Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 8 No. 1, February 2023, pp. 100 – 109
DOI:

Improvement of Data Mining Models using Forward Selection and Backward Elimination with Cryptocurrency Datasets

Indri Tri Julianto*[1], **Dede Kurniadi**[2]
[1,2]*Institut Teknologi Garut, Jalan Mayor Syamsu No. 1, Jagaraya, Garut (0262) 232773*
E-mail : indritrijulianto@itg.ac.id[1], dede.kurniadi.@itg.ac.id[2]*
**Corresponding author*

Fathia Alisha Fauziah[3], **Ricky Rohmanto**[4]
[3]*Universitas Garut, Jl. Raya Semarang, Jl. Hampor Kecamatan No 52A, Garut (0262) 544217,*
[4]*Masoem University, Jl. Raya Cipacing No.22, Cipacing, Jatinangor, Sumedang (022) 7798340*
E-mail : fathiaalisha@uniga.ac.id[3], rickyrohmanto@masoemuniversity.ac.id[4]

Abstract – Cryptocurrency is a digital currency not managed by a state or central bank, and transactions are peer-to-peer. Cryptocurrency is still considered a speculative asset and its price volatility is relatively high, but it is also expected to become an efficient and secure transaction tool in the future. The purpose of this study is to compare and improve the performance of the Data Mining Algorithm model using the Feature Selection-Wrapper with the Binance Coin (BNB) cryptocurrency dataset. The Feature Selection-Wrapper approach used is Forward Selection and Backward Elimination. The algorithms used are Neural Networks, Deep Learning, Support Vector Machines, and Linear Regression. The methodology used is Knowledge Discovery in Databases. The results showed that from a comparison using K-Fold Cross Validation with a value of K=10, the Neural Network Algorithm has the best Root Mean Square Error value of 10,734 +/- 10,124 (micro average: 14,580 +/- 0,000). Then after improving performance using Forward Selection and Backward Elimination in the Neural Network Algorithm, the best performance improvement results are shown by using Backward Elimination with RMSE 5,302 +/- 2,647 (micro average: 5,805 +/- 0,000).

Keywords – algorithms, cryptocurrency, data mining, feature selection-wrapper

1. INTRODUCTION

The cryptocurrency market continues to grow substantially as more public companies incorporate this technology into their product offerings [1], [2]. Cryptocurrencies, often called digital currencies, are becoming increasingly popular because of their decentralization, high level of security, and (partial) anonymity features [3]–[5]. The existence of cryptocurrency is the answer to the need in today's digital era to make transactions that are simple, fast, transparent, and acceptable to both parties [6].

Data mining is extracting information from a data set using machine assistance (Algorithms) [7]–[11]. The models contained in Data Mining are divided into three, Supervised-Learning, Un-Supervised Learning, and Semi-Supervised Learning [12]. This study uses the Supervised Learning model where the dataset already has a label/target. This research uses four algorithms: Neural Network, Deep Learning, Support Vector Machine, and Linear Regression.

Several previous studies have discussed Forward Selection [13], and regarding Backward Elimination [14]–[17]. In general, the previous research succeeded in improving the performance of the Data Mining model using the Forward Selection and Backward Elimination methods. The first research on optimizing the C4.5 Algorithm uses the Forward Selection method for creditworthiness prediction datasets [13]. The study results show that this method's performance of the C4.5 Algorithm has increased by 9.2%. The second study concerns the optimization of the K-Nearest Neighbors Algorithm using the Backward Elimination method for Software Effort Estimation datasets [14]. The study results show that this method performs better when compared to only using the K-Nearest Neighbors Algorithm. The third study regarding the use of Backward Elimination in the K-Nearest Neighbors Algorithm for heart failure datasets [15]. The results showed that using Backward Elimination increased the performance from 94.56% Accuracy to 98.33%, Precision from 93.87% to 97.94%, and Recall from 95.55% to 98.63%. The fourth study concerns the optimization of the K-Nearest Neighbors and Naïve Bayes Algorithms using the Backward Elimination method for customer satisfaction datasets [16]. The results showed that this method works more optimally against the Naïve Bayes Algorithm with an Accuracy of 99.04%, while the resulting Accuracy of the K-Nearest Neighbors Algorithm is 97.28%. Fifth research regarding optimization of the K-Nearest Neighbors, Naïve Bayes, and C4.5 Algorithms using Backward Elimination of the diabetes dataset [17]. The results showed that the Backward Elimination model on the KNN Algorithm had an accuracy of 92.8% and AUC of 0.942, the Naïve Bayes algorithm had an accuracy of 88.0% and AUC of 0.912, the C4.5 algorithm had an accuracy of 96.7% and AUC of 0.956, while the results of the model after optimization is the KNN algorithm with an accuracy of 97.6% and AUC of 0.973, the Naïve Bayes algorithm with an accuracy of 89.4% and AUC of 0.958, the C4.5 algorithm has an accuracy of 97.5% and AUC of 0.988. To make it clearer in understanding previous research, a comparative analysis of the previous technique is presented [5], [18], as shown in Table 1.

Table 1. Comparative Analysis Of Previous Researche

Research	Techniques	Outcome	Dataset
1	C.45	Forward Selection	Creditworthiness Prediction
2	K-NN	Backward Elimination	Software Effort Estimation
3	K-NN	Backward Elimination	Heart Failure
4	K-NN & Naïve Bayes	Backward Elimination	Customer Satisfaction
5	K-NN, Naïve Bayes& C4.5	Backward Elimination	Diabetes
Present	NN, DL, SVM & LR	Forward Selection & Backward Elimination	Cryptocurrecy (BNB)

Note:

(K-NN) K-Nearest Neighbours
(NN) Neural Network
(DL) Deep Learning
(SVM) Support Vector Machine
(LR) Linear Regression

This research fills in the gaps with previous research by using four Supervised Learning Algorithms and using two Feature Selection-Wrapper methods, namely Forward Selection and Backward Elimination. In contrast, in previous studies, each only used one method. Then the dataset used is the BNB cryptocurrency collected from the website www.yahoo.finance.com. The model validation used is the K-Fold Cross Validation with a value of K = 10, a significant test is carried out using the T-Test.

2. RESEARCH METHOD

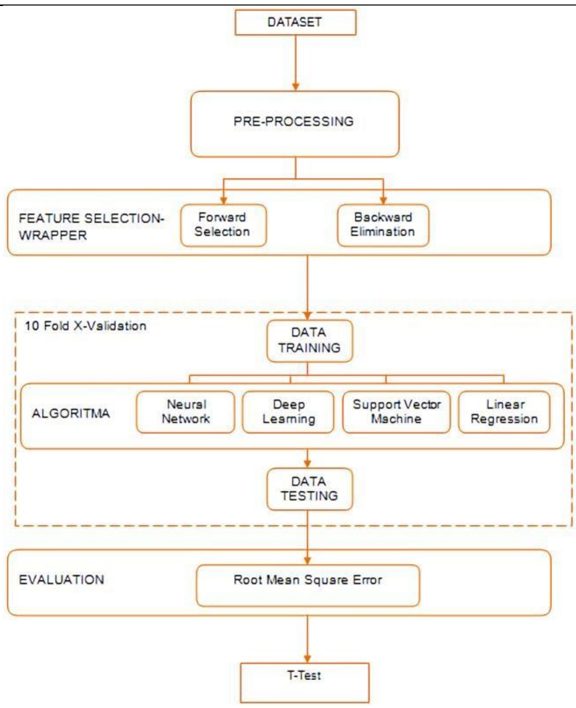


Figure 1. Research Method

Figure 1 explains the flow of the method used in this study. This method consists of four stages: Dataset Collection, Pre-Processing, Modeling, and Evaluation [19], [20].

2.1. Dataset

This stage is carried out by collecting datasets regarding the BNB cryptocurrency through website pages https://finance.yahoo.com/quote/BNB-USD/history? [21]. The dataset collected is population data on the website from July 2017 to January 2023. The increase in the price of BNB occurred in the range 2021-2022, as shown in Figure 2, the initial dataset is shown in Figure 3, and the total data is 469, which consists of seven attributes, as shown in Table 2.

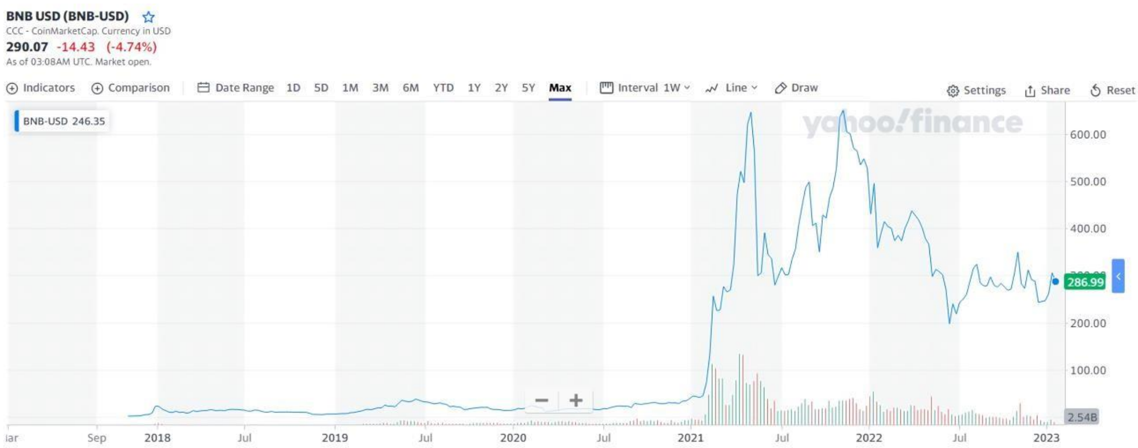


Figure 2. Chart BNB Price [21]

Table 2. Dataset Description [3]

No	Atributes	Description
1	Date	Format Day - Month - Year
2	Open	Opening price in trade
3	High	Highest price in trade
4	Low	Lowest Price in trade
5	Close	Closing Price in trade
6	Volume	Transaction volume is usually in the number of sheets
7	Adjusted Close	Closing price adjusted for corporate actions such as rights issue, stock split or stock reverse

2.2. Pre-Processing

This stage is the dataset preparation stage before the Data Mining process. I'll do data cleaning and select appropriate attributes at this stage. Data preparation in question, namely 1) Data Cleansing is the process of cleaning data from empty values, inconsistent, empty attributes such as missing values and noisy data; 2) Data Integration is the process of merging data into one archive; 3) Data Reduction is the process to eliminate unnecessary attributes [22].

2.3. Feature Selection-Wrapper

This stage is the stage of the method used to improve the performance of the model built using the four Algorithms that have been mentioned. The Feature Selection-Wrappers used are Forward Selection and Backward Elimination. The way Forward Selection works starts with a set of attributes to be deleted. Attributes are tested individually, and the best attribute with the most remarkable accuracy is selected. Then run the next test iteration continuously, pausing until the tested attribute does not significantly affect the accuracy [23], [24]. Whereas the Backward Elimination method selects the future variables by testing all variables and then removing the variables that are considered irrelevant [17].

2.4. K-Fold Cross Validation

K-Fold Cross Validation is a validation method that divides data into k parts and classifies them based on different factors. In each experiment, what used test data and part k-1 was training data. For example, k is used 10, then for data testing 10% of the training data becomes 90% of the total data [17].

2.5. Root Mean Square Error

Root Mean Square Error (RMSE) is an alternative evaluation method on a forecasting technique used to measure the accuracy of a model's forecasting results. The value generated by the RMSE is the root mean square of the number of errors in the forecast model [25].

2.6. T-Test

This parametric statistical test method indicates how far an individual's influence from the independent variable is in explaining the dependent variable. A t-test was performed at a significant level of 0.05 (α = 5%) [26].

3. RESULTS AND DISCUSSION

Date	Open	High	Low	Close	Adj Close	Volume
Aug 1, 2017	0.105	2.955	0.096	2.204	2.204	786763181
Sep 1, 2017	2.202	2.849	0.527	1.284	1.284	221387190
Oct 1, 2017	1.284	2.033	1.096	1.313	1.313	153469470
Nov 1, 2017	2.053	2.174	1.463	1.997	1.997	371780810
Dec 1, 2017	1.997	11.302	1.923	8.636	8.636	1722931600
Jan 1, 2018	8.630	24.912	7.959	11.145	11.145	5901238384
Feb 1, 2018	11.178	11.871	5.590	10.438	10.438	1811644896
Mar 1, 2018	10.448	14.839	7.174	11.056	11.056	2934023616
Apr 1, 2018	11.152	15.926	10.378	14.312	14.312	3189054696
May 1, 2018	14.315	16.221	11.703	14.190	14.190	2260288580
Jun 1, 2018	14.252	17.438	13.544	14.657	14.657	2580554212
Jul 1, 2018	14.676	14.869	11.648	13.775	13.775	1467978204
Aug 1, 2018	13.770	14.455	8.663	11.014	11.014	1080238800

Figure 3. Preliminary Dataset

This initial dataset will be pre-processed by first looking at the level of correlation between attributes using the Correlation Matrix. This is done to determine which attributes will be used in the Data Mining modeling process. The reference level of correlation between attributes is presented as shown in Table 3.

Table 2. Indicator Correlation Coefficient [26]

Coefficient Range	Strange of Association
±0.91to ±1.00	Very Strong
±0.71to ±0.90	High
±0.41to ±0.70	Moderate
±0.21to ±0.40	Small but definite relationship
±0.01to ±0.20	Slight, almost negligible

The results of tests conducted on the BNB cryptocurrency dataset are presented as shown in Figure 4.

Attribut...	Date	Open	High	Low	Close	Adj Close	Volume
Date	1	?	?	?	?	?	?
Open	?	1	0.964	0.949	0.925	0.925	0.716
High	?	0.964	1	0.943	0.976	0.976	0.841
Low	?	0.949	0.943	1	0.968	0.968	0.656
Close	?	0.925	0.976	0.968	1	1	0.793
Adj Close	?	0.925	0.976	0.968	1	1	0.793
Volume	?	0.716	0.841	0.656	0.793	0.793	1

Figure 4. Correlation Matrix Dataset

Based on Figure 4, we can see that the level of correlation between attributes is at the Moderate to Very Strong Association level. This indicates that each attribute in the dataset can be used in the Data Mining modeling process. The next step is to assign a Label/Target to the Close attribute in the BNB dataset, as shown in Figure 5.

Close	Open	High	Low	Adj Close	Volume
2.204	0.105	2.955	0.096	2.204	786763181
1.284	2.202	2.849	0.527	1.284	221387190
1.313	1.284	2.033	1.096	1.313	153469470
1.997	2.053	2.174	1.463	1.997	371780810
8.636	1.997	11.302	1.923	8.636	1722931600
11.145	8.630	24.912	7.959	11.145	5901238384
10.438	11.178	11.871	5.590	10.438	1811644896
11.056	10.448	14.839	7.174	11.056	2934023616
14.312	11.152	15.926	10.378	14.312	3189054696
14.190	14.315	16.221	11.703	14.190	2260288580
14.657	14.252	17.438	13.544	14.657	2580554212
13.775	14.676	14.869	11.648	13.775	1467978204
11.014	13.770	14.455	8.663	11.014	1080238800
10.018	11.027	11.542	9.037	10.018	676271200

Figure 5. Pre-Processing Result

The model built is the fourth comparison stage of the Algorithm which will produce an output in the form of a Root Mean Square Error (RMSE) value. This model was created using the Rapidminer Studio Application. The modeling results are presented as shown in Figure 6.

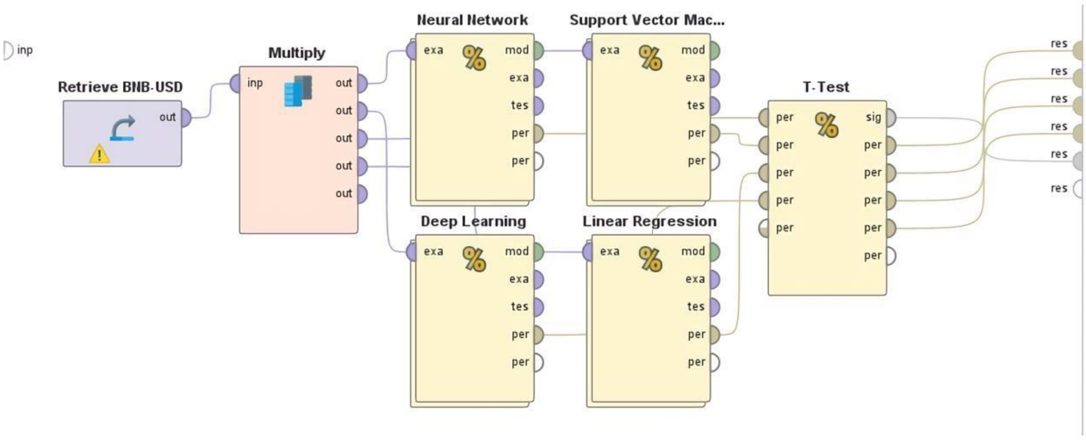


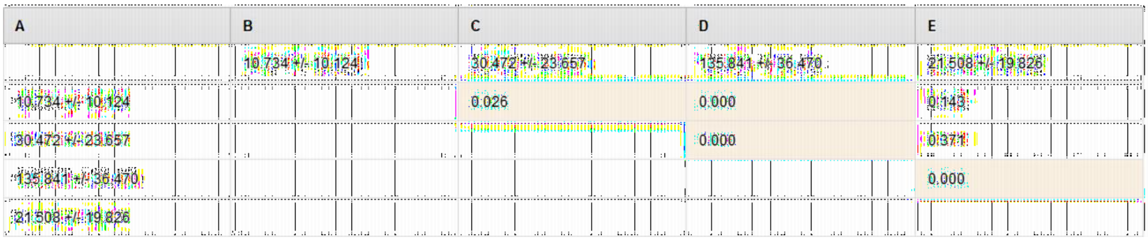
Figure 6. Model Process

After the Algorithm comparison process through the modeling is run, what will obtain the best Algorithm. The indicators used as an assessment are the RMSE value and the significance test through the T-Test. The results of the RMSE values are presented in Table 3, and the results of the T-Test are in Figure 7.

Table 3. Root Mean Sqaure Error

Algorithms	RMSE
Neural Network	10.734 +/- 10.124 (micro average: 14.580 +/- 0.000)
Deep Learning	30.472 +/- 23.657 (micro average: 38.557 +/- 0.000)
Support Vector Machine	135.841 +/- 36.470 (micro average: 141.431 +/- 0.000)
Linear Regression	21.508 +/- 19.826 (micro average: 28.817 +/- 0.000)

Based on Table 3, the algorithm that has the most optimal RMSE value using the BNB dataset is a Neural Network, namely 10,734 +/- 10,124 (micro average: 14,580 +/- 0,000), and an algorithm with the least optimal RMSE value is Support Vector Machine, namely 135,841 +/- 36,470 (micro average: 141,431 +/- 0,000).



A	B	C	D	E
	10.734 +/- 10.124	30.472 +/- 23.657	135.841 +/- 36.470	21.508 +/- 19.826
10.734 +/- 10.124		0.026	0.000	0.143
30.472 +/- 23.657			0.000	0.371
135.841 +/- 36.470				0.000
21.508 +/- 19.826				

Figure 7. Model Process

Note:

B : Neural Network

C : Deep Learning

D : Support Vector Machine

E : Linear Regression

: No Significant Difference

: Significant Difference

It can be seen in Figure 7 that the column that is colored pink shows that there is no significant difference in the relationship between the algorithms. In contrast, the column not colored pink indicates a significant difference because the Alpha value is > 0.050, so the profit ranking of the algorithm is presented as shown in Table 4.

Table 4. Algorithms Rating

No	Algorithms	RMSE	T-Test
1	Neural Network	10.734 +/- 10.124 (micro average: 14.580 +/- 0.000)	No Significant Difference
1	Deep Learning	30.472 +/- 23.657 (micro average: 38.557 +/- 0.000)	No Significant Difference
1	Linear Regression	21.508 +/- 19.826 (micro average: 28.817 +/- 0.000)	No Significant Difference
2	Support Vector Machine	135.841 +/- 36.470 (micro average: 141.431 +/- 0.000)	Significant Difference

After it is known that Neural Network, Deep Learning, and Linear Regression are in the same rank based on the results of the T-Test, the Neural Network Algorithm was chosen because it has the lowest RMSE value of 10,734 +/- 10,124 (micro average: 14,580 +/- 0,000) . What will improve the performance of the Neural Network Algorithm by using the Feature Selection-Wrapper approach with Forward Selection and Backward Elimination techniques. The model built is presented as shown in Figure 8.

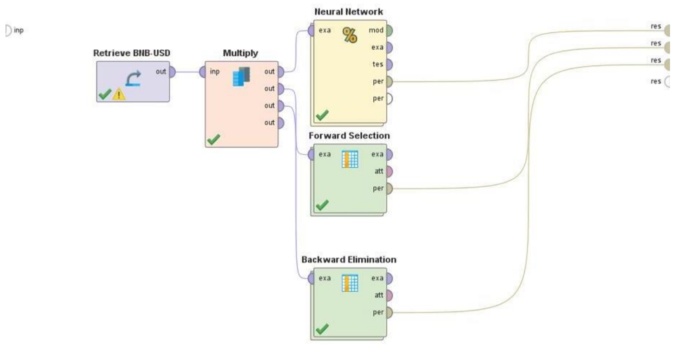


Figure 8. Model Process Feature Selection-Wrapper

The model describes a comparison made to the Neural Network Algorithm to determine performance improvements using Forward Selection and Backward Elimination. The results of these comparisons are presented as shown in Table 5, and the graphs are presented as shown in Figure 9.

Table 5. Optimization Performance using Feature Selection-Wrapper

No	Algorithms	RMSE
1	Neural Network	10.734 +/- 10.124 (micro average: 14.580 +/- 0.000)
2	Neural Network + Forward Selection	5.880 +/- 3.763 (micro average: 6.808 +/- 0.000)
3	Neural Network + Backward Elimination	5.302 +/- 2.647 (micro average: 5.805 +/- 0.000)

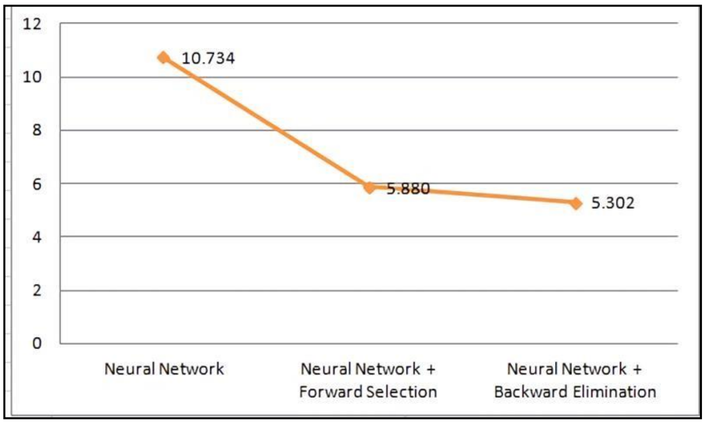


Figure 9. Result Feature Selection-Wrapper

4. CONCLUSION

This study concludes that the results of a comparison of the four Algorithms show that the Neural Network has the most optimal RMSE value, namely 10,734 +/- 10,124 (micro average: 14,580 +/- 0,000), then the significant test results using the T-Test show that the Neural Network, Deep Learning, and Linear Regression are in the same ranking order (No Significant Difference), while the relationship between the Support Vector Machine and the other three Algorithms is Significant Difference. The results of improving the performance of the Neural Network Algorithm using Forward Selection and Backward Elimination show that the optimal improvement value is indicated by Backward Elimination with an RMSE value of 5,302 +/- 2,647 (micro average: 5,805 +/- 0,000), while Forward Selection has an RMSE value of 5,880 +/ - 3,763 (micro average: 6,808 +/- 0,000).

This study has several limitations, and the first is that this study focuses on comparing two methods, namely Forward Selection and Backward Elimination. Future research can add other methods as a comparison, such as using Feature Extraction. Both of these studies used four Supervised Learning Algorithms. Future research can add different algorithms to get more optimal results.

REFERENCES

[1] P. Katsiampa, L. Yarovaya, and D. Zi, "Journal of International Financial Markets , High-frequency connectedness between Bitcoin and other top-traded crypto assets during the COVID-19 crisis," vol. 79, no. November 2021, 2022, doi: 10.1016/j.intfin.2022.101578.

[2] F. Xu, E. Bouri, and O. Cepni, "Blockchain and crypto-exposed US companies and major cryptocurrencies : The role of jumps and co-jumps," *Financ. Res. Lett.*, vol. 50, no. April, p.

103201, 2022, doi: 10.1016/j.frl.2022.103201.
[3] K. Vikram, N. Sivaraman, and D. P. Balamurugan, "Crypto Currency Market Price Prediction Using Data Science Process," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 2, pp. 1451–1454, 2022, doi: 10.22214/ijraset.2022.40521.
[4] Y. Wei, Y. Wang, B. M. Lucey, and S. A. Vigne, "Cryptocurrency Uncertainty and Volatility Forecasting of Precious Metal Futures Markets," *J. ofCommodityMarkets J.*, vol. 29, pp. 1–16, 2023.
[5] P. Solana, F. Orte, and M. J. S, "A Random Forest-Based Model for Crypto Asset Forecasts in Futures Markets with out-of-Sample Prediction," *Res. Int. Bus. Financ.*, vol. 64, no. November 2022, pp. 1–29, 2023, doi: 10.1016/j.ribaf.2022.101829.
[6] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, "Comparison Of Data Mining Algorithm For Forecasting Bitcoin Crypto Currency Trends," *JUTIF*, vol. 3, no. 2, pp. 245–248, 2022.
[7] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data*. 2014.
[8] C. C. Aggarwal, *Data Mining : The Textbook*. New York: Springer, 2015.
[9] A. S. Yudistira and A. Nugroho, "Prediction Of The English Premier League Champion Team For The 2021 / 2022 Season Using The Naïve Bayes Method," *JUTIF*, vol. 3, no. 5, pp. 1239–1243, 2022.
[10] Z. R. S. Elsi *et al.*, "Utilization of Data Mining Techniques in National Food Security during the Covid-19 Pandemic in Indonesia," *J. Phys. Conf. Ser.*, pp. 1–7, 2020, doi: 10.1088/1742-6596/1594/1/012007.
[11] A. A. Argasah and D. Gustian, "Data Mining Analysis To Determine Employee Salaries According To Needs Based On The K-Medoids Clustering Algorithm Analisis Data Mining Untuk Menentukan Gaji Karyawan Sesuai Penilaian Kemampuan Menggunakan Algoritma K-Medoids," *JUTIF*, vol. 3, no. 1, pp. 29–35, 2022.
[12] A. R. Muhajir, E. Sutoyo, and I. Darmawan, "Forecasting Model Penyakit Demam Berdarah Dengue Di Provinsi DKI Jakarta Menggunakan Algoritma Regresi Linier Untuk Mengetahui Kecenderungan Nilai Variabel Prediktor Terhadap Peningkatan Kasus," *Fountain Informatics J.*, vol. 4, no. 2, pp. 33–40, 2019.
[13] I. Ubaedi and Y. M. Djaksana, "Optimasi Algoritma C4.5 Menggunakan Metode Forward Selection Dan Stratified Sampling Untuk Prediksi Kelayakan Kredit," *JSiI (Jurnal Sist. Informasi)*, vol. 9, no. 1, pp. 17–26, 2022, doi: 10.30656/jsii.v9i1.3505.
[14] W. Nugroho, "Optimasi Metode K-Nearest Neighbours dengan Backward Elimination Menggunakan Dataset Software Effort Estimation," *Bianglala Inform.*, vol. 8, no. 2, pp. 129–133, 2020.
[15] I. R. Amilia, H. Oktavianto, and G. Abdurrahman, "Penerapan Backward Elimination Untuk Seleksi Fitur Pada Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Gagal Jantung," *J. Smart Teknol.*, vol. 1, no. 1, pp. 1–9, 2021.
[16] Yunitasari, H. S. Hopipah, and R. Mayasari, "Optimasi Backward Elimination untuk Klasifikasi Kepuasan Pelanggan Menggunakan Algoritme k-nearest neighbor (k-NN) and Naive Bayes," *Technomedia J.*, vol. 6, no. 1, pp. 99–110, 2021, doi: 10.33050/tmj.v6i1.1531.
[17] M. A. Wiratama and W. M. Pradnya, "Optimasi Algoritma Data Mining Menggunakan Backward Elimination untuk Klasifikasi Penyakit Diabetes," *J. Nas. Pendidik. Tek. Inform.*, vol. 11, no. 1, pp. 1–12, 2022, doi: 10.23887/janapati.v11i1.45282.
[18] I. T. Julianto, "Design And Build Virtual Reality Photography Web-Based To Support Tourism," *J. Electr. Electron. Information, Commun. Technol.*, vol. 3, no. 2, p. 58, Oct. 2021, doi: 10.20961/jeeict.3.2.54833.
[19] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, "Comparison Of Classification

Algorithm And Feature Selection In Perbandingan Algoritma Klasifikasi Dan Feature Selection," *JUTIF*, vol. 3, no. 3, pp. 739–744, 2022.

[20] I. T. Julianto, D. Kurniadi, M. R. Nashrulloh, and A. Mulyani, "Data Mining Algorithm Testing For SAND Metaverse Forecasting," *J. Appl. Intell. Syst.*, vol. 7, no. 3, pp. 259–267, 2022.

[21] yahoo finance, "BNB USD (BNB-USD)," *yahoo.finance.com*, 2023. https://finance.yahoo.com/quote/BNB-USD/history?period1=1500940800&period2=1674086400&interval=1mo&filter=history&frequency=1mo&includeAdjustedClose=true.

[22] Mikhael, F. Andreas, and U. Enri, "Perbandingan Algoritma Linear Regression, Neural Network, Deep Learning, Dan K-Nearest Neighbor (K-Nn) Untuk Prediksi Harga Bitcoin," *JSI J. Sist. Inf.*, vol. 14, no. 1, pp. 2450–2464, 2022, [Online]. Available: http://ejournal.unsri.ac.id/index.php/jsi/index.

[23] Han and Kamber, *Data Mining Concepts and Technique*. San Francisco: Diane Cerra, 2006.

[24] D. Kurniadi, F. Nuraeni, and S. M. Lestari, "Implementasi Algoritma Naïve Bayes Menggunakan Feature Forward Selection dan SMOTE Untuk Memprediksi Ketepatan Masa Studi Mahasiswa Sarjana," *J. Sist. Cerdas*, vol. 05, no. 02, pp. 63–82, 2022.

[25] F. I. Sanjaya and D. Heksaputra, "Prediksi Rerata Harga Beras Tingkat Grosir Indonesia dengan Long Short Term Memory," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 2, pp. 163–174, 2020, doi: 10.35957/jatisi.v7i2.388.

[26] T. Sellar and A. A. Arulrajah, "The Role of Social Support on Job Burnout in the Apparel Firm," *Int. Bus. Res.*, vol. 12, no. 1, p. 110, 2018, doi: 10.5539/ibr.v12n1p110.