Comparative Analysis of K-NN and Naïve Bayes Methods to Predict Stock Prices

1Budi Soepriyanto
1STIMIK Sepuluh Nopember Jayapura
1Jl. Ardipura II No 22B Polimak Jayapura Selatan, Papua
1budisoep@gmail.com

Abstract—Buying and selling shares is a transaction that is widely carried out at this time, especially buying and selling stocks online which are widely available in the market, to make buying and selling shares require ability or knowledge so that the buying and selling of shares are profitable, to be able to help economic players predict prices. Profit shares or not purchased in the future, this research will conduct stock price predictions using classification methods, namely K-Nearest Neighbor and Naïve Bayes, to predict the stock price data used for one month in minute levels totalling 39065 data, based on prediction results. The highest results obtained were using Naïve Bayes with an accuracy value of 69.38 then the K-Nearest Neighbor method with a K = 5 value of 67.25%, based on these results it can be concluded that the use of the K-Nearest Neighbor and Naïve Bayes methods for prediction share price not yet owned I high accuracy, so it can be combined with other methods or by using other variable predictors.

Keywords: K-Nearest Neighbor, Naïve Bayes, Prediction, Stock.

I. INTRODUCTION

The progress of information technology is very fast and affects all aspects of life starting from education, government, and the business world, to the development of information technology providing a concept of flexibility in obtaining information and data [1]. One of the business fields that has a positive impact on the development of information technology is investing in gold stocks, investing can certainly bring benefits to business people if done well. Gold is one part that becomes an instrument and people are interested in investing, this can happen because the gold commodity tends to have a relatively stable value, besides that gold also has a high liquidity value, other things that can affect gold are quite attractive because of less risk [2]. Currently, very many entrepreneurs move their investment portfolios into gold investment. However, the problem that occurs if someone who wants to invest in gold stocks does not know how to calculate or predict profit, so it often incurs losses. Careful calculations must be made, at least in investing someone can at least predict the impact and risks that will arise when investing. In this study, the objective of this research is to analyse gold stocks using the K-Nearest Neighbor (KNN) method and the Naïve Bayes method. It is hoped that this research can provide information and how to predict prices. The KNN and Naïve Bayes methods have been used several times in solving predictive cases but on different objects. Research conducted in 2015 used the K-Nearest Neighbor (KNN) method to predict the price of pepper. The research aims to develop a prediction model by combining an attribute selection method, especially forward selection to predict the commodity of pepper, in this study, concluded that the selected feature is a feature that is both in selecting variables, then this research suggests combining the KNN method with other methods to obtain more accurate results, the novelty value in this research is to combine the KNN method and the Naïve Bayes method to predict stock prices [1]. Another study to predict stock prices is by using the naïve Bayes method, this research aims to determine future stock price predictions, by utilizing classification techniques, in this study the prediction results use the naïve Bayes classifier algorithm in implementing the RapidMiner application that has been tested. In this study, it is concluded that the price of gold is influenced by several things such as the US dollar exchange rate against the euro, rupiah, and world crude oil. Furthermore, the prediction using the Naïve Bayes Classifier algorithm is implemented by Rapid Miner from the 16 data tested has an accuracy value of 75%, this is it proves that naïve Bayes can be used to predict stock prices well, the weakness in this study is that the number of variables used is very small, namely 3 variables, the difference in this study is the number of variables is 9 and the method used is different, namely combining the KNN and naïve Bayes methods. [3]. Another research that uses the naïve Bayes method is used to predict the smooth running of payments at banks, this study aims to optimize the naïve Bayes algorithm with the forward selection feature to be able to increase the accuracy or success rate obtained. The study concludes that the naïve Bayes algorithm can be used to predict the smooth payment of credit at a bank, this study suggests combining the naïve Bayes algorithm in predicting the smooth payment of bank credit, the difference in this study is that it combines two methods, namely KNN and Naïve Bayes, besides that the object of research is focused on the prediction of stock prices.

II. RESEARCH METHODS

2.1 Naïve Bayes Method

Naïve Bayes is a statistical classification model that can be used to predict the probability of class membership. Naïve Bayes is based on the Bayes theorem which has similar classification capabilities to decision trees and neural networks [4]. The formula for closeness to Naïve Bayes is shown in equation 1 [5]:

\[ P(H|X) = \frac{P(X|H)P(H)}{P(X)} \] (1)
Information:
X : Is data with an unknown class
H : Is the data hypothesis X is a specific class
P(H|X): Is the probability of hypothesis H based on condition X (posterior probability)
P(H): Is the probability hypothesis H (prior probability)
P(X): Is the Probability of X

2.2 K-Nearest Neighbor Method

The K-Nearest Neighbor (KNN) algorithm is a method for classifying objects based on learning data that is closest to the object.

Learning data is projected into a multi-dimensional space, where each dimension describes a data feature. This space is divided into groups based on the classification of learning data. A point in this space is marked with class c if class c is the classification that is most often found in the k nearest neighbors of that point [9].

Steps to calculate the K-Nearest Neighbor Algorithm method [9]:

a. To select the K Parameter (number of closest neighbors).

b. To calculate the squared difference or Euclidean (query instance) of each object against the given sample data.

c. Next, divide these objects into groups that have the smallest Euclidian distance.

d. Collects category Y (Nearest Neighbor classification)

e. By using the most majority Nearest Neighbor category, the calculated instance query value can be predicted.

To define the distance between two points, namely the point on the training data (x) & the point on the testing data (y), the Euclidean formula [9] is used, as shown in equation (2):

\[ D(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \]  

2.3 Research Method

The research method is carried out in 5 stages, namely: literature study to find literature and references as a reference for conducting research both from books, journals, proceedings, and others. Furthermore, data collection is done by looking for a dataset that will be used in this study, in the form of transaction data at the company, the data which will be processed in the study. Next is the analysis stage to obtain the level of accuracy of the KNN and naive Bayes methods. The fourth stage is the implementation of the method into the program code to make it easier to produce analyzes and predictions. The last step is testing that will be carried out with the existing dataset and then try it into the Weka application to ensure whether the data accuracy level The research flow image is shown in Figure 1.

III. RESULT AND ANALYSIS

3.1 Dataset

To make predictions, the dataset used is divided into 2 types, namely training data and testing data. The training data used is the minute level of stock price data in the span of one month from December 22, 2020, to January 22, 2021, the total data available is 39065. The predictor variable used can be seen in table 1.

Table 1. List of Variables Used

<table>
<thead>
<tr>
<th>No</th>
<th>Variable Name</th>
<th>Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rsi 14</td>
<td>Relative strength index</td>
</tr>
<tr>
<td>2</td>
<td>sma 9</td>
<td>Simple Moving 9 Period</td>
</tr>
<tr>
<td>3</td>
<td>sma 180</td>
<td>Simple Moving 180 Period</td>
</tr>
<tr>
<td>4</td>
<td>vwap</td>
<td>178-236</td>
</tr>
<tr>
<td>5</td>
<td>spread 14</td>
<td>Spread Indicator</td>
</tr>
<tr>
<td>6</td>
<td>volume</td>
<td>Volume Indicator</td>
</tr>
<tr>
<td>7</td>
<td>prev close</td>
<td>Previous close</td>
</tr>
<tr>
<td>8</td>
<td>prev floor</td>
<td>Previous floor</td>
</tr>
<tr>
<td>9</td>
<td>prev ceil</td>
<td>Previous ceil</td>
</tr>
</tbody>
</table>

Figure 1. Research Flow
3.2.3.2 Prediction Using K-Nearest Neighbor

The following is an example of implementing the K-Nearest Neighbor to predict the profit or failure of the stock price using several existing variables. The training data sample used consists of 4 data can be seen in table 2.

Table2. Sample K-NN Training Data

<table>
<thead>
<tr>
<th>Date Time</th>
<th>Rsi 14</th>
<th>Sma 9</th>
<th>vwap</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020-12-22 09:56:00</td>
<td>38.0441504 7683149</td>
<td>0.00368451 0519390338</td>
<td>0.00439352 4465</td>
</tr>
<tr>
<td>2020-12-22 10:05:00</td>
<td>38.8361936 9907474</td>
<td>0.00141821 155769339</td>
<td>0.0039532 757</td>
</tr>
<tr>
<td>2020-12-22 10:33:00</td>
<td>40.9072973 0489165</td>
<td>-0.00737446 7701630083</td>
<td>-0.0082152 485</td>
</tr>
<tr>
<td>2020-12-22 11:45:00</td>
<td>40.4479488 2957466</td>
<td>-0.00156709 9706425212</td>
<td>-0.0078882 594</td>
</tr>
</tbody>
</table>

By using existing sample data, predictions can be done with the following steps:

1. Determine the number of closest neighbors (K value).
   The value of K that will be used to predict the stock price above is to use the value of k = 1, k = 3, k = 5.

2. Calculating the squares of the Euclidean distance of each object against the given sample data. The formula used can be seen in equation 3:

   \[ D(x,y) = \sqrt{\sum_{k=1}^{n}(x_k-y_k)^2} \]  

As an example of the calculation, testing data is taken from the second data then the distance is calculated from the first data:

\[ D = \sqrt{(38.0441504 - 38.8361936)^2 + (7683149 - 9907474)^2 + (0.00368451 - 0.00141821)^2 + (0.00439352 - 0.0039532)^2} \]

So the Euclidean Distance obtained is 792043222243304.0000. The testing data is calculated by the Euclidean Distance throughout the existing training data, then ranked by taking the 5 closest neighbors, namely the 5 smallest calculated values.

3.3 Prediction Using Naïve Bayes

To provide an example of an application using Naïve Bayes, the data used in the K-Nearest Neighbor example in table 2 is rounded up so that it can be seen in Table 3 below.

Table3. Sample Data TrainingNaïve Bayes

<table>
<thead>
<tr>
<th>Date Time</th>
<th>Rsi 14</th>
<th>Sma 9</th>
<th>vwap</th>
<th>Is Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020-12-22 09:56:00</td>
<td>38</td>
<td>-0.004</td>
<td>-0.004</td>
<td>False</td>
</tr>
<tr>
<td>2020-12-22 10:05:00</td>
<td>39</td>
<td>-0.001</td>
<td>0.004</td>
<td>True</td>
</tr>
<tr>
<td>2020-12-22 10:33:00</td>
<td>41</td>
<td>-0.007</td>
<td>-0.008</td>
<td>False</td>
</tr>
</tbody>
</table>

The stages of the Naive Bayes process are:

1. Count the number of classes/labels.
   \[ P(Ci) = P(\text{Is Profit} = "True") = 2/4 \]
   \[ P(\text{Is Profit} = "False") = 2/4 \]

2. Calculating the Number of Cases Per Class
   \[ P(X|Ci) = P(\text{rsi 14} = "38", \text{sma 9} = "-0.004", \text{vwap} = "-0.004", | \text{Is Profit} = "True") = 0/2 * 0/2 * 0/2 = 0 \]
   \[ P(\text{rsi 14} = "38", \text{sma 9} = "-0.004", \text{vwap} = "-0.004", | \text{Is Profit} = "False") = 1/2 * 1/2 * 1/2 = 0.125 \]

3. Multiply All Class Variables
   \[ P(X|\text{Is Profit} = "True") = P(\text{rsi 14} = "38", \text{sma 9} = "-0.004", \text{vwap} = "-0.004", | \text{Is Profit} = "True") = 0 * 0/2 * 0/2 = 0 \]

4. Compare Results Per Class
   Based on this example, it can be concluded that rsi 14 = "38", high school 9 = "-0.004", vwap = "-0.004", entering the class Is Profit = "False"
3.4 Implementation

Tests are carried out using the WEKA application, testing is carried out using 4 scenarios, each test is carried out using 10 Fold Cross-Validation, the first is testing using the K-Nearest Neighbor method using the value of $K = 1$, the second using $K = 3$, the third using $K = 5$, and the fourth uses the Naïve Bayes method.

The test results using WEKA can be seen in the image below.

---

**Figure 2. The test results using the KNN value of $K = 1$**

**Figure 3. The test results using the KNN value of $K = 3$**

**Figure 4. The test results using the KNN value of $K = 5$**

**Figure 5. Test results using Naïve Bayes**

The following can be seen a diagram of the percentage of test results 4 times

---

Based on the picture above, it can be seen that the Naïve Bayes Method has better performance than the K-Nearest Neighbor method, Naïve Bayes has the highest truth
VI. CONCLUSIONS

Based on the results of testing 4 scenarios from 2 methods, namely K-Nearest Neighbor and Naïve Bayes, it is known that the highest value is using Naïve Bayes with the percentage of prediction truth of 68.38% and for the KNN method the value of K which has the highest accuracy is K = 5 with the percentage of the truth value is 67.25%. So it can be concluded that the performance of these two methods to predict stock prices is still not good, so for further research, you can use another classifier method or you can use other predictor variables and with more data.

REFERENCES


