JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i3 (217-223)

ISSN : 1979-9292
E-ISSN : 2460-5611

COMPARATIVE OF ID3 AND NAIVE BAYES IN PREDICTID INDICATORS OF HOUSE WORTHINESS

Ade Clinton Sitepu[1)], Wanayumini[2)] , Zakarias Situmorang[3)]
[1,2]Universitas Potensi Utama, Jl. KL YosSudarso Km 6.5 No 3A, Medan, 59391, Indonesia
email: adeclintonsitepu@gmail.com, wanayumini@gmail.com
[3]Universitas Katolik Santo Thomas,Jl. Setia Budi Kp. Tengah, Medan, 20135, Indonesia
email: zakarias65@yahoo.com

Abstract

Decision making is method of solving problems using certain way / techniques so that can be accepted. After making some calculations and considerations through several stages, the decision have taken that decision maker goes through. This stage will be selected until the best decision has made. Decision-making aims to solve problems that solve problems so that decisions with final goals can be implemented properly and effectively. This study uses a simulation of decision making from seven attributes to the proportion of the feasibility of a house based on data from Central Statistics Agency (BPS). There are several techniques for presenting decision making including: ID3 (decision tree) algorithm concept and Naïve Bayes algorithm. Both classification are learning-supervised data grouping. ID3 algorithm depicts the relationship in the form of a tree diagram whereas Naïve Bayes makes use of probability calculations and statistics. As a result, in data training, decision trees are able to model decision making more accurately. The prediction results using the decision tree model = 90.90%, while Naïve Bayes = 72.73%. Meanwhile, the speed of the Naive Bayes algorithm is better.

Keywords: Decision Tree, Naive Bayes, Confusion Matrix, Binary Classification

INTRODUCTION

The process of finding a model (function) that describes and differentiates class or concept data that aims to be used to predict the class of objects whose class label is unknown. The most widely used classification algorithms, namely decision / classification trees, Bayesian classifiers / Naïve Bayes classifiers, Neural networks, Statistical Analysis, Genetic Algorithms, Rough sets, k-nearest neighbor, Rule Based Methods, Memory based reasoning, and Support vector machines (SVM)[1]. The ID3 algorithm is a classification method using supervised learning concept. Training on the Decision Tree (ID3) uses training data that has been classified in advance. The concept of the Decision Tree is to transform data into a decision tree which then makes it to the rule. Data is input in the form of a table containing instances and attributes. One of the attributes is a class / category itself[2]. Data usually consists of more than one attribute. The problem in decision tree is how to select the attribute of the main anode and choose the next attribute. To select the main node, the gain value of each attribute is used. After that, gain information value of each attribute is determined. The attribute that has the most gain info will be the primary node. The selection of the next attribute that becomes the next node also uses the info gain value. Attributes that have less entropy value become main leaf first. The Naïve Bayes algorithm is among the most popular data mining algorithms [3]. Bayesian classification is a statistical classification that can be used to predict the probability of membership of a

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i3 (217-223)

ISSN : 1979-9292
E-ISSN : 2460-5611

class. Bayesian classification is based on the Bayes theorem which has similar classification capabilities to the decision tree and neural networks. Bayesian classification is proven to have high accuracy and speed when applied to databases with large data [4]. The Bayes method is a statistical approach to induced inference on classification problems. First discussed about the basic concepts and definitions in the Bayes Theorem, then use this theorem to classify in Data Mining.

The ability of the two classification techniques has been carried out by many previous studies. An algorithm needs to be compared to find out which algorithm is suitable in certain cases. Of course, the algorithms being compared must also have proven their ability through previous studies [5]. This study tries to simulate using home feasibility study data and compares the classification results of the two techniques based on previous studies in order to obtain a good understanding of the advantages and disadvantages of this technique.

METHOD

The data source used in this research is secondary data. The data was obtained from the publication of Housing and Health Statistics in North Sumatra 2016-2018. Decision Tree uses data from 2018 as training data, then test data is data for 2016 and 2017 [6]. The attributes that determine a livable house (Y): Average per capita floor area (AFA), proper drinking water source (PDS), proper sanitation (PS), electric lighting source (ELS), Non-leaf roof type (RT), Wall type (WT) and non-ground floor type (GT). The collected secondary data were analyzed and processed using the Decision Tree Algorithm and Naïve Bayes with the help of data processing software. The data processing flow chart uses the Decision Tree Algorithm and Naïve Bayes as follows:

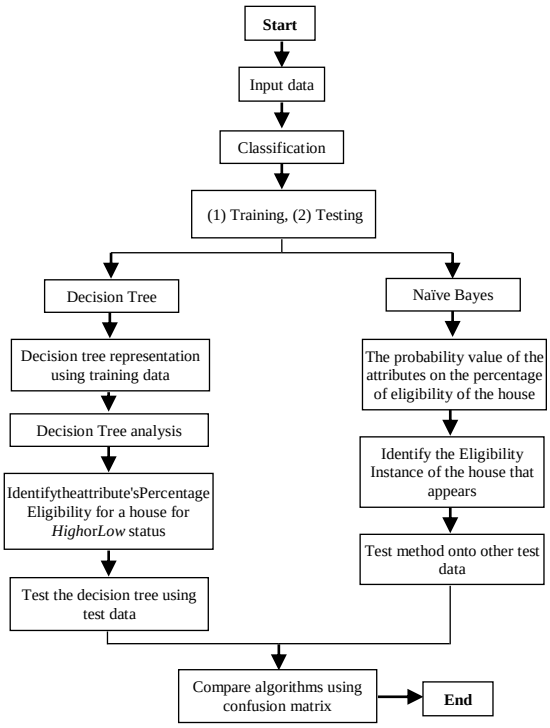

Figure 1. Research method diagram

Decision tree problem is how to select attribute which is the main anode and choose the next attribute. Choose the primary node based on the gain value of each attribute. After that, the gain information value of each attribute is determined. The attribute that has the most gain info will be the primary node. The selection of the next attribute that becomes the decision node also uses the info gain value. Attributes that have lower entropy value become main leaf first. Entropy or Info (I) (a term in J. Han's book) is the estimated number of bits needed to be able to extract a class ($+$ or $-$) from a number of random data in the

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i3 (217-223)

ISSN : 1979-9292
E-ISSN : 2460-5611

sample space S[7]. Entropy can be said as a bit requirement to represent a class. The smaller the Entropy value, the better it is to be used in extracting a class.The amount of Entropy / Info (I) in the sample space is defined by[8]:

a. Entropy for two classes: + and –

$$\text{Info}(D) = -p_+ \log_2 p_+ - p_- \log_2 p_- \quad (1)$$

b. Entropy for class> 2

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log(p_i) \quad (2)$$

The value of the gain info is determined by the equation (3) and (4):

$$\text{Info}_X(D) = \sum_{j=1}^{k} \frac{|D_j|}{|D|} \tilde{A}— I(D_j) \quad (3)$$

$$\text{Gain}(X) = \text{Info}(D) - \text{Info}_X(D) \quad (4)$$

Hypothesis Maximum Appropri Probability (HMAP) states the hypothesis is taken based on the probability value based on known prior conditions. HMAP is a simplified model of the Bayes method called Naive Bayes[9].HMAP is used in machine learning as a method to get a hypothesis for a decision. The Bayes method uses conditional probability as its basis. In science conditional probability is expressed as[10]:

$$P\left(X_k \middle| Y\right) = \frac{P(Y|X_k)}{\sum_i P(Y|X_i)} \quad (5)$$

where $\sum_i P(Y|X_i) > 0$. To explain the Naïve Bayes theorem, it should be noted that the classification process requires a number of clues to determine what class is suitable for the sample being analyzed. Therefore, the Bayes theorem above is adjusted as follows:

$$P(C|F_1 â€¦ F_n) = \frac{P(C)P\left(F_1 â€¦ F_n \middle| C\right)}{P(F_1 â€¦ F_n)} \quad (6)$$

Confusion matrix contains information that compares the classification results performed by the system with the classification results that should be[11]. In confusion matrix there is a type of binary classification which only has 2 class outputs:

Binary Classification		*Predicted Class*	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 1. Binary Classification

This confusion matrix performs calculations that produce 4 outputs:*recall, precision, accuracy,* and *error rate.* Recall is the success rate of the system in recovering information. Precision is the level of accuracy between the information requested by the user and the answers given by the system. Meanwhile accuracy is defined as the level of closeness between the predicted value and the actual value. The confusion matrix formula is as follows[12]:

$$\text{recall} = \frac{TP}{FN + TP} \tilde{A}— 100\% \quad (7)$$

$$\text{precission} = \frac{TP}{FP + TP} \tilde{A}— 100\% \quad (8)$$

accuracy

$$= \frac{(TP + TN)}{(TP + TN + FP + FN)} \tilde{A}— 100\% \quad (9)$$

RESULTS AND DISCUSSION

Naïve Bayes Classification

Steps to process data with Naïve Bayes:

a. Here, we have the training data, the test data classification process will be carried out, namely data001: $AFA =$

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i3 (217-223)

ISSN : 1979-9292
E-ISSN : 2460-5611

low, PDS =*low*, PS =*low*, ELS =*low*, RT = *low*, WT = *low* and GT = *low*.

b. We have 66 training data:
Probability Class Y (Percentage of livable houses) *high*:

$$P(\mathrm{Y_{high}}) = 42/66 = 0.636363636$$

Probability Class Y (Percentage of livable houses) *low*:

$$P(\mathrm{Y_{low}}) = 24/66 = 0.363636364$$

c. Calculates the probability of belonging to the $Y_{\mathbf{high}}$ category:

$$P(\mathrm{Y_{High}}|AFA_{\mathrm{Low}}) = 12/42 = 0.285714286$$
$$P(\mathrm{Y_{High}}|PDS_{\mathrm{Low}}) = 23/42 = 0.547619048$$
$$P(\mathrm{Y_{High}}|PS_{\mathrm{Low}}) = 17/42 = 0.404761905$$
$$P(\mathrm{Y_{High}}|ELS_{\mathrm{Low}}) = 17/42 = 0.404761905$$
$$P(\mathrm{Y_{High}}|RT_{\mathrm{Low}}) = 14/42 = 0.333333333$$
$$P(\mathrm{Y_{High}}|WT_{\mathrm{Low}}) = 10/42 = 0.238095238$$
$$P(\mathrm{Y_{High}}|GT_{\mathrm{Low}}) = 28/42 = 0.666666667$$

Probability included in category Y (Percentage of livable houses) *high*:

$$\mathrm{Y_{High}} = \frac{12}{42}Ã—\frac{23}{42}Ã—\frac{17}{42}Ã—\frac{17}{42}Ã—\frac{14}{42}Ã—\frac{10}{42}Ã—\frac{28}{42}$$
$$\mathrm{Y_{High}} = 0.001356276$$

d. Calculates the probability of belonging to the $Y_{\mathbf{low}}$ category:

$$P(\mathrm{Y}_{Low}|AFA_{\mathrm{Low}}) = 18/24 = 0.75$$
$$P(\mathrm{Y_{Low}}|PDS_{\mathrm{Low}}) = 20/24 = 0.833333333$$
$$P(\mathrm{Y_{Low}}|PS_{\mathrm{Low}}) = 18/24 = 0.75$$
$$P(\mathrm{Y_{Low}}|ELS_{\mathrm{Low}}) = 18/24 = 0.75$$
$$P(\mathrm{Y_{Low}}|RT_{\mathrm{Low}}) = 16/24 = 0.666666667$$
$$P(\mathrm{Y_{Low}}|WT_{\mathrm{Low}}) = 11/24 = 0.458333333$$
$$P(\mathrm{Y_{Low}}|GT_{\mathrm{Low}}) = 23/24 = 0.958333333$$

Probability included in category Y (Percentage of livable houses) *low*:

$$\mathrm{Y_{Low}} = \frac{18}{24}Ã—\frac{20}{24}Ã—\frac{18}{24}Ã—\frac{18}{24}Ã—\frac{16}{24}Ã—\frac{11}{24}Ã—\frac{23}{24}$$
$$\mathrm{Y_{Low}} = 0.102945964$$

e. Probability value of Class Percentage of livable houses is *High*< Class Percentage of livable houses is *Low*, it can be concluded that `data001` is in the category of Low percentage of livable houses.

The results of the same classification process using the naïve Bayes method on data testing totaling 33 districts resulted in the Table 2:

No	District	Y High	Y Low
1	data001	0.001356	0.102945964
2	data002	0.00434	0.121663411
3	data003	0.00434	0.121663411
4	data004	0.001356	0.102945964
5	data005	0.0217	0.020277235
6	data006	0.019384	0.00090121
…	…	…	…
33	data033	0.001995	0.034315321

Table 2. Result of Decision Test Data using Naïve Bayes

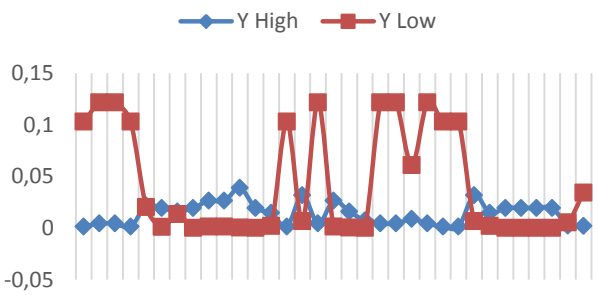

Figure 2. Result of Decision Test Data using Naïve Bayes

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

JURNAL IPTEKS TERAPAN
Research of Applied Science and Education V14.i3 (217-223)

ISSN : 1979-9292
E-ISSN : 2460-5611

ID3 Classification

Steps to convert the data into a tree:

a. Determine the selected node:

- To determine the selected node, use the Entropy value of each criterion with the sample data specified. The selected data is the criterion that has the greatest Entropy Gain. Info (Y) = I (36.30):

$$I(36,30) = -\frac{36}{66}\log_2\frac{36}{66} - \frac{30}{66}\log_2\frac{30}{66} = 0.47698 - (-0.51704) = 0.994030211$$

- Calculating info and gain based on existing attributes:

$$\text{Info}_{AFA}(Y) = -\frac{23}{66}I(21,2) + \frac{43}{33}I(15,28) = \frac{23}{66}\text{Ã—}0{,}426 + \frac{22}{33}\text{Ã—}0{,}933 = 0.756414346$$

So,

$\text{Gain}(AFA) = \text{Info}(Y) - \text{Info}_{AFA}(Y) = 0.994030211 - 0.756414346 = 0.23761586$

.

Following are the results of calculating the gain info for each attribute in Table 3.

Attribute	$I(p_i, n_i)$	Gain
AFA	0.756414	0.237616
PDS	0.722453	0.271578
PS	0.544941	0.449089
ELS	0.772588	0.221442
RT	0.932366	0.061664
WT	0.755277	0.238753
GT	0.822351	0.171679

Table 3. The gain of each attribute in the first iteration

The biggest gain value is attribute PS with value 0.449089 that attribute PS becomes the main node.

Determine the leaf node

The next Leaf Node can be selected on the part that has a value + and -, in this case attribute PS = *Low* which has value + and - so all of them must have leaf nodes. To arrange leaf nodes, do it one by one.

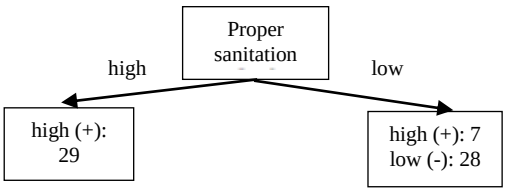

Figure 3. First Decision Tree

Since the values for + and - in the *Low* and *High* instances don't have a value of 0 (zero), we define the next leaf node for both instances. Next we determine the PS leaf node on the *high* instance.

In calculating the Info value, PS attribute is omitted because it has been designated as the primary node. Those that enter the calculation of other attributes are provided that the PS attribute has a *High* instance of 31 data. Info Value (Y):

$$\text{I}(29,2) = -\frac{29}{31}log_2\frac{29}{31} - \frac{2}{31}log_2\frac{2}{31} = 0.345117315$$

In the same way, gain for each attribute has been calculated. The Rule is obtained as follows:

IF(PS = "high", **IF**(GT = "high", "high", **IF**(WT = "high", "high", **IF**(RT = "high", **IF**(PDS = "low", "high", **IF**(AFA = "high", "high", "low")), **IF**(ELS = "high", "high", "low")))), **IF**(PDS = "high", **IF**(GT = "high", "high", "low"), **IF**(RT = "low", "low", **IF**(AFA = "high", "low", "low"))))

Testing Model withConfusionMatrix

The dataset is presented in two parts, 67% of the dataset will be used as training and

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

ISSN : 1979-9292
E-ISSN : 2460-5611

the remaining 33% will be used as testing.Confusion matrix testing uses equations (7), (8) and (9). True Positives (TP): "high" examples *correctly* identified as "high", True Negatives (TN): "low" examples *correctly* identified as "low", False Positives (FP): "low" examples *falsely*identified as "high", and False Negatives (FN): "high" examples *falsely* identified as "low".

Result of confusion matrix test on ID3:

Clasifier levels: High/Low

	Label Positive	Label Negative
Predict Positive	17	1
Predict Negative	2	13

Figure 4. ID3 Confusion Matrix

Thus, accuracy = 0.90909, precision = 0.94444 and recall = 0.89474.Result of confusion matrix test on Naïve Bayes:

Clasifier levels: High/Low

	Label Positive	Label Negative
Predict Positive	16	3
Predict Negative	6	8

Figure 5. Naive Bayes Confusion Matrix

Thus, accuracy = 0.72727, precision = 0.84211 and recall = 0.72727.

CONCLUSIONS

We recommend that study the statistics again to be able to really support mastery of the sciences of Data Mining, Naïve Bayes and Decision Tree.The level of accuracy of the decision tree model is better than the Naïve Bayes model. However, in terms of time, the naïve Bayes process is faster than the decision tree process. This will have a significant impact if the data larger.The formula for calculating Entropy Value has many versions and procedures. However, each of these formulations should give the same (not contradictory) results. For further research, we can using more varied instances to prevent accuracy errors with a wide variety of data models and case-like attributes with more than two instances.

ACKNOWLEDGMENT

Thank you to those who have helped both substantially and financially. To UniversitasPotensiUtama lecturers through the Faculty of Engineering and Computer Science who have provided theories to compile this research and we also like to thank our friends and family who supported us and offered deep insight into the research.

REFERENCES

[1] B. P. Battula and R. S. Prasad, "An overview of recent machine learning strategies in data mining," *Ionosphere*, vol. 351, no. 34, p. 0, 2013.

[2] A. I. Permana, "Accuracy Of C4.5 Algorithm Based Gain Average Values In Predicting Student Values," *J. Ipteks Terap.*, vol. 14, no. 2, pp. 99–105, 2020.

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020

ISSN : 1979-9292
E-ISSN : 2460-5611

[3] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Syst.*, vol. 192, p. 105361, 2020.

[4] M. Boullé, C. Charnay, and N. Lachiche, "A scalable robust and automatic propositionalization approach for Bayesian classification of large mixed numerical and categorical data," *Mach. Learn.*, vol. 108, no. 2, pp. 229–266, 2019.

[5] A. C. Sitepu, "Studi Komparatif Algoritma Genetika dan Simulated Annealing untuk Menyelesaikan Travelling Salesman Problem pada Masalah Transportasi," 2018.

[6] Badan Pusat Statistik Sumatera Utara, "Publikasi Online," *Badan Pusat Statistik*, 2018. [Online]. Available: https://sumut.bps.go.id/publication.html. [Accessed: 15-Apr-2020].

[7] J. Han and M. Kamber, "Classification and prediction," *Data Min. Concepts Tech.*, pp. 347–350, 2006.

[8] G. Ayyappan, D. C. Nalini, and D. A. Kumaravel, "Efficient mining for social networks using Information Gain Ratio based on Academic dataset," *Int. J. Civ. Eng. Technol.*, vol. 8, no. 1, 2017.

[9] Y. S. Belutowe, "Diagnosa Penyakit Septicaemia Epizootica Pada Sapi Ternak Dengan Teorema Bayes," *J. Teknol. Terpadu*, vol. 1, no. 2, 2015.

[10] T. D. Salma and Y. S. Nugroho, "Sistem Rekomendasi Pemilihan Sekolah Menengah Tingkat Atas Menggunakan Metode Naive Bayes," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 2, no. 2, pp. 85–94, 2016.

[11] C. P. Utomo, P. S. Pratiwi, A. Kardiana, I. Budi, and H. Suhartanto, "Best-Parameterized Sigmoid ELM for Benign and Malignant Breast Cancer Detection," in *International Conference on Artificial Intelligence and Pattern Recognition, AIPR 2014*, 2014.

[12] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: a measure driven view," *Inf. Sci. (Ny).*, vol. 507, pp. 772–794, 2020.

AUTHOR(S)BIOGRAPHY

Ade Clinton Sitepu

Ade Clinton Sitepu is an Assistant Lecturer in the Department of Informatics at InstitutTeknologidanBisnis Indonesia (ITBI) North Sumatera, Indonesia. He received anS.Si. fromUniversitas Sumatera Utara in the field of Computational Mathematics and is currently studying Master's Education in the Department of Computer Science at the University of Potential Utama. Its research and teaching focuses on machine learning, artificial intelligence and optimization.

Submitted : 30 Agt 2020 – Accepted : 26 Sep 2020 – Published : 29 Sep 2020