk&HR). As ML algorithms cannot directly work on categorical data, and to convert input features in digital form, we used a hot encoding technique. Hot encoding is a technique that can map categorical data into integers; as a result, the Ml algorithm can produce better results. Hot encoding is useful when there is no relationship between the variables. The 10-fold cross-validation method was used to increase the generalization ability of the models and to ensure that the model behavior was optimal. Furthermore, accuracy, true positive rate (TPR), false-positive rate (FPR), and receiver operating characteristic (ROC) curve were used as evaluation metrics. The accuracy represents the percentage of correct predictions of the model given unseen data. In ML, the TPR is also known as recall or sensitivity and indicates the percentage of actual positive values that are correctly predicted by the model. Finally, the ROC curve plots the TPR of the model [89]. The current study performance metrics are shown below.

$$\text{TPR} = \text{TP}/\text{TP} + \text{FN}$$

$$\text{FPR} = \text{FP}/\text{FP} + \text{TN}$$

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

Notes: TP (true positive), FN (false negative), TN (true negative), FP (false positive).

DT algorithms belong to the supervised category of ML algorithms. A DT is simple and easily understandable. We selected this technique because it is widely used by researchers due to its simplicity. In addition, a DT has some great advantages, such as representing rules that could be easily understood and interpreted by users [81]. This type of algorithm performs well for categorical and numerical attributes and does not require complex data preparation. In short, ML classifiers and their outputs are easy to understand for individuals with a non-analytical background [90]. The default parameters (ccp_alpha = 0.0, criterion = 'gini', min_samples_split = 2) are used to train the DT model, and an accuracy of 55% was obtained by DT using 10-fold cross-validation. Table 3 shows that the DT correctly classifies student specialization with a TPR of 0.87 and an FPR of 0.71.

Table 3. Confusion matrix of the decision tree classifier.

	TN = 10 (0.29)	**FP = 25 (0.71)**
Actually (0)	FN = 4 (0.13)	TP = 26 (0.87)
Actually (1)	**Predicted (0)**	**Predicted (1)**

In the second step of the first experiment, RF classifiers were used to predict student specialization given a student placement dataset. The RF classifier is a supervised learning algorithm that applies to both classification and regression problems [80]. RF creates multiple DTs from random sample data and then gives predictions on high-accuracy trees [91,92]. The RF classifier predicts the student field specialization with the following default parameters (bootstrap = True, ccp_alpha = 0.0, criterion = 'gini', max_depth = 15, max_features = 'auto',max_leaf_nodes = 10). The TPR of RF is 0.70, and the FPR is 0.20, as shown in Table 4.

Table 4. Confusion matrix of the random forest classifier.

	TN = 28 (0.80)	**FP = 7 (0.20)**
Actually (0)	FN = 9 (0.30)	TP = 21 (0.70)
Actually (1)	**Predicted (0)**	**Predicted (1)**

Extra tree classifiers (ETC) are used in the third step to predict the student's field of specialization. ETC is an ensemble learning technique that collects the result of multiple trees. The approach is similar to an RF classifier, but the tree construction method differs. The accuracy of the ETC classifier on the student dataset is 0.52. During training, the ETC Classifier used the default parameters (n_estimators = 100, random_state = 0) to predict student specialization with high accuracy. The TPR and FPR of the ETC classifier were 0.53 and 0.49, respectively, as shown in Table 5.

Table 5. Confusion matrix of the extra tree classifier.

	TN = 18 (0.51)	**FP = 17 (0.49)**
Actually (0)	FN = 14 (0.47)	TP = 16 (0.53)
Actually (1)	**Predicted (0)**	**Predicted (1)**

In the fourth step of the first experiment, we used the SVM classifier. SVM is a supervised learning algorithm that is mostly used for classification problems. SVM finds the hyperplane that divides a dataset into two classes [93]. SVM classifiers perform well on clean and small datasets. Furthermore, SVM is faster than other machine learning techniques [93]. The best accuracy (52%) of SVM was achieved with the following default parameters (random_state = 0, tol = 1×10^{-5}), as shown in Table 6. The TPR and FPR of SVM are 0.53 and 0.49, respectively, as shown in Table 7.

Table 6. Experimental results of the machine learning models on the current study dataset.

Model	**Accuracy**
Decision Tree Algorithm	0.5538
Random Forest Algorithm	0.7538
Extra Trees Classifier	0.5231
Support Vector Machine	0.5231
XGBoost	0.6154

Table 7. Confusion matrix of the support vector machine (SVM).

	TN = 18 (0.51)	**FP = 17 (0.49)**
Actually (0)	FN = 14 (0.47)	TP = 16 (0.53)
Actually (1)	**Predicted (0)**	**Predicted (1)**

Finally, we used the XGBoost classifier to predict student specializations by using default parameters (base_score = 0.5, booster = 'gbtree', colsample_bylevel = 1, learning_rate = 0.1). XGBoost is a popular boosting technique in ensemble ML, and its performance is good on structured and tabular data. XGBoost is also called GBC. XGBoost uses parallel tree boosting to solve real-life data science problems. We used this technique because its impact has been widely recognized in many machine learning and data mining challenges, where it has become a commonly used and popular tool among Kaggle's competitors and data scientists [87]. XGBoost predicts the student's specialization with an accuracy of 61%, and the TPR and FPR of the XGBoost classifier are 0.57 and 0.35, respectively, as shown in Table 8.

Table 8. Confusion matrix of the XGBoost classifier.

	TN = 23 (0.66)	**FP = 12 (0.34)**
Actually (0)	FN = 13 (0.43)	TP = 17 (0.57)
Actually (1)	**Predicted (0)**	**Predicted (1)**

In the first experiment, the TPR and accuracy of the RF and GBC classifiers were higher than DT, SVM, and ETC, and the FPR was lower than that of DT, SVM, and ETC, as shown in Figure 5, which indicates that the performance of these classifiers in predicting field specializations is good, compared to that of the alternatives. Sometimes, accuracy is misleading when the dataset is imbalanced [94–96]. In other words, if the ratio of some classes is less than that of others in the dataset, we used the ROC curve and TPR, which is also called recall. We used ROC to further understand the performance of the models. The ROC is an evaluation metric that represents the performance of an ML model in the form graph [97] by plotting the TPR and FPR of the model. Figure 5 shows that the TPR of the RF and GBC classifiers was high, compared to that of the DT, SVM, and ETC. The results also showed that RF and GBC classifiers are appropriate classifiers to build student field recommendation systems because they can handle ordinal, non-ordinal, and categorical data and are also good choices for skewed and multimodal data [98]. Moreover, the RF and GBC classifier ensemble method outperforms simple DT classifiers. The previous study showed that the performance of SVM in small data is good and faster [93]. Furthermore, DT and ETC are unstable and have high sensitivities for overfitting classifiers [15].

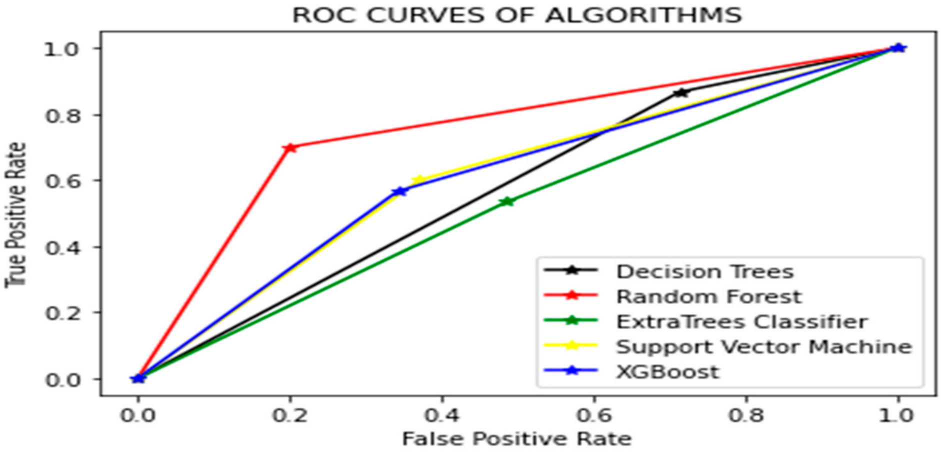

Figure 5. ROC curves of all the classifiers.

Question 2: How is a student's undergraduate major path choice associated with that student's academic performance and the job market?

We performed a second experiment to investigate the second research question. The second experiment investigated how a student's field of specialization was associated with that student's previous academic history, salary, and experience. First, we observed the baseline characteristics of different selected variables, such as secondary education percentage, higher secondary education percentage, degree percentage, MBA percentage, and employability test percentage. Then, we performed statistical analysis (Spearman correlation) to assess the relationship between a student's major and other input variables. The Spearman correlation shows how closely two variables are related. Table 9 shows the Spearman correlation between student field specialization and other input variables of the current study. The statistical results show that higher student marks in higher secondary (hsc_p), university degree (Degree_p), and entry test (etest_p) play a significant role in student field of specialization, and we can easily suggest study fields according to these features. Furthermore, student work experience (work_exp) and job placement (status) also impact student field specialization. Several interesting observations are obtained from the above statistical analysis. First, students at the high secondary stage are very excited about their field in university (undergraduate level) or undergraduate major path choice. At this level, every student wants to go into a good study field. In other words, students place high importance on student field specialization decisions. Second, students who get admitted to their favorite field graduate with higher grades. Third, the student field of specialization also affects student work experience and market salary. Fourth, students who applied for their favorite field may receive high marks on their university entry test. Fifth, students who graduate in their favorite field have a high chance of getting a job. In addition, the student marks percentage in a higher secondary, university degree, and entry test is useful criteria for study field suggestion. The result also demonstrates that student marks percentage in higher secondary, university degree, entry test assignment, and other factors are beneficial to the intelligent recommendation system. Using these variables, the proposed recommender system correctly predicts student field specialization according to their marks and preferences.

Table 9. Spearman correlation between student specialization and other input variables.

Input Features	r	*p* Value	Mean	Std
gender	−0.106	0.12	0.64651	0.47917
ssc_p	−0.17	0.01	67.3034	10.8272
ssc_b	−0.05	0.45	0.46047	0.499598
hsc_p	−0.24	0.00	66.3332	10.8975
hsc_b	0.002	0.97	0.6093	0.48905
hsc_s	0.17	0.01	1.37209	0.58098
degree_p	−0.21	DOT00	66.3702	7.35874
degree_t	0.08	0.21	0.6	0.89024
workex	−0.19	0.00	0.34419	0.47621
etest_p	−0.23	0.00	72.1006	13.276
mba_p	−0.1	0.12	0.44186	5.83339
status	−0.25	0.00	0.68837	0.46424
ssc_p_catg	0.15	0.02	0.50698	0.84759
hsc_p_catg	0.16	0.01	0.45116	0.80081
mba_p_catg	0.1	0.14	0.37674	0.4857
degree_p_catg	0.24	0.00	0.3814	0.78158
etest_p_catg	−0.011	0.86	0.641860	0.80716
Salary	−0.14	0.07	288,655.405	93,457.45

The current study results show that we can design a recommendation system for predicting the field of specialization using RF, GBC, and SVM classifiers. The proposed recommendation system will offer a variety of functions to students and college/university staff, such as recommending appropriate fields of study for students, ranking highly demanding fields in the coming and current years, and predicting the future salary of recommended fields. The results also show that higher secondary education is an appropriate stage to enter a good study field. Moreover, having a suitable specialization might affect students' academic performance and job salary, which could assist in lessening their anxiety and confusion and could lead to significantly better study program completion and increase graduation rates in the future. Having early awareness of the estimated number of incoming freshmen per study specialization program could also be of high value to the college administration. With this great insight, they would be able to allocate required resources per specialization field and better prepare schedules.

5. Conclusions

Unsuitable field of specialization selection for new graduate students has serious consequences for students and universities. Choosing an appropriate field of specialization is a critical determinant of a student's future academic and work progression. The current study used a machine learning and statistical approach to investigate the student study field. In this study, we extracted data from the Kaggle repository, which is publicly available for research purposes, and then converted these data into a form that is acceptable for ML models. We then applied several supervised learning techniques (DT, RF, ETC, and GBC) to our dataset and evaluated them using a 10-fold cross-validation method. The findings showed that RF and, GBC predict student study fields with accuracy 0.75% and 0.61 respectively. The results indicate that RF and GBC are the most appropriate, classifiers to integrate into the intelligent field recommender system for predicting suitable fields for students according to the job market because the performance of these classifiers is good on less training data Additionally, the intelligent field recommender system will help educational institutions to suggest study fields according to the current job market and demand. Using this recommendation system, students can select a field that is according to the job market. The study also demonstrated that the student field of specialization selection is mostly dependent on the percentage of marks in higher secondary, university, and entry tests. Student work experience and student job placement also affect the student's field of study. Furthermore, student mark percentage in higher secondary, university, and entry tests are appropriate criteria for all higher education institution admission departments to select the right undergraduate major path choice.

This experiment aims to investigate whether these data could be used to suggest an appropriate study field for students. This study used student academic data and job market data from the Kaggle repository. The student's field of specialization is a complex problem that also depends on other factors, such as country and student family background. Therefore, these factors must be further investigated.

The Current study limitations: There are some limitations, for example, the current study has limited specializations, records, and input features. In the future, we will use design surveys to assess other factors or input features related to the student field of specialization. Additionally, the accuracy of RF, GBC, and SVM models will be further improved by increasing the number of observations and hyperparameter tuning. Then, we will build an intelligent field recommendation system using collaborative filtering to recommend suitable fields to students according to their preferences and the job market. This proposed system will help the university admission system make quick decisions about student field recommendations.

Author Contributions: Conceptualization, A.O.A. and M.S.M.R.; Methodology, M.H., S.H.J. and S.H.; Software, M.H.; Validation, M.H., M.S.M.R. and M.L.J.; Formal Analysis, A.O.A. and M.H., S.H.J.; Investigation, A.O.A. and S.H.J.; Resources, A.O.A.; Data Curation, M.H.; Writing—Original Draft Preparation, A.O.A. and N.A.; Writing—Review & Editing, A.O.A., S.H.J. and M.H.; Visualization, M.H.; Supervision, M.S.M.R., I.A. and M.H.; Project Administration, A.O.A., M.H. and N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research work is funded under the project number IFPIP:536-830-1442.

Institutional Review Board Statement: Data is being taken from an authorized open source site (Kaggle).

Informed Consent Statement: Not applicable.

Data Availability Statement: The current study data are publicly available online (https://www.kaggle.com/benroshan/factors-affecting-campus-placement 7 November 2021) for research purposes. No participants' personal information (e.g., name or address) was included in this study.

Acknowledgments: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number IFPIP:536-830-1442 and King Abdulaziz University, DSR, Jeddah, Saud Arabia.

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

References

1. Mengash, H.A. Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. *IEEE Access* **2020**, *8*, 55462–55470. [CrossRef]
2. Fong, S.; Si, Y.-W.; Biuk-Aghai, R. Applying a hybrid model of neural network and decision tree classifier for predicting university admission. In Proceedings of the 2009 7th International Conference on Information, Communications and Signal Processing (ICICS), Macau, China, 8–10 December 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 1–5.
3. Alshaikh, K.; Bahurmuz, N.; Torabah, O.; Alzahrani, S.; Alshingiti, Z.; Meccawy, M. Using Recommender Systems for Matching Students with Suitable Specialization: An Exploratory Study at King Abdulaziz University. *Int. J. Emerg. Technol. Learn.* **2021**, *16*, 316–324. [CrossRef]
4. Aulck, L.; Velagapudi, N.; Blumenstock, J.; West, J. Predicting student dropout in higher education. In Proceedings of the 33rd International Conference on Machine Learning (ICML) Workshop on #Data4Good: Machine Learning in Social Good Applications, New York, NY, USA, 24 June 2016; pp. 16–20.
5. Elbadrawy, A.; Polyzou, A.; Ren, Z.; Sweeney, M.; Karypis, G.; Rangwala, H. Predicting Student Performance Using Personalized Analytics. *Computer* **2016**, *49*, 61–69. [CrossRef]
6. Leu, K. *Beginning College Students Who Change Their Majors within 3 Years of Enrollment*; NCES: Washington, DC, USA, 2017.
7. Atuahene, F. An analysis of major and career decision-making difficulties of exploratory college students in a Mid-Atlantic University. *SN Soc. Sci.* **2021**, *1*, 80. [CrossRef]
8. Yeyie, P. Selecting Program of Study for Undergraduate Students in the Valley View University, Kumasi. *Soc. Educ. Res.* **2021**, *2*, 315–330. [CrossRef]
9. Ezz, M.; Elshenawy, A. Adaptive recommendation system using machine learning algorithms for predicting student's best academic program. *Educ. Inf. Technol.* **2019**, *25*, 2733–2746. [CrossRef]
10. Bautista, R.; Dumlao, M.; Ballera, M. Recommendation system for engineering students' specialization selection using predictive modeling. In Proceedings of the Third International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM2016), Thessaloniki, Greece, 13–15 May 2016; SDIWC: Lodz, Poland, 2016; pp. 34–40.
11. Al-Shalabi, L. A Data Mining Model for Students' Choice of College Major Based on Rough Set Theory. *J. Comput. Sci.* **2019**, *15*, 1150–1160. [CrossRef]
12. Eydi, M.; Moradi, Z.; Randian, R.; Rahdari, A.; Aliabadi, A. A Model to Determine Effective Factors on Pharmacy Major Selection (A Case Study: Students of Zabol University of Medical Sciences). *J. Pharm. Res. Int.* **2017**, *17*, 1–8. [CrossRef]
13. Reddy, M.Y.S.; Govindarajulu, P. College recommender system using student'preferences/voting: A system development with empirical study. *Int. J. Comput. Sci. Netw. Secur.* **2018**, *18*, 87–98.
14. Alghamdi, S.; Alzhrani, N.; Algethami, H. Fuzzy-Based Recommendation System for University Major Selection. In Proceedings of the 11th International Joint Conference on Computational Intelligence, Vienna, Austria, 17–19 September 2019; SCITEPRESS—Science and Technology Publications: Setúbal, Portugal, 2019; pp. 317–324.
15. Hattie, J.; Timperley, H. The Power of Feedback. *Rev. Educ. Res.* **2007**, *77*, 81–112. [CrossRef]
16. Kazi, A.S.; Akhlaq, A. Factors affecting students' career choice. *J. Res. Reflect. Educ.* **2017**, *2*, 187–196.
17. Astorne-Figari, C.; Speer, J.D. Are changes of major major changes? The roles of grades, gender, and preferences in college major switching. *Econ. Educ. Rev.* **2019**, *70*, 75–93. [CrossRef]
18. Bettinger, E.P.; Baker, R.B. The Effects of Student Coaching. *Educ. Eval. Policy Anal.* **2014**, *36*, 3–19. [CrossRef]
19. Long, M.; Ferrier, F.; Heagney, M. *Stay, Play or Give It Away? Students Continuing, Changing or Leaving University Study in First Year*; Centre for the Economics of Education and Training, Monash University: Melbourne, Australia, 2006.
20. Sharma, S.; Gupta, Y.K. Predictive analysis and survey of COVID-19 using machine learning and big data. *J. Interdiscip. Math.* **2021**, *24*, 175–195. [CrossRef]
21. Ağbulut, Ü.; Gürel, A.E.; Biçen, Y. Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renew. Sustain. Energy Rev.* **2021**, *135*, 110114. [CrossRef]
22. Haupt, S.E.; Cowie, J.; Linden, S.; McCandless, T.; Kosovic, B.; Alessandrini, S. Machine Learning for Applied Weather Prediction. In Proceedings of the 2018 IEEE 14th International Conference on e-Science (e-Science), Amsterdam, The Netherlands, 29 October–1 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 276–277.
23. Stefanovič, P.; Štrimaitis, R.; Kurasova, O. Prediction of Flight Time Deviation for Lithuanian Airports Using Supervised Machine Learning Model. *Comput. Intell. Neurosci.* **2020**, *2020*, 8878681. [CrossRef] [PubMed]

24. Subudhi, S.; Verma, A.; Patel, A.B.; Hardin, C.C.; Khandekar, M.J.; Lee, H.; McEvoy, D.; Stylianopoulos, T.; Munn, L.L.; Dutta, S.; et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *Npj Digit. Med.* **2021**, *4*, 87. [CrossRef]
25. Gogas, P.; Papadimitriou, T.; Agrapetidou, A. Forecasting bank failures and stress testing: A machine learning approach. *Int. J. Forecast.* **2018**, *34*, 440–455. [CrossRef]
26. Alam, T.M.; Shaukat, K.; Hameed, I.A.; Luo, S.; Sarwar, M.U.; Shabbir, S.; Li, J.; Khushi, M. An investigation of credit card default prediction in the imbalanced datasets. *IEEE Access* **2020**, *8*, 201173–201198. [CrossRef]
27. Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Xu, M. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access* **2020**, *8*, 222310–222354. [CrossRef]
28. Alam, T.M.; Shaukat, K.; Mushtaq, M.; Ali, Y.; Khushi, M.; Luo, S.; Wahab, A. Corporate bankruptcy prediction: An approach towards better corporate world. *Comput. J.* **2020**, *65*. [CrossRef]
29. Javed, U.; Shaukat, K.; Hameed, I.A.; Iqbal, F.; Alam, T.M.; Luo, S. A review of content-based and context-based recommendation systems. *Int. J. Emerg. Technol. Learn.* **2021**, *16*, 274–306. [CrossRef]
30. Shin, J.C.; Harman, G. New challenges for higher education: Global and Asia-Pacific perspectives. *Asia Pac. Educ. Rev.* **2009**, *10*, 1–13. [CrossRef]
31. Anoopkumar, M.; Rahman, A.M.J.M.Z. A review on data mining techniques and factors used in educational data mining to predict student amelioration. In Proceedings of the International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, India, 16–18 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 122–133.
32. Isma'Il, M.; Haruna, U.; Aliyu, G.; Abdulmumin, I.; Adamu, S. An Autonomous Courses Recommender System For Undergraduate Using Machine Learning Techniques. In Proceedings of the 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS), Ayobo, Nigeria, 18–21 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
33. Tan, L.; Main, J.B.; Darolia, R. Using random forest analysis to identify student demographic and high school-level factors that predict college engineering major choice. *J. Eng. Educ.* **2021**, *110*, 572–593. [CrossRef]
34. Dhar, J.; Jodder, A.K. An Effective Recommendation System to Forecast the Best Educational Program Using Machine Learning Classification Algorithms. *Ingénierie Des Syst. Egrave Mes Inf.* **2020**, *25*, 559–568. [CrossRef]
35. Meng, Y.; Fu, M. CMRS: Towards Intelligent Recommendation for Choosing College Majors. In Proceedings of the 2020 4th International Conference on Advances in Image Processing, Chengdu, China, 13–15 November 2020; pp. 152–157.
36. Baskota, A.; Ng, Y.-K. A Graduate School Recommendation System Using the Multi-Class Support Vector Machine and KNN Approaches. In Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 6–9 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 277–284.
37. Baker, R.S.; Yacef, K. The State of Educational Data Mining in 2009: A Review and Future Visions. *J. Educ. Data Min.* **2009**, *1*, 3–17. [CrossRef]
38. Shaukat, K.; Nawaz, I.; Aslam, S.; Zaheer, S.; Shaukat, U. Student's performance in the context of data mining. In Proceedings of the 2016 19th International Multi-Topic Conference (INMIC), Islamabad, Pakistan, 5–6 December 2016; pp. 1–8.
39. Alam, T.M.; Mushtaq, M.; Shaukat, K.; Hameed, I.A.; Umer Sarwar, M.; Luo, S. A Novel Method for Performance Measurement of Public Educational Institutions Using Machine Learning Models. *Appl. Sci.* **2021**, *11*, 9296. [CrossRef]
40. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R. Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. *Comput. Intell. Neurosci.* **2018**, *2018*, 6347186. [CrossRef] [PubMed]
41. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R.; Ali, S. Using machine learning to predict student difficulties from learning session data. *Artif. Intell. Rev.* **2018**, *52*, 381–407. [CrossRef]
42. El-Qulity, S.A.; Mohamed, A.; Bafail, A.O.; Abdelaal, R.M.S. A Multistage Procedure for Optimal Distribution of Preparatory-Year Students to Faculties and Departments: A Mixed Integer Nonlinear Goal Programming Model with Enhanced Differential Evolution Algorithm. *J. Comput. Theor. Nanosci.* **2016**, *13*, 7847–7863. [CrossRef]
43. Sahin, A.; Waxman, H.C.; Demirci, E.; Rangel, V.S. An Investigation of Harmony Public School Students' College Enrollment and STEM Major Selection Rates and Perceptions of Factors in STEM Major Selection. *Int. J. Sci. Math. Educ.* **2020**, *18*, 1249–1269. [CrossRef]
44. Powar, V.; Girase, S.; Mukhopadhyay, D.; Jadhav, A.; Khude, S.; Mandlik, S. Analysing recommendation of colleges for students using data mining techniques. In Proceedings of the 2017 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 1–2 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–5.
45. Xiao, M.; Yi, H. Building an efficient artificial intelligence model for personalized training in colleges and universities. *Comput. Appl. Eng. Educ.* **2021**, *29*, 350–358. [CrossRef]
46. Awaliyah, M.M.; Kurniawati, A.; Rizana, A.F. Profile matching for students specialization in industrial engineering major. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *830*, 032063. [CrossRef]
47. Pertiwi, D.A.; Daniawan, B.; Gunawan, Y. Analysis And Design of Decision Support System in Major Assignment at Buddhi High School Using AHP and SAW Methods. *Tech-E* **2019**, *3*, 13–21. [CrossRef]
48. Bhargava, N.; Sharma, G.; Bhargava, R.; Mathuria, M. Decision tree analysis on J48 algorithm for data mining. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2013**, *3*, 1114–1119.
49. Hsu, C.W.; Chang, C.C.; Lin, C.J. *A Practical Guide to Support Vector Classification*; Department of Computer Science and Information Engineering, National Taiwan University: Taipei, Taiwan, 2003; pp. 1–16.

50. Pupara, K.; Nuankaew, W.; Nuankaew, P. An institution recommender system based on student context and educational institution in a mobile environment. In Proceedings of the 2016 International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 14–17 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–6.
51. Salaki, R.J.; Kawet, C.R.; Manoppo, R.; Tumimomor, F. Decision support systems major selection vocational high school in using fuzzy logic android-based. In Proceedings of the International Conference on Electrical Engineering, Informatics, and Its Education (CEIE) 2015, Malang, Indonesia, 3 October 2015; CEIE: Fairfax, Virginia, 2015; pp. 1–6.
52. Fiarni, C.; Sipayung, E.M.; Tumundo, P.B. Academic Decision Support System for Choosing Information Systems Sub Majors Programs using Decision Tree Algorithm. *J. Inf. Syst. Eng. Bus. Intell.* **2019**, *5*, 57–66. [CrossRef]
53. Kularbphettong, K.; Tongsiri, C. Mining educational data to support students' major selection. *Int. J. Educ. Pedagog. Sci.* **2014**, *8*, 21–23.
54. Wei, Y.; Ni, N.; Liu, D.; Chen, H.; Wang, M.; Li, Q.; Cui, X.; Ye, H. An Improved Grey Wolf Optimization Strategy Enhanced SVM and Its Application in Predicting the Second Major. *Math. Probl. Eng.* **2017**, *2017*, 9316713. [CrossRef]
55. Sethi, K.; Jaiswal, V.; Ansari, M.D. Machine Learning Based Support System for Students to Select Stream (Subject). *Recent Adv. Comput. Sci. Commun.* **2020**, *13*, 336–344. [CrossRef]
56. Samra, G.E.A.; Faloudah, A. Machine Learning based Marks Prediction to Support Recommendation of Optimum Specialization and Study Track. *Int. J. Comput. Appl.* **2019**, *181*, 15–25. [CrossRef]
57. Latifah, S.N.; Andreswari, R.; Hasibuan, M.A. Prediction Analysis of Student Specialization Suitability using Artificial Neural Network Algorithm. In Proceedings of the 2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC), Bandung, Indonesia, 20–22 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 355–359.
58. Zubaedah, R.; Lintang, M.; Putra, N.P. Decision Support System for Departemen Selection for Prospective Students using the Naïve Bayes Method and Analytical Hierarchy Process Model at Faculty of Engineering Universitas Musamus. *IOP Conf. Series: Mater. Sci. Eng.* **2021**, *1125*, 012030. [CrossRef]
59. Tamiza, L.; Shahin, G.; Tahboub, R. Intelligent Model for Suitable University Specialization Selection in Palestine. In Proceedings of the 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), Aqaba, Jordan, 28 October–1 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–8.
60. Iyer, S.; Variawa, C. 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, ON, Canada, 8–12 June 2019; Queen's University Library: Kingston, ON, Canada, 2019; pp. 1–17.
61. AymanAlAhmar, M. A Prototype Rule-based Expert System with an Object-Oriented Database for University Undergraduate Major Selection. *Int. J. Appl. Inf. Syst.* **2012**, *4*, 38–42. [CrossRef]
62. Kamal, N.; Sarker, F.; Mamun, K.A. A Comparative Study of Machine Learning Approaches for Recommending University Faculty. In Proceedings of the 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 19–20 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
63. Ben, R. Placement_Data_Full_Class.csv. Available online: https://www.kaggle.com/benroshan/factors-affecting-campus-placement (accessed on 22 July 2021).
64. Beggs, J.M.; Bantham, J.H.; Taylor, S. Distinguishing the factors influencing college students' choice of major. *Coll. Stud. J.* **2006**, *42*, 381–395.
65. Strange, C.; Gordon, V.N. The Undecided College Student: An Academic and Career Advising Challenge. *J. High. Educ.* **1986**, *57*, 113. [CrossRef]
66. Damayanti, A.S.; Wibawa, A.P.; Pujianto, U.; Nafalski, A. The Use of Adaptive Neuro Fuzzy Inference System in Determining Students' Suitable High School Major. In Proceedings of the 2018 4th International Conference on Education and Technology (ICET), Malang, Indonesia, 26–28 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
67. Stein, S.A.; Weiss, G.M.; Chen, Y.; Leeds, D.D. A College Major Recommendation System. In Proceedings of the Fourteenth ACM Conference on Recommender Systems, Virtual Event, Brazil, 22–26 September 2020; pp. 640–644.
68. Chen, L.; Pratt, J.A.; Cole, C.B. Factors Influencing Students' Major and Career Selection in Systems Development: An Empirical Study. *J. Comput. Inf. Syst.* **2016**, *56*, 313–320. [CrossRef]
69. Crampton, W.J.; Walstrom, K.A.; Schambach, T.P. Factors influencing major selection by college of business students. *Issues Inf. Syst.* **2006**, *7*, 226–230.
70. Han, S. Korean Students' Attitudes toward STEM Project-Based Learning and Major Selection. *Educ. Sci. Theory Pract.* **2017**, *17*, 529–548. [CrossRef]
71. Kim, Y.-J.; Yoo, H.; Park, M. Effect of Motive for Major Selection on Major Satisfaction, Campus-life Satisfaction, and Self-directed Learning Ability among Nursing Students. *J. Korea Acad. Coop. Soc.* **2016**, *17*, 261–270. [CrossRef]
72. Khasanah, F.N.; Permanasari, A.E.; SuningKusumawardani, S. Fuzzy MADM for major selection at senior high school. In Proceedings of the 2015 2nd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia, 16–18 October 2015; pp. 41–45.
73. Rabani, R.; Rabiei, K. Evaluation of Major Selection and its Impact on Educational Satisfaction among Isfahan University Students. *IRPHE* **2011**, *17*, 99–120.
74. Lobb, W.B.; Shah, M.; Kolassa, E.M. Factors Influencing the Selection of a Major: A Comparison of Pharmacy and Nonpharmacy Undergraduate Students. *J. Pharm. Teach.* **2004**, *11*, 45–64. [CrossRef]

75. Ullah, Z.; Saleem, F.; Jamjoom, M.; Fakieh, B. Reliable Prediction Models Based on Enriched Data for Identifying the Mode of Childbirth by Using Machine Learning Methods: Development Study. *J. Med. Internet Res.* **2021**, *23*, e28856. [CrossRef]
76. Nti, I.K.; Adekoya, A.F.; Weyori, B.A. Efficient Stock-Market Prediction Using Ensemble Support Vector Machine. *Open Comput. Sci.* **2020**, *10*, 153–163. [CrossRef]
77. Bresfelean, V.P. Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment. In Proceedings of the 2007 29th International Conference on Information Technology Interfaces, Cavtat, Croatia, 25–28 June 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 51–56.
78. Dervisevic, O.; Zunic, E.; Eonko, D.; Buza, E. Application of KNN and Decision Tree Classification Algorithms in the Prediction of Education Success from the Edu720 Platform. In Proceedings of the 2019 4th International Conference on Smart and Sustainable Technologies (SpliTech), Split, Croatia, 18–21 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
79. Qu, W.; Tan, G.; Zeng, Q.; Xu, X. Based on the SVM university education's quality regression analysis. In Proceedings of the Third International Symposium on Intelligent Information Technology Application, Nanchang, China, 21–22 November 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 306–309.
80. Beaulac, C.; Rosenthal, J.S. Predicting University Students' Academic Success and Major Using Random Forests. *Res. High. Educ.* **2019**, *60*, 1048–1064. [CrossRef]
81. Patil, R.; Tamane, S. A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes. *Int. J. Electr. Comput. Eng.* **2018**, *8*, 3966–3975. [CrossRef]
82. Hämäläinen, W.; Vinni, M. Classifiers for educational data mining. In *Handbook of Educational Data Mining*; Romero, C., Pechenizkiy, M., Baker, R.S.J.D., Ventura, S., Eds.; Chapman & Hall/CRC Press: Boca Raton, FL, USA, 2010; pp. 57–74.
83. Hern, A. Why Data Is the New Coal. Available online: https://www.theguardian.com/technology/2016/sep/27/data-efficiency-deep-learning (accessed on 24 September 2021).
84. Jeay, S.; Gaulis, S.; Ferretti, S.; Bitter, H.; Ito, M.; Valat, T.; Murakami, M.; Ruetz, S.; Guthy, D.A.; Rynn, C.; et al. A distinct p53 target gene set predicts for response to the selective p53–HDM2 inhibitor NVP-CGM097. *eLife* **2015**, *4*. [CrossRef] [PubMed]
85. Cao, L.; Tay, F. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Netw.* **2003**, *14*, 1506–1518. [CrossRef] [PubMed]
86. Huang, S.; Fang, N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Comput. Educ.* **2013**, *61*, 133–145. [CrossRef]
87. Asselman, A.; Khaldi, M.; Aammou, S. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interact. Learn. Environ.* **2021**. [CrossRef]
88. Huo, H.; Cui, J.; Hein, S.; Padgett, Z.; Ossolinski, M.; Raim, R.; Zhang, J. Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach. *J. Coll. Stud. Retent. Res. Theory Pract.* **2020**. [CrossRef]
89. Kohavi, R.; Provost, F. Glossary of terms. Machine learning—Special issue on applications of machine learning and the knowledge discovery process. *Mach. Learn.* **1998**, *30*, 271–274.
90. Song, Y.-Y.; Lu, Y. Decision tree methods: Applications for classification and prediction. *Shanghai Arch Psychiatry* **2015**, *27*, 130–135. [CrossRef] [PubMed]
91. Jin, Z.; Shang, J.; Zhu, Q.; Ling, C.; Xie, W.; Qiang, B. RFRSF: Employee turnover prediction based on random forests and survival analysis. In *Web Information Systems Engineering—WISE WISE Lecture Notes in Computer Science*; Huang, Z., Beek, W., Wang, H., Zhou, R., Zhang, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 503–515.
92. Cheng, C.; Yan, X.; Sun, F.; Li, L.M. Inferring activity changes of transcription factors by binding association with sorted expression profiles. *BMC Bioinform.* **2007**, *8*, 452. [CrossRef] [PubMed]
93. Shahiri, A.M.; Husain, W.; Rashid, N.A. A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Comput. Sci.* **2015**, *72*, 414–422. [CrossRef]
94. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
95. Khushi, M.; Shaukat, K.; Alam, T.M.; Hameed, I.A.; Uddin, S.; Luo, S.; Yang, X.; Reyes, M.C. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access* **2021**, *9*, 109960–109975. [CrossRef]
96. Alam, T.M.; Shaukat, K.; Mahboob, H.; Sarwar, M.U.; Iqbal, F.; Nasir, A.; Hameed, I.A.; Luo, S. A machine learning approach for identification of malignant mesothelioma etiological factors in an imbalanced dataset. *Comput. J.* **2021**. [CrossRef]
97. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
98. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]