p–ISSN: 2723 - 6609 e-ISSN: 2745-5254
Vol. 5, No. 7 July 2024 http://jist.publikasiindonesia.id/

Shallot Production Prediction System Using the C.45 Decision Tree Algorithm

Aghnie Kurnia Fadhila
Universitas Jendral Achmad Yani, Indonesia
Email: Aghniekurnia2@gmail.com
*Correspondence

Keywords: C4.5 algorithm; shallot; decision tree; data mining.

ABSTRACT

This research applies the C4.5 algorithm, which is a machine learning algorithm for classification using decision trees, in a case study for predicting the performance of shallot production. The data used includes attributes such as production yield, land area, and productivity. The C4.5 Decision Tree algorithm is utilized to build an accurate prediction model after going through data cleaning and training processes. This study results in an application that can perform the entire process of initial data processing to data analysis using the aforementioned technique, making it efficient and effective in analyzing large amounts of data to obtain optimal prediction results.

Introduction

Shallots (Allium ascalonicum L) are vegetables that have high commercial value both in terms of economy and nutrition (Damayanti, 2022). The health benefits of shallots have been widely felt, and the related industry is booming, leading to an increase in demand in the domestic market. (Nurhayati, Sibuea, Kusbiantoro, Silaban, & Wanto, 2022) The demand for shallots in Indonesia, both as a vegetable and as a seed, continues to increase by 5% every year as the population grows and consumer interest increases (Baihaqi, Handayani, & Pujianto, 2019).

Shallot production is one of the crucial agricultural sectors for the Indonesian economy. Shallots, as one of the main horticultural commodities, have a significant role in meeting the food needs of the community and making an important contribution to farmers' income (Priyaungga, Aji, Syahroni, Aji, & Saifudin, 2020). However, fluctuations in production caused by various factors such as climate change, pest attacks, and suboptimal cultivation techniques are often a major challenge for farmers and stakeholders (Hana, 2020).

To overcome these problems, a production prediction system is needed that can provide accurate and reliable information. With a prediction system, farmers can better plan their cultivation activities, optimize the use of resources, and minimize the risk of losses due to production that is not by estimates (Kesuma & Kholifah, 2019). Decision Tree C4.5 works by dividing the dataset into several subsets based on the attributes that are most significant in influencing the target variable, in this case, the production of

shallots. Each branch in the decision tree represents a specific condition or decision that leads to the outcome of the prediction.

Based on the trend of increasing demand for shallots in the domestic market, it is important to analyze this problem to predict production that is likely to increase or decrease (Wajhillah & Yulianti, 2017). Previous research (Zulkarnain & Marciano, 2022) has been conducted on the Prediction of Shallot Harvest Production Using a Simple Linear Regression Method. The results show a Mean Squared Error (MSE) of 2073311, a Root Mean Squared Error (RMSE) score of 45533, and an R2 score of 0.98% or 98% accuracy (Ghozali & Wibowo, 2019). Although previous studies predicted increased production, evaluations with different analysis methods could provide a new perspective on accuracy. Therefore, this study will compare the prediction results with previous studies to determine significant differences in accuracy or other factors that need to be considered (Maulana, Martanto, & Ali, 2023).

Research Methods

This research involves several stages in the data mining process. First, data was included as a research subject. Next, preprocessing is carried out to prepare the data. Then, the prediction process is carried out by applying the C4.5 Decision Tree algorithm for testing. Finally, the results are analyzed and reported in the form of publications. The diagram in Figure 1 shows the overall stages of the research.

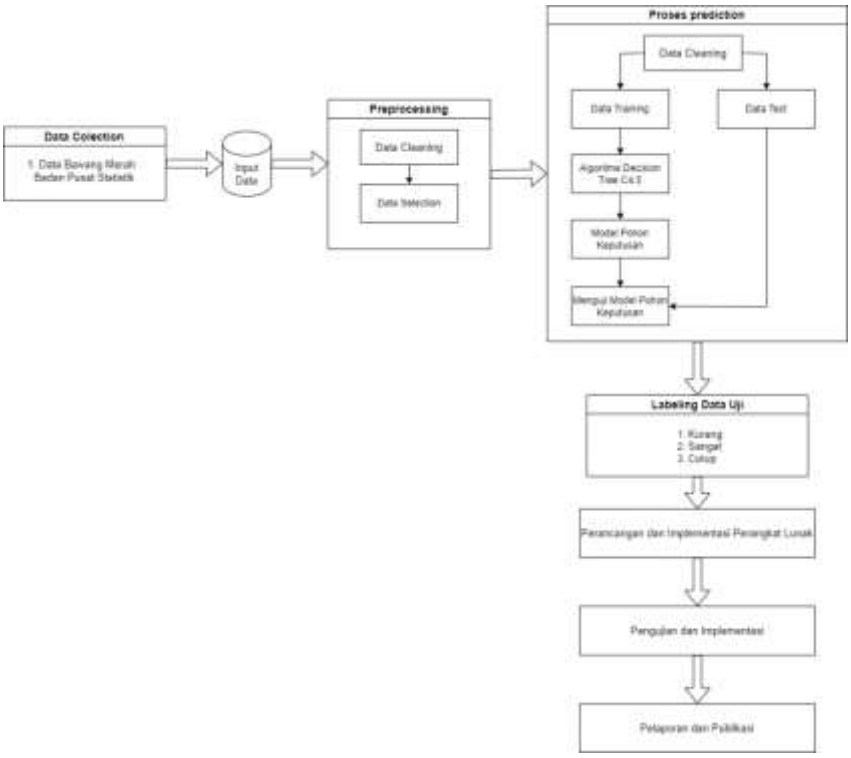

Figure 1 Description of the research flow

Data Collection

In this study, data was obtained from BPS (Central Statistics Agency).

Pre-Processing

At this stage, the existing data is transformed into a suitable format to be used as an object in the research. There are two stages of pre-processing carried out in this study, namely Data Cleaning and Data Selection.

1. Data Cleaning

Data Cleaning is an important first step to maintaining the quality of data from non-compliant data sources. This step involves correcting and removing inappropriate data to make the information obtained more relevant. Incomplete, redundant, or inconsistent data will be eliminated to make the analysis process easier.

2. Data Selection

The second stage after Data Cleaning is Data Selection, where data that already has complete information is selected according to the needs of the necessary information. In this stage, the data is selected from the overall 8 required attributes.

Prediction Process using Decision Tree C4.5

At this stage, data derived from the latest BPS and assessment sources that have been processed into a data set during the previous pre-processing stage will be examined. The process and steps of this method are as follows:

Data Training

Record data from BPS and assessments that have gone through the pre-processing stage and combined into one dataset will be used as training data to train the C4.5 decision tree algorithm. It aims to produce an accurate model to make predictions.

Algoritma Decision Tree C4.5

In this stage, the rules of the C4.5 Decision Tree algorithm are used to form a decision tree model based on training data.

Decision Tree Model Creation

The decision tree model is made based on the classification results using the C4.5 Decision Tree algorithm, by the predetermined tree formation rules.

Data Testing

The data is used to test the performance of a previously trained algorithm when faced with new data that has never been seen before. The test data will be classified using the tree model that has been created.

Decision Tree Model Performance Testing

Testing is carried out on the decision tree model by entering test data into the tree model that has been created, to evaluate the performance of the algorithm that has been trained.

Labelling Data Uji

Once the test data has been classified using the prediction model, the next step is to label the test data results between the categories of "less", "very", and "adequate".

Software Design and Development

This stage of software design and development is carried out by applying a prediction tree model to the software creation process.

Testing and Evaluation

The evaluation is carried out by calculating the level of accuracy and assessing the extent to which the system has succeeded in generating correct information based on the model that has been created.

Results and Discussion

Analysis of Prediction of Shallot Production Using Decision Tree C4.5 Algorithm

Analysis using the Decision Tree C4.5 algorithm was carried out to predict the yield of shallots based on historical data and factors that affect it such as weather, soil type, and cultivation techniques. The methodology includes data collection, data cleaning, sharing of training and test data, C4.5 model development, model evaluation, optimization, and interpretation of results. This analysis aims to identify the main factors that affect shallot production so that it can help farmers and stakeholders in making decisions related to agricultural practices to increase production efficiently.

Data Collection

In this study, secondary data is taken from various websites to find data that is the focus of the research. The purpose of these efforts is to ensure that the data used is valid and supports the smooth and successful conduct of the research. After various search efforts, finally, the dataset used consisted of 243 data records taken from the Central Statistics Agency (BPS). By using this dataset, it is hoped that research can be carried out well and produce accurate and relevant results.

Shallot Dataset

The following is a view of the dataset taken from the Central Statistics Agency with the number of 8 attributes which can be seen in Table 1 of the Shallot Daset.

A

Id	Provincial Code	Name Provinsi	District Code Kota	Broad	Onion Production Red	Year	Parameter
1	32	West Java	3	20	10	2013	Less
2	32	West Java	2	24	285	2013	Very
3	32	West Java	2	31	183	2013	Enough
4	32	West Java	1	2915	31682	2013	Very
5	32	West Java	3	1967	19728	2013	Very
6	32	West Java	1	13	90	2013	Enough
7	32	West Java	2	0	0	2013	Less
8	32	West Java	3	237	2218	2013	Enough
9	32	West Java	3	3658	36449	2013	Enough
10	32	West Java	2	2150	23683	2013	Very
11	32	West Java	1	27	204	2013	Enough
12	32	West Java	1	197	950	2013	Less
13	32	West Java	2	0	0	2013	Less
14	32	West Java	3	0	0	2013	Less
15	32	West Java	1	0	0	2013	Less
16	32	West Java	2	3	21	2013	Enough
17	32	West Java	2	15	83	2013	Enough
18	32	West Java	3	0	0	2013	Less
19	32	West Java	3	0	0	2013	Less
20	32	West Java	1	0	0	2013	Less

Data Cleaning

In the data cleaning stage, identification is an important step to ensure the quality of the data used in the analysis. Therefore, in this study, the data-cleaning process is carried out by identifying empty values. In the Shallot data obtained at the Central Statistics Agency, the dataset has been checked that there is a missing value in the data. As well as the removal of attributes on the shallot dataset. By carrying out this data-cleaning process, the data used in the analysis becomes more accurate and reliable. In

addition, the proper data cleaning process also ensures that the results of the analysis produced are of good quality and can be accounted for.

City district code	Broad	Shallot production	Parameter
3	20	10	less
2	24	285	very
2	31	183	enough
1	2915	31682	very
3	1967	19728	very
1	13	90	enough
3	237	2218	enough
3	3658	36449	enough
2	2150	23683	very
1	27	204	enough
1	197	950	less
2	3	21	enough
2	15	83	enough

Transformation Data

In this data analysis process, Data Transformation is carried out to facilitate data processing and analysis more effectively. Data Transformation is carried out by changing data variables into numerical data forms.

The dataset used in this study consists of 8 attributes which include ID, province code, province name, city district code, Area, shallot production, year, and Parameters. Of the datasets, 6 attributes have numerical data types, while the other two attributes have nominal data types. To overcome this, this study transforms data by converting two nominal attributes into numerical ones.

The nama_provinsi attribute and the Parameter attribute which originally consisted of less, sufficient, and very were also changed to numeric by replacing less with a value of 0, sufficient with a value of 1, and very with a value of 2. This altered data is then used in research to facilitate analysis and testing.

Table 3
Transformation Data

City district code	Broad	Shallot production	Parameter
3	20	10	0
2	24	285	2
2	31	183	1
1	2915	31682	2
3	1967	19728	2
1	13	90	1
3	237	2218	1

3	3658	36449	1
2	2150	23683	2
1	27	204	1
1	197	950	0
2	3	21	1
2	15	83	1

In Table 3. Data Transformation is an example of a dataset that has been pre-processed with Data Cleaning and Data Transformation, the dataset is ready to be carried out for the Prediction process with the C4.5 Model.

Prediction Process with Model C4.5

Of all the data generated through preprocessing, this study divided the data into two parts, namely training and testing data, where 171 or 70% was partitioned for training data and 72 or 30% for testing data.

Table 4
Proportion of each class

Status	Code Kabupaten_kota		Broad		Shallot production		Parameter	
	Sum	Proposition	Sum	Proposition	Sum	Proposition	Sum	Proposition
Less	90	0.37	73	0.30	56	0.23	61	0.25
Enough	66	0.27	85	0.35	134	0.55	61	0.25
Very	88	0.36	85	0.35	53	0.22	121	0.50
Total(s)	**243**	**1.00**	**243**	**1.00**	**243**	**1.00**	**243**	**1.00**

$pCapEquals = \sum_{i=1}^{n} -pi\ x\ \log 2\ (pi)$(Parameter 0,1,2) = $\left(-\left(\frac{61}{243}\right).\log_2\left(\frac{61}{243}\right)\right) + \left(-\left(\frac{61}{243}\right).\log_2\left(\frac{61}{243}\right)\right) + \left(-\left(\frac{121}{243}\right).\log_2\left(\frac{121}{243}\right)\right) = 1.557$

1. Calculating Information Gain

This calculation is intended for each attribute used with S as a class set of less, sufficient, and very much. Classes are less with code 0, classes are sufficient with ode 1 and classes are very with code 2 s reobtain. Cap$CapEntropy(Si) = \sum_{i=1}^{n} -pi\ x\ \log 2\ (pi)$

Table 5
Entropy (City District Code)

S	City code	district	Less	Enough	Very
S1	0		23	15	31
S2	1		37	32	29

S3	2	30	18	28

$Entropy\ (0) = \left(-\left(\frac{23}{69}\right).\log_2\left(\frac{23}{69}\right)\right) + \left(-\left(\frac{15}{69}\right).\log_2\left(\frac{31}{69}\right)\right) + \left(-\left(\frac{28}{69}\right).\log_2\left(\frac{28}{69}\right)\right) =$
$1.537 Entrop\ (1) = \left(-\left(\frac{37}{98}\right).\log_2\left(\frac{37}{98}\right)\right) + \left(-\left(\frac{32}{98}\right).\log_2\left(\frac{32}{98}\right)\right) +$
$\left(-\left(\frac{29}{98}\right).\log_2\left(\frac{29}{98}\right)\right) = 1.576 Entropy\ (2) = \left(-\left(\frac{30}{76}\right).\log_2\left(\frac{30}{76}\right)\right) +$
$\left(-\left(\frac{18}{76}\right).\log_2\left(\frac{18}{76}\right)\right) + \left(-\left(\frac{28}{76}\right).\log_2\left(\frac{28}{76}\right)\right) = 1.552 GA(S) -$
$\sum_{i=1}^{n} \frac{|Si|}{|S|}\ x\ Entropy\ (Si)\ GAIN(S, City\ District\ Code) = 1{,}557 -$
$\left(\frac{23+15+28}{69}\ x\ 1.537\right) + \left(\frac{37+32+29}{98}\ x\ 1.576\right) + \left(\frac{30+18+28}{76}\ x\ 1.552\right) =$
-2.319 Calculating Retro Gain

$$SplitInfo(S,A) = \sum_{i=1}^{n} \frac{|Si|}{|S|} \log 2 \frac{|Si|}{|S|} SplitInfo(S,A)$$
$$= \left(\left(\frac{69}{243}\right).\log_2\left(\frac{69}{243}\right)\right) - \left(\left(\frac{98}{243}\right).\log_2\left(\frac{98}{243}\right)\right)$$
$$- \left(\left(\frac{76}{243}\right).\log_2\left(\frac{76}{243}\right)\right) = 1.557\ \ Gainatio(S,A)$$
$$= \frac{Gain(S,A)}{SplitInfo(S,A)} GainRatio(S,A) = \frac{-2319}{1.557} = -1.471$$

From the calculation process, the result of the information gain of the kode_kabupaten_kota attribute was -1,471. The calculation of information acquisition is calculated for all attributes, so the results are shown in Table 6.

Table 6
Attribute calculation results

No.	Attribute	Info Gain	Split Info	Gain Ratio
1.	Code-kabupaten_kota	-2.319	1.557	-1.471
2.	Broad	-0.012	0.981	-0.012
3.	Produksi_bawang_merah	- 1.739	1.482	-1.735

Based on the construction of the model that has been made, a tree model is created that depicts the relationship between attributes.

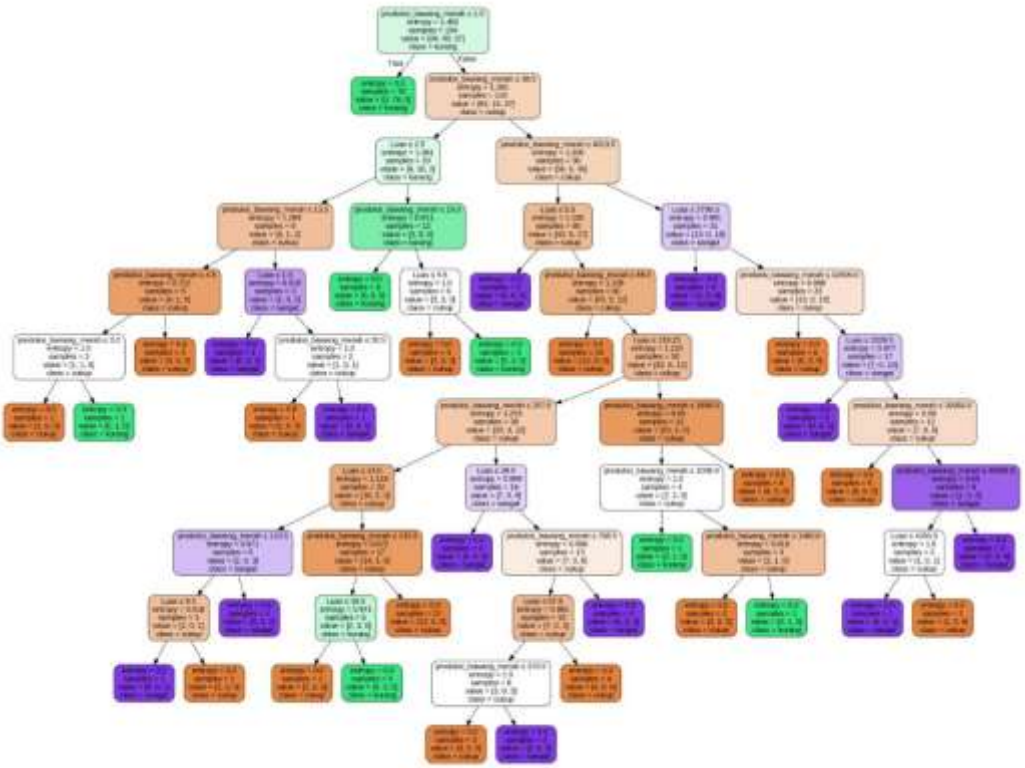

Figure 2
Tree model

The results of the decision tree algorithm for the classification of shallot production show that shallot production can be predicted using the parameters of land area and production amount. This decision tree classifies production into three categories: less, adequate, and very. In this tree, the main variable that affects the classification is the amount of shallot production, with the initial node dividing the data based on whether the shallot production is less than or equal to 1.5. This node has a high entropy, indicating great uncertainty in the early stages. The division continues with additional parameters such as land area, which helps to clarify the classification. For example, shallot production below 1.5 is generally classified as "poor", while production above 1.5 but below 4013.5 is generally classified as "adequate". Overall, this model shows that the increase in the amount of shallot production and the variation in land area significantly affect the classification of shallot production levels.

Conclusion

The above problem about the data of shallot prediction results can be solved by the C4.5 algorithm method using classification rules to develop a prediction model that can predict the quality of shallots. And identify important attributes that affect the quality of shallots.

Bibliography

Baihaqi, Dimas Imam, Handayani, Anik Nur, & Pujianto, Utomo. (2019). Perbandingan metode Naive Bayes dan C4. 5 untuk memprediksi mortalitas pada peternakan ayam broiler. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *10*(1), 383–390.

Damayanti, Desi. (2022). Implementasi Algoritma C4. 5 Prediksi Produksi Komoditas Tanaman Perkebunan Berdasarkan Luas Lahan. *Tin: Terapan Informatika Nusantara*, *2*(10), 571–579.

Ghozali, Muhammad Rizal, & Wibowo, Rudi. (2019). Analisis Risiko Produksi Usahatani Bawang Merah di Desa Petak Kecamatan Bagor Kabupaten Nganjuk. *Jurnal Ekonomi Pertanian Dan Agribisnis*, *3*(2), 294–310.

Hana, Fida Maisa. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4. 5. *Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan)*, *4*(1), 32–39.

Kesuma, Chandra, & Kholifah, Desiana Nur. (2019). Sistem Informasi Akademik Berbasis Web Pada Lkp Rejeki Cilacap. *EVOLUSI: Jurnal Sains Dan Manajemen*, *7*(1).

Maulana, Alfin, Martanto, Martanto, & Ali, Irfan. (2023). Prediksi Hasil Produksi Panen Bawang Merah Menggunakan Metode Regresi Linier Sederhana. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(4), 2884–2888.

Nurhayati, Nurhayati, Sibuea, Mhd Buhari, Kusbiantoro, Dedi, Silaban, Martina, & Wanto, Anjar. (2022). Implementasi Algoritma Resilient untuk Prediksi Potensi Produksi Bawang Merah di Indonesia. *Building of Informatics, Technology and Science (BITS)*, *4*(2), 1051–1060. https://doi.org/10.47065/bits.v4i2.2269

Priyaungga, Bayu Aji, Aji, Dwi Bayu, Syahroni, Mukron, Aji, Nurul Tri Sukma, & Saifudin, Aries. (2020). Pengujian Black Boxpada Aplikasi Perpustakaan Menggunakan Teknik Equivalence Partitions. *Jurnal Teknologi Sistem Informasi Dan Aplikasi ISSN*, *2654*, 3788.

Wajhillah, Rusda, & Yulianti, Ita. (2017). Penerapan algoritma c4. 5 untuk prediksi penggunaan jenis kontrasepsi berbasis web. *Klik-Kumpul. J. Ilmu Komput*, *4*(2), 160.

Zulkarnain, Muhammad, & Marsisno, Waris. (2022). Penerapan Pembelajaran Mesin Untuk Estimasi Luas Lahan Bawang Merah Berdasarkan Data Citra Satelit Resolusi Menengah. *Seminar Nasional Official Statistics*, *2022*(1), 1005–1016. https://doi.org/10.34123/semnasoffstat.v2022i1.1307