Tadulako Science and Technology Journal
Vol. 3, No. 1, September 2022, pp. 35-41, E-ISSN : 2776-4893

Comparative Analysis of C4.5 And Naïve Bayes Algorithms for Classification of Food Vulnerable Areas

Hajra Rasmita Ngemba [a,1,*], Syaiful Hendra [b,2], Kadek Agus Dwijaya[b,3], Herdianto Ladania [b4], Muhammad Aristo Indrajaya [c5]

[an] Information Technology Department, Faculty of Engineering Tadulako University, Palu Indonesia
[b,] Information System Department, Faculty of Engineering Tadulako University, Palu Indonesia
[c,] Electrical Department, Faculty of Engineering Tadulako University, Palu Indonesia

[1]hajra.rasmita@gmail.com*; syaiful.hendra.garuda@gmail.com[2]; [3] kadekagusdwiwijaya@gmail.com; [4] aristo90c@gmail.com

ARTICLE INFORMATION

History of the article
Received : September 03, 2022
Revised : September 03, 2022
Accepted : September 18, 2022

Keywords
Data mining,
Classification,
Food insecurity,
C4.5 Algorithm,
Naïve Bayes.

ABSTRACT

Introduction: Data mining is the process of finding important information or patterns in large databases and is an activity to find useful information or knowledge automatically from large amounts of data. In data mining, large data are processed using certain techniques to obtain new information about the data. One of the techniques commonly used in data mining is classification. Classification is the process of learning a function or model against a set of training data so that the model can be used to predict the classification of the test data. Some of the methods include the C4.5 Algorithm and nave Bayes. Food insecurity is a condition where there is not enough food available for each individual or individual to be able to live a sustainable quality of life. This study aims to compare the C4.5 and Nave Bayes algorithms in terms of accuracy for classifying food insecure areas. The data used in this study is food insecurity data in Central Sulawesi province, with a total of 517 data consisting of data from 2018 to 2020. The results in this study show that the classification algorithm C4.5 has a better accuracy rate of 84% compared to Nave Bayes at 68%.

1. Introduction

Data mining is the process of finding important information or patterns in large databases and is an activity to find useful information or knowledge automatically from large amounts of data. Data mining or often also called *knowledge discovery in database* (KDD), is an activity that includes collecting and using historical data to find patterns of regularity and patterns of relationships in *sets* of large data. In its application, data mining can be used to improve decision-making in the future. In data mining, large data are processed using certain techniques to obtain new information about the data. One of the techniques commonly used in data mining is classification.

Classification is the process of learning a function or model against a set of training data so that the model can be used to predict the classification of the test data. Some of the methods include *a Decision Tree* and *a Bayesian classifier*. A *Decision Tree* is a learning method using training data that has been grouped based on certain classes in the decision tree. Bayesian *classifiers* use statistical methods and are based on Bayes theory [1].

The C4.5 Algorithm is a development of the ID3 Algorithm that can handle the problem of unavailable information and can handle continuous data and pruning by forming a decision tree. The

use of the C.45 algorithm to classify data that has quantitative (numeric) and categorical attributes. The result of this classification process is in the form of rules that are used to make a decision [2].

The Naïve Bayes algorithm is a simple probability-based classification method based on the application of Bayesian theory or rules with the assumption that object attributes are independent. In addition, Naive Bayes can also analyze the variables that most influence it in the form of opportunities [3].

Food insecurity is a condition where there is not enough food available for each individual or individual to be able to live a sustainable quality of life. Some indicators that affect food insecurity are the number of poor people, morbidity rates, access to electricity, clean water, illiterate women, under-five heights (*stunting*), adequate road access, distance from health facilities, and the ratio of normative consumption to clean availability Cereals (NCPR). Food insecurity can be classified into six priorities, namely very, food vulnerability, food vulnerability, moderately food vulnerability, moderately food security, food security, and very food security [4].

Based on the explanation above, in this study, a comparative analysis of the C4.5 and Naive Bayes algorithms will be carried out for the classification of food insecure areas. The choice of using the two algorithms is more because they are very popular and widely used in practice. This research was conducted in order to help provide information to decision-makers in policy making and to find out which Algorithm has the highest accuracy value in classifying food insecure areas.

2. Research Methods Research

2.1. Stages

There are several stages that need to be carried out in research so that research can run well. The stages in the research are as follows:

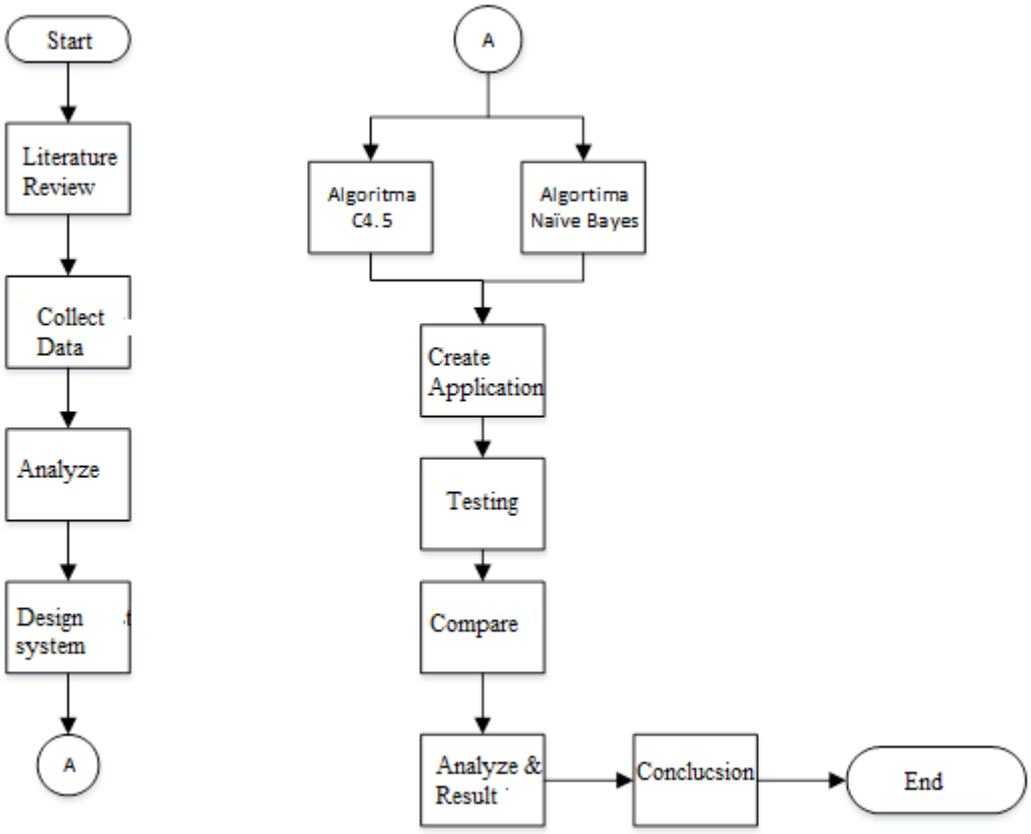

Fig 1 Stages of Research

2.2. Data Collection

Appropriate data collection techniques are by considering used based on the type of data and its source. Data that is objective and relevant to the subject matter of the research is an indicator of the success of a research. The data collection of this research was carried out in the following way:

1. Observation
   Observations were made in this study by observing the flow of data in the Central Sulawesi Regional Food Security Agency to find out existing problems and solutions to solve them and to find out what data will be collected. Needed in this research.
2. Interview
   Interviews were conducted in this study, namely conducting interviews with unstructured methods, which do not use an interview guide that contains specific questions but only contains the important points of the problem to be studied.
3. Literature review.
   In addition to collecting data through observation and interviews, the information collected is also through journals, books, and information on the internet as a reference in supporting the theories in this study.

2.3. Classification Model

The classification models used in this study are as follows:

1. Algorithm C4.5
   The Algorithm is a method for making a decision tree based on the training data that has been provided. The C4.5 Algorithm is the development of ID3. At the same time, the decision tree can be interpreted as a very strong way to predict or clarify. Decision trees can divide large data sets into smaller record sets by applying a set of decision rules [5].
2. Nave Bayes
   Nave Bayes is a classification method in the category of supervised learning, where at the nave Bayes stage, an initial dataset is needed to conduct training to produce a decision. The training determines the probability value as a weighting for each parameter. This nave Bayes method is a classification method that is quite easy and accurate to be applied to a classification problem in data mining [6-10].

2.4. Testing

System testing is divided into several parts consisting of application testing using the Black Box Testing method. Black Box Testing is testing an application which is done by observing the execution results on the application and checking the functionality of the application. At the same time, the algorithm testing used a table to measure the performance of a classification model [11]. *The confusion Matrix* is a table consisting of the number of rows of test data that are predicted to be true and not true by the classification model.

2.5. Comparison

At this stage, the value of *precision, recall, and accuracy* for each Algorithm. After that, the results of the C4.5 and Naïve Bayes algorithms are calculated so that conclusions can be drawn regarding the best Algorithm for classifying food insecure areas [12-16].

3. Results and Discussion

3.1. Classification The

Data used in this study are food insecurity data in Central Sulawesi, as many as 517 data consisting of 175 sub-district data from 2018 - 2020. Food insecurity data in this study can be seen in table 3.1 below:

Table 3.1 Food insecurity

Data	Attribute									Status
	NPCR	A1	A2	A3	P1	P2	P3	P4	P5	
1	1.08	15.92	15.92	15.01	6.4	50.46	3.04	17.4	29.02	Fairly Food Vulnerable
2	0.82	15.92	15,92	28.2	6.4	60.64	4.31	20	26.84	Moderately Food Vulnerable
3	1.5	15.92	15.92	9.39	6	36.6	5.64	30.8	13.76	Moderately Food Vulnerable
4	0.97	15.92	15.92	26.9	6.4	55.62	5.11	40.5	38.06	Fairly Food Vulnerable
5	0.25	15.92	15.92	21.87	6.4	54.5	3.25	36.4	23.47	Enough Food Resistant
…	…	…	…	…	…	…	…	…	50.94	0.220
517	Training	30.68	7,176	6	Resistant	35.99	30.50	43.5	14.98	Enough Food

Data used in both algorithms is 80% of the total data, which is 415 data. Food insecurity data consists of labels/*classes* and attributes, which can be seen in table 3.2 below:

Table 3.2 Class and Attributes

Attribute	Label/*class*
1. NPCR 2. A1 3. A2 4. A3 5. P1 6. P2 7. P3 8. P4 9. P5	1. Very Vulnerable Food 2. Vulnerable Food Vulnerable 3. Moderately Food Vulnerable 4. Enough Food Resistant Food 5. Resistant 6. Very Food Resilience

The results of the classification of the C4.5 and Nave Bayes algorithms can be seen in table 3.3 below:

Table 3.3 Classification results

	Accurate	Not Accurate	Total
C4.5	434 Data	83 Data	517
Nave Bayes	data 354 Data	163	

3.2. Testing

Results of testing *the confusion matrix* algorithm C4. 5 and Naïve Bayes are as follows:

1. Algorithm C4.5

$$\text{Accuracy} = \frac{5+38+57+127+184+23+0}{517} = \frac{434}{517} = 0.84 \; or \; (84\%)$$

$$\textit{Precision} = \frac{0.71+0.84+0.89+0.82+0.93+0.76}{6} = \frac{4.96}{6} = 0.828 \; or \; (83\%)$$

$$\textit{Recall} = \frac{0.71+0.86+0.81+0.91+0{,}88+0.76}{6} = \frac{4.95}{6} = 0.826 \; or \; (83\%)$$

$$\textit{Error Rate} = 1 - 0.84 = 0.16 \; or \; (16\%)$$

2. Naïve Bayes

$$\text{Accuracy} = \frac{6+31+47+96+158+16}{517} = \frac{354}{517} = 0.68 \; or \; (68\%)$$

$$\textit{Precision} = \frac{0.66+0.65+0.69+0.59+0.79+0{,}5}{6} = \frac{3.9}{6} = 0.65 \; or \; (65\%)$$

$$\textit{Recall} = \frac{0.6+0.73+0.60+0.67+0.72+0.59}{6} = \frac{3.93}{6} = 0.66 \; or \; (66\,\%)$$

$$\textit{Error Rate} = 1 - 0.68 = 0.32 \; or \; (32\%)$$

3.3. Comparison

Comparison is made by comparing the values of accuracy, *precision*, *recall*, and *error rate* in each Algorithm. After that, the results of the C4.5 and Naïve Bayes algorithms are calculated so that conclusions can be drawn regarding the best Algorithm for classifying food insecure areas.

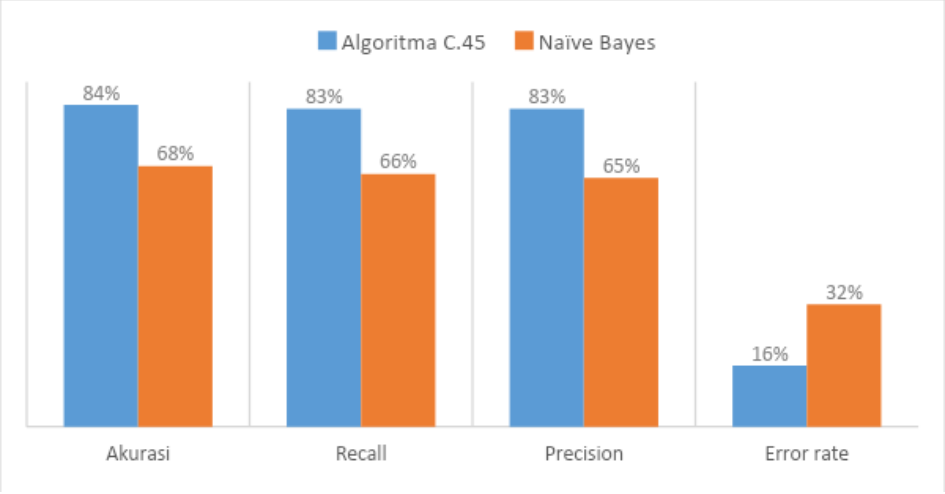

Fig 2 Algorithm comparison graph

Based on the *confusion matrix*, it can be seen that the C4.5 Algorithm has a higher accuracy value than nave Bayes in the case of food insecure area classification. The difference in accuracy values between the C4.5 and nave Bayes algorithms is influenced by several factors, such as many labels/*classes*, the amount of data used for training data and test data, as well as changing patterns or characteristics of the data. In the C4.5 Algorithm, there is one additional label/*class*, namely an *error*

because in the C4.5 algorithm classification process there is data that has not been defined *rule* or decision tree. This affects the *recall* and *precision* of the C4.5 Algorithm. The accuracy of the C4.5 and nave Bayes algorithms decreased in the 2020 data because the data patterns or characteristics were different from 2018 and 2019. Of the total data that has been tested on the food insecurity classification system with a total of 517 data, the C.45 algorithm gets an accuracy value of 84%, while nave Bayes by 68%. The C4.5 Algorithm gets the *recall*, *precision,* and *error rate* of 83%, 83%, and 16%, while Nave Bayes gets the *recall*, *precision,* and *error rate* of 66%, 65%, and 32%.

4. Conclusion

Based on the results of research testing and analysis on the comparison of the C4.5 and nave Bayes algorithms in the classification system of food insecure areas, it can be concluded that:

1. The application of the C4.5 and naive Bayes algorithms can be well applied to the classification system of food insecure areas using a web-based system with discussed PHP programming.
2. Comparison the performance of the C4.5 and nave Bayes algorithms can be compared using the *confusion matrix* by testing accuracy, *recall*, *precision,* and *error rate.*
3. The C4.5 Algorithm has a higher accuracy value of 84% when compared to the Nave Bayes accuracy of 68%. Followed by *recall* 71%, *precision* 66%, and *error rate* for the C4.5 and nave Bayes algorithms 66%, 73%, and 32%.
4. The difference in *recall*, *precision,* and *error rate* in the C4.5 and nave Bayes algorithms is influenced by the amount of training data, the number of labels/*classes*, and the pattern or characteristics of the data. The more training data used, the more the *recall*, *precision,* and *error rate will increase.*
5. The results of the classification using the C4.5 Algorithm and naive Bayes on the classification system for food-insecure areas are more effective using the C4.5 Algorithm. However, these results cannot be used as a benchmark under the C4.5 Algorithm in other cases, and it is better because to determine the best Algorithm, you must pay attention to label/*class*, attributes, and the amount of data in that case.

REFERENCES

[1] M. zhari, Z. Situmorang, and R. Rosnelly, "Comparison of Accuracy, Recall, and Classification Precision on the C4.5 Algorithm, Random Forest, SVM, and Naive Bayes," *J. Media Inform. Budidharma*, vol. 5, no. April, pp. 640–651, 2021, doi:10.30865/mib.v5i2.2937.

[2] SJS Tyas, M. Febianah, F. Solikhah, AL Kamil, and WA Arifin, "Comparative Analysis of Naive Bayes and C.45 Algorithms in Classification of Data Mining to Predict Graduation," *J. Teknol. inf. and Commune.*, vol. 8, no. 1, pp. 86–99, 2021.

[3] FK Pratama, DW Widodo, and N. Shofia, "Implementation of the Naïve Bayes Method in Classifying Recipients of the Family Hope Program (PKH) in Minggiran Kediri Village," 2021.

[4] Food Security Council, *Resilience Map and Indonesian Food Vulnerability 2020*. 2020.

[5] KF Irnanda, D. Hartamas, and AP Windarto, "Analysis of the C4.5 Classification of the Factors Causing the Decline of Student Achievement During the Pandemic," *J. Media Inform. Budidharma*, vol. 5, no. 1, pp. 327–331, 2021, doi:10.30865/mib.v5i1.2763.

[6] I. Rukmana, A. Rasheda, F. Fathulhuda, M. Ri. Cahyadi, and Firtriyani, "Comparative Analysis of the Performance of the Naïve Bayes Algorithm, Decision Tree-J48, and Lazy-IBK," vol. 5, pp. 1038–1044, 2021, doi:10.30865/mib.v5i3.3055.

[7] AD Ashari, "Application of Radial Basis Function (Rbf) for Classification of Food Insecure Areas (Case Study: Food Security Office of Riau Province)," 2019, DOI:

10.31227/osf.io/n4f68.

[8] KF Irnanda, D. Hartamas, and AP Windarto, "Analysis of the C4.5 Classification of the Factors Causing the Decline of Student Achievement During the Pandemic," *J. Media Inform. Budidharma*, vol. 5, no. 1, pp. 327–331, 2021, doi:10.30865/mib.v5i1.2763.

[9] K. Suhada, A. Elanda, and A. Aziz, "Classification of the Predicate Graduation Level of Informatics Engineering Students Using the C4.5 Algorithm (Case Study: STMIK Rosma Karawang)," *Digamaya*, vol. 01, no. 02, pp. 9–11, 2021.

[10] FF Harryanto and S. Hanson, "Application of the C4.5 Algorithm to Predict New Employee Recruitment at PT WISE," *Tek. Information. Dan Sis. inf.*, vol. 3, no. 2, pp. 95–103, 2017, [Online]. Available: http://jurnal.mdp.ac.id/index.php/jatisi/article/view/71

[11] I. Rukmana, A. Rasheda, F. Fathulhuda, M. Ri. Cahyadi, and Firtriyani, "Comparative Analysis of the Performance of the Naïve Bayes Algorithm, Decision Tree-J48, and Lazy-IBK," vol. 5, pp. 1038–1044, 2021, doi:10.30865/mib.v5i3.3055.

[12] MFA Saputra, T. Widiyaningtyas, and AP Wibawa, "Illiteracy classification using K means-nave Bayes algorithm," *Int. J. Informatics Vis.*, vol. 2, no. 3, pp. 153–158, 2018, DOI: 10.30630/joiv.2.3.129.

[13] C. Anam and HB Santoso, "Comparison of C4 Algorithm Performance. 5 and Naive Bayes for Classification of Scholarship Recipients," vol. 8, no. 1, pp. 13–19, 2018.

[14] I. Octavio, "Analysis of the Effectiveness and Contribution of Regional Taxes as a Source of Original Revenue for Batu City (Study on the Batu City Regional Revenue Service 2009-2013)," *J. Adm. Business S1 Univ. Brawijaya*, vol. 15, no. 1, p. 84581, 2014.

[15] Y. Irwan, "Application of the C4 Decision Tree Algorithm. 5 To Predict the Eligibility of Prospective Blood Donors by Classification Data Mining (Application of the C4. 5 Decision Tree Algorithm to Predict the Eligibility of Prospective Blood Donors by Classification Data Mining," *J. Teknol. Inf. and Multimed.*, vol. 2, no. 4, pp. 181–189, 2021.

[16] E. Nuria *et al.*, "Application of Nave Bayes for Classification of Risk Levels, Implementation of Nave Bayes for Classification Dental Diagnosis," vol. 4, no. 2, pp. 127– 132, 2021, DOI: 10.33387/jiko.