ULK Scientific Journal, Vol. 46(1), 2025 ISSN 2308-5843
https://dx.doi.org/10.4314/ulksj.v46i1.6

Predicting School Dropout Using Machine Learning Models: A Case Study of Nyanza District, Rwanda

Jean Bosco Musabe[1*], Emmanuel Byiringiro[2], Uwamahoro Pascaline[3]

[1*]Kigali Independent University (ULK), School of Science and Technology, Department of Computer Science, Kigali, Rwanda
[2]Adventist University of Central Africa (AUCA), Department of Big Data Analytics, Kigali, Rwanda
[3]Authentic World Ministry Zion Temple, AZWM, Kicukiro, Kigali, Rwanda

[*]**Corresponding Author**: bosulus@gmail.com.

Abstract

This study investigates the application of machine learning models to predict school dropouts in Nyanza District, Rwanda, addressing the challenge of early identification of at-risk students. By adopting a classification-based approach, the research analyzes data from parents or guardians, school instructors, and teachers to pinpoint contributing factors such as socioeconomic conditions, academic performance, and family background. The research explores a range of machine learning models, including Logistic Regression, Decision Tree Classifier, Gradient Boosting Regression, Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and Naive Bayes. These models are evaluated using metrics like accuracy, recall, precision, F1 score, and ROC-AUC, with an emphasis on balancing recall (identifying at-risk students) and precision. The study reveals that different models offer varying levels of performance. KNN achieves a notable accuracy of 0.72 and an exceptional recall of 0.91, successfully identifying 91% of at-risk students. Naive Bayes, however, is highlighted as the most well rounded model, balancing precision and recall effectively. This research fills the gap in predictive analytics for dropout prevention in Nyanza District and offers actionable insights for educators and policymakers to enhance student retention through targeted interventions.

***Keywords**:* School Dropout Prediction; Machine Learning; Educational Data; Student Retention.

1. Introduction

The Dropout report by the Ministry of Education MINEDUC and the United Nations International Children's Emergency Fund UNICEF Rwanda (2019) provides an insightful analysis of school dropout rates in Rwanda, revealing a clear gender disparity in dropout experiences. The report highlights that 13.4% of 12-year-old boys have faced at least one dropout, while the figure for girls is considerably lower, at 5.2%. This difference suggests that boys are more likely to drop out of school compared to girls at this age, emphasizing the need for targeted interventions to address the underlying factors contributing to this disparity (RGB & World Health Organisation, 2019)(Card, 2023). The report further delves into the various factors influencing school dropouts in Rwanda, shedding light on socio-economic, cultural, and educational challenges that affect boys and girls differently. It calls for comprehensive strategies to reduce dropout rates, with a focus on ensuring equal opportunities for both genders and tackling the root causes of these disparities. Addressing the dropout issue is crucial for improving educational outcomes and empowering children in Rwanda, particularly in marginalized communities. In Rwanda, the dropout rates reveal a shifting trend based on age and gender. While boys exhibit slightly higher dropout rates in lower secondary schools, the trend reverses after age 16. This change is largely attributed to low primary-to-secondary transition rates, which particularly affect girls.

Despite the government's efforts to address the issue, such as eliminating tuition fees, girls still face significant challenges in continuing their education, leading to higher dropout rates as they transition to secondary school (MINEDUC & UNICEF, 2017). The persistence of high dropout rates in Rwanda can be linked to various systemic challenges. Issues such as low enrollment, overcrowded classrooms, and limited resources continue to affect the education system. While these obstacles hinder educational progress for all students, they disproportionately influence girls, making it more difficult for them to remain in school beyond the lower secondary level. Despite the government's initiatives, these challenges remain a significant barrier to achieving universal education and reducing dropout rates in the country. Recognizing the multifaceted nature of these issues, a researcher focuses on investigating and proposing solutions for Nine Years of Basic Education and upper secondary schools in Nyanza District.

Machine learning has emerged as a promising tool to address the complex socio-economic barriers faced by students in developing countries(Goran et al., 2024). Although initiatives like the Twelve Years Basic Education (12YBE) program have aimed to reduce dropout rates, these challenges remain a significant hurdle in achieving a dropout-free education system. This study examines the application of machine learning in understanding and predicting school dropouts in Rwanda, with a particular focus on Nyanza District.

Guided by human capital theory—which underscores the value of education as an investment in future income—the research seeks innovative solutions to persistent dropout problems. By leveraging machine learning models, the study aims to provide actionable insights into dropout factors, enabling educators and policymakers to implement targeted interventions(Raju & Eid, n.d.)(Syed Mustapha, 2023). This survey serves as a valuable resource for exploring the role of technology in combating dropout challenges, emphasizing the potential of machine learning to support education as a cornerstone of social and economic development.

The paper is organized as follows: Section II presents the literature review carried out in this field using Machine learning. Section III describes our proposed method for predicting school dropouts. Section IV describes different experiments carried out (interpretation of our results) and the results obtained. Finally in section V summarizes the main conclusions and future research.

2. Literature Review

To successfully reduce student attrition, it is imperative to understand which students are at risk of dropping out. We develop an early detection system (EDS) to predict student success in tertiary education as a basis for a targeted intervention. The EDS uses regression analysis, neural networks, decision trees, and the AdaBoost algorithm to identify student characteristics that distinguish potential dropouts from graduates(et al., 2023). The developed method can be implemented in every German university, as it uses student performance and demographic data collected and maintained by legal mandate(Eckhoff et al., 2023). Therefore, the EDS self-adjusts to the university where it is employed. The EDS is tested and applied at a state university and a private university of applied sciences. Both institutes of higher education differ considerably in their organization, tuition fees, and student-teacher ratios. Our results indicate a prediction accuracy at the end of the first semester of 79% for the state university and 85% for the private university of applied sciences. After the fourth

semester, the accuracy improves to 90% for the state university and 95% for the private university of applied sciences (Berens et al., 2019).

Early prediction of student dropout can assist academic institutions in providing timely intervention as well as suitable planning and training to improve students' success rates. This study used a variety of machine-learning techniques to predict academic dropout among students(Villar & de Andrade, 2024). The model was trained and tested using DT, LR, NB, SVM, KNN, and ANN(Cam et al., 2024). With the use of the suggested prediction approach, course advisors, organizations, and the university will be able to assess students' performance and put effective interventions in place to raise their academic performance in advance. This study discovered that the Logistic Regression Model outperformed the other models employed in this investigation in predicting student dropouts. To increase accuracy, the proposed model may need to be re-evaluated using additional datasets, perhaps drawn from academic Big Datasets(Technique, 2023)

Finding probable dropout students is a wonderful way to improve retention strategies like help programs, training, or mentoring. Thus, a quantitative evaluation of the chances of failure can help allocate instructional, psychological, and administrative resources effectively. If captured, data produced within the academic setting can provide amazing insight that could also help identify and manage student dropouts. The study's primary goal is to design a machine-learning model for student dropout prediction. To achieve this, academic data of first-year undergraduate Computer Science at the University of Benin between the years 2016 to 2020 were considered (Technique, 2023).

Student dropout is a serious issue in that it not only affects the individual students who drop out but also has negative impacts on the former university, family, and society together(Aina et al., 2022). To resolve this, various attempts have been made to predict student dropout using machine learning. Academic records collected from 20,050 students of the university were analyzed and used for learning. Various machine learning models(Arya et al., 2024) were used to implement the model, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Deep Neural Network, and LightGBM (Light Gradient Boosting Machine)(Kanber et al., 2024), and their performances were compared through experiments(Rufo et al., 2021)(McCarty et al., 2020). We also discuss the influence of oversampling used to resolve data imbalance issues in the dropout data. For this purpose, various oversampling algorithms such as SMOTE, ADASYN, and Borderline-SMOTE were tested. Our experimental results showed that the proposed model implemented using LightGBM provided the best performance with an F1-score of 0.840, which is higher than the results of previous studies discussing the dropout prediction with the issue of class imbalance (Cho et al., 2023).

Predicting school dropout involves identifying students at risk of discontinuing their education based on a variety of factors. Studies have shown that socioeconomic background, academic performance, attendance records, and family circumstances are among the key indicators influencing dropout rates (Pokhrel, 2024)(Paul & Thapa, 2024). Predictive analytics can offer early warnings, enabling timely interventions. Research highlights the importance of robust models to differentiate between students likely to drop out and those who are not, minimizing false positives(Pazukhina et al., 2024)(Gordon et al., 2024).

Machine learning (ML) has gained traction as an effective tool for analyzing complex educational datasets. Algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Gradient Boosting are commonly employed in dropout prediction studies(Kummaraka & Srisuradetchai, 2024)(Rohani et al., 2024). These models help identify patterns and relationships within data that traditional statistical approaches might miss.

Recent advancements have focused on deep learning techniques(Nithya & Umarani, 2023), which offer enhanced predictive accuracy but often require larger datasets and computational power. The availability of educational data, including student demographics, academic records, attendance, and behavioral logs, has been pivotal in advancing dropout prediction models. Open datasets from global education initiatives and school systems enable researchers to refine and validate their models (Luong et al., 2024)(Dinh et al., 2025). However, challenges such as data privacy, collection inconsistencies, and missing values remain significant barriers. The integration of additional contextual data, such as community-level socioeconomic indicators, has been suggested to improve prediction accuracy.

The ultimate goal of dropout prediction is to enhance student retention. The literature emphasizes that data-driven insights should inform targeted interventions, such as academic support programs, counseling services, and family engagement initiatives(Smith et al., 2024). Some studies have proposed adaptive learning systems that use machine learning predictions to personalize educational content, fostering engagement and reducing dropout risks. Moreover, longitudinal studies indicate that sustained intervention efforts yield better retention outcomes than one-time actions. While machine learning offers promising avenues for dropout prediction, challenges remain. Many studies focus on short-term datasets, limiting the generalizability of their findings. Additionally, the interpretability of complex ML models, such as neural networks, often presents obstacles for educators and policymakers who require actionable insights. Future research should emphasize the development of explainable AI (XAI) models tailored to the educational sector

3. Methodology

Student behaviors and related data are collected from Nyanza District, Southern Province, Rwanda Country. The dataset is investigated by different kinds of students they differ in their nature of behavior, family type, interest, and social impacts. Some (student) of the real-time data are multidimensional, they can be rebalanced and converted into normal and processed data with the help of data cleansing and sampling techniques, https://github.com/bosuluss/School-Dropout_Dataset.

https://dx.doi.org/10.4314/ulksj.v46i1.6

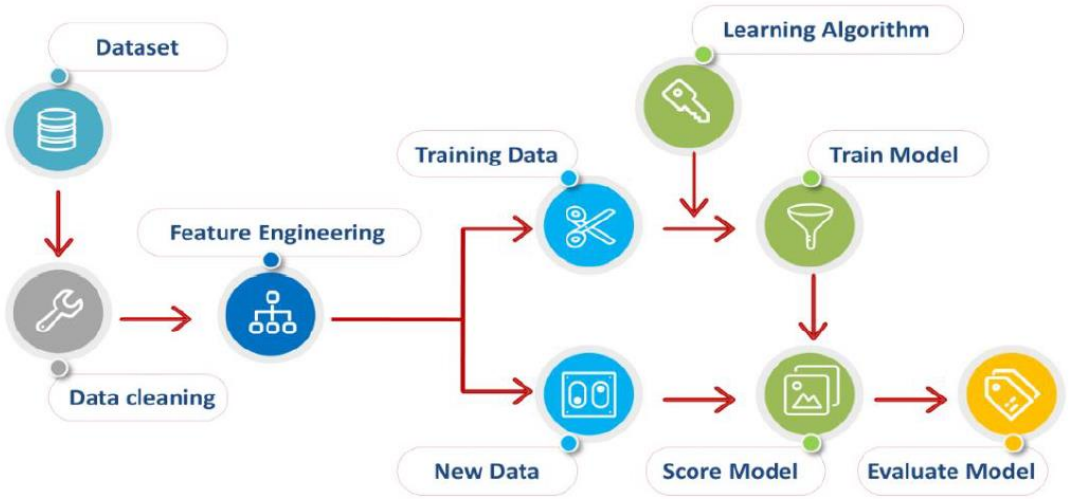

Figure 1: Architecture for machine learning model follows different feature engineering.

Figure 1 illustrates the typical workflow of a machine-learning project, starting with the collection of a dataset, which is then cleaned to remove noise and handle missing values, ensuring data quality. Following this, feature engineering is conducted to enhance the data by selecting or creating new features that improve model performance. The prepared data is split into training and testing sets, with the training data used to train a selected learning algorithm, such as decision trees or neural networks, which identify patterns and relationships. After training, the model is tested on new data to make predictions, which are then scored against actual results. The final step involves evaluating the model's performance using metrics like accuracy, precision, recall, and F1 score to assess its effectiveness and identify potential improvements.

3.1 Data Preprocessing

Student data samples were gathered from various colleges, aiming to analyze and predict student dropouts. The reasons behind dropouts are diverse, making the prediction process complex. A dataset was constructed by integrating 300 samples, focusing on scenarios that lead to dropout success. Since real-time datasets are often multi-dimensional, rebalancing the data was necessary. Data mining techniques were employed to evaluate and filter these samples effectively, ensuring a robust dataset for analysis. This dataset contains various attributes of students, including age group, gender, performance metrics, and behavioral factors, to predict school dropout rates. Each column represents specific features such as age group (AG), gender (GN), performance ranges (PFRM), and participation in activities (SACT), with the target variable being dropout status (DRP), indicating whether a student has dropped out (1) or not (0). The dataset comprises both categorical and binary indicators that collectively help in analyzing the likelihood of students dropping out based on different attributes Table 1.

The dropout analysis focused on 18 reduced features, emphasizing the selection of the most relevant attributes for predicting dropouts. Feature extraction was employed to minimize redundant data while preserving essential information, ensuring efficient processing. This process was categorized into filter-based methods, which independently evaluate attributes, and wrapper-based methods, which use learning algorithms to determine sample desirability. By finalizing the best 18 attributes, the study addressed various factors influencing student performance, providing a comprehensive approach to predicting dropouts.

3.2 Performance Metrics

Predictive modeling tackles classification and regression problems, each with distinct evaluation metrics due to the nature of variables involved. Regression handles continuous variables by calculating errors between actual and predicted values, while classification categorizes outputs as correct or incorrect. Employing multiple metrics is essential for a comprehensive model evaluation, as single metrics can be misleading. Key metrics like accuracy, precision, recall, and F1 score evaluate classification models, such as Logistic Regression, Decision Tree, Gradient Boosting, SVM, ANN, Naïve Bayes, and KNN. Accuracy measures correct classifications, precision focuses on minimizing false positives, recall assesses the ability to identify all positives, and F1 score balances precision and recall, crucial for imbalanced datasets. Data cleansing, the first step in data science and machine learning workflows, ensures data is clean, making exploration and model training more effective.

Table 1: Sample of dataset

	AG	GN	LSMA	LSFE	CL	PR	FCN	DAB	LMOT	ILLN	OTHR	ABSNT	DISPL	PFRM	SACT	YEOS	DRP
0	17 - 19	Male	1	0	0	0	0	0	0	0	0	0	0	41-50	None	Upper Secondary	1
1	06 - 13	Male	1	0	0	0	0	0	0	0	0	1	1	41-50	Other	Primary	1
2	06 - 13	Male	1	0	0	0	0	0	0	0	0	1	1	71-100	Other	Primary	0
3	06 - 13	Male	0	0	0	0	0	0	0	1	0	1	0	51-60	Dance	Primary	1
4	06 - 13	Male	0	0	0	0	0	0	0	0	0	1	0	71-100	Other	Primary	0
5	06 - 13	Male	0	0	0	0	0	0	0	0	0	1	0	41-50	Dance	Primary	1
6	17 - 19	Male	0	0	0	0	0	0	0	1	0	1	0	41-50	None	Upper Secondary	1
7	06 - 13	Male	0	0	0	0	0	0	0	0	0	0	0	51-60	None	Primary	1

3.2.1 Handling the Missing Data

The data received is often non-uniform, with missing values that need to be addressed to avoid negatively impacting machine learning model performance. This is achieved by replacing missing values with the mean or median of the corresponding column, a process facilitated by the Imputer class in the sklearn preprocessing library. The Imputer class accepts parameters such as missing values, where placeholders like integers or 'NaN' are identified and completed; strategy, which specifies the imputation approach (e.g., replacing with mean, median, or the most frequent value); and axis, which determines whether imputation is applied along columns (axis=0) or rows (axis=1).

Once configured, the imputer object is fitted to the dataset, ensuring consistent handling of missing data.

3.2.2 Data Transformation

Data transformation involves converting the data into a format suitable for analysis. The acquired data was not directly compatible with the model, necessitating transformation to ensure its validity for modeling purposes. One-hot encoding was employed to convert categorical features such as Child_age (a range format), Child Gender, social activity, year of study, and performance (also in range format) into numerical representations, with 0 representing female and 1 representing male, and similar encoding for other categories. Additionally, to ensure high precision for the KNN models, the input dataset underwent standardization using the StandardScaler function. JupyterLab was utilized to accomplish the data transformation tasks. The transformed dataset was subsequently saved. Table 2 below showcases a sample of the dataset after the transformation process was completed.

Table 2 displays a dataset with various attributes that detail different student-related aspects, such as age group, gender, class level, family support, drug abuse, motivation, health issues, absenteeism, disciplinary issues, academic performance, participation in school activities, and years of education. Each row represents an individual student, with the values indicating specific characteristics or conditions. The data is primarily categorical or ordinal, likely intended for use in predictive modeling or statistical analysis of student outcomes.

Table 2: Sample of Encoded Dataset

	AG	GN	LSMA	LSFE	CL	PR	FCN	DAB	LMOT	ILLN	OTHR	ABSNT	DISPL	PFRM	SACT	YEOS
293	2	0	1	1	1	0	1	1	1	0	1	1	1	3	3	2
93	0	0	1	0	0	0	1	0	0	0	0	1	0	3	0	1
47	2	1	1	0	0	0	0	0	0	0	0	0	0	1	3	2
175	2	1	0	0	0	0	0	1	0	1	0	1	0	2	3	2
84	3	1	1	0	1	0	1	0	1	0	0	0	1	0	1	2
16	3	1	1	1	0	0	0	0	0	0	0	1	0	2	3	2
83	1	1	1	0	0	0	1	0	0	0	1	0	0	4	2	0
292	3	1	0	1	1	0	1	0	1	0	1	0	1	0	3	2

3.2.3 Splitting the Data Set and Feature Engineering

Now we divide our data into two sets, one for training our model called the training set, and the other for evaluating the performance of our model called the test set. The split is generally 80/20, which means 80% for the training, and 20 % for the data testing. To do this we import the "train_test_split" method of "sklearn.model_selection" library. Then split the data into two sets. One for training the model called the training set, and another for evaluating the performance of the

model, called the test set. The split is usually 80/20. To do this, import the "train_test_split" method from the "sklearn.model_selection" library. Feature scaling is essential in machine learning as most models use Euclidean distance between data points, which means features with larger magnitudes influence the distance calculations more than those with smaller magnitudes. To mitigate this, techniques like standardization or z-score normalization are applied. During the exploratory data analysis phase, the dataset is examined to identify issues, uncover patterns, detect outliers, and reveal relationships between variables. Various visualizations, such as plots, charts, and graphs, are used to explore the data, providing valuable insights and a deeper understanding of its structure.

4. Results and Discussion

4.1 Performance Measurements

Student dropout prediction can be framed as a binary classification problem, where instances are categorized into two classes: dropout (positive class) and non-dropout (negative class). When evaluating the prediction results, a correct prediction is labeled as True, while an incorrect prediction is labeled as False. In this context, there are four possible outcomes for assessing the performance of the binary classification model, True Positive (TP): The model correctly predicts that a student will drop out. False Positive (FP), The model incorrectly predicts that a student will drop out when they do not. False Negative (FN), The model incorrectly predicts that a student will not drop out when they do. True Negative (TN): The model correctly predicts that a student will not drop out. A commonly used metric to evaluate the performance of a binary classification model is accuracy, which is calculated as the ratio of the number of correct predictions (TP + TN) to the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + FP + FN - TN} \qquad Eq1.$$

Note that when the data are skewed toward one class, accuracy cannot be used properly. For example, the dropout rate of four-year universities in South Korea is about 5%, where the data are highly skewed to the N class. In this case, simply predicting that no one will drop out would yield 95% accuracy. However, if the opposite prediction is made, the accuracy is significantly lowered to 5%.

The above problem indicates that FP and FN should be used together when measuring the prediction performance. Precision can be used to measure the performance from the perspective of FP and is defined as $\Pr ecision = \frac{TP}{TP + FP}$ *Eq2.*

Similarly, recall can be used to measure the performance from the perspective of FN, which is defined as

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \qquad Eq3.$$

https://dx.doi.org/10.4314/ulksj.v46i1.6

In this paper, the F1-score was used for evaluating the performance of prediction models since it can properly reflect the data imbalance in the performance evaluation:

$$F1\_score = \frac{\Pr ecison * \Re call}{\Pr ecision + \Re call} \qquad Eq4.$$

In this paper, the F1-score was used for evaluating the performance of prediction models since it can properly reflect the data imbalance in the performance evaluation.

Table 3: Dataset Table in the study

Dataset	**Dropout (0)**	**Successful (1)**	**Total**
Train	76	164	240
Test	16	45	61
SUM	92	209	301

Three hundred data points that were collected between 2024 are used in the experiments that were conducted. Learning on Decision Tree, KNN, NB, LR, ANN, and SVM was achieved by dividing collected and pretreated data into learning and test datasets in ratio 8:2 ratio, we applied a tester to the trained model to evaluate the accuracy of the prediction of the following are the evaluation findings for every model.

4.2 Machine learning models

Table 4: Summary of the Performance Evaluation

Algorithms	**Accuracy**	**Precision**	**Recall**	**F1 Z Score**
Logistic Regression	0.69	0.76	0.84	0.80
Decision Tree Classifier	0.66	0.77	0.76	0.76
Gradient Boosting Classifier	0.62	0.76	0.71	0.74
Support Vector Machine	0.70	0.78	0.84	0.81
Artificial Neural Network	0.69	0.82	0.73	0.78
Naïve Bayes	**0.74**	**0.81**	**0.84**	**0.83**
K-Nearest neighbors	**0.72**	**0.76**	**0.91**	**0.83**

Figure 2 is a correlation heatmap illustrating the relationships between various factors affecting school dropouts. Each cell in the heatmap shows the correlation coefficient between two variables, with values ranging from -1 to 1. Positive values indicate a direct relationship, while negative values

indicate an inverse relationship. Darker colors represent stronger correlations, either positive or negative. For example, "Lack of School Fees" negatively correlates with "Dropped out," suggesting that students struggling with school fees are more likely to drop out. The heatmap helps identify which factors are most strongly associated with dropping out of school.

Figure 3 comprises four stacked bar charts that provide insights into school dropouts. The top left chart shows the distribution of students by age group and gender, revealing more males in each group and the highest student numbers in the 06-13 age range. The top right chart illustrates dropout status across age groups, with the 06-13 group having the highest dropout rates. The bottom left chart highlights a gender disparity in dropouts, with more males dropping out than females. The bottom right chart examines the relationship between age groups and drug abuse, showing higher prevalence in the 17-19 and 20-21 age groups compared to younger groups.

Figure 4 This figure displays a bar chart comparing the performance metrics of various machine learning models used to predict school dropouts, including Logistic Regression, Decision Tree, Gradient Boosting Classifier, Support Vector Machine, Artificial Neural Network, Naïve Bayes, and K-Nearest Neighbors (KNN). The models are evaluated based on Accuracy, Precision, Recall, and F1 Score. Naïve Bayes and KNN stand out with higher F1 scores, indicating they are the best performers, with KNN excelling in recall and Naïve Bayes showing balanced performance across all metrics. In contrast, Logistic Regression and Decision Tree exhibit moderate performance, while Support Vector Machine and Artificial Neural Network have slightly lower scores. This chart aids in comparing the models to select the most appropriate one for dropout prediction, depending on the prioritized performance criteria.

https://dx.doi.org/10.4314/ulksj.v46i1.6

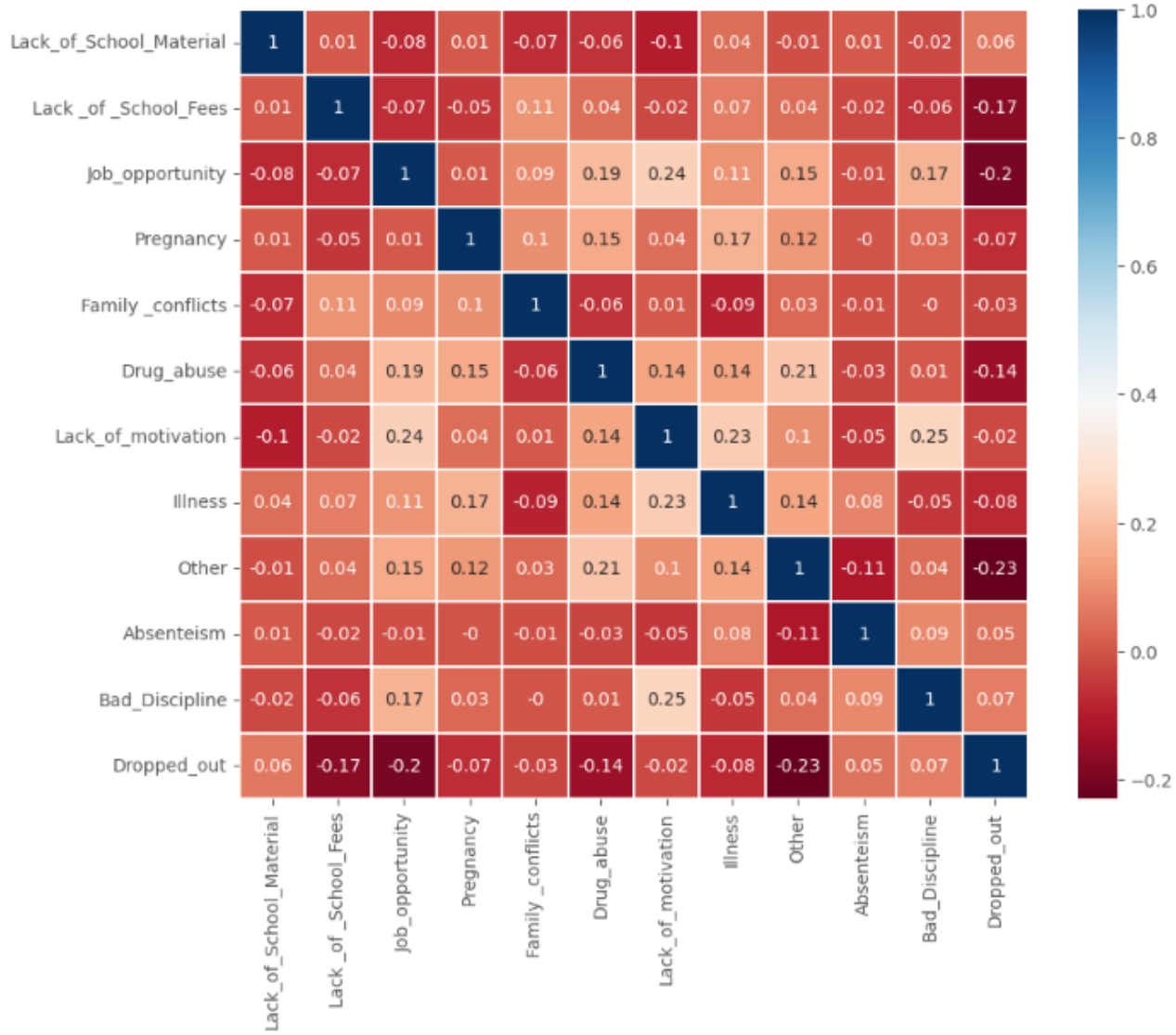

Figure 2: Confusion matrix of different features of school dropout

https://dx.doi.org/10.4314/ulksj.v46i1.6

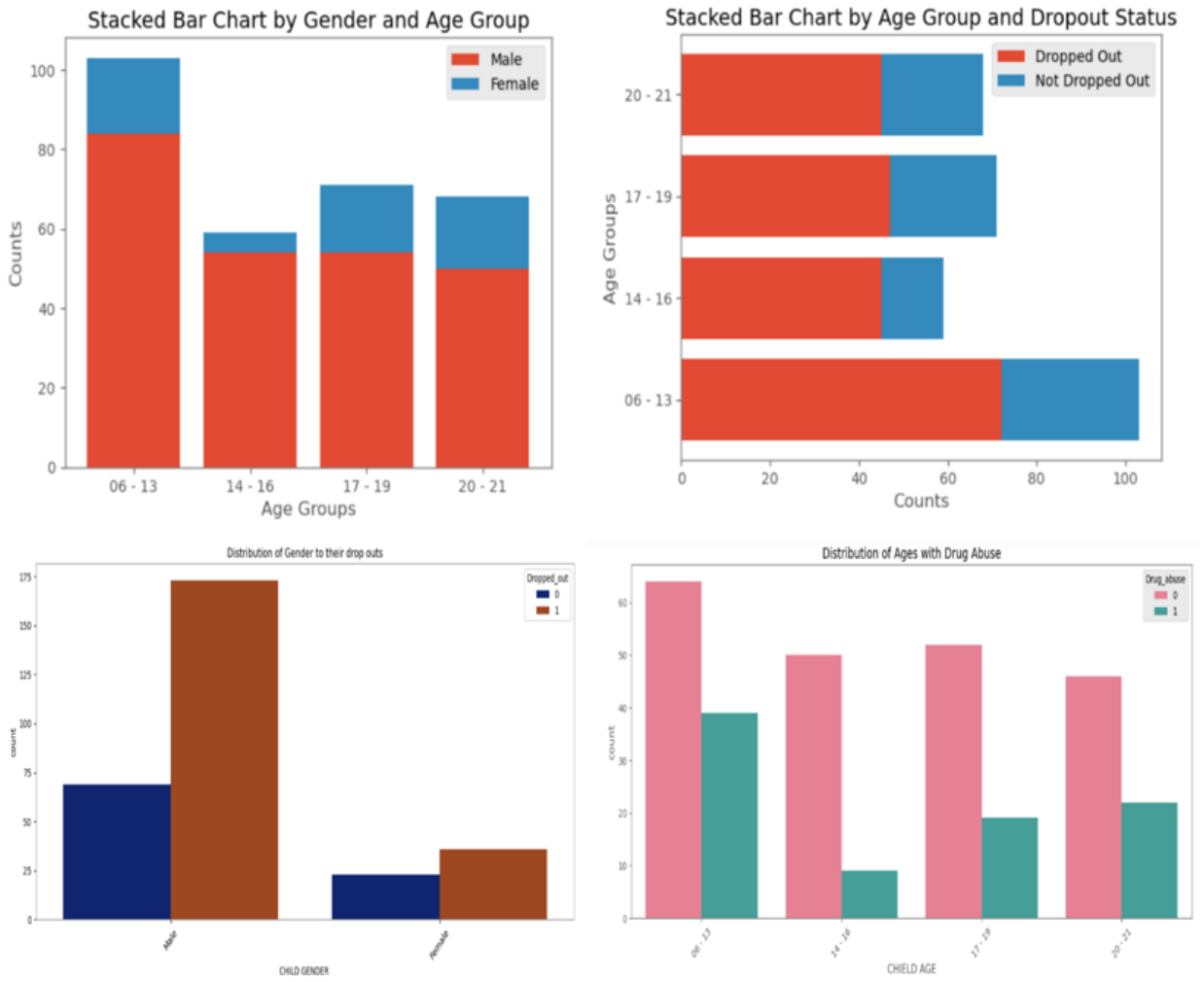

Figure 3: Data Visualization using a different metric

https://dx.doi.org/10.4314/ulksj.v46i1.6

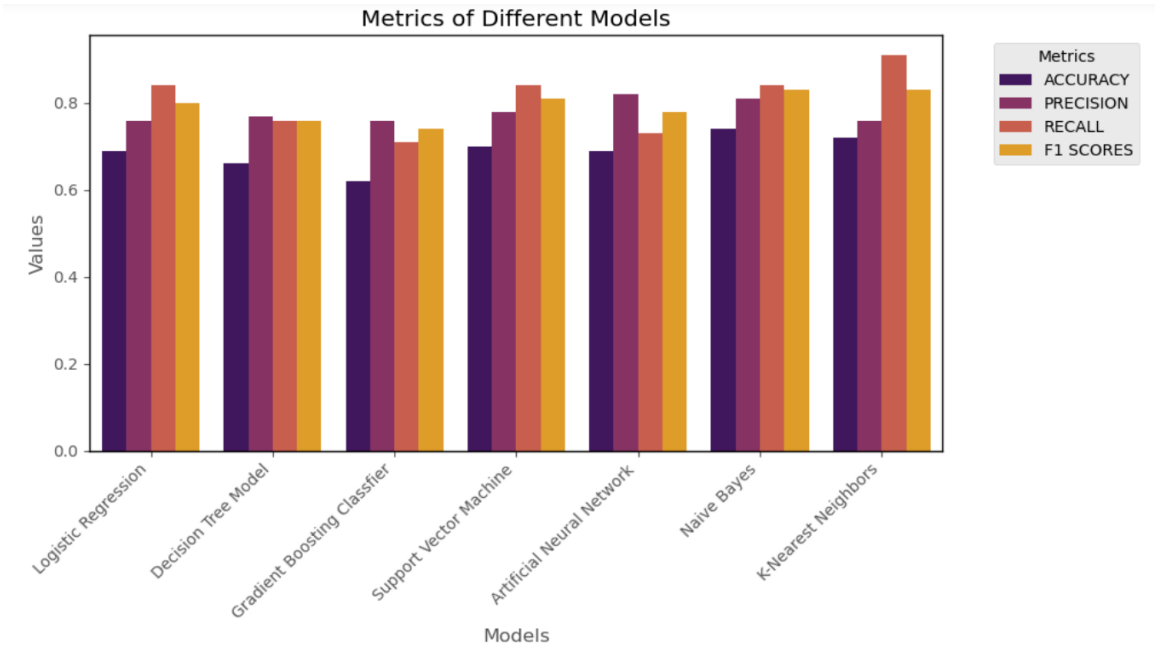

Figure 4: Representation of knowledge discovered from different modules

The comparison of various machine-learning models revealed that Naïve Bayes and K-Nearest Neighbors (KNN) were the top performers, each achieving the highest F1 score of 0.83. Naïve Bayes demonstrated a more balanced performance with an accuracy of 0.74, precision of 0.81, and recall of 0.84, making it the most reliable model overall. In contrast, KNN achieved an accuracy of 0.72 and an exceptional recall of 0.91, successfully identifying 91% of positive cases, although its precision was slightly lower at 0.76. These findings indicate that while both models are effective at capturing positive instances, Naïve Bayes is preferable in scenarios that require a balanced trade-off between precision and recall, whereas KNN is advantageous when the priority is to maximize recall. Other models, including Logistic Regression, Decision Tree Classifier, Gradient Boosting, Support Vector Machines (SVM), and Artificial Neural Networks (ANN), exhibited varied performance metrics. Although these models provided solid predictions with reasonable levels of precision and recall, their F1 scores and overall accuracy were lower compared to Naïve Bayes and KNN. This suggests that while these models are competent, they may not be as effective in consistently identifying at-risk students as the top two models. In summary, both Naïve Bayes and KNN are excellent choices depending on the specific needs for balancing precision and recall. However, Naïve Bayes stands out as the most well rounded model, offering the best overall performance for predicting school dropouts. These results highlight the importance of selecting the appropriate machine learning model based on the desired balance between different performance metrics to effectively address the issue of student retention.

5. Conclusion

This study compared machine-learning models for predicting school dropouts, with Naïve Bayes and K-Nearest Neighbors (KNN) performing the best, both achieving F1 scores of 0.83. Naïve Bayes

was the most balanced model with high precision and recall, while KNN excelled in recall. Other models showed solid performance but were less accurate. The results suggest that Naïve Bayes is ideal for balancing precision and recall, while KNN is better for maximizing recall. Future work should focus on enhancing model scalability, incorporating advanced techniques, and integrating socio-economic and behavioral data augmentation techniques for improved performance.

Conflict of Interest

This work has no Conflict of interest

Data Availability

All data are available from the link https://github.com/bosuluss/School-Dropout_Dataset.

Acknowledgment

Kigali Independent University ULK, ULK AJOL fellowship granted by the School of Science and Technology, for their support, supports this research.

References

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, *79*(June 2020), 101102. https://doi.org/10.1016/j.seps.2021.101102

Arya, S., Anju, A., & Azuana Ramli, N. (2024). Predicting the stress level of students using Supervised Machine Learning and Artificial Neural Network (ANN). *Indian Journal of Engineering*, *21*(56), 1–24. https://doi.org/10.54905/disssi.v21i55.e9ije1684

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, *11*(3), 1–41.

Cam, H. N. T., Sarlan, A., & Arshad, N. I. (2024). A hybrid model integrating recurrent neural networks and the semi-supervised support vector machine for identification of early student dropout risk. *PeerJ Computer Science*, *10*, 1–31. https://doi.org/10.7717/peerj-cs.2572

Card, C. R. (2023). Citizen Report Card Crc 2023 Ishusho Y'Uko Abaturage Babona Imiyoborere N'Imitangire Ya Serivisi Mu Nzego Zibegereye. *Rgb*, 2307–2431.

Cho, C. H., Yu, Y. W., & Kim, H. G. (2023). A Study on Dropout Prediction for University Students Using Machine Learning. *Applied Sciences*, *13*(21), 12004. https://doi.org/10.3390/app132112004

Dinh, T. N. T., Van Nguyen, H., Vu, A. T. L., Nguyen, P. M., Nguyen, T. T. A., & Phan, L. T. (2025).

The capacity of primary school inclusive teachers meets the requirements of the 2018 general education program. *Multidisciplinary Science Journal*, *7*(3). https://doi.org/10.31893/multiscience.2025170

Eckhoff, J. A., Rosman, G., Altieri, M. S., Speidel, S., Stoyanov, D., Anvari, M., Meier-Hein, L., März, K., Jannin, P., Pugh, C., Wagner, M., Witkowski, E., Shaw, P., Madani, A., Ban, Y., Ward, T., Filicori, F., Padoy, N., Talamini, M., & Meireles, O. R. (2023). SAGES consensus recommendations on surgical video data use, structure, and exploration (for research in artificial intelligence, clinical quality improvement, and surgical education). *Surgical Endoscopy*, *37*(11), 8690–8707. https://doi.org/10.1007/s00464-023-10288-3

Goran, R., Jovanovic, L., Bacanin, N., Stankovic, M., Simic, V., Antonijevic, M., & Zivkovic, M. (2024). Identifying and understanding student dropouts using metaheuristic optimized classifiers and explainable artificial intelligence techniques. *IEEE Access*, *12*(July). https://doi.org/10.1109/ACCESS.2024.3446653

Gordon, H., Salim, N., Tong, S., Walker, S., De Silva, M., Cluver, C., Mehdipour, P., Hiscock, R., Sutherland, L., Doust, A., Bergman, L., Wikström, A. K., Lindquist, A., Hesselman, S., & Hastie, R. (2024). Metformin use and preeclampsia risk in women with diabetes: a two-country cohort analysis. *BMC Medicine*, *22*(1), 418. https://doi.org/10.1186/s12916-024-03628-0

Kanber, B. M., Smadi, A. Al, Noaman, N. F., Liu, B., Gou, S., & Alsmadi, M. K. (2024). LightGBM: A Leading Force in Breast Cancer Diagnosis Through Machine Learning and Image Processing. *IEEE Access*, *12*(February), 39811–39832. https://doi.org/10.1109/ACCESS.2024.3375755

Kummaraka, U., & Srisuradetchai, P. (2024). Time-Series Interval Forecasting with Dual-Output Monte Carlo Dropout: A Case Study on Durian Exports. *Forecasting*, *6*(3), 616–636. https://doi.org/10.3390/forecast6030033

Luong, N. Van, Thuy, L. T. N., Tinh, T. T., Yen, N. T. H., & Thuy, D. T. (2024). Integrating Open Knowledge and Administrative Management in the Digital Transformation Model of Education Institutions: An Effective Approach. *International Journal of Religion*, *5*(7), 290–302. https://doi.org/10.61707/2vywvv49

McCarty, D. A., Kim, H. W., & Lee, H. K. (2020). Evaluation of light gradient boosted machine learning technique in large scale land use and land cover classification. *Environments - MDPI*, *7*(10), 1–22. https://doi.org/10.3390/environments7100084

MINEDUC, & UNICEF. (2017). Understanding dropout and repetition in Rwanda full report. *Report*, 2–244. http://www.rencp.org/wp-content/uploads/2018/09/DROPOUT-STUDY-FULL-REPORT.pdf

Nithya, S., & Umarani, S. (2023). an Identification of the Prominent Learner Behavioral Features To Predict Mooc Dropouts Using Hybrid Algorithm. *Journal of Theoretical and Applied Information Technology*, *101*(3), 1261–1274.

Paul, P., & Thapa, S. (2024). *Non-completion of primary and secondary levels of schooling among India ' s youth : evidence from a national survey Non-completion of primary and secondary levels of schooling among India ' s youth : evidence from a national survey*. *November*. https://doi.org/10.1007/s43545-024-01009-1

Pazukhina, E., Garcia-Gallo, E., Reyes, L. F., Kildal, A. B., Jassat, W., Dryden, M., Holter, J. C., Chatterjee, A., Gomez, K., Søraas, A., Puntoni, M., Latronico, N., Bozza, F. A., Edelstein, M., Gonçalves, B. P., Kartsonaki, C., Kruglova, O., Gaião, S., Chow, Y. P., … ISARIC Clinical Characterisation Group and ISARIC Global Covid-19 follow up working group. (2024). Long Covid: a global health issue - a prospective, cohort study set in four continents. *BMJ Global Health*, *9*(10), 1–14. https://doi.org/10.1136/bmjgh-2024-015245

Pokhrel, S. (2024). No TitleΕΛΕΝΗ. *Αγαη*, *15*(1), 37–48.

Raju, S. K., & Eid, M. M. (n.d.). *Predicting Student Adaptability in Online Education : A Comparative Study of Machine Learning Models and Copula-Based Analysis*.

RGB, & World Health Organisation. (2019). Citizen Report Card - Crc 2019 Ishusho Y ' Uko Abaturage Babona Imiyoborere N ' Imitangire. *Global Network of WHO Collaborating Centres for Bioethics*, 2307–2431. http://apps.who.int/iris/bitstream/10665/164576/1/9789240694033_eng.pdf

Rohani, N., Rohani, B., & Manataki, A. (2024). *ClickTree: A Tree-based Method for Predicting Math Students' Performance Based on Clickstream Data*. 1–19. https://doi.org/10.5281/zenodo.13627655

Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (Lightgbm). *Diagnostics*, *11*(9), 1–14. https://doi.org/10.3390/diagnostics11091714

Smith, L. R., Perez-Brumer, A., Nicholls, M., Harris, J., Allen, Q., Padilla, A., Yates, A., Samore, E., Kennedy, R., Kuo, I., Lake, J. E., Denis, C., Goodman-Meza, D., Davidson, P., Shoptaw, S., & El-Bassel, N. (2024). A data-driven approach to implementing the HPTN 094 complex intervention INTEGRA in local communities. *Implementation Science* , *19*(1), 1–14. https://doi.org/10.1186/s13012-024-01363-x

Syed Mustapha, S. M. F. D. (2023). Predictive Analysis of Students' Learning Performance Using Data Mining Techniques: A Comparative Study of Feature Selection Methods. *Applied System Innovation*, *6*(5). https://doi.org/10.3390/asi6050086

U.J, U., & Malik, S. (2023). A Hybrid Weight based Feature Selection Algorithm for Predicting Students' Academic Advancement by Employing Data Science Approaches. *International Journal of Education and Management Engineering*, *13*(5), 1–22. https://doi.org/10.5815/ijeme.2023.05.01

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1). https://doi.org/10.1007/s44163-023-00079-z