SISTEM PREDIKSI RISIKO DIABETES MENGGUNAKAN NEURAL NETWORK (MLPClassifier)

Abstract

This study aims to develop a diabetes risk prediction system based on machine learning using the Multi-Layer Perceptron (MLPClassifier) algorithm. The proposed system is implemented as a Python-based desktop application with a Tkinter graphical user interface to facilitate user interaction and result visualization. Several health-related parameters, including age, blood glucose level, blood pressure, body mass index (BMI), cholesterol level, and smoking status, are utilized as input features for the prediction model.

Prior to model training, the dataset undergoes a normalization process using the StandardScaler method to improve learning performance. The neural network architecture consists of multiple hidden layers with ReLU activation functions and is trained to estimate the probability of diabetes risk. The output of the model is presented as a percentage value, enabling users to better understand their relative risk level. In addition, the system provides categorized risk levels accompanied by personalized health recommendations to support early prevention and awareness.

The experimental results demonstrate that the developed system is capable of producing informative and interpretable diabetes risk predictions. By integrating machine learning techniques with an intuitive user interface, this system has the potential to assist individuals and healthcare practitioners in preliminary diabetes risk assessment and health education.

Keywords

Diabetes Prediction, Neural Network, MLPClassifier, Machine Learning, Health Risk Assessment, Tkinter Application

1. Introduction

Diabetes mellitus is a non-communicable disease that has become a major global health problem, with prevalence rates increasing every year. The International Diabetes Federation (IDF) reported that in 2021, more than 537 million adults worldwide were living with diabetes, and this number is projected to increase significantly in the coming years (IDF, 2021). In Indonesia, diabetes mellitus ranks among the top ten causes of death and imposes a substantial burden on the national healthcare system (Ministry of Health of the Republic of Indonesia, 2022). These conditions indicate that early detection and prevention are strategic measures to reduce the risk of severe diabetes-related complications.

Conventional diabetes detection is generally conducted through clinical and laboratory examinations, such as fasting blood glucose tests, oral glucose tolerance tests, and HbA1c measurements. Although these methods are accurate, limitations related to cost, access to healthcare facilities, and low public awareness cause many individuals to avoid routine

health screenings (American Diabetes Association, 2023). As a result, diabetes is often diagnosed at an advanced stage when complications have already developed.

Advancements in information technology and the availability of large-scale health data have encouraged the adoption of machine learning as an alternative approach in healthcare. Machine learning enables systems to learn patterns from historical data and generate predictions based on complex relationships among variables. Several studies have shown that machine learning approaches can improve accuracy and efficiency in predicting chronic diseases, including diabetes, compared to traditional statistical methods (Rajkomar et al., 2022).

One of the widely used machine learning algorithms for disease prediction is Artificial Neural Networks (ANN), particularly the Multi-Layer Perceptron (MLP). MLP has the capability to model nonlinear relationships among features, making it well suited for complex and heterogeneous medical data. Research conducted by Zou et al. (2021) demonstrated that MLP achieved strong predictive performance in diabetes classification using clinical parameters such as blood glucose levels and body mass index. Other studies have also reported that MLP provides competitive accuracy compared to decision tree and logistic regression algorithms in predicting metabolic diseases (Kavakiotis et al., 2021).

However, most previous studies primarily focus on experimental model evaluation and have not extensively integrated these models into user-friendly application systems for non-technical users. Moreover, prediction results are often presented in binary classification formats (positive or negative), which provide limited quantitative insight into individual risk levels.

Based on these limitations, this study proposes the development of a diabetes risk prediction system using a machine learning approach with the Multi-Layer Perceptron (MLPClassifier) algorithm, implemented as a Python-based desktop application. The system utilizes health parameters such as age, blood glucose level, blood pressure, body mass index (BMI), cholesterol level, and smoking status as input features. Prediction outputs are presented as percentage-based risk values, categorized risk levels, and health recommendations, thereby supporting early risk assessment and improving public health awareness.

2. Research Methodology

2.1 Research Type and Approach

This study employs a quantitative experimental approach using machine learning methods. The objective is to build a diabetes risk prediction model based on patient health data and to evaluate the prediction results generated by the Multi-Layer Perceptron algorithm.

2.2 Dataset

The dataset used in this study consists of patient data with several key health parameters. Each patient record includes the following attributes:

No	Attribute	Description
1	Age	Patient age (years)
2	Blood Glucose	Blood glucose level
3	Blood Pressure	Blood pressure
4	BMI	Body Mass Index
5	Cholesterol	Cholesterol level
6	Smoking Status	0 = Non-smoker, 1 = Smoker

Diabetes risk labels are defined based on logical conditions, where patients are categorized as at risk if their blood glucose level is ≥ 140 mg/dL or BMI is ≥ 30 kg/m^2. This approach is applied to create a controlled training dataset aligned with basic medical indicators.

2.3 Data Preprocessing

Prior to model training, data preprocessing is performed to enhance algorithm performance. The preprocessing steps include:
*Feature selection to identify relevant health parameters.
*Data normalization using the StandardScaler method to standardize numerical feature scales.
*Label generation based on predefined diabetes risk thresholds.

2.4 Model Architecture and Algorithm

The diabetes risk prediction model is built using the Multi-Layer Perceptron (MLPClassifier) algorithm implemented with the Scikit-learn library. The neural network architecture includes:
*Two hidden layers with 32 and 16 neurons
*ReLU activation function
*Maximum iteration parameter set to 3000
*Random state configuration to ensure training consistency
*The model is integrated into a pipeline with data normalization to ensure consistent preprocessing for training and testing data.

2.5 System Implementation

The system is implemented as a Python-based desktop application with a graphical user interface developed using Tkinter. Users can select patient data, analyze diabetes risk, and view prediction results in the form of:

*Diabetes risk percentage
*Bar chart visualization
*Health recommendations based on risk level

2.6 Research Workflow

The overall research workflow consists of:
*Dataset collection and understanding
*Data preprocessing
*MLPClassifier model training
*Diabetes risk prediction
*Result visualization and recommendation generation

2.7 Research Workflow (Textual Description)

The research process begins with collecting patient datasets containing key health parameters. Feature selection is then applied to ensure that only attributes relevant to diabetes risk prediction are used. Selected features include age, blood glucose level, blood pressure, body mass index (BMI), cholesterol level, and smoking status.

Next, the selected data undergo normalization using the StandardScaler method to equalize feature scales and prevent dominance of specific variables during model learning. Label generation is also conducted based on logical thresholds for blood glucose levels and BMI.

After preprocessing, the data are used to train the Multi-Layer Perceptron (MLPClassifier) model until convergence is achieved. The trained model is then utilized to predict diabetes risk for patient data. Prediction results are produced as probability values, converted into percentage risk levels, categorized accordingly, and presented alongside textual recommendations and graphical visualizations.

2.8 Neural Network Architecture (Textual Description)

The diabetes risk prediction model in this study is a Multi-Layer Perceptron (MLP) neural network consisting of multiple processing layers. The input layer receives six patient health parameters: age, blood glucose level, blood pressure, body mass index (BMI), cholesterol level, and smoking status, with each parameter represented by one neuron.

The input layer is connected to two hidden layers. The first hidden layer contains 32 neurons, and the second hidden layer contains 16 neurons. Both hidden layers utilize the Rectified Linear Unit (ReLU) activation function to capture nonlinear relationships among input features.

The output layer consists of a single neuron responsible for producing the probability value of diabetes risk. A sigmoid activation function is applied to ensure output values range between 0 and 1. These values are then converted into percentages for easier

interpretation. All preprocessing and modeling steps are integrated into a Scikit-learn pipeline to ensure consistency across training and prediction phases.

3. System Implementation

3.1 Development Environment

The diabetes risk prediction system is developed using the Python programming language due to its flexibility in data processing and machine learning application development. The primary libraries used include Pandas and NumPy for data handling, Scikit-learn for machine learning modeling, Matplotlib for data visualization, and Tkinter for the desktop-based user interface.

System development is conducted on a personal computer running a Windows-based operating system with Python 3.x. The selection of these libraries ensures efficient development and execution without requiring high computational resources.

3.2 Machine Learning Model Implementation

The machine learning model employed in this system is the Multi-Layer Perceptron (MLPClassifier), implemented using Scikit-learn. The model is constructed within a pipeline that integrates data normalization and model training processes.

Data normalization is performed using the StandardScaler method to standardize feature values to zero mean and unit variance. This step improves training stability and accelerates convergence during the MLP learning process.

The model architecture consists of two hidden layers with 32 and 16 neurons, respectively, and uses the ReLU activation function. The model is trained with a maximum of 3000 iterations to ensure optimal learning performance. Once trained, the model is used to generate probability-based diabetes risk predictions for patient data.

3.3 User Interface Implementation

The user interface is developed using the Tkinter library to facilitate interaction between users and the system. On the main application screen, users can select patient names from the available dataset without manually entering data.

After selecting a patient and initiating analysis, the system processes the data using the trained machine learning model. Prediction results are displayed in a new window in the form of diabetes risk percentages, visual charts, and health recommendations tailored to the patient's risk level.

3.4 Prediction Result Visualization

Prediction result visualization is implemented using the Matplotlib library and integrated into the Tkinter interface. The system displays bar charts representing diabetes risk percentages, providing a simple and intuitive visual representation for users.

In addition to individual predictions, the system offers visualization of diabetes risk distributions across all patients. This feature enables users or data administrators to identify overall risk trends and patients with higher relative risk levels.

3.5 Health Recommendation Implementation

Beyond displaying prediction results, the system provides health recommendations based on predefined rule-based logic. Recommendations are generated according to the categorized diabetes risk levels, offering preventive guidance relevant to each individual's condition.

This recommendation feature enhances the educational value of the system and promotes greater awareness of healthy lifestyle management and regular health monitoring.

4. Results and Discussion

4.1 System Testing Results

System testing is conducted to evaluate the performance of the Multi-Layer Perceptron (MLPClassifier) algorithm in predicting diabetes risk based on selected health parameters. The trained model is tested using the same dataset as an initial internal evaluation.
Prediction results are visualized using a confusion matrix to compare actual conditions with system predictions, as shown in **Figure X.**

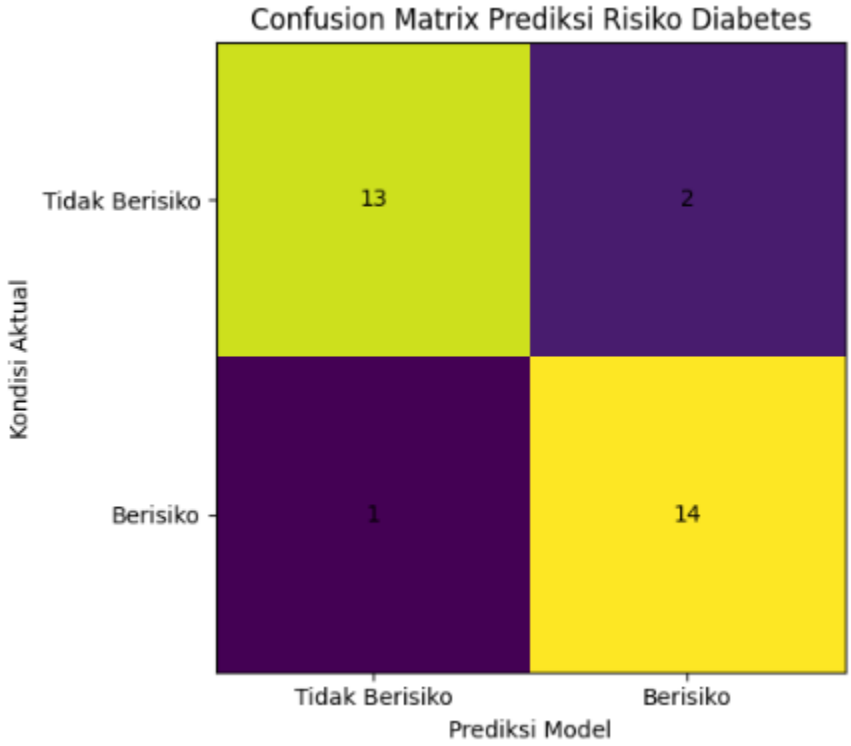

Figure X

Based on the confusion matrix, the following results are obtained:

- True Negative (TN): 13 samples
- False Positive (FP): 2 samples

- False Negative (FN): 1 sample
- True Positive (TP): 14 samples

Out of 30 test samples, the model correctly classified 27 samples.

4.2 Model Performance Analysis

Based on the confusion matrix, model performance metrics are calculated as follows:

Accuracy

Accuracy = (TP + TN) / Total = (14 + 13) / 30 = **90%**

An accuracy value of 90% indicates that the MLPClassifier model demonstrates strong capability in predicting diabetes risk for the given dataset. The high true positive rate suggests effective detection of at-risk individuals, while the low false negative rate indicates a minimal likelihood of missing patients with diabetes risk.

However, the presence of false positives indicates that in some cases, the system incorrectly classifies non-risk individuals as at risk. This limitation may result from the small dataset size, feature distribution imbalance, and limited data diversity.

4.3 Discussion

The experimental results show that the application of the Multi-Layer Perceptron algorithm provides sufficiently accurate and informative diabetes risk estimations. Integrating the machine learning model with a Tkinter-based graphical interface allows users to directly obtain percentage-based risk predictions and health recommendations.

Although the achieved accuracy is relatively high, the evaluation remains preliminary due to the use of the same dataset for training and testing. Future studies are recommended to employ train-test splitting or cross-validation techniques to obtain more objective and representative evaluation results.

5. Conclusion

Based on the conducted research and system testing, it can be concluded that the application of the Multi-Layer Perceptron (MLPClassifier) algorithm in a diabetes risk prediction system produces accurate and informative results. The model utilizes health parameters such as age, blood glucose level, blood pressure, body mass index (BMI), cholesterol level, and smoking status as input features.

Evaluation using a confusion matrix indicates that the system achieves an accuracy of 90%, demonstrating strong performance in distinguishing between at-risk and non-risk individuals. In addition to classification results, the system provides percentage-based risk estimations and health recommendations that are easy for users to understand.

The integration of machine learning models with a Tkinter-based graphical interface enhances system interactivity and applicability, making it a potential early-assessment tool for diabetes risk monitoring and public health awareness.

6. Recommendation

Based on the limitations identified in this study, the following recommendations are proposed for future development:

- Future research should utilize larger and more diverse datasets to improve model generalization across different patient characteristics.

- Model evaluation should apply train-test splitting or cross-validation techniques to obtain more objective performance assessments.
- Additional evaluation metrics such as precision, recall, and F1-score should be considered to provide a more comprehensive performance analysis.
- The system can be further enhanced by incorporating additional relevant features, such as family history of diabetes and physical activity levels, to improve prediction accuracy.
- Deployment of the system on web-based or mobile platforms is recommended to increase accessibility and usability.

REFERENCES

[1] World Health Organization, "Diabetes," *World Health Organization*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[2] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels, Belgium: International Diabetes Federation, 2022.

[3] A. Kumar, S. Jain, and R. Bansal, "Prediction of diabetes disease using machine learning algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 5, pp. 215–221, 2022, doi: 10.14569/IJACSA.2022.0130526.

[4] M. M. Rahman, M. K. Hasan, and M. S. Islam, "A comparative study of machine learning techniques for diabetes prediction," *Healthcare Analytics*, vol. 3, p. 100118, 2023, doi: 10.1016/j.health.2023.100118.

[5] Scikit-learn Developers, "MLPClassifier — scikit-learn documentation," 2024. [Online]. Available: https://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[6] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 3, no. 3, p. 160, 2022, doi: 10.1007/s42979-022-01002-5.

[7] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2021.