Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 1 Iss. 2 August 2024, pp: 54-60
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Comparison of Machine Learning Algorithms in Diabetes Risk Classification

Zairy Cindy Dwinnie[1*], Zaira Cindya Dwynne[2],
Mohammed Jahidul Islam[3], Noviarni[4]

[1,2]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[3]Department of Business, BSc Hon's in Business Studies,
Niels Brock Copenhagen Business College, Denmark
[4]Department of Mathematics Education, Faculty of Tarbiyah and Education,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[1,2]Puzzle Research Data Technology (Predatech), Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

E-Mail: [1]12050324081@students.uin-suska.ac.id, [2]12050320480@students.uin-suska.ac.id,
[3]jahidul.dn@gmail.com, noviarni@uin-suska.ac.id

Received Dec 28th 2023; Revised Jun 13th 2024; Accepted Jul 06th 2024
Corresponding Author: Zairy Cindy Dwinnie

Abstract

Diabetes is a disease in which blood sugar levels are excessive without insulin control so that body functions do not function normally. Diabetes is also a disease that many people suffer from and is one of the main causes of death throughout the world. For this reason, we need to know the factors that are indicators of someone suffering from diabetes. This research compares the Decision Tree, Logistic Regression, and K-Nearest Neighbors algorithms with accuracy and Confusion Matrix parameters to determine diabetes sufferers in 520 data with the main indicator attributes supporting diabetes. From the test results of the three algorithms, the Decision Tree and K-Nearest Neighbors models have the highest accuracy of 86%. The Logistic Regression Algorithm has a fairly good accuracy of 83%.

Keyword: Confusion Matrix, Decision Tree, Diabetes, K-Nearest Neighbors, Logistic Regression

1. INTRODUCTION

Diabetes is one of the main causes of death worldwide in the non-communicable disease category. Data from the World Health Organization (WHO) reports that the number of people suffering from diabetes has increased drastically, rising from 108 million in 1980 to 422 million in 2014. In that decade, the prevalence of diabetes in the adult population (over 18 years) also increased from less than 5% to 8.5%. Diabetes is a disease in which excessive sugar levels in the blood without insulin control cause the body's functions to not function normally [1]. Insulin is a hormone produced by the pancreas, which plays a key role in allowing glucose from food to enter the body's blood cells, where the glucose is then used to produce energy [2]. As time goes on, the risk of complications in diabetes increases, which include cardiovascular problems, stroke, damage to blood vessels, vision, hearing, skin, kidneys, and feet, and can cause depression [3].

It is important to understand the factors that indicate diabetes in a person so that you can take anticipatory steps and undergo further examination. With technological developments, diabetes can be identified earlier through a Data Mining approach [4]. Data mining is a combination of computer science disciplines that aims to discover new patterns from large datasets. The method involves artificial intelligence, machine learning, statistics, and database systems. The advantages of data mining include efficient data processing, and turning past data into new knowledge. The use of data mining is not only limited to technology but also extends to the health sector, where it can be used to predict and diagnose disease with applicable methods [5]. Several methods are generally applied in data mining, such as regression, classification, clustering, association, and various data processing techniques [6].

In research conducted by Yum Thurfa, et al, 2021. This research classified fetal heart rate using seven machine learning algorithms. It was found that the best model was produced by the Random Forest algorithm with an accuracy of 94.5% [7]. Another research by Dewi and Dana 2017, classified clothing pattern selection

DOI: https://doi.org/10.57152/ijatis.v1i2.1141

using three classification algorithms. From the comparison results, it can be concluded that the decision tree has a higher level of accuracy than the naive Bayes and nearest neighbor algorithms, reaching 75.6% [8]. Previous research on the classification of Diabetes patients was carried out by Baiq and Intan in 2021, using the Decision Tree and Naïve Bayes algorithms. The results of this research are a comparison of the two algorithms for determining diabetes sufferers, where the greatest accuracy of 95% was produced by the Decision Tree algorithm model [5].

This research aims to identify the most effective and accurate method for predicting diabetes patients based on a patient dataset. By comparing the performance of these three algorithms, the research seeks to determine which algorithm provides the best results in terms of accuracy, recall, and precision. Additionally, this research is expected to provide insights into the relative effectiveness of K-Nearest Neighbors (K-NN), Decision Tree, and Logistic Regression in the classification of diabetes. The dataset processing and model evaluation will be carried out using the Python programming language on a text editor.

2. MATERIAL AND METHOD

The data used in this research comes from Kaggle, which is a platform for accessing datasets. The following can be seen in Figure 1.

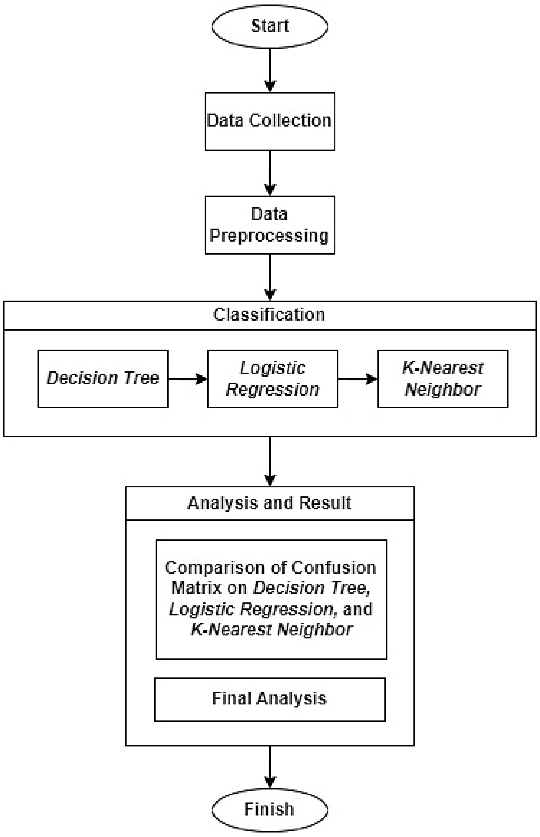


Figure 1. Research Methodology

2.1. Data Collection

The data used in this research is a type of Diabetes Resikp classification data sourced from the Kaggle website. Where the data consists of 17 attributes consisting of Age, Sex, and 15 Diabetes Indicators. Data attributes can be seen in the table 1.

Table 1. Data Attributes

Data Attributes	Information
Age	The age range of the individuals
Sex	Gender information
Polyuria	Presence of excessive urination
Polydipsia	Excessive thirst
Polyphagia	Excessive hunger
Genital Thrush	Presence of genital thrush
Visual Blurring	Blurring of vision
Itching	Presence of itching
Irritability	Display of irritability

Data Attributes	Information
Delayed Healing	Delayed wound healing
Partial Paresis	Partial loss of voluntary movement
Muscle Stiffness	Presence of muscle stiffness
Alopecia	Hair loss
Obesity	Excess body fat
Class	Positive or negative a person experiences certain health problems based on the factors above.

In this research, analysis is carried out by reviewing a set of data so that it can be understood and useful, to find unexpected relationships and make summaries in a different way than before, which is one of the definitions of data mining [9]. This research will explain the things that cause the risk of diabetes. It is a scientific discipline that uses machine learning, pattern recognition, statistics, databases, and visualization techniques to solve problems that arise when retrieving information from large databases [10].

2.2. Data Preprocessing

This stage is a data cleaning step, such as processing inconsistent data, cleaning data from noise, and removing duplicate data [11]. This stage is carried out after the data collection process. Preprocessing also fills data with empty data, copies data, checks for data discrepancies, cleans data, and corrects data errors [12]. In this study, we used clean data, so no cleaning was required. The data type for each indicator attribute is object (yes/no) and the conversion is carried out to an integer data type. The following table is the data that will be used in this research.

Table 2. Before Conversion

Sudden Weight Loss	Weakness	Visual Blurring	Itching	Irritability	Delayed Healing	Partial Paresis	Alopecia	Obesity	Class
No	Yes	No	Yes	No	Yes	No	Yes	Yes	Positive
No	Yes	Yes	No	No	No	Yes	Yes	No	Positive
No	Yes	No	Yes	No	Yes	No	Yes	No	Positive
Yes	Yes	No	Yes	No	Yes	No	No	No	Positive
…	…	…	…	…	…	…	…	…	…
No	No	No	No	No	No	No	No	No	No

Table 3. After Conversion

Sudden Weight Loss	Weakness	Visual Blurring	Itching	Irritability	Delayed Healing	Partial Paresis	Alopecia	Obesity	Class
0	1	0	1	0	1	0	1	1	1
0	1	1	0	0	0	1	1	0	1
0	1	0	1	0	1	0	1	0	1
1	1	0	1	0	1	0	0	0	1
…	…	…	…	…	…	…	…	…	…
0	0	0	0	0	0	0	0	0	0

In Table 3, it can be seen that the dataset transformation that has occurred is that all attributes have become numeric (integer). The data in the table above are the attributes used to support the target label classification. Where the labels in this classification are class attributes. This attribute consists of two classes, namely positive = 1 and negative = 0. The other attributes are classes yes = 1 and no = 0. The data used is 520 rows and 10 attributes including 1 attribute as a label and 9 attributes as the Diabetes indicator which is the most common. experienced (sudden weight loss weakness, visual blurring, Itching, Irritability, delayed healing, partial paresis, Alopecia, Obesity).

2.3. Classification

The process of searching for patterns (or features) that describe and differentiate data classes or concepts to predict object classes whose class labels are unknown [13]. Classification algorithms that are widely used are Decision/classification trees, Bayesian Classifiers/ Naïve Bayes classifiers, Neural Networks, Statistical Analysis, Genetic Algorithms, Rough sets, k-nearest neighbors, Rule-Based Methods, Memory based reasoning, and Support Vector Machines (SVM) [14].

2.4. Decision Trees

A Decision Tree is an algorithm that is commonly used for decision-making. This algorithm seeks solutions to problems by using criteria as nodes that are connected to form a tree structure [8]. Decision Tree or DT, is a Machine Learning algorithm that is often used to solve classification and regression tasks [15]. The model produced by this algorithm has the form of a tree, where the leaf nodes reflect the results of classification or regression, while the internal nodes reflect the evaluation of the attributes [16].

2.5. Logistic Regression

The logistic regression algorithm is mainly used for binary classification, where the category types can be 0 and 1, true or false, large or small [17]. The independent variable in logistic regression is categorical, and this difference differentiates it from multiple regression or other linear regressions [18]. In other words, logistic regression is suitable for situations where the dependent variable is binary or has two distinguishable categories [19].

$$P(Y = 1 \mid \mathrm{X}) = \frac{1}{1 + e^{-\mathrm{x}}} \tag{1}$$

Information:

$P(Y = 1 \mid \mathrm{X})$	: The probability that Y is equal to 1 (success) given the value X.
x	: Input variables
e	: Euler's number (2.71828...), the base of the natural logarithm.

2.6. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a basic and simple approach to object classification. This algorithm has the principle of classifying new data by considering attributes and training data samples [20][21]. K is a positive integer that is determined before running the algorithm [22]. This algorithm is superior when dealing with training data that contains a lot of noise, and its effectiveness is seen when used with large training data sets. However, it also has weaknesses such as the need to determine the value of the parameter k (number of nearest neighbors), difficulties in establishing optimal distance metrics, and challenges in selecting characteristics that produce the best results [23].

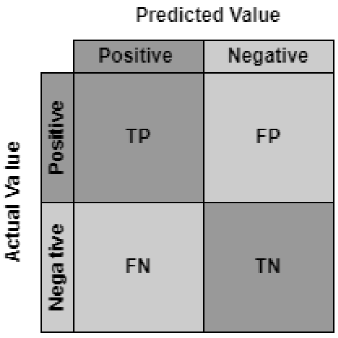
$$\mathrm{d}() = x_i, x_j \sqrt{\textstyle\sum_{r = 1}^{n}} (a_r(x_i) \; - \; (a_r(x_j))^2 \tag{2}$$

Information:

$\mathrm{d}() x_i, x_j$	: Euclidean Distance (Euclidean Distance)
(x_i)	: i-th record
(x_j)	: jth record
$(a \quad)$	: r-th data
i,j	: 1,2,3,….r

2.7. Confusion Matrix

Confusion matrix is a table that displays the number of test data that were classified correctly and the number of test data that were classified incorrectly [24]. For example, Table 4 is a binary classification confusion matrix that shows the distribution of correct and incorrect classification results. The confusion matrix includes four crucial values, such as True Positive (TP) and True Negative (TN), which reflect accurate prediction results, and False Positive (FP) and False Negative (FN), which indicate inaccurate predictions. The figure 2 is a diagrammatic illustration of the Confusion Matrix [7].



		Predicted Value	
		Positive	Negative
Actual Va lue	Positive	TP	FP
	Nega tive	FN	TN

Figure 2. Confusion Matrix

3. RESULT AND DISCUSSION

The comparison between the three algorithm models tested can be seen by looking at the model with the highest levels of Accuracy, Recall, Precision, and F1 Score.

3.1. Confusion Matrix of the Three Algorithms

After all data preparation is complete, testing is carried out on the three algorithm models used using Python programming. Training data and training data are divided in a ratio of 80:20, using random state=42. The parameters in testing the Decision Tree, Logistic Regression, and K-neighbors algorithms are the Confusion Matrix which includes Precision, Recall, and Accuracy values. The following Confusion Matrix from the test results of the three models can be seen in the table 4.

Table 4. Decision Trees

	True Positive	True Negative	Class Precision
Pred. Positive	48	22	69%
Pred. Negative	23	11	32%
Class Recall	68%	33%	

It can be seen in Table 4 of the confusion matrix above, that based on 520 patient data with diabetes indicators used in this study, 48 patients were diagnosed with diabetes correctly. Meanwhile, 22 patients who were predicted not to have diabetes were correct. From the matrix results above, 11 people were predicted not to have diabetes, but 23 people who were predicted to be negative turned out to have diabetes. This model has an accuracy of 86%.

Table 5. Logistic Regression

	True Positive	True Negative	Class Precision
Pred. Positive	48	21	70%
Pred. Negative	23	12	34%
Class Recall	68%	36%	

In Table 5 of the confusion matrix of the Logistic Regression model above, the data used and the diabetes indicators used in this study showed that 48 patients were predicted to have diabetes correctly. Meanwhile, 21 patients who were predicted not to have diabetes were correct. From the matrix results above, 12 people who were predicted not to have diabetes were correct, but 23 people who were predicted to be negative turned out to have diabetes. This model has an accuracy of 83%.

Table 6. K-Nearest Neighbors

	True Positive	True Negative	Class Precision
Pred. Positive	47	21	69%
Pred. Negative	24	12	33%
Class Recall	66%	36%	

In table 76 confusion matrix Logistic Regression model above, the data used and the diabetes indicators used in this study, it was obtained that 47 patients were predicted to have diabetes correctly. Meanwhile, 21 patients who were predicted not to have diabetes were correct. From the matrix results above, 12 people were predicted not to have diabetes, which is correct. However, 24 people who were predicted to be negative turned out to have diabetes. This model has an accuracy of 86%. The performance visualization of the three algorithms can be seen in Figure 3.

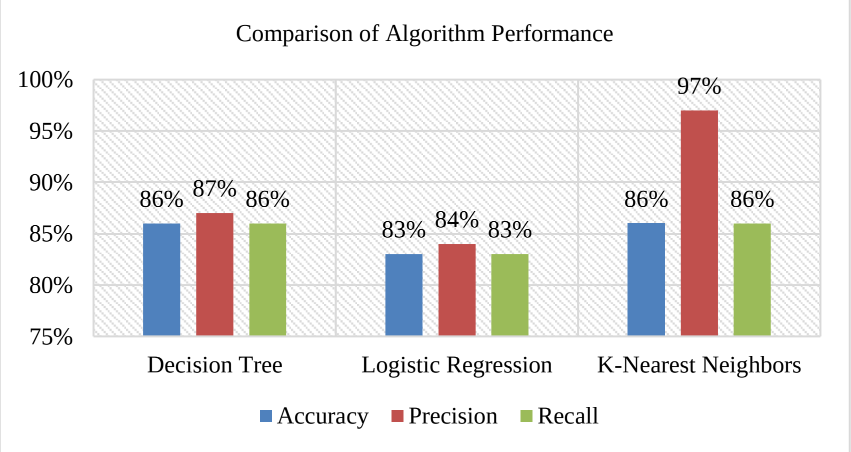


Figure 3. Bar Diagram Comparison of Decision Tree, Logistic Regression, and K-Nearest Neighbors

Comparison of the three algorithms requires the use of the same standards to determine the most optimal algorithm. This process involves measuring the accuracy, memory usage, and precision of the three algorithms [25]. Based on Figure 3, it can be seen that the performance of the Decision Tree algorithm with the Confusion Matrix produces the highest accuracy, recall, and precision compared to the other two algorithms.

4. CONCLUSION

From the test results of the three classification algorithms above, the accuracy, precision, and recall results for each algorithm were obtained, namely Decision Tree of 86%, 87% and 86%. The Logistic Regression Model obtained 83%, 84%, and 83%. Meanwhile, the K-Nearest Neighbors model obtained scores of 86%, 97%, and 86%. From the explanation of the Confusion Matrix and accuracy values above, it can be seen that in testing the three algorithms, the best models used in this testing were the Decision Tree and K-Nearest Neighbors models. So it can be concluded that the classification carried out on diabetes patient data with positive and negative classes is quite accurate with an accuracy value of 86%.

REFERENCES

[1] A. Muliawati and H. Nurramdhani Irmanda, "Penerapan Borderline-SMOTE dan Grid Search pada Bagging-SVM untuk Klasifikasi Penyakit Diabetes," 2022.

[2] S. Anggraini, M. Akbar, A. Wijaya, H. Syaputra, and M. Sobri, "Klasifikasi Gejala Penyakit Coronavirus Disease 19 (COVID-19) Menggunakan Machine Learning," *Journal of Software Engineering Ampera*, vol. 2, no. 1, pp. 57–68, Feb. 2021.

[3] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *Jurnal Sistem Komputer dan Kecerdasan Buatan*, vol. 4, no. 1, pp. 15–21, Oct. 2020.

[4] H. Apriyani, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," 2020. [Online]. Available: https://journal-computing.org/index.php/journal-ita/index

[5] B. A. Candra Permana and I. K. Dewi Patwari, "Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes," *Infotek : Jurnal Informatika dan Teknologi*, vol. 4, no. 1, pp. 63–69, Jan. 2021, doi: 10.29408/jit.v4i1.2994.

[6] H. A. Dwi Fasnuari, H. Yuana, and M. T. Chulkamdi, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Diabetes Melitus," *Antivirus : Jurnal Ilmiah Teknik Informatika*, vol. 16, no. 2, pp. 133–142, Oct. 2022, doi: 10.35457/antivirus.v16i2.2445.

[7] I. F. Nurahmadan, A. Agusta, P. A. Winarno, B. H. Sazali, Y. Thurfah, and A. Rosaliah, *Perbandingan Algoritma Machine Learning Untuk Klasifikasi Denyut Jantung Janin*. 2021.

[8] D. Sartika and D. I. Sensuse, "Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian," *Jatisi*, vol. 1, no. 2, pp. 151–161, Mar. 2017.

[9] A. Ilham Fatimah and S. Saepudin, "Penerapan Data Mining Dengan Metode Apriori Pada Penjualan Sembako (Studi Kasus: Grosir Sembako Lina)," 2022. [Online]. Available: https://rekayasa.nusaputra.ac.id/index

[10] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *Jurnal Media Informatika Budidarma*, vol. 4, no. 2, p. 437, Apr. 2020, doi: 10.30865/mib.v4i2.2080.

[11] I. Kadek, J. Arta, G. Indrawan, G. R. Dantes, P. Studi, and I. Komputer, "Data Mining Rekomendasi Calon Mahasiswa Berprestasi Di Stmik Denpasar Menggunakan Metode Technique For Others Reference By Similarity To Ideal Solution," 2016.

[12] Sri Diantika, Windu Gata, Hiya Nalatissifa, and Mareanus Lase, "Komparasi Algoritma SVM Dan Naive Bayes Untuk Klasifikasi Kestabilan Jaringan Listrik," *Jurnal Ilmiah Elektronika Dan Komputer*, vol. Vol.14, no. No.1, pp. 10–15, Oct. 2021.

[13] S. Ramadani, N. Zannah, S. Ayu, N. Nurhayati, F. Azzahra, and A. P. Windarto, "Analisis Data Mining Naive Bayes Klasifikasi Pada Kelayakan Penerima PKH," *KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer)*, vol. 4, no. 1, pp. 374–381, 2020, doi: 10.30865/komik.v4i1.2725.

[14] M. Gunawan, M. Zarlis, and R. Roslina, "Analisis Komparasi Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu," *Jurnal Media Informatika Budidarma*, vol. 5, no. 2, p. 513, Apr. 2021, doi: 10.30865/mib.v5i2.2925.

[15] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/3820360.

[16] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput Appl*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[17] A. B. Amjoud and M. Amrouch, “Transfer Learning for Automatic Image Orientation Detection Using Deep Learning and Logistic Regression,” *IEEE Access*, vol. 10, pp. 128543–128553, 2022, doi: 10.1109/ACCESS.2022.3225455.

[18] C. Krishna Suryadevara, “Issue 4 Diabetes Risk Assessment Using Machine Learning: A Comparative Study Of Classification Algorithms,” 2023. [Online]. Available: www.iejrd.com

[19] R. A. Husen, R. Astuti, L. Marlia, R. Rahmaddeni, and L. Efrizoni, “Analisis Sentimen Opini Publik pada Twitter Terhadap Bank BSI Menggunakan Algoritma Machine Learning,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 211–218, Oct. 2023, doi: 10.57152/malcom.v3i2.901.

[20] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” *ICT Express*, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

[21] A. Nurjulianty and H. Darwis, “Jurnal Media Informatika Budidarma Perbandingan Metode Naïve Bayes dan K-NN dengan Ekstraksi Fitur GLCM pada Klasifikasi Daun Herbal,” *Jurnal Media Informatika Budidarma*, vol. 7, pp. 1740–1748, 2023, doi: 10.30865/mib.v7i4.6262.

[22] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, “A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic,” *IEEE Access*, vol. 9, pp. 7869–7884, 2021, doi: 10.1109/ACCESS.2020.3047942.

[23] A. W. Sari, T. I. Hermanto, and M. Defriani, “Sentiment Analysis Of Tourist Reviews Using K-Nearest Neighbors Algorithm And Support Vector Machine,” *Sinkron*, vol. 8, no. 3, pp. 1366–1378, Jul. 2023, doi: 10.33395/sinkron.v8i3.12447.

[24] D. Normawati and S. A. Prayogi, “Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter,” 2021.

[25] Y. I. Kurniawan, “Perbandingan Algoritma Naive Bayes dan C.45 dalam Klasifikasi Data Mining,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 5, no. 4, p. 455, Oct. 2018, doi: 10.25126/jtiik.201854803.