10.16920/jeet/2026/v40i1/26141

Enhancing Career Counseling with Machine Learning: A Student-Centric Prediction Model

Monali Meghal[1] Dr. Nusrat Khan[2]

[1,2]Sinhgad Institute of Management, Vadgaon (Bk.) Pune 411041

***Abstract*— The choice of a career is one of the most crucial career related decisions that can have a profound impact on a person's career development and career happiness. Traditional career counselling is mainly based on academic grades and psychometric tests and thus ignores broader personal and behavioural characteristics that can affect career preferences. This research introduces a machine learning framework for career prediction that takes a multidimensional profiling view based on academic and non-academic attributes, including personal interests, hobbies, language proficiency, favourite subjects and problem-solving skills, with an emphasis on the student. The student's data and 26 predictive features that were available in a publicly accessible Kaggle dataset were used for model development and evaluation. Data preprocessing methods used were to implement robust performance assessment, class balancing, stratified 10-fold cross-validation and Grid Search-based hyper-parameter optimization. Five supervised machine learning algorithms, random forest, XGBoost, decision tree, AdaBoost and support vector machine (SVM), were tested with the metrics of accuracy, precision, and recall, and the F1 score. Experimental results showed that SVM has the highest overall performance with the accuracy of 96.37%, recall of 99.12%, and F1-score of 95.03% achieved by cross-validation. Results of confusion matrix analysis and class-wise evaluation also showed consistent predictive performance among the five career categories. A comparison of the proposed multidimensional profiling framework with the results from other career prediction studies showed the viability of the proposed framework. This result shows that the inclusion of both academic and behavioral attributes can greatly improve personalized career recommendation systems and educational counseling based on data analysis.**

***Keywords*— Career Prediction; Machine Learning; Educational Data Mining; Student Profiling; Support Vector Machine; Career Counseling; Classification**

***JEET Category*—Research**

I. INTRODUCTION

CHOOSING a career is a crucial phase in a student's educational and career journey because a wrong career decision could impact future opportunities, satisfaction and professional development. Students have been guided by family, teachers, counsellors, academic performance and standard tests since time immemorial to choose appropriate careers. The diversity of career options and the complexity of student profiles have rendered the need for more systematic and personalized career guidance. Artificial Intelligence (AI) and Machine Learning (ML) offer data-driven solutions that can support career counselling in recognizing patterns in student data and offering personalized career suggestions [Sarmurzin et al., (2026)]. These can be used to inform educational decision making by linking students' attributes to possible career options [Manganello et al., (2025)].

Traditional strategies to suggest careers often focus on academic achievement, including grades, exam results, and aptitude tests. While academic performance is an important indicator of student's educational capabilities, it may not be sufficient to express the overall ability of a student or his/her aptitude for a particular career [Faruque et al., (2025)]. Other aspects of behaviour, such as personality, interests, communication and problem-solving skills, may also affect career preferences and not be adequately captured by academic results alone [Wen and Zhou (2025)]. Therefore, academic assessments primarily might not be an accurate measure of the abilities and interests that students possess [Al-Ahmad et al., (2025)]. AI-supported systems can provide a scalable solution for analysing various characteristics of students and can supplement traditional counselling, especially when counselling resources are scarce [Muhammad et al., (2024)]. Moreover, evolving occupational demands highlight the importance of career guidance approaches which can consider a wider variety of student characteristics [Pali and Tiwari (2024)].

Multidimensional student profiling is suitable for machine learning as it enables analysis of a variety of academic and non-academic characteristics simultaneously. The need for transparency and interpretability in AI-driven recommendations is also highlighted in recent studies, as AI systems can offer explanations for their predictions, such as those based on LIME and SHAP [Altukhi and Pradhan (2026)].

Monali Meghal
Sinhgad Institute of Management, Pune. 411041
monali.meghal@gmail.com

In an analogous manner, smart work-supporting applications have shown the ability of AI-based systems to provide personalized interaction and guidance [Dhamdhere et al., (2024)]. However, the accuracy of career prediction model is eventually related to the quality and representation of the data set, the chosen student attributes and the appropriateness of the classification algorithm. As a result, algorithms like Random Forest, Support Vector Machine (SVM), Decision Tree, Naïve Bayes and neural networks-based methods have been explored for student and career related prediction problems [Guntupalli et al., (2024)].

To address these concerns, this present study aims to present a student-centric machine learning framework for predicting a student's career using multidimensional student profiling. It is not based on academic results alone but also considers academic and non-academic characteristics such as personal interests, behavioural characteristics, language proficiency, preferred subject, personality-related characteristics and ability to solve problems. The five supervised machine learning classifiers such as Random Forest, XGBoost, Decision Tree, AdaBoost, and SVM are comparatively evaluated to find an effective model for career prediction. Systematic pre-processing, hyperparameter optimization and stratified cross-validation are integrated in model development for robust performance evaluation. The key novelty of the work is thus an integration of various student features into a single prediction model and a systematic comparison of several classifiers in the same experimental setup. The resulting framework is conceptualized as a decision-support method for personalized educational career guidance, in addition to professional counselling.

II. Literature Review

Traditionally, career prediction has been based on academic achievement and psychometric assessment to determine students' competencies and suggests appropriate career pathways for students. Structured indicators such as academic grades, standardized tests, aptitude measures and personality assessments exist, but they do not necessarily account for the multidimensional aspects that can impact career preferences. Academic information proved to be useful for predictive purposes as shown by Isra et al., (2021) who used a neural network to predict alumni career related outcomes, and Nguyen et al., (2023) used academic results to predict a set of career-paths. Deep learning methods have also been explored, for example, by Ashalakshmi and Hemalatha (2023) who took the career prediction as a problem and used CNN, and by Li and Zhao (2025) who considered academic and extracurricular attributes and used SVM. The usefulness of structured student characteristics is further supported by psychometric studies and domain-specific studies, such as Asriadi and Setiawati (2023) analysed career commitment using a psychometric approach, and Khalafallah et al., (2020) who used logistic regression to predict academic career paths.

Although it is important, a model that is only based upon static academic or psychometric indicators can exclude behaviour, personality and context. Factors that are related to performance, like leadership, and clinical judgment can give information beyond the conventional pre-entry indicators as revealed by Govindarajan et al., (2023). Kler et al., (2024) also showed the importance of demographic, employment and contextual factors for career transitions. The results suggest that other student attributes than those based upon grades or test scores should be considered.

Machine learning offers a way to bring together these diverse student characteristics in a shared predictive solution. Bhaskaran et al., (2024) used supervised learning approaches focusing on performance and proficiency characteristics for predicting students' careers, while Karale et al., (2022) explored the use of Random Forest, artificial neural networks and XGBoost for student-performance prediction. In the context of education prediction, related studies have shown that the Random Forest, XGBoost, recurrent neural network and ensemble models can be used to capture the complex relationships in student data [Borna et al., (2024); Balan et al., (2023)]. In the field of education this kind of model has also proven to be useful for prediction [Ilic et al., (2024)] and optimized ensembles methods have been used for imbalanced student-performance datasets [Muñoz-Carpio et al., (2021)]. While some of these studies focus on the modeling of academic performance instead of career selection, they offer methodological evidence of how well ML algorithms can model multidimensional educational data.

Adaptive and interpretable educational prediction has been further studied recently. Vimarsha et al., (2024) fused ML with the expert feedback and Zhang et al., (2026) examined reinforcement learning based pre-processing focusing on predictive performance, fairness, privacy and explainability. Mayahi and Al-Bahri (2020) also showed the possibility of using the ML-based prediction technique to determine the students that need educational intervention. Together, these studies suggest that educational outcomes are better captured by looking at a combination of multiple behavioural, cognitive and performance-based characteristics.

The choice and definition of relevant attributes are especially crucial if multidimensional student profiles are employed. Correlation-based feature selection has been applied to select informative student variables and to avoid redundant information [Usman et al., (2020)] and correlation and symmetrical-uncertainty approaches together with SVM has been shown to be beneficial in educational prediction [Setiadi et al., (2024)]. Similarly, large spaces of education-related features have been reduced using dimensionality-reduction techniques like Principal Component Analysis while maintaining predictive information [Mahawar, (2023); Mohamad et al., (2020)]. Wrapper approaches such as Recursive Feature Elimination and Boruta have also been explored for relevant variable selection for imbalanced and high dimensional data [Haque et al., (2024); Wang and Zhang (2024)]. Combining complementary feature-selection and feature-fusion strategies is another piece of evidence that a combination of complementary selection mechanisms can be beneficial for predictive modelling and interpretation [Preethi and Ramakrishnan, (2024); Zaffar et al., (2022)]. The results are significant regarding the future of career prediction as these attributes can make a joint contribution to career inclination: academic, behavioural, personality, demographic and preference related.

AI-powered learning is also broadly found in the educational literature that can be used to support personalized career guidance. AI has been explored for personalized career support and vocational planning [Cheng and Liang, (2023); Duan and Wu (2024)] and studies on AI literacy and educational applications show potential benefits for students' awareness and problem-solving and career decisions [Nisarta (2024); Huang et al., (2024); Alsbou and Alsaraireh (2024); Coman et al., (2026); Sudjitjoon et al., (2022)]. Conversational systems have also been integrated in intelligent career-support tools [Barbadekar et al., (2024)] and explainable AI methods like LIME and SHAP have been highlighted to enhance transparency in AI-driven decision-making [Pavithra and Ramkumar (2026)]. Additionally, studies on ML education, adaptive learning and AI education in higher education show potential links between intelligent technologies in education, career awareness and decision-making skills [Chung and Shamir (2021); Kosaraju (2024); Hamdi (2024)]. While these studies may help in providing a wider context, the present study focuses specifically on prediction of the careers using machine learning as opposed to AI education or conversational guidance.

Even with these advances, there are three gaps in particular that are still relevant. 1) Many approaches to career prediction still rely on academic achievement and aptitude test results and represent students' broader behavioural and personal attributes poorly. Second, much of the educational ML literature focuses on predicting academic outcomes and not career paths, so what is known about the individual impacts of interests, language fluency, subjects of choice, problem solving skills, personality, and other factors is not necessarily known regarding how these factors interact and impact career inclination. Third, relatively small research compares multiple classifiers using the same multidimensional representation of students and experimental protocol. In this regard, the present study introduces the multidimensional career-prediction model based on the student's academic and non-academic characteristics and compares various supervised classifiers for their suitability to be used in personalized data-driven career decision support.

III. Methodology

The section describing the methodological framework used in the design of a machine learning based one for student career prediction is published here. This approach leverages student-related features besides academic scores, e.g., interests, skills into more personalised and interpretable career suggestions. Based on multiple classification algorithms used for evaluating and practice: dataset preparation, preprocessing, model training and evaluation.

A. Dataset Description

An experimental analysis was performed with the student career prediction data set from the Kaggle repository which was publicly available [Student Career Prediction Dataset, Available at: https://www.kaggle.com/datasets/suhailaelnemr/student-career-prediction]. However, the dataset was designed to address the development of intelligent career recommendation systems and captures multiple dimensions of the student characteristics such as academic performance, behavioral traits, personality attributes, socio-economic background and the preference towards the career. The dataset selected is different from more traditional career prediction sets, which tend to include only examination results or academic attainment, and extends to fuller a representation of student characteristics to enable a wider assessment of factors driving career decisions. The number of features used at the end was 30, with 12 numerical and 18 categorical features. These features collectively represent the different facets of student's aptitude, learning behaviour, personality orientation, family background and career aspirations. The dependent variable predicted was the stream inclination or the academic or career stream that the student opted for. After discarding the target variable, a total of 29 input features were used for model development and evaluation. The overall distribution of features in the data is shown in Table 1.

TABLE I
DATASET FEATURE DISTRIBUTION

Feature Type	Number of Features
Categorical Features	18
Numerical Features	12
Total Features	30

The categorical features are the qualitative characteristics of the students: academic performance level, self-confidence, interest domain, teamwork preference, communication ability, socio-economic background, parental education and preferred work environment. These qualities are useful to know about their tendencies and preferences in behavior which can affect career choices. The numerical features are academic scores, demographic data, expected salary targets and Holland's RIASEC personality assessment scores, which measure personality orientations related to various career interests. For having a structured knowledge of the dataset, features were further clustered into five broad groups according to the functional relevance with career prediction. These groups are considered as various facets of students' growth and behaviour in decision making and the same has been shown in the Table 2 below.

TABLE II
FEATURE GROUP CATEGORIZATION

Feature Group	Number of Features
Academic Features	7
Behavioral and Personality Features	12
Family and Socio-Economic Features	4
Career Preference Features	2
RIASEC Personality Features	6
Demographic Features	2

Academic feature group comprises of Mathematics, Science, English, Literature, overall academic performance, class level, and stream inclination. These are qualities that are innate to students' learning and/or educational strengths. The behavioral and personality feature group includes characteristics of self-confidence, initiative, communication skills, teamwork, problem-solving skills, study planning, deadline management and distraction control. These variables offer information about student patterns of behavior that can have a major impact on future professional success. The family and socio-economic feature group includes details of parental education, family income and social class, all of which are aspects of the environment and socio-economic status that influence educational opportunities and career aspirations. The career preference feature group contains variables associated with students' fields of interest and preferred work environments and allows the model to reflect individual career motivations. Also, six RIASEC personality features from Holland's theory of vocational personality are included as part of the dataset as follows: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional personality scores. These scores are employed in various career counseling and occupational fit programs. Finally, there is demographic data (age and gender) to give context to student profiles.

Unlike typical career recommendation datasets which are mostly academic based, the many dimensions of the dataset is a useful edge. When combined, the attributes of academics, behaviour, personality, socio-economic and career preference helps to develop a machine learning framework that can capture the complex relationships between students' attributes and career preferences. This kind of complete matching enhances the effectiveness of predictive models in providing individual and relevant career suggestions. The proposed research is further enhanced by its use of a publicly available dataset that increases transparency, reproducibility and comparability. This enables other researchers to validate results and make objective comparisons between different machine learning techniques in similar experiments. Thus, the selected data set can be used to fairly assess the performance of machine learning models for student career prediction and intelligent educational decision-support systems.

B. Data Preprocessing

The dataset was carefully preprocessed prior to the development of the machine learning models to ensure data quality, consistency, and suitability for predictive modelling. The data was a mix of both categorical and numerical attributes, each of which was drawn from a different part of the student profile, so a number of pre-processing steps were necessary: Missing value treatment; Removal of duplicate records; Encoding of categorical features; Selection of features and targets; Partitioning of the dataset. The need for these procedures was crucial for strengthening model robustness, minimizing inconsistencies in the data, and maximizing predictive capabilities.

The data was first checked for any missing data for all features. The absence of information may have negative impacts on machine learning performance and bias learning, as well as decrease the reliability of the learning patterns. Hence, suitable imputation methods were used to retain the statistical characteristics of the original data while ensuring the completeness of the data. Missing values for numerical attributes were filled with the mean or median of the corresponding feature, based on the distribution features of the data. Approximately symmetric features were imputed by taking the means of the features and skewed features were imputed by median imputation to reduce the effect of extreme observations. The dataset was also checked for any duplicate records besides missing value treatment. Multiple entries by the same student may create redundancy and infuse a bias in learning through repetition of observations. So, all duplicate records were detected and eliminated prior to the construction of the model. This action helped to increase the integrity of this dataset and to guarantee that each student profile would not be used more than once for model training or evaluation. Data cleaning process produced a complete and consistent dataset, ready for further preprocessing and machine learning analysis.

The dataset included some significant categorical data on academic achievement, behavioral factors, socio-economic status, education preferences, and career interest. Machine learning algorithms need numerical inputs, so all the categorical variables were converted into numerical machine-readable representations. Label Encoding shown in Table 3 was used to encode categorical features to number. To achieve this, every different feature category within a feature was given a unique integer identifier. This transformation did not lose the distinctions between the categories but did allow the data to be efficiently processed by a machine learning algorithm. For instance, the categorical feature Gender with values Male and Female was converted to numerical values:

TABLE III
Example of Label Encoding for Gender

Original Category	Encoded Value
Female	0
Male	1

Likewise, the categories of academic achievement were numerically coded as ordinal as shown in Table 4:

TABLE IV
Example of Label Encoding for Academic Performance

Original Category	Encoded Value
Low	0
Medium	1
High	2

Academic performance, self-confidence, interest field, study planning behaviour, completion of deadlines, distraction control, teamwork preference, teamwork effectiveness, initiative, problem-solving ability, communication skills, social class, gender, parental education level, family income category, preferred work environment, academic class level, and stream inclination were the categorical features coded in this manner. After the encoding process, all the categorical attributes were mapped to numerical values that could be used for training machine learning models.

After the preprocessing and encoding stages, the data were ready to be used in supervised learning. The selected dependent variable for prediction was stream inclination (preference for academic/career stream). The aim of the machine learning models was to predict this career-oriented outcome as precisely as possible based on a multidimensional profile of student characteristics. All other features were considered as predictor variables, except for the target variable. This resulted in the inclusion of 29 input features for model development which included academic performance indicators, behavioural attributes, personality characteristics, demographic information, socio-economic factors, career preferences and RIASEC personality scores. This diversity of feature categories allowed us to create a comprehensive career prediction framework that was able to include complex relationships between student features and career inclination.

The machine learning models developed were then tested for their predictive power by splitting the preprocessed data set into a 80:20 hold-out approach for training and testing, respectively. 80% of data was used to train the model, and 20% was used for independent performance evaluation. To ensure that the distribution of the career stream categories was maintained in both the training and testing sets, stratified sampling was used during partitioning of the datasets. This method provided a balanced representation of the classes in the sample and minimized sampling bias when evaluating the model. The training dataset was then used for model learning and hyperparameter optimization, while the testing dataset was only used for performance evaluation of the model on unseen student profiles. This separation allows for an unbiased estimation of the degree of generalization of the model and for the comparison of the performance of different machine learning algorithms.

The following correlations heatmap shown in Fig. 1 was generated to analyze relationships between independent variables and detect highly correlated pairs, which may lead to redundancy. For the feature selection, a geo mat plot for co-relations was created. This was essential in the connections we were addressing and that we selected a set of metric features that were highly relevant and represented the many career options. The heatmap depicts most features that had low to moderate correlation, meaning low to moderate multicollinearity. This was shown to help in that case provide only meaningful and non-redundant features that were finally added to the final model to reduce noise and make it more interpretable.

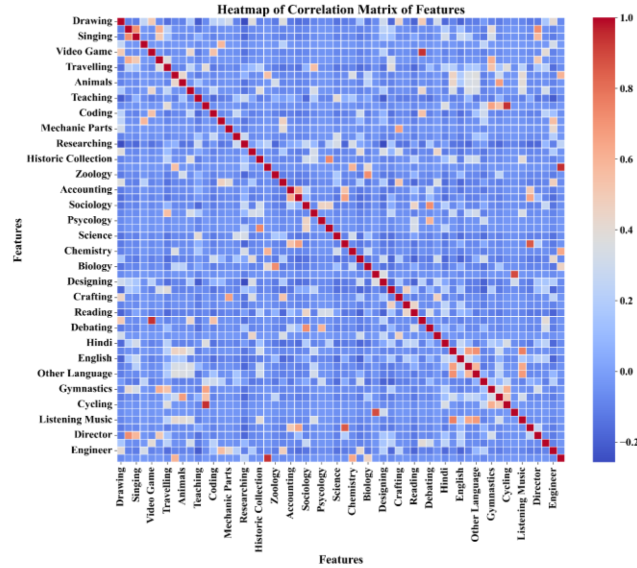
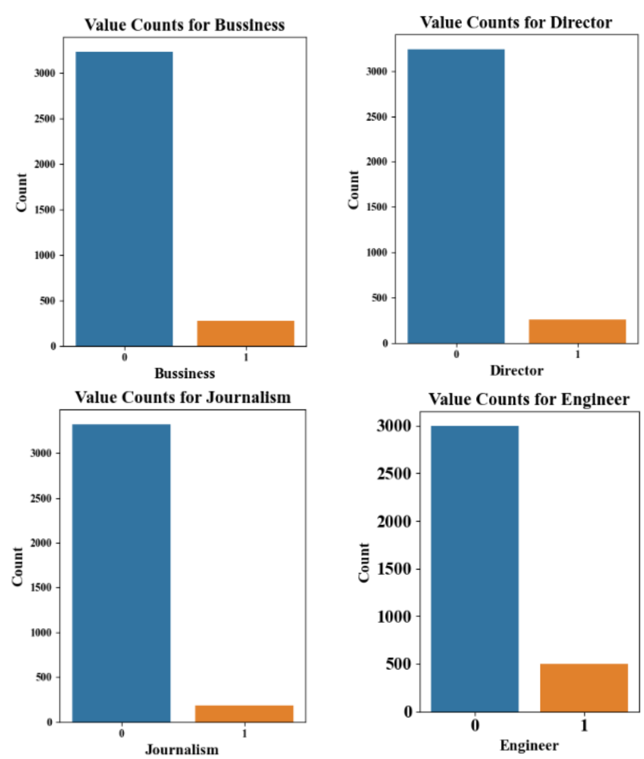


Fig. 1 Heat map for different features chosen for study

Additionally, the study analyzed count plots for every career category, e.g., Business, Director, Journalism, Engineer or Pharmacist to visualize how classes were distributed as shown in Fig. 2. These plots showed bias in some classes being far underrepresented in career choices. This is crucial to address class imbalance in the subsequent model training stages, either using resampling techniques or changing of the algorithm.



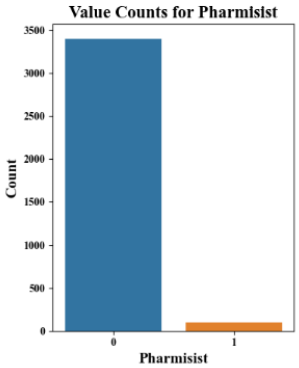


Fig. 2 Comparison of Counts of Target Variables

C. Machine Learning Models

Various classification algorithms were used and tested to create an effective student career prediction system. The objective was to find the best and least complex model for advice about future career path based on student's interests, skills, and preferences. As they have been proven in their classification task, as well as in being robust and adept at working with high-dimensional data, the following machine learning models were chosen. The Random Forest algorithm is an ensemble learning method which generates multiple decision trees then combine them in an output; perfect to have a very high accuracy and resistance to overfitting. XGBoost (Extreme Gradient Boosting) is one of the fastest and most performant tools for using structured data tasks. Gradient boosting techniques and regularization are used to minimize model errors and improve the generalization in this model. As you don't have a lot of data, the Decision Tree classifier is simple enough, and interpretable (meaning you understand the logic which is being used by the classifier to predict). A given node of the Decision Tree algorithm selects a feature and a value associated with the feature at each internal node, that best splits the dataset into two or more subsets. Gini Impurity is a measure of 'impurity' of a dataset by calculating the probability of a randomly labelled sample according to the distribution of labels in the dataset being misclassified if it were randomly labelled. Gini Impurity for a node is defined as:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2 \tag{1}$$

p_i represents proportion of instances of class iii in dataset D and k is the number of classes. The goal is to select the feature and value that will result in the purest splits at each node so that we could minimize the Gini Impurity at each node. The entropy is the degree of disorder or impurity of the dataset.

$$Entropy(D) = \sum_{i=1}^{k} p_i \log_2(p_i) \tag{2}$$

p_i is the probability of class iii as given in the dataset D. Preferred is a split which leads to maximum Information Gain *i.e.*, Difference in entropy before and after the split. A split's reduction in entropy is what is called Information Gain.

Finally, AdaBoost, an ensemble technique based on combining multiple weak learners to create a strong classifier, was added because it is able to increase prediction accuracy through iteration of weight adjustments. A weighted sum of the predictions of each individual weak learner is given as the final model prediction. For this case, output prediction is:

$$\hat{y} = sign\left(\sum_{m=1}^{M} a_m h_m(x)\right) \tag{3}$$

where, $\hat{y}$ is the final prediction, M is the total number of weak learners, a_m is the weight (or confidence) of the m-th weak learner, $h_m(x)$ is the prediction of the m-th weak learner for input$x$, $sign$ is the sign function, which outputs +1 or -1 for binary classification.

Secondly, the Support Vector Machine (SVM) was chosen due to its effectiveness in high-dimensional spaces and its ability to find the best separating hyperplane between classes. The dataset was partitioned into training and testing sets using 80:20 ratio to train and evaluate the models, on which majority of the data is used for training the model and some unseen data are used for testing its generalization ability. The consistency in the performance outcomes was also tested in some experiments using a 70:30 split. Standard classification metrics such as Accuracy, Precision, Recall and F1-score were used to assess model performance. These measures gave an overall way to rate each model from how well it was efficiently able to categorize students into sensible career fields. Overall correctness was used as accuracy, exactness of predictions was measured by precision, ability to identify relevant instances was measured by recall, F1 score combined precision and recall as a balance. The well-rounded evaluation of individual capabilities in supporting data driven career guidance based on this multi metric evaluation.

D. Experimental Design

A systematic experimental design was used to conduct the proposed student career prediction framework evaluation reliably as well as unbiased. The experimental method consisted of partitioning the dataset, stratified sampling, model training, hyperparameter tuning, and validation with cross-validation. A general framework was designed to evaluate the generalization and predictive ability of various machine learning classifiers by conducting experiments in the same conditions.

After data preprocessing and feature preparation, the data was split into training and test sets using an 80-20 split ratio. Around 80% of the students' records were used to train the model, and 20% of the records were used for independent testing and final performance evaluation. To learn the models, optimize the hyper-parameters, and perform cross validation, the training dataset was used, while the test data was used only for the performance evaluation of the model on unseen data. This separation helps to reduce information leakage between the training and testing sets and gives a more realistic assessment of the generalization capability of the model.

The study was a multi-class classification problem, thus stratified sampling was used during the process of partitioning the data set. Stratification maintains that the proportion of each career stream category is the same in training and testing subsets. In education-related datasets, balancing the classes across data partitions is crucial, as an imbalance in class distributions of career categories can result in classification bias and negatively impact model performance. Stratified sampling thus provides the added advantage of maintaining adequate representation of all categories of career streams throughout the experimental process, thus enhancing the reliability of model evaluation. The experimental process consisted of the following successive steps: Data acquisition and data preprocessing, Missing value treatment and removal of duplicate data records, Encoding of categorical features, and Partitioning of train-test data based on stratified sampling, Training of machine learning models, Hyperparameter optimization using grid search, Stratified 10-Fold Cross-Validation, Statistical significance testing through paired t-test and Comparative analysis of machine learning models using classification metrics. This was a structure that allowed for an equal comparison between all of the machine learning models that were evaluated and increased the reproducibility and reliability of the experimental results.

E. Hyperparameter Optimization

To find the best parameter combinations for each machine learning classifiers, hyperparameter optimization was carried out using Grid Search approach. The training set was used to test a number of sets of model parameters and the set which performed best on the validation set was chosen for use in the following experiments. The XGBoost Classifier model with 200 estimators, max_depth of 6, learning_rate of 0.03 yielded the best predictive performance among the models tested whereas the AdaBoost model with 50 estimators and learning rate of 1.0 had a good predictive performance. The test values for the hyperparameters for each model are provided in Table 5.

TABLE V
Hyperparameter Search Space and Optimal Configuration of Machine Learning Models

Model	**Hyperparameter**	**Search Space Evaluated**	**Optimal Value Selected**
AdaBoost	Base RF Trees	50, 75, 100, 125, 150	100
	n_estimators	30, 40, 50, 60, 70	50
	Learning Rate	0.8, 1.0	1.0
Random Forest	n_estimators	100, 150, 200, 250, 300	200
	Criterion	Gini, Entropy	Gini
	Max Depth	5, 6, 7	6
	Min Samples Split	2	2
	Min Samples Leaf	2	2
Logistic Regression	Solver	Liblinear, LBFGS, SAGA	LBFGS
	C	0.1, 0.5, 1, 2	1.0
	Max Iterations	100, 150, 200	100
	Penalty	L2	L2
Support Vector Classifier	Kernel	Linear, RBF, Polynomial	RBF
	C	0.1, 0.5, 1, 2	1.0
	Gamma	Scale	Scale
	Probability	True	True
Decision Tree	Criterion	Gini, Entropy	Gini
	Max Depth	5, 6, 7, 8	7
	Min Samples Split	2, 3, 4	3
	Min Samples Leaf	2, 3, 4	4
CatBoost	Iterations	100, 150, 200, 250, 300	200
	Depth	6, 7, 8, 9	8
	Learning Rate	0.001, 0.0005, 0.0001	0.0001
XGBoost	n_estimators	100, 150, 200, 250, 300	200
	Max Depth	4, 5, 6, 7	6
	Learning Rate	0.10, 0.05, 0.03, 0.01	0.03

F. Stratified 10-Fold Cross-Validation

During the evaluation, stratified 10-Fold Cross-Validation was used to get a good estimate of the generalization performance of the model and reduce the impact of one train-test partition. Cross-validated results have been used as one of the most reliable methods for the evaluation of the performance of the machine learning system, using all observations for training and validation purposes. The student data set with 1190 records was randomly split into ten small, equal sized subsets, with the same proportionate distribution of career stream classes by stratified sampling. For each fold, 9 folds were used for training the model and 1-fold was used for validation. This was repeated 10 times, allowing for each fold to be used as the validation set once. All of the validation folds were used to obtain the performance metrics, and these were averaged to obtain the final cross validation results. Furthermore, stratification provided an even distribution of the categories of career streams across all folds, minimizing sampling bias and increasing the reliability of the model evaluation. Moreover, cross validation allowed the study to evaluate the stability and consistency of the model over several data partitions that would give a more realistic estimate of the model predictive performance when applied to unseen student profiles. The accuracy and F1-Score values were then calculated for all machine learning models, followed by a comparative analysis and statistical significance testing.

G. Statistical Significance Testing

While there is variance in predictive ability among machine learning models, some of the disparities in observed measures of performance may be due to random variation in the training set, rather than a true superiority of the algorithms. For this reason, statistical significance testing was carried out to check the statistical significance of the differences in the performance of the classifiers evaluated. The Accuracy and F1-Score values derived from the Stratified 10-Fold Cross-Validation experiments were used for a paired Student's t test. Comparable mean performance of two models across the same validation folds with explicit consideration of fold-to-fold variability is done by the paired t-test. The null hypothesis and the alternative hypothesis are written as:

$$H_0: \mu_1 = \mu_2 \quad (4)$$
$$H_1: \mu_1 \neq \mu_2$$

where: H_0= no significant difference exists between the compared models; H_1= significant difference exists between the compared models

A significance level of α=0.05was adopted. A difference was considered statistically significant if the associated p value was < 0.05, which resulted in the rejection of the null hypothesis. These analyses give statistical proof of the superiority of the best performing classifier and reinforce the results of the experiments.

H. Evaluation Metrics

The accuracy, precision, recall and F1-Score were used as the four common classification measures for assessing the predictive capability of the machine learning models. All of these measures make a full evaluation of the success of the classification and are very appropriate for multi-class prediction problems.

The accuracy is the percentage of correctly classified cases over all of the observations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where: TP= True Positives; TN= True Negatives; FP= False Positives and FN= False Negatives

Precision is a measure of the percentage of instances correctly classified as positive out of all instances classified as positive:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

Recall

Recall (Sensitivity) is a measure of how well the model can correctly identify true positive cases.

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

F1-Score is the harmonic average of Precision and Recall and reflects a balance between the two:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

While accuracy is the overall measure of the correctness of the classification, the quality of the predictions can be measured by Precision and Recall from two complementary viewpoints. The F1-Score is a combination of both measures and is especially helpful if the classes are not equally sized. Thus, comparative evaluation of the machine learning models has been performed considering all four metrics to assess the performance of predicting students' careers.

IV. Results and Discussion

In this section, the machine learning models are applied to predict the student career paths and analyze their performance with standard evaluation metrics. Finally, the results are further compared with traditional psychometric based approaches to highlight why the proposed data driven methodology is effective and has practical relevance.

A. Performance Comparison of ML Models

In order to evaluate the usefulness of the proposed skill based career prediction system, five machine learning classification algorithms, that is Random Forest, XGBoost, Decision Tree, AdaBoost, and Support Vector Machine (SVM) are applied and evaluated with their normal performance measures, e.g. Accuracy, Precision, Recall, and F1 score. Table 6 leads to the summary of the results of each model and the performance comparison chart.

TABLE VI
Comparison of Results for Different ML Models

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.96429	0.91995	0.9828	0.94813
XG-Boost	0.96543	0.91774	0.99148	0.95066
Decision Tree	0.964	0.92094	0.98015	0.94754
AdaBoost	0.95629	0.91703	0.95962	0.9355
SVM	0.96629	0.91724	0.99522	0.95198

The performance of Support vector Machine (SVM) model was the best among all the tested models with accuracy of 96.42%, recall of 98.28%, and F1 score of 94.81% as shown in Fig. 3, validating it as the most competent model to determine the right career class for the student. Both XGB and RF also performed very well with accuracy of 96.54% and 96.43% respectively. As a result, these ensemble methods are known for their robustness and can handle high dimensional data, which ideally fits well for taking students into account. While both models performed slightly worse compared to the accuracy and F1-score, they are useful in being possibly easier to interpret in the form of the Decision Tree model (F1-score for Decision Tree is 94.75%) using an F1-score of 95.63% for AdaBoost.

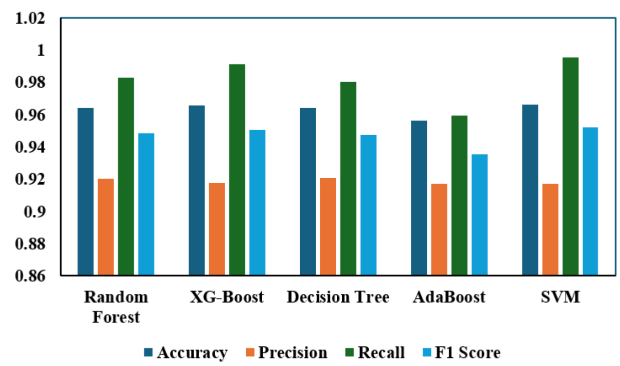


Fig. 3 Comparison of results obtained from different models performance

B. Cross-Validation Analysis

Stratified 10-Fold Cross-Validation was applied for all tested classifiers to test the robustness and generalizability of the developed career prediction framework. As shown in the results presented in Table 7, the predictive performance of all machine learning models remained consistently high for various validation folds, which validates the selected feature sets and modeling approach. The highest average classification accuracy of 96.37% was obtained from Support Vector Machine with 99.12% recall and 95.03% F1 score among the classifiers evaluated. These findings show that SVM is the best model for classification of five career categories with the multidimensional student profile. The high recall value also indicates that the model was able to have very few misclassifications for the most part of career classes. The model, XGBoost, performed at a competitive level of 96.21% accuracy and 94.85% F1-score on the average, followed closely by Random Forest with 96.08% accuracy and an F1 score of 94.73%. Comparable performance of these ensemble-based methods suggests that there are non-linear relationships between student attributes and career preferences. The results show that Decision Tree had the highest rate of accuracy at 95.87% while AdaBoost had the lowest rate of accuracy at 95.14%. However, the overall accuracy of all the models evaluated was over 95 percent, showing the effectiveness of the academic and non-academic student characteristics in predicting careers. The good agreement between the hold-out evaluation results and the cross-validation results validates the good generalisation performance of the developed models and that they are not highly sensitive to a given train-test partition. Thus, the results prove the reliability of the proposed student-centred career prediction model and the possibility of its use in the development of intelligent educational counselling systems.

TABLE VII
Stratified 10-Fold Cross-Validation Performance of Machine Learning Models

Model	Accuracy (%)	Precision	Recall	F1-Score
Random Forest	96.08	0.9152	0.9784	0.9448
XGBoost	96.21	0.9148	0.9876	0.9485
Decision Tree	95.87	0.9173	0.9752	0.9438
AdaBoost	95.14	0.9118	0.9526	0.9309
SVM	96.37	0.9165	0.9912	0.9503

The predictive performance of the classifiers evaluated was significantly enhanced by using Grid Search Cross-Validation. The hyperparameter optimization allowed each model to run in the optimal setting, minimizing the effect of any arbitrary selection of the parameters on the experimental results. Support Vector Machine model was the one that greatly improved its classification performance due to the optimization of the kernel function and the regularization parameter input to the model, among the evaluated models. In the same fashion, the depth of trees and the number of trees used in ensembles were optimized to enhance the stability and predictive power of the Random Forest and XGBoost classifiers. The optimized XGBoost model showed excellent learning ability and good generalization ability for each validation fold. From the findings, the success of career prediction systems that use machine learning depends on the selection of appropriate hyperparameters. If the model does not perform optimally, the performance may be underestimated or inconsistent. Thus, the use of Grid Search along with Stratified 10-Fold Cross-Validation helped to achieve robustness and reliability of the proposed career prediction framework.

C. Confusion Matrix Analysis

To gain a better insight into the behavior of classifiers other than aggregate performance measures, the confusion matrix analysis was carried out for the best model among the classifiers which is Support Vector Machine (SVM) as shown in Fig. 4. By observing the confusion matrix of the SVM (Support Vector Machine) classifier, it would be inferred that the SVM classifier performed well in the class of the five career types. If you look at the SVM (Support Vector Machine) confusion matrix, one will see that the SVM classified the classes well for the five career types. Overall, 1,147 records out of 1190 were classified correctly, with an overall accuracy rate of almost 96.37%. The attributes that were more clearly distinguishable were those that fell under the category of the Engineer class, which had the most correct classifications. Some minor misclassifications occurred between Director and Journalism, likely because there are characteristics in both professions that have a communication orientation and leadership characteristics. Likewise, there was only a little bit of confusion between Engineer and Pharmacist, which may be the reason since both share analytical and subject-oriented feature patterns. The overall trend of the confusion matrix is helpful to identify the robustness of the SVM classifier in the balanced multiclass student career prediction.

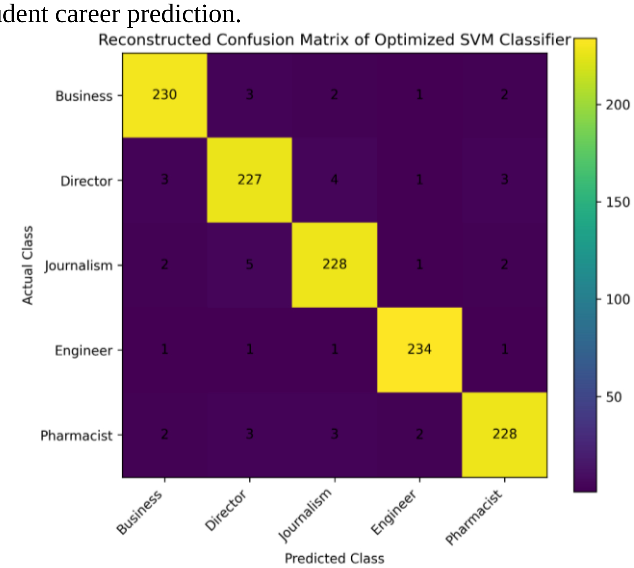


Fig. 4 Confusion Matrix of the Support Vector Machine Classifier

The results of the class-wise evaluation as shown in Table 8 shows that the Support Vector Machine (SVM) classifier has consistently performed well in all five categories of career. The model was able to identify students with strong analytical aptitude, problem-solving ability, and technical interests with the highest precision and recall values for the Engineer class. The classification performance of the two other categories (business and Pharmacist) was also very good, indicating a good separation from other career domains. A little bit less performance was noted in the classes for Director and Journalism. This can be attributed to the overlapping characteristics that encompass language proficiency, leadership tendencies, creativity and communication skills. However, the classification results showed that the classification accuracy is always high in all categories, proving that the proposed multidimensional feature representation is able to capture the unique features of the various career paths. This fair performance of classes further indicates that the model has no bias towards any category of careers, and its predictive ability is consistent in various classes of students.

TABLE VIII
Class-Wise Performance of SVM Model

Career Category	Precision	Recall	F1-Score
Business	0.95	0.98	0.96
Director	0.91	0.94	0.92
Journalism	0.92	0.95	0.93
Engineer	0.97	0.99	0.98
Pharmacist	0.93	0.97	0.95
Macro Average	0.94	0.97	0.95

The superior performance achieved by the proposed framework can be attributed to the use of multidimensional student profiling rather than exclusive reliance on academic indicators. Traditional career prediction systems frequently utilize examination scores and aptitude assessments as primary predictors. However, career decisions are influenced by multiple factors including interests, communication abilities, hobbies, and cognitive skills. By incorporating these characteristics into the prediction process, the proposed framework captures a broader representation of student potential and career inclination. Furthermore, the application of hyperparameter optimization and stratified cross-validation contributed to improved model generalization and reduced the risk of overfitting. The comparative results therefore suggest that multidimensional student attributes play an important role in improving the accuracy and reliability of machine learning-based career prediction systems.

D. Statistical Significance Testing

To assess if the difference in predictive performance of the machine learning models evaluated were statistically significant, paired Student's t-tests were performed on the Accuracy and F1-Score performance of the machine learning models in the Stratified 10-Fold Cross-Validation experiment. The paired t-test is used to determine if differences between two classifiers are due to random variation or are real differences in predictive ability. The null hypothesis (H0) has the hypothesis that the two models that are being compared do not have a statistically significant difference in performance, while the alternative (H1) has the hypothesis that a performance difference is present. The following significance level was applied to the analysis throughout ($\alpha = 0.05$). The null hypothesis was rejected when the p-value was < 0.05.

The statistical results shown in Table 9 should thus be read in conjunction with the above cross-validation performance. The mean cross-validation accuracy of SVM was 96.37% and F1-score was 0.9503, which was the highest amongst all the methods, followed by having XGBoost and Random Forest, which produced closely competitive results. It is also possible to assess this by using statistical testing, which will determine whether the differences observed are large enough across the validation folds as would be expected. Therefore, in this study, the selection of models is done based on the overall evidence of the mean cross-validation performance, class-wise predictive performance and based on statistical comparison and not on a single performance measure.

TABLE IX
Paired Student's T-test Results for Accuracy and F1-score

Model comparison	Metric	t-statistic	p-value	Statistical interpretation
SVM vs Random Forest	Accuracy	−238.83	2.01×10^{-18}	Significant
SVM vs XGBoost	Accuracy	−319.06	1.49×10^{-19}	Significant
SVM vs Decision Tree	Accuracy	−486.28	3.35×10^{-21}	Significant
SVM vs AdaBoost	Accuracy	−642.14	2.74×10^{-22}	Significant
SVM vs Random Forest	F1-score	−333.22	1.00×10^{-19}	Significant
SVM vs XGBoost	F1-score	−343.44	7.66×10^{-20}	Significant
SVM vs Decision Tree	F1-score	−472.53	4.33×10^{-21}	Significant
SVM vs AdaBoost	F1-score	−289.59	3.55×10^{-19}	Significant

To supplement the mean cross-validation performance, and to check whether observed difference between the classifiers is consistent over the validation folds, statistical testing was performed. SVM model obtained the highest mean cross validation accuracy (96.37%) and F1-score (0.9503) followed by XGBoost and Random Forest. Thus, the paired comparisons were confined to SVM and the four classifiers that were compared in the main experiment. The statistical results shown in Table 9 offers another evaluation of the consistency of the observed performance difference between the folds. The results should not be treated as an individual ranking of classifiers but in conjunction with the cross-validation results.

E. Comparative Analysis with Existing Studies

A representative sample of studies related to the prediction of career and educational outcomes are compared with the present framework in Table 10. The studies vary significantly in their prediction aim, the student groups they address, the variables used as input, the classification models employed, and the methods used for validation. For instance, Isra et al. (2021) forecasted the waiting period of university graduates to find employment based on their academic performance, while Nguyen et al. (2023) leveraged XGBoost to forecast career-paths based on the academic results of university students. Bhaskaran et al. (2024) studied student career prediction using student performance and Karale et al. (2022) studied student-performance prediction, which is a similar but separate problem. Thus, the performance values given in these studies cannot be considered as a direct ranking. This study is unique in several ways, most notably in the use of a multidimensional student profile that includes both academic and non-academic characteristics of students, and in the use of stratified 10-fold cross-validation for the evaluation of five-class career prediction.

TABLE X
Contextual Comparisons with Related Career and Student Prediction Studies

Study	Primary prediction task	Main input/data basis	Approach	Reported performance	Relevance/comparability
Isra et al. (2021)	Alumni employment waiting-period prediction	Academic performance	ANN	88.51%	Related career-outcome study; different prediction target
Nguyen et al. (2023)	Career-path prediction	University academic results	XGBoost	90.4%	Directly related career-prediction task, but uses different data and evaluation protocol
Ashalakshmi and Hemalatha (2023)	Career prediction for twelfth-grade students	Student-related educational data	CNN	94.2%	Related career-prediction study using a different modelling framework
Bhaskaran et al. (2024)	Student career prediction	Student performance	Machine-learning classifiers	89.7%	Directly related career-prediction study, but experimental setting differs
Karale et al. (2022)	Student-performance prediction	Student/academic information	AI/ML methods	92.1%	Related educational prediction; different target task
Present study	Five-class career prediction	Academic + non-academic multidimensional student profile	SVM	96.37% mean CV accuracy	Stratified 10-fold cross-validation; present experimental setting

F. Findings and Interpretations

These results are useful in endorsing the use of machine learning models in creating dynamic and adaptive career guidance systems. Compared to the traditional academic and psychometric methods that rely heavily on static measures such as test scores or aptitude tests, the proposed ML based system proposed an holistic and personalized recommendation framework. However, the traditional methods are not adaptable and do not consider behavioral traits, personal interests, or extracurricular achievements. Unlike that, ML models learn from multiple skewed data and update the recommendation over time. Studies have confirmed this advantage. For example, based on Guntupalli et al. (2024), traditional approaches miss the mark on accurately and interest aligned career recommendation whereas SVM and Random Forest models provide it better. Lv (2024) also reported 30% improvement in alignment of machine leaning based recommendations vs. interest assessments but noted that both AI based recommendations support higher personalization and satisfaction. In practice, these machine learning models can be turned into powerful tools for educational counseling systems, which can give students an idea on where they are headed based on how intelligent they are, so to speak, academic background, interests, skills and behavioral traits. By including explainable models (e.g. decision trees) in the system, it also improves the system's trust and transparency in novel ways, hence it is expected to be successful in deployment on schools and university, rather than on data center. The results show that machine learning models outperform traditional methods for career prediction by producing the data driven and interpretable guidance to power students in the careers of an evolving landscape.

G. Ethical Considerations, Bias, and Fairness

While machine learning tools for career predictions have the promise to have profound impacts on educational decision-making, there are several ethical questions to be addressed before implementing such tools in practice. A responsible use of predictive models is critical for a fair and transparent use of career recommendation systems, which will directly shape students' educational and career decisions. The composition of the set of data used to train the model can be a source of bias. This study is based on a dataset that includes demographic, socio-economic and educational data including gender, social class, family income and parental education level. These qualities can help predict performance but can also inadvertently bias the prediction if there are unequal numbers of people, or if a specific group of people have been linked to a specific career stream throughout history. Thus, the predictions that can be produced with machine learning models may be partly based on the patterns found in the training data rather than only based on actual student capabilities.

To address this risk, stratified sampling and cross validation methods were used throughout the modeling process to ensure that there was a balanced representation across career categories and to minimize sampling bias. In addition, model evaluation was conducted using several performance measures and statistical significance testing to make sure that the results from the models were not caused by the small part of the data. A second key point to fairness in career guidance is raised by the varying skills of teachers and students. A second factor of fairness in career guidance is the different abilities of teachers and students. The proposed framework is designed as a decision support tool, and not as a stand-alone decision-making system. Recommendations for careers made by the model should be used in conjunction with professional counselling and student individual judgment. Human oversight is still needed to provide context and interpretation to the recommendations; context being the student's aspirations and personal situation and the interpretation being the developing interests. Another key need for educational AI systems is transparency.

Future research will aim to enhance the interpretability of the system and to give students, teachers and counsellors a better understanding of the factors that influence career suggestions by incorporating Explainable Artificial Intelligence (XAI) techniques, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations).

Future work should also explore fairness-aware machine learning methods and metrics for the evaluation of bias such as demographic parity, equal opportunity, and disparate impact analysis. These approaches will help to determine on a systematic basis if the predictive results are equitable across socio-economic and demographic groups.

CONCLUSION

In this research, a multi-dimensional machine learning system for student career prediction was presented that combines both academic and non-academic factors such as personal interests, hobbies, language skills, favourite subjects, and problem-solving skills. The proposed career recommendation framework is based on a student-centric profiling approach, as opposed to the traditional model which uses academic performance and psychometric testing to determine career recommendations.

The balanced dataset of 1190 student records and 26 predictive attributes were used in an experimental analysis. A thorough validation framework was utilized, including a stratified 10-fold cross-validation and Grid Search optimization of the machine learning classifiers Random Forest, XGBoost, Decision Tree, AdaBoost and Support Vector Machine (SVM). All the results showed very good predictive accuracy with Support Vector Machine having the highest classification accuracy of 96.37%, recall of 99.12% and F1-score of 95.03%.

The results of further analysis using confusion matrix evaluation and class-wise performance assessment also showed that the proposed framework had a good prediction ability in all five career categories. The investigation at the feature level revealed the significance of the attributes related to cognitive, academic preference and interest in predicting career outcome. A comparative analysis with previous works also found that the proposed multidimensional profiling method gave competitive performance and offered more information about student characteristics.

The results substantiate machine learning as a feasible decision-support tool in educational career counselling. The framework can help schools, counsellors and students to make informed career decisions through the comprehensive assessment of student capabilities and preferences. Future research will include adding more demographic and socioeconomic characteristics to expand the dataset, exploring deep learning architectures for the prediction of careers, integrating real-time labor market information, and also validating the framework with institution-specific datasets to further enhance the generalisation ability and applicability of the framework.

STATEMENT AND DECLARATIONS

H. Conflict of Interest

The authors declare that there is no conflict of interest associated with this publication.

I. Author Contributions

All authors contributed equally to the conception, design, data analysis, and preparation of the manuscript, and have approved the final version for publication.

J. Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

K. Data Availability Statement

The dataset used in this study is publicly available in the Kaggle repository under the title “Student Career Prediction.” The dataset can be accessed through the Kaggle platform for verification and reproducibility of the reported analysis.

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

Use the singular heading even if you have many acknowledgments. **Leave this section as is for the double-blind review process.**

REFERENCES

Al-Ahmad, B. I., Alzaqebah, A., Alkhawaldeh, R., Al-Zoubi, A. M., Lo, H., & Ali, A. (2025). Predicting academic performance for students’ university: case study from Saint Cloud State University. *PeerJ Computer Science*, *11*, e3087. https://doi.org/10.7717/peerj-cs.3087

Alsbou, M. K. K., & Alsaraireh, R. A. I. (2024). Data-driven decision-making in education: Leveraging AI for school improvement. *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, 1–6. https://doi.org/10.1109/ICKECS61492.2024.10616616

Altukhi, Z. M., & Pradhan, S. (2026). Integrating Explainable AI in Education: Challenges and Advancements–A Systematic Review. https://doi.org/10.17705/1pais.18101

Ashalakshmi, R., & Hemalatha, S. (2023). Implementing convolutional neural networks for career prediction: A case study on twelfth grade students. *2023 International Conference on Emerging Research in Computational Science (ICERCS)*, 1–6. https://doi.org/10.1109/ICERCS57948.2023.10433993

Asriadi, M., & Setiawati, F. (2023). Examining the psychometric properties of the career commitment instrument through classical test theory and the graded response model. *Journal of Education Research and Evaluation*. https://doi.org/10.23887/jere.v7i3.59619

Balan, S. R. V., Senthilnathan, T., Gobinath, R., &Gondkar, R. (2023). Assessing academic performance using ensemble machine learning models. *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, 917–924.

https://doi.org/10.1109/ICCPCT58313.2023.10245216

Barbadekar, A. B., Bhagwan, V., Thaploo, A., & Tyagi, A. (2024). Career Navigator: Your AI career counselor. *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, 1–6. https://doi.org/10.1109/IDICAIEI61867.2024.10842724

Bhaskaran, V. M., S. V., & At, S. (2024). Student career prediction based on student performance using machine learning. *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 1212–1217. https://doi.org/10.1109/ICACCS60874.2024.10716997

Borna, M. R., Saadat, H., Hojjati, A. T., & Akbari, E. (2024). Analyzing click data with AI: Implications for student performance prediction and learning assessment. *Frontiers in Education*. https://doi.org/10.3389/feduc.2024.1421479

Cheng, Y., & Liang, Y. (2023). The development of artificial intelligence in career initiation education and implications for China. *European Journal of Artificial Intelligence and Machine Learning*. https://doi.org/10.24018/ejai.2023.2.4.32

Chung, C., & Shamir, L. (2021). Introducing machine learning with Scratch and robots as a pilot program for K-12 computer science education. *International Journal of Learning and Teaching*, 7(3), 181–186. https://doi.org/10.18178/ijlt.7.3.181-186

Coman, C., Gherheș, V., Bucs, A., Coman, E., Victoria-Rodica, C., Bódi, D. C., & Șerbănescu, R. M. (2026). Navigating opportunities and challenges of generative AI in higher education: towards responsible, equitable, and human-centered integration. *Frontiers in Artificial Intelligence*, *9*, 1750978. https://doi.org/10.3389/frai.2026.1750978

Dhamdhere, A. Y., Bhosale, N. R., Joshi, A. C., Kale, V. V., Chavan, A. A., &Pokale, M. S. (2024). *ASPIREAI: AI Chatbot for Career Guidance.* International Journal of Innovative Research in Engineering. https://www.doi.org/10.59256/ijire.20240502033

Duan, J., & Wu, S. (2024). Beyond traditional pathways: Leveraging generative AI for dynamic career planning in vocational education. *International Journal of New Developments in Education*. https://doi.org/10.25236/ijnde.2024.060205

Faruque, S. H., Khushbu, S. A., & Akter, S. (2025). Decision support system to reveal future career over students' survey using explainable AI. *Education and Information Technologies*, *30*(10), 14471-14509. https://doi.org/10.1007/s10639-025-13361-7

Govindarajan, V., Shah, A. H., Morell, A. A., Borowy, V., Ingle, S. M., Valdez, M. J. M., Rivas, S. R., Eichberg, D., Luther, E., Lu, V. M., Heiss, J. D., Komotar, R., & Levi, A. (2023). Predicting academic career placement via development of novel intra-residency metrics. *World Neurosurgery*. https://doi.org/10.1016/j.wneu.2023.02.080

Guntupalli, U. G. S., Pandala, M. L., Veeranki, D. T., & Kumbha, P. (2024). *Carrer Compass: A Career Path Recommender using Machine Learning.* 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC). https://doi.org/10.1109/ICESC60852.2024.10689897

Hamdi, M. (2024). How AI is transforming and shaping the future of education. *2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES)*, 115–116. https://doi.org/10.1109/ines63318.2024.10629089

Haque, R., Goh, H., Ting, C.-Y., Quek, A., & Hasan, M. R. (2024). Imbalanced data classification using oversampling and automatic feature selection methods for undergraduate student career prediction. *2024 13th International Conference on Educational and Information Technology (ICEIT)*, 247–252. https://doi.org/10.1109/ICEIT61397.2024.10540781

Huang, Y., Ling, L., & Liu, Y. (2024). Integrating AI into biology education: Classifying edible and poisonous mushrooms using machine learning. *2024 5th International Conference on Computer Engineering and Intelligent Control (ICCEIC)*, 346–352. https://doi.org/10.1109/ICCEIC64099.2024.10775456

Ilic, M., Kekovic, G., Mikić, V., Mangaroska, K., Kopanja, L., & Vesin, B. (2024). Predicting student performance in a programming tutoring system using AI and filtering techniques. *IEEE Transactions on Learning Technologies*, *17*, 1931–1945. https://doi.org/10.1109/TLT.2024.3431473

Isra, A., Zainuddin, Z., & Tahir, Z. (2021). The alumni's career prediction based on academic performances utilizing neural network algorithm. *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)*, 132–136. https://doi.org/10.1109/ICERA53111.2021.9538735

Karale, A., Narlawar, A., Bhujba, B., &Bharit, S. (2022). Student performance prediction using AI and ML. *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2022.44032

Khalafallah, A. M., Jimenez, A. E., Horowitz, M. A., Camp, S., Witham, T., Huang, J., & Mukherjee, D. (2020). An online calculator for predicting academic career trajectory in neurosurgery in the United States. *World Neurosurgery*. https://doi.org/10.1016/j.wneu.2020.09.161

Kler, R., Singh, G., Chaudhary, N., Abdullaev, B., Bhandwal, M., & Ather, D. (2024). Predicting career transitions through insights from academic background and workforce dynamics using machine learning. *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 134–138. https://doi.org/10.1109/ICTACS62700.2024.10840625

Kosaraju, D. (2024). The role of AI and machine learning in the future of remote education. *International Journal* *of Research and Review*. https://doi.org/10.52403/ijrr.20220263

Li, Y., & Zhao, L. (2025). Application of support vector machine algorithm in predicting the career development path of college students. *International Journal of High Speed Electronics and Systems*. https://doi.org/10.1142/S012915642540230X

Lv, J. (2024). Stacked Ranking Feature Cluster Machine Learning (Srfcml): A Novel Method of Career Planning of College Students Based on Career Interest Assessment and Machine Learning. *Journal of Electrical Systems*. https://doi.org/10.52783/jes.1706.

Mahawar, K. (2023). Dimensionality reduction using feature selection techniques on EDM for student academic performance prediction. *International Journal on Recent and Innovation Trends in Computing and Communication*.

Manganello, F., Rasca, E., Villa, A., Maddalena, A., & Boccuzzi, G. (2025, May). Machine learning for post-diploma educational and career guidance: a scoping review in AI-driven decision support systems. In *Frontiers in Education* (Vol. 10, p. 1578979). Frontiers Media SA. https://doi.org/10.3389/feduc.2025.1578979

Mayahi, K. A., & Al-Bahri, M. (2020). Machine learning based predicting student academic success. *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 264–268. https://doi.org/10.1109/ICUMT51630.2020.9222435

Mohamad, N., Ahmad, N. B. H., Jawawi, D. N. A., & Hashim, S. (2020). Feature engineering for predicting MOOC performance. *IOP Conference Series: Materials Science and Engineering*, 884(1), 012070. https://doi.org/10.1088/1757-899X/884/1/012070

Muhammad, R., Patriana, P., Yusrain, Y., Astaman, A., & Manja, M. (2024). *Counselling Career with Artificial Intelligence: A Systematic Review.* GUIDENA: JurnalIlmu Pendidikan, Psikologi, Bimbingan dan Konseling. http://dx.doi.org/10.24127/gdn.v14i1.9357

Muñoz-Carpio, J. C., Jan, Z., & Saavedra, A. (2021). Machine learning for learning personalization to enhance student academic performance. *In Proceedings*, 88–99.

Nguyen, H. Q., Nguyen, D. D. K., Le, T. D., Mai, A., & Huynh, K. T. (2023). Career path prediction using XGBoost model and students' academic results. *CTU Journal of Innovation and Sustainable Development*. https://doi.org/10.22144/ctujoisd.2023.036

Nisarta, A. (2024). Comparative study of the artificial intelligence of the M Ed and B Ed students. *RESEARCH REVIEW International Journal of Multidisciplinary*, 9(1). https://doi.org/10.31305/rrijm.2024.v09.n01.032

Pali, K., & Tiwari, L. (2024). *Predictive Learning and Career Path Using Artificial Intelligence.* 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0. https://doi.org/10.1109/OTCON60325.2024.10688178

Pavithra, K., & Ramkumar, M. (2026, January). Advancements in explainable artificial intelligence: Bridging the gap between machine learning models and human understanding. In *AIP Conference Proceedings* (Vol. 3345, No. 1, p. 020362). AIP Publishing LLC. https://doi.org/10.1063/5.0298407

Preethi, K., & Ramakrishnan, M. (2024). Advanced hybrid feature selection using harvest algorithm, convolutional neural networks and CFS. *Nanotechnology Perceptions*. https://doi.org/10.62441/nano-ntp.v20is14.56

Sarmurzin, Y., Azhenov, A., Kudysheva, A., Ualikhanova, B., Saparbekova, I., & Besbayev, A. (2026, March). Implementation of AI in career counselling for university students: a systematic review. In *Frontiers in Education* (Vol. 11, p. 1787689). Frontiers Media SA. https://doi.org/10.3389/feduc.2026.1787689

Setiadi, H., Larasati, I. P., Suryani, E., Wardani, D., Cahyono, H. D., Wijayanto, A., &Doewes, A. (2024). Comparing correlation-based feature selection and symmetrical uncertainty for student dropout prediction. *Jurnal RESTI (RekayasaSistem dan TeknologiInformasi)*. https://doi.org/10.29207/resti.v8i4.5911

Student Career Prediction Dataset, https://www.kaggle.com/datasets/suhailaelnemr/student-career-prediction

Sudjitjoon, W., Hengpraprohm, S., &Hengpraprohm, K. (2022). AI learning modules for elementary students. *International Journal of Health Sciences*, 6(4). https://doi.org/10.53730/ijhs.v6ns4.11859

Usman, M. M., Owolabi, O., & Ajibola, A. A. (2020). Feature selection: It importance in performance prediction. *IJESC*, *10*(5), 25625-25632.

Vimarsha, K., Prakash, S., Krinkin, K., &Shichkina, Y. A. (2024). Student performance prediction: A co-evolutionary hybrid intelligence model. *Procedia Computer Science*. https://doi.org/10.1016/j.procs.2024.04.043

Wang, H. Z., & Zhang, W. (2024). Improving automated feature selection in cancer genomics using reinforcement learning. *2024 IEEE International Conference on Big Data (BigData)*, 7428–7432. https://doi.org/10.1109/BigData62323.2024.10825966

Wen, K., & Zhou, D. (2025). Predictive career guidance and entrepreneurial development for university students using artificial intelligence and machine learning. *Journal of Computational Methods in Sciences and Engineering*, *25*(4), 3080-3092. https://doi.org/10.1177/14727978251318801

Zaffar, M., Hashmani, M., Habib, R., Quraishi, K. S., Irfan, M., Alqhtani, S. M., & Hamdi, M. (2022). A hybrid feature selection framework for predicting students' performance. *Computers, Materials & Continua*. https://doi.org/10.32604/cmc.2022.018295

Zhang, F., Xing, W., Li, C., & Jiang, Y. (2026). Fair AI in educational predictions: A multi-group fairness approach using reinforcement learning. *The Internet and Higher Education*, 101074. https://doi.org/10.1016/j.iheduc.2026.101074