DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE PRODUCTION USING CRISP-DM METHOD

by Ali Khumaidi

Submission date: 10-Aug-2022 09:39PM (UTC+0700)
Submission ID: 1881026989
File name: ICTING_THE_AMOUNT_OF_COFFEE_PRODUCTION_USING_CRISP-DM_METHOD.pdf (933.03K)
Word count: 4664
Character count: 25689

DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE PRODUCTION USING CRISP-DM METHOD

Ali Khumaidi

Informatics Engineering
Universitas Krisnadwipayana, Jakarta, Indonesia
www.unkris.ac.id
alikhumaidi@unkris.ac.id

Abstract—*The production of coffee plantations has become the leading plantation commodity with the export value of the fourth rank after oil palm, rubber and coconut. The number of coffee needs for export every year always increases, therefore it is necessary to predict the yield of coffee plants to estimate planting and anticipation that will be done so as to achieve the target. Coffee plant productivity is influenced by internal and external factors, namely the quality of the plant itself, soil, altitude and climate. The method used in this study is the CRISP-DM method and multiple linear regression algorithm to predict the amount of coffee production and determine the relationship between the variables. The steps taken are business understanding, data understanding, data preparation, modeling and evaluation. The data set that is used as many as 170 data after going through the data preparation stage produced 150 data with 5 attributes in the table. With calculations using tools, the coefficient of determination is 91.96%. That the variation in the value of the production of coffee plants is influenced by independent variables, namely the area of plantations, rainfall, air pressure and solar radiation by 91.96% and 8.04% influenced by other variables not measured in this study. The results of the evaluation and validation of predictions produce good accuracy with an RMSE value of 0.3477.*

Keyword: data mining, coffee plants, crisp-dm, multiple linear progression.

Intisari—*Produksi hasil perkebunan kopi menjadi komoditas perkebunan unggulan dengan nilai ekspor urutan ke empat setelah komoditas kelapa sawit, karet, dan kelapa. Jumlah kebutuhan kopi untuk ekspor tiap tahun selalu mengalami peningkatan oleh karena itu diperlukan prediksi hasil produksi tanaman kopi untuk memperkirakan penanaman dan antisipasi yang akan dilakukan sehingga dapat mencapai target. Produktivitas tanaman kopi dipengaruhi oleh faktor internal dan eksternal yaitu kualitas tanaman itu sendiri, tanah, ketinggian dan ilkim. Metode yang digunakan dalam penelitian ini adalah metode CRISP-DM dan algoritma multiple linear regression untuk prediksi jumlah produksi kopi dan mengetahui hubungan antar variabelnya. Tahapan yang dilakukan adalah business understanding, data understanding, data preparation, modeling dan evaluation. Data set yang digunakan sebanyak 170 data setelah melalui tahap data preparation dihasilkan 150 data dengan 5 atribut pada tabel. Dengan perhitungan menggunakan tools dihasilkan koefisien determinasi sebesar 91,96%. Bahwa variasi nilai jumlah produksi tanaman kopi dipengaruhi oleh variabel independen yaitu luas areal perkebunan, curah hujan, tekanan udara dan penyinaran matahari sebesar 91,96% dan 8,04% dipengaruhi oleh variabel lain yang tidak diukur pada penelitian ini. Hasilevaluasi dan validasi prediksi menghasilkan nilai akurasi yang baik dengan nilai RMSE sebesar 0,3477.*

Kata kunci: *data mining, tanaman kopi, crisp-dm, multiple linier progression.*

INTRODUCTION

The production of Indonesian coffee plantations ranks fourth in the world after Brazil, Vietnam and Colombia. As for the largest export commodity in Indonesia, coffee is ranked fourth with a total trade reached 1.19 billion US $ in 2017 so that coffee becomes one of the leading plantation commodities after the commodities of oil palm, rubber, and coconut. Coffee commodity by the Directorate General of Plantations is placed in the strategic plan as the main target and priority agenda for improving the agro-industry sector, namely increasing the amount of production and exports and developing agro-industry in the village (Direktorat Jenderal Perkebunan, 2014). To support this, control and monitoring need to be done so that coffee productivity tends to increase.

Indonesia ranks second only to Brazil in the area of coffee plantations. The area of coffee plantations tends to increase 1.6% per year during the period 1980 to 2018. However, the data for the last 10 years the area of coffee plantations has decreased 0.05% per year with an area of 1.27

million hectares in 2009 to 1.26 million hectares in 2009 2018 (Direktorat Jenderal Perkebunan, 2014). Although there was a reduction in the area of coffee plantations, in terms of productivity there was an increasing tendency with growth of 1.14% per year. During the period 1980 to 2017 the number of coffee exports was quite volatile with a tendency to increase 3.93% per year. The period of 2012 to 2016 export volume experienced slowing growth, even in 2014 there was a decline in the value of coffee exports up to 11.47% due to the decline in coffee plantation production (Badan Pusat Statistik, 2018).

Climate change can cause negative impacts on plants, so that it can affect its productivity, including coffee(Iscaro, 2014). In Indonesia, climate factors are part of natural phenomena whose changes are influenced by nature and human intervention. The development of coffee plants is influenced by temperature which is able to control root growth, respiration, absorption of nutrients and water, photosynthesis and reaction speed (Lenisastri, 2000). Enzyme systems can function well and are stable at optimum temperatures and at cold temperatures the system is stable but reduces the function so that the enzyme can be damaged (Setiawan, 2009). In addition to temperature, rainfall and humidity factors influence the growth of coffee plants. High levels of rainfall, flowering process can be disrupted and the level of humidity affects generative and vegetative growth (Ashari, 2004). Climate affects coffee productivity, temperature is directly related but humidity and rainfall are not directly related. Good coffee cultivation management techniques can be used to anticipate climate change (Prasetyo et al., 2017). The presence of diseases and pest attacks can be caused by climate change, which previously was only affected by low altitude (Widayat, Anhar, & Baihaqi, 2015).

The condition of the declining area of coffee plantations, increasing export needs and volatile climate conditions as well as targets and priorities for increasing coffee production, it is necessary to predict coffee productivity so that the predicted results can be input for making the best policy. Prediction is a pattern that can be identified by data mining (Maulida, 2018). Data mining is a very large data extraction or extraction process and can be used to assist in making important decisions (Soni & Ganatra, 2012). In previous studies, data mining has been widely used for predictions, research for book sales predictions using data mining in PT. Niaga Swadaya (Kamal, Hendro, & Ilyas, 2017), research for predicting the number of student registrations per semester using linear regression at Ichsan University Gorontalo (Bengnga & Ishak, 2018), research related to the application of data mining to determine the estimated productivity of sugarcane by using an algorithm linear regression in Rembang Regency (Warih & Rahayu, 2015).

In using data mining, it is necessary to use the right method and the appropriate algorithm. In the prediction of coffee plant productivity there are several factors involved, namely the area of plantation crops, the amount of production, rainfall, solar radiation and air pressure. Multiple linear regression algorithm is a regression analysis that explains the relationship between dependent variables with factors that affect more than one. This multiple linear regression algorithm has been used in several studies with a fairly high accuracy rate of up to 95% (Kamal et al., 2017). Therefore, the multiple linear regression algorithm is expected to obtain the results of the productivity of coffee plants as a consideration in making further policies.

Analysis of coffee plant productivity needs to be done in depth in order to obtain hidden patterns and the discovery of knowledge related to production predictions. The appropriate method for conducting this research is the Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a method that provides standard processes in data mining for solving problems in business. CRISP-DM is easier to apply because each stage or phase is clearly defined and structured and has a complete and well-documented data mining methodology.

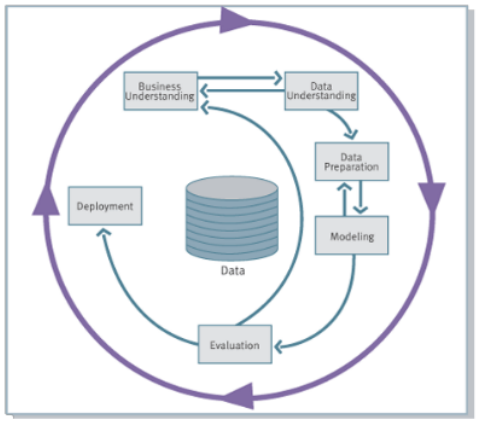
MATERIALS AND METHODS

The stages in this study are adjusted to the phase of the CRISP-DM method. CRISP-DM is a data mining standard that contains a framework for data mining tasks. In the data mining process based on CRISP-DM there are 6 phases (Colin Shearer, 2000) according to Figure 1.

Business Understanding is the phase of understanding the substance of data mining activities that will be carried out. Its activities include: determining business goals or objectives, understanding business situations, and determining data mining goals.

Data Understanding is the initial data collection phase. Its activities include: studying and describing data, exploiting data and identifying problems related to data quality, and detecting an interesting subset of data as an initial hypothesis.

Data Preparation is an activity carried out to prepare data. Its activities include data selection, data building, data integration and data cleaning.



Source: (Colin Shearer, 2000)

Gambar 1. Model CRISP-DM

Modeling is the phase of determining data mining techniques. Data mining techniques use multiple linear regression algorithm. Regression is a model development technique that can be used for prediction, looking for the relationship between the dependent variable and the independent variable. Multiple linear regression is a regression analysis that links the dependent variable with more than one independent variable that affects it (Ngumar, 2008). Multiple linear regression will work optimally if the input used is numeric. The formula used for multiple linear regression is expressed in equation (1).

$$Y = a + b1X1 + b2X2 + \cdots + bnXn \quad \text{..... (1)}$$

Information:

Y	= dependent variable
X1, X2 .... Xn	= independent variable
a	= constant
b1, b2 .... bn	= regression coefficient

According to equation (1), to calculate the amount of coffee crop production using multiple linear regression algorithm, equation (2) is used.

$$Y = a + b1X1 + b2X2 + b3X3 + b4X4 \quad \text{..... (2)}$$

Information:

Y = Number of Production
X1 = Area of plantation area
b1 = Variable coefficient of plantation area
X2 = Rainfall
b2 = coefficient of rainfall variable
X3 = Barometric pressure
b3 = coefficient of variable air pressure
X4 = Solar Light
b4 = coefficient of solar irradiation variables

The first stage is the formation of a model, by finding the values of a, b1, b2, b3, and b4 using the least square with the general equation referring to equation (3).

$$\begin{aligned}
\Sigma Y &= an + b1\Sigma X1 + b2\Sigma X2 + b3\Sigma X3 + b4\Sigma X4 \\
\Sigma X1Y &= a\Sigma X1 + b1\Sigma(X1)2 + b2\Sigma(X1X2 + b3\Sigma(X1X3) \\
&\quad + b4\Sigma(X1X4) \\
\Sigma X2Y &= a\Sigma X2 + b1\Sigma(X1X2) + b2\Sigma(X2)2 \\
&\quad + b3\Sigma(X2X3) + b4\Sigma(X2X4) \\
\Sigma X3Y &= a\Sigma X3 + b1\Sigma(X1X3) + b2\Sigma(X2X3) + b3\Sigma(X3)2 \\
&\quad + b4\Sigma(X3X4) \\
\Sigma X4Y &= a\Sigma X4 + b1\Sigma(X1X4) + b2\Sigma(X2X4) + \\
&b3\Sigma(X3X4) + b4\Sigma(X4)2 \quad \text{..... (3)}
\end{aligned}$$

Next the acquisition of inversion results is obtained, then the determinant matrix multiplication is carried out with ΣY, $\Sigma X1Y$, $\Sigma X2Y$, $\Sigma X3Y$, $\Sigma X4Y$. Then the determinants of matrices A, A0, A1, A2, A3, and A4 are calculated. The following equation in calculation (4):

$$\begin{vmatrix} \Sigma Y \\ \Sigma X1Y \\ \Sigma X2Y \\ \Sigma X3Y \\ \Sigma X4Y \end{vmatrix} = \begin{vmatrix} N & \Sigma X1 & \Sigma X2 & \Sigma X3 & \Sigma X4 \\ \Sigma X1 & \Sigma X1X1 & \Sigma X1X2 & \Sigma X1X3 & \Sigma X1X4 \\ \Sigma X2 & \Sigma X2X1 & \Sigma X2X2 & \Sigma X2X3 & \Sigma X2X4 \\ \Sigma X3 & \Sigma X3X1 & \Sigma X3X2 & \Sigma X3X3 & \Sigma X3X4 \\ \Sigma X4 & \Sigma X4X1 & \Sigma X4X2 & \Sigma X4X3 & \Sigma X4X4 \end{vmatrix} x \begin{vmatrix} a \\ b1 \\ b2 \\ b3 \\ b4 \end{vmatrix} \text{.... (4)}$$

Information:

N = amount of data
Σ Y = the sum of the variable production quantities
ΣX1 = number of variable area
ΣX2 = number of rainfall variables
ΣX3 = amount of variable air pressure
ΣX4 = number of solar irradiation variables

Next, the results of the calculation of the determinant matrix are used to find the values of a, b1, b2, b3, and b4 by referring to equation (5).

$$\begin{aligned}
a &= (Det\,(A0)) \,/\, (Det A) \\
b1 &= (Det\,(A1)) \,/\, (Det A) \\
b2 &= (Det\,(A2)) \,/\, (Det\,A) \quad \text{..... (5)} \\
b3 &= (Det\,(A3)) \,/\, (Det A) \\
b4 &= (Det\,(A4)) \,/\, (Det A)
\end{aligned}$$

Next, a partial correlation test will be calculated to see the degree of interrelation of the independent variable with the dependent variable. The related values are rx1Y, rX2Y, rX3Y, rX4Y, rX1X2, rX1X3, rX1X4, rX2X3, rX2X4, and rX3X4. Calculation of correlation values uses equation (6).

$$\begin{aligned}
\Sigma X1Y &= \Sigma X1Y - ((\Sigma X1).(\Sigma Y))/n \\
\Sigma X2Y &= \Sigma X2Y - ((\Sigma X2).(\Sigma Y))/n \\
\Sigma X3Y &= \Sigma X3Y - ((\Sigma X3).(\Sigma Y))/n \\
\Sigma X4Y &= \Sigma X4Y - ((\Sigma X4).(\Sigma Y))/n \\
\Sigma X1X2 &= \Sigma X1X2 - ((\Sigma X1).(\Sigma X2))/n \quad \text{..... (6)} \\
\Sigma X1X3 &= \Sigma X1X3 - ((\Sigma X1).(\Sigma X3))/n \\
\Sigma X1X4 &= \Sigma X1X4 - ((\Sigma X1).(\Sigma X4))/n \\
\Sigma X2X3 &= \Sigma X2X4 - ((\Sigma X2).(\Sigma X3))/n \\
\Sigma X2X4 &= \Sigma X2X4 - ((\Sigma X2).(\Sigma X4))/n
\end{aligned}$$

Calculations for finding R2 refer to (7).

$$R2 = 1 - \frac{SS\ Error}{nSS\ Total} = 1 - \frac{\Sigma(yi-yi)2}{\Sigma(yi-yi)2} \quad \text{..... (7)}$$

Information:
y = forecast i-response
y = average
yi = observation of the i-response

For the coefficient of determination the calculation is according to equation (8).
Information:
y = forecast i-response
y = average
yi = observation of the i-response

For the coefficient of determination the calculation is according to equation (8).

$$Kd = R2\ x\ 100\% \quad \text{..... (8)}$$

Information:
R2 = correlation coefficient value.
Kd = large coefficient of determination

The accuracy of a prediction is an important aspect and objective of a prediction, so that the error rate is sought as small as possible. There are 3 models for measuring errors in concluding historical errors, namely mean squared error, mean absolute deviation, and mean absolute percent error (William J. Stevenson, 2014). Testing the accuracy of predictions in this study by measuring the value of root mean squared error (RMSE) according to equation (9). RMSE is a method for evaluating prediction results by measuring the accuracy of the model (Chai & Draxler, 2014). The lower RMSE value concludes that the predicted results approach the true value (Choirunisa, 2019). The greater the RMSE value indicates the worse the accuracy level.

$$RMSE = \frac{\sum_i^n = \sqrt{(yi-yi)2}}{n} \quad \text{..... (9)}$$

Evaluation is the phase of interpreting the results of data mining produced. Evaluation is carried out in depth aiming to obtain a model that is in accordance with the objectives.

Deployment is the phase of compiling a report of the knowledge gained at the evaluation stage.

In CRISP-DM there are three stages that are worked on, the first is the data collection stage, the second is the data understanding and business stage and the third is the modeling and evaluation stage.

1. Data Collection Stage
   At this stage, conducting a literature study and collecting coffee plant data includes: plantation area, total production, rainfall, air pressure and solar radiation. The data is collected from BPS.
2. Business Understanding and Data Understanding
   In this stage consists of 3 stages, namely business understanding, data understanding, and data preparation.
3. Modeling and Evaluation stage
   At this stage consists of 2 stages, namely modeling and evaluation.

RESULTS AND DISCUSSIONS

Coffee plant production prediction models include five phases which are the stages of business understanding and data as well as the modeling and valuation stages, as follows:

Business Understanding

This stage refers to the prediction of coffee plant production. At this stage an understanding of the background and objectives of the business processes related to coffee plants is needed, including:

1. Determine Business Goals
   The business objective of doing research is to recognize coffee plant production to determine the prediction of coffee plant production so that policies and actions can be taken if the production predictions are not appropriate and to know the relationship of other factors that affect the amount of production
2. Assess the situation
   The process of developing coffee plants is influenced by internal and external factors. Internal factors, namely the quality of coffee plants and external factors, namely climate, temperature, air pressure, humidity, rainfall, altitude, plantation area, solar radiation and others.
3. Determine the Purpose of Data Mining
   The purpose of data mining or the purpose of this study is to explore knowledge about the patterns of influence of external factors on the production of coffee plants for prediction of the amount of coffee production.

Data Understanding

The understanding of the data stage begins with the collection of preliminary data and the results of activities in order to identify data problems, to determine the first insight into the data or detect interesting subsets to form hypotheses for hidden information.

1. Initial Data Collection
   Data collection is done by taking data from BPS 34 provinces, including:
   - Data on coffee plantation area by province in 2011-2018
   - Data on coffee plantation production by province in 2008-2018
   - Rainfall data for 2000-2015
   - Air pressure data for 2003-2015
   - Data on solar radiation from 2003-2015
2. Describe Data
   This stage is for analysis to understand the data obtained from the initial data collection results.
   - Data on the area of coffee plantations in units of thousands of hectares and consists of data from 34 provinces plus total totals (Indonesia).
   - Production data of coffee plantations in thousand tons and consists of data from 34 provinces plus total totals (Indonesia).
   - Rainfall data in millimeters (mm) taken from 34 provincial BMKG stations including data on the number of rainy days in a year.
   - Air pressure data in units of millibars (mb) taken from BMKG stations in each province are 34.
   - Solar irradiance data in annual percentages based on 34 provincial BMKG station data.
3. Exploring DataThe data exploration process is carried out on four tables, namely:
   - Data on coffee plantation area consisting of attributes of the province, year and amount.
   - Coffee plantation production data consisting of attributes of the province, year and amount.
   - Rainfall data consisting of provincial attributes, BMKG stations, year, amount of rainfall and number of rainy days.
   - Data on air pressure and solar radiation consisting of attributes of the province, BMKG station, year, air pressure and solar radiation.
4. Verifying Data Quality
   The process of verifying data quality by looking at the structure of the table, the data contained in the table then integrates between tables. Since the data collected with different number of years is then sliced between the four data tables, the 2011-2015 period is used. The total data used is 170 data.

Data Preparation

The data preparation stage is the preparation of the dataset, the data used for modeling. This stage includes the following activities:

1. Select and build data
   At this stage related tables are chosen to simplify the process of selecting data. There are 4 tables involved, a table of production quantities, a table of crop area, a table of rainfall, a table of air pressure and solar radiation. Each table contains 170 data with variations in the attributes in the tables below.

Table 1. Total of Coffee Production

Year	Province	Total of Production (Thousand of Tons)
2011	Aceh	52.3
2011	Sumatera Utara	56.8
2011	Sumatera Barat	30.8
2011	Riau	1.9
...	...	...
2015	Papua	2

Source: (Badan Pusat Statistik, 2018)

Table 2. Plantation Area (Thousand Hectares)

Year	Province	Plantation Area
2011	Aceh	120.7
2011	Sumatera Utara	80.6
2011	Sumatera Barat	40.3
2011	Riau	4.7
...	...	...
2015	Papua	10

Source: (Badan Pusat Statistik, 2018)

Table 3. Total Rainfall (millimeters)

Year	Province	BMKG Station	Rainfall
2011	Aceh	Sultan Iskandar Muda	1268.00
2011	Sumatera Utara	Kualanamu	2042.00
2011	Sumatera Barat	Sicincin	
2011	Riau	Sultan Syarif Kasim II	2405.00
...	...	...	...
2015	Papua	Angkasapura	1265.90

Source: (Badan Pusat Statistik, 2018)

Table 4. Air Pressure and Solar Light

Year	Province	BMKG Station	Air pressure	Solar radiation
2011	Aceh	Sultan Iskandar Muda	1009.40	52.20
2011	Sumatera Utara	Kualanamu		44.40
2011	Sumatera Barat	Sicincin	990.80	32.80
2011	Riau	Sultan Syarif Kasim II	1008.70	42.30
...	...	...		...
2015	Papua	Angkasapura	1011.10	64.47

Source: (Badan Pusat Statistik, 2018)

2. Integrating Data
This stage combines four tables into one table into a new data set. The results of the transformation in the previous table are integrated into table 5.

Table 5. Combined tables

Year	Province	BMKG Station	Production	Plantation	Rainfall	Air pressure	Solar radiation
2011	Aceh	Sultan Iskandar Muda	52.3	120.7	1268.00	1009.40	52.20
2011	Sumatera Utara	Kualanamu	56.8	80.6	2042.00		44.40
2011	Sumatera Barat	Sicincin	30.8	40.3		990.80	32.80
2011	Riau	Sultan Syarif Kasim II	1.9	4.7	2405.00	1008.70	42.30
...	...	...	...	...	...		...
2015	Papua	Angkasapura	2	10	1265.90	1011.10	64.47

Sumber: (Badan Pusat Statistik, 2018)

3. Clearing Data
This process removes some attributes or variables so that the final results for the variables will be processed in table 6. There are only 5 variables used for the modeling stage.

Table 6. Variables for Modeling

Production	Plantation Areal	Rainfall	Air pressure	Solar radiation
52.3	120.7	1268.00	1009.40	52.20
56.8	80.6	2042.00		44.40
30.8	40.3		990.80	32.80
1.9	4.7	2405.00	1008.70	42.30
...	...	...		...
2	10	1265.90	1011.10	64.47

Source: (Khumaidi, 2019)

In addition to adjusting the variables, cleaning of the empty data is also done, so that initially there were 170 data to 150 data.

Modelling

Based on the data set prepared by doing calculations with several equations, the results obtained are a, b1, b2, b3, and b4, as shown in table 7.

Table 7. Calculation Results

a	*b1*	*b2*	*b3*	*b4*
19266.64013	0.57369	0.514468	-15.385717	-111.240876

Source: (Khumaidi, 2019)

Based on the results obtained for the coefficients a, b1, b2, b3, and b4, the multiple linear regression model is as follows.

$$\Sigma Y = 19266.64013 + 0.573069\Sigma X1 + 0.514468\Sigma X2 - 15.3857\Sigma X3 - 111.24087\Sigma X4$$

Next, test the coefficient of determination to measure the effect of the dependent variable on the independent variable, to determine the degree of compatibility of the multiple linear regression model. The coefficient of determination has a value between 0 to 1, for the value of R2 = 0 meaning that the influence between the dependent variable and the independent variable does not exist, when the value of R2 is close to 1, the stronger the effect of the independent variable on the dependent variable. The results of testing the coefficient of determination in table 8 is 0.919622326, it means that the multiple linear regression model has a match of 91.96%. The dependent variable, the amount of coffee plant production is influenced by the independent variable of 91.96%, namely the area, rainfall, air pressure, and solar radiation. The remaining 8.04% is influenced by other variables not measured in this study.

Table 8. Determination Coefficient Test Results

R2	Kd
0.919622326	91,96%

Source: (Khumaidi, 2019)

Evaluation

Based on the results of the validation carried out using the RapidMiner software the calculation of the average value of the accuracy of the RMSE is equal to 0, 3477. This shows that the variation in values generated by the model has predictive results with a good degree of accuracy.

CONCLUSION

Based on the results of the analysis and discussion of data processing and modeling it can be concluded that the CRISP-DM method and

multiple linear regression algorithm can be applied to predict the amount of coffee crop production. After the data preparation stage with 150 data obtained from BPS, using the multiple linear regression model produces a coefficient of determination of 91.96%. That the variation in the value of the number of coffee plantations is influenced by the independent variables namely the area of plantations, rainfall, air pressure and solar radiation by 91.96% and 8.04% influenced by other variables not measured in this study. The results of the prediction evaluation produce good accuracy values with an RMSE value of 0.3477.

REFERENCES

Ashari, S. (2004). *Biologi reproduksi tanaman buah-buahan komersial*. Malang : Bayu Media Pub.

Badan Pusat Statistik. (2018). Data Online Badan Statistik: Tabel Dinamis Subjek Perkebunan. Retrieved from https://www.bps.go.id/subject/54/perkebunan.html#subjekViewTab6

Bengnga, A., & Ishak, R. (2018). Prediksi Jumlah Mahasiswa Registrasi Per Semester Menggunakan Linier Regresi Pada Universitas Ichsan Gorontalo. *ILKOM Jurnal Ilmiah*, *10*(2), 136–143. Retrieved from http://jurnal.fikom.umi.ac.id/index.php/ILKOM/article/view/274

Colin Shearer. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13–22.

Direktorat Jenderal Perkebunan. (2014). *Statistik Perkebunan Indonesia Komoditas Kopi 2013 - 2015*. Jakarta. Retrieved from https://bulelengkab.go.id/assets/instansikab/126/bankdata/data-statistik-perkebunan-indonesia-2013-2015-kopi-93.pdf

Iscaro, J. (2014). The Impact of Climate Change on Coffee Production in Colombia and Ethiopia. *Global Majority E-Journal*, *5*(1), 33–43. Retrieved from http://www.bangladeshstudies.org/files/Global_Majority_e_Journal_5_1.pdf#page=33

Kamal, I. M., Hendro, T., & Ilyas, R. (2017). Prediksi Penjualan Buku Menggunakan Data Mining Di Pt. Niaga SwadayaPrediksi Penjualan Buku Menggunakan Data Mining Di Pt. Niaga Swadaya. In STMIK AMIKOM (Ed.), *Seminar Nasional Teknologi Informasi dan Multimedia* (pp. 49–54). Yogyakarta: STMIK AMIKOM.

Khumaidi, A. (2019). *Final Research Report*. Jakarta.

Lenisastri, L. (2000). *Penggunaan Metode Akumulasi Satuan Panas (Heat Unit) sebagai Dasar Penentuan Umur Panen Benih Sembilan Varietas Kacang Tanah (Arachis hypogaea L.)*. Institut Pertanian Bogor. Retrieved from https://repository.ipb.ac.id/handle/123456789/20324

Maulida, L. (2018). Kunjungan Wisatawan Ke Objek Wisata Unggulan Di Prov . Dki Jakarta Dengan K-Means. *JISKa*, *2*(3), 167–174. Retrieved from http://ejournal.uin-suka.ac.id/saintek/JISKA/article/view/1200

Ngumar, Y. H. (2008). Aplikasi Metode Numerik Dan Matrik Dalam Perhitungan Koefisien-Koefisien Regresi Linier Multiple Untuk Peramalan. In *Konferensi Nasional Sistem dan Informatika* (pp. 157–162). Bali: STIKOM Bali. Retrieved from https://yudiagusta.files.wordpress.com/2009/11/157-162-knsi08-029-aplikasi-metode-numerik-dan-matrik-dalam-perhitungan-koefisien-koefisien-regresi-linier-multiple-untuk-peramalan.pdf

Prasetyo, S. B., Aini, N., Dawam, M., Jurusan, M., Pertanian, B., & Pertanian, F. (2017). Dampak Perubahan Iklim Terhadap Produktivitas Kopi Robusta (Coffea Robusta) Di Kabupaten Malang. *Jurnal Produksi Tanaman*, *5*(5), 805–811.

Setiawan, E. (2009). Kajian Hubungan Unsur Iklim Terhadap Produktivitas Cabe Jamu (Piper Retrofractum Vahl) Di Kabupaten Sumenep. *AGROVIGOR*, *2*(1), 1–7.

Soni, N., & Ganatra, A. (2012). Categorization of Several Clustering Algorithms from Different Perspective: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, *2*(8), 63–68. Retrieved from https://www.researchgate.net/profile/Neha_Soni8/publication/267368768_Categorization_of_Several_Clustering_Algorithms_from_Different_Perspective_A_Review/links/575a6f7208ae414b8e460fa6/Categorization-of-Several-Clustering-Algorithms-from-Different-Perspective-A-Review.pdf

Warih, E. I. A., & Rahayu, Y. (2015). *Penerapan Data Mining Untuk Menentukan Estimasi*

Produktivitas Tanaman Tebu Dengan Menggunakan Algoritma Linier Regresi Berganda Di Kabupaten Rembang. Semarang. Retrieved from http://eprints.dinus.ac.id/16925/1/jurnal_16115.pdf

Widayat, H. P., Anhar, A., & Baihaqi, A. (2015). Dampak Perubahan Iklim Terhadap Produksi, Kualitas Hasil Dan Pendapatan Petani Kopi Arabika Di Aceh Tengah. *Agrisep*, *16*(2), 8–16. Retrieved from http://www.jurnal.unsyiah.ac.id/agrisep/article/view/3041

DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE PRODUCTION USING CRISP-DM METHOD

ORIGINALITY REPORT

10%	7%	6%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijicic.org Internet Source	1%
2	Submitted to University of Surrey Student Paper	1%
3	Dedi, Muhammad Iqbal Dzulhaq, Kartika Wulan Sari, Syaipul Ramdhan, Rahmat Tullah, Sutarman. "Customer Segmentation Based on RFM Value Using K-Means Algorithm", 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019 Publication	1%
4	www.science.gov Internet Source	<1%
5	ojs.unm.ac.id Internet Source	<1%
6	www.researchgate.net Internet Source	<1%
7	Kaimuddin, K Mustari, I Ridwan, F Natasya, A Yassi, A H Bahrun. "Effect of climatic factors	<1%

	on the level of Coffee berry borer (Hypothenemus hampei Ferr) attack on smallholder coffee plantation in Tana Toraja Regency", IOP Conference Series: Earth and Environmental Science, 2021 Publication	
8	Moch Fachri, Ali Khumaidi. "Positioning Accuracy of Commercial Bluetooth Low Energy Beacon", 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019 Publication	<1 %
9	archive.org Internet Source	<1 %
10	quizlet.com Internet Source	<1 %
11	Submitted to South Bank University Student Paper	<1 %
12	Steven Johan, Friska Natalia, Ferry Vincenttius Ferdinand, Sud Sudirman. "Image-Based Skin Cancer Early Detection using CNN Algorithm", 2021 6th International Conference on New Media Studies (CONMEDIA), 2021 Publication	<1 %
13	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %

14	thesis.binus.ac.id Internet Source	<1 %
15	Ervan Triyanto, Heri Sismoro, Arif Dwi Laksito. "IMPLEMENTASI ALGORITMA REGRESI LINEAR BERGANDA UNTUK MEMPREDIKSI PRODUKSI PADI DI KABUPATEN BANTUL", Rabit : Jurnal Teknologi dan Sistem Informasi Univrab, 2019 Publication	<1 %
16	ebin.pub Internet Source	<1 %
17	kc.umn.ac.id Internet Source	<1 %
18	Enrica Ciucci, Pamela Calussi, Ersilia Menesini, Alessandra Mattei, Martina Petralli, Simone Orlandini. "Weather daily variation in winter and its effect on behavior and affective states in day-care children", International Journal of Biometeorology, 2010 Publication	<1 %
19	I Ridwan, A Ala, Irfansyah T, Rafiuddin, M Farid BDR, F Haring. " Good Agriculture Practice (GAP) of arabica coffee (L.): Implementation on the smallholder estate in Enrekang Regency ", IOP Conference Series: Earth and Environmental Science, 2020 Publication	<1 %

20	N Almumtazah, N Azizah, Y L Putri, D C R Novitasari. "Prediksi Jumlah Mahasiswa Baru Menggunakan Metode Regresi Linier Sederhana", JURNAL ILMIAH MATEMATIKA DAN TERAPAN, 2021 Publication	<1 %
21	aristiareyhan.home.blog Internet Source	<1 %
22	repository.ub.ac.id Internet Source	<1 %
23	www.dmst.aueb.gr Internet Source	<1 %
24	J. M. Mesa, C. Menendez, F. A. Ortega, P. J. Garcia. "A smart modelling for the casting temperature prediction in an electric arc furnace", International Journal of Computer Mathematics, 2009 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE

by sky high

Submission date: 27-Jul-2022 11:13AM (UTC-0700)
Submission ID: 1875883288
File name: ICTING_THE_AMOUNT_OF_COFFEE_PRODUCTION_USING_CRISP-DM_METHOD.pdf (933.03K)
Word count: 4664
Character count: 25689

DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE PRODUCTION USING CRISP-DM METHOD

Ali Khumaidi

Informatics Engineering
Universitas Krisnadwipayana, Jakarta, Indonesia
www.unkris.ac.id
alikhumaidi@unkris.ac.id

Abstract—*The production of coffee plantations has become the leading plantation commodity with the export value of the fourth rank after oil palm, rubber and coconut. The number of coffee needs for export every year always increases, therefore it is necessary to predict the yield of coffee plants to estimate planting and anticipation that will be done so as to achieve the target. Coffee plant productivity is influenced by internal and external factors, namely the quality of the plant itself, soil, altitude and climate. The method used in this study is the CRISP-DM method and multiple linear regression algorithm to predict the amount of coffee production and determine the relationship between the variables. The steps taken are business understanding, data understanding, data preparation, modeling and evaluation. The data set that is used as many as 170 data after going through the data preparation stage produced 150 data with 5 attributes in the table. With calculations using tools, the coefficient of determination is 91.96%. That the variation in the value of the production of coffee plants is influenced by independent variables, namely the area of plantations, rainfall, air pressure and solar radiation by 91.96% and 8.04% influenced by other variables not measured in this study. The results of the evaluation and validation of predictions produce good accuracy with an RMSE value of 0.3477.*

Keyword: data mining, coffee plants, crisp-dm, multiple linear progression.

Intisari—*Produksi hasil perkebunan kopi menjadi komoditas perkebunan unggulan dengan nilai ekspor urutan ke empat setelah komoditas kelapa sawit, karet, dan kelapa. Jumlah kebutuhan kopi untuk ekspor tiap tahun selalu mengalami peningkatan oleh karena itu diperlukan prediksi hasil produksi tanaman kopi untuk memperkirakan penanaman dan antisipasi yang akan dilakukan sehingga dapat mencapai target. Produktivitas tanaman kopi dipengaruhi oleh faktor internal dan eksternal yaitu kualitas tanaman itu sendiri, tanah, ketinggian dan ilkim. Metode yang digunakan dalam penelitian ini adalah metode CRISP-DM dan algoritma multiple linear regression untuk prediksi jumlah produksi kopi dan mengetahui hubungan antar variabelnya. Tahapan yang dilakukan adalah business understanding, data understanding, data preparation, modeling dan evaluation. Data set yang digunakan sebanyak 170 data setelah melalui tahap data preparation dihasilkan 150 data dengan 5 atribut pada tabel. Dengan perhitungan menggunakan tools dihasilkan koefisien determinasi sebesar 91,96%. Bahwa variasi nilai jumlah produksi tanaman kopi dipengaruhi oleh variabel independen yaitu luas areal perkebunan, curah hujan, tekanan udara dan penyinaran matahari sebesar 91,96% dan 8,04% dipengaruhi oleh variabel lain yang tidak diukur pada penelitian ini. Hasilevaluasi dan validasi prediksi menghasilkan nilai akurasi yang baik dengan nilai RMSE sebesar 0,3477.*

Kata kunci: *data mining, tanaman kopi, crisp-dm, multiple linier progression.*

INTRODUCTION

The production of Indonesian coffee plantations ranks fourth in the world after Brazil, Vietnam and Colombia. As for the largest export commodity in Indonesia, coffee is ranked fourth with a total trade reached 1.19 billion US $ in 2017 so that coffee becomes one of the leading plantation commodities after the commodities of oil palm, rubber, and coconut. Coffee commodity by the Directorate General of Plantations is placed in the strategic plan as the main target and priority agenda for improving the agro-industry sector, namely increasing the amount of production and exports and developing agro-industry in the village (Direktorat Jenderal Perkebunan, 2014). To support this, control and monitoring need to be done so that coffee productivity tends to increase.

Indonesia ranks second only to Brazil in the area of coffee plantations. The area of coffee plantations tends to increase 1.6% per year during the period 1980 to 2018. However, the data for the last 10 years the area of coffee plantations has decreased 0.05% per year with an area of 1.27

million hectares in 2009 to 1.26 million hectares in 2009 2018 (Direktorat Jenderal Perkebunan, 2014). Although there was a reduction in the area of coffee plantations, in terms of productivity there was an increasing tendency with growth of 1.14% per year. During the period 1980 to 2017 the number of coffee exports was quite volatile with a tendency to increase 3.93% per year. The period of 2012 to 2016 export volume experienced slowing growth, even in 2014 there was a decline in the value of coffee exports up to 11.47% due to the decline in coffee plantation production (Badan Pusat Statistik, 2018).

Climate change can cause negative impacts on plants, so that it can affect its productivity, including coffee(Iscaro, 2014). In Indonesia, climate factors are part of natural phenomena whose changes are influenced by nature and human intervention. The development of coffee plants is influenced by temperature which is able to control root growth, respiration, absorption of nutrients and water, photosynthesis and reaction speed (Lenisastri, 2000). Enzyme systems can function well and are stable at optimum temperatures and at cold temperatures the system is stable but reduces the function so that the enzyme can be damaged (Setiawan, 2009). In addition to temperature, rainfall and humidity factors influence the growth of coffee plants. High levels of rainfall, flowering process can be disrupted and the level of humidity affects generative and vegetative growth (Ashari, 2004). Climate affects coffee productivity, temperature is directly related but humidity and rainfall are not directly related. Good coffee cultivation management techniques can be used to anticipate climate change (Prasetyo et al., 2017). The presence of diseases and pest attacks can be caused by climate change, which previously was only affected by low altitude (Widayat, Anhar, & Baihaqi, 2015).

The condition of the declining area of coffee plantations, increasing export needs and volatile climate conditions as well as targets and priorities for increasing coffee production, it is necessary to predict coffee productivity so that the predicted results can be input for making the best policy. Prediction is a pattern that can be identified by data mining (Maulida, 2018). Data mining is a very large data extraction or extraction process and can be used to assist in making important decisions (Soni & Ganatra, 2012). In previous studies, data mining has been widely used for predictions, research for book sales predictions using data mining in PT. Niaga Swadaya (Kamal, Hendro, & Ilyas, 2017), research for predicting the number of student registrations per semester using linear regression at Ichsan University Gorontalo (Bengnga & Ishak, 2018), research related to the application of data mining to determine the estimated productivity of sugarcane by using an algorithm linear regression in Rembang Regency (Warih & Rahayu, 2015).

In using data mining, it is necessary to use the right method and the appropriate algorithm. In the prediction of coffee plant productivity there are several factors involved, namely the area of plantation crops, the amount of production, rainfall, solar radiation and air pressure. Multiple linear regression algorithm is a regression analysis that explains the relationship between dependent variables with factors that affect more than one. This multiple linear regression algorithm has been used in several studies with a fairly high accuracy rate of up to 95% (Kamal et al., 2017). Therefore, the multiple linear regression algorithm is expected to obtain the results of the productivity of coffee plants as a consideration in making further policies.

Analysis of coffee plant productivity needs to be done in depth in order to obtain hidden patterns and the discovery of knowledge related to production predictions. The appropriate method for conducting this research is the Cross-Industry Standard Process for Data Mining (CRISP-DM). CRISP-DM is a method that provides standard processes in data mining for solving problems in business. CRISP-DM is easier to apply because each stage or phase is clearly defined and structured and has a complete and well-documented data mining methodology.

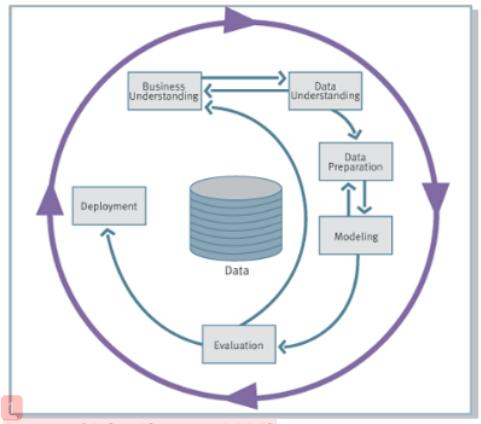
MATERIALS AND METHODS

The stages in this study are adjusted to the phase of the CRISP-DM method. CRISP-DM is a data mining standard that contains a framework for data mining tasks. In the data mining process based on CRISP-DM there are 6 phases (Colin Shearer, 2000) according to Figure 1.

Business Understanding is the phase of understanding the substance of data mining activities that will be carried out. Its activities include: determining business goals or objectives, understanding business situations, and determining data mining goals.

Data Understanding is the initial data collection phase. Its activities include: studying and describing data, exploiting data and identifying problems related to data quality, and detecting an interesting subset of data as an initial hypothesis.

Data Preparation is an activity carried out to prepare data. Its activities include data selection, data building, data integration and data cleaning.



Source: (Colin Shearer, 2000)

Gambar 1. Model CRISP-DM

Modeling is the phase of determining data mining techniques. Data mining techniques use multiple linear regression algorithm. Regression is a model development technique that can be used for prediction, by looking for the relationship between the dependent variable and the independent variable. Multiple linear regression is a regression analysis that links the dependent variable with more than one independent variable that affects it (Ngumar, 2008). Multiple linear regression will work optimally if the input used is numeric. The formula used for multiple linear regression is expressed in equation (1).

$$Y = a + b1X1 + b2X2 + \cdots + bnXn \quad \text{..... (1)}$$

Information:

Y	= dependent variable
X1, X2 .... Xn	= independent variable
a	= constant
b1, b2 .... bn	= regression coefficient

According to equation (1), to calculate the amount of coffee crop production using multiple linear regression algorithm, equation (2) is used.

$$Y = a + b1X1 + b2X2 + b3X3 + b4X4 \quad \text{..... (2)}$$

Information:

Y = Number of Production
X1 = Area of plantation area
b1 = Variable coefficient of plantation area
X2 = Rainfall
b2 = coefficient of rainfall variable
X3 = Barometric pressure
b3 = coefficient of variable air pressure
X4 = Solar Light
b4 = coefficient of solar irradiation variables

The first stage is the formation of a model, by finding the values of a, b1, b2, b3, and b4 using the least square with the general equation referring to equation (3).

$$\begin{aligned}
\Sigma Y &= an + b1\Sigma X1 + b2\Sigma X2 + b3\Sigma X3 + b4\Sigma X4 \\
\Sigma X1Y &= a\Sigma X1 + b1\Sigma(X1)2 + b2\Sigma(X1X2 + b3\Sigma(X1X3) \\
&\quad + b4\Sigma(X1X4) \\
\Sigma X2Y &= a\Sigma X2 + b1\Sigma(X1X2) + b2\Sigma(X2)2 \\
&\quad + b3\Sigma(X2X3) + b4\Sigma(X2X4) \\
\Sigma X3Y &= a\Sigma X3 + b1\Sigma(X1X3) + b2\Sigma(X2X3) + b3\Sigma(X3)2 \\
&\quad + b4\Sigma(X3X4) \\
\Sigma X4Y &= a\Sigma X4 + b1\Sigma(X1X4) + b2\Sigma(X2X4) + \\
&b3\Sigma(X3X4) + b4\Sigma(X4)2 \quad \text{..... (3)}
\end{aligned}$$

Next the acquisition of inversion results is obtained, then the determinant matrix multiplication is carried out with ΣY, $\Sigma X1Y$, $\Sigma X2Y$, $\Sigma X3Y$, $\Sigma X4Y$. Then the determinants of matrices A, A0, A1, A2, A3, and A4 are calculated. The following equation in calculation (4):

$$\begin{vmatrix} \Sigma Y \\ \Sigma X1Y \\ \Sigma X2Y \\ \Sigma X3Y \\ \Sigma X4Y \end{vmatrix} = \begin{vmatrix} N & \Sigma X1 & \Sigma X2 & \Sigma X3 & \Sigma X4 \\ \Sigma X1 & \Sigma X1X1 & \Sigma X1X2 & \Sigma X1X3 & \Sigma X1X4 \\ \Sigma X2 & \Sigma X2X1 & \Sigma X2X2 & \Sigma X2X3 & \Sigma X2X4 \\ \Sigma X3 & \Sigma X3X1 & \Sigma X3X2 & \Sigma X3X3 & \Sigma X3X4 \\ \Sigma X4 & \Sigma X4X1 & \Sigma X4X2 & \Sigma X4X3 & \Sigma X4X4 \end{vmatrix} x \begin{vmatrix} a \\ b1 \\ b2 \\ b3 \\ b4 \end{vmatrix} \text{ .... (4)}$$

Information:

N = amount of data
Σ Y = the sum of the variable production quantities
ΣX1 = number of variable area
ΣX2 = number of rainfall variables
ΣX3 = amount of variable air pressure
ΣX4 = number of solar irradiation variables

Next, the results of the calculation of the determinant matrix are used to find the values of a, b1, b2, b3, and b4 by referring to equation (5).

$$\begin{aligned}
a &= (Det\,(A0)) \,/\, (Det A) \\
b1 &= (Det\,(A1)) \,/\, (Det A) \\
b2 &= (Det\,(A2)) \,/\, (Det\,A) \quad \text{..... (5)} \\
b3 &= (Det\,(A3)) \,/\, (Det A) \\
b4 &= (Det\,(A4)) \,/\, (Det A)
\end{aligned}$$

Next, a partial correlation test will be calculated to see the degree of interrelation of the independent variable with the dependent variable. The related values are rx1Y, rX2Y, rX3Y, rX4Y, rX1X2, rX1X3, rX1X4, rX2X3, rX2X4, and rX3X4. Calculation of correlation values uses equation (6).

$$\begin{aligned}
\Sigma X1Y &= \Sigma X1Y - ((\Sigma X1).(\Sigma Y))/n \\
\Sigma X2Y &= \Sigma X2Y - ((\Sigma X2).(\Sigma Y))/n \\
\Sigma X3Y &= \Sigma X3Y - ((\Sigma X3).(\Sigma Y))/n \\
\Sigma X4Y &= \Sigma X4Y - ((\Sigma X4).(\Sigma Y))/n \\
\Sigma X1X2 &= \Sigma X1X2 - ((\Sigma X1).(\Sigma X2))/n \quad \text{..... (6)} \\
\Sigma X1X3 &= \Sigma X1X3 - ((\Sigma X1).(\Sigma X3))/n \\
\Sigma X1X4 &= \Sigma X1X4 - ((\Sigma X1).(\Sigma X4))/n \\
\Sigma X2X3 &= \Sigma X2X4 - ((\Sigma X2).(\Sigma X3))/n \\
\Sigma X2X4 &= \Sigma X2X4 - ((\Sigma X2).(\Sigma X4))/n
\end{aligned}$$

Calculations for finding R2 refer to (7).

$$R2 = 1 - \frac{SS\ Error}{nSS\ Total} = 1 - \frac{\Sigma(yi-yi)2}{\Sigma(yi-yi)2} \quad \ldots\ldots (7)$$

Information:
y = forecast i-response
y = average
yi = observation of the i-response

For the coefficient of determination the calculation is according to equation (8).
Information:
y = forecast i-response
y = average
yi = observation of the i-response

For the coefficient of determination the calculation is according to equation (8).

$$Kd = R2\ x\ 100\% \quad \ldots\ldots (8)$$

Information:
R2 = correlation coefficient value.
Kd = large coefficient of determination

The accuracy of a prediction is an important aspect and objective of a prediction, so that the error rate is sought as small as possible. There are 3 models for measuring errors in concluding historical errors, namely mean squared error, mean absolute deviation, and mean absolute percent error (William J. Stevenson, 2014). Testing the accuracy of predictions in this study by measuring the value of root mean squared error (RMSE) according to equation (9). RMSE is a method for evaluating prediction results by measuring the accuracy of the model (Chai & Draxler, 2014). The lower RMSE value concludes that the predicted results approach the true value (Choirunisa, 2019). The greater the RMSE value indicates the worse the accuracy level.

$$RMSE = \frac{\sum_{i}^{n}=\sqrt{(yi-yi)2}}{n} \quad \ldots\ldots (9)$$

Evaluation is the phase of interpreting the results of data mining produced. Evaluation is carried out in depth aiming to obtain a model that is in accordance with the objectives.

Deployment is the phase of compiling a report of the knowledge gained at the evaluation stage.

In CRISP-DM there are three stages that are worked on, the first is the data collection stage, the second is the data understanding and business stage and the third is the modeling and evaluation stage.

1. Data Collection Stage
   At this stage, conducting a literature study and collecting coffee plant data includes: plantation area, total production, rainfall, air pressure and solar radiation. The data is collected from BPS.
2. Business Understanding and Data Understanding
   At this stage consists of 3 stages, namely business understanding, data understanding, and data preparation.
3. Modeling and Evaluation stage
   At this stage consists of 2 stages, namely modeling and evaluation.

RESULTS AND DISCUSSIONS

Coffee plant production prediction models include five phases which are the stages of business understanding and data as well as the modeling and valuation stages, as follows:

Business Understanding

This stage refers to the prediction of coffee plant production. At this stage an understanding of the background and objectives of the business processes related to coffee plants is needed, including:

1. Determine Business Goals
   The business objective of doing research is to recognize coffee plant production to determine the prediction of coffee plant production so that policies and actions can be taken if the production predictions are not appropriate and to know the relationship of other factors that affect the amount of production
2. Assess the situation
   The process of developing coffee plants is influenced by internal and external factors. Internal factors, namely the quality of coffee plants and external factors, namely climate, temperature, air pressure, humidity, rainfall, altitude, plantation area, solar radiation and others.
3. Determine the Purpose of Data Mining
   The purpose of data mining or the purpose of this study is to explore knowledge about the patterns of influence of external factors on the production of coffee plants for prediction of the amount of coffee production.

Data Understanding

The understanding of the data stage begins with the collection of preliminary data and the results of activities in order to identify data problems, to determine the first insight into the data or detect interesting subsets to form hypotheses for hidden information.

1. Initial Data Collection
   Data collection is done by taking data from BPS 34 provinces, including:
   - Data on coffee plantation area by province in 2011-2018
   - Data on coffee plantation production by province in 2008-2018
   - Rainfall data for 2000-2015
   - Air pressure data for 2003-2015
   - Data on solar radiation from 2003-2015
2. Describe Data
   This stage is for analysis to understand the data obtained from the initial data collection results.
   - Data on the area of coffee plantations in units of thousands of hectares and consists of data from 34 provinces plus total totals (Indonesia).
   - Production data of coffee plantations in thousand tons and consists of data from 34 provinces plus total totals (Indonesia).
   - Rainfall data in millimeters (mm) taken from 34 provincial BMKG stations including data on the number of rainy days in a year.
   - Air pressure data in units of millibars (mb) taken from BMKG stations in each province are 34.
   - Solar irradiance data in annual percentages based on 34 provincial BMKG station data.
3. Exploring DataThe data exploration process is carried out on four tables, namely:
   - Data on coffee plantation area consisting of attributes of the province, year and amount.
   - Coffee plantation production data consisting of attributes of the province, year and amount.
   - Rainfall data consisting of provincial attributes, BMKG stations, year, amount of rainfall and number of rainy days.
   - Data on air pressure and solar radiation consisting of attributes of the province, BMKG station, year, air pressure and solar radiation.
4. Verifying Data Quality
   The process of verifying data quality by looking at the structure of the table, the data contained in the table then integrates between tables. Since the data collected with different number of years is then sliced between the four data tables, the 2011-2015 period is used. The total data used is 170 data.

Data Preparation

The data preparation stage is the preparation of the dataset, the data used for modeling. This stage includes the following activities:

1. Select and build data
   At this stage related tables are chosen to simplify the process of selecting data. There are 4 tables involved, a table of production quantities, a table of crop area, a table of rainfall, a table of air pressure and solar radiation. Each table contains 170 data with variations in the attributes in the tables below.

Table 1. Total of Coffee Production

Year	Province	Total of Production (Thousand of Tons)
2011	Aceh	52.3
2011	Sumatera Utara	56.8
2011	Sumatera Barat	30.8
2011	Riau	1.9
...	...	...
2015	Papua	2

Source: (Badan Pusat Statistik, 2018)

Table 2. Plantation Area (Thousand Hectares)

Year	Province	Plantation Area
2011	Aceh	120.7
2011	Sumatera Utara	80.6
2011	Sumatera Barat	40.3
2011	Riau	4.7
...	...	...
2015	Papua	10

Source: (Badan Pusat Statistik, 2018)

Table 3. Total Rainfall (millimeters)

Year	Province	BMKG Station	Rainfall
2011	Aceh	Sultan Iskandar Muda	1268.00
2011	Sumatera Utara	Kualanamu	2042.00
2011	Sumatera Barat	Sicincin	
2011	Riau	Sultan Syarif Kasim II	2405.00
...	...	...	...
2015	Papua	Angkasapura	1265.90

Source: (Badan Pusat Statistik, 2018)

Table 4. Air Pressure and Solar Light

Year	Province	BMKG Station	Air pressure	Solar radiation
2011	Aceh	Sultan Iskandar Muda	1009.40	52.20
2011	Sumatera Utara	Kualanamu		44.40
2011	Sumatera Barat	Sicincin	990.80	32.80
2011	Riau	Sultan Syarif Kasim II	1008.70	42.30
...	...	...		...
2015	Papua	Angkasapura	1011.10	64.47

Source: (Badan Pusat Statistik, 2018)

2. Integrating Data
This stage combines four tables into one table into a new data set. The results of the transformation in the previous table are integrated into table 5.

Table 5. Combined tables

Year	Province	BMKG Station	Production	Plantation	Rainfall	Air pressure	Solar radiation
2011	Aceh	Sultan Iskandar Muda	52.3	120.7	1268.00	1009.40	52.20
2011	Sumatera Utara	Kualanamu	56.8	80.6	2042.00		44.40
2011	Sumatera Barat	Sicincin	30.8	40.3		990.80	32.80
2011	Riau	Sultan Syarif Kasim II	1.9	4.7	2405.00	1008.70	42.30
...	...	...	...	...	...		...
2015	Papua	Angkasapura	2	10	1265.90	1011.10	64.47

Sumber: (Badan Pusat Statistik, 2018)

3. Clearing Data
This process removes some attributes or variables so that the final results for the variables will be processed in table 6. There are only 5 variables used for the modeling stage.

Table 6. Variables for Modeling

Production	Plantation Areal	Rainfall	Air pressure	Solar radiation
52.3	120.7	1268.00	1009.40	52.20
56.8	80.6	2042.00		44.40
30.8	40.3		990.80	32.80
1.9	4.7	2405.00	1008.70	42.30
...	...	...		...
2	10	1265.90	1011.10	64.47

Source: (Khumaidi, 2019)

In addition to adjusting the variables, cleaning of the empty data is also done, so that initially there were 170 data to 150 data.

Modelling

Based on the data set prepared by doing calculations with several equations, the results obtained are a, b1, b2, b3, and b4, as shown in table 7.

Table 7. Calculation Results

a	*b1*	*b2*	*b3*	*b4*
19266.64013	0.57369	0.514468	-15.385717	-111.240876

Source: (Khumaidi, 2019)

Based on the results obtained for the coefficients a, b1, b2, b3, and b4, the multiple linear regression model is as follows.

$$\Sigma Y = 19266.64013 + 0.573069\Sigma X1 + 0.514468\Sigma X2 - 15.3857\Sigma X3 - 111.24087\Sigma X4$$

Next, test the coefficient of determination to measure the effect of the dependent variable on the independent variable, to determine the degree of compatibility of the multiple linear regression model. The coefficient of determination has a value between 0 to 1, for the value of R2 = 0 meaning that the influence between the dependent variable and the independent variable does not exist, when the value of R2 is close to 1, the stronger the effect of the independent variable on the dependent variable. The results of testing the coefficient of determination in table 8 is 0.919622326, it means that the multiple linear regression model has a match of 91.96%. The dependent variable, the amount of coffee plant production is influenced by the independent variable of 91.96%, namely the area, rainfall, air pressure, and solar radiation. The remaining 8.04% is influenced by other variables not measured in this study.

Table 8. Determination Coefficient Test Results

R2	Kd
0.919622326	91,96%

Source: (Khumaidi, 2019)

Evaluation

Based on the results of the validation carried out using the RapidMiner software the calculation of the average value of the accuracy of the RMSE is equal to 0, 3477. This shows that the variation in values generated by the model has predictive results with a good degree of accuracy.

CONCLUSION

Based on the results of the analysis and discussion of data processing and modeling it can be concluded that the CRISP-DM method and

multiple linear regression algorithm can be applied to predict the amount of coffee crop production. After the data preparation stage with 150 data obtained from BPS, using the multiple linear regression model produces a coefficient of determination of 91.96%. That the variation in the value of the number of coffee plantations is influenced by the independent variables namely the area of plantations, rainfall, air pressure and solar radiation by 91.96% and 8.04% influenced by other variables not measured in this study. The results of the prediction evaluation produce good accuracy values with an RMSE value of 0.3477.

REFERENCES

Ashari, S. (2004). *Biologi reproduksi tanaman buah-buahan komersial*. Malang : Bayu Media Pub.

Badan Pusat Statistik. (2018). Data Online Badan Statistik: Tabel Dinamis Subjek Perkebunan. Retrieved from https://www.bps.go.id/subject/54/perkebunan.html#subjekViewTab6

Bengnga, A., & Ishak, R. (2018). Prediksi Jumlah Mahasiswa Registrasi Per Semester Menggunakan Linier Regresi Pada Universitas Ichsan Gorontalo. *ILKOM Jurnal Ilmiah*, *10*(2), 136–143. Retrieved from http://jurnal.fikom.umi.ac.id/index.php/ILKOM/article/view/274

Colin Shearer. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13–22.

Direktorat Jenderal Perkebunan. (2014). *Statistik Perkebunan Indonesia Komoditas Kopi 2013 - 2015*. Jakarta. Retrieved from https://bulelengkab.go.id/assets/instansikab/126/bankdata/data-statistik-perkebunan-indonesia-2013-2015-kopi-93.pdf

Iscaro, J. (2014). The Impact of Climate Change on Coffee Production in Colombia and Ethiopia. *Global Majority E-Journal*, *5*(1), 33–43. Retrieved from http://www.bangladeshstudies.org/files/Global_Majority_e_Journal_5_1.pdf#page=33

Kamal, I. M., Hendro, T., & Ilyas, R. (2017). Prediksi Penjualan Buku Menggunakan Data Mining Di Pt. Niaga SwadayaPrediksi Penjualan Buku Menggunakan Data Mining Di Pt. Niaga Swadaya. In STMIK AMIKOM (Ed.), *Seminar Nasional Teknologi Informasi dan Multimedia* (pp. 49–54). Yogyakarta: STMIK AMIKOM.

Khumaidi, A. (2019). *Final Research Report*. Jakarta.

Lenisastri, L. (2000). *Penggunaan Metode Akumulasi Satuan Panas (Heat Unit) sebagai Dasar Penentuan Umur Panen Benih Sembilan Varietas Kacang Tanah (Arachis hypogaea L.)*. Institut Pertanian Bogor. Retrieved from https://repository.ipb.ac.id/handle/123456789/20324

Maulida, L. (2018). Kunjungan Wisatawan Ke Objek Wisata Unggulan Di Prov . Dki Jakarta Dengan K-Means. *JISKa*, *2*(3), 167–174. Retrieved from http://ejournal.uin-suka.ac.id/saintek/JISKA/article/view/1200

Ngumar, Y. H. (2008). Aplikasi Metode Numerik Dan Matrik Dalam Perhitungan Koefisien-Koefisien Regresi Linier Multiple Untuk Peramalan. In *Konferensi Nasional Sistem dan Informatika* (pp. 157–162). Bali: STIKOM Bali. Retrieved from https://yudiagusta.files.wordpress.com/2009/11/157-162-knsi08-029-aplikasi-metode-numerik-dan-matrik-dalam-perhitungan-koefisien-koefisien-regresi-linier-multiple-untuk-peramalan.pdf

Prasetyo, S. B., Aini, N., Dawam, M., Jurusan, M., Pertanian, B., & Pertanian, F. (2017). Dampak Perubahan Iklim Terhadap Produktivitas Kopi Robusta (Coffea Robusta) Di Kabupaten Malang. *Jurnal Produksi Tanaman*, *5*(5), 805–811.

Setiawan, E. (2009). Kajian Hubungan Unsur Iklim Terhadap Produktivitas Cabe Jamu (Piper Retrofractum Vahl) Di Kabupaten Sumenep. *AGROVIGOR*, *2*(1), 1–7.

Soni, N., & Ganatra, A. (2012). Categorization of Several Clustering Algorithms from Different Perspective: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*, *2*(8), 63–68. Retrieved from https://www.researchgate.net/profile/Neha_Soni8/publication/267368768_Categorization_of_Several_Clustering_Algorithms_from_Different_Perspective_A_Review/links/575a6f7208ae414b8e460fa6/Categorization-of-Several-Clustering-Algorithms-from-Different-Perspective-A-Review.pdf

Warih, E. I. A., & Rahayu, Y. (2015). *Penerapan Data Mining Untuk Menentukan Estimasi*

Produktivitas Tanaman Tebu Dengan Menggunakan Algoritma Linier Regresi Berganda Di Kabupaten Rembang. Semarang. Retrieved from http://eprints.dinus.ac.id/16925/1/jurnal_16115.pdf

Widayat, H. P., Anhar, A., & Baihaqi, A. (2015). Dampak Perubahan Iklim Terhadap Produksi, Kualitas Hasil Dan Pendapatan Petani Kopi Arabika Di Aceh Tengah. *Agrisep*, *16*(2), 8–16. Retrieved from http://www.jurnal.unsyiah.ac.id/agrisep/article/view/3041

DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE

ORIGINALITY REPORT

94%	94%	6%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ejournal.nusamandiri.ac.id Internet Source	92%
2	Submitted to Asia e University Student Paper	1%
3	Submitted to Sabanci Universitesi Student Paper	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

DATA MINING FOR PREDICTING THE AMOUNT OF COFFEE