International Journal of Informatics and Computation (IJICOM)
Vol. 1, No.2, 2019
ISSN: 2685-8711, E-ISSN: 2714-5263

Evaluation of Naive Bayes and Chi-Square performance for Classification of Occupancy House

Nurhadi Wijaya[1], I Wayan Ordiyasa[2] Anang Faktchur Rachman[3]

Abstract
Occupancy status is one indicator of the rehabilitation and reconstruction program to support eruption victims in Indonesia. It needs to establish rehabilitation and reconstruction in a digital system with a structured database. In this paper, we provide dataset 2,146 occupied and 370 unoccupied houses. We utilize a naive Bayes classifier to classify the objects and implement a chi-square algorithm to measure comparison data to actual observed data. This research uses a combination of Naive Bayes and Chi-Square by applying weighting to the dataset attributes. Our study concludes that the combination of the algorithms can achieve a promising result in classifying the occupancy status. The combination of the techniques gains 89.59% accuracy and a ROC-AUC value of 0.839. Therefore, our approach is better than the standard Naive Bayes without combination with the Chi-Square approach.

Keywords
Data Mining, Naive Bayes, Chi-Square, Classification, Rehabilitation Housing.

1. Introduction

Mount Merapi disaster in 2010 destroyed various settlements in Yogyakarta and Magelang Regency, Central Java, Indonesia. Based on the head regulation of the National Disaster Management Agency regarding regional action plans, we undergo a study for the rehabilitation and reconstruction scheme for housing and settlements based on communities. This scheme aims to move settlements physically as well as to move their lives and livelihoods. The rehabilitation and reconstruction of settlements on new sites or land are carried out through a community empowerment approach by promoting a combination of development and community-based values [1]. In this paper, we conduct a study to measure the indicators of the successful performance of the rehabilitation and reconstruction program. Based on the performance indicators with local communities, the status of the habitable house is one of the main parts as a performance indicator of success. The more homes are occupied, the better its performance [2].

Currently, machine learning is a comprehensive method to deal with many problems [8][9][14][15]. The classification research of the status of occupied homes for rehabilitation and reconstruction after the Merapi disaster. We conduct the data analysis by using learning approaches like data mining to optimize classification results on the problem. A study proposed the classification of occupancy status by using Naive Bayes. The experiment utilized the Naive Bayes to classify the data for inhabitant's case with an

Corresponding Author: Nurhadi Wijaya[1]
1 Nurhadi Wijaya, Departement Of Informatics, Faculty of science and Technology, University of Respati Yogyakarta, Nurhadi@respati.ac.id
2 I Wayan Ordiyasa, Departement Of Informatics, Faculty of science and Technology, University of Respati Yogyakarta, wayanordi@respati.ac.id
3 Anang Faktchur Rachman, Departement Of Informatics, Faculty of Engineering, University of Madura, Pamekasan, East Java, Indonesia anang@unira.ac.id

accuracy of classification, reaching 89.59% and with the ROC-AUC reaching 0.826 [3]. Naive Bayes is a part of supervised learning that applies Bayes theorem with the naive assumption of conditional independence between every pair of features given the value of the class variable [13]. Various studies utilize the NB algorithm to address the computer application case [12].

Based on the previous research, we experiment with a combination of classification algorithms with the same dataset. Our proposed method combines the Naive Bayes with the Chi-Square algorithm. To train the model with those algorithms, we construct a dataset by collecting occupancy data from the work unit of the Government. To experiment, we gather 2,146 housing data and 370 houses unoccupied. It is a new classification technique to deal with the occupancy house problem by collecting a massive dataset from the real environment and combining the learning algorithm.

2. Methodology

In this paper, we collect primary data from the eruption environment as our benchmark dataset. Moreover, we conduct literature studies to enrich the sources of knowledge before conducting experiments. This research involves data analysis and training process by taking into account parameters, attributes, and variables. We take the sample population with probability sampling, where each element of the population has the same opportunity to be selected as subjects of the sample. We undergo sampling technique by using Stratified Random Sampling to reduce the influence of various factors and to divide the elements of the population into strata.

Evaluation and validity tests in this study were conducted by manually calculating data samples using the equation:

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)} \quad (1)$$

Note:

X: Data with unknown class

P (X) : Probability X

H: Hypothesis data X is a specific class

P (H | X): Hypothesis probability based on conditions (posterior probability)

P (H): Probability hypothesis H (prior probability)

P (X | H): Probability of X based on the conditions in hypothesis H

Classification with continuous data used the Gauss Density equation:

$$P(X_i = x_i \mid Y = Y_j = \frac{1}{\sqrt{2\pi\sigma ij}} e^{\frac{(xi-\mu ij)^2}{2\sigma^2 ij}} \quad (2)$$

Description

P : Opportunity

x_i : Attribute Value to i

X_i : Attribute to i

Y_j : Subclass Y sought

μ : Mean (average of all attributes)

Y : Class sought

σ : Standard deviation (variants of all attributes)

This study uses the Chi-Square selection feature Algorithm to evaluate the value of

features based on the calculation of the statistical value [4]. Chi-Square is one type of non-parametric comparative test conducted on two variables, where the scale of the data for both variables is nominal. (If of 2 variables, there is 1 variable with a nominal scale then a chi-square test is performed concerning the test that must be used at the lowest degree).

In this paper, we propose a new technique for evaluation of the validation of the housing dataset by combining the Naive Bayes (NB) with the Chi-Square algorithm. We utilize 10-fold cross-validation validation with Stratified Random Sampling by dividing data randomly into k sections and classifying each part. In the Cross-validation technique, it does as many experiments as k. For instance, the technique can process 10 times the k-value test to estimate accuracy [5]. Each trial uses one testing data, and the k-1 part becomes training data, then the testing data be exchanged for one training data to produce different testing data. We need the training data in learning to teach the model, and we feed the testing dataset as unseen information in the learning model. Finally, we can measure how effective the model in producing an accuracy of learning outcomes.

We provide classification accuracy, confusion matrix tables to depict the classification performance. This paper also presents the AUC curve to measure performance by estimating the output probability from randomly selected samples. In the model performance measuring, the greater the AUC score represents a better classification model. In this paper, we separate the performance of AUC accuracy into five groups [5], namely:

a. 0.90 - 1.00 = excellent classification (excellent classification).
b. 0.80 - 0.90 = good classification.
c. 0.70 - 0.80 = fair classification.
d. 0.60 - 0.70 = poor classification.
e. 0.50 - 0.60 = failure classification.

The Confusion matrix is one of the measurement tools to get the amount of accuracy of the dataset classification for active and inactive classes. Evaluation of the classification model is based on testing to estimate the right and wrong objects. We provide the test sequence in a confusion matrix form to show the predicted class result. Each cell has a number that indicates how many actual cases of the class are observed to be predicted [6]. Confusion matrix provides decisions result in training and testing and provides an assessment of the performance classification based on the object correctly or incorrectly [7]. We present the confusion matrix table as follows:

TABLE I
CONFUSION MATRIX

Classification		**Predicted Class**	
		Class = Yes	**Class = No**
Observed Class	**Class = Yes**	a (True Positive-TP)	b (False Negative-FN)
	Class = No	c (False Positive-FP)	d (True Negative-TN)

Note:
TP = Positive positive prediction FN = Positive negative prediction
FP = Positive negative prediction TN = negative negative prediction

For the calculation of the value of the results of the classification in this study be calculated the value of the accuracy of the classification results with the following calculation equation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

The detailed flow of the research scheme carried out is illustrated as follows:

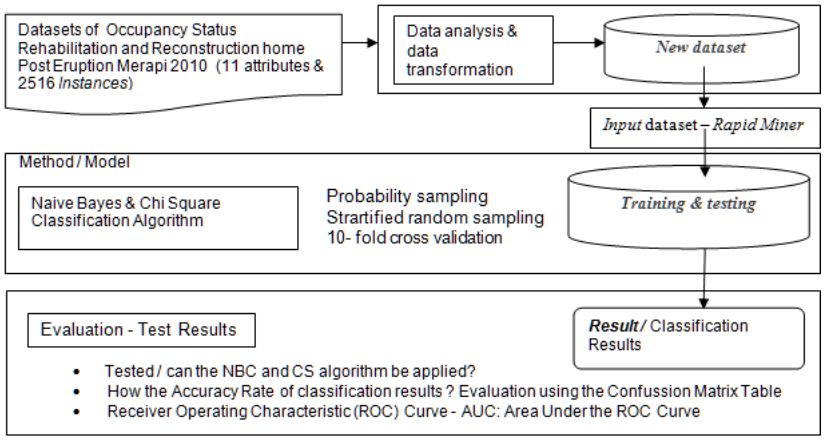

Fig.1. Details of the research scheme.

3. Dataset

In this paper, we gather dataset from the Secondary Database Management Information System (MIS) for the Merapi eruption in 2010. The experiment gathers 2,516 records and 11 attributes. The data is a real dataset that depicts residential status based on rehabilitation status by using the dataset as Table II:

TABLE II
DATASET ANALYSIS

No	Attributes	Information	Missing Values
1	Budget	Data *Poly nominal*, having Entities: BNPB 2011; JRF 2011; PSF 2011; PSF 2012; PSF 2013	No.Missing Values
2	Gender	data, *binominal* has the entity: Man and Woman	No.Missing Values
3	Number of Household Members	Data *Poly nominal.* With entities: Small (1-4 people); Medium (5-7 people); Large (more than 7 people)	No Missing Values
4	Prospective Female Occupants	Continuous Data - *Numeric-Integer*	No Missing Values
5	WaterWorks	Data *Binominal*. Has Entity Already and Not	No Missing Values
6	Electrical Works	Data *Binominal*. Has entities Already and Not	No Missing Values
7	Damage Level	Binominal Data has entities: Missing / Severely & Moderately Damaged / Minor Damage	No Missing Values
8	Non-Governmental Organizations	Binominal Data with entities: ≥10% & <10%	No Missing Values
9	Land Status	Poly nominal Data owns: independent; Collective Independent; Ground the village treasury (GVT)	No Missing Values
10	Building Area	Data Poly nominal which has the entity: 36 m2> 36 m2; <36 m2	Number of Missing Values
11	Occupancy Status	Data binominal that has entities. Already and yet, this data is a class from the dataset used.	No Missing Values

4. Result and Discussion

At the initial phase of the experiment, this study applies 10-fold cross-validation for data validation, where 10-fold cross-validation has a relatively low variant bias value. 10-fold cross-validation, the data is divided into 10 parts randomly in advance with the same comparison. We calculate the error rate for each section after calculating the average error rate of all data sections [4]. The study tries to address the case by manual calculations at the initial. However, we feed all of the datasets into RapidMiner data processing tools to compute 2516 samples. The modeling of the primary feature selection process using the Chi-Square algorithm on rapid miners can be seen in Fig. 2.:

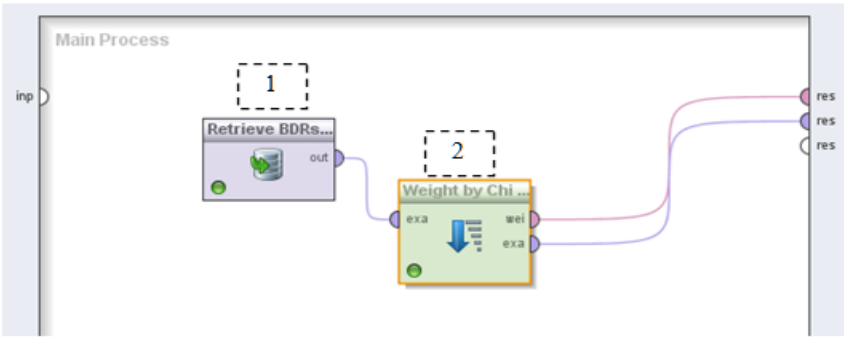

Fig. 2. Primary process feature selection data for occupancy using the Chi-Square algorithm

Explanation of numbers in Fig. 2.:

1. Data Set (Reconstruction funds recipient about 2516 records)
2. The Chi-Square algorithm as feature selection by optimizing of attribute dataset status of house funds benefit

From the modeling experiments in Fig. 2., we obtain the results of weighting attributes/variables as Table III:

TABLE III
ATTRIBUTE WEIGHTS WITH CHI ALGORITHM SQUARE

Attributes	Weight
Damage Level	0
Gender	0.004
Prospective Female Occupants	0.009
Building Area	0.014
Number of Household Members	0.015
Non-Governmental Organizations	0.090
Land Status	0.401
WaterWorks	0.425
Budget	0.569
Electricity Work	1

We compute the dataset with Naive Bayes and Chi-Square combination by using application tools, Rapid Miner. The modeling of this study can be seen in Fig. 3. & Fig. 4:

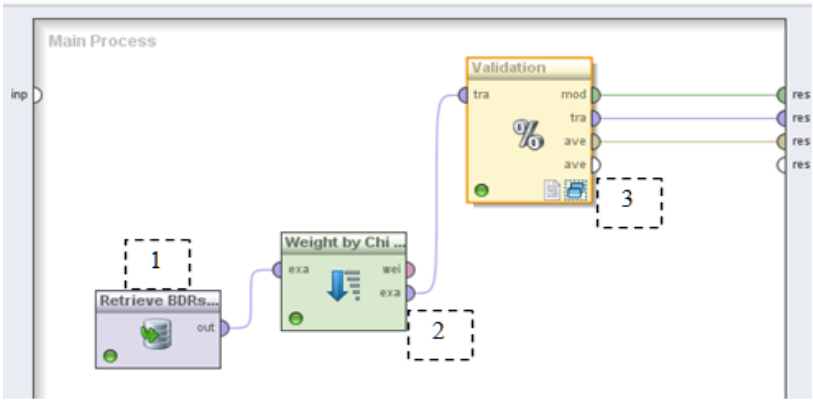

Fig. 3. Classification process for occupancy data using Naive Bayes and Chi-Square algorithm

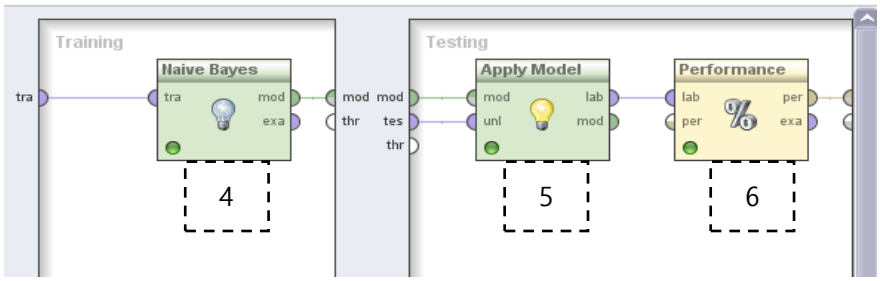

Fig. 4. Training and Testing data classification by using Naive Bayes and Chi-Square algorithm

Description number in Fig. 3. & Fig. 4. above:

1. Data Set (Rehabilitation and Reconstruction Houses Fund Recipient - as many as 2516 (record/instances) and 10 variables/attributes
2. Algorithm Chi-Square to select features by applying weighting attributes/variables
3. Validation (validating data with 10-fold cross-validation)
4. Algorithm Naive Bayes
5. Apply model
6. Performance measurement

The accuracy and performance results of applying the algorithms (Naive Bayes combine with Chi-Square) can be shown in Table IV, as follows:

TABLE IV
ACCURACY AND PERFORMANCE RESULTS

Naive Bayes Classification & Chi-Square	
Accuracy Value	AUC Performance
89.59%	0.839

The experimental results of Table IV show the results of the classification accuracy with the confusion matrix. It is a table to describe the performance of a classification model (classifier) on the dataset of test data to display which the true values are known.

The results of the confusion matrix table can be shown in the following Table V:

TABLE V
CONFUSION MATRIX TABLE RESULTS

Classification		Predicted Class	
		Class = Yes	Class = No
Observed Class	Class = Yes	a 2053	b 169
	Class = No.	d 93	d 201

In the testing process, we also calculate the confusion matrix score to produce the accuracy value based on equation (3). We calculate the accuracy to evaluate classification models. Informally, accuracy is how to determine the predictions our model got right or several correct predictions in our model. The results of the calculation are as follows

$$Accuracy = \left(\frac{2053+201}{2053+169+93+201}\right) = \left(\frac{2254}{2516}\right) x100\% = 89,59\% \quad :$$

Based on the above results of the accuracy, the level of classification accuracy with the Naive Bayes combined with the Chi-Square for occupancy status can achieve a promising result value of 89.59%.

As the final part, we test the combination algorithms to achieve a useful classification. The accuracy results in table IV and table V show that the accuracy of about 89.59% and the classification performance in the ROC-AUC curve in 0.839. We calculate the AUC score in classification analysis to determine which of the used models predicts the classes best. Fig. 5. depicts the visualization of AUC curves to show the classification performance:

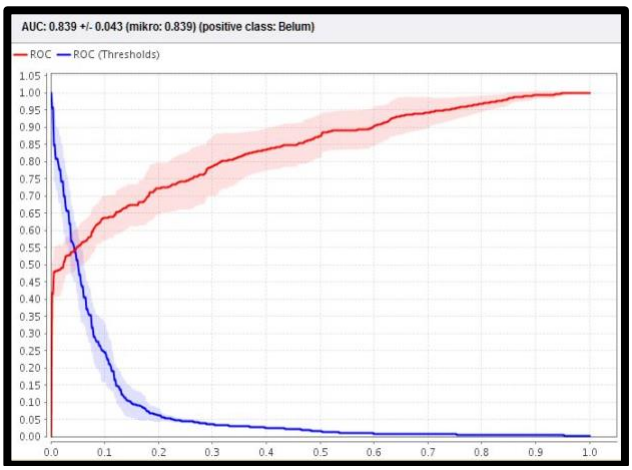

Fig. 5. ROC-AUC curves results

By demonstrating extensive experiments, we can achieve classification accuracy and performance up to 0.839. It produces a better accuracy than Naïve Bayes' standard implementation with the same dataset. Therefore, the proposed technique by combining Naive Bayes and Chi-Square algorithms can attain the classification performance for this case and can be a useful approach to deal with the post-eruption rehabilitation approach. So, the government or organization can apply the model to classify occupancy houses in a real condition [5].

5. Conclusion

The learning algorithms can be a new method to classify the occupancy house status after the Indonesia eruption. In this paper, we propose Naive Bayes and Chi-Square to classify the data on the occupancy status for the reconstruction of rehabilitation centers of the Merapi eruption. Based on the experiment, the classification accuracy reaching a value amounted to 89.59%. While the results of the classification accuracy performance obtained an AUC = 0.839. Further research development needs to combine classification algorithms with feature selection and weight optimization algorithms, for example, Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Ant Colony Optimization (ACO), and AdaBoost learning algorithm.

Acknowledgment

I would like to thank CSRRP, the World Bank, the permanent resident at Sleman Regency, and Magelang Regency, who helped researchers to complete this study.

References

[1] S. Bekti, *Enlightening Accompaniment*, 1st ed., Ministry of Public Works, Directorate General of Human Settlements of the Republic of Indonesia, 2013.

[2] KPI, "Key Performance Indicators: Community-based Settlements Rehabilitation and Reconstruction Project (CSRRP)," KPI CSRRP JRF-PSF Status, Nov. 2014. [Online]. Available: http://www.rekompakciptakarya.org/KPI. [Accessed: May 18, 2016].

[3] N. Wijaya, "Application of Naive Bayes Classification Algorithm for Occupancy Status Data, Home Fund Rehabilitation and Reconstruction Post Disasters of Merapi Eruption 2010," in *National Multidisciplinary Science Seminar*, University of Respati Yogyakarta, 2019.

[4] J. Indriyato, "Chi Square-Based K-Nearest Neighbor Algorithm for Insurance Customer Prediction," Postgraduate Thesis, University of Dian Nuswantoro, Semarang, Central Java, Indonesia, 2014.

[5] F. Gorunescu, *Data Mining: Concept, Models and Techniques*, 12th ed., Prof. L. C. Jain and Prof. J. Kacprzyk, Eds., Craiova, Romania: Springer, 2011.

[6] D. R. L. Polczynski, "WEKA Classification Using Decision Trees," Lecture Notes, Computer Science: College of Engineering & Applied Sciences, 2010.

[7] M. Bramer, *Principles of Data Mining*, London: Springer, 2011.

[8] P. Wanda, "A Survey of Intrusion Detection System," *International Journal of Informatics and Computation*, vol. 1, no. 1, pp. 1-10, Jan. 2020.

[9] S. Lestari, M. Diqi, and R. Widyaningrum, "Measurement of Maximum Value of Dental Radiograph to Predict the Bone Mineral Density," in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2017, pp. 1-4.

[10] M. Martinez-Arroyo and L. E. Sucar, "Learning an Optimal Naive Bayes Classifier," in *18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, 2006, pp. 1236-1239.

[11] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature Selection for Text Classification with Naïve Bayes," *Expert Systems with Applications*, Elsevier, 2009.

[12] Y. Qiu, G. Yang, and Z. Tan, "Chinese Text Classification Based on Extended Naïve Bayes Model with Weighed Positive Features," in *2010 First International Conference on Pervasive Computing, Signal Processing and Applications*, Harbin, 2010, pp. 243-246.

[13] Y. Huang and L. Li, "Naive Bayes Classification Algorithm Based on Small Sample Set," in *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, Beijing, 2011, pp. 34-39.

[14] P. Wanda and H. J. Jie, "URLDeep: Continuous Prediction of Malicious URL with Dynamic Deep Learning in Social Networks," *I. J. Network Security*, vol. 21, pp. 971-978, 2019.

[15] M. Diqi and M. Mujastia, "Design and Building Javanese Script Classification in The State Museum of Sonobudoyo Yogyakarta," *International Journal of Informatics and Computation*, vol. 1, no. 2, pp. 35-45, Feb. 2020.