Knowledge-Based, Context-Aware AI for Accurate Learner Classification with Resampling Techniques in Low-Resource environments

Sameher Ajili[1,*], Rym Cheour[1,2], Mariem Abid[3], Mouna Baklouti[1] and Richard Hotte[3]

[1] Computer & Embedded Systems Laboratory (Ces.lab), National Engineering School of Sfax (ENIS), University of Sfax, Tunisia
[2] Higher Institute of Applied Science and Technology of Kasserine (ISSAT), University of Kairouan, Tunisia
[3] Applied Artificial Intelligence Institute (I2A), TELUQ University, Montréal, QC, Canada

Abstract

AI-driven prediction models face significant challenges in low-resource educational contexts, where cold start conditions and severe class imbalance data are prevalent. This study addresses this problem by proposing a methodological framework that integrates NLP-based preprocessing for feature extraction, advanced imbalance-aware resampling, and a comprehensive multi-model comparison. We evaluated the framework using a small, authentic dataset of learners from Mali, applying ten classifiers, including SVM, Random Forest, XGBoost, and ensemble methods. Model performance was assessed using stratified 10-fold cross-validation, with and without resampling via Random Oversampling (ROS) and SMOTE strategies, with evaluation based on accuracy, F1-score, and AUC-ROC. Results show that ensemble methods, specifically Random Forest, Gradient Boosting and XGBoost, achieved over 90% accuracy, and superior F1-scores, after SMOTE, significantly outperforming baselines and ROS.

Keywords

Learning analytics, Class imbalance, Low-resource education, Ensemble Learning, NLP Preprocessing, SMOTE

1. Introduction

In the 21st century, education is transforming beyond traditional classrooms through the rapid rise of online learning. While e-learning provides access to diverse educational content, it also highlights growing learner heterogeneity. Students nowadays enter the learning environment with unique prior knowledge, needs, and goals, making standardized pedagogical approaches increasingly inadequate [1]. This challenge has defined learner modeling as the core focus of Educational Data Mining (EDM). Learner models represent cognitive characteristics, competencies, and preferences to enable personalized educational pathways. These models serve as simplified representations of complex learning phenomena or datasets, designed to both understand and explain observable patterns and predict future learning dynamics [2]. A major axis of this modeling concerns competences [3], defined as the articulation between knowledge, practical skills and behavioral dispositions [4]. The literature generally identifies three fundamental dimensions: (i) declarative Knowledge (know-what), (ii) Strategies and procedural Skills (know-how), and (iii) socio-affective and motivational Dispositions (know-why). The multidimensional K-S-D model necessitates context-aware learning modeling, incorporating factors such as sociodemographic variables and technological conditions [5] that directly affect progression [4]. This need is reflected in the diverse techniques developed in this field, which [3] categorized into five main families: clustering and classification, predictive modeling, uncertainty-based approaches, overlay models, and ontology-driven methods. However, the cold-start problem [6], the lack of available data for new learners, remains a significant challenge, as it limits initial model adaptation and risks a negative early user experience. In this context, classification and clustering approaches are particularly critical, enabling the rapid estimation of a learner's needs and goals from minimal initial data [3].

Cooperative Information Systems – Early Research Achievement and Demos, 2025
*Corresponding author.
sameher.ajili@enis.tn (S. Ajili); rym.cheour@enis.tn (R. Cheour); mariem.abid@teluq.ca (M. Abid); Mouna.baklouti@enis.tn (M. Baklouti); Richard.Hotte@teluq.ca (R. Hotte)

This study introduces a methodological framework for predicting learners' knowledge levels in low-resource educational settings. This approach addresses two major challenges: overcoming class imbalance to improve the representation of minority learner groups and conducting a rigorous comparative analysis of machine learning models to identify the most effective models within data constraints. The remainder of this paper is organized as follows. Section 2 reviews related studies on imbalance mitigation methods and ML education. Section 3 describes the dataset, preprocessing, and experimental setup. Section 4 reports the results, with an emphasis on the performance gains for minority-class predictions. Section 5 provides conclusions and future directions.

2. Related Work

Recent research on learner modeling and cold-start remains a significant problem in adaptive learning environments. On the modeling side, [7] structured learner modeling into initial assessment, early categorization, and continuous adaptation for feedback loops. In this context, classification and clustering techniques, along with predictive modeling approaches, have been widely used to initialize learner states from first interactions [3]. This approach has been extended using knowledge graphs for scalability [8] and refined using data-driven methods that exploit learner attributes and advanced algorithms, such as deep learning and matrix factorization [6].

Within this context, external (environment, educational resources, social interactions) and internal (demographics, motivation, emotional state, competencies, etc.) factors shape the modeling space [9]. Hybrid pipelines combine collaborative filtering, rule-based reasoning, and K-nearest neighbors to deliver context-aware personalization based on demographic, behavioral, and self-reported indicators for multi-faced learner modeling [10]. Moreover, the combination of K-means clustering and Random Forest achieved an 86.1% prediction accuracy on Open University data, thus highlighting the effectiveness of hybrid unsupervised-supervised methods [11]. Another challenge that amplifies the cold-start problem in educational data is class imbalance, which biases predictive models towards majority classes and compromises fairness. Data-level resampling is widely used; however, its effectiveness is highly context-sensitive. [12], using the U.S. High School Longitudinal Study dataset, found that Random Oversampling (ROS) yielded the highest accuracy for moderately imbalanced tasks (accuracy 85%), whereas the hybrid SMOTE-NC + RUS combination produced superior results in extreme imbalance scenarios (10-15% accuracy gains). Similarly, [13] showed that oversampling techniques such as SMOTE and SMOTE-ENN significantly boosted performance, achieving high AUC values for both SVM and Random Forest, and proposed Equi-Fused-Data-SMOTE, a hybrid approach, which achieved high accuracy, F1-scores, and AUC. These studies underscore that both the selection of a classifier algorithm and the strategy for handling class imbalances critically influence predictive outcomes.

Existing work addresses cold-start via side-information, hybrid recommendation pipelines, and deep sequence models, and reduces imbalance primarily through resampling with diverse classifiers, validated on large-scale or synthetically balanced datasets [12] [11]. However, the convergence of these two challenges: data scarcity and severe imbalance in authentic, low-resource educational contexts with small, authentic data remains underexplored. Our study targets this gap by integrating NLP-based feature extraction with imbalance-aware training and a deep multi-model comparison on a real-world dataset from Mali.

3. Methodology

This study uses a dataset from the AMI project, a child-centered intelligent learning system designed for semi-nomadic communities in sub-Saharan Africa in the Kalani region. The project employs learner modeling and dynamic adaptation to personalize instruction according to prior knowledge, socio-cultural context, and lived experiences [14]. The dataset was collected from 100 children aged 6-11, capturing demographic, knowledge and performance, economic, sociocultural, and motivational dimensions.

3.1. Data Preprocessing

In this study, the AMI dataset was systematically pre-processed to convert unstructured textual responses into a structured, analyzable format. We employed a comprehensive NLP pipeline encompassing stop-word removal and space normalization to reduce noise, along with tokenization to segment text into linguistically meaningful units. Word frequency analysis was then conducted to identify dominant terms and thematic patterns, which were subsequently organized into categories using automated and semi-automated methods, producing a structured representation of both learner objectives and socio-cultural contextual information. Learners' objectives were grouped into twelve categories encompassing foundational cognitive skills (e.g., reading, writing, counting) and context-specific practices (e.g., food preparation, animal care, health, hygiene), Such categorization provides a structured lens for analyzing how children's self-declared goals align with their socio-cultural environments in which learning takes place. Figure 1 illustrates the frequency distribution of these objective categories across the dataset. To address multivalued attributes (goals and interests), a column explosion was used, mapping each learner-category association to an independent record. This expansion increased the dataset from 100 learner profiles (26 variables) to 227 learner models (30 variables).

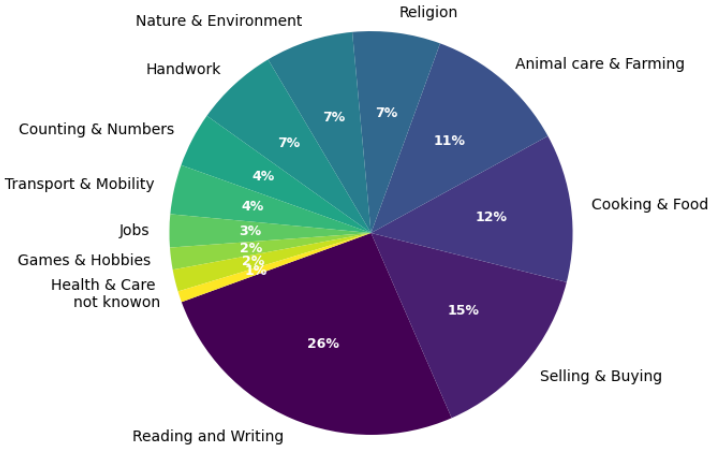


Figure 1: Distribution of Learners' Objectives Across Categories

3.2. Skill Distribution and Learning Goals

A visual analysis was conducted to examine the core competencies. Five key skills were analyzed: number and letter recognition, mathematical sign recognition, arithmetic operations, color recognition, and geometric shape recognition. Figure 2 summarizes the competency levels, highlighting the predominance of beginner-level proficiency.

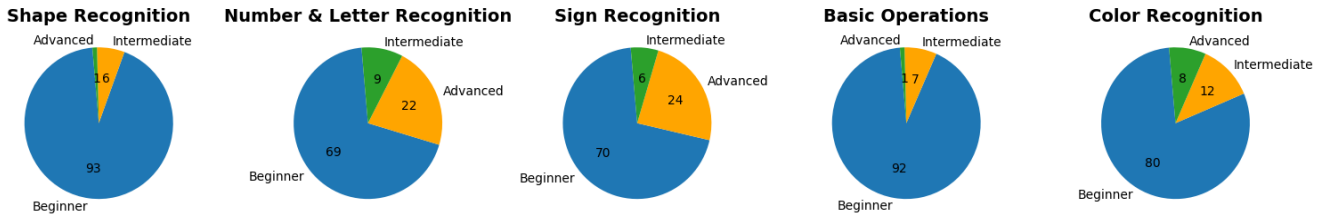


Figure 2: Frequency Distribution of Competency Factors.

3.3. Feature Engineering

An aggregate proficiency label, referred to as (`Overall_Level`), was established as the classification target. This label was calculated by majority vote through a rule-based knowledge approach across the five skill-specific variables. As shown in Figure 3, the target is highly imbalanced, with beginner at 86.3%, advanced at 9.7%, intermediate at 4.0%, yielding an approximate ratio of 21:2:1. The predictive features selected were the learners' objectives and their living environment: the objectives were multi-label binarized, the environment one-hot encoded, and the target label-encoded. All the encoders were fitted only in the training folds to avoid leakage and ensure the validity of subsequent evaluations.

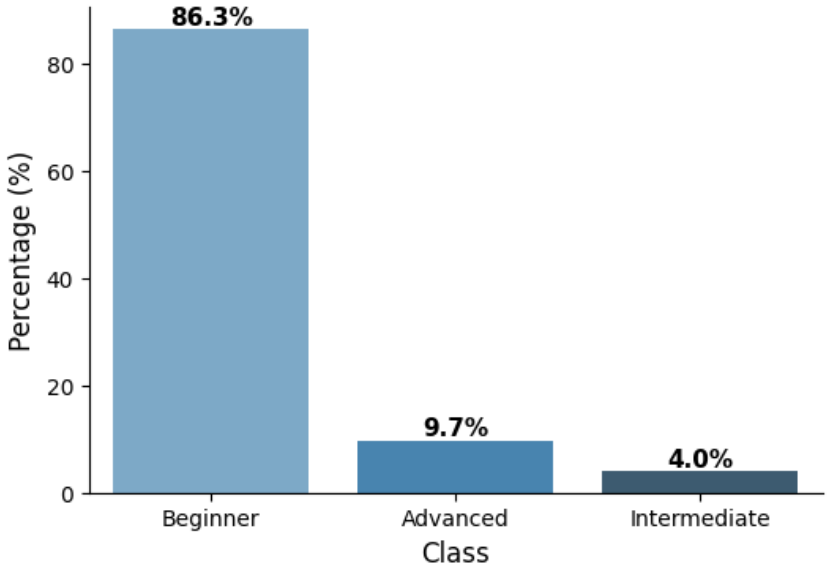


Figure 3: Overall Class Distribution and Total Unique Children

3.4. Imbalance-Aware Training

To address the problem of class imbalance, we applied both over- and undersampling strategies that are widely used in educational data mining. Oversampling approaches included the Random Over-Sampler (ROS), which duplicates minority instances, and SMOTE/SMOTEN, which generates synthetic samples for numeric and categorical features [15]. Under-sampling used Random Under-Sampling (RUS) and NearMiss to reduce the majority-class prevalence or focus on decision boundaries [12].

3.5. Model Selection and Evaluation

To evaluate resampling effectiveness, we benchmarked a set of algorithms, including linear models (Logistic Regression, linear SVM, SGD), nonlinear models (SVM-RBF, K-Nearest Neighbors), and tree-based ensembles (Random Forest, Gradient Boosting, XGBoost), probabilistic (Naïve Bayes), neural network (multi-layer perceptron (MLP)), and Ensemble Learning methods with voting classifiers. Each classifier was evaluated using stratified 10-fold cross-validation with standard metrics: accuracy, precision, recall, and F1-score, to ensure a comprehensive performance comparison.

4. Results

4.1. Baseline Performance on Imbalanced Data

Initial classification using the raw dataset without any resampling technique revealed performance variations across algorithms (Figure 4). The overall accuracy varied between 71.74% and 89.13% with

linear approaches (Logistic Regression, Linear SVM, and SGD Classifier), consistently achieving the highest scores (≈ 89%). Tree-based models and ensemble methods stabilized at ≈80.43%, whereas Naive Bayes displayed the lowest performance at 71.74%. These results suggest a clear advantage for linear models.

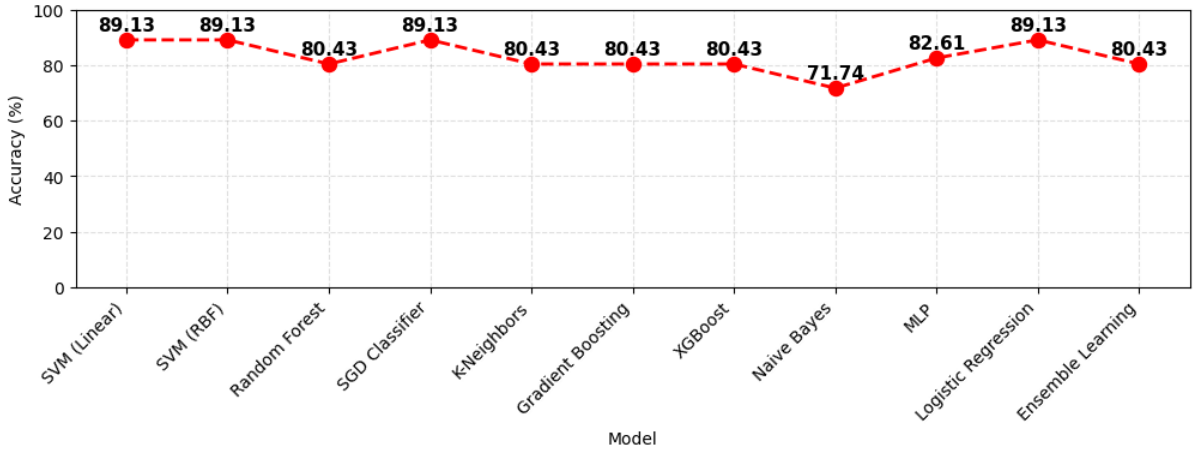


Figure 4: Overall accuracy of tested classifiers before resampling

A detailed analysis of class-level metrics in Figure 5 highlights critical weaknesses. Several classifiers were unable to recognize minority categories, with recall and F1-scores dropping to zero. Their predictions were driven almost entirely by the dominant class, resulting in systematic misclassification of underrepresented learners. This result illustrates a known problem of imbalanced datasets, where accuracy alone provides a distorted and overly optimistic view of model performance.

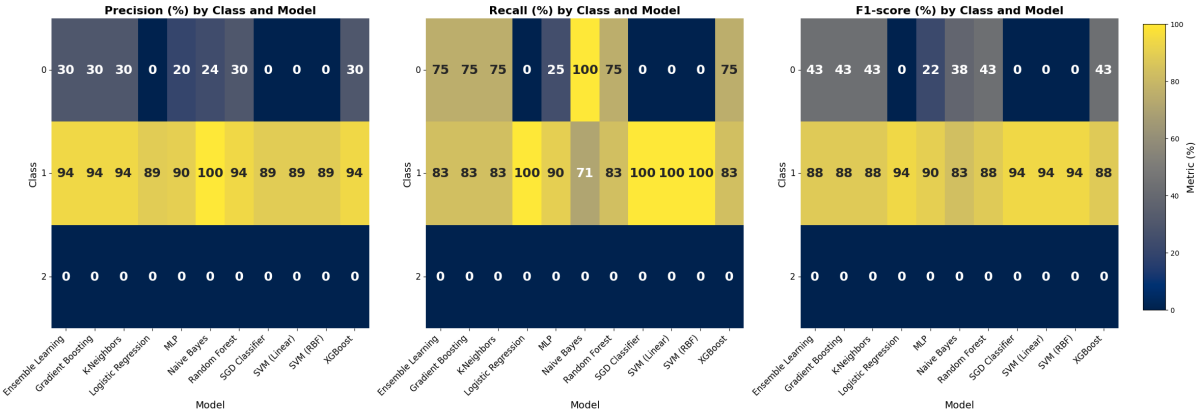


Figure 5: Class-specific performance metrics before resampling

4.2. Comparative Analysis of Resampling Strategies

We applied and compared five resampling strategies to examine their impact, as shown in Figure 6. The original data revealed a stark imbalance, with 196 beginner learners compared to only 22 advanced and 9 intermediate learners. Oversampling approaches, such as ROS, SMOTE, and SMOTEN, generated perfectly balanced classes of 196 learners each, whereas undersampling techniques, such as RUS and NearMiss, reduced all classes to nine. For small categorical datasets, such as AMI, oversampling appears more appropriate from both methodological and practical standpoints.

The evaluation of the resampling strategies highlighted the superiority of SMOTE over ROS in terms of overall accuracy (Figure 7). SMOTE consistently enhanced performance, with models such as Random

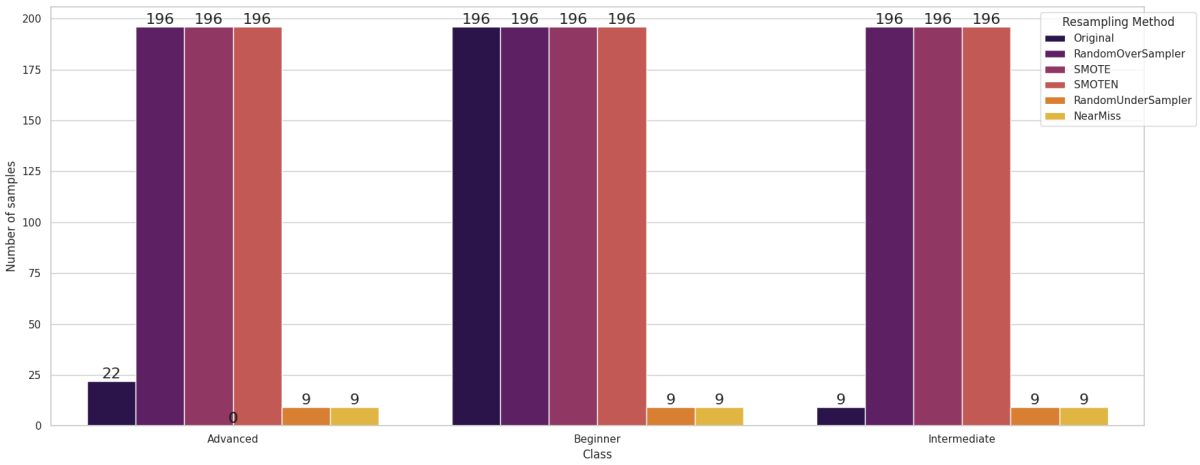


Figure 6: Comparison of resampling strategies for learner classes

Forest, Gradient Boosting, and XGBoost exceeding 90% accuracy. While ROS produced only moderate gains, some classifiers (e.g., K-neighbors, SGD) remained limited owing to the small size of the dataset.

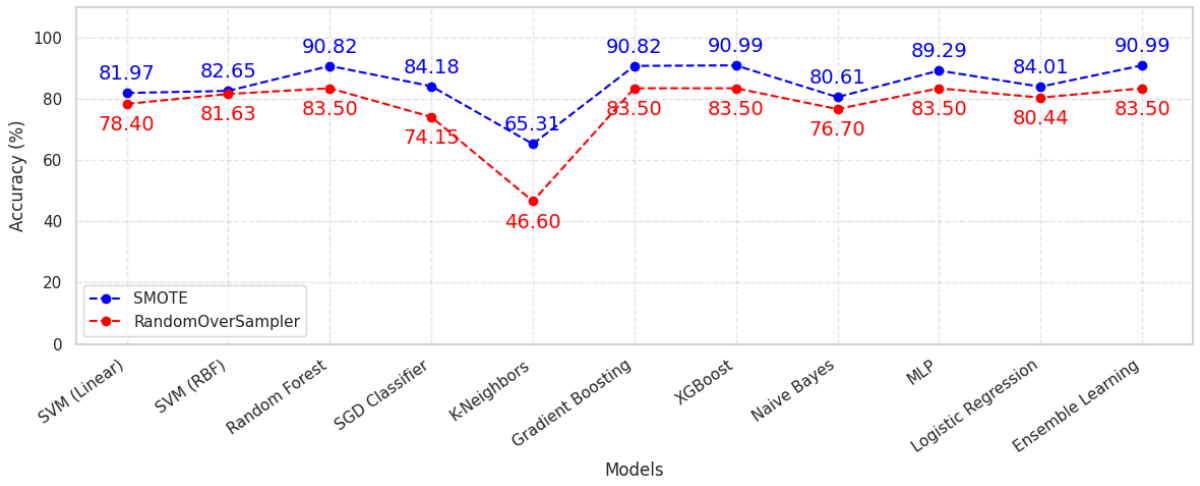
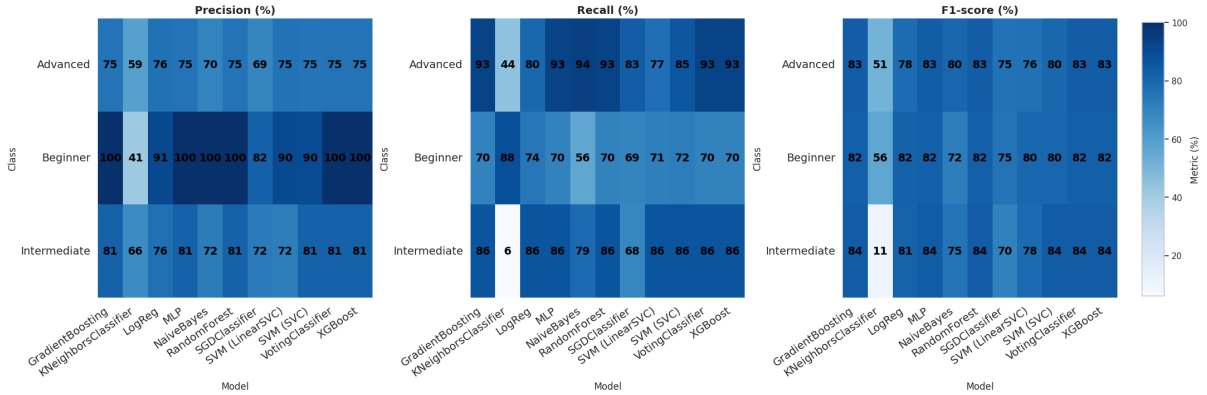
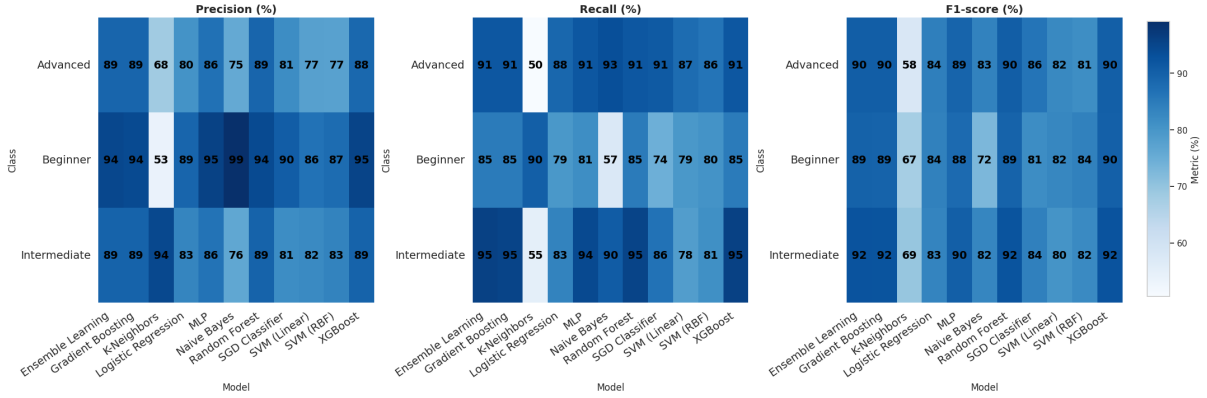


Figure 7: Overall classifier accuracy after applying resampling techniques: SMOTE and ROS.

A detailed comparison of the two resampling strategies reveals substantial differences in classifier performance across prior knowledge levels. Under Random Over-Sampling (ROS, Figure 8a), ensemble-based models—Random Forest, Gradient Boosting, XGBoost, and MLP—achieved high F1-scores for advanced (83.56%), intermediate (84.16%), and beginner learners (82.63%), whereas models such as K-Neighbors and the SGD Classifier displayed pronounced deficiencies, particularly for intermediate learners (F1-score 11.21%). Application of SMOTE (Figure 8b) markedly enhanced performance and produced more balanced results across all classes: advanced learners reached F1-scores between 90.23% and 90.45%, intermediate learners between 90.35% and 92.61%, and beginners between 88.15% and 89.78%. Even previously underperforming models, such as K-Neighbors and Naive Bayes, benefitted from SMOTE, although their scores remained below those of ensemble-based classifiers. These findings demonstrate that SMOTE effectively mitigates class imbalance by generating synthetic minority samples, thereby enabling classifiers to recognize patterns in underrepresented learner categories. Collectively, this evidence supports the conclusion that ensemble methods combined with SMOTE provide the most reliable strategy for predicting learner proficiency in small-scale, low-resource educational datasets.



(a) Results with ROS



(b) Results with SMOTE

Figure 8: Comparison of model performance after resampling with (a) ROS and (b) SMOTE.

5. Conclusion

This study presented an NLP-based preprocessing pipeline for the AMI dataset (N=100 learners from Mali), capturing both skill-related and contextual factors. Our comprehensive evaluation demonstrated that SMOTE resampling enhanced the classifier performance across all classifiers, with the best models (Random Forest, Gradient Boosting, and XGBoost) exceeding 90% accuracy, outperforming ROS. These results confirm the effectiveness of synthetic oversampling in predicting knowledge levels in low-resource educational settings. Despite these promising results, the study presents certain limitations. The relatively small sample size, composed of children aged 6 to 11 with heterogeneous skill profiles, may constrain the generalizability of the findings. Furthermore, the data were collected within a specific cultural and educational environment, which suggests that the models' applicability to other populations and learning contexts remains to be empirically verified.

Future research directions should include the exploration of additional evaluation metrics beyond accuracy, precision, recall, and F1-score, such as model calibration and fairness assessments, to provide a more comprehensive understanding of model performance. Moreover, extending the pipeline to incorporate advanced NLP techniques beyond classical preprocessing, age-stratified modeling, and enhanced data augmentation strategies could further improve predictive reliability. Such advancements would contribute to developing robust, context-aware learning analytics frameworks capable of supporting personalized education in diverse and resource-constrained settings.

6. Declaration on Generative AI

During the preparation of this work, the authors used standard proofreading services (e.g., Overleaf/ Grammarly) solely for grammar and spelling checks. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Afzal, F. Gul, I. Muzzamil, Rethinking smart learning environments: addressing equity, engagement, and future challenges, Contemporary Journal of Social Science Review 3 (2025) 1418–1432.

[2] J. Hallström, P. Norström, K. J. Schönborn, Authentic stem education through modelling: an international delphi study, International Journal of STEM education 10 (2023) 62.

[3] A. Abyaa, M. Khalidi Idrissi, S. Bennani, Learner modelling: systematic review of the literature from the last 5 years, Educational Technology Research and Development 67 (2019) 1105–1143.

[4] M. Sabin, J. Impagliazzo, H. Alrumaih, C. Tang, M. Zhang, It2017 report: Implementing a competency-based information technology program, in: Proceedings of the 49th acm technical symposium on computer science education, 2018, pp. 1045–1046.

[5] S. Ajili, R. Cheour, M. Abid, M. Baklouti, R. Hotte, Predicting digital literacy gains in rural contexts using multidimensional data: A machine learning approach, in: 2025 22th IEEE ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), IEEE, 2025.

[6] H. Yuan, A. A. Hernandez, User cold start problem in recommendation systems: A systematic review, IEEE access 11 (2023) 136958–136977.

[7] A. Kumar, N. J. Ahuja, An adaptive framework of learner model using learner characteristics for intelligent tutoring systems, in: Intelligent Communication, Control and Devices: Proceedings of ICICCD 2018, Springer, 2019, pp. 425–433.

[8] D. Baig, D. Nurbakova, B. Mbaye, S. Calabretto, Knowledge graph-based recommendation system for personalized e-learning, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 561–566.

[9] Y. M. Hemmler, D. Ifenthaler, Indicators of the learning context for supporting personalized and adaptive learning environments, in: 2022 international Conference on advanced learning technologies (ICALT), IEEE, 2022, pp. 61–65.

[10] D. F. Murad, Y. Heryadi, S. M. Isa, W. Budiharto, Personalization of study material based on predicted final grades using multi-criteria user-collaborative filtering recommender system, Education and Information Technologies 25 (2020) 5655–5668.

[11] Y. Bai, M. Bosu, D. Abuaiadah, Predicting learning outcomes in an online learning platform, in: H. Sharifzadeh (Ed.), Proceedings: CITRENZ 2023 Conference, Auckland, 27–29 September, ePress, Unitec, 2024, pp. 78–90. URL: https://doi.org/10.34074/proc.240111. doi:10.34074/proc.240111.

[12] T. Wongvorachan, S. He, O. Bulut, A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining, Information 14 (2023) 54.

[13] Y. Chachoui, N. Azizi, R. Hotte, T. Bensebaa, Enhancing algorithmic assessment in education: Equifused-data-based smote for balanced learning, Computers and Education: Artificial Intelligence 6 (2024) 100222.

[14] R. Hotte, A. Masmoudi, A. Jaballah, O. Masmoudi, A. A. Maïga, Work-in-progress about dynamicity as a foundation for ami, a mobile intelligent and adaptive learning system, in: Interactive Mobile Communication, Technologies and Learning, Springer, 2021, pp. 111–119.

[15] T. Watthaisong, K. Sunat, N. Muangkote, Comparative evaluation of imbalanced data management techniques for solving classification problems on imbalanced datasets, Statistics, Optimization & Information Computing 12 (2024) 547–570.