Implementation of Data Mining with Naive Bayes Algorithm for Eligibility Classification of Basic Food Aid Recipients

Yamato Shino[1], Yusuf Durachman[2], Nana Sutisna[3]
University of Miyazaki[1], Syarif Hidayatullah State Islamic University Jakarta[2],
Buddhi Dharma University[3]
Miyazaki, Gakuenkibanadainishi[1], Jl. Ir H. Juanda Cemp. Putih, Kota Tangerang Selatan[2],
Jl. Imam Bonjol No.41, Karawaci, Kota Tangerang[3]
Japan[1], Indonesia[2,3]
e-mail: yamatoshino@yahoo.com[1], yusuf_durachman@uinjkt.ac.id[2],
nana.sutisna@ubd.ac.id[3]

Author Notification
October 2022
Final Revised
October 2022
Published
October 2022

To cite this document:
Shino, Y., Durachman, Y., & Sutisna, N. (2022). Implementation of Data Mining with Naive Bayes Algorithm for Eligibility Classification of Basic Food Aid Recipients. *International Journal of Cyber and IT Service Management (IJCITSM)*, 2(2), 154-162. Retrieved from https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/114

DOI:
https://doi.org/10.34306/ijcitsm.v2i2.114

Abstract

One of the primary issues that the government of a nation concentrates on is poverty. The provision of precise and focused data on poverty is a crucial component of the Poverty Reduction Strategy. One technique for classifying data is Naïve Bayes. The aid manager will subsequently use the categorization findings to inform judgments about categorizing and determining who should get basic food assistance. Predictions for those who get basic food assistance fall into two categories: eligible and ineligible. Sample data from the hamlet of XYZ used as the basis for the forecast. In this study, a web-based application is used to construct and assess the Naïve Bayes method. The accuracy for 135 training data, 40 test data, and seven characteristics employed generates 86 percent accuracy, 85 percent recall, and 88 percent precision according to the assessment findings using the confusion matrix.

Keywords *: Naive Bayes, Confusion Matrix, Data Mining, Algorithm.*

1. Introduction

One of the primary issues that the government of a nation concentrates on is poverty. The provision of precise and focused data on poverty is a crucial component of the Poverty Reduction Strategy. The Indonesian people's nine fundamental requirements, known as *sembako*, are met through food and beverages that are consumed on a daily basis. Based on this, the government often develops fundamental food aid programs for those in need [1]. The government provides beneficiary families with non-cash food aid through the Non-Cash Food Assistance/*Sembako* Program each month via a banking system. More accurate data collection is required for families that are eligible or ineligible to receive basic food assistance since, in practice, the supply of basic food assistance falls short of expectations. Data mining techniques can be employed to establish the eligibility of beneficiaries of basic food. Data mining is a practical technique for extracting meaningful information from a variety of data utilizing statistical, mathematical, and pattern recognition expertise [2]. Big data is extracted from and identified through data mining to locate information that is beneficial to the business. To get relevant information, data mining may be used to categorize, forecast, and estimate. Planning is aided by data mining, which offers accurate information for making forecasts based on historical patterns and present circumstances [4]. Because automated decision-making can result in lower costs, data mining enables businesses to allocate financial resources more effectively.

A function called prediction may extract a certain pattern from data. The numerous data variables may be used to find these patterns. Once a pattern is identified, it is possible to forecast other variables whose value or type is unknown using the pattern that was discovered. The Naïve Bayes algorithm is one of the prediction techniques that may be applied to data mining. One technique for classifying data is Naïve Bayes. A statistical classification procedure called Bayesian classification is used to forecast the likelihood that a class will include members [5]. The Naïve Bayes algorithm has been successfully used in a number of previous studies to classify data, including those that determined credit distribution, where water source development would occur, how well students performed in school, how many goods would be produced, how to categorize the poor, and who would be eligible for help from the Hopeful Family Program. On the assumption that the decision value is true and based on object knowledge, the Naïve Bayes class of choices uses mathematical probability computations [6]. The outcomes of the categorization exercise will subsequently assist aid managers in making choices about how to classify and identify those who will get basic food assistance.

This study uses the Naive Bayes method to forecast the categorization of people receiving food help [7]. Finding information or patterns of distinctive similarity in a certain group or class is one of the functions of the Naïve Bayes algorithm. The basic food aid clients' predictions are split into two categories, namely possible and not feasible. Sample data from the hamlet of XYZ used as the basis for the forecast. In this study, a web-based application is used to build and assess the Naïve Bayes method.

2. Method

2.2 Research Stages

In research, planning and structured steps are needed so that research can run well. The stages of the research carried out can be seen in Figure 1.

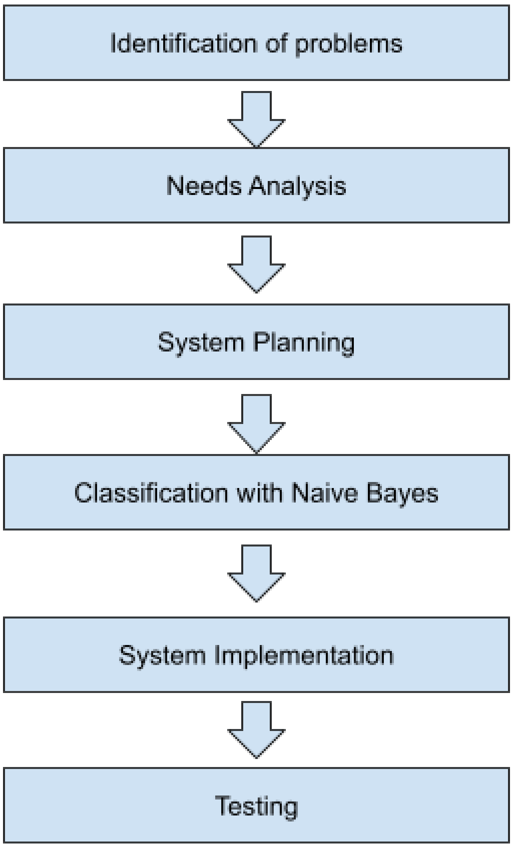

Figure 1. Research Stages

a. Identification of Problems

Generally speaking, problem identification is a step in the research process that may be seen as an effort to characterize already-existing problems and turn them into measurable, testable issues [8]. Simply simply, problem identification involves figuring out what the main focus of a study will be. Beginning with data collection through interviews or user needs, it is helpful to identify information and issues that need to be resolved. How to create a system that can categorize household heads who are qualified to get basic food assistance is the key challenge in this research.

b. Needs Analysis

The functional requirements that the user requires are included in the needs analysis. To determine what functions the system can do or what features it has, as well as who can utilize the system that was constructed, functional requirements are necessary [9]. The system that has to be constructed must comply with the following functional requirements:

One person, either admin or user, will utilize the system. The procedure that an administrator or user can carry out is:

1. Administrators can control training data.
2. The admin can input test data.
3. Test results are visible to admin

c. System Planning

Identification and description of a system abstraction based on its connections constitute system design [10]. Use case diagrams, a type of Unified Modeling Language (UML) diagram, were employed by the researchers throughout the design phase. Use case diagrams show a relationship between one or more actors and the future information system.

d. Classification with Naive Bayes

One technique that may be used to categorize data is Naïve Bayes. Bayesian classification is a statistical classification that may be used to forecast the likelihood that a class will contain a given individual [11]. On the assumption that the decision value is true and based on object knowledge, the Naïve Bayes class of choices uses mathematical probability computations.

The Naïve Bayes approach proceeds as follows:

1. Review practice data
2. If the data is numerical, determine the mean and standard deviation of each parameter, which is numerical data.
   a) Calculate the number and probability.
   b) Determine the probability value by dividing the total number of relevant data in a category by the total number of data in that category.
3. Calculate the mean, standard deviation, and probability of the values in the table.

Figure 2 displays the Naive Bayes scheme.

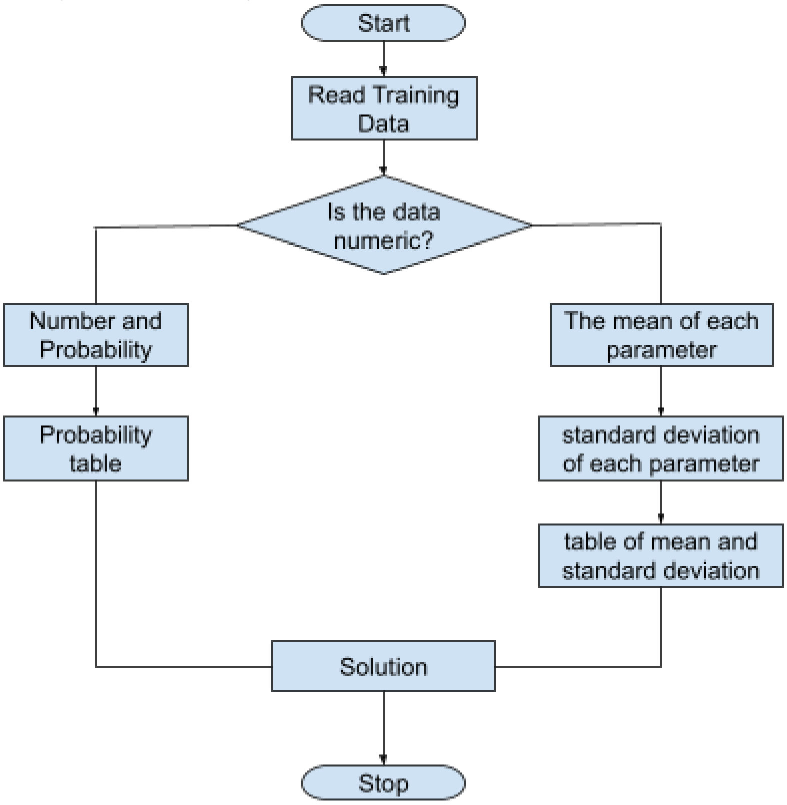
Start
Read Training Data
Is the data numeric?
Number and Probability
Probability table
The mean of each parameter
standard deviation of each parameter
table of mean and standard deviation
Solution
Stop

Figure 2. Naive Bayes Scheme

e. System Implementation

Coding is done at the implementation stage using the preceding design and analytical work as a foundation. Coding is the process of altering an existing design and putting it into action as a computer-understandable programming language. The PHP programming language, the Visual Studio Code compiler, and the MySQL DBMS are used to create the code for this study.

f. Testing

This study's testing procedure makes use of a confusion matrix to determine the precision, recall, and accuracy scores [12]. To determine precision, recall, and accuracy, the confusion matrix includes true positive, false positive, true negative, and false negative. The degree of accuracy between the user's information request and the system's response is known as precision. Recall measures how well a system is able to retrieve information. The degree of agreement between the anticipated value and the actual value is measured by accuracy. You may use the following equation to calculate precision, recall, and accuracy:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

A true positive (TP) is a result of the quantity of positive data that is projected to be accurate. The number of negative data that are projected to be true is used to calculate the true negative, or TN. A false positive (FP) is a result of a large quantity of negative data that are projected to be positive data. While FN is a false negative (false negative) that is projected to be true based on the number of positive data.

3. Results

3.1 Data Analysis

Data analysis is being done right now by:

1. Reduce noise (inconsistent data or irrelevant data).

Remove irrelevant or indirect data that isn't directly connected to the data mining process and purpose. Redundant data must be eliminated, inconsistent data must be checked for, and data errors like printing mistakes must be fixed.

2. Data grouping.

Several factors that are used to map or categorize the receivers of Basic Food Aid are converted into variables by the Naive Bayes Classification approach, including:

a) Name
Is the identity variable of the name of the head of the family.

b) Status of the Family Hope Program (PKH)
It is a PKH or non PKH family status variable.

c) Number of Dependents
Is a variable that contains the number of dependents borne by the head of the family.

d) Head of Household
Is a gender status variable of the head of the household, male or female.

e) House Condition
Is a variable condition of the house occupied, permanent stone, woven stone or boards.

f) Total Income
Is a variable amount of income of the head of the family.

g) Homeowner Status
It is a variable of home ownership status which is grouped into two categories, namely self-owned or rented.

Sample training data can be seen in table 1.

Training ID	Name	PKH	The number of dependents	Head of household	amount of income	eligibility status
1	Beni	Non	1	Male	IDR. 100.000	worthy
2	Afra	Non	4	Male	IDR. 300.000	worthy
3	Angga	Non	3	Male	IDR. 1.600.000	not feasible
4	Nining	1	1	Female	IDR. 100.000	worthy

Table 1. Sample Data Training

3.2 Design

Identification of issues is the first step in determining needs in order to create a categorization system for the viability of getting basic food assistance. These specifications are transformed into working systems [13]. Afterward, the system is created based on the needs analysis. In this study, a use case diagram was used to construct the system. The system under development's use case diagram is shown in Figure 3.

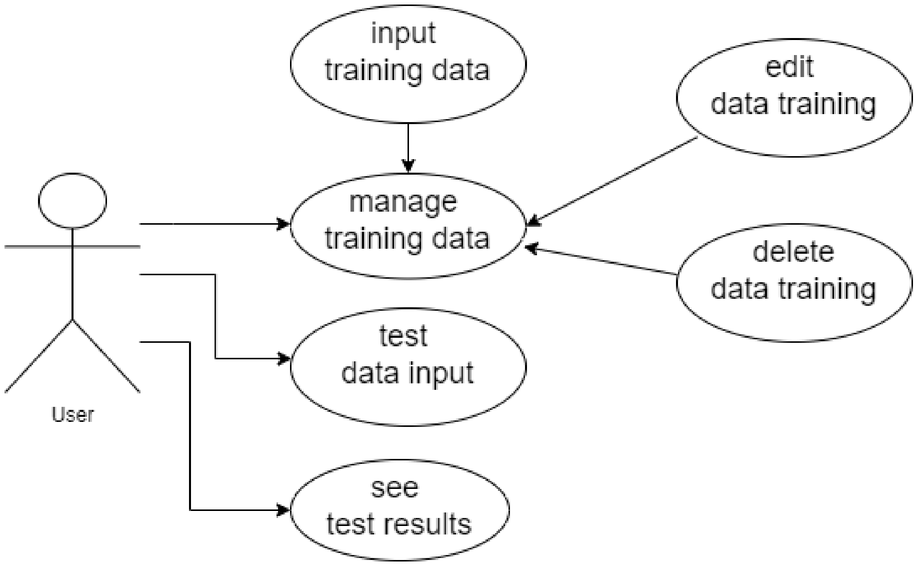

Figure 3. Use Case Diagram of the Classification System for the Receipt of Basic Food Aid

In Figure 3 it can be seen that the system will be used by one user or user. Users can manage training data, input test data and view test results.

3.3 Testing

In order to assess the degree of accuracy of the model created, the system must also be examined. Using a confusion matrix, the data correctness is determined [14]. The confusion matrix is essentially the accuracy data that arises from comparing the system's generated classification results with the expected results. Table 2 displays the outcomes of the system assessment.

Evaluation	Percentage
Accuracy	86%
Recall	85%
Precision	88%

Table 2. Results of Evaluation

To summarize the outcomes of the assessment and classification procedure utilizing 135 records for training and 40 records for testing [15], based on the information shown by the confusion matrix. 86 percent of the findings from the test were accurate, which indicates that the classification's accuracy value is extremely good. Calculating the accuracy is done by dividing the total amount of data by all the right forecasted values. While the recall (Sensitivity) results from the test likewise indicated a value of 85% [16]. By dividing the total number of positive classes by the number of valid positive predictions, recall or sensitivity is determined. The accuracy (precision) of the test's correctly identified findings is quite good. The total number of accurate positive prediction scores serves as the basis for calculating precision.The entire number of accurate positive predictive scores divided by the total number of accurate class predictions is used to assess precision [17].

The training data, testing data, and class or classification label all have an impact on or depend on the precision number in this study, which is 88 percent. The more training data, testing data, and accurate classes there are, the higher the degree of accuracy will be[18].

4. Conclusions

In this study, predictions are given regarding the Naive Bayes algorithm's categorization of the determination of the beneficiaries of needs. The Naive Bayes algorithm is one technique for categorizing data. Bayesian classification is a statistical classification that may be used to forecast the likelihood that a class will contain a given individual. Finding information or patterns of distinctive similarity in a certain group or class is one of the functions of the Naive Bayes algorithm. For the acceptance rate of basic food aid, there are two categories of predictions: realistic and unrealistic. Data from a sample of XYZ village residents were used to make the prediction. The accuracy for 135 training data with 40 testing data and seven characteristics resulted in an accuracy of 86 percent, recall of 85 percent, and precision of 88 percent according to the evaluation using the confusion matrix. A number of variables, including the quantity of training data, testing data, and characteristics employed, might affect accuracy. To create the most accurate model possible for further study, many combinations of training data, testing data, and characteristics can be employed.

References

[1] E. Dolan and R. Widayanti, “Implementation Of Authentication Systems On Hotspot Network Users To Improve Computer Network Security,” *International Journal of Cyber and IT Service Management*, vol. 2, no. 1, pp. 88–94, Mar. 2022, doi: 10.34306/IJCITSM.V2I1.93.

[2] M. R. Anwar, R. Panjaitan, and R. Supriati, “Implementation Of Database Auditing By Synchronization DBMS,” *International Journal of Cyber and IT Service Management*, vol. 1, no. 2 SE-Articles, pp. 197–205, Oct. 2021, [Online]. Available: https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/53

[3] A. Williams and C. S. Bangun, “Artificial Intelligence System Framework in Improving The Competence of Indonesian Human Resources,” *International Journal of Cyber and IT Service Management*, vol. 2, no. 1, pp. 82–87, Mar. 2022, doi: 10.34306/IJCITSM.V2I1.91.

[4] M. R. Anwar and S. Purnama, “Boarding House Search Information System Database Design”, IJCITSM, vol. 2, no. 1, pp. 70–81, Mar. 2022.

[5] M. R. Anwar and S. Purnama, “Boarding House Search Information System Database Design,” *International Journal of Cyber and IT Service Management*, vol. 2, no. 1, pp. 70–81, Mar. 2022, doi: 10.34306/IJCITSM.V2I1.89.

[6] A. Dudhat and T. Mariyanti, “Indoor Wireless Network Coverage Area Optimization,” *International Journal of Cyber and IT Service Management*, vol. 2, no. 1, pp. 55–69, Mar. 2022, doi: 10.34306/IJCITSM.V2I1.86.

[7] A. Dudhat and T. Mariyanti, “Indoor Wireless Network Coverage Area Optimization”, IJCITSM, vol. 2, no. 1, pp. 55–69, Mar. 2022.

[8] W. Setyowati, R. Widayanti, and D. Supriyanti, “Implementation Of E-Business Information System In Indonesia : Prospects And Challenges,” *International Journal of Cyber and IT Service Management*, vol. 1, no. 2 SE-Articles, pp. 180–188, Oct. 2021, [Online]. Available: https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/49

[9] D. Immaniar, N. Azizah, D. Supriyanti, N. Septiani, and M. Hardini, “PoTS: Proof of Tunnel Signature for Certificate Based on Blockchain Technology,” *International Journal of Cyber and IT Service Management*, vol. 1, no. 1 SE-Articles, pp. 101–114, May 2021, [Online]. Available: https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/28

[10] N. . Azizah, V. Hartajaya, and S. Riady, “Comparison Of Replication Strategies On Distributed Database Systems”, IJCITSM, vol. 2, no. 1, pp. 20–29, Jan. 2022.

[11] A. Williams and C. S. Bangun, “Artificial Intelligence System Framework in Improving The Competence of Indonesian Human Resources”, IJCITSM, vol. 2, no. 1, pp. 82–87, Mar. 2022.

[12] T. Ayuninggati, N. Lutfiani, and S. Millah, “CRM-Based E-Business Design (Customer Relationship Management) Case Study : Shoe Washing Service Company S-Neat-Kers,” *International Journal of Cyber and IT Service Management*, vol. 1, no. 2 SE-Articles, pp. 216–225, Oct. 2021, [Online]. Available: https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/58

[13] D. Apriani, M. Aan, and W. E. Saputra, “Data Visualization Using Google Data Studio”, IJCITSM, vol. 2, no. 1, pp. 11–19, Jan. 2022.

[14] I. Y. Ruhiawati, A. P. Candra, and S. N. Sari, “Design and Build a Multimedia System for Indonesian Religious Activities Based on Android,” *International Journal of Cyber and IT Service Management*, vol. 1, no. 2 SE-Articles, pp. 233–239, Oct. 2021, [Online]. Available: https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/64

[15] P. Edastama, A. Dudhat, and G. Maulani, “Use of Data Warehouse and Data Mining for Academic Data : A Case Study at a National University,” *International Journal of Cyber and IT Service Management*, vol. 1, no. 2 SE-Articles, pp. 206–215, Oct. 2021, [Online]. Available: https://iiast-journal.org/ijcitsm/index.php/IJCITSM/article/view/55

[16] B. Rawat and S. Purnama, “MySQL Database Management System (DBMS) On FTP Site LAPAN Bandung,” *International Journal of Cyber and IT Service Management*, vol. 1, no. 2, pp. 173–179, 2021.

[17] D. . Rustiana, J. D. . Pratama, T. . Mudabbir, M. A. . Fahmi, and . G. A. . Rofei, “Adoption Computerized Certificate Transparency And Confidentiality”, IJCITSM, vol. 2, no. 1, pp. 1–10, Jan. 2022.

[18] E. Dolan and R. Widayanti, “Implementation Of Authentication Systems On Hotspot Network Users To Improve Computer Network Security”, IJCITSM, vol. 2, no. 1, pp. 88–94, Mar. 2022.