Journal of Social Research
P-ISSN:2827-9832
E-ISSN: 2828-335x
Volume 5, Number 8, July 2026
http://ijsr.internationaljournallabs.com/index.php/ijsr

Analysis of the Accuracy of Machine Learning Models Using the Confusion Matrix Method for the Naive Bayes and CatBoost Algorithms in Diagnosing Diabetes

Basyach Aryodito Tranggono Heryanto*, Onny Marleen
Universitas Gunadarma, Indonesia
Email: basyach@gmail.com*, onny_marleen@staff.gunadarma.ac.id

Keywords:
Diabetes, Accuracy Rate, Naive Bayes, CatBoost, Analysis.

ABSTRACT

Diabetes is a chronic metabolic disease characterized by high blood glucose levels that can lead to severe complications affecting various organs. The increasing prevalence of diabetes requires an accurate and efficient diagnostic approach to support early detection and medical decision-making. The development of machine learning technology provides opportunities to assist disease diagnosis by analyzing patient health data and identifying patterns associated with diabetes conditions. This study aims to compare the accuracy performance of the Naive Bayes and CatBoost algorithms in diagnosing diabetes using the confusion matrix evaluation method. The research method consists of several stages, including data collection, data preprocessing, model selection, model training, and model evaluation. The dataset used was obtained from Kaggle, consisting of 768 records with nine attributes, including pregnancies, glucose level, blood pressure, skin thickness, insulin, body mass index, diabetes pedigree function, age, and outcome classification. The models were trained using 80% training data and 20% testing data. The evaluation results showed that the Naive Bayes algorithm achieved an accuracy of 79%, while the CatBoost algorithm obtained a higher accuracy of 81%. These findings indicate that CatBoost provides better classification performance compared to Naive Bayes for diabetes diagnosis in this dataset. Therefore, CatBoost can be considered a more effective machine learning approach to support diabetes prediction. Future research is recommended to utilize larger datasets, optimize model parameters, and explore other advanced algorithms to improve diagnostic accuracy.

INTRODUCTION

Diabetes is a chronic metabolic disease characterized by increased blood glucose (or blood sugar) levels, which over time causes serious damage to the heart, blood vessels, eyes, kidneys, and nerves (Zuhri, 2024). Diabetes is known as the silent killer because this disease can affect all organs of the body and cause all kinds of complaints (Putra, Rihadini & Wijayanti Fuad, 2024)

Diabetes is a condition in which the body cannot produce the hormone insulin as needed or the body cannot optimally utilize the insulin produced, there is a spike in blood sugar levels that exceed normal. Diabetes is a chronic hyperglycemia condition accompanied by various metabolic disorders due to hormonal disorders (Esti, Ludiana & Tri, 2023). Hypoglycemia occurs when blood glucose drops below 70mg/dl (Wabula, 2021). Meanwhile, hyperglycemia occurs as a result of a decrease in insulin secretion, either completely or partially (Yeanneke & Jelita, 2023). There are many factors that affect people suffering from diabetes, some of which are high blood pressure, excess sugar levels, weight, history of diabetes hereditary, age, number of pregnancies, thickness of skin folds, and the amount of insulin levels in the body (Cahyani, 2022)

According to World Health Organization data 2021, Diabetes is among the 10 highest causes of death in Indonesia (World Health Organization, 2021; Sun et al., 2022), while International Diabetes Federation (IDF) data found that the number of diabetics in 2021 in Indonesia has increased rapidly in the last ten years (International Diabetes Federation, 2021; Saeedi et al., 2019). This number is estimated to reach 28.57 million by 2045 or 47% greater than the number of 19.47 million in 2021. From these data, early detection and accurate diagnosis are very important in preventing further complications of this disease (Rachman, Bhatara, & Hendryanny, 2021; American Diabetes Association, 2024; Tinajero & Malik, 2021).

With the development of information technology, machine learning can be used to help diagnose diseases (Topol, 2019; Esteva et al., 2019). Machine learning can help doctors analyze medical data and make predictions more accurate (Muhammad, 2023; Rajpurkar et al., 2022). Machine learning is a system that can learn to make its own decisions without having to be programmed repeatedly by humans so that computers can become smarter and learn from their experiences with data (Goodfellow, Bengio, & Courville, 2023). Based on the learning technique, supervised learning can be distinguished by using labeled datasets (Wijoyo, 2024; Géron, 2022). There are various algorithms used in the diagnosis process (Junaidi, 2024; Kaur, Kumari, & Kumar, 2023). So knowing which algorithms have higher accuracy in building models is very important.

A number of previous studies have also been conducted to analyze the use of machine learning algorithms in diagnosing diabetes with various algorithms (Tan et al., 2023; Liu et al., 2023; Mackenzie, Sainsbury, & Wake, 2024; Alanazi et al., 2023; Oikonomou & Khera, 2023). Research by Achmad (2020) shows that the Naive Bayes algorithm with a total of 520 lines of dataset is able to achieve an accuracy level of 90.20% in classifying diabetes. Another study conducted by Erlin (2022) showed that the logistic regression algorithm with a dataset of 768 lines was able to achieve an accuracy rate of 82%. Meanwhile, research by Mulyani (2025) compared the algorithm performance between K-Nearest Neighbors and the Support Vector Machine with a lot of data of 768 lines was able to achieve an accuracy level in the algorithm of 83.33%, while the Support Vector Machine algorithm was able to reach 81.67%. In another study conducted by Akbar (2025), comparing the algorithm performance between Support Vector Machine and Logistic Regression, it was able to obtain an accuracy level in the Support Vector Machine algorithm of 77.24% and Logistic Regression of 79.31%. Meanwhile, a study by Osama (2025) who compared the performance of the Naive Bayes and KNN algorithms by having 768 lines of data managed to get an accuracy level in the Naive Bayes algorithm of 76.6% and an accuracy level in the K-Nearest Neighbors algorithm of 79%.

This study will focus on a comparative analysis of the accuracy level of machine learning algorithms, namely CatBoost and Naive Bayes in diagnosing diabetes. Both algorithms were chosen because they have their own advantages, CatBoost excels in handling category data without requiring extensive pre-processing and has good performance on unbalanced data and compared to other methods, CatBoost has the ability to reduce overfitting through gradient-based boosting algorithms and faster processing on large datasets

(Mohamad, Suhartono&Totok, 2025), while Naive bayes was chosen for its ability to handle many textual features and perform classification quickly (Tri, Nana & Willy, 2024). This research is expected to make a significant contribution in selecting the best methods that can be applied to improve the accuracy of diagnosis and accelerate decision-making in the treatment of diabetic patients.

METHOD

In this study, the research process will go through several systematic stages such as the drawing of the research scheme below.

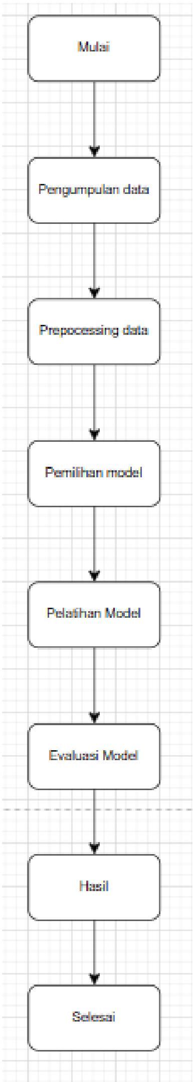

Figure 1. Research Scheme

Data collection

The dataset used in this study is sourced from https://www.kaggle.com This data is in the form of a file with the csv extension has a total of 768 records and 9 columns consisting of (Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome) this data will be used as data mining material, data definition will be carried out through data visualization to better understand the content of the dataset used.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
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

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Figure 2. Dataset Contents

Preprocessing data

Data preprocessing is a fairly important process before the process of selecting models and training moodel at this stage aims to ensure that the data that will be used is clean data, so that there is no data that does not have a value or commonly called missing values, and also deletes data that contains duplicates, at this stage also a data separation process is carried out, namely data testing and also data training with this process, it is expected that The training of the model becomes more accurate

Model Selection

In this study, the selection of models was carried out because it had the purpose of comparing the performance of several machine learning algorithms in detecting diabetes, as for some of the models chosen, namely Naive Bayes and CatBoost, the two algorithms were chosen because they are algorithms for classifying data and the selection of these algorithms is based on different approaches in processing data and determining classification, so that a more effective analysis can be obtained on each - Each algorithm in detecting diabetes

Model Training

At this stage, the two algorithms that have been selected are trained using 614 training data. The training data was used to form a prediction model by finding patterns of relationships between input variables and classification targets. The features of the dataset are defined as x_train, while labels or targets are defined as y_train. The goal of this process is for the model to learn from the data and be able to make predictions on the test data accurately.

Model Evaluation

At this stage, the performance measurement of the model that has been created using the confusion matrix is carried out. Using a confusion matrix, the performance or correctness of the diabetes diagnosis of the created model can be calculated. The results of the model process in the confusion matrix are True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Here is an example of implementation in the case of diabetes diagnosis.

l True Positive (TP) is positive data that is predicted to be true.
l True Negative (TN) is negative data that is predicted to be true.
l False Positive (FP) is negative data but is predicted to be positive data.
l False Negative (FN) is positive data but is predicted to be negative data.

Results

At this stage after the training and evaluation process of the model has been carried out, the last stage is the result of the model that has been built. This result is an overview of the performance of the model that has been determined, At this stage, the results displayed are the basis for assessing whether the model is already working well or still needs improvement. Then the results of a good level of accuracy will determine whether a person has diabetes or not.

RESULTS AND DISCUSSION

Data collection results

The data used in this study was obtained from public datasets available on the kaglle platform. This dataset is widely used in diabetes-related research has a total of 768 records and 9 columns consisting of (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome) data collection results, it is known that there are 268 patient data with positive diabetes diagnosis results and 500 diabetic negative patient data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
```

Figure 3. Data collection results

Data prepocessing results

The first step in creating a model to assess the level of accuracy in this study after obtaining data is to process the data or called prepocessing the data obtained. It is a stage to prepare data so that the data can be used in the model training process. The data used is data sourced from Kaggle with a total of 768 data with 9 abel data features used are the "Outcome" column which shows whether a person has diabetes or not. In the process of preposing the data, it is done by separating the feature as (x) and the label as (y), then the data is divided into data testing and data training. Training data is used as data that trains models. Meanwhile, the tesing data is used to evaluate the performance of the model resulting from this stage as follows.

```
Jumlah data training: 614
Jumlah data testing: 154
```

Figure 4. Data prepocessing results

Model Selection Results

After the data collection and prepocesing process, the next stage is the selection of the model to be used in this process, in this study, two algorithms were selected, namely naive bayes and catboost.

Model Training results

The model training stage is carried out after the algorithm is selected and the data is divided into two parts, namely training data and test data. In this study, data sharing was carried out with a ratio of 80% training data and 20% test data.

```
# Latih model
naive_bayes_classifier.fit(X_train, Y_train)
catboost_classifier.fit(X_train, Y_train)

✓ 2.4s
<catboost.core.CatBoostClassifier at 0x226d417fdd0>
```

Figure 5. Model Training Results

Results of model evaluation

After selecting and training the models, namely naive bayes and cat boost, the model evaluation was carried out on the algorithmic model using the confusion matrix method. So the results of the algorithm evaluation are obtained, namely

```
Model: Naive Bayes
              precision    recall  f1-score   support

           0       0.81      0.88      0.84       100
           1       0.73      0.61      0.67        54

    accuracy                           0.79       154
   macro avg       0.77      0.75      0.75       154
weighted avg       0.78      0.79      0.78       154


Model: CatBoost
              precision    recall  f1-score   support

           0       0.81      0.92      0.86       100
           1       0.80      0.59      0.68        54

    accuracy                           0.81       154
   macro avg       0.80      0.76      0.77       154
weighted avg       0.80      0.81      0.80       154
```

Figure 6. Model Evaluation Results

Based on the results of his evaluation, the model that uses the naive bayes algorithm gets an accuracy rate of 79% while the catboost algorithm gets an accuracy rate of 81%

Model Prediction Results

In this section, after obtaining the results of the model evaluation, it is tested by input data to get the results of the model's prediction using the model that gets the best evaluation results, namely catboost the input value:

Pregnancies = 0, Glucose = 84, BloodPressure = 82, Skin Thickness = 31, Insulin = 125, BMI = 38.2, Diabetes Pedigree Function = 0.233, Age = 23

After the input was processed by the model, the prediction results were obtained, namely: The patient did not have diabetes

CONCLUSION

This study concludes that machine learning algorithms can be utilized as an approach to support diabetes diagnosis through classification models based on patient health-related data. The comparison between the Naive Bayes and CatBoost algorithms using the confusion matrix evaluation method shows that both models are capable of performing diabetes classification, but with different levels of accuracy. Based on the evaluation results, the Naive Bayes algorithm achieved an accuracy rate of 79%, while the CatBoost algorithm obtained a higher accuracy rate of 81%. These results indicate that CatBoost provides better predictive performance compared to Naive Bayes in classifying diabetes conditions using the dataset consisting of 768 records with nine features, including glucose level, blood pressure, body mass index, insulin level, age, and other related factors. Therefore, CatBoost can be considered a more optimal machine learning model for diabetes prediction in this study because it produces higher classification accuracy and better prediction capability.

For future research, further studies are recommended to improve model performance by using larger and more diverse medical datasets from different populations to increase the generalizability of the prediction model. Future researchers may also explore other machine learning and deep learning algorithms, such as Random Forest, Support Vector Machine, XGBoost, or Artificial Neural Networks, and compare their performance with CatBoost and Naive Bayes. In addition, optimization techniques such as hyperparameter tuning, feature selection, and handling of class imbalance can be applied to improve accuracy and reduce prediction errors. The development of an intelligent diagnostic system integrated with healthcare platforms may also become a potential direction to support medical professionals in early diabetes detection and clinical decision-making.

.

REFERENCES

American Diabetes Association. (2024). *Standards of care in diabetes—2024*. *Diabetes Care, 47*(Supplement_1), S1–S350. https://doi.org/10.2337/dc24-SINT

Alanazi, N., Alruwaili, Y., Alazmi, A. S. M., Alazmi, A. M. M., Alanazi, M., & Alruwaili, W. (2023). A systematic review of machine learning and artificial intelligence for diabetes care. *Journal of Health Informatics in Developing Countries, 17*(1).

Des Suryani, D., Nasution, S. D., Fathurrahman, H., Febriyanti, E., Arifin, H., Nurallita, N. S., & Siregar, P. P. (2024). *Buku ajar kedokteran Islam*. UMSU Press.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Goodfellow, I., Bengio, Y., & Courville, A. (2023). *Deep learning* (Updated ed.). MIT Press.

Indriyani, E., Ludiana, & Dewi, T. K. (2024). Analisis perbandingan algoritma klasifikasi untuk identifikasi diabetes dengan menggunakan metode Random Forest dan Naïve Bayes. *Jurnal Inovasi Teknologi dan Sains (JINTEKS).* https://jurnal.uts.ac.id/index.php/JINTEKS/article/view/5146

International Diabetes Federation. (2021). *IDF diabetes atlas* (10th ed.). International Diabetes Federation. https://diabetesatlas.org/

Junaidi. (2024). *Buku ajar machine learning*. Sonpedia.

Kaur, P., Kumari, S., & Kumar, A. (2023). Machine learning techniques for disease prediction: A systematic review. *Healthcare Analytics, 3*, 100186. https://doi.org/10.1016/j.health.2023.100186

Liu, K., Li, L., Ma, Y., Jiang, J., Liu, Z., Ye, Z., Liu, S., Pu, C., Chen, C., & Wan, Y. (2023). Machine learning models for blood glucose level prediction in patients with diabetes mellitus: Systematic review and network meta-analysis. *JMIR Medical Informatics, 11*, e47833. https://doi.org/10.2196/47833

Mackenzie, S. C., Sainsbury, C. A. R., & Wake, D. J. (2024). Diabetes and artificial intelligence beyond the closed loop: A review of the landscape, promise and challenges. *Diabetologia*, *67*, 223–235. https://doi.org/10.1007/s00125-023-06038-8

Muhammad. (2023). *Dasar-dasar pembelajaran mesin*. Sada Kurnia Pustaka.

Oikonomou, E. K., & Khera, R. (2023). Machine learning in precision diabetes care and cardiovascular risk prediction. *Cardiovascular Diabetology, 22*, Article 259. https://doi.org/10.1186/s12933-023-01985-3

Rachman, S. K., Bhatara, T., & Hendryanny, E. (2021). Scoping review: Hubungan kontrol glikemik (HbA1C), durasi penyakit, dan profil lipid pada pasien diabetes melitus tipe II dengan kejadian neuropati diabetik. *Journal of Health and Science Integration, 3*(2), 207–214.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, *28*(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A. A., Ogurtsova, K., Shaw, J. E., Bright, D., & Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas (9th ed.). *Diabetes Research and Clinical Practice, 157*, 107843. https://doi.org/10.1016/j.diabres.2019.107843

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C. N., Mbanya, J. C., Pavkov, M. E., Ramachandaran, A., Wild, S. H., James, S., Herman, W. H., Zhang, P., Bommer, C., Kuo, S., Boyko, E. J., & Magliano, D. J. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, *183*, 109119. https://doi.org/10.1016/j.diabres.2021.109119

Tan, K. R., Seng, J. J. B., Kwan, Y. H., Chen, Y. J., Zainudin, S. B., Loh, D. H. F., Liu, N., & Low, L. L. (2023). Evaluation of machine learning methods developed for prediction of diabetes complications: A systematic review. *Journal of Diabetes Science and Technology*, *17*(2), 474–489. https://doi.org/10.1177/19322968211056917

Tinajero, M. G., & Malik, V. S. (2021). An update on the epidemiology of type 2 diabetes: A global perspective. *Endocrinology and Metabolism Clinics of North America, 50*(3), 337–355. https://doi.org/10.1016/j.ecl.2021.05.013

Tinungki, Y. L., & Hinonaung, J. S. H. (2023). *Deteksi dini penyakit diabetes melitus (DM) dan obat tradisional DM pada lansia di Kabupaten Kepulauan Sangihe*. PT Sonpedia Publishing Indonesia.

Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Tri Arifqi, Suarna, N., & Prihartono, W. (2024). Penggunaan Naïve Bayes dalam menganalisis sentimen ulasan aplikasi McDonald's di Indonesia. *JATI*. https://ejournal.itn.ac.id/index.php/jati/article/download/8740/5313/

World Health Organization. (2021). *Global health estimates 2021: Leading causes of death*. World Health Organization. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates

Zuhri, M. R., Kusrini, & Ariatmanto, D. (2024). Penerapan senam kaki diabetes melitus terhadap kadar glukosa darah pada penderita diabetes mellitus di Puskesmas Yosomulyo. *Jurnal Wacana Cendekia*. https://jurnal.akperdharmawacana.ac.id/index.php/JWC/article/download/466/300