Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm

Erlin[1*], Yulvia Nora Marlim[2], Junadhi[3], Laili Suryati[4], Nova Agustina[5]

***Abstract*—Diabetes is one of the deadliest diseases in the world, including in Indonesia. It can cause complications in numerous body parts and increase the overall risk of death. One way to detect diabetes is to use machine learning algorithms. Logistic regression is a classification model in machine learning widely used in clinical analysis. In this paper, a predictive model was created in Python IDE using logistic regression to conduct an early detection if a person has diabetes or not depending on the initial data provided. The experiment was carried out using a dataset from the Pima Indians Diabetes Database, which consisted of 768 patient data with eight independent variables and one dependent variable. Exploratory data analysis was applied to obtain maximum insight of the datasets owned by using statistical assistance and presenting them through visual techniques. Some dataset variables contained incomplete data. Missing data values were replaced with the median value of each variable. Unbalanced data was handled using the synthetic minority over-sampling technique (SMOTE) to increase the minority class through synthetic data sampling. The model was evaluated based on the confusion matrix, which showed a reasonably good performance with an accuracy value of 77%, precision of 75%, recall of 77%, and F1-score of 76%. In addition, this paper also used the grid search technique as a hyperparameter tuning that could improve the performance of the logistic regression model. The primary model performance with the model after applying the grid search technique was tested and evaluated. The experimental results showed that the hyperparameter tuning-based model could improve the performance of the logistic regression algorithm for prediction with an accuracy value of 82%, precision of 81%, recall of 79%, and F1-score of 80%.**

***Keywords*—Early Detection, Diabetes, Machine Learning, Logistic Regression, Grid Search.**

I. Introduction

Diabetes is a chronic metabolic disease characterized by elevated levels of glucose (blood sugar) that are higher than average, caused by impaired insulin secretion or impaired biological effects [1]. Diabetes can cause complications in many body parts and increase the overall risk of premature death. Possible complications include kidney failure, leg amputation, vision loss, and nerve damage. Adults with diabetes also have a two to triple increased heart attack and stroke risk. During pregnancy, poorly controlled diabetes will increase the risk of fetal death and other complications [2].

The number of people with diabetes increases yearly, both from the number of cases and the prevalence. In 2019, globally, the number of people with diabetes reached 463 million people. This number is predicted to grow up to 700 million people by 2045. Most diabetics live in low and middle-income countries, while 1.6 million deaths are directly caused by diabetes each year. It makes diabetes one of the ten leading causes of death worldwide [3].

In 2019, Indonesia ranked 7th in the world, after China, India, the United States, Pakistan, Brazil, and Mexico, as the country with the highest number of diabetics, with 10.7 million people. In 2020, this number increased to 10.8 million, with a prevalence rate of patients with diabetes reaching 6.2%. It is estimated that the number of diabetics in Indonesia will increase to 16.7 million in 2045 [4]-[7].

Regarding the relationship between the risk of developing complications from diabetes and the effects of death caused by this disease, early detection of diabetes is essential. Patients can delay and prevent the disease progression to acute diabetes when detected early. Disease prevention is significantly cheaper and more accessible than treating hyperglycemia and diabetes complications. Therefore, identifying, diagnosing, and analyzing diabetes quickly and accurately is a beneficial and crucial research topic. In the medical field, the diagnosis of diabetes is based on blood sugar levels, including current blood sugar levels, fasting blood sugar levels, and blood sugar tolerance levels [8]-[9]. The results of measuring blood sugar levels will indicate whether a person has diabetes. The earlier the diagnosis and detection are made, the easier it is for diabetes to be controlled and treated.

One way to detect diabetes is to use machine learning algorithms [10]-[12]. This algorithm has been widely used in various fields, including the health sector [13]-[14]. Logistic regression is one of the popular machine learning algorithms used for classification problems and is a predictive analysis algorithm based on the concept of probability. Logistic regression provides a better level of accuracy compared to k-nearest neighbor (k-NN) [15], decision tree [16], or other classifier models [17].

Several other studies also strengthen the results of previous studies regarding the reliability of logistic regression in predicting various types of disease. The use of logistic

[1,2] *Program Studi Teknik Informatika, Institut Bisnis dan Teknologi Pelita Indonesia, Jl. Jend. Ahmad Yani, Pekanbaru, 28127, INDONESIA (e-mail: [1]erlin@lecturer.pelitaindonesia.ac.id; [2]yulvia.nora@lecturer.pelitaindonesia.ac.id)*

[3] *Program Studi Teknik Informatika, STMIK Amik Riau, Jl. Purwodadi Indah, Km. 10 Pekanbaru, 28294, INDONESIA (e-mail: junadhi@sar.ac.id)*

[4] *Accounting Study Program, Universitas Persada Indonesia, Jl. Diponegoro No. 74 Jakarta Pusat, INDONESIA (e-mail: lailisuryati61@gmail.com)*

[5] *Program Studi Teknik Informatika, Sekolah Tinggi Teknologi Bandung, Jl. Soekarno-Hatta No. 378, Kidul Bandung, INDONESIA (e-mail: nova@sttbandung.ac.id)*

[*] *Corresponding Author*

[Received: 11 December 2021, Revised: 24 February 2022]

regression to predict cardiovascular disease achieved an accuracy of 85% [18]. Logistic regression was also used to predict chronic kidney disease and showed that logistic regression tended to have lower overfitting than random forests and neural networks [19]. Identification and prediction of liver disease by comparing the logistic regression algorithm with four other algorithms also showed that the accuracy of logistic regression was better than k-NN, support vector machine (SVM), decision tree, and random forest [20]. Another study also compared three machine learning models, namely neural network, naïve Bayes, and logistic regression, to detect diabetes in 768 data derived from Kaggle data. The experimental results showed that the logistic regression algorithm was better than the other two algorithms, with an accuracy of 75.78%, compared to Nave Bayes with an accuracy of 74.87% and neural networks with an accuracy of 69.27% [21]. The evaluation of logistic regression performance was also compared with other machine learning algorithms. It was proven that logistic regression was as good as the neural network algorithm and SVM in predicting chronic diseases such as kidney disease, heart disease, diabetes, and hypertension [22].

Based on several advantages of logistic regression, this paper used a logistic regression algorithm for early detection by predicting whether a person had diabetes or not based on the initial data provided. The research showed that the logistic regression algorithm performed well in accuracy, precision, recall, and F1-score. In addition, this study also conducted experiments using the grid search technique, which is an approach contained in the packet selection model from scikit-learn that can be used to improve the performance of the resulting model [23]-[24]. Grid search will automate setting hyperparameters, which will take a lot of time and resources [25]. The use of grid search in this study is proven to be able to improve the performance of the logistic regression model with better accuracy values.

II. Methodology

A logistic regression algorithm was used to predict diabetes in 768 available data. Prediction using logistic regression requires steps shown in Fig. 1, starting from determining the dataset to evaluating and deploying the model. This logistic regression implementation used the sci-kit learn library from Python, which eased data manipulation, visualization, and analysis easier.

A. Dataset Determination

The dataset used in this study was taken from the National Institute of Diabetes and Digestive and Kidney Diseases as part of the Pima Indians Diabetes Database [26]. The dataset consisted of several medical predictor variables (independent) and one target variable (dependent), namely Target (outcome), as shown in Table I.

B. Load and Read the Data

Datasets in .csv format were loaded into independent variables. There were 768 patient data, all of whom were women aged 21 years and above, consisting of nine variables: eight independent variables, namely Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction/DPF, and Age; and one dependent variable, namely Target. The results of checking the dataset using data.head() showed that several variables had a value of 0, which indicated a missing value.

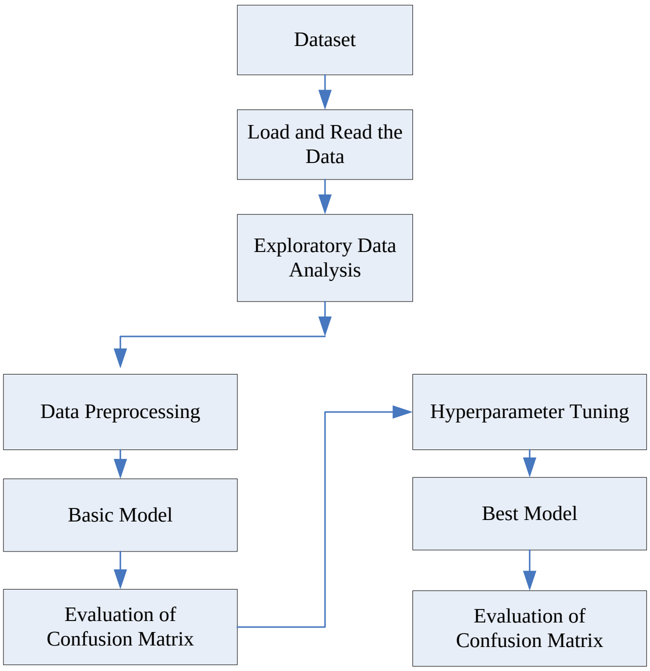


Fig. 1 Research design.

TABLE I
VARIABLES USED IN THE DIABETES PREDICTION

No.	Variable	Information
1	Pregnancies	Pregnancy: the number of times that the patient has been pregnant
2	Glucose	Two-hour plasma glucose concentration in an oral glucose tolerance test
3	BloodPressure	Blood pressure: diastolic blood pressure (mmHg)
4	SkinThickness	Triceps skinfold thickness (mm)
5	Insulin	Two-hour serum insulin (µlU/mL)
6	BMI	body mass index (kg/m^2)
7	DiabetesPedigreeFunction/ DPF	A function that assesses the likelihood of diabetes based on family history
8	Age	Age in this year
9	Outcome/Target	Result: class variable (0 if non-diabetic, 1 if diabetic)

C. Data Exploration Analysis

Data exploration analysis aims to analyze the dataset used to summarize the main characteristics of the dataset using the help of statistics and presenting it through visual techniques. At this stage, the data were checked before the model was built, so maximum insight was obtained from the owned dataset.

TABLE II
CONFUSION MATRIX FOR BINARY CLASSIFICATION

Actual Class	**Prediction Class**	
	Positive	***Negative***
Positive	tp	fp
Negative	fn	tn

D. Data Preprocessing

At this stage, checking was carried out on missing data values since the dataset might contain incomplete data. The missing data values were replaced with the median value of each variable so that each data in the dataset variable had an absolute value. At this stage, imbalanced data were checked. Unbalanced data was handled using a synthetic minority over-sampling technique (SMOTE). This technique increases the number of minority classes by synthesizing data samples while maintaining the number of majority classes.

E. Building a Model Using Logistic Regression Algorithm

Logistic regression models the relationship between categorical and covariate response variables. Specifically, there is a linear combination of the independent variables with the log probability of an event's probability. Logistic regression is a linear model that is more suitable for problems classification than its use for regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt), or log-linear classifier. In logit, the probabilities describing the possible outcomes of a single experiment are modeled using the logistic function.

Logistic regression models can be binary, one-vs-rest, or multinomial logistic regression with l_1, l_2 or elastic-net regulation [27]. Binary logistic regression estimates the probability of the availability of a binary variable characteristic, given the covariate value. For example, Y is a binary response variable with $Y_i = 1$ if the character is available, and $Y_i = 0$ if the character is unavailable and data $[Y_1, Y_2, \dots, Y_n]$ are independent. The value of π_i can be used to be a successful probability of a logistic regression. In addition, $x = (x_1, x_2, \dots, x_p)$ value is also considered a set of variables that can be discrete, continuous, or a combination of both. The π_i logistic function is given by (1) and (2).

$$logit\ (\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \quad (1)$$

with

$$\pi_i = \frac{exp\ (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}}{1 + exp\ (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}} = \frac{exp\ (x_i'\beta)}{1 + exp\ (x_i'\beta)} = \Lambda(x_i'\beta). \quad (2)$$

In this equation, π_i represents the probability that the sample falls under a particular category of the dichotomous response variable, commonly referred to as the probability of success. It is obvious that $0 \leq \pi_i \leq 1. \Lambda\ (.)$ is a logistic cumulative distribution function (CDF) with $\lambda(z) = \frac{e^z}{(1+e^z)} = 1/(1+e^z)$ and β^s represents parameters' vector to be estimated. The equation $\frac{\pi_i}{1-\pi_i}$ is called the odds ratio or relative risk [28].

This study used a binary logistic regression algorithm since the output was a value of 0 and 1, which was used to detect whether a person has diabetes. An output value of 0 suggests that a person does not have diabetes and vice versa; an output of 1 indicates that the person has diabetes.

F. Model Evaluation Using Confusion Matrix

The evaluations used to measure the performance of the algorithm/model were accuracy, precision, recall, and the F1-score in the form of a confusion matrix that other researchers have widely used. The confusion matrix is a table that compares the number of correct and incorrect predictions contained in each class; thus, it provides insight into the model's errors. Table II is a confusion matrix with a size of 2 × 2, which is used to present the actual class and the predicted class. Several measurements in the field of information retrieval and machine learning have been identified based on the classification of the confusion matrix as contained in (3) to (6).

$$Accuracy = \frac{tp+tn}{tp+fp+fn+tn} \quad (3)$$

$$Precision = \frac{tp}{tp+fp} \quad (4)$$

$$Recall/Sensitivity = \frac{tp}{tp+fn} \quad (5)$$

$$F1 - score = \frac{2\ (recall.precision)}{(recall+precision)} \quad (6)$$

In (3) to (6), *tp* (true positive) is the number of patients who are predicted to have diabetes and indeed have diabetes; *tn* (true negative) is a person who is expected to be non-diabetes and is indeed non-diabetes; *fp* (false positive) is a person who is predicted to be non-diabetes, but the person has diabetes; and *fn* (false negative) is a person who is predicted to have diabetes, but the person does not have diabetes.

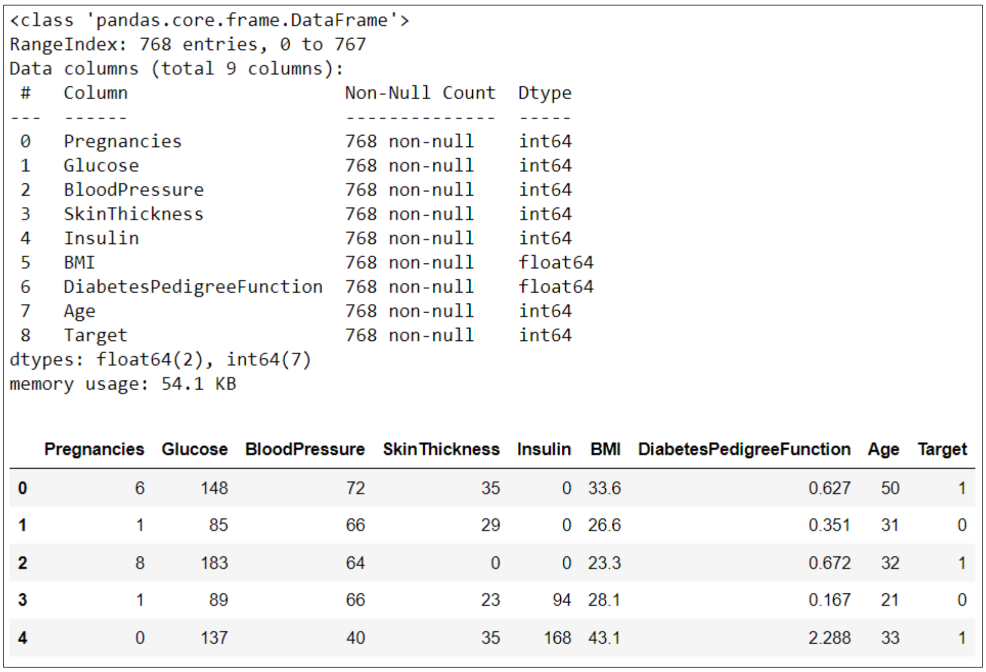
G. Hyperparameter Tuning

Hyperparameter tuning is used to select the optimal set of hyperparameters to improve model performance. Hyperparameters are in the form of model arguments whose values are set before the learning process starts. Hyperparameter tuning is the key to the success of a model/algorithm. In this paper, a grid search technique combining input values in hyperparameters was used. The grid search technique would search for all possible combinations and choose the best combination based on the highest cross-validation value. Two hyperparameters were applied in this study, namely penalty and C value. Penalty used the l1 and l2 regulations (the default value was l2), while the C value was the inverse of the regularization strength.

III. RESULTS AND DISCUSSION

A. Loading and Reading Data

Fig. 2 shows the results of loading and reading the dataset of 768 patient data. There were seven integer data types, namely



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Target                    768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Target
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. 2 General information of dataset.

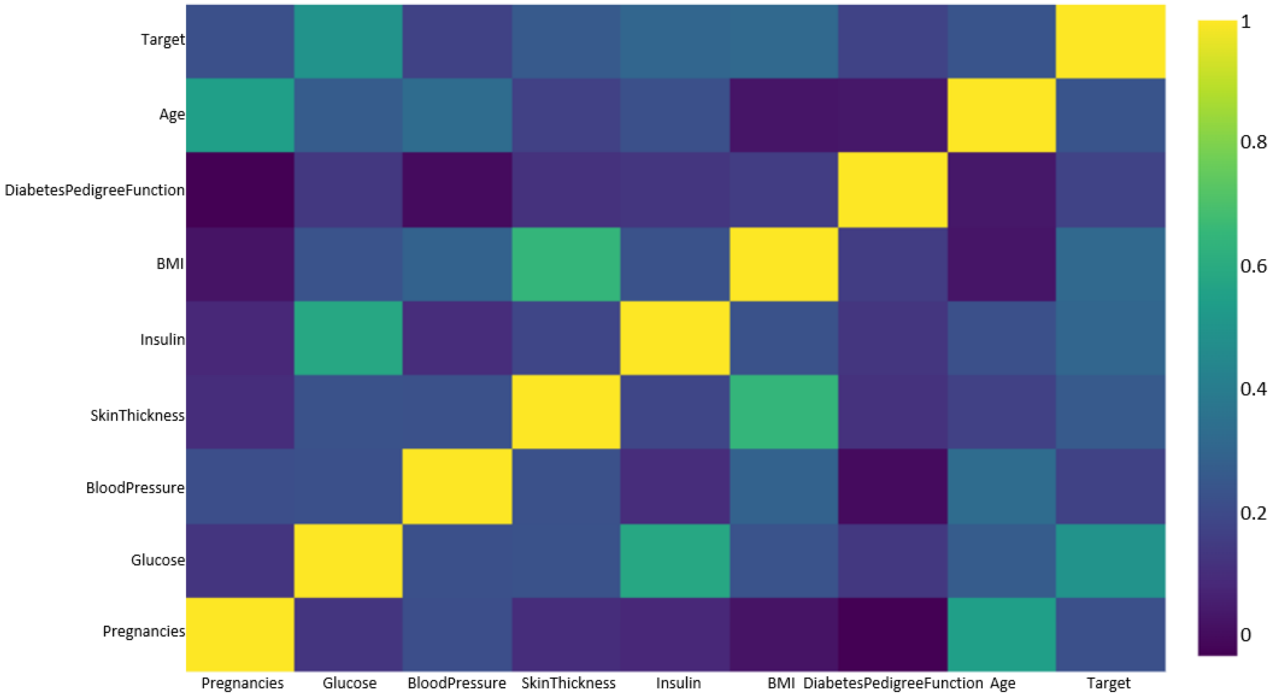


Fig. 3 Correlation matrix.

Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, Age, and Target; and two float data types, namely BMI and DiabetesPedigreeFunction. The data consisted of nine variables, namely eight independent variables and one dependent variable.

The figure also shows that there are empty data values for several variables. The top five data show that the Pregnancies and Insulin variables have empty data values (0). The empty data values would be replaced with the median value of each variable to facilitate the data manipulation process.

B. Exploratory Data Analysis

Understanding the existing dataset was essential before carrying out the following process. Fig. 3 is a correlation matrix that displays the correlation coefficient between a set of variables. Each independent variable (*Xi*) in the table is correlated with the other values in the table (*Xj*). The correlation matrix shows the variable pairs with the highest correlation. Insulin, Glucose, BMI, Pregnancies, and Age has a reasonably strong correlation with the Target variables.

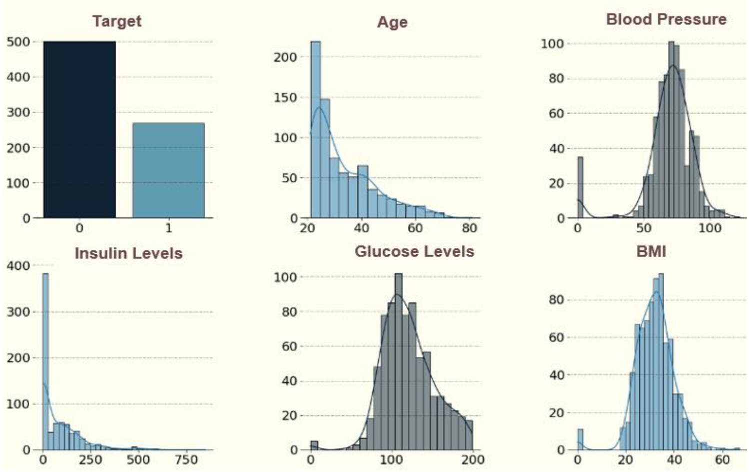
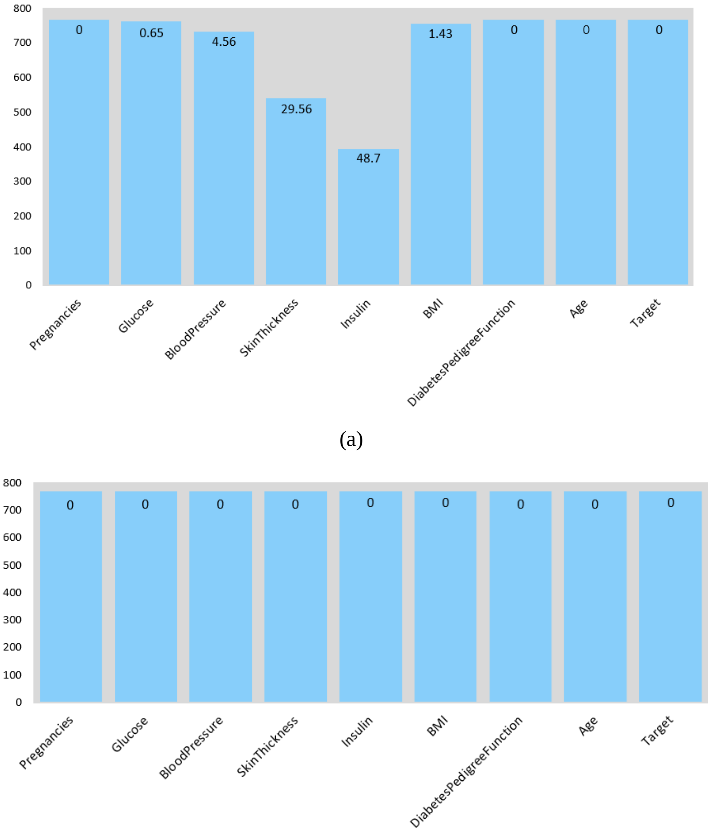


Fig. 4 Plot distribution for each variable.



(a)

(b)

Fig. 5 Amount of missing data in the dataset, (a) before normalization, (b) after normalization.

The distribution of plots for each attribute/variable is shown in Fig. 4. It can be seen that the Age and Insulin columns are very skewed to the right. For this reason, a normalization process was needed before being used for the modeling process.

The dataset indicated that many people were between 20-40 years old; most people had blood pressure between 50-100 mmHg and had insulin of 0. Most people also had glucose levels between 140 mg/dL to 199 mg/dL and were considered prediabetes patients. BMI values ranged from 20 to 50; meanwhile, healthy adults should have a BMI between 18.5-24.9. This dataset suggests that many people were overweight or obese.

C. *Preprocessing Data*

1) Missing Value: The results of checking the dataset showed several missing data values, as shown in Fig. 5(a). The insulin variable is a variable that has the most missing data values, which is 374 data or 48.7%, followed by SkinThickness with 227 data or 29.56%, Blood Pressure with 35 data or 4.56%, BMI with 11 data or 1.43%, and Glucose with 5 data or 0.65%.

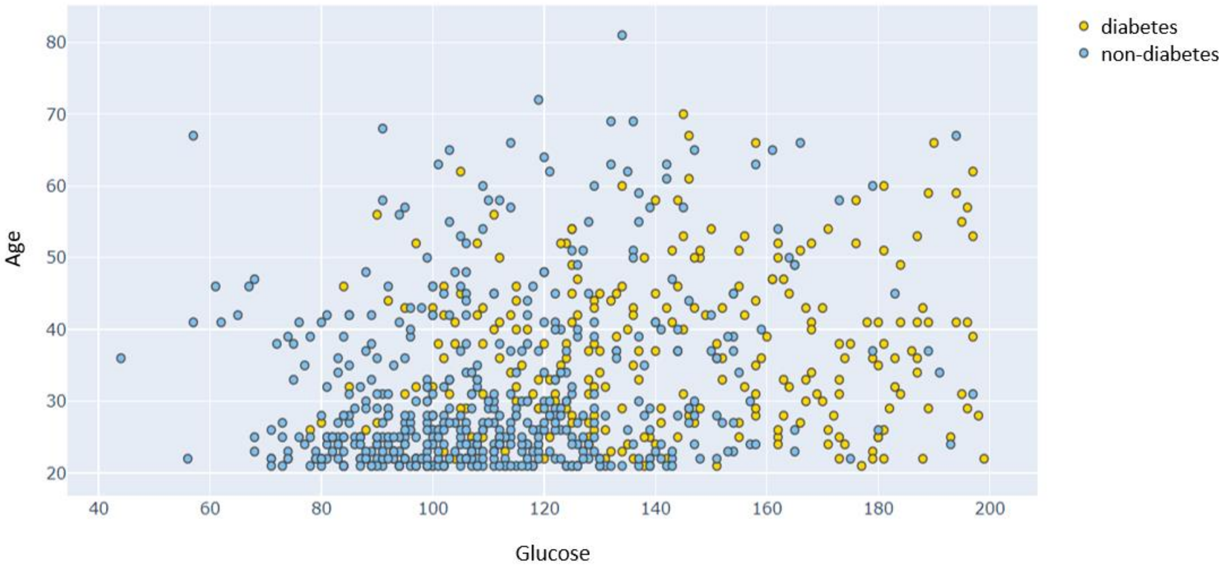


Fig. 6 Scatter plot of the relationship between age and glucose variables.

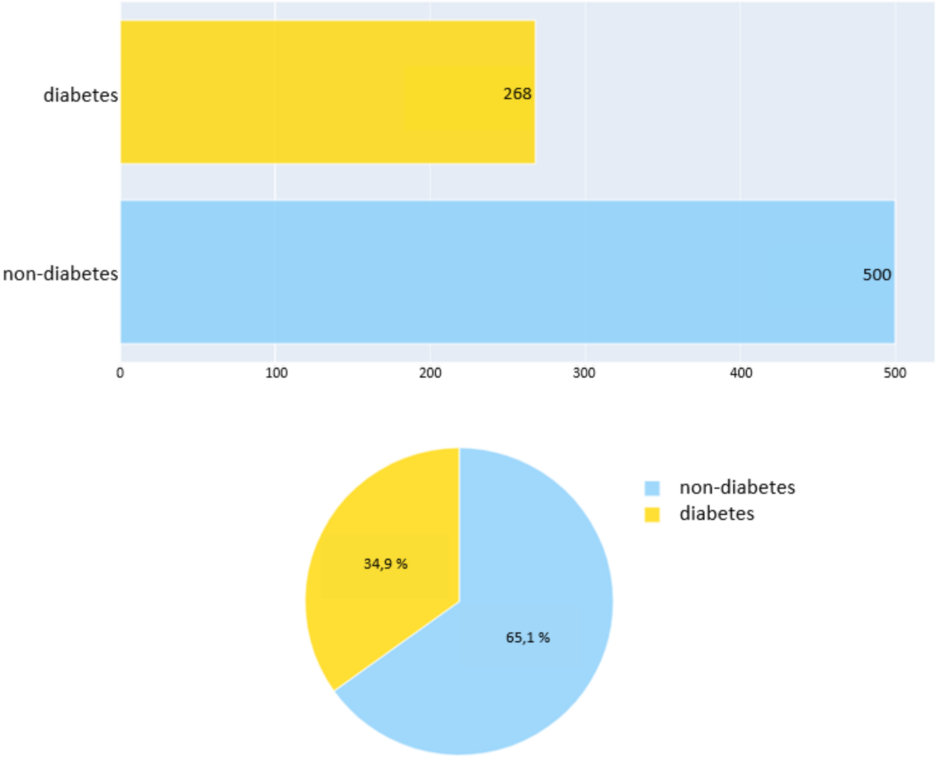
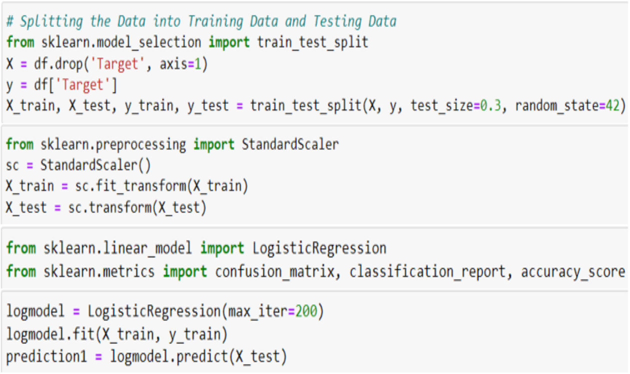


Fig. 7 Unbalanced number of target datasets.

Other variables have complete data. All missing data values were replaced with the median value of each variable so that there were no more empty variable values, as shown in Fig. 5(b). Handling missing data values was used to facilitate the process of data manipulation.

After all variables/attributes were completed, a new feature was created. This feature was a combination of several variables to see whether a person had diabetes, based on the relationship between one variable with another. Fig. 6 shows one of the new attributes: the relationship between the Age variable and the Glucose variable, namely non-diabetes people under 30 years with glucose levels below 120 mg/dL. Healthy people were in a concentrated area at age ≤ 30, and glucose was ≤ 120 mg/dL.

2) Imbalanced Dataset: The graph resulting from checking the imbalanced data is shown in Fig. 7. The number of diabetic patients is 268 people (34.9%), while the number of non-diabetes patients is 500 (65.1%). SMOTE technique was used to overcome the imbalanced data. SMOTE is an oversampling



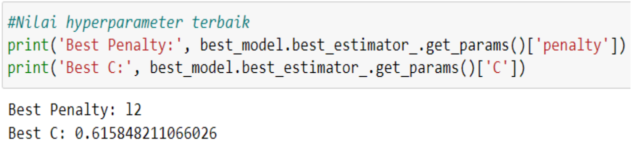
```
# Splitting the Data into Training Data and Testing Data
from sklearn.model_selection import train_test_split
X = df.drop('Target', axis=1)
y = df['Target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score

logmodel = LogisticRegression(max_iter=200)
logmodel.fit(X_train, y_train)
prediction1 = logmodel.predict(X_test)
```

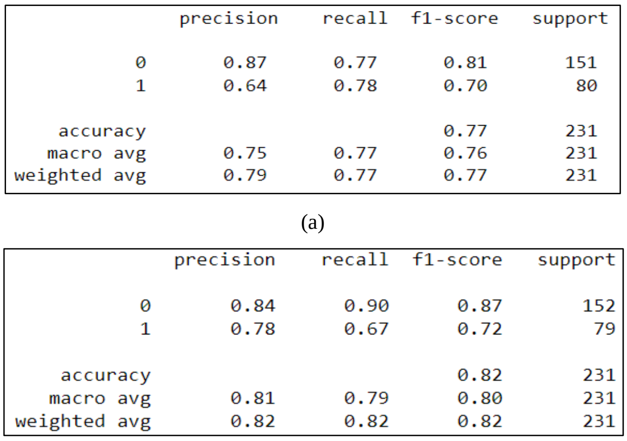
Fig. 8 Python program snippet for building models.



```
#Nilai hyperparameter terbaik
print('Best Penalty:', best_model.best_estimator_.get_params()['penalty'])
print('Best C:', best_model.best_estimator_.get_params()['C'])

Best Penalty: l2
Best C: 0.615848211066026
```

Fig. 9 Selection of the best parameters.



```
              precision    recall  f1-score   support

           0       0.87      0.77      0.81       151
           1       0.64      0.78      0.70        80

    accuracy                           0.77       231
   macro avg       0.75      0.77      0.76       231
weighted avg       0.79      0.77      0.77       231
```

(a)

```
              precision    recall  f1-score   support

           0       0.84      0.90      0.87       152
           1       0.78      0.67      0.72        79

    accuracy                           0.82       231
   macro avg       0.81      0.79      0.80       231
weighted avg       0.82      0.82      0.82       231
```

(b)

Fig. 10 Model performance, (a) before hyperparameter tuning, (b) after hyperparameter tuning.

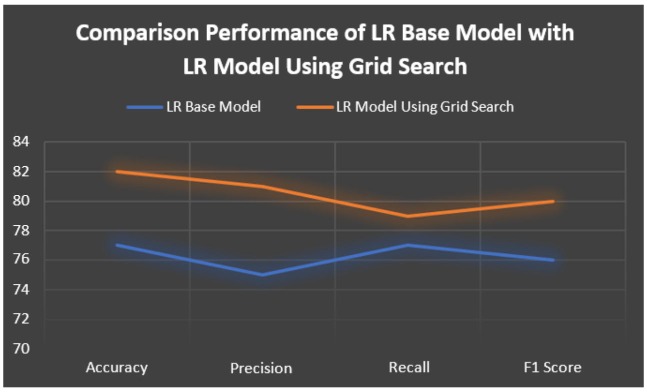


Fig. 11 Comparison of the basic model performance using the tuning hyperparameter model.

TABLE III
CLASSIFICATION RESULT OF CONFUSION MATRIX

Actual Class	**Prediction Class**	
	0 (Non-diabetes)	***1 (Diabetes)***
0 (Non-diabetes)	116	35
1 (Diabetes)	18	62

TABLE IV
PERFORMANCE MEASUREMENT

Precision (%)		**Recall (%)**		**F1-Score (%)**	
0	1	0	1	0	1
87	64	77	78	81	70
Average: 75		Average: 77		Average: 76	

technique in which a synthetic sample is generated for the minority class to help overcome the overfitting problem found in random oversampling.

After the preprocessing stage, the next step was modeling using logistic regression. The data were first divided into two parts, namely training and test data. The comparison of training data and test data was 70:30. This process was done using the scikit-learn library from Python. A snippet of the script is shown in Fig. 8, which shows the distribution of the dataset for training data and test data, followed by modeling using the logistic regression algorithm.

D. Model Evaluation

Model evaluation was carried out after the model was formed. Table III shows the evaluation results of the model in confusion matrix format. One hundred seventy-eight data were in the proper classification (true positive and true negative), consisting of 116 data predicted to be diabetic and, in fact, had diabetes and 62 people were expected to be non-diabetes and, in fact, were non-diabetes. A total of 53 other data were false positive and false negative, namely, 35 non-diabetes people were predicted to have diabetes, and eighteen people with diabetes were expected to be non-diabetes.

Table IV shows model performance measurement based on accuracy, precision, recall, and F1-score values. The precision value obtained for the non-diabetes class was 87%, and for diabetes was 64%, with an average precision value of 75%. The recall value for non-diabetes was 77% and for diabetes was 78%, with an average recall value of 77%. The F1-score showed a number that was not much different from precision and recall. The F1-score for non-diabetes was 81% and for diabetes was 70%, with an average of 76%. The accuracy value was 77%. Based on the calculation of the importance of the four evaluated variables, the model was categorized as having a pretty good performance.

E. Hyperparameter Tuning Using Grid Search Technique

To improve the performance of the model, the parameters used were tuning. Fig. 9 is a snippet of a Python script used to select the best parameters through a grid search technique. L2 was chosen as the best penalty regulation, with the most optimal C value being 0.6158.

F. Evaluation of Model Performance Before and After Hyperparameter Tuning

A comparison of model performance before and after applying the grid search technique was analyzed to measure the impact of using hyperparameter tuning. Fig. 10(a) shows the model's performance before applying the hyperparameter tuning. Fig. 10(b) shows the model's performance after using hyperparameter tuning. From the two images, it is clear that there is performance improvement on all confusion matrix values, starting from the importance of accuracy, precision, recall to the F1-score. In the basic model, the average accuracy value was 77%, precision was 75%, recall was 77%, and F1-score was 76%. On the other hand, the accuracy value increased to 82% in the upgraded model, precision was 81%, recall was 79%, and the F1-score was 80%.

Fig. 11 shows that the logistic regression model using grid search is better than the basic model for all confusion matrix values (accuracy, precision, recall, and F1-score). This experiment proves that the linear regression algorithm can be used for predictions in the clinical or health field with a good level of accuracy. The performance of the logistic regression algorithm was further improved when it was inserted with the grid search technique, which can increase all the values of the confusion matrix, thus increasing the overall performance of the logistic regression algorithm.

IV. Conclusion

This study successfully implemented the logistic regression algorithm in predicting diabetes with a good accuracy. Understanding the data was done through data exploration and research to analyze pairs of variables that had a reasonably strong correlation to the determination of the target value through visualization techniques in the form of distributions and scatter plots. The performance of the basic model of the logistic regression algorithm was improved using the grid search technique. Evaluation of the model using the confusion matrix showed an increase in model performance after implementing hyperparameter tuning. Hence, the experimental results of this study prove that the logistic regression algorithm with the grid search technique is one of the most efficient algorithms in building prediction models. Future research can use deep learning algorithms on larger datasets, including combining logistic regression algorithms with other classification algorithms, such as random forests, support vector machines, k-nearest neighbor with ensemble techniques.

Conflict of Interest

The authors whose names are listed in the article entitled "Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm" state that there is no conflict of interest.

Author Contribution

Conceptualization, Erlin and Yulvia Nora; research methodology, Erlin, Junadhi, Laili Suryati; software, Erlin and Nova Agustina; validation, Erlin, Yulvia Nora, Junadhi, Laili Suryati, Nova Agustina; formal analysis, Erlin; source, Erlin; writing—preparation of the original draft, Erlin; writing—review and editing, Erlin.

References

[1] American Diabetes Association (2020) "Diabetes Overview The path to understanding diabetes starts here." [Online], https://www.diabetes.org/diabetes, access date: 19-Nov-2021.

[2] World Health Organization (2020) "Diabetes," [Online], https://www.who.int/health-topics/diabetes#tab=tab_1, access date: 19-Nov-2021.

[3] International Diabetes Federation (2020 "Diabetes facts & figures," [Online], https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html, access date: 19-Nov-2021.

[4] J. Elflein (2019) "Number of people with diabetes, by country 2019," [Online], https://www.statista.com/statistics/281082/countries-with-highest-number-of-diabetics/, access date: 6-Dec-2021.

[5] H. Nurhayati-Wolff (2020) "Projected number of people with diabetes Indonesia 2017-2024," [Online], https://www.statista.com/statistics/1052625/indonesia-diabetes-projection/, access date: 6-Dec-2021.

[6] B. Hardhana, F. Sibuea, and W. Widiantini, Eds., *Profil Kesehatan Indonesia Tahun 2019*, Jakarta, Indonesia: Kementerian Kesehatan Republik Indonesia, 2020.

[7] Badan Litbangkes Kemenkes RI (2018), "Hasil Utama Riskesdas 2018," [Online], https://drive.google.com/file/d/1MRXC4lMDera5949ezbbHj7UCUj5_EQmY/view, access date: 6-Dec-2021.

[8] Diabetes UK (2018) "Diabetes the Basics," [Online], https://www.diabetes.org.uk/diabetes-the-basics, access date: 8-Dec-2021.

[9] M.C. Riddle, Ed., "Standards of Medical Care in Diabetes—2022," *Diabetes Care*, Vol. 45, Supp. 1, pp. 125-143, Jan. 2022.

[10] D.J. Reddy, *et al.*, "Predictive Machine Learning Model for Early Detection and Analysis of Diabetes," *Mater. Today: Proc.*, akan diterbitkan.

[11] L.V.R. Kumari, *et al.*, "Machine Learning based Diabetes Detection," *Proc. 6th Int. Conf. Commun. Electron. Syst. (ICCES 2021)*, 2021, pp. 1-5.

[12] N. Abdulhadi and A. Al-Mousa, "Diabetes Detection Using Machine Learning Classification Methods," *Proc. 2021 Int. Conf. Inf. Technol. ICIT 2021*, 2021, pp. 350–354.

[13] R. Krishnamoorthi, *et al.*, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques," *J. Healthc. Eng.*, Vol. 2022, pp. 1–10, 2022.

[14] U.M. Butt, *et al.*, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *J. Healthc. Eng.*, Vol. 2021, pp. 1–17, 2021.

[15] P. Arsi and O. Somantri, "Deteksi Dini Penyakit Diabetes Menggunakan Algoritma Neural Network Berbasiskan Algoritma Genetika," *J. Inform. J. Pengemb. IT*, Vol. 3, No. 3, pp. 290–294, 2018.

[16] A.B. Wibisono and A. Fahrurozi, "Perbandingan Algoritma Klasifikasi dalam Pengklasifikasian Data Penyakit Jantung Koroner," *J. Ilm. Teknol. dan Rekayasa*, Vol. 24, No. 3, pp. 161–170, 2019.

[17] J.J. Khanam and S.Y. Foo, "A Comparison of Machine Learning Algorithms for Diabetes Prediction," *ICT Express*, Vol. 7, No. 4, pp. 432–439, 2021.

[18] T. Ciu and R.S. Oetama, "Logistic Regression Prediction Model for Cardiovascular Disease," *IJNMT (Int. J. New Media Technol.)*, Vol. 7, No. 1, pp. 33–38, 2020.

[19] R. Thammasudjarit, *et al.*, "Comparison of Machine Learning with Logistic Regression for Prediction of Chronic Kidney Disease in the Thai Adult Population," *Ramathibodi Med. J.*, Vol. 44, No. 4, pp. 1–12, 2021.

[20] N. Varshney and A. Sharma, "Identification and Prediction of Liver Disease Using Logistic Regression," *Eur. J. Mol. Clin. Med.*, Vol. 7, No. 4, pp. 106–110, 2020.

[21] D.Y. Utami, E. Nurlelah, and F.N. Hasan, "Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to Find the Highest Accuracy in Diabetes," *J. Inform. Telecommun. Eng.*, Vol. 5, No. 1, pp. 152–159, 2021.

[22] S. Nusinovici, *et al.*, "Logistic Regression was As Good As Machine Learning for Predicting Major Chronic Diseases," *J. Clin. Epidemiol.*, Vol. 122, pp. 56–69, 2020.

[23] S. Mezzatesta, *et al.*, "A Machine Learning-based Approach for

Predicting the Outbreak of Cardiovascular Diseases in Patients on Dialysis," *Comput. Methods, Programs Biomed.*, Vol. 177, pp. 9–15, 2019.

[24] S. Ambesange, *et al.*, "Multiple Heart Diseases Prediction Using Logistic Regression with Ensemble and Hyper Parameter Tuning Techniques," *Proc. World Conf. Smart Trends Syst. Secur. Sustain. WS4 2020*, 2020, pp. 827–832.

[25] L. Lama, *et al.*, "Machine Learning for Prediction of Diabetes Risk in Middle-aged Swedish People," *Heliyon*, Vol. 7, No. 7, pp. 1–6, 2021.

[26] (2016) "Pima Indians Diabetes Database," [Online], https://www.kaggle.com/uciml/pima-indians-diabetes-database, access date: 23-Oct-2021.

[27] F. Pedregosa, *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, Vol. 12, No. 85, pp. 2825–2830, 2011.

[28] R.D. Joshi and C.K. Dhakal, "Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches," *Int. J. Environ. Res. Public Health*, Vol. 18, No. 14, pp. 1-17, 2021.