Faktor Exacta

Vol. 17, No. 3, Sept 2024, pp. 221~227
eISSN: 2502-339X, pISSN: 1979-276X, DOI: 10.30998/faktorexacta.v17i3.24388

Comparison of Diabetes Disease Classification Models Using Logistic Regression and Random Forest Algorithms

Putri Nabila[1], Amril Mutoi Siregar[2], Sutan Faisal[3], Adi Rizky Pratama[4]
[1,2,3]Departement of Informatics, Faculty of Computer Science, Buana Perjuangan University, Karawang, Indonesia

Article Info

Article history:

Received 5 Jun 2024
Revised 15 Jun 2024
Accepted 27 Aug 2024

Keywords:

Classification
Confussion Matrix
Diabetes
Machine Learning
ROC Curve

ABSTRACT

Diabetes is a lifelong chronic disease that disrupts blood sugar regulation. Diabetes is a life-threatening condition that, if left untreated, can lead to death and other health problems. Several medical tests, including the glycated hemoglobin (A1C) test, blood sugar test, oral glucose tolerance test, and fasting blood sugar test, can detect diabetes. According to statistics, high glucose levels are one of the problems associated with diabetes. This study aims to categorize patients into diabetic and non-diabetic groups using specific diagnostic metrics included in the dataset. The researchers used one thousand five hundred patient records with 9 attributes and 2 classes. The study used machine learning techniques, including Logistic Regression, Random Forest, and Confusion Matrix and Receiver Operating Characteristics (ROC) assessment. The Random Forest method produced 97% accuracy, 97% precision, 100% recall, and 98% f1-score, indicating that the accuracy level seems good but can still be improved. Based on the accuracy findings, random forest is the most effective strategy for logistic regression.

Corresponding Author:

Amril Mutoi Siregar,
Computer Science,
Buana Perjuangan University,
Jl. Jalan HS. Ronggo Waluyo, Telukjambe Timur, Puseurjaya, Telukjambe Timur, Kabupaten Karawang, Jawa Barat 41361
Email: amrilmutoi@ubpkarawang.ac.id

1. Introduction

Diabetes is a lifelong chronic disease that interferes with the regulation of blood sugar. Diabetes is a life-threatening condition that, if left untreated, can lead to serious health problems and even death [1]. The World Health Organization (WHO) estimates that the number of people with diabetes worldwide increased from 108 million in 2013 to 422 million in 2014. By 2045, this number is expected to increase to 629 million. Diabetes is estimated to be the cause of 1.6 million deaths [2].

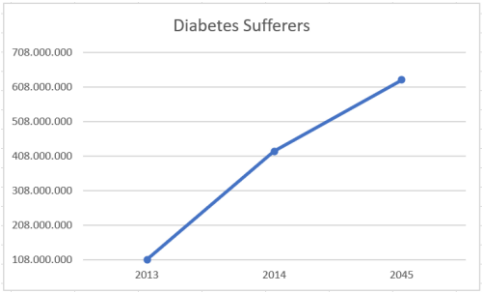

Figure 1. Graph of People With Diabetes According to WHO

A condition characterized by increased blood sugar levels in the body, caused by a disorder that significantly affects the body's ability to remove glucose from the blood. This is caused by a hormone called "insulin", which controls the body's blood sugar levels [3].

Many medical tests, including the glycated hemoglobin (A1C) test, fasting sugar test, oral glucose tolerance test, and random blood sugar test, can be used to detect diabetes. When polyuria, polydipsia, and polyphagia symptoms of diabetes occur, people tend to forget small everyday risks, such as consuming too many sweet foods, which can increase blood sugar levels. This large healthcare dataset can be leveraged to extract knowledge using artificial intelligence, especially Machine Learning (ML) methods. This can lead to early detection and prevention of chronic diseases that reduce quality of life, such as diabetes. [4].

A statistical subfield of artificial intelligence machine learning, concerned with the creation of methods and algorithms that enable computers to learn and gain intelligence through the examination of historical data. Machine learning methods help in solving problems involving pattern identification, prediction, and categorization. Applications for it include facial tagging and recognition, email filtering, search engine optimization, web page ranking, robotics, traffic control, and disease prediction and categorization [5]. Many experiments have been conducted recently to accurately detect diabetes using various ML models on various dataset.[6]. Here are a few instances of deep learning and machine learning algorithms: Decision Tree (DT), Random Forest Classifier (RFC), Naive Bayesian Classifier (NBC), K-Nearest Neighbor Classifier (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest Classifier (NBC). Several prior research have successfully used this strategy to categorize utilizing diabetes data [7].

In their first study [8] introduced a genetic programming based classifier that predicts diabetes using a limited set of functions and various selection strategies. At a maximum accuracy of 89%, the best results were achieved using a tournament selection strategy. Using the PID dataset, a classification method utilizing Gaussian Process Classification (GPC) by [9]. was developed to predict diabetes, using radial, linear, and polynomial basis kernels. [10] (second study) Gaussian Process (GP) technique was used because the medical dataset is non-linear and has intrinsic correlation. Compared to other methods such as NB, Linear Discriminant Analysis (LDA), and Quadratic Covering Discriminant Analysis (QDA), its efficacy is striking. The second trial result showed that the GP using the radial basis kernel performed the best, with an accuracy of 81.5%. A new dataset containing information about individuals who have recently developed diabetes or are at risk of developing the condition in the future was used in a recent study [11] on early diagnosis of diabetes risk. The dataset was provided by patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh. Among the ML classifiers used were NB, LR, and RF. The performance of the classifiers was assessed using train-test split and K-Fold Cross Validation techniques. With an accuracy rating of 97.4% using K-Fold Cross Validation and 99% using train-test split, the RF classifier produced the best results.

A strategy combining ensemble techniques with information retrieval feature selection is described, according to [12]. Bayesian network classifier and multi-layer perceptron form this ensemble. An accuracy of 81.89% was achieved by deriving a subset of 6 features from the entire feature set of 8 features in the PID dataset. [13] proposed a classification scheme designed to provide maximum precision in diabetes risk assessment. First, data pre-processing methods were used on the PID dataset, including replacing missing values and outliers with medians. The next step was feature selection, which involved the use of methods such as Fisher's discriminant ratio, ANOVA, LR, RF, mutual information, and Principal Component Analysis (PCA). Next, a number of classifiers were applied, including NB, LDA, QDA, SVM, ANN, AB, LR, DT, and RF. The use of the RF approach for feature selection and classification resulted in a maximum accuracy of 92.26%.

The aim of this study, however, is to assess and enhance the accuracy results of the algorithms employed in earlier studies, as well as classify diabetes and non-diabetes based on the results of diagnostic tests that have been passed, utilizing a variety of machine learning techniques. The utilization of distinct data and methodologies, namely feature selection to choose the best features from a feature data set, distinguish this study from earlier research. What distinguishes this study from previous studies is in terms of different methods. Researchers use the Random Forest algorithm, and Logistic Regression with a comparison of Confusion Matrix and ROC analysis to improve and find out the accuracy results of the algorithm used in previous studies.

2. METODE (10 PT)

2.1. Dataset

Clinical data of 1500 patients were obtained from kaggle (https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset), which includes two classes and nine attributes: gender, age, blood glucose level, heart disease, hypertension, smoking history, BMI, HbA1c level, and diabetes.

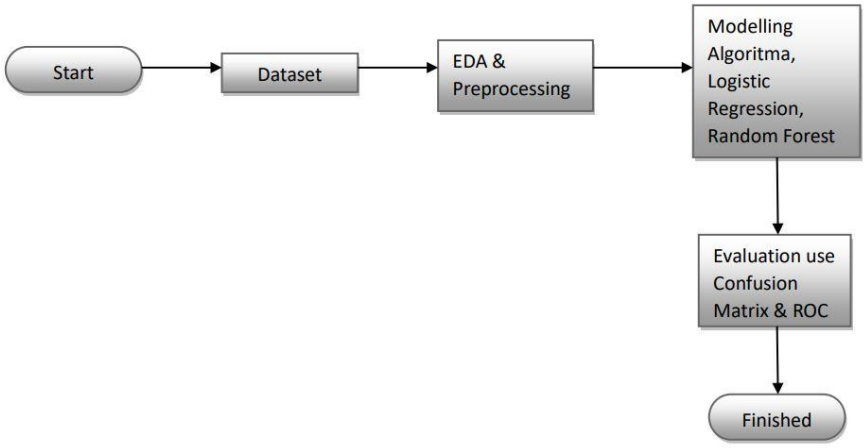

Figure 2. Study Stages

Figure 2 illustrates the many phases, including dataset collection, preparation, algorithm testing, and then its assessment with ROC and confusion matrix. The study used two machine learning techniques: logistic regression, and random forest. The study was divided into two classes: diabetics and non-diabetics.

2.2. Exploratory Data Analysis (EDA)

Data can be investigated and analyzed using exploratory data analysis, which leverages a variety of visual tools. With the help of these techniques, data scientists can find anomalies, question assumptions, and uncover patterns that help them better understand their tasks. [14].

A correlation matrix needs to be calculated before we can ascertain the relationship between each pair of variables in our investigation. If there is a significant correlation between the traits, then the removal of one of them should be considered. To learn more about the traits and the data set, the data balance and attribute distribution are investigated using various data visualization techniques [15]. In addition, the main goal is to find patterns in a very large data set by using visualization techniques and reducing its size.

2.3. Logistic Regression

LR classifies one or more additional data sets (testing set) after the model has been trained on the given data set (training set) [16]. The basic objective of logistic regression is to carefully examine the relationship between predictor and target variables. The positive and negative class variables in the logistic regression model are identified using the sigmoid function [17].

When the dependent (response) variable is binary, logistic regression, a subset of regression analysis, is used. Dichotomous variables, which are typically given a number between 0 and 1, only have two values that represent the presence or absence of an event. [18]. According to research [19] Binary logistic regressions evaluate the chance that a characteristic of a binary variable is present, given the values of the covariates. Specifically, there is a linear combination of independent variables with log-odds of the probability of an event in a logistic model.

2.4. Random Forest

A supervised machine learning algorithm called Random Forest builds many decision trees. The bulk of the decision tree is used to make the ultimate choice. Decision trees have a high variance and a low bias. High variance is converted to low variance using random forests [20].

Among distinct ensemble tree models, RF with drastically diverse tree architectures and splitting factors causes a variety of overfitting and outlier occurrences. Thus, for classification issues, final prediction voting minimizes overfitting, whereas for regression problems, averaging is the solution. The best split in a random sample of predictors is chosen as a candidate split from the entire collection of predictors each time in this individual decision tree generating process. Every split, a fresh predictor sample is collected with a user-specified number of predictors (Mtry). When RF is applied to a user-specified number of trees (Ntree), it generates trees with low bias and high variance.[21]

2.5. Evaluation Matrix

The machine learning model's performance is gauged using the evaluation matrix. After the model is constructed, we must evaluate the machine learning model's performance. To test the model, a number of assessment metrics are offered. Among these are the confusion matrix, ROC-AUC, F1 Score, accuracy, precision, recall, and specificity.

A table that displays the efficacy of the categorization model is known as the confusion matrix, sometimes referred to as the error matrix. AUC-ROC Curve, Precision, Recall, F1 Score, Accuracy, and—most importantly—Accuracy can all be measured using it quite successfully. Table displays four distinct combinations of planned and actual classes [22].

An evaluation tool for classification work at all classification levels is the AUC-ROC curve. The probability curve is represented by the ROC, while the AUC quantifies the degree of separation. The AUC of the model indicates how well it can discriminate between those with diabetes and those without the disease. The True Positive Rate is plotted against the False Positive Rate on the ROC curve. [23].

Several metrics are used by different fields, depending on the objective, to assess a classifier's or predictor's performance. [24]. Analyzing the degree of similarity between the anticipated and actual values is one method of assessing an algorithm's accuracy. In other words, when the algorithm is employed in a study, the accuracy result functions as a gauge of its efficacy. [25] ROC curves and confusion matrices (classification accuracy, recall, and classification precision) were employed to evaluate the study's findings.

3. RESULT AND DISCUSSION

Quality qualities that do not exhibit a strong association with one another are eliminated in order to ascertain the relationship between each pair of variables. Attribute distribution and data balance are examined using a variety of data visualization techniques to enhance comprehension of the dataset and its aspects. The next step is label coding, which simplifies and improves understanding by converting category data to numerical form.

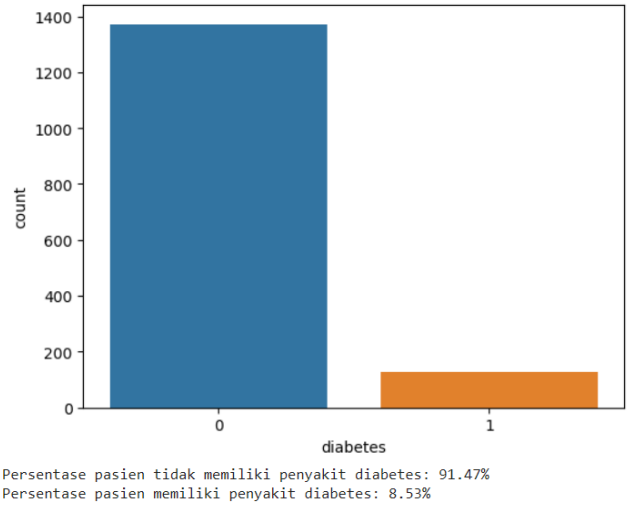

Figure 3. Diagram of Patients With Diabetes and Without Diabetes

Improving understanding of the dataset and its attributes: Before proceeding with additional data processing, attributes such as smoking history that do not contribute to accurate answers were discarded. The visualization of the data presented in Figure 2 shows that 91.47% do not have diabetes and 8.53% do have diabetes, indicated by 1 as having diabetes and 0 as not having diabetes. EDA was performed to identify potential missing values and duplications.

Table 1. Performance Measurement of Multiple Machine Learning Methods

Methode	Accuracy	Precission	Recall	FI-Score
Logistic Regression	96%	96%	99%	98%
Random Forest	97%	98%	100%	98%

Subsequently, the data was partitioned into training and testing sets, and various data splits were tested. Table 1 presents a comparison of the outcomes for f1-score, recall, accuracy, and precision. It also compares the outcomes for Random Forest and Logistic Regression. It demonstrates that the Random Forest method outperforms the Logistic Regression algorithm overall.

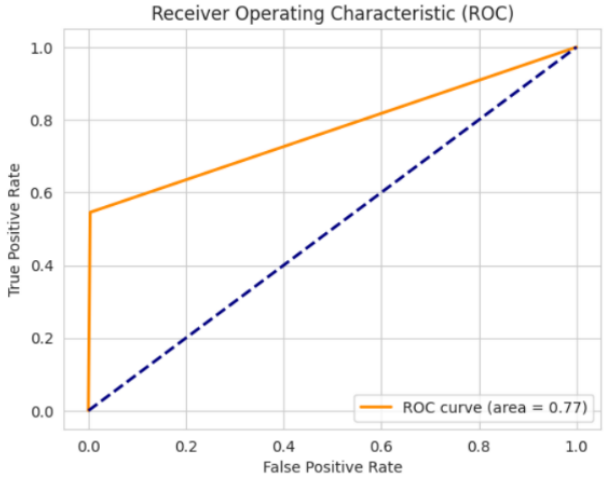

Figure 4. ROC Logistic Regression

Figure 4. Shows the ROC evaluation of the Logistic Regression algorithm with a result of 0.77, this shows that the ROC curve of logistic regression is quite good because it is more than 0.5.

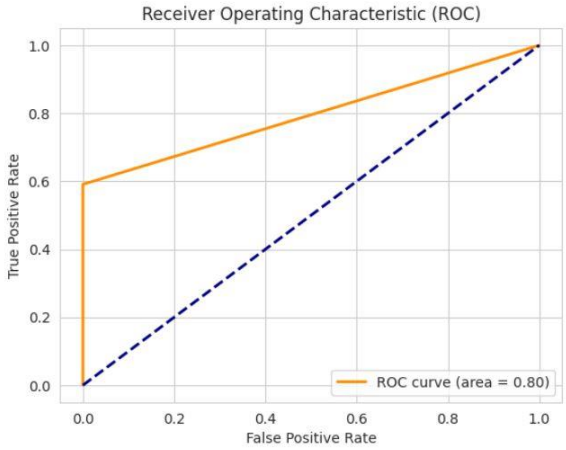

Figure 5. ROC Random Forest

Figure 5. Shows the ROC evaluation of Random Forest with a result of 0.80, this shows that the ROC curve of Random Forest is quite good because it is more than 0.5.

Figures 4 and 5 show the ROC results with 2 different methods. The Logistic Regression value is lower than the Random Forest value, as well as the accuracy results shown in table 1, the accuracy results of Random Forest are better than Logistic Regression.

4. CONCLUSION

Based on the results of previous studies of the two algorithms that have passed the testing process using test data and training data through confusion matrix evaluation, from several of the best experiments used to measure the level of accuracy using the Logistic Regression algorithm with an accuracy of 96%, precision 96%, recall 99% and f1-score 98% while for the results of the Random Forest algorithm obtained an accuracy of 97%, precision 98%, recall 100% and f1-score 98%. So it can be concluded that the classification of diabetes data that has been carried out provides quite good results, and improves the results of previous tests. The accuracy and ROC values differ significantly when evaluated with different machine learning techniques.

The level of accuracy looks good, although there is still room for improvement. Among all these techniques based on the accuracy results, Random Forest as a method is quite good. Researchers hope that further research can use different machine learning methods with the same data to find out the best results of each method.

REFERENCES

[1] A. Armghan, J. Logeshwaran, S. M. Sutharshan, K. Aliqab, M. Alsharari, and S. K. Patel, 'Design of biosensor for synchronized identification of diabetes using deep learning', *Results in Engineering*, vol. 20, Dec. 2023, doi: 10.1016/j.rineng.2023.101382.

[2] F. Anwar, Qurat-Ul-Ain, M. Y. Ejaz, and A. Mosavi, 'A comparative analysis on diagnosis of diabetes mellitus using different approaches – A survey', *Inform Med Unlocked*, vol. 21, Jan. 2020, doi: 10.1016/j.imu.2020.100482.

[3] S. C. Gupta and N. Goel, 'Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques', in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 1257–1269. doi: 10.1016/j.procs.2023.01.104.

[4] F. Navazi, Y. Yuan, and N. Archer, 'An examination of the hybrid meta-heuristic machine learning algorithms for early diagnosis of type II diabetes using big data feature selection', *Healthcare Analytics*, vol. 4, Dec. 2023, doi: 10.1016/j.health.2023.100227.

[5] T. Mora, D. Roche, and B. Rodríguez-Sánchez, 'Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms', *Diabetes Res Clin Pract*, vol. 204, Oct. 2023, doi: 10.1016/j.diabres.2023.110910.

[6] S. Samreen, 'Memory-efficient, accurate and early diagnosis of diabetes through a machine learning pipeline employing crow search-based feature engineering and a stacking ensemble', *IEEE Access*, vol. 9, pp. 134335–134354, 2021, doi: 10.1109/ACCESS.2021.3116383.

[7] S. Chu *et al.*, 'Machine learning algorithms for predicting the risk of fracture in patients with diabetes in China', *Heliyon*, vol. 9, no. 7, Jul. 2023, doi: 10.1016/j.heliyon.2023.e18186.

[8] G. Aguilera-Venegas, A. López-Molina, G. Rojo-Martínez, and J. L. Galán-García, 'Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus', *J Comput Appl Math*, vol. 427, Aug. 2023, doi: 10.1016/j.cam.2023.115115.

[9] N. P. Tigga and S. Garg, 'Prediction of Type 2 Diabetes using Machine Learning Classification Methods', in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 706–716. doi: 10.1016/j.procs.2020.03.336.

[10] T. Mora, D. Roche, and B. Rodríguez-Sánchez, 'Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms', *Diabetes Res Clin Pract*, vol. 204, Oct. 2023, doi: 10.1016/j.diabres.2023.110910.

[11] A. Hennebelle, H. Materwala, and L. Ismail, 'HealthEdge: A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes in an Integrated IoT, Edge, and Cloud Computing System', in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 331–338. doi: 10.1016/j.procs.2023.03.043.

[12] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba, and S. A. Bahaj, 'Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model', *IEEE Access*, vol. 10, pp. 73079–73086, 2022, doi: 10.1109/ACCESS.2022.3189374.

[13] 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.

[14] R. Marzouk, A. S. Alluhaidan, and S. A. El Rahman, 'An Analytical Predictive Models and Secure Web-Based Personalized Diabetes Monitoring System', *IEEE Access*, vol. 10, pp. 105657–105673, 2022, doi: 10.1109/ACCESS.2022.3211264.

[15] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, 'Diabetes prediction using ensembling of different machine learning classifiers', *IEEE Access*, vol. 8, pp. 76516–76531, 2020, doi: 10.1109/ACCESS.2020.2989857.

[16] H. El Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, 'Diabetes Prediction Using Machine Learning Algorithms and Ontology', *Journal of ICT Standardization*, vol. 10, no. 2, pp. 319–338, 2022, doi: 10.13052/jicts2245-800X.10212.

[17] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba, and S. A. Bahaj, 'Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model', *IEEE Access*, vol. 10, pp. 73079–73086, 2022, doi: 10.1109/ACCESS.2022.3189374.

[18] D. Y. Utami, E. Nurlelah, and F. N. Hasan, 'Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to predict diabetes', *JOURNAL OF INFORMATICS AND*

TELECOMMUNICATION ENGINEERING, vol. 5, no. 1, pp. 53–64, Jul. 2021, doi: 10.31289/jite.v5i1.5201.

[19] R. D. Joshi and C. K. Dhakal, 'Predicting type 2 diabetes using logistic regression and machine learning approaches', *Int J Environ Res Public Health*, vol. 18, no. 14, Jul. 2021, doi: 10.3390/ijerph18147346.

[20] M. Pal and S. Parija, 'Prediction of Heart Diseases using Random Forest', in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Mar. 2021. doi: 10.1088/1742-6596/1817/1/012009.

[21] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, 'Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13. Institute of Electrical and Electronics Engineers Inc., pp. 6308–6325, 2020. doi: 10.1109/JSTARS.2020.3026724.

[22] 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

[23] C. Maisyarah, E. Haryatmi, R. Y. Fajriatifah, and Y. H. Puspita, 'Prediksi Penyakit Diabetes menggunakan Algoritma Artificial Neural Network', vol. 2, no. 1, pp. 46–52, 2022.

[24] A. M. Siregar, 'Klasifikasi Untuk Prediksi Cuaca Menggunakan Esemble Learning', *PETIR*, vol. 13, no. 2, pp. 138–147, Sep. 2020, doi: 10.33322/petir.v13i2.998.

[25] S. Faisal and U. Buana Perjuangan Karawang Karawang, 'TechnoXplore Jurnal Ilmu Komputer & Teknologi Informasi Klasifikasi Data Minning Menggunakan Algoritma C4.5 Terhadap Kepuasan Pelanggan Sewa Kamera Cikarang, 2019.