Sains Malaysiana 53(7)(2024): 1715-1728
http://doi.org/10.17576/jsm-2024-5307-18

Machine Learning for Mapping and Forecasting Poverty in North Sumatera: A Data-Driven Approach

(Pembelajaran Mesin untuk Pemetaan dan Ramalan Kemiskinan di Sumatera Utara: Pendekatan Dipacu Data)

ARNITA*, FARIDAWATY MARPAUNG, FANNY RAMADHANI & DEWAN DINATA

Department of Mathematics, Universitas Negeri Medan, Jl. Williem Iskandar Pasar V, Medan, Indonesia

Received: 20 August 2023/Accepted: 13 May 2024

ABSTRACT

Discussing poverty is crucial because it affects many facets of society, including socioeconomic disparity, crime, and the inability to obtain high-quality education. One of the provinces with the highest poverty rate in Indonesia is North Sumatra. A strategy is required to gather accurate data to effectively reduce poverty. Poverty mapping and prediction were conducted in North Sumatra to get a precise spatial distribution of poverty, the operation of the poverty model, and forecasting using machine learning (ML). Poverty prediction was conducted using a random forest (RF) algorithm and poverty mapping was conducted using the K-Means algorithm. The poverty mapping showed a significant inertia value decline in the third and fourth clusters of the elbow graph. The third cluster (0.313) was superior to the fourth cluster (0.244) in the silhouette index. Thus, there were three poverty clusters - low, medium, and high - that were used in the model. The best model was created using the grid search cross-validation, while the best prediction results were created using the RF algorithm, with the following parameters: n-estimator = 50, max depth = 10, min samples split = 2, and min samples leaf = 1. The mean squared error (MSE) of the RF model's predictions was 0.002617, or satisfactory precision.
Keywords: Cross validation, grid search; K-Means; poverty; random forest regression

ABSTRAK

Isu kemiskinan merupakan isu penting untuk dibincangkan kerana kemiskinan mempengaruhi pelbagai aspek kehidupan seperti jurang sosio-ekonomi, jenayah serta akses yang terhad kepada pendidikan berkualiti. Sumatera Utara merupakan salah satu daripada 5 wilayah teratas dengan jumlah kemiskinan tertinggi di Indonesia. Suatu strategi diperlukan untuk mendapatkan maklumat kemiskinan yang tepat supaya pengurusan kemiskinan disasarkan dan berkesan. Oleh itu, pemetaan dan ramalan kemiskinan dijalankan bagi mendapatkan maklumat yang lebih terperinci tentang taburan reruang kemiskinan dan apakah model kemiskinan di Sumatera Utara. Pendekatan yang diambil untuk memetakan dan meramalkan kemiskinan di Sumatera Utara ialah dengan menggunakan pembelajaran mesin (ML). Pemetaan kemiskinan dijalankan dengan menggunakan algoritma K-Means, manakala ramalan kemiskinan dijalankan menggunakan algoritma hutan rawak (RF). Hasil yang diperoleh daripada pemetaan kemiskinan di Wilayah Sumatera Utara jika dilihat daripada graf siku menunjukkan graf tersebut masih mengalami penurunan nilai inersia yang mendadak pada kelompok ke-3 dan ke-4. Manakala jika dilihat dari nilai indeks Siluet, kelompok ke-3 adalah lebih tinggi daripada kelompok ke-4 dengan nilai indeks Siluet masing-masing adalah 0.313 dan 0.244. Maka dapat disimpulkan bahawa kluster kemiskinan yang digunakan ialah 3 dengan label rendah, sederhana dan tinggi. Manakala, hasil ramalan menggunakan algoritma hutan rawak dengan teknik keesahan silang carian grid memperoleh model terbaik dengan parameter n penganggar = 50, kedalaman maks = 10, min pecahan sampel = 2 dan min sampel daun = 1. Peramalan model RF menghasilkan ketepatan tinggi yang mencukupi dan Min Ralat Kuasa Dua (MSE) ialah 0.002617.

Kata kunci: Carian grid; keesahan silang; kemiskinan; K-Means; regresi hutan rawak

INTRODUCTION

One of the fundamental issues in every nation is poverty. Poverty has negative effects on people, society, and the economy as a whole, endangering lives and providing an unstable and unsafe environment. Poverty and crime have a strong link (Lilik et al. 2023). Inequalities in social and economic conditions brought on by poverty also limit access to vital resources, such as healthcare, employment, and education (Knifton & Inglis 2020).

North Sumatra is one of the provinces in Indonesia with various challenges related to poverty. Based on the poverty data released by the Central Bureau of Statistics (BPS 1967), North Sumatra is the 4th province with the largest number of people living below the poverty line in Indonesia in the second semester of 2022. However, it was improving in the first semester of 2023, placing in the 5th place in Indonesia. Although it was slightly improved, North Sumatra's poverty issue remains a critical one that requires in-depth analysis and solution. There are many locations in North Sumatra where it is difficult to access adequate education and healthcare, moreover, it is combined with the high unemployment rates. Some other factors affecting poverty are the human development index, the average duration of schooling, minimum earnings, unemployment, life expectancy, and per capita spending (Pratama 2015).

Developing an effective plan to combat poverty in North Sumatra requires in-depth information. One strategy is to use machine learning (ML) to map and forecast poverty levels in North Sumatra's cities and regencies. Poverty mapping gathers the geographic distribution of poverty, allowing for more effective and targeted poverty reduction efforts. On the other hand, poverty prediction gathers more detailed data, in the form of a poverty model, to carry out proper preventive measures or interventions. Grouping district/city poverty levels into different categories, such as high, medium, and low poverty areas, allows the government to appropriately plan and execute poverty eradication plans based on those categories. Meanwhile, poverty prediction is required to monitor how effective the eradication is in reducing the poverty index.

In North Sumatra, the clustering process facilitates the identification of socioeconomic and spatial patterns. Classifying and clustering areas according to their relevant shared traits can be done to understand the patterns of poverty, the relationship between certain causes of poverty, and the distribution of poverty at the regional level (Han 2022). The primary reasons that lead to poverty in each group can be determined by grouping regions based on shared social and economic characteristics (Ade et al 2022; Lipesa et al. 2023).

Creating algorithms and computer models to learn from data is the focus of the artificial intelligence discipline of ML. ML can be characterized as computer programs and mathematical formulas that are used to learn from data and make future predictions (Nichols, Chan & Baker 2019). There are two types of learning processes in this approach, supervised and unsupervised learning. Supervised learning uses data that has been tagged by the human developer or has human involvement in the data (Taye 2023). Regression analysis, neural networks, random forest (RF), and support vector machines (SVM) are among the algorithms in supervised learning. Unsupervised learning employs algorithms to evaluate and uncover patterns in data without any human involvement (Watson 2023). K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Hierarchical Clustering are some algorithms in unsupervised learning, particularly for clustering. Both supervised and unsupervised learning will be applied in this study. A method known as RF regression was used to forecast poverty. The K-Means clustering approach was employed in the interim for poverty mapping.

RF was first introduced by Leo Breiman (Breiman 2001). It is a supervised algorithm that uses ensemble learning methods for regression and classification. The ensemble learning method combines several ML algorithms to make accurate predictions (Schonlau & Zou 2020). The RF works in two phases. The first phase involves combining a few N decision trees to create an RF. Then, the second phase is to make predictions for each tree made in the first phase. Consequently, the RF algorithm consists of many decision trees, and the forest is built using an algorithm that is trained through bagging and bootstrap aggregation (Kullarni & Sinha 2013). This algorithm is suitable to learn from a small sample and it has a realistic hypothetical space (Ao et al. 2019). Moreover, it is more resistant to outlier data, performs well on non-linear data, has a lower risk of overfitting, operates effectively on big data sets, and has better accuracy than other supervised algorithms (He et al. 2018). These factors should suit the requirement to forecast poverty.

Arranging data into a limited number of groups with shared characteristics is known as clustering (Omran, Engelbrecht & Salman 2007). Creating clustering algorithms involves a variety of techniques. Clustering using a partition and clustering using a hierarchy are the two basic approaches (Pitafi, Anwar & Sharif 2023). K-Means is a non-hierarchical clustering technique that uses centroid-based grouping. The K-means algorithm separates an unlabeled dataset into a few K-clusters and then continues the process until the algorithm is unable to identify the optimal cluster. In particular, large-scale data containing outliers can be grouped effectively and accurately using the K-Means algorithm (Kaushik & Mathur 2014; Nowak-Brzezińska & Gaibei 2022). The K-Means algorithm's ability to quickly converge

with each iteration is one of its benefits (Pérez-Ortega, Almanza-Ortega & Romero 2018). Therefore, the K-Means algorithm was selected to do poverty mapping in North Sumatra.

Materials and Methods

Data and Variables

The data of several poverty indicators (Table 1) in North Sumatra from January to December 2022 were gathered from the Central Statistics Agency (www.bps.sumut.go.id) (BPS 1967). The data consists of 9 variables related to 33 districts in North Sumatra (Anggi, Rulyanti & Muhammad Faisal 2021; Pratama 2015).

K-Means Clustering

K-Means clustering is a commonly used algorithm for information retrieval, computer vision, and pattern recognition. With K-Means grouping, n-object points are divided into k-clusters so multiple objects with similar features are grouped. The process is repeated until a cluster is created with the center of mass (centroid) being closest to other points (Shukla 2014). Large-scale datasets can be effectively managed by the K-Means technique as it is typically effective in handling datasets with several dimensions. The K-means algorithm frequently reaches quick convergence (Ikotun et al. 2023). Data that share similar features are grouped into one cluster while data that differ from one cluster due to those differences are associated with other clusters (Ade, Harun & Gigin 2018). As a result, the data within a cluster have little variance (Figure 1).

TABLE 1. Collected poverty data and its scale

No	Variables	Unit	Scale
1	Number of Poor Population by Regency/City	Thousand Souls	Interval
2	Poor Population Percentage by District/City	Percent	Interval
3	Poverty Depth Index (P1) by Regency/City		Interval
4	Poverty Severity Index (P2) by District/City		Interval
5	Expenditures per Capita (Adjusted)	Thousand Rupiah	Interval
6	Average Length of School	Years	Interval
7	Life expectancy	Years	Interval

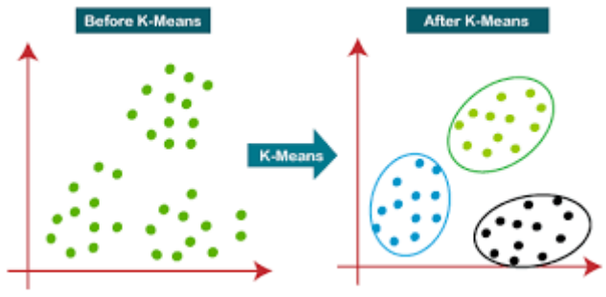

FIGURE 1. The K-means clustering algorithm's operation

Figure 1 shows how the K-Means clustering algorithm functions. The stages for clustering using the K-Means approach are as follows (Tri & Titi 2022):

1) The value of k or the required number of clusters is established

2) The beginning of each cluster is established

3) The separation between each input data and each centroid is calculated using the Euclidean Distance formula until the closest distance to each data from the centroid is found:

i. $$D(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (1)$$

where D is the distance; x and y is the data

4) The information is sorted according to how near it is to the centroid

5) The cluster center is updated to reflect the most recent cluster members. The average value of all the data objects in each cluster is known as the cluster center

6) Each object's calculation is updated to use the updated centroid. The clustering procedure is finished if the centroid remains constant. If the centroid keeps shifting, the process is repeated from step 3 until it stops

SILHOUETTE VALIDITY INDEX

The silhouette validity index is a statistical metric to narrow down the difficulty of figuring out how many clusters are ideal (Rousseeuw 1987). It produces a graph to show the effectiveness of each object position in a cluster, assuming that the data has been divided into clusters. If a(i) is the average object distance between each object and all other objects in the same cluster and b(i) is the average minimum object distance between each object and all other objects in the cluster that are not members of the cluster, then, the following equation can express the silhouette validity index (Nicolaus, Ivy & Hendra 2016):

$$S(i) = \frac{b(i)-a(i)}{Max\{a(i),b(i)\}} \quad (2)$$

where

$$S(i) = \{1 - \frac{a(i)}{b(i)}, if\ a(i) < b(i)\ 0, if\ a(i) = b(i)\ \frac{b(i)}{a(i)} - 1, if\ a(i) > b(i) \quad (3)$$

The average S(i) of all the objects in a cluster shows the similarity of the objects to others and how successfully the objects have been grouped. The grouping is better the closer S(i) is to 1 and poorer the closer S(i) is to -1. If there is just one cluster with a single object as the member, the average value of S(i) will be 0. The optimal number of k-clusters is an estimate of the k value that maximizes the average value of S(i) (Rousseeuw 1987). Additionally, the silhouette index is categorized into four groups (Kaufman & Rousseeuw 1990):

a. $0.7 < IS \leq 1$ *strong structure*

b. $0.5 < IS \leq 0.7$ *medium structure*

c. $0.25 < IS \leq 0.5$ *weak structure*

d. $IS \leq 0.25$ *strong structure*

RANDOM FOREST

The Random Forest method was employed in this study to forecast poverty in North Sumatra. RF is resistant to data anomalies, has better accuracy, resistant to overfitting, and performs well with non-linear data compared to other algorithms (Liu et al. 2021). The RF algorithm comprises two stages. Several N decision trees are combined in the initial stage to construct a random forest. As shown in Figure 2, the second phase entails making predictions for each tree created in the first phase.

The following steps can be used to explain how the RF algorithm functions (Breiman 2001; Schonlau & Zou 2020).

1) A random sample is chosen by the algorithm from the supplied dataset

2) A decision tree is created for each sample that was chosen. Next, each decision tree that was created will yield its forecast findings

3) A voting procedure is used for each predicted outcome. The mean is used for regression problems and the mode is used for classification problems

4) The algorithm will select the final forecast based on the one with the most support

GRID SEARCH CROSS-VALIDATION

An ML technique called grid search cross-validation is used to select the best set of models and hyperparameters. Hyperparameter tuning, or the search for suitable hyperparameter values, is the goal of grid search cross-validation which aims to enhance the performance of ML models (Reliusman, Didi & Joko 2022). K-fold cross-validation is a technique used in grid search cross-validation to assess model performance and adjust hyperparameters more precisely. Grid search cross-validation is used to identify the best combinations of hyperparameters, while k-fold cross-validation is used to assess the model's performance on new data and prevent overfitting.

K-fold cross-validation is a statistical technique to assess the performance of a developed model or algorithm. The evaluation procedure is started by splitting the dataset into training and validation data. The model is then trained using the training data and validated using k-fold validation data (Yunendah et al. 2022). Each stage in the k-fold cross-validation scenario is described as the following:

1. The dataset is divided into k-equal-sized subsets or folds

2. If i = 1, 2, ..., k, then for each subset

   a. Subset i is used as test data

   b. Subgroups were assembled further with training data

   c. The training data is used to train the model

   d. The model's assessment score is calculated based on subset I

   e. The k subgroups' average evaluation score is determined after that

The k-fold cross-validation can be utilized to get around small datasets. The accuracy of ML algorithms is significantly impacted by the amount of supplied data. The ML technique may produce unreliable prediction results if there are fewer than 100 occurrences of data and it is advised to utilize more than 1000 instances to increase the prediction's accuracy (Alpayidin 2004; Sena 2018). Although k = 10 has been utilized in several experiments, adopting k = 5 is preferable since it can shorten calculation times without compromising the model's performance (Marcot & Hanea 2021).

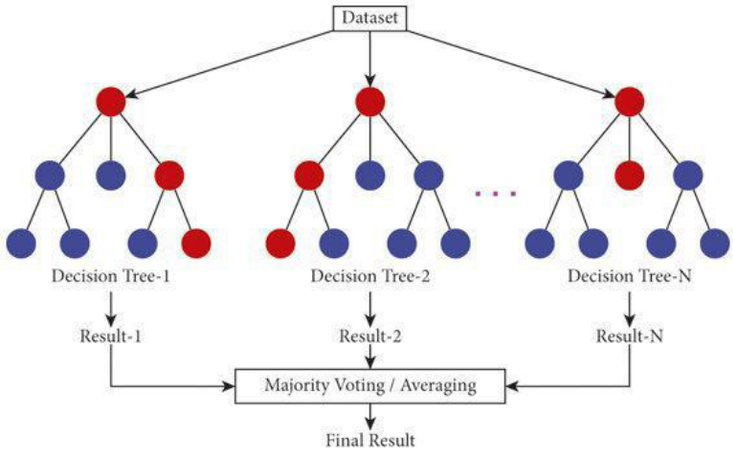

FIGURE 2. The working process of the RF algorithm

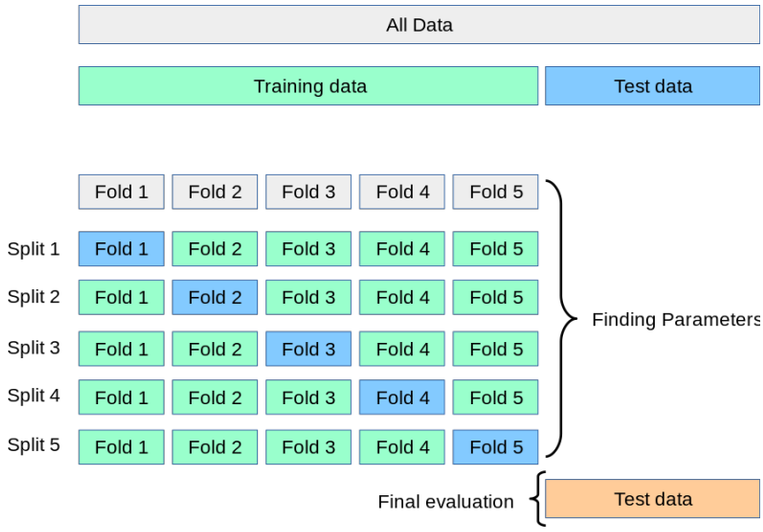

FIGURE 3. The scenario of k-fold cross-validation

Results and Discussion

The acquired data was whole without any missing values, indicating that the data cleaning procedure was not performed. Some variables have different units of measurement, so a normalization process was necessary to reduce data redundancy (repetition) and normalize information for improved data workflow (Singh & Singh 2022).

Jupyter Notebook was used to build the poverty mapping ML model, with Pandas library for data loading, scikit-learn for normalization, data splitting, and prediction processes, and Seaborn and Matplotlib pyplot for data visualization.

K-Means Clustering for Mapping

North Sumatra province is located between 10 – 40°N, and 980 – 1000°E. Its territory is bordered by Aceh province and the Sumatra Strait at the north, West Sumatra and Riau provinces to the west, and Sumatra Strait on the east. North Sumatra Province has 25 regencies, 8 cities, 325 sub-districts, and 5,456 sub-districts and villages. It is one of the provinces with a lower average poverty rate compared to the national average in Indonesia, (8.63% to 9.22% Indonesia average) and the poverty rate fluctuated in numbers and percentages during the 2012–2020 period.

Table 2 shows that the average number of poor people in North Sumatra was 47,0275.58 thousand ± 6,305.26, which was more than the national average number (26,500 thousand souls). Meanwhile, the average percentage of poor people in North Sumatra was 10.32%, the P1 was 1.44, and the P2 was 0.33. The adjusted average per capita expenditure in North Sumatra was 10,716.09 thousand rupiah and it was lower than the national average of 11,480 thousand rupiah. Meanwhile, the average length of schooling in North Sumatra was 9.29 years, higher than the national average of 8.69 years. Lastly, the life expectancy of North Sumatra was 69.53 and it was also higher than the national average life expectancy of 68.25.

The data are extremely homogeneous (Figure 4). This is because the data uses different measurement units and there are outliers in several variables. Therefore, proper data handling is required to produce a more accurate analysis. Normalization is a data management technique for data preparation. This method aids in scaling up the values of numerical fields in datasets for ML and data mining. The minimum-maximum normalization formula (Peshawa Jamal & Rezhna 2014) was used as the normalization. Figure 5 illustrates how the distribution of the data seems more uniform after normalization.

After data normalization, the K-Means algorithm was used to evaluate the data via clustering. To determine the ideal cluster number to reflect poverty in North Sumatra, the elbow diagram was created from the analysis's findings. The elbow diagram showed a

declined value in clusters 3 and 4 (Figure 6). This showed that if more than three or four clusters are created, little information may be produced. When using the K-means approach, the elbow diagram is a supporting technique to choose the best K-value. This approach concentrates on the variance percentage as a function of the cluster count. The K-values will be examined one at a time and the SSE (sum square error) value will be noted, to determine the ideal K-value (Syakur et al. 2018).

TABLE 2. Poverty data in North Sumatra from January to December 2022

Variables	Mean	Standard Deviation	Minimum	Maximum
Number of Poor Population by Regency/City (Thousand Souls)	470,275.58	63,052.26	350,452	631,886
Percentage of Poor Population by District/City (Percent)	10.32	4.45	4	25
Poverty Depth Index (P1) by Regency/City	1.44	0.82	0	5
Poverty Severity Index (P2) by District/City	0.33	0.24	0	1
Expenditures per Capita Adjusted (Thousand Rupiah)	10716.09	2092.46	6152	15503
Average School Length (Years)	9.29	1.37	6	12
Life Expectancy (Years)	69.53	2.51	63	74

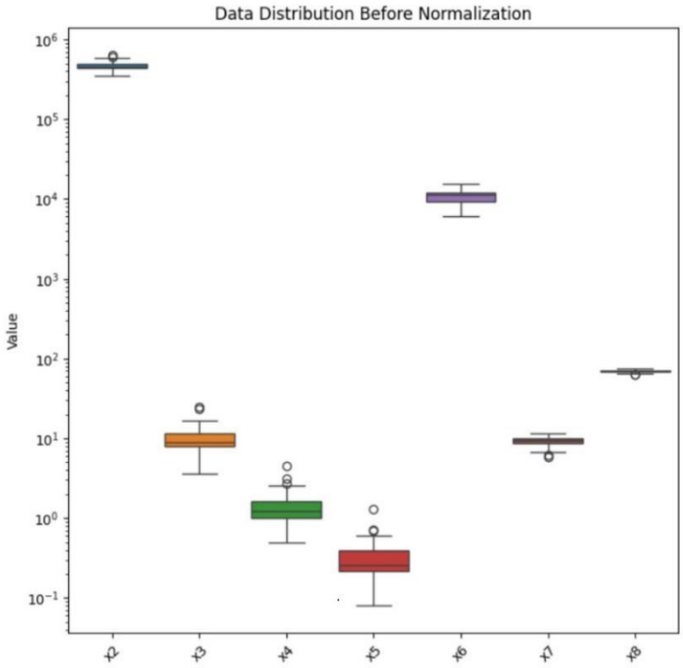

FIGURE 4. Poverty data in North Sumatra before normalization

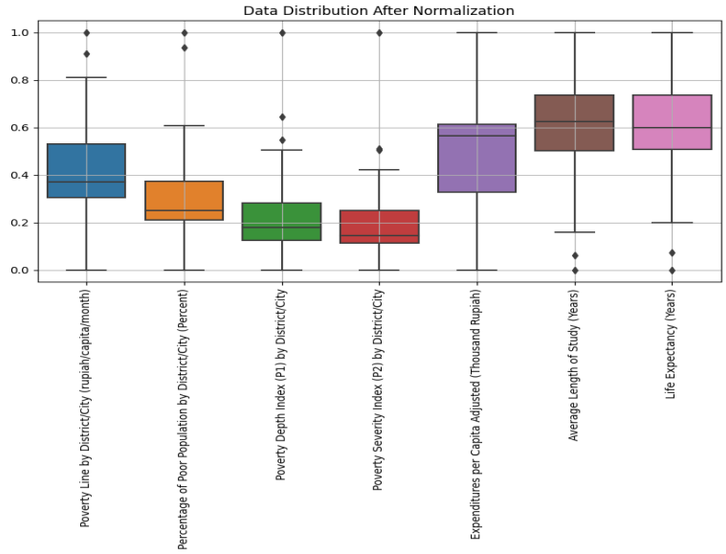

FIGURE 5. Poverty data in North Sumatra after normalization

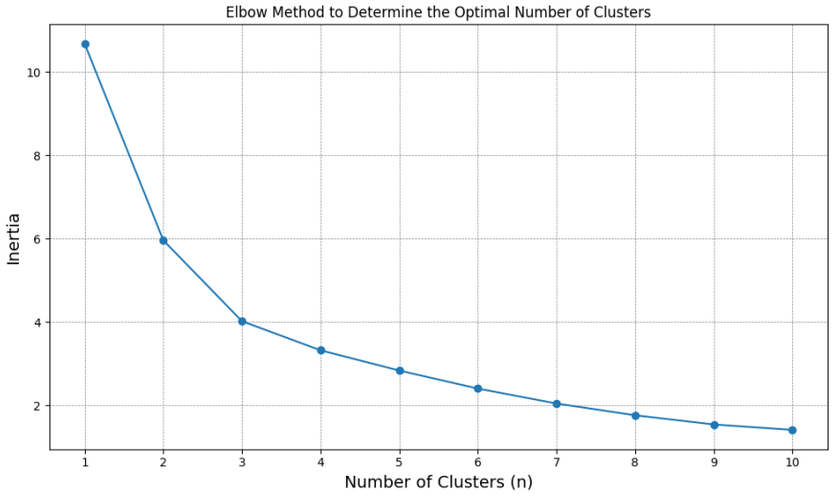

FIGURE 6. Elbow diagram determining the optimal number of clusters

The silhouette index was employed to assess how many ideal clusters are utilized in the formation of clusters. Table 2 shows that there is little difference between the silhouette index values in the third and fourth clusters. The inertia value has also decreased, although not significantly. It was considered that three clusters are the ideal number.

TABLE 3. Silhouette index values

Number of clusters	Silhouette Index
1	10.68
2	5.96
3	4.02
4	3.32
5	2.84
6	2.41
7	2.05
8	1.77
9	1.55
10	1.42

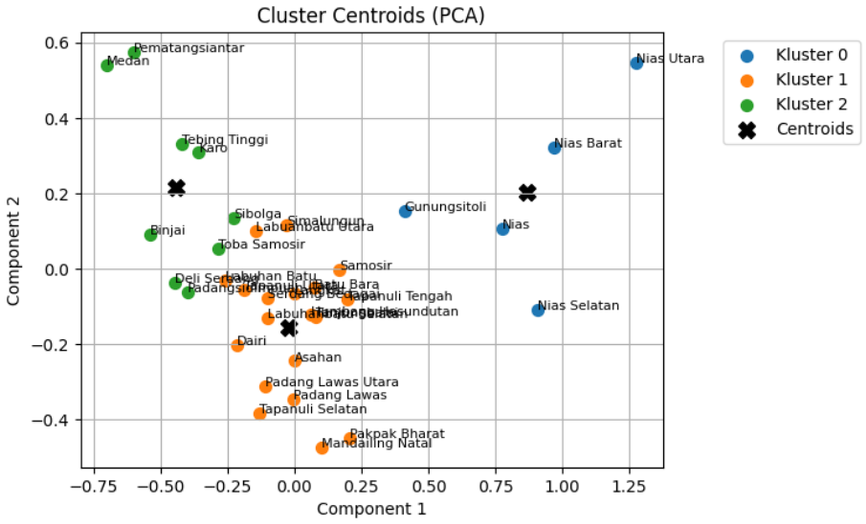

FIGURE 7. The clusters of poverty distribution by city or district in North Sumatra

Figure 7 illustrated the formation of three clusters; 0, 1, and 2. The five districts or cities that included in Cluster 0 were Gunungsitoli, Nias, South Nias, Nias Barat, and Nias Utara. While cluster 1 consisted of 18 regencies or cities: Mandailing Natal, Pakpak Bharat, Tapanuli Selatan, Padang Lawas, Padang Lawas Utara, Asahan, Batubara, Dairi, Humbang Hasundutan, Labuhan Batu, Langkat, Labuhan Batu Selatan, Simalungun, Samosir, Batubara, Tapanuli Tengah, Serdang Bedagai, and Tanjung Balai. Medan, Tebing Tinggi, Karo, Binjai, Sibolga, Toba Samosir, Deli Serdang, Padang Sidempuan, Pematang Siantar, and Labuhan Batu Utara were the ten urban districts categorized in Cluster 2.

Random forest prediction

Apart from the poverty severity index variable, which was utilized as the dependent variable, all variables used for mapping were used as the independent variables to predict poverty in North Sumatra. For forecasting, the previously normalized data from the mapping process was used. The cross-validation method was then used to divide the data into two halves. Data splitting was used to understand how well the model in generalizing the data. Additionally, data splitting supports better model parameters and hyperparameters management. The model will detect potential issues or unexpected patterns not showing up in the training data by being tested on data that has not yet been known via training (Syakur et al. 2018). In contrast, cross-validation is a statistical technique that divides the data into two subsets, learning process data, and validation data, that can be used to assess the effectiveness of a model or algorithm. In addition, the size of the dataset can be used to determine the type of cross-validation. The reason cross-validation K-fold is frequently employed is that it can shorten computation times while retaining estimation accuracy (Berrar 2018). In this study, the k-fold was k = 5, because it can reduce computing time without reducing model performance (Marcot & Hanea 2021).

Grid search cross-validation was used in this study's simulations. By testing and evaluating each combination of models and hyperparameters separately, the grid search cross-validation method offers a way to choose the best combinations. The objective was to identify the combination with the highest model performance and can be used as a prediction model. The grid search cross-validation technique is thorough in looking at all potential combinations of the predefined parameters. Grid search also has a relatively straightforward and repeatable notion, which means that both its processes and outcomes can be repeated. For each parameter, a combination of various values is used to run the simulation. The variables under test are:

The best model was produced from the simulations run with the following parameters: n-estimator = 50, max-depth = 10, min-samples split = 2, and min-samples leaf = 1. These factors led to MSE and RMSE values of 0.002617 and 0.1563, respectively. Even though no reference source states the best RMSE value, a close to zero RMSE value shows that the model has good performance because it will produce a small error (Chai & Draxler 2022). The RMSE number is deemed favorable if it is less than the two standard deviation values of model and observation (Liemohn et al. 2021).

P1, Percentage of poor population by district or city, adjusted expenditures per capita, average length of study, life expectancy, and the poverty line by district or city were the factors with the biggest impact on poverty level in North Sumatra (Figure 8). According to studies (Muhammad & Indah 2023; Spada, Fiore & Galati 2023), access to high-quality education can have a significant impact on economic opportunities and the likelihood to break the cycle of poverty. People with less education tend to have fewer job opportunities and lower incomes. Population growth has been shown to have a positive impact on poverty. The degree of poverty in a society is greatly impacted by income. Several other factors frequently interact with the complex relationship between income and poverty. Access to essentials including food, housing, healthcare, education, and sanitation is improved with higher salaries. It can be challenging for individuals or families to meet these demands when their income is insufficient, which can lead to subpar living conditions (The World Bank 2022).

TABLE 4. Tested parameters in grid search cross-validation

Parameter	Values
n-estimators	50, 100, 200
Max-depth	10, 20, 30
Min-samples split	2, 5, 10
Min-samples leaf	1, 2, 4

The distribution of poverty in 33 urban areas in North Sumatra was predicted in Figure 9. The result suggested that the RF algorithm-based prediction model is capable of accurately predicting poverty in almost all urban areas. In Nias Barat and Nias Utara, there was a large discrepancy between the actual and expected numbers. Bias can be decreased by using the RF algorithm in conjunction with the splitting grid search cross-validation technique. Following the findings of the study in Hao, Luo and Pan (2022), it was claimed that the RF algorithm could deliver experimental results with good performance and create a promising predictive model.

It is also clear why the city or district of Binjai had the lowest percentage of poor population, while Nias Utara and Nias Barat have the largest. Cities and districts on Nias Island generally have greater poverty

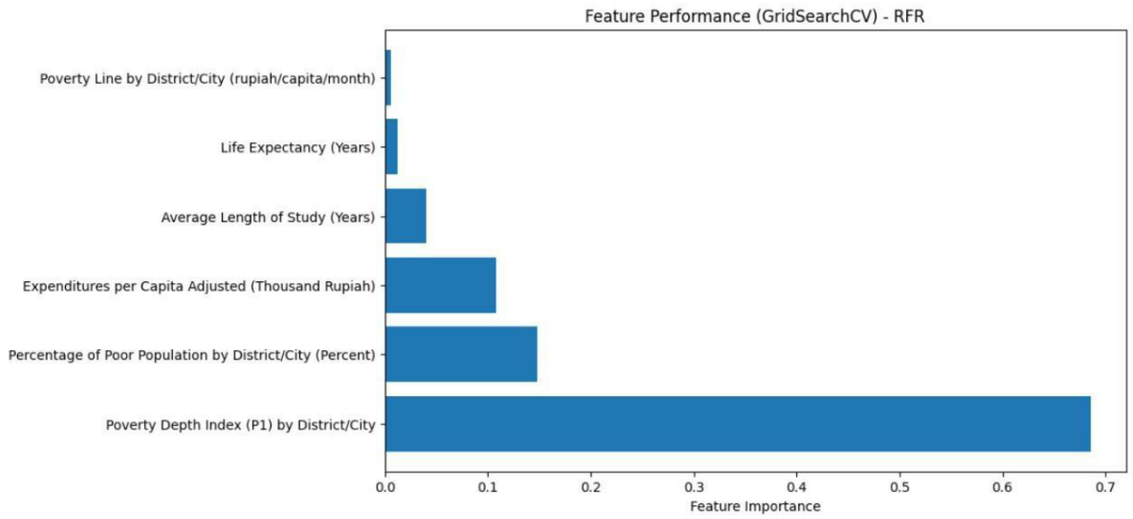

FIGURE 8. The order of feature importance

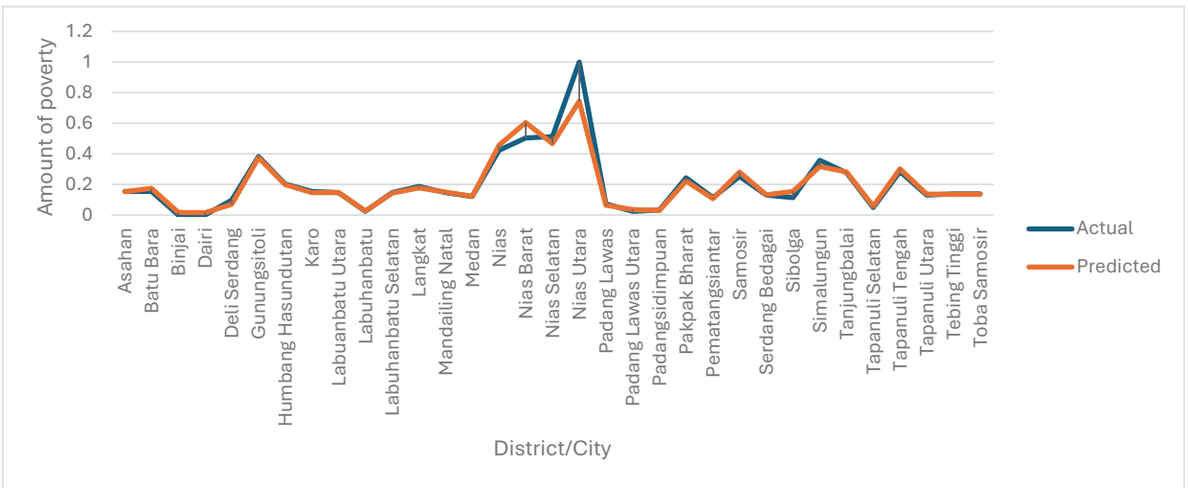

FIGURE 9. Total poverty level based on actual and predicted data

rates than other districts. The classification results also demonstrated that all the Nias Islands' regencies and cities belong to the same cluster and have higher than the national average poverty rates. According to research findings (Reliusman, Didi & Joko 2022), the poverty levels in different Nias Island regions are comparable. In the Nias districts and cities, particularly in Nias Utara, Nias Barat, and Nias Selatan, there is a spatial influence of poverty between regions. Therefore, it is important to focus on the provincial and local government's efforts in combating poverty in urban districts that are included in cluster 0.

Conclusion

Using the K-Means algorithm, the North Sumatra poverty map identified three clusters based on the poverty indices. While the prediction model used the splitting grid search cross-validation algorithm and the random forest algorithm. The settings n-estimator = 50, max depth = 10, min samples split = 2, and min samples leaf = 1 enable the simulation to generate the optimal model. The model's accuracy had MSE and RMSE values of 0.002617 and 0.1563, respectively. Based on the forecast, the district or city of Binjai had the fewest people living in poverty, while Nias Utara, followed by Nias Barat, had the greatest number.

Acknowledgments

We would like to thank Universitas Negeri Medan for funding this study. Especially the rector, LPPM, the dean of FMIPA Universitas Negeri Medan, and the department head who have provided support and facilities for this series of studies.

References

Ade Bastian, Harun Sujadi & Gigin Febrianto. 2018. Penerapan algoritma k-means clustering analysis pada penyakit menular manusia (studi kasus kabupaten Majalengka). *Jurnal Sistem Informasi* 14(1): 26-32. https://doi.org/10.21609/jsi.v14i1.566

Ade Syahputra, Mulyanto, Agustinus Suryantoro & Lukman Hakim. 2022. Analisis deskriptif potensi daerah dan tingkat kemiskinan di Sumatera Utara. *Prosiding Seminar Nasional Universitas Abdurachman Saleh Situbondo*, September 2022. pp. 38-47. https://unars.ac.id/ojs/index.php/prosidingSDGs/article/view/2308%0Ahttps://unars.ac.id/ojs/index.php/prosidingSDGs/article/download/2308/1629

Alpayidin, E. 2004. *Introduction to Machine Learning (Adaptive Computation and Machine Learning Series)*. Vol. 14. Massachusetts: The MIT Press. https://doi.org/10.1017/s1351324906004438

Anggi Aprillia, Rulyanti Susi Wardhani & Muhammad Faisal Akbar. 2021. Analysis of factors affecting poverty in the province of the Bangka Belitung Islands. *Jurnal Ilmu Ekonomi Terapan* 6(2): 188-201. https://doi.org/10.20473/jiet.v6i2.29184

Ao, Y., Li, H., Zhu, L., Ali, S. & Yang, Z. 2019. The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering* 174: 776-789. https://doi.org/10.1016/j.petrol.2018.11.067

Berrar, D. 2018. Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology* 1: 542-545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

BPS. 1967. Jumlah Penduduk Miskin (Ribu Jiwa) Menurut Provinsi dan Daerah 2022-2023. https://www.bps.go.id/indicator/23/185/1/jumlah-penduduk-miskin-menurut-provinsi.html

Breiman, L. 2001. Random forests. *Machine Learning* 45: 5-32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Chai, T. & Draxler, R.R 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Development* 7(3): 1247-1250. https://doi:10.5194/gmd-7-1247-2014

Han, S. 2022. Spatial stratification and socio-spatial inequalities: The case of Seoul and Busan in South Korea. *Humanities and Social Sciences Communications* 9: 23. https://doi.org/10.1057/s41599-022-01035-5

Hao, J., Luo, S. & Pan, L. 2022. Rule extraction from biased random forest and fuzzy support vector machine for early diagnosis of diabetes. *Scientific Reports* 12: 9858. https://doi.org/10.1038/s41598-022-14143-8

He, L., Levine, R.A., Fan, J., Beemer, J. & Stronach, J. 2018. Random forest as a predictive analytics alternative to regression in institutional research. *Practical Assessment, Research and Evaluation* 23(1): 1-16.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. & Heming, J. 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences* 622: 178-210.

Kaufman, L. & Rousseeuw, P.J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis.* New York: Wiley.

Kaushik, M. & Mathur, B. 2014. Comparative study of k-Means and hierarchical clustering techniques. *International Journal of Software & Hardware Research in Engineering* 2(6): 93-98.

Knifton, L. & Inglis, G. 2020. Poverty and mental health: Policy, practice and research implications. *BJPsych Bulletin* 44(5): 193-196. https://doi.org/10.1192/bjb.2020.78

Kullarni, V.Y. & Sinha, P.K. 2013. Efficient learning of random forest classifier using disjoint partitioning approach. *Proceedings of the World Congress on Engineering 2013*. Vol II. July 3-5, London.

Liemohn, M.W., Shane, A.D., Azari, A.R., Petersen, A.K., Swiger, B.M. & Mukhopadhyay, A. 2021. RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar–Terrestrial Physics* 218: 105624. https://doi.org/10.1016/j.jastp.2021.105624

Lilik Sugiharti, Rudi Purwono, Miguel Angel Esquivias & Hilda Rohmawati. 2023. The nexus between crime rates, poverty, and income inequality: A case study of Indonesia. *Economies* 11(2): 62. https://doi.org/10.3390/economies11020062

Lipesa, B.A., Okango, E., Omolo, B.O. & Omondi, E.O. 2023. An application of a supervised machine learning model for predicting life expectancy. *SN Applied Sciences* 5(7): 189. https://doi.org/10.1007/s42452-023-05404-w

Liu, M., Hu, S., Ge, Y., Heuvelink, G.B.M., Ren, Z. & Huang, X. 2021. Using multiple linear regression and random forests to identify spatial poverty determinants in rural China. *Spatial Statistics* 42: 100461. https://doi.org/10.1016/j.spasta.2020.100461

Marcot, B.G. & Hanea, A.M. 2021. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? *Computational Statistics* 36(3): 2009-2031. https://doi.org/10.1007/s00180-020-00999-9

Muhammad Al Faruq & Indah Yuliana. 2023. The effect of population growth on poverty through unemployment in East Java Province in 2017-2021. *Journal of Social Research* 2(6): 1900-1915. https://doi.org/10.55324/josr.v2i6.872

Nichols, J.A., Chan, H.W.H. & Baker, M.A.B. 2019. Machine learning: Applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews* 11(1): 111-118. https://doi.org/10.1007/s12551-018-0449-9

Nicolaus, Evy Sulistianingsih & Hendra Perdana. 2016. Penentuan jumlah *cluster* optimal pada *median linkage* dengan indeks validitas *silhouette*. *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)* 05(2): 97-102.

Nowak-Brzezińska, A. & Gaibei, I. 2022. How the outliers influence the quality of clustering? *Entropy* 24(7): 917. https://doi.org/10.3390/e24070917

Omran, M.G.H., Engelbrecht, A.P. & Ayed Salman. 2007. An overview of clustering methods. *Intelligent Data Analysis* 11(6): 583-605. https://doi.org/10.3233/ida-2007-11602

Pérez-Ortega, J., Almanza-Ortega, N.N. & Romero, D. 2018. Balancing effort and benefit of k-means clustering algorithms in big data realms. *PLoS ONE* 13(9): e0201874. https://doi.org/10.1371/journal.pone.0201874

Peshawa Jamal Muhammad Ali & Rezhna Hassan Faraj. 2014. Data normalization and standardization: A technical report. *Machine Learning Technical Reports* 1(1): 1-6. https://docs.google.com/document/d/1x0A1nUz1WWtMCZb5oVzF0SVMY7a_58KQulqQVT8LaVA/edit#

Pitafi, S., Anwar, T. & Sharif, Z. 2023. A taxonomy of machine learning clustering algorithms, challenges, and future realms. *Applied Sciences (Switzerland)* 13(6): 3529. https://doi.org/10.3390/app13063529

Pratama, Y.C. 2015. Analisis faktor-faktor yang mempengaruhi kemiskinan Di Indonesia. *Esensi* 4(2): 45-53. https://doi.org/10.15408/ess.v4i2.1966

Reliusman Dachi, Didi Nuryadin & Joko Susanto. 2022. Determinan tingkat kemiskinan di Kepulauan Nias tahun 2011 - 2019: Pendekatan regresi spasial. *Syntax Literate: Jurnal Ilmiah Indonesia* 7(7): 8994-9008.

Rousseeuw, P.J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20(1): 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Schonlau, M. & Zou, R.Y. 2020. The random forest algorithm for statistical learning. *Stata Journal* 20(1): 3-29. https://doi.org/10.1177/1536867X20909688

Sena, S. 2018. Pengenalan deep learning Part 8: Gender classification using pre-trained network (transfer learning). *Medium*. https://medium.com/@samuelsena/pengenalan-deep-learning-part-8-gender-classification-using-pre-trained-network-transfer-37ac910500d1

Shukla, S. 2014. A review on k-means data clustering approach. *International Journal of Information & Computation Technology* 4(17): 1847-1860. http://www.irphouse.com

Singh, D. & Singh, B. 2022. Feature wise normalization: An effective way of normalizing data. *Pattern Recognition* 122: 108307. https://doi.org/10.1016/j.patcog.2021.108307

Spada, A., Fiore, M. & Galati, A. 2023. The impact of education and culture on poverty reduction: Evidence from panel data of European countries. *Social Indicators Research* https://doi.org/10.1007/s11205-023-03155-0

Syakur, M.A., Khotimah, B.K., Rochman, E.M.S. & Satoto, B.D. 2018. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering* 336: 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Taye, M. 2023. Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions. *Computers* 12: 91. https://doi.org/10.3390/computers12050091

The World Bank. 2022. Poverty. יראה. 2022. https://www.worldbank.org/en/topic/poverty/overview

Tri Wahyudi & Titi Silfia. 2022. Implementation of data mining using k-means clustering method to determine sales strategy in S&R baby store. *Journal of Applied Engineering and Technological Science* 4(1): 93-103. https://doi.org/10.37385/jaets.v4i1.913

Watson, D.S. 2023. On the philosophy of unsupervised learning. *Philosophy & Technology* 36: 28. https://doi.org/10.1007/s13347-023-00635-6

Yunendah Nur Fuadah, Ibnu Dawan Ubaidillah, Nur Ibrahim, Fauzi Frahma Taliningsing, Nidaan Khofiya SY & Muhammad Adnan Pramuditho. 2022. Optimasi convolutional neural network dan k-fold cross validation pada sistem klasifikasi glaukoma. *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika* 10(3): 728-741. https://doi.org/10.26760/elkomika.v10i3.728

*Corresponding author; email: arnita@unimed.ac.id