Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

Predicting Diabetes with Machine Learning: Evaluating Tree-Based and Ensemble Models with Custom Metrics and Statistical Validation

Gregorius Airlangga*

Engineering Faculty, Information System Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Email: [1]gregorius.airlangga@atmajaya.ac.id
Correspondence Author Email: gregorius.airlangga@atmajaya.ac.id
Submitted: **07/12/2024**; Accepted: **24/12/2024**; Published**: 25/12/2024**

Abstract−Diabetes diagnosis remains a critical challenge in healthcare, with significant clinical implications arising from false negatives that delay treatment. This study aims to address this challenge by evaluating the predictive performance of machine learning models on the Pima Indians Diabetes Dataset to enhance diagnostic accuracy while prioritizing recall, given its importance in minimizing undetected cases. Seven models, including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, Stacking Classifier, and Voting Classifier, were assessed using a 10-fold cross-validation strategy to ensure robust performance evaluation. Standard metrics such as accuracy, precision, recall, F1 score, and ROC AUC were utilized alongside a custom metric specifically designed to align evaluation with clinical priorities. Results indicated that LightGBM and Random Forest were the top-performing individual models, with ensemble methods like the Stacking Classifier leveraging complementary model strengths to achieve robust results. Statistical validation via the Friedman test (test statistic: 22.77, p-value: 0.00088) confirmed significant differences among models, though pairwise comparisons using the Wilcoxon signed-rank test showed no significant difference between top models. These findings underscore the efficacy of tree-based and ensemble methods in addressing diagnostic challenges and highlight the importance of aligning evaluation metrics with clinical priorities. Future research should explore hybrid approaches and larger datasets to further enhance predictive performance and generalizability in real-world healthcare applications.

Keywords: Machine Learning; Diabetes Prediction; Ensemble Models; Statistical Validation; Custom Metric Design

1. INTRODUCTION

Diabetes mellitus remains a pressing global health challenge, with its prevalence steadily increasing worldwide [1][3]. According to the International Diabetes Federation, over 463 million adults were living with diabetes in 2019, a figure projected to escalate to 700 million by 2045 [4][6]. This chronic condition, if left undiagnosed or poorly managed, can lead to severe complications, including cardiovascular disease, kidney failure, and neuropathy, resulting in significant socioeconomic and health system burdens [7]. Early and accurate diagnosis is, therefore, critical to mitigating these adverse outcomes [8]. However, traditional diagnostic approaches are often constrained by variability in clinical settings and a reliance on subjective interpretations [9]. In this context, machine learning (ML) emerges as a transformative approach capable of providing robust, scalable, and precise diagnostic solutions [10]. The application of ML models to diabetes diagnosis has gained significant traction in recent years, fueled by the increasing availability of high-quality medical datasets [11]. The Pima Indians Diabetes Dataset, curated by the National Institute of Diabetes and Digestive and Kidney Diseases, serves as an important benchmark for developing and testing ML models [12]. This dataset includes diagnostic measurements from female patients aged 21 and above of Pima Indian heritage and contains predictive variables such as the number of pregnancies, BMI, insulin levels, and age. Despite its simplicity, the dataset poses challenges such as class imbalance and feature correlations, which demand sophisticated preprocessing and modeling strategies.

Previous research has explored various ML techniques to predict diabetes onset. Early studies applied traditional models like Logistic Regression and Decision Trees, which provided foundational insights but were limited in handling complex interactions within data[13]. For instance, [14] demonstrated the utility of statistical models in forecasting diabetes onset using the Pima dataset. However, as computational capabilities improved, ensemble methods such as Random Forest and Gradient Boosting began outperforming these traditional models due to their ability to capture non-linear relationships and interactions. Studies by [15] showcased the superiority of Random Forest and XGBoost in achieving high accuracy and recall for diabetes prediction tasks. These studies emphasized the importance of hyperparameter tuning and feature selection to optimize model performance. Recent advancements have also highlighted the role of hybrid and ensemble techniques in pushing the boundaries of predictive accuracy. For example, [16] employed a Stacking Classifier that integrated multiple base learners, including Decision Trees and Gradient Boosting, and demonstrated improved performance compared to individual models. Similarly, [17] explored Voting Classifiers and reported significant gains in recall and F1 scores, particularly in datasets with imbalanced class distributions. Despite these advancements, a critical gap remains in evaluating models across multiple dimensions, including custom metrics tailored to the domain's specific diagnostic needs[18].

Another prominent challenge in this domain is handling imbalanced datasets, where minority class predictions are often overlooked by models optimized for accuracy. Techniques like the Synthetic Minority Oversampling Technique (SMOTE) have been employed to address this issue. Studies by [19][21] highlighted the efficacy of SMOTE in improving the recall and precision of minority classes in medical datasets. However, there is limited work integrating these techniques into comprehensive ML pipelines that include advanced ensemble models and custom metric evaluation [22][24]. This research addresses these gaps by proposing a robust comparison framework for

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

diabetes prediction using the Pima dataset. Specifically, the study evaluates individual models such as Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and LightGBM, alongside advanced ensemble techniques like Stacking and Voting Classifiers. The framework integrates SMOTE for handling class imbalance and employs a novel custom metric to assess the models' diagnostic utility comprehensively. This custom metric is designed to align with the clinical relevance of diabetes diagnosis, ensuring that the models provide actionable insights for healthcare practitioners. The contributions of this research are threefold. First, it systematically evaluates a wide range of machine learning models, highlighting their comparative strengths and limitations in predicting diabetes. Second, it introduces a custom metric to complement traditional metrics such as accuracy, precision, recall, and F1-score, offering a nuanced perspective on model performance. Third, it emphasizes the integration of SMOTE and rigorous cross-validation techniques, ensuring that the findings are robust and generalizable. Together, these contributions aim to advance the state-of-the-art in diabetes diagnosis using machine learning and establish a benchmark for future studies. The remainder of this article is structured as follows: Section 2 outlines the methodology, including data preprocessing, model selection, and evaluation strategies. Section 3 reports the results of the comparative analysis, supported by statistical insights and visualizations. Section 4 discusses the findings in light of existing literature, highlighting their implications for clinical practice and future research. Finally, Section 5 concludes the study, summarizing key contributions and outlining potential directions for further exploration.

2. RESEARCH METHODOLOGY

As presented in the Figure 1, the research methodology employed in this study is systematically designed to ensure a comprehensive evaluation of machine learning models for diabetes prediction. The steps are organized sequentially and outlined in the corresponding activity diagram, providing clarity and consistency across the research process. Each phase is integral to addressing the challenges posed by the dataset's characteristics and achieving the study's objectives

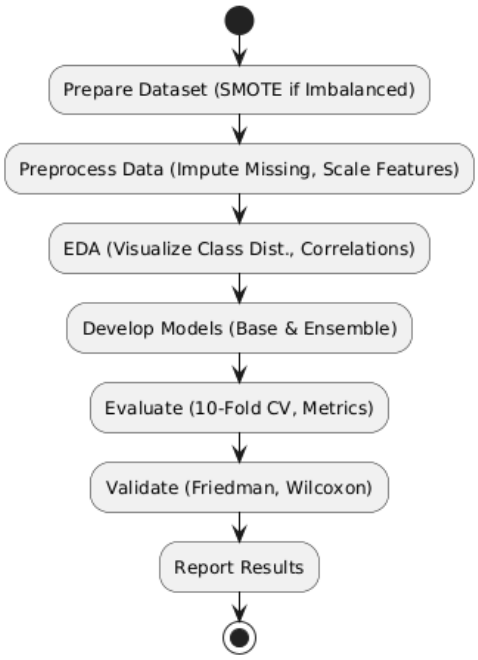

Figure 1. Research Methodology Diagram

2.1 Dataset Preprocessing

Dataset preprocessing is a fundamental step in any machine learning workflow, as it ensures that the data is properly prepared for effective model training and evaluation. For this study, the preprocessing pipeline involves feature scaling and handling class imbalance, both of which are critical to building robust and unbiased machine learning models.

2.1.1 Scaling Features

Feature scaling is essential when working with datasets that have predictor variables with differing ranges. Variables such as *Glucose* and *BMI* may have large numerical ranges compared to variables like *DiabetesPedigreeFunction*, which typically has values between 0 and 1. Without scaling, features with larger ranges may disproportionately influence the learning process, leading to biased models. To address this, a *StandardScaler* function was applied to all predictor variables. The *StandardScaler* function transforms each feature *x* according to the equation 1.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

$$x' = \frac{x-\mu}{\sigma} \quad (1)$$

Here, x' represents the scaled value, μ is the mean of the feature, and σ is the standard deviation of the feature. This transformation ensures that each feature has a mean of 0 and a standard deviation of 1, thereby standardizing all features to a uniform scale. For instance, consider the *Glucose* variable with an example value of 150, a mean of 120, and a standard deviation of 30. The scaled value for this example would be presented in the equation 2.

$$x' = \frac{150-120}{30} = 1.0 \quad (2)$$

Similarly, all other features are transformed to have a mean of 0 and a standard deviation of 1. This standardization ensures that the machine learning models treat all features equally, preventing any single feature from dominating the learning process.

2.1.2 Handling Imbalanced Classes

The dataset used in this study is inherently imbalanced, with approximately 65% of the samples belonging to the non-diabetic class (*Outcome* = 0) and only 35% belonging to the diabetic class (*Outcome* = 1). Such an imbalance poses a significant challenge, as machine learning models trained on imbalanced data are likely to be biased towards the majority class. Consequently, the models may achieve high accuracy by simply predicting the majority class while failing to correctly classify the minority class, which is of greater clinical importance in this study. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE is an oversampling technique that generates synthetic samples for the minority class by interpolating between existing samples in the minority class. The interpolation process involves selecting a random sample (x_i) from the minority class and finding its (k)-nearest neighbors within the minority class. A synthetic sample (x_{new}) is then generated using the equation 3.

$$x_{\text{new}} = x_i + \delta \cdot (x_{\text{nn}} - x_i) \quad (3)$$

In this equation, (x_{nn}) is a randomly selected neighbor of (x_i), and (δ) is a random number between 0 and 1. By repeating this process, SMOTE generates new samples that lie along the line segments connecting existing minority class samples. For example, consider two samples in the minority class: $(x_1 = [150,70])$ and $(x_2 = [155,75])$, representing *Glucose* and *BloodPressure*, respectively. Using SMOTE with ($\delta = 0.5$), the synthetic sample generated would be equation 4.

$$x_{\text{new}} = [150,70] + 0.5 \cdot ([155,75] - [150,70]) = [152.5,72.5] \quad (4)$$

This synthetic sample is then added to the dataset, effectively balancing the number of samples in each class. SMOTE ensures that the machine learning models are trained on a dataset with equal representation from both classes, thereby improving their ability to correctly classify diabetic patients. By combining feature scaling with SMOTE, the dataset is prepared in a way that mitigates bias and ensures that all features and classes contribute equally to the training process. These preprocessing steps form a critical foundation for building accurate and reliable machine learning models for diabetes prediction.

2.2 Model Development

The development of machine learning models in this study involved carefully selecting individual base models and advanced ensemble methods to harness their combined strengths. By integrating diverse algorithms, the approach aimed to optimize predictive performance, balance accuracy with interpretability, and address the inherent complexity of the dataset. The base models included five widely-used machine learning algorithms, each chosen for its unique properties. Logistic Regression, for example, was selected for its simplicity and effectiveness in scenarios where classes are linearly separable. This model computes the probability of the positive class using the sigmoid function, defined as presented in the equation 5.

$$P(y = 1|x) = \sigma(w^{\top}x + b) = \frac{1}{1+e^{-(w^{\top}x+b)}} \quad (5)$$

where (w) represents the weights, (x) is the feature vector, (b) is the bias term, and (σ) is the sigmoid function. This method was configured with a maximum of 1000 iterations, ensuring convergence even in more challenging scenarios. Its inclusion provides a baseline for comparison and sets a foundation of interpretability for other, more complex models. In addition to Logistic Regression, Random Forest was utilized as a robust ensemble model. Unlike Logistic Regression, Random Forest combines multiple decision trees using a bagging technique, whereby each tree is trained on a random subset of the data. The final prediction for classification is made by aggregating the outputs of individual trees, defined mathematically as presented in the equation 6.

$$P(y = 1|x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (6)$$

where (T) is the total number of trees, and $(h_t(x))$ represents the output of the (t)-th tree. This model is particularly effective for handling non-linear relationships and interactions within the dataset, complementing the capabilities of linear models. Gradient Boosting further advanced the predictive capabilities by introducing a

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

sequential learning framework. Unlike Random Forest, which builds trees independently, Gradient Boosting constructs trees iteratively, each correcting the residual errors of the previous tree. This iterative improvement is expressed as equation 7.

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{7}$$

where (F_m) is the prediction at iteration (m), ($h_m(x)$) is the new tree's output, and (η) is the learning rate. The incorporation of Gradient Boosting, along with its optimized implementations in XGBoost and LightGBM, provided efficient and accurate models capable of capturing complex patterns in the data. XGBoost introduced additional computational optimizations, while LightGBM leveraged histogram-based learning to accelerate tree construction further, making it particularly well-suited for large-scale datasets. To enhance the performance of these base models, ensemble methods were employed. The Stacking Classifier aggregated the predictions of base models by combining their outputs as inputs to a meta-learner, specifically a Logistic Regression model. This meta-learner improved prediction accuracy by learning how to reconcile the strengths and weaknesses of the base models. Mathematically, the meta-learner predicts the final class label (y) as presented in equation 8.

$$y = g\big(h_1(x), h_2(x), \dots, h_n(x)\big) \tag{8}$$

where ($h_i(x)$) is the prediction of the (i)-th base model, and (g) represents the meta-learner's function. By exploiting the diversity of base models, Stacking Classifiers effectively addressed challenges such as overfitting and underperformance in specific subspaces of the data. The Voting Classifier provided another layer of ensemble learning by aggregating predictions through a probabilistic soft voting mechanism. This approach computes the final class probabilities by averaging the outputs of the base models as presented in equation 9.

$$P(y = c|x) = \frac{1}{n}\sum_{i=1}^{n} P_i(y = c|x) \tag{9}$$

where ($P_i(y = c|x)$) is the probability predicted by the (i)-th model for class (c), and (n) is the total number of base models. The final class label is assigned based on the highest average probability, effectively leveraging the collective strength of all base models. By combining individual base models with ensemble methods, this approach ensured a balance between simplicity and complexity. The base models provided diverse perspectives on the data, while the ensemble methods unified their strengths to achieve improved predictive performance. Together, these models formed a robust and versatile framework for addressing the challenges of diabetes prediction, ensuring both interpretability and accuracy.

2.3. Evaluation Strategy

This study employed a rigorous evaluation methodology that combines a robust cross-validation framework, diverse performance metrics, and statistical validation to ensure reliable comparisons among machine learning models. The evaluation strategy was specifically designed to align with clinical priorities, including minimizing false negatives, which are critical in identifying diabetic cases, and balancing precision to avoid over-diagnosis. To assess model performance, a 10-fold cross-validation strategy was adopted. In this approach, the dataset, represented as ($\mathcal{D} = \{X, y\}$), where (X) represents the features and (y) the labels, was partitioned into ten equal-sized subsets. During each iteration, one subset was used for testing while the remaining nine were used for training. This iterative process ensures that every data point contributes equally to training and testing, thereby mitigating the risk of overfitting. For each fold (i), the performance metric (M) was calculated, and the final metric was averaged over all ten folds as $M = \frac{1}{10}\sum_{i=1}^{10} M_i$. This framework ensures that the evaluation reflects the model's generalization ability across different subsets of the data.

In addition to the custom metric, the study used standard metrics such as accuracy, precision, recall, F1 score, and ROC AUC to evaluate the models comprehensively. Accuracy was computed as $A = \frac{TP+TN}{TP+TN+FP+FN}$ representing the proportion of correctly predicted instances. Precision, calculated as $P = \frac{TP}{TP+FP}$ measures the proportion of true positives among all positive predictions. Recall, defined as $R = \frac{TP}{TP+FN}$ evaluates the proportion of true positives among all actual positives. The F1 score, expressed as $F1 = 2 \cdot \frac{P \cdot R}{P+R}$ balances the trade-offs between precision and recall. ROC AUC quantifies the model's ability to distinguish between classes by computing the area under the receiver operating characteristic curve. The custom metric introduced in this study was tailored to prioritize recall while maintaining high precision. It was designed to address the clinical imperative of minimizing false negatives, which represent undiagnosed diabetic cases, and to balance this with precision to avoid unnecessary false positives. The custom metric (CM) was defined as $CM = \alpha \cdot \text{Recall} + \beta \cdot \text{Precision}$ where (α) and (β) are weights assigned to recall and precision, respectively, such that ($\alpha + \beta = 1$). In this experiment, (α) was set higher than (β) to reflect the greater importance of identifying diabetic cases. Substituting the definitions of recall and precision, the custom metric can be expressed as $CM = \alpha \cdot \frac{TP}{TP+FN} + \beta \cdot \frac{TP}{TP+FP}$. This metric was implemented as a custom scorer in the cross-validation pipeline, ensuring its seamless integration into the evaluation process.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

To validate the statistical significance of the differences in model performance, the Friedman test was applied. This non-parametric test assesses whether the rankings of models across folds differ significantly. Let (R_{ij}) denote the rank of model (j) for fold (i). The Friedman test statistic, (χ_F^2), is computed as $\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{j=1}^{k} \overline{R_j^2} - \frac{k(k+1)^2}{4}\right)$ where (N) is the number of folds, (k) is the number of models, and $(\bar{R}_j)$ is the average rank of model (j). A significant (p)-value from the Friedman test indicates that at least one model differs significantly in its rankings. For pairwise comparisons, the Wilcoxon signed-rank test was used to evaluate whether two models differ significantly in performance. For models (A) and (B), the test statistic (W) was calculated as $W = \sum_{\text{ranks of } d_i>0} \text{Rank}(|d_i|)$ where (d_i) represents the difference in metric values for fold (i). A significant (p)-value from the Wilcoxon test indicates that the median difference $(\tilde{d})$ is significantly different from zero. In this experiment, the use of the custom metric ensured that the evaluation aligned with clinical priorities, emphasizing the identification of diabetic cases without compromising precision. The integration of rigorous statistical tests validated the significance of observed performance differences, reinforcing the reliability of the results.

3. RESULT AND DISCUSSION

3.1. Dataset Preparation

The Pima Indians Diabetes Dataset, obtained from the National Institute of Diabetes and Digestive and Kidney Diseases, serves as the foundation for this study and can be downloaded from [12]. This dataset comprises 768 records, each corresponding to a female patient of Pima Indian heritage, aged 21 years or older. It aims to predict the likelihood of diabetes onset based on a series of diagnostic measurements. These measurements include variables such as the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, serum insulin levels, body mass index (BMI), diabetes pedigree function, and age. The target variable, denoted as Outcome, is binary, with a value of 1 indicating the presence of diabetes and 0 indicating its absence. A significant challenge inherent in this dataset is the presence of class imbalance. Specifically, approximately 65% of the records belong to the non-diabetic class, while the remaining 35% represent diabetic cases. This imbalance can lead machine learning models to favor the majority class, potentially reducing their ability to correctly classify diabetic instances. Furthermore, several features, such as SkinThickness and Insulin, contain zero values, which likely indicate missing data rather than true measurements. These characteristics of the dataset necessitate careful preprocessing to address missing data and to mitigate the effects of class imbalance, ensuring that the models developed are both robust and unbiased. Detailed information of databases is presented in Table 1.

Table 1. Types of databases

Variable	Description	Example Value	Data Type
Pregnancies	The number of pregnancies the patient has had.	2	Integer
Glucose	Plasma glucose concentration measured two hours after a glucose tolerance test.	148	Integer
BloodPressure	Diastolic blood pressure (mm Hg).	72	Integer
SkinThickness	Triceps skinfold thickness (mm).	35	Integer
Insulin	Serum insulin levels (mu U/ml).	125	Integer
BMI	Body mass index, calculated as weight in kilograms divided by the square of height in meters.	32.5	Float
DiabetesPedigreeFunction	A score representing diabetes hereditary likelihood based on family history.	0.627	Float
Age	Age of the patient in years.	45	Integer
Outcome	Binary classification (1 for diabetic, 0 for non-diabetic).	1	Integer

To provide a comprehensive understanding of the dataset, exploratory data analysis was conducted. The first visualization in Figure 2 illustrates the distribution of the target variable, Outcome. The dataset is clearly imbalanced, with the majority class (non-diabetic) accounting for approximately 65% of the records. This imbalance is evident in the bar chart, which underscores the necessity of addressing this issue during preprocessing to ensure fair and robust model performance.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

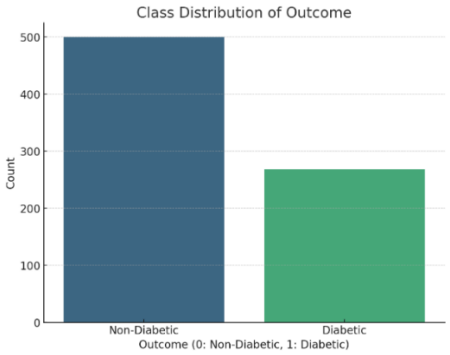

Figure 2. Class Distribution of Dataset

Secondly, as presented in Figure 3, a correlation heatmap was generated to examine the relationships between predictor variables and the target variable. The heatmap reveals that *Glucose*, *BMI*, and *DiabetesPedigreeFunction* exhibit the highest positive correlations with Outcome. These findings align with prior medical studies that emphasize the importance of these variables in predicting diabetes. Additionally, minimal multicollinearity is observed between predictors, suggesting that these features contribute independently to the model.

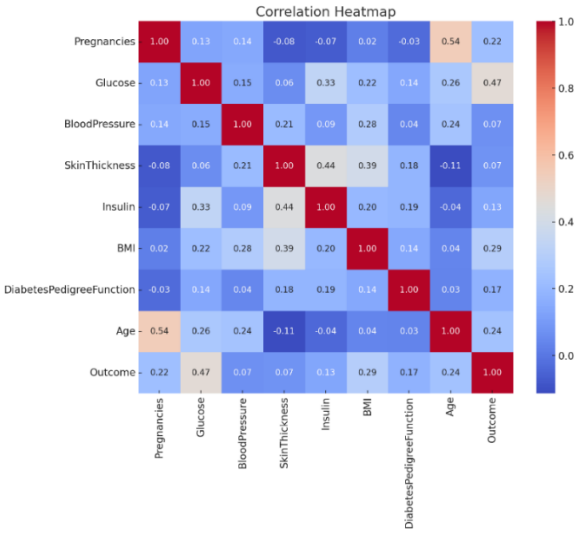

Figure 3. Correlation Heatmap

Third, as presented in Figure 4 a histogram of *Glucose levels*, stratified by Outcome, provides a deeper understanding of its distribution. Diabetic patients tend to have significantly higher glucose levels compared to non-diabetic patients. The kernel density estimate (KDE) further highlights the stark difference between the two classes, reinforcing the critical role of Glucose in distinguishing diabetic from non-diabetic cases.

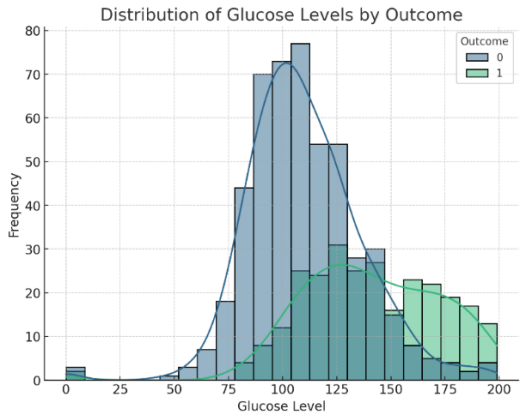

Figure 4. Glucose Level

Lastly, as presented in Figure 5, a scatter plot of Age versus BMI provides insights into the interaction between these two variables for diabetic and non-diabetic individuals. The plot reveals a diverse distribution of BMI across

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

different age groups, with diabetic patients typically showing higher BMI values. This relationship suggests that both age and BMI contribute to the prediction of diabetes, although their influence may vary depending on the patient's profile.

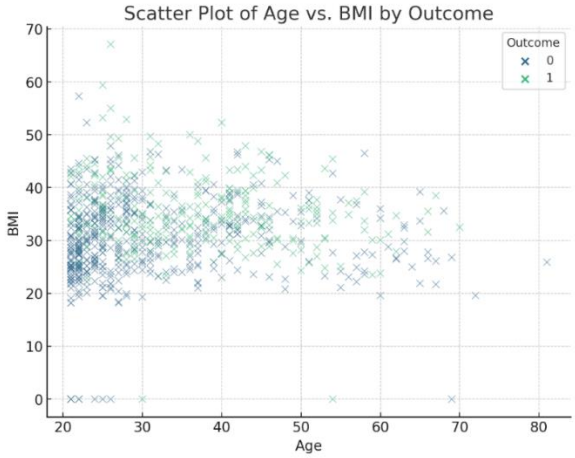

Figure 5. Scatter Plot of Age

3.2 Evaluation Metrics Analysis

The evaluation of the machine learning models on the diabetes dataset provides a comprehensive understanding of their predictive capabilities, robustness, and suitability for clinical application. The models were assessed using key performance metrics, including accuracy, precision, recall, F1 score, ROC AUC, and a custom metric tailored to this study's objectives. As presented in the table 2, the results reveal distinct strengths and limitations across the models, offering valuable insights into their behavior. Logistic Regression, serving as the baseline model, achieved an accuracy of 0.743, precision of 0.764, recall of 0.708, and an F1 score of 0.734, along with an ROC AUC of 0.843. While the model is straightforward and interpretable, its relatively lower performance highlights its limitations in capturing complex, non-linear relationships inherent in the dataset. Despite its simplicity, Logistic Regression remains valuable as a benchmark for evaluating the more advanced models.

Random Forest demonstrated a significant improvement over Logistic Regression, achieving an accuracy of 0.818 and an ROC AUC of 0.904, reflecting its ability to effectively distinguish between diabetic and non-diabetic cases. The model's recall of 0.860 and custom metric score of 0.860 underline its strong performance in identifying diabetic patients while maintaining precision at 0.795. This result emphasizes the power of Random Forest in handling structured data, as it efficiently captures non-linear interactions through its ensemble of decision trees. Gradient Boosting also delivered strong results, with an accuracy of 0.807 and an F1 score of 0.814, supported by a recall of 0.846 and an ROC AUC of 0.887. While Gradient Boosting's precision of 0.786 is slightly lower than Random Forest's, its sequential optimization strategy allows it to correct residual errors effectively. This trade-off indicates that Gradient Boosting may prioritize recall, making it suitable for scenarios where minimizing false negatives is critical.

Table 2. Evaluation Metrics Results

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC	Custom Metric
Logistic Regression	0.743	0.764336	0.708	0.733764	0.84312	0.708
Random Forest	0.818	0.795311	0.86	0.825482	0.90356	0.86
Gradient Boosting	0.807	0.786234	0.846	0.814426	0.8866	0.846
XGBoost	0.816	0.797847	0.854	0.822999	0.89872	0.854
LightGBM	0.826	0.807104	0.86	0.831715	0.8964	0.86
Stacking Classifier	0.826	0.81384	0.85	0.830095	0.90316	0.85
Voting Classifier	0.821	0.800995	0.858	0.827361	0.90036	0.858

XGBoost and LightGBM, both optimized implementations of gradient boosting, further enhanced performance. XGBoost achieved an accuracy of 0.816, recall of 0.854, and an F1 score of 0.823, with an ROC AUC of 0.899. However, LightGBM marginally outperformed XGBoost, achieving the highest accuracy of 0.826 and a custom metric score of 0.860. LightGBM's precision of 0.807 and F1 score of 0.832 underscore its balanced performance, particularly in maintaining high recall and precision simultaneously. These results validate the efficiency of LightGBM's histogram-based learning, which optimizes computational speed without compromising predictive accuracy. The ensemble models, which combine predictions from multiple base models, demonstrated superior overall performance. The Stacking Classifier achieved an accuracy of 0.826, precision of 0.814, recall of 0.850, and an F1 score of 0.830, supported by an ROC AUC of 0.903. This result highlights the Stacking Classifier's ability to refine predictions by leveraging the diversity of base model outputs through its meta-learner. Similarly, the Voting Classifier

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

attained an accuracy of 0.821, precision of 0.801, and recall of 0.858, with an F1 score of 0.827 and an ROC AUC of 0.900. The use of soft voting allowed it to aggregate the strengths of individual models effectively, producing robust predictions across various scenarios.

The custom metric, designed to prioritize recall while maintaining high precision, revealed that LightGBM and Random Forest are particularly effective for diabetes diagnosis, both achieving the highest custom metric score of 0.860. This finding underscores their reliability in minimizing false negatives, a crucial consideration in clinical applications where undiagnosed diabetes can lead to severe complications. The results further highlight the value of ensemble methods. By combining the strengths of base models, the Stacking Classifier and Voting Classifier achieved performance metrics that slightly exceeded those of individual models, indicating their ability to generalize better to unseen data. The Stacking Classifier, in particular, benefited from its meta-learning approach, which reconciled discrepancies in the base models' predictions.

3.2 Friendman Test Analysis

A statistical comparison of the models' performance was conducted using the Friedman test, a non-parametric method designed to evaluate the consistency of rankings across multiple algorithms over different metrics. The test produced a Friedman test statistic of 22.77 and a corresponding p-value of 0.00088. Given that the p-value is significantly below the standard significance threshold of 0.05, we reject the null hypothesis, which posits that all models perform equally well. This result confirms that there are statistically significant differences in the performance of at least one model compared to the others. The rejection of the null hypothesis is crucial as it validates the observed variations in performance metrics across the models, such as accuracy, precision, recall, F1 score, ROC AUC, and the custom metric. These differences highlight that certain models, particularly LightGBM and ensemble techniques like the Stacking Classifier, consistently outperform others in various evaluation criteria. For example, ensemble models demonstrated superior accuracy and ROC AUC, which aligns with their ability to aggregate the strengths of diverse base algorithms. In contrast, Logistic Regression, serving as a baseline, showed relatively lower performance, emphasizing the importance of advanced modeling approaches

3.2 Wilcoxon Signed-Rank Test Results for Pairwise Model Comparisons

The pairwise comparisons of model performances were evaluated using the Wilcoxon signed-rank test, focusing on differences across cross-validation metrics such as accuracy, precision, recall, F1 score, and ROC AUC. As presented in the table 3, the results provide critical insights into the relative strengths of the models and whether the observed performance differences are statistically significant. The statistical significance was determined based on a p-value threshold of 0.05. The comparison between Logistic Regression and all other models, including Random Forest, Gradient Boosting, XGBoost, LightGBM, Stacking Classifier, and Voting Classifier, consistently yielded a Wilcoxon test statistic of 0.0 and a p-value of 0.0625. While the low test statistic suggests a clear performance difference, the p-value is slightly above the significance threshold, indicating that these differences are not statistically significant at the 0.05 level. This result implies that, although Logistic Regression consistently performed worse than the other models in numerical terms, the variations may be attributed to random fluctuations in cross-validation folds rather than inherent model superiority.

When comparing Random Forest to Gradient Boosting, the Wilcoxon test statistic of 0.0 and a p-value of 0.0625 again indicate a lack of statistical significance despite numerical differences. The pairwise comparisons of Random Forest with XGBoost (p=0.3125) and LightGBM (p=0.3573) also failed to reject the null hypothesis, suggesting comparable performances. Interestingly, comparisons between Random Forest and ensemble models like Stacking Classifier (p=0.625) and Voting Classifier (p=0.625) indicate that ensemble methods do not provide statistically significant improvements over Random Forest. This result highlights the strong baseline performance of Random Forest and suggests that while ensemble methods are often effective, their benefits may not always be pronounced depending on the dataset and evaluation metrics. Gradient Boosting displayed a similar pattern of non-significant differences when compared to XGBoost, LightGBM, Stacking Classifier, and Voting Classifier, with p-values consistently at 0.0625. These results underscore the competitive performance of Gradient Boosting and its variants, particularly XGBoost and LightGBM, in structured data tasks like this one. While LightGBM consistently demonstrated the highest accuracy and custom metric scores, the lack of statistical significance in its comparisons with XGBoost (p=0.125) and Voting Classifier (p=0.1875) suggests that the observed differences may not generalize beyond this specific dataset.

The Stacking Classifier and Voting Classifier, as ensemble methods, are often expected to outperform base models due to their ability to combine diverse algorithmic strengths. However, the comparisons between these two ensemble methods (p=0.4375) and their comparisons with LightGBM (p=1.000 and p=0.1875, respectively) indicate no statistically significant advantage. These findings suggest that while ensemble models demonstrate robust performance, their improvements over strong individual models like LightGBM are not always statistically distinct. The results of the Wilcoxon signed-rank tests reinforce several important conclusions. First, while advanced models and ensemble methods demonstrate superior numerical performance compared to simpler models like Logistic Regression, these differences are not statistically significant at conventional thresholds. Second, the performance differences between strong individual models, such as Gradient Boosting, XGBoost, and LightGBM, are subtle and often statistically indistinguishable. Finally, the ensemble methods, despite their theoretical advantages, do not exhibit

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

statistically significant improvements over the best-performing individual models in this study. These findings emphasize the importance of considering both numerical results and statistical validation when evaluating model performance. While ensemble methods and advanced boosting techniques are valuable tools, their effectiveness should be weighed against the computational complexity they introduce, particularly when their advantages are not statistically significant. Future work could explore larger datasets or alternative metrics to further validate these observations and investigate whether these trends persist in other clinical prediction tasks.

Table 3. Wilcoxon Test Results

Model Comparison	Wilcoxon Test Statistic	p-value
Logistic Regression vs Random Forest	0.0	0.0625
Logistic Regression vs Gradient Boosting	0.0	0.0625
Logistic Regression vs XGBoost	0.0	0.0625
Logistic Regression vs LightGBM	0.0	0.0625
Logistic Regression vs Stacking Classifier	0.0	0.0625
Logistic Regression vs Voting Classifier	0.0	0.0625
Random Forest vs Gradient Boosting	0.0	0.0625
Random Forest vs XGBoost	3.0	0.3125
Random Forest vs LightGBM	2.5	0.357273
Random Forest vs Stacking Classifier	5.0	0.625
Random Forest vs Voting Classifier	5.5	0.625
Gradient Boosting vs XGBoost	0.0	0.0625
Gradient Boosting vs LightGBM	0.0	0.0625
Gradient Boosting vs Stacking Classifier	0.0	0.0625
Gradient Boosting vs Voting Classifier	0.0	0.0625
XGBoost vs LightGBM	1.0	0.125
XGBoost vs Stacking Classifier	1.5	0.125
XGBoost vs Voting Classifier	0.0	0.0625
LightGBM vs Stacking Classifier	5.0	1.0
LightGBM vs Voting Classifier	2.0	0.1875
Stacking Classifier vs Voting Classifier	4.0	0.4375

4. CONCLUSION

1This study evaluated the performance of various machine learning models for predicting diabetes using the Pima Indians Diabetes Dataset, analyzing seven models including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, Stacking Classifier, and Voting Classifier. The models were assessed using multiple metrics such as accuracy, precision, recall, F1 score, ROC AUC, and a custom metric prioritizing recall to minimize false negatives. Tree-based models like Random Forest and LightGBM excelled, particularly in recall and ROC AUC, which are critical for clinical applications, with LightGBM achieving the highest overall scores due to its efficiency and predictive power for structured datasets. Ensemble methods, such as the Stacking Classifier and Voting Classifier, further improved performance by leveraging the strengths of multiple base models, though statistical tests revealed that performance differences among top-performing models were not always significant. While Logistic Regression provided interpretability, it was outperformed by advanced models that captured complex non-linear relationships. The Friedman test confirmed significant differences among models, emphasizing the importance of model selection, though pairwise comparisons indicated comparable performance between LightGBM and Random Forest, underscoring the robustness of both approaches. This study highlights the value of advanced and ensemble methods in achieving high accuracy and reliability in predicting diabetes, particularly in balancing precision and recall, which is crucial for clinical applications. Future research should explore incorporating additional features, larger datasets, and hybrid modeling approaches to enhance predictive performance and generalizability, providing a strong foundation for deploying machine learning in real-world healthcare scenarios for early detection and management of diabetes.

REFERENCES

[1] M. L. Avilés-Santa, A. Monroig-Rivera, A. Soto-Soto, and N. M. Lindberg, “Current state of diabetes mellitus prevalence, awareness, treatment, and control in Latin America: challenges and innovative solutions to improve health outcomes across the continent,” *Curr. Diab. Rep.*, vol. 20, pp. 1–44, 2020.
[2] Z. L. Teo *et al.*, “Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis,” *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591, 2021.
[3] L. Jiang *et al.*, “A global view of hypertensive disorders and diabetes mellitus during pregnancy,” *Nat. Rev. Endocrinol.*, vol. 18, no. 12, pp. 760–775, 2022.
[4] J. L. Harding, M. B. Weber, and J. E. Shaw, “The Global Burden of Diabetes,” *Textb. Diabetes*, pp. 28–40, 2024.
[5] U. Ramraj, “Living with diabetes: managing treatment and the psycho-social aspects of the disease,” 2023.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1818−1827
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/ bits.v6i3.6419

[6] H. Wang, S. Akbari-Alavijeh, R. S. Parhar, R. Gaugler, and S. Hashmi, "Partners in diabetes epidemic: A global perspective," *World J. Diabetes*, vol. 14, no. 10, p. 1463, 2023.
[7] M. Zakir *et al.*, "Cardiovascular complications of diabetes: from microvascular to macrovascular pathways," *Cureus*, vol. 15, no. 9, 2023.
[8] D. Crosby *et al.*, "Early detection of cancer," *Science (80-.).*, vol. 375, no. 6586, p. eaay9040, 2022.
[9] T. De Francesco, J. Bacharach, O. Smith, and M. Shah, "Early diagnostics and interventional glaucoma," *Ther. Adv. Ophthalmol.*, vol. 16, p. 25158414241287430, 2024.
[10] S. Asif *et al.*, "Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision," *Arch. Comput. Methods Eng.*, pp. 1–31, 2024.
[11] E. Afrifa-Yamoah *et al.*, "Pathways to chronic disease detection and prediction: Mapping the potential of machine learning to the pathophysiological processes while navigating ethical challenges," *Chronic Dis. Transl. Med.*, 2024.
[12] K. Contributors, "Pima Indians Diabetes Database." 2016. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.
[13] V. G. Costa and C. E. Pedreira, "Recent advances in decision trees: An updated survey," *Artif. Intell. Rev.*, vol. 56, no. 5, pp. 4765–4800, 2023.
[14] M. S. Reza, U. Hafsha, R. Amin, R. Yasmin, and S. Ruhi, "Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset," *Comput. Methods Programs Biomed. Updat.*, vol. 4, p. 100118, 2023.
[15] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes \& Metab. Disord.*, vol. 19, pp. 391–403, 2020.
[16] A. D. Waberi, R. W. Mwangi, and R. M. Rimiru, "Advancing Type II Diabetes Predictions with a Hybrid LSTM-XGBoost Approach," *J. Data Anal. Inf. Process.*, vol. 12, no. 02, pp. 163–188, 2024.
[17] N. Javaid, M. Akbar, A. Aldegheishem, N. Alrajeh, E. A. Mohammed, and others, "Employing a machine learning boosting classifiers based stacking ensemble model for detecting non technical losses in smart grids," *IEEE Access*, vol. 10, pp. 121886–121899, 2022.
[18] K. A. Reed *et al.*, "Metrics as tools for bridging climate science and applications," *Wiley Interdiscip. Rev. Clim. Chang.*, vol. 13, no. 6, p. e799, 2022.
[19] S. Gholampour, "Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 827–841, 2024.
[20] A. M. Sowjanya and O. Mrudula, "Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms," *Appl. Nanosci.*, vol. 13, no. 3, pp. 1829–1840, 2023.
[21] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: review of a decade of research," *Artif. Intell. Rev.*, vol. 57, no. 10, p. 273, 2024.
[22] V.-E. Baciu, J. Stiens, and B. da Silva, "MLino bench: A comprehensive benchmarking tool for evaluating ML models on edge devices," *J. Syst. Archit.*, vol. 155, p. 103262, 2024.
[23] N. L. Rane, S. K. Mallick, O. Kaya, and J. Rane, "Tools and frameworks for machine learning and deep learning: A review," *Appl. Mach. Learn. Deep Learn. Archit. Tech.*, pp. 80–95, 2024.
[24] N. O. Nikitin *et al.*, "Automated evolutionary approach for the design of composite machine learning pipelines," *Futur. Gener. Comput. Syst.*, vol. 127, pp. 109–125, 2022.