Online Version Available at: http://bsti.ubd.ac.id/e-jurnal

JOURNAL OF TECH-E

|2581-1916 (Online) | 2598-7585 (Printed) |

Article

Implementation of Random Forest Algorithm on Palm Oil Price Data

Arif Rahman Hakim[1], *Dewi Marini Umi Atmaja*[2], *Amat Basri*[3], *Muhamad Syafii*[4]

[1, 2,] *Medika Suherman University, Digital Business, Jawa Barat, Indonesia*
[3] *Buddhi Dharma University, Information Systems, Bbanten, Indonesia*
[4] *Budi Luhur University, Information System, Jakarta Selatan, Indonesia*

SUBMISSION TRACK

Recieved: January 15, 2023
Final Revision: February 19, 2023
Available Online: February 25, 2023

KEYWORD

Palm Oil, Random Forest, Data Mining

CORRESPONDENCE

E-mail: arif@medikasuherman.ac.id

A B S T R A C T

One of the potential commodities that are widely cultivated in Indonesia is palm oil, palm oil or commonly referred to as crude palm oil is one of the processed palm oil which produces the most important foreign exchange for Indonesia. Data mining is a process that utilizes mathematical techniques, statistics, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from large databases, including palm oil price data. Random Forest is one of the methods in decision trees. A decision tree is a flowchart shaped like a tree with a root node used to collect data used to solve problems and make decisions. In this study, the random forest algorithm was used to classify palm oil price data from 2014 to 2019. The classification method using the random forest algorithm on palm oil data using the Mtry parameter of 1 and the Ntree parameter of 500 produces a percentage accuracy rate of 100%. The most influential variable (importance variable) in the classification model using the random forest algorithm produced is the palm oil variable.

INTRODUCTION

One potential commodity that is widely cultivated in Indonesia is palm oil, and palm oil or commonly referred to as palm oil is one of the most important foreign exchange earners for Indonesia. The demand for palm oil continues to grow in line with the increasing consumption of vegetable oil in the world [1] . Palm oil is even being converted into biofuel to curb the effects of global warming. The market opportunity for palm oil is very promising, as annual demand is increasing at a high rate. The main factor behind the increased demand for palm oil is its relatively lower price compared to competitors such as soybean oil and Canada oil/canola [2] .

Data mining is a data mining process that utilizes techniques such as mathematics, statistics, artificial intelligence, and machine learning techniques to extract and identify useful information and relevant knowledge from large databases [3] , including palm oil price data. The price of palm oil can increase, decrease or remain the same every day, because there are several factors that affect the price of palm oil, such as the price of other vegetable oils (soybean oil and canada oil/canola), the world crude oil price, and the dollar exchange rate against the exchange rate of producing countries or the exchange rate of consumer countries [4] . Therefore, it is necessary to classify/categorize palm oil prices and what factors affect palm oil prices [5] .

Random Forest is a decision tree method. A decision tree is a flowchart shaped like a tree with a root node used to collect data used to solve problems and make decisions [6] . Classification is the process of finding patterns or functions that describe or distinguish classes in data or a concept with the aim of inferring the class of an object whose label is unknown [7] .

In this research, the random forest algorithm is used to classify palm oil price data from 2014 to 2019. With the classification and knowledge of the factors that affect the price of palm oil is expected to help palm oil industry in making decisions.

I. LITERATURES REVIEW

Literature studies are used to find novelty comparisons from similar studies that have previously been conducted. A similar study discussing the price of palm oil using the Support Vector Machine algorithm has the highest accuracy, precision, and recall compared to the Naïve Bayes algorithm and the K-Nearest Neighbor algorithm. The highest accuracy value in this study is 82.46% with the highest precision of 86% and the highest recall of 89.06% [8]. Other studies that apply data mining to analyze production stocks that will be predicted using a decision tree produce an accuracy value above 90% [9]. In addition to the decision tree algorithm, research has also been carried out using a multiple linear regression algorithm in predicting the price of crude palm oil (CPO). not with 100% accuracy value but still within the margin of error in predicting [10].

II. FRAMEWORK

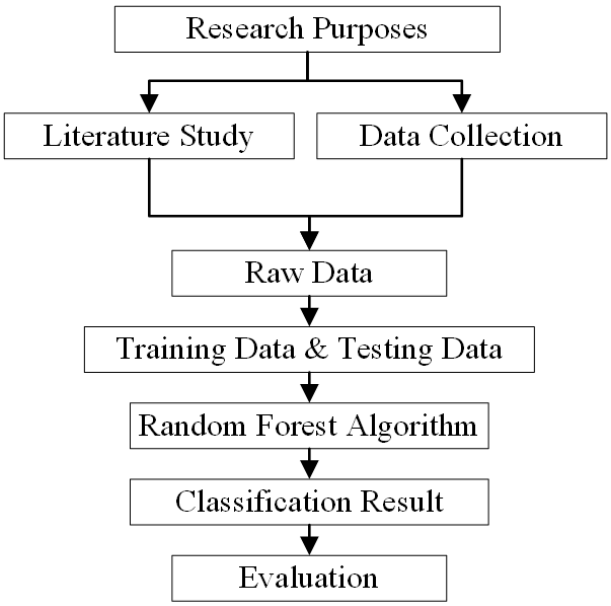

Figure 1. KDD Process Stages

This study uses an approach quantitative and research procedures using Knowledge Discovery in Databases (KDD) stages. The purpose of this study was to classify palm oil price data, then collect various sources of literature in the form of national and international journals to support the research. Simultaneously with the collection of literature

studies, data collection was also carried out to be processed in the research. The data used in this study were taken from the official investing.com website. The raw data collected will then be processed into training and testing data, then a model will be formed using the random forest algorithm to produce a classification model. This model will be used to classify palm oil price data, so that the evaluation results are obtained in the form of an accuracy value.

III. METHODS

In this research, the stages used in the classification of palm oil price data are identifying attributes to facilitate research, so that research is carried out systematically and accurately to achieve the desired goals. In this research, the subject matter is palm oil price data by comparing other variables taken from 2014 to 2019 from the investing.com website.

Research Variable

As for the research variables applied in this research are divided into two types of variables, namely predictors / independent variables consisting of dates (not used), palm oil, WTI crude oil, canoil oil, soybean oil, USD_IDR, USD_INR, USD_MYR, USD_CAD and response variables / dependent variables, namely prediction results. USD_IDR, USD_INR, USD_MYR, USD_CAD variables are currency variables with units of American Dollars (USD), while IDR, INR, MYR and CAD are currencies for Indonesia, India, Malaysia and Canada which are crude oil producers.

In the process of designing a system, it must be done in a structured manner to reduce inefficiency and ineffectiveness. When designing this system, the researcher used the Knowledge Discovery in Databases (KDD) method [8] .

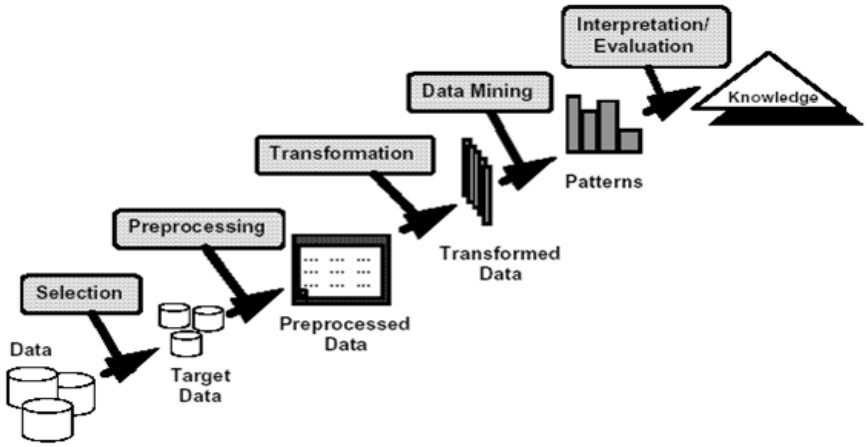

Figure 2 . KDD Process Stages

The stages of the Knowledge in Discovery Database (KDD) process include:

Selection, which is the selection of data from the active dataset, must be done before starting the information extraction stage in Knowledge Discovery in Databases. The selected data used for data mining is stored in a file, separate from the operational database [9]

Preprocessing is the process of preparing data before further processing. Before data mining can be done, it is necessary to perform data cleaning in accordance with Knowledge Discovery in Databases (KDD). The cleaning process includes removing duplicate data, checking for data inconsistencies, and correcting errors in the data [10] . Therefore, it is necessary to handle especially data that contains noise, the noise is a random error or variance in the measured variable [11] . Proprecessing is the process of preparing raw data before further processing that is no longer relevant and modifying the data to facilitate the processing system [12] .

Transformation is a conversion or encoding process that converts data into a specific format for later use and tracking [13] . Encoding is the

process of transforming the selected data so that it is suitable for data mining. The encoding process in Knowledge Discovery in Databases (KDD) is a creative process and depends heavily on the type or pattern of information being sought from the database.

Data mining is the process of finding interesting patterns or information in selected data using certain methods or techniques. Data mining methods, techniques or algorithms are very diverse. The choice of an appropriate method or algorithm is highly dependent on the objectives and the overall Knowledge Discovery in Databases (KDD) process.

Evaluation / Interpretation, information generated by the data mining process must be displayed in a form that is easily understood by the stakeholders [14] . This step is part of the Knowledge Discovery in Databases (KDD) process called interpretation. This step involves checking whether the discovered model or information contradicts pre-existing facts or assumptions.

Analysis Method

The analysis tool used by researchers in this research is Microsoft Excel 2016 software and RStudio IDE 1.3.1092. The method applied in this research is descriptive analysis to get an overview of the research data and use the random forest method to compare the classification analysis results.

The stages of research, especially in this research, namely, the initial stage begins with the identification of the selected subject. Then, identify and formulate the problems that will be investigated in this study. In addition, identify the methodology and literature review of previous studies. The next step is to collect or retrieve data, then clean the data so that it is easy to use properly in this study. The next step is to conduct a descriptive analysis of the data under study. Researchers divide the data into two, namely training data and test data that will be used in the machine learning model. In the classification stage, the method used is random forest. After the model is trained, the accuracy of the test data is calculated. Then, the comparison of accuracy between models is then interpreted and concluded from the research results.

IV. RESULT

Research Result

The dataset used in this study consists of 1233 objects and 9 variables which are divided into two predicted/dependent variables and independent variables (palm oil, WTI crude oil, canoil oil, soybean oil, USD_IDR, USD_INR, USD_MYR, USD_CAD). The categorical dependent variable includes two classes of data, up and down. While the independent variables are integers and are divided into nine variables, namely palm oil, WTI crude oil, canoil oil, soybean oil, USD_IDR, USD_INR, USD_MYR, USD_CAD.

```
> str(sawit)
'data.frame':   1233 obs. of  9 variables:
 $ Minyak.Sawit     : num  514 511 522 519 526 ...
 $ Minyak.Kedelai   : num  28.6 28.4 28.1 28 28.3 ...
 $ Minyak.Canola    : num  448 448 448 446 448 ...
 $ Minyak.Mentah.WTI: num  55.1 56.7 55.8 54.9 53.6 ...
 $ USD_IDR          : num  14185 14238 14255 14255 14240 ...
 $ USD_MYR          : num  4.21 4.22 4.21 4.2 4.2 4.19 4.19 4.18 4.18 4.18 ...
 $ USD_INR          : num  71.5 71.7 71.8 71.5 72 ...
 $ USD_CAD          : num  1.33 1.33 1.33 1.33 1.33 1.33 1.33 1.33 1.33 1.33 ...
 $ Prediksi         : chr  "Naik" "Turun" "Naik" "Turun" ...
```

Figure 2. Oil Palm Data Structure

Before performing classification, the data used must first be divided into training and testing data. The division of training data is very important to train the performance of the random forest algorithm in machine learning while testing data is used to measure the accuracy of the model trained in the training data [15] . The total data used in this study is 1233 data with a division of the Up class as much as 583 data and the Down class as much

as 650 data. Here researchers divide the data with a ratio of 70% for training data (training data) and the remaining 30% for test data (testing data). This data division is done randomly using the Rstudio IDE software, so that the amount of training data obtained is 862 data, while for test data (testing data) obtained as much as 371 data, to see the distribution results can be seen in Table 1 below.

Table 1. Distribution of training and testing data

	Up	Down	Total
Training Data	411	451	862
Testing Data	172	199	371
Total	583	650	1233

Determine Mtry and Ntree (number of trees)

To classify using random forest, the first step that must be done is to determine the Mtry and Ntree parameters, these parameters are needed to build a model to get the best model with the minimum possible error value. The Mtry parameter is the number of independent variables used to build the tree at each repetition [16], There are three ways to select the value of the Mtry parameter, namely:

$$Mtry = \frac{1}{2}\sqrt{total\ independent\ variabel} \quad (1)$$

$$Mtry = \frac{1}{2}\sqrt{8} = 1{,}4 \approx 1$$

$$Mtry = \sqrt{total\ independent\ variable} \quad (2)$$

$$Mtry = \sqrt{8} = 2{,}8 \approx 3$$

$$Mtry = 2 \times \sqrt{total\ independent\ variable} \quad (3)$$

$$Mtry = 2 \times \sqrt{8} = 2 \times 2{,}8 = 5{,}6 \approx 6$$

After getting the three Mtry values, do the testing then find the smallest error value by classifying it according to the default model in the RStudio IDE. After testing with each Mtry value, the best Mtry parameter value is 1 with an OOB error value on the original data of 45.28%. The following is Table 2 which contains a comparison of OOB error values for each Mtry value on the original data.

Table 2. Error Value for each Mtry

Mtry	OOB Error
1	45.28%
3	46.36%
6	48.79%

The results of the three Mtry parameter values on the original data above can be seen in the graph below. Based on these results, the best Mtry parameter value to be used in the random forest model in the model built is the Mtry parameter with a value of 1.

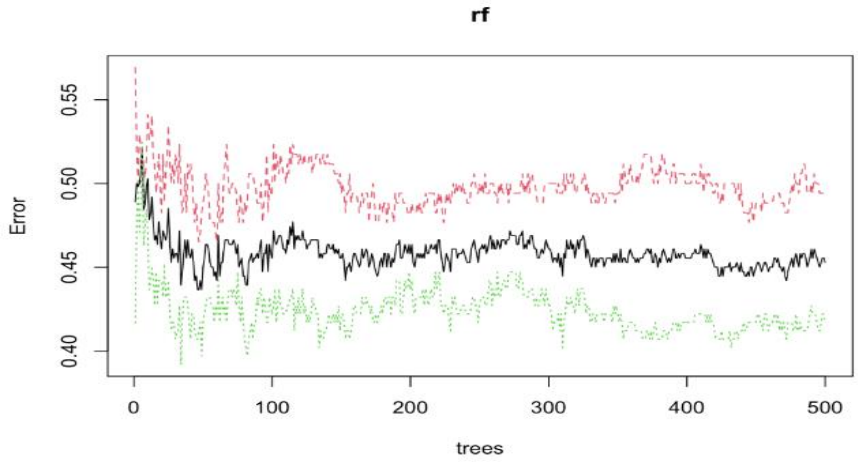

Figure 3. Graph n Error value every Mtry set

Next, we need to find the best number of Ntree (trees) with the smallest error value using the best Mtry parameter value obtained previously. Determining which Ntree parameter value to test is up to the researcher and the values to be tested in this study are Ntree=100, Ntree=200, Ntree=300, Ntree=400, Ntree=500, Ntree=600, Ntree=700, Ntree=800, Ntree=900 and Ntree=1000. The error for each Ntree value can be seen in Table 3 below, where the best Ntree parameter value is 500 with an OOB error value on the original data of 45.28%.

Table 3. Error values set per Ntree

Ntree	OOB Error
100	45.55%
200	45.82%
300	46.09%
400	45.55%
500	45.28%
600	45.82%
700	46.63%
800	45.55%
900	45.82%
1000	46.09%

The Process of Training and Model Building

After determining the best parameter value, namely the value of Mtry = 1 and the parameter value of Ntree = 500, these parameter values will be used in the random forest model for the classification process.

```
randomForest(formula = as.factor(Prediksi) ~ ., data = test,
             mtry = 1, ntree = 500, importance = T, promimity = T)
             Type of random forest: classification
             Number of trees: 500
No. of variables tried at each split: 1
             OOB estimate of error rate: 45.28%
```

Figure 5. Results of the classification of random forest

As can be seen in Figure 5, the random forest model formed is a classification with 500 trees and the number of variables used in each iteration is 1 with an estimated OOB error rate on the training data used is 45.28%.

V. DISCUSSION

Model Accuracy Test

After the model is trained on the training data, the next step is to test the testing data to see the accuracy of the model that has been generated. The prediction results on the testing data can be seen in the table below.

Table 4. Confusion matrix prediction data testing

Predictions	Actual		Precision
	Up	Down	
Up	172	0	1
Down	0	199	1
Recall	1	1	

Each classifier produces a prediction table which is a confusion matrix table that compares the amount of predicted data with the actual data [17] .To measure/evaluate the performance of the prediction data results in the confusion matrix, there are several types of measurements that can be used as benchmarks to see how well the model has been produced in predicting each class, including recall and precision. Recall is the ratio between the number of correct predicted data and the total amount of actual data in a class [18] , while precision is the ratio between the number of correctly predicted data and the total number of predicted data in a class [19]. Table 4 shows the recall and precision for each class, in general it can be seen that the trained model can classify well for each class. The total accuracy of the prediction data results on the test data is formulated as follows:

$$Total\ Accuracy = \frac{\Sigma(true\ prediction)}{\Sigma(all\ predictions)} \quad (4)$$

$$Total\ Accuracy = \frac{\Sigma(172 + 199)}{\Sigma(172 + 0 + 0 + 199)}$$

$$Total\ Accuracy = \frac{1}{1}$$

$$Total\ Accuracy = 1$$

Overall, the accuracy of the random forest model that has been formed using palm oil data in classifying the test data prediction results is 1 or 100%.

Importance Variable

For further analysis, namely sorting the most important independent variables (importance variables), the results of sorting the most important variables can be seen in Table 5 below.

Table 5. Importance Variable

Independent Variable (X)	*Mean Decrease Accuracy*
Palm oil	8.701966
WTI Crude Oil	7.917038
USD_CAD	4.547921
Soybean Oil	3.620018
Canola Oil	2.167288
USD_IDR	2.136008
USD_INR	1.331644
USD_MYR	0.577348

The most important variable of the classification method using random forest algorithm determines how important a variable is in the resulting classification model. One measure to determine significance is the Mean Decrease Accuracy (MDA) which indicates how many additional observations are misclassified if one of the independent variables is not included in the testing process [20] . The higher the Mean Decrease Accuracy (MDA) value of an independent variable, the greater its influence on the accuracy of the classification model. As can be seen in Table 5, the most important independent variable in the resulting random forest classification model is Palm Oil, followed by WTI Crude Oil, USD_CAD, Soybean Oil, Canola Oil, USD_IDR, USD_INR and closed by USD_MYR, all sorted from the largest value to the smallest value.

The results of the decision tree from the classification model using the random forest algorithm on palm oil data can be observed in Figure 6 below.

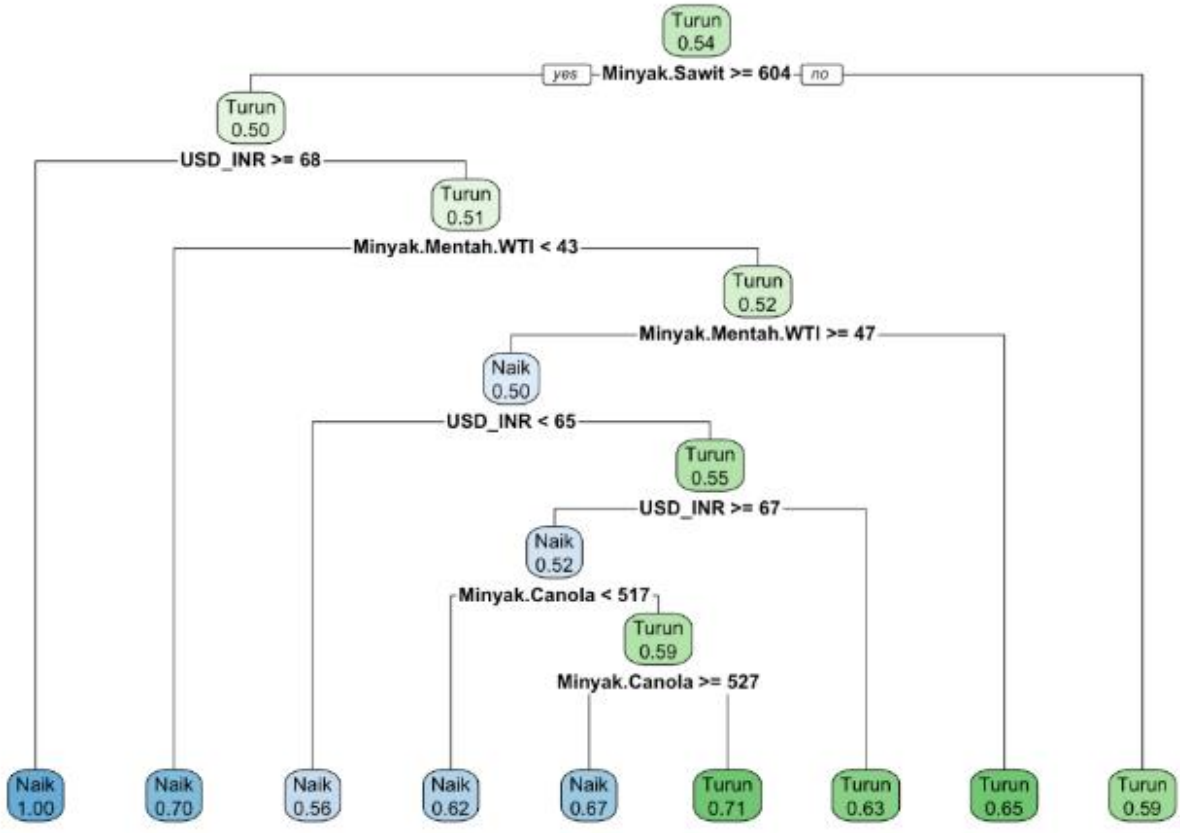

Figure 6. Results of the random forest model decision tree

The blue plot represents the Up class data and the green plot represents the Down class data. For example, from Figure 6 we will classify Up if it has the first condition/conditions that the value of Palm Oil is greater than or equal to 604 and then the second condition/conditions that if the value of USD_INR is greater than or equal to 68 has a 100% chance of Up. The same applies to every other case of Up and Down results, as shown in Figure 6 above.

VI. CONCLUSION

The results of the implementation of the Random Forest algorithm for the application of palm oil data classification, it can be concluded that:

1. To make predictions with the random forest algorithm there are several stages, namely, the first stage of preprocessing, the process of modeling and evaluating the random forest algorithm.
2. The application of the random forest algorithm to palm oil price data using the Mtry parameter of 1 and the Ntree parameter of 500 produces a percentage accuracy rate of 100%.
3. The most influential variable (importance variable) in the classification model using the random forest algorithm produced is the palm oil variable.

REFERENCES

[1] Z. Ismail, A. Khamis, and R. Ali, "Rangkaian Neural dalam Peramalan Harga Minyak Kelapa Sawit," *J. Teknol.*, vol. 39, no. 1, pp. 17–28, 2003, doi: 10.11113/jt.v39.452.

[2] R. E. Caraka, H. Yasin, and A. W. Basyiruddin, "Peramalan Crude Palm Oil (CPO) Menggunakan Support Vector Regression Kernel Radial Basis," *J. Mat.*, vol. 7, no. 1, p. 43, 2017, doi: 10.24843/jmat.2017.v07.i01.p81.

[3] N. Suwaryo, D. Haryadi, D. Marini, U. Atmaja, and A. R. Hakim, "Analisa Data Mining Menggunakan Algoritma Apriori Untuk Mencari Pola Pemakaian Obat," vol. 1, no. November, pp. 1208–1217, 2021.

[4] D. Rahayu, R. C. Wihandika, and R. S. Perdana, "Implementasi Metode Backpropagation Untuk Klasifikasi Kenaikan Harga Minyak Kelapa Sawit," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 4 e-ISSN: 2548-964X, pp. 1547–1552, 2018.

[5] D. I. Puspitasari, "Penerapan Data Mining Menggunakan Perbandingan Algoritma Greedy Dengan Algoritma Genetika Pada Prediksi Rentet Waktu Harga Crude Palm Oil," *Elinvo (Electronics, Informatics, Vocat. Educ.*, vol. 2, no. 1, pp. 21–26, 2017, doi: 10.21831/elinvo.v2i1.13033.

[6] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, vol. 14, no. 2, p. 217, 2018, doi: 10.33480/pilar.v14i2.926.

[7] S. Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan," *Fakt. Exacta*, vol. 11, no. 3, pp. 266–274, 2018, doi: 10.30998/faktorexacta.v11i3.2777.

[8] D. Haryadi and R. Mandala, "Prediksi Harga Minyak Kelapa Sawit Dalam Investasi Dengan Membandingkan Algoritma Naïve Bayes, Support Vector Machine dan K-Nearest Neighbor," *IT Soc.*, vol. 4, no. 1, pp. 28–38, 2019, doi: 10.33021/itfs.v4i1.1181.

[9] A. Kurnia, A. H. Mirza, and A. Andri, "Penerapan Decision Tree Data Mining Pada Produksi Kelapa Sawit PT Hindoli Di Sungai Lilin Kabupaten Musi Banyuasin," *J. Pengemb. Sist. Inf. dan Inform.*, vol. 1, no. 2, pp. 84–99, 2020, doi: 10.47747/jpsii.v1i2.168.

[10] A. Fitri Boy, "Implementasi Data Mining Dalam Memprediksi Harga Crude Palm Oil (CPO) Pasar Domestik Menggunakan Algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatera Utara)," *J. Sci. Soc. Res.*, vol. 4307, no. 2, pp. 78–85, 2020, [Online]. Available: http://jurnal.goretanpena.com/index.php/JSSR.

[11] D. M. U. Atmaja, "Penerapan Algoritma K-Nearest Neighbor Untuk," vol. 1, no. November, pp. 199–208, 2019.

[12] D. Haryadi, A. Rahman, D. Marini, U. Atmaja, and S. Nurgaida, "Implementation of Support Vector Regression for Polkadot Cryptocurrency Price Prediction," vol. 6, no. May, pp. 201–207, 2022.

[13] N. P. A. Widiari, I. M. A. D. Suarjaya, and D. P. Githa, "Teknik Data Cleaning Menggunakan Snowflake untuk Studi Kasus Objek Pariwisata di Bali," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 8, no. 2, p. 137, 2020, doi: 10.24843/jim.2020.v08.i02.p07.
[14] B. S. Ashari and S. C. Otniel, "Jurnal Siliwangi Vol . 5 . No . 2 , 2019 Seri Sains dan Teknologi PERBANDINGAN KINERJA K-MEANS DENGAN DBSCAN Seri Sains dan Teknologi P-ISSN 2477-3891 E-ISSN 2615-4765," vol. 5, no. 2, pp. 64–67, 2019.
[15] A. R. Hakim, D. Marini, U. Atmaja, D. Haryadi, and N. Suwaryo, "Twitter Sentiment Analysis Terhadap Pengguna E-Commerce Menggunakan Text Mining," *SNTEM Semin. Nas. Teknol. Energi dan Miner.*, vol. 1, no. November, pp. 1227–1237, 2021.
[16] A. Supoyo and P. T. Prasetyaningrum, "Analisis Data Mining Untuk Memprediksi Lama Perawatan Pasien Covid-19 Di DIY," *Bianglala Inform.*, vol. 10, no. 1, pp. 21–29, 2022, doi: 10.31294/bi.v10i1.11890.
[17] N. Nosiel, S. Sriyanto, and F. Maylani, "Perbandingan Teknik Data Mining Untuk Prediksi Penjualan Pada UMKM Gerabah," *Pros. Semin. Nas. Darmajaya*, vol. 1, no. 0, pp. 72–86, 2021.
[18] D. Haryadi, D. Marini, U. Atmaja, A. R. Hakim, and N. Suwaryo, "IDENTIFIKASI TINGKAT RESIKO PENYAKIT STROKE MENGGUNAKAN ALGORITMA REGRESI LINEAR BERGANDA," vol. 1, no. November, pp. 1198–1207, 2021.
[19] I. W. Saputro and B. W. Sari, "Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa," *Creat. Inf. Technol. J.*, vol. 6, no. 1, p. 1, 2020, doi: 10.24076/citec.2019v6i1.178.
[20] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.

BIOGRAPHY

Arif Rahman Hakim, Graduated from the Informatics Engineering study program at Pelita Bangsa University in 2017 (S1), and graduated from the Informatics Engineering Masters at the President University in 2020. Currently working as a lecturer at Medika Suherman University.

Dewi Marini Umi Atmaja, Graduated from Informatics Study Program, Jenderal Achmad Yani University in 2018 (S1), and graduated from the Informatics Engineering Masters at the President University in 2020. Currently working as a lecturer at Medika Suherman University.

Amat Basri, Graduated in the Information Technology Study Program (S1) in April, 2002, continued his Masters in Information Systems in Januari, 2015 and graduated in September, 2016. Currently, I am a lecturer in information systems program at Buddhi Dharma University.

Muhamad Syafii, Graduated from the Informatics Engineering study program at Pelita Bangsa University in 2020 (S1), and is pursuing a Masters in Computer Science at Budi Luhur University**.**