Comparative analysis of naïve bayes and knn on prediction of forex price movements for gbp/usd currency at time frame daily

K S Y Pande$^{1,a}$, D G H Divayana$^{1,b}$, G Indrawan$^{1,c}$

$^1$Computer Science, Magister Program, Ganesha University of Education, Bali, Indonesia – 81116

E-mail: $^a$sudana.yasa.pande@gmail.com, $^b$hendra.divayana@undiksha.ac.id, $^c$gindrawan@undiksha.ac.id

Abstract. This study aims to analyze the comparison of the Naïve Bayes and kNN on the Prediction of Forex Price Movements for GBP / USD on Time Frame Daily. The data used is taken from the metatrader-4 application which is often used by forex traders when making transactions. There are 2,145 data rows consisting of the date, hour, open price, high, low, close, and transaction volume columns. From this data, a column for the target class is created with the name 'result'. The result column is filled with increasing or decreasing values. The value of increase or decrease is obtained from the comparison of the previous closing price with the closing price of the next day. This study analyzes the results of the comparison of the data mining classification of algorithm between the Naïve Bayes algorithm and kNN. The 2,145 data were divided into 2 parts, namely 80% for training data and 20% for testing data. The analysis is done by comparing the precision, recall, and accuracy test results for each algorithm. The conclusion of this study is that the kNN algorithm is better than the Naïve Bayes algorithm in case of predicting forex price movements for GBP/USD currency at time frame daily.

1. Introduction

Forex trading is a type of transaction that trades the currency of a country against the currencies of other countries with the aim of making profit from differences in currency values [1]. A trader usually performs analytical techniques to be able to generate profitable returns, this can determine the right momentum to enter the market. The analysis technique used is fundamental analysis and technical analysis. Fundamental Analysis is the method to understand market trends [2]. Fundamental analysis is an analysis that relies on news that is happening on world markets or is currently circulating in the market. This news content drives market traders’ emotions to determine the value of a currency, stock, or other instrument [3]. Technical analysis is an analysis approach to past price movements to predict future price movements [4][5].

Traders who use technical analysis will process historical price movements through data visualization in the form of charts or graphs, then predict price movements on certain time frames. There are 5 main components of the charts, namely: the formation price (open), the highest price (high), the lowest price (low), the closing price (close), and the transaction volume (V) or abbreviated as OHLCV. The OHLCV is generated every second and produce a very abundant data line. This enormous data is
very capable of being processed using data mining techniques so that new knowledge can be extracted [6] to predict future price movements.

Forex traders often use technical analysis to predict price movements of currencies being traded. The technical analysis used usually applies analytical indicators such as Moving Average, Stochastic, Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI) and others. The number of indicators in modern technical analysis is in the hundreds. Each indicator has different characteristics, but only one characteristic is certain and will not change is that price movements are always faster than any other indicator [4]. The technical analysis applied is still limited to the use of indicators that cannot predict future price movements. The pattern of price formation in forex price movements for certain currencies can be explored by utilizing data mining. The process deals with records such that historical records are used to make prediction about an uncertain future [7].

The classification method is a process of training (learning) an objective function (target) which is used to map each set of attributes of an object to one from a certain class label defined previously [8]. Classification techniques are very appropriate for processing data sets in the form of binary data or nominal data. Several algorithms are including C4.5 algorithm classification method, the K-Nearest Neighbor (kNN) algorithm, ID3, Naïve Bayes classification, and Classification and Regression Tree (CART). The classification method can be used to look for patterns [9] in predicting an increase or decrease in the price of a country's currency in forex trading. Some of the available classification algorithms certainly have different effectiveness in the case of the GBP / USD forex currency. Therefore, a study was conducted to analyze the comparison of the Naïve Bayes algorithm and kNN on the prediction of forex price movements.

2. Literature Study
A research was conducted by Shahbazi, Memarzadeh and Gryz in 2016 with the title "Forex Market Prediction Using NARX Neural Network with Bagging". In this study, the researchers proposed a new method for predicting price movements in the forex market based on the artificial neural network method using realtive strength index and stochastic indicators. When compared with the static neural networks, the method significantly reduce the error rate of the response and improves the performance of the prediction [10]. Another research was conducted by Findawati et al. in 2019 entitled "Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast". The researchers tried to compare data mining algorithms, especially the classification method to predict weather forecasts based on previous data. The algorithms being compared are the Naïve Bayes algorithm, K Nearest Neighbor and C4.5. The researcher compared the three algorithms by analyzing the levels considered accurate, precision, recall and f-measure. The researchers succeeded in drawing the conclusion that the KNN algorithm has the highest level of accuracy among others in making weather forecast predictions [11].

3. Classification Algorithm
This section describes the classification of algorithm used in the current study.

3.1 Naïve Bayes
The Naïve Bayes Classifier is an algorithm that can be used to solve problems in data mining, especially for the classification method. In particular, the construction of Naïve Bayes is very simple [12]. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class [13]. Naïve Bayes uses probability theory in solving a case Supervised Learning or self-learning by algorithms based on previous data. The probability value can be obtained by Formula (1).

\[
P(H|X) = \frac{P(X|H)P(H)}{P(X)}
\]  

(1)
Description for the formula:
X: Label or class to search
H: The hypothesis that x is a class or label data
P (H): Probability hypothesis H
P (X): Opportunity sample data
P (X | H): Chance of sample data X if the hypothesis assumption is true

3.2 K-Nearest Neighbor Algorithm (kNN)
K-Nearest Neighbor Algorithm (kNN) is one of the best and most widely used classification algorithms with a variety of applications [14][15]. kNN algorithm has the feature of using an approach to find values that are close to the results by calculating the value of the proximity of new cases to old cases. If the classification of a sample is unknown, then it could be predicted by considering the classification of its nearest neighbour [16]. kNN uses a technique that is based on the weight of the existing case object. The similarity value can be obtained by Formula (2).

\[
similarity(T,S) = \frac{\left(\sum_{i=1}^{n} f(T_i,S_i) \ast W_i \right)}{W_i}
\]

Description for the formula:
T: Number of new cases
S: Case in storage
n: The number of attributes for each case
i: Individual attribute between 1 and n
f: The similarity attribute function between case T and case S
W: The weight assigned to the i-th attribute

4. Classification Performance
Formulations of evaluation metrics for classification methods are based on true positive, true negative, false positive, and false negative [17]. The formulations can be obtained from confusion matrix. Confusion matrix is simple a tabular representation of true and predicted class of each case in the test set [18] as shown in Table 1.

|            | Positive | Negative | Sum      |
|------------|----------|----------|----------|
| Positive   | TP       | FP       | TP+FP    |
| Negative   | FN       | TN       | FN+TN    |
| Sum        | TP+FN    | FP+TN    | N        |

These four factors are used for evaluating recall, precision, accuracy, and F-Score, as follows [19][20][21]:
- Recall is the ratio of true positives to the sum of true positives and false negative. The recall value is obtained by formula (3).
- Precision represents the fraction of correctly predicted outlier points among all the predicted outliers. The precision value is obtained by formula (4).
- Accuracy is the percentage of observations correctly classified. The accuracy value is obtained by formula (5).
- F-score is measure strikes a balance between precision and recall. The f-score value is obtained by formula (6).

\[\text{Recall} = \frac{TP}{TP+FN} \ast 100\% \quad (3)\]
5. Research Methodology

This section describes the steps of the research.

5.1 Steps of research

The steps of the research carried out are shown in Figure 1. The research began with data collection. The data used was the GBP/USD currency price movement in daily time frame. GBP/USD was chosen because GBP and USD are a major currency [22], they have fluctuating movements, there was no big news impact that occurs on the GBP/USD currency, and in terms of the GBP / USD currency pair, there were times when both currency market positions are in a position open. Daily Time frame was chosen because trading with shorter time frames are more risky and provide low liquidity to the traders, as compared to longer Time Frames [23]. The data were obtained from the metatrader4 application, which is often used by traders. This data consists of OHLCV (open, high, low, close, and volume) and result column as class attributes. Data were collected every trading day from 4 October 2012 to 14 July 2020 totalling 2,145 rows of data. Example dataset can be seen in Table 2.

![Figure 1. Steps of research.](image)

After the data were collected, data pre-processing was carried out [24][25]. The data obtained must be checked by way of selecting data types, checking missing values, removing unnecessary columns such as date columns and checking for duplicate data [26][27]. The pre-processing data result were divided into 2 parts. The first part was used as training data, and the other part was used as testing data. As much as 80% of the data were used for training data, while the rest were used for testing data. Training data were processed with the Naïve Bayes algorithm and the kNN algorithm. The implementation of the algorithm application uses the python programming language. Python programming was chosen because The Python programming language is establishing itself as one of the most popular languages for scientific computing [28].

After the data had been successfully trained, the next step is to test the training data. The test is carried out by using the test indicators for accuracy, recall, precision, and f-score. These indicators can
be calculated after the confusion matrix has been successfully created. The results of data testing in the form of a percentage of each indicator include the percentage of accuracy, percentage of recall, percentage of precision, and percentage of f-score. The percentage results are then analyzing by comparing the indicator results for both Naïve Bayes algorithm and the kNN algorithm.

6. Result and Discussion

The implementation of the Naïve Bayes algorithm and the kNN algorithm was carried out using the Python programming language. The results of the implementation produce confusion matrix which can be seen in Table 3 and Table 4.

6.1 Naïve bayes testing result

Based on the naïve bayes confusion matrix data, the value of accuracy, recall, precision and f-score can be calculated as follows:

\[
\text{Recall} = \frac{TP}{TP + FN} \times 100%
\]
\[
= \frac{99}{99 + 81} \times 100%
\]
\[
= 55\%
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \times 100%
\]
\[
= \frac{99}{99 + 132} \times 100%
\]
\[
= 43\%
\]

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \times 100%
\]
\[
= \frac{99 + 117}{99 + 81 + 132 + 117} \times 100%
\]
\[
= 50\%
\]

\[
\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100%
\]
\[
= \frac{2 \times 0.43 \times 0.55}{0.43 + 0.55} \times 100%
\]
\[
= 49\%
\]

The value 1 in the result column represents that the price has increased as seen from the comparison of today’s closing price with the next day's closing price. Meanwhile, the value of 0 represents the price has decreased as seen from the comparison of today’s closing price with the next day's closing price. For example, in the 3rd row of the result column is 0 because today’s closing price (30387) had decreased price on the next day (29476).
6.2 kNN testing result

Based on the kNN confusion matrix data, the value of accuracy, recall, precision and f-score can be calculated as follows:

Recall = \frac{TP}{TP + FN} \times 100% \\
= \frac{124}{124 + 97} \times 100% \\
= 56%

Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \times 100% \\
= \frac{124 + 101}{124 + 97 + 107 + 101} \times 100% \\
= 53%

Precision = \frac{TP}{TP + FP} \times 100% \\
= \frac{124}{124 + 107} \times 100% \\
= 54%

F-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100% \\
= \frac{2 \times 0.54 \times 0.56}{0.54 + 0.56} \times 100% \\
= 55%

6.3 Discussion

The data from the calculation of precision, accuracy, recall, and f-score from the Naïve Bayes algorithm and the kNN were then compared with each other which can be seen in Figure 4. The graph shows that the kNN algorithm always produces a higher value than the Naïve Bayes algorithm both from recall, precision, accuracy, and f-score testing.

Naïve bayes and kNN were chosen because the dataset was modified by adding a result column which became a class. Researchers wanted to find out whether the classification method in datamining was able to predict forex price movements. From the results of this study, the kNN algorithm was better than the Naïve Bayes algorithm. However, the classification performance test still shows results below 60 percent, so further research needs to be done such as adding other types of algorithms in datamining or implementing a better pre-processing algorithm so that the test results are close to perfect.

| Positive | Negative | Sum |
|----------|----------|-----|
| Positive | 99       | 132 | 231 |
| Negative | 81       | 117 | 198 |
| Sum      | 180      | 249 | 429 |

| Positive | Negative | Sum |
|----------|----------|-----|
| Positive | 124      | 107 | 231 |
| Negative | 97       | 101 | 198 |
| Sum      | 221      | 208 | 429 |
7. Conclusion

The conclusion obtained from this study is that the kNN algorithm is better than the Naïve Bayes algorithm in testing recall, precision, accuracy, and f-score in the case of predictions of forex price movements on GBP/USD time frame daily. The Naïve Bayes algorithm test gets an accuracy value of 50%, precision of 43%, a recall of 55% and a f-score of 49%. Meanwhile, the kNN algorithm test gets an accuracy value of 53%, precision of 54%, a recall of 56% and a f-score of 55%.

References

[1] Busman B, Nurhayati N, Amali F and Muttaqin Z 2017 Penerapan Big Data Pada Forex Trading Menggunakan Analisa Statistik Dengan Breakout Strategy Pseudocode 4 137–43
[2] Milani E 2019 Fundamental Analysis - Forex Trading Techniques (Amazon Digital Services LLC)
[3] Drakopoulou V 2016 A Review of Fundamental and Technical Stock Analysis Techniques J. Stock Forex Trading 05 1–8
[4] T Suharto F 2013 Investasi Secara Benar: Mengungkap Rahasia Forex (Jakarta: Kompas Gramedia)
[5] Schiavone D 2019 Technical Analysis - Trading Indicators and Charting (Independently Published)
[6] Hamsathvani M, Rajeswari D and Kalaiselvi R 2014 Survey on Infrequent Weighted Item set Mining Using FP Growth Int. J. Adv. Res. Electr. 3
[7] Jayne C, Yue S and Lazaros S I 2013 Engineering Applications of Neural Networks (Springer)
[8] Nofriansyah D and Nurcahyo G W 2015 Algoritma Data Mining dan Pengujian (Deepublish)
[9] Jadhav K R and Nimblekar P 2020 Survey On Mining Frequent Pattern On Big Data Using Map Reducton Technique 5 119–22
[10] Shahbazi N, Memarzadeh M and Gryz J 2016 Forex Market Prediction Using NARX Neural Network with Bagging MATEC Web Conf. 68 6–10
[11] Findawati Y, Astutik I R I, Fitroni A S, Indrawati I and Yuniasih N 2019 Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast J. Phys. Conf. Ser. 1402
[12] Amor N Ben, Benferhat S and Elouedi Z 2004 Naive Bayes vs decision trees in intrusion detection systems Proc. ACM Symp. Appl. Comput. 1 420–4
[13] Han J, Pei J and Kamber M 2011 Data Mining: Concepts and Techniques (Elsevier) p 744
[14] Kuhkan M 2016 A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm Int. J. Comput. Eng. Inf. Technol. 8 90–5
[15] Jopri M H, Ab Ghaiii M R, Abdullah A R, Manap M, Sutikno T and Too J 2020 K-nearest neighbor and naïve bayes based diagnostic analytic of harmonic source identification Bull. Electr. Eng. Informatics 9 2650–7
[16] Mucherino A, Papajorgji P and Pardalos P 2009 Data Mining in Agriculture (Springer Science & Business Media)
[17] Misra S, Li H and He J 2019 Machine Learning for Subsurface Characterization (Gulf Professional Publishing)
[18] Rhys H I 2020 Machine Learning with R, the tidyverse, and mlr (Simon and Schuster)
[19] Bugnion P, Manivannan A and Nicolas P R 2017 Scala: Guide for Data Science Professionals (Packt Publishing Ltd)
[20] Rahmad F, Suryanto Y and Ramli K 2020 Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification IOP Conf. Ser. Mater. Sci. Eng. 879
[21] Düntsch I and Gediga G 2019 Confusion Matrices and Rough Set Data Analysis J. Phys. Conf. Ser. 1229
[22] Doman M and Doman R 2020 The Dynamics of Linkages Between Major Currencies: How Does It Change Depending on the Time of Day? SSRN Electron. J. 00
[23] Mphil O D and Kirichenko L 2009 Investigation of Time Series of Original Values of Currency Rates Measured on Small Time Frames on FOREX Using Methods of Chaos Theory Радиоэлектроника И Информатика 18–24
[24] Divayana D G H, Adiarta A and Abadi I B G S 2018 Initial draft of CSE-UCLA evaluation model based on weighted product in order to optimize digital library services in computer college in Bali IOP Conf. Ser. Mater. Sci. Eng. 296
[25] Ariawan I P W, Simatupang W, Ishak A M, Agung A A G, Suratmin, Adiarta A and Divayana D G H 2018 Development of ANEKA evaluation model based on TOPSIS in searching the dominant aspects of computer learning quality determinants J. Theor. Appl. Inf. Technol. 96 6580–96
[26] Mahayukti G A, Dantes N, Candiasa I M, Marhaeni A A I N, Gita I N and Divayana D G H 2018 Computer-based portfolio assessment to enhance students’ self-regulated learning J. Theor. Appl. Inf. Technol. 96 2351–60
[27] Sugiharni G A D, Setiasih N W, Mahendra I W E, Ardana I M and Divayana D G H 2018 Development of alkin model instruments as evaluation tools of blended learning implementation in discrete mathematics course on STIKOM Bali J. Theor. Appl. Inf. Technol. 96 5803–18
[28] Varoquaux G, Buining L, Louppe G, Grisel O, Pedregosa F and Mueller A 2015 Scikit-learn: Machine Learning in Python GetMobile Mob. Comput. Commun. 19 29–33