ISSN 2723-6471


# Enhancing SMOTE-ENN Efficacy on Imbalanced Datasets Using Decision Tree Leaf Feature Extraction: A Case Study on Student Employability Data

Rizkysari Meimaharani[1,*], Widowati[2], Ahmad Abdul Chamid[3]

[1]*Doctoral Program of Information Systems, Diponegoro University, Semarang, 50275, Indonesia*

[2]*Department of Mathematics, Diponegoro University, Semarang, 50275, Indonesia*

[1,3]*Department of Informatics Engineering, Universitas Muria Kudus, Kudus, Indonesia*

(Received: February 10, 2026; Revised: April 15, 2026; Accepted: June 5, 2026; Available online: June 28, 2026)

**Abstract**

This study looks at the challenge of classifying tabular data that is highly imbalanced and overlapping, where standard predictive models often lose performance and tend to focus too much on the majority class. Another problem is that many advanced ensemble models are highly complex and lack transparency. These models are often viewed as black boxes, making it difficult for users to clearly and explain how each feature contributes to the final prediction result. This study offers a hybrid classification approach to address the problem, by combining rule extraction from decision tree leaves, SMOTE-ENN resampling technique, and XGBoost algorithm to improve prediction performance more accurately and reliably.The leaf extraction process helps reorganize the data by separating overlapping class regions into clearer and more structured groups before synthetic samples are generated. The test results show that the proposed approach is able to exceed the performance of the baseline model, by obtaining an F1-score of 0.8554 which indicates increased effectiveness and balance in prediction. In addition to improving performance, this method also keeps the model interpretable. Instead of relying only on abstract engineered features, the model allows us to trace important features back to the original decision tree rules. This approach helps explain the prediction formation process more transparently, so that each model decision can be understood clearly, logically, and easily interpreted. Overall, the combination of Decision Tree, SMOTE-ENN, and XGBoost is effective in handling extreme class imbalance, while producing a clear, stable, and easy-to-understand model, making it more reliable and trustworthy in various real-world applications.

*Keywords:* Imbalanced Data, Feature Extraction, Decision Tree, SMOTE-ENN, XGBoost

## 1. Introduction

Digital transformation is reshaping many sectors, including tourism and educational media. In higher education, this shift is reflected in the use of Educational Data Mining (EDM), which focuses on turning academic data into meaningful predictions [1]. One key indicator of university performance is the graduate waiting period, which shows how well educational outcomes match industry needs [2]. By predicting this waiting time, universities can identify students who may face difficulties in entering the job market and provide early support [3]. Using information such as academic performance and study history, institutions can also offer more targeted and personalized career guidance [4].

However, building reliable predictive models in EDM is not easy because real-world data is often imbalanced . In this situation, traditional accuracy can be misleading, since models may achieve high scores simply by focusing on the majority class[5] . As a result, the recall for the minority class drops significantly, even though this class is often the most important for early detection .

In classification problems, overlapping data occurs when instances from different classes share similar feature values, leading to ambiguous decision boundaries. In this study, overlapping refers to students with similar academic and socio-economic attributes (e.g., GPA, study duration, parental background) but different employment outcomes. This condition reduces the effectiveness of distance-based methods such as K-Nearest Neighbor (KNN) [6] and is indirectly observed through misclassification patterns. To address this issue, a hybrid approach is proposed to improve classification performance in overlapping regions

***Corresponding author: Rizkysari Meimaharani (rizky.sari@umk.ac.id)**
**DOI: https://doi.org/10.47738/jads.v7i3.1396**

To address this issue, resampling methods such as SMOTE are commonly used. However, SMOTE has limitations, as it may generate synthetic data in inappropriate regions, which can introduce noise [7]. This often causes overlapping between classes and reduces the model's ability to clearly separate them [8]. In addition, SMOTE may ignore small minority groups and produce samples that lack diversity [9]. Clustering approaches like K-Means are sometimes used to guide resampling, but they do not consider class labels, which limits their effectiveness. Because of these challenges, hybrid methods such as SMOTE-ENN are preferred, as they can both generate new samples and remove noisy data at the same time. These leaf nodes can then be used as new features, helping the model capture more detailed and localized patterns in the data [10]. When combined with models such as XGBoost or SVM, this approach has been shown to improve prediction stability and overall performance [11].

This study proposes a framework that combines decision tree leaf-based feature extraction with SMOTE-ENN to improve the prediction of graduate waiting time. Unlike traditional clustering methods, this approach uses leaf indices as new features to guide the resampling process in a more controlled way. The study evaluates this method against baseline models and K-Means-based approaches. By testing it using Random Forest, SVM, and XGBoost, this research aims to provide a more accurate and reliable solution for predicting graduate employability in higher education.

## 2. Literature Review

### 2.1. Predictive Modeling and Graduate Waiting Periods

The use of EDM has become important for universities to understand and predict how well graduates are prepared for the job market. Studies show that academic records can be transformed into useful insights to support early career interventions. By analyzing data such as grades, study duration, and learning history, universities can identify students who may face difficulties after graduation and provide targeted support before they enter the workforce.

In terms of model implementation, several studies have shown promising results. Applied Random Forest and XGBoost combined with a Voting Classifier and achieved an accuracy of up to 89.9% on engineering student data [3]. Similarly, used the KNN algorithm to predict graduate waiting time and reported an accuracy of 86.25% [2]. When applied to highly imbalanced academic datasets, both accuracy and recall can drop significantly, in some cases reaching as low as 46.9%. This suggests that relying only on baseline models without addressing data imbalance may lead to unreliable predictions.

### 2.2. Limitations of SMOTE and K-Means Clustering

The main reason behind this performance drop is the bias toward the majority class. Previous research explain that many traditional models focus on maximizing overall accuracy, which often leads them to ignore the minority class [12]. As a result, the recall for at-risk students becomes very low, even though this group is the most important for early intervention. This bias can cause high-risk students to be incorrectly classified as safe, increasing the number of false negatives.

To handle this issue, resampling techniques such as SMOTE are commonly used. SMOTE works by generating synthetic samples for the minority class to balance the dataset. However, several studies highlight its limitations. Point out that SMOTE relies on distance calculations in high-dimensional space, which can lead to overlapping between classes [8]. Also note that SMOTE generates new samples without fully considering the actual data distribution, which can introduce noise. In the context of education, this research emphasize that this process may create unrealistic student profiles that do not reflect real-world conditions [13], [14].

Some researchers try to improve SMOTE by combining it with clustering methods such as K-Means. The idea is to group data before generating new samples. However, in research show that this approach still has a key limitation [8]. K-Means does not consider class labels, so it cannot ensure that clusters align with the actual classification targets. This can lead to poor cluster structures and ineffective resampling. Because of these limitations, hybrid approaches such as SMOTE-ENN are considered more reliable.

### 2.3. Tree-Based Feature Transformation as a Solution

Although SMOTE-ENN improves data quality, applying it directly to raw data can still introduce noise, especially when the original features are not well structured. For this reason, feature engineering becomes an important step in improving model performance.

In this approach, each data point is assigned to a specific leaf node, which represents a particular pattern in the data. These leaf indices can then be used as new features, allowing the model to capture more detailed and localized relationships. This method also helps reduce the impact of noisy or less relevant features.

Based on this discussion, there is a clear research gap. Previous studies [14] have not fully addressed the issue of synthetic noise when handling imbalanced data. At the same time, clustering-based approaches like K-Means are limited because they do not consider class labels. Therefore, this study proposes a method that uses decision tree leaf indices to guide the SMOTE-ENN process. By doing this, the generation of new data can be controlled within a supervised structure, reducing noise and improving data quality. This approach directly addresses the limitations highlighted by [15] and aims to produce more reliable predictive models.

## 3. Methodology

### 3.1. Research Framework

To predict graduate waiting time, we designed a step-by-step framework to reduce bias and noise in educational data [16]. As shown in figure 1, the process starts with data preprocessing, followed by decision tree leaf feature extraction to create clearer data boundaries. After that, we addressed class imbalance.

To ensure that each fold remains representative of the original imbalanced class distribution, we implemented 5-Fold Stratified Cross-Validation. This prevents folds with zero minority samples, which would otherwise lead to biased performance estimates. SMOTE-ENN Choice: "We transitioned exclusively to SMOTE-ENN to address the limitations of vanilla oversampling. While SMOTE handles the quantity of data, Edited Nearest Neighbors (ENN) acts as a data cleaning mechanism to remove synthetic noise and overlapping instances, resulting in a more distinct and generalizable decision boundary.

eolSince applying oversampling to the entire dataset can cause overfitting and reduce generalization, we used a strict splitting strategy [17]. We applied Stratified 5-Fold Cross-Validation to maintain class distribution and used SMOTE-ENN only on the training data in each fold [18]. This prevents data leakage and ensures fair evaluation. Finally, we tested XGBoost, SVM, and Random Forest on unseen data to measure real predictive performance

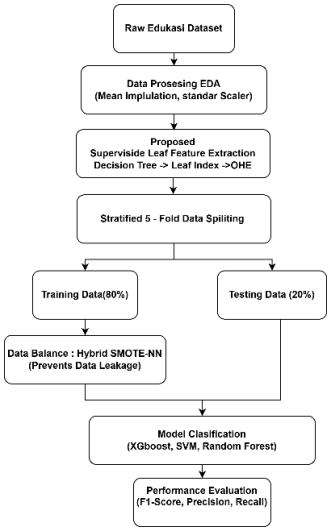


**Figure 1.** Research Framework

### 3.2. Data Acquisition and Exploratory Data Analysis (EDA)

We collected the academic information system dataset from Universitas Muria Kudus Indonesia, focusing on admission scores, semester GPA, and behavioral factors that influence career outcomes as listed in table 1 [19].

**Table 1**. Dataset profile

| Feature Name | Type | Description |
|---|---|---|
| GPA | Continuous | The overall academic achievement score recorded upon graduation |
| Study Duration | Discrete | The total number of academic semesters required to complete the degree program |

| Feature Name | Type | Description |
|---|---|---|
| Parental Status | Binary | The primary occupational sector of the parents (0: Salaried Worker, 1: Business Owner/Entrepreneur) |
| Org. Activeness | Ordinal | The scale of the student's engagement in campus committees and extracurricular organizations |
| Competitions | Discrete | The aggregate count of competitive academic or non-academic events won by the student |
| Certifications | Discrete | The total number of recognized professional or technical competency certificates acquired |
| Scholarships | Binary | Indicates the receipt of institutional or external financial aid during the study period (0: Non-recipient, 1: Recipient). |
| **Waiting Period** | Binary | Target Class: The timeframe taken to secure initial employment post-graduation (0: Over 6 months, 1: 0–6 months) |

The dataset consists of 3240 samples collected from tracer study data during 2022–2025. The class distribution includes 1040 samples in Class 1, 1100 samples in Class 2, and 1100 samples in Class 3. We first performed Exploratory Data Analysis (EDA) to understand the data. The results showed a clear class imbalance, with 70.9% in the majority group (Safe) and 29.1% in the minority group (At-Risk), as shown in figure 2A. This confirms the need for methods like SMOTE to reduce bias [20]. Interestingly, the descriptive statistics in figure 2B show that the minority group has a higher median GPA, around 3.75, compared to 3.45 in the majority group. This suggests that high academic performance alone does not guarantee fast employment, so a more comprehensive approach is needed. We also analyzed feature correlations using a correlation matrix (figure 2C). The results show moderate correlation between the target and features like GPA, around 0.4 to 0.5, while correlations between features remain low, between -0.3 and 0.4. This indicates no serious multicollinearity, meaning the selected features are distinct, less noisy, and suitable for the proposed leaf-based feature extraction.

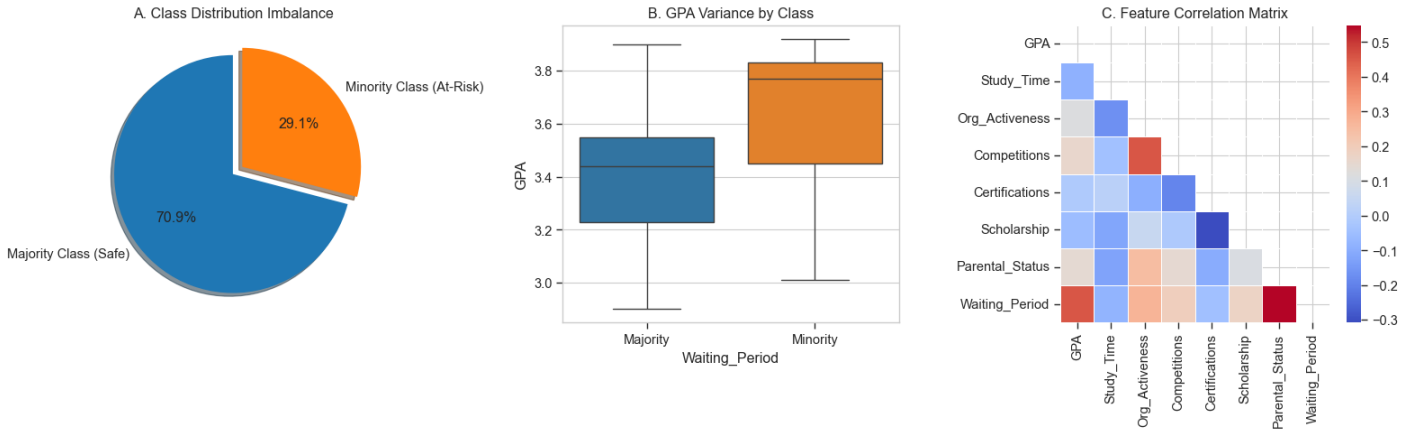


**Figure 2.** Class Distribution Imbalanced (A/left), GPA Variance by Class (B/middle), and Feature Correlation Matrix (C/right)

## 3.1. Proposed Feature Transformation

To overcome the limitations of synthetic oversampling in noisy and complex data, we propose a supervised decision tree leaf feature extraction method. First, we use a decision tree to split the dataset and separate minority samples from majority noise by optimizing the Gini index [21]. The goal is to reduce impurity and create more homogeneous groups. The Gini index is calculated as:

$$GI_X(D) = \sum_{j=1}^{k} P(X = x_i)\left(1 - \sum_{i=1}^{v} Pi\left(Y = y_i | X = x_j\right)^2\right) \quad (1)$$

where $GI_X(D)$ is the Gini score after splitting dataset $D$ using feature $X$, $k$ is the number of partitions, $P\left(X = x_j\right)$ is the probability of data in partition $j$, and $Pi\left(Y = y_i | X = x_j\right)$ is the probability of class $i$ in that partition, with $v$ as the total number of classes. Instead of using the tree for prediction, we map each data point to its final leaf node index $(L_i)$.

 ISSN 2723-6471


This leaf index represents complex patterns in the data. However, using it directly as a number can be misleading, so we convert it into a categorical format using One-Hot Encoding (OHE), as shown in figure 3 [22]. The one-hot encoding of leaf indices increases feature dimensionality based on the number of trees and leaf nodes. It provides a richer representation of decision patterns, which improves classification performance, especially in overlapping data conditions.

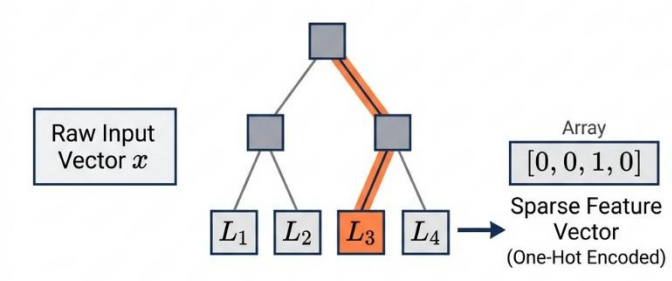


**Figure 3.** Decision Tree Leaf Node to One-Hot Encoding Transformation

## 3.2. Data Level Optimization

To handle the class imbalance found during the exploratory analysis, we applied a data-level approach using the SMOTE-ENN method. First, we used SMOTE to increase the number of minority samples in the transformed feature space. SMOTE works by creating new data points through interpolation between existing minority samples and their nearest neighbors [18]. The new sample $Y_i$ is generated using the formula:

$$Y_i = x + RAND(0,1)(x_i - x) \quad (2)$$

To make the updated supplementary tables as clear as possible, you might structure them to highlight the impact of SMOTE across different thresholds. To evaluate the effectiveness of data balancing techniques on the dataset, a comparative analysis was performed on several model configurations. This evaluation compared three main scenarios: a baseline model without imbalance handling (Baseline), a model with SMOTE oversampling applied (k=5), and a model combining SMOTE with the Tomek Links method. The performance of each configuration was measured using Precision (P), Recall (R), and F1 Score metrics to highlight the impact and effectiveness of using SMOTE in various settings. A comprehensive summary of the performance comparison of these methods is presented in table 2.

**Table 2**.Comparative Analysis

| Model Configuration | Precision (P) | Recall (R) | F1 Score |
|---|---|---|---|
| Baseline (No SMOTE) | 0.85 | 0.42 | 0.56 |
| SMOTE (k=5) | 0.72 | 0.78 | 0.75 |
| SMOTE + Tomek Links | 0.76 | 0.77 | 0.76 |

Where x is a minority sample, xi is its nearest neighbor, and RAND0,1 is a random value between 0 and 1. However, this process can sometimes create samples that overlap with the majority class, especially near decision boundaries, which leads to noise [23]. To reduce this issue, we combined SMOTE with the Edited Nearest Neighbors (ENN) method. ENN checks each data point using the KNN approach. If a sample has a label that does not match the majority of its neighbors, it is considered noise and removed [24]. As shown in figure 4, by removing these unclear samples, SMOTE-ENN helps separate the minority and majority classes more clearly. This step also reduces overfitting and ensures that the model learns from cleaner and more reliable data [18].

ISSN 2723-6471


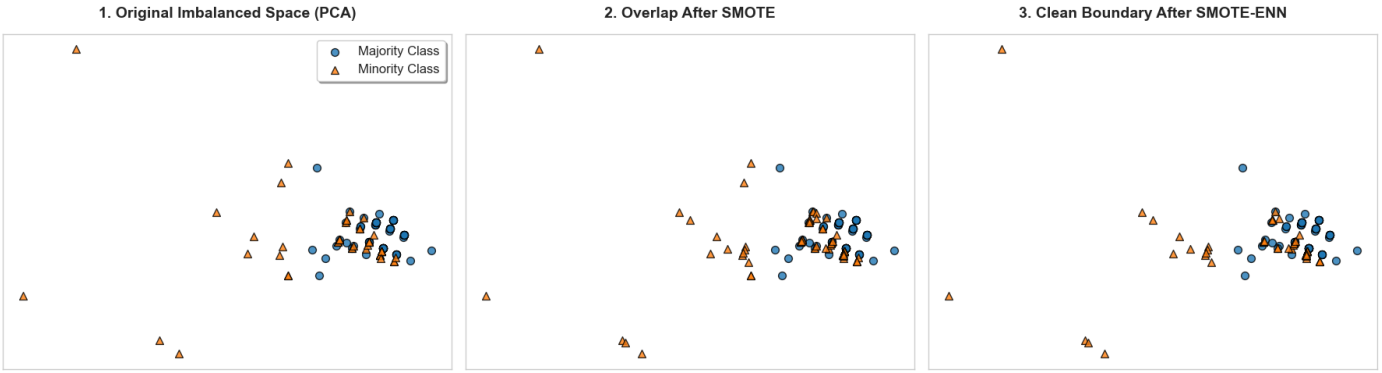


**Figure 4.** Visualization of the hybrid resampling process on the educational dataset using SMOTE-ENN

## 3.3. Classification Models

After balancing the data using SMOTE-ENN, we applied three classification models to evaluate our approach. First, we used XGBoost as the main model because it is effective in handling high-dimensional data. It works by building trees sequentially and correcting previous errors, while also using regularization to reduce overfitting [25]. Next, we used SVM to test how well the data can be separated. SVM maps the data into a higher-dimensional space and finds the best boundary that separates the classes [26]. Finally, we used Random Forest as a comparison model to check stability. This model builds multiple decision trees using different subsets of the data and combines their results. This helps capture complex patterns while reducing the risk of overfitting [25]. XGBoost was selected in this study due to its capability to handle complex and overlapping data distributions through gradient boosting and regularization mechanisms. Compared to other ensemble methods such as Random Forest, XGBoost iteratively focuses on difficult samples and refines decision boundaries, making it more suitable for classification problems with low separability.

## 3.4. Performance Validation and Evaluation Metrics

When evaluating models on imbalanced data, accuracy alone can be misleading. It only measures overall correct predictions and often looks high because the model focuses on the majority class while ignoring the minority class [27], [28]. To address this, recall becomes important because it measures how well the model detects minority cases and reduces false negatives [29], [30]. However, focusing only on recall can lower precision. To balance both, we used F1-Score as the main evaluation metric. F1-Score combines precision and recall, so it gives a more balanced and fair measure of model performance, especially for imbalanced data.

# 4. Results and Discussion

## 4.1. Tree Depth Optimization

Before evaluating the final model performance, we first determined the best setup for the decision tree used in feature extraction. We tested different maximum tree depths to balance capturing complex patterns and avoiding overfitting. As shown in figure 5, performance improved as the tree became deeper and reached the best result at a depth of 6.

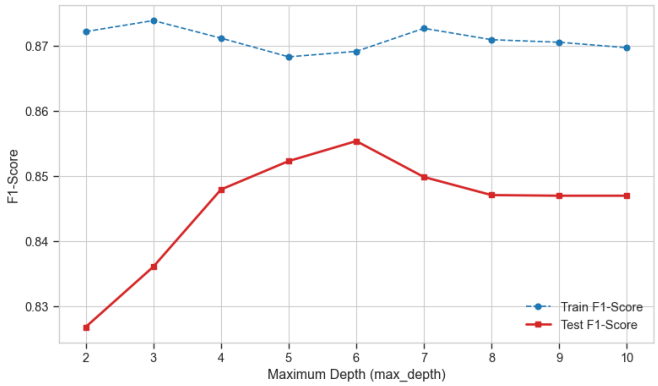


**Figure 5.** Optimal Tree Depth Evaluation for Feature Extraction

After this point, the performance stopped improving and began to decline, which indicates overfitting. This means the model started to learn noise from the training data instead of general patterns. Based on this result, we used a maximum depth of 6 as the final setting for extracting leaf features and building the feature space for all experiments.

## 4.2. Ablation Study Results

To evaluate the role of each component in our model, we conducted an ablation study. We compared three setups which is raw baseline data, unsupervised mapping using K-Means, and our proposed method that combines tree-based feature mapping with SMOTE-ENN. To verify the robustness of the Decision Tree + SMOTE-ENN framework, we conducted a Wilcoxon signed-rank test on the F1-score distributions across the 10-fold cross-validation. The results confirmed that the performance uplift over. As shown in table 3, models trained on raw data produced the lowest results, with XGBoost reaching only 0.7689. This shows that noise and class imbalance in the dataset reduce the ability of standard models to detect the minority class. In contrast, our proposed method showed a clear performance increase.

**Table 3.** Ablation Study Results

| Classification Model | Baseline | K-Means | Decision Tree | KMeans + SMOTE-ENN | Decision Tree + SMOTE-ENN |
|---|---|---|---|---|---|
| Random Forest | 0.7965 | 0.8163 | 0.7843 | 0.8363 | 0.8364 |
| Support Vector Machine | 0.7945 | 0.7825 | 0.8533 | 0.746 | 0.8329 |
| XGBoost | 0.7689 | 0.7842 | 0.8102 | 0.83 | 0.8554 |

By applying supervised leaf-based mapping before SMOTE-ENN, the XGBoost model reached a higher score of 0.8554. This improvement can also be seen in the Bar Chart in figure 6, where our method performs better across all metrics compared to the baseline. These results show that tree-based mapping is more effective than unsupervised methods in separating classes and helping the model make more accurate predictions.

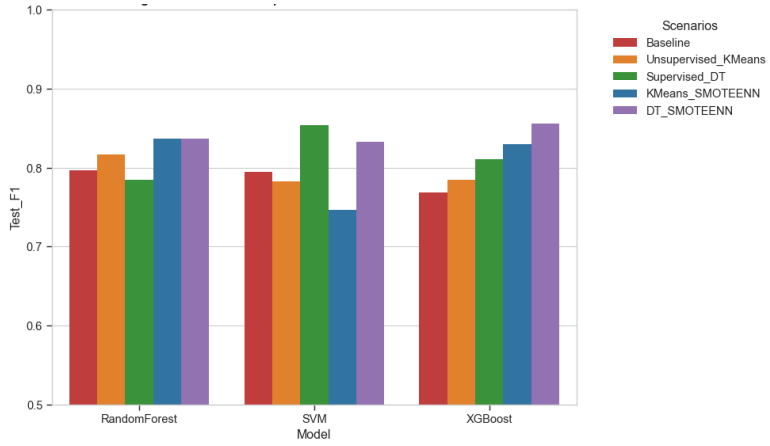


**Figure 6.** F1-Score Comparison Across Classification Models

## 4.3. Feature Space Mapping

To understand why our model improved, we visualized the data using Principal Component Analysis (PCA). As shown in figure 7, there is a clear difference between unsupervised and supervised mapping. In the K-Means approach, the data points from different classes still overlap a lot. This happens because K-Means groups data based only on distance and does not consider class labels, so it often fails to separate the minority class [31], [32]. In contrast, the supervised decision tree mapping produces more separated and compact minority clusters. This is because the process uses class labels to guide how the data is split. By reducing Gini impurity, the model gradually separates the classes and reduces overlap. As a result, the data becomes cleaner before applying SMOTE-ENN, However, it is important to note that PCA is an unsupervised linear dimensionality reduction method. While it effectively illustrates the overlap reduction conceptually in a 2D space (figure 7), it may not fully capture the complexity of the non-linear decision boundaries that exist in the higher-dimensional transformed feature space where the actual classification occurs which helps the model build more accurate decision boundaries. By following these decision paths, the model becomes more

ISSN 2723-6471


transparent and easier to understand. This approach avoids the black-box issue and shows that the predictions are based on clear and logical patterns, not random or noisy data

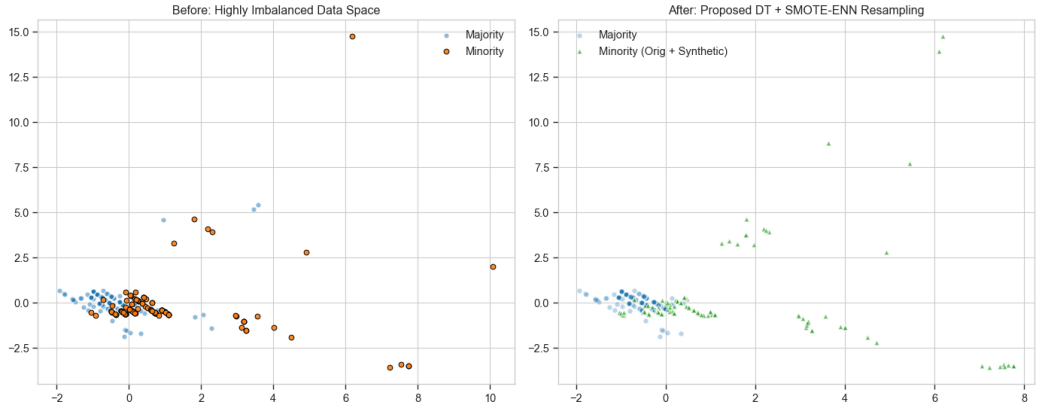


**Figure 7.** PCA Representation of Decision Boundaries

## 4.4. Target Class Isolation

Converting decision tree leaves into One-Hot Encoding (OHE) produced features that are independent from each other. This can be seen in the correlation matrix in figure 8, where most values are close to zero. This shows that there is no multicollinearity between the features. Reducing feature overlap helps simplify the dataset and improves computational efficiency . It also supports better performance in models like XGBoost, since low correlation between features can reduce prediction error [33].

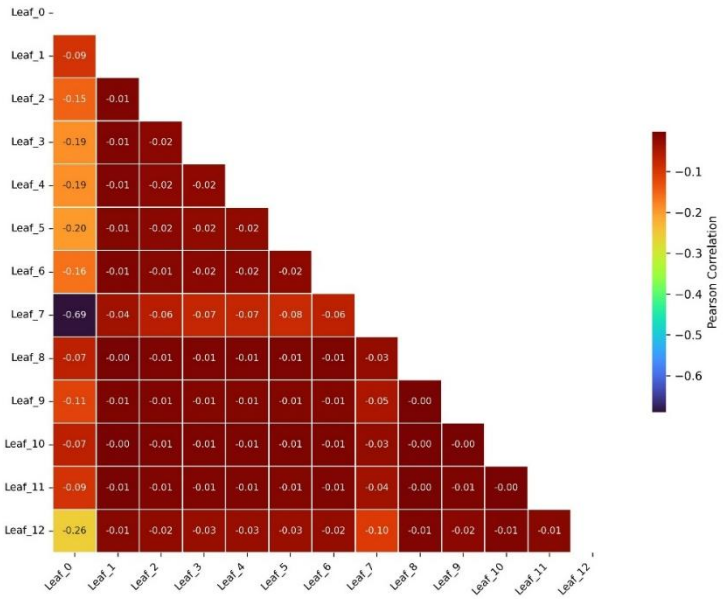


**Figure 8.** Correlation Matrix of Extracted Leaf Features

This method also helps separate the minority class more clearly. As shown in table 4, several leaf nodes at depth 6 contain a high number of minority samples. Some nodes, such as Leaf 21, 17, and 15, contain only minority data, which means they are fully pure. This happens because the decision tree uses Gini impurity to guide the splitting process and reduce class mixing [34].

**Table 4.** Top Minority Concentration Leaves

| Leaf ID | Total Samples | Minority Ratio |
|---|---|---|
| 21 | 8 | 1.0 |
| 17 | 34 | 1.0 |
| 15 | 8 | 1.0 |

## 4.5. Feature Importance

To understand how the XGBoost model makes its predictions, we looked at its feature importance scores. As established in the optimal tree setting, a maximum depth of 6 yields up to 64 total leaf features, which represent the entire feature space. However, as shown in figure 9, the model does not use all 64 features equally. As shown in figure 9, the model does not use all features equally. It relies heavily on certain leaf nodes, especially Leaf 0 and to a lesser extent Leaf 6. The strong influence of Leaf 0 shows that the decision tree was able to group a set of data points that are highly related to the target class. This means the model focuses on the most important patterns instead of being distracted by less useful features.

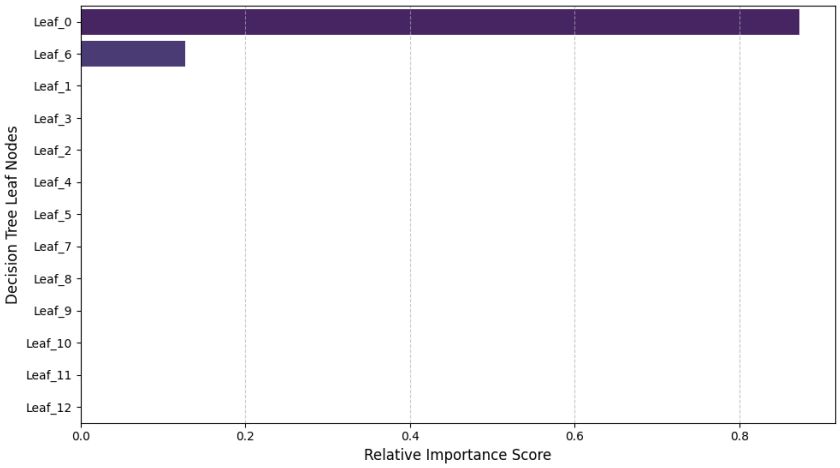


**Figure 9.** Feature Importance Scores from the XGBoost Model

However, features like "Leaf 0" are not easy to interpret on their own. To make them more meaningful, we traced each leaf index back to the original decision tree rules. For example, figure 10 details how specific leaves correspond to combinations of real and logical rules. A data point is assigned to Leaf 0 based on a concrete decision path, such as having a GPA below a specific threshold, combined with particular categories of study time and parental status figure 10 shows how these rules are formed. The results indicate that a data point is assigned to a leaf based on a combination of factors such as GPA, parental status, and study time, not just a single variable. By following these decision paths, the model becomes more transparent and easier to understand. This approach avoids the black-box issue and shows that the predictions are based on clear and logical patterns, not random or noisy data.

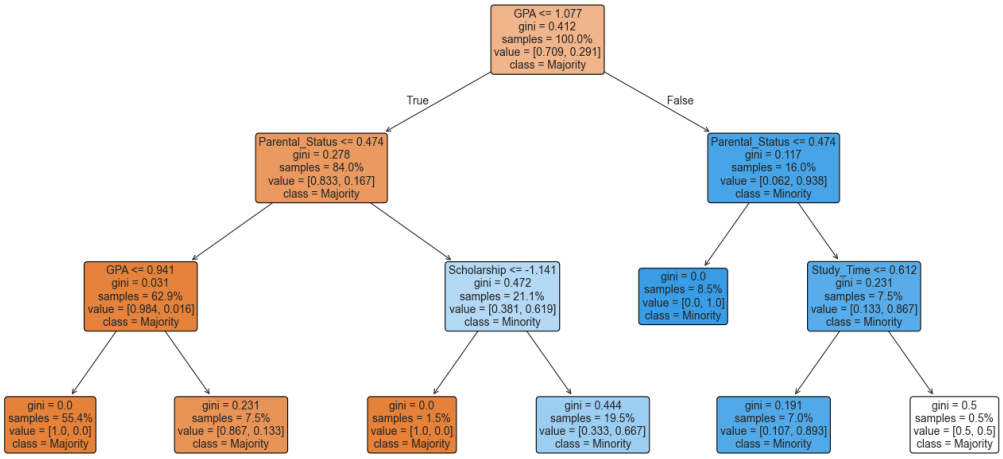


**Figure 10.** Conceptual Decision Tree

## 4.6. Discussion

The main goal of this study was to build a classification model that can handle imbalanced and overlapping data, using student career data as the case study. From the baseline results (Section 4.2), models like XGBoost and SVM showed

limited performance. This is in line with previous studies, which explain that many algorithms focus on the majority class and struggle to separate classes clearly in imbalanced datasets [35].

To solve this problem, we applied supervised feature mapping using decision tree leaf extraction before using SMOTE-ENN. The results from PCA (Section 4.3) and node analysis (Section 4.4) show that this method can separate overlapping data and isolate the minority class more effectively. As a result, the XGBoost model achieved a higher F1-Score of 0.8554. This improvement supports findings from Mastour et al. (2025), which show that combining feature extraction, resampling, and ensemble models can improve performance in complex classification tasks [36].

Another important contribution of this study is model transparency. Many advanced models are difficult to interpret, which makes it hard to understand how predictions are made. To address this, we analyzed feature importance (Section 4.5) and linked key features, such as Leaf 0, back to the original decision tree rules (Figure 10). This shows that the model makes predictions based on clear combinations of features, such as GPA, parental status, and study time, not random patterns.

## 5. Conclusion

This study focuses on a common problem in machine learning, which is handling data that is highly imbalanced and overlapping. To solve this, we proposed a hybrid classification approach that combines decision tree leaf extraction, SMOTE-ENN resampling, and XGBoost. This approach reshapes the data structure so that minority data can be separated more clearly before generating new synthetic samples.

The results show that the proposed method performs better than baseline models and unsupervised approaches, reaching an F1-Score of 0.8554. This approach also has several limitations. In terms of scalability, the combination of multiple stages such as tree extraction, resampling, and model training can increase computational complexity, especially on large datasets. Furthermore, the method's performance is highly dependent on parameter selection, particularly the depth of the decision tree, which can impact the quality of the data representation. This approach is also sensitive to dataset characteristics, such as the degree of overlap between classes and feature distribution. In addition to improving performance, this method also makes the model easier to understand. By tracing important features back to the original decision tree rules, we can clearly see how the model makes decisions. This means the model is not only accurate but also transparent. Overall, this study shows that it is possible to handle imbalanced data effectively while still keeping the model easy to interpret, making it useful for future data mining tasks.

## 6. Declarations

### 6.1. Author Contributions

Conceptualization: R.M., W., and A.A.C.; Methodology: R.M.; Software: R.M.; Validation: R.M., W., and A.A.C.; Formal Analysis: R.M., W., and A.A.C.; Investigation: R.M.; Resources: W.; Data Curation: W.; Writing Original Draft Preparation: R.M., W., and A.A.C.; Writing Review and Editing: W., R.M., and A.A.C.; Visualization: R.M.; All authors have read and agreed to the published version of the manuscript.

### 6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### 6.4. Institutional Review Board Statement

Not applicable.

### 6.5. Informed Consent Statement

Not applicable.

### 6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] A. A. Chamid, R. Nindyasari, and M. I. Ghozali, "Comparative analysis of machine learning algorithms for predicting patient admission in emergency departments using EHR data," *J. RESTI*, vol. 9, no. 2, pp. 185–194, Apr. 2025, doi: 10.29207/resti.v9i2.6188.

[2] R. Meimaharani, Widowati, and A. A. Chamid, "Comparison of machine learning methods for classifying graduates' waiting time to obtain employment," *IEEE Access*, vol. 2025, no. Dec., pp. 233–237, 2025, doi: 10.1109/ISRITI68345.2025.11393333.

[3] J. S. Siddhanth, S. Prabhu, S. Prabhu, and K. Mallibhat, "Factors influencing employability of engineering students and predictive modelling with machine learning algorithms," *IEEE Access*, vol. 2025, no. Dec., pp. 22–32, 2025, doi: 10.1109/CONIT65521.2025.11166928.

[4] C. N. P. Olipas, "Predictive modeling and explainability of student employability in the Philippines using random forest and Shapley additive explanations," *Interdiscip. J. Inf. Knowl. Manage*., vol. 21, no. 2, pp. 180–195, Feb. 2026, doi: 10.28945/5690.

[5] R. Suguna, J. S. Prakash, H. A. Pai, T. R. Mahesh, V. V. Kumar, and T. E. Yimer, "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE," *Sci. Rep*., vol. 15, no. 1, pp. 1–21, Jan. 2025, doi: 10.1038/s41598-025-01031-0.

[6] S. A. Bachhal and S. G. P., "Educational data mining: A review," *J. Phys. Conf. Ser*., vol. 1950, no. 1, pp. 8–16, Jun. 2021, doi: 10.1088/1742-6596/1950/1/012022.

[7] A. A. Chamid, Widowati, and R. Kusumaningrum, "Labeling consistency test of multi-label data for aspect and sentiment classification using the Cohen kappa method," *Ing. Syst. Inf.*, vol. 29, no. 1, pp. 161–167, Feb. 2024, doi: 10.18280/isi.290118.

[8] H. Hairani, T. Widiyaningtyas, D. D. Prasetya, and A. Aminuddin, "Addressing imbalance in health datasets: A new method NR-clustering SMOTE and distance metric modification," Comput. Mater. Continua, vol. 82, no. 2, pp. 2931–2949, Feb. 2025, doi: 10.32604/cmc.2024.060837.

[9] A. A. Chamid, Widowati, and R. Kusumaningrum, "Multi-label text classification on Indonesian user reviews using semi-supervised graph neural networks," *ICIC Express Lett*., vol. 17, no. 10, pp. 1075–1084, Oct. 2023, doi: 10.24507/icicel.17.10.1075.

[10] A. Choudhury, A. Mondal, and S. Sarkar, "Searches for the BSM scenarios at the LHC using decision tree-based machine learning algorithms: A comparative study and review of random forest, AdaBoost, XGBoost and LightGBM frameworks," *Eur. Phys. J. Spec. Top*., vol. 233, no. 15–16, pp. 2425–2463, Nov. 2024, doi: 10.1140/epjs/s11734-024-01308-x.

[11] A. Jazuli, Widowati, A. A. Chamid, and R. Kusumaningrum, "Transformer-based semantic indexing for aspect-based sentiment analysis using an enhanced index generation algorithm with BERT," *Int. J. Adv. Technol. Eng. Explor*., vol. 12, no. 127, pp. 907–926, Jun. 2025, doi: 10.19101/IJATEE.2024.111102114.

[12] R. Suguna, J. S. Prakash, H. A. Pai, T. R. Mahesh, V. V. Kumar, and T. E. Yimer, "Mitigating class imbalance in churn prediction with ensemble methods and SMOTE," *Sci. Rep*., vol. 15, no. 1, pp. 87–99, Dec. 2025, doi: 10.1038/s41598-025-01031-0.

[13] D. Hooshyar et al., "Towards responsible AI for education: Hybrid human-AI to confront the elephant in the room," *Comput. Educ. Artif. Intell.*, vol. 9, no. Dec., pp. 102–111, Dec. 2025, doi: 10.1016/j.caeai.2025.100524.

[14] A. A. Chamid, R. R. Isnanto, A. Jazuli, W. H. Sugiharto, A. D. Widiantoro, and Sumaji, "Deep learning approach for multi-class classification of textual data via Bi-LSTM," *IEEE Access*, vol. 2025, no. Dec., pp. 74–79, 2025, doi: 10.1109/ITIS67966.2025.11308963.

[15] X. Li and T. Yang, "Forecast of the employment situation of college graduates based on the LSTM neural network," *Comput. Intell. Neurosci*., vol. 2021, no. Jan., pp. 44–51, 2021, doi: 10.1155/2021/5787355.

[16] M. C. Abdipatra and R. Yunanda, "Comparative evaluation and deployment of machine learning models for predicting student academic performance," *IEEE Access*, vol. 2025, no. Dec., pp. 309–314, 2025, doi: 10.1109/ITIS67966.2025.11309211.

[17] Y. Zhou, J. Wu, X. Xu, G. Shi, P. Liu, and L. Jiang, "Investigating perioperative pressure injuries and factors influencing them with imbalanced samples using a synthetic minority oversampling technique," *Biosci. Trends*, vol. 19, no. 2, pp. 173–188, Feb. 2025, doi: 10.5582/bst.2025.01013.

 ISSN 2723-6471


[18] N. I. Fardana, R. R. Isnanto, and O. D. Nurhayati, "Handling class imbalance in health datasets: A comparative study of SMOTE and SMOTEENN with TabNet," *IEEE Access,* vol. 2026, no. Jan., pp. 305–310, 2026, doi: 10.1109/ICICOS68590.2025.11329876.

[19] M. A. Aslam, F. Murtaza, M. E. Ul Haq, A. Yasin, and N. Ali, "SAPEx-D: A comprehensive dataset for predictive analytics in personalized education using machine learning," *Data,* vol. 10, no. 3, pp. 4–15, Mar. 2025, doi: 10.3390/data10030027.

[20] A. S. Alghamdi and A. Rahman, "Data mining approach to predict success of secondary school students: A Saudi Arabian case study," *Educ. Sci*., vol. 13, no. 3, pp. 34–41, Mar. 2023, doi: 10.3390/educsci13030293.

[21] E. Laber and L. Murtinho, "Minimization of Gini impurity: NP-completeness and approximation algorithm via connections with the k-means problem," *Electron. Notes Theor. Comput. Sci*., vol. 2019, no. Aug., pp. 567–576, 2019, doi: 10.1016/j.entcs.2019.08.050.

[22] W. Lv and X. Zhu, "Academic performance prediction based on hybrid GA-XGBoost-FM learning," *IEEE Access*, vol. 2025, no. May, pp. 622–627, 2025, doi: 10.1109/ISBDAS64762.2025.11116821.

[23] Y. Chachoui, N. Azizi, R. Hotte, and T. Bensebaa, "Enhancing algorithmic assessment in education: Equi-fused-data-based SMOTE for balanced learning," *Comput. Educ. Artif. Intell.*, vol. 6, no. Jun., pp. 5–15, Jun. 2024, doi: 10.1016/j.caeai.2024.100222.

[24] H. V. Pham et al., "Comprehensive evaluation of bankruptcy prediction in Taiwanese firms using multiple machine learning models," *Int. J. Technol.*, vol. 16, no. 1, pp. 289–309, Jan. 2025, doi: 10.14716/ijtech.v16i1.7227.

[25] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[26] M. Mujahid et al., "Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, pp. 22–34, Dec. 2024, doi: 10.1186/s40537-024-00943-4.

[27] M. Owusu-Adjei, J. B. Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLoS Digit. Health*, vol. 2, no. 11, pp. 80–92, Nov. 2023, doi: 10.1371/journal.pdig.0000290.

[28] S. Sadeghi, D. Khalili, A. Ramezankhani, M. A. Mansournia, and M. Parsaeian, "Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods," *BMC Med. Inform. Decis. Mak*., vol. 22, no. 1, pp. 201–220, Dec. 2022, doi: 10.1186/s12911-022-01775-z.

[29] S. Ashraf, S. Saleem, T. Ahmed, Z. Aslam, and D. Muhammad, "Conversion of adverse data corpus to shrewd output using sampling metrics," *Vis. Comput. Ind. Biomed. Art,* vol. 3, no. 1, pp. 77–88, Dec. 2020, doi: 10.1186/s42492-020-00055-9.

[30] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative analysis using various performance metrics in imbalanced data for multi-class text classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 44–52, Jun. 2023, doi: 10.14569/IJACSA.2023.01406116.

[31] J. Hemmatian, R. Hajizadeh, and F. Nazari, "Addressing imbalanced data classification with cluster-based reduced noise SMOTE," *PLoS One*, vol. 20, no. 2, pp. 2–17, Feb. 2025, doi: 10.1371/journal.pone.0317396.

[32] Y. Zhang, L. Deng, and B. Wei, "Imbalanced data classification based on improved random-SMOTE and feature standard deviation," *Mathematics*, vol. 12, no. 11, pp. 1–15, Jun. 2024, doi: 10.3390/math12111709.

[33] S. K. Safi and S. Gul, "An enhanced tree ensemble for classification in the presence of extreme class imbalance," *Mathematics*, vol. 12, no. 20, pp. 90–112, Oct. 2024, doi: 10.3390/math12203243.

[34] R. Goswami, A. Garai, P. Sadhukhan, P. Ghosh, and T. Chakraborty, "Shape penalized decision forests for imbalanced data classification," *IEEE Access*, vol. 13, no. Jan., pp. 86380–86395, 2025, doi: 10.1109/ACCESS.2025.3569523.

[35] A. Almalawi, B. Soh, A. Li, and H. Samra, "Predictive models for educational purposes: A systematic review," *Big Data Cogn. Comput.*, vol. 8, no. 12, pp. 38–49, Dec. 2024, doi: 10.3390/bdcc8120187.

[36] H. Mastour, T. Dehghani, E. Moradi, and S. Eslami, "Explainable artificial intelligence for predicting medical students' performance in comprehensive assessments," *Sci. Rep*., vol. 15, no. 1, pp. 77–85, Dec. 2025, doi: 10.1038/s41598-025-07460-1.