Jurnal Penelitian Sains Teknologi

Early Classification of Diabetes Risk in Productive Age Groups Using Machine Learning

Muhammad Ghifari Zaki[1]✉, Helmi Imaduddin[2]

[1,2]Faculty of Communication and Informatics, Universitas Muhammadiyah Surakarta, Indonesia

doi: 10.23917/saintek.v2i1.16430
Received: January 15th, 2026 | Revised: February 20th, 2026 | Accepted: February 23rd, 2026
Available Online: February 25th, 2026 | Published Regularly: March, 2026

Abstract

Diabetes mellitus is a chronic non-communicable disease with a rapidly increasing prevalence, including among productive-age individuals (18–44 years), potentially affecting socio-economic productivity. Early risk identification is essential to prevent severe complications and reduce healthcare burden. This study aims to develop an early diabetes risk classification model for the productive-age population using machine learning techniques. The research employed the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework and utilized the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset from the Centers for Disease Control and Prevention (CDC). After age filtering and preprocessing, 48,867 observations were analyzed. Three algorithms Logistic Regression, Random Forest, and XGBoost were compared, with recall prioritized as the primary evaluation metric due to its importance in health screening contexts. Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE). The results indicate that Logistic Regression achieved the highest recall (75.06%) and ROC-AUC (83.62%), demonstrating superior capability in detecting high-risk individuals compared to the other models. Feature selection using mutual information identified ten significant predictors, with general health status, age, and body mass index emerging as dominant risk factors. The selected model was implemented into a Flask-based web prototype to support real-time early screening. These findings demonstrate that machine learning can effectively support early diabetes risk detection among productive-age populations and provide a practical, data-driven approach for preventive public health strategies.

Keywords: CRISP-DM, diabetes risk prediction, health screening, logistic regression, machine learning, productive-age population

✉**Corresponding Author:**
Muhammad Ghifari Zaki, Faculty of Communication and Informatics, Universitas Muhammadiyah Surakarta, Indonesia
Email: l200220149@student.ums.ac.id

Introduction

The Diabetes mellitus is a chronic non-communicable metabolic disease that remains a major global public health challenge [1]. This condition occurs due to impaired glucose regulation mechanisms in the body, either because the pancreas is unable to produce adequate insulin or because the available insulin cannot be utilized effectively by body tissues [2]. This imbalance results in prolonged hyperglycemia. If not detected and managed early, diabetes mellitus can lead to severe complications, including cardiovascular disease, kidney dysfunction (nephropathy), retinal damage (retinopathy), and an increased risk of premature mortality.

According to the 2025 report from the International Diabetes Federation (IDF), the global number of individuals living with diabetes has surpassed 589 million and is projected to reach 853 million by 2050 [3]. Indonesia is among the countries with a high diabetes burden, with more than 19 million cases, ranking fifth globally. The significant rise in diabetes prevalence among the productive-age population has far-reaching impacts, not only on national healthcare systems but also on economic and social productivity. The productive-age group, generally defined as individuals aged 18–44 years, represents a strategically important phase of life in economic and social activities, as widely discussed in epidemiological studies of diabetes, including research on young-onset Type 2 Diabetes Mellitus (T2DM) [4].

Rapid urbanization over recent decades has further contributed to lifestyle changes associated with an increased risk of diabetes, particularly among the productive-age population [5]. Modern lifestyles are characterized by reduced physical activity due to sedentary occupations, increased consumption of high-sugar and high-fat foods, and low awareness of routine health screenings [6]. These factors collectively contribute to insulin resistance, one of the primary mechanisms in the pathogenesis of Type 2 diabetes. This condition underscores the importance of preventive strategies focused on early detection to identify high-risk individuals before the disease progresses to more severe stages and causes complications. Accurate identification of diabetes risk is essential to ensure that preventive and therapeutic interventions are implemented effectively and appropriately [7]. Data-driven early detection approaches are considered capable of reducing diabetes-related morbidity and mortality while maintaining societal productivity, as prevention is generally more efficient and cost-effective than treating diabetes complications [8].

The application of Artificial Intelligence (AI), particularly machine learning, has been widely adopted in healthcare to support the analysis and classification of chronic disease risks. Machine learning is capable of processing complex and multidimensional data, such as demographic variables, clinical indicators, and lifestyle patterns, which often exhibit nonlinear relationships [9]. This approach has been shown to improve analytical accuracy and clinical decision-making efficiency compared to conventional statistical methods [10]. By utilizing mathematical algorithms to learn patterns from historical data, machine learning can generate inferences without explicit programming [11], making it highly promising for identifying diabetes risk before clinical symptoms emerge.

Analysis of large-scale health data (big data) enables the discovery of predictive indicators that may be difficult to identify manually [12]. The integration of statistical methods and AI serves as a crucial foundation for extracting meaningful knowledge from such data [13]. Although previous studies indicate that ensemble algorithms such as Extreme Gradient Boosting (XGBoost) often outperform Logistic Regression and Random Forest, variations in results are still observed due to differences in dataset characteristics and preprocessing stages. Therefore, a comparative evaluation was conducted among the productive-age group (18–44 years) to assess the performance of these three algorithms in supporting early diabetes risk screening. This evaluation utilized health indicator data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), collected by the Centers for Disease Control and Prevention (CDC), which includes demographic variables, health conditions, and behavioral patterns relevant to diabetes risk in young adult populations.

This study was structured using the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which provides systematic research stages from business understanding, data understanding, data preparation, modeling, evaluation, to deployment. The use of this framework ensures that the research process is systematic, well-documented, and reproducible. Furthermore, CRISP-DM offers a clear foundation for analyzing variables influencing diabetes risk, thereby enhancing the validity and reliability of the resulting model to support preventive decision-making in public health.

The primary objective of this study is to develop a diabetes risk prediction system for the productive-age population (18–44 years) by comparing the performance of three machine learning algorithms: Logistic Regression, Random Forest, and XGBoost, with recall prioritized as the main evaluation metric for health screening applications. The best-performing model was subsequently implemented into a web-based prototype system designed as a self-assessment tool for the general public. Previous studies have demonstrated that these algorithms can effectively identify diabetes risk factors at early stages and achieve high predictive accuracy on population health data [14]. Model performance was evaluated using quantitative metrics including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). In the context of health screening, recall (sensitivity) was prioritized because the model’s ability to identify all high-risk individuals is more critical than minimizing false positives. This is particularly important given that missed diagnoses (false negatives) in chronic diseases such as diabetes may lead to fatal consequences. In addition to quantitative evaluation, an analysis of each variable’s contribution to diabetes risk was conducted to enhance model interpretability. Ultimately, the findings of this study are expected to provide practical benefits in increasing public awareness of diabetes risk while contributing scientifically to the development of data-driven prediction methods for non-communicable diseases, particularly within the productive-age population.

Method

This study was designed using the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which provides a structured research workflow from problem understanding to predictive model implementation. CRISP-DM was selected because it ensures that the research process is systematic, well-documented, and reproducible, thereby producing a valid, accurate, and applicable early classification model for diabetes risk [15].

The CRISP-DM framework consists of six main phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment [16]. The methodological workflow based on CRISP-DM is illustrated in Figure 1, which depicts the cyclical relationship between each phase in the machine learning model development process.

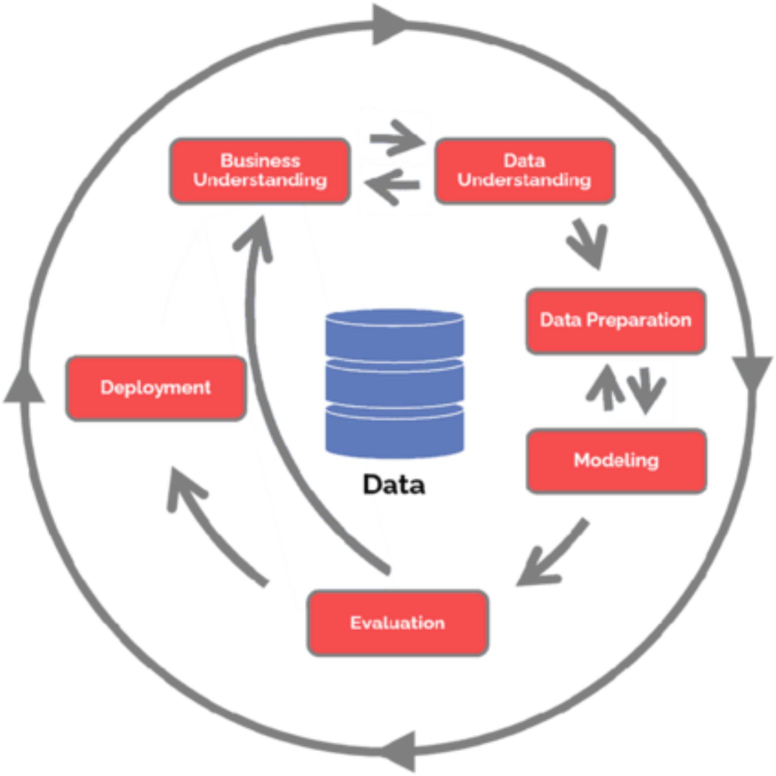


Figure 1. Research Methodology Framework Based on CRISP-DM

a. Business Understanding

The initial phase of this study focused on understanding the problem of diabetes mellitus among the productive-age population, defined in this research as individuals aged 18–44 years. This age group plays a strategic role in socio-economic activities but has shown an increasing trend in diabetes risk associated with modern lifestyle changes. The high prevalence of diabetes and delayed diagnosis within this age group highlight the need for a rapid, accurate, and data-driven early screening method.

This study aims to develop an early diabetes risk classification model using machine learning algorithms based on population health data, including demographic variables, clinical indicators, and health behavior factors. The developed model is expected to provide practical benefits as an initial screening tool to increase awareness among the productive-age population, while also contributing scientifically by establishing a methodological framework for machine learning-based prediction of non-communicable diseases.

b. Data Understanding

This study utilized the Diabetes Health Indicators Dataset, sourced from the Centers for Disease Control and Prevention (CDC) through the 2015 Behavioral Risk Factor Surveillance System (BRFSS), accessed via the Kaggle platform [17]. The dataset originates from an annual telephone-based health survey in the United States designed to identify risk factors for chronic diseases, including diabetes.

The dataset consists of 253,680 observations with 22 features, including health indicators (HighBP, HighChol, BMI), lifestyle behaviors (Smoker, PhysActivity), healthcare access (AnyHealthcare), physical and mental health status (GenHlth, PhysHlth, MentHlth), and demographic factors (Sex, Age, Income, Education). The target variable is Diabetes_binary, coded as 1 for diabetes and 0 for non-diabetes.

To align with the study focus, data were filtered to include only individuals aged 18–44 years, and duplicate records were removed. From the initial 253,680 observations, age filtering resulted in 54,401 observations (21.45%). After removing 5,534 duplicate records (10.17%), a clean dataset of 48,867 observations was obtained, consisting of 46,661 non-diabetes cases (95.49%) and 2,206 diabetes cases (4.51%). This class imbalance reflects the lower prevalence of diabetes in the productive-age population.

Feature selection was conducted using mutual information, resulting in 10 features with the strongest association with diabetes: HighBP, HighChol, BMI, AnyHealthcare, GenHlth, PhysHlth, DiffWalk, Sex, Age, and Income. Although the dataset is based on the U.S. population, the selected health indicators are considered universal and relevant for global diabetes risk prediction research.

The dataset was deemed adequate and appropriate for developing a diabetes risk classification model in the productive-age population.

c. Data Preparation

The data preparation phase aimed to ensure data consistency and quality prior to modeling. The cleaned dataset (48,867 observations) was divided into a training set (39,093 observations, 80%) and a test set (9,774 observations, 20%) using stratified sampling to preserve class proportions.

Class distribution analysis of the training set revealed significant imbalance, with 37,328 negative cases and 1,765 positive cases, resulting in an imbalance ratio of 21.15:1. Such imbalance is common in healthcare datasets, particularly among productive-age populations.

All features were standardized using StandardScaler, transforming them to a mean of 0 and a standard deviation of 1 to prevent dominance by features with larger numeric ranges. Standardization was applied after dataset splitting to prevent data leakage; the scaler was fitted only on the training set and subsequently applied to the test set.

To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training set. SMOTE generates synthetic minority samples through k-nearest neighbors interpolation. The diabetes class increased from 1,765 to 37,328 samples, resulting in a balanced training dataset of 74,656 observations with a 1:1 class ratio.

The test set retained its original distribution (9,333 non-diabetes and 441 diabetes cases) to provide realistic evaluation under imbalanced conditions. This approach ensures that the model is trained on balanced data but evaluated under real-world class distribution. Class balancing was intended to improve model sensitivity (recall), which is critical in health screening systems.

d. Modelling

The modeling phase aimed to construct machine learning models capable of classifying diabetes risk among productive-age individuals. Model training was performed using cross-validation to minimize overfitting and ensure prediction stability.

Three algorithms were implemented:

Logistic Regression models binary outcome probability using the logit function:

$$P(Y=1\mid X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_nX_n)}}$$

where $P(Y=1\mid X)$ represents the probability of diabetes risk, $\beta_0$is the intercept, $\beta_i$are regression coefficients, and $X_i$are predictor variables. Parameters were estimated using Maximum Likelihood Estimation (MLE).

Random Forest is an ensemble algorithm that constructs multiple decision trees using bootstrap sampling. Final predictions are obtained through majority voting:

$$\hat{y}=\text{mode}(h_1(X),h_2(X),\dots,h_M(X))$$

where $h_i(X)$represents the prediction from the i-th tree. This algorithm is robust to noise and capable of capturing complex patterns while providing feature importance.

XGBoost is a boosting algorithm that sequentially builds trees to correct previous errors. The objective function is:

$$\text{Obj}=\sum_{i=1}^{n}l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k)$$

where l denotes the loss function, Ω is the regularization term, and K is the number of trees. XGBoost is known for high efficiency and predictive performance.

e. Evaluation

Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC.

Accuracy:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

Precision:

$$Precision=\frac{TP}{TP+FP}$$

Recall:

$$Recall=\frac{TP}{TP+FN}$$

F1-score:

$$F1=2\times\frac{Precision\times Recall}{Precision+Recall}$$

In health screening contexts, recall (sensitivity) was prioritized to minimize false negatives, as missed diagnoses in chronic diseases such as diabetes may have severe consequences.

Feature importance analysis was conducted on ensemble-based models to identify the most influential variables, enhancing model interpretability and clinical relevance.

f. Deployment

The deployment phase involved implementing the best-performing model (Logistic Regression) into a web-based prototype system. The model, evaluated using accuracy, precision, recall, F1-score, and AUC-ROC, was integrated into a Flask-based web application to support early diabetes risk screening.

The prototype was developed using Python and Flask, providing an HTML form interface for user input based on the 10 selected features. Input data undergo preprocessing (standardization using StandardScaler), followed by real-time risk prediction. The model was saved using joblib for efficient loading, and the system architecture, routing structure, and workflow were documented to ensure reproducibility and future development.

Result and Discussion

a. Classification Model Evaluation Results

This study aimed to identify the most suitable classification algorithm for early diabetes risk detection among individuals of productive age. Three machine learning algorithms were evaluated, namely Logistic Regression, Random Forest, and XGBoost, using the complete set of 21 features. The dataset was divided into training and testing sets with an 80:20 ratio. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied only to the training data.

Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC in order to provide a comprehensive assessment of classification capability [18]. The evaluation results are presented in Table 1.

Table 1. Classification model evaluation results

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0,7769	0,1414	0,7778	0,2393	0,8385
Random Forest	0,9445	0,3413	0,1293	0,1875	0,8059
XGBoost	0,9510	0,4031	0,1791	0,2480	0,8327

The evaluation results presented in Table 1 indicate the presence of a trade-off among evaluation metrics across the three models. Logistic Regression achieved the highest recall at 77.78% with a ROC-AUC of 83.85%, although its precision was relatively low at 14.14%. On the other hand, Random Forest and XGBoost demonstrated very high accuracy, exceeding 94%; however, their recall values were considerably lower, at 12.93% and 17.91%, respectively. This condition indicates that although these two models are highly accurate overall, many diabetes cases were not successfully detected.

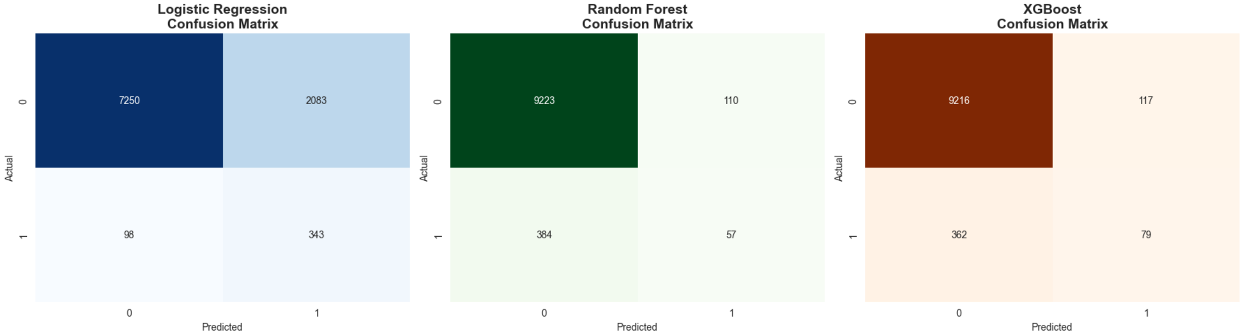


Figure 2. Confusion Matrix Visualization

The confusion matrix visualization in Figure 2 shows that Logistic Regression successfully detected 343 true positives out of 441 diabetes cases (77.78%), with 98 false negatives. The model also produced 7,250 true negatives and 2,083 false positives. Random Forest had the lowest number of false positives (110) but detected only 57 true positives, with 384 false negatives. XGBoost demonstrated a more balanced performance, with 79 true positives and 362 false negatives. The high number of false negatives in Random Forest and XGBoost represents a significant risk in health screening contexts.

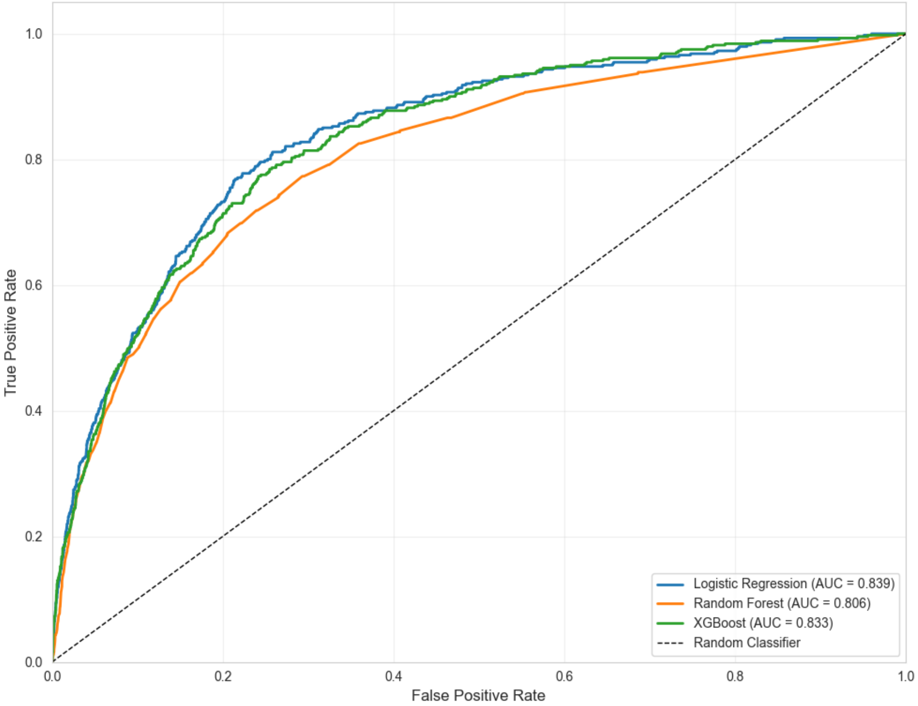


Figure 3. ROC Curve Visualization

The ROC curve visualization in Figure 3 shows that Logistic Regression (blue) achieved the highest AUC (0.839), outperforming XGBoost (green, 0.833) and Random Forest (orange, 0.806). All three curves lie well above the diagonal line of a random classifier, indicating significant predictive capability. The Logistic Regression curve is closest to the top-left corner, demonstrating the model's effectiveness in maximizing the true positive rate while minimizing the false positive rate. Therefore, it was selected for the feature selection stage to optimize system usability without sacrificing predictive performance.

b. Selected Feature Analysis

After Logistic Regression was selected based on the highest recall (77.78%) and superior ROC-AUC (83.85%), feature number optimization was conducted to balance model performance with system usability. Although using 21 features yielded good performance, a high input burden may reduce the application's effectiveness in real-world implementation.

Experiments were conducted with variations in the number of features (5, 7, 10, 12, 15, 18, and 21) using elbow curve analysis, with recall as the primary evaluation metric. Logistic Regression was retrained under each configuration to evaluate the trade-off between complexity and performance.

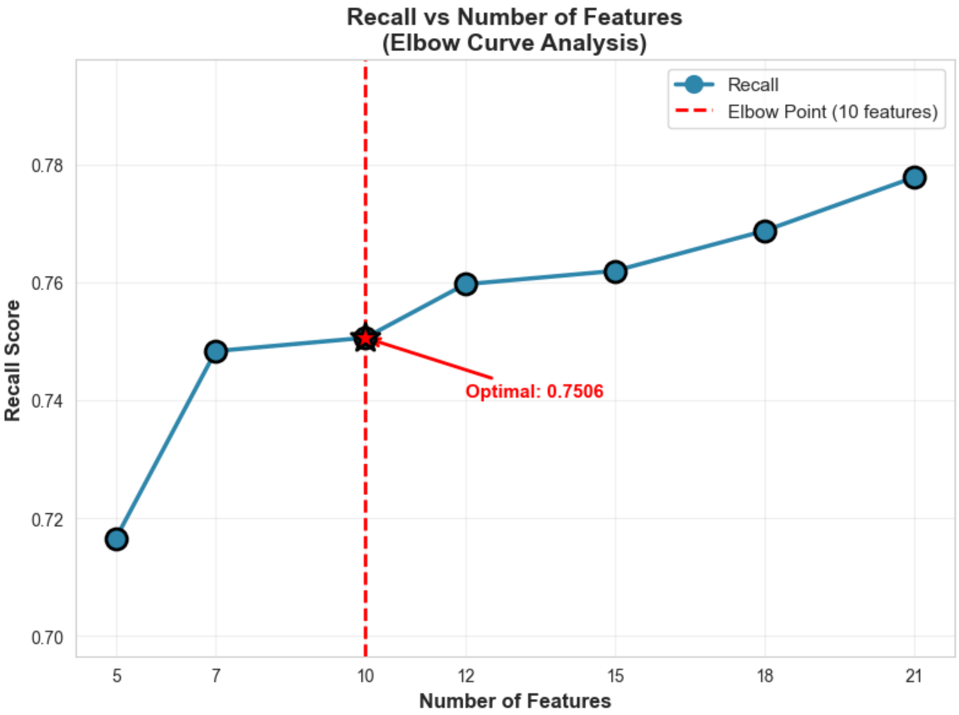


Figure 4. Elbow Curve Analysis and Marginal Performance Gain

The elbow curve analysis results in Figure 4 indicate that increasing the number of features from 5 to 7 resulted in a significant recall improvement (71.86%). However, adding features from 7 to 10 increased recall by only a marginal gain of 3.17%, and further additions resulted in improvements of less than 1%. The optimal point (elbow point) was identified at 10 features with a recall of 75.06%, where additional features were not proportional to the added complexity, in line with the principle of parsimony [19].

The top 10 features were selected using Mutual Information (MI) due to its ability to measure non-linear dependency. The selected features were HighBP, HighChol, BMI, AnyHealthcare, GenHlth, PhysHlth, DiffWalk, Sex, Age, and Income. These features represent a combination of health indicators, functional status, healthcare access, as well as clinically relevant demographic and socioeconomic factors.

c. Interpretation of Feature Contribution

To understand the contribution of each variable to diabetes risk prediction, an analysis of feature coefficients in the Logistic Regression model was conducted. The model coefficients were transformed into odds ratios (OR) to provide more intuitive interpretation, where OR > 1 indicates increased diabetes risk and OR < 1 indicates decreased risk. This analysis aims to identify the most influential features in the model’s decision-making process.

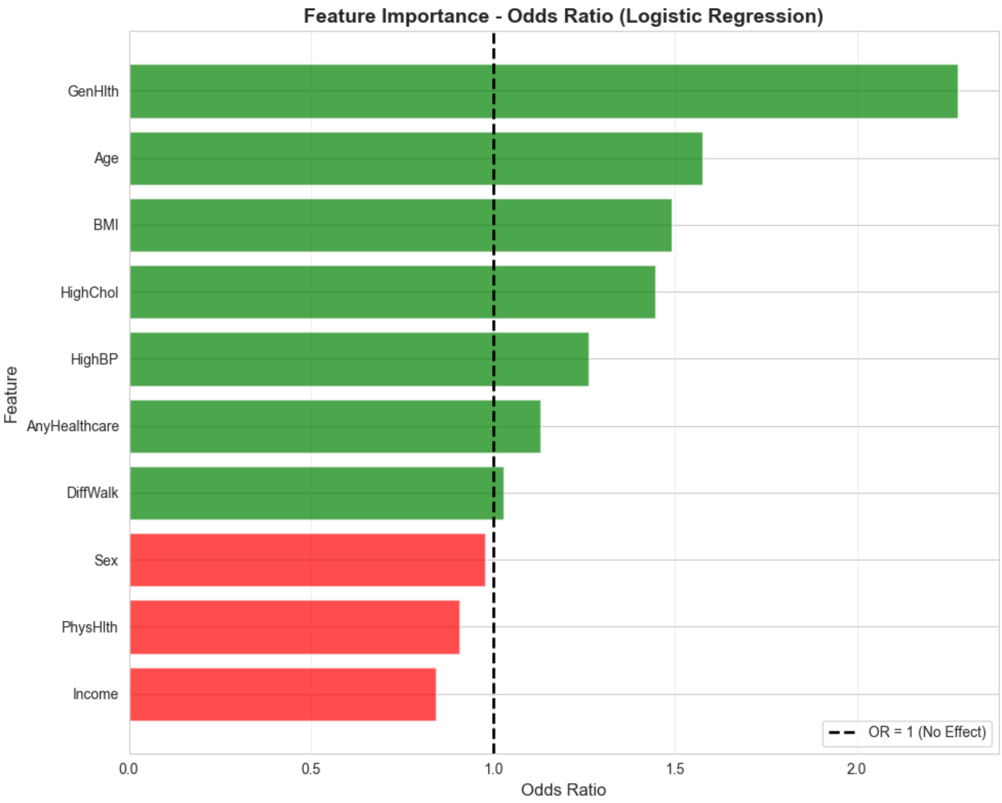


Figure 5. Feature Importance Visualization Odds Ratio (Logistic Regression)

Based on the analysis results in Figure 5, the variable GenHlth (general health status) showed the highest odds ratio of 2.28, meaning that each one-unit increase in poorer health status increases diabetes risk by 128%. This finding is consistent with public health literature stating that poor perceived general health strongly correlates with chronic disease prevalence. The variable Age ranked second with an OR of 1.58, indicating a 58% increase in risk for each increase in age category, consistent with epidemiological understanding that diabetes prevalence increases with age. BMI showed an OR of 1.49, reinforcing the role of obesity as a major risk factor contributing to insulin resistance.

Metabolic condition variables such as HighChol and HighBP showed OR values of 1.45 and 1.26, respectively, reflecting the role of metabolic syndrome as a predictor of diabetes. AnyHealthcare (access to healthcare services) with an OR of 1.13 indicates that individuals with healthcare access are more likely to be diagnosed, not necessarily due to increased risk, but due to better detection through routine examinations.

Meanwhile, the variables Sex, PhysHlth, and Income showed odds ratios below 1, indicating a negative association with diabetes risk. Income had an OR of 0.84, meaning that increased income correlates with a 16% reduction in diabetes risk. This finding aligns with literature on social determinants of health, where higher socioeconomic status provides better access to healthy nutrition, physical activity, and preventive healthcare services.

This odds ratio analysis strengthens the interpretative validity of the model, as the key variables align with medical and epidemiological literature on diabetes risk factors.The model is not only quantitatively accurate but also captures clinically relevant patterns. The high interpretability of Logistic Regression makes it an appropriate choice for screening system implementation, where transparency in decision-making is essential for clinical trust and adoption by healthcare professionals.

d. Model Validation

Model validation was conducted to ensure that model performance does not depend on a single data-splitting scheme and that it possesses good generalization ability. Validation was performed using Stratified k-Fold Cross-Validation with five folds (5-fold), where class proportions were maintained in each fold. In each fold, preprocessing steps (StandardScaler and SMOTE) were applied exclusively to the training data to prevent data leakage.

Table 2. Stratified k-Fold Cross-Validation Evaluation Results (5-Fold)

Fold	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1	0.7715	0.1343	0.7460	0.2276	0.8361
2	0.7802	0.1389	0.7421	0.2340	0.8424
3	0.7777	0.1340	0.7188	0.2259	0.8222
4	0.7781	0.1439	0.7914	0.2435	0.8548
5	0.7767	0.1380	0.7528	0.2333	0.8330
Mean	**0.7768**	**0.1378**	**0.7502**	**0.2328**	**0.8377**
Std	**0.0029**	**0.0036**	**0.0235**	**0.0062**	**0.0107**

Table 2. Stratified k-Fold Cross-Validation Evaluation Results (5-Fold). The validation results show that the Logistic Regression model achieved stable performance with an average recall of 75.02% ± 2.35% and ROC-AUC of 83.77% ± 1.07%. The low standard deviation indicates consistent performance across different data subsets without overfitting. The recall range across folds (71.88%–79.14%) indicates that the model consistently detects 7–8 out of 10 diabetes cases.

The high recall confirms the model's ability to identify the majority of at-risk individuals with a low false negative rate, while the low precision (13.78%) reflects a relatively high number of false positives. This characteristic aligns with the priority of health screening systems, where detecting as many at-risk cases as possible for further examination is more important. The recall variation of 2.35% across folds indicates stable detection performance across different subsets of the productive-age population.

Overall, the cross-validation results reinforce the selection of Logistic Regression as a model with good stability and generalization capability for implementation in a diabetes risk classification system for the productive-age group.

The analysis results demonstrate that the machine learning approach is capable of producing a diabetes risk classification model with good performance for the productive-age group. The Logistic Regression model with 10 selected features achieved a recall of 75.06% and a ROC-AUC of 83.62%, indicating good discriminative ability in distinguishing between at-risk and non-risk individuals. These findings highlight the potential of the model as an early screening tool to identify diabetes risk before the onset of more severe clinical symptoms.

Differences in performance among the algorithms reveal a clear trade-off between evaluation metrics. Logistic Regression excels in recall (75.06%), whereas Random Forest and XGBoost achieve higher accuracy (>93%) and precision but exhibit low recall (<30%). In health screening contexts, the primary objective is to detect as many at-risk individuals as possible for further examination; therefore, recall becomes the most critical metric. Based on this consideration, Logistic Regression was selected because it consistently detects three out of four diabetes cases, as confirmed by cross-validation results (recall 75.02% ± 2.35%).

The confusion matrix results on the test set show that the model successfully identified 331 out of 441 diabetes cases (true positives), with 110 false negatives. Although 2,101 false positives were observed, this characteristic is acceptable in initial screening, as individuals predicted to be at risk will undergo further examination for diagnostic confirmation. Missing diabetes cases (false negatives) is considered more critical than unnecessary referrals (false positives) in a public health context.

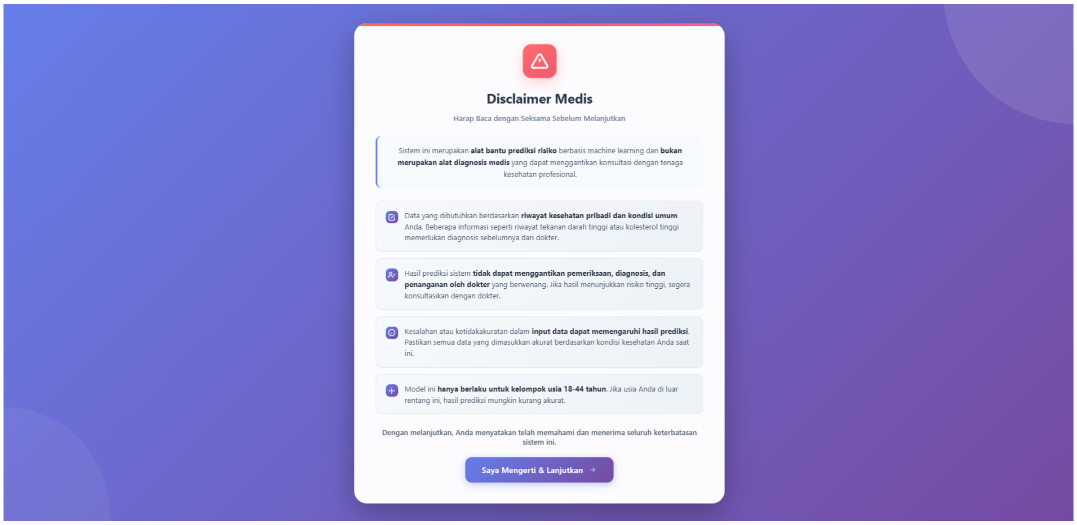
Feature coefficient analysis shows that GenHlth (OR 2.28), Age (OR 1.58), and BMI (OR 1.49) are dominant predictors of diabetes risk, consistent with epidemiological literature on metabolic disease risk factors. The finding that Income is negatively associated with diabetes risk (OR 0.84) aligns with research on social determinants of health, where higher socioeconomic status provides better access to healthy lifestyles and preventive services.

This study has several limitations. First, the use of secondary data from the United States population limits the generalizability of findings to the Indonesian population, which has different demographic and lifestyle characteristics. Second, the dataset does not include detailed lifestyle factors such as dietary patterns, physical activity intensity, and smoking duration, which could further improve predictive accuracy. Third, the model is designed as an initial screening tool rather than a substitute for clinical diagnosis; therefore, individuals predicted to be at risk still require laboratory examination (HbA1c, fasting blood glucose) for medical confirmation.

e. System Implementation

The selected Logistic Regression model was subsequently implemented into a web-based prototype application using the Flask framework. The system was designed to allow users to perform early diabetes risk screening directly through an interactive interface. The model, StandardScaler, and feature configuration were stored using joblib and automatically loaded when the application runs.

The application consists of three main components. The first page is a medical disclaimer page that users must pass before accessing the prediction form. This page explains that the system serves as a screening tool and does not replace clinical diagnosis, that prediction accuracy depends on the correctness of the input data, and that individuals predicted to be at high risk should undergo laboratory examination for medical confirmation. The disclaimer page interface is shown in Figure 6.



Disclaimer Medis

Harap Baca dengan Seksama Sebelum Melanjutkan

Sistem ini merupakan **alat bantu prediksi risiko** berbasis machine learning dan **bukan merupakan alat diagnosis medis** yang dapat menggantikan konsultasi dengan tenaga kesehatan profesional.

Data yang dibutuhkan berdasarkan **riwayat kesehatan pribadi dan kondisi umum** Anda. Beberapa informasi seperti riwayat tekanan darah tinggi atau kolesterol tinggi memerlukan diagnosis sebelumnya dari dokter.

Hasil prediksi sistem **tidak dapat menggantikan pemeriksaan, diagnosis, dan penanganan oleh dokter** yang berwenang. Jika hasil menunjukkan risiko tinggi, segera konsultasikan dengan dokter.

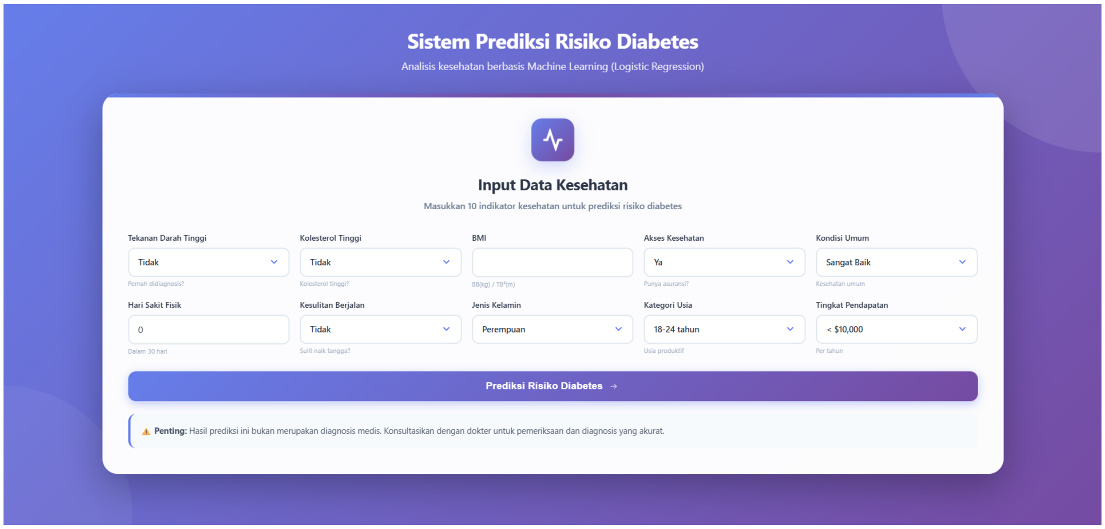
Kesalahan atau ketidakakuratan dalam **input data dapat memengaruhi hasil prediksi**. Pastikan semua data yang dimasukkan akurat berdasarkan kondisi kesehatan Anda saat ini.

Model ini **hanya berlaku untuk kelompok usia 18-44 tahun**. Jika usia Anda di luar rentang ini, hasil prediksi mungkin kurang akurat.

Dengan melanjutkan, Anda menyatakan telah memahami dan menerima seluruh keterbatasan sistem ini.

Figure 6. Disclaimer Page Interface

The second page is the input form, which provides ten fields corresponding to the variables used in the modeling process: HighBP (high blood pressure), HighChol (high cholesterol), BMI (body mass index), AnyHealthcare (healthcare access), GenHlth (general health status), PhysHlth (number of physically unhealthy days), DiffWalk (difficulty walking), Sex (gender), Age (age category), and Income (income level). After users complete all fields and submit the data, the system performs input validation and preprocessing, including feature standardization using the same StandardScaler applied during training, and then generates a prediction using the Logistic Regression model. The form interface is shown in Figure 7.



Sistem Prediksi Risiko Diabetes

Analisis kesehatan berbasis Machine Learning (Logistic Regression)

Input Data Kesehatan

Masukkan 10 indikator kesehatan untuk prediksi risiko diabetes

Tekanan Darah Tinggi: Tidak — Pernah didiagnosis?

Kolesterol Tinggi: Tidak — Kolesterol tinggi?

BMI: — BB(kg) / TB^2(m)

Akses Kesehatan: Ya — Punya asuransi?

Kondisi Umum: Sangat Baik — Kesehatan umum

Hari Sakit Fisik: 0 — Dalam 30 hari

Kesulitan Berjalan: Tidak — Sulit naik tangga?

Jenis Kelamin: Perempuan

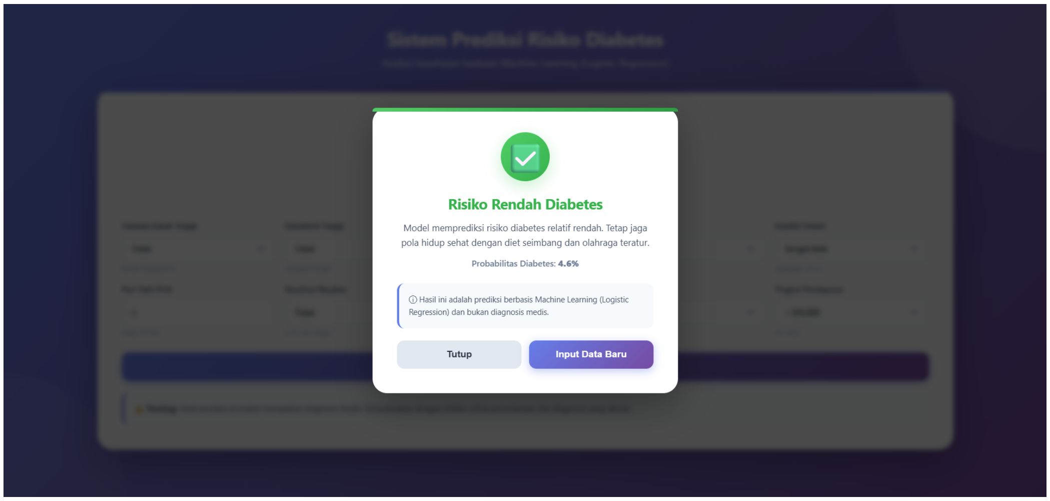
Kategori Usia: 18-24 tahun — Usia produktif

Tingkat Pendapatan: < $10,000 — Per tahun

⚠ **Penting:** Hasil prediksi ini bukan merupakan diagnosis medis. Konsultasikan dengan dokter untuk pemeriksaan dan diagnosis yang akurat.

Figure 7. Health Data Input Form Interface

The third page is a prediction result modal displaying the risk classification (high risk or low risk), the model confidence level expressed as a percentage, and follow-up recommendations tailored to the classification outcome. The result modal also reiterates the medical disclaimer to ensure users understand that the results do not constitute a medical diagnosis and require confirmation through laboratory testing (HbA1c, fasting blood glucose) by healthcare professionals. The high-risk and low-risk prediction result interfaces are shown in Figures 8 and 9, respectively.



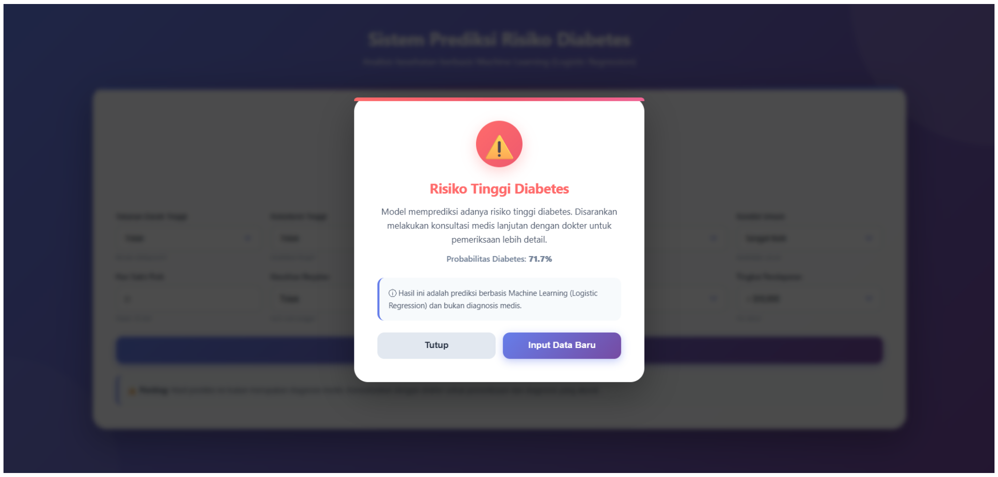
Risiko Rendah Diabetes

Model memprediksi risiko diabetes relatif rendah. Tetap jaga pola hidup sehat dengan diet seimbang dan olahraga teratur.

Probabilitas Diabetes: **4.6%**

ⓘ Hasil ini adalah prediksi berbasis Machine Learning (Logistic Regression) dan bukan diagnosis medis.

Figure 8. Low-Risk Diabetes Prediction Result Interface



Risiko Tinggi Diabetes

Model memprediksi adanya risiko tinggi diabetes. Disarankan melakukan konsultasi medis lanjutan dengan dokter untuk pemeriksaan lebih detail.

Probabilitas Diabetes: **71.7%**

ⓘ Hasil ini adalah prediksi berbasis Machine Learning (Logistic Regression) dan bukan diagnosis medis.

Figure 9. High-Risk Diabetes Prediction Result Interface

The system was designed with usability and accessibility considerations, where all input fields use simple and easily understandable formats for general users. Implementing Logistic Regression provides additional advantages in terms of high prediction speed and strong interpretability, enabling real-time responses and more transparent explanation of prediction results.

Conclusion

Based on all research stages, including data understanding, data preparation, feature selection, modeling, evaluation, and validation, it can be concluded that the machine learning approach successfully produced a diabetes risk classification model with good performance for the productive-age group. The evaluation results indicate that the three algorithms

(Logistic Regression, Random Forest, and XGBoost) exhibit different performance characteristics with clear trade-offs among evaluation metrics.

In the context of health screening, where the primary objective is to detect as many at-risk individuals as possible for further examination, Logistic Regression was selected as the best model with a recall of 75.06% and a ROC-AUC of 83.62%. The model consistently detects three out of four diabetes cases, as confirmed by cross-validation results showing an average recall of 75.02% ± 2.35% across five folds. The low standard deviation indicates good generalization ability without overfitting.

Feature selection using elbow curve analysis and mutual information identified 10 optimal features that balance model performance and system usability. Feature contribution analysis shows that general health status (GenHlth), age (Age), and body mass index (BMI) are dominant predictors of diabetes risk, consistent with epidemiological literature on metabolic disease risk factors. The model has been implemented into a Flask-based web prototype that provides an interactive interface for early diabetes risk screening in the productive-age population.

Reference

[1] J. Homepage, A. Oktaviana, D. Puspasari Wijaya, A. Pramuntadi, and D. Heksaputra, “MALCOM: Indonesian Journal of Machine Learning and Computer Science Prediction of Type 2 Diabetes Mellitus Using The K-Nearest Neighbor (K-NN) Algorithm Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN),” vol. 4, no. 3, pp. 812–818, 2024.

[2] T. Palabaş, “Early-Stage Diabetes Risk Prediction Using Machine Learning Techniques Based on Ensemble Approach,” *Eskişehir Tek. Üniversitesi Bilim ve Teknol. Derg. - C Yaşam Bilim. Ve Biyoteknoloji*, vol. 13, no. 2, pp. 74–85, 2024, doi: 10.18036/estubtdc.1320922.

[3] International Diabetes Federation, *IDF Diabetes Atlas. In IDF Diabetes Atlas*, vol. 11th editi. 2025. [Online]. Available: https://www.idf.org/aboutdiabetes/type-2-diabetes.html

[4] R. Dhadse *et al.*, “Clinical Profile, Risk Factors, and Complications in Young-Onset Type 2 Diabetes Mellitus,” *Cureus*, vol. 16, no. 9, 2024, doi: 10.7759/cureus.68497.

[5] G. S, R. Venkata Siva Reddy, and M. R. Ahmed, “Exploring the effectiveness of machine learning algorithms for early detection of Type-2 Diabetes Mellitus,” *Meas. Sensors*, vol. 31, no. December 2023, p. 100983, 2024, doi: 10.1016/j.measen.2023.100983.

[6] F. M. Gilang, R. Ferdiansyah, and N. E. Ardian, “Prediksi Resiko Diabetes,” *Semin. Nas. Amikom Surakarta*, no. November, pp. 14–24, 2024.

[7] H. Imaduddin, W. Widayat, and F. Y. A’Ia, “Classification of Diabetes Using Ensemble and Individual Machine Learning Algorithms,” *2025 2nd Int. Conf. Adv. Innov. Smart Cities, ICAISC 2025*, 2025, doi: 10.1109/ICAISC64594.2025.10959526.

[8] Erlin, Yulvia Nora Marlim, Junadhi, Laili Suryati, and Nova Agustina, “Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 11, no. 2, pp. 88–96, 2022, doi: 10.22146/jnteti.v11i2.3586.

[9] D. Gunawan, “Classification of Privacy Preserving Data Mining Algorithms: A Review,” *J. Elektron. dan Telekomun.*, vol. 20, no. 2, p. 36, 2020, doi:

10.14203/jet.v20.36-46.

[10] K. R. Ummah, T. Karlita, R. Sigit, E. M. Yuniarno, I. K. E. Purnama, and M. H. Purnomo, "Effect of Image Pre-Processing Method on Convolutional Neural Network Classification of Covid-19 Ct Scan Images," *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 6, pp. 1895–1912, 2022, doi: 10.24507/ijicic.18.06.1895.

[11] O. O. Oladimeji, A. Oladimeji, and O. Oladimeji, "Classification models for likelihood prediction of diabetes at early stage using feature selection," *Appl. Comput. Informatics*, vol. 20, no. 3–4, pp. 279–286, 2024, doi: 10.1108/ACI-01-2021-0022.

[12] A. Ali Linkon *et al.*, "Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus," *IEEE Access*, vol. 12, no. September, pp. 165425–165440, 2024, doi: 10.1109/ACCESS.2024.3488743.

[13] Z. Amri, M. Rodi, M. N. Wathani, A. Bagja, and V. No, "Infotek : Jurnal Informatika dan Teknologi Prediksi Diabetes Menggunakan Algoritma K-Nearest (KNN) Teknik SMOTE-ENN Infotek : Jurnal Informatika dan Teknologi," vol. 8, no. 1, pp. 193–204, 2025.

[14] N. R. Panda, J. N. Mohanty, R. Bhuyan, P. K. Raut, and Manulata, "Exploring machine learning approaches for early diabetes risk prediction: A comprehensive examination of health indicators and models," *J. Assoc. Med. Sci.*, vol. 57, no. 3, pp. 155–165, 2024, doi: 10.12982/JAMS.2024.057.

[15] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[16] A. A. Putri Lo and V. J. E. Tjioe, "Penerapan Model CRISP-DM untuk Prediksi Penyakit Diabetes Menggunakan Metode K-Nearest Neighbor dan Logistic regression," *Pros. SENAM 2024 Semin. Nas. Sist. Inf. Inform. Univ. Ma Chung*, vol. 4, pp. 48–57, 2024, [Online]. Available: https://ocs.machung.ac.id/index.php/seminarnasionalmachung/article/view/452

[17] T. Alex, "Diabetes Health Indicators Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[18] M. C. Peréz, M. B. Calisto, and D. Riofrío, "Application of Machine Learning algorithms for the prediction of payment by agreement in a debt collection company with the CRISP-DM methodology," vol. 2020, no. February 2021, pp. 474–485, 2023, doi: 10.46254/sa03.20220112.

[19] A. Rianti, N. Wachid, A. Majid, and A. Fauzi, "CRISP-DM: Metodologi Proyek Data Science," *Pros. Semin. Nas. Teknol. Inf. dan Bisnis*, pp. 107–114, 2023.