ISSN: 2654-3788
e-ISSN: 2654-4229
DOI: 10.32493/jtsi.v7i1.38258

Comparative Analysis of Logistic Regression, SVM, Xgboost, and Random Forest Algorithms for Diabetes Classification

Rahmat Hidayat[1], Deni Mahdiana[2], Anggun Fergina[3]

[1*,2]Master of Computer Science, Budi Luhur University, Jl. Ciledug Raya, North Petukangan, South Jakarta, Jakarta, Indonesia, 12260
e-mail: [*1]2311600031@student.budiluhur.ac.id, [2]deni.mahdian@email.ac.id

[3]Informatics Engineering, Nusa Putra University, Jl. Raya Cibolang No.21 Cisaat Sukabumi, Sukabumi, 43152
e-mail: [3]anggun.fergina@nusaputra.ac.id

Submitted Date: January 23rd, 2024
Reviewed Date: January 26th, 2024
Revised Date: January 27th, 2024
Accepted Date: January 29th, 2024

Abstract

Diabetes is a disease that can attack anyone, where this disease occurs because there is excessive sugar content in the human body. Therefore, prevention of diabetes is necessary so that preventive measures can be given as early as possible. In this research, a classification process will be carried out using the Random Forest algorithm, Support Vector Classification and XGBoost. This research will use a dataset which consists of 768 total data with a distribution of non-diabetic data of 500 and a distribution of diabetes data of 268. For the classification results after testing, the results were that classification using random forest obtained a testing accuracy of 79.22%, with using support vector classification gets a testing accuracy of 76.62%, using XGBoost gets a testing accuracy of 79.22% using Logistic Regression gets a testing accuracy of 80.52%. The best classification value is obtained when using the Logistic Regression algorithm, namely with a precision of 79.00%, recall of 77.00% and F1-Score of 78.00%.

Keywords: Diabetes; Random forest; SVC; XGBoost; Classification

1. Introduction

High blood sugar levels are a hallmark of diabetes, a chronic metabolic condition brought on by the body's ineffective production or utilization of insulin (Pelegrín & Hospitaleche, 2022). Insulin is a hormone that helps body cells absorb glucose, which helps to control blood sugar levels (Gray, 1996). Diabetes comes in two basic forms: type 1, where the immune system assaults the pancreatic cells that create insulin, preventing the body from creating any, and type 2, where the body either produces insufficient amounts of insulin or uses it inefficiently (Clucas et al., 2022). Genetics, weight, and way of life are risk factors. Diabetic consequences that persist over time include issues with the heart, eyes, nerves, and kidneys (Fattorini & Olmastroni, 2021). Changes in lifestyle, food, regular exercise, and frequently the use of insulin or medications are all part of managing diabetes. Controlling blood sugar levels is essential to treating this illness and avoiding major consequences. Indonesia is rated fifth in the world among countries with the highest number of diabetes sufferers, with 19.47 million people expected to have the disease in 2021, according to data from the International Diabetes Federation (IDF). According to IDF predictions, the number of Indonesians with diabetes is expected to rise further, perhaps reaching 28.57 million in 2045—a 47% increase from 2021. In addition, Indonesia has the highest number of type 1 diabetes patients in ASEAN, with 41.8 thousand cases in 2022, according to the IDF research. One of the chronic illnesses that kills people the most frequently in Indonesia is diabetes mellitus.

Data science and statistical techniques are applied in data mining, which is a component of machine learning. According to Farahani et al (F Shahrabi Farahani, M Alavi, M Ghasem, 2020), the volume and complexity of data are growing, therefore appropriate tools are required to process data, analyze existing data, and extract valuable knowledge or information from the data. Data mining techniques have significantly expanded as a result of this part, which helps us uncover information hidden in data. As data mining techniques have grown, research is moving away from traditional statistical methodologies that were thought to be standard procedures since they were thought to be less effective at processing larger volumes and more complicated data.

Random forest is a straightforward model that was first presented by Breiman in 2001. It generates anticipated outcomes by binaryly separating predicted variables. Despite all of their advantages, decision trees can be an inaccurate algorithm when used with more complex data. To generate the model's output, the random forest approach employs several classification and regression trees constructed from a random portion of the training dataset and a random subset of the prediction variables. Predictions for observations are generated for each iteration based on the combination of the trees' results. Consequently, random forests outperform decision tree models in terms of accuracy, and they also regularly outperform other models in the data categorization domain in terms of prediction accuracy. The Support Vector Machine (SVM), which operates on the structured risk minimization concept, includes Support Vector Classification (SVC) (Robles-Velasco et al., 2020). Support Vector Machine and Support Vector Classification both work by reducing the distance between the sample (maximum margin) and the decision border (support vector) (Rákos et al., 2020). Consequently, for every class sample, a hyperplane will be searched during the procedure (Liu & Rao, 2020), Alternatively put, this approach will seek out a hyperplane that divides the positive class and the negative class in the best possible way utilizing the maximum possible margin (Djedidi et al., 2021). An algorithm known as XGBoost is a Gradient Boosting Decision Tree realization (Zhang et al., 2019). A number of decision tree models are combined in the ensemble method, which also includes this approach (Cherif & Kortebi, 2019). XGBoost is utilized in the procedure to enhance the decision tree so that overfitting of the constructed tree model is prevented (Thongsuwan et al., 2021). Since supervised classification is the nature of logistic regression (Shah et al., 2020), labels must be used as targets. When used correctly, logistic regression performs exceptionally well at predicting discrete probabilities (only has two classes) (Samsudin et al., 2019). This is possible because the logistic function is used in logistic regression to determine the probability value of an occurrence (Thongsuwan et al., 2021).

The random forest and SVC algorithms will be used in this study to classify diabetes cases. The random forest algorithm is used because it is an ensemble learning technique that creates a forest by combining multiple decision tree models. In the meanwhile, SVC is being used since it is a component of the SVM method, which is well-known for its capacity to categorize data. The rationale behind utilizing XGBoost is its ability to handle data with imbalanced values with good accuracy. Meanwhile, since this model is straightforward and utilized in the binary classification procedure, logistic regression is employed. The intention behind utilizing many algorithms for diabetes classification is to enable comparative analysis and identify the optimal algorithm for diabetes classification.

It is crucial to do this research to make the process of forecasting diabetes simpler so that values or factors like blood sugar and others can be used in the analytic process. In order for you to receive the proper medical care after receiving a diabetes diagnosis and for it to be effectively treated and healed. This research is novel because it makes use of a diabetes dataset that hasn't been used much in other studies. As a result, it can give a general overview of the machine learning method used in the diabetes classification process.

Radja et al.'s earlier study (Pratomo et al., n.d.) describes how machine learning is used to classify diabetes cases. The purpose of this study is to use machine learning to the categorization of diabetic diseases. The research's findings, which have a 77.3% accuracy rate, represent the best diabetes classification method utilizing the SVM algorithm. Thaiyalnayaki's 2021 (Thaiyalnayaki, 2021) research examines the process of classifying diabetes through the application of deep learning and machine learning techniques. Finding an algorithm that can carry out the diabetes categorization procedure as efficiently as possible is the goal of this research. According to the study's findings, classification using SVM yields an accuracy of 65.102% while MLP deep learning yields an accuracy of 77.474%. Describes how the C4.5 algorithm, SVM, and linear regression are used to classify diabetes. According to the research findings, the linear regression method has an MSE of 0.216, the SVM algorithm has an accuracy of 82%, and the C4.5 algorithm has a 75% accuracy rate. Dewi et al. (Dhita Diana Dewi, Nurul Qisthi, Siti Sarah Sobariah Lestari, 2023) performed research comparing neural network and SVM approaches for diabetes categorization. According to the study's findings, the neural network approach produced 77.60% accuracy and 65.24% SVM. Bayesian and MLP approaches were used in research by Rasna et al (Rasna & Matdoan, 2022) to explore the diabetes categorization procedure. Using the Bayesian and MLP classification techniques, the research's findings yielded an accuracy of 81.89%. Research conducted by Hunafa et. All (Hunafa & Hermawan, 2023) in 2023 discussed the comparison of the Naive Bayes algorithm and KNN in the classification of diabetes. The research yielded the best classification, 83.02%, when the KNN algorithm was used without the SMOTE approach. A study published in 2022 by Rinanda et al (Rinanda et al., 2022) examines the categorization of diabetes using the KNN and Naive Bayes algorithms. According to the research findings, classification using KNN yields a testing accuracy of 74.48% and using Naive Bayes yields a testing accuracy of 75.78%. A study published in 2022 by Cahyani et al (Cahyani et al., 2022) examines the use of logistic regression to classify diabetes. According to the research's findings, 76% of the samples were accurate following testing. In 2021, Soleh at. All (Soleh et al., 2021) did research on the use of logistic regression to classify diabetes. Test accuracy in this study was 80%, according to the testing data.

According to a review of the literature, there aren't many studies that address the classification process using the XGBoost classifier method to be able to conduct a comparative classification process when compared with the classification process other studies have used, which makes the XGBoost classifier method the novel feature of this research. in order for this research to benefit the world. Information and communication technology can provide additional understanding of how well the Random Forest, SVM, XGBoost, and Logistic Regression algorithms perform in the model built for diabetes classification.

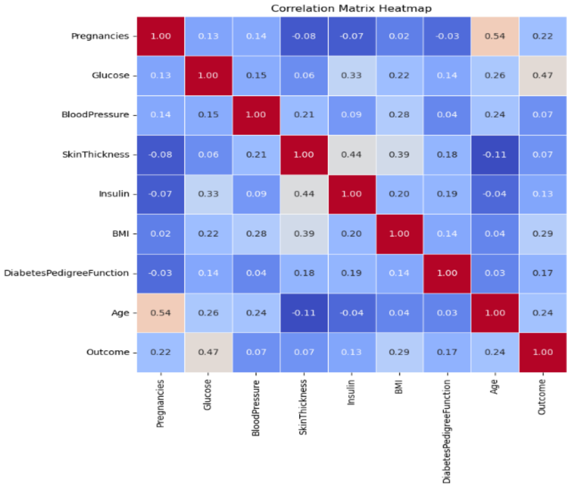
2. Methodology

A diabetic dataset with the.csv extension that we downloaded from the website kaggle.com will be used in this study. The 768 total data utilized in this study are divided into two classes: the non-diabetic class and the diabetes class. There are 268 diabetes data and 500 non-diabetic data in the data distribution. The 768 total data are split into 90% training data and 10% testing data for the purpose of training and testing the model. Testing data is used to evaluate a model that has already been trained to perform data recognition, whereas training data is used to enable the model to learn patterns from the data. Table 1 provides the dataset for visualization.

Table 1 . Visualization of Datasets

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Predigree	Age	Outcome
2	128	78	37	182	43.3	1.224	31	1
3	111	56	39	0	30.1	0.557	30	0
9	89	62	0	0	22.5	0.142	33	0
2	108	62	32	56	25.2	0.128	21	0
10	125	70	26	115	31.1	0.205	41	1

Pregnancies or the total number of pregnancies experienced are the factors associated to carrying out the diabetes categorization procedure. glucose, or the body's amount of glucose; blood pressure, or the body's blood pressure; skin thickness, or the body's thickness of skin; and insulin, or the body's amount of insulin. body mass index (BMI) is a metric that expresses the relationship between a person's height and body weight. A person's age is their age, their diabetes predigree is their genetic score for having the disease owing to inherited causes, and their result is a variable that represents the dataset's aim or class, such diabetes classification or No. Of course, the dataset contains a variable that affects the target class or result; the correlation matrix makes this information clear. Figure 1 displays the correlation matrix from this study.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
SkinThickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1.00

Figure 1. Correlation Matrix

The correlation matrix results between each variable and the outcome or target variable are displayed in Figure 1. Figure 1 illustrates how the glucose variable, which represents the amount of glucose or blood sugar in the human body, substantially affects the result value or goal variable.

An unequal distribution of the data utilized means that a normalizing technique must be performed. One method to make sure that the data in a dataset has the same range of values (i.e., no data gaps) is to normalize the data. Therefore, normalization is a crucial step when working with unstructured data that has a wide range of values[13]. When we attempt to train the model, issues will arise if there are gaps in the data. One normalizing approach is MinMaxScaler. A sort of normalization called MinMaxScaler may arrange the values in a dataset so that every piece of data has a value range of 0 to 1. Points 1 and 2 provide the normalizing approach for computations.

$$Z_{std} = \frac{Z - Z.minimun}{(Z.maximum - Z.minimum)} \qquad 1)$$

$$Z_{scale} = Z_{std} * (Z.maximum - Z.minimum) + Z.minimum \qquad 2)$$

One machine learning technique that falls within the ensemble learning category is called Random Forest. This means that in order to increase performance and prediction accuracy on data, a random forest will integrate many models, typically the same model. The capabilities of many decision tree algorithms are combined in Random Forest. As a consequence, the random forest method creates several decision trees from randomly selected data, combining the outcomes of each tree's forecast to create the final prediction. The random forest pseudocode and computations are provided below.

$$Predict(Z) = \frac{(deciTree\_1(Z) + deciTree\_2(Z) + \ldots + deciTree\_N(Z)}{N} \qquad 3)$$

```
Pseudocode
Define N
(1) Select randomly X feature from data
(2) For each I in X
    (1) Ent (Z)= -∑_(x=1)^n P_x 〖log〗_2 〖(P〗_x)
    (2) Ent (Z,I)= ∑_(m ∈ I) P(m)E(m)
    (3) InfoGain (C,Z)=Ent (C)- Ent(Z,I)
    (4) Select node X which has the highest information gain
    (5) Split node into sub node
Repeat steps 1 to 5 until construct tree and reach minimum number of sample that required
(3) Repeat step 1 to 2 for N times until building forest of N trees.
```

One machine learning classification approach is called Support Vector Classification (SVC). The process is determining which hyperplane (separation plane) is best for classifying the data. This hyperplane is selected to optimize the distance between it and the data points for both classes. SVC can handle nonlinear data by using kernel functions and can be applied to binary or multiclass classification jobs. You may maximize SVC performance in your classification jobs by adjusting settings. Point 4 provides the mathematical computation for SVC.

$$\omega * Z + b = 0 \qquad (4)$$

Where, ω represents the hyperplane weight vector, X represents the features used, and b is the bias (offset) value.

Gradient Boosting Decision Trees are realized in the method known as XGBoost (Zhang et al., 2019). A number of decision tree models are combined in the ensemble method, which also includes this approach (Cherif & Kortebi, 2019). XGBoost is utilized in the procedure to enhance the decision tree so that overfitting of the constructed tree model is prevented (Thongsuwan et al., 2021). Thus, it is envisaged that the model developed throughout this procedure would be able to produce accurate and ideal forecasts. Here is the pseudocode for XGBoost.

```
Define model = [], num boost (iteration), x, y
(1) Repeat for iteration
    (1) gradientMin = $-\nabla L(y, predict1)$
    (2) base = train decision tree classifier (x,y)
    (3) base = predict(GradientMin)
    (4) model.append(base)
    (5) predict1 = $predict1 + learning\ rate * softmax(base\_predict)$
```

When used correctly, supervised classification techniques like logistic regression (Shah et al., 2020), do exceptionally well at predicting discrete probabilities (Purnamasari & Syakti, 2020). This is possible because the logistic function is used in logistic regression to calculate the probability value of an occurrence (Kumar & Ramamoorthy, 2022). Thus, the output of the logistic function is either 0 or 1 (Shah et al., 2020). For this reason, binary class classification benefits greatly from the logistic regression approach. The pseudocode for logistic regression is provided below.

```
Define w (weight), b (bias), ∝ (learn rate), iteration
(1) $\sigma(x) = \frac{1}{(1+e^{-z})}$
(2) Repeat for iteration
        (1) x = w * feature + b
        (2) prob = $\sigma(x)$
        (3) loss = $-(target * log\ log\ (prob) + (1 - target) * log\ log\ (1 - prob))$
        (4) dw = $\frac{1}{total\ data} * \sum ((prob - target) * feature)$
        (5) db = $\frac{1}{total\ data} * \sum (prob - target)$
        (6) w = w * $\alpha$ * dw
        (7) b = b * $\alpha$ * db
(3) if prob > 0.5 then 1 else 0
```

After the data is trained and evaluated using the random forest method and the SVC algorithm, the research employs a confusion matrix to determine the algorithm's performance outcomes. The output variables that are produced include precision, recall, f1-score, and support per class. There are four values in the confusion matrix: True Positive, True Negative, False Positive, and False Negative. These four values can be used to determine the model's performance value: recall, which measures how well the mode finds all positive cases from all of the model's iterations; precision, which measures how well the model can identify when there are positive cases; and f1-score, which is the harmonic value (average calculation) between recall and precision. Points 5, 6, and 7 provide the values for f1-score, recall, and accuracy.

$$Precision = \frac{TP}{(TP + FP)} \tag{5)}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{6)}$$

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{7)}$$

This study employs Jupyter Notebook as an IDE and the Python programming language with built-in Python libraries like Numpy, Pandas, and Sklearn for developing code in the process of categorizing diabetes. The random forest algorithm, SVC, and XGBoost are the three algorithms used in the testing and training phases of the classification process. Figure 2 shows the workflow procedure for employing algorithms for both training and testing.

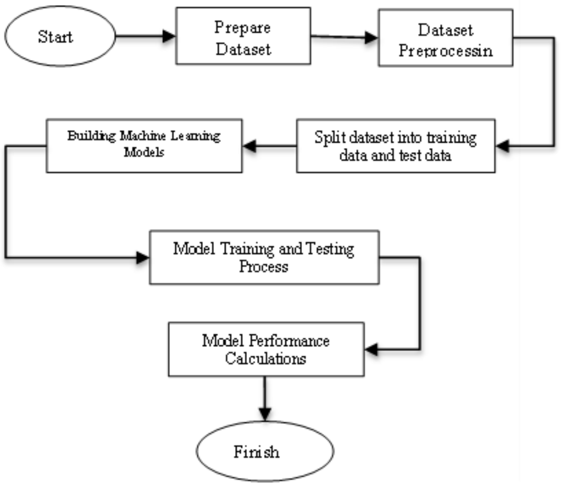

Start
Prepare Dataset
Dataset Preprocessin
Split dataset into training data and test data
Building Machine Learning Models
Model Training and Testing Process
Model Performance Calculations
Finish

Figure 2. Workflow classification

Initially, the input data for the diabetes categorization step comes from what we were able to get from Kaggle.com. After that, the downloaded data may be handled. The steps in the diabetes categorization procedure are as follows:

a. Read the CSV data to be processed.
b. Normalize the data using a min max scaler, because the input data still has gaps in the attribute mass so it can have an impact on computing using decision tree and random forest algorithms. When using the min max scaler, the data will have a range between 0 and 1, so the data is more evenly distributed.
c. Setelah melakukan normalisasi, kita harus melakukan split data. Pada penelitian ini akan menggunakan split data 90:10, yang berarti 90% menjadi data training dan 10% menjadi data testing.
d. After splitting the data, we build models for random forest, SVC, XGBoost and Logistic Regression classifiers, for the parameters shown in tables 1, 2, 3 and 4:

Table 2. Parameter for Random Forest

Parameter	Values
N estimators	100
Criterion	gini
Max depth	None
Min samples split	2
Min samples leaf	1
Min weight fraction leaf	0.0
Max features	auto
Max leaf nodes	None
Min impurity decrease	None
Bootstrap	True
Oob score	False
N jobs	-1
Random state	42

Table 3. Parameter for SVC

Parameter	Values
C	1.0
kernel	sigmoid
degree	3
gamma	1.0
coef0	0.0
shrinking	True
probability	False
tol	0.001
class_weight	balanced
verbose	False
max_iter	-1
decision_function_shape	ovr
random_state	None

Table 4. Parameter for XGBoost

Parameter	Values
objective	Binary:logistic
n_estimators	100
learning_rate	0.1
max_depth	3
random_state	42
booster	GBtree
gamma	0
min_child_weight	1
max_delta_step	0
subsample	1
colsample_bytree	1
colsample_bylevel	1
colsample_bynode	1

Table 5. Parameter for Logistic Regression

Parameter	Values
objective	Binary:logistic
n_estimators	100
learning_rate	0.1
max_depth	3
random_state	42
booster	GBtree
gamma	0
min_child_weight	1
max_delta_step	0
subsample	1
colsample_bytree	1
colsample_bylevel	1
colsample_bynode	1

e. After the model is created, the data will be trained using a decision tree or random forest algorithm. Then, the algorithm that has been trained will be tested using testing data.

f. After testing, the results of the machine performance training report can be viewed using the confusion matrix.

3. Result And Analysis

Using the Jupyter Notebook IDE and the Python programming language, the decision tree and random forest algorithms were employed in this study to carry out the penguin categorization procedure. In order to improve accuracy and efficiency when working on the algorithm, the process begins with the data being supplied and read by the software. Next, the read data is normalized to make the existing data more uniformly distributed. The data will be normalized and then divided into 90% training data and 10% testing data. Following the data's breakdown, a model—either a decision tree or random forest model—is constructed, and training is done using an algorithm on the previously broken-down training set of data. Using the available testing data, the model is evaluated once it has been trained. Testing yielded findings that indicated the following: 79.22% accuracy for the random forest method, 76.62% accuracy for SVC, 79.22% accuracy for XGBoost, and 80.52% accuracy for logistic regression. The pictures 3, 4, 5, and 6 display the test results obtained from the confusion matrix.

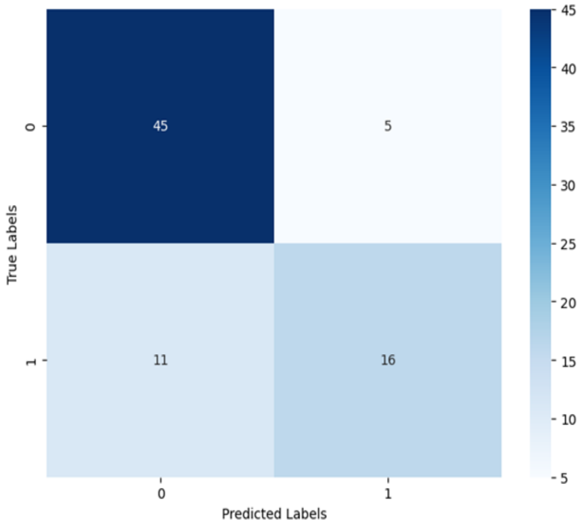


Figure 3. Classification report for Random Forest

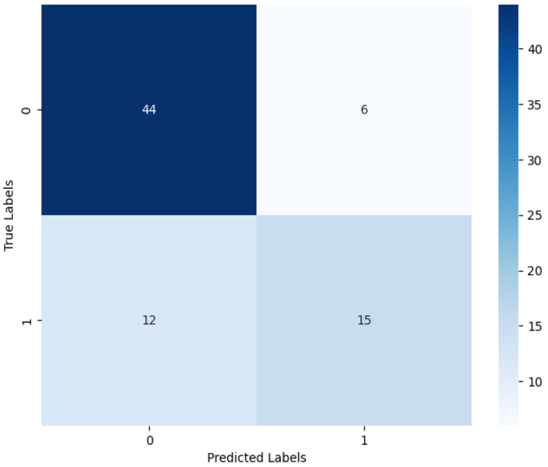


Figure 4. Classification report for SVC

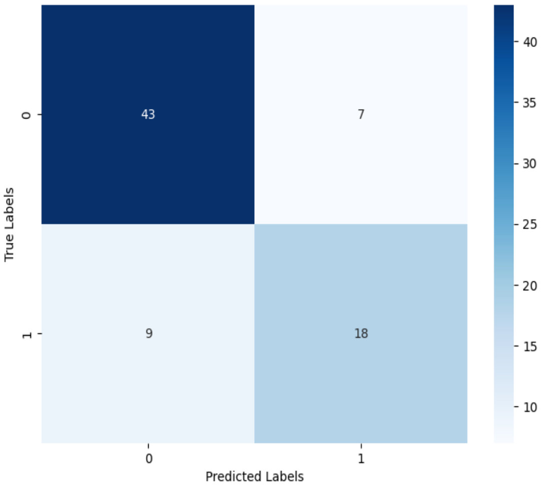


Figure 5. Classification Report for XGBoost

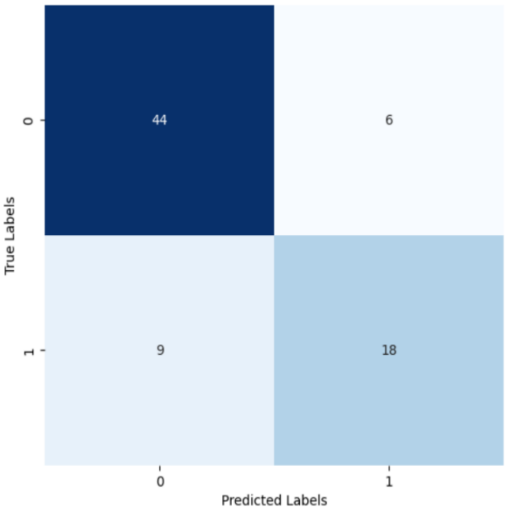


Figure 6. Classification Result for Logistic Regression

The model-built categorization results are shown in Figures 3, 4, 5, and 6. In comparison to the support vector classification (SVC) technique, it is evident that the confusion matrix results produced by the random forest and XGBoost algorithms yield higher accuracy, recall, precision, and F1-score. The random forest, XGBoost, SVC, and logistic regression methods have average precisions of 78.00%, 75.00%, 77.00%, and 79.00%, respectively. This demonstrates how accurate random logistic regression is in predicting data.

Additionally, the random forest, XGBoost, SVC, and logistic regression algorithms had average recalls of 75.0%, 75.0%, 72.00%, and 77%, respectively. This demonstrates how effectively and accurately the logistic regression method can categorize and forecast all classes. Furthermore, 76.00%, 77.00%, 73.00%, and 78.00% are the average F1-scores for the random forest, XGBoost, SVC, and logistic regression methods, respectively. This suggests that the designed logistic regression approach has a decent recall to precision ratio. Since the support value (77 for all models) indicates the quantity of data points utilized to test the model, it is the same for all models. Table 1 contains the table categorization report.

Table 6. Classification Result from Model

Model	Precision	Recall	F1-Score	Accuracy
RF	78.00%	75.00%	76.00%	79.22%
SVC	75.00%	72.00%	73.00%	76.62%
XGBoost	77.00%	76.00%	77.00%	79.22%
LR	79.00%	77.00%	78.00%	80.52%

The categorization report's results are displayed in Table 6 as f1-score, recall, and accuracy values. Based on these findings, the greatest accuracy value that could be attained by logistic regression was 80.52%. There is an increase in accuracy in this research as compared to other studies that used the same dataset, namely from kaggle.com with a total of 768 data. There was a 3.22% gain in comparison to research 18 (Rahayu et al., 2023), which used SVM to get the greatest accuracy of 77.3%. Meanwhile, there was a 3.10% improvement in accuracy in this study as compared to research, which used MLP to reach the greatest accuracy, which was 77.42%. This figure indicates that, in comparison to earlier studies, this study was successful in raising accuracy.

4. Conclusion

Logistic regression is strongly supported by the confusion matrix results, which include measures like accuracy, recall, precision, and F1-score. Compared to SVC's 75.00% average precision, its astounding average precision of 79.00% demonstrates a remarkable capacity for precise prediction-making. This higher accuracy shows that random forests perform quite reliably in classification, which is crucial for a variety of applications.

Average recall, a metric that assesses the model's accuracy in class identification, also shows that logistic regression performs significantly better than SVC, with a value of 77.00%, while SVC comes in last at 72.00%. This difference shows that all classes are consistently and successfully classified by the random forest method, which is a crucial feature in situations when all classes are equally relevant. The average F1-score, which illustrates the capabilities of logistic regression by balancing the trade-off between accuracy and recall, was 78.00%, whereas SVC yielded a score of 73.00%. The random forest's higher F1 value highlights its capacity to maintain a balanced approach, offering high accuracy while guaranteeing thorough class coverage. The amount of data points utilized for testing was indicated by the support value, which was constant at 77 for all models. Together, these findings support the hypothesis that, for this specific classification problem, logistic regression performs better than the SVC method. This discussion's primary finding is that random forests are not the best option for this classification problem because of their greater accuracy, recall, precision, F1-score, and balance between precision and recall.

The study's experimental findings indicate that the logistic regression technique works better than Support Vector Classification (SVC). Numerous variables contribute to this exceptional performance. Popular machine learning techniques include logistic regression and Support Vector Machines (SVM), particularly its SVC (Support Vector Classification) form for classification problems. Each has pros and cons. Since logistic regression is easy to understand, straightforward, and effective when handling data that can be divided into linear segments, it is frequently preferred over SVC in specific situations. A linear model called logistic regression forecasts the likelihood that an instance will fall into a specific class. Binary classification challenges are a good fit for logistic regression, which is also easily expandable to address multiclass issues. Its interpretability is one of its key benefits; in logistic regression, the coefficients show how each factor affects the log-odds of the anticipated result. This facilitates practitioners' comprehension and communication of model results, which can be crucial in situations where transparency is crucial. As the dimensionality of the data rises, SVC can become more computationally expensive and challenging to comprehend, even if it can accommodate non-linear correlations by using kernel functions. SVC can capture complicated decision boundaries and is especially useful in situations when the data cannot be separated linearly.

It has been concluded from the test results that the Random Forest classification method, These results suggest that the logistic regression model is a superior and more accurate option for classifying diabetes. It is advised that future study investigate and create models for diabetes categorization using multiple algorithms so that different models may be compared to identify the most effective model. Furthermore, future studies should take into account parameter tweaking as a means of optimizing model parameters in order to further enhance classification performance.

References

Cahyani, Q. R., Finandi, M. J., Rianti, J., Arianti, D. L., Dwi, A., Putra, P., & Artikel, G. (2022). Prediksi Risiko Penyakit Diabetes menggunakan Algoritma Regresi Logistik Diabetes Risk Prediction using Logistic Regression Algorithm Article Info ABSTRAK. *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, *1*(2), 2828–9099. https://doi.org/10.55123/jomlai.v1i2.598

Cherif, I. L., & Kortebi, A. (2019). 2019 Wireless Days, WD 2019. *IFIP Wireless Days*, *2019-April*, 1–6.

Clucas, G. V., Warwick-Evans, V., Hart, T., & Trathan, P. N. (2022). Using habitat models for chinstrap penguins, Pygoscelis antarctica, to inform marine spatial management around the South Sandwich Islands during the penguin breeding season. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *199*(March), 105093. https://doi.org/10.1016/j.dsr2.2022.105093

Dhita Diana Dewi, Nurul Qisthi, Siti Sarah Sobariah Lestari, Z. H. S. P. (2023). *Perbandingan Metode Neural Network Dan Support Vector Machinedalam Klasifikasi Diagnosa Penyakit Diabetes*. *3*(September), 828–839. https://cerdika.publikasiindonesia.id/index.php/cerdika/article/view/662/866

Djedidi, O., Djeziri, M. A., Morati, N., Seguin, J. L., Bendahan, M., & Contaret, T. (2021). Accurate detection and discrimination of pollutant gases using a temperature modulated MOX sensor combined with feature extraction and support vector classification. *Sensors and Actuators, B: Chemical*, *339*(March), 129817. https://doi.org/10.1016/j.snb.2021.129817

F Shahrabi Farahani, M Alavi, M Ghasem, Bt. (2020). Scientific Map of Papers Related to Data Mining in Civilica Database Based on Co-Word Analysis. *International Journal of Web Research*, *3*(1), 11–18.

Fattorini, N., & Olmastroni, S. (2021). Pitfalls and advances in morphometric sexing: insights from the Adélie penguin Pygoscelis adeliae. *Polar Biology*, *44*(8), 1563–1573. https://doi.org/10.1007/s00300-021-02893-6

Gray, O. (1996). Review Article. *Caribbean Quarterly*, *42*(4), 70–74. https://doi.org/10.1080/00086495.1996.11672093

Hunafa, M. R., & Hermawan, A. (2023). KLIK: Kajian Ilmiah Informatika dan Komputer Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Imbalace Class Dataset Penyakit Diabetes. *Media Online*, *4*(3), 1551–1561. https://doi.org/10.30865/klik.v4i3.1486

Kumar, K. V., & Ramamoorthy, M. (2022). Machine Learning-based spam detection using Naïve Bayes Classifier in comparison with Logistic Regression for improving accuracy. *Journal of Pharmaceutical Negative Results*, *13*(SO4), 548–554. https://doi.org/10.47750/pnr.2022.13.s04.061

Liu, W., & Rao, Z. (2020). Road Icing Warning System Based on Support Vector Classification. *IOP Conference Series: Earth and Environmental Science*, *440*(5). https://doi.org/10.1088/1755-1315/440/5/052071

Pelegrín, J. S., & Hospitaleche, C. A. (2022). Evolutionary and Biogeographical History of Penguins (Sphenisciformes): Review of the Dispersal Patterns and Adaptations in a Geologic and Paleoecological Context. *Diversity*, *14*(4), 1–20. https://doi.org/10.3390/d14040255

Pratomo, A. H., Universitas Pembangunan Nasional “Veteran” Yogyakarta, Universitas Pendidikan Indonesia, Institute of Electrical and Electronics Engineers. Indonesia Section, & Institute of Electrical and Electronics Engineers. (n.d.). *2019 5th International Conference on Science in Information Technology (ICSITech) : proceeding : October 23-24, 2019, Yogyakarta, Indonesia*.

Purnamasari, S. D., & Syakti, F. (2020). Implementasi Usability Testing dalam Evaluasi Website Sekolah. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *9*(3), 420–426. https://doi.org/10.32736/sisfokom.v9i3.1000

Rahayu, D. S., Afifah, J., & Intan, S. (2023). Classification of Diabetes Mellitus Using C4 . 5 Algorithm , Support Vector Machine (SVM) and Linear Regression Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma C4 . 5 , Support Vector Machine (SVM) dan Regresi Linear. *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, *1*(1 SE-), 56–63. https://journal.irpi.or.id/index.php/sentimas/article/view/550

Rákos, O., Aradi, S., & Bécsi, T. (2020). Lane change prediction using Gaussian classification, support vector classification and neural network classifiers. *Periodica Polytechnica Transportation Engineering*, *48*(4), 327–333. https://doi.org/10.3311/PPTR.15849

Rasna, & Matdoan, M. R. I. (2022). Metode Bayesian dan Multilayer Percepton dalam Mengklasifikasi Diabetes Mellitus. *Jurnal Sistim Informasi Dan Teknologi*, *4*, 82–86. https://doi.org/10.37034/jsisfotek.v4i2.132

Rinanda, P. D., Delvika, B., Nurhidayarnis, S., Abror, N., & Hidayat, A. (2022). Perbandingan Klasifikasi Antara Naive Bayes dan K-Nearest Neighbor Terhadap Resiko Diabetes pada Ibu Hamil. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(2), 68–75. https://doi.org/10.57152/malcom.v2i2.432

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Onieva, L. (2020). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. *Reliability Engineering and System Safety*, *196*. https://doi.org/10.1016/j.ress.2019.106754

Samsudin, N. M., Mohd Foozy, C. F. B., Alias, N., Shamala, P., Othman, N. F., & Wan Din, W. I. S. (2019). Youtube spam detection framework using naïve bayes and logistic regression. *Indonesian Journal of Electrical Engineering and Computer Science*, *14*(3), 1508–1517. https://doi.org/10.11591/ijeecs.v14.i3.pp1508-1517

Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1). https://doi.org/10.1007/s41133-020-00032-0

Soleh, M., Ammar, N., & Sukmadi, I. (2021). Website-Based Application for Classification of Diabetes Using Logistic Regression Method. *Jurnal Ilmiah Merpati (Menara Penelitian Akademika Teknologi Informasi)*, *9*(1), 23. https://doi.org/10.24843/jim.2021.v09.i01.p03

Thaiyalnayaki, K. (2021). Classification of diabetes using deep learning and svm techniques. *International Journal of Current Research and Review*, *13*(1), 146–149. https://doi.org/10.31782/IJCRR.2021.13127

Thongsuwan, S., Jaiyen, S., Padcharoen, A., & Agarwal, P. (2021). ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost. *Nuclear Engineering and Technology*, *53*(2), 522–531. https://doi.org/10.1016/j.net.2020.04.008

Zhang, R., Li, B., & Jiao, B. (2019). Application of XGboost Algorithm in Bearing Fault Diagnosis. *IOP Conference Series: Materials Science and Engineering*, *490*(7). https://doi.org/10.1088/1757-899X/490/7/072062