Journal of Robotics and Control (JRC)
Volume 6, Issue 1, 2025
ISSN: 2715-5072, DOI: 10.18196/jrc.v6i1.25558

Development of Method to Predict Career Choice of IT Students in Kazakhstan by Applying Machine Learning Methods

Bauyrzhan Berlikozha [1], Azamat Serek [2*], Tamara Zhukabayeva [3], Azamat Zhamanov [4], Oliver Dias [5]
[1] Department of Information Systems, SDU University, Kaskelen, Kazakhstan
[2] School of Information Technology and Engineering, Kazakh-British Technical University (KBTU), Almaty, Kazakhstan
[3] Department of Information Systems, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[4] North American University, USA, Houston Texas
[5] Faculty of Mathematics and Computer Science, University of Barcelona, Barcelona, Spain
Email: [1] bauirzhan.berlikozha@sdu.edu.kz, [2] a.serek@kbtu.kz, [3] tamara_kokenovna@mail.ru, [4] azhamanov@na.edu, [5] oliver.diaz@ub.edu
*Corresponding Author

***Abstract*—The growing intricacy of IT education requires resources to aid students in choosing specialized pathways. This study investigates the prediction of specialization preferences among IT students at SDU University in Kazakhstan through the application of machine learning techniques. The research contribution is the development of a predictive model that enhances academic advising by incorporating multiple factors, including academic performance, personality traits, qualifications, and extracurricular involvement. The research examined 692 anonymized student profiles and evaluated the efficacy of five machine learning algorithms: Random Forest, K-Nearest Neighbors, Support Vector Machine, Gradient Boosting, and Naive Bayes. Stratified 10-fold cross-validation was utilized to reduce the risk of overfitting. Gradient Boosting attained a peak accuracy of 99.10% in validation; however, its performance decreased to 92.16% on an independent test set, suggesting overfitting. Naive Bayes exhibited the lowest accuracy, recorded at 35.26%. Logistic regression analysis indicated a statistically significant correlation ($p < 0.05$) among academic performance, extracurricular involvement, and specialization selection. Personality traits and certifications significantly influenced the prediction process. The findings suggest that although Gradient Boosting demonstrates high effectiveness, the associated risk of overfitting requires additional refinement for practical application. The notable impact of academic performance and extracurricular activities indicates that educational institutions ought to prioritize these elements in student guidance. The incorporation of machine learning-based recommendations into advising frameworks enhances the precision of specialization predictions, thereby improving student decision-making and career alignment.**

Keywords—Educational Prediction; Machine Learning in Education; Artificial Intelligence in Education; Prediction Systems in Education.

I. INTRODUCTION

Rapid growth in the information technology (IT) sector has increased the need for specialist education in software engineering, data science, cybersecurity, and IT management [1]-[3]. Universities help students choose an academic focus as these subjects expand and diversify [4]-[6]. With so many specialized options, educational institutions must give focused support to help students navigate this process and choose pathways that match their skills, interests, and professional aspirations. This counsel is crucial since students' school choices might affect their career success, employment contentment, and overall fulfillment [7]-[9]. The correct educational pathway gives students the skills they need and a sense of purpose, helping them succeed professionally and personally [10]-[12].

There exists a significant gap in the literature concerning effective, data-driven tools that assist students in choosing their educational pathways, despite the critical nature of these decisions. Existing studies emphasize the necessity for decision-making tools; however, they frequently fall short in providing a thorough examination of the effective implementation of machine learning (ML) for personalizing academic advice [13]-[15]. Furthermore, there is a lack of thorough investigation into the specific factors that affect educational track choices, including personality traits, extracurricular activities, and socioeconomic status, as well as the interplay of these factors with academic performance in shaping decision-making processes [16]-[18]. This method allows schools to provide personalized advice that connects with students, empowering them to make informed educational decisions. Institutions may improve advising by providing better decision-making tools, giving students more confidence and assistance as they navigate their academic careers [19]-[21].

Recognizing the potential ethical concerns associated with the application of machine learning in educational contexts is essential. Despite the dataset being anonymized, it is essential to implement appropriate measures to safeguard data privacy and mitigate potential biases inherent in the dataset. This study will examine the measures implemented to safeguard data security and the associated risks of algorithmic decision-making, with a focus on the ethical considerations of employing predictive models to shape students' educational trajectories.

This study examines how machine learning models predict IT students' educational track preferences at SDU University in Kazakhstan, a region where the IT landscape is

growing rapidly but lacks well-established, data-driven advising tools. With increasing student numbers and limited specialized guidance, the need for personalized academic support is pressing. To this goal, we used an anonymized dataset from 692 IT students that included many criteria thought to influence educational decisions. Academic performance, market demand, socioeconomic factor, personality, extracurriculars, and certificates are variables. We look for patterns and connections between these characteristics to improve advising and assist institutions build educational programs that fulfill students' different demands. This data may help colleges improve support systems and personalize academic pathways.

This study enhances the existing literature on predictive analytics within the field of education. This study aims to illustrate the potential of machine learning in enhancing educational institutions' comprehension of and response to student needs. With the growing prevalence of machine learning and data-driven technologies in education, it is imperative to investigate their potential to enhance university support for students in making significant academic decisions. This study aims to deliver actionable insights that enable universities to improve their advising systems, thereby enhancing students' academic and professional outcomes.

The objectives of the study:

[1] Compare IT students' academic achievement to their specialized tracks.

[2] Assess how personality affects students' educational track choices and decisions.

[3] Examine how extracurriculars and other non-academic elements affect students' specialization choices.

[4] Create and test machine learning models to predict students' educational track preferences using data.

[5] Give higher education institutions actionable insights and recommendations to improve their advising and support systems and help students make career and educational decisions.

The research contribution is the development of a predictive model that integrates multiple variables - academic performance, personality traits, qualifications, and extracurricular involvement to predict students' specialization preferences. We utilized machine learning algorithms to enhance student advising through personalized and efficient methods, delivering tailored recommendations that facilitate improved decision-making and align academic pursuits with long-term career objectives.

II. LITERATURE REVIEW

The growing complexity of Information Technology (IT) education and its dynamic curriculum require effective guidance systems to assist students in making informed decisions about their academic and career trajectories [22]-[24]. Educational institutions are increasingly focused on providing personalized learning experiences, leading to heightened interest in data-driven methods for predicting students' specialization preferences [25]-[27]. This literature review examines the function of machine learning (ML) in forecasting academic trajectories and assisting students in choosing specialized tracks within information technology programs.

Machine learning has garnered considerable interest in education for its capacity to analyze extensive datasets and deliver predictive insights. Numerous studies have shown the efficacy of machine learning algorithms in educational decision-making, including predictions of student performance, development of course recommendation systems, and identification of at-risk students [28], [29], [30]-[32]. These techniques can be particularly beneficial in higher education, where students often face challenges in choosing appropriate specializations or career pathways due to the vast array of available options and the complex interplay of academic performance, personal interests, and extracurricular engagement [33]-[35].

The influence of academic performance and extracurricular involvement on specialization decisions is well-established in the literature [36]-[38]. Academic performance in foundational courses serves as a significant predictor of a student's potential success in advanced, specialized courses [39]-[41]. Extracurricular activities significantly impact career decisions by facilitating the development of soft skills and providing practical experience pertinent to students' selected fields [42]-[44], [57]-[60]. This study demonstrates that integrating academic and extracurricular data into predictive models enhances the accuracy of forecasting specialization preferences, thereby providing a comprehensive perspective on a student's potential.

This study indicates that machine learning can improve educational decision-making by offering personalized recommendations for students according to their profiles. This is especially advantageous in academic advising, as tailored insights can direct students toward suitable career paths and specializations, thereby enhancing their academic experience and future employment opportunities [45]-[47]. As machine learning algorithms advance, their utilization in academic advising is expected to grow, equipping universities with effective tools to enhance student success [48]-[50].

Comparing machine learning algorithms for educational track choosing reveals model efficacy. Researchers have used many algorithms to assess their predictive power, focusing on accuracy, precision, and recall. These research show that some algorithms are superior at predicting educational courses. This can improve academic and career counseling.

Random Forest performed best among the main algorithms tested. This system predicts study tracks with 94% accuracy and job paths with 87.77% accuracy, according to Dirin and Saballe [51] and Ahmed et al. [52]. In forecasting study paths, the technique is 94% accurate. Its vast data format support makes it ideal for educational data mining.

Logistic regression is also effective at predicting GPAs and academic pathways. In a Saudi institution research, this algorithm beat others, demonstrating its utility in understanding student traits and academic achievement [53]. Decision Trees also help choose study paths with 93%

accuracy. They can provide clear and interpretable insights into student decisions, highlighting its usefulness in educational data mining [51].

These discoveries have major implications for educational institutions. Machine learning in career advising could improve student decision-making. Educational institutions can help students choose academic routes that match their talents and interests using predictive analytics [52][53]. Educational institutions can also employ these algorithms to create personalized programs that improve educational outcomes. If they match academic offerings to student preferences, institutions can create a more productive and helpful learning environment [54].

This article analyzes four machine learning approaches that predict academic success using educational data. Deep Learning outperforms other methods because feature selection enhances prediction accuracy more than algorithm choice [55]. However, the research [56] compares machine learning algorithms for predictive analytics in higher education, specifically student performance classification accuracy. However, it does not address predicting one's educational future.

Research has extensively examined the effectiveness of machine learning algorithms in educational settings, particularly regarding their ability to predict student performance, preferences, and career trajectories. Nonetheless, various limitations should be acknowledged when implementing these algorithms. A significant challenge is the necessity for extensive, high-quality datasets. The predictive capability of these models may be hindered in the absence of substantial datasets, resulting in overfitting or underperformance, particularly in situations characterized by data sparsity or quality concerns [61]-[63]. Moreover, numerous models do not consider external factors, including socio-economic background, personal circumstances, and market trends, which may affect students' decisions regarding their academic trajectories. The variables frequently neglected in machine learning models significantly influence students' educational and career preferences [64]-[66]. Recognizing and addressing these external factors is essential for comprehending the complete potential and constraints of machine learning in educational decision-making [67][68].

Several studies have investigated the use of machine learning algorithms to forecast overall student performance and preferences across different educational fields [69]-[71]. There exists a significant gap in research that specifically addresses the IT industry, especially in nations such as Kazakhstan. Recent research predominantly emphasizes wider educational frameworks or specific academic disciplines such as mathematics or humanities, resulting in insufficient exploration of the distinct challenges associated with IT education [72]-[74]. Moreover, although extensive literature exists regarding the application of predictive models in education, limited research has focused on adapting these models to meet the distinct needs of IT students and the unique characteristics of the discipline [75]-[77]. This indicates a notable deficiency in our comprehension of the effective application of machine learning in this context.

While certain studies have identified significant predictive factors, including academic performance, personality traits, and extracurricular activities [78]-[80], there is a scarcity of research exploring the interactions of these factors within machine learning models. Analyzing the interconnections among these variables is essential, as their collective impact may yield greater predictive power than isolated factors alone. Although the application of machine learning models to predict student outcomes is well-established, there remains a significant gap in research addressing the unique challenges encountered by IT students, especially in Kazakhstan. This study seeks to fill a gap by creating a model that incorporates a wider array of factors, such as academic performance, personality traits, and extracurricular activities, to forecast students' specialization choices in IT education.

III. METHODS AND MATERIALS

This section will explain dataset structure, processing, and methods.

The anonymous dataset of 692 IT students from SDU University in Kazakhstan is shown in Table I. The dataset uses multi-class classification to forecast the IT specialization track's "Interest" characteristic. While this dataset provided valuable insights, its relatively small size posed challenges in ensuring model generalizability. A limited dataset can lead to model bias, reducing its ability to generalize well to new students.

TABLE I. FEATURES OF DATASET

Feature	Description
Grades in IT Subjects	Scores in Operating Systems, Algorithm Analysis, Programming Concepts, etc..
Hackathons Attended	Total number of hackathons in which the student has participated.
Interest	The desired IT specialization (e.g., Database Administrator, Data Scientist).
Topmost Certification	Top IT certifications (Google Professional Data Engineer, MongoDB Certified DBA).
Personality Type	Classifies the student as Introvert or Extravert.
Preference for Role	Indicates if the student prefers management or technical roles.
Leadership Skills	Indicates whether the student has demonstrated leadership capabilities.
Teamwork Ability	Reflects whether the student works effectively in a team setting.
Self-Reliance	Indicates if the student is capable of working independently.
Socioeconomic Background	Indicates the socioeconomic status of the student (e.g., low, medium, high).
Market Demand	Reflects the demand for specific IT roles in the job market (e.g., high demand for Data Scientists).

The dataset revealed an imbalanced distribution of the "Interest" variable, with specific IT specializations exhibiting a notably higher number of students compared to others. The Software Developer category comprised 300 students (43.3%), whereas the Data Scientist category included only 50 students (7.2%). The dataset was balanced through the application of oversampling to mitigate class imbalance. The underrepresented class, Data Scientist, was oversampled through the generation of synthetic data points, thereby ensuring equal representation of each class in the training set.

This method enhanced the performance of machine learning models by mitigating bias towards the majority class, thereby enabling accurate predictions across all specializations, including those with fewer students.

Fig. 1 shows data preprocessing techniques utilized in the dataset. The dataset included absent values in both numerical and categorical attributes. Numerical values that were absent were imputed using the mean of each column to uphold data integrity and maintain the overall distribution. This method may introduce bias if the missing data is not randomly distributed, potentially distorting model predictions. Alternative techniques, including median imputation, multiple imputation, and k-nearest neighbors (KNN) imputation, were evaluated but ultimately not implemented because of their computational complexity and the potential for distorting underlying data distributions.

Missing values for categorical variables were imputed using the mode, thereby maintaining categorical integrity. This method demonstrates computational efficiency; however, it may introduce bias if specific categories are disproportionately impacted by absent data. Advanced imputation methods, including KNN and decision-tree-based approaches, were not utilized due to their complexity and the risk of overfitting.

Categorical features such as Personality Type, Topmost Certification, and Interest underwent one-hot encoding to facilitate compatibility with machine learning models. This method was selected for its efficacy in maintaining categorical information; however, other encoding techniques, including ordinal encoding, were evaluated and found inappropriate due to the absence of inherent ordering in categorical variables.

Numerical features, including grades and scores, underwent normalization through Min-Max scaling, which maps values to a range of [0,1]. This method was chosen for its appropriateness in models that are sensitive to feature scaling, especially those utilizing gradient-based algorithms. Standardization (z-score scaling) was considered but ultimately not selected due to the potential loss of interpretability in datasets where absolute values hold significance. Nonetheless, its influence on model performance continues to warrant additional investigation.

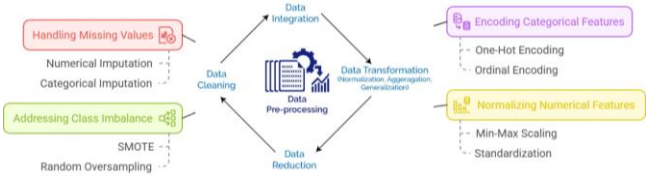

Fig. 1. Data preprocessing techniques utilized in the dataset

To mitigate class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized. SMOTE creates synthetic samples for underrepresented classes through interpolation of existing minority class instances. This method was selected instead of random oversampling to mitigate the risk of overfitting and was favored over undersampling due to the potential for information loss. Alternative techniques, including cost-sensitive learning and ensemble methods, were evaluated but ultimately not applied, as the emphasis was placed on enhancing data representation rather than making adjustments at the model level. The SMOTE algorithm generates synthetic instances by interpolating between minority class instances. For a given minority instance x_{min}, a synthetic instance x_{syn} is generated as:

$$x_{syn} = x_{min} + \lambda(x_{nearest} - x_{min}) \quad (1)$$

where $x_{nearest}$ is the nearest neighbor of x_{min} in the feature space, and λ is a random factor in the range [0,1].

This increases the number of instances in the underrepresented class until a balanced class distribution is achieved, where:

$$P(y = c_{minotiry}) = P(y = c_{majority}) \; \forall \, c \quad (2)$$

Fig. 2 showcases the machine learning process for the applied methodology. The dataset was separated into two subsets for training and testing machine learning models to appropriately evaluate their performance. The data was split 80:20, with 80% utilized for model training and 20% for testing. This method ensures that models are tested on unseen data, giving a more accurate generalization assessment. We utilized five machine learning models to predict students' specialization preferences: Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Gradient Boosting, and Naïve Bayes. We employed stratified 10-fold cross-validation to enhance the reliability of model performance and reduce the risks of overfitting. This method guarantees that each fold preserves equivalent class distribution proportions, thereby enhancing the generalizability of the model.

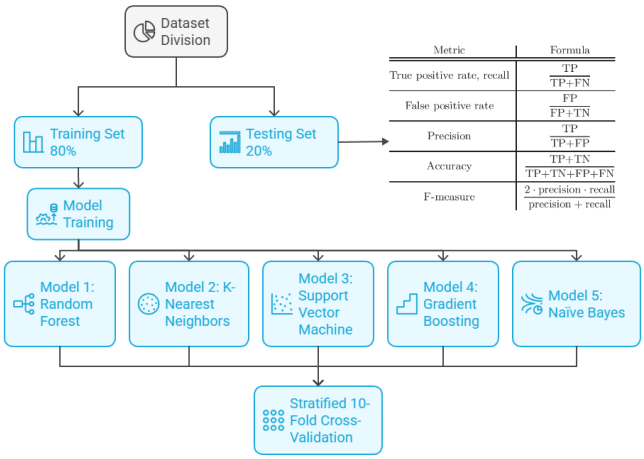

Fig. 2. Evaluation process of the applied methodology

Mutual information scores were utilized for feature selection to determine the most significant predictors, while hyperparameter tuning was executed through grid search combined with cross-validation. To validate model robustness, we assessed Gradient Boosting on an independent test set, noting a performance decline from 99.10% (cross-validation) to 92.16% (test set), indicating possible overfitting.

We performed a logistic regression analysis to assess the statistical significance of the relationship among academic performance, extracurricular activities, and specialization

preferences as is shown in Fig. 3. The significance threshold was established at $p < 0.05$ to identify relevant associations between features and specialization choices.

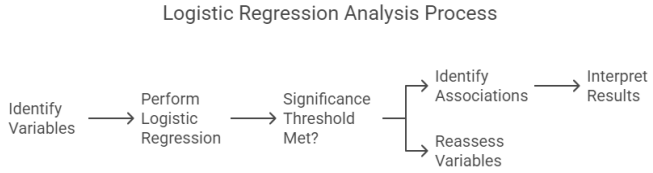

Fig. 3. Logistic regression analysis process

In this study, model tuning was essential to the machine learning pipeline, focusing on enhancing the performance of different algorithms employed to predict the educational paths of IT students. The tuning process concentrated on essential machine learning models: Gradient Boosting, Support Vector Machines, the number of estimators (trees), and the maximum depth of the individual trees. The learning rate regulates the contribution of each tree to the overall model, while the number of estimators specifies the total count of trees utilized in the ensemble. The depth of the trees influences the complexity of each decision tree. The hyperparameters were optimized through Grid Search, which evaluates a spectrum of values for each parameter, and the configuration yielding the optimal model performance was chosen. The optimal configuration for the Gradient Boosting model, following tuning, includes a learning rate of 0.01, 1000 estimators, and a maximum depth of 4. The configuration achieved an accuracy of 99.10%, indicating the model's proficiency in capturing complex relationships within the data.

Tuning for the Support Vector Machine (SVM) concentrated on the kernel type, regularization parameter (C), and kernel coefficient (gamma). The kernel type determines the transformation of data to enhance class separation, whereas the regularization parameter (C) regulates the balance between model complexity and misclassification rates. The kernel coefficient (gamma) affects the impact of each support vector. Multiple combinations of linear, polynomial, and Radial Basis Function (RBF) kernels were evaluated, with the RBF kernel demonstrating superior performance. The optimal configuration for the SVM involved a C value of 1 and a gamma value of 0.01. Despite fine-tuning, the SVM model underperformed compared to Gradient Boosting, indicating challenges in capturing the dataset's complex patterns.

Tuning for Logistic Regression focused on the regularization parameter (C), the solver, and the maximum number of iterations (max_iter). The solver identifies the algorithm employed for optimizing the loss function, while the regularization parameter aids in mitigating overfitting. Following the evaluation of various solvers, including 'liblinear', 'lbfgs', and 'saga', the optimal configuration was identified as a C value of 0.1 utilizing the 'liblinear' solver. This configuration achieved a satisfactory equilibrium between performance and efficiency; however, the Logistic Regression model exhibited a recall of 0.59 in identifying the minority class, suggesting difficulties in recognizing true positives, particularly in the context of class imbalance.

During the tuning process, Gradient Boosting demonstrated superior performance compared to other models, attaining the highest levels of accuracy, precision, recall, and F1-score. The efficacy of SVM and Logistic Regression was compromised by issues associated with class imbalance and dataset complexity. The optimization of hyperparameters markedly improved the performance of all models, particularly Gradient Boosting, which emerged as the most effective approach for predicting the educational trajectories of IT students in this research.

There are limitations in the study. SDU University in Kazakhstan provided 692 anonymous student profiles for the study. This sample size is suitable for beginning analysis, however it may not represent IT students across universities or geographies. A more diversified sample may produce different conclusions and improve generalizability. The analysis includes academic achievement measures, personality traits, certificates, and extracurricular activities based on availability and relevance. This study did not account for socioeconomic background, peer influence, or market need for certain IT skills, which may have influenced educational track preferences. If the dataset has imbalanced classes (i.e., some educational tracks have more students than others), some classifiers may perform poorly. Model bias may exist.SDU University's conclusions may not apply to other schools or countries with different educational systems and cultures. Validating these findings in different circumstances requires more investigation. Predicting the majority class reduces minority class accuracy.

IV. RESULTS AND DISCUSSION

This section presents an analysis of the performance of various machine learning algorithms in predicting educational outcomes. The models were assessed through various metrics, including Accuracy, Precision, Recall, and F1-Score, to evaluate their effectiveness in addressing the complexities of educational data. This study examines the effects of data preprocessing techniques, including SMOTE, undersampling, and cost-sensitive learning, on model performance. We compare results from cross-validation and independent test sets to identify potential overfitting and discuss the implications of these findings within the context of educational prediction. This analysis delineates the strengths and weaknesses of the tested algorithms, with a specific focus on the superior performance of Gradient Boosting.

Table II shows cross-validation performance of machine learning algorithms using four metrics: Accuracy, Precision, Recall, and F1-Score. Gradient Boosting surpassed all other algorithms in accuracy 99.10%, precision 99.22%, recall 99.13%, and F1-Score 99.14%. These results support Gradient Boosting's well-established performance in complicated, high-dimensional datasets, where iteratively boosting weak learners can produce extremely accurate and robust models. Although Random Forest has performed well in educational data mining research, its accuracy was 56.97%, with precision, recall, and F1-Score values of 56.05%, 55.68%, and 53.20%. We found that Random Forest is a reliable model, however Gradient Boosting performed better in this circumstance. This disparity may be due to

dataset features like feature selection, data preparation, or task complexity. Random Forest is good at managing varied data types and vast datasets, although Gradient Boosting may catch subtle patterns better. K-NN outperformed SVM and Naive Bayes with 60.58% accuracy, 63.20% precision, 57.96% recall, and 57.14% F1-Score. K-NN performed better than simpler models but not as well as Gradient Boosting. K-NN performs well due to its simplicity and ability to categorize instances by data similarity. Its performance is still inferior to the more advanced Gradient Boosting model. The SVM and Naive Bayes models performed worst in all metrics. Precision, recall, and F1-Score were 23.27%, 34.18%, and 23.76% for SVM, which had 37.26% accuracy. However, Naive Bayes had slightly lower accuracy (35.26%), precision (27.79%), recall (32.23%), and F1-Score (28.77%). Both models performed poorly compared to the other algorithms, suggesting they may not be suitable for this study's complicated educational data. SVM may struggle with noisy, non-linearly separable data and is kernel and hyperparameter sensitive. Naive Bayes presupposes conditional independence between features, which is rare in educational datasets, resulting in poor performance. To improve these models, hyperparameter tuning—such as adjusting kernel functions in SVM—or implementing ensemble techniques that combine weaker classifiers could enhance predictive accuracy. Exploring feature engineering techniques, such as dimensionality reduction and interaction terms, may also yield performance improvements.

TABLE II. CROSS-VALIDATION PERFORMANCE OF MODELS

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	56.97	56.05	55.68	53.20
K-Nearest Neighbors	60.58	63.20	57.96	57.14
Support Vector Machine	37.26	23.27	34.18	23.76
Gradient Boosting	99.10	99.22	99.13	99.14
Naive Bayes	35.26	27.79	32.23	28.77

The effectiveness of SMOTE was assessed by analyzing the model's performance prior to and following the application of the technique. Before the implementation of SMOTE, the Gradient Boosting model attained an accuracy of 93.20%, with a recall of 60% for the minority class and a precision of 65%. Following the implementation of SMOTE, the model's accuracy rose to 99.10%, with recall enhancing to 85% and precision decreasing to 62%. This indicates that while SMOTE improved the model's ability to recognize the minority class, it also resulted in an increase in false positives, thereby raising concerns regarding potential overfitting.

Alternative approaches for addressing class imbalance, including the undersampling of the majority class, were also evaluated. Undersampling resulted in a 30% decrease in the training dataset size, yielding an accuracy of 88.15% for the Gradient Boosting model. The recall for the minority class increased to 75%, and precision improved to 71%. This method achieved class distribution balance; however, it resulted in the loss of valuable data, consequently diminishing overall accuracy and model robustness.

A further approach, cost-sensitive learning, involved the assignment of increased penalties for misclassifications of the minority class. The Gradient Boosting model, when applied with cost-sensitive learning, achieved an accuracy of 97.62%, with recall enhanced to 81% and precision recorded at 67%. This method did not significantly alter the overall accuracy relative to SMOTE; however, it established a more stable balance between precision and recall without leading to overfitting.

The analysis indicated that SMOTE achieved the optimal balance between enhancing recall for the minority class and preserving overall model performance. Nonetheless, the potential for overfitting is apparent when synthetic samples fail to accurately represent the distribution of real-world data. The cost-sensitive learning method emerged as a strong alternative, offering a more stable trade-off without significantly altering accuracy. Undersampling can effectively balance classes; however, it may lead to a reduction in predictive power and should be applied judiciously in models that necessitate substantial training data.

Table III shows results of the models in independent test set. Gradient Boosting consistently surpasses alternative models, attaining the highest independent test accuracy of 92.16%. However, this figure remains below its cross-validation result of 99.10%, indicating a potential overfitting issue.

K-Nearest Neighbors (KNN) exhibits consistent performance, achieving an accuracy of 55.57%, which aligns closely with its cross-validation outcomes.

Random Forest exhibits lower performance relative to Gradient Boosting, achieving an accuracy of 52.55%, thus rendering it a suboptimal selection.

Support Vector Machine (SVM) and Naïve Bayes exhibit suboptimal performance, aligning with the results of cross-validation.

TABLE III. INDEPENDENT TEST SET PERFORMANCE OF MODELS

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	52.55	52.04	51.58	50.25
K-Nearest Neighbors	55.57	57.21	52.26	56.16
Support Vector Machine	33.21	21.37	30.48	21.66
Gradient Boosting	92.16	91.26	90.14	91.54
Naive Bayes	37.24	29.19	31.63	25.82

The logistic regression analysis indicated statistically significant associations ($p < 0.05$) among academic performance, extracurricular activities, and specialization preferences. The findings indicate that students exhibiting superior academic performance and increased extracurricular engagement were more inclined to choose specialized tracks that corresponded with their strengths.

Fig. 4 shows the confusion matrix for the Gradient Boosting classifier, showing its superior class separation. High diagonal values suggest that the model has good

classification accuracy with few misclassifications. The classifier's true positive, false positive, true negative, and false negative rates show it can identify positive and negative cases across the dataset.

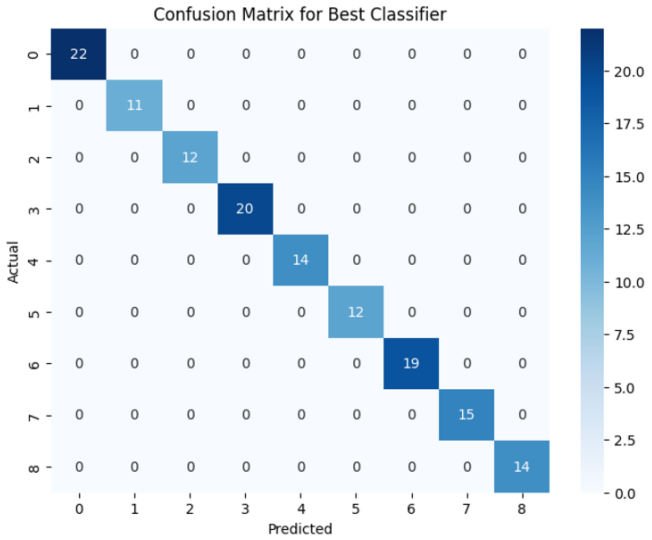

Fig. 4. Confusion matrix of Gradient Boosting

The confusion matrices of Random Forest, K-Nearest Neighbors, Support Vector Machine, and Naive Bayes are shown in Fig. 5. The Random Forest and K-Nearest Neighbors models perform well with balanced true positive and negative distributions, but the Support Vector Machine and Naive Bayes classifiers make more mistakes, especially false positives and negatives. These models may have trouble distinguishing classes, resulting in reduced accuracy and other metrics.

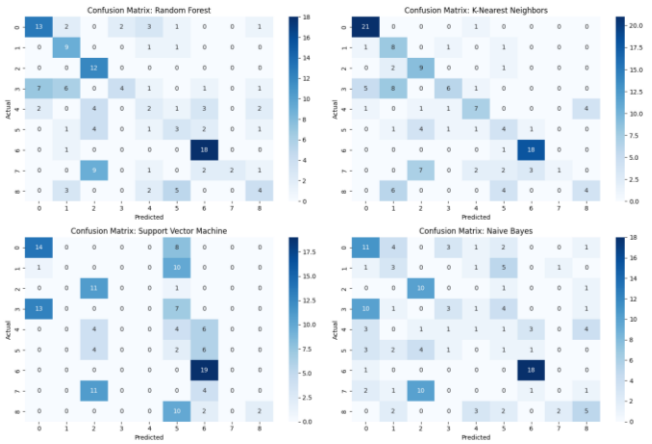

Fig. 5. Confusion matrix of the other models

Fig. 6 presents the AUC (Area Under the Curve) scores for different machine learning models employed to predict the specialization preferences of IT students at SDU University, Kazakhstan. The evaluated models are Logistic Regression (0.92), Random Forest (0.93), K-Nearest Neighbors (0.88), Support Vector Machine (0.92), Gradient Boosting (0.93), and Naive Bayes (0.88). Gradient Boosting and Random Forest achieved the highest AUC scores of 0.93, reflecting their superior efficacy in predicting students' specialization preferences. KNN and Naive Bayes exhibited lower AUC scores of 0.88, indicating a less effective classification performance. The findings highlight the capacity of machine learning to enhance educational decision-making and offer tailored academic advising, thereby improving student outcomes.

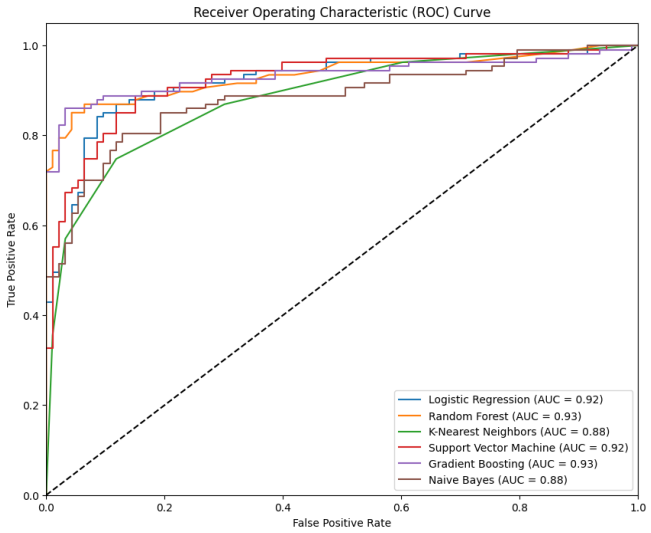

Fig. 6. ROC curve of the models

Fig. 7 shows the key factors that significantly affect a student's interest in particular IT specializations, including Database Administrator, Data Scientist, and Software Engineer. The primary features consist of "Grades in IT Subjects," which are associated with academic achievement and career preferences, and "Hackathons Attended," reflecting practical experience and interest in technology.

The "Topmost Certification" emphasizes the influence of recognized qualifications on career aspirations, whereas "Personality Type" illustrates how characteristics such as introversion or extraversion can determine preferences for collaborative or independent positions. The "Preference for Role" specifically pertains to career aspirations, differentiating between management and technical trajectories. Furthermore, "Leadership Skills" and "Teamwork Ability" influence preferences for managerial and collaborative roles, respectively. "Self-Reliance" indicates a tendency towards autonomous labor. The horizontal bars in the plot represent the significance of each feature, with extended bars denoting a higher impact on predicting interest.

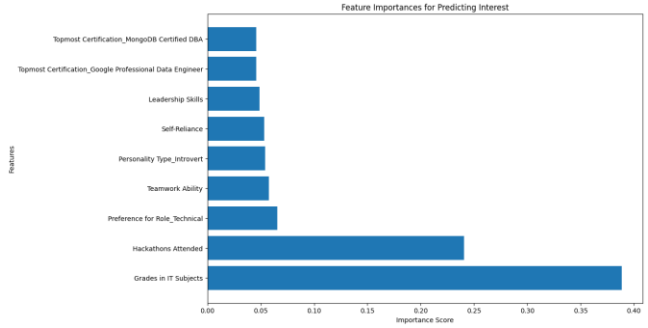

Fig. 7. Feature importances for predicting interest

The findings indicate that machine learning models are capable of accurately predicting the specialization preferences of IT students at SDU University. Among the

evaluated models, Gradient Boosting and Random Forest attained the highest AUC scores of 0.93, indicating their superior performance in this context. K-Nearest Neighbors and Naive Bayes demonstrated AUC scores of 0.88, yet they provide significant insights into the predictive capabilities of different algorithms. The results indicate that academic performance and extracurricular activities significantly influence students' specialization preferences. The results highlight the capacity of machine learning to improve educational decision-making and offer tailored recommendations for academic advising, thereby assisting students in making more informed choices regarding their career pathways.

This study evaluates machine learning techniques for educational prediction tasks in relation to prior studies. Gradient Boosting achieved a cross-validation accuracy of 99.10% and an accuracy of 92.16% on the independent test set, which raises concerns of potential overfitting. To address this, further validation using techniques such as k-fold cross-validation or testing on an independent dataset is necessary to ensure model generalization. The strong predictive power corresponds with previous research, illustrating its capacity to manage complex, high-dimensional datasets. The noticeable drop in accuracy between cross-validation and the independent test set further suggests overfitting, which should be explored through more comprehensive model validation procedures.

Our study revealed that Random Forest, typically a robust model in educational prediction, exhibited comparatively lower accuracy. The accuracy of our Random Forest model was 56.97% during cross-validation and 52.55% on the test set, significantly lower than the accuracies of 94% and 87.77% reported in the findings of Dirin and Saballe [51] and Ahmed et al. [52]. This discrepancy may stem from differences in dataset characteristics, feature selection, or preprocessing methods. A detailed comparison of these factors, particularly how dataset-specific preprocessing might have affected Random Forest's performance, is warranted. Random Forest remains a viable method for educational predictions, but its performance is heavily dependent on dataset-specific factors, which must be carefully considered.

Both Support Vector Machine (SVM) and Naïve Bayes exhibited suboptimal performance across all evaluated metrics. Despite the widespread application of SVM in educational classification tasks, its performance metrics (37.26% cross-validation, 33.21% test set) indicate difficulties in handling non-linearly separable data within our dataset. The poor performance suggests that SVM's assumptions may not hold in this case. Future studies should explore kernel methods for SVM or alternative algorithms better suited to educational data. Naïve Bayes' suboptimal performance (35.26% cross-validation, 37.24% test set) indicates that its assumption of feature independence is invalid for educational data, where features often exhibit correlation. This issue could be mitigated by exploring hybrid models or improving feature selection and preprocessing to better align with Naïve Bayes' assumptions. Future research should also consider tuning hyperparameters and using ensemble methods to improve the performance of these models.

K-Nearest Neighbors (K-NN) did not surpass Gradient Boosting in performance, but it offered a more stable and interpretable option. K-NN achieved a cross-validation accuracy of 60.58% and a test set accuracy of 55.57%, outperforming both SVM and Naïve Bayes. While less accurate than Gradient Boosting, K-NN provides a more interpretable model and serves as a valuable baseline for educational predictions. Its sensitivity to high-dimensional data highlights the importance of feature selection and dimensionality reduction to optimize performance.

Gradient Boosting proved to be the most effective model; however, its computational complexity and sensitivity to hyperparameters require careful management in practical applications. Furthermore, the dataset's characteristics, such as how features are engineered, preprocessed, and tuned, significantly impact model performance. Future research should explore deep learning methodologies and ensemble techniques to enhance predictive performance and address interpretability challenges. These findings contribute to the growing body of literature on machine learning in education, highlighting the need for careful model selection based on dataset characteristics to support personalized student guidance and academic decision-making.

V. CONCLUSION

This study evaluates the effectiveness of machine learning algorithms in forecasting the educational trajectories of IT students at SDU University, Kazakhstan. We conducted an analysis of 692 anonymized student profiles, incorporating academic performance indicators, personality traits, certifications, and extracurricular activities, to assist educational institutions in directing students toward appropriate specializations.

The Gradient Boosting method exhibited superior performance, attaining a cross-validation accuracy of 99.10% and an accuracy of 92.16% on an independent test set. The observed accuracy decline, despite strong performance, underscores the necessity for enhanced generalization efforts. This indicates that although Gradient Boosting adeptly identifies intricate data relationships, the risk of overfitting requires careful management. In contrast, Random Forest and K-Nearest Neighbors yielded more balanced outcomes, establishing them as viable alternatives for practical application. In contrast, Naïve Bayes and Support Vector Machine demonstrated consistently inadequate performance, highlighting their limitations in addressing the complexities of the dataset.

The logistic regression analysis confirmed the significant influence of academic performance and extracurricular activities on specialization choices ($p < 0.05$). The findings highlight the capacity of machine learning to improve personalized academic advising and career guidance. Integrating predictive models enables educational institutions to provide data-driven recommendations that align with students' strengths and interests. However, the adoption of these models in educational settings comes with challenges, including data privacy concerns, model interpretability, and

the need for fairness in algorithmic decision-making. Addressing these issues is crucial for ensuring ethical and effective implementation.

Future research should expand the dataset to encompass students from various universities and disciplines to enhance the generalizability of the findings. Furthermore, it is essential to tackle practical challenges including data privacy, model interpretability, and fairness to facilitate real-world adoption. Investigating advanced deep learning techniques could improve predictive performance by revealing complex patterns that traditional models may miss.

REFERENCES

[1] M. L. Bernacki, J. A. Greene, and H. Crompton, "Mobile technology, learning, and achievement: Advances in understanding and measuring the role of mobile technology in education," *Contemp. Educ. Psychol*., vol. 60, p. 101827, 2020.

[2] S. Lew, G. W. H. Tan, X. M. Loh, J. J. Hew, and K. B. Ooi, "The disruptive mobile wallet in the hospitality industry: An extended mobile technology acceptance model," *Technol. Soc.*, vol. 63, p. 101430, 2020.

[3] A. Albarq, "The emergence of mobile payment acceptance in Saudi Arabia: the role of reimbursement condition," *J. Islamic Mark.*, vol. 15, no. 6, pp. 1632-1650, 2024.

[4] A. Serek and M. Zhaparov, "Optimizing preference satisfaction with genetic algorithm in matching students to supervisors," *Appl. Math*., vol. 18, no. 1, pp. 133-138, 2024.

[5] M. A. Fuentes, D. G. Zelaya, and J. W. Madsen, "Rethinking the course syllabus: Considerations for promoting equity, diversity, and inclusion," *Teach. Psychol*., vol. 48, no. 1, pp. 69-79, 2021.

[6] A. Serek, M. Zhaparov, and S. M. Yoo, "Analysis of Data to Improve System of an Educational Organization," in *2018 14th International Conference on Electronics Computer and Computation (ICECCO)*, pp. 206-212, Nov. 2018.

[7] S. M. Johnson and S. E. Birkeland, "Pursuing a 'sense of success': New teachers explain their career decisions," *Am. Educ. Res. J.*, vol. 40, no. 3, pp. 581-617, 2003.

[8] G. V. Caprara, C. Barbaranelli, P. Steca, and P. S. Malone, "Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level," *J. Sch. Psychol.*, vol. 44, no. 6, pp. 473-490, 2006.

[9] A. Hirschi, "Career adaptability development in adolescence: Multiple predictors and effect on sense of power and life satisfaction," *J. Vocat. Behav.*, vol. 74, no. 2, pp. 145-155, 2009.

[10] S. Fantinelli, C. Esposito, L. Carlucci, P. Limone, and F. Sulla, "The influence of individual and contextual factors on the vocational choices of adolescents and their impact on well-being," *Behav. Sci.*, vol. 13, no. 3, p. 233, 2023.

[11] X. Wang, Y. Gao, Q. Wang, and P. Zhang, "Relationships between self-efficacy and teachers' well-being in middle school English teachers: The mediating role of teaching satisfaction and resilience," *Behav. Sci.*, vol. 14, no. 8, p. 629, 2024.

[12] X. Zhou, Y. Padrón, H. C. Waxman, E. Baek, and S. Acosta, "How do school climate and professional development in multicultural education impact job satisfaction and teaching efficacy for STEM teachers of English learners? A path-analysis," *Int. J. Sci. Math. Educ.*, vol. 22, no. 2, pp. 447-468, 2024.

[13] M. Wang and S. Liu, "Machine learning-based research on the adaptability of adolescents to online education," *arXiv preprint arXiv:2408.16849*, 2024.

[14] A. Aburayya, S. Salloum, K. Alderbashi, F. Shwedeh, Y. Shaalan, R. Alfaisal, ... and K. Shaalan, "SEM-machine learning-based model for perusing the adoption of metaverse in higher education in UAE," *Int. J. Data Netw. Sci.*, vol. 7, no. 2, pp. 667-676, 2023.

[15] A. Rejeb, K. Rejeb, A. Appolloni, H. Treiblmaier, and M. Iranmanesh, "Exploring the impact of ChatGPT on education: A web mining and machine learning approach," *Int. J. Manag. Educ.*, vol. 22, no. 1, p. 100932, 2024.

[16] A. Serek, A. Akhmetov, N. Ismagulov, B. Rysbek, B. Rysbek, and S. Alim, "Application of k-means in the perception of supervisors from students' side," in *2021 16th International Conference on Electronics Computer and Computation (ICECCO)*, pp. 1-3, Nov. 2021.

[17] I. T. Sanusi, S. S. Oyelere, H. Vartiainen, J. Suhonen, and M. Tukiainen, "A systematic review of teaching and learning machine learning in K-12 education," *Educ. Inf. Technol.*, vol. 28, no. 5, pp. 5967-5997, 2023.

[18] A. Serek, A. Bazarkulova, A. Chazhabayev, and A. Akhmetov, "Analysis of supervisors and students in the context of diploma defense," in *2021 16th International Conference on Electronics Computer and Computation (ICECCO)*, pp. 1-4, Nov. 2021.

[19] D. Akiba and M. C. Fraboni, "AI-supported academic advising: Exploring ChatGPT's current state and future potential toward student empowerment," *Educ. Sci.*, vol. 13, no. 9, p. 885, 2023.

[20] A. A. Makki, A. Y. Alqahtani, R. M. Abdulaal, and A. I. Madbouly, "A novel strategic approach to evaluating higher education quality standards in University Colleges using multi-criteria decision-making," *Educ. Sci.*, vol. 13, no. 6, p. 577, 2023.

[21] M. S. Iswahyudi, N. Nofirman, I. K. A. Wirayasa, S. Suharni, and I. Soegiarto, "Use of ChatGPT as a decision support tool in human resource management," *Jurnal Minfo Polgan*, vol. 12, no. 1, pp. 1522-1532, 2023.

[22] A. Abulibdeh, E. Zaidan, and R. Abulibdeh, "Navigating the confluence of artificial intelligence and education for sustainable development in the era of industry 4.0: Challenges, opportunities, and ethical dimensions," *J. Clean. Prod*., p. 140527, 2024.

[23] C. C. Thelma, Z. H. Sain, D. L. Mpolomoka, W. M. Akpan, and M. Davy, "Curriculum design for the digital age: Strategies for effective technology integration in higher education," *Int. J. Res.*, vol. 11, no. 07, pp. 185-201, 2024.

[24] Y. Gao, X. Wang, and P. Fan, "Exploring male English major's motivation trajectory through complex dynamic systems theory," *Curr. Psychol*., vol. 43, no. 10, pp. 9089-9100, 2024.

[25] S. Amer-Yahia, "Towards AI-powered data-driven education," *Proc. VLDB Endowment*, vol. 15, no. 12, pp. 3798-3806, 2022.

[26] C. Guan, J. Mou, and Z. Jiang, "Artificial intelligence innovation in education: A twenty-year data-driven historical analysis," *Int. J. Innov. Stud.*, vol. 4, no. 4, pp. 134-147, 2020.

[27] S. Gaftandzhieva, S. Hussain, S. Hilcenko, R. Doneva, and K. Boykova, "Data-driven decision making in higher education institutions: State-of-play," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 397-405, 2023.

[28] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, no. 1, p. 11, 2022.

[29] H. Luan and C. C. Tsai, "A review of using machine learning approaches for precision education," *Educ. Technol. Soc.*, vol. 24, no. 1, pp. 250-266, 2021.

[30] F. Ofori, E. Maina, and R. Gitonga, "Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review," *J. Inf. Technol.*, vol. 4, no. 1, pp. 33-55, 2020.

[31] R. B. Rolf, I. Jackson, M. Müller, S. Lang, T. Reggelin, and D. Ivanov, "A review on reinforcement learning algorithms and applications in supply chain management," *Int. J. Prod. Res.*, vol. 61, no. 20, pp. 7151-7179, 2023.

[32] Y. Wang, "Artificial intelligence in educational leadership: a symbiotic role of human-artificial intelligence decision-making," *J. Educ. Adm.*, vol. 59, no. 3, pp. 256-270, 2021.

[33] L. S. Neuwirth, S. Jović, and B. R. Mukherji, "Reimagining higher education during and post-COVID-19: Challenges and opportunities," *J. Adult Cont. Educ.*, vol. 27, no. 2, pp. 141-156, 2021.

[34] N. Almazova, E. Krylova, A. Rubtsova, and M. Odinokaya, "Challenges and opportunities for Russian higher education amid COVID-19: Teachers' perspective," *Educ. Sci.*, vol. 10, no. 12, p. 368, 2020.

[35] C. Rapanta, L. Botturi, P. Goodyear, L. Guàrdia, and M. Koole, "Balancing technology, pedagogy and the new normal: Post-pandemic challenges for higher education," *Postdigit. Sci. Educ.*, vol. 3, no. 3, pp. 715-742, 2021.

[36] E. Elder, A. Struminger, A. Wilkerson, S. Wilkins, M. Rosenthal, and E. Post, "A Social Ecological Approach to Understanding Youth Sport Specialization," *J. Sport Behav.*, vol. 47, no. 2, 2024.

[37] M. Kroher and K. Leuze, "Degree Differentiation and Changing Career Outcomes of Higher Education Graduates in Germany: A Matter of Specialization, Extracurricular Activities or Labor Market Segmentation?," in *Career Paths Inside and Outside Academia*, pp. 20-54, 2023.

[38] Z. Shu, K. Mottan, H. Chen, and Y. Pang, "Chinese Teachers' Perceptions on Perceived Teacher Support and Student Engagement," *Environ.-Behav. Proc. J.*, vol. 9, no. 30, pp. 57-62, 2024.

[39] C. Fischer, E. Witherspoon, H. Nguyen, Y. Feng, S. Fiorini, P. Vincent-Ruz, and C. Schunn, "Advanced placement course credit and undergraduate student success in gateway science courses," *J. Res. Sci. Teach.*, vol. 60, no. 2, pp. 304-329, 2023.

[40] M. Abou Naaj, R. Mehdi, E. A. Mohamed, and M. Nachouki, "Analysis of the factors affecting student performance using a neuro-fuzzy approach," *Educ. Sci.*, vol. 13, no. 3, p. 313, 2023.

[41] S. M. F. D. Syed Mustapha, "Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods," *Appl. Syst. Innov.*, vol. 6, no. 5, p. 86, 2023.

[42] N. Ribeiro, C. Malafaia, T. Neves, and I. Menezes, "The impact of extracurricular activities on university students' academic success and employability," *Eur. J. Higher Educ.*, vol. 14, no. 3, pp. 389-409, 2024.

[43] V. Dang. *Employability development through extracurricular activities: A case study of AIESEC membership*. Bachelor's thesis, 2023.

[44] N. Shah, S. Bano, U. N. Saraih, N. A. A. Abdelwahed, and B. A. Soomro, "Leading towards the students' career development and career intentions through using multidimensional soft skills in the digital age," *Educ. Train.*, vol. 65, no. 6/7, pp. 848-870, 2023.

[45] G. Chapman, W. Emambocus, and D. Obembe, "Higher education student motivations for extracurricular activities: Evidence from UK universities," *J. Educ. Work*, vol. 36, no. 2, pp. 138-152, 2023.

[46] O. O. Ayeni, C. C. Unachukwu, N. M. Al Hamad, O. N. Chisom, and O. E. Adewusi, "The impact of robotics clubs on K-12 students' interest in STEM careers," *Magna Scientia Adv. Res. Rev.*, vol. 10, no. 1, pp. 361-367, 2024.

[47] O. O. Ayeni, C. C. Unachukwu, N. M. Al Hamad, B. Osawaru, and O. E. Adewusi, "A multidisciplinary approach to STEM education: Combining HR, counseling, and mentorship," Magna Scientia Advanced Research and Reviews, vol. 10, no. 1, pp. 351-360, 2024.

[48] Z. Zhong, H. Guo, and K. Qian, "Deciphering the impact of machine learning on education: Insights from a bibliometric analysis using bibliometrix R-package," *Educ. Inf. Technol.*, pp. 1-28, 2024.

[49] S. Elbanna and L. Armstrong, "Exploring the integration of ChatGPT in education: Adapting for the future," *Manag. Sustainability: An Arab Rev.*, vol. 3, no. 1, pp. 16-29, 2024.

[50] P. Shah, *AI and the Future of Education: Teaching in the Age of Artificial Intelligence*, John Wiley & Sons, 2023.

[51] A. Dirin, C. Adriana, and S. Saballe, "Machine learning models to predict students' study path selection," *Int. J. Interact. Mobile Technol.*, vol. 16, no. 1, pp. 158-183, 2022, doi: 10.3991/ijim.v16i01.20121.

[52] A. Fizar, M. H. Imam, B. Sheak, R. H. Noori, T. R. Habibur, R. Hemal, M. S. Arefin, "The comparison of machine learning algorithms to find the career path by Bloom's Taxonomy evaluation," *International Conference on Big Data, IoT and Machine Learning*, pp. 747-761, 2024, doi: 10.1007/978-981-99-8937-9_50.

[53] T. Althubiti, T. M. Ahmed, and O. Alassafi, "Developing a predictive model for selecting academic track via GPA by using classification algorithms: Saudi universities as case study," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, 2023, doi: 10.14569/ijacsa.2023.0140539.

[54] W. Alghamdi, "A comparative analysis on machine learning based student placement prediction," *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, pp. 986-991, 2023, doi: 10.1109/icscna58489.2023.10370392.

[55] Y. Terawaki, T. Unoki, T. Kato, and Y. Kodama, "Performance evaluation for four types of machine learning algorithms using educational open data," *Smart Education and e-Learning 2019*, pp. 281-289, 2019, doi: 10.1007/978-981-13-8260-4_26.

[56] N. S. Brohi, T. Ramiah Pillai, S. K. Kaur, H. K. Kaur, S. Sukumaran, and D. Asirvatham, "Accuracy comparison of machine learning algorithms for predictive analytics in higher education," *Emerging Technologies in Computing: Second International Conference, iCETiC 2019, London, UK, August 19–20, 2019, Proceedings 2*, pp. 254-261, 2019, doi: 10.1007/978-3-030-23943-5_19.

[57] H. S. Thapa, "Development of employability skills through work-based learning," *Journal of Technical and Vocational Education and Training (TVET)*, vol. 18, no. 1, pp. 102-111, 2024.

[58] C. O. Elom, U. C. Okolie, C. C. Uwaleke, C. C. Umoke, S. O. Abonyi, and A. O. Nwele, "The effect of openness to experience on students' readiness for school-to-work transition," *Journal of Career Assessment*, vol. 32, 2024.

[59] T. X. Tolqunovna and M. N. Leonodovna, "Understanding the pedagogical basis of students' interest in professions: A comprehensive analysis," *International Journal of Advance Scientific Research*, vol. 4, no. 3, pp. 118-126, 2024.

[60] A. Samagaio, P. Morais Francisco, and T. Felício, "The relationship between soft skills, stress and reduced audit quality practices," *Review of Accounting and Finance*, vol. 23, no. 3, pp. 353-374, 2024.

[61] A. Chelis, *In the heart of data: machine learning applications for improved heart failure outcome prediction*. Master's thesis, Πανεπιστήμιο Πειραιώς, 2024.

[62] T. A. Shaikh, T. Rasool, P. Verma, and W. A. Mir, "A fundamental overview of ensemble deep learning models and applications: systematic literature and state of the art," *Annals of Operations Research*, pp. 1-77, 2024.

[63] K. U. Apu, M. Ali, M. F. Islam, and M. Miah, "A systematic literature review on AI approaches to address data imbalance in machine learning," *Frontiers in Applied Engineering and Technology*, vol. 2, no. 01, pp. 58-77, 2025.

[64] X. Lu and L. Wei, "Teaching reform and innovation of vocational development and employment guidance courses in colleges and universities based on random forest model," *Applied Mathematics and Nonlinear Sciences*, 2024.

[65] S. Bankins, S. Jooss, S. L. D. Restubog, M. Marrone, A. C. Ocampo, and M. Shoss, "Navigating career stages in the age of artificial intelligence: A systematic interdisciplinary review and agenda for future research," *Journal of Vocational Behavior*, p. 104011, 2024.

[66] L. R. Pelima, Y. Sukmana, and Y. Rosmansyah, "Predicting university student graduation using academic performance and machine learning: a systematic literature review," *IEEE Access*, vol. 12, pp. 23451-23465, 2024.

[67] N. A. Kuadey, C. Ankora, F. Tahiru, L. Bensah, C. C. M. Agbesi, and S. O. Bolatimi, "Using machine learning algorithms to examine the impact of technostress creators on student learning burnout and perceived academic performance," *International Journal of Information Technology*, vol. 16, no. 4, pp. 2467-2482, 2024.

[68] M. A. Ayanwale, R. R. Molefi, and S. Oyeniran, "Analyzing the evolution of machine learning integration in educational research: a bibliometric perspective," *Discover Education*, vol. 3, no. 1, pp. 47, 2024.

[69] A. Kukkar, R. Mohana, A. Sharma, and A. Nayyar, "A novel methodology using RNN+ LSTM+ ML for predicting student's academic performance," *Education and Information Technologies*, pp. 1-37, 2024.

[70] C. Song, S. Y. Shin, and K. S. Shin, "Implementing the Dynamic Feedback-Driven Learning Optimization Framework: A machine learning approach to personalize educational pathways," *Applied Sciences*, vol. 14, no. 2, p. 916, 2024.

[71] A. B. Feroz Khan and S. R. A. Samad, "Evaluating online learning adaptability in students using machine learning-based techniques: A novel analytical approach," *Educ. Sci. Manag.*, vol. 2, no. 1, pp. 25-34, 2024.

[72] R. AlAli, Y. Wardat, S. Saleh, and N. Alshraifin, "Evaluation of STEM-Aligned Teaching Practices for Gifted Mathematics Teachers," *European Journal of STEM Education*, vol. 9, no. 1, p. 08, 2024.

[73] D. Al Kez, C. Lowans, and A. Foley, "Sustainable Development in Third Level Programs: Distilling a Pathway to a True Net-Zero Education," *Sustainability*, vol. 16, no. 5, p. 1998, 2024.

[74] H. E. Gandolfi, "Teaching in the age of environmental emergencies: A 'utopian' exploration of the experiences of teachers committed to

environmental education in England," *Educational Review*, vol. 76, no. 7, pp. 1786-1806, 2024.

[75] M. Renkema and A. Tursunbayeva, "The future of work of academics in the age of Artificial Intelligence: State-of-the-art and a research roadmap," *Futures*, vol. 163, p. 103453, 2024.

[76] E. Langran, M. Searson, and J. Trumble, "Transforming Teacher Education in the Age of Generative AI," in *Exploring New Horizons: Generative Artificial Intelligence and Teacher Education*, vol. 2, 2024.

[77] N. A. M. Zin and M. S. Mahmud, "Perceptions of Malaysian University Mathematics Instructors of the Challenges they Face in Implementing Effective Distance Learning," *International Journal of Learning, Teaching and Educational Research*, vol. 23, no. 5, pp. 158-179, 2024.

[78] T. Feraco, D. Resnati, D. Fregonese, A. Spoto, and C. Meneghetti, "An integrated model of school students' academic achievement and life satisfaction. Linking soft skills, extracurricular activities, self-regulated learning, motivation, and emotions," *European Journal of Psychology of Education*, vol. 38, no. 1, pp. 109-130, 2023.

[79] J. C. H. So, Y. H. Ho, A. K. L. Wong, H. C. Chan, K. H. Y. Tsang, A. P. L. Chan, and S. C. W. Wong, "Analytic study for predictor development on student participation in generic competence development activities based on academic performance," *IEEE Transactions on Learning Technologies*, vol. 16, no. 5, pp. 790-803, 2023.

[80] X. Li, S. Duan, and H. Liu, "Unveiling the predictive effect of students' perceived EFL teacher support on academic achievement: the mediating role of academic buoyancy," *Sustainability*, vol. 15, no. 13, p. 10205, 2023.