International Journal of Informatics and Computation (IJICOM)
Vol. 4, No.2, December 2022
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Implementation of KNN Algorithm for Occupancy Classification of Rehabilitation Houses

Nurhadi Wijaya[1], Joko Aryanto[2], Kasmawaru Kasmawaru[3], Anang Faktchur Rachman[4]

Abstract The

The 2010 eruption of Mount Merapi and the resulting rain lava in Central Java's Kab. Sleman DIY and Magelang Regency damaged homes and infrastructure. According to the Head of BNPB Regulation No. 5, the Community Rehabilitation and Reconstruction and Community-Based Settlement program plan is utilized to repair and rebuild properties damaged by the 2011 Merapi eruption. Two thousand five hundred sixteen residences that will stay in the area have been built permanently due to this initiative. Occupancy rates (permanent occupancy) are used by the World Bank's Key Performance Indicators (KPI) to gauge a program's effectiveness. The database has information on how the software was used and proved successful. Databases, essential tools for introducing new data patterns and revealing previously hidden information, are used in data mining. This study applies the KNN algorithm to classify the house's occupancy status data after Mount Merapi's eruption. The accuracy results obtained from the classification of 82.03%, and the performance of the results through the AUC obtained a value of 0.935.

Keywords:

Data Mining, KNN, Classification, Rehabilitation, Occupancy Status, Merapi.

1. Introduction

Mount Merapi eruption caused damage to housing in Central Java Province and Yogyakarta Province. A total of 2,682 houses in the Sleman Regency area were heavily damaged; in Central Java (Central Java), 174 homes were severely damaged. The subsequent flow of rainwater or rain lava from Mount Merapi resulted in 2,856 primary-affected housing units and 1,087 secondary-affected units [1]. The regulation of the Head of the National Disaster Management Agency (BNPB) No. 5, 2011, stipulates the rehabilitation and reconstruction of the housing and settlement sector after the eruption of Merapi using the Rehabilitation and Reconstruction of Community and Community-based Settlement schemes [1]. This scheme has built 2,516 permanent housing units in Sleman Regency, Yogyakarta Province, 2,040 shelters, and Magelang Regency, Central Java Province, and as many as 476 units of Housing Fund Assistance (BDR) [2].

Based on the Key Performance Indicators: Community-based Settlements Rehabilitation and Reconstruction Project issued by The World Bank, the occupancy status of a built house called a permanent residence (huntap), is one indicator of the successful performance of rehabilitation and reconstruction after the eruption of Merapi. The more huntaps are inhabited, the better the indicators of successful implementation are [3].

After the disaster, those who carried out rehabilitation and reconstruction support operations kept records of the rehabilitation and reconstruction of homes, starting with the

Corresponding Author:
1 Nurhadi Wijaya[1], Universitas Respati Yogyakarta (nurhadi@respati.ac.id),
2 Joko Aryanto[2], Universitas Teknologi Yogyakarta (joko.aryanto@uty.ac.id),
3 Kasmawaru[3], Universitas Dipa Makassar (kasmawaru@undipa.ac.id),
4 Anang Faktchur Rachman[4], Universitas Madura (anang@unira.ac.id).

recipient's name and ending with the home's occupancy status (huntap). 2,146 residences have the classification of permanent housing, but it is unclear at this time whether all of them are occupied. [3].

Today, many studies have been carried out in data mining. Data Mining is a process that aims to find patterns from the data contained in the database. The discovery of this pattern can be used for grouping objects or data based on the similarity of the data. Classification technique in data mining is a learning technique used to predict the value of the target category attribute. Classification aims to divide objects assigned only to one number of categories called classes [4].

Research on the classification of occupancy data status of permanent houses after the 2010 eruption of Mount Merapi in Central Java and Yogyakarta has been conducted. In a research article in 2018, it was stated that the Naive Bayes algorithm could be used to classify data on the occupancy status of permanent residences for rehabilitation and reconstruction houses after the eruption of Mount Merapi [5]. In 2019, the Naive Bayes algorithm combined with the chi-square feature selection algorithm succeeded in classifying data on the occupancy status of permanent residences after the eruption of Mount Merapi [6].

In machine learning studies, classification can help researchers develop learning in various fields, such as agriculture, namely the classification of leaf types using CNN with an accuracy rate of 86% [7]. Research entitled Design and Building Javanese Script Classification in The State Museum of Sonobudoyo Yogyakarta proves that classification can produce an accurate level of accuracy [8]. In addition, classification with the RNN algorithm also produces an accuracy rate of 91%, which can be expressed as an accurate level of accuracy [9].

Based on the results of the studies that have been achieved above raise ideas or ideas and questions about whether the methods or algorithms used from previous studies can be replaced or changed with other algorithms, such as the KNN algorithm. Can modifications impact the accuracy and performance of data categorization on the occupancy status of permanent residences for restoration houses after Mount Merapi's 2010 eruption to the model and the algorithm approach used?

2. Related Works

The Ministry of Public Works, Directorate General of Human Settlements, is in charge of the housing and settlement sector's post-disaster rehabilitation and reconstruction program. This plan prioritizes a combination of development based on community groups and development based on values, carrying out rehabilitation and reconstruction works using an empowerment approach. Communities are viewed as the primary players in development because it is considered that they can make important decisions about their own lives and carry out recovery with the right support. Development is dependent on community groupings. Value-based development mandates that community growth serves as a vehicle for instilling high ideals such as respect for others, compassion, cooperation, and others to foster the growth of social capital. [2].

A paper presented a study on the classification of occupancy status data for permanent residences for rehabilitation houses after the eruption of Mount Merapi using the Naive Bayes Algorithm [5]. In 2019 research related to the combination of the Naive Bayes algorithm with Chi-Square for data classification of permanent residential occupancy status. An article discussed the implementation of the Decision Tree C.45 Algorithm for Classification of Occupant Status Data for Post-Merapi Eruption Rehabilitation Houses has also been carried out, wherefrom previous studies, the classification technique with the algorithm has been successfully carried out. In this study, the KNN algorithm will be used to classify data on the occupancy status of permanent residences for rehabilitation houses after the eruption of Mount Merapi [6].

3. Proposed Method

KNN is one of the methods used in classification [10]. The steps for using the KNN method are explained as follows: [11]

1. Determine the K parameter
2. Calculate the distance between the data to be evaluated with all training
3. Sort the distance formed (in ascending order)
4. Determine the closest distance to the sequence K
5. Pair the appropriate classes
6. Find the number class from the nearest neighbor and assign the class

As the data class to be evaluated, there are several KNN Approach Methods [12]. In this research, the approach method used is the Euclidean Distance Approach. Here is the formula for finding the square root of two vectors:

$$d_i = \sqrt{\sum_{i-1}^{p}(x_{2i} - x_{1i})^2} \qquad (1)$$

Table 1. Variables and Description from Formulas

Variable	Description
x^1	Sample Data
x^2	Test Data or Testing Data
I	Variable data
D	Distance
P	Amount of Training Data

4. Experimental Setup

1. Dataset

   Data collection in this study used data from Secondary Private Database Management Information System (MIS) rehabilitation and reconstruction after the eruption of Merapi and rain lava in 2010. Data were collected and obtained through the rehabilitation and rehabilitation work unit (Satker) in the form of books, notes (FactSheets), and Websites related to this research. The number of records or instances is 2,516. There are 11 attributes or variables used. Data was used as datasets with class residential status (class already inhabited and not inhabited) permanent housing rehabilitation homes after the Merapi eruption and lava raining 2010.

2. Proposed Classification Model

   This research will use the KNN algorithm to classify the occupancy status data for permanent housing rehabilitation houses after the 2010 eruption of Mount Merapi. As for the attributes that will be used, among others, such as budget, gender, number of household members, prospective female occupants, water work, electrical work, level of damage, community self-help, land status, building area, and occupancy status as a class to be classified.

   A technique for categorizing objects based on the learning data that is most similar to the object is the KNN algorithm. Data from learning is projected onto a multidimensional space in which each dimension corresponds to a different data attribute. This area has been divided into divisions based on how learning data is categorized. If class c is the most prevalent classification among the point's k closest neighbors, then the point belongs to class c in this space. The Euclidean distance is typically used to calculate the proximity of two neighbors using the formula in Equation 2. [13].

$$distance = \sqrt{\sum_{i=1}^{n}\left(X_{training}^{i} - X_{testing}\right)^2} \quad (2)$$

Table 2. Variables and Description from Formulas

Variable	Description
$X_{training}^{i}$	i training data
$X_{testing}$	Test Data
i	record (row) from the table
n	Amount of training data.

During training, this method just stores feature vectors and classifies training data. During the classification stage, the same features are computed for the test and training data (whose classification is unknown).

The closest K integers are chosen after calculating the distance between this new vector and all learning data vectors. Predictions indicate that most of these classifications will include the newly categorized points. The data will determine the ideal K value for this approach. In general, a large value of K will lessen the impact of noise on classification, but it will also exacerbate the blurring of classification lines. By employing cross-validation, for instance, parameter optimization can choose an appropriate value for K. The nearest neighbor algorithm refers to the specific situation where the classification is predicted based on the closest learning data (i.e., K = 1).

The existence or absence of irrelevant features or whether the weight assigned to these characteristics does not correspond to their importance for classification significantly impacts the KNN algorithm's accuracy. The KNN algorithm has several benefits, including resistance to noisy training data and effectiveness with significant training sets. However, KNN's main flaw is that it requires the user to determine the value of the parameter K (the number of closest neighbors) and that distance-based training lacks clarity over which distance type and attribute should be employed for optimal results. That is money for computers because calculations are involved. Distance between every query instance and the entire training sample.

With the K-Nearest Neighbor (KNN) method, which employs a supervised algorithm, the outcomes of a new query instance are categorized using most of the KNN's categories. This algorithm aims to categorize new objects using attributes and training data. The classifier relies entirely on memory and does not use matching models. It will locate numerous K objects or (training points) that are close to a given query point. Among the K object categories, this categorization receives the most votes. The KNN algorithm employs nearby classification to determine the new query instance's prediction value.

5. Result & Analysis

1. Dataset Analysis

We gather a rehabilitation and reconstruction dataset after the eruption to conduct our study. The data obtained through the rehabilitation work unit in the form of Database queries, interviews, and field surveys are also used as additional facts in this study. The data sets analyzed in this study are data or classes on the occupancy status of houses receiving housing grants after the eruption of Mount Merapi with the label class "not yet" inhabited, which means the house is not yet occupied or not yet occupied (there are 370 houses) and the label "Already occupied". " (the house has been settled or occupied) as many as 2146 of the total 2516 records dataset entity. The analysis results of the dataset dwelling after the eruption of Mount Merapi rehabilitation can be shown in Table 3 below.

Table 3. Results of analysis

Attribute Name	Type	Range	Missings
Status Huni	binomial	Not (370), It (2146)	no missing values
Budget	Polynomial	Fund BNPB 2011 (947), Fund JRF 2011 (172), Fund PSF 2011 (368), Fund PSF 2012 (921), Fund PSF 2013 (108)	no missing values
Gender	Binominal	Is Male (2014), is Female (502)	no missing values
Number of Members Households	Polynomial	Large (13), Small (2194), Medium (309)	no missing values
Prospective Female Occupants	Integer	0 (50) 1 (1257) 2 (865) 3 (272) 4 (63) 5 (7) 6 (1) 7 (1)	no missing values
for Water Works	Binominal	Not Yet (334), Completed (2182)	no missing values
for electrical Works	Binominal	Not Yet (229), Completed (2287)	no missing values
Damage Level	Binominal	Lost/Heavy Damage (2116). Moderate/Slightly damaged (400)	no missing values
Community Self-Help	Polynomial	< 10 % (296), 10 % (2220)	no missing values
Land status	Polynomial	Independent (541), Independent Collective (415), TKD (1560)	no missing values
Building Area	Polynomial	< 36 m2 (4) = 36 m2 (8) > 36 m2 (2504)	no missing values

2. Experimental results of the KNN Algorithm

The KNN algorithm was utilized for testing and experimental validation in this study. Modeling algorithms using the software Rapid Miner. Creating a dataset as a table in an MS Excel spreadsheet and instances or records is the first step in modeling the KNN algorithm. 2516 The next step is to validate the data using 10-fold cross-validation. In 10-fold cross-validation, the data is randomly divided into ten equal-sized parts, and the error rate for each part is then calculated. The overall error rate is then determined by averaging the error rates of all data portions [14]. The value of the number of validations (10) and the sampling type are then entered (stratified sampling).

The dataset validation dataset and the KNN method are connected in the next step. After choosing the KNN algorithm, it connects to validation data, proceeds with the model application, and concludes by displaying the correctness of the KNN model performance results. Figure 1 shows the outcomes of the application accuracy algorithm KNN data status dwelling repair following Mount Merapi's eruption.

accuracy: 82.03% +/- 2.44% (mikro: 82.03%)

	true Sudah	true Belum	class precision
pred. Sudah	1924	230	89.32%
pred. Belum	222	140	38.67%
class recall	89.66%	37.84%	

Fig 1. The results of the performance of the accuracy of the KNN data status dwelling rehabilitation after the eruption of Mount Merapi

In Figure 1 above, the accuracy obtained from the experimental results of applying the algorithm KNN for the residential data status of the rehabilitation house after the eruption of Mount Merapi reached 82.03%.

Then the accuracy results above are tested again using the table confusion matrix. The table confusion matrix based on the accuracy results can be displayed as follows:

Table 4. The state of residential data for house rehabilitation following Mount Merapi's eruption is shown in the table for the confusion matrix application of the method on KNN.

	True Already	True Not
Pred. Already	a	b
	1924	230
Pred Not yet	C	d
	222	140

Then the table confusion matrix above is tested by calculating the accuracy value. The equation is as follows:

$$\text{Accuracy} = \left(\frac{a+d}{a+b+c+d}\right) = \frac{TP+TN}{TP+TN+FP+FN} \text{ x}100\% \qquad (3)$$

Calculation Results:

$$\text{Accuracy} = \left(\frac{1924+140}{1924+230+222+140}\right) = \left(\frac{2064}{2516}\right) \text{x}100\% = 82.03\% \qquad (4)$$

Based on the results of the above calculation, the percentage level of accuracy of the application of the algorithm on KNN the status of residential data for rehabilitation houses after the eruption of Mount Merapi reaches 82.03%. Thus, testing the accuracy results shown in Figure 1 above with a confusion matrix using equations 3 and 4 are the same and appropriate.

The next step is the Receiver Operating Characteristic (ROC)-Curve Area Under the Curve AUC test. The AUC curve evaluates the performance by calculating the likelihood that a randomly chosen sample will produce positive or negative results. The classification utilized is stronger the bigger the AUC. Five categories of AUC accuracy performance can be distinguished [15], including:

a. 0.90 to 1.00 = excellent categorization (excellent classification).

b. Classification between 0.80 and 0.90 is good (good classification).

c. 0.70 to 0.80 = a mediocre classification (appropriate classification).

d. A classification of 0.60 to 0.70 is subpar.

e. 0.50 to 0.60 is the incorrect classification (failure).

The results of the AUC curve for the application of the algorithm on KNN and the status of residential data for rehabilitation houses after the eruption of Mount Merapi can be shown in the following figure below:

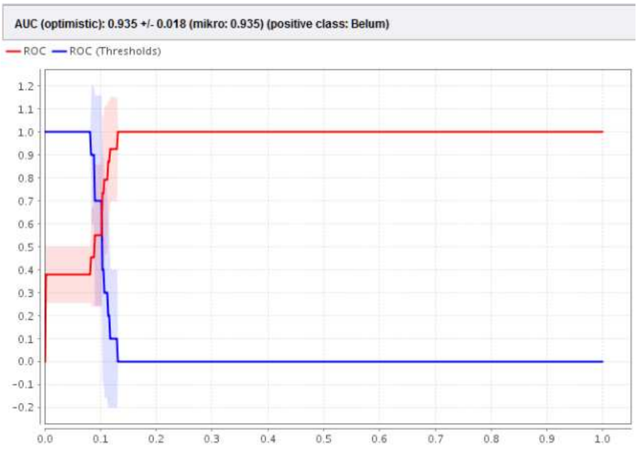

Fig 2. ROC – AUC curve accuracy

Based on Fig. 2 above, it can be seen that the AUC number is 0.935, then according to [14], the performance of the application of the algorithm KNN for the residential data status of rehabilitation houses after the eruption of Mount Merapi can be applied and is classified in the excellent classification

6. Conclusion

Based on the experimental results, it is concluded that the KNN Algorithm can be applied to classify the data on the occupancy status of the rehabilitation houses after the eruption of Mount Merapi. The classification accuracy results using the KNN Algorithm obtained an accuracy value of 82.03%, and the classification performance obtained an AUC (Area Under the ROC curve) value of 0.935. Based on the results of classification accuracy and classification performance values, the KNN Algorithm can be applied to classify the data on the occupancy status of post-eruption rehabilitation houses of Mount Merapi in the excellent classification category.

Acknowledgment

The Department of Informatics at Universitas Respati Yogyakarta, Indonesia, is responsible for conducting this study.

References

[1] Setiawan, B., "Enlightening assistance," First Printing, Ministry of Public Works, Directorate General of Human Settlements of the Republic of Indonesia, Jakarta, 2013.

[2] Satker Rehabrekon., "Sustainable Settlement Development: livable and sustainable, Satker Rehabrekon Houses Post Earthquake DIY & Central Java," Directorate General of Human Settlements. Jakarta, 2014.

[3] KPI-CSSRP., "Key Performance Indicators: Community-based Settlements Rehabilitation and Reconstruction Project," KPI Rekompak Central Java & DIY. JRF-PSF Status November 2014.

[4] Bramer, M., "Principles of data mining," Springer, London 2011

[5] Nurhadi W., "Application of Naive Bayes Classification Algorithm for Data Status Huni Home Rehabilitation and Reconstruction Assistance Fund Post-eruption of Mount Merapi in 2010," Proceedings of the national seminar, Multidisciplinary Approach to Disaster Management. Yogyakarta Respati University. Special Region of Yogyakarta, Vol. 1, No. 1, 2018.

[6] Nurhadi W., and Wayan O., "Evaluation of Naïve Bayes and Chi-Squar performance for Classification of Occupancy House," International Journal of Informatics and Computation (IJICOM), Vol.1, No. 2, pp. 46-54, 2019.

[7] Paryadi. C, M. Diqi, and SH Mulyani., "Implementation of CNN for Plant Leaf Classification," International Journal of Informatics and Computation (IJICOM), Vol. 2, No.2, 2020.

[8] M. Diqi, and M. F Muhdalifah., "Design and Building Javanese Script Classification in The State Museum of Sonobudoyo Yogyakarta," International Journal of Informatics and Computation (IJICOM), Vol. 1, No.2, 2019.

[9] T. A Utami, W. Ordiyasa, and Hamzah., "Sentiment Analysis of Hotel User Review using RNN Algorithm," International Journal of Informatics and Computation (IJICOM) Vol. 3, No.1, 2021.

[10] Sumarlin, "Implementation of the K-Nearest Neighbor Algorithm as Support for Classification Decisions for PPA and BBM Scholarship Recipients," Journal of Business Information Systems, 2015.

[11] H Risman, D. Nugroho, and YR WU, "Application of the K-Nearest Neighbor Method in Applications," TIKOMSIN Journal, Vol. 3, No. 2, pp. 19–25, 2013.

[12] Yeni Kustiyah ningsih and N. Syafa'ah, “Decision Support System for Determining Majors in High School Students Using KNN and Smart Methods,” Jsii, Vol. 1, No. 1, pp. 19–28, 2014.
[13] W. Yustanti., "K-Nearest Neighbor Algorithm to Predict Land Selling Price," Journal of Mathematics, Statistics and Computing, Vol.9, No.1, pp.57-68, 2012.
[14] Han, J. and Kamber, M., “Data Mining: Concepts, Models, and Techniques,” Springer, Verlag Berlin Heidelberg, 2011.
[15] Gorunescu, F., “Data Mining Concept, Models and Techniques, Springer, Craiova, Romania, 2011.