Comparing k-nearest neighbor and k-means methods for clustering Indonesian farmers’ welfare

L V Pawa¹ and Kismiantini²

¹ Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
² Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

E-mail: lucky.virgian2015@student.uny.ac.id; kismi@uny.ac.id

Abstract. A farmer’s welfare classification can be performed to accommodate all significant issues that will assist policymakers, government, and scientists. This study aims to compare K-Nearest Neighbor (K-NN) and K-Means methods for clustering Indonesian farmers’ welfare using the fifth wave of Indonesia Family Life Survey (IFLS 5) data. The K-Means method is an unsupervised learning algorithm by classifying the data according to the closest distance between observed and centroids. The K-NN method is a supervised learning algorithm by classifying most of the nearest neighbor data. This study used fifteen factors affecting farmers' welfare including land area, type of water, type of rice, income, expenditure, loan, mobile phone use, harvest frequency, crop failure, land ownership, gender, age, level of education, home ownership, and ownership of health insurance. The K-NN performed well to classify farmers' welfare as the K-Means methods in the district data, with an accuracy of 89.8% compared to 53.7%. The K-NN classification results in provinces data showed that the provinces of Bali, East Java, South Kalimantan, Lampung, West Nusa Tenggara, South Sulawesi, and South Sumatra were included as prosperous provinces; while the provinces of Banten, DI Yogyakarta, West Java, Central Java, West Sumatra, and North Sumatra were included as non-prosperous provinces.

1. Introduction

The agricultural sector makes a significant contribution to the economic growth and welfare of farmers throughout the developing countries. The agricultural sector is a source of basic needs, food, housing, job creation, contributing to high domestic income, and foreign exchange for the country. Farmers’ economic growth and welfare depends on farmers’ income levels and the benefits derived from the farming sector itself. The agricultural sector is a key element in improving many Indonesians’ well-being because most Indonesians live in rural areas and work in the agricultural sector [1].

The agricultural sector is the second largest contributor to the Gross Domestic Product (GDP) which drives domestic economic growth in Indonesia. The Indonesian Government is launching programs aimed at increasing the production capacity of agricultural commodities, including the Integrated Farmers Empowerment Movement (GPPT) and Farmer Regeneration, in an effort to support one of the Nawacita (nine hopes), namely the realization of food self-sufficiency in Indonesia [2].
An indicator that can be used to measure the level of welfare of farmers is income. There are many factors affecting farmers’ income levels, including land area, crop types, water sources, and several other factors [3]. Factors affecting crop production are technical culture, plant varieties, as well as fertilizer and fertilization. According to [4] the factors that influence agricultural audiences are the communication tools of the internet.

Based on the factors that may affect the welfare of farmers, it is in interest of determining the region to be classified as prosperous or non-prosperous by the use of Artificial Intelligence (AI). The AI is a branch of computer science that stimulates intelligence on computers. One application of AI is machine learning, which is one technique for inferring data by modeling through a mathematical approach that reflects data patterns. Machine learning algorithms can be grouped into three categories, namely supervised learning, unsupervised learning, and reinforcement learning [5].

One of the unsupervised learning methods commonly used in grouping data is clustering, which aims to group several objects into a cluster (region) based on the similarities they have. K-Means is a non-hierarchical clustering method that separate data into K clusters (regions or groups) of data in such a way that the data is grouped according to the same characteristics. The K-Means method has the ability to group large amounts of data with reasonably fast and efficient computing time. Therefore, this method is the simplest and most widely used clustering method [5]. Lack of classification using the K-Means method is not a guarantee the results of the analysis are the most optimal results due to the random selection of K centroids so that the grouping obtained is different [5].

K-Nearest Neighbor (K-NN) is one of the most widely used for classification. The simplicity and relatively high speed of convergence make this method commonly used. The K-NN method is simple to use in a variety of situations because this method uses only one parameter, called neighbor [6]. The weakness of the K-NN method is that this method cannot determine which variables are significant in the classification [6].

In this study, we would like to compare the performance of the K-Means and K-NN methods to assess the wellbeing of farmers in Indonesia. The best method is then used to construct a farm welfare map based on some of the variables of interest from the fifth wave of Indonesia Family Life Survey data.

2. Materials

2.1. Data

Data used in this study were secondary data from the fifth wave of Indonesia Family Life Survey (IFLS 5) in 2014-2015 which was conducted by RAND and Survey Meter. The household survey data of the IFLS 5 were collected from 16,204 households in 13 of 27 provinces in Indonesia including: four provinces on Sumatra (North Sumatra, West Sumatra, South Sumatra, and Lampung), all five provinces on Java (DKI Jakarta, West Java, Central Java, DI Yogyakarta, and East Java), and four provinces on major island groups (Bali, West Nusa Tenggara, South Kalimantan, and South Sulawesi) [7]. The IFLS5 is a continuing longitudinal socioeconomic and health survey. The sampling scheme in the IFLS5 was stratified on provinces and urban/rural location and then randomly sampled within these strata. The variables of interest for this study include land area, type of water, type of rice, income, expenditure, loan, mobile phone use, harvest frequency, crop failure, land ownership, gender, age, level of education, home ownership, health insurance ownership, and life satisfaction.

2.2. Measures

In this study, we focus on information from the head of household in the farmer family. The size of land for farming was measured in square meter. Income was measured in rupiah, where the respondent was asked to provide an approximate amount of total production from the farm business during the past 12 months. Expenditure was also measured in rupiah, where the respondent was asked to provide an approximate amount of total expenses spent for the farm business during the past 12 months. Loan was measured on the current total loan amount. Harvest frequency was measured per season for planting paddy.
Socio-demographic factor questions included gender (male or female), age (≥18), education levels (in five categories of elementary school, middle school, high school, Islamic school, and college/university).

Farm business factor questions included type of water, type of rice, crop failure, land ownership, mobile phone use. Type of water was assessed with a question on, “What was the main source of water for the […] crop?” Responses were coded as 1 = rain, 2 = irrigation canal, 3 = pump/tube well, 5 = others. Type of rice was assessed with a question on, “What was the main variety of rice did you plant?”. Responses were coded as 1 = Cianjurkepala, 2 = Cianjurslyp, 3 = Setra, 4 = Saigon Bandung, 5 = Muncil, 6 = IR64, 7 = IR42, 8 = White sticky rice, 9 = Black sticky rice, 95 = others. Crop failure was assessed with a question on “Did this household experienced crop loss in the past 12 months?”. Responses were grouped as 1 = Yes and 3 = No. Land ownership was assessed with a question on, “Does the household farm business own the farm land?”. Responses were coded as 1 = Yes and 3 = No. Mobile phone use was assessed with a question on, “Do you or your household member use cell-phone for the farm business?”. Responses were coded as 1 = Yes and 3 = No.

Household characteristics factor questions included health insurance ownership and home ownership. Health insurance ownership was assessed with a question on, “Does this household have a Health Card, ASKESKIN, JAMKESMAS, BPJS or JKN card?”. Responses were coded as 1 = Yes and 3 = No. Home ownership was assessed with a question on, “What is the status of this house?”. Responses were coded as 1 = Self-owned, 2 = Occupying, 5 = Rented/contracted, 95 = Other.

Life satisfaction was assessed with a question on “Please imagine a six-step ladder where on the bottom (the first step), stand the poorest people, and on the highest step (the sixth step), stand the richest people. On which step are you today?”. Responses were grouped into poor (1 to 3) and not poor (4 to 6).

The continuous variables in this study (land area, income, expenditure, age, loan, harvest frequency) were averaged by district and province. The categorical variables were transformed into percentages by selecting a common/particular category. Thus, the categorical variables were the percentage of males, the percentage of respondents using irrigation for the crop, the percentage of respondents planting IR64 rice type, the percentage of respondents with the most recent elementary school education, the percentage of respondents owning their homes, the percentage of respondents using cell-phone for the farm business, the percentage of respondents experiencing crop loss, the percentage of respondents owning their farm lands, the percentage of respondents having health insurance, and the percentage of respondents who were prosperous.

2.3. K-means method
The main idea of the k-means clustering algorithm is minimizing the sum of squared Euclidean distances between the districts or provinces and their cluster means [8]. Since the variables were measured on different scales, before cluster analysis each of the variables was standardized to unit variance. Below are the steps for grouping data based on some variables into clusters using the K-means method.
1. Determine the number of clusters desired. In this study, the districts/provinces are grouped into prosperous and non-prosperous resulting two clusters exist, k = 2.
2. Generate two-centroid (center point on each of the two clusters) randomly from the dataset, then calculate the Euclidean distance between data points and every centroid. Then, the two centroids draw a straight line. The Euclidean distance between entity points \( x_i \) and the centers \( c_k \) for \( m \)-dimensional space as follows

\[
d(x_i, c_k) = \sqrt{\sum_{h=1}^{m}(x_{ih} - c_{kh})^2}
\]
3. Divide data points into two clusters based on the perpendicular bisector or the boundary line. Data points were grouped into cluster 1 or cluster 2 based on their closest Euclidean distances to the centroids.
4. Determine a new centroid from each new cluster. The new centroid value is taken from averaging all the data points in a cluster.
5. Repeat steps 2 to 4 until the centroids stop moving.
6. Gather data points using the two final centroids.

2.4. K-NN method
The K-NN algorithm uses feature similarity to predict the values of the new data points, which means that the new data point will be assigned a value based on how closely the points in the training set match. Given two data sets of testing and training, this algorithm is finding the k-nearest neighbor from training data set for each object in the testing data set [9]. Similar to the k-means method, the variables of interest are normalized. Following are steps for clustering the data using the K-NN method.
1. Define training and testing data. These data are randomly selected from the normalized data with 80% for training data set and 20% for the testing data set.
2. Determine the value of k, $k = \sqrt{\frac{n}{2}}$ where n is a number of observations.
3. Calculate the Euclidean distance between testing and each training data points. The Euclidean distance between entity testing points $x_i$ and the training points $y_i$ for $m$-dimensional space as follows
$$d(x_i, y_i) = \sqrt{\sum_{h=1}^{m} (x_{ih} - y_{ih})^2}$$
4. Sort the distance values in ascending order and determine the first k points.
5. Gather a class to the testing points based on the majority of classes present in the selected points.

In this study, we apply, first, the two methods of clustering the farmers' welfare in Indonesia using district data. We compare the performance of both methods. The best method is then used to classify farmers' welfare in Indonesia by province.

3. Results and discussion
This section presents the results of the study, including the description of the sample and the classification of the welfare of farmers by district and province.

3.1. Sample description
The samples used in this study were 1,515 respondents (1,392 males and 123 females). Their age ranges from 18 to 79 years. According to their life satisfaction responses, 76.2% respondents reported to fall in non-prosperous category and 23.8% in prosperous category. Most farmers reported to use irrigation canal (51.9%) followed by rain (34.5%), others (9.5%), and pump/tube well (4.1%). Most farmers reported planting others type of paddy (64.6%). About 30.2% of farmers chose IR64 paddy type, 3.4% for IR42, and <1% for Cianjur Kepala, Cianjur Slyn, Setra, Saigon Bandung, Muncil, white sticky rice, and black sticky rice. A half of farmers reported twice a year harvesting paddy, while 24.1% reported harvesting three times a year, 22.8% reported harvesting once a year, and 3.1% reported harvesting more than three times a year. Almost all farmers (79.2%) reported not using a cell-phone for the farm business. Most farmers reported not experiencing crop loss (66.5%). In Indonesia, the majority of farmers owned their farm lands (72.8%) and homes (92.4%). It was reported that about half of the respondents had the highest education level of elementary school, about 20% had middle and high schools, and less than 6% had at least diploma degree. A slightly more than 50%, the respondents were reported to have health insurance. The average farm size was 4,161 m² ($SD = 6,738$ m²), income was 11,400,247 in rupiah ($SD = 13,738,450$), expenditure was 3,805,826 in rupiah ($SD = 6,510,616$), and loan was 5,006,231 in rupiah ($SD = 21,570,510$).

3.2. Districts data classification
A total of 108 districts were used in 13 provinces in this study. The district distributions in each province were 4 in Bali, 3 in Banten, 4 in Special Region of Yogyakarta, 11 in West Java, 17 in Central Java, 22 in East Java, 6 in Lampung, 6 in West Nusa Tenggara, 10 in South Sulawesi, 8 in West Sumatra, 3 in
South Sumatra, and 7 in North Sumatra. Using 15 variables of the average of land area \( \text{m}^2 \), income (in rupiah), expenditure (in rupiah), loan (in rupiah), harvest frequency, age (in years), males (%), irrigation for the crop (%), planting IR64 rice (%), elementary school degree (%), owning homes (%), owning farm land (%), using cell-phone for the farm business (%), experiencing crop loss (%), owning health insurance (%), we classify the welfare of farmers into two categories using the K-means and K-NN methods.

For the K-NN, we used the \( k = \sqrt{108/2} \approx 5 \), thus we consider 5 of the nearest points and take the label of majority of these 5 points as the predicted label.

**Table 1. Results of district data classification**

| Method  | Classification |
|---------|----------------|
|         | Prosperous     | Non-prosperous |
| K-Means | 44             | 64             |
| K-NN    | 1              | 107            |

3.3. Accuracy of district data classification

The performance of the K-Means and K-NN methods shall be reported using the accuracy of classification. We compare the results of the classification with the answers of the respondents regarding their satisfaction with their lives. The evaluation metrics are sensitivity, specificity, true/false positive (TP/FP), true/false negative (TN/FN), and accuracy. The sensitivity is the proportion of a test to correctly classify a district as prosperous and the specificity is the proportion of a test to correctly a district as non-prosperous [10].

\[
\text{Sensitivity} (\%) = \frac{TP}{TP + FN} \times 100 \\
\text{Specificity} (\%) = \frac{TN}{TN + FP} \times 100 \\
\text{Accuracy} (\%) = \frac{TP + TN}{N} \times 100
\]

**Table 2. The confusion matrix of the district results from the K-Means and K-NN methods**

|       | K-Means | K-NN |
|-------|---------|------|
| Predicted Classification | Prosperous | Non-prosperous | Total |
| Prosperous                 | 3        | 41    | 44   |
| Non-prosperous             | 9        | 55    | 64   |
| Total                      | 12       | 96    | 108  |

|       | K-Means | K-NN |
|-------|---------|------|
| Predicted Classification | Prosperous | Non-prosperous | Total |
| Prosperous                 | 1        | 0     | 1    |
| Non-prosperous             | 11       | 96    | 107  |
| Total                      | 12       | 96    | 108  |

From table 2, the sensitivity of the district to prosperity is determined by \( 3/(3+9) = 0.25 \) for K-Means and \( 1/(1+11) = 0.083 \) for K-NN. The specificity of the district to non-prosperous is determined by \( 55/(55+41) = 0.573 \) for K-Means and \( 96/(96+0) = 1 \) for K-NN. The accuracy values are \( (3+55)/108 = 0.537 \) for K-Means and \( (1+96)/108 = 0.898 \) for K-NN. The K-NN has the highest accuracy value than K-Means indicating the best method for classifying districts as prosperous and non-prosperous.

3.4. Provincial data classification

A total of 13 provinces were identified in IFLS 5. Using the K-NN method with \( k = 3 \), it was determined that 7 provinces (Bali, East Java, South Kalimantan, Lampung, West Nusa Tenggara, South Sulawesi and South Sumatra) belonged to prosperous and 6 (Banten, DI Yogyakarta, West Java, Central Java, West Sumatra, and North Sumatra) to non-prosperous (figure 1).
3.5. Accuracy of provincial data classification
According to the K-NN results from provincial data, the sensitivity of the province to prosperity is determined by $6/(6+1) = 0.857$, the specificity of the province to non-prosperous is determined by $5/(5+1) = 0.833$, and the accuracy is $(6+5)/13 = 0.846$. Table 3 shows that there are two misclassifications in the provinces of Central Java and South Sulawesi.

| No. | Province          | Original classification | Prediction         |
|-----|-------------------|-------------------------|--------------------|
| 1   | Bali              | Prosperous              | Prosperous         |
| 2   | Banten            | Non-prosperous          | Non-prosperous     |
| 3   | DI Yogyakarta     | Non-prosperous          | Non-prosperous     |
| 4   | West Java         | Non-prosperous          | Non-prosperous     |
| 5   | Central Java      | **Prosperous**          | **Non-prosperous** |
| 6   | East Java         | Prosperous              | Prosperous         |
| 7   | South Kalimantan  | Prosperous              | Prosperous         |
| 8   | Lampung           | Prosperous              | Prosperous         |
| 9   | West Nusa Tenggara| Prosperous              | Prosperous         |
| 10  | South Sulawesi    | **Non-prosperous**      | **Prosperous**     |
| 11  | West Sumatra      | Non-prosperous          | Non-prosperous     |
| 12  | South Sumatra     | Prosperous              | Prosperous         |
| 13  | North Sumatra     | Non-prosperous          | Non-prosperous     |
3.6. Effect of variables on the classification of provinces
Analyzing variables that have an influence on the prosperous and non-prosperous provinces are useful for deeper understanding. Figure 2(a) shows that the prosperous provinces tend to have higher income than the non-prosperous provinces. The province with the highest income is South Sulawesi and the lowest income is Banten. Figure 2(b) shows that the prosperous provinces tend to have higher size of land than the non-prosperous provinces. The province with the highest size of land is South Sulawesi and the lowest is DI Yogyakarta.

![Figure 2. (a) Income for each province based on the welfare category, (b) Area for each province based on the welfare category](image)

Figure 2 (a) Income for each province based on the welfare category, (b) Area for each province based on the welfare category

Figure 3 (a) shows that the prosperous households have a higher percentage of owning house and a higher percentage of owning farm land than the non-prosperous households. Figure 3 (b) shows that the prosperous households have a higher percentage of harvest more than 2 times and the percentage of crop failure is lower than non-prosperous households.

![Figure 3. (a) The percentages of farmers owning house and land, (b) the percentage of farmers harvesting their paddy crops](image)

Figure 3. (a) The percentages of farmers owning house and land, (b) the percentage of farmers harvesting their paddy crops
3.7. Discussion

The results of the classification of provincial data using the K-NN method indicate that Central Java and South Sulawesi are misclassified. The Central Java is originally belonged to a prosperous province but after classification it is classified as a non-prosperous province. According to [1] wider land ownership, the greater the contribution of agricultural sector income to the total income of farm households. The Central Java has an average of size of land of 2,607 m², annual income of 8,239,017 in rupiah, expenditures of 2,527,057 in rupiah and loan of 4,475,258 in rupiah. The Central Java Province is rated as no-prosperous because its size of land and income are below the average income of prosperous province of 5581 m² and 15,105,113 in rupiah, respectively. The study of [11] in a family welfare in the province of Central Java found that 35 out of 64 households were classified as non-prosperous. This result is comparable to our finding that the province of Central Java is rated as non-prosperous.

South Sulawesi is originally belonged to non-prosperous province, but after classification it is categorized as a prosperous province. This province has an average of size of land of 9,555 m², income of 22,170,483 in rupiah, expenditure of 6,669,319 in rupiah and loan of 11,919,747 in rupiah. The size of land and income of this province are above the average of size of land and loans of all prosperous provinces. Thus, the South Sulawesi can be classified as a prosperous province.

The provinces which are located in the east (East Java, Bali, South Kalimantan, West Nusa Tenggara, South Sulawesi) are classified as prosperous. As these provinces have a high rice productivity in the period 2014-2018 [12].

4. Conclusions

The K-Means and K-NN methods have been used to cluster the farmers’ welfare in Indonesia using 108 districts with 15 variables. The number of clusters desired is two, representing a prosperous and non-prosperous district or province. The K-NN performs well than the K-means with a higher accuracy value. The K-NN is then used to construct a farm welfare map of provinces in Indonesia. Characteristics of each cluster are described using bar charts for the average percentage of variables in each cluster. The map shows that more provinces in the east are classified as prosperous than those in the west. A prosperous province has a larger farmland and a higher income than a non-prosperous province. This evidence indicates that farming in wider area will further increase income.

References
[1] Putri CK and Noor TI 2018 Analisis pendapatan dan tingkat kesejahteraan rumah tangga petani padi sawah berdasarkan luas lahan di desa Sindangsari, kecamatan Banjarsari, kabupaten Ciamis, provinsi Jawa Barat [Analysis of income and household welfare levels of paddy rice farmers based on land area in Sindangsari village, Banjarsari sub-district, Ciamis district, West Java province] J. Ilmiah Mahasiswa Agroinfo Galuh 4 927-935 doi:https://dx.doi.org/10.25157/jimag.v4i3.1678.
[2] Badan Pusat Statistika (BPS) 2018 Statistik Kesejahteraan Rakyat [Welfare Statistics] Available from: https://www.bps.go.id/publication
[3] Bakker MM and Doom AM 2009 Farmer-specific relationships between land use change and landscape factors: Introducing agents in empirical land use modelling. J. Land Use Policy 26 809-817 doi:https://doi.org/10.1016/j.landusepol.2008.10.010
[4] Irani T 2000 Prior experience, perceived usefulness and the web: factors influencing agricultural audiences’ adoption of internet communication Tools J. Applied Communications 84 49-63 doi:https://doi.org/10.4148/1051-0834.2151
[5] Russell S and Norvig P 2010 Artificial Intelligence A Modern Approach. New Jersey: Upper Saddle River.
[6] Coomans D and Massart DL 1982 Alternative K-nearest neighbour rules in supervised pattern recognition part 2. Probabilistic classification on the basis of the KNN method modified for direct density estimation J. Analytica Chimica Acta 138 153-165 doi:https://doi.org/10.1016/S0003-2670(01)85298-3
[7] Strauss J, Witoelar F and Sikoki B. 2016. The Fifth Wave of the Indonesia family life survey (IFLS5): Overview and field report. WR-1143/1-NIA/NICH.

[8] Everitt BS, Landau S, Leese M and Stahl D. 2011. "Cluster analysis 5th edition." West Sussex: John Wiley & Sons.

[9] Mucherino A, Papajorgji PJ and Pardalos PM 2009 k-nearest neighbor classification. Springer optimization and its Applications, 83-106. doi:10.1007/978-0-387-88615-2_4

[10] Tharwat A 2018 Classification assessment methods. Applied Computing and Informatics. doi:10.1016/j.aci.2018.08.003

[11] Agustina S, Agoestanto A and Hendikawat P 2017 Klasifikasi tingkat kesejahteraan keluarga provinsi jawa tengah tahun 2015 menggunakan metode regresi logistik ordinal dan support vector machine (SVM) [Classification level of family welfare in Central Java in 2015 using ordinal logistic regression method and support vector machine (SVM)] UNNES J. of Mathematics 6 59-69 doi:https://doi.org/10.15294/ujm.v6i1.11678

[12] Kementerian Pertanian Republik Indonesia 2018 Produktivitas padi menurut provinsi tahun 2014-2018 [Paddy productivity based on province in 2014-2018] Available from: https://www.pertanian.go.id/Data5tahun/TPATAP-2017(pdf)/30-ProdtpPadi.pdf