International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research

COMPARING SCALING METHODS AND DIMENSIONALITY REDUCTION TECHNIQUES: THEIR INFLUENCE ON CLASSIFICATION ALGORITHMS

Anu Priya J[*,†] M.Sc. Information Technology
Priti Mondal, M.Sc. Applied Mathematics
*Correspondence to: Anu Priya J.
Author E-mail IDs: anumap2497@gmail.com, pritihaldar2014@gmail.com
[†]Anu Priya J is considered as the first author and corresponding author.

Abstract
Machine learning in predictive analysis can be highly sensitive to its preprocessing methods, especially with scaling and dimensionality reduction. This study aims to create a unified and reproducible framework for supervised classification pipelines under varying preprocessing configurations. The framework is designed to integrate multiple classifiers with scaling methods and principal component analysis (PCA), enabling structured experimentation across different classification datasets. The models are evaluated using accuracy, precision, recall and F1-score, combined into a composite performance metric for comparison. The proposed approach is believed to produce a reproducible and extensible methodology for benchmarking classification pipelines in practical machine learning applications.

Keywords
Composite Performance Metric, HistGradientBoosting, Logistic Regression, Naive Bayes, Predictive Analytics, Principal Component Analysis (PCA), Random Forest, Scaling Methods, Supervised earning, Support Vector Machine (SVM)

1. Introduction

Machine learning has become a keystone for predictive analysis, with multiple compounded applications from healthcare diagnosis prediction to finance forecasting. The major influence on the predictive results is based on the data preprocessing techniques used which vary from dataset to dataset and highly influence the predictive accuracy.

There are numerous studies that have evaluated individual preprocessing methods with classifiers. While they analyse the methods independently, few frameworks systematically analyse scaling techniques and dimensionality reduction methods and how it interacts across multiple classifiers. This scarcity of unified approach bounds reproducibility and comparability of the results.

This study aims to cultivate a framework that is reproducible and integrates multiple classification algorithms with various scaling and PCA configurations. This would enable structured evaluation of their combined impact on model performance. We implement pipelines combining variables of standard scaling, min-max scaling and PCA over five popular classifiers including Logistic Regression, Random Forest, HistGradientBoosting, SVM and Naïve Bayes. The evaluation of the models is arbitrated based on accuracy, precision, recall and F1-score, culminating in composite performance metrics. This would give us a unified and

extendable framework while demonstrating reproducibility and scalability.

2. Literature review

2.1 Importance of Data Preprocessing

Data preprocessing provides a vital impact in machine learning performances by highly influencing accuracy, behaviour and contributing to computational efficiency. As we know, raw data frequently contain redundant features, disparities over scaling and scaling-major factors impacting classifiers performance in a negative way. Therefore, preprocessing techniques such as feature scaling, encoding and dimensionality reduction are essential to improve the data quality and to optimize model learning.

Research over these factors demonstrates evidently that preprocessing decisions significantly influences predictive outcomes and should be considered central to the model design rather than accessary.

2.2 Feature scaling and its impact on performance.

Feature scaling standardizes ranges of the features for consistent model behaviour. This is vivacious for the algorithms that relies on distance calculations or gradient- based optimization.

Empirical analysis of preprocessing pipelines indicates that scaling and encoding strategies substantially affect predictive performance, particularly for linear and distance-based algorithms, whereas inadequate preprocessing may diminish model accuracy (Toer & Kim, 2023).

Experimental evidence shows that scaling makes algorithms like Support Vector Machines and Logistic Regression work much better. This is because these models are sensitive to differences in feature magnitude, while tree-based models are not very sensitive because they are based on splits (Pinheiro, 2025).

2.3 Principal Component Analysis (PCA) for Dimensionality Reduction

Principal Component Analysis is often used to lower the number of dimensions while keeping the variance the same. This makes computations faster and lessens the effects of multicollinearity. As it changes correlated variables into orthogonal components, which cuts down on redundancy and makes classifiers more stable and efficient (Zheng & Rakovski, 2021). PCA reduces the number of dimensions by keeping the most useful components and getting rid of less important variation (Bharadiya, 2023). This makes it easier to understand and reduces noise.

PCA can also help find the most important features and make it easier to recognize patterns and understand them when classifying (Younes & Kharboutly, 2023).

2.4 Dimensionality reduction in high-dimensional data.

To handle high-dimensional datasets while keeping predictive accuracy and computational efficiency, we need to be able to reduce the number of dimensions quickly and easily. Structured reduction methods have been demonstrated to surpass random projections, enhancing classification efficacy and scalability (Wojnowicz et al., 2019). Cutting down on unnecessary features makes things more efficient, but cutting too much can get rid of important information that is needed for accurate classification (Mantripragada et al., 2022; Taloba, 2018). These results underscore the necessity of reconciling dimensionality reduction with information preservation.

2.5 Preprocessing and Dimensionality Reduction in Applied Classification

For large-scale classification jobs with complicated datasets, strong preprocessing and feature engineering are important. Scalable machine learning pipelines show that good preprocessing makes real-world classification applications more reliable, easier to understand and faster (Miric, 2021). In all fields, proper preprocessing makes models more stable and scalable.

2.6 Identified Research Gap

There is a need for a systematic evaluation technique with scaling and PCA configurations within unified pipelines to determine how preprocessing methods influence performance across varied data conditions.

3. Background

3.1 Feature Scaling

Feature scaling is vital when the data is comprised of different magnitude. Since, ML algorithms rely on distance measurements which makes them sensitive to differences in the scaling, we are considering 2 scaling techniques.

3.1.1 Standardization

StandardScaler transforms data so that features have a mean of zero and a standard deviation of one. This transformation allows algorithms such as logistic regression and support vector machines to converge more efficiently during training. Studies have shown that this significantly improves the stability and performance. (Toer & Kim, 2023; Pinheiro, 2025).

3.1.2 Min Max Scaling

MinMaxScaler completely rescales the features into a fixed range (0 and 1). This way the relative distribution of values is preserved, and all features contribute with propotion during model training. Research demonstrates normalization techniques such as this can improve predictive performance for models, particularly in classification and regression pipelines (Pinheiro, 2025).

3.2 Principal Component Analysis

PCA is a widely used dimensionality reduction technique that transforms correlated features into a smaller set of uncorrelated variables called principle components. This is normally used to reduce computational complexity, removing redundant contents and avoid multicollinearity among features. Research has shown that PCA can significantly enhance classification performance in high-dimensional datasets by improving feature representation and reducing redundancy (Taloba, 2018; Buatoom, 2023).

3.3 Machine Learning Pipelines

Machine learning pipelines make it easier to combine preprocessing steps and model training into one workflow. Pipelines make sure that all changes made during training are also made during testing and prediction. Studies show that structured machine learning pipelines are important for reproducibility and scalability, especially when testing different preprocessing methods and model architectures (Toer & Kim, 2023). Pipelines make it easier to experiment and lower the risk of data leakage by combining preprocessing and model training into a single structure.

4. Methodology

4.1 Dataset Preparation

This study uses 3 datasets, to identify and evaluate the impact of preprocessing with ML pipelines. These are publicly available and chosen for its varying features, structure, dimensionality and data distributions.

4.1.1 California Housing Dataset

The California Housing dataset contains census-based information related to housing and population characteristics. It includes numerical features such as median income, housing age and average number of rooms, along with a categorical feature describing ocean proximity. The target variable is the ocean proximity category so this would be a multiclass prediction.

4.1.2 Adult Income Dataset

The Adult Income dataset, obtained from the UCI Machine Learning Repository, contains demographic and employment-related attributes such as age, education level, occupation and working hours per week. The target variable is the income, as this would be a binary classification.

4.1.3 Bank Marketing Dataset

The Bank Marketing dataset, also from the UCI Machine Learning Repository, consists of customer information collected during direct marketing campaigns conducted by a financial institution. It includes features such as age, job type, marital status, education and previous campaign outcomes. The target variable is the outcome of the client, making it a binary classification.

4.2 Pipeline Architecture

A modular machine learning pipeline was constructed for standardization, preprocessing and model training. This is done by integrating

multiple transformation steps using column transformer.

Numerical features are processed through a configurable pipeline that may include scaling and dimensionality reduction. Categorical features are transformed using **one-hot encoding**, which converts categorical values into binary indicator variables. The encoder is configured to ignore unseen categories during prediction to maintain robustness when encountering new data. This ensures that all preprocessing are applied across training and testing phases.

4.3 Preprocessing Configurations

The experimental framework evaluates several preprocessing configurations to examine how feature scaling and dimensionality reduction influence classification performance. These preprocessing techniques are applied to numerical attributes within the pipeline to ensure consistent data transformation prior to model training. Three scaling setups are used for numerical features:

- Standard Scaling (StandardScaler) takes away the mean and scales to unit variance to make features more normal.
- Min–Max Scaling (MinMaxScaler) changes the range of features to a fixed range between 0 and 1.
- No Scaling (Baseline) means that numerical features are used as they are, without any changes.

The preprocessing function builds pipelines on the fly with the chosen scaler and an optional PCA step. This makes it easy to compare different preprocessing combinations.

4.4 Classification Models

Five machine learning classifiers are evaluated within the pipeline framework to represent different algorithmic families:

- Logistic Regression—a linear classification model suitable for establishing a baseline.
- Random Forest—an ensemble learning method based on multiple decision trees.
- Histogram Gradient Boosting—a gradient boosting algorithm optimized for structured datasets.
- Support Vector Machine (SVM)—a margin-based classifier capable of modeling nonlinear relationships using the radial basis function kernel.
- Gaussian Naïve Bayes—a probabilistic classifier that assumes conditional independence among features.

These models were selected because they represent diverse learning paradigms and exhibit varying sensitivity to feature scaling and dimensionality reduction.

4.5 Pipeline Combinations

The study evaluates multiple combinations of preprocessing methods and classifiers to investigate their interaction effects. For each algorithm, pipelines are constructed using:

- No preprocessing
- Standard scaling
- Min–Max scaling
- Standard scaling with PCA
- Min–Max scaling with PCA

This systematic combination results in a collection of pipelines that allow direct comparison of preprocessing strategies across multiple algorithms.

4.5 Model Training and Evaluation

Each pipeline is trained using the training dataset and evaluated on the testing dataset. Predictions generated by the models are assessed using several performance metrics commonly used in classification tasks:

- Accuracy—proportion of correctly classified instances
- Precision—proportion of correctly predicted positive observations
- Recall—proportion of actual positives correctly identified
- F1-score—harmonic mean of precision and recall

Because relying on a single metric may not fully capture model performance, a composite

evaluation score is introduced. The composite score combines the four metrics using weighted contributions:

- F1-score: 35%.
- Precision: 25%.
- Recall: 25%.
- Accuracy: 15%.

This weighted formulation prioritizes balanced classification performance while still accounting for overall predictive accuracy.

4.7 Parallel Processing and result aggregation

To improve efficiency, pipeline training and evaluation are executed using parallel processing. Each pipeline configuration is trained independently across multiple processing cores, enabling simultaneous experimentation with different preprocessing and model combinations. This execution significantly reduces total runtime while maintaining reproducible results across experiments.

The performance metrics for all pipelines are collected and stored in a structured dataframe. The dataframe is then sorted using Composite performance metrics and visualized for interpretation.

5. Results and Discussion

The experiments analyse how scaling methods and dimensionality reduction influence classification performance across multiple machine learning algorithms. Performance is evaluated using accuracy, precision, recall and F1-score, along with a composite performance metric that provides an overall ranking of pipeline configurations.

5.1 Adult Income Dataset

For the adult income, HGB achieved the highest accuracy without any preprocessing. This is because of the inherited robustness to features scaling of the algorithm. (Figures 1 and 2) The boosting mechanism sequentially corrects errors from previous trees, enabling the model to learn patterns without the need of dimensionality reduction or normalization. As the tree-based models are scale-invariant and it has the ability to capture complex non-linear feature interactions in the adult income, the model has achieved high accuracy with raw data. (Figure 3)

5.2 California Housing

For this dataset, Random Forest achieved the highest score because the algorithm captures nonlinear relationships and interactions amongst numerical features (Figures 4 and 5). Random Forest is insensitive to scaling, and normalization places all features within the range and reduces the need of extreme value captures. This leads to more stable tree splits and improved prediction. (Figure 6)

5.3 Bank Marketing Dataset

HGB achieved the best score once again, no matter how it was preprocessed, this means that the dataset has complicated nonlinear relationships that boosting-based trees can pick up without normalization features(Figure 7). The next best method was Random Forest (Figures 8 and 9), which further proved that tree-based methods are good for this problem because they can handle different feature scales and model interactions between variables.

Linear models such as Logistic Regression and Support Vector Classification performed comparatively lower even with scaling. This suggests that linear separation does not do a good job of showing the decision boundaries. The results indicate that ensemble tree methods are the most suitable for this dataset, and preprocessing has minimal effect on their performance.

6. Conclusion

The results indicate that the choice of model has a greater impact on performance than preprocessing alone. Tree-based ensemble methods, particularly Histogram Gradient Boosting and Random Forest, consistently produced the highest accuracy and performed well even without scaling. Linear models improved with normalization but were less effective on more complex datasets. Overall, ensemble tree methods proved to be the most reliable and practical option for the structured datasets used in this study.

References

Bharadiya, J.P., 2023. *A tutorial on principal component analysis for dimensionality reduction in machine learning.*

Buatoom, U., 2023. Improving classification performance with statistically weighted dimensions and dimensionality reduction. *Applied Sciences.*

Mahapatra, A. & Patra, P.K., 2023. A unified and scalable machine learning framework for feature fusion in object classification using weighted PCA with adaptive concatenation and dynamic scaling.

Mantripragada, K., Dao, P.D., He, Y. & Qureshi, F.Z., 2021. The effects of spectral dimensionality reduction on hyperspectral pixel classification: A case study.

Miric, M., 2021. Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Journal of Management Analytics.*

Pinheiro, J.M.H., 2025. The impact of feature scaling in machine learning: Effects on regression and classification tasks.

Taloba, A.I., 2018. A comparative study on using principal component analysis with different text classifiers.

Toer, G.A. & Kim, G., 2023. An empirical study on the impact of feature scaling and encoding strategies on machine learning regression pipelines.

Wojnowicz, M., Zhang, D., Chisholm, G., Zhao, X. & Wolff, M., 2019. Projecting "better than randomly": How to reduce the dimensionality of very large datasets. *Information Sciences.*

Younes, K. & Kharboutly, Y., 2023. Application of unsupervised machine learning for the evaluation of aerogels' efficiency towards ion removal—A principal component analysis (PCA) approach.

Zheng, A. & Rakovski, C., 2021. On the application of principal component analysis to classification problems. *Data Science Journal.*

Datasets:

Moro, S., P. Rita, and P. Cortez. 2014. Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306. (Accessed: 13 March 2026).

Becker, Barry and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20. (Accessed: 13 March 2026).

Pace, R.K. and Barry, R. (1997) *California Housing Dataset*. Available at: OpenML: https://www.openml.org (Accessed: 13 March 2026).

Figures

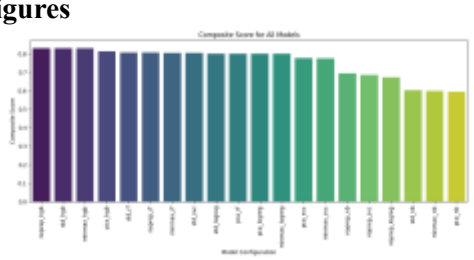


Figure 1. Composite score for all model configuration with Adult Income Dataset

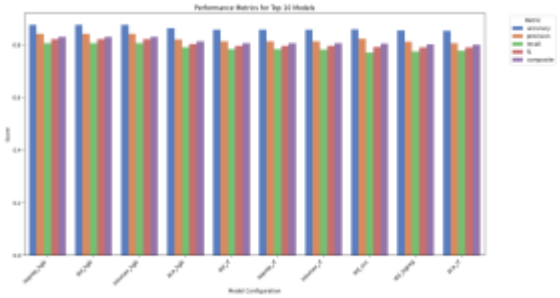


Figure 2. Accuracy, Precision, Recall, F1, and composite score for all model configuration with Adult Income Dataset

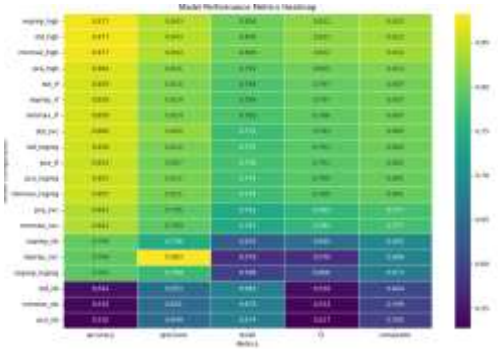


Figure 3. Heatmap comparison of performance metrics with Adult Income Dataset.

International Journal of Engineering Research and Science & Technology

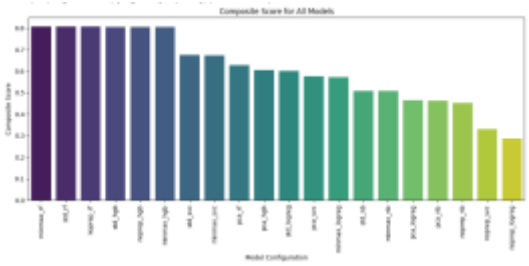


Figure 4. Composite score for all model configuration with California Housing Dataset

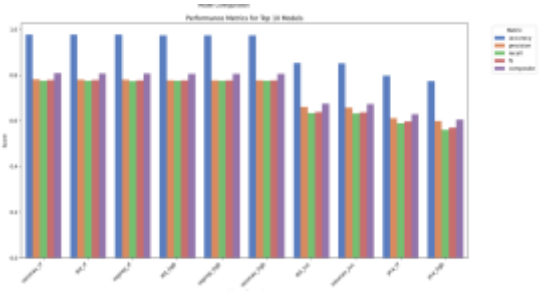


Figure 5. Accuracy, Precision, Recall, F1, and composite score for all model configuration with California Housing Dataset.

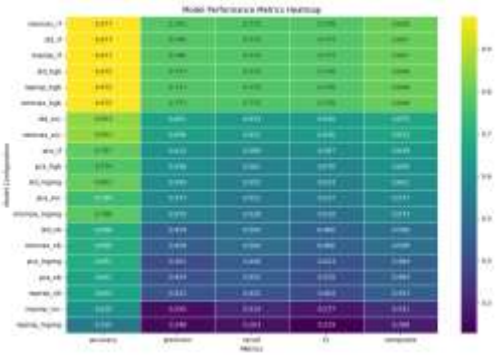


Figure 6. Heatmap comparison of performance metrics with California Housing Dataset.

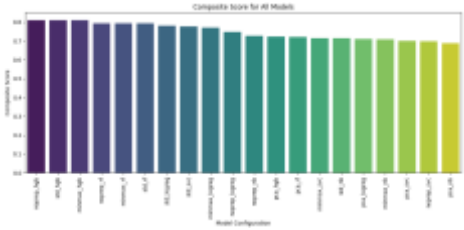


Figure 7. Composite score for all model configuration with Bank Marketing Dataset

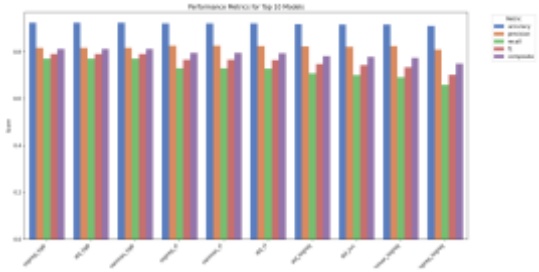


Figure 8. Accuracy, Precision, Recall, F1, and composite score for all model configuration with Bank Marketing Dataset

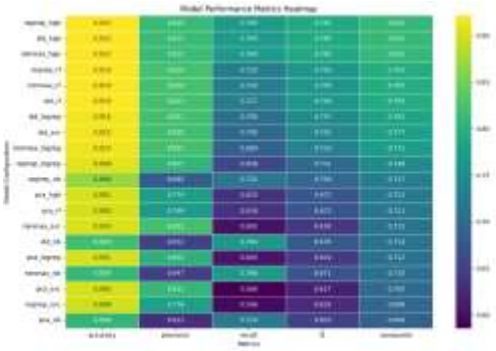


Figure 9. Heatmap comparison of performance metrics with Bank Marketing Dataset.