

Diabetes Disease Prediction on Unbalanced Data Using SMOTE-Tomek Links and Random Forest Algorithm

Titis Fatmah Sukamto^{1✉}, Cathy Lintang Prameswary², Dedi Royadi³, Detin Sofia⁴

^{1,2} Information System, Faculty of Information and Communication Technology, Global Institute, Indonesia

³, Digital Business, Faculty of Business and Management, Global Institute, Indonesia

⁴, Informatics Engineering, Faculty of Information and Communication Technology, Global Institute, Indonesia

✉ **Corresponding Author** : Titis Fatmah Sukamto (e-mail: titisfatmah.253@gmail.com)

Article Information	ABSTRACT
Article History Received : May 27, 2025 Revised : June 3, 2025 Published : July 04, 2025	Diabetes, a chronic condition caused by the body's incapacity to generate or apply insulin as it should, is characterized by elevated blood sugar. If not treated early, the disease can lead to serious complications. This research aims to implement a machine learning-based classification model to predict diabetes, applying the methodology known as CRISP-DM (Cross-Industry Standard Process for Data Mining). The dataset was obtained from the Health Center of Sukatani Village, Rajeg, with a total of 2,075 records and 21 columns. The SMOTE-Tomek Links resampling technique was used to resolve the data's class imbalance. Five classification algorithms, Naive Bayes, Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbor (KNN), were compared in this study. Experiments revealed that the Random Forest algorithm performed the best with 97% accuracy, which increased to 99.64% after the application of SMOTE-Tomek Links. This best model was implemented in a web-based application using the Streamlit framework. The combination of the CRISP-DM approach, Random Forest algorithm, and SMOTE-Tomek Links proved to be effective in predicting diabetes, so that it can help medical personnel and the community in preventing, managing, and monitoring diabetes optimally.
Keywords: <i>Diabetes;</i> <i>Machine Learning;</i> <i>CRISP-DM;</i> <i>Random Forest;</i> <i>SMOTE-Tomek Links</i>	

INTRODUCTION

The WHO report (2024) shows that 800 million individuals worldwide have diabetes. In 2021, there were 19.5 million sufferers in Indonesia between the ages of 20 and 79 (World Health Organization, 2024). Unhealthy lifestyle, genetic factors, obesity, consumption of foods high in calories and fat, lack of exercise, smoking, alcohol, and stress are all factors that can cause diabetes to become dangerous (Ratnasari, 2023). Diabetes results in hyperglycemia, or elevated levels of blood sugar, when the pancreas fails to produce insulin or the body doesn't react well to it (Thakkar et al., 2021). Insulin is in charge of permitting glucose to enter cells so that it can be converted into energy (Jaiswal et al., 2021). This disorder is the main sign of diabetes mellitus (Ema et al., 2023). Symptoms of diabetes mellitus include excessive hunger, constant thirst, and frequent urination (Khanam & Foo, 2021).

Classification models that use the CRISP-DM data mining approach are needed because diabetes symptoms are diverse and difficult to identify. Mawarni et al. (2023) showed that this method with the Random Forest algorithm can detect early symptoms of diabetes with an accuracy

of 90.38%. Similarly, Arminarahmah and Mahalisa (2024) found that the KNN algorithm, which is used in the Streamlit application, is very effective in the classification of diabetes status. Several comparative studies have also been conducted. Emmanuel et al. (2021) evaluated multiple algorithms, SVM, Decision Tree, Naive Bayes, and KNN, within the CRISP-DM approach. They showed that KNN has the highest accuracy of 75%. However, a separate study by Febrian et al. (2022) stated that Naive Bayes has an accuracy of 76.07%, which is higher than KNN. Research conducted by Or. (2025) showed that the combination of Logistic Regression and SMOTE-Tomek Links algorithm improved the classification accuracy from 58.31% to 76.16%, which is an effective improvement in overcoming class imbalance.

The technique of finding patterns and important information in enormous amounts of data is known as data mining (Han et al., 2023). In addition, machine learning is used for data analysis and prediction (Cady, 2020). Classification method is one of the applications of machine learning, which serves to assign target categories to database objects (Rastogi and Bansal, 2023). However, classification accuracy can differ, especially in imbalanced data (Apriliyani and Salim, 2022). This imbalance is a big problem for machine learning, especially for medical and financial applications, because the error of the minority class has a significant effect (Jang, 2025).

This study aims to evaluate the best machine learning algorithm for predicting diabetes with imbalanced data using the CRISP-DM approach and the SMOTE-Tomek Links technique.

RESEARCH METHODS

This study employs a quantitative methodology with CRISP-DM, SMOTE-Tomek Links, and machine learning for diabetes prediction. In addition, this research applies Streamlit for user-friendly web applications (Prasetyo and Laksana, 2022). A Python package called Streamlit uses a straightforward interface to make web development easier (Ali et al., 2024). Data analysis was conducted with CRISP-DM, a structured method that supports the continuous development of predictive models (Montes, 2023). This method offers a systematic design approach. CRISP-DM has 6 main stages, Figure 1.

The CRISP-DM stage in diabetes prediction starts from business understanding, which identifies the needs and goals for building an accurate diabetes prediction system. Furthermore, in the data understanding stage, dataset collection and exploration are carried out, including statistical descriptions. The Data Preparation stage includes the process of data cleaning, normalization, visualization, splitting, and handling data imbalance with SMOTE-Tomek Links. Modeling is done by applying machine learning, training, testing, and hyperparameter tuning. Evaluation is done to evaluate the model using recall, precision, accuracy, F1-score, and AUC to ensure fit for purpose. Finally, the deployment stage is to implement the model into a web application based on the Streamlit framework.

In the medical world, the SMOTE (Synthetic Minority Oversampling Technique) technique has been widely used to handle unbalanced data (Ishaq et al., 2021). By balancing and filtering the data, SMOTE-Tomek Links improves the model's ability to classify classes (Or, 2025). As shown in the following formula:

$$Y' = Y_i + (Y_j - Y_i) * \gamma \quad (1)$$

In the SMOTE formula, Y' represents the synthetic data added from the minority class. Y_i is the original data from the minority class, while Y_j is the random data from the K-nearest neighbors of Y_i . The γ is a random value between 0 and 1 that is used to determine the interpolation position between Y_i and Y_j (Hairani and Priyanto, 2023). An ensemble system called Random Forest makes predictions based on voting choices (Cady, 2020). Another powerful machine learning approach for regression is Random Forest (Apriliyani and Salim, 2022). Random Forest excels in accuracy, handles large and unbalanced data, and assesses variables objectively (Salman et al., 2024). The Following is the Random Forest formula:

$$Gini = 1 - \sum_{i=1}^C (P_i)^2 \quad (2)$$

The symbol C represents the total number of classes in the dataset, while P_i indicates the proportion of samples in the i -th class (Saputra, 2023).

Five machine learning algorithms: Random Forest, Decision Tree, KNN, Naive Bayes, and Logistic regression were selected for their common use in medical classification and diverse learning approaches, allowing a comprehensive performance comparison on imbalanced health data.

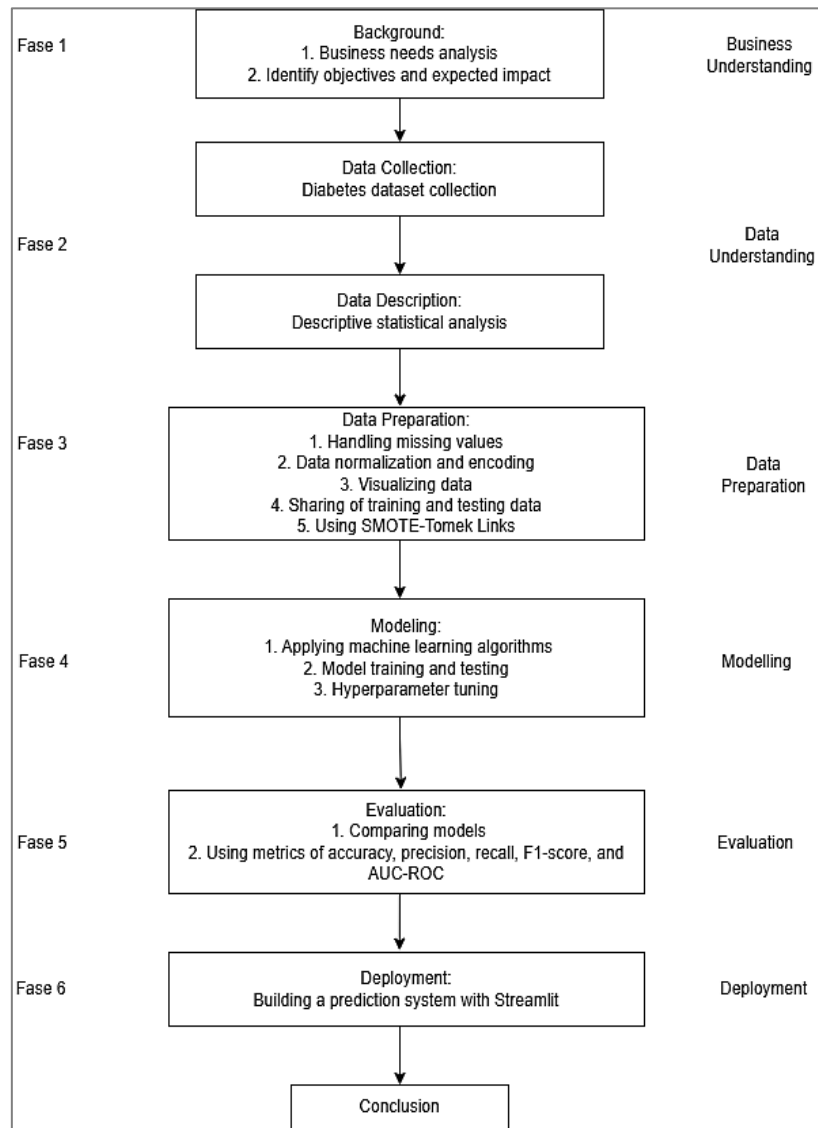


Figure 1. Stages of CRISP-DM

RESULTS AND DISCUSSION

The author applies the CRISP-DM method by going through the following steps:

Business Understanding:

The problem at UPTD Puskesmas Sukatani is the high number of diabetes cases that are not detected early due to a lack of routine checks, low risk awareness, lack of education about healthy lifestyles, and limited facilities for detection and monitoring.

Data Understanding

```

RangeIndex: 2075 entries, 0 to 2074
Data columns (total 21 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   TANGGAL_PEMERIKSAAN                     2075 non-null   object
1   TAHUN                                    2075 non-null   int64
2   TANGGAL_LAHIR                           2075 non-null   object
3   USIA                                     2075 non-null   int64
4   JENIS_KELAMIN                           2075 non-null   object
5   RIWAYAT_PENYAKIT_TIDAK_MENULAR_PADA_KELUARGA  75 non-null     object
6   RIWAYAT_PENYAKIT_TIDAK_MENULAR_PADA DIRI SENDIRI  75 non-null     object
7   MEROKOK                                  2075 non-null   object
8   KURANG_AKTIFITAS_FISIK                  2075 non-null   object
9   GULA_BERLEBIHAN                         2075 non-null   object
10  GARAM_BERLEBIHAN                       2075 non-null   object
11  LEMAK_BERLEBIHAN                       2075 non-null   object
12  KURANG_MAKAN_BUAH_DAN_SAYUR            2075 non-null   object
13  KONSUMSI_ALKOHOL                       2075 non-null   object
14  TEKANAN_DARAH_SISTOL                    2075 non-null   int64
15  TEKANAN_DARAH_DIASTOL                  2075 non-null   int64
16  TINGGI_BADAN(CM)                       2075 non-null   object
17  BERAT_BADAN(KG)                        2075 non-null   object
18  LINGKAR_PERUT(CM)                      2075 non-null   int64
19  PEMERIKSAAN_GULA                       2075 non-null   int64
20  DIAGNOSIS                               2075 non-null   object

```

Figure 2. Diabetes Prediction Dataset Info

The data source in this study came from UPTD Puskesmas Sukatani, with as many as 2075 records and 21 attribute columns based on Figure 2. Data analysis and processing were carried out using Google Colab with the Python programming language.

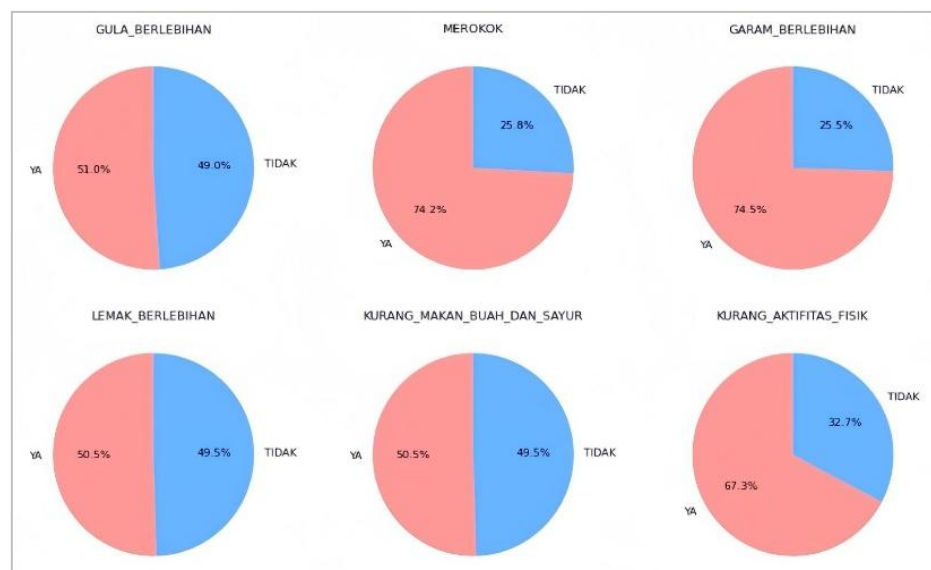


Figure 3. Visualization of Diabetes Lifestyle

Figure 3 shows the lifestyle habits of people with diabetes, with 51% consuming excessive sugar, 74.2% excess salt, and 67.3% lack of physical activity. Meanwhile, 50.5% consume excessive fat, and another 50.5% eat fewer fruits and vegetables. This visualization emphasizes the dominance of lifestyle risk factors in the increase of diabetes cases.

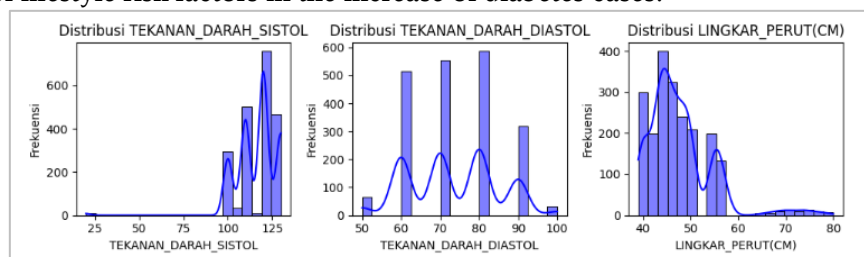


Figure 4. Visualization of Health Conditions

Figure 4 shows that people with diabetes have blood pressure and abdominal circumference within the normal range of 110-130 mmHg (systole) and 60-90 mmHg (diastole), and an abdominal circumference of 40-50 cm. Provides a basis for further risk analysis.

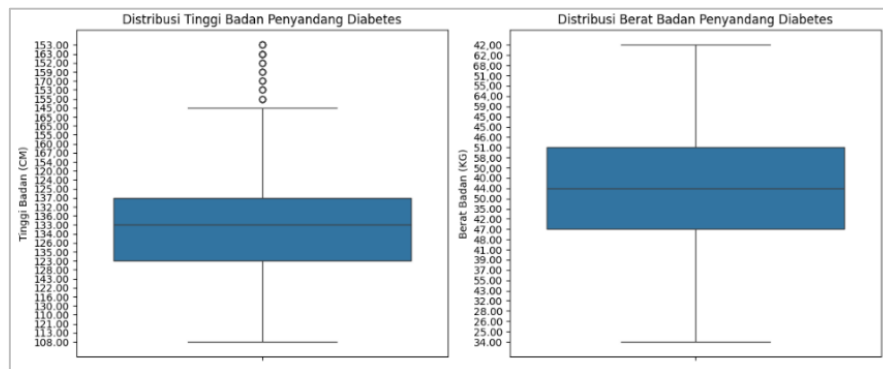


Figure 5. Visualization of Height and Weight of Diabetics

The visualization in Figure 5 shows that most diabetics have a height of 145-160 cm and a body weight of 50-56 kg, indicating a medium body size.

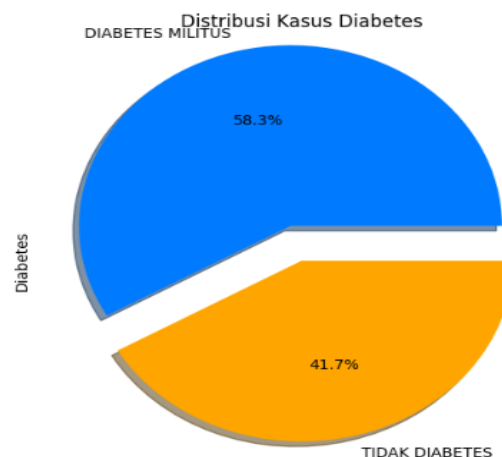


Figure 6. Visualization of Diabetes Case Distribution

The visualization of the chart diagram in Figure 6 shows that 58.3% are people with diabetes mellitus, 41.7% do not have diabetes, which means that more than half of the patients in this data are people with diabetes.

Data Preparation

RangeIndex: 2075 entries, 0 to 2074
Data columns (total 13 columns):

#	Column	Non-Null Count	Dtype
0	JENIS_KELAMIN	2075 non-null	object
1	KURANG_AKTIFITAS_FISIK	2075 non-null	object
2	GULA_BERLEBIHAN	2075 non-null	object
3	GARAM_BERLEBIHAN	2075 non-null	object
4	LEMAK_BERLEBIHAN	2075 non-null	object
5	KURANG_MAKAN_BUAH_DAN_SAYUR	2075 non-null	object
6	TEKANAN_DARAH_SISTOL	2075 non-null	int64
7	TEKANAN_DARAH_DIASTOL	2075 non-null	int64
8	TINGGI_BADAN(CM)	2075 non-null	object
9	BERAT_BADAN(KG)	2075 non-null	object
10	LINGKAR_PERUT(CM)	2075 non-null	int64
11	PEMERIKSAAN_GULA	2075 non-null	int64
12	DIAGNOSIS	2075 non-null	object

Figure 7. Dataset after data cleaning

Figure 7 displays the attributes retained after the data cleaning process. This stage involved merging several datasets from UPTD Puskesmas Sukatani and performing comprehensive cleaning procedures, including the removal of duplicate records, empty or missing data, outliers, and irrelevant attributes. The selection of relevant columns was also guided by expert interviews to ensure that only variables with analytical value were retained. Initially, the dataset contained 21 attributes and 2,075 records. After cleaning, the final dataset consisted of 12 selected attributes, while maintaining the total number of records at 2,075, and was then used for the modeling process.

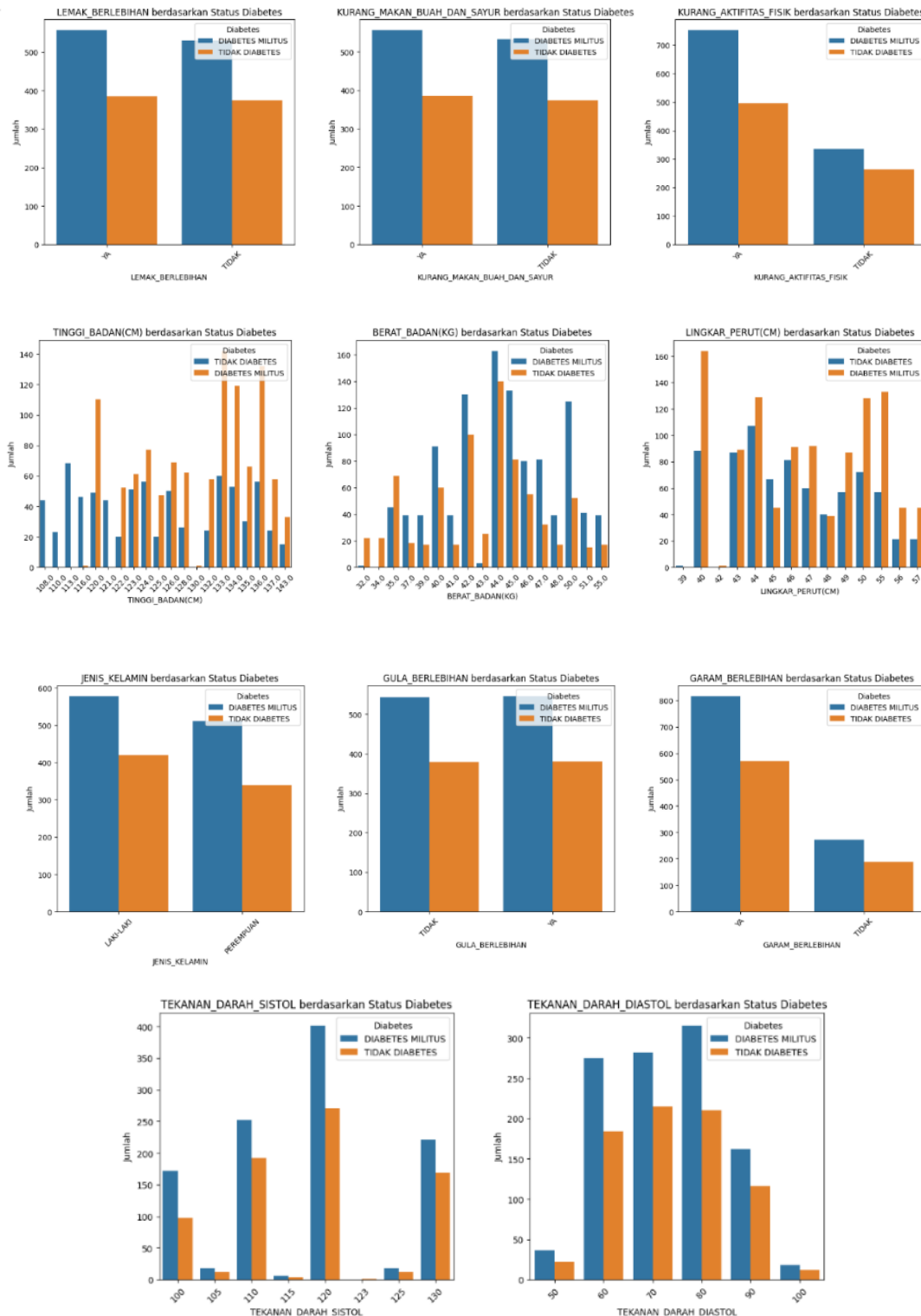


Figure 8. Visualization after Cleaning Data

Figure 8 illustrates the link between lifestyle, body metrics (such as weight and abdominal circumference), and blood pressure with diabetes status. Diabetic individuals typically follow an unhealthy diet, high in fat and low in fruits and vegetables, engage in less physical activity, and show higher body weight and abdominal measurements. They also tend to have elevated systolic and diastolic blood pressure. These patterns highlight the significant role of lifestyle and physical condition in increasing diabetes risk.

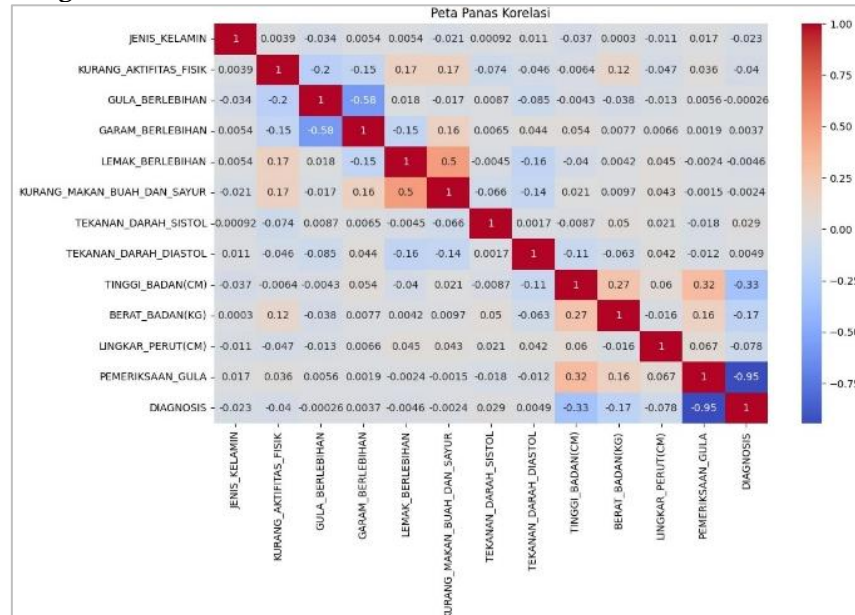


Figure 9. Heatmap Visualization

Figure 9 shows that the attributes of the sugar test and abdominal circumference (cm) are correlated with diagnosis, while the other variables tend to have low and less linear effects.

Modelling

After the data is cleaned, Encoding and normalization are performed on numeric attributes except diagnosis with MinMaxScaler. Then, splitting data with a ratio of 70:30 for test data shows that splitting data 1292 training data and 555 test data. This process is done to avoid overfitting. Resampling is done using the SMOTE-Tomek Links method to overcome imbalanced data between classes. This method helps to improve the training quality of the model, creating a balanced class distribution.

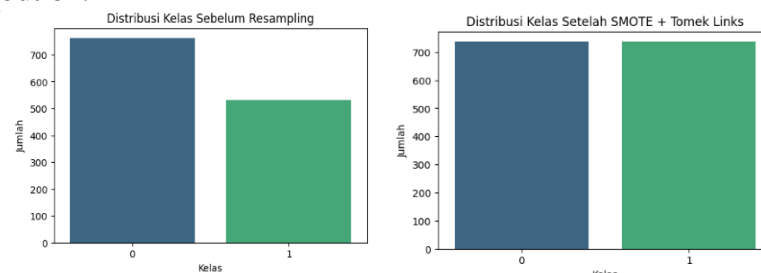


Figure 10. Before and After Resampling

Figure 10 shows that the imbalance of the data was successfully resolved, resulting in optimized training of the Random Forest algorithm for machine learning. The model was optimized with GridSearchCV and k-fold cross-validation to reduce overfitting.

Table 1. Model Comparison Table

Model	ACC	Prec	Rec	F1-score	AUC
Random Forest	99.64%	99.64%	99.64%	99.64%	99.68%
KNN	90.27%	90.25%	90.27%	90.24%	97.32%

Model	ACC	Prec	Rec	F1-score	AUC
Decision Tree	99.64%	99.64%	99.64%	99.64%	99.56%
Naive Bayes	97.84%	97.91%	97.84%	97.83%	99.35%
Logistic Reg	98.56%	98.57%	98.56%	98.56%	99.50%

Table 1 displays the accuracy, precision, recall, F1-score, and AUC values of the various models. With an accuracy of 99.64% and an AUC of 99.68%, Random Forest is the most accurate. Decision Tree comes in second with similar accuracy but slightly lower AUC, 99.56%. Although Decision Tree has similar accuracy, Random Forest was chosen because it generalizes better, is more resistant to overfitting, and has a higher AUC score. As an ensemble method, Random Forest produces more stable and reliable predictions than a single decision tree.

Evaluation

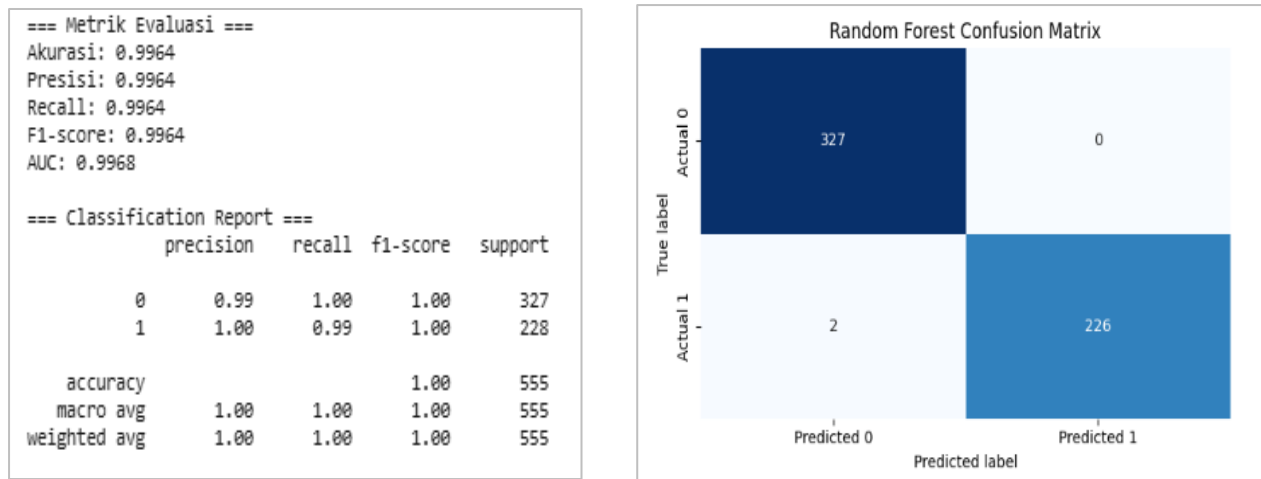


Figure 11. Random Forest Output and Confusion Matrix

Figure 11 shows the evaluation of Random Forest with very high performance, accuracy 99.64%, precision 99.64%, recall 99.64%, F1-score 99.64%, and AUC 99.68%. The confusion matrix reveals 226 True Positives, 327 True Negatives, no False Positives, and only 2 False Negatives, indicating excellent classification with minimal error. Although the Decision Tree had similar accuracy, Random Forest was chosen for better generalization, less overfitting, and higher AUC. As an ensemble, it produces more stable and reliable results than a single tree.

Deployment

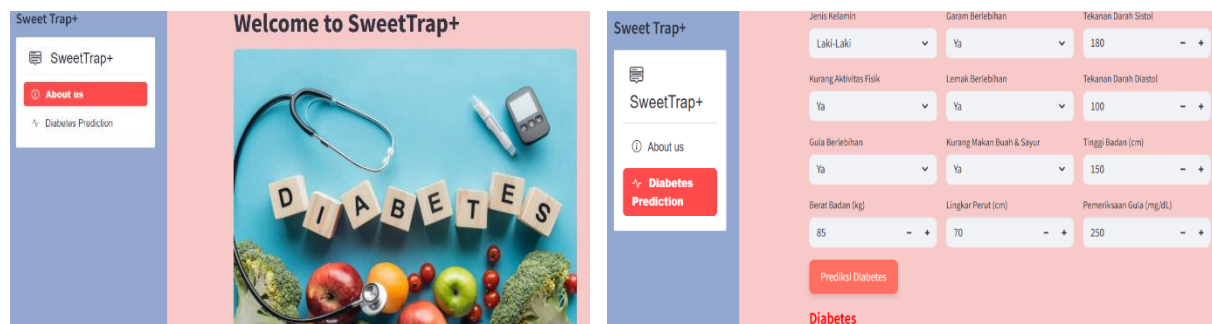


Figure 12. Diabetes Prediction Web View

Figure 12 displays the SweetTrap+ web application streamlined using Random Forest. Users enter data such as gender (Male), excessive salt (Yes), lack of activity (Yes), excessive fat (Yes), excessive sugar (Yes), lack of eating fruits and vegetables (Yes), blood pressure (180/100),

as well as body weight (85 kg), height (150 cm), abdominal circumference (55 cm), and blood sugar (250 mg/dl). Based on these inputs, the prediction result is “Diabetes.”

CONCLUSION

This study presents an accurate and accessible diabetes prediction tool by integrating SMOTE-Tomek Links, the CRISP-DM methodology, and Random Forest into a single pipeline. Implemented as a Streamlit web app, the model improved accuracy from 97% to 99.64% and supports early diagnosis in real-world healthcare settings.

Future research development is recommended to expand the scope of data collection, add additional features, and integrate medical records to make the prediction system more accurate and applicable. In addition, deep learning methods should be considered to strengthen the model's ability to process complex patterns from large and diverse data.

ACKNOWLEDGEMENTS

The authors would like to express their deepest gratitude to all those who have provided support for this research, especially to UPTD Puskesmas Sukatani, who has become the object of research as well as the dataset provider. As well as to the supervisor who provided important advice throughout the implementation of the research. The author also gives appreciation to parents and closest friends who always provide support and good wishes to complete this research.

REFERENCES

- Ali, I. T., Rahayu, Y., & Setiawan, A. (2024). Web Application Based on Machine Learning for Diabetes Detection using Microstrip Resonator and Streamlit. *Jurnal Teknologi Terapan*, 6(3), 120–131.
- Arminarahmah, N., & Mahalisa, G. (2024). Implementasi Model Machine Learning pada Klasifikasi Status Penyakit Diabetes Berbasis Streamlit. *Smart Comp: Jurnalnya Orang Pintar Komputer*, 13(3), 470–475. <https://doi.org/10.30591/smartcomp.v13i3.5866>
- Ema, P., Ludiana, & Immawati. (2023). Penerapan Senam Kaki Diabetes Untuk Meningkatkan Sensitivitas Kaki Pasien Diabetes Mellitus Tipe Ii Di Puskesmas Metro. *Jurnal Cendikia Muda*, 3(2), 235–244.
- Emmanuel, G., Hungilo, G. G., & Emanuel, A. W. R. (2021). Performance evaluation of machine learning classification techniques for Diabetes disease. *IOP Conference Series: Materials Science and Engineering*, 1098(5), 052082. <https://doi.org/10.1088/1757-899x/1098/5/052082>
- Ericha Apriliyani, & Salim, Y. (2022). Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset. *Indonesian Journal of Data and Science*, 3(2), 47–54. <https://doi.org/10.56705/ijodas.v3i2.45>
- Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2022). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216(2022), 21–30. <https://doi.org/10.1016/j.procs.2022.12.107>
- Field Cady. (2020). *Data Science: The Executive Summary - A Technical Book for Non-Technical People*. Wiley.
- Hairani, H., & Priyanto, D. (2023). A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data. *International Journal of Advanced Computer Science and Applications*, 14(8), 585–590. <https://doi.org/10.14569/IJACSA.2023.0140864>

- Han, J., Pei, J., & Tong, H. (2023). *Data Mining Concepts and Techniques*. In *Sustainability (Switzerland)*.
- Irwansyah Saputra. (2023). *Belajar Mudah Data Mining Untuk Pemula*. Informatika Bandung.
- Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. *IEEE Access*, 9, 39707–39716. <https://doi.org/10.1109/ACCESS.2021.3064084>
- Jaiswal, V., Negi, A., & Pal, T. (2021). A review on current advances in machine learning based diabetes prediction. *Primary Care Diabetes*, 15(3), 435–443. <https://doi.org/10.1016/j.pcd.2021.02.005>
- Jang, Y. (2025). Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE , RUS , and random forest methods : a prediction study. 48(2).
- Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432–439. <https://doi.org/10.1016/j.ict.2021.02.004>
- Mawarni, A. C., Rusdah, R., Hin, L. L., & Anubhakti, D. (2023). Deteksi Dini Gejala Awal Penyakit Diabetes Menggunakan Algoritma Random Forest. *IDEALIS : Indonesia Journal Information System*, 6(2), 165–171. <https://doi.org/10.36080/ideal.2023.v6i2.3018>
- Montes-bustamante, J. (2023). Diabetes Predictive Model Using the CRISP-DM Process for Disease Prevention Implementing Machine Learning Modelo Predictivo de Diabetes Utilizando el Proceso CRISP-DM para la Prevención de la Enfermedad Implementando Aprendizaje Automatizado. 4(2), 17–37.
- Or, B. (2025). Improving Requirements Classification with SMOTE-Tomek Preprocessing. 1–8. <http://arxiv.org/abs/2501.06491>
- Prasetyo, A. B., & Laksana, T. G. (2022). Optimasi Algoritma K-Nearest Neighbors dengan Teknik Cross Validation Dengan Streamlit (Studi Data: Penyakit Diabetes). *Journal of Applied Informatics and Computing (JAIC)*, 6(2), 194. <http://jurnal.polibatam.ac.id/index.php/JAIC>
- Rastogi, R., & Bansal, M. (2023). Diabetes prediction model using data mining techniques. *Measurement: Sensors*, 25(November 2022), 100605. <https://doi.org/10.1016/j.measen.2022.100605>
- Ratnasari, D. (2023). Comparison of Performance of Four Distance Metric Algorithms in K-Nearest Neighbor Method on Diabetes Patient Data. *Indonesian Journal of Data and Science*, 4(2), 101–112. <https://doi.org/10.56705/ijodas.v4i2.71>
- Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. <https://doi.org/10.58496/bjml/2024/007>
- Thakkar, H., Shah, V., Yagnik, H., & Shah, M. (2021). Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. *Clinical EHealth*, 4(2021), 12–23. <https://doi.org/10.1016/j.ceh.2020.11.001>
- World Health Organization. (2024). *Global Report on Diabetes 2024*. <https://www.who.int/publications>.