International Journal of Basic and Applied Sciences, 14 (4) (2025) 602-608

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/70586478
Research paper

Evaluating The Impact of Dimensionality Reduction Techniques for Decision Tree Performance in Multiclass Imbalanced Datasets

S.Sridhar [1] *, Sridevi Srinivasan [2]

[1] *Assistant Professor, Department of Computer Science and Engineering (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani campus, Chennai,TN, India.*
[2] *Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, TN, India.*
**Corresponding author E-mail: sridhars2@srmist.edu.in*

Received: June 18, 2025, Accepted: July 7, 2025, Published: August 21, 2025

Abstract

Imbalanced datasets are a common challenge in real-world applications, where the class of interest is often a minority. Addressing class imbalance in multi-class datasets receives less attention compared to binary datasets due to the increased complexity. This complexity arises from varying class frequencies and associated costs. High-dimensional datasets, with numerous features, pose another challenge in machine learning. Feature selection techniques help mitigate dimensionality, improving classifier efficiency in accuracy and computation. These techniques involve creating new features or selecting subsets from the original set. Effective strategies for imbalanced learning aim to retain minority class concepts by leveraging informative features. This study investigates the impact of dimensionality reduction techniques, such as Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), on multi-class imbalanced datasets using decision trees, a challenge commonly encountered in high-dimensional domains such as bioinformatics and medical diagnostics. While datasets with clear class boundaries may reduce the effectiveness of dimensionality reduction, PCA could be more effective in cases of class overlap, where the majority class has more samples. Experimental results support these conclusions.

Keywords: *Classification; Feature selection; LDA; Multi-class imbalanced datasets; PCA*

1. Introduction

In imbalanced learning scenarios, high-dimensional data is frequently encountered. A significant feature of such data is that the number of variables is much larger than the number of available samples. Traditional classifiers are usually trained on datasets with non-uniform numbers of samples per class. However, researchers have delved into the potential challenges posed by high dimensionality in achieving balanced predictions across different classes [10].

In the area of high-dimensional data classification, a key requirement is the need for dimensionality reduction, achieved through either variable selection or variable extraction. Feature selection involves investigating a subset of features for training the classification model, which can occur either before the classifier's development or within the classification method itself [3]. On the other hand, feature extraction involves generating a condensed set of synthetic attributes derived from the most informative aspects of the original features. This new set replaces the original features for conducting learning tasks. Research indicates that utilizing classifiers directly on imbalanced high-dimensional data can lead to significantly biased classification outcomes toward the majority class [1]. The level of bias varies depending on factors such as the particular classification method used, the extent of class imbalance, and the degree of difference between classes [15]. Moreover, this bias tends to increase further when standard feature selection methods are applied. Unsupervised feature reduction techniques can reveal a way of representing the data that more clearly separates items from different classes [9].

Many studies have concentrated on learning from imbalanced datasets using standard supervised learning techniques, such as Support Vector Machines (SVMs), decision trees, the nearest neighbor rule, and others [2]. However, none of these studies have investigated the impact of dimensionality reduction techniques on improving the classification of multiclass imbalanced datasets. Feature extraction reduces the computational resources needed to analyze large datasets by reducing the original feature space and thereby reducing data dimensionality [11]. Feature extraction is well-known for its ability to greatly reduce the dimensionality of the feature space compared to feature selection. When dealing with complex data, a significant challenge stems from the abundance of features, which frequently requires substantial memory and computational resources. Consequently, classification algorithms may overfit to training samples, leading to poor performance on new samples [14]. To mitigate these issues, feature extraction involves employing various methods to generate

combinations of variables that accurately capture the information. Within the field of data science, effectively optimizing feature extraction is widely regarded as essential for developing efficient models.
The paper's structure is as follows: Section 2 provides background information on Dimensionality reduction techniques, specifically principal component analysis (PCA) and Linear Discriminant Analysis (LDA). Section 3 presents the specifics of the datasets utilized in the experiments, along with an analysis of the results. These results are further discussed in Section 4. Finally, section 5 includes the references for this work.

2. Materials and Methods

2.1 Principal Component Analysis (PCA)

PCA offers a significant advantage by simplifying complex datasets into lower-dimensional spaces while retaining their overall structure intact. Its widespread applicability stems from its straightforwardness and non-parametric nature. Principal Component Analysis (PCA) finds separate groups of the original variables and combines them using linear combinations to form new components that represent the data [12]. According to Jolliffe (2002), PCA operates under several key assumptions. Firstly, it assumes that the mean and variance are adequate statistics for representing the probability distributions of the entire dataset. Secondly, PCA presumes linearity within the dataset, which is characterized by a combination of specific basis vectors. Lastly, PCA acknowledges that directions with maximum variance typically signify important dynamics, while directions with lower variance are often indicative of noise. Leveraging these principal directions can potentially enhance the accuracy of the classification process [5].
The process begins with the initial dataset, which undergoes dimensionality reduction to produce data with fewer dimensions. This involves steps such as standardizing the data, computing eigenvectors and eigenvalues, and decomposing the covariance matrix to understand its structure better. Following this, principal components are selected based on their significance. A projection matrix is then constructed to facilitate the transformation of the data onto a new feature space. Finally, the transformed data is obtained, resulting in a reduced-dimensional representation of the original dataset. PCA is a method that simplifies data by reducing its dimensions, but it might miss important details if the data has complex patterns in higher dimensions [6].

2.2 Linear Discriminant Analysis (LDA)

LDA is a method utilized for reducing dimensionality. It works by identifying a linear combination of features that effectively distinguishes between multiple classes. LDA is specifically designed for scenarios where there are continuous independent variables and a categorical dependent variable. The primary goal of LDA is to condense a larger dataset into a lower-dimensional space while retaining crucial discriminant features. This technique was formulated by Ronald A. Fisher, who demonstrated its effectiveness as a classifier. The original concept of Linear Discriminant Analysis (LDA) was initially outlined for a two-class problem. Later on, in 1948, C. R. Rao extended this concept, naming it "multi-class Linear Discriminant Analysis" or "Multiple Discriminant Analysis" [13].
Initially, compute the mean vectors for each class in d dimensions within the dataset. Following this, compute the scatter matrices, including both the between-class and within-class scatter matrices. Next, determine the eigenvectors (eigenvector_1, eigenvector_2, ..., eigenvector_d) and their corresponding eigenvalues (eigenvalue_1, eigenvalue_2, ..., eigenvalue_d) for these scatter matrices. Then, arrange the eigenvectors in descending order based on their eigenvalues and select k eigenvectors with the largest eigenvalues to build a d×k-dimensional matrix W, where each column represents an eigenvector. Lastly, utilize this d×k eigenvector matrix to transform the samples into the new subspace. This transformation can be concisely depicted by the matrix multiplication: transformed_samples = original_samples × W, where original_samples denotes the n×d dimensional matrix containing n samples, and transformed_samples represents the transformed n×k dimensional samples in the newly established subspace. A variant of LDA, namely LADA, attempts to find the optimal way to transform the data so that similar points are moved closer together and dissimilar points are pushed farther apart, without making any assumptions about the data's distribution [7]. Dimensionality reduction typically serves not only to reduce computational costs in a given classification task but also to lessen the overfitting issue by minimizing errors in parameter estimation

2.3 Decision Tree Model

Decision trees are a fundamental tool in the field of data mining, with C4.5 being one of the most widely used algorithms. However, for binary class datasets characterized by class imbalance, Hellinger distance decision trees (HDDTs) have emerged as an alternative approach. HDDTs are particularly suited for handling class imbalance in binary class datasets. They offer a distinct method for constructing decision trees, which considers the imbalance between classes. Provost and Domingos proposed a modification to C4.5, termed C4.4, which incorporates Laplace smoothing at the leaves of unpruned and uncollapsed decision trees. Empirical evidence suggests that this approach outperforms other configurations, leading to its adoption in various experiments. An essential aspect of building decision trees is the choice of splitting criterion, which dictates how the data should be divided to maximize performance. In C4.4, the splitting criterion is the gain ratio, which assesses the purity of splits based on entropy. Conversely, HDDTs utilize Hellinger distance as the splitting criterion, offering an alternative measure for assessing split quality [4].

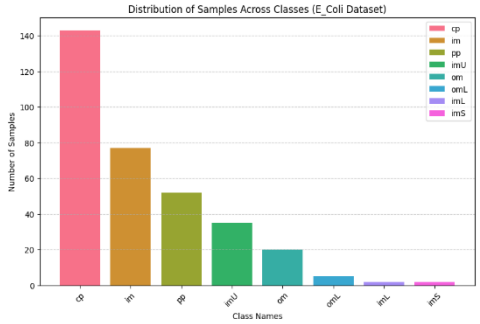
3. Results and Discussion

3.1 Datasets Used

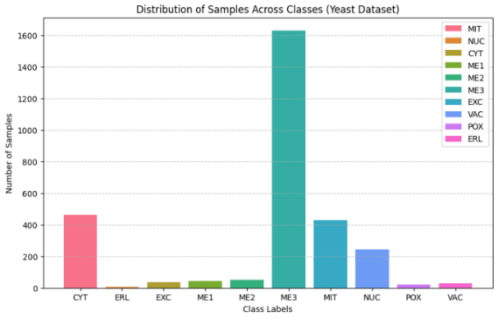
The paper investigates the effects of various data preprocessing techniques on multiclass imbalanced datasets. Five such datasets were selected from the UCI repository, namely E-coli, Yeast, Wine Quality, Hypo_Thyroid, and CTG

Table 1: Multiclass Imbalanced Dataset Summary with Class Distribution Details

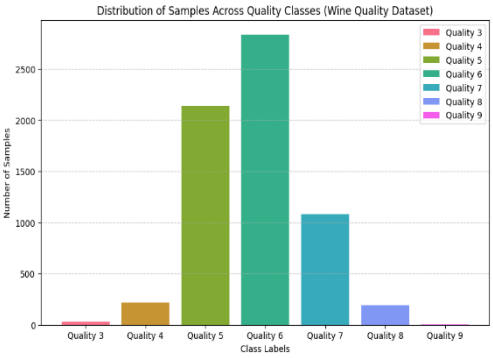
Dataset	Number of Samples	Number of Features	Number of Classes	# Majority Classes	# Minority Classes
Wine Quality	6,497	12	7	3	4
Hypo Thyroid	2,800	27	4	1	3
Yeast	1,484	8	10	3	7
CTG	2,126	33	3	1	2
E. coli	336	7	8	2	6



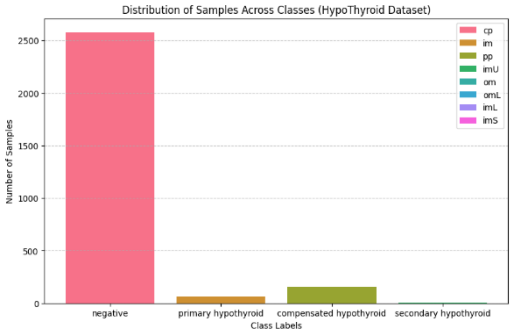
E-coli Dataset:



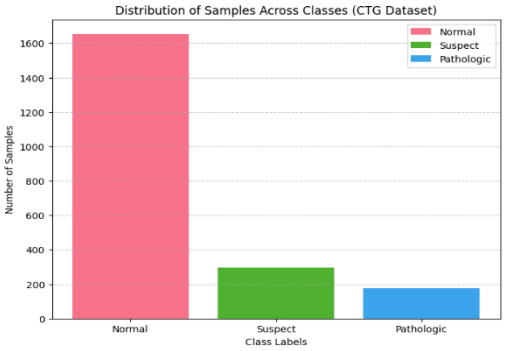
Yeast Dataset:



Wine Quality Dataset:



Hypothyroid Dataset:



CTG Dataset:

Fig. 1: Datasets used in the assessment.

These datasets were chosen specifically for their multiclass imbalanced nature, and the paper aims to assess how dimensionality reduction techniques affect the performance of machine learning models trained on such datasets. The dataset covers both multi-majority and multi-minority cases, which describe the distribution of classes based on their frequencies.
Multi-Majority refers to the classes in the dataset that have a higher number of samples compared to other classes. In datasets with multiple majority classes, each majority class holds a significant proportion of the total samples relative to the dataset size. Multi-Minority refers to the classes in the dataset that have a lower number of samples compared to other classes. In datasets with multiple minority classes, each minority class contains a relatively smaller number of samples compared to the majority classes

3.2 Model Used

A Decision Tree Classifier is used with "gini" as the criterion, the "best" splitter strategy, and no maximum depth limit. The algorithm will recursively split nodes based on the Gini impurity until each leaf node contains fewer samples than the minimum required for splitting. The dataset is divided into training and testing sets, with proportions of 75% and 25%, respectively.

3.3 Measures Used

- **Accuracy and the confusion matrix** are fundamental metrics used to evaluate classification models. Accuracy indicates the proportion of correctly classified instances out of the total instances in the dataset, offering an overall assessment of the model's perfor-

mance across all classes. On the other hand, the confusion matrix provides a detailed breakdown of the model's performance. It displays the number of correct and incorrect predictions for each class, including true positives, true negatives, false positives, and false negatives. By examining accuracy and the confusion matrix together, we gain valuable insights into the model's overall effectiveness and its performance on individual classes [8].

- The **F1-score**, also referred to as the F-measure, serves as a widely used metric for assessing classification model performance. It offers a balanced evaluation by combining precision and recall. The F1-score is computed as the harmonic mean of precision and recall, expressed by the formula:

$$F1 = 2 \times precision \times recall / precision + recall \quad (1)$$

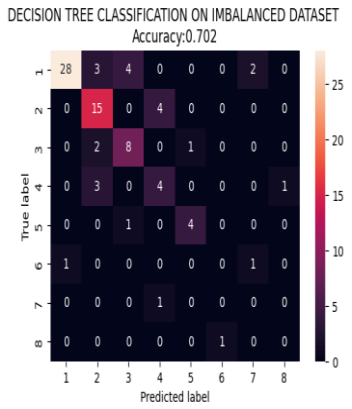
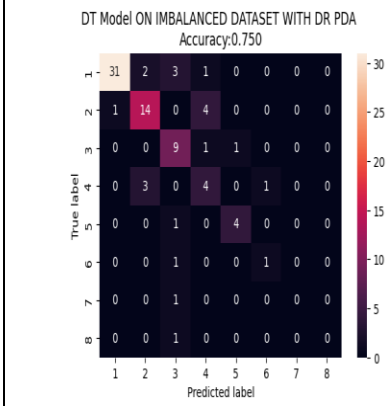
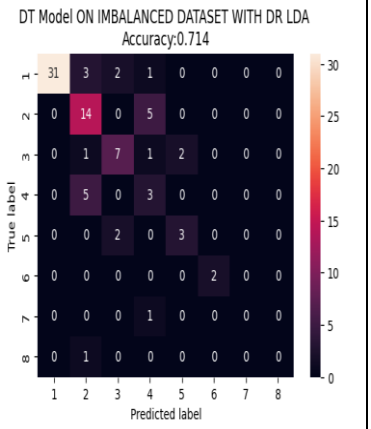
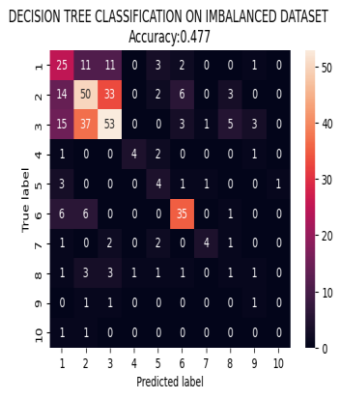
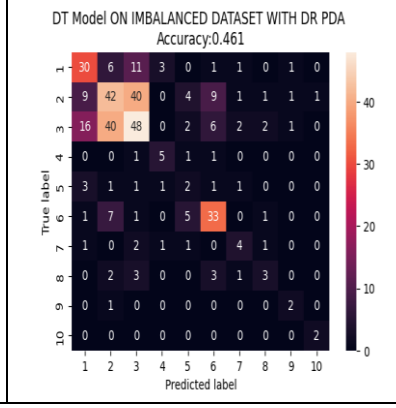
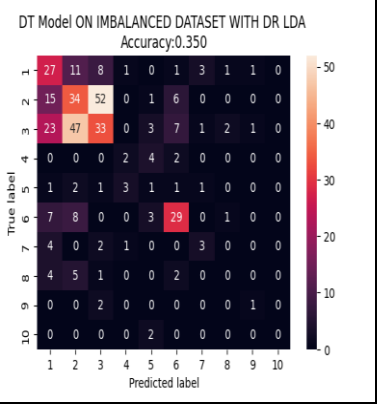
- The **G-Mean (Geometric Mean)** is a metric commonly utilized in classification tasks, particularly when handling imbalanced datasets. It offers valuable insights into a model's performance across various classes, especially in the presence of significant class imbalances. To compute the G-Mean, the square root of the product of class-wise sensitivities (true positive rate or recall) is taken. This calculation is expressed by the formula:

$$\text{G-Mean} = \text{Sqrt}\left(\text{Sensitivity}_1 \times \text{Sensitivity}_2 \times \ldots \times \text{Sensitivity}_n\right) \quad (2)$$

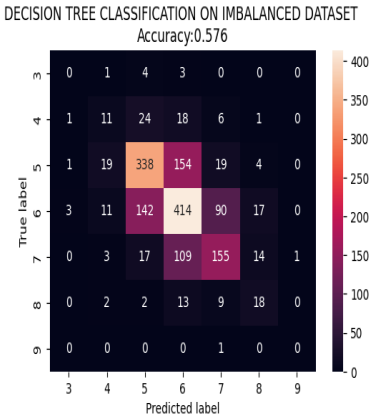
3.4 Experimental Results and Analysis

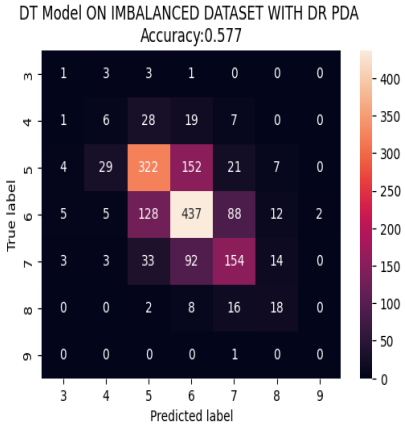
Table 2 below presents the results of the Decision Tree models' accuracy and confusion matrix with and without dimensionality reduction methods, namely PCA and LDA. Additionally, their classification performance across various classes is illustrated in Figure.
PCA demonstrates a significant positive impact on enhancing the overall accuracy of multiclass imbalanced data classification compared to LDA. This effect is particularly pronounced in datasets featuring both multi-majority and multi-minority classes. However, both LDA and PCA exhibit a detrimental effect on classification performance in datasets characterized by a single majority and multiple minority classes.

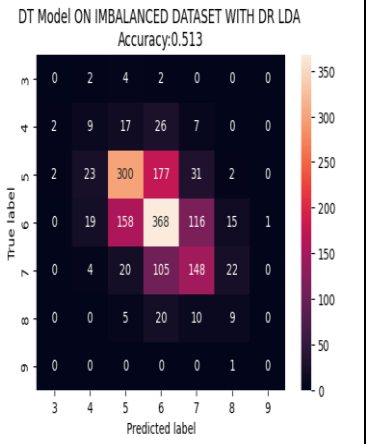
Table 2: Accuracy and Confusion Matrix of DT model with and without Dimensionality Reduction Techniques

E-coli Dataset:		
		
Yeast Dataset:		
		

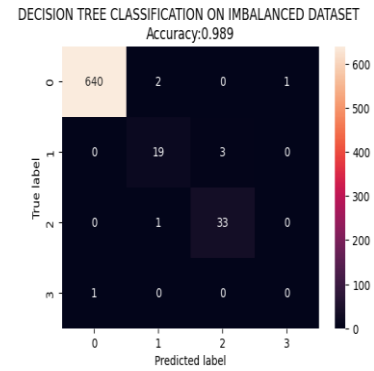
Wine Quality Dataset:

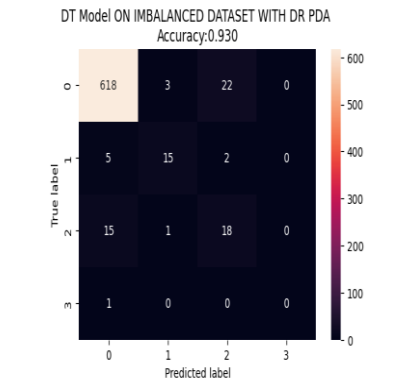


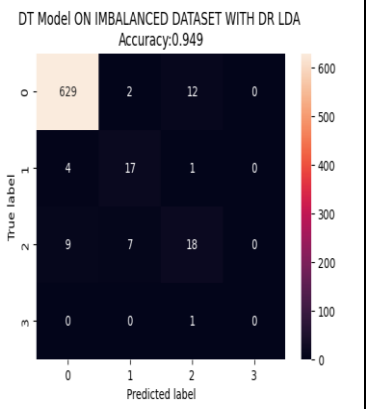




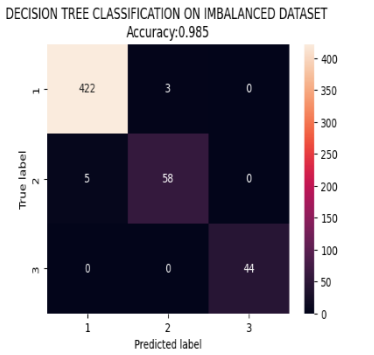
Hypothyroid Dataset:

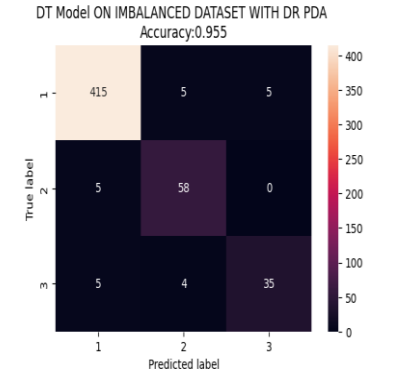


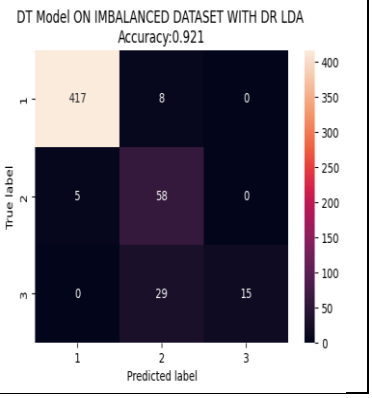


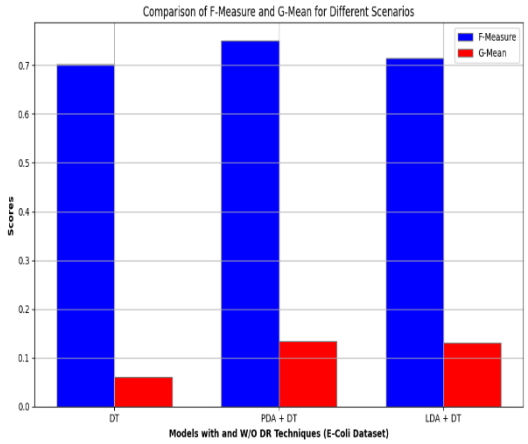


CTG Dataset:

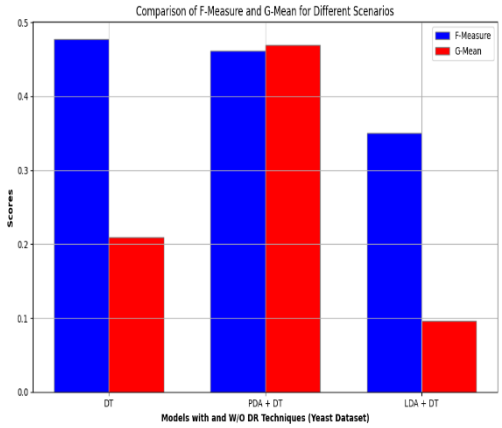




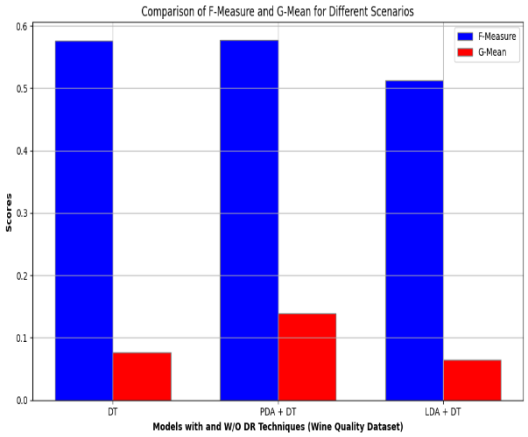




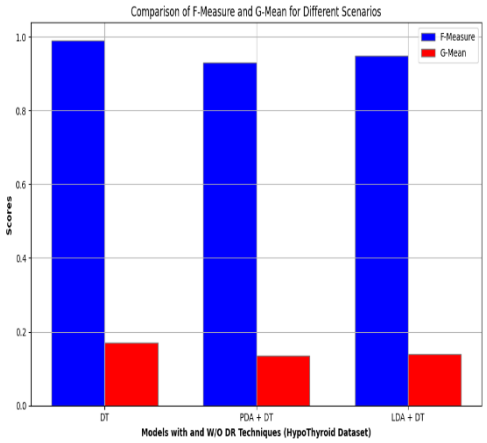
E-coli Dataset



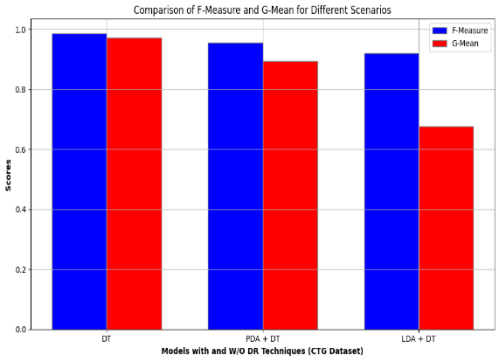
Yeast Dataset



Wine Quality Dataset



Hypothyroid Dataset



CTG Dataset:

Fig. 2: Models' performance analysis with and without dimensionality reduction techniques

3.4.1 Figure captions

The following table 3 presents a comparative analysis of Decision Tree (DT) models across five datasets under three scenarios: baseline DT without dimensionality reduction, DT with PDA, and DT with LDA.G-Mean is an important metric to assess how fairly a model performs across different classes, especially when the class distribution is not balanced. PDA helps improve both class-wise balance (measured by G-Mean) and macro F1-score in most datasets, showing its effectiveness in handling imbalanced data. In contrast, LDA often performs worse, especially on datasets with high class imbalance like Yeast. For datasets that already perform well, such as CTG and Hypothyroid, using dimensionality reduction techniques like PDA or LDA gives little to no improvement and can sometimes even lower the performance.

Table 3: Performance comparison of the DT model with and without dimensionality reduction techniques

Dataset	Model	Accuracy	F1-Score	G-Mean
E. coli	DT	0.702	0.72	0.08
	PDA + DT	**0.75**	**0.745**	**0.13**
	LDA + DT	0.714	0.725	0.11
Yeast	DT	0.477	0.48	0.21
	PDA + DT	0.461	0.45	**0.45**
	LDA + DT	0.35	0.35	0.1
Wine Quality	DT	0.576	0.57	0.1
	PDA + DT	0.577	**0.575**	**0.13**
	LDA + DT	0.513	0.51	0.09
Hypothyroid	DT	**0.989**	**0.98**	0.12
	PDA + DT	0.93	0.93	0.11
	LDA + DT	0.949	0.94	0.1
CTG	DT	**0.985**	**0.97**	**0.94**
	PDA + DT	0.955	0.95	0.9
	LDA + DT	0.921	0.92	0.75

4. Conclusion

Our analysis shows that the Decision Tree model faces challenges when dealing with imbalanced datasets containing multiple classes. It tends to favor the majority class, often misclassifying the minority classes. However, applying dimensionality reduction techniques like PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) can improve the model's performance. Specifically, using PCA with the Decision Tree model led to better results than using the model alone or with LDA. This combination helped the model better identify patterns in the data. PCA, being an unsupervised technique, selects components that capture the highest variance in the data, without considering class labels, and is particularly effective in datasets with overlapping class distributions. In contrast, LDA is a supervised method that seeks to maximize class separability, and its effectiveness declines when the classes are not linearly separable, as observed in datasets like Yeast. On the other hand, in situations where there is one dominant class and several smaller ones, applying PCA and LDA appears to reduce the model's accuracy due to their neglect of low-variance components and the generalization of the dominant class. In the future, PCA and LDA can be explored in combination with other classifier models on a larger number of multiclass imbalanced datasets to evaluate broader applicability. Additionally, non-linear dimensionality reduction techniques such as t-SNE and UMAP should be investigated, particularly for handling multiclass imbalanced datasets.

References

[1] R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data", BMC Bioinformatics, vol. 11, no. 1, (2010), p. 523.

[2] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, Learning from Imbalanced Data Sets, Springer, (2018), ISBN: 978-3-319-98073-7.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", J. Mach. Learn. Res., vol. 3, (2003), pp. 1157–1182.

[4] T. R. Hoens, Q. Qian, N. V. Chawla, and Z. H. Zhou, "Building decision trees for the multi-class imbalance problem", in Advances in Knowledge Discovery and Data Mining, PAKDD 2012, Springer, vol. 7301, (2012), pp. 122–133.

[5] I. T. Jolliffe, Principal Component Analysis, 2nd ed., Springer, New York, (2002).

[6] A. D. Kalian, E. Benfenati, O. J. Osborne, J. L. C. M. Dorne, D. Gott, C. P. Potter, M. Guo, and C. Hogstrand, "Improving accuracy scores of neural networks driven QSAR models of mutagenicity", in Proc. 33rd Eur. Symp. on Computer Aided Process Engineering (ESCAPE-33), Elsevier, Athens, Greece, (2023), p. 846.

[7] X. Li, Q. Wang, F. Nie, and M. Chen, "Locality Adaptive Discriminant Analysis Framework", IEEE Trans. Cybern., vol. 52, (2022), pp. 7291–7302.

[8] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix", Pattern Recognition, vol. 91, (2019), pp. 216–231.

[9] S. Lusito, A. Pugnana, and R. Guidotti, "Solving imbalanced learning with outlier detection and features reduction", Mach. Learn., vol. 113, (2024), pp. 5273–5330.

[10] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance", Neural Netw., vol. 21, no. 2–3, (2008), pp. 427–436.

[11] R. E. Nogales and M. E. Benalcázar, "Analysis and evaluation of feature selection and feature extraction methods", Int. J. Comput. Intell. Syst., vol. 16, (2023), p. 153.

[12] T. Parhizkar, E. Rafieipour, and A. Parhizkar, "Evaluation and improvement of energy consumption prediction models using principal component analysis-based feature reduction", J. Clean. Prod., vol. 279, (2021), p. 123866.

[13] C. R. Rao, "The utilization of multiple measurements in problems of biological classification", J. R. Stat. Soc. Ser. B (Methodol.), vol. 10, (1948), pp. 159–193.

[14] S. Wang and X. Yao, "Multiclass imbalance problems: analysis and potential solutions", IEEE Trans. Syst. Man Cybern. B Cybern., vol. 42, no. 4, (2012), pp. 1119–1130.

[15] J. H. Xue and D. M. Titterington, "Do unbalanced data have a negative effect on LDA?", Pattern Recognition, vol. 41, no. 5, (2008), pp. 1558–1571.