

*Research Article*

# Comparison of Classification Algorithm Performance for Diabetes Prediction Using Orange Data Mining

Aditya Nugraha Yesa<sup>1</sup>, Hafiz Aryan Siregar<sup>2</sup>, Muhammad Zacky Raditya<sup>3</sup>, Inggih Permana<sup>4</sup>

<sup>1</sup> Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia, 12150312429@students.uin-suska.ac.id

<sup>2</sup> Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia, 12150310904@students.uin-suska.ac.id

<sup>3</sup> Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia, 12150311608@students.uin-suska.ac.id

<sup>4</sup> Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia, inggihpermana@uin-suska.ac.id

Correspondence should be addressed to Aditya Nugraha Yesa; 12150312429@studets.uin-suska.ac.id

Received 18 October 2023; Accepted 10 December 2023; Published 31 December 2023

Copyright © 2023 Indonesian Journal of Data and Science. This scholarly piece is accessible under the Creative Commons Attribution Non-commercial License, permitting dissemination and modification, conditional upon non-commercial use and due citation

## Abstract:

Diabetes is a disease that contributes to a relatively high mortality rate. The human death rate due to diabetes is a widespread issue globally. The primary goal of this research is to predict individuals suffering from diabetes using a publicly available dataset from the UCI Repository with the Diabetes Disease dataset. To obtain the best classification algorithm, a comparison is made among three algorithms: KNN, Naive Bayes, and Random Forest, commonly used for predicting diabetes. The comparison results indicate that the Random Forest algorithm is the appropriate and accurate algorithm for predicting individuals with diabetes, with an accuracy rate of 97%.

**Keywords:** Data Mining, Diabetes, KNN, Naive Bayes, Random Forest.

**Dataset Link:** Diabetes

## 1. Introduction

Diabetes is a chronic disease caused by a disruption in the secretion of endogenous insulin. Diabetes is divided into type 1 diabetes and insulin-dependent diabetes mellitus (IDDM) and type 2 diabetes, known as non-insulin-dependent diabetes [1]. Diabetes is a chronic disease characterized by higher than normal blood sugar levels. If not managed properly, diabetes can lead to various complications in organs such as the eyes, kidneys, heart, blood vessels, and nerves, posing a threat to life and affecting the quality of life. Complications can be either acute or chronic. Acute complications involve sudden increases or decreases in blood sugar, while chronic complications are the long-term effects of elevated blood sugar. These complications can lead to a reduced life expectancy, disability, and increased financial burden on clients and their families. Diabetes is a lifelong condition and has a significant impact on quality of life if not well-managed. Complications of diabetes, such as chronic diabetic ulcers, disrupt the quality of life [2].

The International Diabetes Federation (IDF) reported a global prevalence of diabetes at 1.9%, making diabetes the seventh leading cause of death worldwide, with 382 million deaths globally in 2013. Those affected by diabetes constitute 95% of the world's population with type 2 diabetes. The prevalence of type 2 diabetes is 85-90%. According to the WHO report in 2003, only 50% of diabetic patients in developed countries adhere to prescribed treatment. Complications can occur if diabetes is not controlled, impacting both quality of life and the economy. In 2013, the prevalence of diabetes in Indonesia was 2.1%, a higher value compared to 2007.

The impact of diabetes includes physical, psychological, and social aspects on patients. The physical impact involves changes in the body's ability to control blood sugar, increasing the risk of both microvascular and macrovascular

complications. Psychological impact includes feelings of anxiety related to the patient's health condition. Social impact involves the need for support to enable patients to manage the disease [3].

## Literature Review

### A. Data Mining

Data mining is a process that utilizes statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from various large databases. The term data mining essentially represents a discipline whose primary objective is to discover, dig, or mine knowledge from the data or information available to us. Data mining is often also referred to as Knowledge Discovery in Databases (KDD). KDD is an activity that involves the collection and utilization of historical data to uncover regularities, patterns, or relationships within large-sized datasets [4].

### B. K-Nearest Neighbor

K-Nearest Neighbor (K-NN) belongs to the instance-based learning group. This algorithm is also a type of lazy learning technique. K-NN is performed by identifying a group of  $k$  objects in the training data that are most similar to the object in the new or testing data. A classification system is needed for K-NN to function as a system capable of seeking information. For example, in a case where a solution for a new patient's problem is sought using solutions from past patients, K-NN can be applied. [5].

### C. Naive Bayes Classifier (NBC)

Naive Bayes Classifier (NBC) is one of the algorithms in data mining that applies Bayesian theory to classification. The Bayes decision theorem is a fundamental statistical approach in pattern recognition. Naive Bayes is based on the simplifying assumption that attribute values are conditionally independent given the output value. In other words, given the output value, the probability of observing them together is the product of the individual probabilities. [4]

### D. Random Forest

Random Forest is one of the data mining methods used for classification. The Random Forest method itself consists of a classification tree with a large number of connections to seek maximum accuracy [6], [7]. Random Forest can also be described as a combination of individual trees used in a single model [8]. Random Forest is a classifier that consists of trees  $\{h(x, k), k = 1, \dots\}$  where  $k$  is a randomly distributed vector independently, and each component tree selects the most popular class with the input value  $x$  [9].

## 2. Method:

The stages of this research are explained as follows:

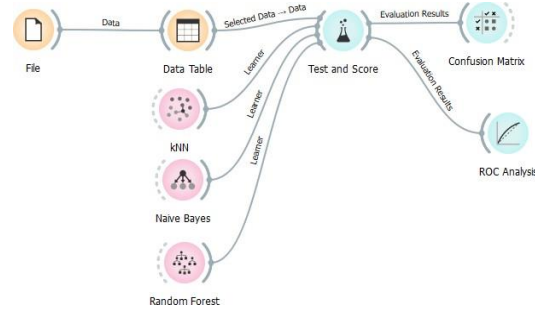
1. Dataset Collection: The dataset to be tested in this study is the diabetes disease dataset from the UCI Repository.
2. Literature Review: This literature review involves gathering articles and research journals related to the subject of the paper, specifically focusing on predicting diabetes.
3. Research Model Selection: In this paper, the research will utilize the best-performing models as determined by previous researchers. These models include the KNN, Naive Bayes algorithm, and Random Forest.
4. Learner or Training from the Dataset: The selected models will undergo training using the collected dataset.
5. Prediction Evaluation: The predictions will be evaluated using metrics such as AUC (Area Under the Curve), CA (Classification Accuracy), F1 score, Precision, Recall, Confusion Matrix, and ROC(Receiver Operating Characteristic) analysis [10]–[12].

Analysis and Conclusion: This stage involves analyzing the results of the conducted tests and drawing conclusions based on the analysis **Figure 1**.



**Figure 1.** Data Mining Process

As for the data mining process, here we use the data mining tool, namely Orange Python, with the following workflow as **Figure 2**.



**Figure 2.** Diabetes Prediction Process Workflow

Some performance metrics that are common and often used are as follows.

### 1) Accuracy

In the context of classification models, accuracy is the main indicator to evaluate the extent to which the model is able to classify data correctly [13]–[15]. Accuracy is measured as a comparison between the number of true positive data and true negative data with the total existing data. Mathematically, accuracy can be illustrated by Equation (1), which reflects the degree of closeness of the model's predicted values to the actual values of the observed data. Therefore, the higher the accuracy value, the better the model's performance in producing predictions that match the actual situation. Understanding and measuring good accuracy is key in evaluating the reliability and effectiveness of classification models in the context of data processing.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

### 2) Precision

In evaluating model performance, precision plays a crucial role as an indicator of the extent to which the model is able to provide accurate predictions for the requested data [13]–[15]. Precision can be interpreted as the degree of accuracy of positive predictions by a model, measured through the comparison between the number of true positive predictions and the overall positive predicted outcomes. In other words, precision describes how many of all positive classes are predicted correctly by the model. Equation (2) reflects the precision value, which is an important parameter in evaluating the reliability of the model in providing results in accordance with the required criteria. High precision measurements demonstrate the model's ability to minimize errors in identifying and classifying positive data, strengthening confidence in the interpretation of predictions provided by the model.

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

### 3) Recall

In model performance analysis, recall or sensitivity is an important parameter that reflects the model's success in detecting relevant information [13]–[15]. Recall can be interpreted as the ratio of true positive predictions divided by the overall true positive data, highlighting the extent to which the model is able to identify all instances of true positive data. In other words, recall describes the model's ability to obtain relevant information and avoid the error of ignoring positive data. Equation (3) provides the recall value, which is a vital benchmark in measuring the model's sensitivity to positive cases. A high recall value indicates that the model is able to effectively capture most of the existing positive data, providing confidence that the model can detect significant information with high

accuracy.

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

#### 4) Specificity

In the context of model evaluation, specificity describes the level of accuracy in predicting negatives, measured as the ratio of true negative predictions divided by the total number of true negative data. In other words, specificity reflects the model's ability to identify and minimize errors in classifying data as negative. Equation (4) provides the value of specificity, which is an indicator of the model's performance in correctly predicting negative cases. A high specificity value indicates that the model has a good ability to avoid errors in predicting negative data, thus providing confidence that the model can provide accurate and reliable results in classifying various cases. Specificity evaluation is important in understanding the extent to which the model is able to provide precise predictions on existing negative data.

$$Specificity = \frac{TN}{(TN + F)} \quad (4)$$

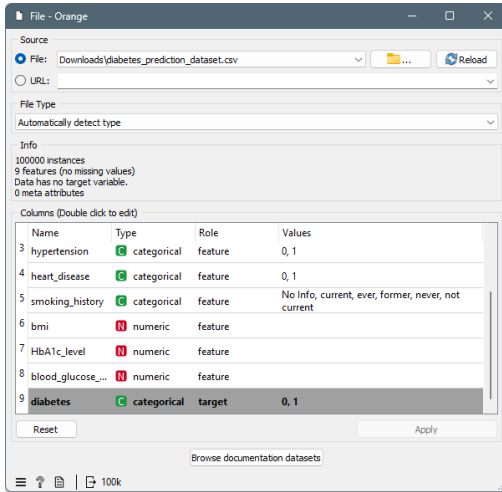
#### 5) F1 Score

The F1 score is an important parameter in evaluating model performance, summarizing the average comparison between precision and recall by providing balanced weights [13]–[15]. The F1 score can be interpreted as a measure of the overall success of the model in classifying positive data, describing the balance between precision and recall in the system. Equation (5) shows the value of the F1 score, and this score achieves its best performance when there is an optimal balance between precision and recall. Therefore, the F1 score provides a comprehensive picture of the model's effectiveness in handling both aspects in a balanced manner. A high F1 score indicates that the model is able to provide predictions with high accuracy while minimizing errors in ignoring or misidentifying positive data, creating a reliable and balanced classification system.

$$F - measure = \frac{2(precision \times recall)}{(precision + recall)} \quad (5)$$

### 3. Result and Discussion:

This study utilizes a classification algorithm, employing Naive Bayes, Random Forest, and KNN algorithms, to develop a prediction system used for analyzing and predicting the likelihood of Diabetes. The Diabetes Disease dataset, publicly available on Kaggle, is used to assess the performance of these classification algorithms. This dataset comprises nine attributes, consisting of both numerical and nominal data. The absence of diabetes and the presence, show in Figure 3.



Name	Type	Role	Values
3 hypertension	categorical	feature	0, 1
4 heart_disease	categorical	feature	0, 1
5 smoking_history	categorical	feature	No info, current, ever, former, never, not current
6 bmi	numeric	feature	
7 HbA1c_level	numeric	feature	
8 blood_glucose...	numeric	feature	
9 diabetes	categorical	target	0, 1

Figure 3. Diabetes Dataset

This system consists of several steps. Firstly, data is preprocessed, removing both redundancy and noise, and then classified using KNN, Random Forest, and Naive Bayes algorithms. Subsequently, validation is performed to assess the performance of each algorithm, including AUC, CA, F1, Precision, and Recall. This is done to determine the algorithm with the best accuracy. The experiment is conducted using the Orange Python application, which has a dataset consisting of fourteen attributes comprising both numerical and nominal data. KNN, Naive Bayes, and Random Forest are the classification algorithms under examination. The following are the confusion matrices for Naive Bayes, Random Forest, and Random Forest. The values in the confusion matrix calculate the accuracy of data mining concepts or decision support systems, serving to analyze whether the classifier is proficient in recognizing various classes.

		Predicted		Σ
		0	1	
Actual	0	91301	199	91500
	1	2693	5807	8500
Σ		93994	6006	100000

**Figure 4.** Confusion Matrix Random Forest

**Figure 4** is Random Forest algorithm exhibits strong predictive performance, accurately identifying 91,301 instances and 5,807 instances with precision, while making 2,693 correct negative predictions and 199 incorrect ones out of a total of 100,000 test data points. Consequently, the algorithm achieves an impressive accuracy rate of 97.1%. This high level of accuracy underscores the effectiveness of the Random Forest algorithm in making precise predictions based on the given dataset.

		Predicted		Σ
		0	1	
Actual	0	88404	3096	91500
	1	4238	4262	8500
Σ		92642	7358	100000

**Figure 5.** Confusion matrix Naive Bayes

**Figure 5** is Naive Bayes algorithm demonstrates solid predictive capabilities, accurately predicting 88,404 instances and 4,262 instances with precision. It also makes 4,238 correct negative predictions but has 3,096 incorrect ones out of a total of 100,000 test data points. As a result, the algorithm achieves a commendable accuracy rate of 93%. This accuracy underscores the reliability of the Naive Bayes algorithm in making precise predictions based on the provided dataset.

		Predicted		Σ
		0	1	
Actual	0	91309	191	91500
	1	2659	5841	8500
Σ		93968	6032	100000

**Figure 6.** Confusion matrix KNN

**Figure 6** is KNN algorithm demonstrates robust predictive performance, accurately forecasting 91,309 instances and 5,841 instances with precision. It also makes 2,659 correct negative predictions, with 191 incorrect predictions, out of a total of 100,000 test data points. Consequently, the algorithm achieves an impressive accuracy rate of 97%. This high accuracy underscores the efficacy of the KNN algorithm in making precise predictions based on the given

dataset.

#### 4. Conclusion:

From the stages carried out, this research reached a conclusion, namely to predict individuals who suffer from diabetes using the diabetes dataset that is publicly available from the UCI Repository. Three classification algorithms, namely KNN, Naive Bayes, and Random Forest, were evaluated to determine the best algorithm. The comparison results show that the Random Forest algorithm has the highest level of accuracy, reaching 97%, so it is considered a precise and accurate algorithm for predicting individuals.

who suffer from diabetes. These conclusions may serve as a basis for the development of further diagnostic approaches in the management of diabetes.

#### References:

- [1] [Longmore, D. K., Barr, E. L., Wilson, A. N., Barzi, F., Kirkwood, M., Simmonds, A., ... & Maple-Brown, L. J. \(2020\). "Associations of gestational diabetes and type 2 diabetes during pregnancy with breastfeeding at hospital discharge and up to 6 months: the PANDORA study." \*Diabetologia\*, 63,2571-2581.](#)
- [2] [A. R. P. Abimanyu et al., "Pengaruh Terapi Pada Penderita Diabetes Mellitus Sebagai Penurunan Ka dar Gula Darah: Review Artikel," \*Innovative: Journal Of Social Science Research\*, vol. 3, no. 2, pp. 8931-8949, 2023.](#)
- [3] [Maryati, Y., Alifiar, I., Nurfatwa, M., Nofianti, T., & Rahayuningsih, N. \(2019, July\). "Antlion \(\*Myrmeleon sp.\*\) Infusion as Antidiabetic in Dexamethasone Induced Mice." In \*Journal of Physics: Conference Series\*, vol. 1179, No. 1, p. 012177. IOP Publishing.](#)
- [4] [M. Ridwan, H. Suyono, dan M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," \*Jurnal EECCIS \(Electrics, Electronics, Communications, Controls, Informatics, Systems\)\*, vol. 7, no. 1, pp. 59-64, 2013.](#)
- [5] [D. Cahyanti, A. Rahmayani, dan S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," \*Indonesian Journal of Data and Science\*, vol. 1, no. 2, pp. 39-43, 2020.](#)
- [6] [M. Fithratullah, "Representation of Korean Values Sustainability in American Remake Movies," \*Teknosastik\*, vol. 19, no. 1, p. 60, 2021. \[Online\]. Available: \[<https://doi.org/10.33365/ts.v19i1.874>\]](#)
- [7] [A. Wantoro, A. Syarif, K. N. Berawi, K. Muludi, S. R. Sulistiyanti, U. Lampung, I. Komputer, U. Lampung, K. Masyarakat, F. Kedokteran, U. Lampung, T. Elektro, F. Teknik, U. Lampung, U.Lampung, G. Meneng, dan B. Lampung, "Metode Profile Matching Pada Sistem Pakar Medis Untuk," vol. 15, no. 2, pp. 134–145, 2021.](#)
- [8] [F. Dharma, A. Noviana, M. Tahir, dan N. Hendrastuty, "Prediction of Indonesian Inflation Rate Using Regression Model Based on Genetic Algorithms," \*J. Informatics Optim. Nanotechnol. Mater.\*, vol. 5, no. 1, pp. 45–52, 2020. \[Online\]. Available:\[<https://doi.org/10.15575/join>\]](#)
- [9] [E. D. Listiono, A. Surahman, dan S. Sintaro, "Ensiklopedia Istilah Geografi Menggunakan Metode Sequential Search Berbasis Android Studi Kasus:Sma Teladan Way Jepara Lampung Timur," \*Jurnal Teknologi Dan Sistem Informasi\*, vol.2, no. 1, pp. 35–42, 2021](#)
- [10] [H. Azis, F. Fattah, and P. Putri, "Performa Klasifikasi K-NN dan Cross-validation pada Data Pasien Pengidap Penyakit Jantung," \*ILKOM Jurnal Ilmiah\*, vol. 12, no. 2, pp. 81–86, 2020, \[Online\]. Available: <file:///Users/kbh/Downloads/507-2012-5-PB.pdf>](#)
- [11] [D. Mahapatra, "Handwritten Character Recognition Using KNN and SVM Based Classifier over Feature Vector from Autoencoder," \*Communications in Computer and Information Science\*, vol. 1240, pp. 304–317, 2020, doi: 10.1007/978-981-15-6315-7\\_25](#)

- [12] [H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," \*Techno.Com\*, vol. 19, no. 3, 2020, \[Online\]. Available: \[file:///Users/kbh/Library/Application Support/Mendeley Desktop/Downloaded/Azis, Admojo, Susanti - 2020 - Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah.pdf\]\(file:///Users/kbh/Library/Application%20Support/Mendeley%20Desktop/Downloaded/Azis,%20Admojo,%20Susanti%20-%20Analisis%20Perbandingan%20Performa%20Metode%20Klasifikasi%20pada%20Dataset%20Multiclass%20Citra%20Busur%20Panaah.pdf\)](#)
- [13] [Y. Jusman, "Machine Learnings of Dental Caries Images based on Hu Moment Invariants Features," \*Proceedings - 2021 International Seminar on Application for Technology of Information and Communication: IT Opportunities and Creativities for Digital Innovation and Communication within Global Pandemic, iSemantic 2021\*, pp. 296–299, 2021, doi: 10.1109/iSemantic52711.2021.9573208](#)
- [14] [Y. Jusman, "Classification System for Leukemia Cell Images based on Hu Moment Invariants and Support Vector Machines," \*Proceedings - 2021 11th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2021\*, pp. 137–141, 2021, doi: 10.1109/ICCSCE52189.2021.9530974](#)
- [15] [X. Ye, "Prediction of Breast Cancer of Women Based on Support Vector Machines," \*ACM International Conference Proceeding Series\*, pp. 780–784, 2020, doi: 10.1145/3443467.3443853](#)