# Naive Bayes Classifier Algorithm Approach for Mapping Poor Families Potential

Sri Redjeki, M. Guntara, Pius Anggoro
Informatics Engineering Department
STMIK AKAKOM
Yogyakarta, Indonesia

***Abstract*—The poverty rate that was recorded high in Indonesia becomes main priority the government to find a solution to poverty rate was below 10%. Initial identification the potential poverty becomes a very important thing to anticipate the amount of the poverty rate. Naive Bayes Classifier (NBC) algorithm was one of data mining algorithms that can be used to perform classifications the family poor with 11 indicators with three classifications. This study using sample data of poor families a total of 219 data. A system that built use Java programming compared to the result of Weka software with accuracy the results of classification of 93%. The results of classification data of poor families mapped by adding latitude-longitude data and a photograph of the house of the condition of poor families. Based on the results of mapping classifications using NBC can help the government in Kabupaten Bantul in examining the potential of poor people.**

***Keywords—Data Mining; Naive Bayes; Poverty Potential; Mapping***

## I. INTRODUCTION

Poverty in Indonesia has a number that is still quite high, above 10% [12]. It is becoming a top priority for the government to find solutions to reduce the poverty rate is below 10% [2]. Central Statistics Bureau (BPS) defines poverty as the inability to meet the minimum standards of basic needs that include food and non-food needs. BPS showed that the poverty rate in Indonesia, in September 2014, was still high at about 27.7 million people, or approximately 10.96% [11]. The poverty data graph shown in Figure 1.

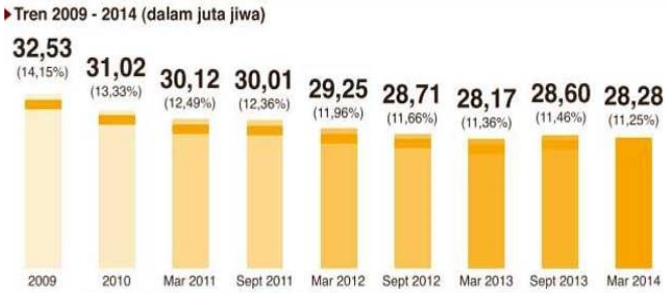


Fig. 1. Poverty Data Graph in Indonesia (bps.go.id)

The number of poor people in Indonesia are mostly locate in Java island with a total of 57.8% of the total number of poor people in Indonesia, and it is located in Yogyakarta province. Poverty measurement in each country or even in every region that does not have the same size [8][9]. The poverty measurement that called as the poverty indicator becomes the most important part in determining poverty status [8][9]. In Bantul, which is one district in Yogyakarta has a fairly high poverty rate, above 14%.

The determination in classifying the poverty status of someone is the tough section that needs hard effort because it must represent the accurate results. Naive Bayes Classifier is one of the data mining algorithms that uses probabilistic approach [1][4][5]. This research will discuss how Naive Bayes Classifier algorithm can classify the status of poor families to identify potential poverty based on existing indicators. There were 11 indicators of poor families used in this study, and each of them has certain value [10]. The indicators were food, clothing, shelter, income, health, education, wealth (rupiah), property (land), water, electricity and the number of family members. While the classification used is very poor, poor and vulnerable poor [10][12].

## II. LITERATURE REVIEW

### A. Poverty

Poverty is a matter of deprivation or problematic deficiencies. Poverty is a condition where a person or a family is in a state of deprivation [2][9]. From these definitions, poverty can be divided into two parts: absolute and relative,

*a) Absolute poverty is defined as the inability to achieve a minimum standard of life. Understanding the needs of different minimum standards in each country.*

*b) Relative poverty, on the other hand, is defined as the inability to achieve the standards of contemporary needs , which is linked to the welfare-rata average or average income community planning at the time.*

Based on the data, the factors are distinguished into the data that affect poverty in the countryside and in urban areas, too. The comparison is important because poverty does not only happen in rural area but also in urban area. Based on this geographical approach, then poverty can be differentiated into poverty in rural and urban areas.

*a) Rural poverty is a poverty which has the characteristics such as: i) limited access to the ground facilities and irrigation, ii) the slow adaptation to modern technology, iii) too large burden borne, iv) limited human capital, v) only concentrated in rural areas and vi) only concentrated on certain ethnic minorities [9].*

*b) Urban poverty is a poverty which has the characteristics such as: i) have limited access to resources and services, ii) limited human resources quality, iii) too large burden borne, iv) the low wages earned, v) big amount of the disorganized small enterprises and vi) a big amount of groups that do not have the capability [9].*

### B. Data Mining

Data Mining as a process to obtain useful information from a data warehouse [6] [7]. The term data mining is often called knowledge discovery. One technique that is made in data mining is to explore existing data to build a model and then use that model in order to identify the pattern of other data that is not stored in the database [6].

### C. Naive Bayes Classifer (NBC)

Naive Bayes Classifier estimates the conditional class opportunities which assume that the attributes are conditionally independent and given the class label *Y* [3][5] Conditional independent assumptions can be expressed in the following form :

$$P(X|Y=y)=\prod_{i=1}^{d}P(X_i|Y=y) \quad \text{..... (1)}$$

each set of attributes consisting of *d* attributes. There is a special treatment before the features with numeric data types are put into Naive Bayes. The first way is to use discretization and the assumption of a Gaussian distribution. Gaussian distribution was chosen to represent the conditional probability that a continuous feature in a class independency P(Xi|Y). This Gaussian distribution approach was used by the researcher to obtain a probability value of each poverty indicator.

Generics Naive Bayes Classifier Algorithms:

*1) Read attributes and class of the data set.*
*2) Calculate the posterior probability of each attribute to an existing class.*
*3) Calculate the probability pior of existing classes.*
*4) Calculate the multiplication value of the posterior probability of each class and the value prior to all existing classes.*
*5) Find the greatest probability value in step four as the final classification.*

## III. METHODOLOGY

The data has used in this study taken from the poor families in the Kabupaten Bantul. The overall system can be seen in the block diagram that existed at Figure 2.

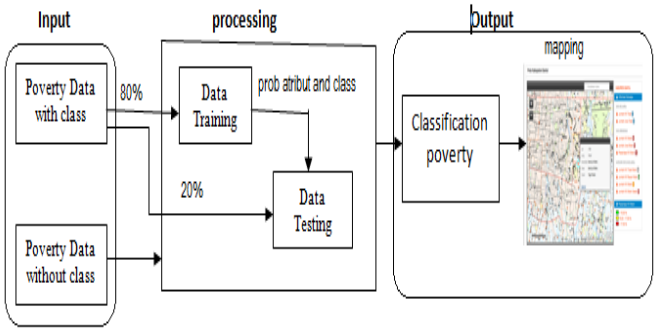


Fig. 2. Diagram Block System

The block diagram system represented in Figure 2 was divided into three parts. The first part was the data input which was consisted of three classes of poverty (poverty status) and the poverty data which would be used for identification. Poverty classes consist of very poor, poor and vulnerable poor. The number of parameters for classification was composed from 11 indicators as presented on Table 1.

TABLE I. POVERTY INDICATOR

| No | Indicator | Indicator Score |
|---|---|---|
| 1 | Food | (0,12) |
| 2 | Clothing | (0,9) |
| 3 | Shelter | (0,9) |
| 4 | Income | (0,35) |
| 5 | Health | (0,6) |
| 6 | Education | (0,6) |
| 7 | Wealth (Rupiah) | (0,5) |
| 8 | Property (Land) | (0,6) |
| 9 | Water | (0,4) |
| 10 | Electricity | (0,3) |
| 11 | Number of Familly members | (0,5) |

There were 219 data which were divided into two parts: 80% (175 data) were used for training data and 20% (approximately 44 data) were used for data testing. The second part was the main process of Naive Bayes Classifier that calculated a probability value to be used for classification. The calculated data was the training data set. The training phase results were in the form of probability values which would be used for testing.

Phase testing was done to see the accuracy of the obtained classification. The third section resulted the classification of the poverty class which would be mapped to see the poverty potential in an area by using Google Maps.

The training process (training) on the algorithm of Naive Bayes Classifier (NBC) can be seen in Figure 3.

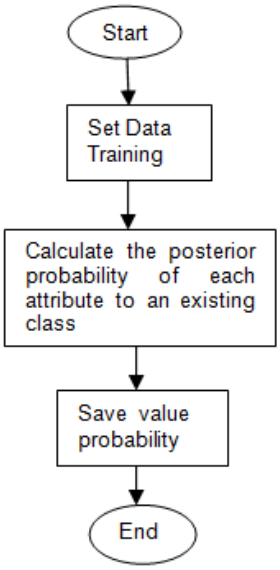


Fig. 3. Training Fase Naive Bayes

Figure 3, showed the step-by-step process of Naïve Bayes Classifier algorithm which included reading the data sets. The display of training data is shown in Figure 4.

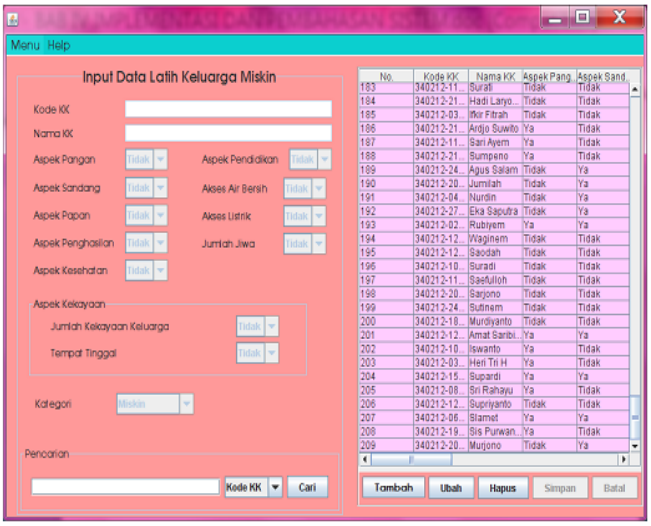


Fig. 4. Data Training Display

The training input from the indicators was Yes and No category which can be seen in Figure 4. Input yes represent value a score largest while value not represent value a score smallest (see. indicator score in table 1). Whereas for the testing process can be shown in Figure 5. The data used for testing were as much as 20% which was approximately 44 data from 219 data.

The results of the testing phase was used to calculate the probability of each classification by using a probability value obtained in the training phase to determine the poverty classification results by taking the smallest probability. In the testing phase, it can be seen that the high accuracy of the identification of the poor people status in Kabupaten Bantul. Figure 6 represented the display of testing menu interface.

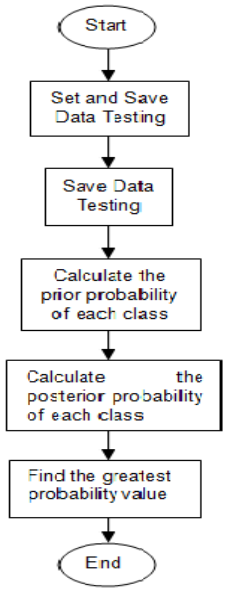


Fig. 5. Testing Fase Naive Bayes

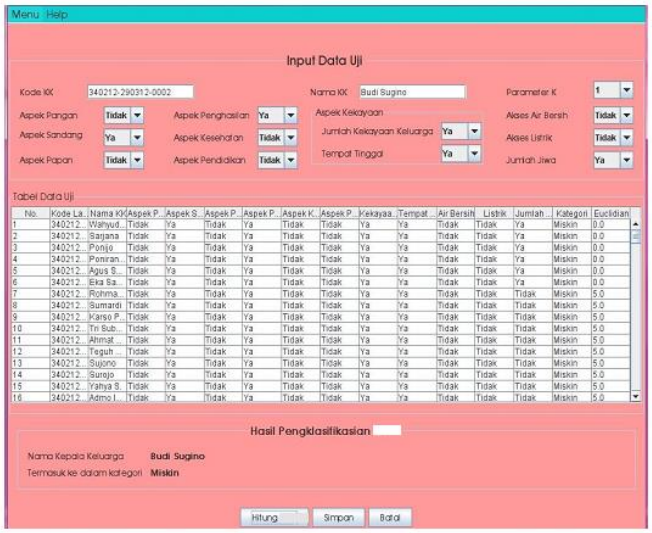


Fig. 6. Data Testing Display

## IV. Result and Discussion

The implementation of Naïve Bayes algorithm to determine the classification of poverty was built using Java. These results were used as the input for mapping poor families.These results would be mapped using Google Maps by adding the data ordinates (latitude and longitude) location of a poor family. The results of testing the classification of the data shown in Figure 7 were presented in the form of recapitulation.

From the results of the data testing, the accuracy of the data was 92.5% which came from the data of 44 poor people from the total amount of 219 poor residents. Results of existing data were also compared with the results from the Weka software. Before the data were being processed, preprocessing the data was done previously. This stage was done to look at the description of the data which needs to be processed using NBC. The description of statistical data

showed that the data to be processed had an average value 1.804 and a standards deviation value of 2.869. From this value, it was indicated that the deviation of the data is very high. After preprocessing being done, then classification analysis was done using Naive Bayes Classifier (NBC) algorithm. From Weka data testing results in Figure 8, it was shown that the classification results had the accuracy of 93.18%.

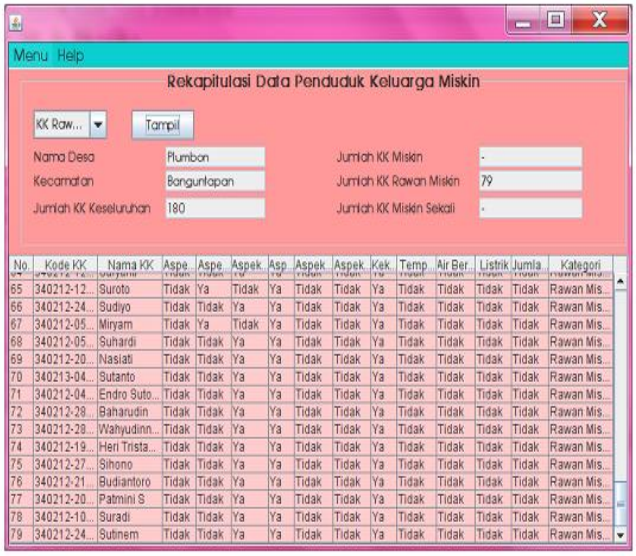


Fig. 7. Result Testing Rekapitulation

From 44 data tested using Weka, there were 41 data that could be recognized correctly, while there were 3 data that could not be identified. The results of the classification in Figure 7 were used as the input for mapping the poor families by adding the data latitude and longitude as well as home photo of the poor families.

```
inst#,     actual, predicted, error, probability distribution
    1        2:M       1:RM      +  *0.817  0.183  0
    2        2:M        2:M          0.003 *0.997  0
    3        2:M        2:M          0.027 *0.973  0
    4        2:M        2:M          0.011 *0.989  0
    5        2:M        2:M          0     *1      0
    6       1:RM       1:RM          *0.824  0.176  0
    7       1:RM       1:RM          *0.84   0.16   0
    8        2:M        2:M          0.191 *0.809  0
    9        2:M        2:M          0.21  *0.79   0
   10       1:RM       1:RM          *0.824  0.176  0
   11       1:RM       1:RM          *0.84   0.16   0
   12       1:RM       1:RM          *0.824  0.176  0
   13        2:M        2:M          0.004 *0.996  0
   14       1:RM       1:RM          *0.824  0.176  0
   15        2:M       1:RM      +  *0.974  0.026  0
   16        2:M        2:M          0.191 *0.809  0
   17        2:M        2:M          0.001 *0.999  0
   18        2:M        2:M          0.011 *0.989  0
   19        2:M        2:M          0.21  *0.79   0
   20        2:M        2:M          0.018 *0.982  0
   21        2:M        2:M          0.184 *0.816  0
   22        2:M        2:M          0.003 *0.997  0
   23        2:M        2:M          0.001 *0.999  0
   24       1:RM       1:RM          *0.84   0.16   0
   25        2:M        2:M          0.001 *0.999  0
   26       1:RM       1:RM          *0.84   0.16   0
   27        2:M        2:M          0.21  *0.79   0
   28       1:RM       1:RM          *0.84   0.16   0
   29        2:M        2:M          0.191 *0.809  0
   30        2:M        2:M          0.001 *0.999  0
   31       1:RM       1:RM          *0.824  0.176  0
   32       1:RM       1:RM          *0.84   0.16   0
   33        2:M        2:M          0.003 *0.997  0
   34       1:RM       1:RM          *0.824  0.176  0
   35        2:M        2:M          0.135 *0.865  0
   36        2:M        2:M          0.001 *0.999  0
   37       1:RM       1:RM          *0.84   0.16   0
   38        2:M        2:M          0.002 *0.998  0
   39       1:RM       1:RM          *0.84   0.16   0
   40        2:M       1:RM      +  *0.817  0.183  0
   41        2:M        2:M          0.001 *0.999  0
   42       1:RM       1:RM          *0.824  0.176  0
   43        2:M        2:M          0.005 *0.995  0
   44       1:RM       1:RM          *0.994  0.006  0

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          41               93.1818 %
Incorrectly Classified Instances         3                6.8182 %
Kappa statistic                          0.8584
Mean absolute error                      0.1021
```

Fig. 8. Weka Output

The mapping displays of the poor families were shown in Figure 9 and Figure 10. Figure 9 shows the location mapping of the poor families for all categories of poverty in a certain area. This mapping information will describe the potential of the existing poverty in a certain region. Figure 10 provides detailed information about a poor family that includes Family Identification Number, Name of the head of the family, the home location and home photos of poor families.

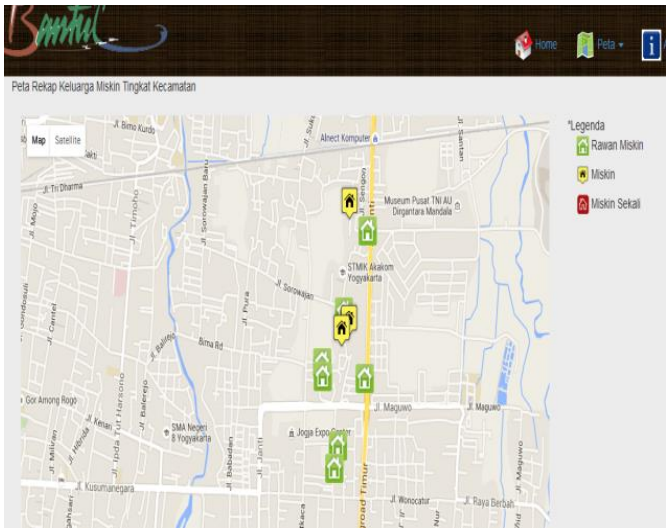


Fig. 9. Poverty Mapping

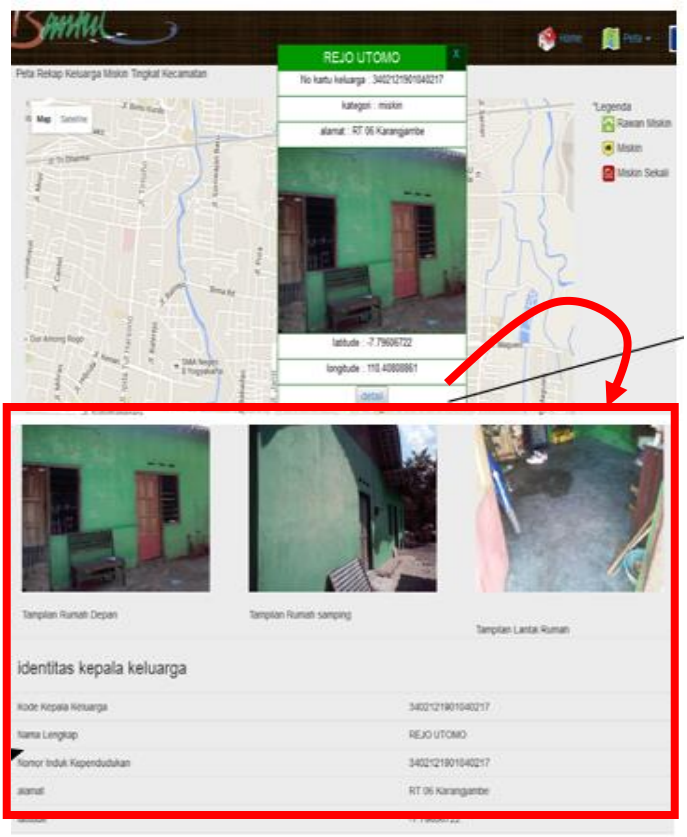


Fig. 10. Detail Information Poor Family

This detailed data will provide benefits for decision makers in providing aid or poverty reduction solutions.

## V. CONCLUSION

From the explanation that is in chapter before it can be taken conclusion on the results of the study among other:

*a) Method Naive Bayes Classifier can do classifications the determination of their position in the family poor with the accuracy 93%.*

*b) Classifications produced on data testing only two classifications that are 25 poor and 19 prone to poor, where 41 data recognizable by right and 3 data could not identified.*

*c) Mapping and detailed information from the classifications into Google Maps can inform us about the potential of poverty in areas.*

*d) The implementation of Naive Bayes Classifier algorithm built use Java used by the decision makers who is in Kabupaten Bantul.*

## ACKNOWLEDGMENT

We want to thank to SKPD BKKPPKB in Kabupaten Bantul which has given the opportunity to do research in the field of poverty.

## REFERENCES

[1] Addin, O., Sapuan, S. M., Mahdi, E., & Othman, M. "A Naive-Bayes classifier for damage detection in engineering materials", Materials and Design, 2007, pp. 2379-2386.

[2] BKKBN,2006, Kependudukan dan Pembangunan, http://www.bkkbn.go.id/news-detail.php?nid790, diakses tanggal 14 Februari 2013.

[3] Chang-Hwan Lee, Fernando G, Dejing D, 2011, Calculating Feature Weights in Naive Bayes with Kulback-Leibler Measure, 11th IEEE International Conference on Data Mining,1550-4786/11.

[4] Congle Zhang,Gui-Rong Xue,Yong Yu and Hongyuan Zha, 2009, web-Scale Classification with Naive Bayes, Poster Season, April 22, 2009, Madrid, Spain. ACM 978-160558-487-4/09/04

[5] Daniela Xhemali, Christopher J. Hinde and Roger G. Stone, Naive Bayes vs Decision Trees vs Neural Network in the Classification of Training Web Pages, IJCSI International Journal of Computer Science Issues, Vol. 4, No. 1, 2009, ISSN (Online): 1694-0784, ISSN (Print): 1694-0814

[6] Eko Prasetyo, 2012, Data Mining : Konsep dan Aplikasi menggunakan Matlab, Penerbit Andi Yogyakarta.

[7] Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, Morgan Kaufmann Publisher, Microsoft research,2007.

[8] Manurung, Martin, 2005, Measuring Poverty: The Prominent and Alternative Indicators, Mimeo, University of East Anglia.

[9] Pernia, Ernesto M dan M.G. Quibra, 1999, Poverty in Developing Countries, Handbook of Regional and Urban Economics Vol 3. Amsterdam: Elseiver

[10] Peraturan Bupati Bantul Nomor 21A Tahun 2007.

[11] Yeffriansyah Salim, 2012, Penerapan Algoritma Naive Bayes untuk Penentuan Status Turn-Over Pegawai. Media Sains, Volume 4, Nomor 2.

[12] Badan Pusat Statistik, 2013, Jumlah Penduduk Miskin, Persentase Penduduk Miskin dan Garis Kemiskinan, http://bps.go.id/linkTabelStatis/view/id/1494

[13] BPS Kabupaten Bantul, Penduduk Miskin, http://bantulkab.bps.go.id/Subjek/view/id/23#subjekViewTab1