DECODE: JURNAL PENDIDIKAN TEKNOLOGI INFORMASI
P-ISSN: 2775-2984 | E-ISSN: 2775-1813
Vol. 5, No 1, Maret 2025, Hal: 54-64
Doi: http://dx.doi.org/10.51454/decode.v5i1.1039
https://journal.umkendari.ac.id/index.php/decode

Diabetes Mellitus Disease Prediction Using Logistic Regression (LR) and Support Vector Machine (SVM) Methods

Akbar Febrian Dwi Hastono[1], Anik Vega Vitianingsih[1*], Pamudi Pamudi[1], Anastasia Lidya Maukar[2], Seftin Fitri Ana Wati[3]

[1]Informatic Engineering Study Program, Dr. Soetomo University, Indonesia.
[2]Department of Industrial Engineering, President University, Indonesia.
[3]Information System Department, Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia

Article Info

***Keywords*:**

Confusion Matrix;
Diabetes Mellitus;
Logistic Regression;
Prediction;
Support Vector Machine.

Kata Kunci*:*

Confusion Matrix;
Diabetes Mellitus;
Logistic Regression;
Prediksi;
Support Vector Machine.

History Article:

Submitted: January 18, 2025
Accepted: March 12, 2025
Published: March 14, 2025

Corresponding Author:
Anik Vega Vitianingsih
Email: vega@unitomo.ac.id

Abstract: Diabetes Mellitus (DM), also known as diabetes or sugar disease, marked by high blood sugar levels and poses a major health issue in Indonesia with the number of cases increasing every year. Often referred to as the silent killer, DM often goes unnoticed due to its subtle symptoms, increasing the risk of severe complications if not treated promptly. The lack of information or awareness about the early symptoms of DM, limited time and cost in conducting health checks, and limited access to health services are challenges in detecting DM disease early. To overcome this problem, the development of a prediction model is essential to prevent serious complications. This study aims to create a predictive model using LR and SVM methods based on parameters such as pregnancy, glucose levels, blood pressure, skin thickness, insulin, BMI, diabetes pedigree, age, and outcome. The dataset used is DM disease risk data collected by Kaggle from the National Institute of Diabetes and Disgetive and Kidney Disease (NIDDK). Based on the research results, the LR method shows a better level of accuracy compared to the SVM method. The accuracy of the model using the Logistic Regression method is 79.31% while the SVM method has an accuracy value of 77.24%, with a difference in accuracy of 2.07%. This research applies hyperparameter tuning with Grid Search to find the best combination of hyperparameter.

INTRODUCTION

Diabetes Mellitus (DM) is a chronic disease characterized by high blood sugar levels. It is included as a Non-Communicable Diseases (NCDs) that pose serious health risks and cause high levels of sugar in body fluids (Marwati & Fauzi, 2024). As stated by the International Diabetes Federation (IDF), number of individuals affected by Diabetes Mellitus (DM) has grown from 382 million in 2013 to over 550 million in 2023, and is expected to continue to grow every year, so real efforts are needed for prevention (Gunawan et al., 2020). DM disease can also be called a silent killer, because DM has the potential to damage the body slowly, so that if it does not immediately get proper treatment it can cause complications (Todkar, 2016). Some of the symptoms experienced by people with DM include frequent urination, easy thirst, easy hunger, significant weight loss, dry skin, slow wound healing, and visual impairment (Mardiana et al., 2020). Indonesia as reported by the IDF, the number of individuals diagnosed with DM has been rising steadily since 2020. In 2020 there were around 10,8 million people infected with DM which increased to 19,5 million people in 2021, this will certainly continue to increase every year and it is estimated that by 2045 it will reach 28,6 million people who will have DM (Aris, 2019).

Given the large number of DM cases, efforts to prevent and control this disease are considered quite important by implementing a healthy lifestyle. Symptoms of DM are often not realized, so people who have risk factors for DM and DM sufferers must take several medical precautions to prevent complications that result in premature death. This is because there are many cases of serious complications that arise for DM sufferers due to delays in diagnosis causes by several factors, including lack of information regarding the early symptoms of DM, limited time in conducting routine health checks, and financial limitations to obtain proper health services. Based on the findings of these conditions, one of the efforts to prevent serious complications due to DM is to predict DM. This prediction capability can be user to detect DM early based on the symptoms found as an effort to prevent serious complications due to DM (Kopitar et al., 2020). Thus, a predictive approach related to Diabetes Mellitus (DM) disease in this study is considered quite important.

Literature study on previous research, namely the results of research (Maulidah et al., 2021) using the Support Vector Machine (SVM) and Naive Bayes methods in predicting DM disease, getting an accuracy value in the SVM method of 78.04% while the Naïve Bayes method gets an accuracy value of 76.98%. Research (Cahyani et al., 2022) predicting the risk of DM disease using the Logistic Regression method obtained an accuracy value of 55% after normalization, while without normalization it was 43%. Research (Oktaviana et al., 2024) using the K-Nearest Neighbor (K-NN) method in predicting type 2 DM disease obtained an accuracy value of 88%, precision 83.54%, recall 87.5% and f1-score 85.36%. Based on the literature study, this research will develop a prediction model with Logistic Regression and SVM methods to predict DM disease. This study chose to use LR and SVM in DM disease prediction although previous studies have shown that K-NN has higher accuracy. One of the main reasons is that LR is easier to interpret as it is able to show the contribution of each variable in prediction (Christodoulou et al., 2019), while SVM has the advantage of handling high-dimensional data with optimal separation margins (Christodoulou et al., 2019) On the other hand, K-NN tends to be more prone to overfitting and requires longer computation time in the inference process (Lee et al., 2024). By considering the aspects of efficiency, stability, and potential optimization through proper preprocessing techniques, the use of LR and SVM is expected to produce a DM prediction model that is not only accurate but also more efficient and easy to understand. Furthermore, the literature study on previous research indicates that hyperparameter tuning has not been extensively explored to optimize LR and SVM performance, by incorporating hyperparameter tuning to determine the best parameter combinations.

This study aims to design a prediction model for early detection of DM disease comparing the performance of LR and SVM approaches. The Logistic Regression method is used to predict binary probabilities, this ensures the prediction results are in accordance with the discrete data in this study, where the predicted data only has two possibilities, namely 1 for positive DM and 0 for negative DM (Pratama et al., 2023). While the SVM approach is utilized in this research to maximize the margin

between classes through hyperplane boundaries, so as to increase accuracy in DM disease prediction (Hovi et al., 2022). However, this study goes beyond simply comparing the two algorithms by integrating hyperparameter tuning using Grid Search to find hyperparameter combinations of the two models. The novelty of this study lies in the detailed performance evaluation using accuracy, precision, recall, and F1-score, as well as in-depth correlation analysis between the predictor variables in Figure 2. The results, presented in Table 7, show the impact of hyper parameter tuning in finding hyperparameter combinations, making this approach more effective in DM prediction.

METHOD

The development of prediction models using the Logistic Regression and SVM methods involves a series of stages described in the form of a flow chart. This flowchart provides a comprehensive explanation of the DM disease prediction process using both methods. The picture shown represents a systematic and continuous flow in each stage. The stages in this study include all steps starting from data preprocessing, processing, postprocessing, to model evaluation using the confusion matrix shown in Figure 1.

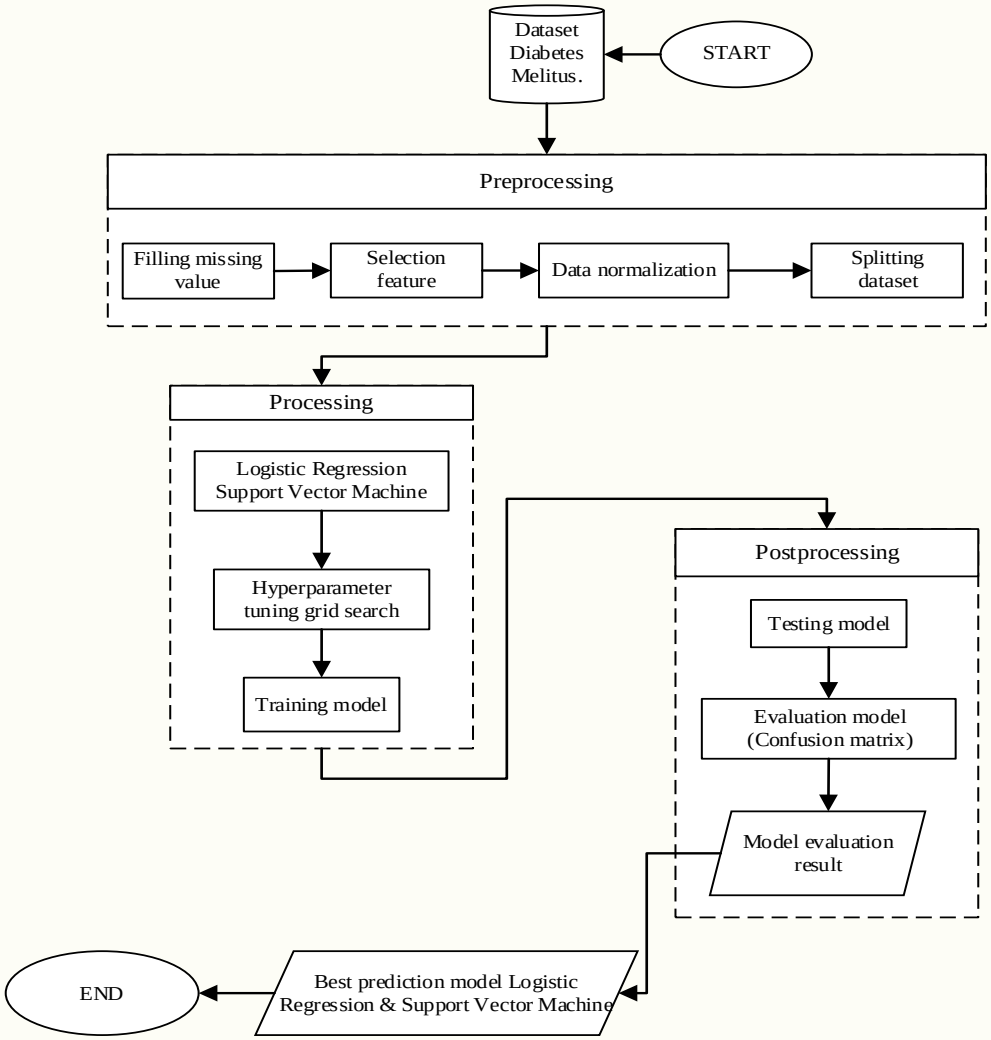

Figure 1. Research Flowchart *Logistic Regression* & SVM

The research utiliez a dataset related to DM data obtained from kaggle (*Diabetes Dataset*, n.d.) and the parameters used refer to independent variables including pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index (bmi), diabetes pedigree, age and the dependent variable outcome. The discussion of methods is grouped based on the techniques used in the development of prediction

models, namely Logistic Regression and SVM. The Logistic Regression model was developed using a sigmoid function to predict binary probabilities (Tangkere, 2024), whereas SVM utilizes the maximum margin to separate data classes through a hyperplane (Hovi et al., 2022). The stages in both methods use a similar approach, as described in Figure 1. These stages begin with preprocessing, which includes filling in missing data, feature selection to determine attributes and targets, normalizing data, then the normalized data will be split into training data and testing dataset. Next, the model training progress (processing), and ends with model evaluation (postprocessing) using confusion matrix as well as matrices such as accuracy, precision, recall, and f1-score.

A. Prediction Model

1) Logistic Regeression method: A machine learning method used to analyze the relationship between dependent variables and independent variables, especially when the independent variables are binary categorical, with predicted outcomes of 0 and 1 (Suprihati, 2021). When it comes to predicting DM illness, LR is used to understand how independent variables behaving as predictors affect the likelihood of a person being diagnosed with DM (Risiko et al., 2023). Through coefficient estimation, Logistic Regression helps quantify the extent to which each predictor contributes to the probability of a person being categorized as having or not having DM (Saepudin et al., n.d.). *Logistic Regression serves to determine a number of linear or logit functions* based on Equation (1), where the variable logit(πi) is the feature probability, β0 is a constant or intercept, *β1, β2...βp* are regression coefficients, *x1, x2...xp* are predictor variables.

$$\boldsymbol{logit}(\boldsymbol{\pi_i}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \cdots + \boldsymbol{\beta_p X_p} \quad (1)$$

2) Support Vector Machine (SVM) method: SVM is a machine learning algorithm that uses a hyperplane as the optimal dividing line to classify data (Firdaus et al., 2023). This algorithm focuses on optimizing the margin, which is the perpendicular distance between the data points of the two classes closest to the hyperplane, this optimization process is carried out through a minimization approach involving Lagrange techniques (Amelia et al., 2022). The data classification process in SVM is done by utilizing Equation (2). In this context, the input x and the predicted class, denoted by the function f(x), are influenced by the weight vector w, which defines the hyperplane's orientation within the feature space. Meanwhile, the input data is represented by the feature vector x, and the bias term b adjust the hyperplane's position away from its original placement.

$$\boldsymbol{f(x)} = \boldsymbol{w.x} + \boldsymbol{b} \quad (2)$$

B. Hyperparameter Tuning Grid Search

The hyperparameter optimization process in this study was carried out using grid search, a simple method for hyperparameter tuning that tests each parameter combination systematically. The advantage is that each data set is evaluated with consistent parameters, allowing the search for which combination produces the best value one by one (Desiani et al., 2022). The scheme and parameters used in this study in the grid search are displayed in Table 1 and Table 2 for each model.

Table 1. Hyperparameter Logistic Regression Scheme

Parameter	Value
C	'0.001', '0.01', '0.1', '1', '10'
Fit_intercept	'True', 'false'
Solver	'lbfgs', 'liblinear', 'newton-cg', 'sag'

Table 2. Hyperparameter SVM Scheme

Parameter	Value
C	'0.001', '0.01', '0.1', '1', '10'
Kernel	'linear', 'rbf', 'poly', 'sigmoid'
Gamma	'scale', 'auto'

C. Evaluation Model

Confusion Matrix is a representation of the classification results that describe how data is predicted correctly or incorrectly (Damayanti et al., 2024). This study contains the classification results of diabetes and non-diabetes. This confusion matrix test is based on Table 3.

Table 3. Elements of a Confusion Matrix

Class		Actual Value	
		Positive	**Negative**
Diabetes	**Positif**	True Positive (TP)	False Negative (FN)
Non-Diabetes	**Negatif**	False Positive (FP)	True Negative (TN)

Based on Table 3, a prediction is considered correct and classified as positive if it lies in (TP). Conversely, if the prediction is incorrect but still classified as positive, then the result is located in the (FP) column. Prediction errors classified as negative will be recorded in the (FN) column, while correct predictions classified as negative will be recorded in the (TN) column (Damayanti et al., 2024). The results of the confusion matrix will be utilized to calculate evaluation metrics such as accuracy, precision, recall, and f1-score, these are used to evaluate how well a model performs overall (Teknik Elektro dan Komputasi et al., 2022), shown in Equation (3) (4) (5) and (6).

$$Accuracy = \frac{TP+TN}{Total\ Data} \tag{3}$$

$$Precission = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$F1 - Score = 2 \times \frac{precission \times recall}{precission + recall} \tag{6}$$

RESULT AND DISCUSSIONS

The dataset utilized in this study is Diabetes Mellitus (DM) disease data obtained from kaggle collected by the National Institute of Diabetes and Kidney Diseases (NIDDK), comprising 768 data samples. Where the data samples used are women of Pima Indian descent with a minimum age of 21 years, which consists of 8 attributes and 1 target including parameters such as pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree, age, with the outcome serving as the target variable. The dataset description is shown in Table 4, and the sample dataset is shown in Table 4.

Table 4. Description Related Attributes

Attribute	Description
Pregnancies *(Pr)*	Total number of pregnancies ever experienced
Glucose *(Gc)*	Plasma glucose levels were measured two hours after performing the oral glucose tolerance test.
Blood Pressure *(BP)*	Diastolik blood pressure (mmHg).
Skin Thickness *(ST)*	Measurment of skinfold thickness at the triceps.
Insulin *(In)*	2-hour serum insulin levels (mu U/ml)
BMI (Body Mass Index)	Body mass index is calculated by dividing body weight (kg) by the square of the height (m)

Attribute	Description
Diabetes Pedigree Function *(DPF)*	Hereditary aspects of diabetes pedigree
Age	Age
Outcome	The outcome variable represents the class, where 0 indicates a negative outome and 1 signifies a positive outcome for DM.

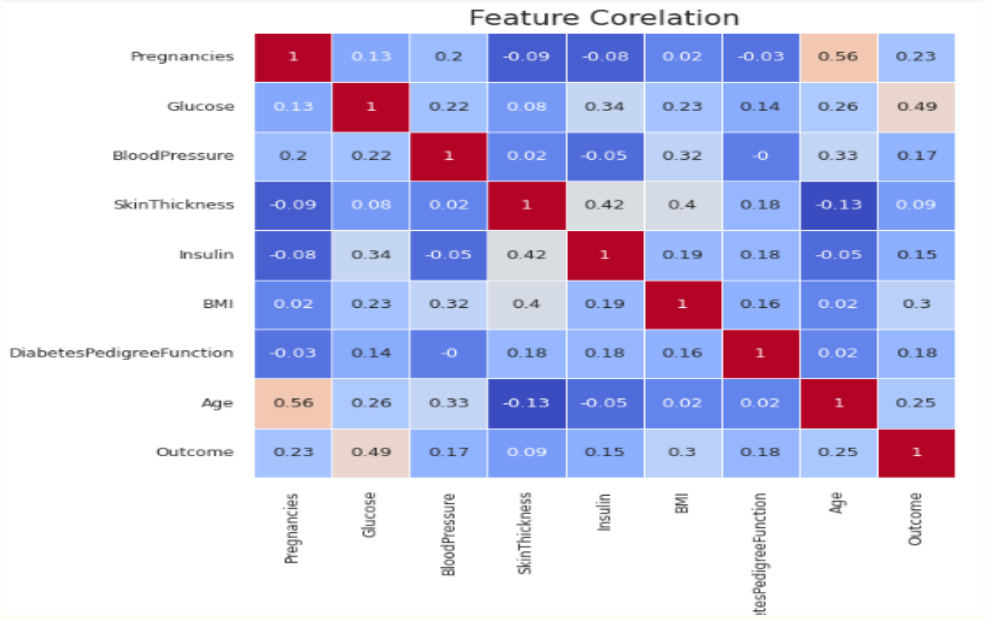

Figure 2. Heatmap Correlation Between Features on Dataset

Figure 2 displays a heatmap showing the correlation between the features of the dataset, highlighting the linear relationships between them, the value ranges from -1 to 1, where the darker color indicates a stronger relationship.

1) The correlation value between the pregnancies features and the outcome of 0.23 interprets the relationship between pregnancies and DM diagnoses is relatively weak, which means that the higher the value of pregnancy does not significantly affect DM diagnoses.

2) The correlation value between the glucose feature and the outcome of 0.49 indicates that the relationship between glucose content in the blood and DM diagnosis is relatively strong, so that the higher the glucose indicates a positive diagnosis of DM. From several correlation features, the glucose feature is the main predictor of DM disease.

3) The correlation value between the blood pressure feature and the outcome of 0.17 interprets the relationship between blood pressure and DM diagnosis is relatively quite weak, which means that the higher the blood pressure does not significantly affect the diagnosis of DM.

4) The correlation value between the skin thickness feature and the outcome of 0.09 means that the relationship between skin thickness and DM diagnosis is relatively weak, which means that the higher the value of skin thickness, the less significant the DM diagnosis.

5) The correlation value between the insulin feature and the outcome of 0.15 indicates that the relationship between insulin levels and DM diagnosis is relatively weak, which means that the higher the value of insulin levels does not significantly affect the diagnosis of DM.

6) The correlation value between the BMI feature and the outcome of 0.3 interprets that the relationship between body mass index and DM diagnosis is relatively weak, which means that the higher the value of body weight or height does not really affect the diagnosis of DM.

7) The correlation value between the diabetes pedigree function feature and the outcome of 0.18 indicates that the relationship between diabetes lineage and DM diagnosis is relatively weak, meaning that the higher the value is not too significant to affect DM diagnosis.

8) The correlation value between the age feature and the outcome of 0.25 interprets that the relationship between age and DM diagnosis is relatively quite weak, which means that and person's age, even at higher values, does not significantly affect the diagnosis of DM.

Table 5. DM Disease Dataset Sample

Pr	Gc	BP	ST	In	BMI	DPF	Age	Outcome
6	148	72	35	0	33,6	0,627	50	1
1	85	66	29	0	26,6	0,351	31	0
8	183	64	0	0	23,3	0,672	32	1
1	89	66	23	94	28,1	0,167	21	0
0	137	40	35	168	43,1	2,288	33	1
5	116	74	0	0	25,6	0,201	30	0
3	78	50	32	88	31	0,248	26	1

Before being processed using the *Logistic Regression* (LR) and SVM methods with hyperparameter tuning techniques through grid search, the data first goes through a preprocessing stage to ensure that each attribute in the dataset does not have variables that interfere with the analysis process. In this research, the preprocessing stage performed is as shown in Figure 1.

A. Preprocessing

The preprocessing stage starts by ensuring that there are no empty values in each parameter, where the test results show that there are no empty values in all parameters. After that, perform feature selection to determine the parameters used as attributes (independent variables) and targets (dependent variables). The attributes used in this study include Pr (pregnancy), Gc (glucose), BP (blood pressure), ST (skin thickness), In (insulin), BMI (body mass index), DP (diabetes pedigree), and age. Meanwhile, the outcome or the result of the prediction is the target feature in this study, which shows whether a person is diagnosed with DM (1) or not diagnosed (0). The next stage is to normalize the data, the results of data normalization are shown in Table 6. After that, dataset separation is carried out, the model is trained using 80% of the dataset, with remaining twenty percent left aside for performance testing.

Table 6. Result After Data Normalization

No.	Pr	Gc	BP	ST	In	BMI	DP	Age
1	0,89930	2,31923	-0,36687	0,43161	-0,71101	0,52546	0,80704	0,67175
2	-0,58834	0,76569	0,27936	0,87296	0,92547	0,86081	-0,42758	-0,37945
3	0,00671	0,18312	0,20531	1,33379	-0,71101	0,29218	-0,50474	-0,81745
4	-1,18340	1,63956	0,44091	0,68381	1,52438	2,07096	2,03573	2,16097
5	0,60177	-1,20858	0,44091	1,33379	-0,71101	-0,15980	-0,27028	0,75935
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
571	0,00671	0,70096	-1,17465	0,43161	0,46995	-0,40766	-0,55223	0,32135
572	1,19683	-0,88493	-0,04375	-1,33379	-0,71101	0,65669	0,03540	2,07337
573	1,19683	2,41633	0,11780	-1,33379	-0,71101	-0,93255	2,13071	0,49655
574	0,30424	1,41300	0,27936	-1,33379	-0,71101	1,95432	-0,39493	-0,55465
575	0,00671	0,28021	-0,36687	-0,00974	0,68927	0,11722	-0,92915	-0,46705

After the preprocessing stage is complete, the data that is ready will be processed using the *Logistic Regression* and SVM methods, hyperparameter tuning is performed using the grid search method to enhance the performance of both models. Grid search helps to find the best hyperparameter combination. The best hyperparameter combination is selected based on the results of evaluating model performance using training data.

B. Processing

The processing stage is to develop the best prediction model using the Logistic Regression and SVM methods. The best prediction model uses a grid search approach to obtain the best parameter combination. The candidate hyperparameter combination of Logistic Regression is shown in Table 1, and SVM is shown in Table 2. The grid search test results for the best prediction model of Logistic Regression are shown in Table 7 and SVM is shown in Table 8.

Table 7. Grid Search Parameter Combination Result for LR

Model	**Parameter**		
	C	**Fit_intercept**	**Solver**
Logistic Regression	0.1	True	Libliner

Table 8. Grid Search Parameter Combination Result for SVM

Model	**Parameter**		
	C	**Gamma**	**Kernel**
SVM	0,1	Scale	Linear

C. Postprocessing

Once the model is trained, a postprocessing step is performed to assess how well the model using a confusion matrix, which gives an idea of how well the model performs classification by evaluating how the predicted outcomes align with the proper labels. The model is evaluated by calculating the resulting values and presenting a diagrammatic visualization of the confusion matrix. The confusion matrix is employed to examine how well the model performs, as well as calculate matrices such as accuracy, precission, recall, and f1-score which are very useful in evaluating the prediction model shown in Table 6. Model performance calculations are based on Equations (3), (4), (5), and (6) whose values are obtained from the confusion matrix visualizations in Figures 3 and 4.

1) Logistic Regression Implementation: Heatmap Figure 3 shows the confusion matrix visualization results of the Logistic Regression model implemented for classification between diabetes and non-diabetes. Based on the visualization shown, there are 85 non-diabetes data that are correctly classified as non-diabetes (TN), 9 non-diabetes data that are misclassified as diabetes (FP), 21 diabetes data that are misclassified as non-diabetes (FN), and 30 diabetes data that are correctly classified as diabetes (TP).

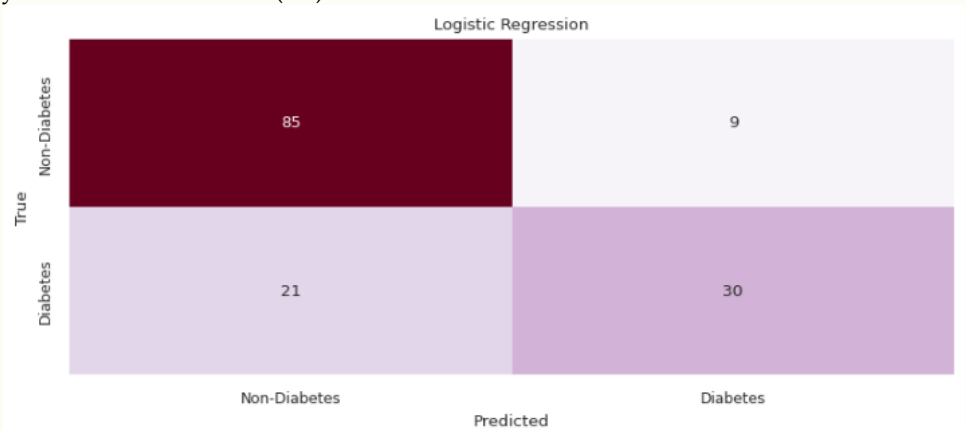

Figure 3. Logistic Regression Visualization Confusion Matrix

$$Accuracy = \frac{30+85}{145} = 0{,}7931$$

$$Precission = \frac{30}{30+9} = 0{,}7692$$

$$Recall = \frac{30}{30+85} = 0{,}5882$$

$$F1 - Score = 2 \times \frac{0{,}7692 \times 0{,}5882}{0{,}7692+0{,}5882} = 2 \times \frac{0{,}4524}{1{,}3574} = 0{,}66656$$

2) Support Vector Machine (SVM) Implementation: Figure 4 presents a heatmap displaying the visualization of the confusion matrix generated by the implemented SVM model. The model successfully classified 84 correctly classified non-diabetes data as non-diabetes (TN), 10 incorrectly classified non-diabetes data as diabetes (FP), 23 incorrectly classified diabetes data as non-diabetes (FN), and 28 correctly classified diabetes data as diabetes (TP).

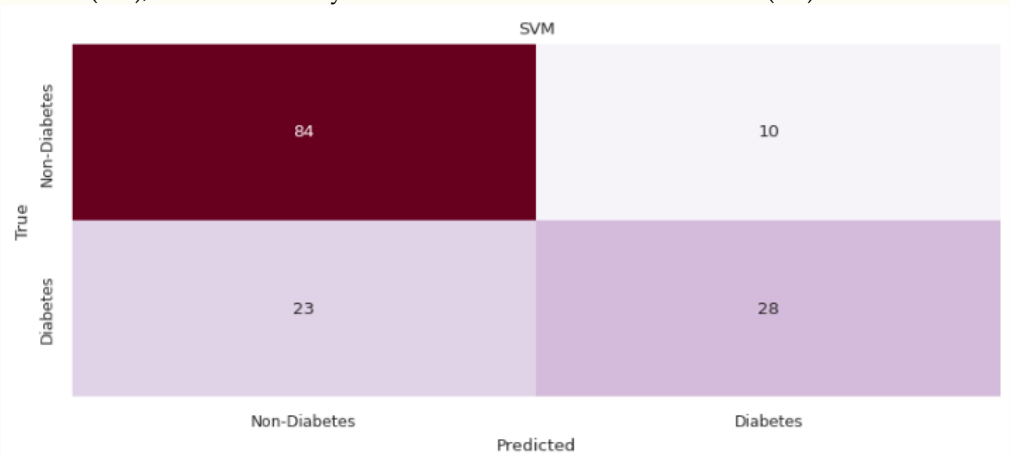

Figure 4. SVM Visualization Confusion Matrix

$$Accuracy = \frac{28+84}{145} = 0{,}7724$$

$$Precission = \frac{28}{28+10} 0{,}7368$$

$$Recall = \frac{28}{28+23} = 0{,}5490$$

$$F1 - Score = 2 \times \frac{0{,}7368 \times 0{,}5490}{0{,}7368 + 0{,}5490} = 2 \times \frac{0{,}4045}{1{,}2853} = 0{,}6294$$

Table 9. Performance Result Accuracy, Precision, Recall dan F1-Score

Value	*Logistic Regression*	SVM
Accuracy	79,31%	77,24%
Precision	76,92%	73,68%
Recall	58,82%	54,90%
F1-Score	66,67%	62,92%

Based on the results in Table 6, LR model shows superior performance because LR works effectively on data that has a linear relationship between the independent and dependent variables, as seen from the correlation between several features in the dataset and the diabetes outcome. In addition, LR is more stable in handling numerical variables and is able to interpret the probability of occurrence well, making it an effective method for binary classification cases, such as DM early detection prediction (Gunawan et al., 2020). The Logistic Regression method has better approach compared to the SVM, where the Logistic Regression produces superior accuracy values of 79.31%, precision 76.92%, recall 58.82%, and f1-score 66.67% while SVM gets 77.24% for accuracy, 73.68% precision, 54.90% recall, and 62.92% for f1-score.

CONCLUSIONS

Based on this research, made using LR and SVM methods. The remaining twenty percent of the data is used for testing to predict Diabetes Mellitus disease based on the attributes in the data, and the remaining eighty percent is used for training. After conducting a series of experiments and model evaluation, the results of the hyperparameter estimation of the Logistic Regression model using grid search the model achieved an accuracy of 79.31%, a precision of 76.92%, a recall of 58.82%, and an f1-score of 66.67%. While the SVM model obtained an accuracy value of 77.24%, precision of 73.68%, recall of 54.90%, f1-score of 62.92%. From the evaluation results, it is evident that the Logistic Regression surpasses the SVM model by a small margin.

REFERENCES

Amelia, U., Indra, J., & Masruriyah, A. F. N. (2022). Implementasi Algoritma Support Vector Machine (SVM) Untuk Prediksi Penyakit Stroke Dengan Atribut Berpengaruh. *Scientific Student Journal for Information, Technology and Science*, *III*(2), 254–259.

Aris, F. (2019). *Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi*. *1*(1), 1–6.

Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., & Pratama, A. D. (2022). *Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm*. *1*(2), 107–114. https://doi.org/10.55123/jomlai.v1i2.598

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Damayanti, E., Vitianingsih, A. V., Kacung, S., Suhartoyo, H., & Lidya Maukar, A. (2024). Sentiment Analysis of Alfagift Application User Reviews Using Long Short-Term Memory (LSTM) and Support Vector Machine (SVM) Methods. *Decode: Jurnal Pendidikan Teknologi Informasi*, *4*(2), 509–521. https://doi.org/10.51454/decode.v4i2.478

Desiani, A., Akbar, M., Irmeilyana, I., & Amran, A. (2022). Implementasi Algoritma Naïve Bayes dan Support Vector Machine (SVM) Pada Klasifikasi Penyakit Kardiovaskular. Jurnal Teknik Elektro Dan Komputasi (ELKOM), 4(2), 207-214.

Diabetes Dataset. (n.d.). Retrieved March 4, 2025, from https://www.kaggle.com/datasets/mathchi/diabetes-data-set

Firdaus, A. A., Yudhana, A., & Riadi, I. (2023). Analisis Sentimen Pada Proyeksi Pemilihan Presiden 2024 Menggunakan Metode Support Vector Machine. Decode: Jurnal Pendidikan Teknologi Informasi, 3(2), 236-245. https://doi.org/10.51454/decode.v3i2.172

Gunawan, M. I., Sugiarto, D., & Mardianto, I. (2020). Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *6*(3), 280. https://doi.org/10.26418/jp.v6i3.40718

Hovi, H. S. W., Id Hadiana, A., & Rakhmat Umbara, F. (2022). Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM). *Informatics and Digital Expert (INDEX)*, *4*(1), 40–45. https://doi.org/10.36423/index.v4i1.895

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-68771-z

Lee, J., Park, H., Kim, M., Yoon, J., Yoo, K., & Byun, S. J. (2024). FastSimiFeat: A Fast and Generalized Approach Utilizing k-NN for Noisy Data Handling. *International Conference on Information and Knowledge Management, Proceedings*, 1143–1152. https://doi.org/10.1145/3627673.3679591

Mardiana, T., Ditama, E. M., & Tuslaela, T. (2020). an Expert System for Detection of Diabetes Mellitus With Forward Chaining Method. *Jurnal Riset Informatika*, *2*(2), 69–76. https://doi.org/10.34288/jri.v2i2.121

Marwati, F., & Fauzi, R. (2024). Prediksi Penyakit Diabetes Melitus Menggunakan Jaringan Syaraf Tiruan Dengan Metode Backpropagation. *Jitu: Jurnal Informatika Utama Hal*, *2*(1), 26–34.

Maulidah, N., Supriyadi, R., Utami, D. Y., & Hasan, F. N. (2021). *Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes*. *7*(1), 63–68.

Oktaviana, A., Wijaya, D. P., Pramuntadi, A., & Heksaputra, D. (2024). *Prediction of Type 2 Diabetes Mellitus Using The K-Nearest Neighbor (K-NN) Algorithm Prediksi Penyakit Diabetes Melitus Tipe 2 Menggunakan Algoritma K-Nearest Neighbor (K-NN)*. *4*(July), 812–818.

Pratama, A., Nurcahyo, A. C., & Firgia, L. (2023). Penerapan Machine Learning dengan Algoritma Logistik Regresi untuk Memprediksi Diabetes. *Prosiding CORISINDO 2023*, 116–121.

Saepudin, A., Faqih, A., & Dwilestari, G. (2024). *Perbandingan Algoritma Klasifikasi Support Vector Machine, Random Forest dan Logistic Regression Pada Ulasan Shopee*. *18*(1), 178–192.

Suprihati, F. R. (2021). Analisis Klasifikasi SMS Spam Menggunakan Logistic Regression. *Jurnal Sistem Cerdas*, *4*(3), 155–160. https://doi.org/10.37396/jsc.v4i3.166

Tangkere, B. B. (2024). Analisis Performa Logistic Regression dan Support Vector Classification untuk Klasifikasi Email Phising. Jurnal Ekonomi Manajemen Sistem Informasi (JEMSI), 5(4), 442–450. https://doi.org/10.31933/jemsi.v5i4.1916

Todkar, S. (2016). Diabetes Mellitus the ‘Silent Killer' of mankind: An overview on the eve of World Health Day! *Journal of Medical and Allied Sciences*, *6*(1), 39. https://doi.org/10.5455/jmas.214333