JOURNAL MEDICAL INFORMATICS TECHNOLOGY
MEDINFTech
medinftech.org ISSN 2988-7003 (online)

Early Diabetes Detection Using Machine Learning Models: A Case Study from Indonesian Clinical Data

Yasrizal[1] and Muhammad Haris[2]*

[1,2] *Universitas Nusa Mandiri, Indonesia*

ARTICLE HISTORY

Received: 22 December 26
Final Revision: 10 March 26
Accepted: 17 March 26
Online Publication: 31 March 26

KEYWORDS

Artificial Neural Network, Diabetes Mellitus, Early Detection, Indonesian Clinical Data, Machine Learning.

CORRESPONDING AUTHOR

14220014@nusamandiri.ac.id

DOI

10.37034/medinftech.v4i1.80

ABSTRACT

Diabetes is a major health problem that can significantly reduce life expectancy and increase the risk of serious complications such as kidney failure, stroke, and cardiovascular disease. Therefore, early detection is essential to prevent the progression of the disease. This study proposes a machine learning-based approach for early diabetes detection using a private dataset obtained from RSUP Persahabatan General Hospital in Jakarta, Indonesia. The dataset consists of 501 patient records with clinical and laboratory features extracted from the hospital's electronic medical record system. Several machine learning algorithms were implemented and compared, including Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, Naïve Bayes, Extreme Gradient Boosting, Ensemble methods, and Artificial Neural Networks. Feature selection was performed using ANOVA, and hyperparameter optimization was applied using GridSearchCV to improve model performance. The experimental results show that the Artificial Neural Network model achieved the best performance with an accuracy of 0.86 (86%). Statistical analysis using logistic regression identified systolic blood pressure, diastolic blood pressure, age, HDL cholesterol, and leukocyte levels as the most significant risk factors associated with diabetes. These findings demonstrate the potential of machine learning techniques to support early diabetes detection using clinical data from Indonesian healthcare settings.

1. Introduction

Along with the increasing prevalence of diabetes mellitus (DM) in Indonesia, the disease has developed into a serious public health problem. According to data from the Ministry of Health, the prevalence of diabetes mellitus in Indonesia increased from 5.7% in 2007 to 6.9% in 2013, and further increased to 8.5% in 2018 (MOH, 2008; MOH, 2013b; MOH, 2019). This steady increase indicates that diabetes has become one of the major health burdens affecting the Indonesian population. Diabetes is currently the third leading cause of death in Indonesia after ischemic heart disease and stroke. The proportion of deaths due to DM reached 7.8% of all causes of death (MOH, 2015a). This figure shows a significant increase compared to 2007 when diabetes ranked fifth with a proportion of 5.7% (MOH, 2008). Additional data revealed that in 2017, deaths due to DM ranked third in Indonesia, representing a dramatic increase from the ninth position in 1990. Among all causes of death, diabetes recorded the largest increase, reaching 162% over this period (IHME, 2018; MOH, 2018) [1]. These statistics highlight the urgent need for effective strategies for early detection and prevention of diabetes in Indonesia.

Early detection plays an important role in preventing complications and reducing mortality caused by diabetes. With the rapid advancement of information technology, artificial intelligence and machine learning techniques have been widely applied in the healthcare sector to assist disease prediction and diagnosis. Machine learning methods are capable of analyzing large and complex medical datasets, identifying hidden patterns, and supporting clinical decision-making processes more efficiently than traditional statistical approaches. Recent studies have highlighted the growing application of machine learning and deep learning models in diabetes prediction and management [2].

In recent years, numerous studies have explored the application of machine learning algorithms for diabetes prediction. Algorithms such as Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting have been widely implemented to classify patients based on their risk of developing diabetes [3], [4]. Furthermore, neural network models have shown promising results in capturing complex nonlinear relationships within medical data and improving prediction performance [5].

Several studies in Indonesia have also attempted to utilize machine learning approaches for diabetes prediction using hospital or public health datasets. These studies generally focus on comparing multiple classification algorithms or improving model performance through data preprocessing techniques. However, many existing studies still rely on limited datasets or emphasize algorithm comparison without extensively optimizing neural network models or exploring comprehensive clinical variables related to diabetes in Indonesian populations.

Therefore, further research is needed to develop an optimized diabetes prediction model that considers various clinical indicators and utilizes neural network approaches. In this study, machine learning approaches and statistical methods are used simultaneously to identify potential risk factors associated with diabetes and to propose a system that can effectively identify diabetic patients [6]. Furthermore, hyperparameter tuning techniques are applied to obtain the best predictive model [7].

Risk factor analysis was conducted using several independent variables derived from clinical and laboratory examinations, including age, systolic blood pressure, diastolic blood pressure, body mass index, weight, height, hemoglobin, leukocytes, platelets, uric acid, HDL, LDL, total cholesterol, creatinine, AST, and ALT. These variables were analyzed to identify the most influential factors associated with diabetes. The dataset used in this research was obtained from RSUP Persahabatan Jakarta, with the aim of generating diabetes prediction results that reflect the characteristics of the Indonesian population.

2. Research Method

Several machine learning classification methods were implemented in this study to obtain the best performance for early diabetes detection. In addition, statistical analysis was conducted to identify the most influential risk factors associated with diabetes. Figure 1 illustrates the overall research workflow. The process begins with dataset collection from RSUP Persahabatan General Hospital, followed by data preprocessing and feature selection to identify relevant variables. In parallel, risk factor analysis is conducted to determine variables associated with diabetes.

Next, hyperparameter tuning is performed to optimize the model parameters. Several machine learning algorithms are then implemented, including Logistic Regression, Random Forest, Decision Tree, SVM, XGBoost, Naïve Bayes, ANN, and Ensemble models. Finally, the models are evaluated and the results are obtained to determine the best-performing model for diabetes prediction.

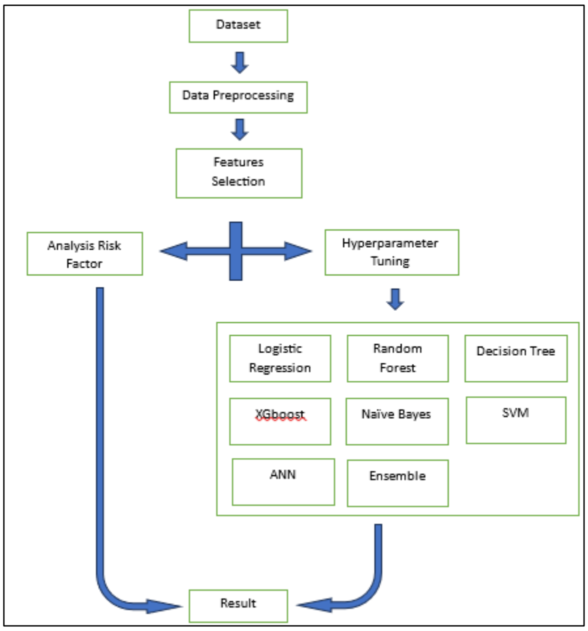

Figure 1. Research methodology for diabetes prediction using machine learning models.

2.1. Data Description

The dataset used in this study was obtained from RSUP Persahabatan General Hospital, which includes data from outpatients, inpatients, prevention programs, medical check-ups (MCU), and early diabetes detection examinations.

The dataset consists of 501 records with 24 variables (features) representing demographic, clinical, and laboratory indicators related to diabetes risk. Private datasets such as hospital data are typically limited in size and are obtained from institutions such as hospitals, banks, factories, or other service providers that serve as the focus of the study [8].

2.2. Data Preprocessing

Data preprocessing was conducted to improve data quality before the modeling process. This stage included data cleansing, handling missing values, and outlier analysis. First, all numerical variables were converted into float64 format to ensure compatibility with statistical analysis and machine learning algorithms. Missing values were handled using the mean imputation method to maintain data completeness without removing records, considering the limited number of observations. In addition, outlier analysis was performed using boxplot visualization to identify extreme values in the dataset. Although several variables contained outliers, these values represent actual laboratory test results; therefore, they were retained to preserve important information in the dataset.

2.3. Cross Validation

A technique for data partitioning called cross-validation (CV) divides the dataset into two groups: training data and test/validation data. The cross-validation technique used in this study is the Shuffle Split function, which produces a user-defined number of training and validation splits. Samples are sorted and separated into training and testing/validation sets using this procedure. Choosing this approach gives you choice over the quantity of data samples on either side of the training/validation set and the number of iterations [9]. In the cross-validation process, the dataset is split into 80% training set and 20% validation set. CV split values of 3 and 5 were used to test the data using the predetermined training and validation split.

2.4. Machine Learning Algorithms

Several machine learning algorithms were implemented and compared in this study to identify the most effective model for diabetes detection. The selection of these algorithms was based on their proven effectiveness in medical classification problems and their ability to handle different data characteristics.

2.4.1. Logistic Regression

In earlier research on diabetes patients risk factor identification, logistic regression was employed [10]. Logistic regression is a statistical classification method commonly used in medical research for binary classification problems. It is particularly useful for identifying relationships between independent variables and disease outcomes.

$$logit(p) = b_a + b_1X_1 + b_2X_2 + b_3X_3 +..+ b_kX_k \quad (1)$$

P is the probability of the characteristic of interest being present. The logit transform is defined as log:

$$odds = p\frac{p}{1-p} \quad (2)$$

$$= \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ presence\ of\ characteristic}$$

And

$$logit(p) = ln\left(\frac{p}{1-p}\right) \quad (3)$$

2.4.2. Support Vector Machines

Prior research on diabetes risk factor identification also made use of support vector machines (SVM) [11]. Support Vector Machine (SVM) is a supervised learning algorithm widely used for classification tasks and pattern recognition. SVM identifies an optimal hyperplane that separates classes in a high-dimensional feature space [12]. The hyperplane's mathematical equation is:

$$W.Y + p = 0 \quad (4)$$

Where b stands for the scalar data and W for the weighted vector

The SVM's Radial Basis Function (RBF) kernel is one of the classification techniques employed in this study. The RBF kernel is represented mathematically as follows:

$$K(x, y) = \exp\left(-\gamma||x - y||^2\right) \quad (5)$$

2.4.3. Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training and combines their predictions to produce a final result [13]. This method improves prediction accuracy and reduces overfitting [14].

2.4.4. Decision Tree

The investigations also employ classification based on decision trees [10]. Decision Tree (DT) is a tree-based classification algorithm that splits data into subsets based on feature values. It is widely used due to its interpretability and ability to model nonlinear relationships between variables [13].

2.4.5. Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a powerful boosting-based ensemble learning algorithm widely used in machine learning competitions and data analysis tasks [15], particularly for jobs like regression, ranking, and classification. It was created in 2014 by Tianqi Chen and gained a lot of popularity because of its effectiveness, quickness, and capacity for precise prediction-making. To increase performance overall, XGB combines predictions from several models using ensemble learning approaches. The method is based on boosting, which combines weak learners typically straightforward decision trees to create a powerful learner.

2.4.6. Naïve Bayes

The Naive Bayes algorithm consists of two names: Naive and Bayes. Naive is defined as This algorithm is called Nave because it contends that the development of one element is unrelated to the occurrence of another element. For instance, when color, shape, and flavor are taken into account, a red, round, and sweet fruit is identified as an apple. Because of this, each feature helps set one apple apart from the others without depending on the others. It is called Bayes because it is predicated on the notion of Bayes' Theorem [16]. The Bayes Theorem, sometimes known as Bayes' Rule or just the "Bayes' rule," is a technique for calculating the probability that an idea would be supported by prior knowledge. The decision is made using conditional probability. The Bayes theorem formula is as follows [17].

$$P(X) = \frac{P(H)P(H)}{P(X)} \tag{6}$$

2.4.7. Artificial Neural Network

Artificial Neural Networks (ANN) are computational models inspired by biological neural networks and are capable of learning complex nonlinear relationships in data [18]. In this study, the ANN model consists of several dense layers and includes a dropout layer to prevent overfitting, as suggested in previous studies [10]. The architecture of the ANN model is illustrated in Figure 2.

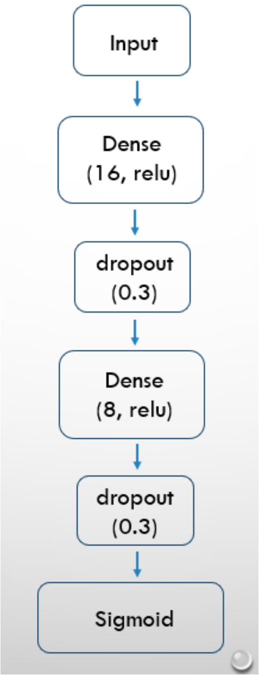

Figure 2. ANN Architectures

2.4.8. Ensemble

The ensemble approach combines two or more categorization methods to improve or enhance the overall performance. These days, boosting and bagging are the two ensemble-based strategies that are most in use. While bagging combines the outputs of several models to produce the final result, boosting is a sequential process in which the subsequent model corrects the inaccuracy of the preceding model [19]. The bagging classifier, which combines or averages predictions from base classifiers with random dataset segments to provide a final prediction, is an illustration of an ensemble meta-estimator [20].

2.5. Feature Analysis Using ANOVA

Analysis of variance, is a straightforward yet effective statistical technique that looks at the means of many groups or variables. It also establishes the degree to which the groupings diverge considerably [21]. The F-score, which indicates the significance or relatedness of a single feature from the independent variables, is measured using the scikit-learn program.

2.6. GridsearchCV

Grid Search is a method that provides a selection of predetermined parameter choices and searches for combinations of hyperparameters to yield the best model performance [22]. According to the formula for calculating the total number of combinations is the product of the number of values for each hyperparameter [23], the advantage of this method is that it can guarantee finding the best hyperparameter combination if the search space is not too large and all combinations are tried. The disadvantage is that it can be very slow and time consuming if the number of hyperparameter combinations is large.

2.7. Evaluation Matrix

The ratio of accurate predictions both positive and negative to the total data is known as accuracy. If our dataset has a relatively close proportion of False Positive and False Negative data (Symmetric), accuracy is a very good reference for algorithm performance. Accuracy responds to the question of what percentage of patients are correctly predicted to be diabetic and not diabetic from the entire data.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{7}$$

where the values for True Positives, True Negatives, False Positives, and False Negatives are represented, respectively, by the symbols TP, TN, FP, and FN [24].

3. Result and Discussion

3.1. Data Exploration

Before beginning the research, the data analysis is crucial. In order for the data to be properly understood, this can include information on the data that will be utilized in this study, as well as how to obtain data, sources, and classifications to distinguish between patients with diabetes and those without. Data exploration will be discussed in this section prior to conducting additional study to produce findings.

The dataset used in this study consists of 501 records with 24 variables (features) representing demographic, clinical, and laboratory indicators related to diabetes.

The data types include 1 datetime variable, 8 floating-point variables, and 15 integer variables. The variables contained in the dataset include sequence number, gender, medical record number, sample date, age, systolic blood pressure, diastolic blood pressure, body mass index, weight, height, hemoglobin, leukocytes, platelets, HbA1c, random glucose, fasting glucose, uric acid, HDL, LDL, total cholesterol, creatinine, AST, ALT, and diabetes diagnosis.

The distribution of the sample population based on gender and diabetes status is shown in Figure 3, while the distribution of all variables in the dataset is illustrated in Figure 4.

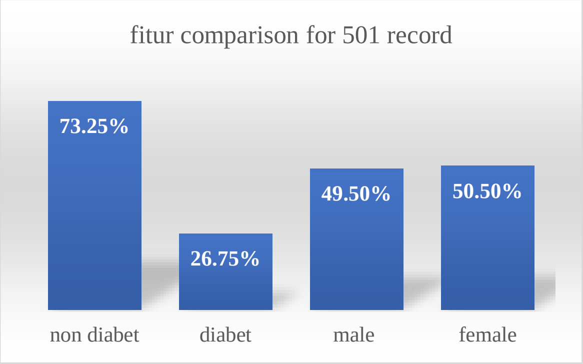

Figure 3. Sample Population

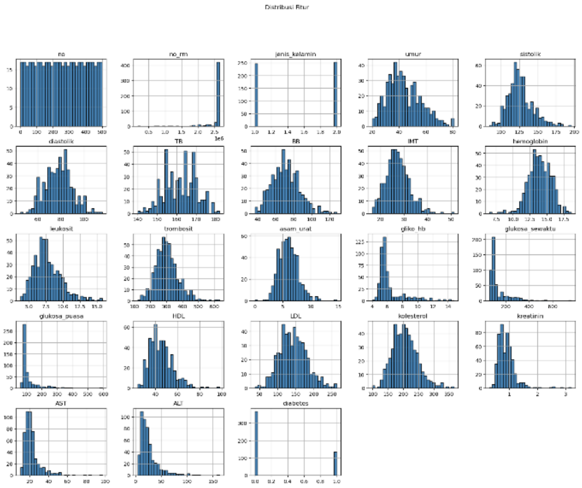

Figure 4. Feature Distribution

3.2. Data Cleansing

Convert Data into float64 format, converting integer data format to float in python is a simple process that has a huge impact on data analysis. Sometimes, a lot of important information is stored in formats that can't be processed right away. In this study there are 13 features with integer data types. When importing data from CSV or Excel files, the numbers in the table will be read as strings or integers, so that calculations and analysis can be carried out, the numerical numbers stored as integers need to be converted into float format to carry out statistical analysis and get the right calculation results.

3.3. Missing Value

Handling missing values is an important step in data preprocessing to prevent bias in model training. The dataset used in this study contains missing values in three variables: HbA1c, AST, and ALT, with the number of missing records ranging from 1 to 7 entries per variable.

Since the dataset size is relatively limited, missing values were handled using mean imputation, where the empty values were replaced with the average value of the respective feature. This approach allows the dataset to remain complete without removing records.

3.4. Data Outlier

Outlier analysis, Basically, outliers occur for a variety of reasons such as measurement errors, data processing errors or true anomalies. Understanding outliers is very important as their presence can have a huge impact on our data analysis.

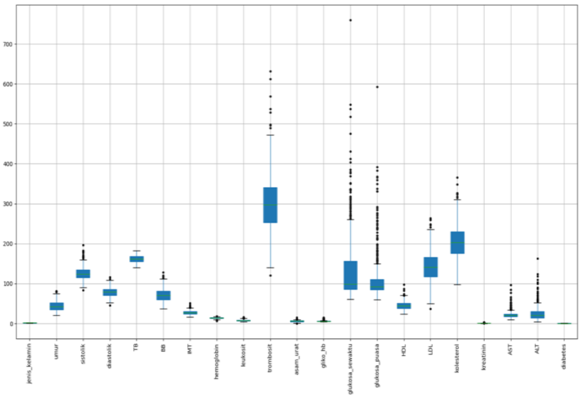

Figure 5. Outlier Analysis

Figure 5 presents the boxplot visualization used to identify potential outliers in each feature of the dataset. Several variables show the presence of outlier values, which may occur due to measurement variability, laboratory testing results, or natural variations in patient health conditions.

Since the dataset is derived from real laboratory examinations and the number of records is relatively limited, these outliers were retained in the dataset to preserve the integrity of the medical information and avoid losing potentially meaningful clinical data.

3.5. Analysis of Variance

Analysis of Variance Based on the calculation of the ANOVA F-score value 17 independent variables were selected from the 23 features available in the dataset. To determine the optimal features in this study, a feature selection algorithm was implemented. The feature selection method can also improve the quality of prediction. Analysis of Variance or ANOVA is a simple and powerful statistical method that examines the means of several groups or more variables [21]. It also determines how much the groups differ significantly from each other. In this study, the top features were selected and sorted based on the ANOVA F-score values shown in Table 1.

Table 1. F score Calculation Result

No	Features	Feature Scores
1	glukosa sewaktu	936,486
2	gliko hb	822,301
3	glukosa puasa	581,239
4	no	129,759
5	umur	100,913
6	no rm	57,496
7	sistolik	44,845
8	leukosit	43,782
9	kreatinin	17,928
10	HDL	7,355
11	AST	3,691
12	kolesterol	3,339
13	hemoglobin	2,575
14	diastolik	1,938
15	tinggi badan	1,909
16	ALT	1,224
17	berat badan	0,572
18	asam urat	0,453
19	jenis kelamin	0,220
20	Indek masa tubuh	0,056
21	LDL	0,034
22	trombosit	0,001

However, six variables were excluded from the modeling process for the following reasons:

1. There are 3 features, namely gliko_hb (HbA1c), glucose fasting and glucose ad random which have a very high correlation with the target or diabetes label and mainly used to indicate diabetes. It can cause the model to overfit to these features and fail to generalize the pattern. So that we avoid using those features with the assumption that the model should work as early detection;
2. There are 3 features, namely sequence number, medical record number and sample date that are not relevant to the research objectives because the information listed is only general and does not have a direct relationship with the target or label of diabetes.

3.6. Model Performance Results

In this study, eight machine learning algorithms were used to find the optimal model for early detection of diabetes. These algorithms are Logistic Regression (LR), Extreme Gradient Boosting (XGB) Support Vector Machine (SVM), Random Forest (RF) Decision Tree (DT), Naive Bayes (NB), Ensemble and Artificial Neural Network (ANN). In addition to determining the fit model of the eight algorithms, an improvement method with hyperparameter tuning techniques is also carried out on the ANN algorithm to get maximum results. Each model was trained using 80% training data and evaluated using 20% testing data from the total dataset of 501 patient records. The performance of each model was evaluated using accuracy, precision, recall, and F1-score, as shown in Table 2.

Table 2. Comparation of Accuracy

No	Model	Accuracy	Precision	Recall	F1 Score
1	Logistic Regression	0,84	0,83	0,84	0,83
2	XGB Classifier	0,84	0,84	0,84	0,84
3	SVM	0,85	0,85	0,85	0,84
4	Random Forest	0,85	0,85	0,85	0,84
5	Decision Tree	0,72	0,74	0,72	0,73
6	Gaussian Naïve Bayes	0,80	0,79	0,80	0,79
7	Ensemble	0,79	0,79	0,79	0,78
8	ANN	0,82	0,81	0,82	0,81

Based on Table 2 show that SVM and Random Forest achieved the highest accuracy of 0.85, followed closely by Logistic Regression and XGBoost with an accuracy of 0.84. Decision Tree produced the lowest performance with an accuracy of 0.72.

3.7. ANN Model Optimization

Searching for the best hyperparameters to obtain optimal prediction results using GridSearchCV on a previously built ANN model resulted in an accuracy increase of 0.4, from 0.82 to 0.86 for a model trained using 80% training data and 20% testing data. The results of the improvement using the best hyperparameters in the ANN algorithm are presented in Table 3.

Table 3. ANN With Hyperparameter Tuning

No	Model	Accuracy	Precision	Recall	F1 Score
1	ANN	0,861	0,858	0,861	0,852

In the ANN algorithm model after hyperparameter search is carried out to get the best accuracy results, a model evaluation is carried out with 20% testing data or as many as 101 records from all patient sample data used as input for analysis and results in a predictive value of the accuracy of the model of 86%. Figure 6. The following will display information on the evaluation results of the ANN model with hyperparameter tuning in confusion matrix format.

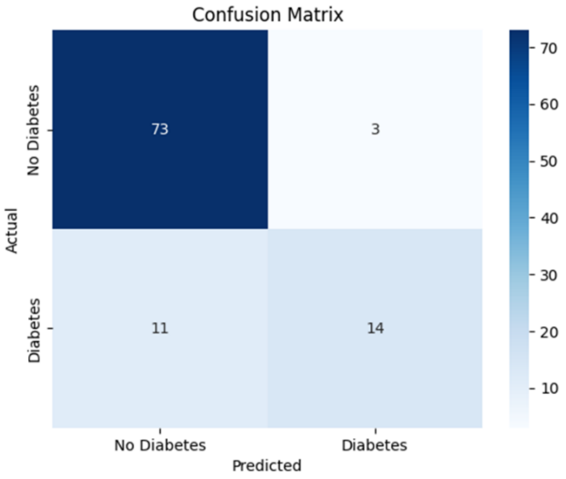

Figure 6. ANN Confusion Matrix

3.8. Risk Factor Analysis

In addition to predictive modeling, statistical analysis was conducted to identify significant risk factors associated with diabetes. The p-value calculation was performed using a logistic regression model.

Variables with p-values less than 0.05 are considered statistically significant predictors of diabetes. The results of the p-value calculation from the dataset used are presented in Table 4.

Table 4. Analysis P value

No	Features	P-value
1	Systolic Blood Pressure	0.000002
2	Age	0.000017
3	Diastolic Blood Pressure	0.001223
4	HDL Cholesterol	0.010946
5	Leukocyte (White Blood Cell Count)	0.010982
6	Body Weight	0.059106
7	Total Cholesterol	0.094952
8	Body Mass Index (BMI)	0.096078
9	LDL Cholesterol	0.113556
10	Height	0.118452
11	ALT (Alanine Aminotransferase)	0.224783
12	Platelet Count	0.437805
13	AST (Aspartate Aminotransferase)	0.496897
14	Hemoglobin	0.619208
15	Gender	0.675760
16	Uric Acid	0.733663
17	Creatinine	0.852175

Based on the results shown in Table 4, the most significant variables included systolic blood pressure, age, diastolic blood pressure, HDL cholesterol, and leukocyte count. These variables showed a statistically significant association with diabetes incidence in the dataset used in this study.

4. Conclusion

This study implemented several machine learning algorithms to develop an early diabetes detection model using a private dataset obtained from RSUP Persahabatan General Hospital. The dataset consisted of 501 patient records with multiple clinical and laboratory features extracted from the hospital's electronic medical record system.

Before model development, the dataset underwent preprocessing, including data cleansing, missing value handling, and feature selection. The data were divided into 80% training data and 20% testing data for model evaluation. Among the evaluated models, the Artificial Neural Network (ANN) achieved the best performance after hyperparameter optimization using GridSearchCV, improving the accuracy from 0.82 to 0.86 (86%).

In addition to predictive modeling, statistical analysis using logistic regression identified several significant risk factors associated with diabetes. Variables with p-values below 0.05 included systolic blood pressure, age, diastolic blood pressure, HDL cholesterol, and leukocyte count, indicating their strong relationship with diabetes occurrence.

However, this study is limited by the relatively small dataset size of 501 patient records. Future research is recommended to utilize larger datasets, potentially involving 2,000 to 10,000 patient records, to further improve model performance and generalization capability. Additionally, further studies may explore hyperparameter optimization for other machine learning algorithms to enhance diabetes prediction accuracy.

References

[1] M. Wahidin, A. M. Latelay, and M. Nitami, “DANA ALOKASI KHUSUS DAN CAPAIAN STANDAR PELAYANAN MINIMAL DIABETES MELITUS DI INDONESIA,” *Indones. J. Nurs. Heal. Sci.*, vol. 8, no. 1, pp. 84–90, 2023, doi: 10.47007/ijnhs.v8i1.6362.

[2] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, “Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review,” *Diabetol. Metab. Syndr.*, vol. 14, no. 1, p. 196, 2022, doi: 10.1186/s13098-022-00969-9.

[3] A. S. Chauhan, M. S. Varre, K. Izuora, M. B. Trabia, and J. S. Dufek, “Prediction of Diabetes Mellitus Progression Using Supervised Machine Learning,” *Sensors*, vol. 23, no. 10, 2023, doi: 10.3390/s23104658.

[4] B. F. Wee, S. Sivakumar, K. H. Lim, W. K. Wong, and F. H. Juwono, “Diabetes detection based on machine learning and deep learning approaches,” *Multimed. Tools Appl.*, vol. 83, no. 8, pp. 24153–24185, 2024, doi: 10.1007/s11042-023-16407-5.

[5] J. J. Sonia, P. Jayachandran, A. Q. Md, S. Mohan, A. K. Sivaraman, and K. F. Tee, “Machine-Learning-Based Diabetes Mellitus Risk Prediction Using Multi-Layer Neural Network No-Prop Algorithm,” *Diagnostics*, vol. 13, no. 4, 2023, doi: 10.3390/diagnostics13040723.

[6] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, “An efficient convolutional neural network for coronary heart disease prediction,” *Expert Syst. Appl.*, vol. 159, p. 113408, 2020, doi: https://doi.org/10.1016/j.eswa.2020.113408.

[7] R. R. Achmad and H. Muhammad, “Hyperparameter Tuning Deep Learning for Imbalanced Data,” *TEPIAN*, vol. 4, no. 2, pp. 90–101, 2023, doi: 10.51967/tepian.v4i2.2216.

[8] B. Sugara and A. Subekti, “PENERAPAN SUPPORT VECTOR MACHINE (SVM) PADA SMALL DATASET UNTUK DETEKSI DINI GANGGUAN AUTISME,” *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 177–182, 2019, doi: 10.33480/pilar.v15i2.649.

[9] C. A. Ramezan, T. A. Warner, and A. E. Maxwell, “Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification,” *Remote Sens.*, vol. 11, no. 2, 2019, doi: 10.3390/rs11020185.

[10] K. A. Hasan and M. A. M. Hasan, “Prediction of Clinical Risk Factors of Diabetes Using Multiple Machine Learning Techniques Resolving Class Imbalance,” in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1–6. doi: 10.1109/ICCIT51783.2020.9392694.

[11] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” *BMC Med. Inform. Decis. Mak.*, vol. 11, no. 1, p. 51, 2011, doi: 10.1186/1472-6947-11-51.

[12] A. Jimeno-Yepes, “Hyperplane bounds for neural feature mappings,” *CoRR*, vol. abs/2201.05799, 2022, [Online].

Available: https://arxiv.org/abs/2201.05799

[13] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, “Classification and prediction of diabetes disease using machine learning paradigm,” *Heal. Inf. Sci. Syst.*, vol. 8, no. 1, p. 7, 2020, doi: 10.1007/s13755-019-0095-z.

[14] H. A. Salman, A. Kalakech, and A. Steiti, “Random Forest Algorithm Overview,” *Babylonian J. Mach. Learn.*, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.

[15] M. Li, X. Fu, and D. Li, “Diabetes Prediction Based on XGBoost Algorithm,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 768, no. 7, p. 72093, Mar. 2020, doi: 10.1088/1757-899X/768/7/072093.

[16] K. L. Priya, M. S. Charan Reddy Kypa, M. M. Sudhan Reddy, and G.s R. Mohan Reddy, “Retracted: A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier,” in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 2020, pp. 603–607. doi: 10.1109/ICOEI48184.2020.9142959.

[17] R. Rachman and S. Moritamil, “SISTEM PAKAR DETEKSI PENYAKIT REFRAKSI MATA DENGAN METODE TEOREMA BAYES BERBASIS WEB,” *J. Inform.*, vol. 7, no. 1, pp. 68–76, 2020, [Online]. Available: https://ojs.bsi.ac.id/index.php/ji/article/view/7267/pdf

[18] N. S. El_Jerjawi and S. S. Abu-Naser, “Diabetes Prediction Using Artificial Neural Network,” *J. Adv. Sci.*, vol. 121, pp. 55–64, 2018, doi: 10.14257/ijast.2018.121.05.

[19] K. Oliullah, M. H. Rasel, M. M. Islam, M. R. Islam, M. A. H. Wadud, and M. Whaiduzzaman, “A stacked ensemble machine learning approach for the prediction of diabetes,” *J. Diabetes Metab. Disord.*, vol. 23, no. 1, pp. 603–617, 2024, doi: 10.1007/s40200-023-01321-2.

[20] T. Manimegalai, J. Manju, M. M. Rubiston, B. Vidhyashree, and R. T. Prabu, “Prediction of OPTIMIZED Stock Market Trends using Hybrid Approach Based on KNN and Bagging Classifier (KNNB),” in *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*, 2022, pp. 257–262. doi: 10.1109/CSNT54456.2022.9787638.

[21] A. Smiley, D. King, J. Harezlak, P. Dinh, and A. Bidulescu, “The association between sleep duration and lipid profiles: the NHANES 2013–2014,” *J. Diabetes Metab. Disord.*, vol. 18, no. 2, pp. 315–322, 2019, doi: 10.1007/s40200-019-00415-0.

[22] M. Azhar and H. F. Pardede, “Klasifikasi Dialek Pengujar Bahasa Inggris Menggunakan Random Forest,” *J. Media Inform. Budidarma*, vol. 5, no. 2, pp. 439–446, 2021, doi: 10.30865/mib.v5i2.2754.

[23] S. Shekhar, A. Bansode, and A. Salim, “A Comparative study of Hyper-Parameter Optimization Tools,” in *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 2021, pp. 1–6. doi: 10.1109/CSDE53843.2021.9718485.

[24] K. A. Hasan and M. A. M. Hasan, “Classification of Parkinson’s Disease by Analyzing Multiple Vocal Features Sets,” in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 758–761. doi: 10.1109/TENSYMP50017.2020.9230842.