Jurnal Varian
Vol. 6, No. 2, April 2023, Hal. 157–166
e-ISSN: 2581-2017

K-Means Resilient Backpropagation Neural Network in Predicting Poverty Levels

Bobby Poerwanto[1]
[1]Universitas Negeri Makassar, Indonesia

Article Info

Article history:

Received : 02-07-2022
Revised : 05-06-2023
Accepted : 05-17-2023

Keywords:

Backpropagation Neural Network;
K-Means;
Poverty level;
Resilient Algorithm.

ABSTRACT

In solving economic problems, the government has implemented several development policies. However, this policy is considered to be too centered on big cities. So, through this research it is hoped that it can provide an overview related to regional groups that fall into the poorer category so that the government can also provide accelerated development policies that are oriented towards improving the economy of residents in the area. This study aims to determine the results of classifying district/city poverty levels in Indonesia as a basis for classification for predictions and to classify district/city poverty levels based on influencing factors. The method used in this study is K-Means Clustering using the poverty depth index and poverty severity index variables, then proceed with using the Backpropagation Neural Network (BNN) algorithm using the GRDP, per capita expenditure, human development index, and mean years of schooling. The results obtained using the K-Means algorithm are that there are 42 districts/cities that belong to cluster 1 where this region has a poverty index depth and severity index value that is higher than the 472 districts/cities in cluster 2. Furthermore, the cluster results are used as response variables for classification with BNN. The accuracy of the model obtained is very high, which is equal to 98.06, so the model is very feasible to be used as a poverty rate prediction model based on the variables used.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v6i2.2756

Corresponding Author:

Bobby Poerwanto,
Department of Statistics, Universitas Negeri Makassar.
Email: bobby_poerwanto@unm.ac.id

A. INTRODUCTION

Machine learning is a method of statistical analysis that is often used in research. There are two categories in machine learning that are quite popular, namely unsupervised learning and supervised learning. One of the unsupervised learning methods is cluster analysis. Cluster analysis is one of the methods used to group data based on certain characteristics (Barchitta et al., 2021) where there are several methods that are often used in this cluster analysis such as k-means (Meng et al., 2018), and fuzzy c- means (Heil et al., 2019; Poerwanto and Ali, 2019). These two methods are quite effective in grouping based on similarities and dissimilarities (Askari, 2021). However, k-means has the advantage of creating groups that are homogeneous within one group, and heterogeneous between groups (Butarbutar et al., 2017).

For supervised learning, one type is classification. There are several methods commonly used in classifying, such as logistic regression logistik (Bustan and Poerwanto, 2021; Tiro et al., 2021), support vector machines (Xu et al., 2020), and neural networks (Pawara et al., 2020). In terms of accuracy, the NN method is superior when the data has large dimensions (Poerwanto and Fajriani, 2020). Neural network (NN) is a method used for analysis related to predictions, both for metric and nonmetric data. In nonmetric data, for example, it is about classification (Zhang et al., 2021). Besides being superior in terms of data with large dimensions, one type of NN is also capable of solving nonlinear cases in classification with high accuracy, namely backpropagation neural network (BNN) (Zhang et al., 2021).

The two methods above, namely k-means and BNN will be integrated in predicting the poverty rate of districts/cities in Indonesia. The results of grouping using optimum k-means will be used as a basis for predicting the poverty rate using the BNN with a resilient algorithm. This poverty rate prediction consists of 2 groups of variables, namely the dependent and independent variables. The dependent variable used is the k-means result, namely the Regency/City group with high poverty rates and the low poverty rate group. The independent variables used are GRDP, per capita expenditure, human development index (IPM), and average length of schooling (Rusdarti and Sebayang, 2013).

In poverty classification research, several publications focus on households, as was done by Suparman (Suparman and Zainuddin, 2019) and (Annur, 2018) according to BPS standard. There is no poverty standard based on district or city. Therefore, the novelty in this study is an approach to districts/cities poverty classification using K-Means to make the clusters based on poverty indicators in Indonesia, namely the poverty depth index and poverty severity index. Meanwhile, BNN is used to predict poverty levels based on GRDP, per capita income, inflation, and education level variables. This study aims to determine the results of grouping poverty rates in districts/cities in Indonesia as a basis for classification for predictions and to determine the results of predicting poverty rates with the highest accuracy based on influencing factors, namely GRDP, per capita expenditure, human development index (IPM), and average length of schooling.

B. RESEARCH METHOD

The data used in this study is secondary data obtained online from the website of the Indonesian Central Statistics Agency (BPS), Provincial BPS in Indonesia, and Regency/City BPS in Indonesia. The variables used in this research are poverty depth and severity indices for cluster analysis, then for the neural network using GRDP, per capita expenditure, human development index (IPM), and average length of schooling, as well as cluster results as a response in 2019. The stages in this study can be seen in the Figure 1:

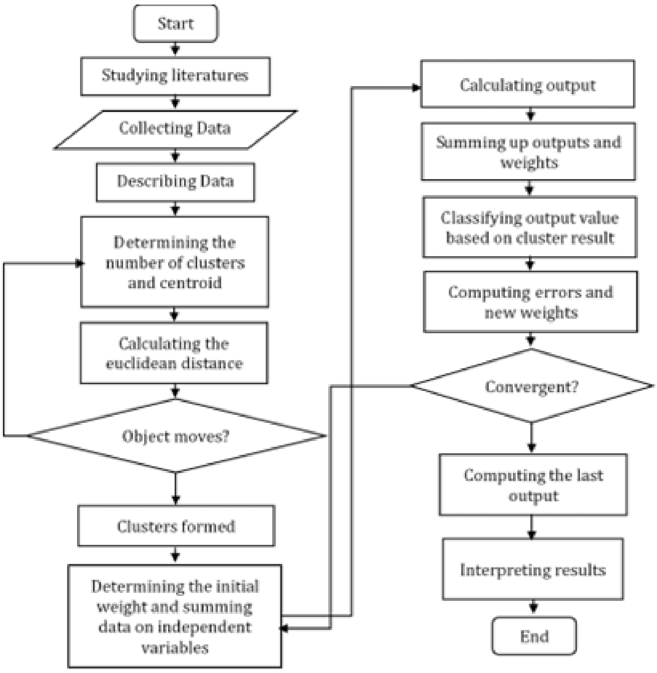
Start
Studying literatures
Collecting Data
Describing Data
Determining the number of clusters and centroid
Calculating the euclidean distance
Object moves?
Clusters formed
Determining the initial weight and summing data on independent variables
Calculating output
Summing up outputs and weights
Classifying output value based on cluster result
Computing errors and new weights
Convergent?
Computing the last output
Interpreting results
End

Figure 1. Flowchart of the study

The following are the research stages used in the two methods above, namely:

1. K-means algorithm

The algorithm in the k-means cluster (Kakushadze and Yu, 2017) is as follows:

1. Determine the number of clusters to be formed
2. Generate an initial cluster center (centroid) randomly

3. Calculate the distance of each data to the center of the cluster using the Euclidean Distance as in the Equation 1 below

$$d(x_i - x_j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2} \tag{1}$$

x_i, x_j is the two data that will be calculated the distance and p is the number of dimensions used.

4. Classify each data based on the closest distance to the cluster center
5. Determine the position of the new centroid using the equation below.

$$C_{m(q)} = \frac{1}{n_m} \sum_{i=1}^{n_m} x_{i(q)} \tag{2}$$

where:

$C_{m(q)}$	:	the center of the m-th cluster variable p
m	:	$1, 2, \ldots, k$
n_m	:	the number of objects in the mth group
k	:	number of cluster
q	:	$1, 2, \ldots, p$
$x_{i(q)}$	:	the observed value of the ith object of the q-th variable
i	:	$1, 2, \ldots, n$

2. Resilient Backpropagation Neural Network

The backpropagation network in the training process has three stages (Vishwakarma et al., 2020), namely:

1. The feedforward stage of the input
2. Calculation stage and backpropagation of the error
3. Stage of updating the weights and biases.

The algorithm of the BNN method is as follows.

Step 0 : Determine the weights

Step 1 : Each pair on the training data, perform steps 2-7.

feedforward

Step 2 : Each input is received by the input layer and forwarded to the hidden layer.

Step 3 : Each hidden layer adds up the weighting results at the input layer plus the bias.

Step 4 : Add up the multiplication results between the output layer and the weight, then add the bias.

Backpropagation of errors :

Step 5 : Each output layer receives the target pattern according to the input pattern.

Step 6 : Calculate on each hidden layer

Update weights and bias:

Step 7 : Each hidden layer updates the weights and biases so that they get new weights and biases.

Step 8 : Test the stop condition (already converged), then the iteration ends

To activate the function in the hidden layer, an activation function is needed. This activation function is a function that processes input that has been brought into the hidden layer in order to get the desired output. This activation function consists of several types, including linear, log-sigmoid, and threshold functions (Annas et al., 2021). This study will use the log-sigmoid activation function, with the following Equation 3:

$$y = \frac{1}{1 + e^{-x}} \tag{3}$$

Predictive accuracy can be seen from three criteria, namely sensitivity, specification, and accuracy. The following is the formula for each criterion (Hamed et al., 2013).

$$\begin{aligned} Sensitivity &= \frac{TP}{TP+FN} \\ Specification &= \frac{TN}{TN+FP} \\ Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \end{aligned}$$

where :

TP	:	True Positive
TN	:	True Negative
FP	:	False Positive
FN	:	False Negative

C. RESULTS AND DISCUSSION

1. Determination of the number of clusters

The data used in this study were taken from district/city data from 34 provinces in Indonesia, namely 514 districts/cities. Based on the attributes of the poverty depth index and poverty severity index, it was found that the optimum number of clusters used was 2 clusters. This analysis can be seen in the Figure 2 below:

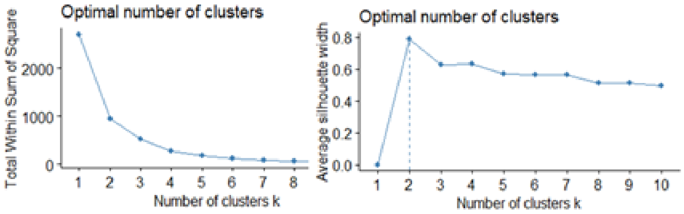

Figure 2. Elbow and Silhouette Index methods in determining the number of clusters

In Figure 2 above it can be seen that using either the elbow method or the Silhouette Index (SI) in determining the optimal number of clusters (Poerwanto, 2021), the results obtained are the same, namely 2 clusters.

2. Cluster Results

By running the K-Means algorithm on R-Studio, the following cluster results are obtained:

Table 1. Membership of each cluster and centroid

Cluster	Amount	Poverty Gap Index	Poverty Severity Index
1	42	7.79	2.81
2	472	1.54	0.372

The number of districts/cities in cluster 1 is 42, and approximately 11 times as many as in cluster 2. The number of iterations needed to achieve local optimum is 11 iterations. Table 1 also shows that the center of cluster 1 is at 7.79 for the poverty gap index and 2.81 for the poverty severity index, while for cluster 2 the average is 1.54 for the poverty gap index and 0.372 for the poverty severity index. This can be interpreted that cluster 1 is the regions with higher poverty rates.

Table 2. Districts/cities Cluster 1

No	Province	District/City	No	Province	District/City
1	Riau	Kepulauan Meranti	22	Papua	Nabire
2	NTB	Lombok Utara	23	Papua	Kepulauan Yapen
3	NTT	Sumba Barat	24	Papua	Biak Numfor
4	NTT	Sumba Timur	25	Papua	Paniai
5	NTT	Timor Tengah Selatan	26	Papua	Puncak Jaya
6	NTT	Lembata	27	Papua	Mimika
7	NTT	Rote Ndao	28	Papua	Asmat
8	NTT	Sumba Tengah	29	Papua	Yahukimo
9	NTT	Sabu Raijua	30	Papua	Pegunungan Bintang
10	Maluku	Seram Bagian Barat	31	Papua	Tolikara
11	Maluku	Maluku Barat Daya	32	Papua	Waropen
12	Papua Barat	Fakfak	33	Papua	Supiori
13	Papua Barat	Teluk Wondama	34	Papua	Mamberamo Raya
14	Papua Barat	Teluk Bintunni	35	Papua	Nduga
15	Papua Barat	Manokwari	36	Papua	Lanny Jaya
16	Papua Barat	Sorong	37	Papua	Mamberamo Tengah
17	Papua Barat	Tambrauw	38	Papua	Yalimo
18	Papua Barat	Maybrat	39	Papua	Puncak Jaya
19	Papua Barat	Manokwari Selatan	40	Papua	Dogiyai
20	Papua Barat	Pegunungan Arfak	41	Papua	Intan Jaya
21	Papua Jayawijaya		42	Papua	Deiyai

Table 2 shows that there are 42 regencies/cities in cluster 1 where in this cluster the poverty gap index and poverty severity index values are high. Of the 42 regencies/cities in cluster 1, only 1 district is from the western part of Indonesia, namely Meranti District from Riau Province. This is supported by research conducted by Chalid and Yusuf (Chalid and Yusuf, 2014) with the results of a study stating that Meranti District is a district in Riau Province with the lowest human development index value and the highest poverty rate.

Visually, the mapping of district/city clusters in Indonesia (Annas et al., 2022) can be seen in Figure 3 .

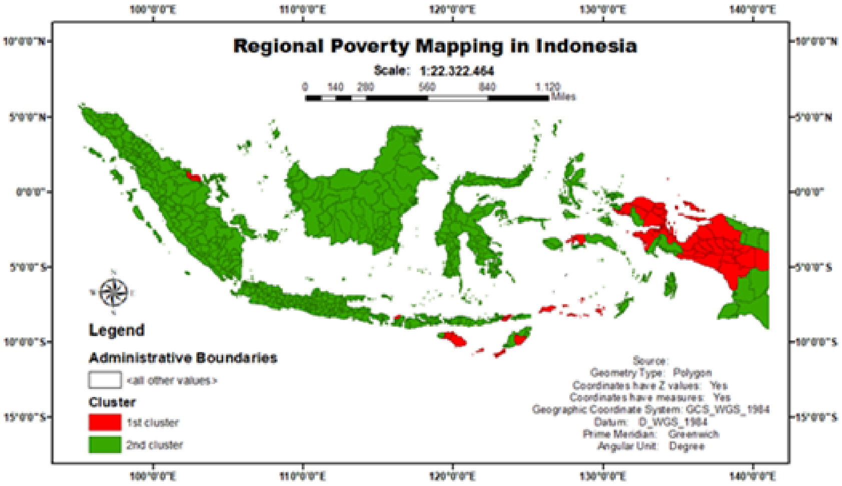

Figure 3. Regional Poverty Mapping in Indonesia

As can be seen in Table 2 and Figure 3, districts/cities in Java Island, Sulawesi Island are all included in the green zone or cluster 2 and for Papua Island where there are West Papua and Papua Provinces, there are only 4 districts/cities in West Papua and 7 Regencies/Cities in Papua Province which are included in cluster 2.

DOI: https://doi.org/10.30812/varian.v6i2.2756

3. Classification Results

Poverty rate prediction based on districts/cities in Indonesia with the independent variables GRDP, per capita expenditure, human development index (IPM), and average length of schooling using the backpropagation neural network (BNN) method with a resilient algorithm. Resilient has a faster calculation process than its predecessor algorithm, namely the backpropagation algorithm. The data in this study is divided into two parts. The first is training data, which is a dataset that is used to form a prediction model, and the second is data testing, which is a dataset that is used to evaluate the prediction model that is formed.

There are 2 scenarios for dividing the datasets in this study, namely scenario 1 with a ratio of training data and data testing is 70:30, and scenario 2 with a ratio of 80:20. From these two scenarios it can be seen which scenario gives better prediction results, seen from the accuracy of the classification of poverty levels which consist of 2 categories, namely 1 for high poverty levels and 2 for low poverty levels. The following are the results of the analysis of each scenario and their comparison.

a. Scenario 1 (training 70: testing 30)

Prediction of poverty rates based on districts/cities in Indonesia on a comparison of training and testing data of 70:30 uses all independent variables with the logistic activation function, and uses a hidden layer of 5 nodes. The iterations used to obtain optimum results in model building on the training data are 162 times with an error in the resulting optimum iteration of 7.94. The visualization of BNN architecture with resilient algorithms on training data as a model for prediction is as follows:

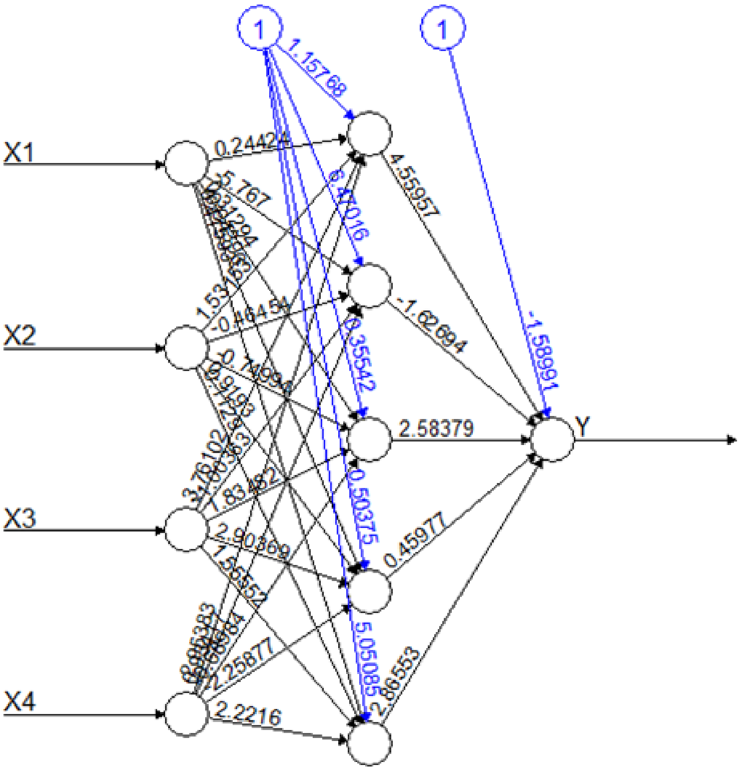

Figure 4. BNN Architecture in Scenario 1

The accuracy results obtained from the training data as a model formation for predictions are show in Table 3:

Table 3. Confusion Matrix for Training Data in Scenario 1

		Prediction	
		1	**2**
Actual	**1**	11	18
	2	1	328

The resulting accuracy based on the confusion matrix is 94.43%. This shows that by using the independent variables GRDP, Per Capita Expenditure, HDI, and average length of schooling, the prediction results correctly classify high and low poverty rates in training data or the process of building a model for predictions of 94.43% of 359 Regencies/Cities. While the results of the accuracy of the data testing used as validation of the formed model is 98.06%. This accuracy is obtained from the confusion matrix in Table 4.

Table 4. Confusion Matrix for Testing Data in Scenario 1

		Prediction	
		1	**2**
Actual	**1**	11	2
	2	1	141

The accuracy obtained in the testing data shows that there is a 1.94% poverty rate that is not correctly classified into the categories of high poverty rates and low poverty rates out of 155 Regencies/Cities, or in other words there are only 3 Regencies/Cities misclassified predictions.

b. Scenario 2 (training 80: testing 20)

The prediction of poverty rates based on districts/cities in Indonesia in scenario 2 is also in the same stages as scenario 1, namely using all independent variables with the logistic activation function, and using a hidden layer of 5 nodes, the difference is the ratio of training and testing data, which is 80: 20. The iterations used to obtain optimum results in model building on the training data are 143 times with an error in the resulting optimum iteration of 7.81. The visualization of BNN architecture with resilient algorithms on training data as a model for prediction is as follows:

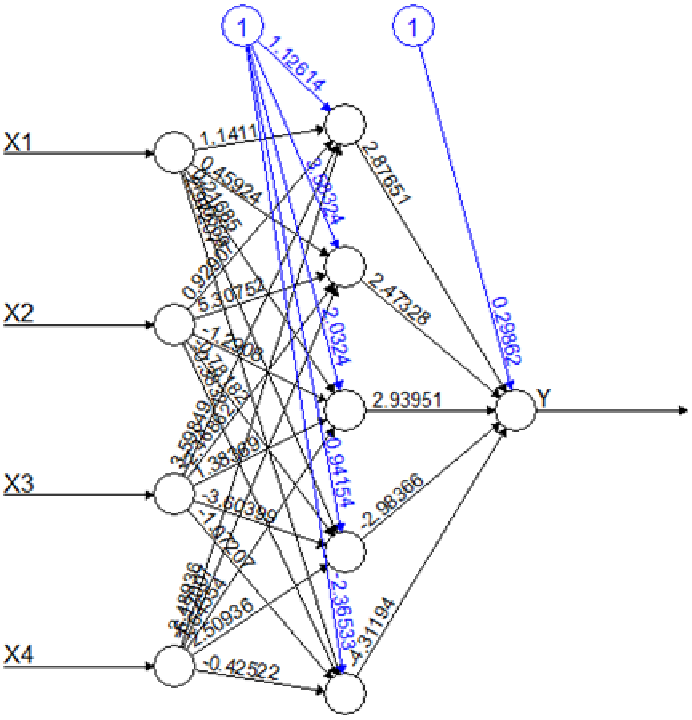

Figure 5. BNN Architecture in Scenario 2

Iteration in scenario 2 is faster when compared to scenario 1 to achieve optimum results, which is the difference of 19 iterations. The accuracy results obtained from the training data as a model formation for predictions are as follows:

Table 5. Confusion Matrix for Training Data in Scenario 2

		Prediction	
		1	**2**
Actual	**1**	17	17
	2	5	373

The resulting accuracy based on the confusion matrix is 94.66%. This shows that by using the independent variables GRDP, Per Capita Expenditure, HDI, and average length of schooling, the prediction results correctly classify high and low poverty rates in training data or the process of building a model for predictions of 94.66% of 412 Regencies/Cities. While the results of the accuracy of the data testing used as validation of the formed model is 92.16%. This accuracy is obtained from the confusion matrix in Table 6 below.

Table 6. Confusion Matrix for Testing Data in Scenario 2

		Prediction	
		1	**2**
Actual	**1**	8	0
	2	8	86

The accuracy obtained from the testing data shows that there is a 7.84% poverty rate that is not correctly classified into the categories of high poverty rates and low poverty rates out of 102 Regencies/Cities, or in other words there are 8 Regencies/Cities misclassified predictions.

c. Comparison between Scenario 1 and Scenario 2

The results of the accuracy comparison for scenarios 1 and 2 on the prediction of the Poverty Level with variables that are thought to influence changes in the poverty rate in the form of GRDP, per capita expenditure, HDI, and average length of schooling are:

Table 7. Accuracy Comparison

	Scenario	Accuracy
1	(70:30)	98.06%
2	(80:20)	92.16%

The two scenarios used in this analysis give very high accuracy, namely 98.06% for scenario 1 and 92.16% for scenario 2. So it can be said that by using 4 independent variables that are thought to influence the accuracy of classification the poverty rate is very high. However, from these two scenarios the one that gives better results is scenario 1, namely by using a 70:30 ratio with a classification accuracy of 98.06

The results of this study are also better when compared to the research conducted by Poerwanto and Fajriani (Poerwanto and Fajriani, 2020) in terms of higher accuracy and coverage of the classification area. In the previous study, the area classified was only Sulawesi Island, whereas in this study it was conducted in all districts and cities in Indonesia.

D. CONCLUSION AND SUGGESTION

In the grouping carried out using the K-Means algorithm, the results showed that there were 42 regencies/cities in the cluster with higher P1 and P2 values compared to the 472 regencies/cities in cluster 2. Furthermore, the results of the classification analysis using the National Narcotics Agency obtained accurate results that on data testing that is equal to 98.06% in scenario 1 and 92.16% in scenario 2 where these two results are very good. Thus, the model obtained in this study is feasible to be used as a model to predict the poverty rate of districts or cities in Indonesia based on the variables used.

REFERENCES

Annas, S., Aswi, A., Abdy, M., and Poerwanto, B. (2021). Stroke classification model using logistic regression. In *Journal of Physics: Conference Series*, volume 2123, page 012016. IOP Publishing.

Annas, S., Poerwanto, B., Sapriani, S., et al. (2022). Implementation of k-means clustering on poverty indicators in indonesia. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 21(2):257–266.

Annur, H. (2018). Klasifikasi masyarakat miskin menggunakan metode naive bayes. *ILKOM Jurnal Ilmiah*, 10(2):160–165.

Askari, S. (2021). Fuzzy c-means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. *Expert Systems with Applications*, 165:113856.

Barchitta, M., Maugeri, A., Favara, G., Riela, P., La Mastra, C., La Rosa, M., San Lio, R. M., Gallo, G., Mura, I., Agodi, A., et al. (2021). Cluster analysis identifies patients at risk of catheter-associated urinary tract infections in intensive care units: findings from the spin-uti network. *Journal of Hospital Infection*, 107:57–63.

Bustan, M. and Poerwanto, B. (2021). Logistic regression model of relationship between breast cancer pathology diagnosis with metastasis. In *Journal of Physics: Conference Series*, volume 1752, page 012026. IOP Publishing.

Butarbutar, N., Windarto, A. P., Hartama, D., and Solikhun, S. (2017). Komparasi kinerja algoritma fuzzy c-means dan k-means dalam pengelompokan data siswa berdasarkan prestasi nilai akademik siswa. *Jurasik (Jurnal Riset Sistem Informasi dan Teknik Informatika)*, 1(1):46–55.

Chalid, N. and Yusuf, Y. (2014). Pengaruh tingkat kemiskinan, tingkat pengangguran, upah minimum kabupaten/kota dan laju pertumbuhan ekonomi terhadap indeks pembangunan manusia di provinsi riau. *Jurnal ekonomi*, 22(2):1–12.

Hamed, A. A., Li, R., Xiaoming, Z., and Xu, C. (2013). Video genre classification using weighted kernel logistic regression. *Advances in Multimedia*, 2013:2–2.

Heil, J., Häring, V., Marschner, B., and Stumpe, B. (2019). Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: A case study with west african soils. *Geoderma*, 337:11–21.

Kakushadze, Z. and Yu, W. (2017). * k-means and cluster models for cancer signatures. *Biomolecular detection and quantification*, 13:7–31.

Meng, Y., Liang, J., Cao, F., and He, Y. (2018). A new distance with derivative information for functional k-means clustering algorithm. *Information Sciences*, 463:166–185.

Pawara, P., Okafor, E., Groefsema, M., He, S., Schomaker, L. R., and Wiering, M. A. (2020). One-vs-one classification for deep neural networks. *Pattern Recognition*, 108:107528.

Poerwanto, B. (2021). Evaluating the k-means analysis in clustering area based on estates productivity in tana luwu using silhouette index. In *Journal of Physics: Conference Series*, volume 1752, page 012014. IOP Publishing.

Poerwanto, B. and Ali, B. (2019). Implementasi algoritma fuzzy c-means dalam mengelompokkan kecamatan di tana luwu berdasarkan produktifitas hasil perkebunan. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 19(1):163–172.

Poerwanto, B. and Fajriani, F. (2020). Resilient backpropagation neural network on prediction of poverty levels in south sulawesi. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, 20(1):11–18.

Rusdarti, R. and Sebayang, L. K. (2013). Faktor-faktor yang mempengaruhi tingkat kemiskinan di provinsi jawa tengah. *Jurnal Economia*, 9(1):1–9.

Suparman, P. and Zainuddin, A. (2019). Implementasi metode k-nearest neighbor untuk klasifikasi penduduk miskin di desa ngemplak kidul kabupaten pati jawa tengah. *Jurnal Informatika SIMANTIK*, 4(1):21–28.

Tiro, M., Poerwanto, B., and Fahmuddin, M. (2021). Logistics regression modelling on student career path choices at the statistics department, fmipa unm makassar. In *Journal of Physics: Conference Series*, volume 2123, page 012002. IOP Publishing.

Vishwakarma, G. K., Paul, C., and Elsawah, A. (2020). An algorithm for outlier detection in a time series model using backpropagation neural network. *Journal of King Saud University-Science*, 32(8):3328–3336.

Xu, J., Tan, W., and Li, T. (2020). Predicting fan blade icing by using particle swarm optimization and support vector machine algorithm. *Computers & Electrical Engineering*, 87:106751.

Zhang, L., Wei, Y., and Chu, E. K.-w. (2021). Neural network for computing gsvd and rsvd. *Neurocomputing*, 444:59–66.