Scientific Journal of Informatics
Vol. 5, No. 1, May 2018

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor

Yofi Firdan Safri[1], Riza Arifudin[2], Much Aziz Muslim[3]

[1,2,3]Computer Science Department, FMIPA, Universitas Negeri Semarang, Indonesia
Email: [1]yofi.firdan26@gmail.com, [2]riza.arifudin@gmail.com,
[3]a212muslim@yahoo.com

Abstract

Health is a human right and one of the elements of welfare that must be realized in the form of giving various health efforts to all the people of Indonesia. Poverty in Indonesia has become a national problem and even the government seeks efforts to alleviate poverty. For example, poor families have relatively low levels of livelihood and health. One of the new policies of the Sakti Government Card Program issued by the government includes three cards, namely Indonesia Smart Card (KIP), Healthy Indonesia Card (KIS) and Prosperous Family Card (KKS). In this study to determine the feasibility of a healthy Indonesian card (KIS) required a method of optimal accuracy. The data used in this study is KIS data which amounts to 200 data records with 15 determinants of feasibility in 2017 taken at the Social Service of Pekalongan Regency. The data were processed using the K-Nearest Neighbor algorithm and the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm. This can be seen from the accuracy of determining the feasibility of K-Nearest Neighbor algorithm of 64%, while the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm is 96%, so the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm is the optimal algorithm in determining the feasibility of healthy Indonesian card recipients with an increase of 32% accuracy. This study shows that the accuracy of the results of determining feasibility using a combination of K-Nearest Neighbor-Naive Bayes Classifier algorithms is better than the K-Nearest Neighbor algorithm.

Keyword: Determination of feasibility, Poverty, K-Nearest Neighbor, Naive Bayes Classifier Algorithm

1. INTRODUCTION

Indonesia is one of the most populous countries in Asia. Economic growth negatively affects poverty in Indonesia. The unemployment rate has a positive effect on poverty in Indonesia. Government spending on poverty alleviation has no effect on poverty in Indonesia. Government policy should encourage economic growth, where high economic growth can increase national income and will directly increase per capita income of every resident [1].

Poverty is a deficiency situation that happens and is not desired by everyone [2]. Poverty can infect every level and sphere of life, from individual to state level [3]. For example, poor families have a relatively low level of livelihood and health compared to people whose lives are sufficient [4]. There are various factors that cause poor households, among others, the termination of employment from the office or company for those who were previously employed to become unemployed and have no income, low education and no skills so difficult to find work, the change of poor criteria from the Central Bureau of Statistics (BPS) [5]. Poverty is multidimensional because the human needs are diverse, seen from the aspect of primary and secondary aspects [6].

There are nine points Nawacita Joko Widodo, of the nine there is nothing specifically related to the field of health. But as long as political action is able to draw attention from the public that is considered spectacular health package KIS. Health is a human right and one of the elements of welfare that must be realized in the form of giving various health efforts to all the people of Indonesia through the implementation of development of quality health and affordable by the community. Development in the health sector is directed towards achieving awareness, willingness and ability to live healthy for every resident [7]. The Sakti Card Program is intended for the underprivileged and underprivileged Indonesians [8]. Some people think that the launching of three magic cards is politically charged. It coincides with the policy plan of fuel price hike (BBM). Others support the program [9]. In general, the problems that will arise in the field are related to the target or category of KIS recipients.

Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases [10, 11]. Activities that include collection and use of historical data to find regularities, patterns or relationships in large data sets [12]. Classification is a new record of data to one of several predefined categories (or classes). Also called supervised learning [13]. The task of classification is to map data into class groups [14]. Data mining classification techniques can be used to determine feasible or not feasible to get a Healthy Indonesia Card. The outputs generated by the data mining classification can be used for knowledge. The classification of community data plays a role to determine who is feasible and who is not objectively and accurately. One of the methods to be used is with data mining.

According to research [15], text classification obtained results of accuracy for the use of Naive Bayes Classifier method 86.7%, and K-Nearest Neighbor (KNN) 87.57%. The combination of Decision Tree and Naive Bayes Classifier is used to overcome the difficulties of continuous attributes, missing attribute values and noise (noise) in the training process. The test results achieved high detection rates and significantly reduced FP (False Positives) for different types of disorders [16]. In 2013 the combination of Naive Bayes Classifier and K-Nearest Neighbor to predict the 12 positions of profitability of financial institutions in Bangladesh Country [17]. The classification algorithm combines several algorithms conducted in 2004, combining Bayesian network algorithms and K-Nearest Neighbors for data analysis, predicting cancer class classes into three DNA microarray datasets namely Colon, Leukemia and NCI-60 [18].

Based on the background, the purpose of this study is to obtain a model and compare the accuracy results using the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm.

2. METHODS

2.1. Data Collection

The data used in this study was taken randomly through the data of the healthy card Indonesia cardee feasibility in 2017 at the Office of Social Affairs of Pekalongan Regency. The amount of data taken as many as 200 records, consisting of 119 records as feasible as a determinant of the community meets the criteria for receiving KIS and 81 records that are not feasible to show that the community does not meet the criteria for receiving KIS.

2.2. Data Processing

In data processing, the process of grouping data to determine the variables to be used, performing data representation into numerical form and doing data sharing into training data and test data [19].

2.3. The Algoritm Used

In this research will be done comparative analysis using two classification algorithm from data mining. The proposed algorithm is a combination of K-Nearest Neighbor algorithm with Naive Bayes Classifier algorithm, then evaluate and validate the result with confusion matrix. The next stage is to compare the results of accuracy and time complexity of each algorithm, to obtain the model of the classification algorithm which obtains the highest accuracy and time complexity.

The combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm is done by finding the probability value of each attribute data to be classified on each attribute $P(x \mid c_i)$, then the data having a greater probability of α will be tested using K-Nearest Neighbor algorithm . Calculate D (x, y) with the K-Nearest Neighbor algorithm for each stored data. The last step determines the order of the minimum value of D (x, y) on the calculation result. The data input comes from the trainer data then the expected output is the result of the prediction based on the closest distance to the K-Nearest Neighbor algorithm. The combination of these two methods is useful for accelerating the performance of the K-Nearest Neighbor algorithm so it is not necessary to calculate the overall data, but to calculate from the probability possible. Flowchart the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm, as shown in Figure 1.

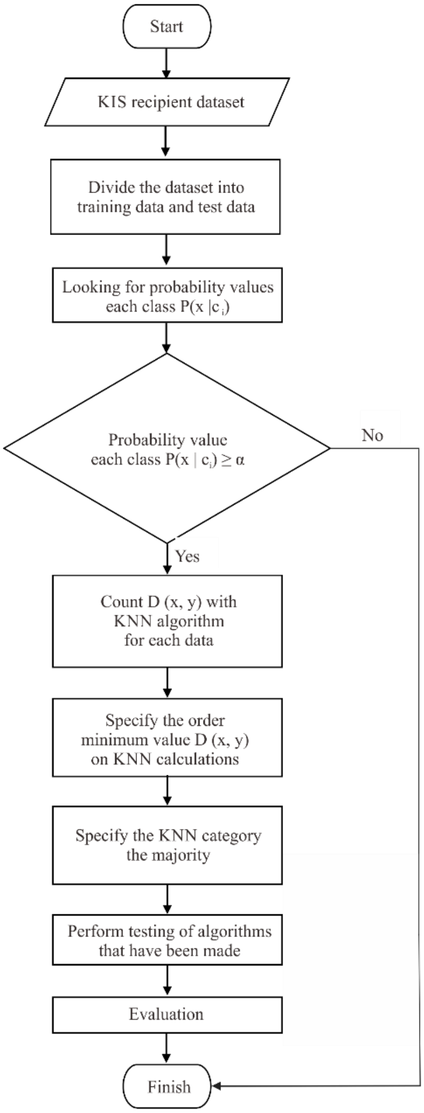
Start
KIS recipient dataset
Divide the dataset into training data and test data
Looking for probability values each class P(x |c i)
Probability value each class P(x | ci) ≥ α
No
Yes
Count D (x, y) with KNN algorithm for each data
Specify the order minimum value D (x, y) on KNN calculations
Specify the KNN category the majority
Perform testing of algorithms that have been made
Evaluation
Finish

Figure 1. Flowchart the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm

3. RESULT AND DISCUSSION

The software used in this research is Matlab R2013b. By utilizing Matlab, can be done data analysis, algorithm development and create models and applications and can also be made visual display of a program so that it can facilitate the user.

This study uses the dataset of healthy card recipients obtained from the Office of Social Affairs of Pekalongan Regency. The data contained 200 data records, had 15 attributes and 1 class attribute. These attributes include age, floor area of the building, type of building floor, type of wall, defecation facility, drinking water source, main household lighting source, daily cooking fuel, meat/chicken/dairy per-week, daily feeding frequency for each ART, the ability to buy new clothes for each ART within one year, the ability to pay for medical treatment at the Puskesmas/Polyclinic, household head's income, highest education of household head, asset/saving [20]. In the class attribute has two values that are feasible and not feasible. The data sharing in this research is 75% for the process with a number of 150 records of training data and 25% for the test process using test data with 50 data records.

Steps to facilitate the mining process in the system then attributes that have category type represented in numeric form 1 and 0, that is 1 for Yes and 0 for not. While class attributes are also represented in numerical form 1 and 0, ie 1 for proper and 0 for improper. Data ready for mining process can be seen in Table 1.

Table 1. Data for mining process

Name	NIK	A1	A2	A3	...	A15	A16
Ruwat	3326080304680021	49	1	1	...	6	1
Kundiyah	3326085205880001	29	0	0	...	18	0
Supardi	3326080107660429	51	1	0	...	10	1
Kasiroh	3326081008860001	31	1	0	...	26	0
Barokah	3326086205730001	44	0	1	...	10	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Wasri	3326086312720001	45	1	1	...	10	1

3.1. Mining process on K-Nearest Neighbor algorithm

The advantage of applying the K-Nearest Neighbor algorithm is that the training process runs faster and is more flexible because it is based on the proximity of existing training data [21,22]. First, attribute grouping is done based on the classification of the feasibility of discrete data and continuous data.

After going through the calculation process using the K-Nearest Neighbor algorithm by using the Euclidean distance squared, then enter the value of [23, 24, 25] the data in the process to get the value of accuracy and execution time. The model obtained from K-Nearest Neighbor algorithm method is then tested using 75% training data and 25% test data. Obtained table confussion matrix as shown in Table 2.

Table 2. Configuration Matrix Test Results 50 Test Data On K-Nearest Neighbor Algorithm

Classification		Predicted Class	
		Class =feasible	Class = not feasible
Observed Class	Class =feasible	25	2
	Class = not feasible	16	7

The accuracy value is the proportion of the correct number of predictions. Can be calculated using Equation 1 [26]:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}\ x\ 100 \qquad (1)$$

The accuracy of the total test data that is correctly classified can be calculated based on the calculation formula of measurement accuracy obtained:

$$Accuracy = \frac{41}{50}\ x\ 100 = 64\%$$

3.2. The mining process of combination K-Nearest Neighbor-Naive Bayes Classifier algorithm

The Naive Bayes Classifier algorithm is a statistical classification method based on the bayes theorem [27]. The Naive Bayes Classifier algorithm model has a very minimum error rate [28] and is known for its simple, fast, and highly accurate calculations [29]. Use of Naive Bayes Classifier would be better if more training data. Required training data as precise as possible and the result will be better [30].

Mining process by applying the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm, firstly calculated using Naive Bayes Classifier algorithm and then proceed with K-Nearest Neighbor algorithm.

The data sharing in this research is 75% for the process with a number of 150 records of training data and 25% for the test process using test data with 50 data records. After the data-sharing process, separate discrete data and continuous data then calculate the mean and standard deviation values of continuous data, the first thing to do is to determine the mean value or mean and standard deviation of the feasible class and not feasible classes in each attribute , ie age, income of head of household, asset / saving ownership. Table 3 shows the mean (μ) and standard deviation (S) results for each feasible and not feasible class of the three attributes.

Table 3. List of calculations mean (μ) and standard deviation (S)

Variable	Mean		Standard Deviation	
	Feasible	Not Feasible	Feasible	Not Feasible
Age	48,87	31,28	9,93	1,63
Source of income	10,25	18,52	3,10	4,12
Savings	10,93	21,24	2,97	5,22

The result of calculating the mean value (μ) and the standard deviation (S) of each attribute that has the continuous data type can be seen in Table 2, then the determination of the healthy card recipients will be calculated using the Naive Bayes Classifier method with the Gauss dentity formula for attributes which has a continuous type, whereas for data the category type is calculated the probability of occurrence of each value for a variable that has category type [22].

Once the probability value of each attribute is known, the next step is to choose a probability value that has a value greater than alpha, for example the alpha value = 0.30. The purpose of choosing the probability value of each attribute is greater than the alpha that is used for later calculations on the K-Nearest Neighbor algorithm. The probability value of each attribute can be seen in Table 4.

Table 4. The probability value of each attribute

Attribute	Probability Value		Value
	Feasible	Not Feasible	
A1	0,12	0,06	Not Used
A2	0,79	0,33	Used
A3	0,93	0,55	Used
A4	0,76	0,45	Used
A5	0,51	0,34	Used
A6	0,70	0,52	Used
A8	0,20	0,60	Used
A9	0,78	014	Used
A10	0,54	0,21	Used
A11	0,73	0,31	Used
A13	0,10	0,13	Not Used
A14	0,86	0,29	Used
A15	0,10	0,06	Not Used

Attributes that have a greater probability value than alpha are sorted then the next step is to calculate with K-Nearest Neighbor algorithm for each data. The calculation formula of K-Nearest Neighbor algorithm using Euclid distance square has Equation 2 as follows:

$$D(x,y) = \sqrt{\sum_{k-1}^{n}(x_k - y_k)^2} \qquad (2)$$

Steps to calculate the K-Nearest Neighbor algorithm. Specifies the parameter of value n (the closest number of neighbors), given the value parameter n=4. Calculate the distance between the new data and all data in the training data. For example, the Euclid distance square is used from the distance between the new data and all data in the training data can be seen in Table 5.

Table 5. Calculation of the square of euclide distance

Attribute							the square of euclide distance
A2	A3	A4	A5	A6	...	A15	(A2=1,A3=1,A4=1,...,A15=14)
1	1	1	1	1	...	8	1,41
0	0	1	0	0	...	16	2,82
1	0	1	0	1	...	10	2,00
⋮	⋮	⋮	⋮	⋮		⋮	⋮
1	1	1	1	1	...	14	2,00

Then sort the records that have the smallest Euclid distance based on the minimum distance to-n. Once sorted by the minimum distance to-n then determine using the category of K-Nearest Neighbor the most majority.

After going through the calculation process then the data in the process to get the value of accuracy and execution time using confussion matrix. Obtained table confussion matrix as shown in Table 6.

Table 6. Confusion Matrix Test Results 50 Test Data On K-Nearest Neighbor Combination Algorithm with Naive Bayes Classifier

Classification		Predicted Class	
		Class =feasible	Class = not feasible
Observed Class	Class =feasible	25	2
	Class = not feasible	0	23

The accuracy of the total test data that is correctly classified can be calculated based on the calculation formula of measurement accuracy obtained:

$$Accuracy = \frac{48}{50} x\ 100\% = 96\%$$

From the results obtained, there is an increase of accuracy and execution time of the K-Nearest Neighbor combination algorithm with Naive Bayes Classifier which can be seen in Table 7.

Table 7. Comparison of Accuracy

Algorithm	Accuracy	Execution Time
KNN algorithm	64%	0,01428 s
Combination of KNN & NB	96%	0,00118 s

By applying the Naive Bayes Classifier algorithm to K-Nearest Neighbor it is evident that the Naive Bayes Classifier is an algorithm to improve the accuracy of the K-Nearest Neighbor algorithm in determining the feasibility of healthy Indonesian card recipients; the Naive Bayes Classifier algorithm aims to minimize variation within an attribute to obtain accurate higher than the K-Nearest Neighbor algorithm alone. For further research is expected to classify more than 2 types of classes and added the number of attributes determining the feasibility in order to get a higher level of accuracy.

4. CONCLUSION

The combination algorithm can overcome the weakness of K-Nearest Neighbor algorithm with faster time process and weakness in Naive Bayes Classifier with higher accuracy percentage. The accuracy of the K-Nearest Neighbor algorithm is 64%, while the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm produces 96% accuracy and execution time at KNN 0,01428 second after using the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm 0,00118 second so that after applying the combination of K-Nearest Neighbor-Naive Bayes Classifier algorithm in determining the classification of KIS for the poor has an accuracy increase of 32% and has an increase in execution time of 0,0131 second.

5. REFERENCES

[1] Ramdani, M. (2015). Determinan Kemiskinan Di Indonesia Tahun 1982-2012. *Economics Development Analysis Journal*, *4*(1), 58-64.

[2] Yuniarti, N. (2012). Eksploitasi Anak Jalanan Sebagai Pengamen dan Pengemis Di Terminal Tidar Oleh Keluarga. *Komunitas*, 4 (2),210 – 217.

[3] Hadim. (2009). Dinamika Kemiskinan Rumah Tangga Di Pedesaan (Studi Kasus Desa Malasari, Kecamatan Nanggung, Kabupaten Bogor, Propinsi Jawa Barat). *Skripsi*. Fakultas Pertanian, Institut Pertanian Bogor. Hal.1 - 203.

[4] Puspita, D. W. (2015). Analisis Determinan Kemiskinan di Provinsi Jawa Tengah. *JEJAK: Jurnal Ekonomi dan Kebijakan*, *8*(1), 101-107

[5] Nugroho, N. A. (2010). “Faktor-Faktor Penyebab Meningkatnya Rumah Tangga Miskin Di Kecamatan Suruh Kabupaten Semarang”. *Skripsi*. FE, Ekonomi Pembangunan, Universitas Negeri Semarang.

[6] Annur, R. A. (2013). Faktor-Faktor yang Mempengaruhi Kemiskinan di Kecamatan Jekulo dan Mejobo Kabupaten Kudus Tahun 2013. *Economics Development Analysis Journal*, *2*(4), 409 – 426.

[7] Ilyas, A., dkk. (2014). Pertanggungjawaban Pidana Bagi Dokter Dalam Malpraktik Medik Di Rumah Sakit.

[8] Hukum Online. (2015, 3 Maret). *Presiden pun Akui Ada Masalah BPJS Kesehatan*. Diperoleh 26 Januari 2017, dari http://www.hukumonline.com/berita/baca/lt54f56e21e54eb/presiden-pun-akui-ada-masalah-bpjs-kesehatan/.

[9] Hukum Online. (2014, 14 November). *Pro Kontra Tiga Kartu Sakti Ala Presiden Jokowi*. Diperoleh 27 Januari 2017, dari http://www.hukumonline.com/berita/baca/lt5465ef9669c1f/pro-kontra-tiga-kartu-sakti-ala-presiden-jokowi/.

[10] Kusrini dan Luthfi, E. T. (2009). *Algoritma Data mining*. Yogyakarta: Penerbit Andi.

[11] Kaur, H. & Kaur, H. (2013). Proposed Work for Classification and Selection of Best Saving Service for Banking Using Decision tree Algorithms. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(9), 680-684.

[12] Santosa, B. (2009). *Data mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.

[13] Hermawati, F.S. (2013). *Data Mining*. Surabaya:Penerbit Andi.

[14] Jain, V., Narula,G.S.,& Singh, M. (2013). Implementation of Data Mining in Online Shopping System using Tanagara Tool. *International Journal of Computer Scienceand Engineering*, 2(1), 47-58.

[15] Danesh, A., Moshiri, B., & Fatemi, O. (2007, July). Improve text classification accuracy based on classifier fusion methods. In *Information Fusion, 2007 10th International Conference on* (pp. 1-6). IEEE.

[16] Farid, D.M., Harbi, N. & Rahman , M.Z. (2010). Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection. *International Journal of Network Security & Its Applications (IJNSA)*, 2(2), 12-25.

[17] Ferdousy, E.Z., Islam, M.M. & Matin, M.A. (2013). Combination of Naive Bayes Classifier and K-Nearest Neighbor (cNK) in the Classification Based Predictive Models. *Computer Science and Information Science*, 6(3), 48-56.

[18] Sierra, B., Lazkano, E., Martínez-Otzeta, J. M., & Astigarraga, A. (2004, December). Combining Bayesian networks, k nearest neighbours algorithm and attribute selection for gene expression data analysis. In *Australasian Joint Conference on Artificial Intelligence* (pp. 86-97). Springer, Berlin, Heidelberg.

[19] Prasetyo, E. (2012). *Data mining Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi.

[20] Skpd Batam Kota. (2014, 25 Agustus). *14 Kriteria Miskin Menurut Standar Bps*. Diperoleh 27 Januari 2017, dari http://skpd.batamkota.go.id/sosial/persyaratan-perizinan/14-kriteria-miskin-menurut-standar-bps/.

[21] Hidayah, M. R., Aklis, I. & Sugiharti, E. (2017). Recognition Number of The Vehicle Plate Using Otsu Method and K-Nearest Neighbour Classification. *Scientific Journal of Informatics*, 4(1), 66 – 75.

[22] Rohmana, I., & Arifudin, R. (2014). Perbandingan Jaringan Syaraf Tiruan dan Naive Bayes dalam Deteksi Seseorang Terkena Penyakit Stroke. *Jurnal MIPA, 37*(2), 178-191.

[23] Bouzalmat, A., Kharroubi, J. & Zarghili, A. (2013). Face Recognition Using SVM Based on LDA. *IJCSI International Journal of Computer Science Issues,* 10(1), 171 – 179.

[24] Kumar, K. S. & Chezian, D. R. M. (2014). Support Vector Machine and K-Nearest Neighbor Based Analysis for the Prediction of Hypothyroid. *International Journal of Pharma and Bio Sciences*, 5(4), 447- 453.

[25] Bharathi, D. A., & Deepankumar, E. (2014). Survey on Classification Techniques in Data Mining. *International Journal on Recent and Innovation Trends in Computing and Communication*, *2*(7), 1983-1986.

[26] Wati, R. (2016). Penerapan Algoritma Genetika untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan Menggunakan Naïve Bayes. *Jurnal Evolusi*. 4(1).

[27] Baby, N & P.L.T. (2012). Customer Classification And Prediction Based On Data Mining Technique. *International Journal of Emerging Technology and Advanced Engineering*, 2(12), 314-18.

[28] Pratiwi, R. W. & Nugroho, Y. S. (2016). Prediksi Rating Film Menggunakan Metode Naïve Bayes. *Jurnal Teknik Elektro*, 8(2), 60 – 63.

[29] Syarifah, A. & Muslim, M. A. (2015). Pemanfaatan Naive Bayes untuk Merespon Emosi dari Kalimat Berbahasa Indonesia. *Unnes Journal of Mathemathic*, 4 (2),147-156.

[30] Sugiharti, E., Firmansyah, S. & Devi, F. R. (2017). Predictive Evaluation Of Performance Of Computer Science Students Of Unnes Using Data Mining Based On Naive Bayes Classifier (Nbc) Algorithm. *Journal of Theoretical and Applied Information Technology.*95(4), 902 – 911.