Jurnal Mantik
Volume 4, Number 3, November 2020, pp. 1899-1905 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Classification of Citizens with Low Economic Level Using Naive Bayes Classification Method

Artika Surniandari[1], Hilda Rachmi[2], Lisda Widiastuti[3]

[1,3]Accounting Information System
[2]Information System
[1,2,3]Universitas Bina Sarana Informatika, Jl. Kramat Raya No 98, Jakarta Pusat

E-mail: artika.ats@bsi.ac.id, hilda.hlr@bsi.ac.id, lisda.ltt@bsi.ac.id

ARTICLE INFO

Article history:
Received: 12/07/2020
Revised: 22/08/2020
Accepted: 30/09/2020

Keywords:
Classification, Economic, Data Mining, Naive Bayes.

ABSTRACT

The effort to reduce poverty, it is hoped that the economic growth of a region can be spread evenly. To equalize the economic level of its citizens, local governments provide many opportunities for their citizens to get help or subsidies and even help open businesses so that the welfare of their citizens can be improved. The availability of accurate and sustainable citizens' economic status data is one of the important instruments for evaluating government policies in alleviating poverty by distributing targeted aid. Therefore, this study will conduct a classification based on data from citizens with low economic levels obtained from Pasar Babakan Sub-district, Bogor. The data used in this study is the data of underprivileged residents who are in Pasar Babakan Sub-district Bogor using data mining techniques. The training data are taken from 214 citizens that has a category of underprivileged citizens with attributes used in the classification of income per day, occupation and number of dependents. The test results using the Naive Bayes Classifier method produce 100% accuracy including the excellent category with 100% precision and 100% recall.

1. Introduction

Poverty is one of the fundamental issue at the center attention in every country. Poverty becomes the main cause of a number of social problems, politics, and economy that happened especially in developing countries including Indonesia [1]. The level of welfare for citizens is often as a measure of the success of leadership both at the central and regional levels. To be able to create a peaceful and prosperous community life is a goal that is always expected to be realized, with the support of the government and the community, of course this can be realized and not just become a dream. Development that has not been equitable and economic conditions of citizens who are not equal are of particular concern to the government and of course should be responded intelligently by residents who need an improvement in the economy or infrastructure.

Government policies in this case the central and regional levels are certainly targeted at the equal distribution of the economy of its citizens as well as at the sub-district level, in one sub-district it can have more disadvantaged citizens so that it gets the attention of the central or private government to assist in determining which citizens are entitled receiving the assistance, of course, the sub-district needs the classification of citizens who qualify as recipients of assistance and should as citizens who are in the category of poor people should rise up and use the assistance to make their lives better so that eventually they can get out of these incapable criteria for classification in this study. Headman of Sub-District of Babakan Pasar Bogor said “We hope that all residents here receive all the assistance allocated from both the APBD and APBN, and what is certain is that the recipients are really on target” [2]

Classification is one of the most important tasks of data mining. A grading system comprises of a set of training instances with class markings. The classification accuracy of a classifier is typically determined by its precision or error rate in test cases. Despite these challenges, classifiers that rely on chance might provide probability estimates or class prediction "confidence” [3].

In this research, Naive Bayes Classification (NBC) method is used for classifying a condition has been carried out widely such as the classification of books conducted by [4] where in the study used 249 samples. Naive Bayes algorithm test results obtained 70 relevant documents and 5 documents are not relevant to the value of each recall of 88, 20%, precision of 94.56%, f-measure of 90.46%, and accuracy of 97.78% so that it can be concluded that the Naive Bayes algorithm can be used in automating the classification of book titles.

Accredited “Rank 4”(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

The classification of poor families can also be used in decision making with different methods such as research conducted [5] in the category of very poor, poor, vulnerable poor and not poor. In another study regarding the determination of poor families using K-Nearest Neighbor, an accuracy of 83% was obtained [6]. While research related to using the ordinal logistic regression method produces an accuracy of 80.47% and Fuzzy K-Nearest Neighbor produces an accuracy of 87.60% [7]. The purpose of this study was to determine the level of accuracy of the data classification of underprivileged residents in the sub-district of Babakan Pasar Bogor market taken from data sets of citizens who are entitled to receive assistance.

Another example of the use of Naive Bayes for classification can be seen in the study conducted [8]. He used 200 data records from the Pekalongan Regency Social Service with 15 attributes including age, building floor area, building floor type, wall type, waste facilities big water, drinking water source, main household lighting source, daily cooking fuel, meat / chicken / milk per week, frequency of daily feeding for each household member, ability to buy new clothing for each household member in one year, ability to pay for medical treatment at the Public Health Center/ Polyclinic, income of the head of the household, the highest education of the head of the household, and assets / savings. This study aims to determine the feasibility of the recipient of the Indonesia Healthy Card (KIS) and show the use of a combination of the K-Nearest Neighbor-Naïve Bayes. The classification algorithm produces 96% accuracy.

Naïve Bayes algorithm, also known as Bayesian Classification is characterized by a supervised learning methodology comprising of a statistical procedure for classification, it also called as Bayes (1702-1761) developed by UN agency [9]. Naïve Bayes is a simple probabilistic classification that calculates a set of probabilities by adding up the frequency and combination of values from a given dataset. The algorithm uses the Bayes theorem and assumes all the independent or non-interdependent attributes given by values to class variables [10]. Review on characterization calculations found that basic Bayesian known as Naive Bayesian classifier can be contrasted in execution and choice trees and different classifiers and displays high precision [11]. A favorable position of the innocent Bayes classifier is that it requires a limited quantity of preparing information to evaluate the boundaries (means and fluctuations of the factors) vital for arrangement. Since free factors are accepted, just the fluctuations of the factors for each class should be resolved and not the whole framework [12].

2. Research Method

In this study, the data obtained from the results of interviews and observations in the community section of sub-district Babakan Pasar, involving data from underprivileged families, including data on income, employment and dependents from these data will be tested whether the data is appropriate in the classification of underprivileged families. Data Mining is an activity that involves collecting, using historical data to find regularities, patterns and relationships in large data sets [13]. One of the things that data mining can do is classification by dividing data into several groups that have been determined [14]. This study uses data sets of residents from the sub-district of Babakan Pasar, Central Bogor. To determine the status used 3 parameters, namely: income, employment, and dependents. The number of dependents affects the welfare of the family if it is not balanced with adequate income [15].

The equation from the Bayes theorem is:

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

Where:

X: Data with unknown classes

H: The data hypothesis is a specific class

P (H | X): H hypothesis probability based on condition X (posterior probability)

P (H): Hypothesis probability H (prior probability)

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

P (H): Probability X

The research phase consisted of library research, data collection, interviews and observations. The input system will then be processed using Naïve Bayes with flow and stages according to the diagram of the following data processing model:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

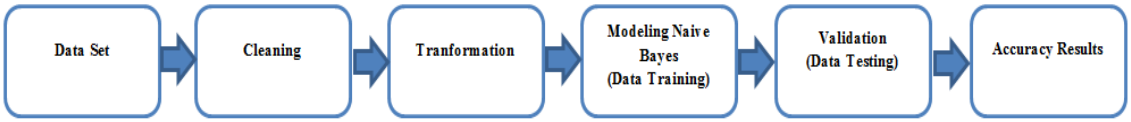

Fig 1. Data Processing Model

3. Result And Discussion

The classification process in this study is based on 4 components, they are: 1) *Class*, it is the dependent variable in the form of categories: very poor, poor, and nearly poor which represents the label on the object of research, 2) *Predictor*, it is an independent variable represented by data characteristics such as family name, income, occupation, and number of family dependents, 3) *Testing*, it is data set containing new data that will be classified using the research model we made and it will be evaluated for classification accuracy, and 4) *Training*, it is data set containing values from predictor and class components which will be used to determine suitable classes based on predictors

In this research, we use data processing steps from raw data to inform that it can be processed. Stages starting from determining the data and variables that we use include 4 predictors and 1 class. This data then undergoes a data cleansing process that is carried out by removing incomplete data, containing errors, and inconsistent data. Furthermore, it will check whether there is repeated data, then this data will be combined. All data will be selected whether relevant to the analysis or not.

A. Testing the Classification Method

We obtained a dataset of 214 data sets of citizens with the attributes Name, monthly income, work and dependents and who acted as a class is information. In this case the existing data is divided into 90% of the data as training data and the other 10% as testing data.

In this study the data processing stages were obtained using the naive bayes theory as follows :

1) Data set is raw data that has not been through the process of cleaning some 214 data
2) The cleaning process cleans up data whose variables do not meet the calculations, in this study all data meet the calculations
3) Transform raw data into categories that correspond to data mining. Transforming the data of this study, namely Income:> 20000, 20000, <20000; Work: Not working, housewife, Labor and Dependents:> = 5,> 2, <2. From the transformation results obtained: Very Poor Category: 110 data, Poor Category: 51 data, and Nearly Poor Category: 53 data.
4) Testing the Classification Method Using Training Data

B. Classification with Naïve Bayes Using Training Data

The results of the collection of datasets / training data from the institution where the research was carried out where data obtained from citizens who meet the criteria of the underprivileged determined from the relevant agencies where the category of disability can be divided into three classification, namely for income less than Rp. 600,000 per month is categorized as Very Poor, earning as much as Rp. 600,000 per month is classified as Poor, and above Rp. 600,000 is categorized as almost Poor, from the data set the classification process will be carried out on the following new data:

Attributes: Income, Work, Dependents, Class: Information
Naïve Bayes process
Probability of Nearly Poor Classes:
P (Nearly Poor) = 47/192 = 0.24
Poor Class Probability:
P (Poor) = 47/192 = 0.24
Very Poor Class Probabilities:
P (Very Poor) = 98/192 = 0.51

Table 1
Class Probability Table with income attributes

Income	Nearly Poor	Poor	Very Poor
>20000	1,00	0,00	0,00
20000	0,00	1,00	0,00
<20000	0,00	0,00	1,00

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Table 2
Class Probability Table with Job attributes

Work	Nearly Poor	Poor	Very Poor
Tidak Bekerja	0,00	0,00	0,01
Housewife	0,89	1,00	0,99
Labor	0,11	0,00	0,00

Table 3
Class Probability Table with dependent attributes

Dependents	Nearly Poor	Poor	Very Poor
>=5	0,15	0,23	0,28
>2	0,72	0,70	0,66
<2	0,13	0,06	0,06

C. Testing the Classification Method Using Training Data

Table 4.
Training Data

No	Name	Income	Work	Dependent	Condition
1	Widia Ningsih	10000	Housewife	5	Very Poor
2	Wina Winarti	20000	Housewife	3	Poor
3	Wiwi	25000	Housewife	2	Nearly Poor
4	Wiwin	25000	Housewife	4	Nearly Poor
5	Yanah Maemunah	15000	Housewife	6	Very Poor
6	Yani	15000	Housewife	5	Very Poor
7	Yani Kustinah	15000	Housewife	4	Very Poor
8	Yanih Soleh	10000	Housewife	4	Very Poor
9	Yanti Susanti	10000	Housewife	3	Very Poor
10	Yanti Yuliani	20000	Housewife	3	Poor
11	Yanti Yulianti	10000	Housewife	4	Very Poor
12	Yati	10000	Housewife	3	Very Poor
13	Yati Maryati	15000	Housewife	3	Very Poor
14	Yayah	15000	Housewife	4	Very Poor
15	Yayah Rokayah	10000	Housewife	2	Very Poor
16	Yayat Aryati	10000	Housewife	3	Very Poor
17	Yeni. Nilawati	20000	Housewife	3	Poor
18	Yeti Agustin	25000	Housewife	5	Nearly Poor
19	Yuniarti	25000	Housewife	2	Nearly Poor
20	Yusi Herawati	20000	Housewife	3	Poor
21	Suwito Winata	25000	Labor	2	Nearly Poor
22	Suryanto	25000	Labor	4	Nearly Poor

D. Calculation of accuracy using rapidminer

In this process several steps are carried out to classify the data and calculate the accuracy of the data, including entering data with the type. Xls, then making a data processing model with Naïve Bayes as shown below:

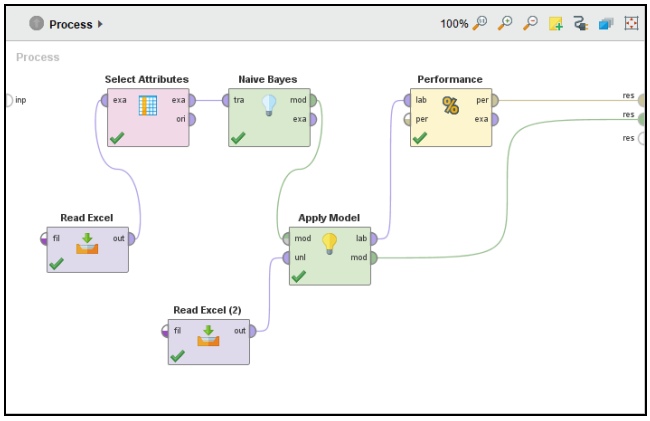

Fig 2. the data accuracy model uses rapid miner

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Then the next experiment is testing the algorithm with folds cross validation technique with data testing ranging from 2 to 10. With the measurement results in the form of the highest accuracy rate of 100%, the highest precision level of 100%, and the largest recall rate of 100% which is included in the excellent category. The results of extensive experiments and theoretical verification are demonstrated by the use of 10-fold cross validation which gets the most accurate validation. 10-fold cross validation performs retest 10 times and displays the average results of all the tests. Evaluation of the results of the performance comparison by measuring the accuracy of the accuracy generated in the 10-fold cross validation and testing the performance of the Confusion Matrix will obtain accuracy results and AUC values that can determine the classification results into very good, good, sufficient, low or wrong classification.

accuracy: 100.00%

	true Sangat Miskin	true Miskin	true Hampir Miskin	class precision
pred. Sangat Miskin	12	0	0	100.00%
pred. Miskin	0	4	0	100.00%
pred. Hampir Miskin	0	0	6	100.00%
class recall	100.00%	100.00%	100.00%	

Fig 3. the data accuracy model uses rapidminer

SimpleDistribution

Distribution model for label attribute Ket

Class Sangat Miskin (0.510)
8 distributions

Class Miskin (0.245)
8 distributions

Class Hampir Miskin (0.245)
8 distributions

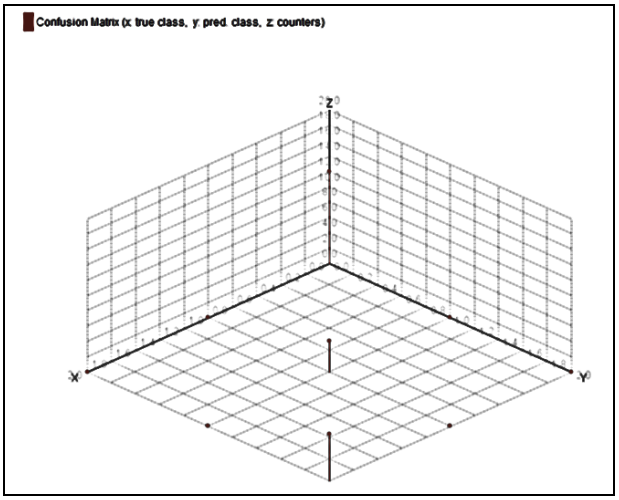

Fig 4. Confusion Matrix

Figure 4 presents the distribution model for label attributes Poor, Very Poor and Nearly Poor that resulted from applying Naïve Bayes Classifier algorithm. In this case the accuracy of this method is used to process data using manual calculations obtained an accuracy of 100%. Confusion matrix yields ideal suite for evaluating Naïve Bayes Classification performance and tell the accuracy, recall, and precision. This is a multi-nominal classification model, the target has three column labeled as Poor, Very Poor and Nearly Poor. We classified the data with True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP). The cell identified by the row and column for the positive class contains the TP where the actual and predicted class is very poor, value TP in this case is 12. Cells identified by the row for the positive class and columns for the negative class contains the FN, where the actual class is very poor and the predicted class is poor or nearly poor, value FN in this case is 0. Cells identified by rows for the negative class and the column for the positive class contain the FP, where the actual class is poor or nearly poor, and the actual class is very poor, value FP in this case is 0. Cells outside the row and column for the positive class contain the TN, where the actual class is poor or nearly poor, and the predicted class is poor or nearly poor, value TN in this case is 10. Now we can calculate the accuracy, recall and precision. We devided the number of TP (12) and TN (10) by all the number of TP (12), TN (10), FP (0) and FN (0) multiplied by 100% to get accuracy. The result of accuracy is 100%. Recall measures how good the model in detecting positive events. The formula is TP/(TP+FN) = 12/(12+0) = 1. Precision shows how accurate the very poor class of citizen economic level is with formula TP/(TP+FP) = 12/(12+0) = 1.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

4. Conclusion

Based on the results of research in sub-district of Babakan Pasar Bogor using the Naive Bayes method, training data are used to produce probabilities from several classes so that conditions can be predicted according to the classification using these methods. Through the classification process of 192 training data and 22 testing data, the accuracy of the method used to process the data is 100%. For next research, by using Naive Bayes Method, data can be collected in more wide area, like in a district. Another suggestion is that Future Selection Method may be as another way to analyze the object of research.

5. References

[1] R. Arifando, N. Hidayat, and A. A. Soebroto, "Klasifikasi Calon Penerima Bantuan Keluarga Miskin Menggunakan Metode Learning Vector Quantization (LVQ) (Studi Kasus : Daerah Kecamatan Mlandingan , Situbondo)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 6, pp. 2173–2181, 2018.

[2] D2N, "Kelurahan Babakan Pasar Distribusikan Bansos Tahap Pertama Dari Pemprov Jawa Barat," 2020. [Online]. Available: http://www.87onlinenews.com/2020/05/kelurahan-babakan-pasar-distribusikan.html.

[3] E. Chandra, V. Thedla, and S. Thota, "A Study of Structure Learning of Bayesian Network using Averaged Extended TreeAugmented Naive Classifier," vol. 7, no. 18, pp. 2875–2880, 2020.

[4] V. Rizqiyani, A. Mulwinda, and D. Mahadji, "Klasifikasi Judul Buku dengan Algoritma Nae Bayes dan Pencarian Buku pada Perpustakaan Jurusan Teknik Elektro," *J. Tek. Elektro*, vol. 9, no. 2, pp. 60–65, 2017.

[5] U. Lestari and M. Targiono, "Sistem Pendukung Keputusan Klasifikasi Keluarga Miskin Menggunakan Metode Simple Additive Weighting (Saw) Sebagai Acuan Penerima Bantuan Dana Pemerintah (Studi Kasus: Pemerintah Desa Tamanmartani, Sleman)," *J. TAM (Technology Accept. Model.*, vol. 8, no. 1, pp. 70–78, 2017.

[6] I. W. Supriana and L. G. Astuti, "Implementasi K-Nearest Neighbor Pada Penentuan Keluarga Miskin Bagi Dinas Sosial Kabupaten Tabanan," *J. Teknol. Inf. dan Komput.*, vol. 5, no. 1, pp. 120–129, 2019, doi: 10.36002/jutik.v5i1.645.

[7] D. Puspita, Suparti, and Y. Wilandari, "Klasifikasi Tingkat Keluarga Sejahtera Dengan Menggunakan Metode Regresi Logistik Ordinal Dan Fuzzy K-Nearest Neighbor (Studi Kasus Kabupaten Temanggung Tahun 2013)," *J. Gaussian*, vol. 3, no. 4, pp. 645–653, 2014.

[8] Y. F. Safri, R. Arifudin, and M. A. Muslim, "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor," *Sci. J. Informatics*, vol. 5, no. 1, p. 18, 2018, doi: 10.15294/sji.v5i1.12057.

[9] P. Patil and S. Shinde, "PERFORMANCE ANALYSIS OF DIFFERENT CLASSIFICATION ALGORITHMS : NAÏVE BAYES , DECISION TREE AND K-STAR," vol. 7, no. 19, pp. 1160–1164, 2020.

[10] T. R. Patil and S. . Sherekar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 6, no. 2, pp. 256–261, 2013, doi: 10.18201/ijisae.2019252786.

[11] D.Saranya, S.Thulasidass, and D.Gomathi, "Automatic Service Discovery using Ontology Learning Semantic Focused Crawler for Mining," vol. 4, no. 10, pp. 3805–3811, 2015.

[12] S. Vanakovarayan, D. Murali, S. Prasanna, and T. Priyaradhikadevi, "J48 , CART AND NAVIE BAYESIAN ALGORTHIM FOR PERFORMANCE ANALYSIS OF SOFTWARE," vol. 7, no. 16, pp. 2435–2440, 2020.

[13] I. M. Kamal, T. Hendro P, and R. Ilyas, "Prediksi Penjualan Buku Menggunakan Data Mining Di Pt. Niaga Swadaya," *Semnasteknomedia Online*, vol. II, no. 1, pp. 49–54, 2017.

[14] H. Naparin, "Klasifikasi Peminatan Siswa SMA Menggunakan Metode Naive Bayes," *Syst. Inf. Syst. Informatics J.*, vol. 2, no. 1, pp. 25–32, 2016, doi: 10.29080/systemic.v2i1.104.

[15] A. Purwanto and B. M. Taftazani, "Pengaruh Jumlah Tanggungan Terhadap Tingkat Kesejahteraan Ekonomi Keluarga Pekerja K3L Universitas Padjadjaran," *Focus J. Pekerj. Sos.*, vol. 1, no. 2, p. 33, 2018, doi: 10.24198/focus.v1i2.18255.