International Journal of
Advanced Science Computing and Engineering

Journal Homepage www.ijasce.org/index.php/ijasce

Classification of Population Data of Nagari Based on Economic Level Using The K-Nearest Neighbor Method

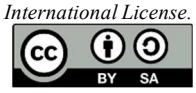
Ainil Mardiah [a], Defni [a,*], Aster Happy Lestari [a], Junaldi [b], Titin Ritmi [c]

[a] *Information Technology Department, Politeknik Negeri Padang, Kampus Limau Manis, Padang, Indonesia*
[b] *Electrical Engineering Department, Politeknik Negeri Padang, Kampus Limau Manis, Padang, Indonesia*
[c] *English Department, Politeknik Negeri Padang, Kampus Limau Manis, Padang, Indonesia*
*Corresponding author: *defni@pnp.ac.id*

***Abstract*— Economy is one of the factors determining the level of prosperity of a region. Changes in economic figures are obtained from data on people's income in the area. This research was carried out in the Koto Baru Simalanggang district, which is one of the district areas in West Sumatra. In this research, authors measure and classify community economic level. The classification of community economic levels aims to facilitate decision making by local governments to recommend communities that are entitled to receive economic assistance from the government. Based on field studies, data collection and grouping process of community economic status levels is still done manually so that errors often occur and the data collection process is inefficient. This research has several stages, namely data collection, system analysis and design, application design, application demo and testing and application implementation. The data collection was done with the assistance of the regional government. At the analysis stage, the process of grouping community economic data is done, while the system is designed using the waterfall method. The design of the community economic level classification application was designed using the K-Nearest Neighbor (KNN) method. There are 4 variables used to group community economic data, namely residential status, income, expenses and amount of responsibility. This application design uses the Laravel framework and MySQL database. The accuracy value produced by this application reaches 81%, the precision level is 84% and the recall level is 94%. So it can be concluded that this application is able to makes it easier for regional government to make decisions based on the classification of community economic levels.**

***Keywords*—Classification; K-Nearest neighbor; economic level.**

Manuscript received 10 Nov. 2023; revised 12 Feb. 2024; accepted 29 Mar. 2024. Date of publication 30 Apr. 2024.



I. INTRODUCTION

The economic level is the level of welfare and strengthening of the economic structure in a region. Economic growth shows the extent to which economic activity will generate additional income in a certain period. In other words, the economy is said to have increased if the real income of the community in a particular year is greater than the real income of the community in the previous year. Economics identifies an individual's place in a group of people based on their economic activities, income, education level, type of work, and status. The data can be used as a benchmark for household opinion in community economic reporting. The data is then compared to the relevant rupiah exchange amount required to cover the minimum cost of living.

The process of collecting data and classifying the level of economic status of Nagari residents is currently manual, so the efficiency in the data collection process is less than ideal. Only when necessary, such as when there is government assistance, is the grouping process carried out. As a result, the citizens of Nagari are not well controlled by government authorities due to the lack of detailed government supervision. Based on the above statement, a specialized application should be created. To be able to classify the economic status of families in Nagari by using the K- Nearest Neighbor (K-NN) algorithm. Based on the economic data of the community, it will be categorized into K-NN.

Based on research conducted it can be concluded that, the algorithm K-NN is very suitable for use in the classification process of society prasejahtera contained in the village sapikerep probolinggo, because the algorithm K-NN has a

good level of accuracy in the process of classification of data used [1]. Research by [2] this study is a study that uses algorithms Fuzzy C-Means and Naive Bayes. The Data used is the database Integrated according to Kepmensos No. 71 / huk/2018 obtained in the field Social protection and Social Security of the District of Buleleng. Data are grouped into 3 groups, namely program beneficiaries Keluarga Harapan (PKH), Bantuan Sosial Pangan (BSP) and beneficiaries Health Insurance (PBI). Data mining accuracy calculation method using the confusion matrix. Based on research conducted can be concluded that, the results implementation of the algorithm with 1350 family data shows the level of accuracy Naive Bayes algorithm is better than Fuzzy C-means. Naive Bayes accuracy value of 74% and Fuzzy C-means accuracy of 67%. From tests that have been done using the calculation confusion matrix obtained the results of an effective algorithm used in determining the family of such beneficiaries is Naive Bayes algorithm.

This research is research[3] using the K-mean algorithm. The data processed is the proposed data recipient of nagari Taluk BLT-DD in 2022.Based on research conducted can be concluded that, data processing using PHP MYSQL Software, from a sample of 25 data then generated 11 data included in the cluster 1 with the status beneficiaries are said to be eligible, 5 data including cluster 2 with status recipients considered and as many as 9 data included to cluster 3 with unworthy status. From the test results obtained the level of accuracy amounted to 83.33% so that it can be recommended to help the government of wali nagari in taking policy.

Research [4]. This study uses the k-Nearest Neighbor (K-NN) algorithm which requires training information to classify objects that are very close. Based on research conducted it can be concluded that, with the existence of community grouping based on the family economy can easily see the economic data of the community below, economy or upper economy. K-NN clustering algorithm method has excellent and accurate performance in the process of grouping the data, especially the grouping of data based on family economics. Based on the results of testing conducted then fold cross validation obtained an accuracy value of 90% .

Research by [5]. This study is a study that uses k-Nearest Neighbor algorithm. Data used in this study a total of 1289 data with 13 Attributes obtained from the Department of Housing People and residential areas of Jepara Regency. Data processing begins from attribute selection, data categorization, outlier data cleaning, normalization data and application of methods.

Based on research conducted it can be concluded that, K-NN can be applied for eligibility classification of rtlh rehabilitation recipients. The highest accuracy of 97.93% with 13 data attributes as variables influential and has been determined by Disperkim, namely safety aspects (covering the foundation, walls, beams and roofs), building structures (house area, roof, floor, and wall materials), and health (lighting and ventilation). Based on related research that has been compiled from several journals in relation to this final project, the k-Nearest Neighbor method is used for grouping data. K-NN can perform data grouping based on the closest distance from the data train with test data.

II. Material and Methods

Figure 1 below explains the methodology stages used in this research.

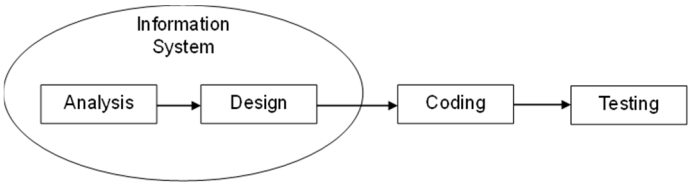


Fig. 1 System Development

The stages of the waterfall method in making classification applications:

A. Software Requirements Analysis

Analyzing software requirements in the waterfall model begins with the step of identifying needs. The classification application will be used to classify data based on economic levels. The classification application has the ability to import and classify data based on housing status, income, expenses, and number of dependents. The classification application can process data quickly and accurately. The data used is from Nagari.

B. Design

Designing the overall structure of the classification application system as needed. The design stage is carried out to create a system design, database design, and interface design. System design includes making use case diagrams, activity diagrams, and sequence diagrams. In the database design, design the database structure that will be used by the classification application.

C. Coding

Coding uses the Laravel framework and the PHP programming language. At the coding stage, the K-NN algorithm is implemented to analyze the classification of economic levels.

The stages performed by the KNN algorithm are:

1. Determine the value of N for the initial data.
2. Determine the value of K (nearest neighbour).
3. Prepare training data in the form of criteria.
4. Determine the status of each training data.
5. Calculating the distance of each training data sample.

D. Testing

Testing the classification results in the application is done by comparing the classification results of the system and the classification results using excel. Calculation of the accuracy of system classification results is by using a confusion matrix.

Confusion matrix is information about the actual classification results that can be predicted by a classification system. The confusion matrix table for 2 x 2 dimensional classes is shown in the following table [1].

TABLE I
CLASSIFICATION CONFUSION MATRIX

Real	**Prediction**	
	Positive	Negative
Positive	TP	FN
Negative	TP	FN

The explanation of Table 1 is as follows:

1. TP (True Positive) is the amount of data in the actual class that is positive and the prediction class is also positive.
2. FN (False Negative) is the amount of data in the actual class that is positive while the prediction class is negative.
3. FP (False Positive) is the amount of data in the actual class is negative while the prediction class is positive.
4. TN (True Negative) is the amount of data in the actual class is negative and the prediction class is also negative.

Accuracy is a testing method based on the level of closeness between the predicted value and the actual value as a whole. Here is the accuracy formulation:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

Precision is the class agreement of the data label with the positive label given by the classifier. Here is the precision formulation:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall is the effectiveness of the classification to identify positive labels. Here is the formula for finding recall:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

III. Result and Discussion

A. Implementation of K-Nearest Neighbor Algorithm

The K-Nearest Neighbor (K-NN) algorithm is a method for classifying objects based on the learning data that is closest to the object. K-NN is a supervised learning algorithm where the results of new query instances are classified based on the majority of categories in the K-NN algorithm. Where the class that appears the most which will be the result class of a classification [2].

Proximity is defined in terms of metric distance, such as Euclidean distance. Euclidean distance can be found by using the following equation:

$$D_{xy} = \sqrt{\sum_n 1(x_i - y_i)^2} \quad (4)$$

Description:

D: proximity distance x: training data
y: testing data
n: number of individual attributes between 1 and n
f: similitary function of attribute i between case X and case Y
i: Individual attributes between 1 to n

In the calculation for the classification of economic levels, 4 criteria are used, namely house status, income, expenses, and dependents. At home status the criteria are not yet in numeric form. So that the house status criteria are converted to numeric to facilitate the calculation to determine the classification.

At the K-NN algorithm implementation stage, the data is normalized, then create a method to calculate the classification. The value of k chosen is k = 3.

Fig. 2 Implementation of the K-NN Algorithm

TABLE II
THE RESULTS OF THE DATA CLASSIFICATION PROCESSING

Name	Home State	Income	Spending	Dependents	Classification
xxxxxx	Rented	1.500.000	1.500.000	7	Less well-off
xxxxxx	owned by the parents	3.500.000	2.000.000	1	Middle
xxxxxx	self-owned	10.000.000	7.000.000	5	Well-off
xxxxxx	Owned by the family	1.500.000	1.300.000	4	Middle
xxxxxx	Use loan	1.200.000	1.200.000	6	Less well-off

In table 2 is 5 data classification results from the implementation of the K-NN algorithm in the application system.

B. Interface Implementation

The implementation of the interface in the classification application to determine the economic level uses templates provided by bootstrap and CSS libraries. The interface display in the application consists of a login page, dashboard (home) page, manage population data page, classification page, and print report page.

C. Comparison of system results with excel

This test aims to compare the results of data classification using the application system with data classification using Microsoft Excel. The comparison also aims to evaluate the performance of the KNN algorithm in classifying data and understanding the impact of the criteria used on the classification results.

Classification testing using the K-Nearest Neighbor (KNN) algorithm in Microsoft Excel is carried out to measure the performance and effectiveness of the algorithm in classifying data based on similarity patterns. The following is the accuracy test of the classification results using excel.

TABLE III
THE RESULTS OF MATRIX CONVOLUTION PROCESSING PROCESSED IN EXCEL

N=221	Positive Real (1)	Negative Real (0)
Positive Prediction (1)	TP: 174	FP: 35
Negative Prediction (0)	FN: 11	TN: 1
	184	36

Based on table 3 above, the accuracy value is

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN) \times 100\%}$$

$$= \frac{(174+1)}{(174+35+11+1)x\ 100\ \%}$$

$$= \frac{(175)}{(221)x\ 100\ \%}$$

= 0.79 x 100%
= 79 %

Based on the evaluation results, the accuracy of using Microsoft Excel in classifying data based on the economic level is 79%.

Testing the classification results of the application system created based on laravel, has the aim of measuring the performance of the K-Nearest Neighbor algorithm in classifying data using a more integrated and complex approach in a system. This test involves the process of implementing the K-NN algorithm in the Laravel framework. The following is an accuracy test of the classification results using the system.

TABLE IV
THE RESULTS OF PROCESSING THE CONFUSION MATRIX USING A WEB-BASED APPLICATION

N=221	Positive Real (1)	Negative Real (0)
Positive Prediction (1)	TP: 175	FP: 31
Negative Prediction (0)	FN: 10	TN: 5
	185	36

Based on table 3 above, the accuracy value is

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN) \times 100\%}$$

$$= \frac{(175+5)}{(175+31+10+5)x\ 100\ \%}$$

$$= \frac{(180)}{(221)x\ 100\ \%}$$

= 0.81 x 100%
= 81 %

Based on the evaluation results, the accuracy of the economic level classification results in the application system is 81%.

IV. CONCLUSION

Based on the results of the discussion on the application of data classification of residents of Nagari based on economic level using the K-Nearest Neighbor algorithm, the following conclusions are obtained: Based on the analysis conducted in Nagari Koto Baru Simalanggang, there are 4 criteria for determining the economic level, namely house status, income, expenses, and dependents. Data is stored using a database, so that in managing all data related to the system application process can be safe and efficient. Application testing is seen from the results of the comparison between manual calculations using Microsoft Excel and calculations using the appropriate system. For the accuracy value on the classification application system calculated using the confusion matrix is 81%. While the accuracy value of the classification results using microsoft excel is 79%.

REFERENCES

[1] A. Khairi, A. F. Ghozali and A. D. N. Hidayah, " implementation of K- Nearest Neighbor (KNN) for the classification of underprivileged people in Sapikerapp Village, Sukarapu Sub-District," Journal Trilogi, vol. 2, no. 3, pp. 319-323, 2021.

[2] P. S. Saputra, " comparison of Fuzzy C-Means algorithm and Naive Bayes algorithm in determining beneficiary families (KPM) based on the lowest socioeconomic Status (SSE)," Journal of Science and Technology (Ann), vol. 10, no. 1, pp. 1-8, 2021.

[3] Y. Filki, " K-Means Clustering algorithm in predicting recipients of Village Fund Direct Cash Assistance (BLT)," Journal of Business Economics Informatics, vol. 4, no. 4, pp. 166-171, 2022.

[4] R. M. A. Thousands Of People, E. Bu'ulolo and S. A. Hutabarat, "K-Nearest Neighbor Clustering algorithm in Medan Area Sub-District Community grouping based on Family Economic level," KOMIK (National Conference of Information and Computer Technology), vol. 6, no. 1, pp. 773-782, 2022.

[5] A.- N. S. Na'iema, H. Cozy and N. A. Widiastuti, "classification of homeless rehabilitation program beneficiaries using k-Nearest Neighbor algorithm," Journal of Technology and Computer Systems, vol. 10, no. 1, pp. 32-37, 2022.

[6] P. Son, A. M. H. Pardede and S. Syahputra, " analysis of the K-Nearest neighbor (KNN) method in the classification of Iris Data," JTIK (Journal of Information Engineering Kaputama), vol. 6, no. 1, pp. 297-305, 2022.

[7] F. D. Son, J. Riyanto and A. F. Zulfikar, "Asset Management Information System Design at Pamulang University WEB-Based," Journal of Engineering, Technology, and Applied Science, vol. 2, no. 1, pp. 32-50, 2020.

[8] H. and R., "Effect of User Profiling on system recommendations using K Means and KNN," Journal of Information System Management (JOISM), vol. 2, no. 1, pp. 13-18, 2020.

[9] L. Farokhah, "K-Nearest Neighbor implementation for Flower classification by RGB color Feature Extraction," Journal of Information Technology and Computer Science, vol. 7, no. 6, pp. 1129-1136, 2020.

[10] R. Y. Endra, Y. April and Y. Yanu, "comparative analysis of PHP Programming Language Laravel with PHP Native in Website development," EXPERT: Journal of Information Systems Management and Technology, vol. 11, no. 1, p. 48, 2021.

[11] S. Manish and G. Parul, "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning", International Journal of Engineering Trends and Technology, vol. 70, no. 7, pp. 43-48, 2022.

[12] M. Bansal, A. Goyal and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning", Decision Analytics Journal, vol. 3, pp. 1-21, 2022.

[13] S. Maohua and Y. Ruidi, "An efficient secure k nearest neighbor classification protocol with high-dimensional features", International Journal of Intelligent System, vol. 35, no. 11, pp. 1791-1813, 2020.

[14] S. R. Cholil, T. Handayani, R. Prathivi and T. Ardianita, "Implementation of K-Nearest Neighbor (KNN) Classification Algorithm for Scholarship Recipient Selection Classification", Indonesian Journal on Computer and Information Technology (IJCIT), vol. 6, no. 2, pp. 118-127, 2021.

[15] V. A. Prilia Putri, A. B. Prasetijo and D. Eridani, "Comparison of the Performance of the Naive Bayes and K-Nearest Neighbor (KNN) Algoritm for House Price Prediction", Scientific Journal of Electrical Engineering, vol. 24, no. 4, pp. 162-171, 2022.