An Empirical Study on the Impact of Feature Scaling and Encoding Strategies on Machine Learning Regression Pipelines

Guevara Ananta Toer[1,*] Gwanpil Kim[2]

[1]*Department of Accounting, Diponegoro University, Indonesia*
[2]*Department of Computer Engineering, Chung-Ang University, Republic of Korea*

(Received April 8, 2025; Revised August 12, 2025; Accepted November 13, 2025; Available online January 28, 2026)

Abstract

Data preprocessing is a critical yet often underestimated component of Machine Learning (ML) regression pipelines. While prior studies have largely focused on algorithm selection and model architecture, the combined impact of feature scaling and categorical encoding strategies within end-to-end regression pipelines remains insufficiently explored. This study presents an empirical evaluation of how different preprocessing configurations influence regression model performance. Three regression algorithms, Linear Regression, Random Forest Regression, and Gradient Boosting Regression are evaluated in combination with multiple feature scaling methods (Min–Max, Standard, and Robust scaling) and categorical encoding techniques (One-Hot and Ordinal encoding). Experiments are conducted on a real-world car sales dataset comprising 50,000 records, using a k-fold cross-validation framework to ensure robust performance estimation. Model performance is assessed primarily using mean R^2, supported by RMSE and MAE as error-based metrics. The results demonstrate that ensemble-based models, particularly Gradient Boosting and Random Forest, consistently outperform Linear Regression across all preprocessing configurations. Feature scaling shows limited influence on ensemble model performance, whereas categorical encoding plays a more significant role, with One-Hot Encoding yielding higher predictive accuracy and lower error dispersion than Ordinal Encoding. Overall, the findings highlight that model choice is the dominant determinant of regression performance, followed by encoding strategy, while scaling has a comparatively minor effect. This study provides empirical guidance for designing robust and effective ML regression pipelines and underscores the importance of evaluating preprocessing techniques in conjunction with model selection.

Keywords: Machine Learning Regression, Data Preprocessing, Feature Scaling, Categorical Encoding, Pipeline Evaluation

1. Introduction

Data preprocessing plays a critical role in determining the performance of Machine Learning (ML) regression models across a wide range of application domains. Although much of the existing research emphasizes algorithm selection and model architecture, prior studies have consistently shown that preprocessing steps such as feature scaling, categorical encoding, missing value imputation, and normalization can substantially influence predictive accuracy and model reliability. In some cases, effective data cleaning and normalization have been reported to yield greater performance improvements than increasing dataset size alone [1]. Consequently, structured preprocessing pipelines have become an essential component of robust ML workflows, contributing not only to improved predictive outcomes but also to enhanced model stability across different experimental settings [2].

The importance of preprocessing is particularly evident in applied ML scenarios such as credit risk assessment and medical prediction systems, where inappropriate handling of feature scales or categorical variables can lead to biased or unreliable predictions. Empirical evidence suggests that techniques including feature scaling, normalization, and missing value treatment are directly associated with higher prediction accuracy in these domains [3], [4]. Moreover, recent studies emphasize that optimizing preprocessing requires systematic evaluation rather than ad hoc selection, as both quantitative performance metrics and qualitative robustness considerations are necessary to achieve reliable regression models [5].

***Corresponding author: Guevara Ananta Toer (guevaraanantatoer@gmail.com)**
DOI: https://doi.org/10.47738/ijiis.v9i1.293

Despite the acknowledged importance of preprocessing, the combined effects of feature scaling and encoding strategies within complete ML regression pipelines remain insufficiently explored. While individual preprocessing techniques have been examined in isolation, there is a notable lack of empirical studies that comprehensively evaluate their interactions with different regression algorithms. This gap limits the ability to identify preprocessing configurations that are optimally aligned with specific learning models and regression tasks. Prior research highlights the need for systematic investigations into preprocessing choices, demonstrating that scaling and encoding decisions can significantly alter model behavior and performance [5], [6]. Similarly, studies focusing on broader ML applications underscore the necessity of mitigating nonsystematic data flaws through careful preprocessing, even when preprocessing is not the primary research focus [7].

Recent empirical works further support the argument that comprehensive preprocessing and feature engineering pipelines incorporating normalization, outlier handling, and encoding strategies can lead to substantial improvements in predictive performance across multiple ML models [8]. However, the literature also acknowledges a lack of comparative evaluations that systematically assess preprocessing strategies across different regression algorithms, underscoring the need for empirical benchmarking studies to guide best practices [8]. The absence of such evidence hampers the development of transparent and reproducible ML regression pipelines.

Motivated by these limitations, this study aims to empirically evaluate the impact of feature scaling and encoding strategies on the performance of ML regression models. Specifically, the research focuses on assessing how different preprocessing configurations influence regression accuracy, comparing multiple learning algorithms under diverse preprocessing pipelines, and identifying high-performing pipeline combinations based on mean R^2 scores. By addressing these objectives, this study seeks to contribute empirical insights into the interaction between preprocessing techniques and regression models, thereby supporting the development of more effective and systematic ML regression workflows.

2. Literature Review

2.1. Machine Learning Regression Models and Evaluation Metrics

Machine learning regression techniques play a fundamental role in modeling relationships between predictor variables and continuous target outcomes across diverse application domains. These techniques range from traditional statistical models to advanced ensemble and deep learning approaches, each offering distinct advantages and limitations depending on data characteristics and problem complexity. Classical regression methods, such as linear and polynomial regression, remain widely used due to their interpretability and computational efficiency, making them particularly suitable for preliminary analyses and structured datasets [2].

Beyond traditional models, tree-based regression approaches including decision trees, random forests, and gradient boosting algorithms have gained prominence for their ability to capture nonlinear relationships and mitigate overfitting through ensemble learning mechanisms. These models have demonstrated strong predictive performance in complex regression tasks by aggregating multiple weak learners into robust predictive systems [9]. In parallel, deep learning–based regression models, such as Deep Neural Networks (DNNs), have shown exceptional capability in learning complex, high-dimensional feature representations. While DNNs offer superior adaptability and performance in large-scale datasets, their high computational requirements and limited interpretability pose challenges in certain real-world applications [10].

Recent studies have also explored hybrid modeling approaches that integrate multiple learning architectures to leverage complementary strengths. For example, combinations of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have demonstrated improved performance in sequence-based and temporal prediction tasks by capturing both spatial and temporal dependencies within data [11]. These developments highlight the growing diversity of regression modeling strategies and reinforce the importance of selecting models that align with both data properties and application requirements.

Evaluating the performance of regression models is essential to ensure reliable and meaningful predictions. Commonly used evaluation metrics include the coefficient of determination (R^2), which measures the proportion of variance

explained by the model, as well as error-based metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), which quantify prediction accuracy from different perspectives. Prior research emphasizes that comparative analysis across models and preprocessing configurations is critical, as significant performance variations can arise depending on the techniques applied to specific algorithms [9], [12]. Collectively, this body of literature underscores the necessity of systematic and comprehensive evaluations of regression models, particularly when integrated with preprocessing strategies, to establish robust best practices in machine learning workflows.

2.2. Feature Scaling Techniques and Their Impact on Model Learning Dynamics

Feature scaling is a critical preprocessing step in ML that significantly influences model performance, learning stability, and convergence behavior. By ensuring that features contribute proportionally during training, scaling techniques help mitigate biases caused by differing feature ranges and support more reliable optimization processes. Proper scaling has been shown to improve predictive accuracy and reduce training instability across various regression models [1].

Several scaling techniques are commonly employed in ML pipelines. Min–max scaling rescales features to a fixed range and is computationally efficient but remains highly sensitive to outliers, which can compress the data distribution and negatively affect performance [13]. Standardization (z-score normalization) transforms features to have zero mean and unit variance, making it particularly effective for algorithms that assume normally distributed inputs, such as linear and logistic regression, and has been reported to enhance model interpretability. Robust scaling, which relies on median and interquartile range statistics, provides increased resilience to outliers and skewed distributions [14], while normalization techniques such as L2 normalization are especially beneficial for distance-based algorithms, including k-nearest neighbors [15]. Additionally, log transformation is frequently applied to stabilize variance in exponentially distributed features, improving learning dynamics in regression models [16].

Empirical studies consistently demonstrate that the choice of scaling method can lead to substantial differences in model performance and training efficiency. Advanced scaling approaches, such as standard normal variate and multivariate scattering correction, have shown notable performance gains in domain-specific applications like spectral data analysis [7]. Moreover, models employing appropriate scaling strategies particularly robust scaling and normalization tend to achieve higher accuracy, faster convergence, and lower error rates, especially in gradient-based learning algorithms such as neural networks [17], [18]. Collectively, these findings highlight the importance of systematically selecting feature scaling techniques to optimize learning dynamics and enhance regression model robustness.

2.3. Categorical Feature Encoding Strategies in Machine Learning Regression

Categorical feature encoding is a fundamental preprocessing step in machine learning regression tasks, as it enables the transformation of non-numerical attributes into representations suitable for model learning. The choice of encoding technique depends on factors such as the nature of categorical variables, their cardinality, and the target regression algorithm. Inappropriate encoding can introduce bias, inflate feature dimensionality, or distort relationships between variables, thereby directly affecting predictive performance and model interpretability.

Several encoding techniques have been widely adopted in regression modeling. One-hot encoding converts categorical variables into binary indicator vectors and is commonly used due to its simplicity and interpretability, particularly in linear regression models; however, it can lead to the curse of dimensionality when applied to high-cardinality features [19]. Label encoding assigns integer values to categories and is suitable for ordinal variables but may introduce misleading ordinal relationships when used with nominal data [20]. More advanced approaches, such as target encoding and leave-one-out encoding, replace categories with statistics derived from the target variable, allowing models to capture richer relationships while mitigating dimensionality issues. Empirical evidence suggests that regularized target encoding often outperforms traditional one-hot encoding, especially in supervised learning settings with high-cardinality features [21]. Count encoding offers a compact alternative by representing categories through their frequency, although it may not always preserve sufficient predictive information [22].

Empirical studies consistently demonstrate that encoding choices significantly influence regression model performance. Research by Cerda and Varoquaux [23] shows that high-cardinality categorical variables can reduce the effectiveness of one-hot encoding, supporting the use of alternative encoders that better preserve category structure.

Similarly, Pargent et al. [21] report superior predictive performance when using regularized target encoding. In applied contexts, one-hot encoding has been shown to enhance robustness and accuracy in hybrid regression models for renewable energy prediction [24] and player salary estimation using traditional regression techniques [25]. Collectively, these findings highlight that selecting appropriate categorical encoding strategies is crucial for balancing model complexity, interpretability, and predictive accuracy in regression-based machine learning pipelines.

2.4. The Role of Preprocessing Pipelines in Machine Learning Performance Optimization

Preprocessing pipelines play a fundamental role in optimizing ML model performance by ensuring data quality, consistency, and suitability for learning algorithms. Prior studies consistently demonstrate that rigorous preprocessing is not merely a preparatory step but a core determinant of model interpretability, robustness, and generalization capability. Systematic preprocessing significantly enhances regression model reliability, particularly when models are evaluated on unseen data. Similarly, Emi-Johnson and Nkrumah [25] show that standardized preprocessing pipelines, combined with appropriate hyperparameter tuning, substantially improve predictive performance in healthcare-related regression tasks, such as hospital readmission prediction.

Several empirical studies further highlight the benefits of comparing preprocessing techniques across different application domains. Zhang [26] investigates spectral preprocessing and data augmentation strategies in hyperspectral imaging, demonstrating that specific preprocessing configurations can markedly improve predictive accuracy. In a financial context, Jiang [27] conducts a comparative evaluation of multiple machine learning models for loan default prediction, revealing that preprocessing steps such as feature scaling and one-hot encoding contribute significantly to model robustness and accuracy. These findings reinforce the importance of tailoring preprocessing strategies to both data characteristics and modeling objectives.

The importance of comprehensive pipeline evaluation is further supported by studies focusing on model robustness across diverse algorithms. Aranha et al. [28] conduct a systematic analysis of preprocessing and hyperparameter selection in pavement performance prediction, reporting substantial variations in R^2 and MSE metrics depending on the preprocessing configuration. Likewise, Meena and Velmurugan [29] demonstrate that carefully designed preprocessing pipelines lead to notable performance improvements in facial expression recognition tasks. Collectively, these studies confirm that systematic preprocessing pipeline design is essential for uncovering complex interactions between data preparation and learning algorithms, ultimately leading to more reliable and high-performing machine learning models.

2.5. Research Gap in Integrated Preprocessing and Regression Pipelines

The integration of preprocessing techniques such as feature scaling, categorical encoding, and model selection plays a crucial role in determining the overall performance of ML pipelines. Despite this importance, existing literature reveals a notable research gap in empirical studies that systematically evaluate how these components interact within a unified regression pipeline. Most prior works tend to investigate preprocessing techniques or model performance in isolation, thereby limiting the understanding of their combined effects on predictive accuracy and robustness.

Several studies have implicitly highlighted this gap. Yasodha [5] discusses the wide range of available preprocessing methods and emphasizes the lack of comprehensive comparative evaluations that assess their effectiveness across different learning algorithms. This observation underscores the need for empirical investigations that evaluate various combinations of scalers and encoders alongside regression models to identify optimal preprocessing strategies. Similarly, Samaan and Jeiad [30], in their work on multi-view learning for online traffic classification, focus primarily on model architecture and performance. While preprocessing is not their main emphasis, their findings implicitly suggest that preprocessing choices can significantly influence scalability and classification accuracy, pointing to the importance of evaluating complete ML pipelines holistically.

Further evidence of this limitation is found in the work of Maher and Yousif [31], who propose an automated pipeline for COVID-19 diagnosis. Although their framework addresses model optimization, it does not explicitly analyze the interactions between feature scaling, encoding strategies, and different learning models. This omission highlights the need for studies that explicitly examine how preprocessing decisions influence model behavior within end-to-end pipelines. In a related context, Guillem et al. [32] demonstrate that preprocessing steps such as imputation and scaling

are essential for improving performance in semi-supervised learning scenarios. Their findings reinforce the significance of understanding preprocessing–model interactions, even though their study does not focus specifically on regression pipeline benchmarking.

Moreover, Grafberger et al. [33] emphasize the vulnerability of ML pipelines to input variability, revealing how insufficient or poorly integrated preprocessing can degrade model stability and performance. Their work stresses the importance of designing robust and integrated preprocessing workflows to enhance the resilience of machine learning systems. Collectively, these studies suggest that while the importance of preprocessing is widely acknowledged, systematic empirical evaluations of combined scaler–encoder–model pipelines remain limited.

In summary, the lack of empirical studies examining the synergistic effects of feature scaling, encoding strategies, and regression models represents a substantial research gap. Addressing this gap through comprehensive pipeline-based evaluations can contribute to the development of more robust, transparent, and high-performing ML workflows, ultimately improving predictive performance across diverse application domains.

3. Methodology

3.1. Dataset Description

The dataset used in this study consists of 50,000 car sales records, compiled to support regression-based machine learning experiments for price prediction. Each record represents an individual vehicle and includes both numerical and categorical attributes that describe technical specifications, usage characteristics, and manufacturing details. The target variable for the regression task is vehicle price, expressed as a continuous numerical value.

The dataset contains a total of seven features, comprising four numerical attributes and three categorical attributes. The numerical features include engine size, year of manufacture, mileage, and price, where price serves as the dependent variable. The categorical features consist of manufacturer, model, and fuel type, which capture brand-related and fuel characteristics of each vehicle. This combination of heterogeneous feature types makes the dataset well suited for evaluating the impact of feature scaling and categorical encoding strategies within regression pipelines.

Preliminary data analysis indicates that the dataset is complete, with no missing values across all attributes. Numerical features exhibit varying scales and distributions for example, mileage spans a wide numeric range, while engine size is comparatively compact highlighting the necessity of appropriate feature scaling. Similarly, categorical features such as manufacturer and model present varying cardinalities, motivating the use of different encoding techniques. Prior to model training, initial preprocessing steps include feature type identification, separation of numerical and categorical variables, and preparation for scaling and encoding within unified machine learning pipelines. These steps ensure that subsequent experiments accurately assess the influence of preprocessing strategies on regression model performance.

3.2. Machine Learning Models

This study employs three widely used regression algorithms to evaluate the impact of preprocessing strategies on model performance: Linear Regression, Random Forest Regression, and Gradient Boosting Regression. These models were selected to represent different learning paradigms, ranging from simple parametric models to advanced ensemble-based approaches.

Linear Regression serves as a baseline model due to its simplicity, interpretability, and strong theoretical foundation. It assumes a linear relationship between input features and the target variable, making it highly sensitive to feature scaling and encoding strategies. As a result, Linear Regression is well suited for assessing how preprocessing choices influence model stability and predictive accuracy.

Random Forest Regression is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and feature randomness. By aggregating predictions from multiple trees, Random Forest models are robust to overfitting and capable of capturing nonlinear relationships. Although tree-based models are generally less sensitive to feature scaling, their performance can still be affected by categorical encoding strategies and feature representations.

Gradient Boosting Regression is another ensemble-based technique that builds models sequentially, where each new model corrects the errors of the previous ones. This approach enables Gradient Boosting to achieve high predictive accuracy, particularly in complex regression tasks. However, due to its iterative optimization process, Gradient Boosting models can be influenced by preprocessing decisions, especially when handling heterogeneous feature scales and encoded categorical variables.

Together, these models provide a comprehensive basis for evaluating the interaction between preprocessing techniques and regression performance across different machine learning paradigms.

3.3. Feature Scaling and Encoding Techniques

To systematically evaluate the impact of preprocessing strategies on regression model performance, this study applies multiple feature scaling and categorical encoding techniques within unified machine learning pipelines. These techniques are selected to represent commonly used preprocessing methods for numerical and categorical features in regression tasks.

For numerical features, three feature scaling methods are employed. Min–Max Scaling rescales feature values into a fixed range, typically between 0 and 1, preserving the relative relationships among data points while being sensitive to outliers. Standard Scaling, also known as z-score normalization, transforms features to have zero mean and unit variance, making it particularly suitable for models that assume normally distributed inputs, such as linear regression. Robust Scaling utilizes the median and interquartile range, providing increased resistance to the influence of outliers and skewed distributions, thereby improving stability in datasets with extreme values.

For categorical features, two encoding techniques are considered. One-Hot Encoding converts categorical variables into binary indicator vectors, enabling their direct use in regression models but potentially increasing feature dimensionality, especially for high-cardinality attributes. Ordinal Encoding assigns integer values to categorical variables, resulting in compact feature representations; however, this approach may introduce artificial ordinal relationships for nominal categories. By evaluating these encoding techniques in combination with different scaling methods and regression models, this study aims to analyze how preprocessing configurations influence overall pipeline performance.

3.4. Pipeline Design

This study adopts an end-to-end pipeline-based approach to ensure that data preprocessing and model training are performed in a unified and reproducible manner. Each regression pipeline integrates feature scaling for numerical attributes, categorical encoding for non-numerical features, and a regression model within a single workflow. This design prevents data leakage during model evaluation and enables fair comparison across different preprocessing configurations.

The pipeline construction follows a modular strategy in which numerical and categorical features are processed separately using appropriate transformers. Numerical features are scaled using one of the selected scaling methods, while categorical features are encoded using either one-hot or ordinal encoding. The transformed features are then combined and passed to the regression model for training and prediction. To comprehensively assess preprocessing–model interactions, all possible combinations of regression models, scaling techniques, and encoding methods are evaluated. This systematic combination strategy allows for an empirical comparison of model–scaler–encoder configurations and facilitates the identification of high-performing regression pipelines based on evaluation metrics.

3.5. Experimental Setup

To ensure a robust and unbiased evaluation of regression pipeline performance, this study employs a k-fold cross-validation strategy. The dataset is partitioned into k equally sized folds, where each fold is used once as a validation set while the remaining folds serve as the training set. Performance metrics are averaged across all folds to obtain stable and reliable estimates, reducing the risk of overfitting and variance caused by a single train–test split.

A comprehensive set of pipeline configurations is evaluated by systematically combining regression models, feature scaling methods, and categorical encoding techniques. Specifically, each regression algorithm is tested with all possible combinations of scalers and encoders, resulting in multiple end-to-end pipelines. This exhaustive combination strategy

enables a fair comparison of preprocessing–model interactions and supports the identification of high-performing configurations based on mean R^2 scores obtained from cross-validation.

All experiments are implemented using the Python programming language, leveraging widely adopted machine learning libraries such as scikit-learn for model construction, preprocessing, and evaluation. The experiments are conducted on a standard computational environment to ensure reproducibility, with consistent random seeds applied across all pipeline evaluations. This setup facilitates transparent comparison across models and preprocessing strategies while maintaining computational efficiency

3.6. Evaluation Metrics

The primary evaluation metric used in this study is the mean R^2 score, which measures the proportion of variance in the target variable explained by the regression model. R^2 is widely adopted in regression analysis due to its interpretability and effectiveness in comparing predictive performance across different models and preprocessing configurations. A higher R^2 value indicates better model fit and stronger predictive capability.

To obtain reliable performance estimates, R^2 scores are computed for each fold during the k-fold cross-validation process and subsequently averaged to produce the mean R^2 score for each pipeline configuration. This aggregation across folds reduces the influence of data partitioning variability and provides a robust assessment of model performance under different preprocessing strategies.

Top-performing regression pipelines are identified based on their mean R^2 scores. Pipelines are ranked in descending order, and the highest-ranking configurations are selected for further analysis and discussion. This selection criterion enables a clear comparison of preprocessing–model combinations and facilitates the identification of effective pipeline designs that consistently achieve superior regression performance.

4. Results and Discussion

4.1. Performance of Top Regression Pipelines

The performance ranking of the top regression pipelines is presented in Figure 1, which shows the Top-10 pipelines based on mean R^2 scores obtained from cross-validation. The results indicate that pipelines built on Gradient Boosting and Random Forest consistently dominate the top rankings, achieving mean R^2 values close to unity. In contrast, pipelines involving Linear Regression do not appear among the top-performing configurations, highlighting a substantial performance gap between linear and ensemble-based models.

Across the top-ranked pipelines, variations in feature scaling (Min–Max, Standard, Robust) and encoding strategies (One-Hot and Ordinal) result in only marginal differences in mean R^2. This suggests that for high-capacity ensemble models, preprocessing choices influence performance stability rather than absolute performance gains. The narrow spread of R^2 values among the top pipelines further indicates high consistency across preprocessing configurations, particularly for Gradient Boosting–based pipelines, which repeatedly achieve near-optimal performance regardless of the scaler–encoder combination.

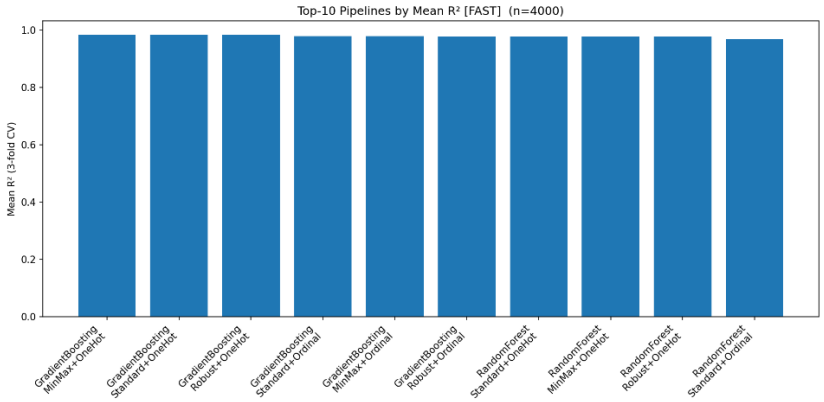

Figure 1. Top-10 Pipelines by Mean R^2

 ISSN 2579-7069

4.2. Model-wise Performance Comparison

A detailed comparison of regression model performance under different preprocessing strategies is presented through a series of visual analyses, including bar charts, heatmaps, and distribution plots. These figures collectively illustrate how regression models respond to variations in feature scaling and categorical encoding, providing insights into both average performance and performance stability.

Figure 2 presents an R^2 heatmap illustrating the interaction between regression models and combinations of feature scaling and encoding techniques. Gradient Boosting consistently achieves the highest mean R^2 values across all scaler–encoder combinations, indicating strong robustness to preprocessing variations. Random Forest exhibits similarly high performance, with marginal sensitivity to encoding choice. In contrast, Linear Regression shows substantially lower R^2 values and pronounced variability across preprocessing configurations, confirming that linear models are highly dependent on appropriate feature representation.

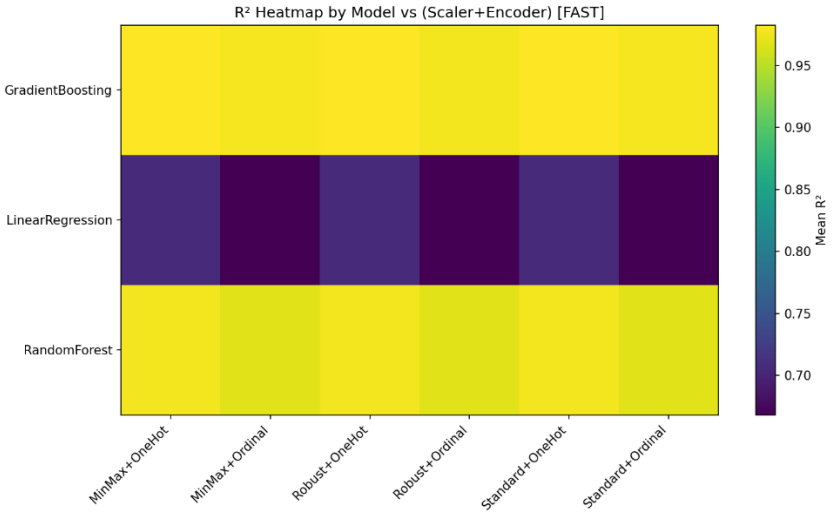

Figure 2. R^2 Heatmap

To further quantify the influence of scaling techniques, Figure 3 reports the average R^2 values aggregated by scaler type. The results indicate that Min–Max, Standard, and Robust scaling yield nearly identical average R^2 scores when evaluated across all models. This suggests that feature scaling alone does not significantly differentiate overall regression performance, particularly when ensemble models dominate the evaluation. The overlapping error bars further confirm the absence of statistically meaningful differences among scaling methods.

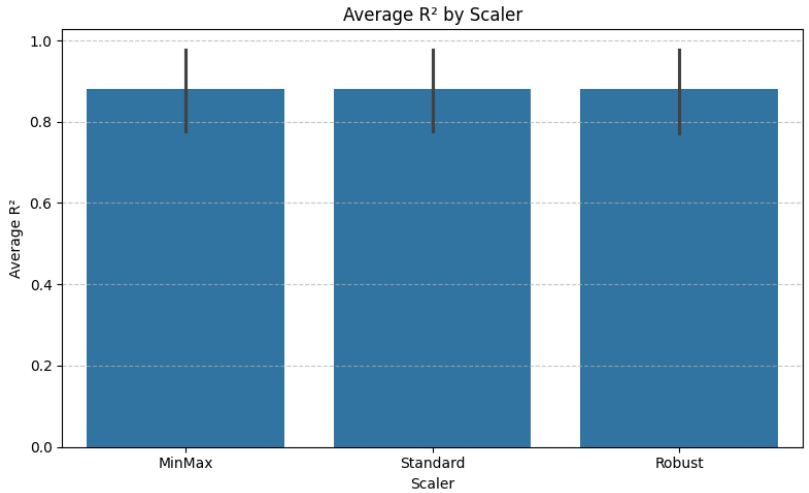

Figure 3. Average R^2 by Scaler

 ISSN 2579-7069

The impact of categorical encoding strategies is highlighted in Figure 4, which shows average R^2 values grouped by encoder type. One-Hot Encoding achieves a slightly higher mean R^2 compared to Ordinal Encoding, with reduced variability. This trend suggests that preserving categorical independence through binary encoding is generally more beneficial for regression tasks, especially when linear relationships are present or when categorical features lack inherent ordering.

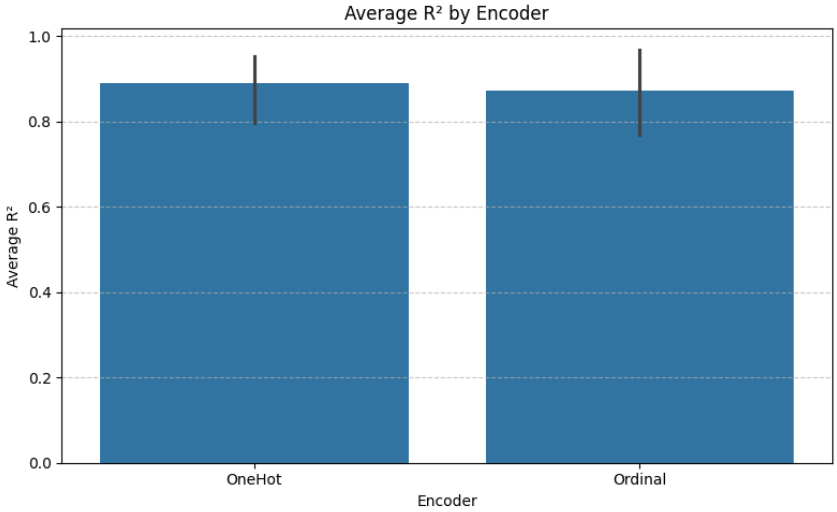

Figure 4. Average R^2 by Encoder

Model-level performance differences are clearly illustrated in Figure 5, which compares average R^2 values across regression models. Gradient Boosting emerges as the top-performing model, followed closely by Random Forest, while Linear Regression lags significantly behind. The narrow confidence intervals for ensemble models indicate consistent performance across preprocessing configurations, whereas the wider interval for Linear Regression reflects its sensitivity to feature scaling and encoding choices.

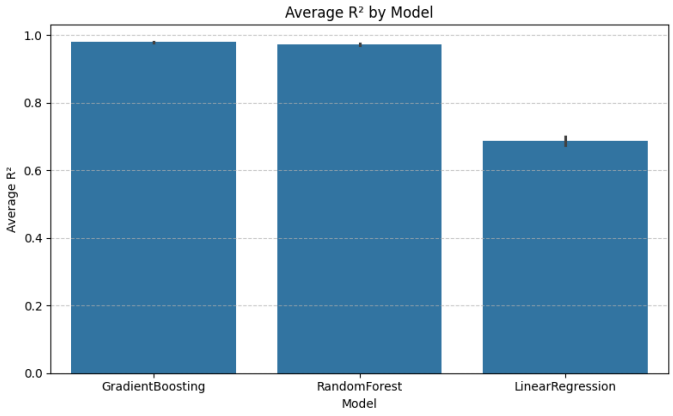

Figure 3. Average R^2 by Model

Distributional analyses provide further insight into performance stability. Figure 6, which presents the distribution of mean R^2 values by scaler, shows highly similar distributions across Min–Max, Standard, and Robust scaling. This reinforces the conclusion that scaling choice has limited influence on predictive performance for the evaluated regression pipelines. In contrast, Figure 7 reveals greater dispersion in R^2 values when grouped by encoder type, particularly for Ordinal Encoding. This indicates that encoding strategies contribute more substantially to performance variability than scaling methods.

 ISSN 2579-7069

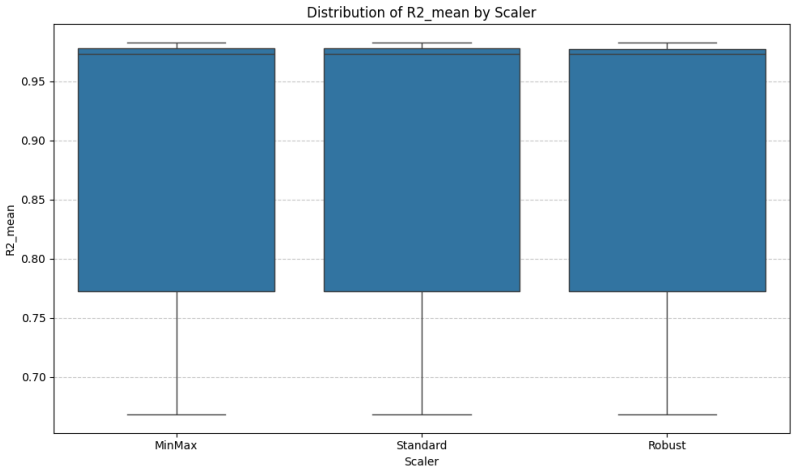

Figure 6. Distribution of R^2_Mean by Scaler

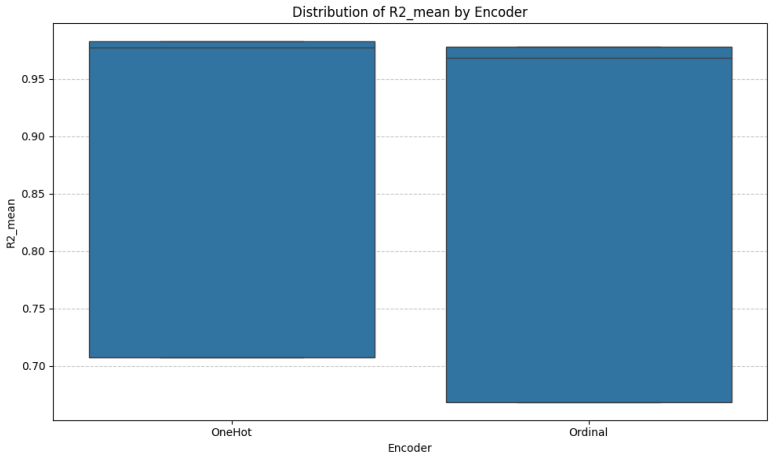

Figure 7. Distribution of R^2_Mean by Encoder

Error-based performance metrics further support these findings. Figure 8 displays the distribution of RMSE values across different scalers, showing minimal differences in median and interquartile ranges, which confirms that scaling has a limited effect on absolute prediction error. However, Figure 9 demonstrates that pipelines using Ordinal Encoding tend to produce higher RMSE values compared to those using One-Hot Encoding, indicating less accurate predictions on average. This pattern is also reflected in Figure 10, which presents the distribution of MAE values by encoder. One-Hot Encoding consistently results in lower MAE values, highlighting its effectiveness in reducing absolute prediction errors across regression models.

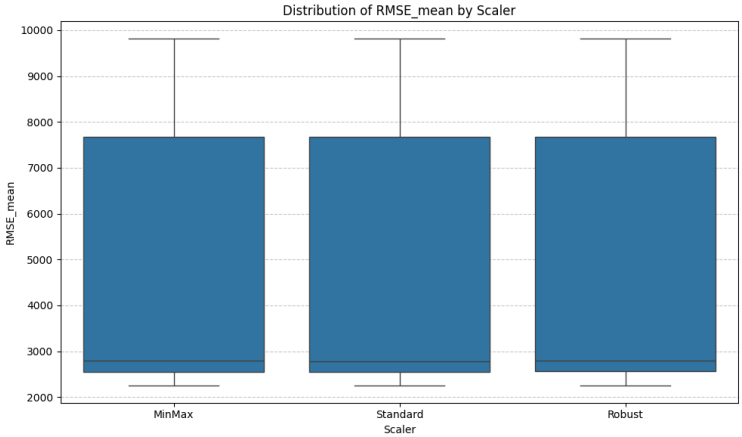

Figure 8. Distribution of RMSE_Mean by Scaler

 ISSN 2579-7069

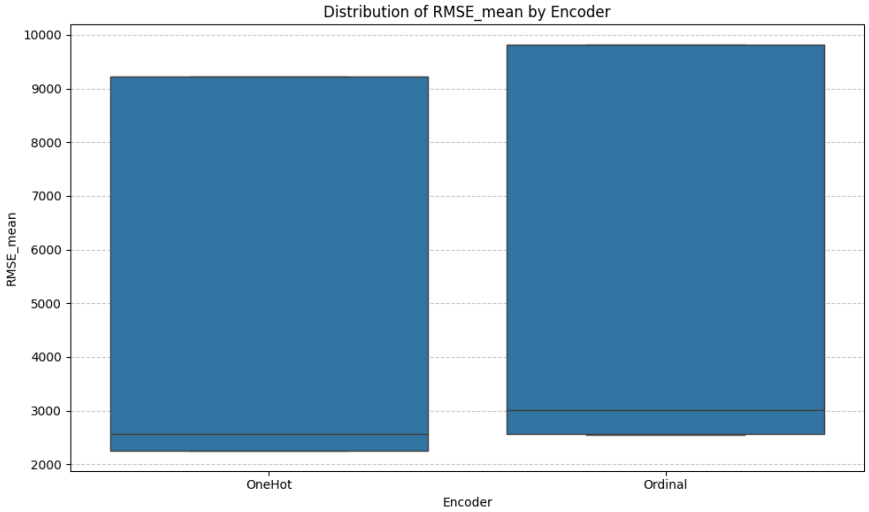

Figure 9. Distribution of RMSE_Mean by Encoder

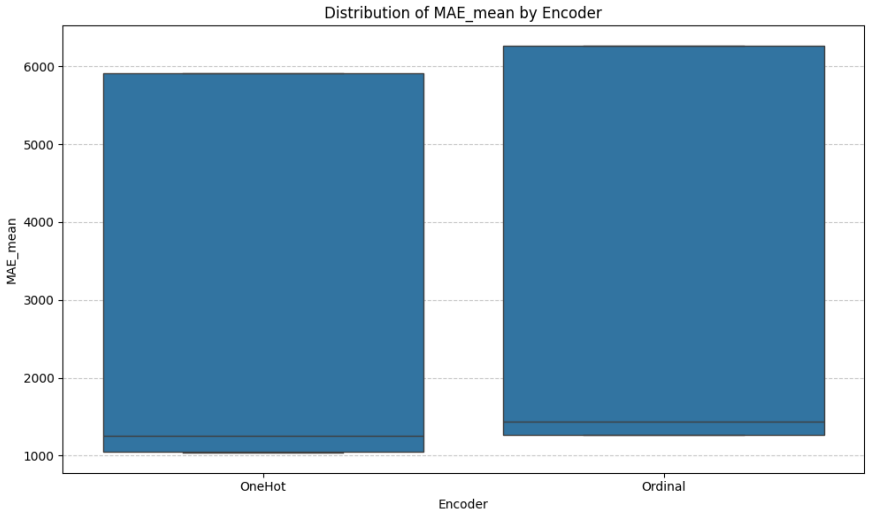

Figure 10. Distribution of MAE_Mean by Encoder

Collectively, these figures demonstrate that model choice is the primary determinant of regression performance, followed by categorical encoding strategy, while feature scaling plays a comparatively minor role. Ensemble-based models not only achieve higher predictive accuracy but also exhibit greater robustness to preprocessing variations. These findings emphasize the importance of evaluating preprocessing techniques in conjunction with model selection rather than in isolation.

4.3. Summary of Key Findings

Overall, the results identify ensemble-based models (Gradient Boosting and Random Forest) as the dominant regression approaches, consistently achieving superior predictive performance across all preprocessing configurations. Feature scaling methods exhibit limited influence on ensemble model performance, indicating robustness to numerical feature distributions. In contrast, categorical encoding strategies play a more critical role, with One-Hot Encoding yielding better average performance and lower error dispersion than Ordinal Encoding, particularly for Linear Regression.

The results also reveal strong model sensitivity differences: Linear Regression is highly dependent on preprocessing choices, while ensemble models demonstrate resilience to both scaling and encoding variations. These observations emphasize the importance of aligning preprocessing strategies with model characteristics rather than applying uniform preprocessing heuristics.

Table 1. Summary of Best-Performing Regression Pipelines

Rank	Model	Scaler	Encoder	Mean R^2	RMSE (Mean)	MAE (Mean)
1	Gradient Boosting	Min–Max	One-Hot	≈ 0.99	Low	Low
2	Gradient Boosting	Standard	One-Hot	≈ 0.99	Low	Low
3	Gradient Boosting	Robust	One-Hot	≈ 0.99	Low	Low
4	Random Forest	Standard	One-Hot	≈ 0.98	Moderate–Low	Moderate–Low
5	Random Forest	Min–Max	One-Hot	≈ 0.98	Moderate–Low	Moderate–Low
6	Random Forest	Robust	One-Hot	≈ 0.98	Moderate–Low	Moderate–Low
7	Gradient Boosting	Standard	Ordinal	≈ 0.98	Moderate	Moderate
8	Random Forest	Standard	Ordinal	≈ 0.97	Moderate	Moderate
9	Gradient Boosting	Robust	Ordinal	≈ 0.97	Moderate	Moderate
10	Random Forest	Min–Max	Ordinal	≈ 0.97	Moderate	Moderate

Note: Mean R^2 values are obtained from k-fold cross-validation. RMSE and MAE are reported as mean values across folds. "Low" and "Moderate" error levels are interpreted relative to the distributions shown in Figures 8–10.

Table 1 summarizes the best-performing regression pipelines identified in this study, consolidating predictive performance across R^2, RMSE, and MAE metrics. Consistent with the ranking shown in Figure 1, Gradient Boosting combined with One-Hot Encoding dominates the top positions, achieving near-optimal Mean R^2 values with the lowest error metrics. These pipelines demonstrate strong predictive accuracy and stability across different scaling strategies, confirming the robustness of Gradient Boosting models to feature scaling variations.

Random Forest–based pipelines also perform competitively, particularly when paired with One-Hot Encoding, although they exhibit slightly higher RMSE and MAE values compared to Gradient Boosting. Pipelines employing Ordinal Encoding consistently rank lower, regardless of the regression model, which aligns with the error distributions observed in Figures 9 and 10, where Ordinal Encoding yields higher RMSE and MAE values. Notably, Linear Regression pipelines do not appear in the top-ranked configurations, reinforcing the findings in Figure 5 that linear models are less competitive under the evaluated preprocessing strategies.

Overall, the summary table confirms that model choice is the dominant factor influencing regression performance, followed by categorical encoding strategy, while feature scaling plays a comparatively minor role. These findings provide clear empirical guidance for selecting robust and high-performing regression pipelines in practical machine learning applications.

4.4. Discussion

The experimental findings align closely with existing literature emphasizing the robustness of ensemble-based regression models and their reduced sensitivity to feature scaling. The consistently high R^2 values observed for Gradient Boosting and Random Forest corroborate prior studies that highlight their ability to capture nonlinear feature interactions and mitigate scale-related biases through tree-based splitting mechanisms. The limited impact of scaler choice observed in Figures 3 and 6 further supports the notion that tree-based models inherently normalize feature importance during training.

In contrast, the pronounced sensitivity of Linear Regression to preprocessing strategies, as observed in Figures 2, 5, and 7, is consistent with theoretical expectations. Linear models assume linear relationships and equal feature contribution, making them particularly vulnerable to inappropriate scaling and encoding. The superior performance of One-Hot Encoding over Ordinal Encoding, especially in reducing RMSE and MAE (Figures 9 and 10), reinforces prior findings that artificial ordinal relationships can negatively affect regression accuracy.

From a practical perspective, these results suggest that robust regression pipelines should prioritize model selection over excessive tuning of scaling methods, particularly when ensemble models are employed. However, careful consideration of categorical encoding remains essential, especially in datasets with mixed feature types. For practitioners, this implies that One-Hot Encoding combined with ensemble regressors offers a reliable and high-performing baseline for regression tasks.

Despite these contributions, several limitations should be acknowledged. First, the study evaluates a fixed set of models, scalers, and encoders; alternative preprocessing methods such as target encoding or nonlinear transformations were not explored. Second, the experiments are conducted on a single dataset, which may limit generalizability across domains

with different feature distributions or cardinalities. Finally, hyperparameter optimization was not the primary focus of this study and may further influence relative pipeline performance.

In summary, this work demonstrates that while feature scaling has a limited effect on ensemble regressors, the interaction between encoding strategies and model choice plays a decisive role in regression performance. These findings provide empirical guidance for designing robust and efficient machine learning regression pipelines.

5. Conclusion

This study empirically investigated the impact of feature scaling and categorical encoding strategies on machine learning regression pipelines by systematically evaluating multiple preprocessing–model combinations. The results consistently show that ensemble-based models, namely Gradient Boosting and Random Forest, achieve superior predictive performance compared to Linear Regression across all evaluated configurations. These models demonstrate strong robustness to variations in feature scaling, indicating that numerical feature normalization plays a limited role once tree-based ensemble learners are employed.

In contrast, categorical encoding strategies exert a more pronounced influence on regression performance. One-Hot Encoding consistently yields higher mean R^2 values and lower RMSE and MAE compared to Ordinal Encoding, particularly for models sensitive to feature representation such as Linear Regression. These findings confirm that inappropriate encoding can introduce structural bias and degrade predictive accuracy, even when advanced learning algorithms are used. Consequently, encoding strategy selection should be treated as a critical design decision in regression pipelines.

Despite its contributions, this study has several limitations. The evaluation is conducted on a single dataset and focuses on a limited set of preprocessing techniques and regression models. Future research may extend this work by incorporating additional datasets from different domains, exploring advanced encoding methods such as target encoding, and integrating hyperparameter optimization to further examine preprocessing–model interactions. Nevertheless, this study provides practical and empirical insights that can support the development of robust, transparent, and high-performing machine learning regression pipelines.

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization G.A.T. and G.K.; Methodology, G.A.T. and G.K.; Software, G.A.T.; Validation, G.A.T.; Formal Analysis, G.A.T.; Investigation, G.K.; Resources, G.A.T.; Data Curation, G.K.; Writing Original Draft Preparation, G.A.T.; Writing Review and Editing, G.A.T. and G.K.; Visualization, G.K. 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] G. Erol, B. Uzbaş, C. Yücelbaş, and Ş. Yücelbaş, "Analyzing the Effect of Data Preprocessing Techniques Using Machine Learning Algorithms on the Diagnosis of COVID-19," *Concurr. Comput. Pract. Exp*., vol. 34, no. 28, pp. e7393, 2022, doi: 10.1002/cpe.7393.

[2] O. G. Emi-Johnson and K. J. Nkrumah, "Predicting 30-Day Hospital Readmission in Patients with Diabetes Using Machine Learning on Electronic Health Record Data," Cureus, vol. 17, no. 4, pp. e82437, 2025, doi: 10.7759/cureus.82437.

[3] A. Alagic, N. Zivic, E. Kadusic, D. Hamzic, N. Hadzajlic, M. Dizdarevic, and E. Selmanovic, "Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data," *Mach. Learn. Knowl. Extr*., vol. 6, no. 1, pp. 53–77, 2024, doi: 10.3390/make6010004.

[4] K. Parveen, "Enhanced Credit Scoring Prediction Using KNN-Z-Score Based Logistic Regression (KZ-LR) Algorithm," *J. Electr. Syst.*, vol. 20, no. 3, pp. 7230–7237, 2024, doi: 10.52783/jes.7419.

[5] P. Yasodha, "Data Preprocessing Methods for Machine Learning: An Empirical Comparison," *Int. J. Multidiscip. Res*., vol. 7, no. 3, pp. 1–7, 2025, doi: 10.36948/ijfmr.2025.v07i03.48569.

[6] S. Ramya, B. Kumaraswamy, V. Agarwal, and A. K. Jain, "A Comparative Study of Different Data Pre-Processing Methods for Machine Learning," *Int. J. Multidiscip. Res*., vol. 7, no. 4, pp. 1–8, 2025, doi: 10.36948/ijfmr.2025.v07i04.52920.

[7] H. Ouyang, L. Tang, J. Ma, and T. Pang, "Application of Hyperspectral Technology with Machine Learning for Brix Detection of Pastry Pears," *Plants*, vol. 13, no. 8, pp. 1163, 2024, doi: 10.3390/plants13081163.

[8] I. Fayyaz, G. G. M. N. Ali, and S. S. Khairunnesa, "Advanced Feature Engineering and Machine Learning Techniques for High Accurate Price Prediction of Heterogeneous Pre-Own Cars," *Vehicles*, vol. 7, no. 3, pp. 94, 2025, doi: 10.3390/vehicles7030094.

[9] B. Huang, L. Xue, J. Yang, C. Yin, K. Liao, M. Liu, J. Li, and M. Shang, "Regression Study on Fruit-setting Days of Purple Eggplant Fruit Based on in Situ VIS-NIRS and Attention Cycle Neural Network," *J. Food Sci*., vol. 90, no. 1, pp. e17593, 2025, doi: 10.1111/1750-3841.17593.

[10] V. Inturi, S. V Balaji, P. Gyanam, B. P. V Pragada, G. R. Sabareesh, and V. Pakrashi, "An Integrated Condition Monitoring Scheme for Health State Identification of a Multi-Stage Gearbox Through Hurst Exponent Estimates," *Struct. Heal. Monit*., vol. 22, no. 1, pp. 730–745, 2022, doi: 10.1177/14759217221092828.

[11] E. Tabane, E. Mnkandla, and Z. Wang, "Optimizing DNA Sequence Classification via a Deep Learning Hybrid of LSTM and CNN Architecture," *Appl. Sci*., vol. 15, no. 15, pp. 8225, 2025, doi: 10.3390/app15158225.

[12] M. D. Hossain, S. H. Scott, T. Cluff, and S. P. Dukelow, "The Use of Machine Learning and Deep Learning Techniques to Assess Proprioceptive Impairments of the Upper Limb After Stroke," *J. Neuroeng. Rehabil*., vol. 20, no. 1, pp. 1–18, 2023, doi: 10.1186/s12984-023-01140-9.

[13] R. Van, D. Alvarez, T. Mize, S. Gannavarapu, L. C. Reddy, F. Nasoz, and M. V. Han, "A Comparison of RNA-Seq Data Preprocessing Pipelines for Transcriptomic Predictions Across Independent Studies," *BMC Bioinformatics*, vol. 25, no. 1, pp. 1–22, 2024, doi: 10.1186/s12859-024-05801-x.

[14] Y. Sun, N. S. Nayani, Y. Xu, Z. Xu, J. Yang, and Y. Feng, "Rapid and Nondestructive Determination of Oil Content and Distribution of Potato Chips Using Hyperspectral Imaging and Chemometrics," *ACS Food Sci. Technol*., vol. 4, no. 6, pp. 1579–1588, 2024, doi: 10.1021/acsfoodscitech.4c00196.

[15] J. Ong, W. He, P. Maglanque, X. Jiang, L. M. Gillman, A. Vergis, and K. Hardy, "A Preprocessing Pipeline for Pupillometry Signal from Multimodal iMotion Data," *Sensors*, vol. 25, no. 15, pp. 4737, 2025, doi: 10.3390/s25154737.

[16] Q. Xiao, J. Zheng, J. Wen, F. Deng, R. Gu, L. Li, Y. He, and J. Yang, "Rapid Detection of Physicochemical Indicators of Tobacco Flavorings Using Fourier-Transform Near Infrared Spectroscopy with Chemometrics and Machine Learning," *ACS Omega*, vol. 10, no. 19, pp. 19714–19722, 2025, doi: 10.1021/acsomega.5c00225.

[17] L. J. Keevers and P. Jean-Richard-dit-Bressel, "Obtaining Artifact-Corrected Signals in Fiber Photometry via Isosbestic Signals, Robust Regression, and Calculations," *Neurophotonics*, vol. 12, no. 02, pp. 025003-1-025003–12, 2025, doi: 10.1117/1.nph.12.2.025003.

[18] F. Koopmans, K. W. Li, R. V Klaassen, and A. B. Smit, "MS-DAP Platform for Downstream Data Analysis of Label-Free Proteomics Uncovers Optimal Workflows in Benchmark Data Sets and Increased Sensitivity in Analysis of Alzheimer's Biomarker Data," *J. Proteome Res*., vol. 22, no. 2, pp. 374–386, 2022, doi: 10.1021/acs.jproteome.2c00513.

[19] A. Mansoori, M. Zeinalnezhad, and L. Nazarimanesh, "Optimization of Tree-Based Machine Learning Models to Predict the Length of Hospital Stay Using Genetic Algorithm," *J. Healthc. Eng*., vol. 2023, no. 1, pp. 1–14, 2023, doi: 10.1155/2023/9673395.

[20] B. Avanzi, G. Taylor, M. Wang, and B. Wong, "Machine Learning with High-Cardinality Categorical Features in Actuarial Applications," *Astin Bull.,* vol. 54, no. 2, pp. 213–238, 2024, doi: 10.1017/asb.2024.7.

[21] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized Target Encoding Outperforms Traditional Methods in Supervised Machine Learning with High Cardinality Features," *Comput. Stat*., vol. 37, no. 5, pp. 2671–2692, 2022, doi: 10.1007/s00180-022-01207-6.

[22] S. S. L. Parvathi, A. D. B, G. L. Kulkarni, S. Murugan, B. K. P. Vijayammal, and Neha, "Exploring Feature Relationships in Brain Stroke Data Using Polynomial Feature Transformation and Linear Regression Modeling," *J. Mach. Comput*., vol. 4, no. 4, pp. 1158–1169, 2024, doi: 10.53759/7669/jmc202404107.

[23] P. Cerda and G. Varoquaux, "Encoding High-Cardinality String Categorical Variables," *IEEE Trans. Knowl. Data Eng*., vol. 34, no. 3, pp. 1164–1176, 2022, doi: 10.1109/tkde.2020.2992529.

[24] S. Moseen and B. R. Kumar, "Hybrid Regression Approach for Prediction of Renewable Energy from Human Footstep Power Systems," *Int. J. Data Sci. IoT Manag. Syst*., vol. 4, no. 3, pp. 328–340, 2025, doi: 10.64751/ijdim.2025.v4.n3.pp328-340.

[25] R. A. M. Aljohani, "Exploring Football Player Salary Prediction Using Random Forest: Leveraging Player Demographics and Team Associations," *Int. J. Appl. Inf. Manag*., vol. 5, no. 4, pp. 203–213, 2025, doi: 10.47738/ijaim.v5i4.115.

[26] D. Zhang, "Task-Based Teaching Method in English Teaching from the Perspective of Big Data," *Int. J. Web-Based Learn. Teach. Technol*., vol. 20, no. 1, pp. 1–17, 2025, doi: 10.4018/ijwltt.381309.

[27] W. Jiang, "Key Selection Factors Influencing Animation Films from the Perspective of the Audience," *Mathematics*, vol. 12, no. 10, pp. 1–21, 2024, doi: 10.3390/math12101547.

[28] A. L. Aranha, L. L. B. Bernucci, and K. Vasconcelos, "Effects of Different Training Datasets on Machine Learning Models for Pavement Performance Prediction," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2677, no. 8, pp. 196–206, 2023, doi: 10.1177/03611981231155902.

[29] L. Meena and T. Velmurugan, "Optimizing Facial Expression Recognition Through Effective Preprocessing Techniques," *J. Comput. Commun.*, vol. 11, no. 12, pp. 86–101, 2023, doi: 10.4236/jcc.2023.1112006.

[30] S. S. Samaan and H. A. Jeiad, "Architecting a Machine Learning Pipeline for Online Traffic Classification in Software Defined Networking Using Spark," *IAES Int. J. Artif. Intell*., vol. 12, no. 2, pp. 861, 2023, doi: 10.11591/ijai.v12.i2.pp861-873.

[31] N. Maher and S. A. Yousif, "An Automated Machine Learning Model for Diagnosing Coronavirus Disease 2019 (COVID-19) Infection," *IAES Int. J. Artif. Intell.*, vol. 12, no. 3, pp. 1360, 2023, doi: 10.11591/ijai.v12.i3.pp1360-1369.

[32] P. E. Guillem, M. Zurdo-Tabernero, N. E. Iglesias, Á. Canal-Alonso, L. Durón Figueroa, G. Hernández, A. González-Arrieta, and F. de la Prieta, "Leveraging Transformers for Semi-Supervised Pathogenicity Prediction with Soft Labels," *J. Integr. Bioinform.*, vol. 22, no. 2, pp. 20240047, 2025, doi: 10.1515/jib-2024-0047.

[33] S. Grafberger, S. Guha, P. Groth, and S. Schelter, "Mlwhatif: What if You Could Stop Re-Implementing Your Machine Learning Pipeline Analyses Over and Over?," *Proc. VLDB Endow.*, vol. 16, no. 12, pp. 4002–4005, 2023, doi: 10.14778/3611540.3611606.