BAREKENG: Journal of Mathematics and Its Applications
March 2023 Volume 17 Issue 1 Page 0467–0474
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss1pp0467-0474

PREDICTION OF THE POOR RATE K-MEANS AND GENERALIZED REGRESSION NEURAL NETWORK ALGORITHMS (CASE STUDY: NORTH SUMATRA PROVINCE)

Nita Suyani [1*], Arnita[2], Rinjani Cyra Nabila[3], Amanda Fitria[4]

[1,2,3] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Medan State University, Williem Iskandar St., Pasar V Medan Estate, Medan, 20221, Indonesia*

*Corresponding author's e-mail: * **arnita@unimed.ac.id***

Article History:

Received: 13th November 2022
Revised: 2nd February 2023
Accepted: 13th February 2023

Keywords:

Prediction;
Poverty Rate;
K-Means;
GRNN.

ABSTRACT

Poverty reduction is a crucial issue and the primary The North Sumatra Provincial government's main concern is lowering the poverty rate, which is a crucial issue. The Province of North Sumatra in Indonesia, one of many nations affected by the Covid-19 pandemic, is particularly troubled economically. In this study, poverty levels were mapped using the K-Means algorithm, and GRNN was then utilized for modeling and prediction. The data source used is time series data from 2010 to 2020 from the Central Statistics Agency (BPS), which includes variables X covering population, health, education, unemployment, and asset ownership and variable Y representing poverty level. The goal of this study is to choose the best model for estimating poverty levels in North Sumatra Province. The districts and cities of Deli Serdang and Medan have the greatest poverty rates, according to the K-means algorithm's mapping of poverty levels. Additionally, the prediction resultss produced MSE values of 0.004659 and RMSE values of 0.00002108. The value of the smoothness parameter is 0.01.

How to cite this article:

N. Suyani, Arnita, R. C. Nabila and A. Fitria., "PREDICTION OF THE POOR RATE K-MEANS AND GENERALIZED REGRESSION NEURAL NETWORK ALGORITHMS (CASE STUDY: NORTH SUMATRA PROVINCE)", *BAREKENG: J. Math. & App.*, vol. 17, iss. 1, pp. 0467-0474, March 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

The COVID-19 pandemic is a catastrophe that is currently affecting the entire world. The COVID-19 outbreak in Wuhan reportedly started on December 30, 2019, when the Wuhan Municipal Health Committee published a "urgent notification on the treatment of pneumonia of unknown source". All around the nation, the coronavirus is rapidly spreading. Acute respiratory syndrome coronavirus 2 (severe acute respiratory syndrome coronavirus 2) is the cause of COVID-19, an infectious disease (SARS-Cov-2) [1]. This epidemic is being propagated, primarily through physical or social exclusion. However, the general fall in economic activity has been impacted by this. The economy's aggregate supply and demand are affected by restrictions on community activities, which causes a decline in supply and demand. The status of those who merely stay at home results in a decline in the production and consumption sectors of the community, paralyzing the economy and decreasing public welfare [2].

According to the North Sumatra Province's Central Bureau of Statistics, the Covid-19 pandemic caused the region's poverty rate to rise to 1.343 million as of March 2021. With this rise, North Sumatra's poverty rate rose to 9.01 percent [2]. Low levels of education, poor health, few job possibilities, and isolating circumstances are the causes of poverty[3]. Low levels of education and technological proficiency, limited natural resources, rapid population increase, and adverse political stability are further factors that contribute to poverty kondusif [4]. The K-Means algorithm divides existing data into a K number of clusters or groups in a non-hierarchical manner [5]. In data mining applications, clustering plays a key role. The K-Means algorithm is a descriptive model. The K-Means algorithm is based on the principal grouping source or the centroid that is closest to the distribution of data [6]. The K-Means algorithm is a component of a distance-based clustering algorithm that separates data into many groups. The K-Means algorithm only functions on numerical qualities [7]. Large numbers can be clustered using this approach [8]. The original cluster center is what leads to the K-Means algorithm's shortcoming. The initial cluster center value provided at the start has a significant impact on the clusters that the K-means algorithm produces. so that the cluster's outcomes are localized ideal solutions. so that the cluster's outcomes are locally ideal answers[9].

There are many techniques to predict, including fuzzy, backpropagation, adaptive neuro-fuzzy inference system (ANFIS), k-nearest neighbor, artificial neural network, and many others. Artificial neural network algorithms are frequently employed to make forecasts or predictions[10]. One artificial neural network model that uses supervised training is the generalized regression neural network (GRNN), where the expected output is instructed to follow the training data output pattern[11]. The GRNN algorithm employs three determining factors throughout the learning process: spread, activation function, and various patterns [12]. Input, pattern, summarization, and output neurons are included in the GRNN's four processing levels [13].The Generalized Regression Neural Network (GRNN) algorithm contains a single parameter, which is known as the smoothing factor or spread [14].

Utilizing data mining is one method for categorizing the level of poverty. Large data stores can be searched through an automated process called data mining to find important information. Data mining's purpose is to forecast future data based on historical data [15]. This study will examine "Poverty Level Prediction Using the Combined K-Means Algorithm and Generalized Regression Neural Network" based on the aforementioned circumstances. In the coming term, it is hoped that this algorithm would help the government execute, enhance, and create policies on poverty.

2. RESEARCH METHODS

Secondary data were employed in this study, and they were sourced from https://www.bps.go.id. 2,541 total data points encompassing 33 regencies and cities in North Sumatra Province make up the time series data from 2010 to 2020. Two research variables—the independent variable and the dependent variable—were used in this study. Population (X1), health (X2), education (X3), unemployment (X4), and asset ownership (X5) are the independent variables. The dependent variable is the degree of poverty (Y).

2.1 Pre-processing Data

To produce a clean dataset for usage in the following stage, the data cleaning process will be carried out now. Data cleaning is done to get rid of duplicate information, look for inconsistencies, and fix mistakes. To improve the precision of the predictions, the data must then be transformed by normalizing it.

2.2 *K-Means Cluster* Process

The average of the variables X, X2, X3, X4, and X5 is what constitutes clustered data. Using the K-means approach, the training data is grouped. Data on population, health, education, unemployment, and asset ownership are used in the clustering process. The study has three clusters, with the first cluster being the highest (C1), the second being the middle (C2), and the third being the lowest (C3). The clustering procedure is as follows:

1. Decide how many groups you wish to search. The statistics on poverty levels are grouped into three clusters, designated C1, C2, and C3, with high, medium, and low values.
2. Identify the central axis (centroid). High clusters (C1), medium clusters (C2), and low clusters (C3) all have randomly chosen cluster sites (C3).
3. Utilizing **Equation (1)**, determine how far the data are from the cluster's centroid.

$$D(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + \cdots + (x_{ki} - x_{kj})^2} \quad (1)$$

Where:

$D(i,j)$ = distance of data i to cluster center j

x_{ki} = data to i in attribute j

x_{ji} = center point to j, at grab k

4. Repeat steps 2 - 4 until no more data is moving to another cluster.

2.3 Sharing of Training and Testing Data

The data is split into training and testing sets before modeling is done. Each year, 80% of the total data are used for training, while 20% are used for testing.

2.4 Creation of the GRNN Model

Using training data, GRNN modeling will be created in this method. Cross-validation will separate training data into learning data and data validation. The smoothing factor is then used in the Generalized Regression Neural Network method. based on the least Mean Square Error, the optimal smoothing factor parameter is chosen (MSE). The Generalized Regression Neural Network is calculated in the following steps:

1. Using the equation's Gaussian function, determine the value of the activation function for each unit pattern neuron (2).

$$\theta_i = e^{-(x-x_i)^T(x-x_i)/2\sigma^2}, \quad (2)$$

where:

X = input vector of the predictor variable for GRNN

X_i = training vector represented by pattern i neurons

σ = smoothing factor parameter

T = transpose operation of the vector

2. Calculating all inputs in the summation unit using two calculations: Ss computes the output pattern layer's arithmetic sum in **Equation (3)**, while Sw computes the output pattern layer's total weight and the interconnection weight of "w" in **Equation (4)**.

$$S_s = \sum_{i=1}^{n} \theta_i. \quad (3)$$

and,

$$S_w = \sum_{i=1}^{n} \theta_i w_i. \quad (4)$$

3. Using equation, calculate each input signal at the output unit by dividing the weighted sum by the total value of the activation function (5).

$$\hat{Y} = \frac{S_w}{S_s} \quad (5)$$

4. Select the GRNN model with the best smoothing factor and the lowest MSE value.

2.5 Testing the GRNN Model

As a testing procedure, the GRNN model with the best smoothing factor parameters is employed. Test the selected GRNN model with the best smoothing factor parameter. The Generalized Regression Neural Network is calculated between the testing data and the selected training data in this process using **Equation (2), Equation (3)**, and **Equation (4)**. The prediction results of the poverty rate of a region (district/city) for the following year will be obtained using **Equation (5)**. The accuracy and errors from this testing process will be calculated using MSE, which is mathematically stated in **Equation (6)**.

2.6 Using the GRNN algorithm for forecasting.

After the model has been tested, forecasting will be done in 2020 using the model that has been tested. The process's end goal is to produce predictions in 2021. The output results from the GRNN are denormalized in the final phase to make them real values. to evaluate the accuracy of forecasts.

2.7 Evaluation

In this study, an evaluation algorithm, namely Mean Square Error (MSE), is used to assess the performance of the poverty rate prediction model (6). MSE is used to assess the accuracy of previous forecasting results. The lower the Mean Square Error (MSE) value, the more accurate the forecasting results were.

$$MSE = \sum_{i=1}^{N} \frac{(x_i - f_i)^2}{N} \quad (6)$$

Where N is the amount of data predicted, x_i is the actual data, and f_i is the predicted data.

3. RESULTS AND DISCUSSION

3.1 The Clustering Of K-mean Poverty Rates By District/city

The first step in the research is to normalize the data. Furthermore, the clustering process averages all variables X 1 to X 5 across all years, so the cluster data only consists of 33 districts/cities based on each variable. The data is divided into three clusters: high (C1), medium (C2), and low (C3) (C3). To determine whether the determined data groups are optimal, the elbow method is used to determine the number of clusters to obtain the optimal K cluster value. **Figure 1** shows the Elbow method, also known as SSE (Sum Square error), for determining the optimal K cluster.

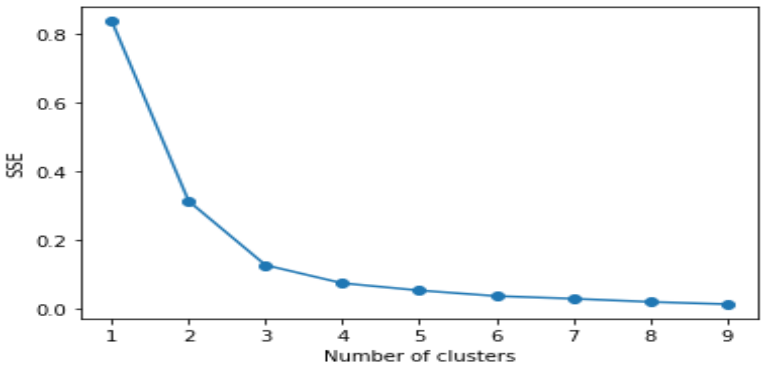

Figure 1. SSE graph for optimal K using the elbow method.

In **Figure 1** ,the optimal K lies at K=3, indicating that K cluster optimum terletak pada K=3, with a 49% improvement in model cluste. As a result, the centroid of each kelompok was determined at random. Nilai centroid can be found in **Table 1**.

Table 1. Random centroid values

Centroid to	*X1*	*X2*	*X3*	*X4*	*X5*
0	0,166144	0,807767	0,260712	0,272902	0,601954
1	0,889402	0,866110	0,629996	0,617926	0,721531
2	0,238882	0,784820	0,441249	0,157927	0,614965

Table 1, the groups have been divided into three, namely high cluster (C1), medium cluster (C2), and low cluster (C3) (C3). South Tapanuli, North Tapanuli, Toba Samosir, Humbang Hasundutan, Pakpak Bharat, Samosir, Sibolga, Pematang Siantar, Binjai, Padangsidempuan, and Gunungsitoli are among the 11 regencies/cities in cluster 0 (C3). Deli Serdang and Medan are the two regencies/cities in cluster 1 (C1). Nias, Mandailing Natal, Central Tapanuli, Labuhan Batu, Asahan, Simalungun, Dairi, Karo, Langkat, South Nias, Medium Bedagai, Batubara, North Padang Lawas, Padang Lawas, North Labuhan Batu, North Nias, West Nias, Tanjung Balai, and Tebing Tinggi are among the 20 regencies/cities in Cluster 2 (C2). The graph below shows the clustering result points.

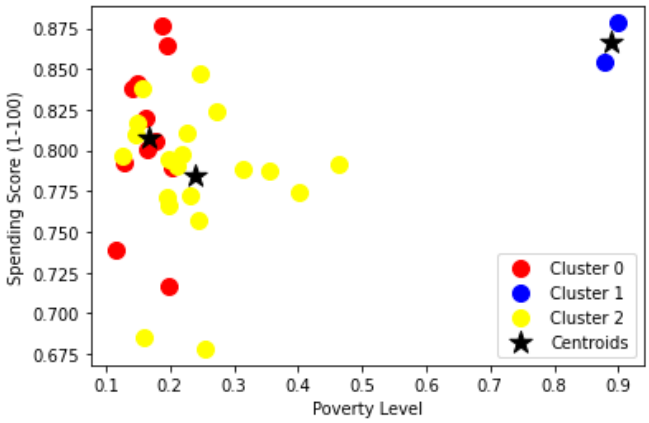

Figure 2. Graph of clustering result points

Figure 2 depicts the results of categorizing poverty levels by district/city. The red cluster is cluster 0 (C3), the blue cluster is cluster 1 (C1), and the yellow cluster is cluster 2 (C2).

3.2 GRNN Modeling and Testing Results

The next step is to create the GRNN network model. A cross-validation process is carried out before forming the training data model by dividing the training data into learning data and validation data. The K value used in this process is K = 10, and the cross-validation process is used to find the best smoothing parameters. According to [16], dividing the data fold 10 (K = 10) is the most commonly used in data mining research. According to research [17], this study used cross validation (K = 10) to achieve a high level of accuracy of 99.168%.

After that, run the training procedure with parameters ranging from 0.01 to 1.0. The number of neurons in the pattern layer is 290, which corresponds to the number of training data input vectors on the X1, X2, X3, X4, X5, and target variables are Y variables. This model formation process is used to select the best smoothing factor parameters and models. best based on MSE (Mean Square Error).

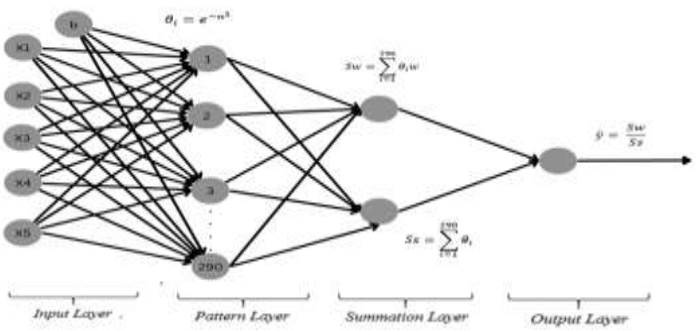

Figure 3. Architecture GRNN poverty rate forecasting

Figure 3 shows that there are 290 neurons in the pattern layer, which corresponds to the number of training data input vectors for each variable X1, X2, X3, X4, X5, and the target variable Y. The GRNN network's network architecture for the training stage is shown below. The following is a comparison of the MSEs of each spread parameter experiment, as shown in Table 2.

Table 2. MSE Values From Experimental Results 0.01 to 0.9

No	*Spread*	MSE
1	0,01	0,000032858
2	0,03	0,000371000
3	0,05	0,000908744
4	0,07	0,001361260
5	0,09	0,001883503
6	0,1	0,002212924
7	0,3	0,007247287
8	0,5	0,014713997
9	0,7	0,017465443
10	0,9	0,018527246

According to Table 2, the lowest MSE value derived from network training results occurs when network training employs a spread of 0.01, namely 0.00003285. The model is then tested by inputting 73 data points for each variable X1, X2, X3, X4, X5, and Y on the GRNN network that has been formed with a smoothing parameter of 0.01. The following is a comparison of the predicted and actual poverty rates using the GRNN network at the testing stage. As shown in Figure 3.4, the GRNN network test has an increase and decrease pattern, and the poverty rate predicted by the GRNN network almost matches the actual data. However, it has decreased to 0.0 in the 39th data point, which is indicated by an arrow. This prediction is performed with a smoothing parameter of 0.01 to yield an MSE of 0.005757 and an RMSE of 0.00003315.

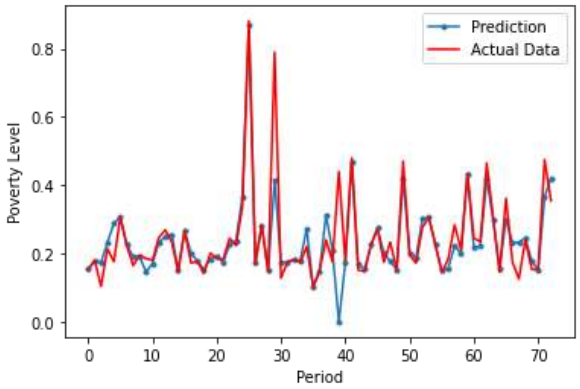

Figure 4. Graph of predicted GRNN

Figure 4 shows that the GRNN network test has an increase and decrease pattern, and the poverty rate predicted by the GRNN network almost matches the actual data. However, it has decreased to 0.0 in the 39th data point, which is indicated by an arrow. This prediction is performed with a smoothing parameter of 0.01 to yield an MSE of 0.005757 and an RMSE of 0.00003315.

3.3 Using the GRNN Algorithm to Forecast Poverty Levels

The next step is to forecast the poverty rate using a tried and true model. Forecasting is performed using 0.01 parameters on input data variables X1, X2, X3, X4, X5, and Y in 2020, with each data being data from 33 districts/cities in North Sumatra Province. After forecasting, there are 5 regencies/cities that will experience changes in poverty levels in 2021, including Toba Samosir Regency/City, which is expected to experience a slight increase of 17.55 people in 2021, compared to only 16.05 people in 2020. Labuhan Batu Regency/City is expected to lose 34.86 inhabitants by 2021, compared to 42.17 in 2020.

Asahan district/city is expected to increase by 73.64 people in 2021, compared to 66.32 people in 2020. Furthermore, it is estimated that the poverty rate in Langkat Regency/city in 2021 will be 73.64 people, a

significant decrease from the poverty rate of 101.87 people in 2020. Furthermore, the Regency/City has changed, with the Medan Regency/City expected to see a decrease of 18.35 people in 2021, compared to the total poverty rate in 2020, which was only 0.460274. The graph below depicts forecasting results using the GRNN network. This forecast has an MSE of 0.004659 and an RMSE of 0.00002108.

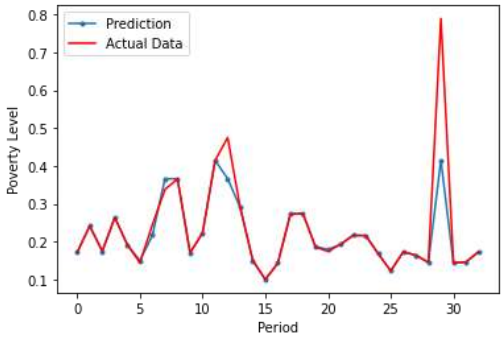

Figure 5. Forecasting the Poverty Rate in 2021

Figure 5 depicts the GRNN network forecasting graph. Although the range of values is slightly wider, the pattern of increase and decrease in the poverty rate predicted by the GRNN network closely resembles the actual data. This prediction was made with a smoothing parameter of 0.01 and yielded an MSE of 0.004659 and an RMSE of 0.00002108.

4. CONCLUSIONS

Based on the research and discussion that has been conducted in predicting the poverty rate using a combined k-means and GRNN algorithm, the following conclusions have been reached:

1. When the GRNNN model is built with the parameter 0.01, it produces a reasonably accurate prediction accuracy with an MSE of 0.005757 and an RMSE of 0.00003315
2. The k-means algorithm generates three poverty level groups with random centroid values. The application of k-means to the overall poverty rate data yields 24 districts/cities classified as low, 11 districts/cities classified as medium, and 2 districts/cities classified as high. whereas applying k-means to the poverty rate in 2021 yields 20 with a good model of 49%. Based on the training and testing process in the 2021 forecasting process using the GRNN model, 5 districts/cities experience increases and decreases, resulting in an MSE value of 0.004659 and an RMSE of 0.00002108.
3. It is estimated that in 2021, Toba Samosir, Asahan, Deli Serdang, Langkat, and Medan will be among the five districts/cities with the highest poverty rates.

.

REFERENCES

[1] S. Hanoatubun, "Dampak COVID – 19 Terhadap Perekonomian Indonesia," *J. Educ. Phychology an Couns.*, vol. 2, no. 1, pp. 146–143, 2020.

[2] F. Rizal and H. Mukaromah, "Filantropi Islam Solusi Atas Masalah Kemiskinan Akibat Pandemi Covid-19," *AL-MANHAJ J. Huk. dan Pranata Sos. Islam*, vol. 3, no. 1, pp. 35–66, 2021.

[3] M. Mulyadi, "Peran Pemerintah dalam Mengatasi Pengangguran dan Kemiskinan dalam Masyarakat," *J. Kaji.*, vol. 21, no. 3, pp. 221–236, 2016.

[4] R. S. Dewi and O. N. I. N. Irama, "Pengaruh Alokasi Dana Desa Terhadap Kemiskinan," *J. Akutansi Dan Bisnis*, vol. 4, no. 2, p. 11, Nov. 2018.

[5] R. Rosmini, A. Fadlil, and S. Sunardi, "Implementasi Metode K-Means Dalam Pemetaan Kelompok Mahasiswa Melalui Data Aktivitas Kuliah," *IT J. Res. Dev.*, vol. 3, no. 1, pp. 22–31, 2018.

[6] Y. R. Sari, A. Sudewa, D. A. Lestari, and T. I. Jaya, "Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer," *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 5, no. 2, pp. 192–198, 2020.

[7] W. M. P. Dhuhita, "Clustering Menggunakan Metode K-Means Untuk," *J. Inform.*, vol. 15, no. 2, pp. 160--174, 2016.

[8] N. D. Saksono, Y. A. Sari, and R. K. Dewi, "Rekomendasi Lokasi Wisata Kuliner di Jakarta Menggunakan Metode K-means Clustering dan Simple Additive Weighting," *J. Ilmu Komput. dan Sist. Inf.*, vol. 7, no. 1, pp. 14–21, 2019.

[9] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, Jan. 2019.

[10] E. Fatchurin, A. Fanani, and M. Hafiyusholeh, "Peramalan Penggunaan Bahan Bakar Pada Pembangkit Listrik Tenaga Gas Uap Menggunakan Metode Backpropagation Neural Network," *J. Ris. dan Apl. Mat.*, vol. 4, no. 2, pp. 1–8, 2020.
[11] M. Alkaff and Y. Sari, "Penerapan Generalized Regression Neural Networks untuk Memprediksi Produksi Padi Terhadap Perubahan Iklim," *J. Teknol. Rekayasa*, vol. 2, no. 2, p. 117, 2017.
[12] S. Herawati, "Peramalan Kunjungan Wisatawan Mancanegara Menggunakan Generalized Regression Neural Networks," *J. Infotel*, vol. 8, no. 1, 2016.
[13] R. Caraka, H. Yasin, and A. Prahutama, "Pemodelan General Regression Nneural Network (GRNN) Pada Data Return Indeks Harga Saham Euro 50," *J. GAUSSIAN*, vol. 4, no. 2, pp. 181–192, 2015.
[14] A. Harianti and N. Widiangga, "Analisis metode RBFNN dan GRNN pada peramalan mata uang EUR/USD," vol. 5, no. 1, pp. 83–90, 2022.
[15] Dewi, Sastradipraja, and Gustian, "Sistem Pendukung Keputusan Kenaikan Jabatan Menggunakan Metode Algoritma Naïve Bayes Classifier," *J. Teknol. dan Inf.*, vol. 11, no. 1, pp. 66–80, 2021.
[16] M. Adib, "Optimasi Parameter K Pada Algoritma Knn Untuk Klasifikasi Heregistrasi Mahasiswa," *J. IC-Tech*, vol. 10, no. 1, 2015.
[17] P. Rosyani, "Pengenalan Citra Bunga Menggunakan Segmentasi Otsu Treshold dan Naïve Bayes," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 1–7, Nov. 2020.