Journal of Applied Intelligent System (e-ISSN : 2502-9401 | p-ISSN : 2503-0493)
Vol. 7 No. 3, December 2022, pp. 259 – 267

Data Mining Algorithm Testing For SAND Metaverse Forecasting

Indri Tri Julianto*[1], Dede Kurniadi[2], Muhammad Rikza Nashrulloh[3], Asri Mulyani[4]
[1,2,3,4]Institut Teknologi Garut, Jalan Mayor Syamsu No 1 Garut, (0262) 232773
E-mail : indritrijulianto@itg.ac.id[1], dede.kurniadi.@itg.ac.id[2], rikza@itg.ac.id[3], asrimulyani@itg.ac.id[4]*
**Corresponding author*

Abstract – Metaverse is a technology that allows us to buy virtual land. In the future life in the real world can be duplicated into the Metaverse to increase efficiency, effectiveness, and a world without being limited by space and time. To buy land in the Metaverse, one can be done by using SAND. SAND is a crypto asset from a game called The Sandbox which functions as a transaction tool where in that game we can buy land and build it for various purposes just like we can store our Non-Fungible Tokens there. Metaverse is a digital business that will promise in the future because it offers easy and fast transactions. This study aims to compare the exact algorithm for making predictions about the SAND cryptocurrency used to buy Metaverse land. 7 algorithms are being compared, namely Deep Learning, Linear Regression, Neural Networks, Support Vector Machines, Generalized Linear Models, Gaussian Process, and K-Nearest Neighbors. The research method used is Knowledge Discovery in Databases. The research results show that the Support Vector Machines Algorithm has the most optimal Root Means Square Error value, root_mean_squared_error: 0.022 +/- 0.062 (micro average: 0.062 +/- 0.000). Based on this comparison, the Support Vector Machines Algorithm is suitable for predicting SAND Metaverse prices.

Keywords – algorithms, data mining, metaverse, SAND

1. INTRODUCTION

The Metaverse technology began to be widely discussed when the founder of Facebook, Mark Zuckerberg, officially transformed his product with the name Meta. His ideas regarding this technology were submitted at the end of June 2021 [1]. Metaverse is a three-dimensional virtual world technology where Avatars or our profile representations in that world can carry out social, political, economic and cultural activities [2]. Even though technology cannot be fully utilized yet, the Metaverse has vast and promising potential in the future, where real and virtual can coexist without any restrictions [2], [3].

The Metaverse concept was adapted from the novel "Snow Crash" by Neal Stephenson in 1992. This concept describes a three-dimensional virtual world where people can interact without the physical boundaries of the real worl [4]. The Metaverse has now become an attraction for the Tech Industry due to its accelerating pace of development Blockchain, Internet of Things (IoT), Virtual Reality/Augmented Reality, Artificial Intelligence (AI), Cloud/Edge Computing, and so on [5].

Land purchases and other business transactions in Metaverse can be done using cryptocurrencies, such as The Sandbox which is a company from the Blockchain game that has SAND crypto currency which is commonly used as a transaction tool in Metaverse [5].

The popularity of Metaverse is the background of this research with the aim of conducting a comparison of the appropriate algorithms for predicting SAND Metaverse prices. This research can be used as a reference for those who have an interest in buying or investing in the Metaverse virtual world.

Data Mining is a process of searching for new knowledge from a large amount of data [6]. Data Mining has five main roles, namely Estimation, Prediction, Classification, Clustering and Association [7]. Forecasting is a process that might occur in the future based on past and present existing datasets [6]. There are 7 algorithms used in this study, namely Deep Learning (DL), Linear Regression (LR), Neural Network (NN), Support Vector Machine (SVM), Generalized Linear Model (GLM), Gaussian Process (GP), K- Nearest Neighbors (K-NN).

Several previous studies have discussed the predictions of a dataset. The first research is about predicting the price of the Bitcoin currency using LSTM and sentiment analysis on social media [8]. The results of this study indicate a Root Means Square Error value of 335.201882 with an epoch of 10. The second study is about Bitcoin price prediction using the Extreme Learning Machine (ELM) method with Artificial Bee Colony (ABC) Optimization [9]. The results show that the ELM-ABC parameters get the best combination, namely the number of features is 12, hidden neurons are 20, bee populations are 20, and iterations are 5. The combination produces an average MAPE value of 1.96983% and an accuracy of 98.03017%, while ELM with a value of 2.70401% and 97.29599%. The third study concerns the application of short-term predictions of Bitcoin prices using the Autoregressive Integrated Moving Average (ARIMA) method [10]. The results show that the ARIMA model (3, 1, 3) produces predictions with the smallest MAPE value compared to other model candidates where the average MAPE value produced is 0.84 and the range of values is 1.34 for predictions on the first day and 0.98 for predictions seventh day. Then the ARIMA model (3, 1, 3) is able to produce predictions with good accuracy and is suitable for use as a Bitcoin prediction method for the next one to seven days. The fourth research regarding bitcoin price predictions uses the Random Forest method in case studies of random data at the start of the Covid-19 pandemic [11]. The results showed a MAPE value of 1.50% with an accuracy of 98.50%, where it is known that the Random Forest Algorithm is a model that can produce good performance in terms of predictions, especially for random data. The fifth research regarding Bitcoin price prediction in Blockchain information uses the Long-Short Term Memory (LSTM) method [12]. The results showed that the model with 20 neurons and 500 epochs had the smallest MSE value so it had a prediction with an accuracy rate of 91.07%. In brief, the five studies are presented in the form of a Research Roadmap, as shown in Table 1.

Table 1. Research Roadmap

Research	Algorithms	Dataset	Outcome
1	LSTM	Bitcoin	Forecasting
2	ELM	Bitcoin	Forecasting
3	ARIMA	Bitcoin	Forecasting
4	Random Forest	Bitcoin	Forecasting
5	LSTM	Bitcoin	Forecasting

Note:
LSTM (Long-Short Term Memory)
ELM (Extreme Learning Machine)
ARIMA (Autoregressive Integrated Moving Average)

This research fills the gap by using the SAND Metaverse dataset, whereas previous research used the Bitcoin dataset. 7 algorithms were compared, where in the previous study only used 1 algorithm. It aims to find the most appropriate algorithm in forecasting SAND Metaverse prices. The validation model that will be used is K-Fold Cross Validation where the K value chosen is 10, then the evaluation model uses Root Mean Square Error (RMSE) and a different test is carried out using the T-Test.

2. RESEARCH METHOD

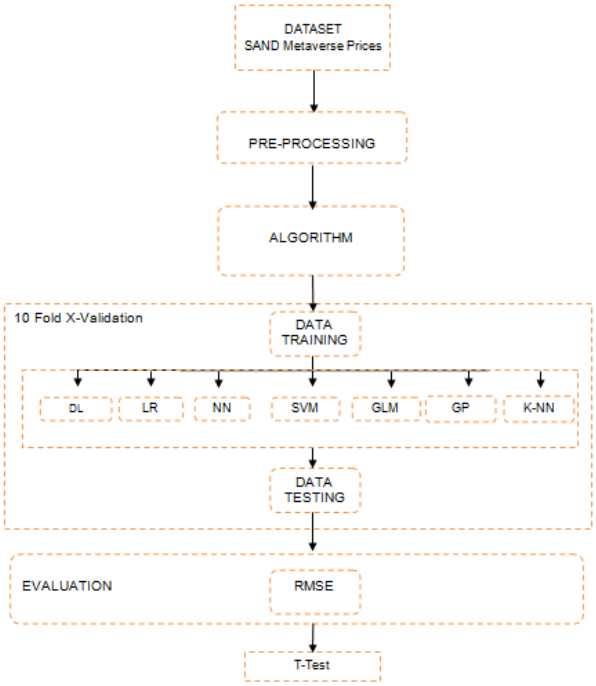


Figure 1. Research Method

The research method uses Knowledge Discovery in Databases, which is a method for searching for knowledge from one or several databases [13]. This method consists of four stages, namely Dataset Collection, Pre-Processing, Modeling, and Evaluation [14].

The first stage is carried out by downloading the dataset from the https://finance.yahoo.com/quote/SAND-USD/history/ page [15]. The dataset is population data owned by Yahoo Finance, starting from November 22 2021 to November 22 2022 with a total of 367 data. This dataset consists of 7 attributes, namely:

1. Date = Date (Format Day - Month - Year);
2. Open = Opening Price;
3. High = Highest Price;
4. Low = Lowest Price;
5. Close = Closing Price;
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.

The second stage is Pre-Processing, which is an important stage before entering the Data Mining modeling process, where data cleaning and attribute selection will be carried out as needed [16]. The method in Pre-Processing is as follows [17]:

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 to merging data into one archive;
3. Data Reduction: is the process to eliminating unnecessary attributes.

The third stage is Modeling using 7 Algorithms which will be compared to find the Algorithm with the lowest RMSE value, so that it can be used for Forecasting SND Metaverse prices. Then proceed with model validation using K-Fold Cross Validation (KVC). KCV will partition k parts of data and do as many k iterations. Whenever a part of the dataset is selected, the first k – 1 are used as learning data while the rest are used as testing data. This process will be repeated k-times and then the average deviation (error) value of the k different test results will be calculated. The illustration for KVC with a value of K-10 is presented in the form of pictures, as shown in Figure 2.

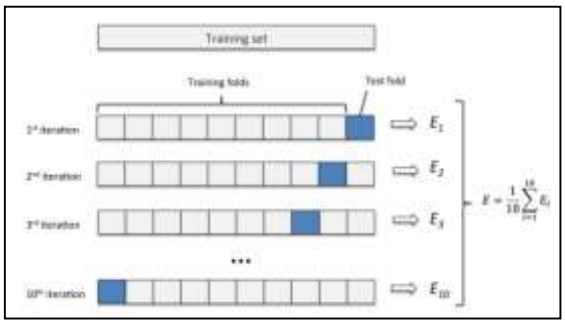


Figure 2. K-Fold Cross Validation (K-10) [17]

The fourth stage is Evaluation, where the measurement of the accuracy of the predictions of the Algorithm being tested is carried out. This research uses Root Mean Square Error evaluation. The RMSE value aims to determine the extent of a model's error rate against the regression line. The smaller the RMSE value, the better [17]. Then proceed with the T-test which is a parametric statistical test method that shows how far the influence of the individual from the independent variable is in explaining the dependent variable. The T-Test was carried out with a significance level of 0.05 (α = 5%) [18]. The following are the criteria for testing the hypothesis T-Test:

1. If the significance value is > 0.05 then the null hypothesis (H0) is accepted and the alternative hypothesis (H1) is rejected. This means that partially the independent variable has no significant effect on the dependent variable;
2. If the significant value is <0.05, then the null hypothesis (H0) is rejected and the alternative hypothesis (H1) is accepted. This means that the independent variable partially has a significant effect on the dependent variable.

3. RESULTS AND DISCUSSION

The results of the Pre-Processing stages are presented as shown in Figure 3.

Date	Open	High	Low	Close	Adj Close	Volume
Nov 22, 2021	4.010	4.971	3.761	4.956	4.956	3130548623
Nov 23, 2021	4.970	5.679	4.879	5.364	5.364	4615278288
Nov 24, 2021	5.357	8.000	5.263	6.334	6.334	5425767404
Nov 25, 2021	7.463	8.442	6.598	8.402	8.402	11077988546
Nov 26, 2021	7.173	7.773	6.384	6.920	6.920	5692742260
Nov 27, 2021	6.989	7.104	6.222	6.556	6.556	3146113846
Nov 28, 2021	6.559	7.594	5.801	7.491	7.491	6242370702
Nov 29, 2021	7.530	7.941	6.947	6.972	6.972	4545318881
Nov 30, 2021	7.000	7.188	6.662	6.773	6.773	2883215549
Dec 1, 2021	6.794	7.033	6.336	6.578	6.578	2104455950
Dec 2, 2021	6.583	6.855	6.024	6.711	6.711	2624486759
Dec 3, 2021	6.696	6.961	5.730	6.035	6.035	2253228786
Dec 4, 2021	6.043	6.195	4.186	6.062	6.062	4922178504
Dec 5, 2021	6.085	6.151	5.226	5.429	5.429	1996358755
Dec 6, 2021	5.419	5.697	4.779	5.509	5.509	2471720119

Figure 3. Preliminary Data

The next step is to find the level of correlation between the attributes in the dataset. The matrix for correlation values between attributes is presented as shown in Table 2.

Table 2. Rule Of Thumb about Correlation Coefficient [18]

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

Then the Correlation Matrix results for the SAND Metaverse dataset are presented as shown in Figure 4.

Attribut...	Date	Open	High	Low	Close	Adj Close	Volume
Date	1	?	?	?	?	?	?
Open	?	1	0.995	0.996	0.993	0.993	0.656
High	?	0.995	1	0.993	0.997	0.997	0.696
Low	?	0.996	0.993	1	0.994	0.994	0.630
Close	?	0.993	0.997	0.994	1	1	0.687
Adj Close	?	0.993	0.997	0.994	1	1	0.687
Volume	?	0.656	0.696	0.630	0.687	0.687	1

Figure 4. Correlation Matrix

In Figure 4 it can be seen that the values for this dataset are in the moderate to the very strong range. This value indicates that the collected dataset has a good level of correlation between its attributes. The next step is to label the close attribute before entering the modeling stage. The results of labeling the close attribute are presented as shown in Figure 5.

Close	Date	Open	High	Low	Adj Close	Volume
4.956	Nov 22, 2021	4.010	4.971	3.761	4.956	3130548623
5.364	Nov 23, 2021	4.970	5.679	4.879	5.364	4615278288
6.334	Nov 24, 2021	5.357	8.000	5.263	6.334	5425767404
8.402	Nov 25, 2021	7.463	8.442	6.698	8.402	11077988546
6.920	Nov 26, 2021	7.173	7.773	6.384	6.920	5692742260
6.556	Nov 27, 2021	6.989	7.104	6.222	6.556	3146113846
7.491	Nov 28, 2021	6.559	7.594	5.801	7.491	6242370702
6.972	Nov 29, 2021	7.530	7.941	6.947	6.972	4545318881
6.773	Nov 30, 2021	7.000	7.188	6.662	6.773	2883215549
6.578	Dec 1, 2021	6.794	7.033	6.336	6.578	2104455950
6.711	Dec 2, 2021	6.583	6.855	6.024	6.711	2624486759
6.035	Dec 3, 2021	6.696	6.961	5.730	6.035	2253228786
6.062	Dec 4, 2021	6.043	6.195	4.186	6.062	4922178604
5.429	Dec 5, 2021	6.085	6.151	5.226	5.429	1996358755
5.509	Dec 6, 2021	5.419	5.697	4.779	5.509	2471720119

Figure 5. The Dataset Pre-Processing Result

The next stage is to create a model in the Rapidminer Studio application. The model building is presented in the form of an image, as shown in Figure 6.

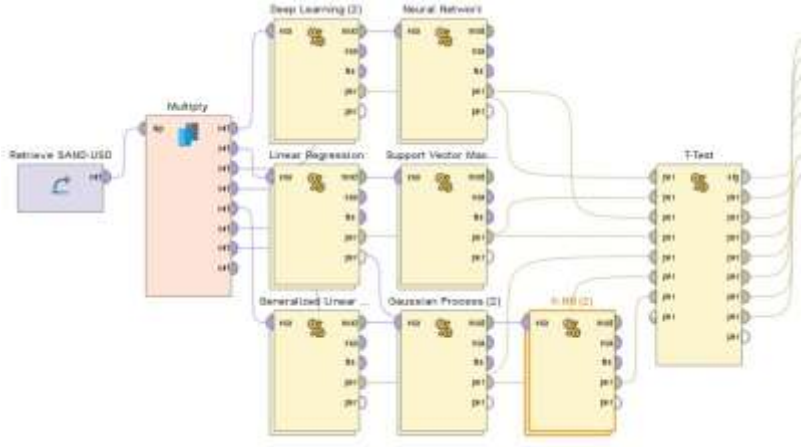


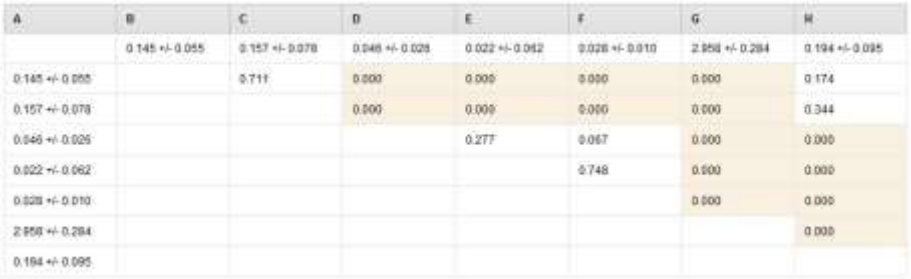
Figure 6. Model Process with T-Test

After executing the model, the RMSE value and also the T-Test will come out. The results of testing the model are presented in tabular form, as shown in Table 3, as well as the results of the T-Test in Figure 7.

Table 3. Root Mean Square Error

No	Algorithms	RMSE
1	Deep Learning	0.145 +/- 0.055 (micro average: 0.154 +/- 0.000)
2	Linear Regression	0.157 +/- 0.078 (micro average: 0.173 +/- 0.000)
3	Neural Network	0.046 +/- 0.026 (micro average: 0.052 +/- 0.000)
4	Support Vector Machine	0.022 +/- 0.062 (micro average: 0.062 +/- 0.000)
5	Generalized Linear Model	0.028 +/- 0.010 (micro average: 0.030 +/- 0.000)
6	Gaussian Model	2.958 +/- 0.284 (micro average: 2.972 +/- 0.000)
7	K-Nearest Neighbours	0.194 +/- 0.095 (micro average: 0.214 +/- 0.000)

Table 3 shows that the RMSE value of the Support Vector Machine Algorithm is the algorithm with the most optimal value among the other algorithms. This can be seen by the lowest RMSE value of 0.022 +/- 0.062 (micro average: 0.062 +/- 0.000).



A	B	C	D	E	F	G	H
	0.145 +/- 0.055	0.157 +/- 0.078	0.046 +/- 0.026	0.022 +/- 0.062	0.028 +/- 0.010	2.958 +/- 0.284	0.194 +/- 0.095
0.145 +/- 0.055		0.711	0.000	0.000	0.000	0.000	0.174
0.157 +/- 0.078			0.000	0.000	0.000	0.000	0.344
0.046 +/- 0.026				0.277	0.057	0.000	0.000
0.022 +/- 0.062					0.748	0.000	0.000
0.028 +/- 0.010						0.000	0.000
2.958 +/- 0.284							0.000
0.194 +/- 0.095							

Figure 7. T-Test Result

Note:
B : Deep Learning
C : Linear Regression
D : Neural Network
E : Support Vector Machine
F : Generalized Linear Model
G: Gaussian Model
H: K-Nearst Neighbours

The results of the T-Test show that the Support Vector Machine, Generalized Linear Model and Neural Network Algorithms have no significant difference because they have an alpha value > 0.050, then the same thing happens to the Deep Learning Algorithm, Linear Regression, and K-Nearest Neighbors with aplha value > 0.050. Then the ranking of the entire algorithm is as shown in Table 4.

Table 4. Algorithms Rating

No	Algorithms	RMSE	T-Test
1	Support Vector Machine	0.022 +/- 0.062 (micro average: 0.062 +/- 0.000)	No Significant Difference
1	Generalized Linear Model	0.028 +/- 0.010 (micro average: 0.030 +/- 0.000)	No Significant Difference
1	Neural Network	0.046 +/- 0.026 (micro average: 0.052 +/- 0.000)	No Significant Difference
2	Deep Learning	0.145 +/- 0.055 (micro average: 0.154 +/- 0.000)	No Significant Difference
2	Linear Regression	0.157 +/- 0.078 (micro average: 0.173 +/- 0.000)	No Significant Difference
2	K-Nearest Neighbours	0.194 +/- 0.095 (micro average: 0.214 +/- 0.000)	No Significant Difference
3	Gaussian Model	2.958 +/- 0.284 (micro average: 2.972 +/- 0.000)	Significant Difference

4. CONCLUSION

The conclusions of this study are as follows, the results of a comparison of the 7 algorithms show that the Support Vector Machine algorithm based on the RMSE value is the best algorithm with a value of 0.022 +/- 0.062 (micro average: 0.062 +/- 0.000). Based on the T-Test it is known that there is no significant difference between the Support Vector Machine Algorithm, the Generalized Linear Model, and the Neural Network, so that if the rankings are sorted, the three are ranked 1st. So the three Algorithms with rank 1 are suitable for use in Forecasting SNDF Metaverse prices.

REFERENCES

[1] A. Ahmad and W. Gata, "Sentimen Analisis Masyarakat Indonesia di Twitter Terkait Metaverse dengan Algoritma Support Vector Machine," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 6, no. 4, pp. 548–555, 2022, doi: 10.35870/jtik.v6i4.569.

[2] S. M. Park and Y. G. Kim, "A Metaverse: Taxonomy, Components, Applications, and Open Challenges," *IEEE Access*, vol. 10, pp. 4209–4251, 2022, doi: 10.1109/ACCESS.2021.3140175.

[3] I. Akbar Endarto and Martadi, "Analisis Potensi Implementasi Metaverse Pada Media Edukasi Interaktif," *J. Barik*, vol. 4, no. 1, pp. 37–51, 2022, [Online]. Available: https://ejournal.unesa.ac.id/index.php/JDKV/.

[4] A. Solechan and T. W. A. Putra, "Literatur Review : Peluang dan Tantangan Metaverse," *J. Tek. Inform. dan Multimed.*, vol. 2, no. 1, pp. 62–70, 2022, [Online]. Available: http://journal.politeknik-pratama.ac.id/index.php/JTIM page62.

[5] Q. Yang, Y. Zhao, H. Huang, Z. Xiong, J. Kang, and Z. Zheng, "Fusing Blockchain and AI With Metaverse: A Survey," *IEEE Open J. Comput. Soc.*, vol. 3, pp. 122–136, 2022, doi: 10.1109/ojcs.2022.3188249.

[6] 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.

[7] 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.

[8] A. D. Arisandi, Ferdiansyah, L. Atika, E. S. Negara, and K. R. N. Wardani, "Prediksi Mata Uang Bitcoin Menggunakan LSTM Dan Sentiment Analisis Pada Sosial Media," *J. Ilm. Komputasi*, vol. 19, no. 4, pp. 559–566, 2020, doi: 10.32409/jikstik.19.4.370.

[9] A. Nurdiansyah, M. T. Furqon, and B. Rahayudi, "Prediksi Harga Bitcoin Menggunakan Metode Extreme Learning Machine (ELM) dengan Optimasi Artificial Bee Colony (ABC)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 6, pp. 5531–5539, 2019.

[10] N. F. B. Pradana and S. Lestanti, "Aplikasi Prediksi Jangka Pendek Harga Bitcoin Menggunakan Metode ARIMA," *J. Ilm. Inform. Komput.*, vol. 25, no. 3, pp. 160–174, 2020, doi: 10.35760/ik.2020.v25i3.3128.

[11] S. Saadah and H. Salsabila, "Prediksi Harga Bitcoin Menggunakan Metode Random Forest (Studi Kasus: Data Acak Pada Awal Masa Pandemic Covid-19)," *J. Komput. Terap.*, vol. 7, no. Vol. 7 No. 1 (2021), pp. 24–32, 2021, doi: 10.35143/jkt.v7i1.4618.

[12] K. D. Larasati and A. H. Primandari, "Forecasting Bitcoin Price Based on Blockchain Information Using Long-Short Term Method," *Param. J. Stat.*, vol. 1, no. 1, pp. 1–6, 2021, doi: 10.22487/27765660.2021.v1.i1.15389.

[13] A. Mulyani *et al.*, "The Prediction Of PPA And KIP-Kuliah Scholarship Recipients Using Naive Bayes Algorithm," *JUTIF*, vol. 3, no. 4, pp. 821–827, 2022.

[14] 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.

[15] yahoo finance, "The Sandbox USD (SAND-USD)," *finance.yahoo.com*, 2022. https://finance.yahoo.com/quote/SAND-USD/history/.

[16] A. D. Savitri, F. A. Bachtiar, and N. Y. Setiawan, "Segmentasi Pelanggan Menggunakan Metode K-Means Clustering Berdasarkan Model RFM Pada Klinik Kecantikan (Studi Kasus : Belle Crown Malang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 9, pp. 2957–2966, 2018.

[17] 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.
[18] 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.