International Journal of Communication Networks and Information Security
2024, 16(4)
ISSN: 2073-607X,2076-0930
https://ijcnis.org/

Research Article

Leveraging Machine Learning and SMOTE for Diabetes Prediction:Implementation of an Application Based on Indonesian Hospital Data

Arief Wibowo[1*], Anis Fitri Nur Masruriyah[2], Selly Rahmawati[1]

[1]Department of Information System, Universitas Budi Luhur, Jakarta, 12260, Indonesia
[2]Department of Informatics, Universitas Pembangunan Nasional Veteran Jakarta, Jakarta, 12450, Indonesia

***Corresponding Author:** arief.wibowo@budiluhur.ac.id

Citation: Arief Wibowo, Anis Fitri Nur Masruriyah and Selly Rahmawati, "Leveraging Machine Learning and SMOTE for Diabetes Prediction: Implementation of an Application Based on Indonesian Hospital Data," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 16, no 4,2024, pp.1033-1041.

ARTICLE INFO

Received: 17 Aug 2024
Accepted: 30 sep 2024

ABSTRACT

Diabetes mellitus is a widespread chronic condition affecting millions globally, including a substantial population in Indonesia. Accurate and early detection is critical for effective management and treatment, and machine learning offers promising solutions for enhancing predictive accuracy. This study evaluates three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, and Naive Bayes, with and without the application of Synthetic Minority Over-sampling Technique (SMOTE) to tackle data imbalance. Data were meticulously collected from an Indonesian regional hospital, including various medical parameters such as age, body mass index (BMI), blood sugar levels, blood pressure, and family history. Our findings reveal that the SVM model, without SMOTE, achieved an accuracy of 95%, precision of 95%, recall of 97%, and an AUC of 98%. With SMOTE, SVM's performance improved to an accuracy of 95.8%, precision of 97%, recall of 94.6%, and an AUC of 99.1%. Logistic Regression without SMOTE demonstrated an accuracy of 94.8%, precision of 96.2%, recall of 96.2%, and an AUC of 98.3%, while with SMOTE, it reached an accuracy of 95.6%, precision of 97.9%, recall of 93.3%, and an AUC of 98.7%. The Naive Bayes model showed an accuracy of 93.5%, precision of 98.5%, recall of 91.9%, and an AUC of 98.1%, improving with SMOTE to an accuracy of 94.3%, precision of 98.3%, recall of 90.2%, and an AUC of 98.6%. The best-performing model, SVM with SMOTE, was implemented into a desktop application. This application successfully validated the model's predictive capabilities, demonstrating effective and accurate diabetes detection in practical scenarios. Our study highlights the significant impact of SMOTE on enhancing model performance and emphasizes the importance of sophisticated machine learning techniques in advancing healthcare diagnostics. This work provides a foundation for further development and deployment of predictive models in clinical settings, contributing to improved patient care and disease management.

Keywords: Clinical Decision Support, Diabetes Prediction, Healthcare Diagnostics, Imbalanced Data, Machine Learning, Synthetic Minority Over-sampling Technique

INTRODUCTION

Diabetes mellitus is a primary public health concern worldwide, affecting millions of individuals and posing a significant burden on healthcare systems [1], [2]. It is characterized by high blood sugar levels that, if left untreated, can lead to severe complications, including cardiovascular diseases, kidney failure, and neuropathy [3]. The prevalence of diabetes has been steadily increasing, especially in developing countries like Indonesia, where lifestyle changes and genetic predispositions contribute to its rise [4].

In the context of health management, early prediction of chronic diseases such as diabetes significantly improves individual patient outcomes. It plays a crucial role in disease control strategies at the population level. Machine

learning technology in health prediction has the potential to transform how health systems respond to and manage chronic diseases. By integrating prediction results into health management systems, hospitals, and clinics can focus resources on population groups most at risk, allowing for earlier and more targeted interventions. This approach reduces the disease burden and optimizes the allocation of often limited health resources [5]. Therefore, this study not only focuses on the technical aspects of diabetes prediction but also considers its broader impact on health management.

Traditional methods of diabetes diagnosis often rely on clinical evaluations and laboratory tests, which, although accurate, can be time-consuming and resource-intensive [2], [3]. The advent of machine learning techniques has opened new avenues for early and accurate prediction of diabetes, offering the potential for timely interventions and personalized treatment plans [6], [7], [8].

Research [6] emphasizes Diabetes Mellitus as a common long-term hormonal disorder, affecting all ages and influenced by genetics and lifestyle choices. It's reported that 68% of the country's population is affected, making early detection crucial to avoid complications. The study compares machine learning methods (K-Nearest Neighbors, Naive Bayes, XGBoost, Decision Tree, and Random Forest) for diabetes prediction. While Random Forest showed initial promise, XGBoost ultimately outperformed others with a 77% precision rate. This demonstrates machine learning's potential for early healthcare intervention and better disease management. Ongoing research in this area is key to improving patient outcomes and healthcare approaches.

Importantly, the research [7] utilized supervised learning and logistic regression to create a diabetes diagnostic model. The study's aim to predict diabetes diagnoses from patient data underscores the potential of data-driven predictions in healthcare. While the logistic regression model achieved a 74% accuracy rate, the study emphasizes the need for more sophisticated models like neural networks and additional features to enhance accuracy. This research underscores the potential of machine learning in disease diagnosis and prevention, suggesting it as a valuable tool for improving healthcare outcomes and reducing costs, while also indicating future research directions to refine the approach.

Moreover, the study [8] investigated the use of machine learning (ML) models to predict complications in adults with Type 2 diabetes, given the increasing prevalence of diabetes and the complexity of related data. The study conducted a systematic review using major databases and PRISMA guidelines to evaluate the performance of ML models specifically designed or validated for predicting these complications. The review included 32 studies and 87 ML models, with neural networks being the most common technique. Common predictors included age, diabetes duration, and body mass index. Performance was assessed using the area under the receiver operating characteristic curve (AUC), with a score above 0.75 indicating effective discrimination. Results showed that 36% of models achieved this level of accuracy, often outperforming non-ML methods, with random forests being the most effective in predicting both microvascular and macrovascular complications. However, the study identified a high risk of bias in most studies, suggesting that most ML models are still exploratory. While random forests showed promise, the study emphasizes the crucial need for extensive external validation before its clinical application.

This study introduces a groundbreaking application of machine learning algorithms—Support Vector Machine (SVM), Logistic Regression, and Naive Bayes—combined with the Synthetic Minority Over-sampling Technique (SMOTE) to predict diabetes mellitus. Utilizing data from an Indonesian regional hospital, this research addresses the critical issue of class imbalance, significantly enhancing model accuracy. By adapting these predictive models to the unique characteristics of the Indonesian population, this study represents a pioneering effort in localized healthcare diagnostics, setting a new standard for machine learning applications in the Indonesian healthcare context.

Data for this research were collected from a regional hospital in Indonesia, including a diverse set of medical parameters critical for accurate diabetes prediction. The study aims to evaluate the effectiveness of these algorithms, both with and without SMOTE, to identify the optimal approach for diabetes detection.

Furthermore, the best-performing model is implemented in a desktop application, providing a practical tool for healthcare professionals to utilize in real-world settings. This application demonstrates the model's predictive accuracy and highlights the potential of integrating machine learning solutions into healthcare diagnostics. Through this research, we seek to contribute to the scientific community's understanding of diabetes prediction and its practical implementation, paving the way for improved patient care and management.

METHODOLOGY

This study applies the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to develop a machine learning-based desktop application for predicting diabetes using hospital data from Indonesia. CRISP-DM (Figure 1) is a widely accepted methodology for data mining and machine learning projects [9]. It provides a structured approach to planning and executing data-driven projects, ensuring the process is systematic and repeatable [10]. The CRISP-DM framework comprises six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase is iterative, allowing for refinement and improvement as the project progresses.

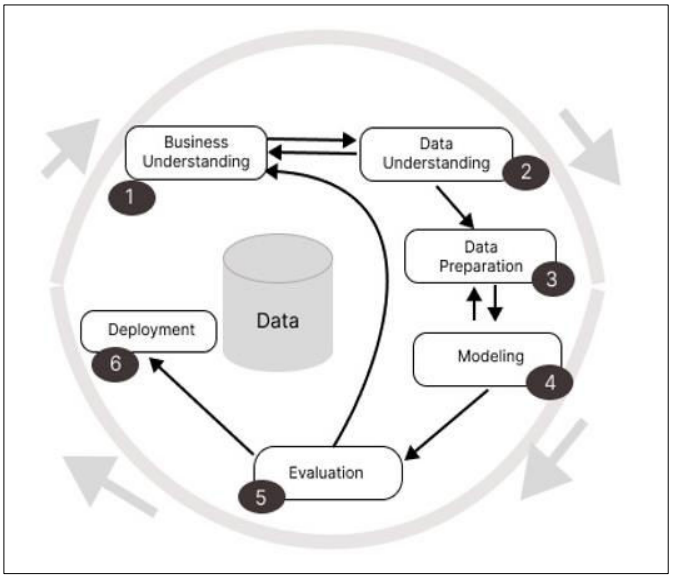

Figure 1. Stages of CRISP-DM

A. Business Understanding

The primary objective of this research is to enhance diabetes prediction accuracy by leveraging machine learning algorithms and synthetic oversampling techniques, specifically SMOTE. Given the rising prevalence of diabetes in Indonesia, accurate prediction models can significantly improve early diagnosis and treatment planning, ultimately reducing the disease burden on healthcare systems. Healthcare professionals can readily access these predictive insights by developing a user-friendly desktop application, facilitating informed decision-making [11].

B. Data Understanding

The dataset utilized in this study was collected from one of the Indonesian general hospitals and comprises demographic and clinical features pertinent to diabetes diagnosis, including age, gender, BMI, blood pressure, glucose levels, and family history of diabetes (Table 1). The dataset was carefully analyzed to understand its structure, distribution, and any inherent class imbalances common in medical datasets. The initial analysis revealed a significant imbalance between diabetic and non-diabetic instances, necessitating the application of oversampling techniques like SMOTE to ensure model robustness and reliability.

Table1. 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.

Attribute	Detail
Sleep Patterns	The duration and quality of the patient's sleep.
Diagnosis	Indicates whether the patient has been formally diagnosed with diabetes.
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.

C. Data Preparation

Data preparation involved cleaning, transforming, and organizing the dataset for analysis [12]. Missing values were handled using imputation techniques, and categorical variables were encoded appropriately [13]. Feature scaling was applied to normalize the numerical features, ensuring compatibility with algorithms like SVM sensitive to feature scaling. To address the class imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic instances for the minority class, effectively balancing the class distribution without merely duplicating existing instances.

D. Modelling

In the modeling phase, three machine learning algorithms were selected for evaluation: Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression. Each algorithm offers unique advantages and is suited for different aspects of classification tasks. SVM, a robust classification algorithm, constructs an optimal hyperplane to separate classes with maximum margin [14], [15], [16]. The model was trained using a radial basis function (RBF) kernel, a powerful tool for handling non-linear relationships in the data. The optimization problem solved by SVM is given by Equation (1). Where w is the weight vector, b is the bias term, xi is a data point and yi is the class label typically -1 or 1. For Naïve Bayes, a Gaussian distribution was assumed for continuous variables, with smoothing parameters set to handle any zero probabilities. The SVM model parameters, such as the regularization parameter (C) and kernel coefficient (γ), were optimized using grid search. Each algorithm's hyperparameters were fine-tuned based on cross-validation results to achieve optimal performance. Decision Trees were pruned to prevent overfitting, and performance was evaluated using accuracy, precision, recall, and F1-score to ensure a robust comparison across techniques.

$$min_{w,b}\frac{1}{2}\|w\|^2, \quad \text{subject to } y_i(w \cdot x_i + b) \geq 1, \forall_i \tag{1}$$

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem, assuming conditional independence among features [17], [18]. The model calculates the posterior probability of each class and assigns the class with the highest probability to the instance. For this study, the Gaussian Naïve Bayes variant was employed, suitable for continuous features. The core idea of Naïve Bayes is to calculate the posterior probability of a class C given a set of feature $X = (x_1, x_2, \ldots, x_n)$ 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 feature x_i given class C and the product $\prod_{i=1}^{n} P(x_i|C)$represents the assumption of independence among features.

Logistic Regression is a linear model for binary classification, estimating the probability of a class using the logistic function [15], [16], [19]. It models the log-odds of the probability as a linear combination of the input features. The logistic function is defined as Equation (3).

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

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

E. Evaluation

The models were evaluated using a variety of metrics to assess their performance comprehensively. The

primary metrics considered were accuracy (4), precision (5), recall (6), and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). A confusion matrix (Table 1) was constructed to gain insights into the distribution of true positives, false positives, true negatives, and false negatives. Additionally, 10-fold cross-validation was implemented to ensure the robustness and generalizability of the results, mitigating the risk of overfitting and bias.

Table2. Confussion Matrix

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

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

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

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

F. Deployment

Upon successful evaluation, the best-performing model was integrated into a desktop application designed for healthcare practitioners in Indonesia. The application, built using a user-friendly interface, allows users to input patient data and receive real-time diabetes risk predictions. This tool supports clinical decision-making, providing actionable insights based on sophisticated machine learning algorithms. The application was tested in a real-world healthcare setting, ensuring its reliability and effectiveness in predicting diabetes outcomes accurately.

RESULTSAND DISCUSSION

The research, conducted with a keen understanding of the local context, utilized the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. This structured approach guided the analysis from problem understanding to deployment, with the primary objective of developing a reliable diabetes prediction model tailored to the Indonesian population. We leveraged medical records from a Regional General Hospital in Indonesia and conducted a rigorous evaluation of three machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, and Naive Bayes. We also explored the impact of the Synthetic Minority Over-sampling Technique (SMOTE) on addressing class imbalance in the dataset.

A detailed comparative analysis was conducted to evaluate the performance of the selected machine learning algorithms, including Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression. Standard performance metrics such as precision, recall, F1-score, and accuracy were used to assess each model's effectiveness. Additionally, confusion matrices were generated for each algorithm to visualize the true positives, false positives, true negatives, and false negatives, providing deeper insight into classification performance. The SVM demonstrated high precision and F1-score, particularly in handling imbalanced classes, whereas Naïve Bayes achieved faster computational time but slightly lower recall. Logistic Regression provided a balanced performance across most metrics but exhibited lower precision in comparison to SVM. This analysis highlights the strengths and weaknesses of each approach, offering valuable insight into their applicability for similar datasets.

In the initial phase of business understanding, the research identified the pressing need for accurate diabetes prediction tools in Indonesia, given the disease's rising prevalence and the local population's specific characteristics. This understanding informed the decision to focus on models that could effectively handle the imbalanced nature of medical datasets, where positive cases of diabetes are often underrepresented. The data used in the study was carefully collected and prepared, ensuring that it was representative of the broader patient population and contained all relevant features, including health history, examination results, and treatment records. Due to the imbalance in the data, as shown in Figure 2, we applied an oversampling technique.

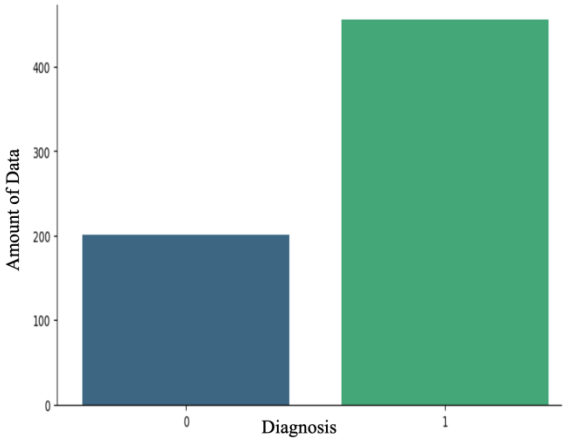

Figure 2. Data Based on Class Target

The data preparation phase was critical, particularly the application of SMOTE, which aimed to balance the dataset by generating synthetic instances of the minority class. This process was essential in mitigating the risk of biased model predictions that could disproportionately favor the majority class (Figure 3).

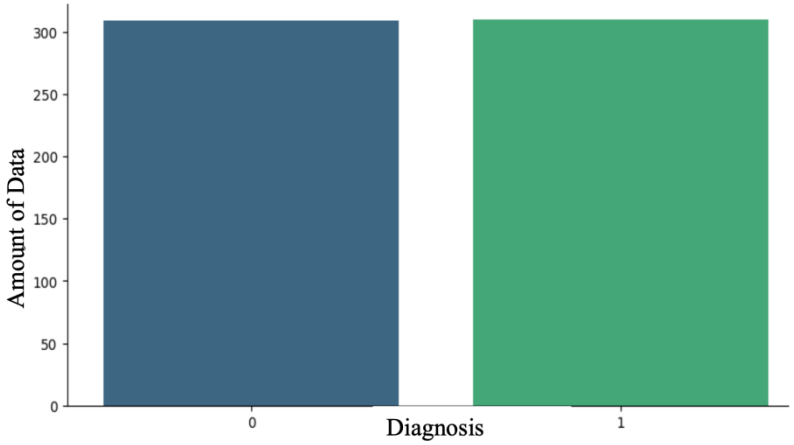

Figure 3. Data After SMOTE

The prepared data was then used to train the selected machine learning models, ensuring that each algorithm had the opportunity to learn from a balanced dataset, which is crucial for improving model generalization and accuracy.

The modelling phase involved the implementation and evaluation of SVM, Logistic Regression, and Naive Bayes algorithms. Each model was trained on both the original and SMOTE-balanced datasets to assess the impact of synthetic oversampling on their predictive performance. The results (Figure 4) demonstrated that SVM, when combined with SMOTE, produced the most accurate predictions, achieving an accuracy of 95.8%, precision of 97%, recall of 94.6%, and an AUC of 99.1. This high level of performance indicates that SVM, supported by SMOTE, is particularly well-suited for the task of diabetes prediction within this specific population. Logistic Regression with SMOTE also showed strong performance, with an accuracy of 95.6%, while Naive Bayes, although precise, had slightly lower overall accuracy.

The evaluation phase was conducted using cross-validation and train-test split techniques, ensuring that the models' performance metrics were reliable and generalizable. The confusion matrix provided a clear depiction of each model's strengths and weaknesses, confirming the superiority of the SVM with SMOTE model in this context. The high AUC score further emphasized the model's ability to discriminate between positive and negative cases effectively, making it a strong candidate for practical deployment in healthcare settings.

In the final deployment phase, the decision was made to implement the SVM with SMOTE model as a desktop application designed for use in Indonesian hospitals (Figure 3). This application will provide healthcare professionals with a user-friendly tool to predict diabetes risk based on patient data, facilitating early diagnosis and intervention.

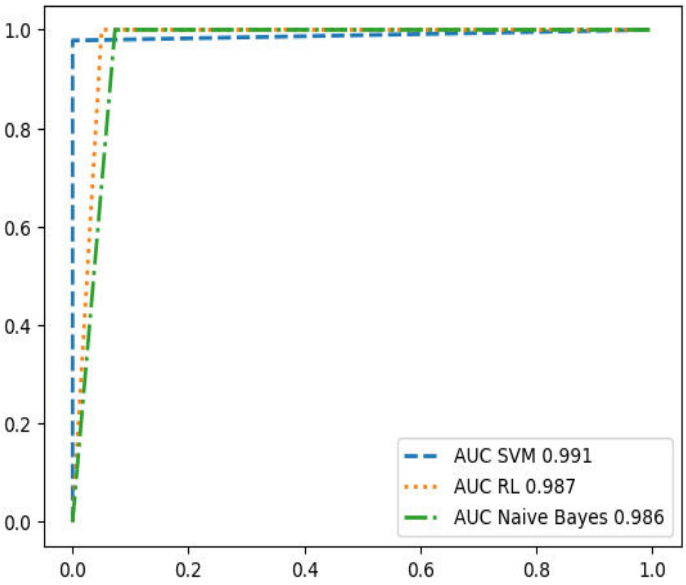

Figure 4. Modelling with SMOTE

The application of machine learning algorithms in diabetes prediction improves the accuracy of early diagnosis and opens up opportunities for broader integration in health management. By incorporating these prediction results into electronic health record (EHR) systems, hospitals and clinics can automate reminders for diabetes control, enabling more effective risk-based care management and personalized care plans. Previous studies have shown that this kind of integration of health information technology can reduce long-term care costs and significantly improve patient outcomes [20]. Therefore, applying the prediction model developed in this study is expected to support national health programs, especially in the context of chronic disease control in Indonesia, more effectively and efficiently.

The introduction of this model as a desktop application marks a significant stride in enhancing diabetes management in Indonesia. It offers a practical solution that can be seamlessly integrated into current healthcare workflows, thereby streamlining the process of diabetes diagnosis and management.

The implementation of the SVM with the SMOTE model as a desktop application, as illustrated in Figure 5, represents a practical outcome of this research designed to facilitate the diagnosis of diabetes within the Indonesian healthcare system. The application allows healthcare providers to input patient data, including age, gender, history of diabetes, body mass index, blood pressure, blood sugar levels, pregnancy status, smoking habits, exercise (sports activity), and sleep patterns. Based on these inputs, the model processes the data and predicts whether the patient is likely to have diabetes, as indicated in the "Result" section.

This interface is straightforward, making it accessible for medical professionals who may not have extensive training in machine learning or data analysis. The application effectively translates complex model predictions into a user-friendly format that provides immediate feedback, assisting clinical decision-making. Including multiple patient parameters in the interface reflects the model's ability to consider a range of factors contributing to diabetes risk, ensuring a comprehensive assessment.

In the broader context of the CRISP-DM framework, this deployment phase underscores the significance of transforming data mining results into actionable tools that can be seamlessly integrated into existing medical workflows. The development of this desktop application demonstrates the successful transition from theoretical model development to practical application, offering a valuable resource for enhancing the accuracy and efficiency of diabetes diagnosis in Indonesia. The model's high accuracy and robust performance, as discussed earlier, ensure that the predictions provided by the application are reliable and can significantly contribute to early detection and intervention efforts.

This practical implementation also sets the stage for future enhancements, such as integrating additional data sources or refining the user interface to include more detailed patient history or lifestyle factors. Moreover, this application could be expanded to other regional hospitals across Indonesia, providing a standardized tool for diabetes prediction that leverages local data, ultimately contributing to improved patient outcomes on a national scale.

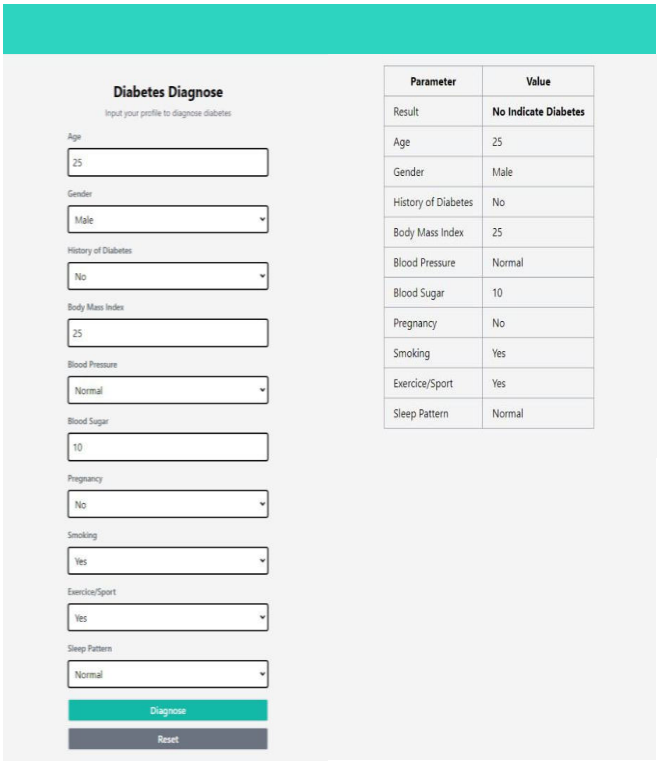

Figure 5. Desktop Application

In addition to providing accurate predictions, the developed application also has the potential to support decision-making in health policy. With integrated prediction data, policymakers can allocate health resources more efficiently, especially in diabetes prevention and management programs in Indonesia. Using machine learning-based prediction models also allows early detection in high-risk populations so interventions can be carried out before severe complications occur. In the long term, this approach can reduce the economic burden on the national health system by reducing the incidence of chronic diseases that require intensive treatment [21]. Thus, implementing this technology is an innovation in medical diagnosis and a strategic tool in broader public health management.

CONCLUSION

This study presents a comprehensive evaluation of machine learning algorithms for diabetes mellitus prediction, utilizing data from an Indonesian regional hospital. Our findings demonstrate the significant impact of SMOTE on enhancing the performance of Support Vector Machine (SVM), Logistic Regression, and Naive Bayes models. The SVM model with SMOTE emerged as the best-performing approach, achieving remarkable accuracy, precision, recall, and AUC scores.

The successful implementation of the SVM with the SMOTE model into a desktop application underscores the practical applicability of machine learning in healthcare settings. This application provides healthcare professionals a reliable tool for early and accurate diabetes detection, facilitating timely interventions and personalized treatment strategies.The research emphasizes the importance of sophisticated data augmentation techniques like SMOTE in addressing class imbalance, a common challenge in medical datasets. By improving model performance, these techniques contribute to more accurate and reliable predictions, ultimately enhancing patient care.

Future work could focus on expanding the dataset to include additional medical parameters and exploring other advanced machine-learning techniques to refine prediction models further. Additionally, the potential for integrating this application into broader healthcare systems is vast, inspiring a vision of a seamless experience for practitioners and a more efficient and effective healthcare delivery.

In conclusion, this study not only highlights the potential of machine learning as a powerful tool in healthcare diagnostics but also reassures the audience of its relevance and impact. By bridging the gap between scientific research and practical application, we aim to make a meaningful impact on diabetes management and patient outcomes in Indonesia and beyond.

REFERENCES

[1] World Health Organization, "Diabetes." Accessed: Jun. 10, 2021. [Online]. Available: https://www.who.int/health-topics/diabetes#tab=tab_1

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

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

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

[5] E. A. Balas, S. Weingarten, C. T. Garb, D. Blumenthal, S. A. Boren, and G. D. Brown, "Improving preventive care by prompting physicians," Arch Intern Med, vol. 160, no. 3, 2000, doi: 10.1001/archinte.160.3.301.

[6] 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.

[7] 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.

[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] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model," Procedia CIRP, vol. 79, pp. 403–408, 2019, doi: 10.1016/j.procir.2019.02.106.

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

[11] 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.

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

[13] M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," IEEE Access, vol. 7, pp. 106111–106123, 2019, doi: 10.1109/ACCESS.2019.2930410.

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

[15] 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.

[16] 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.

[17] M. A. Awal, M. Masud, M. S. Hossain, A. A. M. Bulbul, S. M. H. Mahmud, and A. K. Bairagi, "A Novel Bayesian Optimization-Based Machine Learning Framework for COVID-19 Detection from Inpatient Facility Data," IEEE Access, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3050852.

[18] 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.

[19] 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.

[20] S. Koch, "Home telehealth-Current state and future trends," Int J Med Inform, vol. 75, no. 8, 2006, doi: 10.1016/j.ijmedinf.2005.09.002.

[21] D. Blumenthal and M. Tavenner, "The 'Meaningful Use' Regulation for Electronic Health Records," New England Journal of Medicine, vol. 363, no. 6, 2010, doi: 10.1056/nejmp1006114.