Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.3343

2024, 43(4) 192-213

Career planning matters: Intelligence-based career path predictions using data mining models − A longitudinal study

Khalid Mahboob [a, *], Raheela Asif [b], Najmi Ghani Haider [c]

[a] *College of Computer Science & Information Systems, Institute of Business Management, Karachi, Pakistan*

[b] *Department of Software Engineering, NED University of Engineering and Technology, Karachi, Pakistan*

[c] *Department of Computing Sciences, UIT University, Karachi, Pakistan*

[*] Corresponding author: Khalid Mahboob, Email: khalid.mahboob@iobm.edu.pk

Received: 13 August 2024, Accepted: 21 September 2024, Published: 01 October 2024

KEYWORDS

Career Path

Data Mining

Education

Models

SMOTE

ABSTRACT

It is essential for students to plan their careers because selecting the right career path shapes a person's life. In such disciplines as computer science, information technology, and software engineering, assisting students toward appropriate employment is even more helpful. Thus, throughout the student's education, they must evaluate their strengths to determine which professional sector corresponds to their abilities. It is for this reason that this research introduces an intelligence-based career recommendation system that will incorporate the analysis of such factors as the student's academic performance, economic status, and demographic features to use data mining models in determining the best prospective career path to offer the student a more transparent and more informed vision and course to set on in the future. Three key aspects were addressed: first, one or another model and classifier for assessing the impact of pre-university education on the choice of a profession and, consequently, the selection of technologies were used. Second, these models accurately forecast students' careers using core courses, CGPA, and FYP data. Third, socioeconomic or demographic data was incorporated into the prediction to make it more accurate. Regarding the method of class distribution balancing, the Synthetic Minority Oversampling Technique (SMOTE) approach was used. The study reveals that variables specifying pre-university education directly impact students' career choices and that, employing data mining techniques, career choices could be forecasted considering academic performance and other related factors.

1. Introduction

1.1 Theoretical Background

In this contemporary era, competition is growing day by day. Qualified and trained human resources play a vital role in the strategy of national economic prosperity. Skilled labor is crucial to ensure a good ranking among world nations. To compete and reach the goal, the student must plan and prepare from the inception of education [1]. So, it is crucial to evaluate their performance, identify their interests, assess their understanding and goals, and evaluate whether they are on the right track [2]. This can help students to improve and prepare themselves for a better career according to their aspirations and interests. A clear and stable notion of educational objectives, interests, and career identities can be formed through a necessary career exploration and subsequent participation of students at high school, college, or university. While pursuing an education, many adolescents are still determining their careers. So,

career counseling is essential to help students choose the right career [3].

Career counseling is crucial in higher education to engage graduates in developing their personal and professional skills to benefit their communities [4]. Much of the work done in this area focuses on predicting student performance based on their educational outcomes, encompassing pre-university results or undergraduate semester results [5]. However, choosing the right profession appropriate for graduates is another crucial aspect. Many students need to gain knowledge of the proper career [6]. Additionally, each career has several job roles that graduates can choose based on the strengths of their studies.

However, choosing the right career is still one of the most critical aspects for graduates according to their educational accomplishments and competencies, and it is challenging to select suitable career options when there are so many job roles available [7]. However, lack of good counseling and increased unemployment complicate career choices. Competitions are strengthened in terms of careers in many areas [8]. That is why most graduates looking for a suitable job need clarification and indecision after graduation. Choosing the right career path is important because it depends on success or failure in one's life. Making the right decisions can lead to a rewarding career and a successful life. Instead, choosing the wrong career path can lead to failure, dissatisfaction, and sadness [9].

In this context, educational data mining (EDM) emerges as a beacon of hope, attracting significant interest in improving academic outcomes and decision-making. It focuses on constant search and significant research challenges with data-driven analysis to identify graduates and their educational backgrounds related to career choices. This study aims to harness the power of intelligence-based data mining models to predict appropriate career paths based on graduates' academic, socioeconomic, and demographic information. By understanding the needs of prospective students, their learning attitude, interests, aptitude, difficulties, and suitable job roles, we can pave the way for a more informed and successful career selection process.

1.2 Problem Statement

The role of higher education institutions (HEIs) in career mentoring and professional recruitment is not just significant but integral. HEIs are crucial in involving graduates in their personal and professional development, which benefits their communities. In this context, HEIs strive to gain accreditation from the Higher Education Commission (HEC), Pakistan, to preserve and enhance academic quality. This accreditation represents that the HEC's requirements are met and that educational activities comply with the Commission's regulations. Therefore, it is essential to identify the coherent relationships between the similarities and diversities among the parameters of the features of different HEI datasets. These datasets comprise graduates' educative experiences, outcomes, and socioeconomic or demographic information, and understanding these relationships is crucial for predicting prospective students' career paths and guiding their career selection and direction.

1.3 Research Objectives and Questions

In this study, an intelligent career recommendation system using data mining predictive models is developed based on the combination of the predictions of educational attainments in pre-university (Secondary School Certificate (SSC) / Higher Secondary School Certificate (HSC)) level of education, courses of studies, Cumulative Grade Point Average (CGPA), Final Year Project (FYP), and socioeconomic or demographic information of Software Engineering graduates' belonging to two different HEIs to find out the capacity for satisfying all demands, including providing direction and support to students in selecting a career path that is a good fit according to their interests and skills [15−18]. The rationale of the current study is to predict information about students' career paths, which might assist them in determining whether specific job roles are a better fit for them. The principal objectives of this research are threefold:

1) Firstly, various classifiers are used to identify the role of the pre-university (Secondary School Certificate (SSC) / Higher Secondary School Certificate (HSC)) level of education in influencing students' career choices and correct decision-making on technology selection. Only admission data from SSC and HSC levels of education are employed to develop these classifiers. It is crucial to envisage if acceptable outcomes may be achieved from pre-university data to benefit technology selection decision-making.
2) Secondly, the courses of studies (specifically core courses), CGPA, and FYP information can effectively measure students' achievement in a degree program relating to their career placement using these classifiers. A trade-off between a classifier's predictive

capability and the interpretability of its model may be crucial.

3) Thirdly, the graduates' socioeconomic or demographic information is derived because the importance of socioeconomic or demographic status must be considered in this study since it contributes to determining the suitable careers and qualifications to pursue. The intersection of graduates' career goals in a 4-year degree program with demographic features such as parents'/guardians' occupation, qualification, income, internship experience, graduates' gender, skills, competencies, etc., are mapped to add to the evidence foundation for academic, pedagogical, and administrative effort on higher education access and the policies, procedures, and consequences that may ensue by applying these classifiers.

In light of the above objectives, the following three questions are investigated:

Research Question 1: To what extent does the pre-university (SSC/HSC) level of education influence career choices to correct decision-making on technology selection using data mining models?

Research Question 2: Is it possible to predict a career path in advance using data mining models based on the attainments in the courses of studies, CGPA, and FYP with comparable accuracies?

Research Question 3: Is it possible to predict a career path in advance using data mining models based on socioeconomic or demographic information with comparable accuracies?

The conceptual representation of different possible career paths is depicted in Fig. 1.

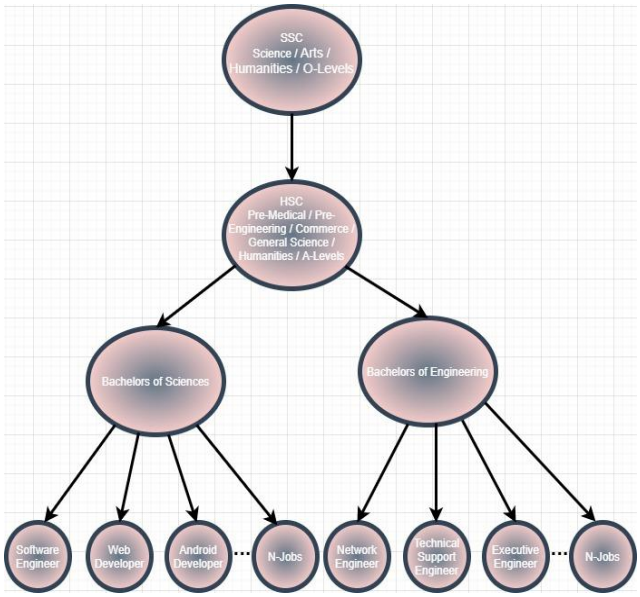

Fig. 1. Conceptual Representation of Career Paths

The remaining sections of the paper are organized as follows. The next section is focused on a literature review, followed by a description of predictive data mining models in section 3. The data preparation and methodology for this investigation are then described in section 4, followed by the results and analysis in section 5 and discussion in section 6. The conclusion is presented in the final section.

2. Literature Review

This research explicitly encompasses educational data mining techniques, concentrating on predicting students' academic performance and correlating it to predicting students' career paths. This review embraces the study's strengths and weaknesses in the current literature and its significant contributions to the domain.

2.1 Related Works on Predicting Students' Educational Performance

A study in [16] used several data mining tools to investigate graduate students' academic performance across four years. First, the study looked at two characteristics of students' performance to predict their accomplishment after a four-year degree program; second, it looked at typical progressions and combined them with predicted outcomes. The research used artificial neural networks, decision tree induction, k-nearest neighbors, naive bayes, random forest trees, rule induction, and clustering approaches such as k-means and x-means, among other classification techniques. Consequently, two distinct groups of students were identified: low achievers and high achievers. According to the findings, only a limited number of courses are indicators of good or bad performance. It may be feasible to offer early warnings, help low-achieving students, counsel, and provide chances to high-performing students.

The research in [19] investigated the effectiveness of deep learning in EDM, particularly in predicting students' academic performance and identifying students at risk of failure. This study used a four-year dataset from a public university to develop predictive models that used a deep neural network (DNN), decision tree, gradient boosting, k-nearest neighbor, logistic regression, random forest, and support vector classifier to predict students' academic performance in upcoming courses based on their grades in previous courses during the first academic year. In addition, it compares resampling methods for handling unbalanced data sets, such as ADASYN, ROS, SMOTE, and SMOTE-ENN. It has been observed that the proposed DNN model can predict students' performance in a data structure course and can also identify students at risk of failure at an early stage of

a semester with an accuracy of 89%, which is higher than other models according to the experimental results.

The data on students at the time of admission and their academic performance was obtained on a semester-by-semester basis in the research presented in [20]. Pre-entry attributes and data about study accomplishments in the first four semesters were harnessed for classification; however, only pre-entry attributes were used for dependency analysis. The study considers interpretable models of data mining to be better. The decision trees and association rules were used to predict the students who successfully finished their studies. The results and findings show that the percentage of lost credit vouchers in the most recent semesters was the most crucial factor. The pre-entry characteristics has a negligible effect. Association rules were constructed to identify features of students who did not complete the first semester of study. The results show that this is due to the time gap between secondary and tertiary education, which is the critical factor that raises the probability of failure.

The k-means and x-means clustering algorithms were used to investigate data to determine the correlation of students' performance in [21] to determine influencing factors. The results of this study show that several personal and societal factors, such as parental employment, parental credentials, and income levels, have substantial implications for students' performance. This research demonstrates that these two algorithms provide the same relevant association between a student's GPA and other characteristics. The results suggest that parental employment, credentials, income level, and the number of hours spent with friends each week significantly affect students' academic achievement. Despite this, the percentage of high school graduates, family size, method of transportation, parental status, and number of friends were insignificant.

Student's academic performance is analyzed and predicted in [22] using the salient theories of clustering, convolution neural network, and discrimination to keep track of the student's future performance in advance. Three datasets, A, B, and C, from a university, were chosen for the study because they all contained students enrolled in the same courses. The first suggestion made in this study is to employ a statistic that has never been used in the K-means method to optimize how the clustering number is determined. Then, using discriminant analysis, the grouping impact of the K-means method is evaluated. Only data from compulsory courses were considered. The convolutional neural network algorithm is used for training and testing data. The model that was created can be used to forecast future performance. Finally, the efficacy of the constructed model is assessed using two metrics in two cross-validation approaches to validate the prediction findings. The experiment's findings show that the statistic makes it easier to estimate the clustering number in the K-means algorithm from a quantitative and objective aspect and increases the accuracy of prediction outcomes.

2.2 Related Works on Predicting Students' Career Paths

A study in [23] proposed a recommendation system for career path selection, including decision trees and linear regression methods. The system uses five modules: students, administrators, suggestions, feedback, and chatbots. According to the research results, a recommendation system was designed for students, and many tests were conducted to help them choose the professional path that best suits their interests. The researchers looked at the students' work and created a graph to show the results.

CareerRec, a recommendation system based on machine learning algorithms presented in [24], is designed to assist IT graduates in choosing a career path primarily based on their skills. A dataset of 2255 employees in Saudi Arabia's IT industry was used to train and evaluate CareerRec. In this study, the accuracy of five machine learning techniques, namely k-nearest neighbors (KNN), decision tree (DT), bagging meta-estimator, gradient boosting, and XGBoost, were compared to predict the best-suited career path among three classes, i.e., Analyst, Developer, and Engineer. The results show that compared with other models, the XGBoost algorithm performs better than other models and provides the highest accuracy rate (70.47%).

A web application system was presented in [25] to provide input to the student and formulate and show the final prognosis using advanced machine learning algorithms consisting of a support vector machine (SVM), XGBoost, and decision trees. Many characteristics are considered for student career predictions, such as academic scores in various disciplines and specialties, personality, memory, programming and analytical ability, personal relationships, hobbies, sports, competitions, interest in hackathons, courses, certifications, books, etc. Those criteria significantly determine a student's academic success in a specific job. Only the classifier's accuracy was compared. The results show that the support vector machine (SVM) gave the highest accuracy, i.e., 90.3%.

The fundamental idea behind the approach in [26] is to determine whether or not a student is interested in continuing their education at a higher level. The study mainly uses machine learning techniques to predict career aspects. The Python programming language is used to implement machine learning techniques. To create the prediction model, 16 features were employed. Age, health, parental position, study time, and other factors are crucial considerations. Depending on the number of features, the prediction accuracy value is altered. The four machine-level classifiers were used to predict the student's career: AdaBoost, SVM, decision tree (DT), and random forest (RF). Compared to other machine learning classifiers, the RF classifier has a higher accuracy of 93%. The prediction model's output is also applied to determine if students are interested in working or continuing their education.

The machine learning (ML) application is used in [27] to increase the accuracy of inventory-based job choice prediction. The prediction accuracy of a new machine-learning augmented method is compared to a traditional interest profile method (profile matching) in predicting occupational membership and vocational aspirations using a large sample (N = 81,267) of working (for employed participants) and jobless participants (for unemployed participants). Results indicate that the machine-learning augmented method produced greater overall accuracy for predicting both types of career choices when compared to the traditional profile method.

The studies in [1-4, 13, 14, 16-22] used no students' pre-university and university educational data, which are used in this study to correlate educational outcomes with career choices. Students' academic performance has been predicted at the degree level in [13, 16, 17, 19], and this information will be employed to forecast students' career paths, considering their pre-university, graduation, and socioeconomic or demographic information using data mining models. The distinctive features of the two HEIs set this work apart from others.

3. Predictive Data Mining Models

Predictive data mining models predict values based on known classes or labels from various data sets. The primary goal of predictive data mining models is to forecast the future using historical data. A data mining task's predictive model includes classification, regression, and prediction. It's a monitoring learning approach that entails explaining how the values of other features influence the values of a few features in the same consequence, as well as the development of a model that can predict these feature values in prior instances [28]. The following are the data mining classification and prediction techniques employed in this study:

3.1 Decision Tree

A decision tree is a more straightforward technique that separates data into nodes depending on class purity. Root nodes, branches, and leaf nodes make up a decision tree. Each internal node represents a feature being tested, each branch represents the result, and each leaf node represents the class label. The uppermost node of the tree is called the root node. Each internal node represents a test on a feature. Each leaf node defines a class. Noise or outliers generate anomalies in the training data; therefore, tree pruning is employed to remove them. Pruning has made the trees more diminutive and less complex. It may be used for both classification and prediction. Humans can better interpret decision trees. [29].

3.2 K-Nearest Neighbor

The k-nearest Neighbor (k-NN) algorithm is a distance calculation method. This data mining approach is reasonable yet highly effective. It may be used for regression as well as classification. However, classification prediction is where it is most typically employed. The spatial domain looks for the k closest training samples and averages them to provide a forecast. The k-nearest neighbor model utilizes the groups of its nearest neighbors to classify new unlabeled input. In the k-NN algorithm, unlabeled data is defined by a constant number of closest neighbors, where k is a positive integer. The accuracy and robustness of the method are determined by the value of k [30].

3.3 Logistic Regression

A classification approach that employs supervised learning to estimate the target variable probability is known as Logistic Regression. The dependent variable is a binary variable, with data represented as 1 (representing success/yes) or 0 (describing failure/no) [31]. The existence of the target or dependent variable is dichotomous, indicating that there are only two groups. As a result, a logistic regression model logically predicts P(Y=1) as a function of X.

3.4 Naïve Bayes

The Naïve Bayes model is one of the most well-known data mining techniques. This simple and rapid

probabilistic classifier is based on Bayes' theorem and features an independence assumption. Naïve Bayes trains a model from the data. As a result, the Naïve Bayes classifier is effective in various real-world scenarios. Furthermore, because only a small quantity of training data is required to estimate the classification parameters, the classifier may be trained gradually with Naïve Bayes. As a result, the approach is advantageous for classification and prediction tasks [32].

3.5 Neural Network

Neural networks are interconnected computing systems that function similarly to neurons in the human brain. A Neural Network is a back-propagating multi-layer perceptron (MLP) algorithm. The model can detect hidden patterns and correlations in raw data, cluster and classify them, and train and improve over time. It interprets and converts a data input of one kind into the desired output using a network of functions [33].

3.6 Stochastic Gradient Descent

The Stochastic Gradient Descent (SGD) model employs stochastic approximation to reduce the size of a loss function to a linear process. By evaluating one sample at a time, the technique approximates a valid gradient while concurrently updating the model based on the slope of the loss function. It returns predictors as sum minimizers, i.e., M-estimators, for regression and is especially beneficial for large-scale and heterogeneous datasets [34].

3.7 Support Vector Machine

Support Vector Machine (SVM) is a data mining method that partitions the attribute space using a hyper-plane and optimizes the margin between groups into distinct classes or class values. An SVM calculates the ideal hyper-plane to maximize the model's generalization potential [35]. As a result, the approach frequently achieves high predictive efficiency.

4. Data Preparation and Methodology

4.1 Data Description

The data of Software Engineering graduates' has been collected from two HEIs encompassing four academic cohorts or batches at Sir Syed University of Engineering and Technology (SSUET), Pakistan (private sector university), and five academic cohorts or batches at NED University of Engineering and Technology (NEDUET), Pakistan (public sector university), for this study. These academic cohorts or batches consist of 250 graduate students of NEDUET who were enrolled in the academic years 2013–14, 2014–15, 2015–16, 2016–17, and 2017–18. Likewise, 250 graduate students of SSUET who were enrolled in the academic years 2014, 2015, 2016, and 2017 have been included. It is noteworthy that both HEIs offer a four-year degree program in Software Engineering; however, the NEDUET Software Engineering degree program is recognized as an engineering discipline as the degree is accredited by the Pakistan Engineering Council (PEC), whereas the SSUET Software Engineering degree program is recognized as a non-engineering discipline as the degree is accredited by the National Computing Education Accreditation Council (NCEAC). The data contains the features related to pre-university (SSC) / (HSC) level of education, attainments in courses of studies (only core or technical courses), Cumulative Grade Point Average (CGPA), Final Year Project (FYP), and socioeconomic or demographic information. The description of features is explained in Table 1–6.

Table 1

List of SSUET Pre-university features with their description

SSUET Pre-university Features	Description		
SSC Majors	Science, O-Level		
HSC Majors	A-Level, Diploma, Diploma-CIT, Pre-Engineering, Science General		
SSC / HSC Grades	Grade	Scale or Percentage	Grade Description
	A-1	80 % or above mark	Outstanding
	A	70 % to 79 % marks	Excellent
	B	60 % to 69 % marks	Very Good
	C	50 % to 59 % marks	Good
	D	40 % to 49 % marks	Fair
	E	33 % to 39 % marks	Satisfactory

Table 2

List of NEDUET Pre-university features with their description

NEDUET Pre-university Features	Description		
HSC MPC	Maths + Physics + Chemistry marks		
HSC Grades	Grade	Scale or Percentage	Grade Description

	A-1	80 % or above mark	Outstanding
	A	70 % to 79 % marks	Excellent
	B	60 % to 69 % marks	Very Good
	C	50 % to 59 % marks	Good
	D	40 % to 49 % marks	Fair
	E	33 % to 39 % marks	Satisfactory

Table 3

List of SSUET CoS, CGPA, and FYP features with their description

SSUET Courses of studies, CGPA, and FYP Features	Description	
ITC	Introduction to Computing	
PF	Programming Fundamentals	
OOP	Object Oriented Programming	
ITSE	Introduction to Software Engineering	
DS&A	Data Structure & Algorithm	
AT&FL	Automata Theory & Formal Languages	
OS	Operating Systems	
SRE	Software Requirement Engineering	
CC&N	Computer Communication & Networks	
ITDBS	Introduction to Database Systems	
SD&A	Software Design & Architecture	
ESE	Enterprise System Engineering	
SQE	Software Quality Engineering	
SEE	Software Engineering Economics	
HCI	Human Computer Interaction	
SPM	Software Project Management	
DWH&DM	Data Ware House & Data Mining	
WE	Web Engineering	
DS&E	Data Security & Encryption	
AI	Artificial Intelligence	
PP	Professional Practice	
FYP	Final Year Project	
FYP Domain	Big Data, Cloud Computing, IoT, Mobile / Android / IOS, Data Mining, Machine Learning, Deep Learning, Networking, Network Security / Cyber Security, Information Security, Image Processing, Web Application, Other	
CGPA	Grade Point	% Marks
	4.00	90-100
	3.7-3.9	85-89
	3.4-3.6	80-84
	3.0-3.3	70-79
	2.5-2.9	60-69
	2.0-2.4	50-59
	0.00	0-49

Table 4

List of NEDUET CoS, CGPA, and FYP features with their description

NEDUET Courses of studies, CGPA, and FYP Features	Description
DS&A	Data Structure & Algorithm
FIT	Fundamentals of Information Technology
PL	Programming Languages
CG	Computer Graphics
DBMS	Database Management Systems
OOC&P	Object Oriented Concepts & Programming
SE	Software Engineering
SRE	Software Requirement Engineering
WE	Web Engineering
AI&ES	Artificial Intelligence & Expert Systems
CCN	Computer Communication Networks
EC	E-Commerce
OS	Operating Systems

SD&A	Software Design & Architecture	
HCI	Human Computer Interaction	
SPM	Software Project Management	
SQE	Software Quality Engineering	
DW&M	Data Warehouse Methods	
N&IS	Network & Information Security	
EP	Entrepreneurship	
EL	Elective Course {Software Testing Strategies & Techniques OR Information Systems Engineering}	
SEP (FYP)	Software Engineering Project	
FYP Domain	Big Data, Cloud Computing, IoT, Mobile / Android / IOS, Data Mining, Machine Learning, Deep Learning, Networking, Network Security / Cyber Security, Information Security, Image Processing, Web Application, Other	
CGPA	Grade Point	% Marks
	4.0	94-100 / 85-93
	3.7	80-84
	3.4	75-79
	3.0	70-74
	2.7	67-69
	2.4	64-66
	2.0	60-63
	1.7	57-59
	1.4	54-56
	1.0	50-53
	0.0	Below 50

Table 5

List of Socioeconomic or Demographic features with their description

Socioeconomic or Demographic Features	Description
Gender	Male, Female
Parent's / Guardian's Qualification	Some Education, Matric (SSC) Pass, Inter (HSC) Pass, Bachelor's Degree, Master's Degree, Doctoral Degree, Diploma / Certificate, Other
Parent's / Guardian's Income	Very High, High, Medium, Low, Very Low
Parent's / Guardian's Occupation	Banking, Government, Private, Construction, General Business / Trade, Education / Teaching, Engineering, Sales, Management, I.T, Medical, Other
Internship Experience	Yes, No
Skills / Competencies	Programming Languages, Algorithms and Complexity, Software Engineering, Database Technologies, Networking and Communications, Digital Marketing, Project Management, Graphics and Visual Computing, Intelligent Systems, Software Development, Technical Writing, Open-Source Technologies, Other

Table 6

List of Target or Class features with their description

Target / Class Feature:	Description / Job Roles
A	{Software Developer, Software Engineer, Programmer}
B	{Web Developer, PHP Developer, Front End Developer, Web Administrator}
C	{IOS Developer, Android Developer}
D	{Graphic Designer, Web Designer}
E	{SQA Engineer, SQA Analyst}
F	{IT Director, IT Manager, MIS Officer, Technical Operations Officer}
G	{Database Administrator, Systems Administrator, Technical Support Engineer / Specialist}
H	{Network Administrator, Network Engineer}
O	{Other}

The data on pre-university, CoS, CGPA, FYP, and gender were extracted from the separate databases of two universities by creating a consolidated data warehouse. The data of the remaining features, such as socioeconomic or demographic features and FYP domain, are collected through emailing and sharing different batch-wise links to graduates using GoogleForm via questionnaires on multiple social media platforms, graduates' profiles, and graduates groups or forums from January 2022 to October 2022.

All sets of features are categorical in Tables 1, 5, and 6. In Table 2, the feature HSC MPC, the sum of the marks attained in mathematics, physics, and chemistry at the HSC level, is numeric, while the feature HSC Grades is categorical. On the contrary, in Tables 3 and 4, all sets of features are numeric except the feature FYP domain, which is categorical.

4.2 Data Pre-processing

Models, algorithms, and statistical inferences are the critical focus of any data analysis. Nevertheless, modeling is typically not performed with raw or dirty data in practical applications. Data pre-processing transforms dirty data (noisy or raw data) into clean data suitable for modeling. Data pre-processing can profoundly impact model outcomes, including removing outliers and imputing missing values [36]. Pre-processing data is, therefore, a crucial step. There might be numerous issues with the data, depending on the circumstances. Before modeling, the data must be cleaned. Additionally, the data needs for various models vary. For instance, some models would need constant scale variables; others might be susceptible to outliers or collinearity; others might be unable to handle categorical variables, and so on. To make the data suitable for the particular model, the data must be pre-processed appropriately [37].

In this research, the data is pre-processed, statistically analyzed, and mined using ORANGE. ORANGE is the ideal software tool used for data mining and machine learning. It is a software tool with components created in Python to support interactive data visualizations. We organized the collected data into two datasets: Dataset I and Dataset II. In Dataset I, we have incorporated the features of SSUET graduates, while in Dataset II, we have incorporated the features of NEDUET graduates. Both datasets are separately designed and prepared by integrating HEIs academic data (retrieved from the HEIs database) and socioeconomic or demographic data (retrieved through an online survey). Data pre-processing uses an Orange tool to normalize features and impute missing data. Removing rows with missing values is employed under the pre-processor for imputing missing values. As both datasets are prepared carefully, no such missing or erroneous values have been found at the pre-processing stage by the ORANGE tool [38].

4.3 Methodology

In this research, we want to investigate the three aspects that influence the career path by considering pre-university features to identify correct decision-making on technology selection, hence opting for the right career path. CoS (core courses), CGPA, and FYP feature to determine any graduate interestingness through their academic achievement that will lead to the right career path. Socioeconomic or demographic features that impact the career decision-making process.

A collection of features describes a data object. A training dataset comprises data objects with a predefined label or class. Under data mining models, classification models (or classifiers) predict the class or label of a data object. A classifier creates a model that most accurately depicts the relationship between the properties of the training dataset and the class labels using a learning method. The class or label of the testing data should be appropriately predicted by the model based on training data. The number of test records that a classification model correctly and incorrectly predicts is a standard measure of a classification model's performance [39].

There are several types of data mining models or classifiers, and none is known to outperform the others consistently. Consequently, assessing if one model or classifier performs better than the others in a specific domain is necessary. The Decision Tree, k-nearest Neighbor (k-NN), Logistic Regression, Naïve Bayes, Neural Network, Stochastic Gradient Descent, and Support Vector Machine data mining predictive models used in this research have produced significant results.

As previously stated, Orange is used to pre-process and analyze the data with the defined features. Since models must be evaluated after training, the data is divided into training and testing parts. Then, using a set percentage of data, each batch or cohort's data belonging to both HEIs (250 occurrences each and separately) is trained using a split of 70% training and 30% testing with stratified sampling [38]. The primary benefit of stratified random sampling is that it accurately represents important population features in the sample. The entire training and testing process have been repeated ten times or cycles to achieve maximum accuracy. The models are trained or labeled in classes, as mentioned in Table 6, according to the graduates' current job roles.

5. Results and Analysis

This section summarizes the findings of using data mining models as described in Section 3, which focused on analyzing the classifiers' performance

metrics and comparing and contrasting their performance.

5.1 Performance or Evaluation Metrics of Data Mining Models

The data mining classification models aim to predict the category or class to which one or more observations belong. Evaluating the model's performance is crucial in any data mining workflow. Here, the predictions can be made on previously unobserved, labeled data using the trained model. Then, we evaluate how many of these predictions the model correctly identified for classification. A classification model's performance is estimated using various methods on average over classes on both datasets, some of which are listed here. Simply dividing the number of predictions by the number of correct predictions gives us a model's overall accuracy [40]. An accuracy score will range from 0 to 1, with 1 being the ideal model. When data is skewed, and one class is significantly bigger than another, this metric should seldom be used alone since the accuracy might need to be corrected.

The model's performance at all potential classification thresholds may be gauged using the AUC, a metric of the complete two-dimensional area under the curve [40]. The model's precision is its ability to identify the positive class with correctness. We would reduce the number of false positives by using this statistic to optimize a model [40]. Recall measures how well the model predicts each dataset's positive observations. Typically, a precision-recall curve is constructed to examine precision and recall simultaneously. This can make the trade-offs between the two metrics at various thresholds simpler to explore [40].

The harmonic mean of recall and precision is the F1 score. The F1 score will provide a number between 0 and 1. Perfect recall and precision are indicated by an F1 score of 1.0. If the precision or recall are both 0, then the F1 score is 0 [40]. When a classification model predicts a probability between 0 and 1, the model's performance is measured by logarithmic loss (also known as log loss). As predicted probability and actual label diverge, log loss tends to rise. Specificity may be calculated and contrasted with recall using the number of false negatives produced by the data mining models. The data mining models or classifiers' performance or evaluation metrics obtained from the analysis of both datasets are enlisted in Table A1.

If a sample from one class is more abundant than another, the data set is said to be highly skewed. In the imbalanced data set, the class with the maximum number of instances is referred to as the major class, while the class with the least number of instances is referred to as the minor class [41]. In such a scenario, most classifiers exhibit extreme bias toward the major classes and have very low classification rates for minor classes. The classifier also likely classifies all classes as major classes while ignoring minor classes [42].

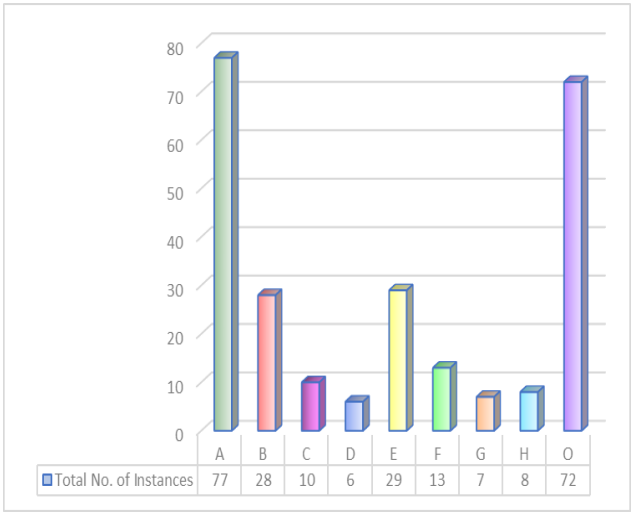

Fig. 2. Information On a Target or Class Feature In Dataset I

Fig. 2 and Fig. 3 show the imbalanced number of instances belonging to each target or class feature of Dataset I and Dataset II, respectively. In this case, Class ‘A’ emerged as the major class, containing the maximum instances comprising job roles {Software Developer, Software Engineer, Programmer}. According to both figures, most models or classifiers performed better when predicting the class ‘A’ feature, our leading class. It is also important to note that in dataset II (see Fig. 3), the instances are not predicted in a “D” class.

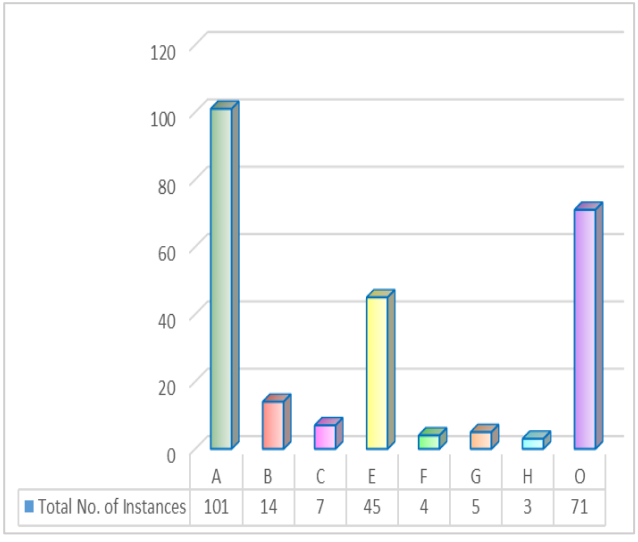

Fig. 3. Information On A Target Or Class Feature In Dataset II

5.2 Applying SMOTE for Imbalanced Classification

In this study, we have fortunately balanced datasets (each dataset with 250 instances), but each dataset's classes are imbalanced. The classes must be balanced, as seen in Figures 2 and 3. All samples from the minority classes (C, D, F, G, and H) in dataset I and (B, C, F, G, and H) in dataset II were deduced in order to balance the classes, and duplicates of each sample were made several times across the dataset by applying SMOTE technique. The Synthetic Minority Oversampling Technique is known as SMOTE. SMOTE is applicable when the data is imbalanced. SMOTE uses a k-nearest neighbor method to generate synthetic data. SMOTE begins by randomly selecting data from the minority class, after which the data's k-nearest neighbors are determined. The k-nearest neighbor was picked randomly, and the random data would then be combined to create synthetic data [43].

SMOTE has augmented up to 693 instances in dataset I and 808 instances in dataset II (collectively 1501 instances), respectively. Regenerated data mining prediction models were compared to the original models in terms of accuracy using the balanced datasets after applying SMOTE. With rebalanced classes in both datasets, the prediction accuracy of each of these models radically increased. The improved data mining models or classifiers' performance or evaluation metrics obtained from the analysis of both datasets after applying SMOTE are enlisted in Table A2.

5.3 Comparing Data Mining Models Accuracy

The results of seven data mining models' comparable accuracies on both datasets are depicted below in Figures 4 and 5. The accuracy results of data mining models that performed better than the benchmark is identified. The results imply that it may be possible to predict the career path in advance with comparable accuracies using the information of pre-university (SSC/HSC), attainments in the courses of studies, CGPA, FYP, and socioeconomic or demographic data. Fig. 4 compares accuracies achieved by the data mining models on each dataset using actual instances, i.e., 250 instances of each HEI data with imbalanced classes. The accuracies are computed as average over nine classes and considered reasonable baseline accuracies. The k-Nearest Neighbor model achieves the highest accuracy on dataset I, i.e., 85.71%, compared to the other six models or classifiers. Similarly, the SVM model ranked high with 86.73% accuracy when analyzing the performance of models or classifiers on dataset II.

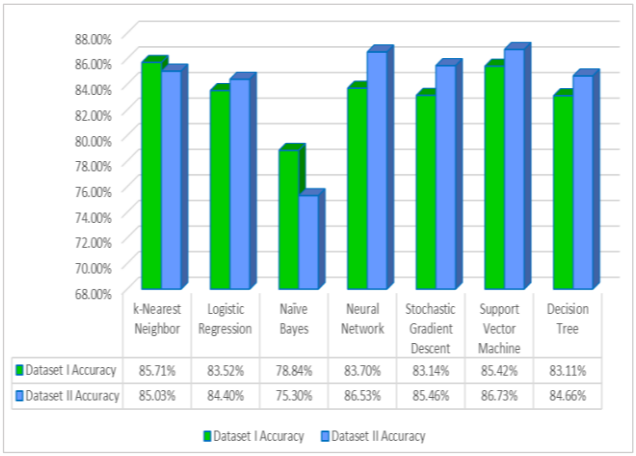

Fig. 4. Comparison Of Models' Accuracy

Accuracy is an effective metric when the target feature classes in the data are reasonably balanced. Fig. 5 depicts the comparison of accuracies achieved by the data mining models after applying the SMOTE technique on each dataset with balanced classes. Each model's prediction accuracy considerably improved with rebalanced classes in both datasets. This time, with a 93.96% accuracy on dataset I, the Neural Network model outperformed the other classifiers or models. Similarly, the Neural Network model performed best with 95.73% accuracy on dataset II.

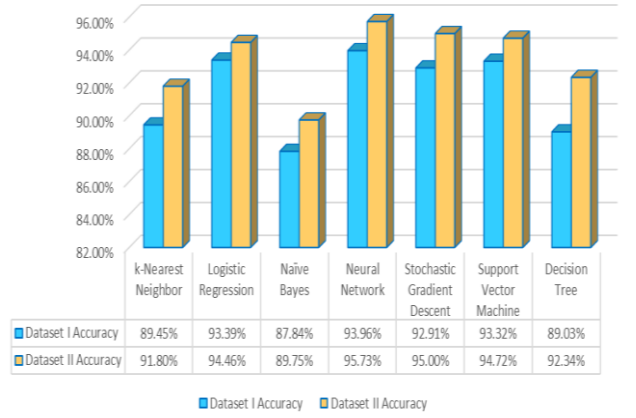

Fig. 5. Comparison Of Models' Accuracy After Applying SMOTE

5.4 Features Selection

Feature selection is the process of minimizing the number of input features. It is essential to eliminate redundant or unnecessary features since they might detriment the model's performance. Limiting the number of input features is preferable to decrease the computational complexity of modeling and, in certain situations, increase the model's performance. Feature selection's primary benefit is that it lessens overfitting [44].

This study used the gain ratio and chi-square (χ^2) feature selection approaches to determine essential features. A ratio between the feature's inherent information and the information gain, which lessens

the information gain's bias towards multivalued features, is referred to as the gain ratio [45]. The chi-square statistic measures the dependency between the feature and the class [46]. The number of selected features has been set at K=15 for each criterion. The best 15 features extracted by feature selection techniques on both datasets are presented in Tables 7 and 8 with their scores.

Table 7

List of features using feature selection techniques in Dataset I

Features	Gain ratio	χ^2
HSC Major	0.190638	0.115883
SEE	0.187124	189.2424
CGPA	0.184446	197.8444
SPM	0.154077	149.7061
CCN	0.150082	139.1507
AI	0.145689	127.072
ITC	0.145099	160.5935
Gender	0.138624	14.27038
ATFL	0.130716	117.2965
PP	0.125279	132.5248
SRE	0.123623	114.9164
FYP	0.123619	114.6338
DWH&DM	0.121293	111.804
Internship Experience	0.115128	18.72466
ESE	0.113734	132.7173

We have set three principal objectives to predict career paths, as mentioned in Section 1. By observing Table 9 based on dataset I features, it is interesting to note that out of four pre-university features, and just a feature, “HSC Major” is extracted to fulfill the first objective, i.e., relating HSC Major influencing the career choices to correct decision-making on technology selection in advance. As far as the second objective is concerned, encompassing core courses of studies, CGPA and FYP, courses “SEE, SPM, CCN, AI, ITC, ATFL, PP, SRE, DWH and DM, and ESE,” and surprisingly, both “CGPA” and “FYP” features have been selected to predict career path in advance. If the third objective is concerned, features “Gender” and “Internship Experience” have been chosen among the other socioeconomic or demographic features to predict career path in advance.

Table 8

List of features using feature selection techniques in Dataset II

Features	Gain ratio	χ^2
MPC Marks	0.294292664	92.12
AI&ES	0.255306706	300.8580858
SE	0.23108595	130.455144
Internship Experience	0.228949806	16.56062909
FIT	0.204856363	210.9570957
SEP	0.204179107	205.8067633
Gender	0.189609959	121.7973381
OOC&P	0.189505504	123.8156797
PL	0.178888808	132.397351
EP	0.173029165	188.8342811
CGPA	0.169264313	113.6369637
DW&M	0.16223072	121.7585644
SPM	0.154529447	143.3562552
SQE	0.15441058	107.5115512
WE	0.148324247	145.1012448

Similarly, by looking at Table 9 based on features from dataset II, it is noteworthy to see that here also, only one pre-university variable, i.e., "MPC Marks," is extracted to accomplish the first objective, which is to relate how the MPC Marks influence career choices to the right choice of technology in advance. For the second objective, which includes core courses of studies, CGPA, and FYP, the courses "AI and ES, SE, FIT, OOC and P, PL, EP, DW and M, SPM, SQE, and WE" have been selected. Surprisingly, both "CGPA" and "SEP" features have also been chosen to predict career paths in advance. Here, the features "Gender" and "Internship Experience" have been selected among the other socioeconomic or demographic features to predict career paths in advance for the third objective.

Furthermore, the analysis of the tables above reveals a significant finding: the presence of common features in both datasets. These features, such as the set of similar courses of studies taught at both HEIs {AI/AI and ES}, {ITC/FIT}, {DWH and DM/DW and M}, {SPM/SPM}, and CGPA, and FYP/SEP, and Gender and Internship Experience, have been identified as crucial using the gain ratio and chi-square (χ^2) features selection approaches.

6. Discussion

As previously mentioned, graduates have a variety of job roles to choose from within each career, depending on their academic strengths. In this context, we delve into the practical implications of the methodology described above, which can significantly impact career decisions.

1) It uses data mining models to determine technology selection that influences career choices based on pre-university features.
2) They are predicting career paths in advance based on features such as attainments in the core courses of studies, CGPA, and FYP using data mining models with comparable accuracies.
3) They are predicting career paths in advance based on socioeconomic or demographic features using data mining models with comparable accuracies.

Below, we present the outcomes of the two HEI datasets in the context of decision tree visualizations. Decision trees are particularly powerful in our analysis due to their ease of comprehension and interpretation. This is crucial in our case, as we need to explain how and why a specific result occurred [47].

6.1 Determining Technology Selection using Data Mining Models

The decision tree (on dataset I) in Fig. 6 below shows the tree is built around the feature HSC Grade with the class or target ‘A’ as the root splitting criterion. One notices that class ‘A’ is predicted when the SSC Grades are A and B, class ‘B’ is predicted when the HSC Grade is A and class ‘C’ is predicted when the SSC Grades are A, C, and D and HSC Grades are B and C respectively, class ‘D’ is predicted when HSC Grades are A-1, B, and C, class ‘F’ is predicted when HSC Grade is B, and class ‘H’ is predicted when HSC Grade is A-1, and so on.

Interestingly, here, the decision tree from the pre-university features has not predicted the features of SSC Majors and HSC Majors in dataset I. This means neither feature plays a vital role in determining technology selection at an early level of education. However, the pre-university majors are equally important in selecting a particular technology, like software engineering, computer engineering, civil engineering, electronic engineering, etc., for enrolment at any HEI.

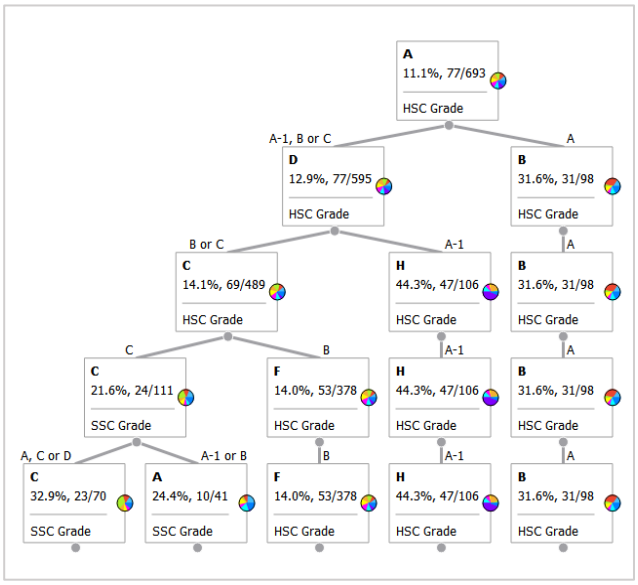

Fig. 6. Decision Tree Representation With Pre-University Features In Dataset I

Similarly, the decision tree (on dataset II) in Fig. 7 below shows the tree is built around the feature MPC Marks with the class or target ‘A’ as the root splitting criterion. One notices that different MPC attainments are predicted in all classes except classes ‘D’ and ‘E.’ Moreover, classes ‘A,’ ‘C,’ and ‘H’ are predicted when the HSC Grade is A-1, classes ‘C’ and ‘H’ are predicted when the HSC Grade is A, and classes ‘A’ and ‘C’ are predicted when the HSC Grades are B and C, and so on. Besides, one may be surprised why the decision tree has not correspondingly predicted the job roles ‘E,’ ‘F,’ and ‘O’ in Dataset I and job roles ‘D’ and ‘E’ in Dataset II.

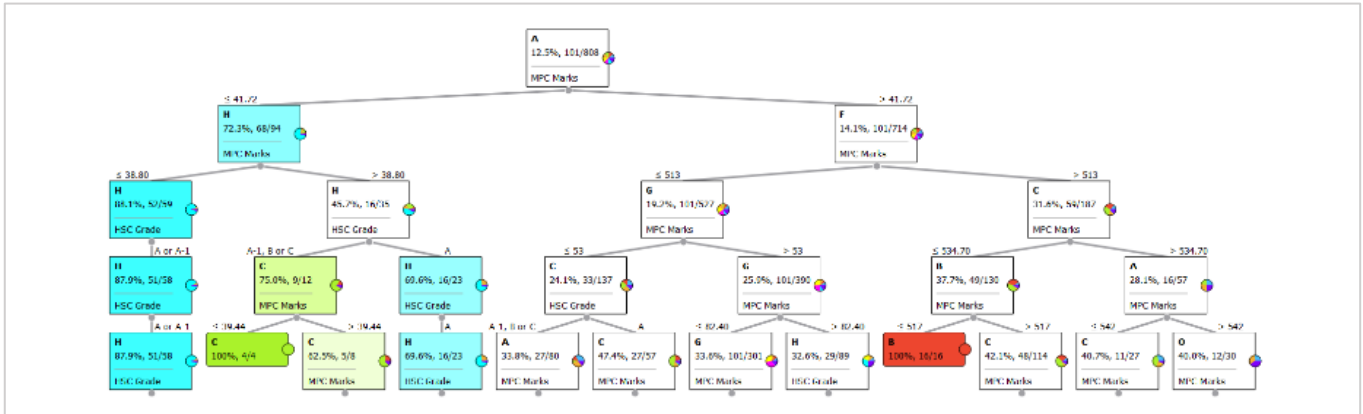

Fig. 7. Decision Tree representation with pre-university features in Dataset II

Thus, Research Question 1: To what extent does the pre-university (SSC/HSC) level of education influence career choices to correct decision-making on technology selection using data mining models? It is answered optimistically. So far, so good; the exciting observations have been brought to light that may guide the forthcoming students to make correct decision-making concerning technology selection (i.e., software engineering technology) using data mining models. Integrating pre-university features in this research is pivotal since early school or college-level education is where students' career path commences and may be helpful in correct technology selection at the university level.

6.2 Predicting Career Path using Data Mining Models based on CoS, CGPA, and FYP

Fig. 8 below presents a decision tree (on dataset I) constructed around the feature CGPA and uses the class or target 'A' as the root-splitting criterion. This decision tree is a vital tool in our research, as it helps us understand how quality points attained in core courses taught in the Software Engineering discipline can influence career paths. It is important to note that core courses of study are mandatory, as well as technology-related courses one must study to meet the requirements of a degree program. Quality Points (QP) refer to the product of a Grade Point (GP) and Credit Hours (CH) in each course. The decision tree predicts that the class 'H' is likely when the CGPA is less than or equal to 2.65, while the class 'B' is likely when the CGPA is more than 2.65.

Similarly, the decision tree predicts that class 'C' is likely when the quality points are greater than 9.38, 11.5, and 11.98 in the PP, ITSE, and ITC courses, respectively. Class 'D' is likely when the quality points are greater than 6.28 in the course SEE and less than or equal to 10.0 and 9.38 in AI and PP courses, respectively. Class 'E' is likely when the quality points are greater than 9.24 in the course ITDBS and less than or equal to 10.8 and 11.98 in the course ITC.

Class 'F' is predicted when the quality points are greater than 7.2, 10.0, 10.8, and 11.69 in the courses SRE, AI, ITC, and WE and less than and equal to 6.35 and 11.95 in the courses AT and FL and ITSE respectively. Class 'G' is predicted when the quality points are greater than 6.35 in the course AT and FL and less than and equal to 11.69 and 9.24 in the courses WE and ITDBS, respectively. Class 'H' is predicted when the quality points are greater than 10.0 in the course ITSE and less than and equal to 6.28 in the course SEE, respectively. Class 'O' is predicted when the quality points are less than and equal to 10.0 and 7.2 in the courses ITSE and SEE, respectively.

Another essential aspect of predicting a career path is the FYP information. One can notice that the C' and 'G' are predicted when the quality points are greater than, less than, and equal to 22.18 in FYP attainment, respectively. Similarly, class 'B' is predicted when the FYP domains are Big Data, Cloud Computing, IoT, Mobile / Android / IOS, Machine Learning, Deep Learning, Networking, Network Security / Cyber Security, Image Processing, Web Application, and Other. At the same time, class 'C' is predicted when the FYP domains are Data Mining, Information Security, and so on.

Surprisingly, the decision tree has not predicted the job roles 'A' and 'B,' which are believed to be the top careers in software engineering.

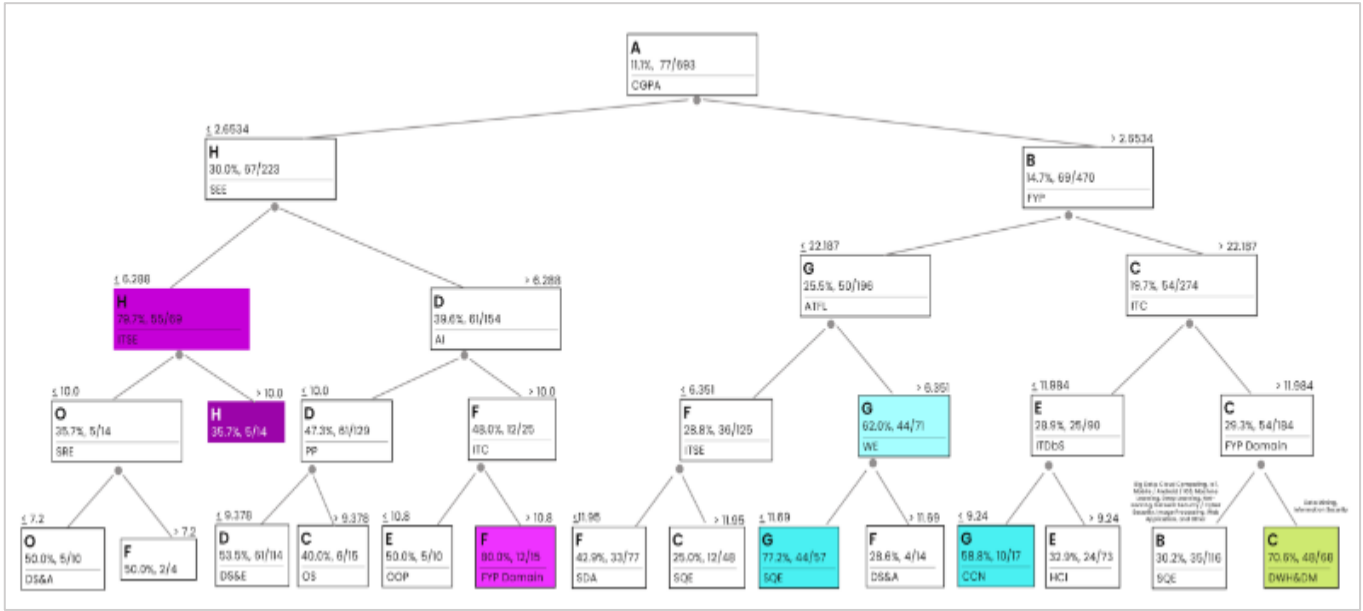

Fig. 8. Decision Tree Representation With Features Cos, CGPA and FYP In Dataset I

Fig. 9 below depicts the decision tree (on dataset II), built around the feature AI and ES, and uses the class or target ‘A’ as the root-splitting criterion. In contrast to dataset I, we used the total marks (obtained out of 100) in core courses taught in the Software Engineering discipline. Class ‘A’ is predicted when the total marks are greater than 80.93 and 81.64 in the courses SE and AI and ES, respectively. Class ‘B’ is predicted when the total marks are greater than 77 and 80 in SE and OS courses and less than and equal to 81.33 and 81.64 in OOC and P and AI and ES, respectively.

Class ‘C’ is predicted when the total marks are greater than 81.33 and 83.37 in the courses OOC and P and FIT and less than and equal to 89.01 and 74.98 in FIT and EP, respectively. Class ‘E’ is predicted when the total marks are greater than 64 in the course SE and less than and equal to 71 in the course HCI, respectively. Class ‘F’ is predicted when the total marks are greater than 78 and 71 in the EP and HCI courses, respectively.

Class ‘G’ is predicted when the total marks are greater than 73.99 and 89.01 in the courses AI and ES and FIT and less than and equal to 80 and 80.93 in the courses OS and SE, respectively. Class ‘H’ is predicted when the total marks are greater than 76.65 in the course DS and A and less than and equal to 73.99 and 78 in the courses AI and ES and EP, respectively. Class ‘O’ is predicted when the total marks are greater than 74.98 in the course EP and less than and equal to 76.65, 64 and 77, and 83.37 in the DS and A, SE, and FIT courses, respectively, and so on.

Regarding the software engineering final year project, the classes ‘B’ and ‘H’ are predicted when the total marks are greater than, less than, and equal to 85.95 in SEP attainment. Here, by observing the decision tree results, the essential features, i.e., CGPA and SEP Domain, have not been predicted in dataset II.

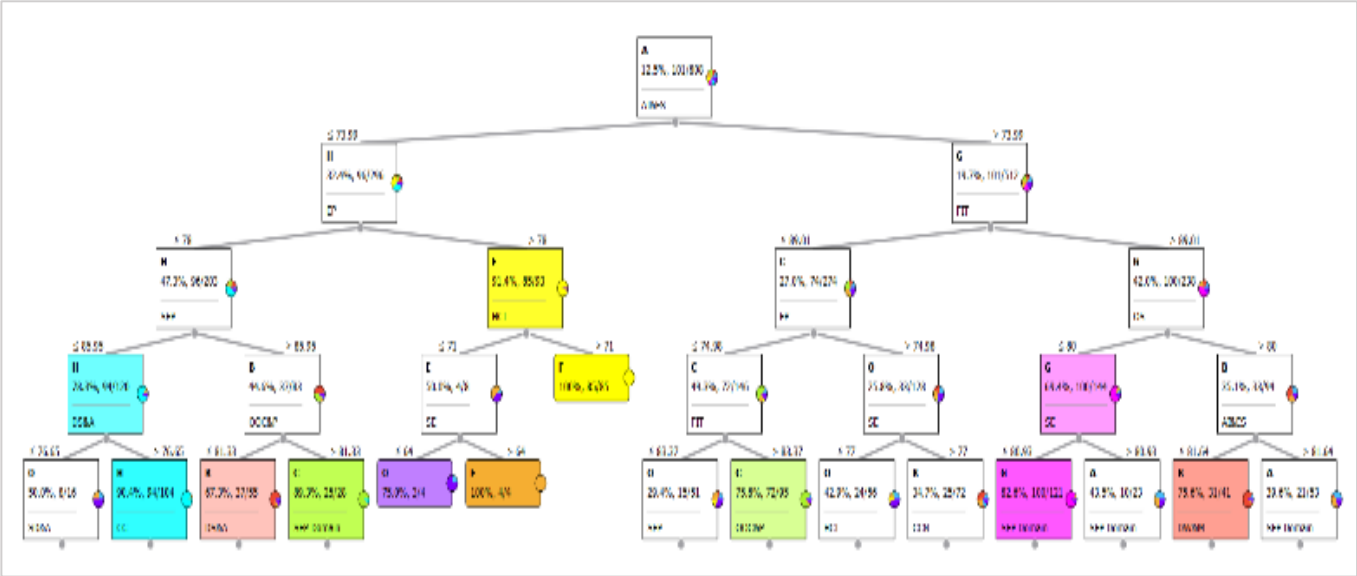

Fig. 9. Decision Tree Representation With Features Cos, CGPA and FYP In Dataset II

Therefore, Research Question 2: Is it possible to predict a career path in advance using data mining models based on the attainments in the courses of studies, CGPA, and FYP with comparable accuracies? It is also answered optimistically. In order to support students in pondering their career opportunities in the future, the second question deals with identifying the CoS, CGPA, and FYP attainments that have emerged as valuable predictors with the help of data mining models in this study that may lead to successful career paths. Hence, to know about a better career path, considering CGPA and the performance in the courses of studies at the university level may be essential to determine how much the students have shown their interest in particular courses and have grabbed the technical skills and practical knowledge in order to succeed in a career.

Also, incorporating FYP information in this study was necessary because FYP helps students develop their fundamental abilities and prepares them for new challenges. A creative and valuable final-year project will reinforce students' ability to solve problems, manage projects, conduct research, and analyze data. An engineering student's final year project marks a significant turning point in their life. It assists in bridging the knowledge gap between theory-based and skill-based learning in careers.

6.3 Predicting Career Path Using Data Mining Models Based on Demography Or Socioeconomics

Fig. 10 below depicts the decision tree (on dataset I), built around the feature Skills/Competencies, and uses the class or target ‘A’ as the root splitting criterion. One can observe that the class ‘A’ is predicted when the Skills/Competencies are Programming Languages, Algorithms and Complexity, Software Engineering, Database Technologies, Digital Marketing, Project Management, Graphics and Visual Computing, Intelligent Systems, Software Development, Technical Writing, Open-Source Technologies, or Other, and the Gender is Male respectively.

Class ‘B’ is predicted when the Parent’s / Guardian’s Occupations are Banking, Private, Construction, General Business / Trade, Education / Teaching, Engineering, Sales, Management, I.T, or Other, and the Skills/Competencies are Programming Languages, Algorithms and Complexity, Software Engineering, Database Technologies, Networking and Communications, Digital Marketing, Intelligent Systems, Software Development, Technical Writing, or Open-Source Technologies respectively. Class ‘D’ is predicted when the Parent’s / Guardian’s Occupation is Government and the Parent’s / Guardian’s Qualifications are Diploma / Certificate, Master’s Degree, or Some Education, respectively.

Class ‘E’ is predicted when the Skills/Competencies are Graphics and Visual Computing, Other, or Project Management. Class ‘G’ is predicted when the Internship Experience is Yes, the Parent’s / Guardian’s Occupations are General Business / Trade, Medical, Other, Private, or Sales, and the Parent’s / Guardian’s Qualifications are Bachelor’s Degree, Doctoral Degree, Inter (HSC) Pass, or Matric (SSC) Pass respectively. Class ‘H’ is predicted when the Skills/Competencies are Networking and Communications. Class ‘O’ is predicted when the Gender is Female, the Internship Experience is No, the Parent’s / Guardian’s Qualifications are all the inputs, and the Parent’s / Guardian’s Occupation is Education / Teaching, Government, I.T, or Management, respectively, and so on.

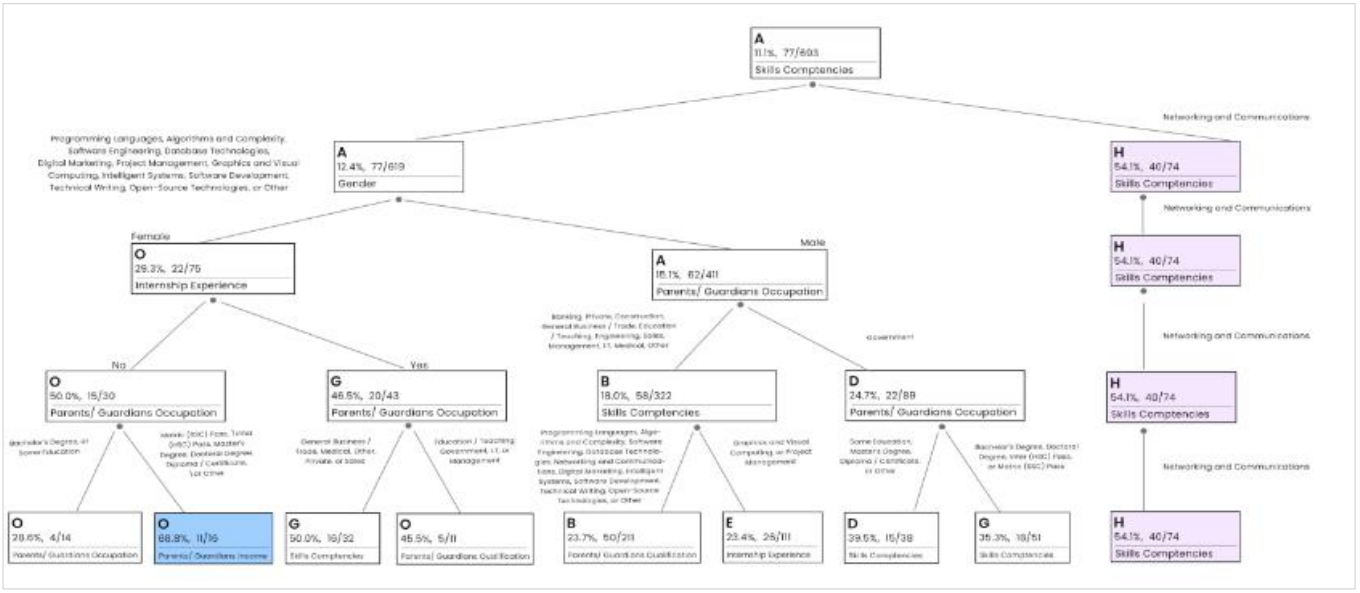

Fig. 10. Decision Tree Representation With Demographic Or Socioeconomic Features In Dataset I

Fig. 11 below depicts the decision tree (on dataset II), built around the feature Internship Experience, and uses the class or target ‘A’ as the root splitting criterion. One can observe that the class ‘A’ is predicted when the Internship Experience is No, the Skills / Competencies are Software Development, Software Engineering, Technical Writing, Database Technologies, or Software Engineering, the Parent’s / Guardian’s Occupations are Banking, Government, Private, Construction, Education / Teaching, Sales, Management, or I.T, the Gender is Male, and the Parent’s / Guardian’s Qualifications are Bachelor’s Degree, Inter (HSC) Pass, or Master’s Degree respectively.

Class ‘B’ is predicted when the Parent’s / Guardian’s Occupations are Engineering, General Business / Trade, Medical, Other, or Government. Class ‘E’ is predicted when the Skills / Competencies are Programming Languages, Algorithms and Complexity, Database Technologies, Networking and Communications, Digital Marketing, Project Management, Graphics and Visual Computing,

Intelligent Systems, Open-Source Technologies, or Other, and the Parent's / Guardian's Occupations are Banking, Government, Private, Construction, General Business / Trade, Education / Teaching, Engineering, Sales, Management, I.T, or Medical respectively.

Class 'F' is predicted when the Internship Experience is Yes, the Skills / Competencies are Programming Languages, Algorithms and Complexity, Networking and Communications, Digital Marketing, Project Management, Graphics and Visual Computing, Intelligent Systems, Software Development, Technical Writing, Open-Source Technologies, or Other, and the Parent's / Guardian's Occupations are General Business / Trade, Government, Medical, or Private respectively.

Class 'G' is predicted when the Gender is Female, the Skills / Competencies are Database Technologies, and the Parent's / Guardian's Qualifications are Some Education, Matric (SSC) Pass, Degree, Doctoral Degree, Diploma / Certificate, or Other, respectively. Class 'H' is predicted when the Parent's / Guardian's Occupations are Banking, Construction, Education / Teaching, Engineering, Sales, Management, I.T, or Other. Class 'O' is predicted when the Parent's / Guardian's Occupations are Medical or Other, and so on.

Nevertheless, the decision tree has not predicted the job roles 'C' and 'F' in dataset I and job role 'C' in dataset II, respectively.

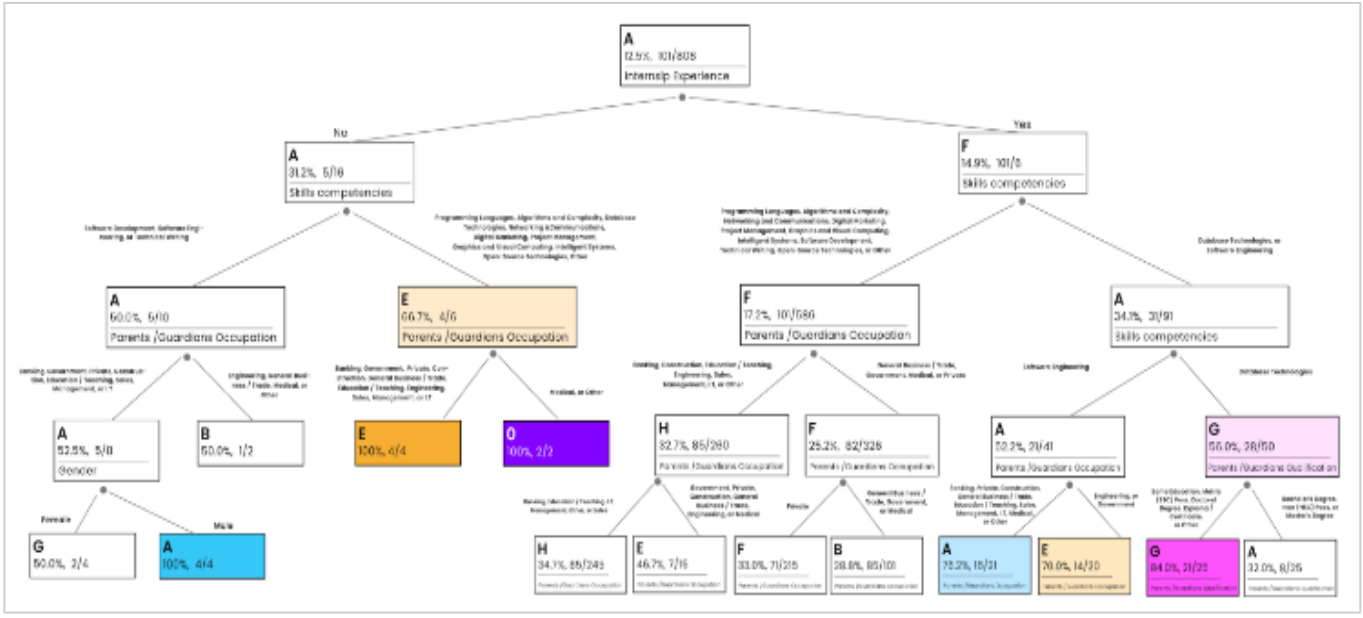

Fig. 11. Decision Tree Representation With Demographic Or Socioeconomic Features In Dataset II

According to prior studies, family demography, gender, skills/competencies, and internship experience significantly impact the dissemination of values, including appropriate career choices, ambition, and career orientation [48]. Thus, Research Question 3: Is it possible to predict a career path in advance using data mining models based on socioeconomic or demographic information with comparable accuracies? It is also answered optimistically. Here, the decision tree has not predicted both datasets' feature Parent's / Guardian's Income.

The quality of interpersonal support and the perception of societal standards, both of which may come from the family background, are related to the quality of interpersonal interactions and the development of career aspirations. Early aspirations influence subsequent career and academic success [49]. Furthermore, the analysis shows that whether or not the students have had an internship experience, the data mining models' findings with comparable accuracies determine that they are likely to have good chances of placement in different careers or job roles according to their specific skills and competencies equally for both males and females [50].

As far as identifying the coherent relationships between the similarities and diversities among the parameters of the features of HEIs datasets is concerned, we have incorporated both SSC and HSC data in SSUET's dataset I. In contrast, if we analyze dataset II, it is worth mentioning that the NEDUET merely focuses on HSC criteria at the time of admission and enrolment, so just HSC data is incorporated. Similarly, from the requirements of relating CoS attainments for predicting career paths, the course information used in both datasets to some

extent are diversified with each other; additionally, for dataset I, the quality points attainments have been used, whereas, on the contrary, the total marks (out of 100) attainments have used in dataset II. Also, it is noteworthy that the CGPA criteria are different in both HEI datasets. Despite this, the features encompassing the FYP Domain and graduates' demographic or socioeconomic information acquired through online surveys are similar in both HEIs datasets.

7. Conclusion

The right career decision is crucially vital for one's success. In this regard, we have determined the correct technology selection decision after the pre-university attempt and predict career paths in advance based on the attainments in the CoS, CGPA, and FYP, as well as integrating demographic or socioeconomics information using data mining models. The present study will be expedient to higher education institutions (HEIs), prospective students who will directly contribute to the teaching and learning process, parents, academicians, degree planners, and practitioners involved in administrative and decision-making processes. The present study has investigated three research questions to provide prospective students with guidance to help them determine how likely they will be to have career placement opportunities in different job roles or careers in the software engineering field based on a specific career path.

Initially, the classes were imbalanced in this exploration. The problem of class imbalance often arises when some classes are more prevalent than others. Standard models or classifiers frequently disregard the minor classes in these situations because they are too overwhelmed with the large classes. So, in order to deal with the class imbalance issue, we have applied the SMOTE technique to balance the classes in both datasets. To have a fair deal, after employing SMOTE, balanced datasets were used to evaluate the accuracy of the regenerated data mining prediction models that were compared to the original models. These models' prediction accuracy drastically improved with rebalanced classes in both datasets.

When comparing the performance of models or classifiers on imbalanced classes, the k-nearest Neighbor model obtained the highest accuracy on dataset I. In contrast, the SVM model ranked the highest accuracy on dataset II. On the contrary, compared to other classifiers or models, the Neural Network model achieved the highest accuracy on both datasets after applying SMOTE. The results and analyses in this study answered all three questions optimistically. Later, the essential features were selected in both datasets using the gain ratio and chi-square (χ^2) feature selection approaches. The features with the Decision Tree model that emerged as useful predictors in both datasets used in this study are SSC/HSC grades, MPC Marks, attainments in the CoS, CGPA, FYP, Parent's / Guardian's Qualification, Occupation, Gender, Internship Experience, and Skills / Competencies. For each criterion, K=15 features have been selected. The primary reason for harnessing academic data like pre-university grades, majors, CoS attainments, CGPA, FYP, etc., as well as demographic or socioeconomics data, is that it has a tangible impact on future educational and career mobility. It may also be seen as a sign of academic potential attained.

In a nutshell, different career paths that lead to specific job roles as a class or target feature have been predicted, indicating prospective students who are likely to place or opt for different careers according to their academic strength and demography or socioeconomic values, as employed in this research. Thus, this research's findings may assist in successfully improving training strategies to anticipate career directions, improving the overall learning system, and organizing a curriculum that follows the job market and environment for the graduates.

8. Acknowledgement

The authors thank the administrations of SSUET and NEDUET for providing the academic data of software engineering graduates.

9. References

[1] M. A. Alimam, H. Seghiouer, M. A. Alimam, and M. Cherkaoui, "Automated system for matching scientific students to their appropriate career pathway based on science process skill model", IEEE Global Engineering Education Conference, EDUCON, no. April, pp. 1591–1599, 2017, doi: 10.1109/EDUCON.2017.7943061.

[2] M. Nie, Z. Xiong, R. Zhong, W. Deng, and G. Yang, "Career choice prediction based on campus big data-mining the potential behavior of college students", Applied Sciences (Switzerland), vol. 10, no. 8, 2020, doi: 10.3390/APP10082841.

[3] M. Nie et al., "Advanced forecasting of career choices for college students based on campus big data", Frontiers of Computer Science, vol. 12, no. 3, pp. 494–503, 2018, doi: 10.1007/s11704-017-6498-6.

[4] A. Bradley, M. Quigley, and K. Bailey, “How well are students engaging with the careers services at university?”, Studies in Higher Education, vol. 46, no. 4. pp. 663–676, 2021, doi: 10.1080/03075079.2019.1647416.

[5] A. Broström, G. Buenstorf, and M. McKelvey, “The knowledge economy, innovation and the new challenges to universities: introduction to the special issue”, Innovation: Organization and Management, vol. 23, no. 2. pp. 145–162, 2021, doi: 10.1080/14479338.2020.1825090.

[6] I. A. Wowczko, “Skills and vacancy analysis with data mining techniques,” Informatics, vol. 2, no. 4, pp. 31–49, 2015, doi: 10.3390/informatics2040031.

[7] E. Pang, M. Wong, C. H. Leung, and J. Coombes, “Competencies for fresh graduates’ success at work: Perspectives of employers”, Industry and Higher Education, vol. 33, no. 1, pp. 55–65, 2019, doi: 10.1177/0950422218792333.

[8] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum, “Fortune teller: Predicting your career path”, 30th AAAI Conference on Artificial Intelligence, AAAI 2016, no. 1, pp. 201–207, 2016, doi: 10.1609/aaai.v30i1.9969.

[9] D. Kurniadi, E. Abdurachman, H. Warnars, and W. Suparta, “A proposed framework in an intelligent recommender system for the college student”, Journal of Physics: Conference Series, IOP Publishing, 2019, p. 066100. doi: 10.1088/1742-6596/1402/6/066100.

[10] G. A. Ansari, “Career guidance through multilevel expert system using data mining technique”, International Journal of Information Technology and Computer Science, vol. 9, no. 8, pp. 22–29, 2017, doi: 10.5815/ijitcs.2017.08.03.

[11] L. I. Jimenez-Raygoza, A. S. Medina-Vazquez, and G. Perez-Torres, “Proposal of a computer system for vocational guidance with data mining”, 2019 IEEE International Conference on Engineering Veracruz, ICEV 2019, pp. 11–15, 2019, doi: 10.1109/ICEV.2019.8920523.

[12] R. H. Rangnekar, K. P. Suratwala, S. Krishna, and S. Dhage, “Career prediction model using data mining and linear classification”, Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018, 2018, doi: 10.1109/ICCUBEA.2018.8697689.

[13] S. Hutt, M. Gardener, D. Kamentz, A. L. Duckworth, and S. K. D’Mello, “Prospectively predicting 4-Year college graduation from student applications”, ACM International Conference Proceeding Series, pp. 280–289, 2018, doi: 10.1145/3170358.3170395.

[14] M. Huptych, M. Hlosta, Z. Zdrahal, and J. Kocvara, “Investigating influence of demographic factors on study recommenders”, vol. 10948 LNAI. Springer International Publishing, 2018, doi: 10.1007/978-3-319-93846-2_27.

[15] I. H. A. H. and N. F. F. S. T. R. Razak, M. A. Hashim, N. M. Noor, “Career path recommendation system for UiTM Perlis students using fuzzy logic” ICIAS2014: 2014 5th International Conference on Intelligent and Advanced Systems (ICIAS): Technological Convergence for Sustainable Future: 3-5 June 2014, Kuala Lumpur Convention Centre: proceedings, 2014, doi: 10.1109/ICIAS.2014.6869553.

[16] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining”, Computers and Education, vol. 113, pp. 177–194, 2017, doi: 10.1016/j.compedu.2017.05.007.

[17] R. Asif, A. Merceron, and M. K. Pathan, “Predicting student academic performance at degree level: a case study”, International Journal of Intelligent Systems and Applications, vol. 7, no. 1, pp. 49–61, 2014, doi: 10.5815/ijisa.2015.01.05.

[18] J. Gore, K. Holmes, M. Smith, E. Southgate, and J. Albright, “Socioeconomic status and the career aspirations of Australian school students: Testing enduring assumptions”, Aust. Educ. Res., vol. 42, no. 2, pp. 155–177, 2015, doi: 10.1007/s13384-015-0172-5.

[19] A. Nabil, M. Seyam, and A. Abou-Elfetouh, “Prediction of students’ academic performance based on courses’ grades using deep neural networks”, IEEE Access, vol. 9, pp. 140731–140746, 2021, doi: 10.1109/ACCESS.2021.3119596.

[20] P. Berka and L. Marek, “Bachelor’s degree student dropouts: Who tend to stay and who tend to leave?”, Educational Evaluation, vol. 70, no. January, 2021, doi: 10.1016/j.stueduc.2021.100999.

[21] M. A. Al-Hagery, M. A. Alzaid, T. S. Alharbi, and M. A. Alhanaya, “Data mining methods for detecting the most significant factors affecting students’ performance”, International Journal of Information Technology and Computer Science, vol. 12, no. 5, pp. 1–13, 2020, doi: 10.5815/ijitcs.2020.05.01.

[22] G. Feng, M. Fan, and Y. Chen, “Analysis and prediction of students’ academic performance based on educational data mining”, IEEE Access, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.

[23] P. D. Dusane, N. V. Bhosale, V. A. Avhad, and P. K. Naikwade, “Recommendation system for career path using data mining approaches”, International Journal of Scientific Research & Engineering Trends, vol. 6, no. 2, pp. 587–589, 2020.

[24] H. Al-Dossari, F. A. Nughaymish, Z. Al-Qahtani, M. Alkahlifah, and A. Alqahtani, “A machine learning approach to career path choice for information technology graduates”, Engineering, Technology & Applied Science Research, vol. 10, no. 6, pp. 6589–6596, 2020, doi: 10.48084/etasr.3821.

[25] K. Sripath Roy, K. Roopkanth, V. Uday Teja, V. Bhavana, and J. Priyanka, “Student career prediction using advanced machine learning techniques”, International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 26–29, 2018, doi: 10.14419/ijet.v7i2.20.11738.

[26] N. VidyaShreeram and D. A. Muthukumaravel, “Student career prediction using machine learning approaches”, Proceedings of the First International Conference on Computing, Communication and Control System, I3CAC 2021, doi: 10.4108/eai.7-6-2021.2308642.

[27] Q. C. Song, H. J. Shin, C. Tang, A. Hanna, and T. Behrend, “Investigating machine learning’s capacity to enhance the prediction of career choices”, Personnel Psychology, no. May, pp. 1–25, 2022, doi: 10.1111/peps.12529.

[28] U. Khurana, H. Samulowitz, and D. Turaga, “Feature engineering for predictive modeling using reinforcement learning”, 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 3407–3414, 2018, doi: 10.1609/aaai.v32i1.11678.

[29] B. Charbuty and A. Abdulazeez, “Classification based on decision tree algorithm for machine learning”, Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[30] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A brief review of nearest neighbor algorithm for learning and classification”, 2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019, pp. 1255–1260, 2019, doi: 10.1109/ICCS45141.2019.9065747.

[31] Z. Guo, P. Rakshit, D. S. Herman, and J. Chen, “Inference for the case probability in high-dimensional logistic regression”, J. Mach. Learn. Res., vol. 22, pp. 1–54, 2021.

[32] H. Zhang, L. Jiang, and L. Yu, “Attribute and instance weighted naive Bayes”, Pattern Recognition, vol. 111, 2021, doi: 10.1016/j.patcog.2020.107674.

[33] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, “A survey on neural network interpretability”, IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 5, no. 5, pp. 726–742, 2021, doi: 10.1109/TETCI.2021.3100641.

[34] E.-M. El-Mhamdi, R. Guerraoui, and S. Rouault, “Distributed Momentum for Byzantine-resilient Stochastic Gradient Descent”, ICLR, pp. 1–37, 2021.

[35] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends”, Neurocomputing, vol. 408, pp. 189–215, 2020, doi: 10.1016/j.neucom.2019.10.118.

[36] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, “A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data”, Front. Energy Res., vol. 9, no. March, pp. 1–17, 2021, doi: 10.3389/fenrg.2021.652801.

[37] S. A. Alasadi and W. S. Bhaya, “Review of data preprocessing techniques in data mining”, Journal of Engineering and Applied Sciences, vol. 12, no. 16, pp. 4102–4107, 2017, doi: 10.3923/jeasci.2017.4102.4107.

[38] U. of L. Bioinformatics Laboratory, “Data mining”, Orange Data Mining - Data Mining.

[Online]. Available: https://orangedatamining.com/. [Accessed: 10-Oct-2022].

[39] D. Papakyriakou and I. S. Barbounakis, “Data mining methods: A review”, International Journal of Computer Applications, vol. 183, no. 48, pp. 5–19, 2022, doi: 10.5120/ijca2022921884.

[40] H. Silva and J. Bernardino, “Machine learning algorithms: An experimental evaluation for decision support systems”, Algorithms, vol. 15, no. 4, 2022, doi: 10.3390/a15040130.

[41] Rushi, Longadge, S. D. Snehlata, and M. Latesh, “Class imbalance problem in data mining: Review”, International Journal of Computer Science and Network (IJCSN), vol. 2, no. 1, pp. 1–6, 2013, doi: 10.1016/j.ejim.2013.08.659.

[42] H. Ali, M. N. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mushtaq, “Imbalance class problems in data mining: A review”, Indonesian Journal of Electrical Engineering and Computer Science, vol. 14, no. 3, pp. 1552–1563, 2019, doi: 10.11591/ijeecs.v14.i3.pp1552-1563.

[43] J. Liu, “Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data”, Soft Computing, vol. 26, no. 3, pp. 1141–1163, 2022, doi: 10.1007/s00500-021-06532-4.

[44] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, “Feature selection: A review and comparative study”, E3S Web of Conferences, vol. 351, p. 01046, 2022, doi: 10.1051/e3sconf/202235101046.

[45] J. Chen, J. Hu, and G. Zhang, “Feature selection based on gain ratio in hybrid incomplete information systems”, 2021 IEEE International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2021, pp. 728–735, 2021, doi: 10.1109/ISKE54062.2021.9755425.

[46] L. J. Cai, S. Lv, and K. B. Shi, “Application of an improved chi feature selection algorithm”, Discrete Dynamics in Nature and Society, vol. 2021, 2021, doi: 10.1155/2021/9963382.

[47] M. Massoudi, S. Ghory, and M. Massoudi, “Career recommender system using decision trees”, 2021 International Conference on Smart Generation Computing, Communication and Networking, SMART GENCON 2021, pp. 29–32, 2021, doi: 10.1109/SMARTGENCON51891.2021.9645805.

[48] M. Yamashita, Y. Li, T. Tran, Y. Zhang, and D. Lee, “Looking further into the future: career pathway prediction”, Proceedings of the 1st International Workshop on Computational Jobs Marketplace at WSDM, February 25, 2022, Tempe, AZ, vol. 1, no. 1. Association for Computing Machinery, 2022.

[49] A. Triayudi and W. O. Widyarto, “Educational data mining analysis using classification techniques”, Journal of Physics: Conference Series, vol. 1933, no. 1, 2021, doi: 10.1088/1742-6596/1933/1/012061.

[50] K. A. Hoff, C. Chu, S. Einarsdóttir, D. A. Briley, A. Hanna, and J. Rounds, “Adolescent vocational interests predict early career success: Two 12-year longitudinal studies”, Applied Psychology, vol. 71, no. 1, pp. 49–75, 2022, doi: 10.1111/apps.12311.

APPENDIX A

Table A1 Summary of models performance or evaluation metrics

Models / Classifiers	Dataset I						Dataset II					
	AUC	Precision	Recall	F1	LogLoss	Specificity	AUC	Precision	Recall	F1	LogLoss	Specificity
k-Nearest Neighbor	0.607	0.358	0.594	0.446	0.337	0.868	0.634	0.368	0.528	0.396	0.354	0.878
Logistic	0.584	0.176	0.175	0.175	0.464	0.893	0.547	0.264	0.264	0.270	0.644	0.895

Regression												
Naïve Bayes	0.573	0.099	0.330	0.082	1.476	0.859	0.535	0.019	0.375	0.034	2.393	0.874
Neural Network	0.598	0.188	0.190	0.186	0.461	0.893	0.592	0.354	0.371	0.357	0.408	0.901
Stochastic Gradient Descent	0.518	0.194	0.184	0.187	5.822	0.894	0.551	0.323	0.323	0.322	5.019	0.901
Support Vector Machine	0.537	0.566	0.383	0.292	0.297	0.865	0.581	0.523	0.442	0.370	0.282	0.896
Decision Tree	0.507	0.159	0.159	0.158	4.683	0.893	0.528	0.252	0.250	0.248	4.167	0.892

Table A2 Summary of model performance or evaluation metrics after applying SMOTE

Models / Classifiers	Dataset I						Dataset II					
	AUC	Precision	Recall	F1	LogLoss	Specificity	AUC	Precision	Recall	F1	LogLoss	Specificity
k-Nearest Neighbor	0.896	0.476	0.527	0.453	0.271	0.940	0.944	0.699	0.670	0.608	0.203	0.952
Logistic Regression	0.916	0.691	0.702	0.695	0.184	0.962	0.945	0.758	0.777	0.764	0.165	0.968
Naïve Bayes	0.752	0.567	0.474	0.480	0.487	0.825	0.900	0.676	0.589	0.602	0.418	0.941
Neural Network	0.927	0.732	0.728	0.729	0.213	0.966	0.965	0.831	0.828	0.829	0.134	0.975
Stochastic Gradient Descent	0.820	0.658	0.681	0.667	2.442	0.959	0.884	0.790	0.799	0.793	1.726	0.971
Support Vector Machine	0.940	0.749	0.699	0.713	0.149	0.853	0.961	0.810	0.788	0.786	0.122	0.969
Decision Tree	0.760	0.502	0.506	0.503	2.402	0.938	0.848	0.723	0.670	0.671	1.208	0.967