Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Predicting First-Year Student Performance with SMOTE-Enhanced Stacking Ensemble and Association Rule Mining for University Success Profiling

Philippe Boribo Kikunda [1,2,3,*], Issa Tasho Kasongo [2], Thierry Nsabimana [3], Jérémie Ndikumagenge [3], Longin Ndayisaba [3], Elie Zihindula Mushengezi [2], and Jules Raymond Kala [1]

[1] Computer Science Department, Faculty of Sciences, Université Catholique de Bukavu (UCB), PO Box 285, Bukavu, Democratic Republic of Congo; e-mail: kikunda.boribo@ucbukavu.ac.cd, raymondkala1@gmail.com
[2] Management Computer Department, Institut Supérieur Pédagogique de Bukavu (ISP/Bukavu), PO Box 854, Bukavu, Democratic Republic of Congo; e-mail: tasho.issa@gmail.com, eliezihindula@yahoo.fr
[3] Doctoral school of the University of Burundi, Center for Research in Infrastructure, Environment and Technology (CRIET), Bujumbura, Burundi;
e-mail: thierry.nsabimana@ub.edu.bi, jeremie.ndikumagenge@ub.edu.bi, longin.ndayisaba@ub.edu.bi
* Corresponding Author : Philippe Boribo Kikunda

Abstract: This study examines the application of Educational Data Mining (EDM) to predict the academic performance of first-year students at the Catholic University of Bukavu and the Higher Institute of Education (ISP) in the Democratic Republic of Congo. The primary objective is to develop a model that can identify at-risk students early, providing the university with a tool to enhance student support and academic guidance. To address the challenges posed by data imbalance (where successful cases outnumber failures), the study adopts a hybrid methodological approach. First, the SMOTE algorithm was applied to balance the dataset. Then, a stacking classification model was developed to combine the predictive power of multiple algorithms. The variables used for prediction include the National Exam score (PEx), the secondary school track (Humanities), and the type of prior institution (public, private, or religious-affiliated schools), as well as age and sex. The results demonstrate that this approach is highly effective. The model is not only capable of predicting success or failure but also of forecasting students' performance levels (e.g., honors or distinctions). Moreover, the use of the Apriori association rule mining algorithm allowed the identification of faculty-specific success profiles, transforming prediction into an interpretable decision-support tool. This research makes several significant contributions. Practically, it provides the University of Bukavu with a tool for student orientation and early risk detection. Methodologically, it illustrates the effectiveness of a combined approach to EDM in an African context. However, the study acknowledges certain limitations, including the non-public nature of the data and the geographical specificity of the sample. It therefore proposes avenues for future research, such as the integration of Explainable AI (XAI) techniques for more refined and transparent analysis of the results.

Keywords: Apriori algorithm; Data mining; Educational data mining; Higher education; Performance prediction; SMOTE; Stacking ensemble; Student success.

Received: August, 6th 2025
Revised: September, 24th 2025
Accepted: September, 26th 2025
Published: September, 30th 2025

1. Introduction

The transition to higher education is a crucial stage in a student's academic career. Failure during the first year often becomes a strong demotivating factor that discourages students from continuing their studies. At the Catholic University of Bukavu and the Higher Institute of Education (ISP), as in most higher education institutions in Bukavu, the first year of university is a critical period that can significantly shape a student's academic path. As shown in [1], a strong correlation exists between first-year results and subsequent-year performance.

Today, almost all universities rely on computerized systems that store large amounts of student data year after year, including academic backgrounds and results. Unfortunately, these

DOI : 10.62411/jcta.14043

systems are generally limited to operational tasks such as automated deliberation, report card printing, and financial management. While such uses are important, they overlook the potential value of these data for deeper analysis.

The growing volume of academic data represents a major asset for data mining technologies [2]. Unlike storage technologies, which are primarily designed to organize and retrieve information, data mining seeks to uncover new knowledge hidden in these datasets [3]. This process provides valuable leverage for decision-making, particularly in guiding program orientation and improving institutional support strategies[4], [5].

Data mining, defined as the extraction of knowledge from large databases [6], is widely applied in various domains. In e-commerce platforms, it contributes to increased productivity, profits, and customer retention. In e-learning environments, it plays a key role in enhancing the quality of teaching and learning [7]. Data mining relies on statistical methods, machine learning (ML), and artificial intelligence (AI) tools. When specifically applied to educational data, the field is referred to as educational data mining (EDM) [8]. Within education, EDM can be applied at several levels: to predict outcomes for individual courses [9], to forecast performance across entire study programs [10], or even to anticipate test results in intelligent tutoring systems [11]. These applications highlight EDM's potential to provide actionable insights into student learning and institutional performance.

However, predicting student outcomes remains a challenging task. The complexity of influencing factors, combined with the natural imbalance in student data—where the number of successful students typically exceeds that of unsuccessful ones, renders many conventional models ineffective. To address these limitations, an advanced combined methodological approach was employed. Specifically, the SMOTE algorithm was applied to correct class imbalance and reduce bias toward the majority class. Additionally, stacking, an ensemble learning technique, was employed to combine the strengths of multiple classifiers. Through this approach, a model was constructed that is both more robust and more accurate than any single classifier.

The primary objective of this research is to apply a combined methodology at the Catholic University of Bukavu and ISP to predict the success of first-year students. The predictive factors include national secondary school examination results, the type of secondary school attended, the field of study pursued at the secondary level, as well as demographic attributes, such as age and gender. More specifically, this study aims to:

- Build a model capable of accurately predicting a student's success or failure in their first year.
- Identify the student's likely performance level (e.g., grades such as "A" or "B") for targeted support.
- Apply association rule mining (Apriori) to identify key factors that explain success or failure within each program of study, thereby transforming prediction into an interpretable decision-making tool.

Such models serve a dual purpose. On the one hand, they can help students align themselves with study programs that best match their profiles. On the other hand, they can support administrators in improving admission policies, while enabling early identification of students at risk of failure so that timely interventions can be provided [12]–[14].

ML is among the most widely used families of techniques in data mining. These tools are generally divided into two main categories: supervised learning and unsupervised learning. Classification tasks fall under supervised learning, whereas rule extraction belongs to unsupervised learning. The purpose of classification is to build a model that assigns a class or label to objects in the dataset. Ideally, the generated model should not only fit the training data but also generalize well to unseen test data [15].

By contrast, the purpose of rule extraction is to uncover relationships among variables that describe a dataset, expressed in the form of implications. In other words, rules connect one or more attributes to another attribute, resulting in "if-then" statements about the data. In EDM, such rules allow researchers to study groups of students and identify common characteristics underlying their behavior. For example, association rules have been applied to admissions data [16] to provide knowledge that supports program-level admission decisions. The remainder of this paper is structured as follows. Section 2 presents the materials and methods, followed by results in Section 3. Section 4 discusses the findings, and Section 5 concludes with final remarks and implications.

2. Literature Review

Research into the use of data mining algorithms in education has shown promising results for predicting students' academic performance. Numerous studies have been conducted using various classifiers, including Bayesian networks, decision trees, random forests, genetic algorithms, k-nearest neighbours (KNN), and ensemble methods. These approaches have been applied to datasets from thousands of students across different educational contexts, consistently demonstrating the potential of data mining in this field.

For example, in a seminal study [17]], student performance was predicted using only academic grades, without including socio-economic variables, and still achieved satisfactory results. The model relied on pre-university grades as well as first- and second-year course results to forecast success in a university program.

Other studies, such as [18]–[20], developed models that incorporated personal, pre-university, and university characteristics, including gender, year and place of birth, place of residence, country, cumulative scores from previous education, current semester performance, and overall university scores. However, despite the wide range of features, these models did not achieve an accuracy rate exceeding 70%.

In addition, [21] extended this line of research by focusing on small-sample modelling. Their analysis, based on 50 student records from the Master of Administration program at the British University in Dubai, revealed that graduate performance could be explained by age, secondary-level GPA, and bachelor-level GPA. This study demonstrated that even limited datasets can yield meaningful insights when analyzed with appropriate methods.

Class imbalance is another frequent challenge in academic databases. Several studies have addressed this issue using oversampling techniques, most notably SMOTE and its variants [22], [23]. On the other hand, research such as [24] integrated broader contextual information—including school of origin, English proficiency, admission test scores, and socio-economic factors—across different universities. Their results achieved prediction rates exceeding 90% for binary classification (Pass/Fail). However, as the number of classes increased, accuracy declined significantly. Decision trees and Bayesian networks were among the models used in this context.

The effectiveness of prediction models has also been enhanced through the use of ensemble learning methods. For instance, [25] proposed a hybrid approach combining multiple classifiers with genetic algorithms to predict final scores. This method achieved improvements of approximately 10–12% compared to traditional aggregation strategies, highlighting the benefit of combining different learners.

Beyond simple prediction, extracting interpretable knowledge from educational data is essential for informed decision-making. The use of association rules—such as those generated by the Apriori algorithm—has been explored to uncover patterns of success and failure [24]. Such interpretability is crucial for educational decision-makers. For example, in [26], Apriori-based analysis was applied to student participation in assignments, assessments, and attendance. The resulting rules enabled the identification of average and below-average students, thereby facilitating targeted interventions aimed at improving academic performance.

Based on this review of prior work, the following research hypotheses were formulated:

- Information related to students' secondary school background, such as state examination scores, field of study, and type of school attended, a significant predictor of first-year university performance.
- The use of a combined methodology involving SMOTE for class imbalance and stacking for ensemble learning will significantly improve predictive performance compared to single-algorithm approaches.
- The Apriori algorithm can be successfully applied to discover student success and failure profiles, providing valuable interpretability and actionable insights that can be directly used by university administration.

3. Proposed Method

3.1. Educational Context

The secondary education system in the Democratic Republic of the Congo (DRC) is a six-year program that places a strong emphasis on early specialization. Unlike many Western systems, where specialisation typically occurs at the university level or near the end of secondary school, Congolese students are required to choose a specific course of study, or option, as early as their third year. After completing a two-year core curriculum, students specialize in one of several streams, which range from science and literature to more technical areas, such as education, business management, or agricultural and veterinary sciences. The most common streams include:

- Scientific: Focused on hard sciences (mathematics, physics, chemistry, biology). It prepares students for studies in engineering, medicine, computer science, and other science-related fields.
- Literary: Oriented toward arts and humanities (philosophy, history, geography). It provides access to fields such as Law, journalism, psychology, and other humanities disciplines.
- Commercial: Designed to prepare students for careers in business (management, accounting, economics). This stream is often chosen by those intending to pursue administration, finance, commerce, or banking.
- Education: Intended to train future primary and preschool teachers, with an emphasis on educational sciences and teaching methodology.
- Technical: Focused on professional and technical fields (electricity, mechanics, construction, tailoring, and related skills). It equips students to work directly as technicians or to pursue further technical studies.
- Agricultural and Veterinary: Concentrated on agriculture and livestock farming. Students acquire practical knowledge of cultivation and animal husbandry, which prepares them for employment in agriculture or for advanced studies in agronomy.

In principle, this structure is designed to prepare young people for either university education or direct entry into the labor market. However, the current realities of the job market in the DRC make a university degree almost indispensable, raising questions about whether early specialization is sufficient to meet professional requirements.

The complexity of this system is further accentuated by its governance. Although the education sector is officially overseen by a centralized national ministry, in practice it is considerably decentralized and involves a wide range of actors, including religious institutions.

The case of Catholic schools is particularly notable. These institutions are public schools under special management (often referred to as "contract schools"), which differentiates them from entirely private or fully religious schools. While they receive state funding, their operational management is entrusted to the Catholic Church. This arrangement has a direct impact on the quality of infrastructure and teacher training. As a result, Catholic schools often achieve better academic outcomes compared to private schools, and especially compared to public schools run directly by the state. This hybrid model creates a highly diverse educational landscape, where the quality of education varies significantly depending on the type of institution (public, private, or private under contract).

Accordingly, the Congolese education system can be categorized by its management structures as follows:

- College: A school managed by the Catholic Church and the Jesuit fathers.
- Lycée: A Catholic school specifically intended for girls.
- Complex School: A private school managed by an individual.
- Institute: A public or private educational establishment managed either by the Congolese state or by private individuals.
- Technique Institute Fundi Maendelo (TIFM): An institution offering technical training (e.g., mechanics, electricity, secretarial studies), managed under either public or private authority.

3.2. Data

The data used in this study consist of personal and academic information from students enrolled in five programs: agronomy, medicine, computer science, Law, and economics at the Catholic University of Bukavu and the Higher Institute of Education (ISP). The sample includes more than 3,000 students drawn from five first-year cohorts between 2015 and 2020. As presented in Table 1, the dataset contains variables related to first-year academic results at the university, including grades (mention), faculty, age, gender, type of school of origin, secondary level section, and the percentage obtained at the end of secondary school.

The national exam score (PEx) is widely used as a proxy for academic achievement in the DRC. However, it does not always fully capture students' competencies due to variations in grading practices, unequal resource distribution across schools, and regional disparities. Furthermore, since national education systems differ worldwide, the PEx score should be contextualized or mapped to equivalent indicators when applying this framework to other countries.

Table 1. Data description.

Variable	Description	Values Taken
Count	Total number of records	3509
SchoolP	Type of school of origin	College, Institute, School Complex, High School, Other
Gender	Sex of student	Male, Female
Age	Age at university entry	≥ 15
SectionH	Section taken for humanitarian studies	Biochemistry, Math-Physics, etc.
PEx	Percentage obtained in state exams	≥ 50
Faculty	Faculty chosen by the student	Medicine, Agronomy, Computer Science, Economics, Law
PG1	Percentage obtained at the end of the first year	≥ 20
Year	Academic year	2015–2020
Decision	Final grade obtained	A, S, D, GD, PGD

Grading in this system follows a categorical structure. An 'A' grade indicates failure, as the student did not meet the deliberation criteria (usually <54%). An 'S' grade means satisfaction, corresponding to a score between 55% and 69%. A 'D' corresponds to 70–79%, while 'GD' represents 80–89%. Finally, 'PGD' is awarded to students achieving 90% or higher. Out of the 2,818 records used in the analysis, 1,959 were labeled 'S', 564 'A', 220 'D', 32 'GD', and 1 'PGD'. For analytical purposes, the categories 'PGD' and 'GD' were merged into a single group, as were 'AA' and 'A'. All data mining techniques in this study were implemented using Python 3.9, while R 4.4.2 was used to create and visualize the decision tree.

3.3. Methodology

Designing a model to predict students' results at the end of the first year of university involves searching for a hypothesis that best fits the data. As noted by Barra [15], this process can be carried out at two levels:

1. Identifying a subset of hypotheses that perform well with the data.
2. Selecting the optimal hypothesis from within that subset.

At the first level, multiple families of classification algorithms, including decision trees, random forests, and neural networks, are explored. In line with the literature review, it is recognized that no single classifier consistently outperforms others across all contexts. Therefore, to address the first and second research objectives, the classifiers most suitable to the dataset are investigated.

Before applying algorithms, however, a comprehensive data preparation phase is required. This phase includes cleaning, attribute generation, handling missing values, transformation, and attribute selection, as shown in Figure 1 (Framework for Developing the Predictive Model). After preprocessing, the dataset is randomly divided into training (70%) and testing (30%) subsets. The training set is used to construct the model, while the test set serves for evaluation to ensure that the model generalizes beyond the training data.

To answer Research Question 1, grades 'A' and 'AA' were grouped as failures, while grades 'S', 'D', 'GD', and 'PGD' were grouped as passes (binary classification). As the descriptive dataset (see Table 1) shows, the class distribution is unbalanced, with success cases dominating. In this context, accuracy alone is insufficient, so precision, recall, and the kappa coefficient are also used as evaluation metrics. For Research Question 2, the same methodology is applied but with multiple categories (A, S, D, GD) instead of just binary Pass/Fail classes.

The second level of analysis focuses on optimizing model parameters. Cross-validation was employed, and the three classifiers that performed best were then selected. These classifiers were subsequently combined using the stacking algorithm, with SMOTE applied to balance the training data. This process is also illustrated in Figure 1.

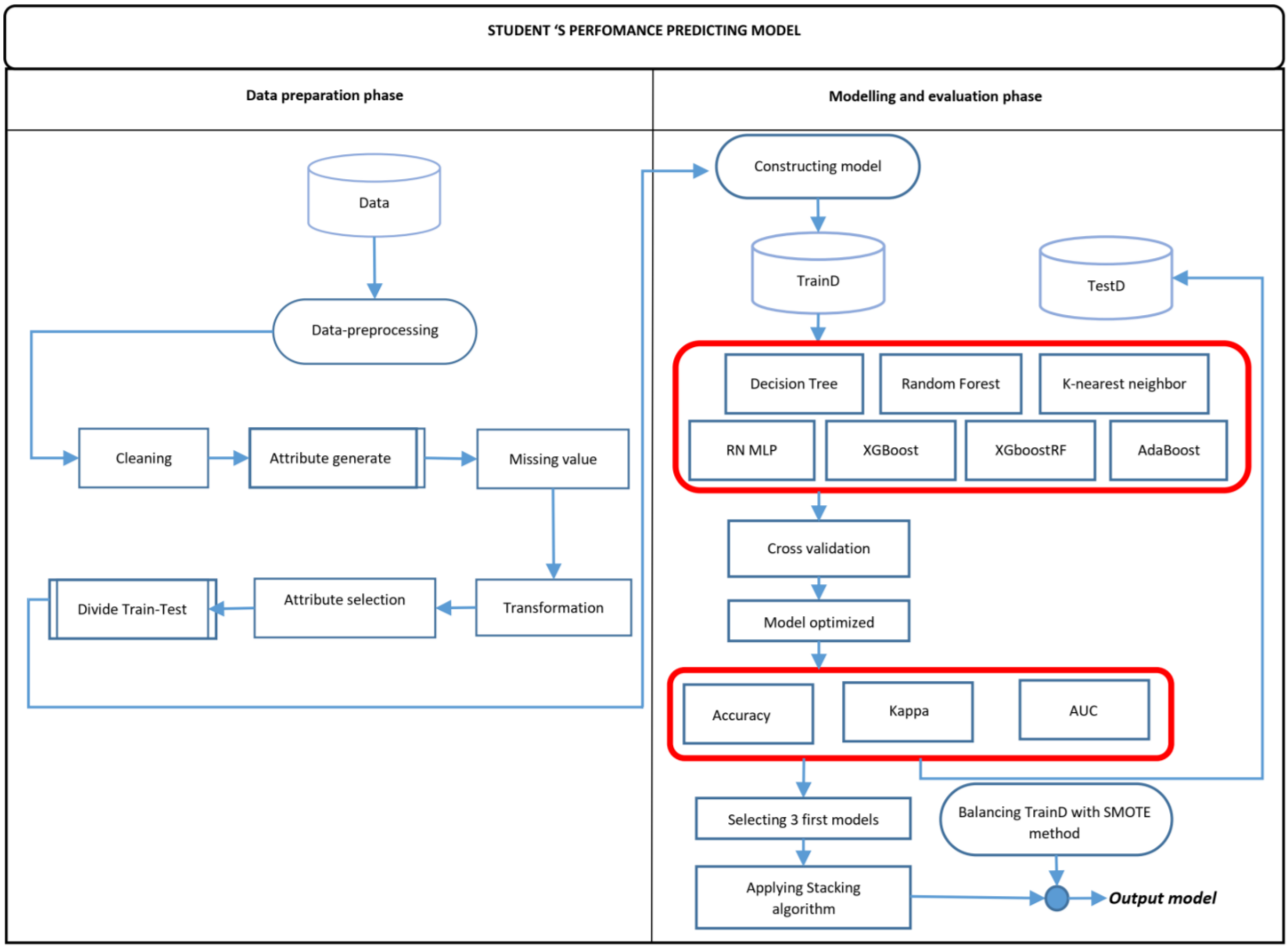

Figure 1. Framework for developing the predictive model

To achieve Objective 1, grades 'A' and 'AA' were grouped as failures, while grades 'S', 'D', 'GD', and 'PGD' were grouped as passes (binary classification). As the descriptive dataset (see Table 1) shows, the class distribution is unbalanced, with success cases dominating. In this context, accuracy alone is insufficient, so precision, recall, and the kappa coefficient are also used as evaluation metrics. To achieve Objective 2, the same methodology is applied but with multiple categories (A, S, D, GD) instead of just binary Pass/Fail classes.

The second level of analysis focuses on optimizing model parameters. Cross-validation was employed, and the three classifiers that performed best were then selected. These classifiers were subsequently combined using the stacking algorithm, with SMOTE applied to balance the training data. An overview of this process is provided in Figure 1.

To achieve Objective 3, attention is directed to the subset of students who were successful (grades A, S, D, GD, PGD). The data are subdivided into groups based on the percentage obtained in the first year. To better capture performance variation, results were reclassified into five categories: Fair (53–59), Fairly Good (60–64), Good (65–75), Very Good (76–80), and Excellent (>80). The Fair category is considered the at-risk group. The Apriori algorithm was then applied to identify the factors explaining success within each program of study. The workflow for this step is presented in Figure 2 (Framework for Identifying Success Profiles).

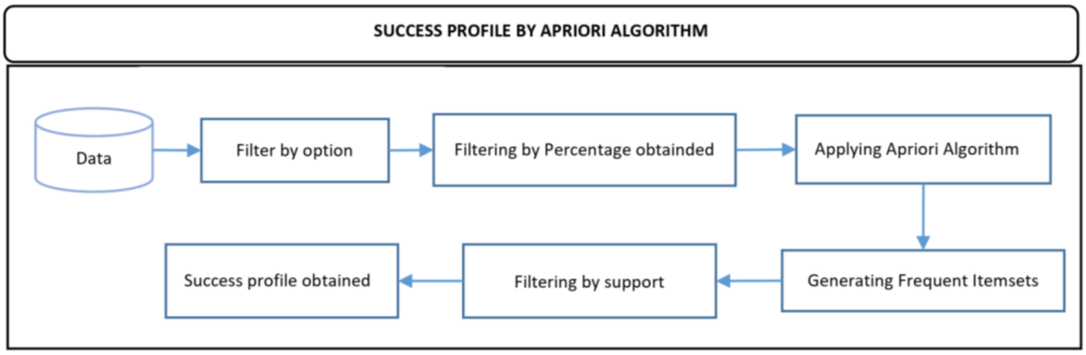

Figure 2. Framework for identifying a success profile

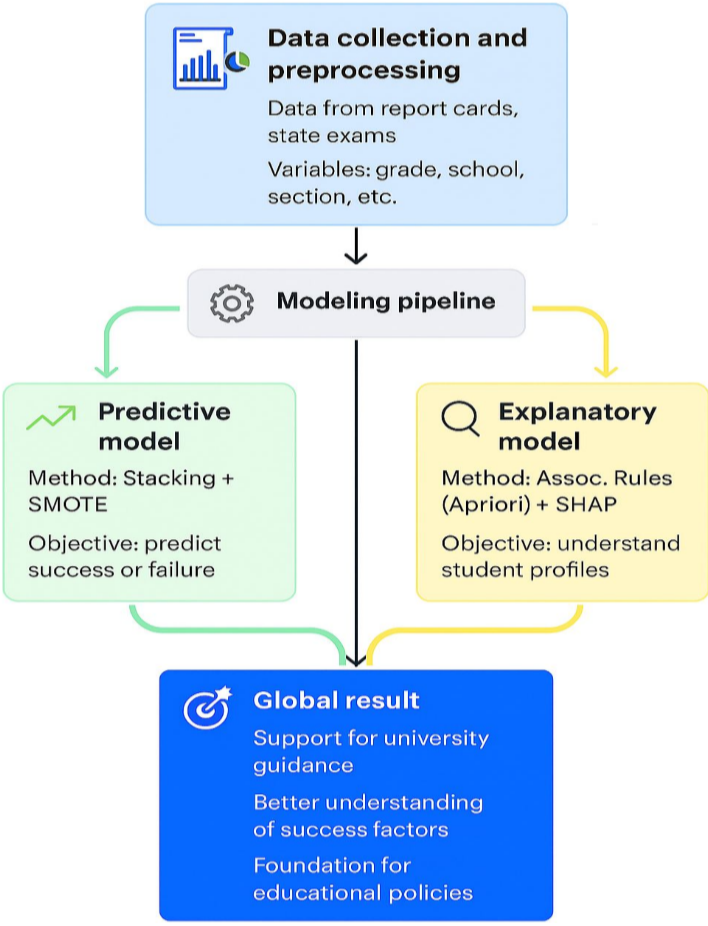

Figure 3. Overall framework

Finally, the overall methodology can be summarized as a two-branch pipeline, as shown in Figure 3 (Overall Framework). The first branch represents the predictive model, which applies SMOTE and stacking to maximize prediction accuracy and identify at-risk students. The second branch represents the explanatory model, which applies the Apriori algorithm to extract interpretable association rules. For example, one discovered rule indicates that "students with an 'A' grade in the State Exam and from a religious school almost always succeed in the Law program." While the predictive branch emphasizes accuracy, the explanatory branch emphasizes interpretability.

In summary, Figure 1 illustrates the predictive modeling process, Figure 2 focuses on extracting success profiles, and Figure 3 integrates both into a comprehensive analytical framework that is both predictive and explanatory. Together, these steps ensure that the model not only predicts performance but also provides actionable insights for academic decision-making.

4. Results and Discussion

4.1. Classification of Pass or Fail

After the preprocessing operations, the sample contained the information presented in Table 1. The first seven algorithms listed in Table 2 were experimented with to create predictive models. From these seven algorithms, the three with the highest accuracy were selected and combined into an eighth model using the stacking algorithm. Since the classes are unbalanced, accuracy alone was not considered sufficient to evaluate model performance. To address this, the SMOTE technique was applied to balance the dataset when implementing the stacking algorithm. Accordingly, Table 2 presents the accuracy, kappa, and AUC values of the eight algorithms with their optimal parameters.

Table 2. Model performance metrics without SMOTE.

Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score	Kappa	AUC
Decision Tree	0.78	0.92	0.78	0.84	0.176	0.77
Random Forest	0.77	0.83	0.77	0.79	0.229	0.76
MLP Classifier	0.79	0.88	0.79	0.82	0.276	0.81
K-Nearest Neighbor	0.77	0.80	0.77	0.78	0.290	0.74
XGBoost Classifier	0.76	0.81	0.76	0.78	0.241	0.76
XGBoostRF Classifier	0.79	0.88	0.79	0.82	0.243	0.81
AdaBoost Classifier	0.77	0.83	0.77	0.79	0.236	0.62
Stacking Classifier	0.79	0.95	0.79	0.85	0.166	0.81

Table 3. Model performance metrics with binary classification using SMOTE.

Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score	Kappa	AUC
Decision Tree	0.74	0.73	0.74	0.73	0.354	0.75
Random Forest	0.75	0.76	0.75	0.76	0.293	0.76
MLP Classifier	0.76	0.76	0.76	0.76	0.337	0.78
K-Nearest Neighbor	0.62	0.67	0.62	0.60	0.247	0.72
XGBoost Classifier	0.76	0.76	0.76	0.76	0.316	0.76
XGBoostRF Classifier	0.77	0.77	0.77	0.77	0.368	0.80
AdaBoost Classifier	0.75	0.75	0.75	0.75	0.297	0.68
Stacking Classifier	0.80	0.80	0.80	0.80	0.407	0.81

In relation to the first objective, namely predicting whether a candidate can succeed in a chosen university program, the results are positive. In addition to the results presented in Table 2, where the stacking classifier achieved a kappa of 0.40 (greater than 0.3), an accuracy of 0.80, and an AUC of 0.81, the confusion matrix in Table 4 also supports this finding. For example, in the first row representing the failure class (class 0), out of 179 students, the classifier correctly predicted 101 as failures, resulting in a recall of 56% (101/179). Similarly, in the first column, out of 196 predicted failures, 101 were actual failures, resulting in a class-specific accuracy of 52% (101/196). This value is much higher than the baseline probability of a student failing (179/844 = 33%). The same interpretation applies to the pass class (class 1).

As mentioned earlier, the kappa metric is particularly important for performance evaluation in cases of class imbalance. Comparing Tables 2 and 3, it is clear that applying the SMOTE technique improves the kappa values across most classifiers. Table 4 presents the ablation study evaluating the effect of removing each variable on the performance of the stacking model. The results indicate that Type of School, Section, and PEx have the most significant impact on prediction quality. The performance metrics (accuracy, precision, recall, kappa, and AUC) illustrate how the model's predictive ability changes when these variables are excluded.

Table 4. Variable ablation and model performance.

Model Variant	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score	Kappa	AUC
With all variables	0.80	0.80	0.80	0.80	0.407	0.81
Without the Type of School	0.76	0.77	0.76	0.77	0.328	0.76
Without Section	0.75	0.76	0.75	0.76	0.279	0.79
Without Age	0.78	0.79	0.78	0.78	0.373	0.80
Without Sex	0.79	0.81	0.79	0.80	0.368	0.80
Without PEx	0.74	0.75	0.74	0.75	0.269	0.75
Without Faculty	0.79	0.79	0.79	0.79	0.406	0.81

4.2. Classification as A, S, D, GD

Table 5 presents the accuracy, precision, recall, and kappa values of eight learning algorithms tested for predictive modeling using the methodology described earlier. It is observed that only the stacking classifier achieves a kappa coefficient of 0.3 or higher, with an overall accuracy of 67%.

In relation to the second objective, which aims to determine the candidate's level of success based on the chosen option, it can be concluded that the stacking model can predict the success level of a candidate with 67% accuracy according to their profile. When comparing the performance on minority classes in the confusion matrix (Table 6), the predictive accuracy for classes A (0.48), D (0.45), and GD (0.11) is notably higher than the proportion of these classes in the entire dataset (A: 0.21, D: 0.11, GD: 0.01). For the Satisfaction class, which represents the majority of students, the predictive accuracy reaches 74.5%.

Table 5. Model performance metrics with multi-class classification using SMOTE.

Model	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score	Kappa
Decision Tree	0.68	0.76	0.67	0.71	0.220
Random Forest	0.68	0.78	0.68	0.71	0.200
MLP Classifier	0.69	0.89	0.69	0.76	0.144
K-Nearest Neighbor	0.67	0.73	0.67	0.69	0.236
XGBoost Classifier	0.68	0.77	0.68	0.71	0.210
XGBoostRF Classifier	0.70	0.86	0.70	0.76	0.231
AdaBoost Classifier	0.67	0.77	0.67	0.71	0.197
Stacking Classifier	0.67	0.66	0.67	0.66	0.308

Table 6. Confusion matrix – stacking classifier

Binary Classification (1,0)	Predicted 0	Predicted 1		
Actual 0	101	78		
Actual 1	95	570		
Multiclass Classification (A, S, D, GD)	**Pred. A**	**Pred. S**	**Pred. D**	**Pred. GD**
Actual A	116	64	3	1
Actual S	74	424	35	1
Actual D	6	55	33	3
Actual GD	0	6	2	1

4.3. Bias Analysis

Table 7 presents an in-depth analysis of model performance across sensitive variables, specifically gender, academic track, and school type, highlighting disparities that range from moderate to substantial. From a gender perspective, the model demonstrates relatively balanced performance, with an F1-score of 78% for female students and 76% for male students, despite the underrepresentation of females in the dataset. This suggests the absence of significant gender bias and indicates that the model generalizes reasonably well across both groups.

Regarding academic tracks, higher F1-scores are observed for students in the literary and scientific streams (≥78%), while the commercial, pedagogical, and agronomy tracks exhibit comparatively lower performance (≤74%). These differences may be attributed to higher data quality or greater internal consistency within the stronger-performing tracks. Class imbalance is considered a less likely explanation, given the overall distribution of the dataset.

Table 7. Bias analysis.

Group	Values	Precision	Recall	F1-score
Gender	Male	0.77	0.76	0.76
	Female	0.77	0.78	0.78
Section	Scientific	0.77	0.79	0.78
	Literary	0.81	0.82	0.81
	Pedagogy	0.75	0.73	0.74
	Commercial	0.73	0.74	0.74
	Agronomy	0.81	0.70	0.73
School Type	Catholic (College and Lycée)	0.90	0.91	0.90
	Public (Institute and Edap)	0.77	0.77	0.77
	Private (CS)	0.68	0.66	0.67

A particularly important observation concerns the relationship between academic track and school type. The literary track is predominantly offered within Catholic schools, which also show the highest overall performance (F1-score of 90%), significantly outperforming public schools (77%) and private schools (67%). This suggests that the model may be capturing more structured and consistent patterns present in Catholic school data.

Such patterns could stem from standardized educational practices, stricter evaluation protocols, or a more homogeneous student population. In contrast, private schools exhibit greater data heterogeneity, which may be attributed to less centralized governance and more variable assessment standards. For instance, in many private institutions, students finance staff salaries directly, which may reduce the imposition of uniform evaluation practices. This variability likely introduces additional noise into the dataset, which may explain the comparatively lower model performance observed for private schools.

4.3. Success Profile by Option

In this section, data from five faculties, including Medicine, Agronomy, Law, Computer Science, and Economics, are analyzed. After the sample data were filtered by study option, the Apriori algorithm was applied to identify recurring patterns associated with student success. In relation to the third objective, which seeks to identify success factors for each program of study, the results confirm that meaningful profiles can indeed be extracted. The findings are summarized in Table 8. For Medicine and Agronomy, completing the Biochemistry section at the secondary level is a major advantage for success.

In contrast, success in Law and Economics is more strongly associated with the Literary and Commercial sections, respectively. In Computer Science, female candidates face greater difficulties compared to male candidates. Excellence in Medicine, Agronomy, and Economics is strongly influenced by achieving a score of at least 70% on the national exam and attending college-type schools. For Computer Science, however, a 60% threshold and an ITFM-type school background are sufficient indicators of strong performance.

4.4. Discussion

This study developed effective predictive models to anticipate the success or failure of first-year university students by leveraging demographic, academic, and institutional data. Results highlight that school type, program section, and secondary exam percentage are among the strongest predictive variables. The integration of machine learning algorithms with the SMOTE rebalancing technique improved predictive accuracy and robustness, while the Apriori algorithm enhanced interpretability by identifying clear success and risk profiles.

Table 8. Success profiles by study program were identified with the Apriori algorithm.

Study Programme	Success Profile (Frequent Patterns)
Agronomy	Biochemistry Section (support = 0.40) Secondary exam score between 60–70% (support = 0.58) Type of school = College (support = 0.50)
Medicine	Biochemistry Section (support = 0.61) Male (support = 0.47) Type of school = College (support = 0.50) Secondary exam score between 70–80% (support = 0.40)
Computer Science	Male (support = 0.65) Type of school = Institute or ITFM (support = 0.51) Secondary exam score between 60–70% (support = 0.50)
Law	Secondary exam score between 60–70% (support = 0.60) Literary Section (support = 0.44) Male (support = 0.77) Type of school = College (support = 0.40)
Economics	Secondary exam score between 60–70% (support = 0.48) Commercial Section (support = 0.51) Female (support = 0.56) Type of school = College or Lycée (support = 0.60)

The model's performance must be interpreted within its educational context. Beyond global metrics, particular attention is given to the recall of the failure class, which is crucial for support policies. A high recall for class "A" (at-risk students) means that the model effectively identifies most struggling students, thus minimizing false negatives. For the Catholic University of Bukavu and the Higher Institute of Education (ISP), this is essential to enable timely interventions and reduce dropout risks.

Conversely, a low recall in class "A" would imply that many at-risk students remain undetected, undermining preventive measures. Precision for class "A" indicates the reliability of predictions: although perfect precision is desirable, slightly lower precision can be acceptable if recall is high, since the cost of unnecessary intervention (false positives) is lower than failing to detect a truly at-risk student (false negatives). In our case, recall for class "A" is 64%, with a precision at 60%. This means the model detects a majority of at-risk students, but about one-third remain unidentified. Meanwhile, 40% of those flagged as at risk may not actually be so, which could lead to unnecessary interventions. Overall, these values reflect an intermediate level of performance, suitable for practical use but with room for improvement.

The application of the Apriori algorithm enabled the identification of variable combinations that contribute to success within specific programs. The resulting profiles reveal strong correlations between secondary-level choices and subsequent academic trajectories. For example, students from the scientific section excel in Medicine and Agronomy, commercial-track students perform better in Economics, literary-track students often succeed in Law, and technical-track students (e.g., electricity, secretarial studies) demonstrate promising outcomes in Computer Science. These findings underscore the long-term impact of secondary-level specialization on higher education outcomes.

Thus, the combination of predictive modeling and association rule mining enhances both performance and interpretability, approaching a valuable decision-support tool. The results align with prior studies emphasizing the importance of academic and socio-demographic variables in student success [20]–[22]. However, this study advances the field by explicitly integrating association rules, which provide more nuanced insights into program-specific success profiles.

The theoretical contribution lies in demonstrating that combining predictive and explanatory methods can strengthen decision-making in higher education. Practically, the models can be used to guide students early, tailor pedagogical interventions, and reduce failure rates. Nonetheless, some limitations remain. The reliance on sensitive variables (type of school, section, PEx) requires vigilance to prevent bias and ensure fairness. The dataset is specific to the DRC, which limits generalizability. Data were collected from a limited number of schools and over a restricted timeframe, which may have affected temporal and geographic representativeness. Moreover, variables such as socio-economic status and student motivation were not available, and privacy concerns limited the inclusion of certain sensitive data. These constraints highlight the need for further studies using broader and more diverse datasets.

5. Conclusions

The potential of EDM to predict student success in the first year at the Catholic University of Bukavu and the Higher Institute of Education (ISP) has been demonstrated in this study. By combining the SMOTE sampling technique with a stacking ensemble model, a robust predictive system was developed that accurately classifies students as successful or unsuccessful, while also addressing multi-class performance prediction. Furthermore, the integration of the Apriori algorithm enabled the extraction of interpretable success profiles tailored to each faculty.

The practical implications of these findings are significant for academic institutions. The proposed model can function as a decision-support tool for student orientation and the early identification of at-risk students. By identifying profiles most likely to succeed in specific academic tracks, universities can implement targeted and personalized support programs, thereby reducing failure and dropout rates. A key finding of the study is the significant impact of three educational background variables: final high school exam score (percentage), academic track followed in secondary school, and type of secondary school attended. These variables consistently emerged as the most important predictors across all modeling approaches. Students with higher exam scores, those from scientific or literary streams, and those who graduated from Catholic schools were more likely to succeed in their first year of university. This highlights the significant impact of prior academic performance, the nature of previous training, and the institutional context on shaping university outcomes. It also suggests that predictive models in education should incorporate contextual variables that reflect differences in preparation and learning environments, rather than treating students as a homogeneous group.

Despite these contributions, the study has notable limitations. The model was trained on a dataset from only two universities, and the data cannot be made publicly available due to confidentiality constraints, which limits replicability. Additionally, the variables considered are primarily academic, excluding socio-economic and behavioral factors that may also impact student performance. These limitations provide avenues for future research. Subsequent studies could expand the dataset to include other Congolese universities to assess the model's generalizability. Moreover, incorporating non-academic variables and adopting explainable AI (XAI) techniques such as SHAP would enhance the model's precision, provide more transparent explanations, and increase its practical utility for academic advisors.

Author Contributions: Conceptualization: P.B.K., I.T.K.; Methodology: P.B.K., T.N.; Software: P.B.K., L.N.; Validation: P.B.K., I.T.K., J.R.K.; Formal analysis: E.Z.M., P.B.K.; Investigation: P.B.K., I.T.K.; Resources: T.N., J.N.; Data curation: J.R.K., L.N.; Writing—original draft preparation: P.B.K.; Writing—review and editing: P.B.K.; Visualization: J.R.K., J.N.; Supervision: J.N., E.Z.M.; Project administration: J.N., J.R.K. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest

References

[1] P. Golding and O. Donaldson, “Predicting Academic Performance,” in *Proceedings. Frontiers in Education. 36th Annual Conference*, 2006, pp. 21–26. doi: 10.1109/FIE.2006.322661.

[2] S. A. Alwarthan, N. Aslam, and I. U. Khan, “Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review,” *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–26, Sep. 2022, doi: 10.1155/2022/8924028.

[3] M. M. Arcinas, una S. Sajja, S. Asif, S. Gour, E. Okoronkwo, and M. Naved, “Role of data mining in education for improving students performance for social change,” *Turkish J. Physiother. Rehabil.*, vol. 32, no. 3, pp. 6519–6526, 2021.

[4] P. B. Kikunda, J. Ndikumagenge, L. Ndayisaba, and T. Nsabimana, “Explainable Bayesian Network Recommender for Personalized University Program Selection,” *J. Comput. Theor. Appl.*, vol. 3, no. 1, pp. 17–33, Jun. 2025, doi: 10.62411/jcta.12720.

[5] K. K. San, H. H. Win, and K. E. E. Chaw, “Enhancing Hybrid Course Recommendation with Weighted Voting Ensemble Learning,” *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 337–347, Jan. 2025, doi: 10.62411/faith.3048-3719-55.

[6] X. Zhang and X. Zhang, “An Overview of Data Mining Techniques for Student Performance Prediction,” in *Artificial Intelligence in Education and Teaching Assessment*, Singapore: Springer Singapore, 2021, pp. 149–159. doi: 10.1007/978-981-16-6502-8_14.

[7] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 40, no. 6, pp. 601–618, Nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[8] M. T. Nguyen and T. T. Nguyen, “Advanced and AI Embedded Technologies in Education: Effectiveness, Recent Developments, and Opening Issues,” *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 191–200, Oct. 2024, doi: 10.62411/faith.3048-3719-19.

[9] M. A. Marjan, M. P. Uddin, and M. Ibn Afjal, “An Educational Data Mining System For Predicting And Enhancing Tertiary Students’ Programming Skill,” *Comput. J.*, vol. 66, no. 5, pp. 1083–1101, May 2023, doi: 10.1093/comjnl/bxab214.

[10] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, “A Comparative Study of Classification and Regression Algorithms for Modelling Students’ Academic Performance,” in *Proceedings of the 8th International Conference on Educational Data Mining*, 2015, pp. 392–395. [Online]. Available: https://files.eric.ed.gov/fulltext/ED560769.pdf

[11] Z. A. Pardos, N. T. Heffernan, B. Anderson, and C. L. Heffernan, “The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks,” in *User Modeling 2007*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 435–439. doi: 10.1007/978-3-540-73078-1_60.

[12] A. Alam and A. Mohanty, “Predicting Students’ Performance Employing Educational Data Mining Techniques, Machine Learning, and Learning Analytics,” in *Communication, Networks and Computing*, 2023, pp. 166–177. doi: 10.1007/978-3-031-43140-1_15.

[13] H. A. Mengash, “Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems,” *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[14] B. I. Igoche, O. Matthew, P. Bednar, and A. Gegov, “Integrating Structural Causal Model Ontologies with LIME for Fair Machine Learning Explanations in Educational Admissions,” *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 65–85, Jun. 2024, doi: 10.62411/jcta.10501.

[15] V. Barra, L. Miclet, and A. Cornuéjols, *Concepts et algorithmes - De Bayes et Hume au Deep Learning*, 4th ed. 2021.

[16] P. Rojanavasu, “Educational Data Analytics using Association Rule Mining and Classification,” in *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*, Jan. 2019, pp. 142–145. doi: 10.1109/ECTI-NCON.2019.8692274.

[17] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining,” *Comput. Educ.*, vol. 113, pp. 177–194, Oct. 2017, doi: 10.1016/j.compedu.2017.05.007.

[18] J. Zimmermann, K. H. Brodersen, H. R. Heinimann, and J. M. Buhmann, “A Model-Based Approach to Predicting Graduate-Level Performance Using Indicators of Undergraduate-Level Performance,” *J. Educ. Data Min.*, vol. 7, no. 3, pp. 151–176, 2015, doi: 10.5281/zenodo.3554733.

[19] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H.-Y. Kwon, and A. Hussain, “Educational data mining to predict students’ academic performance: A survey study,” *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 905–971, Jan. 2023, doi: 10.1007/s10639-022-11152-y.

[20] A. I. Adekitan and O. Salau, “Toward an improved learning process: the relevance of ethnicity to data mining prediction of students’ performance,” *SN Appl. Sci.*, vol. 2, no. 1, p. 8, Jan. 2020, doi: 10.1007/s42452-019-1752-1.

[21] L. M. Abu Zohair, “Prediction of Student’s performance by modelling small dataset size,” *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, p. 27, Dec. 2019, doi: 10.1186/s41239-019-0160-3.

[22] S. Zheng, K. Zhou, and C. Chen, “Perturbation-Based SMOTE for Multi-Class Imbalanced Classification,” in *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*, Oct. 2024, pp. 53–56. doi: 10.1109/ICMLCA63499.2024.10753837.

[23] N. A. Rahmi, S. Defit, and Okfalisa, “Enhancing Classification Performance: A Study on SMOTE and Ensemble Learning Techniques,” in *2024 International Conference on Future Technologies for Smart Society (ICFTSS)*, Aug. 2024, pp. 63–68. doi: 10.1109/ICFTSS61109.2024.10691339.

[24] Nguyen Thai Nghe, P. Janecek, and P. Haddawy, “A comparative analysis of techniques for predicting academic performance,” in *2007 37th annual frontiers in education conference - global engineering: knowledge without borders, opportunities without passports*, Oct. 2007, pp. T2G-7-T2G-12. doi: 10.1109/FIE.2007.4417993.

[25] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, “Predicting student performance: an application of data mining methods with an educational web-based system,” in *33rd Annual Frontiers in Education, 2003. FIE 2003.*, 2003, vol. 1, p. T2A_13-T2A_18. doi: 10.1109/FIE.2003.1263284.

[26] G. V. S. C. S. L. V Prasad and M. Rambabu, “Association Rule Generation for Student Performance Analysis usingApriori Algorithm,” *J. Sci. Technol.*, vol. 7, no. 3, pp. 107–112, 2022, doi: 10.46243/jst.2022.v7.i03.pp107%20-112.