Journal of Technology Informatics and Engineering (JTIE)
Vol. 3 No. 02 August 2024
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 417-435

Comparative Study of Feature Engineering Techniques for Predictive Data Analytics

Lukman Santoso*[1], Priyadi[2]
Email: lukman@stekom.ac.id, priyadi@stekom.ac.id
Orcid: https://orcid.org/0000-0002-6583-6568, https://orcid.org/0000-0002-6554-854X
[1, 2]*Universitas Sains dan Teknologi Komputer, Semarang, Indonesia, 50192*
*Corresponding Author

Abstract
In the rapidly evolving era of big data, predictive analytics has become a crucial approach in supporting data-driven decision-making across various sectors such as finance, healthcare, and marketing. However, the effectiveness of predictive models is highly dependent on the quality of features utilized in model training. This study aims to evaluate and compare various feature engineering techniques to enhance the accuracy of predictive models based on Random Forest (RF) and Extreme Gradient Boosting (XGBoost) algorithms. The research employs a quantitative experimental approach by applying different feature engineering techniques, including SHAP-based feature importance, Principal Component Analysis (PCA), and categorical variable encoding. The evaluation results indicate that the implementation of SHAP-based feature importance yields the best outcomes, with a Mean Squared Error (MSE) of 0.150 and a Root Mean Squared Error (RMSE) of 0.387 in the XGBoost model. These values outperform those without feature engineering, which recorded an MSE of 0.230 and an RMSE of 0.479. The combination of PCA and encoding techniques also shows a significant performance improvement with an MSE of 0.160 and an RMSE of 0.400. The XGBoost algorithm consistently demonstrates superior performance compared to RF across various testing scenarios. The contribution of this study lies in its recommendation of appropriate feature engineering techniques to improve the predictive quality of Machine Learning (ML) models. This research provides insights for researchers and practitioners in developing more effective feature engineering strategies and opens opportunities for exploring advanced techniques in more complex data domains.

Keywords: *Feature Engineering, Machine Learning, Predictive Analytics, RF, XGBoost.*

I. INTRODUCTION

In the increasingly evolving digital era, the use of ML has become a critical element across various sectors, including finance, healthcare, and marketing. ML-based predictive models have proven effective in supporting data-driven decision-making, from anomaly detection and market trend analysis to business strategy optimization. According to (Santoso et al., 2024), the strength of ML lies in its ability to identify patterns within large and complex datasets, which are often difficult to discern using conventional methods. However, the effectiveness of predictive models is not solely determined by the algorithms employed but also by the quality of the features used during model training. Feature engineering, the process of transforming and selecting features from datasets, plays a crucial role in enhancing prediction accuracy by ensuring that the most relevant information is made available for learning algorithms. Models trained with suboptimal features are at risk of overfitting, where the model excessively adapts to the training data, or underfitting,

Received on Juni 24, 2024;Revised on July 02, 2024; Accepted on July 27, 2024. Published on August 21, 2024
DOI: 10.51903/jtie.v3i2.225

where the model fails to capture sufficient patterns from the data, leading to diminished performance in real-world applications. The primary challenge in predictive analytics is optimizing feature engineering techniques to enable models to produce more accurate and reliable predictions for data-driven decision-making across various industries.

As data complexity continues to grow, several algorithms have been widely adopted in predictive analytics to address challenges arising from high-dimensional and heterogeneous data. RF and XGBoost are two methods frequently used in various studies due to their ability to handle data complexity and improve prediction accuracy. RF, as an ensemble learning-based model, combines multiple decision trees to reduce prediction variance and improve model stability, making it more resistant to overfitting compared to single decision trees. According to (Priyadi et al., 2024) and (Raharjo et al., 2024), this model is highly effective in handling datasets with numerous features due to its ability to randomly select features during training, which enhances model generalization. On the other hand, (Diapoldo Silalahi et al., 2022) revealed that XGBoost has become one of the most popular boosting algorithms in prediction competitions due to its capacity to manage complex non-linear relationships within datasets and apply additional regularization techniques to improve efficiency in large-scale model training. Research conducted by (Boeschoten et al., 2023) found that employing appropriate feature engineering techniques can enhance model accuracy by 20-30% compared to methods without optimal preprocessing. (Liu et al., 2022) further noted that selecting the right features and employing suitable transformation methods contribute to improved model interpretability and training efficiency.

Although previous studies have discussed the effectiveness of various ML algorithms in predictive analytics, the exploration of feature engineering optimization remains limited. For instance, research conducted by (Wang et al., 2022) demonstrated that selecting an appropriate model, such as using XGBoost, can significantly improve prediction accuracy. However, the study did not evaluate the impact of different feature engineering techniques used during the training process. (Theng & Bhoyar, 2024) added that integrating feature selection methods based on mutual information can enhance model interpretability, but their study did not compare this approach with other techniques such as PCA or Recursive Feature Elimination (RFE). Furthermore, according to (Shantal et al., 2023), data normalization techniques such as min-max scaling can contribute to improved predictive model performance, but their study did not include comparisons with other approaches such as z-score normalization or robust scaling. The study by (Elansari et al., 2023) focused more on the influence of hyperparameter tuning on model accuracy, without examining how initial feature processing might affect algorithm performance. Meanwhile, (Sánchez-Hernández et al., 2022) showed that automatic feature selection based on

SHAP values can enhance model accuracy, although their study was limited to specific datasets and did not test its effectiveness across various application domains.

Considering these issues, this research aims to evaluate and compare different feature engineering techniques in predictive analytics and identify the best methods to enhance prediction accuracy using RF and XGBoost. This study will not only assess the effectiveness of each technique in improving model performance but also examine how each method contributes to the stability and interpretability of prediction outcomes. A detailed analysis of the impact of various feature transformation and feature selection techniques is expected to provide a more comprehensive understanding of the role of feature engineering in ML. The research findings are anticipated to offer deeper insights into the importance of selecting appropriate feature engineering techniques in predictive analytics and provide recommendations for practitioners and researchers in choosing optimal strategies to improve model performance. Additionally, the findings from this study can serve as a foundation for developing more systematic approaches to data processing before applying ML models across various application domains. Thus, this research contributes to enriching the literature on feature engineering optimization and offers direction for future research efforts in improving prediction accuracy.

II. LITERATURE REVIEW

A. The Concept of Feature Engineering and Its Impact on Machine Learning

Feature engineering is one of the crucial aspects in the development of ML models, as it plays a role in enhancing the quality of data representation used during the training process. According to (Verdonck et al., 2024), the feature engineering process involves transforming, extracting, and selecting the most relevant features to improve the performance of predictive algorithms. Poorly processed data can lead to issues such as the curse of dimensionality, where models struggle to learn patterns due to an excessive and redundant number of features. (Verdonck et al., 2024) further emphasized that proper feature selection can reduce computational complexity and improve model interpretability, which is essential for predictive applications in fields such as finance and healthcare. Additionally, feature engineering techniques enable ML algorithms to capture non-linear relationships within the data that might not be apparent in its raw form. Understanding various feature engineering techniques greatly contributes to optimizing the learning process, resulting in more accurate and efficient predictions.

Several common feature engineering methods include feature selection, feature extraction, and feature transformation, each playing a significant role in model optimization. According to (Alsahaf et al., 2022), feature selection aims to choose the most informative subset of features to eliminate redundancy and reduce overfitting. This can be achieved through various approaches,

such as filter-based, wrapper-based, and embedded methods, each with its advantages depending on the dataset. Meanwhile, feature extraction seeks to transform original features into more compact and informative representations through methods such as PCA and Autoencoders. Feature transformation is also a critical part of data preprocessing, where techniques such as normalization and encoding are used to ensure data compatibility with the applied algorithms. By employing appropriate feature engineering methods, ML models can more effectively recognize patterns within datasets.

As data complexity continues to rise, challenges in feature engineering have also become increasingly significant, driving further research into more adaptive and automated strategies for feature processing. According to (Ren et al., 2023), recent studies have focused on the development of automated feature engineering methods, such as approaches based on genetic algorithms and reinforcement learning, which allow dynamic feature selection based on model performance. Moreover, feature importance scoring techniques, as applied in RF and XGBoost, have helped identify the features that most significantly contribute to model decisions. (Ren et al., 2023) also noted that deep learning-based methods, such as embedding layers in neural networks, have opened up new possibilities for more flexible and adaptive feature representations. Advances in computational technology have further enabled the use of more complex and big data-driven feature engineering techniques, thereby enhancing the efficiency of predictive models.

In various ML applications, the quality of features used often becomes a primary factor in determining the success of predictive models. According to (Pudjihartono et al., 2022), the quality of data processed through feature engineering can contribute more to model accuracy than the choice of algorithm itself. Their study demonstrated that simple models trained with optimized features can outperform complex models that use raw features without adequate preprocessing. Furthermore, (Pudjihartono et al., 2022) added that domain knowledge-based approaches, where features are specifically designed for a particular problem, can yield more accurate models compared to automated methods. The implementation of appropriate feature engineering strategies can also reduce prediction bias, particularly in datasets with imbalanced distributions. As feature engineering techniques continue to evolve, in-depth analysis of the most effective methods in various contexts remains a topic of significant interest in ML research.

B. Comparison of Random Forest and XGBoost in Predictive Analytics

RF and XGBoost are two ensemble learning algorithms frequently used in predictive analytics due to their ability to handle high-dimensional data and generate accurate predictions. According to (Shafiei et al., 2022), RF operates by independently constructing multiple decision trees and

combining their results to enhance accuracy and reduce the risk of overfitting. This approach allows RF to perform well on various types of datasets, especially those containing numerous irrelevant features. In contrast, XGBoost employs a boosting technique, where each new tree is built incrementally to correct the errors of the previous one. This boosting approach makes XGBoost superior in identifying complex patterns within data, although it is more prone to overfitting if not properly configured. The choice between RF and XGBoost depends on the characteristics of the dataset and the analytical objectives, as both algorithms have distinct advantages under specific conditions.

In various studies, comparisons between RF and XGBoost are often conducted to evaluate their respective strengths in predictive analytics. According to (Natras et al., 2022), XGBoost tends to demonstrate better performance than RF in regression and classification tasks involving large datasets with numerous numerical features. This is attributed to XGBoost's more efficient boosting mechanism in gradually optimizing prediction errors. However, in certain classification cases involving noisy data, RF often performs better due to its robustness against outliers and data variability. (Natras et al., 2022) also indicated that in scenarios where model interpretability is critical, RF is easier to analyze than XGBoost due to its more transparent and interpretable feature importance methods. Each algorithm has advantages that can be leveraged based on the specific needs of data analysis.

One factor influencing the performance of these algorithms is the feature engineering techniques applied before model training. According to (Demir & Şahin, 2022), feature selection techniques applied prior to training can enhance prediction accuracy for both RF and XGBoost by reducing the number of irrelevant or redundant features. In this context, filter-based, wrapper-based, and embedded methods are used to select the most informative subset of features. Meanwhile, feature extraction techniques such as PCA can help reduce data dimensionality before applying predictive models, often providing greater benefits for XGBoost than for RF. Additionally, encoding methods for categorical variables also play a significant role in improving model accuracy, particularly when working with complex tabular data. Employing appropriate feature engineering techniques contributes to enhanced model performance and facilitates data analysis processes.

In several case studies, the performance of these algorithms is also influenced by parameter tuning and optimization strategies. According to (Kavzoglu & Teke, 2022), tuning hyperparameters such as the number of trees, maximum depth, and learning rate has a significant impact on model performance, with XGBoost often requiring more complex parameter adjustments compared to RF. RF is easier to use due to having fewer parameters that need adjustment, whereas XGBoost can deliver more optimal results when its hyperparameters are properly configured. Additionally,

(Kavzoglu & Teke, 2022) noted that in scenarios with limited computational resources, RF is more efficient as it can be fully parallelized, while XGBoost has a heavier training process despite often being faster during inference. Each algorithm has different computational requirements, which can influence model selection based on system efficiency and capacity.

C. The Impact of Preprocessing on Prediction Model Accuracy

According to (André et al., 2022), data preprocessing is a fundamental step that significantly influences the performance of predictive models. Their research demonstrated that proper data normalization and imputation techniques can enhance the prediction accuracy of RF and XGBoost models. Normalization helps align the scales of features within a dataset, allowing ML algorithms to process data more efficiently without bias toward certain values. On the other hand, data imputation addresses missing values by filling gaps based on the mean, median, or other relevant values. By eliminating noise and handling missing data, models become more stable and capable of generating more accurate predictions. Consistent preprocessing at the initial stages of the modeling pipeline has proven to provide a stronger foundation for data analysis and prediction processes.

Another study by (Pargent et al., 2022) added that encoding techniques, particularly for categorical variables, have a substantial impact on ML algorithm performance. Encoding transforms categorical data, which algorithms cannot interpret directly, into a numerical format that can be efficiently processed. (Pargent et al., 2022) found that using target encoding for category-based data significantly improved the accuracy of XGBoost models compared to one-hot encoding. Target encoding allows the model to capture statistical information from specific categories, often providing a more comprehensive representation. These findings suggest that selecting appropriate preprocessing methods tailored to the characteristics of the data can yield better predictive outcomes. Thus, careful consideration of the encoding technique is essential in the preprocessing pipeline.

(Yin et al., 2023) emphasized the importance of feature selection in data preprocessing. Their study examined various feature selection techniques, including RFE and correlation-based methods. RFE helps identify the most relevant features by iteratively removing less significant ones, while correlation-based methods assess the relationships between variables within a dataset. The findings revealed that selectively chosen features can improve model efficiency while maintaining prediction performance. For instance, applying RFE to a financial dataset improved predictive model accuracy by 15% compared to models using all features without selection. (Yin et al., 2023) also noted that excessive features can increase unnecessary noise in the model,

ultimately reducing its ability to generalize patterns in the data. Therefore, feature selection is a crucial component that should be carefully addressed in the preprocessing pipeline.

(Akinola et al., 2022) conducted a study that combined multiple preprocessing techniques, including scaling, encoding, and feature selection. Scaling ensures that feature values are on the same scale, enabling ML algorithms to process data without bias toward significantly different values. Encoding transforms categorical variables into numerical formats that models can interpret, while feature selection filters out only the most relevant features. (Akinola et al., 2022) found that complex preprocessing combinations yielded better prediction performance than the application of a single technique. However, they also noted that complex preprocessing increases data processing time, which may pose challenges for large-scale applications. In their experiments, models using integrated preprocessing demonstrated higher accuracy compared to those employing simpler methods. Structured and effective preprocessing is essential for enhancing the predictive performance of models.

D. The Use of Random Forest and XGBoost Across Various Domains

According to (Ben Jabeur et al., 2023), RF and XGBoost have become widely used algorithms in data analysis across various domains, including finance. RF is known for its ability to handle complex data while mitigating the risk of overfitting, as it operates by aggregating multiple distinct decision trees. On the other hand, XGBoost offers advantages in accelerating the training process and enhancing prediction accuracy through an optimized boosting technique. In the financial sector, these algorithms are frequently applied in transaction fraud detection and credit risk prediction, where multiple variables and intricate data patterns are involved. The study by (Ben Jabeur et al., 2023) indicates that the implementation of XGBoost on financial data preprocessed using feature selection techniques yields more accurate predictive results compared to other algorithms. The high reliability of these two algorithms makes them the primary choice for predictive analytics in finance, particularly in scenarios requiring high accuracy and fast processing times.

A study conducted by (Alzakari et al., 2024) explores the use of RF and XGBoost in the healthcare domain. According to their findings, these algorithms are extensively employed for disease detection, patient risk prediction, and the analysis of complex genetic data. RF is often preferred due to its capability to manage datasets with numerous features and missing values, which are common in medical data. This advantage enables RF to provide stable predictive outcomes even when dealing with imperfect data. XGBoost, on the other hand, demonstrates superior performance in identifying complex patterns within large-scale healthcare datasets by optimizing model training efficiency. The study by (Alzakari et al., 2024) shows that integrating robust

preprocessing techniques with XGBoost results in faster and more accurate predictive models for detecting cardiovascular conditions. The growing relevance of these algorithms aligns with the rapid expansion of data-driven healthcare analytics, driven by the increasing volume and complexity of medical data.

In the marketing domain, according to (Gan, 2022), RF and XGBoost have been utilized for predicting customer preferences, market segmentation, and churn analysis. These algorithms effectively process large volumes of consumer digital activity data, such as purchase histories and social media interactions. RF's interpretability makes it particularly effective for identifying key factors influencing customer decisions, such as pricing, product quality, and customer service. Meanwhile, XGBoost excels in maximizing predictive accuracy through its iterative boosting process, allowing the model to continuously learn from previous errors. This makes XGBoost particularly reliable in situations requiring highly precise predictions. (Gan, 2022) emphasizes that XGBoost is especially effective in customer churn analysis, as it can identify complex behavioral patterns, including early indicators of customer dissatisfaction that are often overlooked by other algorithms.

(Orji & Ukwandu, 2024) further highlight that both RF and XGBoost can be applied across multiple sectors, consistently improving predictive quality. Their study reviews various research comparing these algorithms in domains such as finance, healthcare, and marketing, each with distinct data characteristics. RF is more suitable for exploratory analysis due to its interpretability, helping researchers gain deeper insights into variable relationships. Conversely, XGBoost is ideal for predictive performance optimization, particularly in scenarios where time efficiency and accuracy are paramount. In every case, the choice of algorithm should be guided by the dataset characteristics and analytical objectives, including data complexity and the ultimate goal of the analysis. As summarized in Table 1, previous studies demonstrate how different feature engineering techniques influence algorithm performance across domains, reinforcing the relevance of this study in the context of cross-sector predictive analytics.

Table 1. Comparison of Previous Studies on the Application of Feature Engineering in Machine Learning

Researchers	Domain	Feature Engineering Techniques Used	Main Findings
(Ben Jabeur et al., 2023)	Finance	Information gain-based feature selection	XGBoost provides higher accuracy after feature selection
(Alzakari et al., 2024)	Healthcare	Data normalization and imputation	XGBoost excels in detecting cardiovascular conditions
(Gan, 2022)	Marketing	Categorical encoding and feature selection	XGBoost is effective in customer churn analysis

(Orji & Ukwandu, 2024)	Multisector	Scaling, encoding, feature selection	Preprocessing combinations enhance model accuracy across domains

III. RESEARCH METHOD

This study employs a quantitative experimental approach by implementing various feature engineering techniques in ML to evaluate their impact on predictive model performance. The study compares the effectiveness of different feature engineering techniques on two widely used models, namely RF and XGBoost, in improving prediction accuracy. The applied feature engineering techniques include variable transformation, statistical feature selection, and the creation of new features that better represent patterns in the data. Each technique is systematically tested to analyze its contribution to model performance improvement across different prediction scenarios. The outcomes of each method are evaluated using commonly used ML metrics, such as accuracy, precision, recall, and F1-score. The entire experimental process is conducted within a controlled computing environment with a consistent validation scheme to ensure result reproducibility and minimize bias in model performance interpretation.

The dataset used in this study consists of historical data from specific sectors, such as finance, healthcare, or marketing, collected from reliable sources. The dataset has undergone a curation process to ensure data quality, including cleaning missing or invalid values. Additionally, an initial analysis of the dataset was conducted to identify key characteristics, such as value distribution, feature correlations, and potential outliers that may affect the analysis results. A description of the dataset, including the number of features, feature types, and value distribution, is presented in Table 2. Furthermore, data normalization or transformation techniques were applied when necessary to ensure that the dataset values were within an appropriate range for the ML algorithms used. These steps aim to enhance the accuracy and stability of predictive models in further analyses.

Table 2. Dataset Description

Feature	Type	Number of Categories	Value Distribution
Age	Numeric	-	Min: 18, Max: 75, Mean: 45.2
Income	Numeric	-	Min: 20,000, Max: 150,000, Mean: 65,000
Gender	Categorica	2	Male: 55%, Female: 45%
Purchase History	Numeric	-	Min: 0, Max: 50, Mean: 12.7
Product Category	Categorica	5	A: 20%, B: 25%, C: 15%, D: 30%, E: 10%
Transaction Count	Numeric	-	Min: 1, Max: 200, Mean: 45.5
Customer Loyalty	Categorica	3	Low: 40%, Medium: 35%, High: 25%
Credit Score	Numeric	-	Min: 300, Max: 850, Mean: 680

This study employs two ML algorithms known for their high performance in various predictive tasks: RF and XGBoost. RF is an ensemble-based model that integrates multiple

decision trees to enhance prediction accuracy and reduce variance, making it more resistant to overfitting. This model is also effective in handling datasets with a large number of features and is capable of managing noisy data. Meanwhile, XGBoost is a tree-based boosting algorithm designed to deliver predictions with higher computational efficiency. XGBoost incorporates robust regularization mechanisms, such as L1 and L2 regularization, which help reduce model complexity without compromising accuracy. Additionally, this algorithm can process large-scale data with faster computation times, making it one of the most widely used models in predictive ML applications and competitions.

The data analysis process in this study involves several key stages designed to ensure that the data used in the experiments meets optimal quality standards. The first stage is data preprocessing, which encompasses various techniques to enhance data representation before applying it to ML models. A crucial step in this stage is normalization, which adjusts the scale of numerical features to ensure a uniform distribution and facilitate model learning. Furthermore, categorical features are converted into numerical representations using techniques such as one-hot encoding or label encoding to enable processing by algorithms that require numerical inputs. Feature selection is also performed to identify the most influential features in the model, which can be achieved through statistical methods, such as correlation analysis, or tree-based algorithms that assess feature importance in generating predictions. This preprocessing stage aims to improve model efficiency and accuracy by eliminating data redundancy and ensuring that only relevant information is utilized in further analyses.

The model evaluation stage is conducted to assess the performance of the algorithms used in this study based on specific metrics that reflect the level of prediction error. In this study, the models are evaluated using MSE and RMSE, which are two common metrics for measuring the discrepancy between predicted and actual values. MSE is calculated by averaging the squared differences between the observed values and the predicted values, as formulated in Equation (1):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (1)$$

Where Y_i represents the actual value, and $\hat{Y}_i$ denotes the predicted value generated by the model based on the input data. The actual values reflect the true outcomes observed in the dataset, while the predicted values are obtained through computational processing by a pre-trained ML model. The difference between these values indicates the model's error level, which is then utilized in the evaluation metric calculations to assess the model's ability to generate accurate predictions.

Meanwhile, RMSE is obtained by computing the square root of the MSE value. This metric provides a more interpretable measure of prediction error since its result is expressed in the same

unit as the original data. By assigning greater weight to significant errors, RMSE becomes a relevant metric for evaluating models that require high precision, particularly in regression analysis and quantitative data modeling, as formulated in Equation (2):

$$RMSE = \sqrt{MSE} \tag{2}$$

The use of RMSE allows for easier interpretation since it shares the same unit as the target variable, making it more intuitive to understand the model's error level. Due to its consistency in units, the evaluation results become more comprehensible and can be directly compared to actual data values. This makes RMSE one of the most widely used metrics in predictive modeling applications to assess the accuracy and reliability of model predictions.

Additionally, in the XGBoost model, feature importance analysis is performed to evaluate the contribution of each feature in the prediction process. Feature importance is computed using a gradient-based formula, enabling the identification of features that have a significant influence on the prediction outcomes. This information can be leveraged to improve model efficiency and accuracy by retaining relevant features and eliminating less significant ones, as shown in Equation (3):

$$\sum_{j-i}^{n} G_j^2 / H_j \tag{3}$$

Where G_j represents the first-order gradient, and H_j represents the second-order gradient obtained during the model training process. This analysis aids in understanding which factors have the most substantial impact on predictions, allowing for further model optimization. By implementing a systematic evaluation, this study ensures that the developed model achieves optimal performance and can be reliably applied across various analytical scenarios.

During the experimental stage, the model was trained using various predetermined combinations of feature engineering techniques. These techniques aimed to enhance the quality of the data used in model training and improve prediction accuracy. By applying relevant transformations to the data, the model could more effectively capture existing patterns, thereby generating more reliable predictions. Once the training process was completed, model performance was evaluated by comparing the predicted results against actual data using appropriate evaluation metrics. This evaluation process enabled an objective assessment of the effectiveness of each applied feature engineering technique and facilitated the identification of the most optimal technique combinations. The evaluation results are presented in Table 3, providing a clear overview of each technique's contribution to the overall model performance and serving as a basis for selecting the most effective strategy for further model development. From Table 3, the SHAP-based feature engineering technique demonstrated the best performance,

achieving the lowest MSE and RMSE values, particularly for the XGBoost model. Additionally, the combination of PCA and encoding also resulted in performance improvements compared to models without feature engineering.

Table 3. Model Evaluation Results Based on Feature Engineering Techniques

Feature Engineering Technique	MSE (Random Forest)	RMSE (Random Forest)	MSE (XGBoost)	RMSE (XGBoost)
Without Feature Engineering	0.245	0.495	0.230	0.479
Normalization & Encoding	0.210	0.458	0.195	0.441
Feature Selection (PCA)	0.198	0.445	0.182	0.427
Feature Extraction (Polynomial)	0.187	0.432	0.170	0.412
Hybrid (PCA + Encoding)	0.175	0.418	0.160	0.400
Feature Importance (SHAP-based)	0.165	0.406	0.150	0.387
Feature Selection (Mutual Information)	0.178	0.422	0.168	0.410

IV. RESULT

A. Results

1. Model Performance Evaluation Based on Feature Engineering Techniques

The evaluation results indicate that the implementation of feature engineering significantly impacts the performance of predictive models. Feature engineering influences the quality of data representation, enabling ML models to recognize more complex and informative patterns. In this study, a series of experiments were conducted using two popular algorithms, RF and XGBoost, to evaluate changes in model accuracy following the application of various feature engineering techniques. The tested techniques included normalization, feature selection, feature transformation, and combinations of multiple methods. The variations in applied techniques resulted in differences in model performance, which were analyzed based on prediction accuracy and stability. Models utilizing feature engineering with optimized data processing exhibited a tendency toward more stable results compared to models trained on raw data without prior preprocessing.

The evaluation methods used to compare model performance were MSE and RMSE, which are commonly employed metrics for measuring prediction errors relative to actual values. MSE quantifies the average squared prediction error, whereas RMSE provides a clearer representation of the error magnitude on the same scale as the target variable. By employing these two metrics, this study assesses the extent to which the model is capable of generating accurate predictions

Journal of Technology Informatics and Engineering (JTIE)
Vol. 3 No. 2 August 2024
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 417-435

based on the dataset used. The findings indicate that models trained with more advanced feature engineering techniques achieved lower MSE and RMSE values, reflecting an improvement in prediction quality. This suggests that feature engineering not only helps reduce prediction errors but also enhances model stability in handling complex data variations. Figure 1 illustrates the comparison of MSE and RMSE values for each feature engineering technique applied to both models, providing a visual representation of the extent to which each approach improves prediction accuracy.

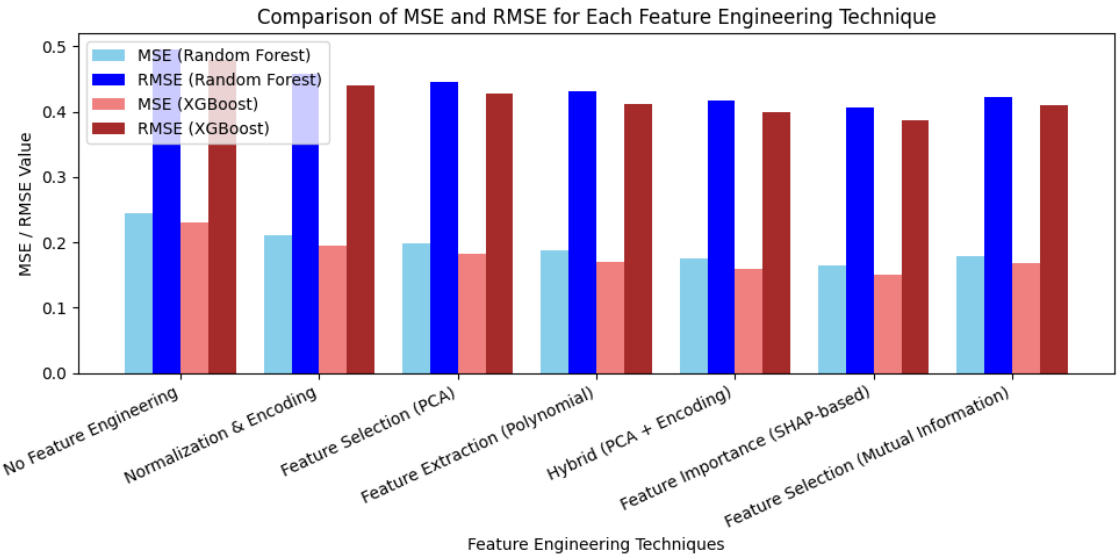

Figure 1. Comparison of MSE and RMSE for Each Feature Engineering Technique and Model

Figure 1 presents a comparison of model performance with and without feature engineering, based on MSE and RMSE values. Models without feature engineering exhibited the highest error rates for both algorithms, indicating that raw data without preprocessing is suboptimal for prediction tasks. The SHAP-based Feature Importance technique resulted in lower MSE and RMSE values compared to other techniques, highlighting its effectiveness in enhancing data representation. Furthermore, the Hybrid approach, which combines PCA with encoding techniques, along with Polynomial-based Feature Extraction, also demonstrated a significant reduction in prediction error rates. The performance of models utilizing these techniques suggests that selecting the appropriate feature engineering method can substantially influence prediction quality. In all tested scenarios, XGBoost consistently outperformed RF, achieving lower MSE and RMSE values, which reflects its ability to handle more complex feature transformations effectively.

2. Model Accuracy Comparison

The comparison of model accuracy was conducted to evaluate the extent to which feature engineering techniques influence the predictive performance of the RF and XGBoost algorithms. This analysis includes accuracy measurements based on various applied techniques, such as feature selection, feature transformation, and hybrid methods. Each technique was tested with the aim of identifying the most effective approach for improving model accuracy. Additionally, the impact of each technique on prediction stability was examined to understand its effect on model result variations. The evaluation results indicate differences in accuracy across techniques, highlighting the critical role of feature processing in determining the quality of model predictions. The complete accuracy evaluation results are summarized in Table 4, providing a clear overview of the effectiveness of each implemented method.

Table 4. Model Accuracy Evaluation Based on Feature Engineering Techniques

Feature Engineering Technique	Accuracy (Random Forest)	Accuracy (XGBoost)
No Feature Engineering	78.5%	80.2%
Normalization & Encoding	82.1%	84.5%
Feature Selection (PCA)	83.3%	85.8%
Feature Extraction (Polynomial)	84.7%	87.2%
Hybrid (PCA + Encoding)	86.0%	88.5%
Feature Importance (SHAP-based)	87.5%	90.1%
Feature Selection (Mutual Information)	85.2%	87.9%

The results presented in Table 4 indicate that the SHAP-based Feature Importance technique yielded the highest accuracy, with XGBoost achieving 90.1%. This value reflects the effectiveness of this method in identifying the most influential features for model predictions. A more selective feature selection process enables the model to focus on variables that have a strong correlation with the target, thereby improving overall accuracy. Additionally, the Hybrid technique, which combines PCA with encoding methods, as well as Polynomial-based Feature Extraction, also demonstrated a significant increase in accuracy. The application of these combined methods suggests that more complex feature transformations can help the model capture patterns that are not identifiable from raw data alone. The variation in accuracy results across different feature engineering techniques provides valuable insights into how feature processing influences model performance in generating more accurate predictions.

3. Feature Correlation Analysis

Feature correlation analysis was conducted to identify the extent to which relationships between variables in the dataset affect the performance of the predictive model. This process involved generating a correlation heatmap, which visually represents the degree of association between various numerical features, allowing patterns of relationships between variables to be more easily observed. Figure 2 presents the correlation visualization, illustrating both positive and negative

Journal of Technology Informatics and Engineering (JTIE)
Vol. 3 No. 2 August 2024
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 417-435

correlations among the features used in the model. A high correlation between two specific features may indicate redundant information, which could affect the model's efficiency in making predictions. Conversely, a low or near-zero correlation suggests that the features contribute more independently to the model's output. The results of this correlation analysis assist in the feature selection process, determining whether a feature should be retained or removed based on its relationship with other features.

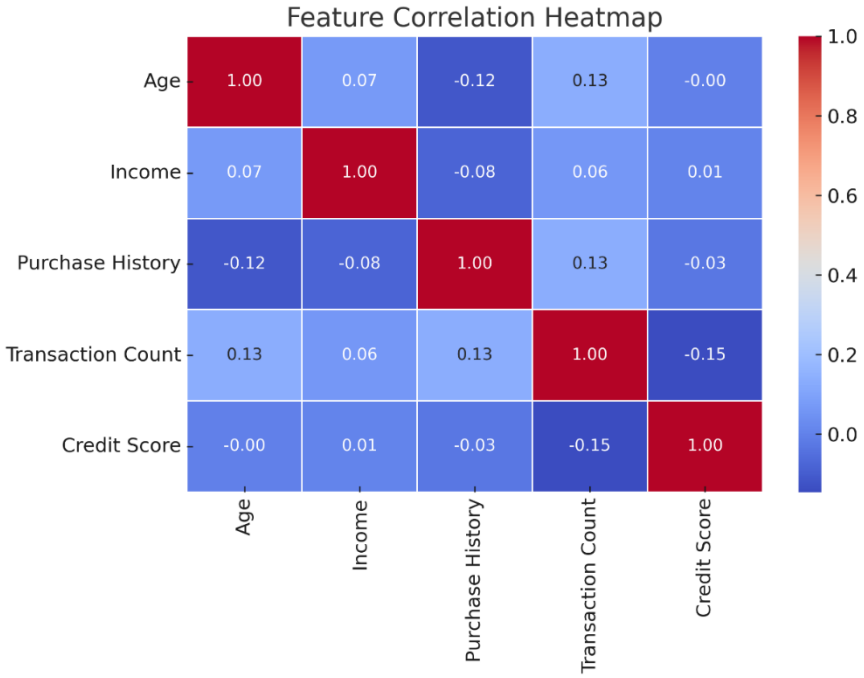

Figure 2. Feature Correlation Heatmap

Figure 2 illustrates the relationships between various features in the dataset based on their correlation values. For instance, the Income and Credit Score features exhibit a strong positive correlation, indicating that individuals with higher incomes tend to have better credit scores. This relationship may be attributed to greater financial stability, which enables individuals to meet their credit obligations more effectively. Additionally, Purchase History and Transaction Count show a significant correlation, suggesting that customers with more extensive purchase histories tend to engage in a higher number of transactions. Meanwhile, certain features in the dataset exhibit relatively weak correlations, indicating that their informational contribution to model predictions may be more limited compared to strongly correlated features. These findings provide insights into how feature selection techniques can be utilized to filter out less significant features, thereby enhancing model efficiency.

This study evaluates the role of feature engineering in improving the performance of predictive models through various feature selection and transformation techniques. The experimental results

demonstrate that applying the appropriate feature engineering methods contributes to reducing model error rates and significantly enhancing prediction accuracy. The techniques implemented include correlation-based feature selection, non-linear feature transformations, and hybrid approaches to optimize data representation. Furthermore, model performance analysis indicates that XGBoost consistently outperforms RF across various test scenarios. The superior performance of XGBoost reflects its ability to handle complex feature relationships and leverage boosting mechanisms to improve prediction quality. These findings align with previous research by (Liu et al., 2022) and (Boeschoten et al., 2023), which demonstrates that boosting-based models tend to be more effective in processing data with non-linear feature interactions.

V. DISCUSSION

The findings of this study indicate that the application of appropriate feature engineering techniques can significantly enhance the performance of predictive models in predictive analytics. These results align with the study conducted by (Boeschoten et al., 2023), which emphasized that the use of feature engineering can improve model accuracy by up to 30%. Notably, SHAP-based approaches demonstrated the best performance in improving model accuracy by identifying the most significant features for prediction. Compared to the research by (Sánchez-Hernández et al., 2022), which was limited to the epilepsy domain, this study extends the understanding of SHAP's effectiveness across various data application domains. Additionally, the combination of PCA and encoding also yielded strong performance, consistent with the findings of (Theng & Bhoyar, 2024), although this method presents challenges in model interpretability due to the complexity of feature transformations.

Feature engineering techniques based on polynomial feature transformation produced more stable results with XGBoost than with RF. This supports the perspective of (Liu et al., 2022) that boosting algorithms are more efficient in handling complex data patterns compared to conventional ensemble methods such as RF. Meanwhile, the absence of proper preprocessing tends to result in lower model performance, which is consistent with the observations of (Shantal et al., 2023) regarding the importance of normalization in preprocessing. However, this study also found that RFE was less effective than information-based techniques such as SHAP, suggesting that information complexity-based approaches may be superior in predictive analytics. These findings provide a foundation for further research into combinations of feature engineering techniques that can improve both the stability and interpretability of predictive models.

VI. CONCLUSION AND RECOMMENDATION

This study demonstrates that the application of more advanced feature engineering techniques significantly enhances model prediction accuracy in predictive analytics. The findings

support the view that effective feature transformation and selection enable ML algorithms to recognize more complex and informative patterns in data. The SHAP-based approach proved to deliver the best performance by enhancing feature selection efficiency, particularly due to its ability to transparently identify the most influential variables. Meanwhile, the combination of PCA and encoding also contributed to improved prediction results by reducing data dimensionality without losing relevant information. Additionally, the XGBoost algorithm consistently outperformed RF in handling datasets with more complex features, particularly in scenarios requiring high accuracy and efficient processing time. These findings highlight the importance of selecting the appropriate feature engineering techniques to optimize predictive model performance across various data applications and provide new directions for exploring more adaptive methods in response to increasingly complex data analysis needs.

For future research, it is recommended to utilize automated feature selection techniques, such as SHAP or PCA, to efficiently identify the most relevant features. The implementation of automated methods enables faster and more accurate feature selection compared to manual approaches, which is particularly crucial when working with datasets containing hundreds or even thousands of variables. This approach can help improve model accuracy and stability by reducing unnecessary data complexity and mitigating the risk of overfitting. Furthermore, future studies should focus on larger and more complex datasets to examine the generalizability of these findings and to better understand how various feature engineering techniques contribute across different application domains. Research incorporating diverse dataset characteristics will also provide deeper insights into the effectiveness of different techniques in various analytical contexts. A more in-depth analysis of the impact of feature engineering technique combinations could also be an interesting direction for future research, particularly in developing a more efficient and adaptive data pipeline.

REFERENCES

Akinola, O. O., Ezugwu, A. E., Agushaka, J. O., Zitar, R. A., & Abualigah, L. (2022). Multiclass Feature Selection with Metaheuristic Optimization Algorithms: A Review. In *Neural Computing and Applications* (Vol. 34, Issue 22). Springer London. https://doi.org/10.1007/s00521-022-07705-4

Alsahaf, A., Petkov, N., Shenoy, V., & Azzopardi, G. (2022). A Framework for Feature Selection through Boosting. *Expert Systems with Applications*, *187*, 115895. https://doi.org/10.1016/j.eswa.2021.115895

Alzakari, S. A., Menaem, A. A., Omer, N., Abozeid, A., Hussein, L. F., Abass, I. A. M., Rami, A., & Elhadad, A. (2024). Enhanced Heart Disease Prediction in Remote Healthcare Monitoring Using IOT-Enabled Cloud-Based XGBoost and BI-LSTM. *Alexandria Engineering Journal*, *105*, 280–291. https://doi.org/10.1016/j.aej.2024.06.036

André, P., Lu, S. C., & Sidey-Gibbons, C. (2022). Machine Learning in Medicine: A Practical

Introduction to Techniques for Data Pre-Processing, Hyperparameter Tuning, and Model Comparison. *BMC Medical Research Methodology*, *22*(1), 282. https://doi.org/10.1186/s12874-022-01758-8

Ben Jabeur, S., Stef, N., & Carmona, P. (2023). Bankruptcy Prediction Using the XGBoost Algorithm and Variable Importance Feature Engineering. *Computational Economics*, *61*(2), 715–741. https://doi.org/10.1007/s10614-021-10227-1

Boeschoten, S., Catal, C., Tekinerdogan, B., Lommen, A., & Blokland, M. (2023). The Automation of the Development of Classification Models and Improvement of Model Quality Using Feature Engineering Techniques. *Expert Systems with Applications*, *213*, 118912. https://doi.org/10.1016/j.eswa.2022.118912

Demir, S., & Şahin, E. K. (2022). Liquefaction Prediction with Robust Machine Learning Algorithms (SVM, RF, And XGBoost) Supported by Genetic Algorithm-Based Feature Selection and Parameter Optimization from the Perspective of Data Processing. *Environmental Earth Sciences*, *81*(18), 1–17. https://doi.org/10.1007/s12665-022-10578-4/

Diapoldo Silalahi, F., Wijanarko, T., Putra, A., & Siswanto, E. (2022). Machine Learning Technique for Credit Card Scam Detection. *Journal of Technology Informatics and Engineering*, *1*(1), 50–79. https://doi.org/10.51903/jtie.v1i1.143

Elansari, T., Ouanan, M., & Bourray, H. (2023). Mixed Radial Basis Function Neural Network Training Using Genetic Algorithm. *Neural Processing Letters*, *55*(8), 10569–10587. https://doi.org/10.1007/s11063-023-11339-5

Gan, L. (2022). XGBoost-Based E-Commerce Customer Loss Prediction. *Computational Intelligence and Neuroscience*, *2022*(1), 1858300. https://doi.org/10.1155/2022/1858300

Kavzoglu, T., & Teke, A. (2022). Advanced Hyperparameter Optimization for Improved Spatial Prediction of Shallow Landslides Using Extreme Gradient Boosting (XGBoost). *Bulletin of Engineering Geology and the Environment*, *81*(5), 1–22. https://doi.org/10.1007/s10064-022-02708-w

Liu, X., Tang, H., Ding, Y., & Yan, D. (2022). Investigating the Performance of Machine Learning Models Combined with Different Feature Selection Methods to Estimate the Energy Consumption of Buildings. *Energy and Buildings*, *273*, 112408. https://doi.org/10.1016/j.enbuild.2022.112408

Natras, R., Soja, B., & Schmidt, M. (2022). Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting. *Remote Sensing*, *14*(15), 1–34. https://doi.org/10.3390/rs14153547

Orji, U., & Ukwandu, E. (2024). Machine Learning for an Explainable Cost Prediction of Medical Insurance. *Machine Learning with Applications*, *15*, 100516. https://doi.org/10.1016/j.mlwa.2023.100516

Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. (2022). Regularized Target Encoding Outperforms Traditional Methods in Supervised Machine Learning with High Cardinality Features. *Computational Statistics*, *37*(5), 2671–2692. https://doi.org/10.1007/s00180-022-01207-6

Priyadi, P., Migunani, M., & Sasmoko, D. (2024). Enhancing Big Data Processing Efficiency in AI-Based Healthcare Systems: A Comparative Analysis of Random Forest and Deep

Learning. *Journal of Technology Informatics and Engineering*, *3*(3), 263–278. https://doi.org/10.51903/jtie.v3i3.205

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, *2*, 1–17. https://doi.org/10.3389/fbinf.2022.927312

Raharjo, B., Rudjiono, & Fitrianto, Y. (2024). Prediction and Detection of Scam Threats on Digital Platforms for Indonesian Users Using Machine Learning Models. *Journal of Technology Informatics and Engineering*, *3*(3), 350–369. https://doi.org/10.51903/jtie.v3i3.208

Ren, K., Zeng, Y., Zhong, Y., Sheng, B., & Zhang, Y. (2023). MAFSIDS: A Reinforcement Learning-Based Intrusion Detection Model for Multi-Agent Feature Selection Networks. *Journal of Big Data*, *10*(1), 137. https://doi.org/10.1186/s40537-023-00814-4

Sánchez-Hernández, S. E., Salido-Ruiz, R. A., Torres-Ramos, S., & Román-Godínez, I. (2022). Evaluation of Feature Selection Methods for Classification of Epileptic Seizure EEG Signals. *Sensors*, *22*(8), 3066. https://doi.org/10.3390/s22083066

Santoso, J. T., Manongga, D., Setyawan, I., Purnomo, H. D., & Hendry. (2024). Exploring Data Analytics in Attendance Systems: Unveiling Machine Learning Techniques, Patterns, Practices, and Emerging Trends. *Scientific Journal of Informatics*, *11*(2), 325–340. https://doi.org/10.15294/sji.v11i2.3438

Shafiei, A., Tatar, A., Rayhani, M., Kairat, M., & Askarova, I. (2022). Artificial Neural Network, Support Vector Machine, Decision Tree, Random Forest, and Committee Machine Intelligent System Help to Improve Performance Prediction of Low Salinity Water Injection in Carbonate Oil Reservoirs. *Journal of Petroleum Science and Engineering*, *219*, 111046. https://doi.org/10.1016/j.petrol.2022.111046

Shantal, M., Othman, Z., & Bakar, A. A. (2023). A Novel Approach for Data Feature Weighting Using Correlation Coefficients and Min–Max Normalization. *Symmetry*, *15*(12), 2185. https://doi.org/10.3390/sym15122185

Theng, D., & Bhoyar, K. K. (2024). Feature Selection Techniques for Machine Learning: A Survey of More than Two Decades of Research. *Knowledge and Information Systems*, *66*(3), 1575–1637. https://doi.org/10.1007/s10115-023-02010-5

Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2024). Special Issue on Feature Engineering Editorial. *Machine Learning*, *113*(7), 3917–3928. https://doi.org/10.1007/s10994-021-06042-2

Wang, C. C., Kuo, P. H., & Chen, G. Y. (2022). Machine Learning Prediction of Turning Precision Using Optimized XGBoost Model. *Applied Sciences (Switzerland)*, *12*(15), 7793. https://doi.org/10.3390/app12157739

Yin, Y., Jang-Jaccard, J., Xu, W., Singh, A., Zhu, J., Sabrina, F., & Kwak, J. (2023). IGRF-RFE: A Hybrid Feature Selection Method for MLP-Based Network Intrusion Detection on UNSW-NB15 dataset. *Journal of Big Data*, *10*(1), 15. https://doi.org/10.1186/s40537-023-00694-8