ISSN 2087-3336 (Print) | 2721-4729 (Online)
TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika
Vol. 11, No. 1, January 2024, page. 132-138
http://jurnal.sttmcileungsi.ac.id/index.php/tekno
DOI: 10.37373

Comparison of naive Bayes and decision tree algorithms to assess the performance of Palembang City fire and Disaster management employees

Dewi Sartika*, Rendra Gustriansyah

* Informatics Engineering, Faculty of Computer Science and Science, Indo Global Mandiri University, Indonesia, 30129. Jalan Jendral Sudirman No. 629 KM 4 Palembang

*✉ dewi.sartika@uigm.ac.id

Submitted: 04/10/2023 Revised: 30/11/2023 Accepted: 21/12/2023

ABSTRACT

The employee performance assessment at the Palembang City Fire and Disaster Management Service (DPKPB) is applied to other than the employee performance assessment implementation team based on the Decree of the Head of the Palembang City DPKPB Number 146 of 2021 concerning the employee performance assessment implementation team and awards for exemplary employees. Subjective assessments are avoided to obtain assessment results that are by the achievements of each employee. The application of data mining can be an alternative to avoid subjectivity in performance assessment. In this research, a comparison of the Naive Bayes and Decision Tree algorithms was carried out to assess the performance of Palembang City DPMPB employees. The results of further research will be used as an alternative solution in conducting performance assessments that are more objective than previous assessments. Both algorithms were evaluated for model performance using the Confusion Matrix. Based on the results of the evaluation carried out, it was stated that the Decision Tree algorithm had better accuracy, namely 91.74% compared to Naïve Bayes which had an accuracy of 88.99% with a test size of 0.4.

Keywords: Accuracy; confusion matrix; decision trees; naïve bayes; test size

1. INTRODUCTION

The Palembang City Fire and Disaster Management Service (DPKPB) conducts employee performance assessments periodically, namely within a quarter (three months). Performance assessment is carried out based on an assessment of several criteria, namely attendance, Employee Work Targets (SKP), work behavior, educational history, and performance reports. Employee performance assessment is applied to other than the employee performance assessment implementation team based on the Decree of the Head of the Palembang City DPKPB Number 146 of 2021 concerning the employee performance assessment implementation team and awards for exemplary employees. The results of the performance assessment are used as a reference in providing allowances. Subjective assessments are avoided to obtain assessment results that are by the achievements of each employee.

The application of data mining can be an alternative to avoid subjectivity in performance assessment. Data mining is carried out by using a set of data to be used as training data in forming a model or pattern. This model or pattern is then used in drawing conclusions and predictions. One data mining algorithm that has many advantages is Naive Bayes. The Naive Bayes algorithm is a simple, fast algorithm [1],[2], doesn't require a lot of training data [3], and is not sensitive to irrelevant data [4],[5],[6]. The Naive Bayes algorithm performs probability calculations based on the number of frequencies and combinations of values from the training data, where each attribute is independent [7],[8]. Apart from Naive Bayes, there is another algorithm, namely Decision Tree. The working principle of the Decision Tree algorithm is to divide the training data into small groups, where each group is divided based on data attributes and directed towards a decision [9][10]. The accuracy of a

TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. **ISSN** 2087-3336 (Print) | 2721-4729 (Online)

Decision Tree model depends on the pruning technique used to reduce model complexity and prevent overfitting [11],[12].

Based on the advantages of each algorithm, in this research, a comparison of the Naive Bayes and Decision Tree algorithms was carried out in assessing the performance of Palembang City DPMPB employees. In this research, we will find out which algorithm has a better level of accuracy. The results of further research will be used as an alternative solution in conducting performance assessments that are more objective than previous assessments. The performance of both algorithms was evaluated using the Confusion Matrix. Confusion matrix can comprehensively describe the performance of a model produced by an algorithm [13],[14].

2. LITERATURE REVIEW

2.1 *Naïve Bayes*.

The naïve Bayes algorithm predicts data using Bayes' Theorem calculations. Bayes' theorem is a statistical approach [15]. Bayes' Theorem calculations are carried out by calculating the probability of a hypothesis based on new evidence and data. The Bayes Theorem equation is stated in Equation 1, while for application in Naive Bayes classification, it can be stated in Equation 2.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (1)$$

$$P(Y|X) = \frac{P(Y)\Pi_{i=1}^{q} P(X_i|Y)}{P(X)} \quad (2)$$

Where :

X : data with unknown labels
Y : label/class
$P(Y|X)$: probability of label Y based on attribute X
$P(X|Y)$: probability of attribute X on label Y
P(X) : probability of attribute X
P(Y) : probability of label Y

2.2 *Decision tree*.

In Decision Tree, training data is arranged into a model that resembles a tree structure. In forming the model, Entropy calculations are carried out (Equation 3) which is useful for measuring the inhomogeneity or uncertainty of the data in each branch separation on the decision tree. Apart from calculating Entropy, when forming the model, Information Gain is also calculated (Equation 4). Information Gain is calculated based on the difference between Entropy before and after splitting the data using a feature or attribute. The greater the Entropy difference, the greater the Information Gain, which means the feature provides better and more informative separation.

$$Entropy(S) = \sum_{i=1}^{c} P_i \log 2^{P_i} \quad (3)$$

$$Gain(S, A) = \sum_{v \in V(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (4)$$

2.3 Confusion matrix.

Confusion Matrix can be used to evaluate model performance in predicting true and false. Confusion Matrix presents information about the classification carried out by the model by comparing the model predictions with the actual values of the data. The Confusion Matrix is arranged based on Table 1.

Table 1. Confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	Correctly predicted positive data (TP)	False predicted positive data (TN)
Negative	False predicted negative data (FP)	Correctly predicted negative data (FN)

The results of the confusion matrix can then be used to calculate accuracy (Equation 5), precision (Equation 6), recall (Equation 7), F1-score (Equation 8) [16],[17].

$$Accuracy = \frac{TP+TN)}{TP+TN+FP+FN} \quad (5)$$

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1 - score = \frac{2\times TP}{2\times TP+FP+FN} \quad (8)$$

3. METHODE

This research was carried out based on several stages which were carried out systematically according to Figure 1:

a. Study of literature
The determination of the two algorithms to be compared was carried out based on algorithms that were relevant for assessing the performance of Palembang City DPMPB employees. Apart from the fact that these two algorithms are relevant for all types of data, these two algorithms also have a high level of accuracy for similar cases based on research that has been carried out to determine employee promotions [18] and teacher performance assessment [19].

b. Data preparation
The data used in this stage is performance assessment data for Palembang City DPMPB employees. The performance data for quarter 1 (January-March) 2023 is 218 data records as training data and the performance assessment data for Palembang City DPMPB employees for quarter 2 (April-June) 2023 is 218 data records as test data. Both data are saved in .csv document format.

c. Model formation
At this stage, it is done by dividing the training data into a ratio of 80% training data-20% test data, 70% training data-30% test data, and 60% training data-40% test data. The level of accuracy for these three comparisons is calculated based on the Confusion Matrix. This stage is carried out in Python.

d. Evaluation of model performance
At this stage, which is also carried out in Python, each model from the two algorithms is evaluated to determine employee performance scores based on test data.

e. Analysis of results
At this stage, an analysis is carried out from the results of the model performance evaluation of the two algorithms, so that the best algorithm is obtained in assessing the performance of Palembang City DPKPB employees.

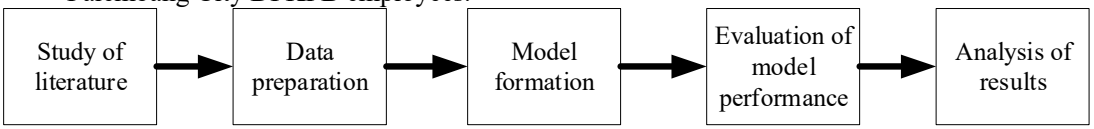

Figure 1. Research stages

4. RESULTS AND DISCUSSION

The number of employees at the Palembang City DPMPB is 218 people. The data used to form the model is performance assessment data for Palembang City DPMPB employees for quarter 1 (January – March) 2023. The attributes used in the performance assessment can be seen in Table 2.

Table 2. Attribute description

Attribute	Type	Value
Presence	Categorical	Discipline Undisciplined

Attribute	Type	Value
Performance Report	Categorical	Fulfill
		Does not meet the
SKP	Categorical	Needs Repair
		Good
		Very good
Performance	Categorical	Needs Repair
Behavior		Good
		Very good
Last education	Categorical	SMP
		SMA/SMK
		D3
		S1
		S2

Model formation for each algorithm using Python was carried out in three comparisons of training and test data (test size). The Confusion Matrix results from the formation of the Naive Bayes model with a comparison of 80% training data-20% test data (test size 0.2) in Table 3, a comparison of 70% training data-30% test data (test size 0.3) Table 4 and the comparison of 60% training data-40% test data (test size 0.4) in Table 5. Meanwhile, the model accuracy for each comparison of training data is in Table 6.

Table 3. Confusion matrix naïve Bayes test size 0.2

Actual	Predicted	
	Positive	Negative
Positive	25	0
Negative	4	15

Table 4. Confusion matrix naïve Bayes test size 0.3

Actual	Predicted	
	Positive	Negative
Positive	37	0
Negative	4	25

Table 5. Confusion matrix naïve Bayes test size 0.4

Actual	Predicted	
	Positive	Negative
Positive	52	0
Negative	5	31

Table 6. Naïve bayes model accuracy

	Test Size		
	0,2	0,3	0,4
Accuracy (%)	90	93	94

Confusion Matrix results from the formation of the Decision Tree model with a comparison of 80% training data - 20% test data (test size 0.2) in Table 7, a comparison of 70% training data - 30% test data (test size 0.3) in Table 8 and the comparison of 60% training data - 40% test data (test size 0.4) in Table 9. Meanwhile, the model accuracy of each training data comparison in Table 10.

Table 7. Confusion matrix decision tree test size 0.2

Actual	Predicted	
	Positive	Negatif
Positive	25	0
Negative	2	17

Table 8. Confusion matrix decision tree test size 0.3

Actual	Predicted

	Positive	Negatif
Positive	37	0
Negative	2	27

Table 9. Confusion matrix decision tree test size 0.4

Actual	Predicted	
	Positive	Negatif
Positive	52	0
Negative	3	33

Table 10. Accuracy model decision tree

	Test Size		
	0,2	0,3	0,4
Accuracy (%)	95	97	96

Each model from the Naive Bayes and decision tree algorithms is used to predict performance assessment data for Palembang City DPMPB employees for the 2nd quarter (April - June) of 2023. The accuracy of predictions using the Naive Bayes algorithm in Table 11, while the accuracy of predictions using the algorithm The decision tree in Table 12. The comparison graph of the accuracy of the two algorithm models is presented in the graph (Figure 2).

Table 11. Prediction accuracy using the naïve Bayes model

	Test Size		
	0.2	0.3	0.4
Accuracy (%)	88,53	88,53	88,99

Table 12. Prediction accuracy using the decision tree model

	Test Size		
	0.2	0.3	0.4
Accuracy (%)	91,74	91,74	91,74

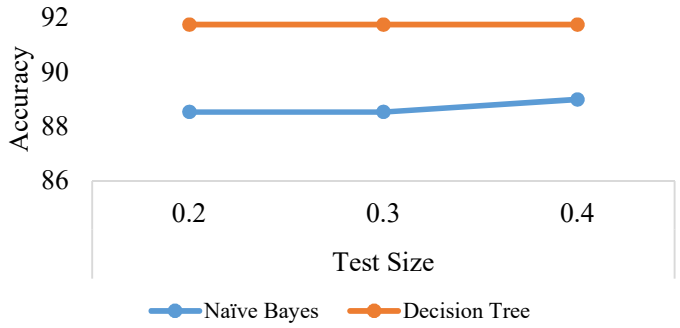

Figure 2. Comparison graph of the accuracy of Naive Bayes and decision trees

5. CONCLUSION

Based on the evaluation results that have been carried out using the Confusion Matrix, it can be concluded that the decision tree algorithm model for the three test sizes (0.2, 0.3, and 0.4) shows the same accuracy for predicting test data (2nd quarter performance values) which is equal to 91.74%, while the Naive Bayes algorithm model for a test size of 0.4 shows higher accuracy than the other two models, namely 88.99%. This is because the three models (test size 0.2, 0.3 and 0.4) used have a relatively close level of accuracy at the model formation stage. Therefore, it can be concluded that the best model that can be used in assessing the performance of Palembang City DPKPB employees is the Decision Tree algorithm model with a test size of 0.4.

ACKNOWLEDGEMENT

Thanks are expressed to Indo Global Mandiri University as the institution where the author resides for support in implementing the tri dharma of higher education. Thanks are also expressed to Intan Cahya Muda, student of the Informatics Engineering Study Program who has assisted in the technical data collection for this research.

REFERENCE

[1] L. Setiyani, M. Wahidin, D. Awaludin, and S. Purwani, "Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Data Mining Naïve Bayes : Systematic Review," *Fakt. Exacta*, vol. 13, no. 1, p. 35, 2020, doi: 10.30998/faktorexacta.v13i1.5548.

[2] K. Rrmoku, B. Selimi, and L. Ahmedi, "Application of Trust in Recommender Systems—Utilizing Naive Bayes Classifier," *Computation*, vol. 10, no. 1, 2022, doi: 10.3390/computation10010006.

[3] F. Harahap, A. Y. N. Harahap, E. Ekadiansyah, R. N. Sari, R. Adawiyah, and C. B. Harahap, "Implementation of Naïve Bayes Classification Method for Predicting Purchase," *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. August 2018, 2019, doi: 10.1109/CITSM.2018.8674324.

[4] P. Al, M. Prasetyo, and A. Hermawan, "Analisis sentimen twitter terhadap pemilihan presiden menggunakan algoritma naïve bayes Sentiment analysis twitter of presidential election using naïve bayes algorithm," vol. 4, pp. 224–233, 2024, doi: 10.37373/infotech.v4i2.863.

[5] Hubert, P. Phoenix, R. Sudaryono, and D. Suhartono, "Classifying Promotion Images Using Optical Character Recognition and Naïve Bayes Classifier," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 498–506, 2021, doi: 10.1016/j.procs.2021.01.033.

[6] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052068.

[7] P. Sainanda Cahyani Moonallika, K. Queena Fredlina, and I. Kresna Sudiatmika, "Penerapan Data Mining Untuk Memprediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes Classifier (Studi Kasus STMIK Primakara)," *Progresif*, vol. 16, no. 1, pp. 47–56, 2020.

[8] S. W. Mafriza, "Klasifikasi karir mahasiswa bidang web developer menggunakan algoritma naïve bayes Career classification of web developer students using the naïve bayes algorithm," vol. 4, pp. 270–280, 2023, doi: 10.37373/infotech.v4i2.907.

[9] H. Syahputra, J. Nainggolan, J. Apriadi, and P. Sirait, "Red Chili Classification Using HSV Feature Extraction and Naive Bayes Classifier," vol. 11, no. 1, pp. 33–40, 2023.

[10] S. R. Jiao, J. Song, and B. Liu, "A Review of Decision Tree Classification Algorithms for Continuous Variables," *J. Phys. Conf. Ser.*, vol. 1651, no. 1, 2020, doi: 10.1088/1742-6596/1651/1/012083.

[11] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.

[12] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "Multi-label Classifier Performance Evaluation with Confusion Matrix," no. June 2023, pp. 01–14, 2020, doi: 10.5121/csit.2020.100801.

[13] B. A. Muktamar, N. A. Setiawan, and T. B. Adji, "Analisis Perbandingan Tingkat Akurasi Algoritma Naive Bayes Classifier dengan Correlated-Naive Bayes Classifier," *Semin. Nas. Teknol. Inf. dan Multimed. 2015*, pp. 49–54, 2015.

[14] T. Puspa, R. Sanjaya, A. Fauzi, A. Fitri, and N. Masruriyah, "Analisis sentimen ulasan pada e-commerce shopee menggunakan algoritma naive bayes dan support vector machine Analysis of review sentiment on shopee e-commerce using the naive bayes algorithm and support vector machine," vol. 4, pp. 16–26, 2023, doi: 10.37373/tekno.v10i2.419.

[15] F. M. J. Mehedi Shamrat, S. Chakraborty, M. M. Billah, P. Das, J. N. Muna, and R. Ranjan, "A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm," *Proc. 5th Int. Conf. Trends Electron. Informatics, ICOEI 2021*, no. April, pp. 1339–1345, 2021, doi: 10.1109/ICOEI51242.2021.9452898.

[16] A. Arias-Duart, E. Mariotti, D. Garcia-Gasulla, and J. M. Alonso-Moral, "A Confusion Matrix for Evaluating Feature Attribution Methods," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2023-June, no. Cv, pp. 3709–3714, 2023, doi: 10.1109/CVPRW59228.2023.00380.

[17] E. R. Arumi, Sumarno Adi Subrata, and Anisa Rahmawati, "Implementation of Naïve bayes Method for Predictor Prevalence Level for Malnutrition Toddlers in Magelang City," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 2, pp. 201–207, 2023, doi: 10.29207/resti.v7i2.4438.

[18] P. S. Dewi, C. K. Sastradipraja, and D. Gustian, "Sistem Pendukung Keputusan Kenaikan Jabatan Menggunakan Metode Algoritma Naïve Bayes Classifier," *J. Teknol. dan Inf.*, vol. 11, no. 1, pp. 66–80, 2021, doi: 10.34010/jati.v11i1.3593.

[19] R. Toyib and S. A. Saputera, "Aplikasi Sistem Penilaian Kinerja Guru Dengan Metode Decision Tree Menggunakan Algoritma ID3 (Studi Kasus SLTP Negeri 3 Marga Sakti Bengkulu Utara)," *J. Technopreneursh. Inf. Syst.*, vol. 2, no. 1, pp. 1–7, 2019, doi: 10.36085/jtis.v2i1.88.