Machine Learning Research
2026, Vol. 11, No. 1, pp. 53–62
https://doi.org/10.11648/j.mlr.20261101.15

Research Article

Predicting Girls' Orientation Toward Technical Tracks in the Republic of Guinea Using Machine Learning: A Comparative Study of Models and Determinants Analysis

Djiba Kourouma[1, *], Mamadou Mouctar Diallo[1], Binko Mamady Toure[2]

[1]Polytechnic Institute, Gamal Abdel Nasser University, Conakry, Guinea
[2]Computer Center, Gamal Abdel Nasser University, Conakry, Guinea

Abstract

The purpose of this study is to develop a machine learning model capable of predicting girls' orientation toward technical tracks in the Republic of Guinea. The dataset was constructed from the university placement records of high school graduates and includes academic and socio-demographic variables related to students' orientation toward technical and non-technical tracks. In addition, the study identifies the variables associated with these orientations. To achieve this goal, four supervised learning algorithms, including Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM) were used. The evaluation of the algorithms' performance was based on the metrics Accuracy, Precision, F1-score, Recall, and AUC. The results show that the Random Forest model performs best, with an accuracy of 87.7%, an F1-score of 82.1%, and an AUC ROC of 0.954. Analysis of the variables reveals that the overall average score is the primary factor guiding girls toward technical tracks. This research highlights the importance of machine learning methods as a decision-making tool for policies aimed at education and the promotion of technical tracks among girls. Models were evaluated using an independent test set and a 5-fold stratified cross-validation procedure to assess robustness.

Keywords

Prediction, Girls' Orientation, Technical Tracks, Machine Learning, Predictive Modeling, Republic of Guinea

1. Introduction

Girls' academic success is a major concern worldwide. Despite progress in this area, disparities persist regarding their access to technical tracks of study, particularly in developing countries.

According to Nuñez et al., in Latin America and the Caribbean, women represent approximately 30% of STEM graduates. In disciplines such as engineering, their participation rarely exceeds 20% [1].

According to UNICEF, young women represent only 25% of students in engineering and ICT fields in countries with available data [2].

This disparity is a complex problem that requires in-depth analysis.

Machine learning has emerged as a powerful tool for analyzing and extracting information from data [3].

Olinmah et al. report that predictive modeling and machine

*Correspondence: Djiba Kourouma (kouroumadji@gmail.com)

Received: 24 May 2026; **Accepted:** 8 June 2026; **Published:** 26 June 2026

learning have seen considerable growth in educational research and practice, particularly for predicting academic outcomes, identifying at-risk students, and personalizing learning pathways [4]. This study aims to develop a predictive model for girls' educational and career paths in technical tracks using machine learning algorithms. This modeling involves a comparative evaluation of frequently used supervised learning algorithms (Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)) to achieve optimal performance. Metrics such as precision, accuracy, f1 score, recall, and AUC are used to evaluate the algorithms' performance. A dataset containing information on 690 students was created for training these algorithms, based on the university placement database for high school graduates. The main activities of this research include dataset creation, data preprocessing, machine learning model design and evaluation, identification of the best model based on metrics, determination of the importance of variables in prediction, and results analysis.

In addition to the conventional train-test evaluation strategy, a stratified cross-validation procedure was employed to assess the robustness and generalization capability of the developed predictive models.

The success of this approach necessarily involves a literature review to understand the state of the art.

Although machine learning has been widely applied to educational prediction problems such as academic performance, dropout detection, and career guidance, very limited research has focused specifically on predicting girls' orientation toward technical tracks in Sub-Saharan Africa. To the best of our knowledge, no previous study has investigated this issue in the context of the Republic of Guinea using machine learning techniques.

This study contributes to the literature in three ways. First, it develops and compares four supervised machine learning algorithms for predicting girls' orientation toward technical tracks. Second, it identifies the most influential factors associated with technical-track orientation. Third, it provides empirical evidence that may support educational policymakers and guidance services in promoting girls' participation in technical education.

2. Literature Review

Several researchers have explored the use of machine learning for prediction in educational settings, and their investigations form the basis of this study. Among these authors, we selected:

1) Rastrollo-Guerrero et al. reviewed 64 studies and reported that approximately 70% focused on predicting student academic performance, while supervised learning approaches accounted for nearly half of the predictive techniques employed. The most frequently used algorithms included Support Vector Machines, Random Forests, Decision Trees, and Naïve Bayes classifiers. These findings highlight the effectiveness of supervised learning methods in educational data mining and support their application to predicting girls' orientation toward technical tracks in Guinea [5].
2) KANDUKI KIVUYIRWA MYSTERE, et al., worked on "predicting students' placement in appropriate fields of study using data mining techniques." They emphasized the application of classic data mining techniques (Naïve Bayes, networks, trees, SVMs) to recommend fields of study. Their central idea is to evaluate several families of algorithms to choose the one that best generalizes placement data. They demonstrate the value of systematic comparison and stress the importance of attribute selection [6].
3) Solemane Coulibaly and Djibril Diarra focused on "A machine learning approach to predict school dropout." These authors apply machine learning methods to detect the risk of school dropout/disengagement; their approach combines behavioral feature extraction and ensemble models (XGBoost, AdaBoost) to prioritize preventive actions. They demonstrate that ensemble models often offer high sensitivity for identifying at-risk students [7].
4) Hussein Altabrawee et al., in their work on "Predicting Students' Performance Using Machine Learning Techniques," compared neural networks, naive Bayes, decision trees, and logistic regression for predicting success. They emphasize the importance of feature engineering and show that, with small datasets, networks do not always outperform simpler methods [8].
5) SAOUABI Mohamed (2021), in his study "Machine Learning for Predicting Employability in Morocco in a Big Data Environment," focused on employability within a Big Data context, combining preprocessing on large volumes of data with classical algorithms (trees, regression, Naive Bayes). The author emphasizes the need for scalable architectures (streaming/pipelines) to leverage diverse data sources [9].
6) Cristóbal Romero, in collaboration with others, offers reviews, methodological frameworks, and best practices for analyzing training data (preprocessing, feature selection, cross-validation adapted to educational data). His major contribution is the formalization of evaluation techniques and the identification of error sources specific to educational data [10].
7) Sebastián Ventura, a regular co-author of Romero's work, studies the practical application of educational data mining algorithms (trees, k-NN, SVM, ensemble methods) and emphasizes the interpretability of models so that educational stakeholders (teachers, policymakers) can act based on predictions [10].
8) Ryan S. J. d. Baker works on detecting learning behaviors (engagement, off task, gaming the system) from interaction traces and on using these signals to improve the prediction of performance and dropout rates. His key contribution is the integration of behavioral variables (logs) in addition to demographic/grade data [11].

9) Joost Dekker (Dekker, Pechenizkiy & Vleeshouwers) [12], in their work on performance and dropout prediction, examine several classical algorithms and discuss sampling problems, class imbalance, and interpretability; they recommend hybrid approaches (metrics + rules) for operational application in schools.
10) Neil Heffernan (and the ASSISTments team) is known for his work on tutoring systems and the accurate prediction of concept mastery: he combines trace models (item response, knowledge tracing) and machine learning algorithms to predict success at the granularity of exercises, providing real-time feedback for retraining/adapting the teaching intervention [13].
11) Matz SC et al. investigated the use of machine learning to predict student retention based on sociodemographic characteristics and measured engagement indicators. They analyzed the records of 50,095 students from four U.S. universities and community colleges and demonstrated that combining these data allows for highly accurate prediction of dropout (average AUC for linear and nonlinear models = 78%; maximum AUC = 88%). They also found that behavioral variables related to engagement, which reflect students' college experience, provided additional predictive value compared to institutional variables such as GPA or ethnicity [14].

The algorithms used by these authors, the performance achieved with the best algorithm, and the data samples on which the models were trained are summarized in the table below.

Table 1. *Summary of algorithms used by the cited authors.*

Authors	Algorithms Used	Best Performance	Sample (typical examples)
MYSTERE, et al.	Naive Bayes, Neural Network, Decision Tree, SVM	SVM: Accuracy (70%)	712 with 28 attributes.
S. Coulibaly & D. Diarra	XGBoost, AdaBoost, Decision Tree, Random Forest	XGBoost: f1-score (92,5%)	3600.
H. Altabrawee et al.	Neural networks, Naive Bayes, Decision tree, Logistic regression	Neural networks: Accuracy (77.04%)	161 with 20 attributes.
M. SAOUABI (2021)	Decision tree, Logistic regression, Naive Bayes	Decision tree: accuracy 81.70%.	1752 with 22 attributes.
Cristóbal Romero	decision tree, Naive Bayes, SVM, ensemble methods (Random Forest, Boosting)	Performance varies depending on the task; assessments often show an accuracy/F1 between 70 and 90% depending on the quality of the features; the strength is the assessment methodology (adapted cross-validation).	Review studies: aggregates of studies — sets of a few hundred to several thousand examples (depending on the case).
Sebastián Ventura	k-NN, Decision Trees, SVM, Random Forest, Bagging/Boosting	Better performance achieved by ensemble methods on many educational tasks; Ventura insists on interpretability (trees) as a practical criterion.	Varied datasets; typically, hundreds to thousands of instances.
Ryan S. J. d. Baker	Features engineering based on logs + Random Forest, SVM, Logistic Regression	Adding behavioral variables significantly improves recall/abandonment detection (notable gains vs. demographic variables only).	Studies with learning logs, samples ranging from a few hundred to thousands.
Joost Dekker et al.	Several common classifiers in Weka: Decision Trees, Naïve Bayes, SVM, ensemble methods; focus on class imbalance	Simple models with accuracy rates ranging from 75% to 80%.	Studies: hundreds of students (depending on the source).
Matz SC et al.	a linear classifier (Elastic Net; implemented in glmnet 4.1–4) 70, 71 and a nonlinear classifier (Random Forest; implemented in randomForest 4.7–1) 72, 73.	high accuracy (average AUC for linear and nonlinear models = 78%; maximum AUC = 88%).	de-identified data from four institutions with a total of 50,095 students (min = 476, max = 45,062).

Authors	Algorithms Used	Best Performance	Sample (typical examples)
Neil Heffernan (ASSISTments)	Knowledge Tracing (Bayesian/Deep), Logistic Regression, Random Forests, sequential models	Very good performance at exercise-to-exercise granularity; shows the value of temporal/tracing models for improving immediate learning predictions.	Tutor system data: tens of thousands of interactions (logs) often; individuals: hundreds to thousands.

The literature demonstrates that machine learning is widely used in the education sector. It offers promising perspectives but remains relatively unexplored in West African countries, particularly Guinea. This research aims to partially fill this gap by developing a predictive model for student guidance.

To fill this gap, the present study seeks to answer the following research question: which machine learning algorithm can most accurately predict the orientation of girls towards technical fields in the Republic of Guinea, and what factors most strongly influence this orientation?

3. Methodology

3.1. Dataset Construction

Raw data on high school graduates' academic tracks served as the basis for constructing the data set. From these data, relevant variables were extracted, including orientation track, school type, area, parental income, baccalaureate average score, and baccalaureate option. The final dataset contains 690 observations, five explanatory variables, and one target variable representing the orientation track, coded as 1 for Technical and 0 for Non-Technical.

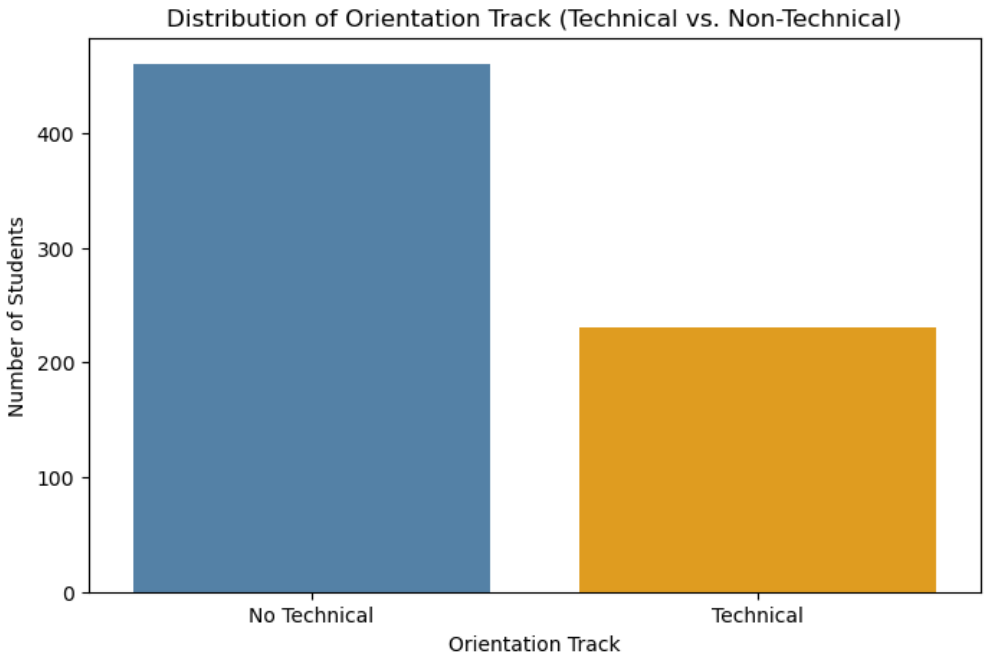

Figure 1. *Distribution of Students by Orientation Track.*

The distribution of students by orientation track is shown in Figure 1, while the correlation matrix of the variables is presented in Figure 2.

This Figure shows the distribution of students according to their orientation track. Approximately two-thirds of the observations belong to the non-technical class, while one-third correspond to technical-track orientations. Although the dataset exhibits a moderate class imbalance, the minority class remains sufficiently represented for supervised learning. Therefore, no resampling technique was applied. To ensure a balanced evaluation of model performance, multiple metrics including Precision, Recall, F1-score, and AUC were considered in addition to Accuracy.

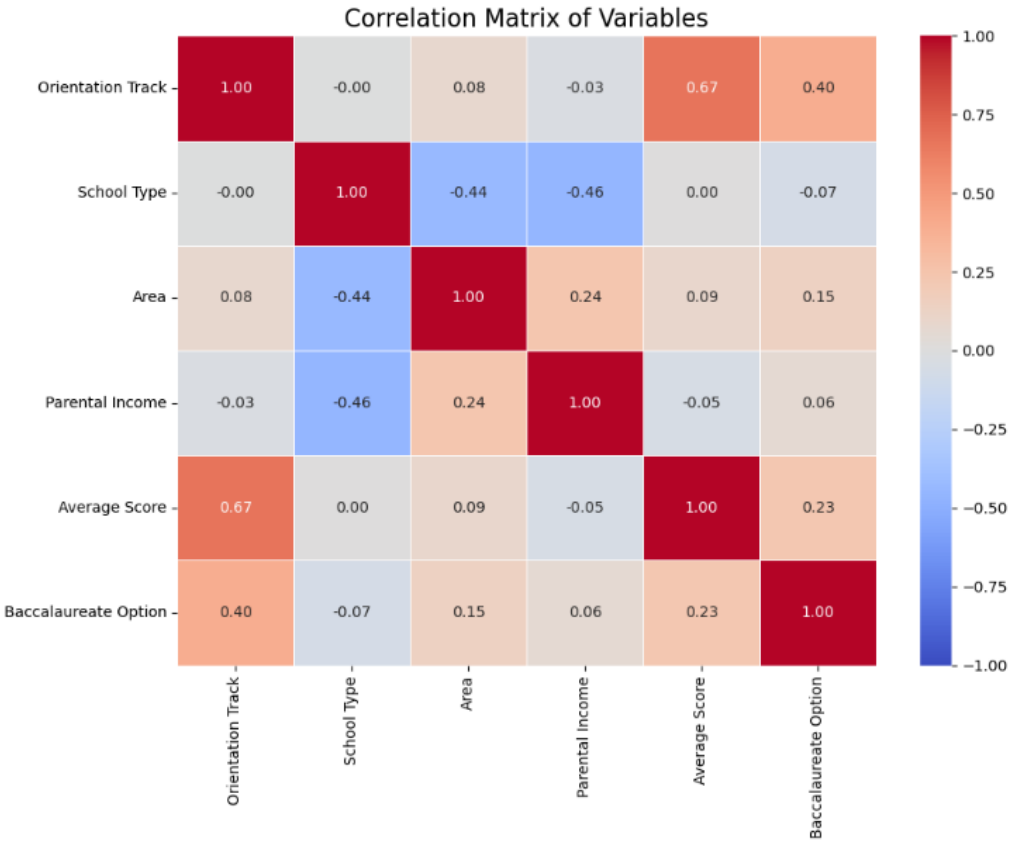

Figure 2. *Correlation Matrix of Variables.*

Table 2. *Class Distribution.*

Orientation Track	Number of Students	Percentage
Non-Technical	460	66.7%
Technical	230	33.3%
Total	690	100%

The correlation matrix provides an exploratory overview of the linear associations between the encoded variables. However, since several variables are categorical, these correlations should be interpreted with caution and complemented by model-based feature importance analysis. To assess the contribution of each explanatory variable to the prediction, the feature importance scores of the Random Forest model were analyzed (Figure 5).

3.2. Machine Learning Techniques

The model development followed these steps:

1) Data loading: from an Excel file;

Data preprocessing: It is essential for building machine learning models, as it ensures that the data is suitable for analysis and prediction [15]. For this study, the dataset was reviewed to verify its quality and consistency. No missing values or significant outliers were identified. Categorical variables were encoded numerically to make them compatible with machine learning algorithms, and the target variable was transformed into a binary variable where 1 corresponds to a technical orientation and 0 to a non-technical orientation;

2) Feature selection: Identification of the most relevant variables for the predictive model using appropriate Python tools.
3) Dataset split: into training data (80%) and test data (20%);
4) Model training: The training data was used to train the model with selected machine learning algorithms (Logistic Regression, Decision Trees, Random Forests, Support Vector Machines);
5) Model evaluation: Five metrics were used to evaluate the model: precision, accuracy, f1 score, recall, and area under the ROC curve (AUC);
6) The Random Forest classifier was implemented using the Scikit-learn library. The model was trained with the default parameter settings provided by the framework, while a fixed random seed (random_state = 42) was used to ensure the reproducibility of the results.

The table below presents the hyperparameters of Random Forest.

Table 3. Random Forest Hyperparameters.

Parameter	Value
n_estimators	100
max_depth	None
min_samples_split	2
min_samples_leaf	1
criterion	gini
random_state	42

The influence of input variables on the prediction was determined.

3.3. Cross-Validation Procedure

To evaluate the stability and robustness of the predictive models, a 5-fold stratified cross-validation was performed on the training dataset. Stratification ensured that the class distribution was preserved across all folds. For each fold, Accuracy, Precision, Recall, F1-score, and AUC were computed and averaged to obtain reliable performance estimates.

4. Results and Analysis

The results of the model evaluation for the different machine learning algorithms tested are presented in the table below:

Table 4. Model evaluation results for the tested algorithms.

Model	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	0.877	0.796	0.848	0.821	0.954
Logistic Regression	0.855	0.810	0.739	0.773	0.943
SVM	0.841	0.962	0.543	0.694	0.944
Decision Tree	0.826	0.696	0.848	0.765	0.842

The results show that the Random Forest model achieves the best overall performance, with an accuracy of 87.7% and an F1-score of 82.1%. The SVM model exhibits high precision but relatively low recall, indicating difficulty in correctly identifying all girls oriented toward technical tracks.

The comparison of the ROC curves of the tested algorithms is shown in the Figure below:

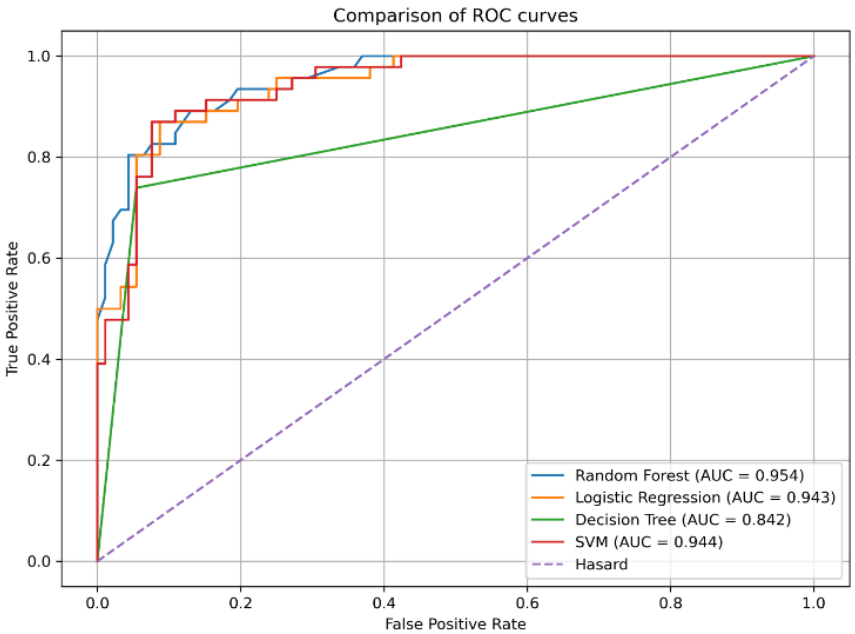

Figure 3. Comparison of ROC curves of the tested algorithms.

The ROC curves show that the Random Forest model achieves the best discriminatory power with an AUC of 0.954, followed by the SVM with an AUC of 0.944 and Logistic Regression with an AUC of 0.943. These results confirm the robustness of the Random Forest model for predicting orientation toward technical tracks.

The confusion matrix of this model is shown in Figure 4 below.

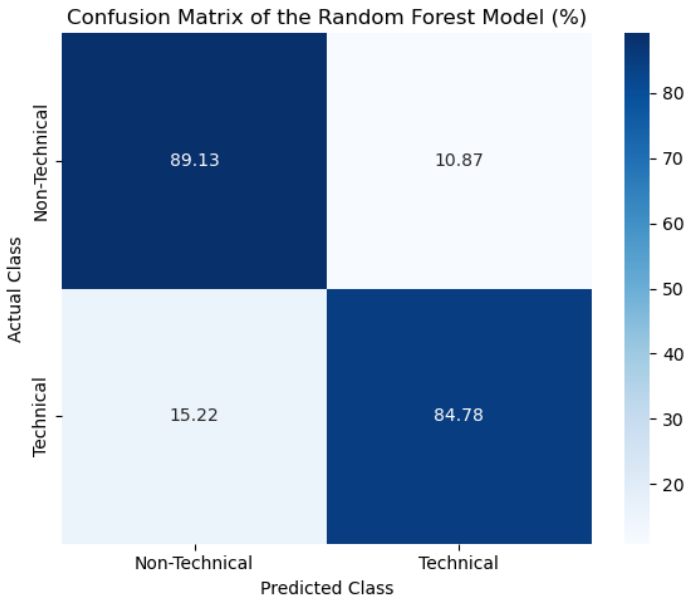

Figure 4. *The confusion matrix of the Random Forest model.*

The confusion matrix for the Random Forest model shows that 89.13% of students in the "non-technical" class are correctly classified, while 84.78% of students in the "Technical" class are correctly identified. These results indicate that the model performs well in distinguishing between the two orientation tracks.

The importance of the variables on the prediction is shown in the Figure below:

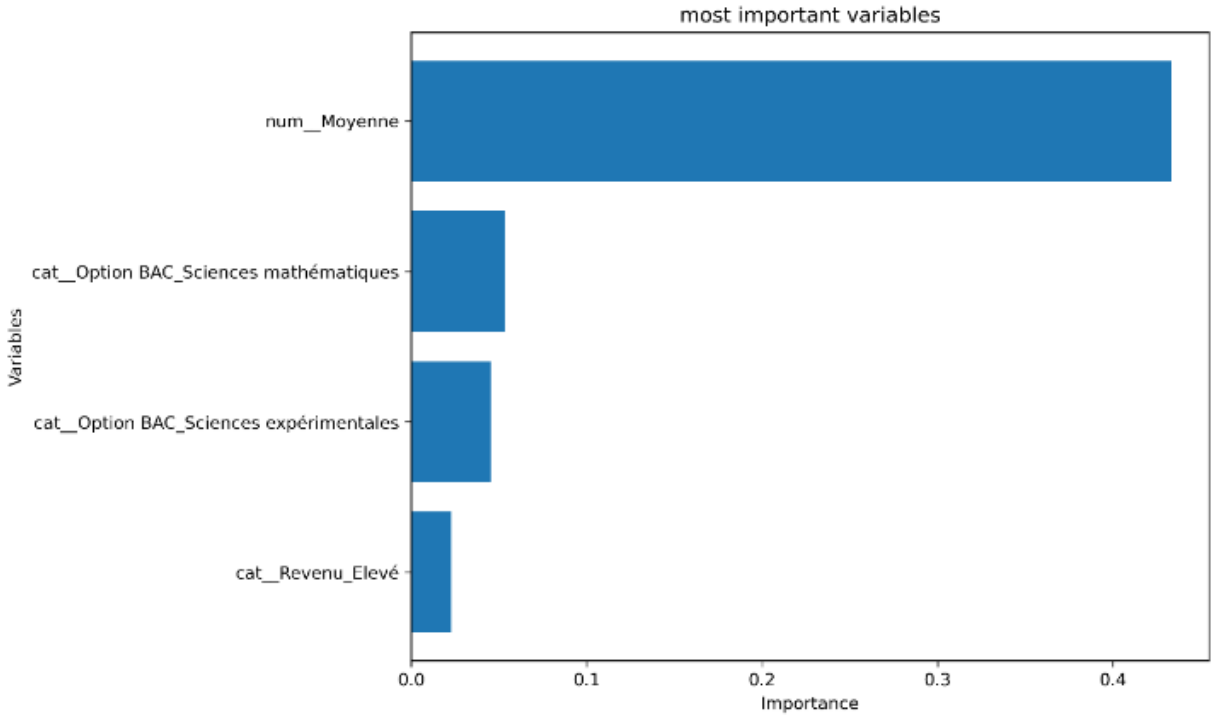

Figure 5. *Feature Importance Scores of the Random Forest Model.*

An analysis of the significance of the variables reveals that the overall grade point average is the most influential factor in predicting girls' orientation toward technical tracks. The subjects chosen for the BAC, particularly Mathematics and Experimental Sciences, also emerge as significant variables. Income level has a more moderate influence.

The cross-validation results confirm the robustness of the Random Forest model, which achieved the highest average Accuracy (0.891), F1-score (0.832), and AUC (0.947). The consistency between cross-validation and test-set results indicates good generalization capability.

Table 5. *The results of the 5-fold cross-validation.*

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.859	0.783	0.805	0.792	0.946
SVM	0.766	0.986	0.304	0.462	0.940
Random Forest	0.891	0.862	0.804	0.832	0.947
Decision Tree	0.870	0.809	0.799	0.803	0.854

5. Discussion

The Random Forest model stands out as the best-performing model among those examined. This performance can be attributed to its ability to handle nonlinear relationships between variables, as well as its capacity to reduce the risk of overfitting by aggregating multiple decision trees.

Although the dataset exhibits a moderate class imbalance, no specific resampling technique was applied. The stratified cross-validation results show stable Recall values across folds, particularly for the Random Forest model (Recall = 0.804), suggesting that the observed imbalance did not substantially affect the predictive performance.

Logistic Regression also performs well, suggesting that the relationships between certain variables and orientation track choice may be partly linear. Nevertheless, tree-based techniques remain more suitable for the complexity of educational data.

The predominance of the average score highlights that academic performance has a significant impact on the orientation toward technical tracks.

This finding aligns with several previous studies showing that students with higher academic performance tend to gravitate more toward scientific and technical tracks.

The impact of Mathematics and Experimental Sciences programs also underscores that prior educational paths significantly influence program selection decisions.

The performance of the Random Forest classifier is consistent with findings reported in recent educational data mining studies. Mutsotsya et al. [16] showed that ensemble-based approaches provide robust prediction performance for academic outcomes, particularly when datasets contain heterogeneous academic and demographic characteristics. Similarly, Islam et al. [17] highlighted that machine learning models combined with explainable AI techniques can effectively support educational decision-making while maintaining strong predictive performance.

The present study also confirms the importance of academic achievement indicators in educational pathway prediction. In particular, the predominance of the average score as the most influential variable is in agreement with the findings of Lau and Abdul Rahman [18], who reported that academic performance remains one of the strongest predictors of future educational trajectories, even when demographic and behavioral variables are considered.

Furthermore, the significant contribution of baccalaureate options related to Mathematics and Experimental Sciences supports previous observations by Wang et al. [19], who demonstrated that prior academic specialization strongly influences career and educational choices. This suggests that girls who follow science-oriented pathways at the secondary level are more likely to pursue technical studies at the university level.

From a practical perspective, the proposed model could support educational authorities, guidance services, and policymakers in identifying students with a higher propensity toward technical tracks. Such predictive tools may contribute to the design of targeted interventions aimed at increasing female participation in technical and STEM-related programs, a challenge that remains important in many developing countries.

Finally, although the proposed model achieved satisfactory predictive performance, the relatively limited dataset remains a constraint. As emphasized by Kemper et al. [20], the integration of larger and more diverse datasets generally improves the generalization capacity of educational prediction models. Future work could therefore incorporate additional contextual variables such as parental support, career aspirations, access

to educational resources, and psychosocial factors.

6. Conclusion

The objective of this research was to explore the use of machine learning models to predict the choice of technical tracks of study among young women in the Republic of Guinea, based on academic and socioeconomic factors.

Following the testing of four classification algorithms namely Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—the results indicate that the Random Forest model stands out with an accuracy of 87.7%, an F1-score of 82.1%, and an AUC ROC of 0.954. The 5-fold stratified cross-validation confirmed the robustness and generalization capability of the Random Forest classifier. An examination of the variables indicates that overall grade point average is the primary factor associated with orientation toward technical tracks.

These findings highlight the importance of academic performance and science-related baccalaureate options in girls' orientation toward technical tracks. They also demonstrate that machine learning methods can help study academic guidance processes and identify the factors influencing educational trajectories.

However, this research has some limitations, particularly regarding the size of the dataset and the type of variables available. Future research could consider a greater number of contextual variables, such as career aspirations, family influence, or access to educational resources.

Among the avenues worth exploring is the possibility of expanding this research by incorporating more sophisticated analytical techniques, including model interpretation methods such as SHAP, to better understand the individual contributions of variables to predictions, the impact of advanced class-balancing techniques such as SMOTE, ADASYN, or cost-sensitive learning to further assess their influence on model performance.

Furthermore, the use of data from various institutions or regions could strengthen the robustness of the models and facilitate better generalization of the results.

Abbreviations

AUC	Area Under the Curve
EDM	Educational Data Mining
ICT	Information and Communication Technology
ML	Machine Learning
RF	Random Forest
ROC	Receiver Operating Characteristic
STEM	Science, Technology, Engineering and Mathematics
SVM	Support Vector Machine
DT	Decision Tree
LR	Logistic Regression
AI	Artificial Intelligence

Author Contributions

Djiba Kourouma: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Resources, Visualization, Writing – original draft

Mamadou Mouctar Diallo: Writing – review & editing, Project Administration, Funding Acquisition, Validation

Binko Mamady Toure: Supervision, Methodology, Validation, Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Y. Nuñez, J. C. Martinez Santos, M. Moreno, et E. Puertas, Predictive Model for STEM Vocational Guidance through Profile Analysis and Information Adaptation with a Gender Perspective. 2025. https://doi.org/10.18687/LEIRD2025.1.1.1122

[2] «Education and gender equality: what you need to know | UNESCO». Consulted April 18, 2026. [Online]. Available on: https://www.unesco.org/fr/gender-equality/education/need-know

[3] C. Meyer, D. Baogui, et M. A. Gouda, «Applying machine learning to gauge the number of women in science, technology, and innovation policy (STIP): a model to accommodate missing data», Humanit. Soc. Sci. Commun., vol. 12, nº 1, p. 1245, august 2025, https://doi.org/10.1057/s41599-025-05610-4

[4] F. Olinmah, B. Otokiti, O. Abiola-Adams, D. Abutu, et I. Okoli, «Integrating Predictive Modeling and Machine Learning for Class Success Forecasting in Creative Education Sectors», Int. J. Adv. Multidiscip. Res. Stud., vol. 3, p. 1796-1802, dec. 2023, https://doi.org/10.62225/2583049X.2023.3.6.4393

[5] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, et A. Durán-Domínguez, «Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review», Appl. Sci., vol. 10, no 3, p. 1042, janv. 2020, https://doi.org/10.3390/app10031042

[6] K. Mystere, H. Mpia, et V. Mutegheki Baraka, «Predicting student orientation into appropriate fields of study using Data Mining techniques», Int. J. Innov. Appl. Stud., vol. 39, p. 193-208, March 2023.

[7] S. Coulibaly and D. Diarra, «A machine learning approach to predicting school dropout, vol. 2, in Conference Proceedings», no. 14, tome 2. Bamako, Mali: MSAS Editions, 2024. Accessed: September 7, 2025. Available at: https://hal.science/hal-05042851

[8] H. Altabrawee, O. Ali, et S. Qaisar, «Predicting Students' Performance Using Machine Learning Techniques», J. Univ. BABYLON Pure Appl. Sci., vol. 27, p. 194-205, avr. 2019, https://doi.org/10.29196/jubpas.v27i1.2108

[9] M. SAOUABI, « Machine learning for predicting employability in Morocco in a Big Data environment », 2021, accessed February 22, 2025. Available at: https://toubkal.imist.ma/handle/123456789/25220

[10] C. Romero, S. Ventura, M. Pechenizkiy, et R. S. J. d Baker, Ed., Handbook of Educational Data Mining. Boca Raton: CRC Press, 2010. https://doi.org/10.1201/b10274

[11] R. S. J. d. Baker, « Modeling and understanding students' off-task behavior in intelligent tutosystems, ms », in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in CHI '07. New York, NY, USA: Association for Computing Machinery, avr. 2007, p. 1059-1068. https://doi.org/10.1145/1240624.1240785

[12] G. W. Dekker, M. Pechenizkiy, et J. M. Vleeshouwers, « Predicting students drop out : a case study », Proc. 2nd Int. Conf. Educ. Data Min. EDM 2009 July 1-3 2009 Cordoba Spain, p. 41-50, 2009.

[13] J. A. Walonoski et N. T. Heffernan, « Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems », in Intelligent Tutoring Systems, vol. 4053, M. Ikeda, K. D. Ashley, et T.-W. Chan, Ed., in Lecture Notes in Computer Science, vol. 4053., Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, p. 382-391. https://doi.org/10.1007/11774303_38

[14] S. Matz, C. Bukow, H. Peters, C. Deacons, A. Dinu, et C. Stachl, using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics, Sci. Rep., vol. 13, avr. 2023, https://doi.org/10.1038/s41598-023-32484-w

[15] R. Guevara-Reyes, I. Ortiz-Garcés, R. Andrade, F. Cox-Riquetti, et W. Villegas-Ch, « Machine learning models for academic performance prediction: interpretability and application in educational decision-making », *Front. Educ.*, vol. 10, août 2025, https://doi.org/10.3389/feduc.2025.1632315

[16] S. P. Mutsotsya, H. N. Mpia, M. M. Nzanzu, I. N. Baelani, and M. K. Kasolene, « Predicting Undergraduate Students' Final Grades: Using Educational Data Mining », Etincelle, vol. 25, no. 2, 2024, Accessed: May 31, 2026. Available from: https://www.academia.edu/download/118334872/RIME_2024_20_manuscrit_Sedar_2024_Etincelle.pdf

[17] Md. M. Islam, F. H. Sojib, Md. F. H. Mihad, M. Hasan, et M. Rahman, « The integration of explainable AI in Educational Data Mining for student academic performance prediction and support system », *Telemat. Inform. Rep.*, vol. 18, p. 100203, juin 2025, https://doi.org/10.1016/j.teler.2025.100203

[18] W. J. Lau et H. Abdul Rahman, « Predicting academic performance through machine learning: integrating demographic, psychological, and behavioral predictors using explainable AI », Asia Pac. Educ. Rev., mai 2026, https://doi.org/10.1007/s12564-026-10124-y

[19] Y. Wang, L. Yang, J. Wu, Z. Song, et L. Shi, « Mining Campus Big Data: Prediction of Career Choice Using Interpretable Machine Learning Method », Mathematics, vol. 10, no 8, p. 1289, janv. 2022, https://doi.org/10.3390/math10081289

[20] L. Kemper, G. Vorhoff, et B. U. Wigger, « Predicting student dropout: A machine learning approach », Eur. J. High. Educ., vol. 10, no 1, p. 28‑47, janv. 2020, https://doi.org/10.1080/21568235.2020.1718520