

The Implementation of Data Mining for Predicting XAU/USD Price Trends in the Forex Market on Meta Trader 5 using Naïve Bayes Method

Sendi Novianto^{1*} and Habib Akbar Wibowo²

Informatic Engineering Program^{1,2}, Faculty of Computer Science, Universitas Dian Nuswantoro

Abstract— The one of the trading instruments is forex or foreign exchange. The forex market provides various commodities, one of which is XAU/USD. XAU/USD dataset with daily, h4, h1 time frames obtained from the MetaTrader 5 application via the FBS broker. Predicting forex prices is difficult because there are various factors that influence it, so data mining methods are needed to predict increases or decreases. Naïve Bayes is a method and logic that can be applied in making predictions. So the research objective of this final project is to apply naïve Bayes methods and logic in predicting the price of XAU/USD on the daily, h4, h1 time frames. The application of the Naïve Bayes method uses several libraries to support research, namely pandas, jcopml, and sklearn. In naïve Bayesian logic research, this is called from the sklearn library using gaussianNB. In the previous study was only done on one time frame, while in this study it was tried on three types of time frames, namely h1, h4 and daily. In this study, the performance reference uses an f1 score matrix because the number of false positives and false negatives is not tight (symmetrical). This study produces values for each time frame obtained from the confusion matrix formula with f-scores of 49.99% (daily), 53.52% (h4), 55.44% (h1).

Keyword—Prediction, XAU/USD, Naïve Bayes

I. INTRODUCTION

The development of technology and information systems is currently growing rapidly, making all activities carried out by humans easier. Within this context, the economy is part of a rapidly expanding sector characterized by the convenience of conducting online business, including trading activities. Forex or foreign exchange market is one of the trading markets available, offering investment products that involve trading in foreign currencies. In Indonesia, the term forex is referred to as forex or foreign exchange. Foreign exchange trading, also known as forex trading, is one of the largest financial markets worldwide, with a turnover of up to 6 trillion US dollars. Retail traders account for 45% of the volume transactions [1].

Forex is currently in great demand because of the huge profits it makes from the forex market [2]. The high profits in

the forex market are by taking advantage of the high level of liquidity and speed of price trends. The profit earned will exceed that of trading in general. As a result of the speed of this movement, trading in forex or foreign exchange is not without high risk [3]. Therefore understanding is very important for a trader or investor, both beginners and those who are already involved in forex trading. Understanding trends in the forex market is very important, because understanding trends will help traders determine buy or sell positions in a currency in order to get big profits or reduce trading losses in the forex market.

Additionally, analytical skills are a must-have for a trader. This ability is an important point for traders to get big profits or avoid losses. In practice, there are two types of analysis used in the forex market, namely technical analysis and fundamental analysis. Technical analysis looks at market changes technically by analyzing market price charts and indicators, then predicting them to make buy or sell decisions. Meanwhile, fundamental analysis is the ability to predict by studying the basic elements that influence the economy, such as stocks or currencies or analyzing economic indicators, government policies, society through news reports, economic data and political events from a country [4].

By definition, technical analysis is more suitable for short-term trading purposes because not much time is needed in the observation process. It is different from the fundamental analysis used in trading which has the aim of being a long-term investment because it takes a relatively long time to make field observations regarding various social, economic and political news information in a country. So that this study has a focus on technical prediction analysis by means of data processing, which is widely referred to as data mining.

The market for foreign currency values or forex is a time series that is volatile and very complex. Gold is one of the commodities found in the forex market. As a commodity, gold is considered the most important purchasing instrument. Countries and multinational corporations use the exchange rate as one of the most important economic variables and gold reserves are the variables used to ensure their relationship with the outside world. This is what makes the gold market one of the largest and most important financial markets in the world.

*Corresponding author

Email address: sendi.novianto@dsn.dinus.ac.id

One that pays in the forex market is XAU/USD or which means how much money in US dollars is needed to buy one ounce of gold [5]. Predicting gold price trends is a difficult and challenging job because it is full of tension factors that affect the rate of forex price trends. So that data mining methods are needed in the process of processing data.

A process in finding patterns and relationships in a data with a large size is the meaning of data mining. Data mining has basic techniques derived from statistical and mathematical techniques [6]. The process of automatic analysis in order to find a pattern or an important trend whose existence is not visible or realized in the data complex and on a large scale is the short definition of data mining. Data mining has several uses, namely classification, association, regression, deviation analysis, sequence analysis, forecasting and grouping. Classification is a technique that aims at finding patterns and depictions in data or classes. This technique is the basis of the research. The process of data mining broadly has 3 important stages, namely data preparation, which includes the preprocessing process. Then the implementation of the algorithm or logic used. The last is to analyze the results of the implementation of the algorithm [7].

One of the methods that can be used in data processing or data mining is to use Bayes' theory for its classification by using the Naive Bayes method. The classification process of the Naive Bayes method is widely used and preferred because of the simplicity and speed of the method [8]. The naïve Bayes method comes from a theory put forward by Thomas Bayes in the 18th century and is called Bayes' theorem. Bayes fundamental theory comes from a statistical approach based on classification decisions using probability.

In particular, the application of the Naive Bayes method can be used in predictions based on classification and class probabilities. Prediction using the Naive Bayes method is an application that can be used as an analysis to predict forex prices to rise or fall. The data taken in this study is data taken from the MetaTrader 5 application which is XAU/USD data for 2022. In this study the proposed method uses the Naive Bayes method which will be used to estimate the rising or falling trend of XAU/USD gold prices. By considering these problems. So that this research can help in predicting practice using the Naïve Bayes method on forex price trend data.

II. LITERATURE REVIEW

A. Related Work

In research conducted by Guntur [8]. This research discusses the use of the Naïve Bayes method to predict gold prices in investments. This study used gold data from December 1 2017 to January 1 2018. The research was processed using the Rapidminer software. Based on this research, the Naïve Bayes method is able to predict gold prices in the next 14 days. The study used 16 test data and obtained an accuracy rate of 75%.

Then research with the title "Implementation of the Naive Bayes Method for Gold Price Prediction" by Ristianto [9]. This study discusses the prediction of gold prices through a data

mining process using the Naïve Bayes method and using the Rapidminder software to perform calculations. The dataset used in this study is data sourced from historical data on the bps.go.id website. The dataset used is from January 2016 - December 2019 and predicts one year ahead with a record that prices remain stable. In this study, the naïve Bayes algorithm was able to obtain an accuracy rate of 95.92% in predicting investment profit.

Next by Sitorus & Tarihoran,[7]. Research conducted by Sitorus and Tarihoran discusses the analysis of stock prices from PT Astra International Tbk taken from the Indonesia Stock Exchange using the Naïve Bayes and Decision Tree-J48 methods. This study compared the naïve Bayes method and the decision tree-J48. The data used in this study amounted to 1,195 and the testing data was 20% of the total data used. Obtaining the level of accuracy from this study resulted in an accuracy of 92.0502% in data testing using the Naïve Bayes method and 98.7448% accuracy in data testing using the J-48 decision tree method.

Research conducted by Sri Indriyani [10] discusses the prediction of precious metal prices using naïve Bayes. Tests carried out in research by Sri Indriyani used 120 sample data. The test results are a comparison of the naïve Bayes method and the naïve Bayes method using the basis of particle swarm optimization features. This research resulted in naïve Bayes accuracy of 84.17% and naïve Bayes pso accuracy of 88.33%.

Lastly is research conducted by Pande [11]. This study has a discussion about comparative analysis from Naïve Bayes and KNN on predictions of forex movements in the value of the GBP / USD currency in the daily time frame. The amount of data in this study was 2145. From this data a column class target was created with the name 'result'. The data in the study were divided into 2 parts, namely training by 80% and testing by 20%. The accuracy results obtained in this study were divided into 2, namely naïve Bayes and knn. The research obtained an accuracy of 50% on the Naïve Bayes method and 53% accuracy on the KNN method.

B. XAU/USD

There are three types of gold investment, namely gold investment in real or physical form, gold in jewelery and gold investment in the form of trading. One of the stable trading instruments is gold besides that gold is also an effective instrument. Gold in XAU/USD is simply the price of gold in dollars (USD). As a trading instrument, gold is referred to as a safe haven, namely a stable investment asset because the value of gold can be stable or even increase even when the market is unstable, therefore gold investment is called safe [12].

C. Naïve Bayes

The Naïve Bayes method comes from a theory discovered by Thomas Bayes in the 18th century, namely Bayes' theorem. In data mining, Bayes' theory is fundamental to the statistical approach, this approach is based on qualifying trade offs between many classification decisions using probabilities. Bayes is a prediction technique with a simple probability basis based on the application of the Bayes theorem with strong independence rules [13]. The general Bayes formula and explanation of the Bayes formula are as follows.

$$\frac{P(H|E) = P(E|H)x P(H)}{P(E)}$$

1. $P(H|E)$ is the conditional probability that a hypothesis occurs if given evidence that E occurs.
2. $P(E|H)$ is the probability that a proof E will occur will affect hypothesis H.
3. $P(H)$ is the initial probability (priori) of hypothesis H occurring regardless of any evidence.
4. $P(E)$ Probability of the initial (priori) evidence E occurs regardless of the hypothesis/other evidence.

D. Classification Performance

The evaluation matrix formulation or formulation for the classification method is based on true positives, true negatives, false positives, and false negatives. Formulation can be obtained from the confusion matrix. The simple confusion matrix is a tabular representation of the true and predicted class of each case in the test set. As shown in the table below [11].

Tabel 1. Representation True and Predicted Class

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

There are four factors used in the evaluation, namely recall, precision, accuracy and F-Score. The following is an explanation of these factors.

1. Recall is the ratio of true positives to true positives plus false negatives. Recall value is obtained by formula.

$$\frac{TP}{TP + FN} \times 100$$

2. Precision represents the fraction that was correctly predicted among all the predicted outliers. The precision value is obtained by formula.

$$\frac{TP}{TP + FP} \times 100$$

3. Accuracy is the percentage of observations that are classified correctly. The accuracy value is obtained by the formula

$$\frac{TP + TN}{TP + FP + FN + TN} \times 100$$

4. The F-score measures the percentage of balance between precision and recall. The f-score value is obtained by the formula $(2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \times 100\%$.

$$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100$$

III. METHODOLOGY

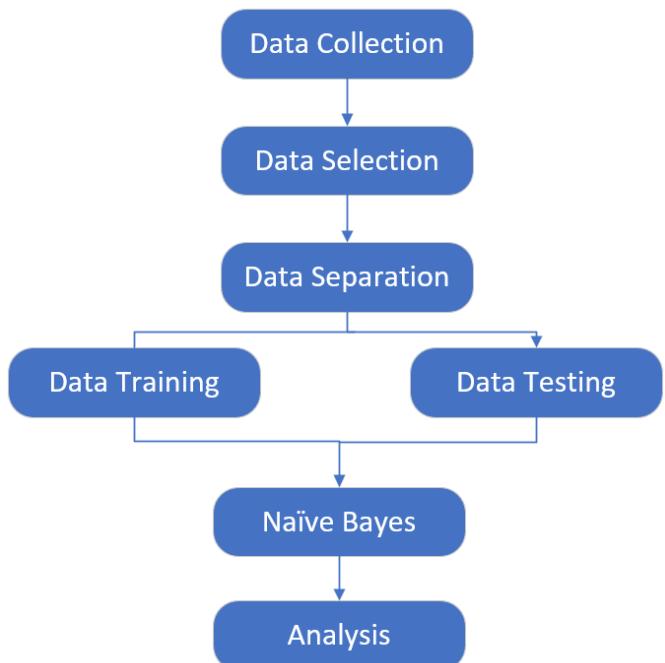


Fig 1. Research Stages

A. Data Collection

In this research begins with data collection. In the previous study was only done on one time frame, while in this study it was tried on three types of time frames, namely h1, h4 and daily. The data used is gold data against the dollar or XAU/USD on the respective time frames, namely H1 (1 hour), H4 (4 hours), and Daily (Daily) in the period from January to December 2022. The data is public data which is download from MetaTrader 5 application through FBS broker on FBS-Demo server with csv format. The data we use is real data in the past. we do not use realtime data because it will take a long time in this study.

B. Data Selection

The data from FBS broker contain date, open, high, low, close, tick volume, volume, spread field. From this field, we just use field date, open and close.

C. Data Separation

Separation of data or data separation is the stage of splitting the data. The data is broken down for different purposes,

Tabel 2. Data Separation

*Corresponding author
Email address: sendi.novianto@dsn.dinus.ac.id

OPEN	HIGH	LOW	CLOSE	VOL	RESULT
182838	183163	179844	180128	58726	0
180114	181682	179854	181439	65323	1
181455	182972	180832	181065	60714	0
181017	181154	178635	178966	78961	0
179079	179855	178254	179627	64993	1
179432	180233	179027	180166	61132	1
180119	182337	180003	182150	62864	1
182108	182818	181477	182621	60384	1
182569	182827	181241	182272	62948	0
182235	182927	181476	181763	68366	0
181917	182330	181318	181878	36387	0
181806	182288	180562	181368	79184	0
181159	184392	181021	184125	72301	1
184034	184837	183641	184004	78307	0

namely training and testing with the respective percentages being 80% training data and 20% testing data.

1. Data Training

Data Training is data used to train algorithms in making predictions and running algorithm functions.

2. Data Testing

Data Testing is data used to see performance and also scores.

D. Naïve Bayes

At this stage it is an implementation of the Naive Bayes algorithm using the Pandas and Sklearn libraries with import GaussianNB. In this study the implementation of Naive Bayes uses the Python programming language. Implementation of Naive Bayes is carried out during training and testing. At this stage it will display the performance classification by displaying the score results which are formulated in the confusional matrix formula as follows.

$$Recall = \frac{TP}{TP+FN} \times 100\%$$

$$Precision = \frac{TP}{TP+FP} \times 100\%$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100\%$$

$$F1-score = \frac{2(Recall \times precision)}{recall+Precision}$$

E. Analysis

This stage is the result of a comparative analysis or comparison of each data on the time frame H1 (1 hour), H4 (4 hours), and Daily (daily). The results of this analysis produce the effectiveness of Naïve Bayes for each time frame.

IV. RESULT AND DISCUSSION

A. Collection

The first stage in this research is data collection. This research data is downloaded via the MetaTrader application with the FBS broker using the FBS-Demo server. The data is data on the daily time frame, h4, h. The daily dataset is 259 data, the h4 dataset is 1547 data, and the h1 dataset is 5911 data.

B. Selection

Based on previous research by Pande [11] in this study the data attributes involved were Open, High, Low, Close, and TickVol. Through this research, in this study the data was selected and left Open, High, Low, Close and Tick Vol and added the Result attribute.

The result attribute is a condition that is obtained from the conditional column process in Microsoft Excel. The numbers 0 and 1 are the result of conditioning between Open and Close. Number 0 if Open is greater than Close and number 1 if Open is less than Close.

C. Data Separation

Before entering the data separation stage, the first is calling several libraries that will be needed in the data mining process. Then upload the dataset into the research project using the pandas library and display the research dataset. After the data is uploaded, the next process is to divide the data into two variables X and y. The X variable displays the open, high, low, and close attributes. While y displays the result attribute as the variable to be predicted. The next process is the data separation stage which is separated into training data and testing data with a percentage of 8:2. The data separation process is modeled using the sklearn library.

D. The implementation of Naïve Bayes

This stage is the implementation of Naïve Bayes on separated data. The use of the Naïve Bayes method uses the sklearn library with import GaussianNB and uses a pipeline to automatically accommodate processes from Naïve Bayes logic. Then after GaussianNB is already in the pipeline adjusted to the X_train and y_train variables, the process continues by scoring scores for X_train, y_train and X_test, y_test. In the daily dataset (daily) obtained a train score of 54.85 and 53.84 for the test score. In the h4 dataset (4 hours) obtained a train score of 53.59 and 49.03 for the test score. In the h1 dataset (1 hour) obtained a train score of 51.18 and 49.87 for the test score.

E. Confusional Matrix Plots

The last process is to make a confusional matrix plot using the Jcopml.plot library by importing plot_confusional_matrix for X_train, y_train, X_test, y_test in the pipeline. The conventional matrix plot process produces a table that contains predicted data and displays train scores and test scores. By default, the confusional metrix plot is arranged in ascending order. Therefore, the position settings are adjusted so that classification performance calculations can be carried out by using the sklearn import confusion matrix library and setting the labels at position [1,0]. Then displays the classification

report. After the data is plotted on the confusional matrix, the next step is to calculate recall, accuracy, precision, and F1-score from the Naïve Bayes method on testing scores. Through calculations based on the confusional matrix formula.

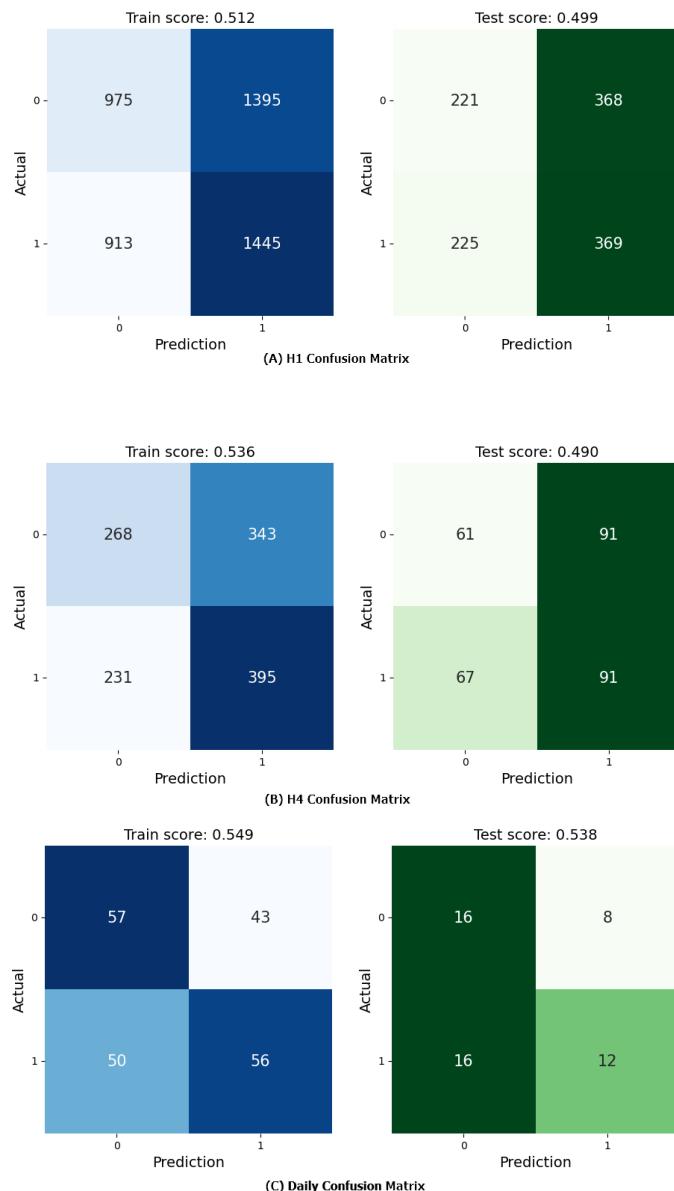


Fig 2. Confusion Matriks Results Stages

F. Analysis

The last stage in the research is to analyze how effective the Naïve Bayes method is on the dataset used in making predictions. In this study the matrix used as a reference for the performance of Naïve Bayes against the dataset used is the f1-score. Because based on a literature study which explains that the accuracy matrix is used if the number of false negatives and false positives is close to symmetrical, but if it is not close to symmetrical then use the f1-score as a performance reference. In this study the number of false negatives and false positives is not close to symmetrical so that the f1-score becomes a reference for the performance of the Naïve Bayes method for

the dataset being examined. The following is a comparison of f1-scores for each time frame.

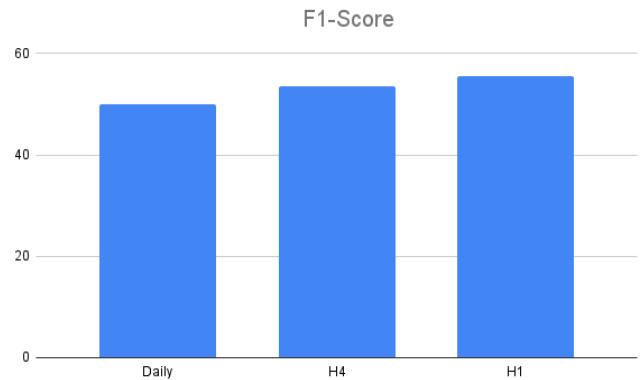


Fig 3. Diagram Result F1-Score

Through the calculation process with the classification report, it produces recall and precision values. This value is used to find the f1-score value. In this study the f1-score values of the daily, h4, h1 time frames were 49.99%, 53.52%, 55.44%, respectively. So that in reference to performance, the performance of Naïve Bayes on the H1 dataset has the greatest value. The factor that differentiates the percentage is the amount of data that differs from each time frame. The amount of data trained respectively is 80% of the total data. Then the amount of data from each time frame is 258 data in the daily dataset, 1547 data in the h4 dataset, and 5911 data in the h1 dataset. The difference in the amount of data is enough to affect the performance of the Naïve Bayes method.

V. CONCLUSION

A. Conclusion

The conclusions obtained from research related to the Naïve Bayes method on the XAU/USD dataset on the time frame, daily, h4, h1 are as follows:

1. The implementation of data mining using the naïve Bayes method in predicting forex datasets which are XAU/USD data can be applied through Jupyter's WIDE (Web Integrated Development Environment). In this study several libraries were involved, namely pandas, sklearn, jcopml. The application of the Naïve Bayes method uses GaussianNB which is imported from the sklearn library. Numbers 0 and 1 are predictions resulting from research.
2. Naïve Bayes logic can be applied and results in a score obtained with the confusional matrix formula. The score of the naïve Bayes logic is displayed using a classification report. The results of the classification report are recall, precision, and accuracy. The daily, h4, and h1 datasets are positive, obtaining recall and precision scores of 42.85%, 57.59%, 62.12% and 60%, 50%, 50.06%. Also, negative values obtain scores and precisions of 66.66%, 40.13%, 37.52% and 50%, 47.65%, 49.55%. Then the results of the accuracy score for daily data, h4, h1 are 53.84%, 49.03%, 49.87% respectively. Overall the use of the Naïve Bayes method for the daily dataset has a higher accuracy value compared to the h4 and h1 datasets.

*Corresponding author

Email address: sendi.novianto@dsn.dinus.ac.id

3. By comparing the data with three different time frames, we can be sure that the daily data plays an important role in making decisions to buy or sell.

B. Future work

In the XAU/USD price prediction research using the Naïve Bayes method, there are still some shortcomings and weaknesses. So this research needs to be developed in further research. The following are suggestions that can be used for further research, including:

1. Predictions using the Naïve Bayes method have a low classification report value. This value can be increased by adding several attributes that can be used as prediction value requirements. These attributes can be in the form of market trend analysis.
2. This research is only in the form of numerical data attributes that are processed using naïve Bayes. In order to increase the accuracy value, you can use attribute categorical data obtained from fundamental analysis by analyzing news about economic indicators, government policies, etc.
3. In this research, the naïve Bayes method can be combined with the time series method or it can be developed using a forecasting method such as the fb-prophet model. So that future research not only predicts but can also forecast data outside of historical data.

REFERENCES

- [1] M. M. Huda and R. D. R. Yusron, "Kombinasi Naive Bayes dan Metode Time Series Sebagai Peramalan Pergerakan Harga pada Perdagangan Valuta Asing," *ilkomnika*, vol. 2, no. 2, pp. 151–155, Aug. 2020, doi: 10.28926/ilkomnika.v2i2.186.
- [2] N. Riha, T. Matulatan, and R. Hidayat, "Penerapan Fuzzy Time Series Stevenson Porter pada Peramalan Pergerakan Nilai Forex," 2020, doi: <https://doi.org/10.29407/inotek.v4i3.83>.
- [3] M. Lutfi, "Prediksi Harga Terendah dan Harga Tertinggi dengan Menggunakan Metode Anfis untuk Analisa Teknikal pada FBS Forex Market," 2019, doi: <https://doi.org/10.35746/jtim.v1i3.40>.
- [4] L. Abednego, "Forex Trading Robot with Technical and Fundamental Analysis," *JCP*, pp. 1089–1097, 2018, doi: 10.17706/jcp.13.9.1089-1097.
- [5] Z. H. Kilimci, "Ensemble Regression-Based Gold Price (XAU/USD) Prediction," 2022, [Online]. Available: https://www.researchgate.net/publication/361109538_Ensemble_Regres_sion-Based_Gold_Price_XAUUSD_Prediction
- [6] A. Nastuti and S. Z. Harahap, "TEKNIK DATA MINING UNTUK PENENTUAN PAKET HEMAT SEMBAKO DAN KEBUTUHAN HARIAN DENGAN MENGGUNAKAN ALGORITMA FP-GROWTH (STUDI KASUS DI ULFAMART LUBUK ALUNG)," *INFORMATIKA*, vol. 7, no. 3, pp. 111–119, Sep. 2019, doi: 10.36987/informatika.v7i3.1381.
- [7] H. Sitorus and Y. Tarihoran, "Analisis Harga Saham PT Astra Internasional Tbk Menggunakan Data Dari Bursa Efek Indonesia dalam Jangka Waktu Pendek Menggunakan Metode Naïve Bayes dan Decision Tree-J48," *TelKa*, vol. 8, no. 1, pp. 21–33, Apr. 2018, doi: 10.36342/teika.v8i1.2239.
- [8] M. Guntur, J. Santony, and Y. Yuhandri, "Prediksi Harga Emas dengan Menggunakan Metode Naïve Bayes dalam Investasi untuk Meminimalisasi Resiko," *RESTI*, vol. 2, no. 1, pp. 354–360, Apr. 2018, doi: 10.29207/resti.v2i1.276.
- [9] F. Ristianto, N. Nurmalasari, and A. Yoraeni, "Implementasi Metode Naïve Bayes Untuk Prediksi Harga Emas," *co-science*, vol. 1, no. 1, pp. 62–71, Jan. 2021, doi: 10.31294/coscience.v1i1.201.
- [10] Sri Indriyani, Muhamad Fatchan, and Andri Firmansyah, "PREDIKSI HARGA LOGAM MULIA DENGAN PENDEKATAN ALGORITHMA NAÏVE BAYES DAN PSO," *JINTEKS*, vol. 5, no. 1, pp. 179–182, Feb. 2023, doi: 10.51401/jinteks.v5i1.2230.
- [11] K. S. Y. Pande, D. G. H. Divayana, and G. Indrawan, "Comparative analysis of naïve bayes and knn on prediction of forex price movements for gbp/usd currency at time frame daily," *J. Phys.: Conf. Ser.*, vol. 1810, no. 1, p. 012012, Mar. 2021, doi: 10.1088/1742-6596/1810/1/012012.
- [12] E. Emilda, "Adakah Pengaruh Event dalam Economic Calendar terhadap Gold Price (XAU/USD)?" *JIEGMK*, vol. 11, no. 1, p. 29, Jul. 2020, doi: 10.36982/jiegmk.v11i1.1058.
- [13] D. Haryadi and R. Mandala, "Prediksi Harga Minyak Kelapa Sawit Dalam Investasi Dengan Membandingkan Algoritma Naïve Bayes, Support Vector Machine dan K-Nearest Neighbor," *itfs*, vol. 4, no. 1, Mar. 2019, doi: 10.33021/itfs.v4i1.1181.