International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 5, Number 2, July 2026, *Page. 2346 - 2354*
Email : editorijhess@gmail.com

Explainable AI-Based Diabetes Mellitus Risk Prediction Using Naive Bayes and Support Vector Machine

Putu Hawariyah[1)]*, Sudin Saepudin[2)], Gina Syabani Yuda[3)]
[1,2,3)] Department of Information Systems, Faculty of Engineering, Computers and Design, Nusa Putra University, Sukabumi, Indonesia

*Corresponding Author
E-mail: putu.hawariyah_si22@nusaputra.ac.id

Abstract

Diabetes Mellitus (DM) is a chronic metabolic disease that continues to increase globally and requires effective early detection to reduce the risk of serious complications. Machine learning has been widely adopted as an approach for predicting diabetes risk; however, most existing models are still black-box in nature, making them difficult to interpret and less useful for clinical decision-making. In addition, the problem of class imbalance in medical datasets often causes models to be biased toward the majority class, reducing their sensitivity in detecting high-risk patients. This study aims to develop and compare diabetes risk prediction models using the Naive Bayes and Support Vector Machine (SVM) algorithms with an Explainable Artificial Intelligence (XAI) approach. Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE) applied to the training data. Interpretability was analyzed using SHAP (SHapley Additive exPlanations) for global feature importance and LIME (Local Interpretable Model-agnostic Explanations) for local instance-level explanations. The dataset used was the Diabetes Health Indicators Dataset from the BRFSS 2015 survey, publicly available on Kaggle, with a sample of 50,000 records and 22 variables. Evaluation results showed that SVM achieved an accuracy of 84.17%, while Naive Bayes achieved a higher recall of 77.76%, indicating better sensitivity in detecting diabetes cases. SHAP analysis identified GenHlth, HighBP, BMI, HighChol, and Age as the most influential risk factors globally, while LIME provided individual-level explanations. This research contributes a prediction model that is not only accurate but also transparent and clinically interpretable.

Keywords: Diabetes Mellitus, Explainable Artificial Intelligence, Naive Bayes, Support Vector Machine.

INTRODUCTION

Diabetes Mellitus (DM) is a chronic metabolic disease characterized by persistently elevated blood glucose levels and has become a significant public health problem globally, including in Indonesia (Nasution 2021). Based on a report by the Ministry of Health of the Republic of Indonesia in 2023, the prevalence of diabetes in the population aged 15 years and above reached 11.7% and is projected to continue increasing in the coming years (Ade Nasihudin, 2024). DM is often referred to as a silent killer because in the early stages it does not cause obvious symptoms, but can develop into serious complications such as cardiovascular disease, kidney failure, retinopathy, neuropathy, and stroke, which reduce the quality of life while increasing the burden of healthcare costs (Erlin 2022). Therefore, early detection and identification of high-risk individuals is a crucial step in diabetes prevention and management.

The development of machine learning has encouraged the use of various algorithms, such as Naive Bayes and Support Vector Machine (SVM), to estimate the risk of diabetes based on health indicators and lifestyle factors (Wantoro 2021). A number of studies have proven that machine learning models are capable of producing high predictive performance, making them potential tools for early diabetes detection (Erlin 2022). However, there are two main challenges that still limit their clinical application. First, class imbalance in medical datasets causes models to be less sensitive to the minority class, risking the misclassification of patients who actually have diabetes (Dikan Ismafillah 2023). Second, the black-box nature of machine learning models makes it difficult for medical professionals to validate and trust predictive results as a basis for clinical decision-making (Handayani 2025).

In an effort to overcome these two challenges, this study integrates class imbalance handling with the Explainable Artificial Intelligence (XAI) approach using SHAP (SHapley Additive

exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) methods in a diabetes risk prediction model based on Naive Bayes and SVM. The novelty of this study lies in the simultaneous combination of two XAI methods to increase model sensitivity toward the minority class while producing a comprehensive and consistent interpretation of the factors determining diabetes risk (Erlin 2022). Thus, this study is expected to produce a prediction model that is not only accurate but also transparent and clinically accountable.

Previous studies have explored the use of machine learning for diabetes prediction; however, most focus solely on predictive accuracy without addressing interpretability or class imbalance simultaneously (Dikan Ismafillah 2023; Rahmawati 2024). This study fills that gap by combining SMOTE for class imbalance handling with both SHAP and LIME for dual-layer interpretability, providing a more holistic framework for diabetes risk prediction in clinical settings.

This paper is structured as follows: Section 2 presents the literature review, Section 3 describes the research method, Section 4 presents the results and discussion, and Section 5 provides the conclusion.

Literature Review

Diabetes Mellitus

Diabetes mellitus is a syndrome characterized by hyperglycemia due to insulin deficiency and is one of the major health problems worldwide (Askandar Tjokroprawiro, 2015). Risk factors include unhealthy eating habits, lack of physical activity, genetic factors, obesity, and age (biofarma, 2025). Early detection allows for faster preventive action, reducing serious complications such as heart disease, kidney failure, and neuropathy (American Diabetes Association 2023).

Naïve Bayes

Naive Bayes is a probability-based classification method that applies Bayes' Theorem with the assumption that each attribute is independent of the predicted class (Vernanda 2024). Posterior probability is calculated using the following equation:

$$P(C \mid X) = P(X \mid C) \times P(C) / P(X) \quad (1)$$

where P(C | X) is the class C probability given data X, P(X | C) is the likelihood, P(C) is the prior class probability, and P(X) is the probability of the data. Its advantages include low computational complexity and ease of implementation, making it commonly used as a baseline model in disease prediction. However, the assumption of feature independence is a limitation for medical data, which generally exhibits correlations between variables (BINUS UNIVERSITY, 2025).

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm that forms an optimal hyperplane as a separator between data classes by maximizing the margin between support vectors (Junus, 2023). The SVM decision function is defined as:

$$f(x) = w^T \cdot x + b \quad (2)$$

where w is the weight vector, x is the input feature vector, and b is the bias. For non-linear data, SVM uses a kernel function $\Phi(x)$ to map data to higher-dimensional spaces, with the RBF kernel: $K(xi, xj) = \exp(-\gamma \|xi - xj\|^2)$. SVM effectively avoids overfitting but requires precise kernel and parameter selection and longer computation time on large datasets (Siregar, 2024)

Data Imbalance

Data imbalance occurs when the number of instances in one class significantly outweighs that of the other, causing the machine learning model to be biased toward the majority class and less sensitive to the clinically more important minority class. This condition generally results in high accuracy but low recall and F1-score values, making its handling an important part of building reliable disease prediction models (Rahmawati, 2024).

Explainable Artificial Intelligence (XAI)

XAI is an approach that aims to increase the transparency of decisions in black-box machine learning models (Mienye, 2024). In diabetes prediction, XAI is important to identify key risk factors and support evidence-based medical decision-making. XAI enables an understanding of the role of

each feature in predictive outcomes globally and locally, thereby increasing the trust and accountability of the model (Hasan, 2024).

SHAP (Shapley Additive exPlanations)

SHAP calculates the contribution of each feature to the prediction outcome based on the Shapley value concept from game theory (Muliani, 2025).

LIME (Local Interpretable Model-agnostic Explanations)

LIME provides a local explanation of model predictions by building simple interpretable models around a specific predicted instance (Hasan, 2024). This method is model-agnostic, meaning it can be applied to various types of models without depending on their internal structure. The limitation of LIME is that its explanations are local and do not always reflect the overall global behavior of the model.

RESEARCH METHODS

Data Collection

The dataset used in this study is the Diabetes Health Indicators Dataset sourced from the Behavioral Risk Factor Surveillance System (BRFSS) in 2015, publicly accessible through Kaggle. This dataset consists of 253,680 records with 22 variables, including 1 target variable (diabetes status) and 21 feature variables comprising health indicators, behavioral factors, and demographic information. For the purpose of this research, a sample of 50,000 records was drawn from the 2015 BRFSS dataset. Table 1 presents the complete list of variables used in this study.

Table 1. Variable Dataset

Variable	Description	Data Type
Diabetes_binary	Diabetes status	Biner
HighBP	High blood pressure	Biner
HighChol	High cholesterol	Biner
CholCheck	Cholesterol check	Biner
BMI	Body mass index	Numerik
Smoker	Smoking habit	Biner
Stroke	History of stroke	Biner
HeartDiseaseorAttack	Heart disease or heart attack	Biner
PhysActivity	Physical activity	Biner
Fruits	Fruit consumption	Biner
Veggies	Vegetable consumption	Biner
HvyAlcoholConsump	Heavy alcohol consumption	Biner
AnyHealthcare	Access to healthcare	Biner
NoDocbcCost	No doctor visit due to cost	Numerik
GenHlth	General health condition	Biner
MentHlth	Mental health	Numerik
PhysHlth	Physical health	Numerik
DiffWalk	Difficulty walking	Biner
Sex	Gender	Biner
Age	Age	Biner
Education	Education level	Biner
Income	Income level	Biner

Pre-Processing Data

The pre-processing stage includes missing value handling, deletion of duplicate data, data normalization, and attribute transformation to match the needs of the classification algorithm. The data are then divided into training data and test data with an 80:20 ratio, yielding 40,000 training records and 10,000 test records.

Data Imbalance Handling

Class imbalances were addressed using the Synthetic Minority Over-sampling Technique (SMOTE) applied only to the training data prior to model development, to improve the model's ability to recognize minority classes and reduce bias against the majority class.

Model Development

The prediction model was built using two classification algorithms, namely Naive Bayes and Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel to handle non-linear data. Both models were trained using training data that underwent SMOTE and evaluated on test data.

Interpretability Analysis (XAI)

Model interpretability was analyzed using SHAP to identify the contribution of each feature globally to the overall model, and LIME to explain predictions locally on a given instance. The combination of both methods is used to obtain a comprehensive understanding of the risk factors that most influence the predictive outcomes.

Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, and F1-score metrics calculated based on the confusion matrix as follows:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

$$Precision = TP / (TP + FP) \quad (4)$$

$$Recall = TP / (TP + FN) \quad (5)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (6)$$

Where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) are components of the confusion matrix of the classification results.

RESULTS AND DISCUSSION

Data Pre-Processing Results

Pre-processing of the data was carried out to prepare the dataset for use in the model training stage. This process includes removing duplicate data and correcting column names. The results of the pre-processing are shown in Table 2.

Table 2. Pre-Processing Results

Remarks	Before	After
Amount of Data	253,680	229,474
Number of Duplicates	24,206	0

Handling Data Imbalance

Analysis of the distribution of data before imbalance handling showed that class 0 (non-diabetes) amounted to 34,477 samples (86.19%) and class 1 (diabetes) amounted to 5,523 samples (13.81%), indicating a class imbalance condition. After the implementation of SMOTE, the distribution of the two classes became balanced with 34,477 samples (50%) each, so the model was expected to be more sensitive in detecting minority classes. Visualizations before and after the implementation of SMOTE can be seen in Figures 1 and 2.

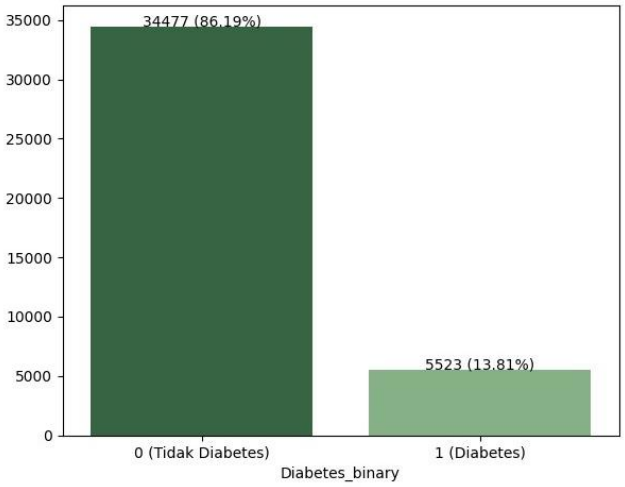

Figure 1. Before SMOTE (Synthetic Minority Over-sampling Technique)

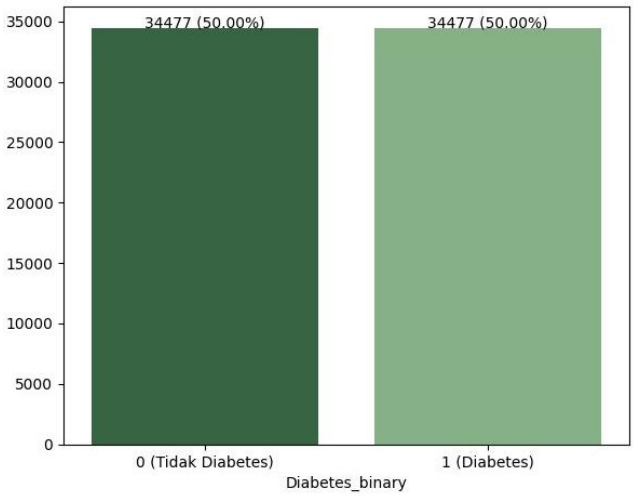

Figure 2. After SMOTE (Synthetic Minority Over-sampling Technique)

Model Performance Evaluation

The training results showed that Naive Bayes achieved a training accuracy of 74.14%, while SVM with the RBF kernel obtained a training accuracy of 89.34%. Evaluation on the test data using a confusion matrix resulted in a comparison of the performance of the two models as shown in Table 3.

Table 3. Evaluation Results

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.686	0.2748	0.7776	0.4062
SVM	0.8417	0.4189	0.3779	0.3974

Naive Bayes produced the highest recall value of 77.76%, indicating good ability in detecting diabetes cases, but low precision values (27.48%) indicated a high false positive. In contrast, SVM showed a more balanced performance with an accuracy of 84.17% and an F1-score of 39.74%, although the recall value was lower than that of Naive Bayes. These differences in characteristics reflect the trade-off between recall and precision in both models.

Confusion Matrix

Naive Bayes confusion matrix analysis yielded 5,786 True Negative (TN), 1,074 True Positive (TP), 2,833 False Positive (FP), and 307 False Negative (FN). In the SVM model, 7,895 TN, 522 TP, 724 FP, and 859 FN were obtained. The high FP value in Naive Bayes is consistent with the low precision value, while the high FN value in SVM explains the lower recall value as shown in the Fig

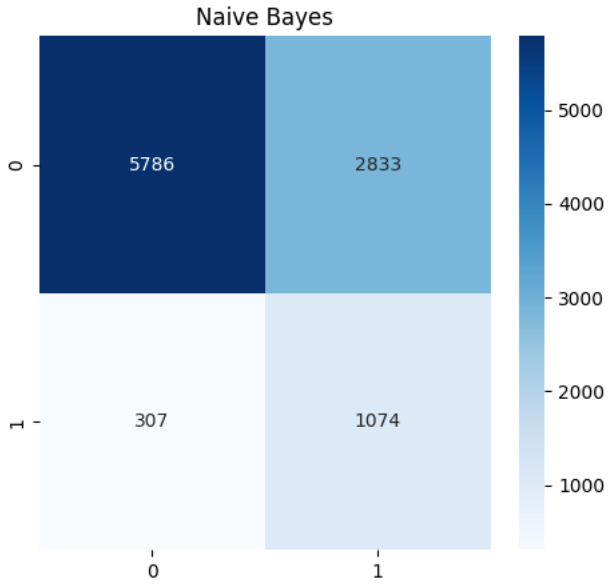

Figure 3. Confusion Matrix Naïve Bayes

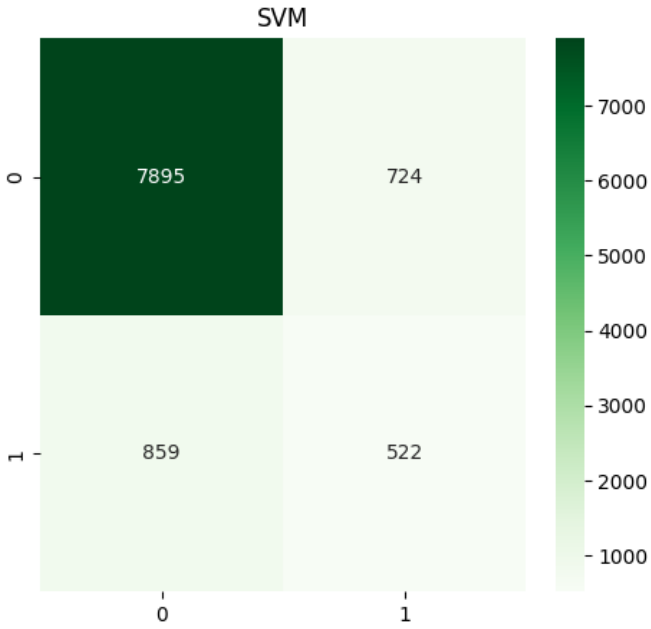

Figure 4. Confusion Matrix Support Vector Machine

Interpretability Analysis (XAI)

The SHAP analysis globally identified five features with the greatest influence on diabetes risk prediction: GenHlth (general health condition), HighBP (high blood pressure), BMI (body mass index), HighChol (high cholesterol levels), and Age. The GenHlth feature was the most dominant, where higher values correlated with an increased risk of diabetes. These findings are clinically consistent because general health conditions, blood pressure, and BMI are indeed major indicators of diabetes risk (Hasan et al., 2024). Visualization of SHAP results can be seen in Figure 5.

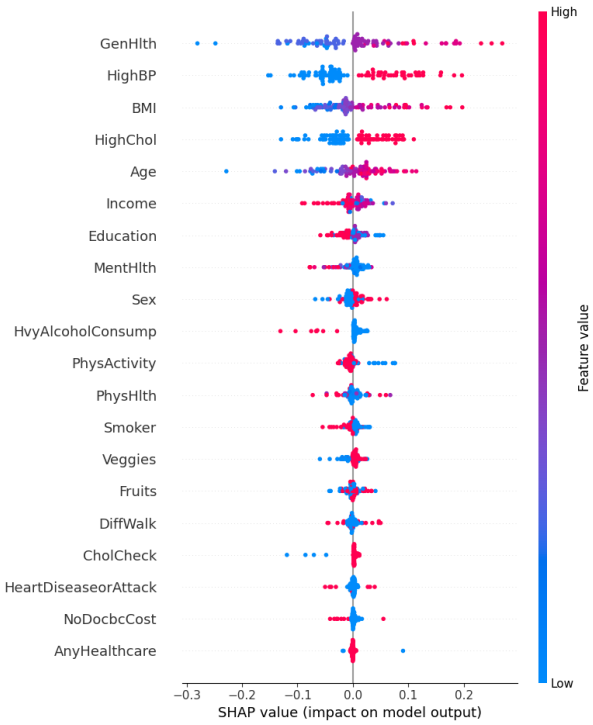

Figure 5. SHAP features the most influential features globally

Analysis of LIME locally showed that predictions at the individual level are influenced by different combinations of features. In the case study example, the HvyAlcoholConsump (excessive alcohol consumption) feature made the largest positive contribution to the diabetes prediction, while MentHlth (mental health), GenHlth, and Income made negative contributions that tended to lower the risk. These results prove that LIME is able to provide specific and contextual explanations for each individual. The LIME visualization can be seen in Figure 6.

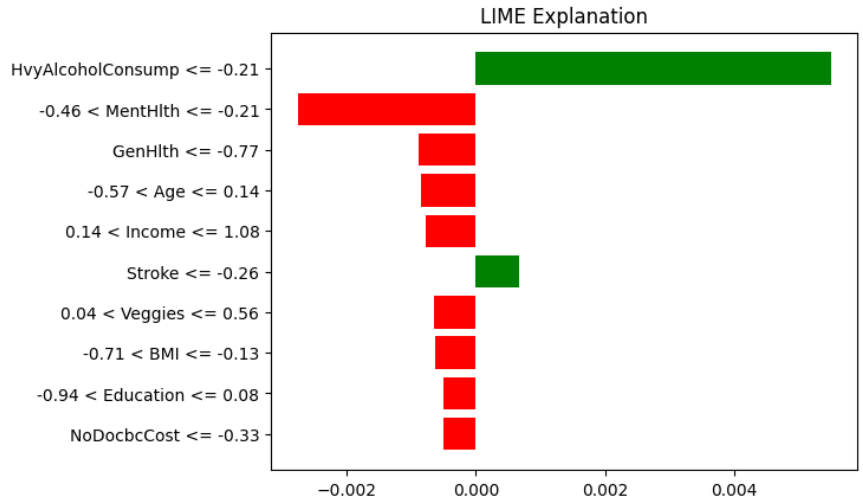

Figure 6. LIME Explanation

Discussion

Overall, SVM outperforms Naive Bayes in terms of accuracy and balance of performance, but Naive Bayes is more sensitive in detecting the diabetes class as indicated by its higher recall value. The application of SMOTE has been shown to be effective in overcoming class imbalance by balancing the distribution of training data, which has an impact on increasing the sensitivity of the model to minority classes. The integration of SHAP and LIME provides a complementary layer of interpretability: SHAP identifies dominant risk factors globally, while LIME describes model

decisions at the individual level. The combination of the two increases transparency and trust in prediction models in clinical contexts.

CONCLUSIONS

This study produced a risk prediction model for Diabetes Mellitus using Naive Bayes and Support Vector Machine algorithms by addressing class imbalance through SMOTE and enhancing model interpretability through Explainable Artificial Intelligence (XAI). Evaluation results showed that SVM achieved a better accuracy of 84.17%, demonstrating stronger overall predictive capability, while Naive Bayes achieved a higher recall of 77.77%, making it more effective in detecting diabetes cases. The application of SMOTE effectively addressed class imbalance, enabling the model to improve detection of the minority class. Furthermore, interpretation using SHAP and LIME successfully explained the model's prediction results, identifying GenHlth, HighBP, BMI, HighChol, and Age as the most influential factors in diabetes risk prediction. The main contribution of this research is the combination of machine learning methods with dual XAI interpretation that succeeded in making the black-box model clearer and more clinically understandable, supporting data-driven decision-making in the health sector.

REFERENCES

Ade Nasihudin Al Ansori. (2024, November 18). *Kemenkes Prediksi Angka Diabetes di Indonesia Naik hingga 28,5 Juta pada 2045, Deteksi Dini dan Mandiri Amat Penting*. Liputan6.Com. liputan6.com/health/read/5793225/kemenkes-prediksi-angka-diabetes-di-indonesia-naik-hingga-285-juta-pada-2045-deteksi-dini-dan-mandiri-amat-penting?

American Diabetes Association Releases 2023 Standards of Care in Diabetes to Guide Prevention, Diagnosis, and Treatment for People Living with Diabetes. (n.d.).

Askandar Tjokroprawiro, dkk. (2015). Buku ajar Ilmu Penyakit Dalam Edisi 2. In Prof. Dr. Askandar Tjokroprawiro (Ed.), *Buku Ajar Ilmu Penyakit Dalam (Edisi 2)* (p. 71). Airlangga University Press.

BINUS UNIVERSITY School of Information Systems. (2025, January 15). *Apa itu Naïve Bayes? Algoritma Klasifikasi Sederhana yang Kuat dalam Machine Learning*. Https://Sis.Binus.Ac.Id/2025/01/15/Apa-Itu-Naive-Bayes-Algoritma-Klasifikasi-Sederhana-Yang-Kuat-Dalam-Machine-Learning/.

Biofarma. (2025, March 25). *Bahaya Diabetes Mellitus dan Faktor Risiko yang Meningkatkan Kemungkinannya*. Https://Www.Biofarma.Co.Id/Id/Announcement/Detail/Bahaya-Diabetes-Mellitus-Dan-Faktor-Risiko-Yang-Meningkatkan-Kemungkinannya.

Dikan Ismafillah, Tatang Rohana, & Yana Cahyana. (2023). Analisis algoritma pohon keputusan untuk memprediksi penyakit diabetes menggunakan oversampling smote. *INFOTECH : Jurnal Informatika & Teknologi*, *4*(1), 27–36. https://doi.org/10.37373/infotech.v4i1.452

Erlin, Yulvia Nora Marlim, Junadhi, Laili Suryati, & Nova Agustina. (2022). Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *11*(2), 88–96. https://doi.org/10.22146/jnteti.v11i2.3586

Handayani, O. P., Purwono, Ashari, I. A., & Ardianto, R. (2025). Systematic Literature Review: Penerapan Machine Learning dalam Diagnosis dan Prediksi Penyakit Diabetes. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, *14*(2), 108–118. https://doi.org/10.34010/komputa.v14i2.16642

Hasan, R., Dattana, V., Mahmood, S., & Hussain, S. (2024). Towards Transparent Diabetes Prediction: Combining AutoML and Explainable AI for Improved Clinical Insights. *Information*, *16*(1), 7. https://doi.org/10.3390/info16010007

Junus, C. Z. V., Tarno, T., & Kartikasari, P. (2023). KLASIFIKASI MENGGUNAKAN METODE SUPPORT VECTOR MACHINE DAN RANDOM FOREST UNTUK DETEKSI AWAL RISIKO DIABETES MELITUS. *Jurnal Gaussian*, *11*(3), 386–396. https://doi.org/10.14710/j.gauss.11.3.386-396

Mienye, I. D., Obaido, G., Jere, N., Mienye, E., Aruleba, K., Emmanuel, I. D., & Ogbuokiri, B. (2024). A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. In *Informatics in Medicine Unlocked* (Vol. 51). Elsevier Ltd. https://doi.org/10.1016/j.imu.2024.101587

Muliani, S., Sukma Negara, B., Irsyad, M., & Iskandar, I. (2025). Application of Shapley Additive Explanations (SHAP) in Deep Learning for Lung Disease Detection Using X-ray Images. *Journal of Artificial Intelligence and Software Engineering*, *5*(2), 709–719. https://doi.org/10.30811/jaise.v5i2.7044

Nasution, F., Andilala, A., & Siregar, A. A. (2021). FAKTOR RISIKO KEJADIAN DIABETES MELLITUS. *Jurnal Ilmu Kesehatan*, *9*(2), 94. https://doi.org/10.32831/jik.v9i2.304

Rahmawati, S., Wibowo, A., & Masruriyah, A. F. N. (2024). Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of SVM, Logistic Regression, and Naive Bayes with SMOTE and ADASYN. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *8*(5), 607–614. https://doi.org/10.29207/resti.v8i5.5980

Siregar, M. M., Hizria, R., & Pardede, D. (2024). Perbandingan Kinerja Kernel SVM dalam Klasifikasi Kategori Kanker Kulit Menggunakan Transfer Learning. *Data Sciences Indonesia (DSI)*, *4*(1), 83–90. https://doi.org/10.47709/dsi.v4i1.4665

Vernanda, D., Nurizati, Z., & Hidayat, A. (2024). KLASIFIKASI BUAH NANAS MENGGUNAKAN METODE NAÏVE BAYES. *Melek IT : Information Technology Journal*, *10*(2), 111–120. https://doi.org/10.30742/melekitjournal.v10i2.355

Wantoro, A., Fitria Yulia, A., Yana Ayu, D., & Mustofa, S. (n.d.). *JIP (Jurnal Informatika Polinema) EVALUASI KINERJA ALGORITMA MACHINE LEARNING (ML) MENGGUNAKAN SELEKSI FITUR PADA KLASIFIKASI DIABETES*. Retrieved www.kaggle.com/uciml/pima-