Analytics-Based on Classification and Clustering Methods for Local Community Empowerment in Indonesia

Dyah Yuniati[1,2] and Kristina Pestaria Sinaga[2](✉)

[1] Dkatalis, Jakarta, Indonesia
dyah.yuniati@binus.ac.id
[2] Department of Master in Information System Management, Bina Nusantara University, Jakarta, Indonesia
kristina.sinaga@binus.edu

Abstract. West Papua is reportedly the second-most populous province in Indonesia. The United Nations International Children's Emergency Fund (UNICEF) highlights Papua's performance in selecting the Sustainable Development Goals (SDG) indicators compared to other provinces in the country. The data shows that food, nutrition, health, education, housing, water, sanitation, and protection are defined as multidimensional child poverty. Population statistics and poverty figures show that inter-provincial equity in Indonesia needs to be re-measured. In 2008, the Regional Governments of Papua and West Papua Provinces implemented a Community Empowerment Program called "PNPM RESPEK", which provided direct community assistance for IDR 100 million per village. To determine the people's level of understanding and perception towards this program, PNPM RESPEK, in collaboration with the Central Statistics Agency, conducted an integrated PNPM RESPEK Evaluation Survey in July 2009. Based on the survey results, this paper identifies a model (pattern) of understanding the people of Papua and West Papua towards the program and finds the best method to build this model through classification techniques. Then the data model was also tested using unsupervised learning, the clustering method. The experimental results show that the J48 decision tree produces the highest accuracy compared to the others. As for clustering, the clustering hierarchy provides the best accuracy. Decision Tree J48 has the best accuracy with an accuracy of 97.31%. In this case, 97.31% of the people of Papua and West Papua who receive direct community assistance meet the level of understanding and perception of the PNPM RESPEK Program.

Keywords: SDG · Poverty · Equality · Machine learning · J48 Decision Tree · Hierarchical clustering

1 Introduction

Indonesia's easternmost provinces of Papua and West Papua generally referred to as Papua, are the country's most violent and resource-rich areas [1]. However, health care

A. Mohamed et al. (Eds.): SCDS 2021, CCIS 1489, pp. 133–145, 2021.
https://doi.org/10.1007/978-981-16-7334-4_10

standards are lower in West Papua than in other regions of Indonesia [2]. World Health Organization (WHO) reported that poverty is a significant cause of ill health and a barrier to accessing health care when needed. This relationship is financial: the poor cannot afford to purchase things needed for good health, including sufficient quality food and health care. However, the relationship is also related to other factors related to poverty, such as lack of information on appropriate health-promoting practices or lack of voice needed to make social services work for them [3].

In 2007, the Government of Indonesia launched the Mandiri National Program for Community Empowerment (PNPM), which aims to reduce poverty, strengthen local government and community institutions' capacity, and improve local government governance. In 2008, this program covered approximately 40,000 villages in Indonesia and was expected to cover nearly 80,000 villages by 2009 [4]. In line with this, the regional governments of Papua and West Papua Provinces in 2008 implemented a Community Empowerment Program called "PNPM RESPEK." RESPEK is funded by the Provincial Expenditure Budget (APBD Propinsi), and it provides 100 million IDR directly to every village in the province [1]. The Regional Governments of Papua and West Papua provide direct community assistance (Indonesian: Bantuan Langsung Masyarakat) of IDR 100 million per village for 3,923 villages in 388 sub-districts. Meanwhile, the Ministry of Home Affairs provides more than 1,000 facilitators through PNPM [4].

The main component of PNPM is its approach called Community-Driven Development (CDD) [5]. Adopting a community-driven development (CDD) approach and with technical financial assistance from the International Bank for Reconstruction and Development, the PNPM is now a national program covering all villages and cities in the country [6, 7]**.** To determine the level of understanding and perception of Papua and West Papua's people towards the PNPM RESPEK Program, PNPM RESPEK, in collaboration with BPS-Statistics Indonesia, conducted the PNPM RESPEK Evaluation Survey, which was integrated through the National Socio-Economic Survey (SUSENAS) in July 2009. This research aims to identify a model (pattern) for understanding the people of Papua and West Papua towards the program and find the best method for building this model through experimental classification and clustering techniques. The primary goals of this research are to help the PNPM RESPEK improving the remote area from a data perspective and understanding principles of extracting valuable knowledge from data.

2 Methodology

Data Mining is the process of extracting and identifying patterns from large sets of data to produce output in the form of useful information or knowledge that was not previously known manually on the raw data. Data mining is carried out using statistical methods, mathematical algorithms, artificial intelligence or machine learning. In general, the stages carried out in Data Mining include data selection, pre-processing data, transformation data, modeling, and interpretation data [8].

2.1 Classification

Classification is the supervised learning technique in data mining. In supervised learning, the data label has already been defined. Classification is used to classify each item in a data set into one of a predefined set of classes or groups. The data analysis task classification is where a model or classifier is constructed to predict class labels. So, the classification technique will assign items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Algorithms for classification include J48 [9] and Logistic Regression [10].

- The J48 algorithm is an algorithm derived from C4.5 [10]. This algorithm generates decision trees based on rules to classify. Each aspect of information is divided into several small subsets to form the basis of decisions. J48 looks at standard data, which results in the separation of information by selecting attributes [11]. Mathematically, J48 algorithm uses the concept of entropy and information gain (IG). The information gain rate (IGR) is the splitting criterion (SplitInfo) to make the J48 decision tree. The IG, SplitInfo, and IGR are formulated as follows

$$IG(S,j) = Entropy(S) - Entropy(S\lfloor j) \tag{1}$$

$$SplitInfo_j(S) = -\sum_{k=k_0}^{k_c}\left(\frac{\left|S_j(k)\right|}{|S|} + \log_2\frac{S_j(k)}{S}\right) \tag{2}$$

$$IGR(j) = \frac{IG(S,j)}{SplitInfo_j(S)} \tag{3}$$

where S is a parent node, j represent the j-th attribute of sample x in one class label, $Entropy(S) = -\sum_{j=1}^{d} x_j \log_2 x_j$, $Entropy(S\lfloor j)$ is the conditional entropy with $Entropy(S\lfloor j) = \sum_{k=k_0}^{k_c} \frac{|S_j(k)|}{|S|} \cdot Entropy\left(S_j(k)\right)$, and $S_j(k) = \left\{x \in S | x_j = k\right\}$.

- Logistic regression is an approach to creating predictive models using equations that describe the relationship between two or more variables [13]. The dependent variable for logistic regression has a dichotomy scale. The dichotomy scale is a nominal data scale with two categories: Yes and No, Success and Failure or High and Low [12]. We often named these two categories as binary-valued labels which the correct label y values is denoted either 0 or 1 $\left(y^{(i)} \in \{0, 1\}\right)$. Mathematically, the probability that data samples belong to the "Yes" class versus the probability that it belongs to the "No" class defined as follows

$$P(y = Yes|x) = h_\theta(x) = \frac{1}{1 + \exp\left(-\theta^T x\right)} \equiv \sigma\left(\theta^T x\right) \tag{4}$$

$$P(y = No|x) = 1 - P(y = Yes|x) = 1 - h \tag{5}$$

where $\sigma(r) = \frac{1}{1+\exp(-r)}$ is the sigmoid or logistic function, and $\theta^T x \in [0, 1]$ is the gradient for linear regression. The cost function for a set of training examples with binary labels $\left\{\left(x^{(i)}, y^{(i)}\right) : i = 1, 2, \ldots, n\right\}$ to measure how close a given h_θ to the correct output y is expressed as below

$$J(\theta) = -\sum_{i=1}^{n}\left(y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right) \tag{6}$$

If we plug in the definition of $h_\theta(x) = \sigma\left(\theta^T x^{(i)}\right)$ into (6), we will get the loss function as below

$$J(\theta) = -\sum_{i=1}^{n}\left(y^{(i)}\log\left(\sigma\left(\theta^T x^{(i)}\right)\right) + \left(1 - y^{(i)}\right)\log\left(1 - \sigma\left(\theta^T x^{(i)}\right)\right)\right) \quad (7)$$

To be noted, the smaller the values of cost function the better the model. In this sense, the model with bigger cost function clearly predict the un-great solution of $y^{(i)}$.

2.2 Clustering

Clustering is a powerful tool in data analysis. It is used for discovering the cluster structure in data sets with the most remarkable similarity within the same cluster but the most noteworthy dissimilarity between different clusters. Generally, cluster analysis became a multivariate statistical analysis branch, and it is an unsupervised learning approach to machine learning [13, 14]. Algorithms for clustering include K-Means [15], Hierarchical clustering (HCA) [16, 17], and DBSCAN algorithms [18].

- K-Means
  K-means is the simplest and most common clustering method. It is because K-means can classify large amounts of data with fast and efficient computation time. K-Means divides n data points in d dimensions into a number of k clusters where the clustering process is carried out by minimizing the sum squares distance between the data and each cluster center [15]. In its implementation, the K-Means method requires three parameters that are entirely user-defined, namely the number of clusters (# of *k)*, cluster initialization and system distance. The objective function of K-Means is formulated as

$$J_{K-Means}(U, V) = \sum_{k=1}^{c}\sum_{i=1}^{n}\mu_{ik}\sum_{j=1}^{d}\left(x_{ij} - v_{kj}\right)^2 \quad (8)$$

$$s.t.,\ \mu_{ik} \in \{0, 1\},\ i = 1, \ldots, n,\ k = 1, \ldots, c \quad (9)$$

The objective function in (8) is optimized by using the Lagrange multipliers and obtained the updating equations of μ_{ik} and v_{kj} as follows

$$v_{kj} = \sum_{i=1}^{n}\mu_{ik}x_{ij} \Bigg/ \sum_{i=1}^{n}\mu_{ik} \quad (10)$$

$$\mu_{ik} = \begin{cases} 1 \ \ if \ \sum_{j=1}^{d}\left(x_{ij} - v_{kj}\right)^2 = \min_{1\le k\le c}\left(\sum_{j=1}^{d}\left(x_{ij} - v_{kj}\right)^2\right) \\ 0, \ \text{otherwise.} \end{cases} \quad (11)$$

- Hierarchical clustering
  Hierarchical clustering (HCA) groups data through a hierarchical chart [16]. In the initial step, the hierarchical clustering identifies the data that has the closest distance,

then associated it into one cluster. Furthermore, hierarchical clustering calculates the distance between the clusters [17]. There are seven hierarchical clustering methods including single link, complete link, group average link, McQuitty's method, median, centroid, and Ward's method. These seven hierarchical clustering methods defines a new relation from datasets to hierarchies by using different *Lance-Williams dissimilarity update formula*. However, the suitable iteration among these seven hierarchical clustering methods similar to each other, they are all carried out until all are connected. Mathematically, if points x and y are agglomerated into cluster $x \cup y$, then the *Lance-Williams dissimilarity update formula* is expressed as below:

$$d(x \cup y, k) = \alpha_x d(x, k) + \alpha_y d(y, k) + \beta d(x, y) + \gamma |d(x, k) - d(y, k)| \tag{12}$$

where α_x, α_y, β, γ define the agglomerative criterion, α_y with index y is defined identically to coefficient α_x with index x. The formulation of *Lance-Williams dissimilarity* in (12) can be expressed as follow

$$d_{x\cup y,k} = \alpha_x d_{xk} + \alpha_y d_{yk} + \beta d_{xy} + \gamma \left| d_{xk} - d_{yk} \right| \tag{13}$$

- Density-Based Spatial Clustering of Application with Noise
  Density-Based Spatial Clustering of Application with Noise (DBSCAN) is a clustering algorithm developed by density-based. DBSCAN separates high-density clusters from low-density clusters. This algorithm will start by dividing the data into d dimensions, then iteratively count the number of data points close to each other [18]. The DBSCAN relay on two parameters, called MinPts and Epsilon ($\varepsilon - neighborhood$). The $\varepsilon - neighborhood$ is a distance measure that will be used to locate the points or to check the density in the neighbourhood of any point x, formulated as

  $$N_\varepsilon(x) = \{y \in d \,|\, \|y - x\| \le \varepsilon\} \tag{14}$$

  Here points x is a points inside of the cluster (MinPts) if the $\varepsilon - neighborhood$ $N_\varepsilon(x)$ of point x greater than or equal to the least number of neighbors v, denoted as $|N_\varepsilon(x)| \ge v$. A point x is directly density-reachable from a point y with respect to $\varepsilon - neighborhood$ and the minimum number of points required to form a dense region if $x \in N_\varepsilon(y)$, and $|N_\varepsilon(y)| \ge$ MinPts.

3 Data Set

This research takes a case study to identify a model for the understanding of the people of Papua and West Papua towards the National Program for Community Empowerment, Strategic Plan of "Kampung" Development (PNPM RESPEK) [7]. The dataset was obtained from the results of the PNPM RESPEK survey in collaboration with BPS-Statistics Indonesia to conduct the PNPM Evaluation Survey, which was integrated through the National Socio-Economic Survey (Susenas) July 2009. The source dataset is openly accessible at https://microdata.worldbank.org/index.php/catalog/1801/ study-description.

This data initially contains 3937 Papua and West Papua people who received support from PNPM RESPEK. Since there are 2041 missing values, the data we used in this research only contains 1896 samples of Papua and West Papua people who have benefited from the PNPM RESPEK project, with 31 attributes. Table 1 shows the data type for each attribute. As our goal is to identify the accuracy of classifiers in predicting people who are likely to get the understanding and perception towards the PNPM RESPEK program, we measured the performances by only using a single evaluation metric, called accuracy rates. It is a common known that accuracy rate is devoted to simultaneously visualize and associate the structure of data based on their similarities. Furthermore, we notice that the high accuracy rate is more important than the resources. The calculation of accuracy rate is based on the percentage of error, expressed as

$$Error\ rate = 1 - \sum\nolimits_{k=1}^{c} n(c_k) \Big/ n \tag{15}$$

where $n(c_k)$ is the number of training data that obtain correct classification/clustering. All the procedures for classification and clustering including pre-processing and processing final input data will be done by using Waikato environment for knowledge analysis (WEKA).

Table 1. Data type of the attributes

No	Characteristics		
	Attribute	*Data Type*	*Values*

1	Age	Numeric	Min = 11
			Max = 99

2	Gender	Nominal	1 = Male 2 = Female
3	Marital Status	Nominal	1 = Single 2 = Married 3 = Divorced 4 = Death divorce
4	Heard about PNPM	Nominal	1 = Yes, spontaneous 2 = Yes, after the interviewer explained 3 = No
5	Has been a PNPM actor	Nominal	1 = Yes 2 = No
6	Previously worked at PNPM	Nominal	1 = Yes 2 = No
7	Present at the PNPM meeting	Nominal	1 = Yes 2 = No
8	Become an SPP member	Nominal	1 = Yes 2 = No

(*continued*)

Table 1. (*continued*)

No	Characteristics		
	Attribute	*Data Type*	*Values*
9	Know the number of funds budgeted	Nominal	1 = Yes 2 = No
10	Information from "Dusun" or "RT" meetings	Nominal	1 = Yes 2 = No
11	Information from the community of mothers	Nominal	1 = Yes 2 = No
12	Information from community group meetings	Nominal	1 = Yes 2 = No
13	Information from friends/neighbors by verbal	Nominal	1 = Yes 2 = No
14	Information from the village apparatus met on the road	Nominal	1 = Yes 2 = No
15	Information from "Kampung" officials who came to the house	Nominal	1 = Yes 2 = No
16	Information from community leaders	Nominal	1 = Yes 2 = No
17	Information from religious figures	Nominal	1 = Yes 2 = No
18	Information from project companion	Nominal	1 = Yes 2 = No
19	Information from sub-district employees	Nominal	1 = Yes 2 = No
20	Information from district/city employees	Nominal	1 = Yes 2 = No
21	Announcement	Nominal	1 = Yes 2 = No
22	Information boards	Nominal	1 = Yes 2 = No
23	Written report	Nominal	1 = Yes 2 = No
24	Do you get information about the use of the "Kampung" budget?	Nominal	1 = Yes 2 = No
25	Do you get information about the use of self-help funds?	Nominal	1 = Yes 2 = No
26	Do you get info on the use of other project budgets?	Nominal	1 = Yes 2 = No

(*continued*)

Table 1. (*continued*)

No	Characteristics		
	Attribute	*Data Type*	*Values*
27	Was there a meeting the last year?	Nominal	1 = Yes 2 = No 3 = Do not know

28	How many times are these meetings?	Numeric	Min = 1
			Max = 12

29	What level is the meeting?	Nominal	0 = Do not know 1 = “Desa/ Kelurahan” level 2 = “Dusun” level 4 = “RT” level 8 = Others
30	Are you looking for info?	Nominal	1 = Yes 2 = No
31	Understand the program objectives (Do you know what the funds are used for?)	Boolean	1 = Yes 2 = No

4 Result and Discussion

4.1 Classification (Supervised Learning)

We use the PNPM RESPEK 2009 data as a study case by Decision Tree J48 and Logistic regression. Our target variable is to predict whether Papua and West Papua people who received direct community assistance meet or do not meet the level of understanding and perception of funds purposes. The results of classification using Logistic regression and Decision Tree J48 are shown in Table 2.

Table 2. Class Distribution

	Class 0	Class 1
Original Data	1835	61
Logistic Regression	1886	10
Decision Tree J48	1860	36

As can be seen in Table 2, the original distribution of Papua and West Papua people who met and who did not meet the level of understanding and perception towards the PNPM RESPEK Program is *1835:61*. Here “Class 0” is defined as people who met the level of understanding and perception towards the PNPM RESPEK Program. While “Class 1” is defined as people who did not meet the level of understanding and perception towards the PNPM RESPEK Program. From Table 2, we can see the Logistic Regression and Decision Tree J48 predictions did not produce perfect results. In this sense, both classifiers suggested some false negative (FN) and false positive (FP). The FN and FP refers to people that incorrectly classified as "Class 0" or "Class 1". Specifically, there are people who was originally met the level of understanding and perception predicted as the opposite and vice versa.

The accuracy and error rates of Logistic regression and Decision Tree J48 are shown in Table 3. Table 3 demonstrated that the Decision Tree J48 is superior to the Logistic Regression with 97.31% accuracy. In contrast, Logistic Regression accuracy's 96.78%, with a 3.22% error rate.

Table 3. Classification performance results

	Accuracy	Error Rate
Logistic Regression	0.9678	0.0322
Decision Tree J48	**0.9731**	**0.0269**

The modeling result of the J48 Decision Tree method is shown in Fig. 1. Figure 1 represents that the meeting number is the most crucial variable to build people's understanding of the program. Information from sub-district employees also helps Papua and West Papua people to understand the program. Another variable that affected Papua and West Papua people's understanding and perception towards the PNPM RESPEK Program are information from the announcement, have been PNPM actors, became SPP members, received information from Dusun or RT meetings, information from the community of mothers, and previously worked at PNPM program. In comparison, people's understanding is not influenced by age and gender factors.

4.2 Clustering (Unsupervised Learning)

Because Decision Tree J48 does not make apparent immediately how they can be used for unsupervised learning, we further used the trick is to call the data of Papua and West Papua people who met the level of understanding and perception towards the PNPM RESPEK Program as "Class 1" and people who did not meet the level of understanding and perception as "Class 2." We used the Papua and West Papua people who received direct community assistance data to demonstrate these unsupervised learning of K-Means, single-linkage hierarchical clustering, and DBSCAN clustering. Since DBSCAN clustering requires two input parameters, we set the value of *Epsilon* = 2.0 and *minPts* = 35. The results of these three clustering techniques are presented in Table 4.

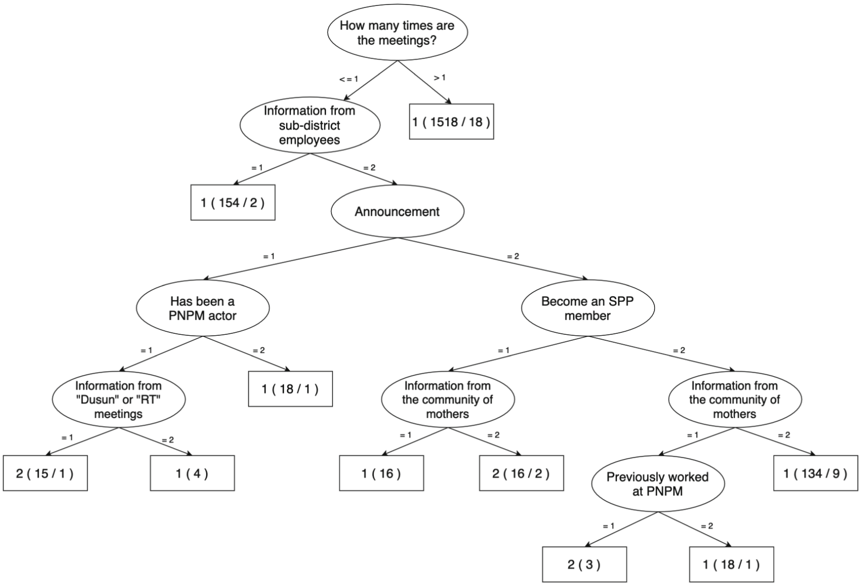
How many times are the meetings?
<= 1
> 1
Information from sub-district employees
1 (1518 / 18)
= 1
= 2
1 (154 / 2)
Announcement
= 1
= 2
Has been a PNPM actor
Become an SPP member
= 1
= 2
Information from "Dusun" or "RT" meetings
1 (18 / 1)
= 1
= 2
Information from the community of mothers
Information from the community of mothers
= 1
= 2
2 (15 / 1)
1 (4)
= 1
= 2
1 (16)
2 (16 / 2)
= 1
= 2
Previously worked at PNPM
1 (134 / 9)
= 1
= 2
2 (3)
1 (18 / 1)

Fig. 1. Decision tree model

Table 4. Clustering performance results

	# of c	Class 1	Class 2	Accuracy	Error Rate
K-Means	2	1197	699	0.6245	0.0375
HCA	**2**	**1895**	**1**	**0.9673**	**0.0327**
DBSCAN	2	1270	626	0.9293	0.0707

Table 4 represented that the Hierarchical clustering (HCA) technique is superior to K-means and DBSCAN clustering techniques, with 96.73% accuracy. Since misclassifying a minority class instance is usually more severe than misclassifying a majority class one, it is clear that class imbalance does not affect the performance of hierarchical clustering (HCA). Figure 2, using HCA, demonstrated that 1895 of Papua and West Papua people who received direct community assistance met the level of understanding and perception towards the PNPM RESPEK. In contrast, Hierarchical clustering (HCA) represented 1 Papua and West Papua people who received direct community assistance did not meet the level of understanding and perception towards the PNPM RESPEK.

Figure 3 visualizes the distribution of Papua and West Papua people who received direct community assistance who met and did not meet the level of understanding and perception towards the PNPM RESPEK Program generated by the K-Means technique. K-means clustering obtained 62.45% accuracy with 1197 of Papua and West Papua people who received direct community assistance met the level of understanding and perception towards the PNPM RESPEK. In contrast, K-means represented 699 Papua

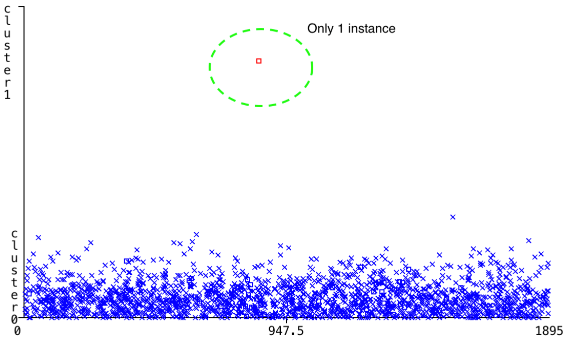

Fig. 2. HCA clustering result

and West Papua people who received direct community assistance did not meet the level of understanding and perception towards the PNPM RESPEK.

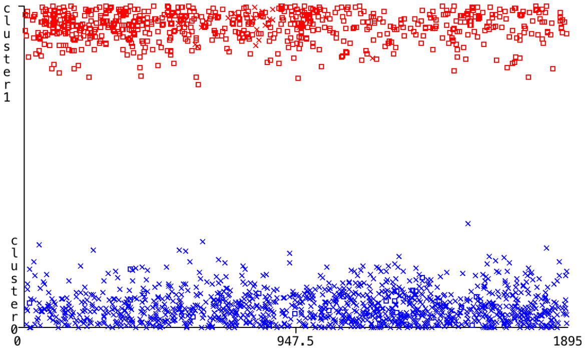

Fig. 3. K-means clustering result

Meanwhile, using the DBSCAN technique, the distribution of instances is shown in Fig. 4. DBSCAN represented 1270 out of 1896 of Papua and West Papua people who received direct community assistance met the level of understanding and perception towards the PNPM RESPEK Program. If we analyze Fig. 4 deeply, the red color distributions are well-separated. In this sense, these points that belong to the red color can be distinguished into two different clusters. These two clusters are 533 un-clustered people, and 93 Papua and West Papua people who received direct community assistance did not meet the level of understanding and perception towards the PNPM RESPEK Program. However, DBSCAN obtained a competitive accuracy of 92.93%, as shown in Table 4.

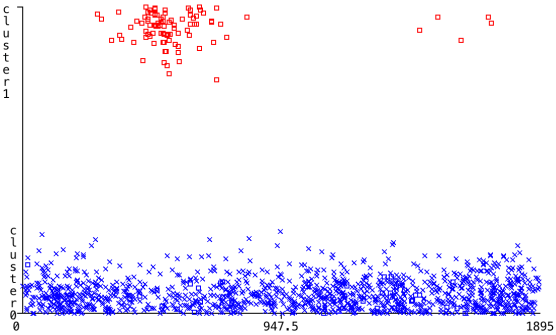

Fig. 4. DBSCAN clustering result

5 Conclusion

Conclusions are drawn based on the output of supervised (classification) and unsupervised learning (clustering). The best performance of classification techniques is generated by the J48 Decision Tree method with 97.31% accuracy. On the other hand, the best performance of clustering techniques is generated by hierarchical clustering (HCA) with 96.73% accuracy. These results are relatively high. In this sense, all 30 variables work satisfactorily in measuring whether 1896 of Papua and West Papua people who received direct community assistance met or did not meet the level of understanding and perception towards the PNPM RESPEK Program. However, these k-means and DBSCAN output can be used as a consideration to address the poverty issues in Papua and West Papua. As DBSCAN represented 533 of Papua and West Papua people who received direct community assistance are still questionable, it is recommended to evaluate this phenomenon to improve decision-making in the future. As data can help accelerate a high performance, we encourage the government to investigate these 2041 out of 3937 original data (known as missing values). These 2041 partial data are essential in providing the right insights to drive better strategic, scenario, and situational decisions.

Overall, we have implemented machine learning techniques to Papua and West Papua people who received support from PNPM RESPEK by simultaneously using two supervised and three unsupervised learning based on data collected from National Socio-Economic Survey (Susenas) July 2009. Future work is intended to conduct the update data so that an optimal result will be form appropriately. We also consider further analysis based on more supervised learning approaches to generate comprehensive results.

References

1. BPS, B.P.S. (2013): Indonesia - Survei Evaluasi Program Nasional Pemberdayaan Masyarakat Rencana Strategis Pembangunan Kampung (2009)
2. Anderson, B.: Papua's Insecurity: State Failure in the Indonesian Periphery. East-West Center, Honolulu (2015)
3. Diani, H.: Health, a specter for Irian Jaya. The Jakarta Post 2000, 21 Aug 5. http://www.library.ohiou.edu/indopubs/2000/08/20/0022.html. Accessed Nov 2008

4. World Bank: Poverty and Health (2014). https://www.worldbank.org/en/topic/health/brief/poverty-health
5. Akatiga: A technical evaluation of PNPM-RESPEK infrastructure built by the barefoot engineers technical facilitator training program in Papua (2015). https://www.akatiga.org/wp-content/uploads/2018/05/Barefoot-Technical-Evaluation-Final-Report-2015.pdf
6. Susilo, A., Trisnanto, A.: The Indonesian national program for community empowerment (PNPM)–Rural: decentralization in the context of neoliberalism and world bank policies. International Institute of Social Studies, **2**(1) (2012)
7. World Bank: Indonesia: Evaluation of the Urban Community Driven Development Program: Program Nasional Pemberdayaan Masyarakat Mandiri Perkotaan (PNPM-Urban) (2013)
8. Rodrigues, I.: CRISP-DM methodology leader in data mining and big data (2020). https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781. Accessed 13 Feb 2021
9. Irwansyah, E.: Clustering. https://socs.binus.ac.id/2017/03/09/clustering/. Accessed 6 Mar 2021
10. Quinlan, J.R.: C4. 5: Programs for Machine Learning. Elsevier (2014)
11. Saravanan, N., Gayathri, V.: Performance and classification evaluation of J48 algorithm and Kendall's Based J48 algorithm (KNJ48). Int. J. Comput. Trends Technol. **59**(2), 73–80 (2018). https://doi.org/10.14445/22312803/ijctt-v59p112
12. Cabrera, A.F.: Logistic regression analysis in higher education: an applied perspective. High. Educ. Handbook theory Res. **10**, 225–256 (1994)
13. Hidayat, A.: Regresi Logistik (2015). https://www.statistikian.com/2015/02/regresi-logistik.html
14. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Inc. (1988)
15. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344. Wiley, Hoboken (2009)
16. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, no. 14, pp. 281–297 (1967)
17. Johnson, S.: Hierarchical clustering schemes. Psychometrika **32**(3), 241–254 (1967). https://doi.org/10.1007/BF02289588
18. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, no. 34, pp. 226–231 (1996)