Early Detection of Diabetes Mellitus in Women via Machine Learning

Ahmad Zaki Arrayyan, Sisdarmanto Adinandra*

Master Program of Electrical Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia
*Corresponding author, e-mail: s.adinandra@uii.ac.id

Abstract – *Diabetes Mellitus (DM) is a major global health concern, responsible for 6.7 million deaths in 2021, equivalent to one death every five seconds. In Indonesia, it was the third leading cause of death in 2019, with a mortality rate of approximately 57.42 per 100,000 people. This study focuses on developing a diabetes prediction model using machine learning, aiming for an accuracy of at least 85%, and incorporates a chatbot-based system to identify potential diabetes in women. The research utilizes primary data, including glucose levels, blood pressure, body mass index, and age, as well as secondary data, such as pregnancy-related metrics, from the UCI Pima Indians Diabetes Database, which contains 768 records with eight attributes. The study evaluates the performance of three machine learning algorithms: Decision Tree, Logistic Regression, and Random Forest, using metrics such as accuracy, precision, recall, and F1-score. Among these models, the Decision Tree demonstrates excellent performance for Class 0, with precision, recall, and F1-score all at 0.97. However, its performance for Class 1, while decent, leaves room for improvement, achieving a precision of 0.80 and a recall of 0.84, resulting in an F1-score of 0.82. Logistic Regression also performs well for Class 0, with a precision of 0.95 and a recall of 0.99, yielding an F1-score of 0.97. Yet, it struggles with Class 1, where its precision is high at 0.93, but its recall drops significantly to 0.68, producing an F1-score of 0.79. Lastly, Random Forest emerges as the best-performing model overall, achieving an accuracy of 0.96. It excels for Class 0, with a precision of 0.96 and a recall of 0.99, leading to an F1-score of 0.97. For Class 1, it maintains high precision (0.93) but exhibits moderate recall (0.74), resulting in an F1-score of 0.82.*

***Keywords**: early detection; diabetes mellitus; public health; machine learning; prediction*

I. Introduction

Diabetes Mellitus (DM) is a significant global health concern, with a marked increase in incidence and mortality. The International Diabetes Federation reported that in 2021, diabetes was responsible for 6.7 million deaths worldwide, equating to one death every five seconds[1], [2]. In Indonesia, diabetes was the third leading cause of death in 2019, with a mortality rate of approximately 57.42 per 100,000 people[3]. DM is classified into two types: Type 1, which is autoimmune in nature, and Type 2, primarily linked to unhealthy lifestyles[4][5].

The rising number of diabetes cases is exacerbated by insufficient public awareness and understanding of early detection[6]. According to the Riskesdas 2018 survey, the prevalence of diabetes among women, diagnosed by a doctor, was 1.78%, with a confidence interval of 1.73 - 1.84[7]. The detection and management of diabetes remain challenging [8], particularly due to the inefficiency and potential errors associated with manual data processing, as well as the high costs of diagnosis and treatment.

To address these challenges, there is a need for predictive models to facilitate early screening and detection of diabetes. This study aims to develop a machine learning-based prediction model for diabetes, targeting an accuracy of at least 85%. Additionally, the research seeks to implement a chatbot-based prediction tool specifically for identifying diabetes risk in women. Utilizing machine learning techniques, including Decision Tree, Logistic Regression, and Random Forest algorithms, this study processes both primary and

Manuscript received September 2024, revised November 2024

secondary data from the UCI Pima Indians Diabetes Database. The goal is to evaluate and compare the performance of these algorithms to develop an effective predictive model accessible via a chatbot, enhancing early detection and management of diabetes.

II. Research Method

The research process outlined in this paper can be illustrated visually in Fig.1.

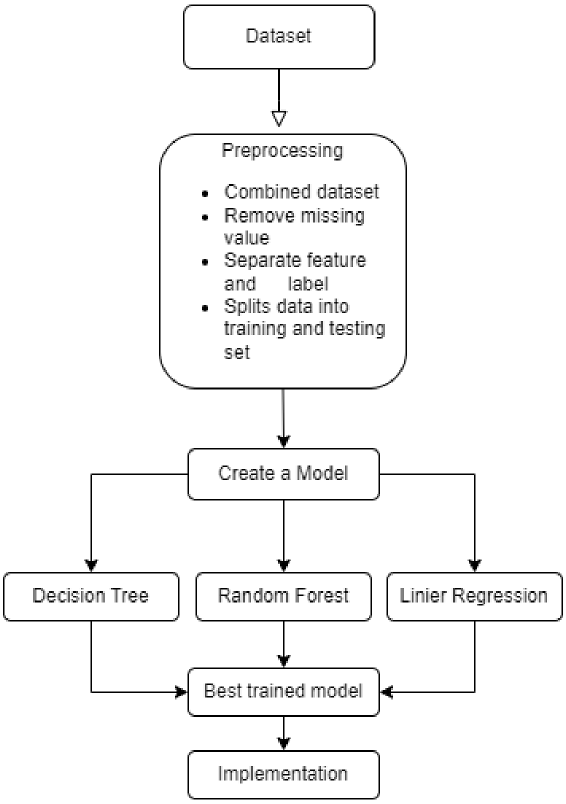


Fig.1. Flow of the methodology

II.1. *Dataset*

Primary dataset used for training was collected directly from 134 subjects, including features such as glucose levels, blood pressure, body mass index (BMI), and age. The inclusion criteria for participants in this study were women between the ages of 21 and 80. Additionally, a secondary dataset was sourced from the UCI Pima Indians Diabetes Database. This dataset consists of 768 samples and includes 9 attributes: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome..

II.2. *Preprocessing*

The raw data loaded from the CSV files (`data_uji.csv` and `diabetes.csv`) undergoes an essential preprocessing step to select specific columns deemed relevant for analysis. The selected columns, namely `Glucose`, `BloodPressure`, `BMI`, and `Age`, are designated as features that will serve as the input variables for the machine learning models. Additionally, the `Outcome` column is separated as the target label, representing the classification results that the models aim to predict. This step ensures that the data is structured in a format suitable for machine learning workflows, focusing only on the relevant information to optimize the training and evaluation processes. By isolating features and labels, the preprocessing lays a foundation for effective model development, facilitating tasks such as training, validation, and performance assessment.

II.3. *Learning Model*

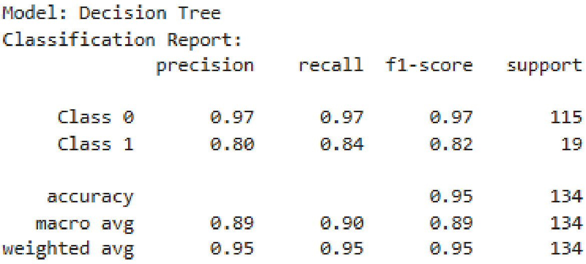
The progress on learning models involves the development and evaluation of three machine learning algorithms: Decision Tree, Logistic Regression, and Random Forest. These models are carefully defined and initialized with specific configurations to suit the data and ensure reproducibility, such as setting the random state and adjusting parameters like the maximum number of iterations for Logistic Regression. To assess their performance, the models are evaluated using Stratified K-Fold cross-validation, a robust technique that splits the training data into multiple folds while maintaining the class distribution across splits. This ensures that each model is tested on diverse subsets of data, reducing bias and variance in the evaluation process. The performance of the models is analyzed through various classification metrics, including precision, recall, and F1-score, as detailed in classification reports for each class (Class 0 and Class 1). Additionally, confusion matrices are generated to provide a comprehensive understanding of the models' ability to correctly classify outcomes, visualizing true positives, true negatives, false positives, and false negatives. These matrices are presented using heatmaps, which enhance interpretability and highlight areas where the models excel or require improvement. By combining robust cross-validation techniques, detailed metrics, and visual analysis, the progress establishes a strong foundation for selecting the most suitable algorithm to predict diabetes outcomes accurately.

III. Results and Discussion

As previously discussed, the performance of each classification model is evaluated by comparing key metrics such as accuracy and precission for class 0 and class 1.

III.1. *Decision Tree Result Analysis*

The performance of the Decision Tree model is evaluated using classification metrics, revealing high overall accuracy at 95%.



```
Model: Decision Tree
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.97      0.97      0.97       115
     Class 1       0.80      0.84      0.82        19

    accuracy                           0.95       134
   macro avg       0.89      0.90      0.89       134
weighted avg       0.95      0.95      0.95       134
```

Fig.2. Classification report of Decision Tree

For Class 0, which represents the majority class, the model demonstrates excellent precision, recall, and F1-score, all at 0.97, indicating its strong ability to correctly classify samples from this class. In contrast, the metrics for Class 1, the minority class, are lower, with precision at 0.80, recall at 0.84, and an F1-score of 0.82. This suggests that while the model performs well in identifying most instances of Class 1 (high recall), it occasionally misclassifies instances as Class 1 when they belong to Class 0 (lower precision). The macro average, which provides an unweighted mean of the metrics across both classes, reflects a balanced view of performance at 0.89 for precision, recall, and F1-score. However, the weighted average, which accounts for class imbalance, aligns closely with the overall accuracy, showing strong performance metrics of 0.95 across precision, recall, and F1-score. This highlights that the model is heavily influenced by the majority class (Class 0), performing exceptionally well for it while slightly underperforming for the minority class (Class 1).

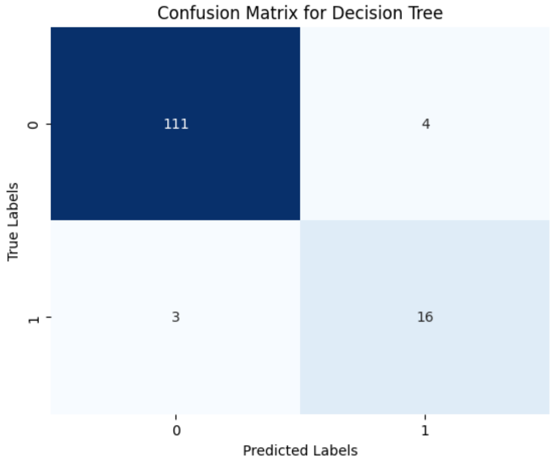
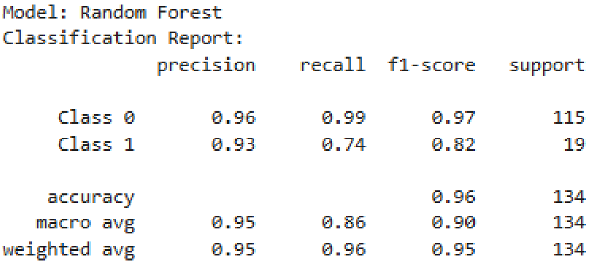


Fig.3. Confussion matrix of Decision Tree

The confusion matrix for the Decision Tree model illustrates the relationship between the true labels and the predicted labels, providing insight into the model's classification performance. The model correctly classifies 111 instances of Class 0 as Class 0, demonstrating its strong ability to identify the majority class with a high degree of accuracy. Additionally, it correctly identifies 16 instances of Class 1 as Class 1, showcasing reasonable performance for the minority class. However, there are 4 instances of Class 0 that are misclassified as Class 1, indicating some errors in precision for the majority class. Similarly, 3 instances of Class 1 are incorrectly predicted as Class 0, reflecting a slight limitation in the model's sensitivity toward the minority class. Despite these misclassifications, the model exhibits strong overall performance, excelling in accurately predicting the majority class while maintaining a good balance in identifying the minority class. The confusion matrix highlights both the model's strengths and areas for potential improvement, particularly in enhancing its sensitivity to the minority class.

III.2. *Random Forest Result Analysis*

The performance of the Random Forest model, as reflected in the classification report, demonstrates high overall accuracy at 96%.



```
Model: Random Forest
Classification Report:
              precision    recall  f1-score   support

     Class 0       0.96      0.99      0.97       115
     Class 1       0.93      0.74      0.82        19

    accuracy                           0.96       134
   macro avg       0.95      0.86      0.90       134
weighted avg       0.95      0.96      0.95       134
```

Fig.4. Classification report of Random Forest

For Class 0, which constitutes the majority class, the model achieves exceptional performance, with a precision of 0.96, a recall of 0.99, and an F1-score of 0.97. This indicates that the model is highly effective at correctly identifying Class 0 instances while minimizing false positives. For Class 1, which represents the minority class, the model achieves a precision of 0.93, showing its capability to accurately classify most instances predicted as Class 1. However, the recall for Class 1 is lower at 0.74, suggesting that the model struggles to identify some true instances of Class 1, resulting in a moderate F1-score of 0.82. The macro average, which provides an unweighted mean of precision, recall, and F1-scores across both classes, highlights a slight imbalance, with a recall of 0.86 compared to a higher precision of 0.95. The weighted average, which accounts for the class distribution, aligns closely with the overall accuracy, maintaining strong performance metrics of 0.95 for precision, recall, and F1-score. These results indicate that the Random Forest model excels in handling the majority class but has some limitations in detecting the minority class. While the model achieves robust overall performance, improvements in recall for Class 1 could enhance its ability to correctly identify minority class instances, ensuring more balanced predictions.

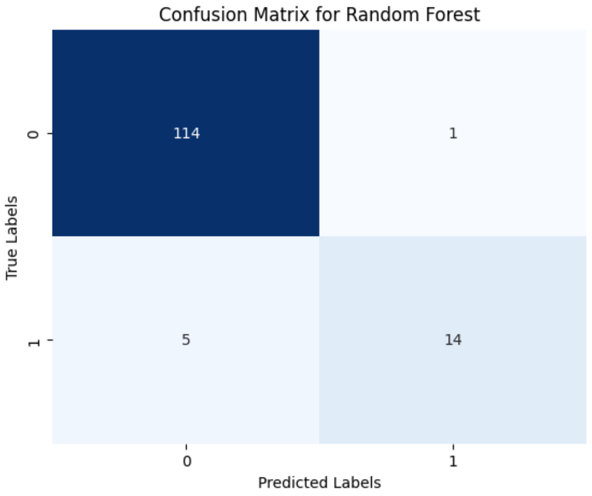


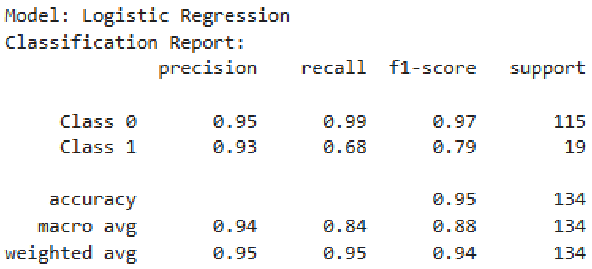
Fig.5. Classification report of Random Forest

The confusion matrix for the Random Forest model provides a detailed breakdown of the model's classification performance. It shows that 114 instances of Class 0 are correctly classified as Class 0, with only 1 instance misclassified as Class 1. This demonstrates the model's strong precision and recall for the majority class, reflecting its effectiveness in minimizing errors for Class 0 predictions. For Class 1, the model correctly classifies 14 instances as Class 1, but 5 instances are misclassified as Class 0. This indicates a limitation in the model's ability to detect all instances of the minority class, leading to a lower recall for Class 1.

Overall, the confusion matrix highlights that the model performs exceptionally well for the majority class while showing moderate performance for the minority class. The imbalance in misclassification rates suggests the model is more confident and accurate in identifying Class 0 but could benefit from improvements in its sensitivity toward Class 1. This analysis aligns with the classification report, which indicates high precision but relatively lower recall for Class 1, emphasizing the need for potential adjustments to better balance predictions across both classes.

III.3. *Logistic Regression Result Analysis*

The performance of the Logistic Regression model, as shown in the classification report, demonstrates high overall accuracy at 95%.



```
Model: Logistic Regression
Classification Report:
               precision    recall  f1-score   support

     Class 0       0.95      0.99      0.97       115
     Class 1       0.93      0.68      0.79        19

    accuracy                           0.95       134
   macro avg       0.94      0.84      0.88       134
weighted avg       0.95      0.95      0.94       134
```

Fig.6. Classification report of Logistic Regression

For Class 0, the majority class, the model performs exceptionally well, achieving a precision of 0.95, a recall of 0.99, and an F1-score of 0.97. This indicates that the model is highly effective at correctly identifying Class 0 instances, with very few false positives and almost no false negatives. For Class 1, the minority class, the model achieves a precision of 0.93, indicating that it classifies most instances predicted as Class 1 accurately. However, the recall for Class 1 is lower at 0.68, suggesting that the model struggles to correctly identify many true instances of Class 1, which results in a lower F1-score of 0.79. The macro average, which provides an unweighted mean of precision, recall, and F1-scores across both classes, shows a slight imbalance, with a recall of 0.84 compared to a higher precision of 0.94. The weighted average, which takes class distribution into account, closely aligns with the overall accuracy, maintaining strong performance metrics of 0.95 for precision, 0.95 for recall, and 0.94 for F1-score. These results suggest that the Logistic Regression model excels at handling the majority class but faces challenges in detecting the minority class.

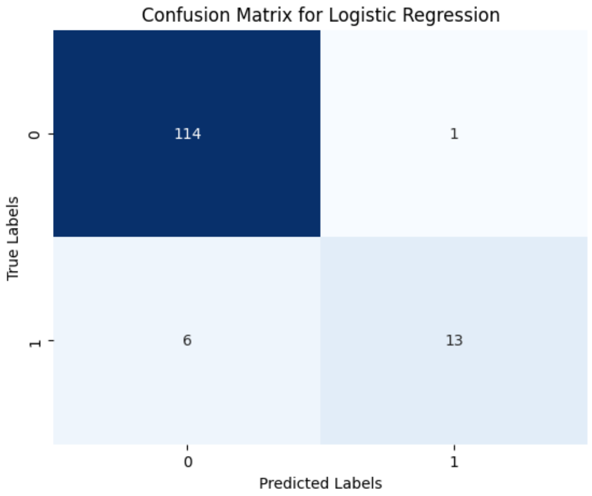


Fig.7. Classification report of Logistic Regression

The confusion matrix for the Logistic Regression model provides a detailed breakdown of its classification performance. It shows that 114 instances of Class 0 are correctly classified as Class 0, with only 1 instance misclassified as Class 1. This demonstrates the model's strong precision and recall for the majority class, reflecting its effectiveness in minimizing errors for Class 0 predictions. For Class 1, the model correctly classifies 13 instances as Class 1, but 6 instances are misclassified as Class 0. This indicates a limitation in the model's ability to detect all instances of the minority class, leading to a lower recall for Class 1.

Overall, the confusion matrix highlights that the model performs exceptionally well for the majority class while showing moderate performance for the minority class. The imbalance in misclassification rates suggests that the model is more confident and accurate in identifying Class 0 but could benefit from improvements in its sensitivity toward Class 1. This analysis aligns with the classification report, which would likely indicate high precision but relatively lower recall for Class 1, emphasizing the need for potential adjustments, such as rebalancing the dataset or modifying the decision threshold, to better balance predictions across both classes.

III.4. *Implementation*

At the core of this system lies the high-performance predictive model, meticulously developed and tested to ensure the highest level of accuracy. This model forms the basis of the chatbot's prediction capabilities, guaranteeing users receive dependable and precise results. By incorporating advanced normalization techniques alongside a robust predictive model, the chatbot offers an effective and accessible solution for users seeking analytical insights. This method combines technical rigor with a focus on user experience, making the chatbot a valuable tool for predictive analysis.

The process begins with a login page, providing a secure entry point for users to access the system.

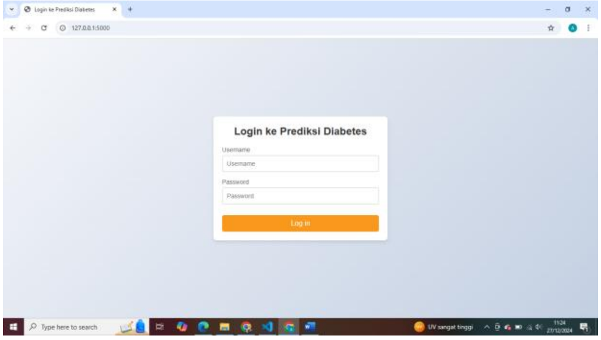
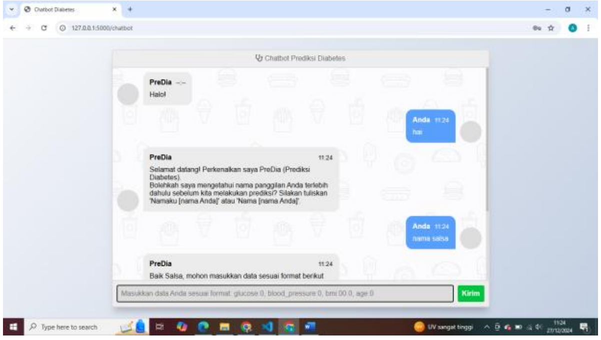


Fig.8. Login page before user enters the chatbot

Upon logging in, users are asked to provide their name, which facilitates a personalized interaction.



PreDia --:--
Halo!

Anda 11:24
hai

PreDia 11:24
Selamat datang! Perkenalkan saya PreDia (Prediksi Diabetes).
Bolehkah saya mengetahui nama panggilan Anda terlebih dahulu sebelum kita melakukan prediksi? Silakan tuliskan 'Namaku [nama Anda]' atau 'Nama [nama Anda]'.

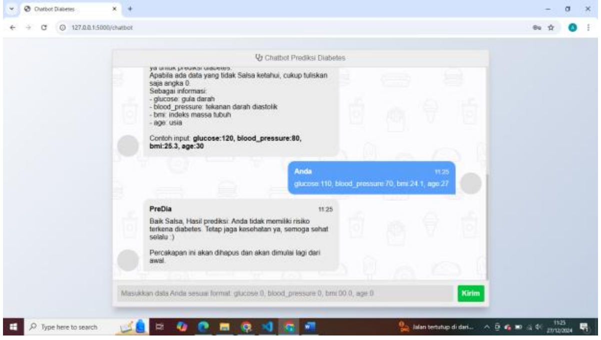
Anda 11:24
nama salsa

PreDia 11:24
Baik Salsa, mohon masukkan data sesuai format berikut

Masukkan data Anda sesuai format: glucose 0, blood_pressure 0, bmi 00.0, age 0

Fig.9. User asked to enter the name

Next, the system collects specific numerical data from the user, which is essential for generating predictions. Before the data is processed, a normalization step is performed, ensuring that the input values are scaled consistently, thereby improving the reliability and accuracy of the predictive model.



ya untuk prediksi diabetes.
Apabila ada data yang tidak Salsa ketahui, cukup tuliskan saja angka 0.
Sebagai informasi:
- glucose: gula darah
- blood_pressure: tekanan darah diastolik
- bmi: indeks massa tubuh
- age: usia

Contoh input: **glucose:120, blood_pressure:80, bmi:25.3, age:30**

Anda 11:25
glucose:110, blood_pressure:70, bmi:24.1, age:27

PreDia 11:25
Baik Salsa, Hasil prediksi: Anda tidak memiliki risiko terkena diabetes. Tetap jaga kesehatan ya, semoga sehat selalu :)

Percakapan ini akan dihapus dan akan dimulai lagi dari awal.

Masukkan data Anda sesuai format: glucose:0, blood_pressure:0, bmi:00.0, age:0

Fig.10. User asked to enter the name

Once normalized, the data is input into the chatbot’s predictive algorithm, which calculates a prediction score based on the processed values. This score represents the result of the system's analysis. To showcase its capabilities, the chatbot offers examples with varying input values, demonstrating how different data can lead to different prediction scores. This structured, user-friendly approach emphasizes the chatbot's ability to deliver actionable insights effectively.

IV. Conlcusion

Based on the analysis of the performance of each model, several key conclusions can be drawn regarding their strengths and weaknesses in handling the given classification task. Each model has demonstrated varying levels of effectiveness, particularly when it comes to accuracy, precision, recall, and the handling of class imbalances, which ultimately influence their overall performance and suitability for the task at hand.

(1) Accuracy: The Random Forest model achieves the highest overall accuracy at 96%, followed by the Decision Tree and Logistic Regression models at 95%.

(2) Class 0 (Majority class): All models perform well on Class 0, with precision, recall, and F1-scores close to 0.97-0.99. However, the Random Forest model slightly outperforms the others with a precision of 0.96 and recall of 0.99.

(3) Class 1 (Minority class): The Random Forest model has the highest precision for Class 1 (0.93), but its recall (0.74) is lower than the Decision Tree (0.84). The Logistic Regression model has a recall of 0.68, which is the lowest among the three models.

(4) Class 1 performance is the area where all models face challenges, but the Decision Tree slightly outperforms the others in terms of recall.

(5) Macro Average: The Decision Tree model performs best here with balanced precision, recall, and F1-scores of 0.89 across both classes, indicating a better balance in handling both classes equally.

(6) Weighted Average: The Random Forest model comes closest to the overall accuracy, with precision, recall, and F1-scores of 0.95, aligning well with the overall performance.

In conclusion, the Random Forest model proves to be the most effective choice for this system. With an overall accuracy of 96%, it outperforms both the Decision Tree and Logistic Regression models. Its performance is particularly strong for the majority class (Class 0), delivering high precision. While the model faces some challenges with the minority class (Class 1) in terms of recall, its higher precision for this class, combined with its overall strength, establishes it as the most robust model. This makes the Random Forest model the optimal foundation for the chatbot’s predictive capabilities, ensuring reliable and accurate results for users.

Acknowledgements

The work is funded by The Minister of Education, Culture and Higher Education Republic Indonesia through Hibah Penelitian Tesis Magister Scheme Contract Number: 064/STDirDPPM/70/DPPM/PTM-KEMDIKBUDRISTEK/ VI/2024.

References

[1] H. Sun *et al.*, “IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045,” *Diabetes Res. Clin. Pract.*, vol. 183, p. 109119, 2022.

[2] R. Vaishya, A. Misra, M. Nassar, and A. Vaish, “Global trend of research and publications in endocrinology, diabetes, and metabolism: 1996–2021,” *Int. J. Diabetes Dev. Ctries.*, vol. 44, no. 3, pp. 419–425, 2024.

[3] E. S. Darmawan, V. Y. Permanasari, L. V. Nisrina, D. Kusuma, S. R. Hasibuan, and N. Widyasanti, “Behind the Hospital Ward: In-Hospital Mortality of Type 2 Diabetes Mellitus Patients in Indonesia (Analysis of National Health Insurance Claim Sample Data),” *Int. J. Environ. Res. Public Health*, vol. 21, no. 5, p. 581, 2024.

[4] R. Khursheed *et al.*, “Treatment strategies against diabetes: Success so far and challenges ahead,”

Eur. J. Pharmacol., vol. 862, p. 172625, 2019.

[5] W. Bielka, A. Przezak, P. Molęda, E. Pius-Sadowska, and B. Machaliński, “Double diabetes—when type 1 diabetes meets type 2 diabetes: definition, pathogenesis and recognition,” *Cardiovasc. Diabetol.*, vol. 23, no. 1, p. 62, 2024.

[6] R. Patil and J. Gothankar, “Risk factors for type 2 diabetes mellitus: An urban perspective,” *Indian J. Med. Sci.*, vol. 71, no. 1, pp. 16–21, 2019.

[7] P. T. Rosha and S. Oksidriyani, “Integrated Healthcare Center in the elderly community in Indonesia: An analysis of Indonesian Family Life Survey Data,” in *Proceedings of International Conference on Physical Education, Health, and Sports*, 2023, pp. 355–372.

[8] T. Zhu, K. Li, P. Herrero, and P. Georgiou, “Deep learning for diabetes: a systematic review,” *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 7, pp. 2744–2757, 2020.

Authors’ information

Ahmad Zaki Arrayyan earned his Bachelor of Engineering (S.T.) from Universitas Muhammadiyah Yogyakarta, in March 2023. He is currently pursuing the master’s degree with the Islamic University of Indonesia, with a focus on the concentration of smart systems based on the Internet of Things.

RM Sisdarmanto Adinandra is an expert in Linear and Non-Linear Control Systems, Robotics, Signal Processing, and Instrumentation. He earned his Bachelor of Engineering (S.T.) from Universitas Gadjah Mada, Indonesia, and later pursued advanced studies in the Netherlands. He received his Master of Science (M.Sc.) from Technische Universiteit Delft and his Doctor of Philosophy (Ph.D.) from Technische Universiteit Eindhoven. His research and academic interests focus on developing innovative control systems and their applications in various fields, including robotics and instrumentation.