International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol :Vol. 02, Issue 02, May 2021
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

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

1stBudi Soepriyanto
[1]Program Studi Sistem Informasi
STIMIK Sepuluh Nopember Jayapura
[1]Jl. Ardipura II No 22B Polimak Jayapura Selatan, Papua
[1]budisoep@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 KKN 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 Naive BayesMethod

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]:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (1)$$

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 Euclidian (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):

$$D\,(x,y) = \sqrt{\Sigma_{k-1}^{n}(x_{k-}y_k)^2} \quad (2)$$

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 naïve 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.

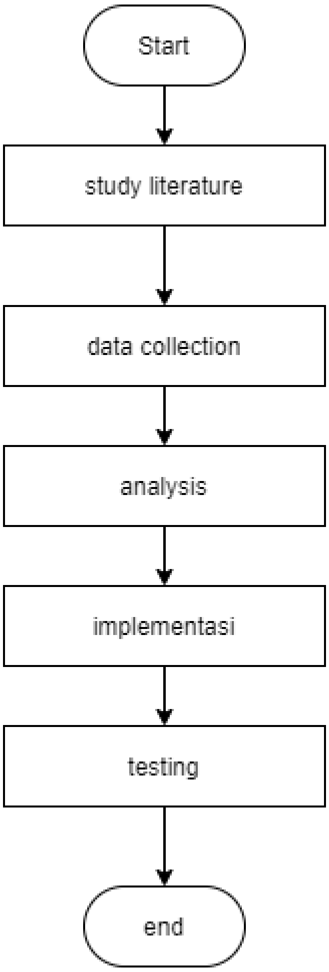


Figure 1. Research Flow

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

No	Variable Name	Detail
1	Rsi 14	Relative strength index
2	sma 9	Simple Moving Average 9 Period
3	sma 180	Simple Moving Average 180Period
4	vwap	178-236
5	spread 14	Spread Indicator
6	volume	Volume Indicator
7	prev close	Previous close
8	prev floor	Previous floor
9	prev ceil	Previous ceil

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

Date Time	Rsi 14	Sma 9	vwap
2020-12-22 09:56:00	38.04415047683149	-0.0036845105193903382	-0.0043935246662244465
2020-12-22 10:05:00	38.83619369907474	-0.00141821155769339	0.003953296551812757
2020-12-22 10:33:00	40.90729730489165	-0.007374467701630083	-0.008215281107206485
2020-12-22 11:45:00	40.44794882957466	-0.0015670997064252123	-0.007888201198146594

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. 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:

$$D\ (x,y) = \sqrt{\Sigma_{k-1}^{n}}(x_{k-}y_k)^2 \quad (3)$$

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

$$D = \sqrt{(38.04415047683149\ -\ 38.83619369907474\,)^2 +} \\ (-0.0036845105193903382\ - -0.00141821155769339\,)^2 + \\ (-0.0043935246662244465\ - 0.003953296551812757\,)^2 \\ =792043222243304{,}00000$$

So the Euclidean Distance obtained is 792043222243304.00000. 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

Date Time	Rsi 14	Sma 9	vwap	Is Profit
2020-12-22 09:56:00	38	-0.004	-0.004	False
2020-12-22 10:05:00	39	-0.001	0.004	True
2020-12-22 10:33:00	41	-0.007	-0.008	False
2020-12-22 11:45:00	40	-0.002	-0.008	True

There is new data that will be used as testing data as shown in Table 4 below.

Table4. Sample Data Testing Naïve Bayes

Date Time	Rsi 14	Sma 9	vwap	Is Profit
2020-12-22 11:46:00	38	-0.004	-0.008	?

The stages of the Naive Bayes process are:

1. Count the number of classes/labels.

P(Ci)
P(Is Profit = “True”) = 2/4
P(Is Profit = “False”) = 2/4

2. Calculating the Number of Cases Per Class

P(X|Ci)
P(rsi 14 = “38” | Is Profit = “True”) = 0/2
P(rsi 14 = “38” | Is Profit = “False”) = 1/2

P(sma 9 = “-0.004” | Is Profit = “True”) = 0/2
P(sma 9 = “-0.004” | Is Profit = “False”) = 1/2

P(vwap = “-0.008” | Is Profit = “True”) = 0/2
P(vwap = “-0.008” | Is Profit = “False”) = 1/2

3. Multiply All Class Variables

P(X|Is Profit = “True”)
= P(rsi 14 = “31”, sma = “-0.004”, vwap = ““-0.001”, | Is Profit = “True”)
= 0/2 * 0/2 * 0/2
= 0

P(X|Is Profit = “False”)
= P(rsi 14 = “31”, sma = “-0.004”, vwap = ““-0.001”, | Is Profit = “False”)
= 1/2 * 1/2 * 1/2
= 0.5 * 0.5 * 0.5
= 0,125

P(X|Ci)*P(Ci)
P(X|Is Profit = “True”)*P(Is Profit = “True”) =
0 * 0/2 = 0

P(X|Is Profit = “False”)*P(Is Profit = “False”) = 0.125 * 2/4
= 0,0625

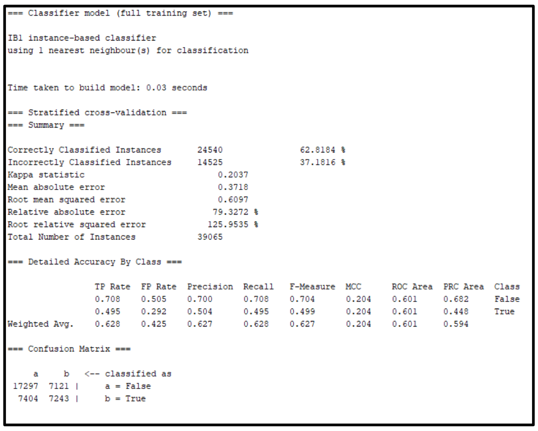
4. Compare Results Per Class

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

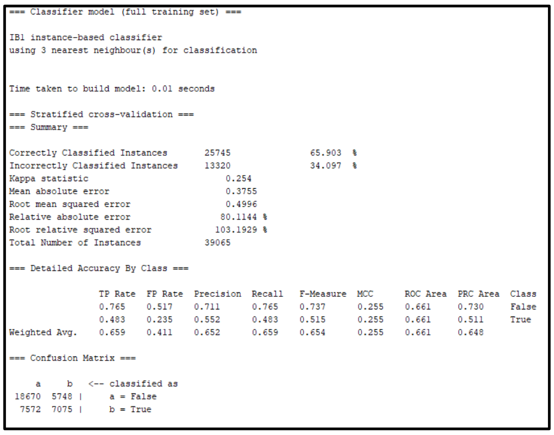
Tests are carried out using the WEKA application, testing is carried out using 4 scenarios, each test is carried out using 10 Fold Cross-Validation, the first is testing using the K-Nearest Neighbor method using the value of K = 1, the second using K = 3, the fourth using K = 5, and the fourth uses the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

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



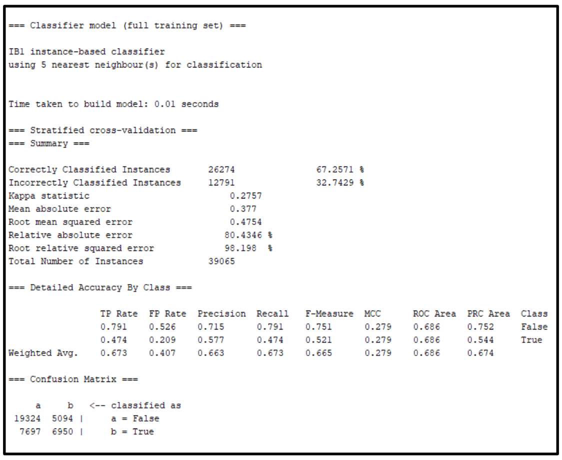
```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       24540               62.8184 %
Incorrectly Classified Instances     14525               37.1816 %
Kappa statistic                          0.2037
Mean absolute error                      0.3718
Root mean squared error                  0.6097
Relative absolute error                 79.3272 %
Root relative squared error            125.9535 %
Total Number of Instances            39065

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.708    0.505    0.700      0.708    0.704      0.204    0.601     0.682     False
                 0.495    0.292    0.504      0.495    0.499      0.204    0.601     0.448     True
Weighted Avg.    0.628    0.425    0.627      0.628    0.627      0.204    0.601     0.594

=== Confusion Matrix ===

     a     b   <-- classified as
 17297  7121 |     a = False
  7404  7243 |     b = True
```

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



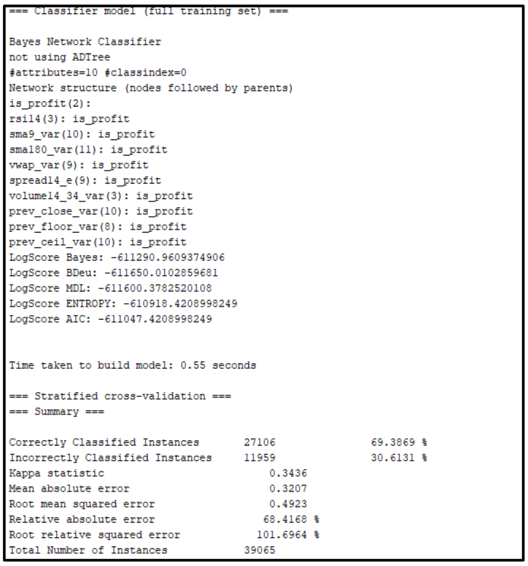
```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       25745               65.903  %
Incorrectly Classified Instances     13320               34.097  %
Kappa statistic                          0.254
Mean absolute error                      0.3755
Root mean squared error                  0.4996
Relative absolute error                 80.1144 %
Root relative squared error            103.1929 %
Total Number of Instances            39065

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.765    0.517    0.711      0.765    0.737      0.255    0.661     0.730     False
                 0.483    0.235    0.552      0.483    0.515      0.255    0.661     0.511     True
Weighted Avg.    0.659    0.411    0.652      0.659    0.654      0.255    0.661     0.648

=== Confusion Matrix ===

     a     b   <-- classified as
 18670  5748 |     a = False
  7572  7075 |     b = True
```

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



```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 5 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       26274               67.2571 %
Incorrectly Classified Instances     12791               32.7429 %
Kappa statistic                          0.2757
Mean absolute error                      0.377
Root mean squared error                  0.4754
Relative absolute error                 80.4346 %
Root relative squared error             98.198  %
Total Number of Instances            39065

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.791    0.526    0.715      0.791    0.751      0.279    0.686     0.752     False
                 0.474    0.209    0.577      0.474    0.521      0.279    0.686     0.544     True
Weighted Avg.    0.673    0.407    0.663      0.673    0.665      0.279    0.686     0.674

=== Confusion Matrix ===

     a     b   <-- classified as
 19324  5094 |     a = False
  7697  6950 |     b = True
```

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



```
=== Classifier model (full training set) ===

Bayes Network Classifier
not using ADTree
#attributes=10 #classindex=0
Network structure (nodes followed by parents)
is_profit(2):
rsi14(3): is_profit
sma9_var(10): is_profit
sma180_var(11): is_profit
vwap_var(9): is_profit
spread14_e(9): is_profit
volume14_34_var(3): is_profit
prev_close_var(10): is_profit
prev_floor_var(8): is_profit
prev_ceil_var(10): is_profit
LogScore Bayes: -611290.9609374906
LogScore BDeu: -611650.0102859681
LogScore MDL: -611600.3782520108
LogScore ENTROPY: -610918.4208998249
LogScore AIC: -611047.4208998249


Time taken to build model: 0.55 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       27106               69.3869 %
Incorrectly Classified Instances     11959               30.6131 %
Kappa statistic                          0.3436
Mean absolute error                      0.3207
Root mean squared error                  0.4923
Relative absolute error                 68.4168 %
Root relative squared error            101.6964 %
Total Number of Instances            39065
```

Figure 5. Test results using Naïve Bayes

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

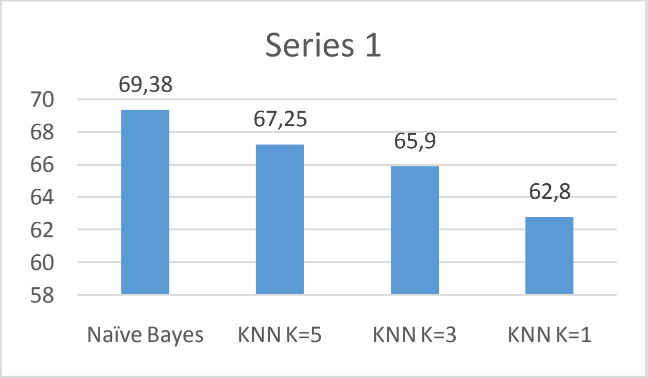


Figure 6. Comparison diagram of the test results of the two methods

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 accuracy value, namely 69.38% then KNN with a K = 5 value has a truth accuracy value of 67.25.

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

[1] M. Nanja and P. Purwanto, "METODE K-NEAREST NEIGHBOR BERBASIS FORWARD SELECTION UNTUK PREDIKSI HARGA KOMODITI LADA," *Pseudocode*, vol. 2, no. 1, pp. 53–64, Aug. 2015.

[2] M. Jannah and N. Nurfauziah, "ANALISIS PENGARUH NILAI TUKAR RUPIAH, TINGKAT SUKU BUNGA SBI (BI RATE) DAN HARGA EMAS DUNIA TERHADAP INDEKS LQ45 DI BURSA EFEK INDONESIA," *J. Manaj. Maranatha*, vol. 17, no. 2, p. 103, May 2018.

[3] M. Guntur, J. Santony, and Y. Yuhandri, "Prediksi Harga Emas dengan Menggunakan Metode Naïve Bayes dalam Investasi untuk Meminimalisasi Resiko," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 2, no. 1, pp. 354–360, 2018.

[4] Liao. 2007. Recent Advances in Data Mining of Enterprise Data: Algorithms and Application. Singapore: World Scientific Publishing.

[5] Budi Santosa. 2007. Data Mining Teknik Pemanfaatan Data UntukKeperluanBisnis. GrahaIlmu.

[6] K. Kaharuddin, K. Kusrini, V. Wati, E. Pawan, and P. Hasan, "Classification of Spice Types Using K-Nearest Neighbor Algorithm," in *2019 International Conference on Information and Communications Technology (ICOIACT)*, 2019, pp. 285–290.

[7] Siregar, A. M., &Puspabhuana, A., 2002. Data Mining Pengolahan Data MenjadiInformasidengan RapidMiner. Sukoharjo: CV Kekata Group.

[8] Muqorobin, M., Rokhmah, S., Muslihah, I., & Rais, N. A. R. (2020). Classification of Community Complaints Against Public Services on Twitter. *International Journal of Computer and Information System (IJCIS)*, *1*(1).

[9] Muqorobin, M., Kusrini, K., Rokhmah, S., & Muslihah, I. (2020). Estimation System For Late Payment Of School Tuition Fees. *International Journal of Computer and Information System*, *1*(1), 341475.