Journal of Biomedical Science and Bioengineering

The Performance Comparison of Machine Learning Models for COVID-19 Classification Based on Chest X-ray

Elvira Sukma Wahyuni*, Eka Putra Prasetya

Universitas Islam Indonesia, Indonesia

ABSTRACT. COVID-19 has become a pandemic spread to nearly all countries in the world. This virus has caused many deaths. Screening using a chest X-ray is an alternative to find out positive COVID-19 patients. Chest X-ray is advantageous because every hospital must have an X-ray device so that hospitals do not need additional equipment to detect COVID-19-positive patients. This study aims to compare the machine learning models of Naive Bayes, Decision Tree, K-Nearest Neighbor, and Logistic Regression to predict COVID-19 positive patients. The stages of the research carried out by this study are the Pre-process stage, feature extraction, and classification. The results showed that the Naïve Bayes classification method got the highest performance with an accuracy of 95.24%.

Keywords: COVID-19, Chest X-Ray, Machine Learning, Naive Bayes, Logistic Regression, K-NN, Decision Tree

Article History: Received 24 May 2022; Received in revised form 1 July 2022; Accepted 3 July 2022; Available online: 27 July 2022 (7.5 pt)

DOI: 10.14710/jbiomes.jbiomes.2.1.1-6

1. INTRODUCTION

Corona Virus Disease or COVID-19 is a virus that can infect the respiratory system. This virus first discovered in Wuhan city, China on the end of December 2019. This virus spread vastly to various countries within a few months [1]. As of December 22, 2021, there were 275 million positive cases of COVID-19 and 5.3 million people had died worldwide [2]. The symptoms of people affected by COVID-19 include cough, fever, shortness of breath. In more serious cases, COVID-19 can cause lungs inflammation or commonly known as pneumonia [3].

One way to detect COVID-19 is to use X-ray. X-ray is used to scan the patient's lungs. The results of scanning images of the patient's lungs are needed to analyze whether the patient's lungs are detected to have been infected by COVID-19. The use of X-ray as a COVID-19 detection tool can be used as an alternative solution since every hospital must have X-ray so no additional equipment is needed [4].

Similar studies on the comparison of machine learning models for the classification of COVID-19 on chest X-ray have been carried out before. The machine learning models used are Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). This research resulted in SVM being more reliable than KNN. SVM produces a precision value of 97%, while KNN produces a precision value of 86% [5].

This study aims to compare machine learning models to classify patients positive for COVID-19 and normal patients using X-ray images of lung scans. The machine learning models that are compared in this study include Naive Bayes, Decision Tree, K-NN, and Logistic Regression. The X-ray scan image was preprocessed to make it easier for the model to classify COVID-19 patients and normal patients. U-Net was used to segment the patient's lungs from objects other than lungs and U2Net was used to remove objects other than lungs. Feature extraction was also used to obtain information from X-ray images of lung scans.

2. LITERARURE REVIEW

2.1 Database

The dataset used is a free dataset from Kaggle. This dataset contains chest X-ray images with a size of 2746 x 2382 pixels. A total of 317 images have been divided into training and test data. There are 3 classes, namely Covid, normal, and viral pneumonia. Covid class shows a chest X-ray of a Covid patient. Normal class shows chest X-ray images of normal people. The viral pneumonia class shows a chest X-ray of a patient with pneumonia. In the test section, there are 26 images in the covid class, 20 images in the normal class, and 20 images in the viral pneumonia class. In the training section, there are 111 images in the covid class, 70 images

* Corresponding author: elvira.wahyuni@uii.ac.id

in the normal class, and 70 images in the viral pneumonia class [6]. In this study, 2 classes were used, namely normal and COVID-19 using 70 X-ray images of normal patients and 70 X-ray images of positive COVID-19 patients.

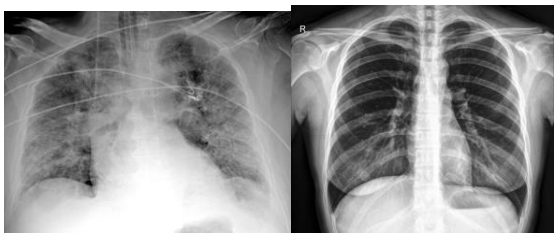

Fig. 1 X-ray images of a) Covid-19 b) Normal

2.2 U-Net

The U-Net architecture is composed of a Fully Convolutional Network (FCN) and modified in such a way to produce better segmentation in medical imaging. U-Net applies elastic deformation to the available training images. This allows the network to study the invariance of these deformations without the need to look at the transformation in the image used [7].

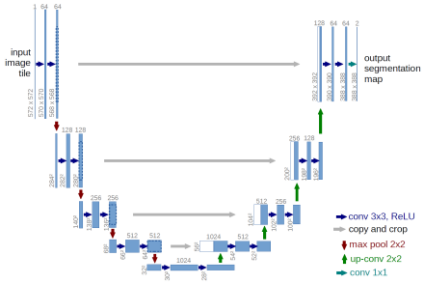

Fig. 2 U-Net Architecture [7]

2.3 U2-Net

U2-Net is commonly used for Salient Object Detection (SOD). U2-Net will make the area you want to detect to be white, whilst the area around it to be black. The U2-Net architecture is composed of a two-level nested U-structure. Nested U-structure allows the network to capture more information from local and global both shallow and deep layers regardless of the image resolution [8].

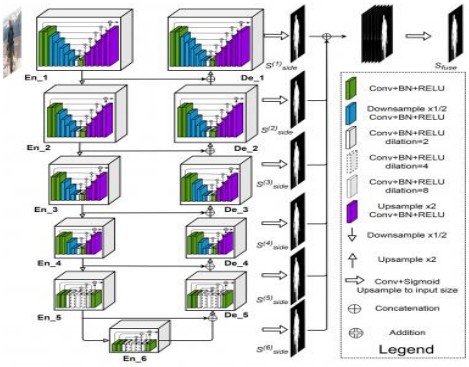

Fig. 3 U2-Net Architecture [8]

2.4 Naive Bayes

Naive Bayes is a classification method using statistics and probability. This method predicts future opportunities based on past experiences. Naive Bayes has a characteristic that is a very strong assumption or can be called naive about the independence of each condition or event. The advantage of this method over other classification methods is that Naive Bayes only requires a small amount of training data to generate predictions from the classification process. This method makes assumptions on the independent variables so that only the variance of a variable in a class is needed to determine the classification, not the entire covariance matrix [9].

2.5 Decision Tree

Decision Tree is a machine learning model that has a tree structure for making decisions. In the Decision Tree there are several elements including root nodes, twigs, and leaf nodes. The root node is at the very top which serves to represent the final goal or major decisions to be made. Twigs serve as branches from the roots that represent different choices. The leaf node is the end of the branch in the Decision Tree that contains the action made [10].

2.6 Logistic Regression

Logistic Regression is a machine learning model that uses the principle of probability. Although there are words regression, Logistic Regression is mostly used to solve classification problems. The output of this method is a probability value with a range of 0 to 1. If the estimated probability value is greater than 50 percent, it is called a positive class, whereas otherwise it is called a negative class [11].

2.6 K-Nearest Neighbor

K-Nearest Neighbor is a machine learning model that uses the principle of taking the nearest K data or its neighbors as a parameter to determine a class of new data. The training data is projected to a larger dimension. Each dimension carries features from the data. The best K value is determined from the data, a high K value will reduce noise in the classification process, the classification boundaries will become more blurred [12].

3. MATERIALS AND METHODS

3.1 Design

In this study, several stages were carried out to process X-ray images of the lungs from Kaggle so that they could be classified between Covid and normal images using Naive Bayes. The stages included preprocessing, feature extraction, data split, performance evaluation. The preprocessing stage was done using python. The feature extraction stage to performance was carried out using RapidMiner Studio.

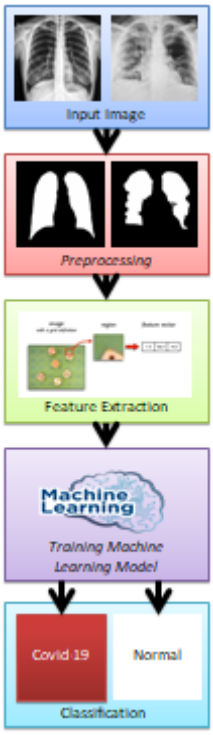

Fig. 4 Research flowchart

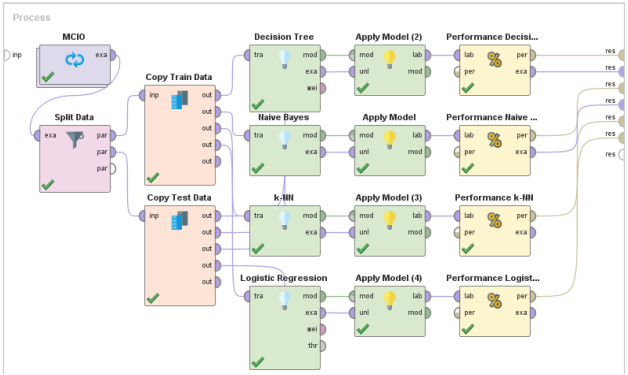

Fig. 5 Process architecture in RapidMiner Studio

3.2 Preprocessing

Preprocessing is needed to process X-ray images of the lungs so that they are easily understood by the Naive Bayes model so as to produce good classification results. The preprocessing process included segmentation using U-Net and removing unnecessary objects using U2-Net.

U-Net was used to mark the lung area. Before entering the U-Net process, the image was resized to 230 x 230 pixels to speed up image processing. Areas of the lungs were marked with a pink area, while areas other than the lungs colored blue.

U2-Net is used for Salient Object Detection (SOD) which serves to remove images other than the image of the lungs. The pink areas of the lungs will turn white, while the areas other than the lungs or blue areas will turn black.

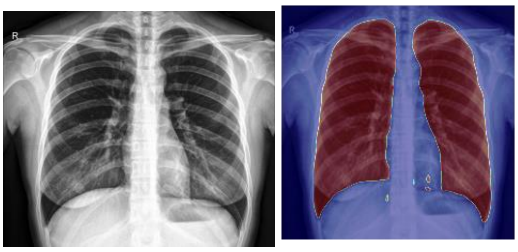

Fig. 6 X-ray images of normal patient a) Before segmentation (left) b) After segmentation (right)

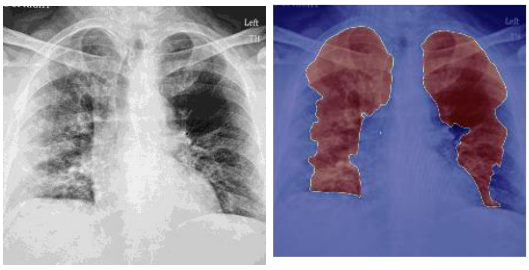

Fig. 7 X-ray images of COVID-19 patient a) Before segmentation (left) b) After segmentation (right)

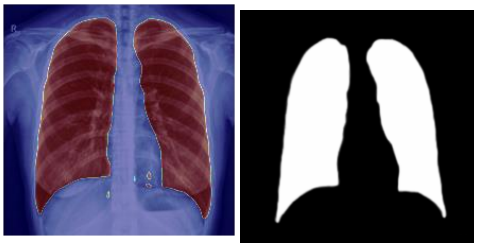

Fig. 8 X-ray image of normal patient a) Before SOD (left) b) After SOD (right)

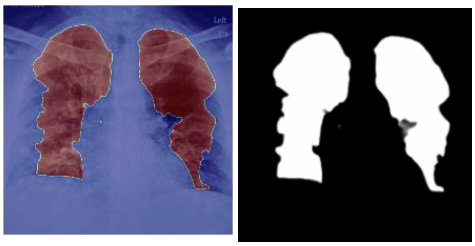

Fig. 9 X-ray image of normal patient a) Before SOD (left) b) After SOD (right)

3.3 Feature Ectraction

Feature extraction is useful for obtaining characteristics from preprocessed lung X-ray images. This study used 4 methods that are already available in RapidMiner Studio including Global statistics, histogram, dLog, and Order-based Block Color (OBCF).

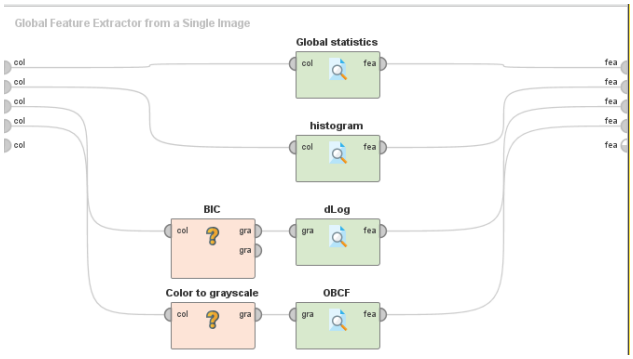

Fig. 10 Extraction architecture feature of the RapidMiner Studio

Global Statistics is useful for extracting global features from images. A global feature is a one-dimensional vector that contains a description of the characteristics of the image at all pixels. In global statistics, the extracted global features are Area Fraction, Edginess, Kurtosis, Max Gray Value, Mean, Median, Minimum Gray Value, Normalized Center of Mass, Peak, Skewness, and Standard Deviation.

The histogram is useful for showing the distribution of pixels based on the intensity of the gray level of each pixel. In this experiment, 10 bins were used. The bin is defined as the number of color bars that form on the histogram.

dLog is useful for calculating the dLog distance from an image. The input for dLog must be a Border/Interior Classification (BIC). dLog will calculate the logarithm distance between the border and interior pixels. BIC serves to classify between the border and the interior. The interior is classified when the 4 neighbors (top, bottom, right, and left) have the same quantized color, while the border is the opposite [12].

Order-based Block Color (OBCF) is useful for extracting colors from images. OBCF calculates the average, minimum, and maximum, gray values of each cell defined with row and column dimensions. Since the processed color is gray, the input image to the OBCF process needs to be passed from Color to Grayscale to convert the image to grayscale.

3.4 Data Split

Data split is useful for separating images between training and testing images for each class. This separation is done by random sampling. Training images account for 70% of the total number of images from each class or 49 images of COVID-19 patients and 49 images of normal patients. Testing images account for 30% of the total number of images from each class or 21 images of COVID-19 patients and 21 images of normal patients.

3.5 Performance Evaluation

The evaluation of classification performance in machine learning commonly used is the Confusion Matrix. The confusion matrix is a table that contains a description of the performance of the classification model against a series of test data, based on the actual values known in advance. The values in the confusion matrix can be used to calculate accuracy, precision, recall, specificity, and F1-Score. The calculation requires 4 elements of the confusion matrix consisting of true positive (TP), true negative (TN), false positive (FP), and false-negative (FN).

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

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

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

$$F1 - Score = 2 \times \frac{(presisi \times recall)}{(presisi + recall)} \quad (4)$$

In this study, TP was defined as a chest X-ray of a patient who was successfully classified by the model as positive for COVID-19. TN was defined as a chest X-ray that was successfully classified by the model as negative for COVID-19 or normal. FP is defined as a model for classifying X-ray images as positive for COVID-19, but in reality, they are negative for COVID-19. FN is defined as a model for classifying X-ray images as negative for COVID-19, but in fact the image is positive for COVID-19.

4. RESULT AND DISCUSSION

Chest X-ray images were used to predict COVID-19 positive patients and normal or COVID-19 negative patients. General machine learning models to solve classification problems such as Decision Tree, Naive Bayes, K-NN, and Logistic Regression are used to train and test chest X-ray images that have gone through the preprocessing stage using U-Net and U2-Net. The performance of the four models will be compared to find out the most reliable machine learning model to solve the problem.

Table 1
Confusion Matrix

Model	TP	FP	TN	FN
Decision Tree	16	0	23	3
Naive Bayes	19	2	21	0
K-NN	12	0	23	7
Logistic Regression	14	2	21	5

Table 1 shows the results of the Decision Tree, Naive Bayes, K-NN, and Logistic Regression test models using the confusion matrix. The test used 30% of the data or a total of 42 images which were divided into 21 X-ray images of positive COVID-19 patients and 21 X-ray images of COVID-19 negative patients. From Table 1 it can be seen that the Naive Bayes model has the highest TP value and TN value which corresponds to the total number of test images of negative COVID-19 patients. This model is best at predicting both positive and negative COVID-19 patients based on chest X-ray images. The K-NN model is the least preferred in predicting both positive and negative COVID-19 patients because it has the least number of TP and the number of TN exceeds the number of chest X-ray images available on the test.

Table 2
Performace result of machine learning model (%)

Model	Accuracy	Precision	Recall	Specificity	F1 Score
Decision Tree	92,86	100	84,21	100	91,43
Naive Bayes	95,24	90,48	100	91,30	95,00
K-NN	83,33	100	63,16	100	77,42
Logistic Regression	83,33	87,50	73,68	91,30	80,00

Table 2 shows the performance of the Decision Tree, Naive Bayes, K-NN, and Logistic Regression models in terms of accuracy, precision, recall, specificity, and F1-score derived from the calculation of TP, TN, FN, FP confusion matrix. This table shows the performance advantages and disadvantages of the four models used to solve the problem of predicting positive and negative COVID-19 patients.

The highest accuracy is owned by the Naive Bayes model with a value of 95.24%. The Naive Bayes model interpreted as having the highest ability to determine between patients who are confirmed positive or negative for COVID-19 among other models.

The highest precision is owned by the K-NN and Decision Tree models with a value of 100%. Both models could correctly classify patients who were confirmed positive for COVID-19 from all predictions of positive patients for COVID-19 carried out by the two models.

The highest recall is owned by the Naive Bayes model with a value of 100%. This Naive Bayes model can be interpreted that the predictions of this model are all correct from a chest X-ray image labeled as a positive patient for COVID-19.

The highest specificity is owned by the K-NN and Decision Tree models with a value of 100%. Both models were able to correctly predict COVID-19 negative patients from the overall chest X-ray image labeled as COVID-19 negative.

The highest F1-Score is owned by Naive Bayes with a value of 95%. Naive Bayes has the best precision and recall values among all models. The F1-Score value that is close to 100% indicates that both the precision and recall values are close to 100%. The precision and recall values of 100% indicate that the model is perfect, without errors in classifying positive COVID-19 patients and COVID-19 negative patients based on chest X-ray images.

Data visualization for table 2 can be seen in Figure 11. Figure 11 makes it easy to see the most superior model for each parameter of machine learning model performance measurement using the confusion matrix.

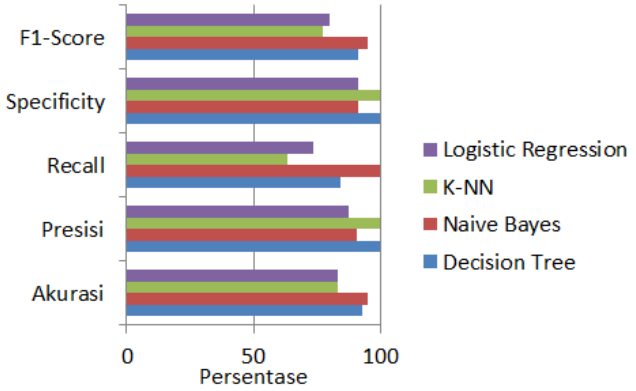

Fig. 11 Machine learning model performance in the form of bar chart

Based on the analysis in Tables 1, 2, and Figure 11, the Naive Bayes Model is the best model among the other three models to predict COVID-19 positive patients and COVID-19 negative patients. There are several reasons that make Naive Bayes the best model to solve this problem.

The first reason is that the True Positive (TP) value of this model is 19. This TP value is the highest value among other models so that Naive Bayes is the model that predicts the most positive patients with COVID-19.

The second reason is the False Negative (FN) value of this model is 0. In solving this problem, FN is more of a consideration than the FP value. FN can be interpreted that the model predicts that the X-ray image is a negative patient, but the chest X-ray image of the patient should be a positive patient for COVID-19. FN value 0 means that there was no error in determining the negative prediction of COVID-19 so that there are no patients confirmed positive for COVID-19 hanging around due to machine learning misclassification.

The third reason is the accuracy value, and F1 – Naive Bayes Score is the highest value among the other three models, which is 95.24%. High accuracy can be interpreted as having the highest ability to determine between patients who are confirmed positive or negative for COVID-19 among other models. The F1 – High score of 95%, it can be concluded that Naive Bayes precision and recall are the best among the other three models.

5. CONCLUSION

Machine learning models Decision Tree, Naive Bayes, K-NN and Logistic Regression were used and compared their performance to predict COVID-19 patients based on chest X-ray images. The performance test of the four models uses accuracy, precision, recall, specificity, and F1-Score from the confusion matrix calculation. Of the four machine learning models, Naive Bayes is the best model to solve this problem. Naive Bayes got the highest accuracy score of 95.24%, had no false negatives (there were no errors in predicting negative COVID-19 patients), and had the highest F1-Score of 95%. In future research, it is recommended to add other machine learning models such as neural networks, Random Forests, and Support Vector Machines. In addition, it is also possible to use different preprocessing methods and extraction features to get the best machine learning model

performance in solving the problem of predicting positive COVID-19 patients.

ACKNOWLEDGMENTS

You can put your acknowledgements here

REFERENCES

[1] M. L. Parwanto, "Virus Corona (2019-nCoV) penyebab COVID-19," J. Biomedika Dan Kesehatan., vol. 3, pp. 1–2, Mar. 2020, doi: 10.18051/JBiomedKes.2020.v3.1-2.

[2] "WHO Coronavirus (COVID-19) Dashboard", *Covid19.who.int*, 2021. [Online]. Available: https://covid19.who.int. [Accessed: 23- Dec- 2021].

[3] Roudsari, P. P., Alavi-Moghadam, S., Payab, M. and et al.,"Auxiliary role of mesenchymal stem cells as regenerative medicine soldiers to attenuate inflammatory processes of severe acute respiratory infections caused by COVID-19," Cell Tissue Bank, 21:405-425, 2020.

[4] Buyut Khoirul Umri, Ema Utami, Mei Parwanto Kurniawan, "Comparative Analysis of CLAHE and AHE on Application of CNN Algorithm in the Detection of COVID-19 Patients", *Information and Communications Technology (ICOIACT) 2021 4th International Conference on*, pp. 203-208, 2021.

[5] S. Samsir, J. Sitorus, Z. Ritonga, F. Nasution and R. Watrianthos, "Comparison of machine learning algorithms for chest X-ray image COVID-19 classification", *Journal of Physics: Conference Series*, vol. 1993, 2021. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1933/1/012040. [Accessed 29 December 2021].

[6] P. Raikote, "COVID-19 Image Dataset", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/pranavraikokte/covid19-image-dataset. [Accessed: 23- Dec- 2021].

[7] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, no. 234, p. 241, 2015. Available: https://arxiv.org/abs/1505.04597. [Accessed 23 December 2021].

[8] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. Zaiane and M. Jagersand, "U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection", *Pattern Recognition Elsevier BV*, vol. 106, 2020. Available: https://arxiv.org/pdf/2005.09007v2.pdf. [Accessed 23 December 2021].

[9] A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan", *Jurnal Ilmiah Ilmu Komputer*, vol. 2, no. 1, pp. 22-26, 2016. Available: https://media.neliti.com/media/publications/283828-metode-naive-bayes-untuk-prediksi-kelulu-139fcfea.pdf. [Accessed 23 December 2021].

[10] L. Rokach and O. Maimon, "Decision Trees", *The Data Mining and Knowledge Discovery Handbook*, pp. 165-192, 2005. Available: https://www.researchgate.net/publication/225237661_Decision_Trees. [Accessed 27 December 2021].

[11] R. Stehling, M. Nascimento and A. Falcao, "A Compact and Efficient Image Retrieval Approach Based on Border/Interior Pixel Classification", *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean*, pp. 102-109, 2002. Available: https://www.researchgate.net/publication/221615004_A_compact_and_efficient_image_retrieval_approach_based_on_borderinterior_pixel_classification. [Accessed 26 December 2021].

[12] "Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM", *Jurnal Sistem Informasi Bisnis*, vol. 1, pp. 52-62, 2015. Available: https://www.researchgate.net/publication/304217190_Implementasi_Algoritma_K-Nearest_Neighbor_Sebagai_Pendukung_Keputusan_Klasifikasi_Penerima_Beasiswa_PPA_dan_BBM. [Accessed 27 December 2021].