Accredited SINTA 2 Ranking
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online at: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 8 No. 5 (2024) 607 - 614 e-ISSN: 2580-0760

Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of SVM, Logistic Regression, and Naive Bayes with SMOTE and ADASYN

Selly Rahmawati[1], Arief Wibowo[2*], Anis Fitri Nur Masruriyah[3]
[1,2]Department Information System, Faculty Information Technology, Universitas Budi Luhur, Jakarta, Indonesia
[3]Department Informatics, Faculty Computer Science, Universitas Buana Perjuangan Karawang, Indonesia
[1]2211601055@student.budiluhur.ac.id, [2]arief.wibowo@budiluhur.ac.id, [3]anis.masruriyah@ubpkarawang.ac.id

Abstract

This study aims to enhance the accuracy of diabetes prediction models in Indonesia by comparing the performance of Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes algorithms, both with and without synthetic oversampling techniques such as SMOTE and ADASYN. The research addresses the issue of imbalanced datasets in medical diagnostics, specifically in predicting diabetes among Indonesian patients, where such imbalance often leads to biased predictions. A comprehensive dataset comprising 657 patient records from a Regional General Hospital in Indonesia was used, with 70% of the data allocated for training and 30% for testing. The results indicate that the SVM model combined with SMOTE achieved the highest accuracy of 95.8% and an AUC of 99.1, underscoring the effectiveness of these techniques in improving prediction performance. The findings of this study highlight the importance of selecting appropriate oversampling methods and algorithms to optimize diabetes prediction accuracy in the Indonesian context, providing valuable insights for future healthcare strategies.

Keywords: adaptive synthetic sampling; diabetes mellitus; logistic regression; naïve bayes; synthetic minority over-sampling technique; support vector machine

How to Cite: S. Rahmawati, A. Wibowo, and A. F. N. Masruriyah, "Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of SVM, Logistic Regression, and Naive Bayes with SMOTE and ADASYN", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 8, no. 5, pp. 607 - 614, Oct. 2024.
DOI: https://doi.org/10.29207/resti.v8i5.5980

1. Introduction

Diabetes mellitus is a widely prevalent chronic metabolic disorder and poses a significant public health challenge, particularly in countries with a high disease burden, including Indonesia. [1]. This condition has a substantial impact on the population and is frequently linked to severe complications such as cardiovascular disease, kidney failure, and neuropathy [2], [3]. In Indonesia, rapid urbanization and lifestyle changes have aggravated the prevalence of diabetes, necessitating a more tailored predictive approach for risk mitigation. The incidence of diabetes in Indonesia has been escalating at an alarming rate, largely due to urbanization, lifestyle changes, and specific genetic predispositions unique to the Indonesian population. This swift urbanization has particularly led to increased consumption of high-calorie foods and decreased physical activity, which, combined with specific genetic factors, has exacerbated the diabetes situation in the country.

This alarming trend necessitates the development of precise predictive models to facilitate early diagnosis and intervention, thereby mitigating the disease's impact on individuals and healthcare systems. As such, accurate prediction and early diagnosis of diabetes are essential to mitigate its adverse impact on individuals and healthcare systems, emphasizing the need for robust predictive models in the Indonesian context [2], [3], [4]. The primary goal of this study is to develop and evaluate a diabetes prediction model that accurately reflects the unique risk factors in Indonesia, while also enhancing prediction accuracy through the application of suitable oversampling techniques. This study seeks to explore how demographic and lifestyle factors in Indonesia influence the accuracy of diabetes prediction models and to identify which algorithms and

Received: 09-08-2024 | Accepted: 01-10-2024 | Published Online: 07-10-2024

oversampling methods are most effective in improving prediction accuracy within the Indonesian context.

Despite the advancements in medical diagnostics, conventional approaches often fall short in accurately predicting diabetes, especially within heterogeneous populations like Indonesia. This discrepancy stems from the unique interplay of genetic, environmental, and lifestyle factors prevalent among Indonesians, which are not adequately captured by generic models. Therefore, there is a critical need for tailored predictive models that can address these nuances.

Recent advancements in machine learning (ML) have demonstrated considerable potential in enhancing the accuracy of disease prediction and classification. The study [5] underscores Diabetes Mellitus as a prevalent chronic endocrine disease influenced by both genetic and lifestyle factors, affecting a diverse age group. It reports that 68% of the population in the country is impacted, highlighting the critical necessity for early prediction to prevent complications. The study evaluates various machine learning techniques, including classifiers such as K-Nearest Neighbors, Naive Bayes, XGBoost, Decision Tree, and Random Forest, for their efficacy in diabetes prediction. Initially, Random Forest was identified as particularly precise; however, the study concludes that XGBoost, with a precision rate of 77%, surpassed other classifiers in predicting diabetic outcomes. This highlights the potential of machine learning in healthcare for early intervention and improved disease management, underlining the importance of continuous research in this field to enhance patient outcomes and healthcare strategies. Additionally, study [6] employed supervised learning and logistic regression to construct a model for diagnosing diabetes. The study utilizes logistic regression to analyze a dataset of patient information, aiming to predict diabetes diagnoses and emphasizing the significance of data-driven predictions in healthcare. The conclusion reveals that the logistic regression model achieved an accuracy rate of 74%, which, while better than random guessing, is not adequate as a standalone diagnostic tool. To improve accuracy, the authors suggest exploring more advanced models, such as neural networks, and incorporating additional features. This study underscores the potential of machine learning in disease diagnosis and prevention, proposing it as a valuable tool for enhancing healthcare outcomes and reducing costs, while also pointing to future research directions for refining the approach.

Furthermore, The research [7] explored the utilization of machine learning and artificial intelligence for early diabetes prediction and diagnosis, focusing on Diabetes Mellitus as a complex polygenic disorder that can lead to multiple organ failures if not properly monitored. The study delves into various methodologies for diabetes detection, concentrating on six critical areas: datasets, preprocessing techniques, feature extraction, machine learning-based analysis, classification and prediction models, and evaluation of results. It emphasizes that automated methods for diabetes prediction and diagnosis using machine learning offer superior accuracy and efficiency compared to manual methods, highlighting the importance of this technology in the research community. The study describes the analytical process, which includes data cleaning, managing missing values, exploratory analysis, and culminates in model creation and testing. The achieved high accuracy on public testing datasets showcases the approach's potential in predicting diabetes outcomes. However, it mentions an apparent error in the results currently related to Google Play Store review classification, suggesting future research to enhance machine learning models by integrating them with cloud technologies and optimizing them for artificial intelligence applications, reinforcing the potential of these techniques in advancing healthcare diagnostics and research.

The study [8] investigated the application of machine learning (ML) models in predicting microvascular and macrovascular complications in adults with Type 2 diabetes, addressing the rising prevalence of diabetes and the complexity of involved data. The study conducts a systematic review using major databases, adhering to the PRISMA guidelines, to assess the performance of ML models specifically developed or validated for predicting these diabetes-related complications. The review comprises 32 studies and 87 ML models, with neural networks being the most frequently employed technique, followed by other methods such as random forests. Common predictors across models include age, duration of diabetes, and body mass index. Performance evaluation is based on the area under the receiver operating characteristic curve (AUC), where a score above 0.75 signifies effective discrimination. Results indicate that 36% of models reached this accuracy level, and ML models often outperformed non-ML methods, with random forests demonstrating the highest effectiveness in predicting both microvascular and macrovascular complications. However, the study identifies a high risk of bias in the majority of included studies, indicating that most ML models are still in exploratory stages. While random forests showed promising results, extensive external validation is necessary before the clinical application of these ML models.

In [9], an interdisciplinary approach is explored to improve disease prediction and diagnosis, specifically targeting diabetes. The study introduces an innovative method utilizing machine learning techniques, particularly stochastic gradient descent for logistic regression, to predict diabetes in patients. A dataset comprising eight original features collected from patients prior to diabetes diagnosis is utilized. The study applies rough set theory (RST) to select the most pertinent features, finding that this selection considerably enhances prediction accuracy. It highlights the importance of precise disease prediction in healthcare, demonstrating through the Pima Indian

Diabetes dataset that stochastic gradient descent combined with RST-enhanced logistic regression surpasses traditional methods. This approach underscores the potential of integrating RST with machine learning algorithms to refine prediction models, suggesting that analogous methodologies could be applied to other algorithms for improved outcomes. The research contributes to ongoing efforts to leverage interdisciplinary techniques in healthcare, advocating further exploration of RST in combination with other machine learning methods to augment predictive accuracy in diabetes and potentially other diseases.

The study in [10] examined diabetes mellitus patient classification at a hospital in Palembang, Indonesia, using machine learning algorithms to enhance data management. It compared Naive Bayes and Support Vector Machine (SVM) using WEKA software, employing tools like cross-validation and confusion matrix to evaluate accuracy. Diabetes, characterized by high blood glucose due to insulin deficiency, complicates patient classification. The study found that SVM with a polynomial kernel achieved the highest accuracy at 96.27%, surpassing Naive Bayes at 92.07%, highlighting SVM's superior performance. The conclusion emphasizes SVM with a polynomial kernel as the best algorithm for diabetes classification in this setting. It references previous research favoring Naive Bayes due to preprocessing steps like stopword removal, which may affect data context. The study concludes that manual calculations do not always ensure classification accuracy, reinforcing the SVM approach's reliability in healthcare.

Amidst various ML algorithms, Support Vector Machines (SVM), Logistic Regression, and Naive Bayes have emerged as leading contenders for classification tasks, particularly in medical diagnostics. However, a predominant challenge associated with these algorithms is managing imbalanced datasets, where the minority class (e.g., diabetes cases) is significantly underrepresented compared to the majority class. This imbalance often leads to biased predictions, as models tend to favour the majority class, resulting in reduced sensitivity and an increased risk of misdiagnosis. To counter this issue, oversampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling have been devised. These methodologies aim to artificially balance the dataset by generating synthetic examples of the minority class, thereby enhancing the model's capability to accurately identify true positive cases. While these techniques have demonstrated significant improvements across various domains, their comparative efficacy in the context of diabetes prediction, particularly using Indonesian data, remains largely unexplored. This study conducts an in-depth analysis of the SVM, Logistic Regression, and Naive Bayes algorithms, integrating SMOTE and ADASYN to improve diabetes prediction accuracy in Indonesia. The novelty of this research lies in its focus on Indonesian diabetes data, presenting unique characteristics and patterns distinct from datasets in other regions. Through a systematic evaluation of these algorithms in conjunction with synthetic oversampling techniques, the study seeks to ascertain the most effective strategy for optimizing diabetes prediction in Indonesia. Additionally, this research contributes to a broader understanding of how oversampling methods can be effectively employed to address class imbalance issues in medical datasets, potentially informing future studies across various diseases and populations.

2. Research Methods

In this study, the research methodology adheres to the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, a comprehensive and systematic approach for data mining endeavors. The research workflow, as illustrated in Figure 1, encompasses several critical phases integral to the study's progression.

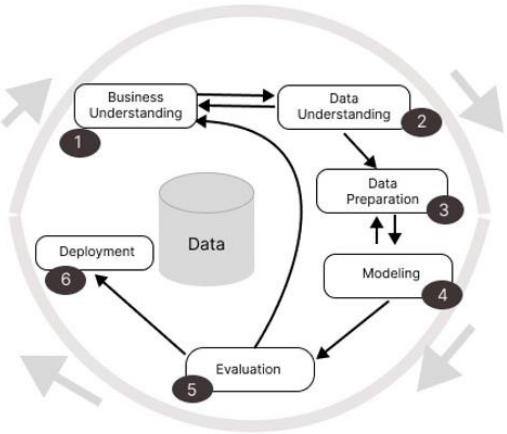

Figure 1. Stages of CRISP-DM

The CRISP-DM methodology, depicted in Figure 1, is divided into six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Prototyping [11]. This study, however, focuses specifically on the Evaluation phase, deliberately excluding the Prototyping stage to concentrate solely on assessing model performance.

The initial phase, Business Understanding, is dedicated to delineating project objectives and requirements from a business perspective [12], [13]. This stage involves defining project goals, establishing criteria for success, and ensuring that data mining efforts align with overarching business objectives to maximize relevance and impact. A thorough analysis of business needs is imperative, alongside a strategic mapping of how data mining can address identified challenges and capitalize on potential opportunities.

The Data Understanding phase entails the preliminary collection and exploration of data to garner insights into the dataset [12], [14]. This stage involves evaluating

data quality, detecting potential issues, and forming initial hypotheses regarding data relationships.

The dataset utilized in this study is a meticulously curated collection of patient data from a Regional General Hospital in Indonesia, focusing on individuals diagnosed with diabetes mellitus. This hospital's extensive reach and diverse patient demographics ensure that the data reflects a broad spectrum of cases, thus providing a comprehensive basis for predictive modelling. For this study, the dataset covers the period from 2022 to 2023, with a total of 657 patient records. The dataset includes attributes such as age, gender, family history of diabetes, Body Mass Index (BMI), blood pressure, blood sugar levels, pregnancy status, smoking habits, physical activity, and sleep patterns, as detailed in Table 1. Each attribute contributes valuable insights into the patients' health profiles, facilitating a comprehensive analysis.

Table 1. Attributes of Data

Attribute	Detail
Age	The patient's age, expressed in years, at the time of data collection.
Gender	The classification of the patient as either male or female.
Family History of Diabetes	The presence or absence of diabetes among the patient's immediate family members, such as parents or siblings.
Body Mass Index (BMI)	A metric that assesses body fat based on the patient's height and weight.
Blood Pressure	The measurement of the force exerted by blood against the arterial walls.
Blood Sugar Levels	The concentration of glucose present in the blood.
Pregnancy Status	Indicates whether the patient has ever been pregnant.
Smoking Habits	Indicates whether the patient uses tobacco products.
Physical Activities	The frequency and intensity of the patient's regular physical exercise or activities.
Sleep Patterns	The duration and quality of the patient's sleep.
Diagnosis	Indicates whether the patient has been formally diagnosed with diabetes.

In the Data Preparation phase, the focus is on data cleaning and transformation to render the dataset suitable for modelling [12], [15]. This involves addressing missing values, eliminating outliers, and performing feature engineering to enhance predictive capabilities. The data preparation process is illustrated in Figure 2. In medical datasets, particularly those related to disease diagnosis, class imbalance is a prevalent challenge, often leading to biased predictions that favour the majority class. Our dataset exhibited a significant class imbalance, with the number of diabetic patients substantially lower than non-diabetic patients. This imbalance could severely impact the model's ability to correctly identify diabetes cases. To address this, we selected SMOTE and ADASYN as our oversampling techniques. These methods were chosen because of their proven effectiveness in generating synthetic samples that enhance model performance in imbalanced scenarios, particularly in medical datasets. SMOTE was selected for its ability to generate new instances that are more representative of the minority class, while ADASYN was chosen for its capacity to adaptively create synthetic samples in areas where the decision boundary is more complex. These characteristics align well with the specific challenges of our dataset, making these techniques particularly suitable for improving the accuracy and robustness of our predictive models.

Moreover, these methods were chosen over others, such as simple random oversampling or undersampling, because they offer a more sophisticated approach that better aligns with the dataset's complexity and the need for precise, reliable predictions in a healthcare setting.

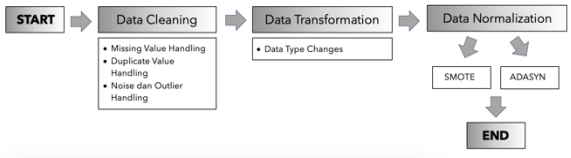

Figure 2. Data Preparation Process

SMOTE operates by selecting a minority class sample and identifying its k-nearest neighbors [16], [17]. New synthetic samples are then created along the line segments connecting the sample to its neighbors, thereby balancing the dataset and improving model performance. The SMOTE pseudocode (Pseudocode 1) details the steps for generating synthetic samples, highlighting the process of neighbor identification and interpolation.

Pseudocode 1 SMOTE

```
def smote(X, y, k_neighbors, N):
    # X: Feature matrix
    # y: Target vector
    # k_neighbors: Number of neighbors to use for generating synthetic samples
    # N: Amount of synthetic samples to generate for each minority class instance

    # Step 1: Identify minority class samples
    minority_class = get_minority_class(y)
    minority_samples = get_samples(X, y, minority_class)

    synthetic_samples = []

    for sample in minority_samples:
        # Step 2: Find k-nearest neighbors for each minority class sample
        neighbors = find_nearest_neighbors(sample, X, k_neighbors)

        for _ in range(N):
            # Step 3: Randomly select a neighbor
            neighbor = random.choice(neighbors)

            # Step 4: Generate a synthetic sample
            synthetic_sample = generate_synthetic_sample(sample, neighbor)
            synthetic_samples.append(synthetic_sample)

    # Step 5: Combine original dataset with synthetic samples
    X_augmented = concatenate(X, synthetic_samples)
    y_augmented = concatenate(y, [minority_class] * len(synthetic_samples))

    return X_augmented, y_augmented
```

Unlike other oversampling methods like SMOTE, ADASYN adaptively focuses on generating more synthetic samples in regions where the decision boundary between classes is more complex or where the minority class is underrepresented [17], [18]. This helps improve the model's ability to learn from difficult examples. ADASYN helps improve the performance of classifiers by focusing on regions of the feature space where the model has difficulty learning, thus enhancing the model's ability to generalize better to unseen data. The detailed stages of ADASYN are shown in Pseudocode 2.

Pseudocode 2 ADASYN

```
def adasyn(X, y, k_neighbors, beta):
    # X: Feature matrix
    # y: Target vector
    # k_neighbors: Number of neighbors to use for generating synthetic samples
    # beta: Control parameter for the number of synthetic samples

    # Step 1: Identify minority and majority class samples
    minority_class = get_minority_class(y)
    majority_class = get_majority_class(y)

    # Step 2: Find k-nearest neighbors for each minority class sample
    minority_samples = get_samples(X, y, minority_class)
    synthetic_samples = []

    for sample in minority_samples:
        neighbors = find_nearest_neighbors(sample, X, k_neighbors)

        # Step 3: Compute the density of the sample's neighborhood
        density = compute_density(neighbors, y, majority_class)

        # Step 4: Generate synthetic samples based on the density
        num_synthetic = int(beta * density)
        for _ in range(num_synthetic):
            synthetic_sample = generate_synthetic_sample(sample, neighbors)
            synthetic_samples.append(synthetic_sample)

    # Step 5: Combine original dataset with synthetic samples
    X_augmented = concatenate(X, synthetic_samples)
    y_augmented = concatenate(y, [minority_class] * len(synthetic_samples))

    return X_augmented, y_augmented
```

The Modeling phase involves the implementation of various algorithms to develop and validate predictive or descriptive models. This iterative process encompasses three key algorithms: Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression. Each algorithm will be utilized to build and test models, with an emphasis on their respective methodologies and performance metrics.

SVM is a powerful supervised machine learning algorithm primarily used for classification tasks [19]. SVM works by finding the optimal hyperplane that separates data points from different classes with the maximum margin [20]. This optimal hyperplane is defined as the one that maximizes the distance between the nearest data points of each class, known as support vectors. The mathematical formulation of SVM involves solving the following optimization problem as shown in Equation 1.

$$min_{w,b}\,\frac{1}{2}\|w\|^2, \tag{1}$$

subject to $y_i(w \cdot x_i + b) \geq 1, \forall_i$

w is the weight vector, b is the bias term, x_i is a data point and y_i is the class label typically -1 or 1.

Naïve bayes is a probabilistic classification algorithm based on Bayes' Theorem [16]. It assumes that the features are conditionally independent given the class label, which simplifies the computation of probabilities. Despite its simplicity, Naïve Bayes is highly effective and is commonly used in text classification tasks, such as spam detection and sentiment analysis. The core idea of Naïve Bayes is to calculate the posterior probability of a class C given a set of feature $X = (x_1, x_2, \dots, x_n)$ using Bayes' Theorem in Equation 2.

$$P(C|X) = \frac{P(C)\cdot P(X|C)}{P(X)} \tag{2}$$

Since P(X) is constant for all classes, the equation can be simplified to $P(C|X) \propto P(C) \cdot \prod_{i=1}^{n} P(x_i|C)$. Where P(C) is the prior probability of class C, $P(x_i|C)$ is the likelihood of a feature x_i given class C and the product $\prod_{i=1}^{n} P(x_i|C)$ represents the assumption of independence among features.

[9]Logistic Regression is a widely used statistical method for binary classification problems. Unlike linear regression, which predicts continuous values, logistic regression predicts the probability of a binary outcome, using the logistic (sigmoid) function to map predictions to the range [0, 1]. The model estimates the probability that a given input x belongs to class 1 as follows in Equation 3.

$$P(y = 1|x) = \sigma(w \cdot x + b) = \frac{1}{1+e^{-(w\cdot x_i+b)}} \tag{3}$$

$\sigma(z) = \frac{1}{1+e^{-(z)}}$ is the sigmoid function, w is the weight vector and b is the bias term.

The Evaluation phase is crucial for assessing the performance of classification models like Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression. Key performance metrics include Accuracy, Precision, Recall (Sensitivity), and the Area Under the ROC Curve (AUC-ROC). Accuracy (Equation 4) represents the proportion of correctly predicted instances, giving a general overview of the model's performance [21], [22]. Precision (Equation 5) measures the accuracy of positive predictions by calculating the ratio of true positives to the sum of true positives and false positives [8], [17]. Recall (Equation 6) evaluates the model's ability to identify all true positives, calculated as the ratio of true positives to the sum of true positives and false negatives. AUC-ROC assesses the model's discriminative ability by plotting the true positive rate against the false positive rate, with a higher AUC indicating better performance.

To validate our models, we used k-fold cross-validation, a robust statistical method for estimating model skill. This method divides the dataset into k subsets (folds), trains the model on k−1 folds, and tests it on the remaining fold. This process repeats k times, allowing each fold to serve as a test set once. By averaging performance across all iterations, we obtain an unbiased estimate of the model's effectiveness. Combining cross-validation with a train-test split ensures that the models are not overfitted to a specific dataset partition and can generalize to new, unseen data.

Cross-validation is complemented by a train-test split, where the dataset is divided into separate training and testing sets. This method provides an initial evaluation of model performance and allows for rapid tuning and parameter adjustments. Together, these techniques create a balanced assessment framework, leveraging the strengths of each method to ensure prediction reliability and validity.

The results, derived from these validation techniques and represented in the confusion matrix (Table 2), offer a detailed breakdown of predictions, including true positives, true negatives, false positives, and false negatives. This comprehensive analysis highlights different dimensions of prediction accuracy and reliability, providing insights into the model's strengths and weaknesses.

Table 2. Confusion Matrix

	Actual	
Predicted	*Positive*	*Negative*
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

3. Results and Discussions

In our comparative analysis of SVM, Logistic Regression, and Naive Bayes algorithms for diabetes prediction, we evaluated each model with and without the application of synthetic oversampling techniques SMOTE and ADASYN. The results reveal significant variations in model performance metrics, including accuracy, precision, recall, and AUC, highlighting the impact of these techniques on predictive accuracy.

Based on Figure 3, the baseline SVM model achieved an accuracy of 95%, a precision of 95%, a recall of 97%, and an AUC of 98. These results indicate strong predictive performance, with high sensitivity to true positives. However, the introduction of SMOTE and ADASYN further enhanced the model's performance. SVM + SMOTE exhibited the highest accuracy (95.8%) and the best AUC (99.1), reflecting improved discrimination ability. Meanwhile, SVM + ADASYN showed an accuracy of 95.3% and an AUC of 99, underscoring its effectiveness in addressing class imbalance.

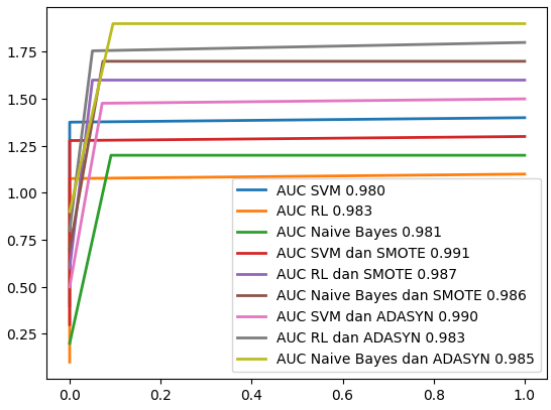

Figure 3. Result of Algorithm Evaluation

Logistic Regression demonstrated an accuracy of 94.8%, a precision of 96.2%, a recall of 96.2%, and an AUC of 98.3. This model also benefited from the use of SMOTE and ADASYN, although the improvements were not as pronounced as with SVM. Logistic Regression + SMOTE achieved an accuracy of 95.6% and an AUC of 98.7, indicating enhanced model performance compared to the baseline. In contrast, Logistic Regression + ADASYN had an accuracy of 94.5% and an AUC of 98.3, with slightly lower recall compared to SMOTE.

Naive Bayes, while exhibiting the highest precision of 98.5% in its base form, had lower recall (91.9%) and accuracy (93.5%) compared to the other algorithms. The application of SMOTE slightly improved the accuracy to 94.3% and the AUC to 98.6%. However, the recall was reduced to 90.2%, indicating that while precision remained high, the model struggled with identifying all positive cases effectively. ADASYN did not significantly enhance the model's performance, with accuracy at 93.8% and an AUC of 98.5%.

SMOTE consistently improved the performance of SVM and Logistic Regression models, particularly by addressing the class imbalance more effectively. The superior performance of SMOTE can be attributed to its ability to generate synthetic samples that closely mimic the characteristics of the minority class, which is crucial in datasets like ours, where the diabetic cases are underrepresented. For SVM, the high-dimensional feature space benefits from SMOTE's synthetic sample generation, which creates a more balanced decision boundary, leading to the highest observed accuracy and AUC. In contrast, ADASYN, while effective, focuses on more complex regions near the decision boundary, which may introduce noise in Logistic Regression models, particularly in less complex feature spaces, resulting in slightly lower improvements. The modest improvements observed in Naive Bayes with SMOTE and ADASYN suggest that probabilistic models may not fully leverage the benefits of oversampling

techniques, potentially due to their reliance on feature independence assumptions. This analysis underscores the importance of selecting oversampling techniques that align with the specific characteristics and complexity of the dataset and the models being used.

ADASYN also contributed to improved model performance, particularly with SVM, which showed a notable increase in precision compared to the baseline. However, its impact on Logistic Regression was less pronounced, with a marginal decrease in recall but similar AUC compared to SMOTE. Naive Bayes showed modest improvements with ADASYN, though not as impactful as SMOTE, suggesting that the choice between SMOTE and ADASYN may vary depending on the specific algorithm and dataset characteristics.

SVM and Logistic Regression demonstrated superior overall performance compared to Naive Bayes, with higher accuracy, recall, and AUC values. SVM + SMOTE achieved the highest metrics across all evaluated criteria, indicating its robustness in handling imbalanced datasets. Logistic Regression + SMOTE also showed strong performance but slightly lower than SVM + SMOTE. Naive Bayes, despite high precision, struggled with lower recall and accuracy, making it less effective for this specific application. Precision and recall are critical metrics for evaluating the trade-offs in predictive models. While Naive Bayes achieved the highest precision, it suffered from lower recall, indicating a tendency to miss positive cases. In contrast, SVM and Logistic Regression models, particularly with SMOTE, balanced high precision with better recall, ensuring a more comprehensive identification of diabetes cases. The AUC metric provides insight into the model's ability to distinguish between classes. SVM + SMOTE achieved the highest AUC, demonstrating the best performance in distinguishing between diabetic and non-diabetic cases. This metric underscores the importance of using synthetic oversampling techniques to enhance the model's discriminative power and overall effectiveness.

The practical implications of our research underscore the potential of machine learning models to revolutionize diabetes care in Indonesia. By enhancing early detection and identifying high-risk individuals based on historical and clinical data, these models can serve as a decision-support tool for healthcare providers, enabling more accurate assessments of diabetes risk. This capability is particularly crucial in Indonesia, where diabetes prevalence is rising rapidly due to lifestyle changes and genetic predispositions specific to the population. Early diagnosis can lead to better management of the disease, reducing the likelihood of complications such as cardiovascular diseases, neuropathy, and kidney failure.

4. Conclusions

In conclusion, this study underscores the critical role of tailored oversampling techniques, such as SMOTE, in enhancing the predictive accuracy of machine learning models for diabetes diagnosis in Indonesia. The superior performance of the SVM model with SMOTE suggests that such techniques can effectively address the class imbalance inherent in medical datasets, thereby improving the robustness of predictions in real-world applications. These results have significant implications for healthcare strategies in Indonesia, where early and accurate diagnosis is essential for managing the growing burden of diabetes. However, this study is not without limitations. The dataset used, while comprehensive, is limited to a specific regional hospital, which may not fully capture the broader diversity of the Indonesian population. Additionally, the reliance on synthetic oversampling techniques like SMOTE and ADASYN, while effective, may introduce synthetic data points that do not fully reflect the complexity of real-world scenarios. Future research should explore the integration of more advanced techniques, such as hybrid sampling methods or deep learning approaches, to further refine prediction accuracy. Moreover, it is recommended that future studies validate these models using datasets from multiple regions and healthcare settings across Indonesia to ensure the generalizability of the results. Further investigation into the impact of other demographic and lifestyle factors specific to Indonesia could also provide deeper insights into the nuances of diabetes prediction in this unique context. By addressing these limitations and expanding the scope of research, the potential for developing robust, context-specific predictive models can be further realized, ultimately contributing to better healthcare outcomes.

Acknowledgements

This study was funded by the Directorate of Technology Research and Community Service (DRTPM) Ministry of Education Culture and Research and Technology of the Republic of Indonesia (Kemendikbud RI).

References

[1] World Health Organization, "Diabetes." Accessed: Mar. 24, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] R. Walker, *Take Control of Your Diabetes*. 2020.

[3] Kementrian Kesehatan Republik Indonesia, "Diabetes." Accessed: Mar. 24, 2024. [Online]. Available: https://p2ptm.kemkes.go.id/informasi-p2ptm/penyakit-diabetes-melitus

[4] L. Poretsky, "Diabetes Management in Hospitalized Patients," 2023. doi: https://doi.org/10.1007/978-3-031-44648-1.

[5] D. M. S. Rao and D. Sai. Sridhathri, "Diabetes Mellitus Prediction Using Ensemble Machine Learning Techniques," *ITM Web of Conferences*, vol. 56, 2023, doi: 10.1051/itmconf/20235605015.

[6] Y. Granillo and G. H. Goldsztein, "Machine Learning as a Tool to the Diagnosis of Diabetes," *Journal of Student Research*, vol. 11, no. 1, 2022, doi: 10.47611/jsrhs.v11i1.2513.

[7] M. Ranjit Reddy, P. Lakshmi Sagar, and N. S. Shaik, "Diabetes Mellitius Detection and Self Management based on Machine Learning," *J Pharm Negat Results*, vol. 13, no. 4, 2022, doi: 10.47750/pnr.2022.13.04.138.

[8] K. R. Tan *et al.*, "Evaluation of Machine Learning Methods Developed for Prediction of Diabetes Complications: A

Systematic Review," *J Diabetes Sci Technol*, vol. 17, no. 2, 2023, doi: 10.1177/19322968211056917.

[9] K. M. Kaka-Khan, H. Mahmud, and A. A. Ali, "Rough Set-Based Feature Selection for Predicting Diabetes Using Logistic Regression with Stochastic Gradient Decent Algorithm," *UHD Journal of Science and Technology*, vol. 6, no. 2, 2022, doi: 10.21928/uhdjst.v6n2y2022.pp85-93.

[10] H. Apriyani and K. Kurniati, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," *Journal of Information Technology Ampera*, vol. 1, no. 3, 2020, doi: 10.51519/journalita.volume1.isssue3.year2020.page133-143.

[11] A. F. N. Masruriyah, C. E. Sukmawati, and B. A. Dermawan, "Memahami Data Mining dengan Python: Implementasi Praktis."

[12] R. Raja, K. K. Nagwanshi, S. Kumar, and K. R. Laxmi, *Data Mining and Machine Learning Applications*. 2022.

[13] M. J. Zaki and M. Wagner Jr, *Data Mining and Machine Learning Fundamental Concepts and Algorithms*. 2020.

[14] X.-S. Yang, *Introduction to Algorithms for Data Mining and Machine Learning*. 2019.

[15] J. N.P. and R. Aruna, "Big data analytics in health care by data mining and classification techniques," *ICT Express*, no. xxxx, 2021, doi: 10.1016/j.icte.2021.07.001.

[16] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[17] A. F. N. Masruriyah, H. Y. Novita, C. E. Sukmawati, A. Fauzi, D. Wahiddin, and H. H. Handayani, "Thorough Evaluation of the Effectiveness of SMOTE and ADASYN Oversampling Methods in Enhancing Supervised Learning Performance for Imbalanced Heart Disease Datasets," in *International Conference on Informatics and Computing (ICIC)*, Institute of Electrical and Electronics Engineers, 2023.

[18] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Scientific Journal of Informatics*, vol. 8, no. 2, 2021, doi: 10.15294/sji.v8i2.32484.

[19] D. A. Pisner and D. M. Schnyer, *Support vector machine*. Elsevier Inc., 2020. doi: 10.1016/B978-0-12-815739-8.00006-7.

[20] H. Hikmayanti, A. F. Nurmasruriyah, A. Fauzi, N. Nurjanah, and A. Nur Rani, "Performance Comparison of Support Vector Machine Algorithm and Logistic Regression Algorithm," *International Journal of Artificial Intelegence Research*, vol. 7, no. 1, p. 1, 2023, doi: 10.29099/ijair.v7i1.1.1114.

[21] A. F. N. Masruriyah, H. Y. Novita, and C. E. Sukmawati, "Performance Evaluation of Popular Supervised Learning Algorithms Towards Cardiovascular Disease," vol. 8, no. 3, pp. 420–426, 2023, doi: 10.32493/informatika.v8i3.34103.

[22] S. Hameetha Begum and S. N. Nisha Rani, "Model Evaluation of Various Supervised Machine Learning Algorithm for Heart Disease Prediction," in *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021*, 2021. doi: 10.1109/ICSECS52883.2021.00029.

