APPLYING MACHINE LEARNING ALGORITHMS FOR CAREER ORIENTATION PREDICTION IN MIDDLE SCHOOL STUDENTS

Hoang Duy Kien[1], Vu Van Thanh[1], Nguyen Quoc Khanh[1], Bui Van Dat[1], Nguyen Thi Kim Son[1,*], Phan Tien Phuong[1]

DOI: https://doi.org/10.57001/huih5804.2026.075

ABSTRACT

In the context of a constantly evolving labor market, career counseling at the middle school level plays a pivotal role in guiding students' early future orientations. From a computer science perspective, this problem can be formulated as a multi-class classification task. The primary challenge lies in the sheer volume, complexity, and heterogeneity of career orientation data - encompassing multidimensional features related to academic performance, psychological profiles, and demographics - which necessitate rigorous data preprocessing procedures and model selection strategies. This study proposes a systematic approach to data processing and the evaluation of machine learning architectures, aiming to optimize robustness on career orientation data, which predominantly exists in tabular form. As a primary contribution, this research introduces a comprehensive empirical evaluation framework to identify the optimal machine learning algorithm tailored to the specificities of career orientation educational data. This establishes a solid foundation for the development of potentially reliable decision-support systems. Experiments conducted on the standardized VCS-024 dataset demonstrate that the XGBoost model achieves an accuracy of up to 92%, outperforming other baseline models. These results affirm the robust performance of ensemble learning methods, particularly XGBoost, in mining tabular career data. Consequently, this study provides robust groundwork for developing intelligent decision support systems within school counseling environments.

Keywords: *Machine Learning, Multi-class Classification, Ensemble Learning, XGBoost, Career Counseling, Tabular Data.*

[1]School of Information and Communications Technology, Hanoi University of Industry, Vietnam

[*]Email: sonntk@haui.edu.vn

Received: 20/01/2026

Revised: 23/3/2026

Accepted: 30/3/2026

1. INTRODUCTION

In the era of Industry 4.0, the proliferation of Big Data has driven profound transformations across various domains. However, within the educational sector, the massive volumes of data collected daily remain significantly underutilized, particularly in the realm of career counseling for middle school students. Against the backdrop of a volatile labor market and an increasingly prevalent career orientation crisis among the youth, early guidance assumes paramount importance. It not only empowers students to comprehend their core competencies and mitigate peer pressure but also establishes a robust foundation for them to construct tailored learning pathways and select appropriate subject combinations from the outset.

From a data science perspective, educational datasets utilized for career guidance predominantly exist in tabular formats. They are characterized by complex correlations among features and severe class imbalance stemming from real-world disparities in occupational distribution. Although traditional and modern machine learning classification algorithms - such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and XGBoost - have proven their efficacy in processing tabular data, their performance varies significantly depending on the specific distributional characteristics of each dataset. Consequently, investigating, evaluating, and identifying an optimal approach not only addresses the immediate problem but also paves the way for strategic research directions concerning AI-integrated school counseling systems.

Driven by the aforementioned practical challenges, this research establishes a core objective: to identify and optimize an appropriate machine learning architecture for mining career orientation data, thereby laying the groundwork for future intelligent Decision Support Systems (DSS).

To address existing gaps in Educational Data Mining, this study makes several key contributions. First, we

propose a systematic educational data processing framework specifically tailored for high-dimensional, noisy, and heterogeneous datasets. This framework incorporates techniques for data cleaning, feature standardization, categorical encoding, and the decomposition of composite multiple-response attributes to maximize information extraction. Second, moving beyond conventional fixed data partitioning, we introduce a meta-heuristic optimization pipeline. By coupling a Stratified 10-Fold Cross-Validation strategy with the Grey Wolf Optimizer (GWO) algorithm, our approach explicitly prioritizes model stability and generalizability over standard empirical accuracy. Third, we conduct a comprehensive empirical evaluation across diverse machine learning families- including linear models, distance-based algorithms, neural networks, and ensemble methods. Utilizing a multi-metric evaluation strategy (combining Accuracy, Macro-Precision, Macro-Recall, Macro F1-Score, and AUC), we empirically demonstrate the superior performance of Ensemble Learning, particularly XGBoost, in handling severe class imbalances and complex socio-psychological factors. Finally, this research presents a practical application by utilizing a real-world career orientation dataset of middle school students in Vietnam (VCS-024). By integrating Educational Data Mining with machine learning optimization, this study establishes a stable algorithmic foundation for the development of future intelligent career counseling systems.

The remainder of this paper is organized as follows. Section 2 reviews related studies in Educational Data Mining and career prediction. Section 3 presents the proposed methodology and optimization framework. Section 4 discusses experimental results and performance evaluation. Finally, Section 5 concludes the study and outlines future research directions.

2. RELATED WORK

In recent years, Educational Data Mining (EDM) has witnessed a significant paradigm shift from traditional statistical analyses to advanced predictive modeling. Particularly at the middle school level - a pivotal stage for cultivating career awareness - the application of machine learning algorithms demands not only high accuracy but also the capability to process highly noisy, multidimensional datasets.

Historically, in educational research, traditional statistical methods such as Logistic Regression and Factor Analysis have served as foundational tools. However, Bzdok et al. [1] conducted a landmark analysis elucidating the fundamental differences between statistics and machine learning. The authors highlighted that traditional statistics predominantly focuses on inference aiming to uncover causal relationships based on stringent assumptions regarding data distributions. Conversely, Machine Learning prioritizes prediction and can autonomously capture complex non-linear interactions that linear models frequently overlook.

This transition is prominently exemplified by the work of Trujillo et al. [2]. In a comprehensive Systematic Literature Review (SLR), the authors examined hundreds of studies published between 2015 and 2025, confirming that Machine Learning is steadily superseding traditional psychometric approaches. The study highlighted that despite ML's high performance, the majority of current research still encounters barriers related to model interpretability and frequently neglects the nuanced qualitative attributes of students.

To bridge the educational theory gap, Ajgaonkar et al. [3] developed the EduKrishnaa system. Rather than relying solely on academic scores, the authors integrated the Multiple Intelligences theory to construct student competency profiles encompassing eight distinct dimensions. By employing a multi-class classification algorithm, the study demonstrated that incorporating psychological theories into AI renders the counseling system more human-centric and practically relevant. Concurrently, Muhammad et al. [4] proposed an end-to-end predictive framework centered on standardizing the behavioral feature extraction process to reduce subjective human bias in predicting student career pathways. Geared towards actionability, Balasubramanian [5] advanced this domain by introducing a hybrid machine learning approach for dynamic, personalized pathway recommendations that align students' current competencies with broader labor market demands.

Real-world educational datasets frequently suffer from severe class imbalance. Sarwat et al. [6] addressed this challenge by integrating Conditional Generative Adversarial Networks (CGANs) to generate synthetic data for minority occupational classes. This approach effectively balances the training set and enhances the performance of the Deep SVM model in predicting academic outcomes.

Nevertheless, concerning tabular data characterized by a high-dimensional feature space, tree-based models

continue to maintain strong performance. Mounika and Persis [7] conducted a comparative study of various machine learning algorithms on educational data. Their findings revealed that certain classifiers provide more stable and reliable predictions when handling student performance data, which inherently harbors inconsistent variables and domain-specific noise.

Among these algorithms, XGBoost (Extreme Gradient Boosting) has emerged as a robust approach. Yan [8] demonstrated the efficacy of XGBoost through a macro-perspective study on student performance prediction, reporting that this algorithm not only excels in accuracy but also inherently handles complex data structures efficiently. The technical foundation underlying this superiority was comprehensively detailed by Chen and Guestrin [9] through the *sparsity-aware split finding* technique, which enables the model to identify optimal split points even in the presence of sparse data. Notably, Shwartz-Ziv and Armon [10] conducted rigorous benchmarking and affirmed that for tabular data, gradient boosted tree models like XGBoost consistently maintain superior performance over complex Deep Learning architectures such as the Multi-layer Perceptron (MLP), owing to their superior ability to capture the underlying feature structures of tabular data.

Although prior studies [2] to [8] have achieved substantial progress, a distinct research gap persists. The majority of existing literature predominantly focuses on university students or the active workforce, demographics characterized by well-defined and structured data. Conversely, the middle school phase in Vietnam - characterized by highly variable sample cohorts and profound influences from local cultural and familial factors - remains inadequately explored using modern algorithmic approaches. Furthermore, applying conventional statistical models to datasets encompassing as many as 248 features frequently induces overfitting or fails to capture underlying latent interactions.

This study aims to bridge this gap by proposing an optimized predictive model based on XGBoost. Unlike traditional approaches predicated on simplistic linear assumptions, this work focuses on harnessing the power of XGBoost in processing high-dimensional and severely imbalanced tabular data. The findings yield a highly reliable career counseling instrument, effectively facilitating the post-middle school tracking and streaming of students within the contemporary educational landscape of Vietnam.

3. PROPOSED METHOD

The overall methodological workflow of this study is illustrated in Figure 1. The proposed pipeline consists of four main phases: (i) data preprocessing and cleaning, (ii) exploratory data analysis, (iii) feature extraction, standardization, and encoding, and (iv) model training, validation, and optimization. This structure was designed to ensure that the raw VCS-024 survey dataset was systematically transformed into a clean, consistent, and machine-readable representation before being used for multi-class career orientation prediction.

A rigorous data preprocessing pipeline was implemented to transform the raw dataset - characterized by diverse data types ranging from Vietnamese text strings to numerical data, as well as inconsistencies inherent in manual data entry - into an optimal standardized format for the predictive model. The initial phase concentrated on noise elimination and dimensionality reduction through stringent feature screening. Specifically, 153 entirely empty columns, along with identifiers lacking statistical significance such as survey timestamps and specific middle school names, were completely discarded. This data pruning process not only mitigates systemic noise but also prevents overfitting, ensuring that the model focuses exclusively on features with genuine predictive value.

Before feature removal, an initial data inspection step was conducted to examine the dataset size, data types, missing-value ratios, and duplicated records. Completely duplicated rows were removed to avoid repeated observations affecting the learning process and distorting the empirical distribution of the dataset. This inspection step also provided the basis for identifying irrelevant, empty, or low-informative attributes before applying subsequent preprocessing operations.

After the removal of confounding factors, the data underwent further standardization and aggregation to address the open-ended nature of the survey questions. A prime example is the "Age" attribute, which initially appeared in various heterogeneous formats - such as birth years or age ranges freely inputted by users - and was subsequently standardized into a uniform age unit. Following this, the binning technique was applied to categorize these values into strategic demographic intervals; for instance, the 30-39 age bracket was encoded

into a single group identifier. This method stabilizes manually entered data and enhances the model's capability to detect characteristic trends corresponding to distinct stages of personal development.

Missing values were handled in a column-wise manner according to the available valid values in each attribute. In the implemented preprocessing procedure, numerical and categorical/object variables were not imputed using two separate strategies such as mean or median for numerical attributes and mode for categorical attributes. Instead, a unified most-frequent-value strategy was applied to each column: if a column contained valid observations, missing entries were replaced by the mode of that column. If no valid mode could be identified, such as in the case of a column with no usable values after inspection, the missing entries were filled with 0 as a fallback value. This strategy was adopted because the survey dataset mainly consists of categorical, ordinal, and encoded tabular attributes after standardization, while qualitative object-type responses were normalized and later converted into numerical representations before model training.

Concurrently with the standardization process, we executed Feature Decomposition to maximize the extraction of information from multiple-response questions. For attributes containing composite information - such as reasons for job transitions - the data was parsed into distinct binary features, each delineating a specific factor like excessive workload or task difficulty. This transformation of a composite variable into individual attributes empowers the algorithm to delve deeper into the nuanced interactions among influencing factors, thereby enhancing the granularity and interpretability of the career orientation prediction task.

Ultimately, the entirely refined and decomposed dataset was converted into a numerical format via the Label Encoding technique. Qualitative categorical values, such as gender or occupational clusters, were mapped to corresponding integers to ensure absolute compatibility with the mathematical architecture of the XGBoost algorithm. This final transformation serves as the concluding link in the preprocessing pipeline, enabling the model to effectively process non-linear relationships across the 248 features and generate precise predictions on the dataset comprising 1,067 real-world observations.

After categorical values were encoded, feature scaling was applied to ensure that numerical attributes were represented on a comparable scale. StandardScaler was used to standardize numerical feature values, thereby reducing the influence of differences in measurement ranges among attributes. Although tree-based ensemble models such as XGBoost are generally less sensitive to feature scale, this step was retained to maintain a consistent preprocessing pipeline across all comparative machine learning models, including distance-based and neural-network-based classifiers. An exploratory data analysis phase was also incorporated to analyze the structure of the VCS-024 dataset before model training. This phase focused on examining feature distributions, the proportion of categorical and numerical attributes, and the distribution of target labels across career orientation groups. The analysis revealed a clear class imbalance among the six target categories, which motivated the use of stratified validation and macro-averaged evaluation metrics in the subsequent experimental design.

To address career orientation prediction on a high-dimensional and severely imbalanced dataset, we employed XGBoost. Compared with conventional learning algorithms, XGBoost optimizes the objective in the function space using second-order approximation and regularization to reduce overfitting.

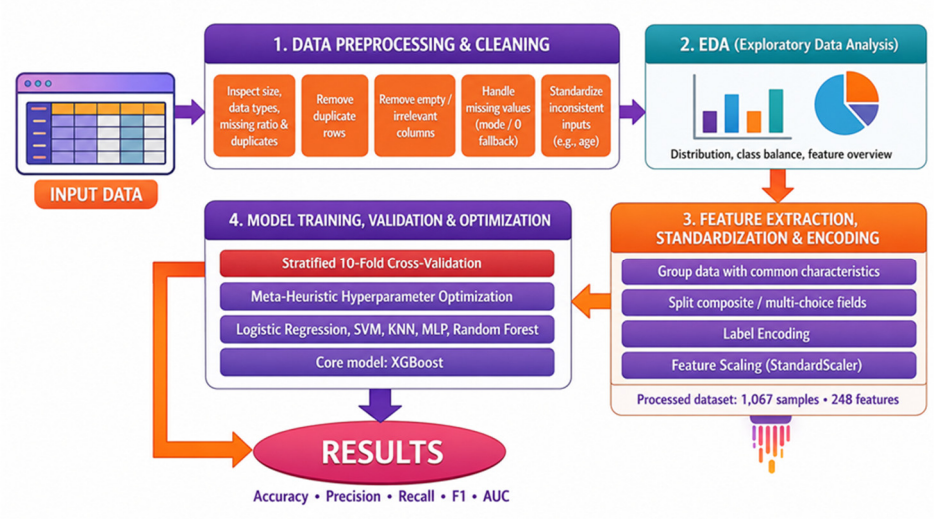

Figure 1. Overview of the research methodology and data processing pipeline

At iteration t, the objective function is defined as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \quad (1)$$

where $l(y_i, \hat{y}_i)$ is the loss function and $\Omega(f_k)$ is the regularization term of the k-th tree:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (2)$$

Here, γ penalizes the number of leaves T, while λ controls the L_2 regularization on leaf weights w_j. To accelerate optimization, XGBoost uses the second-order Taylor expansion:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \quad (3)$$

where g_i and h_i denote the first-order and second-order gradients of the loss function, respectively. The split quality is evaluated by the gain function:

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \quad (4)$$

The split with the highest gain is selected, and splits with negative gain are discarded.

In this study, XGBClassifier was configured with the multi:softprob objective for six-class career prediction. This setting allows the model to estimate the probability distribution over all classes. Early stopping was also applied to improve generalization and prevent overfitting.

Building upon this foundational architecture, to evaluate the predictive models and establish a fitness landscape for hyperparameter tuning, we replaced the conventional train-test split approach. Such conventional methods are susceptible to data variance, especially when modeling high-dimensional tabular data representing complex psychological and academic profiles. Instead, our methodology incorporates a Stratified 10-Fold Cross-Validation (10-fold CV) strategy. The preprocessed dataset is partitioned into K = 10 mutually exclusive subsets, preserving the class distribution of the target variable across all folds. During each iteration, one fold serves as the validation set while the remaining nine constitute the training set. The average performance across all ten iterations provides an objective estimation of the model's generalization capabilities.

Furthermore, the performance of machine learning architectures is contingent upon their hyperparameters. Given the non-differentiable nature of the hyperparameter search space, exhaustive methods such as Grid Search are computationally prohibitive. To address this limitation, we formulated hyperparameter tuning as a global optimization problem utilizing Meta-Heuristic algorithms. Specifically, the models were optimized using the Grey Wolf Optimizer (GWO) implemented via the mealpy library. In this continuous process, the GWO algorithm iteratively explores the search space to maximize the mean classification accuracy obtained from the integrated 10-fold CV process. The selection of GWO over other traditional evolutionary algorithms is primarily motivated by its mathematical modeling of the leadership hierarchy and cooperative hunting mechanism of grey wolves in nature. GWO maintains an adaptive balance between exploration (global search) and exploitation (local refinement) phases. This characteristic is particularly advantageous for hyperparameter optimization, as it ensures rapid convergence while mitigating the risk of entrapment in local optima—a frequent challenge when optimizing the non-convex, multidimensional parameter spaces of complex ensemble models like XGBoost. The complete computational pipeline of the proposed method is formalized in Algorithm 1:

Algorithm 1: XGBoost-based Predictive Model for Career Trends

Input:

- Raw dataset $\mathcal{D}_{raw}$ (1,067 samples, 307 features), target label y (6 career classes)

- Hyperparameter Search Space $\mathcal{S}$

- Number of folds $K = 10$, GWO parameters (Population size N, Max Iterations M)

Output:

- Optimal Hyperparameters P_{best}, Trained final model $\mathcal{M}_{final}$

Phase 1 & 2: Data Preprocessing, Feature Extraction and Encoding

1. $\mathcal{D} \leftarrow$ Remove_Duplicate_Rows($\mathcal{D}_{raw}$)
2. $\mathcal{D} \leftarrow$ Drop_Columns($\mathcal{D}$,153 empty and noisy columns)
3. **for each** column $c \in \mathcal{D}$ **do**
4. **if** c contains missing values **then** Fill_NaN(c,mode_val or 0)
5. **end for**
6. $\mathcal{D} \leftarrow$ Group_Features($\mathcal{D}$,Age)
7. $\mathcal{D} \leftarrow$ Split_Complex_Features($\mathcal{D}$,Job Change Reasons)

8. **for each** categorical column $c \in \mathcal{D}$ **do**
9. $\mathcal{D}[c] \leftarrow$ LabelEncoding($\mathcal{D}[c]$)
10. **end for**
11. $X, y \leftarrow$ Split_Features_Target($\mathcal{D}$)
12. scaler $\leftarrow$ StandardScaler()
13. $X_{scaled} \leftarrow$ scaler.Fit_Transform(X)

Phase 3: Meta-Heuristic Optimization (GWO) with 10-Fold CV

14. Initialize a population of N wolves randomly within $\mathcal{S}$
15. $P_{best} \leftarrow \emptyset$,MaxFitness $\leftarrow -\infty$
16. **for** $iter = 1$ **to** M **do**
17. **for each** candidate P_i in Population **do**
18. Total_Score $\leftarrow 0$
19. Partition $\{X_{scaled}, y\}$ into K stratified folds $\{F_1, F_2, \ldots, F_K\}$
20. **for** $k = 1$ **to** K **do**
21. Train_Set $\leftarrow \{X_{scaled}, y\} \setminus F_k$
22. Validation_Set $\leftarrow F_k$
23. $\mathcal{M}_{temp} \leftarrow$ XGBClassifier(params $= P_i$,objective $=$ multi:softprob)
24. $\mathcal{M}_{temp}.Fit$(Train_Set)
25. Score$_k \leftarrow$ Calculate_Accuracy($\mathcal{M}_{temp}$,Validation_Set)
26. Total_Score $\leftarrow$ Total_Score + Score$_k$
27. **end for**
28. Fitness$_{P_i} \leftarrow$ Total_Score$/K$
29. **if** Fitness$_{P_i} >$ MaxFitness **then**
30. MaxFitness $\leftarrow$ Fitness$_{\mathrm{P_i}}$
31. $\mathrm{P_{best}} \leftarrow \mathrm{P_i}$
32. **end if**
33. **end for**
34. Update wolves' positions based on GWO rules
35. **end for**

Phase 4: Final Model Generation

36. $\mathcal{M}_{final} \leftarrow$ XGBClassifier(params $= P_{best}$,objective $=$ multi:softprob)
37. $\mathcal{M}_{final}.Fit(X_{scaled}, y)$
38. **return** $P_{best}, \mathcal{M}_{final}$

Following the execution of this automated optimization pipeline, the optimal hyperparameter configurations for each evaluated model were identified. The optimal hyperparameters for XGBoost obtained via GWO were configured as follows: n_estimators: 77, max_depth: 12, learning_rate: 0.2263, subsample: 1.0, colsample_bytree: 0.6393, reg_alpha: 0.6607, reg_lambda: 1.7344. These specific configurations, which directly dictate the model's learning capacity and regularization constraints, along with those derived for the comparative baseline models, are detailed in Table 1.

Table 1. Optimized hyperparameter configurations of evaluated models

Model	Optimized HyperParameter Configuration
Logistic Regression	C = 0.01, max_iter = 593
SVM	C = 39.1122, gamma = 0.00016, degree = 5
KNN	n_neighbors = 12
MLP	hidden_layer_size = (70,), alpha = 0.0237, learning_rate_init = 0.0278
Random Forest	n_estimators = 178, max_depth = 30, min_samples_split = 9, min_samples_leaf = 2
XGBoost	n_estimators = 77, max_depth = 12, learning_rate = 0.2263, subsample = 1.0, colsample_bytree = 0.6393, reg_alpha = 0.6607, reg_lambda = 1.7344

4. RESULTS

4.1. Data Descriptions

The dataset utilized in this study was acquired through an online survey administered via the Google Forms platform. The questionnaire was explicitly tailored to gather both quantitative and qualitative data from individuals who had completed their middle school education. The primary objective was to analyze the factors influencing their personal, career, and educational development from the middle school period to the present. The dataset encompasses a comprehensive array of features pertaining to demographics, academic performance, and career guidance activities, along with the extent of their influence on future career orientations, as well as the impacts stemming from educational, familial, and social environments.

The survey was conducted in the Vietnamese context and targeted respondents who had completed lower secondary education. The participants were individuals of different ages and career experiences, with information related to their learning process, career choice, career

transition, or career orientation after the middle school stage. The use of Google Forms enabled the collection of responses from a diverse group of participants, reflecting various personal, educational, familial, school-related, and social factors that may influence career orientation. Accordingly, the collected questionnaire data included both quantitative and qualitative information, covering demographic characteristics, 9th-grade academic performance, school-based career guidance activities, family background, work experience, reasons for career changes, and current career orientation.

Following the data cleaning phase and the exclusion of invalid samples, the final dataset comprises 1,067 valid instances, encompassing a total of 248 features formatted as tabular data. Descriptive statistics reveal a heterogeneous dataset structure, consisting of 227 categorical variables and 19 numerical variables. The questionnaire is structurally categorized into four primary domains: (1) general information (demographics); (2) 9th-grade academic performance (grades and academic rankings); (3) school-based career guidance activities (counseling programs and student participation); and (4) factors influencing career orientation (internal and external determinants originating from the family, school, and societal environments).

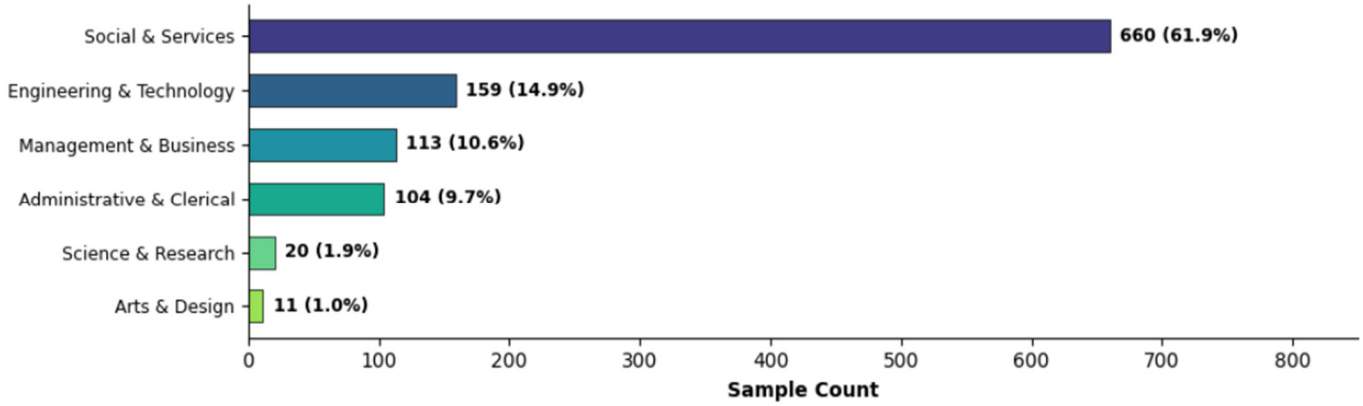

Figure 2. Sample distribution across career orientation categories

The target variable of this study is career orientation, which is categorized into six primary clusters to formulate a multi-class classification task: (1) Social & Services; (2) Engineering & Technology; (3) Management & Business; (4) Administrative & Clerical; (5) Science & Research; and (6) Arts & Design. Figure 1 illustrates the distribution of the target variable within the dataset.

The dataset exhibits pronounced class imbalance, accurately reflecting the real-world occupational distribution inherent in the surveyed sample. Specifically, the "Social & Services" category dominates with 660 instances (61.9%). Subsequent categories comprise "Engineering & Technology" with 159 instances (14.9%), "Management & Business" with 113 instances (10.6%), and "Administrative & Clerical" with 104 instances (9.7%). The two minority classes exhibiting the lowest proportions are "Science & Research" (1.9%) and "Arts & Design" (1.0%). This severe class imbalance characteristic is subsequently addressed during the data preprocessing phase employing appropriate class-balancing techniques. The dataset holds substantial scientific value as it encapsulates a retrospective real-world perspective on the middle school phase - a pivotal period for cultivating career awareness. Furthermore, by providing empirical data from Vietnam, it significantly enriches the broader literature regarding Artificial Intelligence applications within educational and career counseling domains.

4.2. Performance evaluation

To evaluate the efficacy of machine learning approaches in the career orientation classification task, empirical experiments were conducted on the VCS-024 dataset. Initially, the raw input data underwent a rigorous preprocessing pipeline designed for data cleaning, noise reduction, and feature scaling, thereby generating a high-quality input vector space for the predictive models.

The comparative experiments were executed across six representative machine learning models, encompassing linear, distance-based, neural network, and Ensemble Learning architectures: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest, and eXtreme Gradient Boosting (XGBoost).

To ensure an objective evaluation and mitigate the risk of zero-variance bias inherent in fixed data partitioning, we replaced the conventional train-test split approach. Instead, we employed a Stratified 10-Fold Cross-Validation strategy, integrated with the

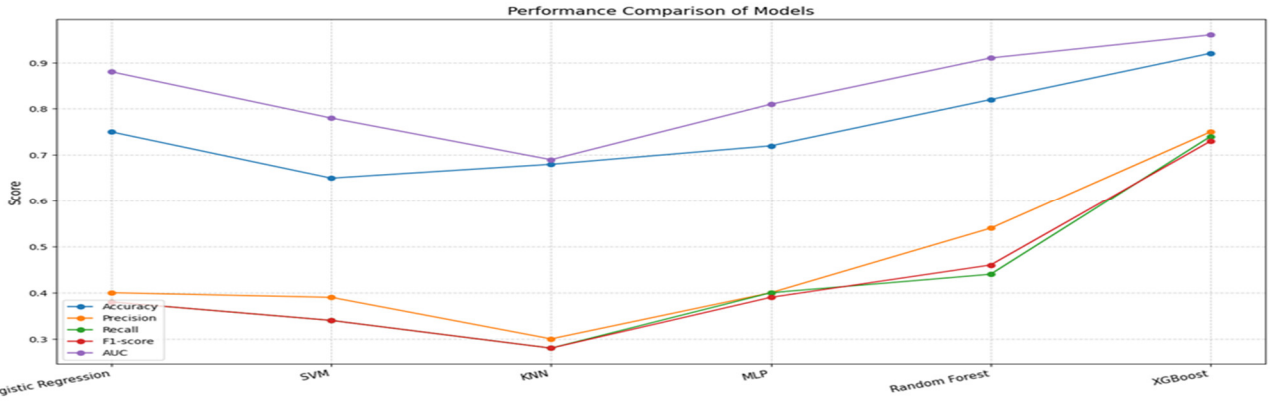

Figure 3. Model performance comparison

hyperparameter optimization framework detailed in Section 3.

Furthermore, given the severe class imbalance present in the dataset, relying solely on standard Accuracy is statistically inadequate and does not fully reflect the model's performance on minority classes. Therefore, the evaluation criteria were expanded to encompass Accuracy, Macro-Precision, Macro-Recall, Macro F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC).

The macro-averaging strategy was selected to compute metrics independently for each class before calculating the unweighted mean. This approach provides a balanced assessment across all career orientation classes by penalizing models that perform poorly on minority groups, thereby mitigating potential bias. The final comparative performance metrics, reported as the Mean ± Standard Deviation across the 10 folds, are summarized in Table 2.

Table 2. Performance metrics of machine learning models on the preprocessed dataset

Model	Accuracy	Macro Precision	Macro Recall	Macro F1 - score	AUC
Logistic Regression	0.75 ± 0.04	0.40 ± 0.05	0.38 ± 0.05	0.38 ± 0.05	0.88 ± 0.06
SVM	0.65 ± 0.01	0.39 ± 0.08	0.34 ± 0.03	0.34 ± 0.03	0.78 ± 0.01
KNN	0.68 ± 0.03	0.30 ± 0.06	0.28 ± 0.02	0.28 ± 0.03	0.69 ± 0.03
MLP	0.72 ± 0.04	0.40 ± 0.08	0.40 ± 0.07	0.39 ± 0.07	0.81 ± 0.06
Random Forest	0.82 ± 0.05	0.54 ± 0.07	0.44 ± 0.07	0.46 ± 0.07	0.91 ± 0.06
XGBoost	**0.92 ± 0.02**	**0.75 ± 0.08**	**0.74 ± 0.09**	**0.73 ± 0.08**	**0.96 ± 0.05**

The empirical results presented in Table 2 reveal differences in performance among the evaluated model families on the VCS-024 dataset. Traditional classifiers and the basic neural network demonstrated limitations when handling this high-dimensional, tabular dataset characterized by a severe class imbalance, where the dominant class accounts for 61.9% of the total observations.

For instance, while Logistic Regression achieved an Accuracy of 0.75 ± 0.04, its Macro F1-score decreased to 0.38 ± 0.05. A similar trend was observed in SVM (Accuracy: 0.65 ± 0.01, F1-score: 0.34 ± 0.03), KNN (Accuracy: 0.68 ± 0.03, F1-score: 0.28 ± 0.03), and MLP (Accuracy: 0.72 ± 0.04, F1-score: 0.39 ± 0.07). This discrepancy between standard Accuracy and the Macro F1-score highlights the "Accuracy Paradox," wherein these algorithms exhibit local overfitting toward the majority class (e.g., "Social & Services") and misclassify instances of minority classes, such as "Science & Research" or "Arts & Design."

Tree-based Ensemble Learning architectures showed improved performance in handling these data distributions. The Random Forest model achieved an Accuracy of 0.82 ± 0.05 and an AUC of 0.91 ± 0.06. However, its Macro F1-score of 0.46 ± 0.07 indicates that while bagging techniques reduce variance, they remain sensitive to severe class imbalances within educational datasets.

In contrast, the eXtreme Gradient Boosting (XGBoost) model yielded the highest performance across all evaluated metrics. XGBoost achieved an Accuracy of 0.92 ± 0.02, outperforming the baseline architectures. Furthermore, it mitigated the majority-class bias, yielding a Macro-Precision of 0.75 ± 0.08 and Macro-Recall of 0.74 ± 0.09, resulting in a Macro F1-score of 0.73 ± 0.08, an improvement of approximately 27% over Random Forest.

From a machine learning perspective, XGBoost's performance in this context can be fundamentally attributed to its sequential boosting mechanism. Unlike bagging or traditional empirical risk minimization, XGBoost iteratively constructs decision trees by calculating the first- and second-order gradients of the loss function. This mechanism causes the subsequent trees to place higher optimization weights on previously misclassified samples - a property that is particularly advantageous for learning the hidden characteristics of minority classes. Furthermore, its built-in regularization penalties effectively constrain model complexity, preventing local overfitting within the sparse, high-dimensional tabular feature space of the career survey data.

To further analyze the predictive behavior of the optimized XGBoost classifier, a visual analysis was conducted utilizing the confusion matrix and Receiver Operating Characteristic (ROC) curves.

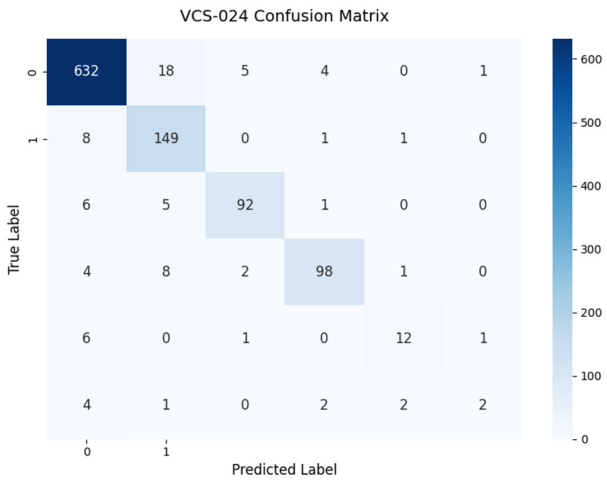

Figure 4. Confusion Matrix of the optimized XGBoost model on the dataset

As illustrated in Figure 4, the concentration of values on the main diagonal confirms the model's precision in predicting the correct career orientations across all categories. Minor off-diagonal overlaps are primarily observed between adjacent career groups, such as Technology/Engineering and Natural Sciences. From a psychological and academic perspective, this alignment is consistent, as students leaning towards these fields frequently share analogous cognitive tendencies and academic performance indicators, particularly high proficiency scores in Mathematics and Physics. XGBoost's ability to isolate these intra-class boundaries further validates its discriminative capacity. This multi-class classification performance is further substantiated by the One-vs-Rest (OvR) ROC curves displayed in Figure 5.

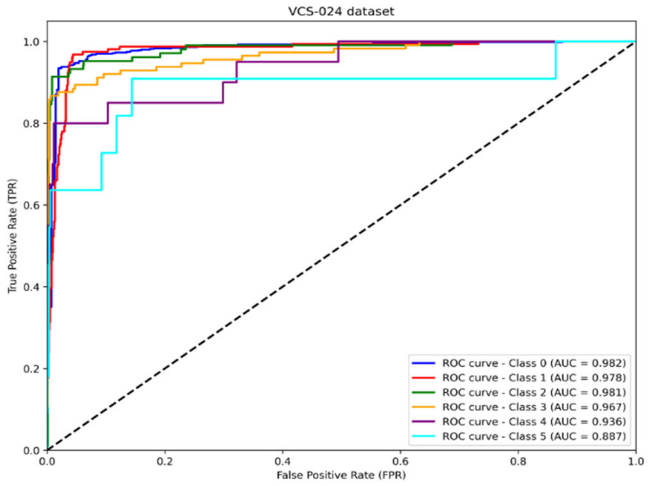

Figure 5. Multi-class ROC curves and AUC scores for the optimized XGBoost model

As depicted in the curves, the true positive rates for most career classes approach the top-left corner, demonstrating high sensitivity alongside a minimized False Positive Rate. The Area Under the Curve (AUC) score of 0.95 ± 0.05 for XGBoost (with individual class AUC values ranging from 0.887 to 0.982, averaging at approximately 0.955) indicates that the model possesses a stable probabilistic boundary. It effectively distinguishes a specific career orientation from other available options. Additionally, the low standard deviations observed across all metrics during the 10-fold cross-validation empirically demonstrate the model's stability and robustness to data partitioning variance. Consequently, these findings establish the optimized XGBoost model as a robust algorithmic foundation for intelligent career counseling support systems.

Although the study achieved highly promising results regarding the performance of early career prediction models, the mechanism for analyzing feature importance within the dataset has not yet been integrated into the proposed system. This aspect is particularly important in career counseling and guidance systems, where transparency and interpretability play a critical role in enhancing user trust and decision-making processes. Recent studies have demonstrated that the integration of Explainable Artificial Intelligence (XAI) techniques into recommendation and prediction systems can significantly improve the effectiveness and interpretability of career guidance outcomes. Therefore, incorporating XAI mechanisms into the proposed framework represents a key direction for future research in this study.

5. CONCLUSION

This study has proposed a systematic approach to evaluating and identifying the optimal machine learning model for career orientation prediction among middle school students. Addressing the specific challenges of career educational data - which is characterized by high-dimensional tabular structures (248 features) and severe class imbalance - this research conducted a comprehensive empirical evaluation using the standardized VCS-024 dataset.

Experimental results demonstrate that the XGBoost ensemble learning algorithm outperforms both traditional classification models and basic neural networks. Specifically, XGBoost achieved an accuracy of 92% and an F1-score of 73%, markedly higher than the second-best model, Random Forest, which reached only 82% accuracy and a 46% F1-score. Beyond high performance, XGBoost exhibited superior generalization capabilities and high stability, as evidenced by rigorous evaluation metrics and negligible standard deviations across multiple independent runs. The model effectively handles data imbalance, maintaining accurate predictions for minority occupational classes while mitigating the local overfitting common in other algorithms.

These findings not only address the existing research gap regarding AI applications for post-middle school student streaming in Vietnam, but also offer significant practical implications. The outstanding performance of XGBoost demonstrates its potential as a reliable core algorithm, thereby establishing a solid foundation for the development of intelligent decision-support systems in educational contexts. Future research could further extend this work by integrating the model into automated career guidance platforms and incorporating XAI techniques to provide detailed interpretations of the factors influencing students' career orientation decisions. In addition, the inclusion of psychological variables and labor market dynamics would contribute to the development of more personalized, comprehensive, and adaptive career guidance pathways for students.

ACKNOWLEDGEMENT

This research is a product of the research group on Intelligent Systems and Applications. The Intelligent Systems and Applications research group (Code: 01-2025-NCM) gratefully acknowledges the support provided by Hanoi University of Industry for the implementation of this research

REFERENCES

[1]. D. Bzdok, N. Altman, M. Krzywinski, "Statistics versus machine learning," *Nature Methods,* 15, 4, 233-234, 2018. doi: 10.1038/nmeth.4642.

[2]. F. Trujillo, M. Pozo, G. Suntaxi, "Artificial Intelligence in Education: A Systematic Literature Review of Machine Learning Approaches in Student Career Prediction," *Journal of Technology and Science Education* (JOTSE), 2025.

[3]. S. Ajgaonkar, P. Tale, Y. Joshi, P. Jore, M. Jakate, S. Lavangare, D. Kadam, "EduKrishnaa: A Career Guidance Web Application Based on Multi-intelligence Using Multiclass Classification Algorithm," in *Multi-disciplinary Trends in Artificial Intelligence*, R. Morusupalli et al., Eds., *Lecture Notes in Computer Science*, 14078, Springer, Cham, 601-610, 2023. doi: 10.1007/978-3-031-36402-0_56.

[4]. B. M. Muhammad, A. A. Lawan, J. Bala, T. S. Abdulrauf, B. I. Muhammad, "Predictive modeling of student career pathways using machine learning techniques," *Journal of Statistical Sciences and Computational Intelligence,* 1, 3, 166-174, 2025. doi: 10.64497/jssci.95.

[5]. A. Balasubramanian, "Personalized Career Pathway: A Hybrid Machine Learning Approach for Dynamic Recommendations," *Journal of Artificial Intelligence, Machine Learning and Data Science,* 1, 1, 1999-2003, 2023. doi: 10.51219/JAIMLD/abhinav-balasubramanian/440.

[6]. S. Sarwat, N. Ullah, S. Sadiq, R. Saleem, M. Umer, A. A. Eshmawi, A. Mohamed, I. Ashraf, "Predicting Students' Academic Performance with Conditional Generative Adversarial Network and Deep SVM," *Sensors,* 22, 13, 4834, 2022. doi: 10.3390/s22134834.

[7]. B. Mounika, V. Persis, "A Comparative Study of Machine Learning Algorithms for Student Academic Performance," *International Journal of Computer Sciences and Engineering*, 7, 4, 721-725, 2019. doi: 10.26438/ijcse/v7i4.721725.

[8]. K. Yan, "Student Performance Prediction Using XGBoost Method from A Macro Perspective," in *Proc. 2021 2nd International Conference on Computing and Data Science* (CDS), 2021. doi: 10.1109/CDS52072.2021.00084.

[9]. T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785-794, 2016. doi: 10.1145/2939672.2939785.

[10]. R. Shwartz-Ziv, A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion,* 81, 84-90, 2022. doi: 10.1016/j.inffus.2021.11.011.