Eduvest – Journal of Universal Studies
Volume 3 Number 6, June, 2023
p- ISSN 2775-3735- e-ISSN 2775-3727

APPLICATION OF J48 AND NAÏVE BAYES ALGORITHMS TO PREDICT REAM BOOKINGS AT PT. NIPPON PRESISI TEKNIK

Panji Satrio Bakti, Eliyani
Universitas Mercubuana, Indonesia
Email: 41519210078@student.mercubuana.ac.id, eliyani@mercubuana.ac.id

ABSTRACT

In the field of goods production, demand prediction is important. By doing sales predictions, companies can make calculations and forecasts for what raw materials are mostly ordered. J48 and Naïve Bayes algorithm are two popular machine learning technique. By using these two algorithms, this study aims to develop an accurate and more reliable predictive model that help the company to make data driven decision. This study focuses on the application of quantitative methods, specifically the J48 algorithm and Naïve Bayes algorithm. This research conducted 4 times testing on each algorithm. This study produces high accuracy values with the Naïve Bayes and J48 algorithms. Both algorithm results have a fairly high accuracy value of 94% for Naïve Bayes and 98% for J48. The findings of this study implicate that by using J48 and Naïve Bayes algorithm, company can make informed decisions lead to improved operational efficiency, cost-effective, and resource utilization.

KEYWORDS *Naïve Bayes; J48; Decision Tree; Algorithm*

INTRODUCTION

In the field of goods production, demand prediction is important. Predicting the sale of an item is a necessity in doing business. Sales results are important in the sustainability of the company. However, not every company can run smoothly and stably. Sales results will show whether the company will run well or not.

By doing sales predictions, companies can make calculations and forecasts for what raw materials are mostly ordered. Prediction systems that use good algorithms are the key to improving production and revenue systems in a company. With the

How to cite:	Bakti, P. S., & Eliyani. (2023). Application of J48 and Naïve Bayes Algorithms to Predict Ream Bookings at PT. Nippon Presisi Teknik. *Journal Eduvest. 3* (6): 1047-1060
E-ISSN:	2775-3727
Published by:	https://greenpublisher.id/

ability to analyze data, the algorithm can provide accurate predictions according to future market needs. With accurate estimates, companies can plan the right steps in increasing operational efficiency, optimizing resources, and also reducing production costs.

Based on research conducted by Kaunang (2018) using the J48 algorithm to analyze the poverty level of the population in Indonesia. The data used was obtained from the Indonesian Central Bureau of Statistics (BPS). The accuracy obtained in this study showed a result of 88.6%, which is a good result. The prediction model used helps policy makers to make decisions.

Then on to research Ronaldo (2021) using the Naïve Bayes algorithm to make predictions in pesticide sales to companies. The results of this prediction produce an accuracy rate of 94.53%, where these results can help companies to increase sales turnover in order to achieve targets.

Then on to research Rukmana (2021) compares how the Naïve Bayes, Decision Tree-j48, and Lazy-IBK algorithms perform. This study aims to determine the highest accuracy of the three algorithms being compared. The results of the research show that Decision Tree-j48 has very high accuracy compared to the Naïve Bayes and Lazy-IBK algorithms.

Sinaga (2022) analyzing the feasibility of borrowing capital in MSMEs with the J48 algorithm. This research builds an application that implements the J48 algorithm which is able to answer problems related to the feasibility analysis of MSME capital loans at PT. Pawnshop.

Then Cendana (2019) conducted a comparison of the Naïve Bayes, J48, and Random Forest Tree algorithms to increase MSME customer loyalty. This study aims to help decision making in giving shopping vouchers so that MSMEs can run to get optimal benefits.

On research Hayuningtyas (2019) the Naïve Bayes algorithm is used to recommend women's clothing. The results of this study resulted in Naïve Bayes helping to recommend women's clothing based on predetermined attributes.

Amillina (2021) carry out the application of Naïve Bayes to classify the level of student satisfaction with online learning. This study gives the result that Naïve Bayes is appropriate for measuring the level of student satisfaction in online learning, where the accuracy level reaches 100% and the precision and recall values reach 100%.

Studysardi (2020) classify the graduation rate of electronics students using the Naïve Bayes algorithm. This study uses 20 attributes with semester 1-3 scores. This study found that students in 2014 entered had an accuracy of 79.07% and students in 2015 had an accuracy of 68%.

Pakpahan (2021) implementing the J48 algorithm in determining the buyer's shopping itemset pattern. This research with the J48 algorithm generates shopping patterns of buyers, items that are frequently purchased, and helps supermarkets to make decisions to add to the increase in these items.

The J48 algorithm is one of the most commonly used machine learning algorithms in decision making. In many applications, the J48 has proven effective in producing accurate predictive models using easy-to-interpret decision tree structures (Quinlan, 2014). The Naïve Bayes algorithm is a popular probability classification method in analysis. This algorithm assumes of feature independence which enables efficient calculation of class probabilities. Naïve Bayes has proven successful in a variety of applications (Rish, 2001).

The aims of this study are to explore the potential of J48 and Naïve Bayes algorithm in predicting ream bookings at PT. Nippon Presisi Teknik. By using these two popular machine learning, we aim to develop an accurate and more reliable predictive model that help the company to make data driven decision.

RESEARCH METHOD

This research consists of several stages as shown in Figure 1.

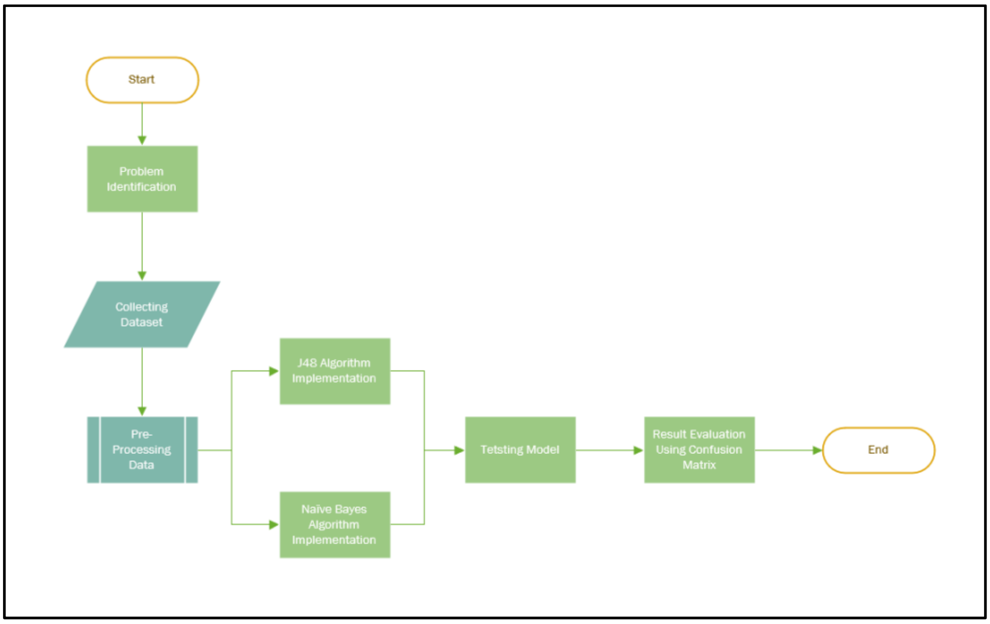

Figure 1. Research Stages

An explanation regarding the stages of the research in Figure 1 is explained as follows:

1) Identification of problems
   this research is to make predictions to find out what items are mostly ordered with the J48 and Naïve Bayes Algorithms.
   Datasets
2) The dataset used in this study is a purchase order dataset at PT Nippon Presisi Teknik in 2019-2022.
3) Pre-processing
   The dataset obtained needs to be pre-processed. Pre-processing is done to convert the raw dataset into a format that can be applied to data mining techniques. In pre-processing there are 2 stages, namely data cleansing and data transform.

4) Algorithm Implementation
 a) J48
 The decision tree consists of decision nodes and leaf nodes, where the decision node determines the test of one of the attributes and the leaf node represents the class value (Ruggieri, 2002). This study uses the J48 algorithm because it has a higher level of accuracy (Jains, 2012). To determine the root decision, the root will be taken from the selected attribute, by calculating the gain value of the attributes, the highest gain value will be the first root. To calculate the gain value, it is necessary to calculate the entropy value first, with the formula:

$$Entropy(S) = \Sigma_{i=1}^{n} - pi * log_2\, pi$$

Information:
S = Case Set
n = Number of sth partitions
pi = Number of cases on the i-th partition

Then calculate the gain value, with the formula:

$$Gain(S, A) = S - \Sigma_{i=1}^{n} \frac{|S_i|}{|S|} * S_i$$

 b) Naïve Bayes
 Naïve Bayes is an algorithm that is included in the top 10 algorithms in data mining (Wu et al., 2008). The Naïve Bayes algorithm is an easy probability classification. This algorithm calculates a set of probabilities by calculating the frequencies and combinations of values in a given data set (Saritas & Yasar, 2019).

Calculation of the Naïve Bayes algorithm as follows:

$$P(H|E) = \frac{P(E|H) * P\ (H)}{P(E)}$$

Information:
H = Unknown class data
E = Hypothesis on data H which is a special class
P(E|H)= Probability value E based on the condition of the hypothesis H
P(H)= Probability value in hypothesis H
P(E)= probability value E

5) Model Testing

Model testing was carried out by testing to calculate the values of precision, recall, accuracy, and f1-score from the J48 and Naïve Bayes algorithms. The dataset that has gone through pre-processing is divided into 2, training data and testing data. In dividing the data into training data and testing data, different data comparisons were made. Differences in the distribution of data carried out are shown in table 1.

Table 1. Differences in the Distribution of Data Training and Data Testing

Testing.	Training Data.	Data Testing.
Testing 1	60%	40%
Test 2	70%	30%
Test 3	80%	20%
Testing 4	90%	10%

In each test, calculations are made for the precision, recall, accuracy, and f1-score values of the J48 and Naïve Bayes algorithms. This test is done to find out which test gets the best value.

6) Outcome Evaluation

The last stage of the research will evaluate the results on precision, recall, accuracy, and f1-score calculations from the J48 and Naïve Bayes algorithms to find out how the results of the two algorithms differ.

RESULTS AND DISCUSSION

Datasets

The dataset used in this study was found to have PO No, Item Name, Type, Size, Material, Quantity, PO Month, etc. with a total of 635 data.

```
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Nomor PO     635 non-null    int64
 1   Nama Barang  635 non-null    object
 2   Tipe         635 non-null    object
 3   Ukuran       635 non-null    int64
 4   Material     635 non-null    object
 5   Quantity     635 non-null    int64
 6   Bulan PO     635 non-null    object
 7   PT           635 non-null    object
 8   Status       635 non-null    object
dtypes: int64(3), object(6)
```

Figure 2. Dataset Attributes

Data Cleaning

The dataset that has been entered is carried out at the pre-processing stage, the data will be cleaned/cleansing first. Data cleansing is cleaning data, the cleaning that is meant here is the process of repairing or deleting data that is not needed and also lost data. Checking first whether there are missing data from the dataset, the following are the results of checking:

```
dataset.isnull().sum()

nomor_po            0
nama_barang         0
ukuran              0
quantity            0
status              0
tipe_encoded        0
material_encoded    0
bulan_encoded       0
pt_encoded          0
dtype: int64
```

Figure 3. Checking Missing Data

Based on figure 3, there is no missing data for each attribute, which means that data processing is not required to fill in the blank data. Next, check the correlation between attributes to find out whether there are attributes that are not needed. The following code is used to check the correlation

```
plt. figure(figsize=(15, 10))
sns.heatmap(dataset.corr(), annot=True)
plt. title('Heat Map', size=20)
plt. yticks(rotation = 0)
plt. show()
```

Outputs:

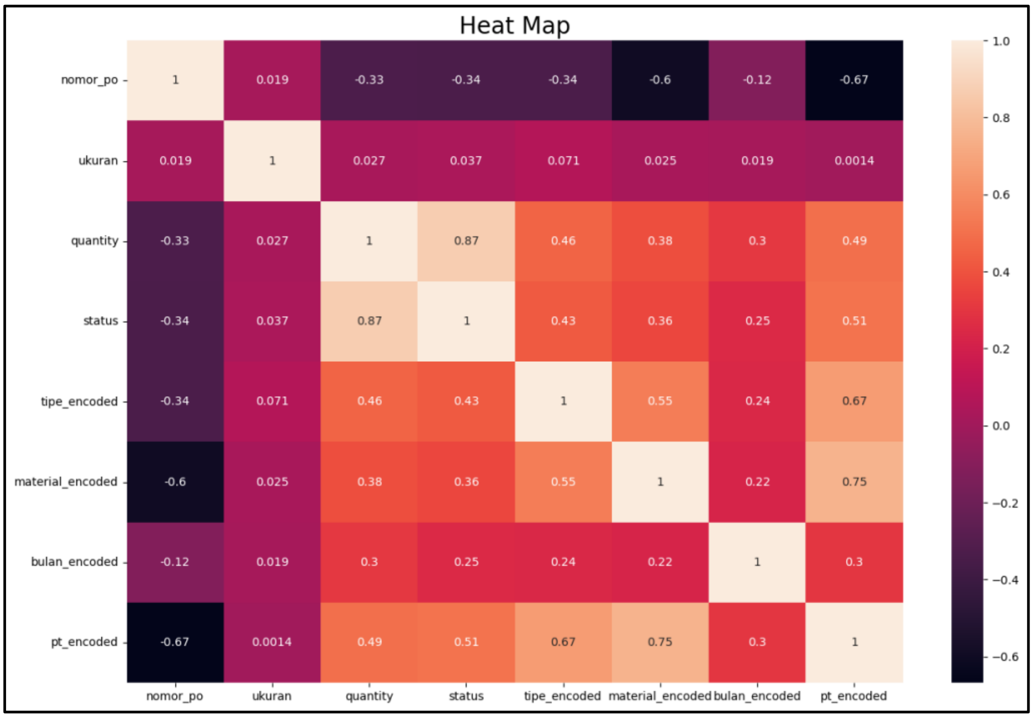

Figure 4. Heatmap Correlation

Based on figure 4, the numer_po attribute has a negative value with 6 other attributes, so the numer_po attribute is discarded because it does not have a high correlation with other attributes.

Transformation Data

After the data is cleaned in data cleansing, the next step is data transformation. Data transformation transforms data so that data can be read for data processing and predictions. The following code is used to perform the transformation:

```
from sklearn. preprocessing import LabelEncoder
le = LabelEncoder()
dataset['tipe_encoded'] = le.fit_transform(dataset.tipe)
dataset['material_encoded'] = le.fit_transform(dataset.material)
dataset['month_encoded'] = le.fit_transform(dataset.month_po)
dataset['pt_encoded'] = le.fit_transform(dataset.pt)
dataset.drop(['type'], axis = 1, inplace = True)
dataset.drop(['material'], axis = 1, inplace = True)
dataset.drop(['month_po'], axis = 1, inplace = True)
dataset.drop(['pt'], axis = 1, inplace = True)
varlist = ['status']
```

Output datasets:

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 635 entries, 0 to 634
Data columns (total 8 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   nama_barang       635 non-null    object
 1   ukuran            635 non-null    int64 
 2   quantity          635 non-null    int64 
 3   status            635 non-null    int64 
 4   tipe_encoded      635 non-null    int32 
 5   material_encoded  635 non-null    int32 
 6   bulan_encoded     635 non-null    int32 
 7   pt_encoded        635 non-null    int32 
dtypes: int32(4), int64(3), object(1)
memory usage: 29.9+ KB
```

Figure 5. Dataset After Transform

Based on figure 5, the dataset which was previously an object type has been changed to an int type using the label encoder but the nama_item attribute is still an object. The item name attribute will also be changed to type int with the following code:

```
x = dataset.iloc[:, :8].values
y = dataset. status
[:,0] = le.fit_transform(x[:,0])
```

Outputs:

```
print(x)
```

```
[[203 10 1 ... 0 3 1]
 [309 10 2 ... 0 3 1]
 [318 10 2 ... 0 3 1]
 ...
 [207 17 2 ... 0 2 1]
 [221 19 1 ... 0 2 1]
 [215 19 1 ... 0 2 1]]
```

Figure 6. Dataset for X

Based on Figure 6 and based on the code that is run, labeling is done to make predictions, where labeling will be used on the status attribute to find out which item predictions are ordered the most. Then also the item name attribute is also transformed into an int form.

Naïve Bayes implementation

The results of applying Naïve Bayes in the 1st test get the results:

	precision	recall	f1-score	support
banyak	0.66	1.00	0.79	19
sedikit	1.00	0.96	0.98	235
accuracy			0.96	254
macro avg	0.83	0.98	0.88	254
weighted avg	0.97	0.96	0.96	254

Figure 7. First Test Naïve Bayes

The results of applying Naïve Bayes in the 2nd test get the following results:

	precision	recall	f1-score	support
banyak	0.54	1.00	0.70	13
sedikit	1.00	0.94	0.97	178
accuracy			0.94	191
macro avg	0.77	0.97	0.84	191
weighted avg	0.97	0.94	0.95	191

Figure 8. Second Test Naïve Bayes

The results of applying Naïve Bayes in the 3rd test get the results:

	precision	recall	f1-score	support
banyak	0.53	1.00	0.70	8
sedikit	1.00	0.94	0.97	119
accuracy			0.94	127
macro avg	0.77	0.97	0.83	127
weighted avg	0.97	0.94	0.95	127

Figure 9. Third Test Naïve Bayes

The results of applying Naïve Bayes in the 4th test get the results:

	precision	recall	f1-score	support
banyak	0.43	1.00	0.60	3
sedikit	1.00	0.93	0.97	61
accuracy			0.94	64
macro avg	0.71	0.97	0.78	64
weighted avg	0.97	0.94	0.95	64

Figure 10. Fourth Test Naïve Bayes

Based on the 4 tests carried out with different training data values and testing data when made in table form it will be as follows:

Table 2. Differences in the Distribution of Data Training and Data Testing

Testing	Testing 1	Test 2	Test 3	Testing 4
accuracy	96%	94%	94 %	94%
Precision	0:1	0:1	0:1	0:0.97
	1:0.66	1:0.54	1: 0.53	1:0.6
recall	0:0.96	0:0.94	0:0.94	0:0.93
	1:1	1:1	1:1	1:1
F-1 Score	0:0.98	0:0.97	0:0.97	0:0.97
	1:0.79	1:0.7	1:0.7	1:0.6

Based on table 2, it can be concluded that the results from test 1 have a better value. It can also be concluded that the larger the testing data used, the greater the evaluation results, and conversely, the smaller the testing data, the less the evaluation results. Then visualize the confusion matrix data in the image below.

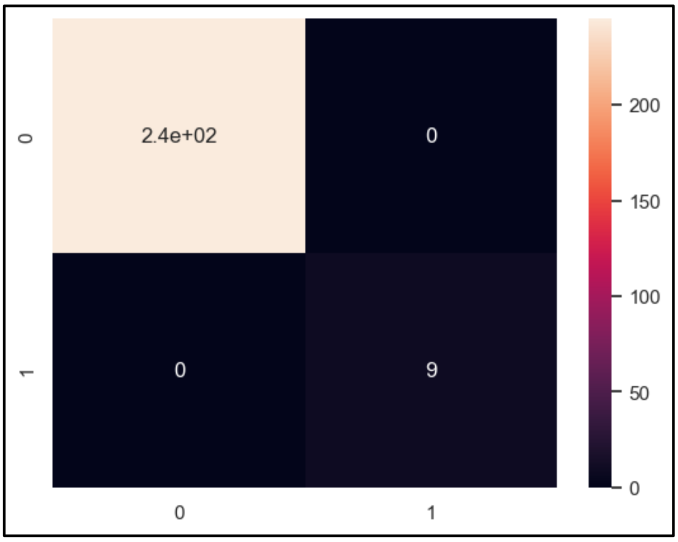

Figure 11. Confusion Matrix Naïve Bayes

The explanation of figure 11 is as follows:

1) The value 245 is in the TRUE NEGATIVE (TN) value position which is negative data that is predicted correctly.
2) Value 9 is in the TRUE POSITIVE (TP) value position which is positive data that is predicted correctly.
3) The value 0 is in the FALSE POSITIVE (FP) position which is negative data but is predicted to be positive
4) The value 0 is in the FALSE NEGATIVE (FN) position which is positive data but is predicted to be negative.

The value 0 of the explanation is data with many statuses and the value 1 is data with little status.

J48 implementation

The results of implementing J48 in the 1st test get the results:

	precision	recall	f1-score	support
banyak	0.94	0.89	0.92	19
sedikit	0.99	1.00	0.99	235
accuracy			0.99	254
macro avg	0.97	0.95	0.96	254
weighted avg	0.99	0.99	0.99	254

Figure 12. First Test J48

The results of implementing J48 in the 2nd test get the results:

	precision	recall	f1-score	support
banyak	0.92	0.92	0.92	13
sedikit	0.99	0.99	0.99	178
accuracy			0.99	191
macro avg	0.96	0.96	0.96	191
weighted avg	0.99	0.99	0.99	191

Figure 13. Second Test J48

The results of implementing J48 in the 3rd test get the results:

	precision	recall	f1-score	support
banyak	0.88	0.88	0.88	8
sedikit	0.99	0.99	0.99	119
accuracy			0.98	127
macro avg	0.93	0.93	0.93	127
weighted avg	0.98	0.98	0.98	127

Figure 14. Third Test J48

The results of implementing J48 in the 4th test get the results:

	precision	recall	f1-score	support
banyak	0.75	1.00	0.86	3
sedikit	1.00	0.98	0.99	61
accuracy			0.98	64
macro avg	0.88	0.99	0.92	64
weighted avg	0.99	0.98	0.99	64

Figure 15. Fourth Test J48

Based on the 4 tests carried out with different training data values and testing data when made in table form it will be as follows:

Table 3. Differences in the Distribution of Data Training and Data Testing

Testing	Test 1	Test 2	Test 3	Testing 4
Accuracy	99%	99%	98%	98%
Precision	0:0.99	0:0.99	0:0.99	0:1
	1:0.94	1:0.92	1:0.88	1:0.75
Recall	0:0.89	0:0.99	0:0.99	0:0.98
	1:0.89	1:0.92	1:0.88	1:1
F-1 Score	0:0.92	0:0.90	0:0.99	0:0.99
	1:0.92	1:0.92	1:0.88	1:0.86

Based on table 3, it can be concluded that the results of the data testing by 30% have a better value. Based on the results of 4 tests, each test has different results. As in many statuses, the testing data is 10%, 20%, and 30% overall have a good value, but in many statuses, data 30% and 40% overall have better results than the others. It was concluded that the results of the 30% testing data had a better value than the others in testing the J48 algorithm. Then visualize the confusion matrix data in the image below.

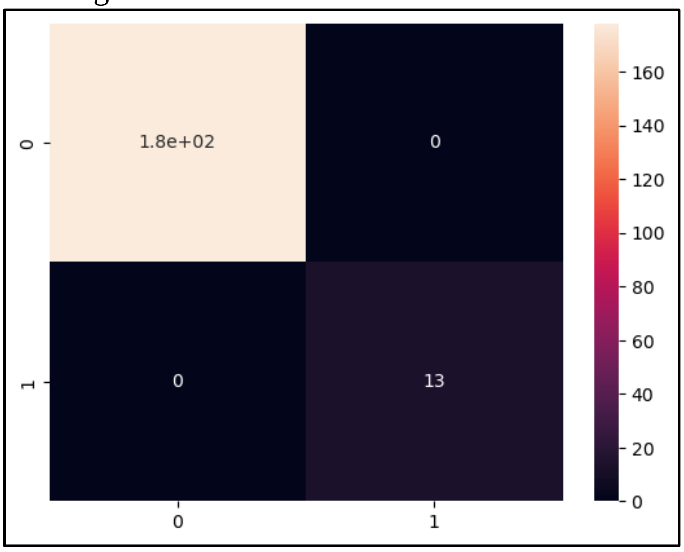

Figure 16. Confusion Matrix Naïve Bayes

The explanation of figure 16 is as follows:

1) The value 178 is in the TRUE NEGATIVE (TN) value position which is negative data that is predicted correctly.
2) The value 13 is in the TRUE POSITIVE (TP) value position which is positive data that is predicted correctly.
3) The value 0 is in the FALSE POSITIVE (FP) position which is negative data but is predicted to be positive
4) The value 0 is in the FALSE NEGATIVE (FN) position which is positive data but is predicted to be negative.

The value 0 of the explanation is data with many statuses and the value 1 is data with little status. Then for the decision tree display as follows:

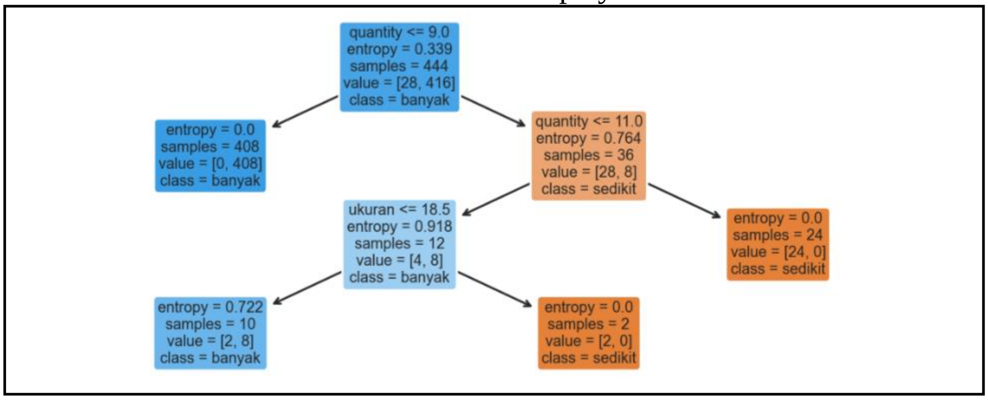

Figure 17. Decision Tree Results

CONCLUSION

This research was conducted using 2 algorithms, namely Naïve Bayes and J48. This research was conducted 4 times testing on each algorithm. Based on the results of the research that has been done, the Naïve Bayes algorithm gives the best results at 60% training data and 40% testing data, while the J48 algorithm gives the best results at 70% training data and 30% testing data. Based on the results of the two algorithms that have the best results, the evaluation value of the J48 algorithm has better precision, recall, f1-score, and accuracy than the Naïve Bayes algorithm. Based on the results of this study it was concluded that the J48 algorithm produces better performance than the Naïve Bayes algorithm. Both algorithm results have a fairly high accuracy value, namely 94% for Naïve Bayes and 98% for J48.

REFERENCES

Amillina, I., & Qoiriah, A. (2021). Penerapan Algoritma Naïve Bayes dalam Klasifikasi Tingkat Kepuasan Siswa terhadap Pembelajaran Daring. *Jurnal Ilmiah Teknologi Informasi Dan Robotika*, *3*(2), 16–23.

Cendana, M., & Permana, S. D. H. (2019). Analisis Perbandingan Algoritma Naive Bayes, J48, Dan Random Forest Tree Dalam Peningkatan Loyalitas Pelanggan Umkm Dengan Voucher Belanja. *Jurnal Integrasi*, *11*(2), 140–145.

Hayuningtyas, R. Y. (2019). Penerapan Algoritma Naïve Bayes untuk Rekomendasi Pakaian Wanita. *Jurnal Informatika*, *6*(1), 18–22.

Jain, Y. K. (2012). Upendra,"An Efficient Intrusion Detection based on Decision Tree Classifier Using Feature Reduction,." *International Journal of Scientific and Research Publication*, *2*(1), 1–6.

Kaunang, F. J. (2018). Penerapan algoritma J48 decision tree untuk analisis tingkat kemiskinan di Indonesia. *Cogito Smart Journal*, *4*(2), 348–357.

Pakpahan, N. S. (2021). Implementasi Data Mining Menggunakan Algoritma J48 Dalam Menentukan Pola Itemset Belanja Pembeli (Study Kasus: Swalayan Brastagi Medan). *Journal of Computing and Informatics Research*, *1*(1), 7–13.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46.

Ronaldi, A. A., & Hunaifi, N. (2021). Implementasi Data Mining Untuk Prediksi Penjualan Pestisida Pada CV Mitra Artha Sejati Menggunakan Algoritma Naive Bayes. *EProsiding Teknik Informatika (PROTEKTIF)*, *1*(1), 250–257.

Ruggieri, S. (2002). Efficient C4. 5 [classification algorithm]. *IEEE Transactions on Knowledge and Data Engineering*, *14*(2), 438–444.

Rukmana, I., Rasheda, A., Fathulhuda, F., Cahyadi, M. R., & Fitriyani, F. (2021). Analisis Perbandingan Kinerja Algoritma Naïve Bayes, Decision Tree-J48 dan Lazy-IBK. *Jurnal Media Informatika Budidarma*, *5*(3), 1038–1044.

Sardi, H. Y. (2020). Klasifikasi Tingkat Kelulusan Mahasiswa Elektronika Menggunakan Algoritma Naïve Bayes Classifier (Studi Kasus: Pendidikan Teknik Informatika FT-UNP). *Voteteknika (Vocational Teknik Elektronika Dan Informatika)*, *8*(4), 147–151.

Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), 88–91.

Sinaga, D. M., Yetri, M., & Mahyuni, R. (2022). Analisa Kelayakan Peminjaman Modal Pada Usaha Mikro Kecil Dan Menengah (UMKM) Pada Produk Krasida di PT. Pegadaian Menggunakan Algoritma J48. *Jurnal Cyber Tech*, *1*(7).

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Yu, P. S. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*, 1–37.