Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Application of Data Mining with K-Nearest Neighbors Algorithm for Shallot Price Prediction

Yuana Inka Dewi Br Sinulingga[1] ✉ and Donny Avianto[2]
[12]Informatics Study Program, Faculty of Science and Technology, Universitas Teknologi Yogyakarta, Indonesia
Corresponding Author: Yuana Inka Dewi Br Sinulingga, **E-mail**: inkasnlg54321@gmail.com

| ABSTRACT
Shallots are an important and widely consumed bulb crop in Indonesia, both for medicinal and culinary purposes. However, shallot yield is substantially affected by its supply, often leading to significant price fluctuations that greatly impact consumers and producers, especially farmers. Farmers who cannot accurately predict shallot prices often incur losses when selling to shallot distributors. If this problem is not resolved, it may discourage farmers from cultivating shallots. Therefore, a prediction system is needed to forecast shallot prices in the future, thus helping farmers make the right decisions. This research uses the K-Nearest Neighbors (KNN) algorithm for shallot price prediction. KNN classifies data into specific categories based on the closest distance to a set of k patterns for each category, using the Euclidean distance formula to calculate the distance. The dataset consists of 303 entries with five features: farmer price, seller price, retail price, seed price, and yield. The test results of the Shallot Price Prediction System in North Sumatra Province, Indonesia, using the K-Nearest Neighbors Algorithm, showed the best performance when using 80% training data and 20% testing data, with a value of k = 2, resulting in a Mean Absolute Error (MAE) of 25,786 and a Mean Squared Error (MSE) of 72. This system empowers farmers to predict the future price of shallots before selling their crops to distributors.

| KEYWORDS

Shallot, Data Mining, Prediction, K-Nearest Neighbor

| ARTICLE INFORMATION

ACCEPTED: 12 October 2023 **PUBLISHED:** 01 November 2023 **DOI:** 10.32996/jcsts.2023.5.4.5

1. Introduction

Indonesian spice commodities represent a significant subset of commodities within the plantation sub-sector, presenting a substantial potential for worldwide market penetration [Anggrasari, 2019]. The investigation of the impact of antioxidants found in spices on human health science is currently a prominent area of research within the field of health studies. Cabbage, potatoes, tomatoes, shallots, and red chilies are among the most extensively cultivated vegetables in Indonesia, exhibiting high levels of output.

Shallots hold a prominent position as a primary culinary enhancer in Indonesian cuisine, rendering them irreplaceable by alternative ingredients. The demand for shallots has been steadily increasing over the years, in line with the growth of the food sector and the expanding global population [Basuki, 2021]. The unbalanced supply and demand of shallot in each province of Indonesia result in trade movements from regions with excess to those with deficits, hence indicating the presence of market integration [Rahmawati, 2018]. North Sumatra is recognized as a prominent province in the production of shallots, and it is also acknowledged as one of the places in Indonesia with the greatest per capita consumption of shallots [Pane, 2020]. Despite the substantial level of output, a significant disparity persists between the quantity of shallots produced and consumed in North Sumatra, primarily attributable to the region's elevated consumption patterns. Therefore, it is imperative to ensure the continuous maintenance of the shallot supply chain in North Sumatra. This is crucial in order to effectively distribute shallots, ensuring price stability and

meeting the demands of consumers [Rahmawati, 2018]. Despite the community's consistent per capita use of shallots, the volatile variations in shallot prices remain an ongoing challenge. Indeed, it is observed that the price of shallots tends to grow during the harvest season, a period characterized by an increase in supply.

The provided visual representation depicts the distribution of shallot prices for the period of 2021-2022.

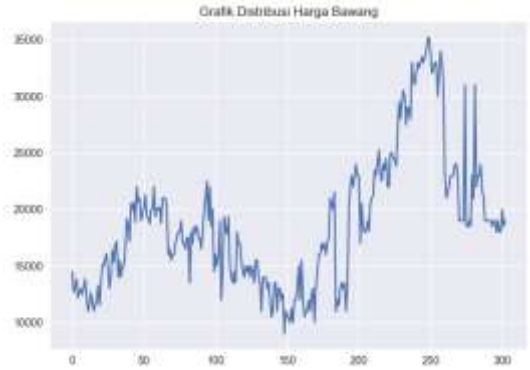

Figure 1.1 Price chart of shallots

Figure 1 shows the fluctuations that occur in shallot prices, if shallot price fluctuations continue to occur, it will cause losses for both producers and consumers. This situation makes shallot farmers in North Sumatra, as one of the shallots producing areas, suffer losses because the price is sometimes unstable, so there is often a gap between capital and profit. If this situation is not resolved immediately, it will cause shallot farmers to lose money and reduce planting in the next planting period because they cannot predict the price of shallots in the future, resulting in a decrease in shallot production. According to Dahlianawati et al., the higher the selling price and the greater the amount of shallot production, the greater the amount of revenue received by shallot farmers, on the contrary, if the amount of shallot production is small coupled with a low selling price, the revenue of shallot farmers will be small which causes losses for shallot farmers [Pane, 2020]

Data mining refers to the systematic collection and subsequent processing of data with the objective of extracting significant and valuable information from the dataset. The acquisition and extraction of data can be accomplished through the utilization of software, which incorporates statistical computations, mathematical algorithms, and Artificial Intelligence (AI) methodologies. Numerous firms are currently adopting its utilization. Data mining is utilized as a valuable tool in addressing the challenges posed by the competitive landscape, namely in the realms of data analysis and the assessment of diverse market trends and patterns. The objective is to examine a range of market trends and patterns in order to produce prompt and efficient market trend analysis. Numerous organizations have adopted the utilization of Data Mining as a strategic tool to address the competitive landscape for data analysis. This approach enables them to assess diverse market trends and patterns, thus facilitating the generation of prompt and efficient market trend analyses [Naik, 2016].

This study uses data mining techniques to develop the K-Nearest Neighbor (KNN) method. The KNN algorithm is well recognized as one of the most effective and commonly utilized classification algorithms [Khaleel, 2017]. The K-Nearest Neighbor algorithm is a prominent illustration of instance-based learning, wherein the training dataset is retained to facilitate the categorization of new, unclassified records. This classification process involves comparing the new record to the most comparable records inside the training set [Larose, 2005]. In order to provide predictions, it relies solely on the training data. Forecasts are generated for each new data point by doing a comprehensive search inside the summary of the output variable over the K examples in the training set, specifically focusing on the K most similar examples (referred to as neighbors) [Kalyani, 2022]. The selection of the distance between data, also known as the K factor, is a crucial parameter that significantly impacts the effectiveness of the Nearest Neighbor algorithm. When selecting the optimal K value that minimizes mistakes in classification or estimation and maximizes accuracy, the data analyzer must carefully balance these concerns [Larose, 2005]. The K factor will be computed via the Euclidean distance formula. The Euclidean distance metric is commonly employed in K-nearest neighbors (KNN) classification [Pulungan, 2019]. The subsequent points outline the primary benefits associated with the K-Nearest Neighbors (KNN) algorithm : (i) The implementation of K-nearest neighbors (KNN) is characterized by its simplicity and ease of justifying the resulting outcomes. (ii) The model exhibits resilience to noisy training data, particularly when employing the Inverse Square of weighted distance as the metric for measuring "distance". (iii) The model demonstrates effectiveness when trained on a large dataset. Despite the advantages of KNN, it is important to acknowledge its associated drawbacks. (i)There is no established guideline to identify the optimal value for the parameter K. (ii) The computational cost of KNN is considerable due to the necessity of calculating the distance between each test instance and all training samples. Lastly, (iii) The accuracy rate in multidimensional data sets is low when irrelevant features are included. The user has provided a numerical reference, indicating the presence of a citation or source.

Based on data from the Food and Agriculture Organization, Indonesia is ranked fourth in terms of shallot exports. Given the contextual backdrop of these issues, there is a pressing need for a system that can aid farmers in forecasting forthcoming shallot prices. To address this, the utilization of data mining, specifically the k-nearest neighbor algorithm, holds promise in mitigating the challenges at hand. By leveraging historical data, this approach seeks to extract pertinent information and discern patterns to effectively address prevailing predicaments.

2. Methodology

2.1 Research Stages

Before starting a research system, steps are needed that have been designed previously. The stages of this research use a flowchart to facilitate the delivery of information on the steps taken in this study. Figure 2.1 shows the stages of research conducted on the shallot price prediction system.

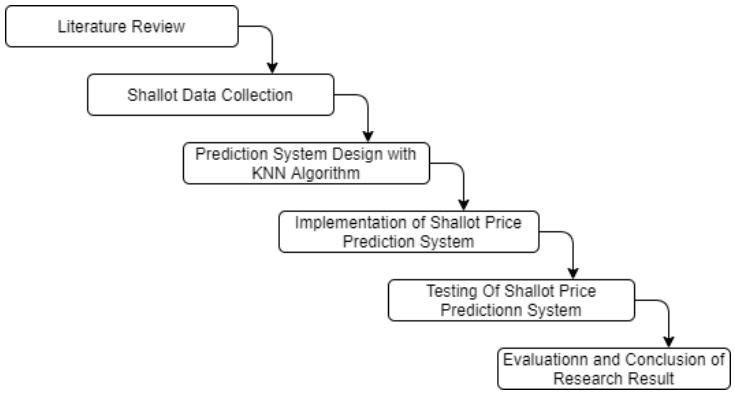

Figure 2.1 Research Framework

1. *Literature Review*

A literature study is carried out for problem identification where the process of understanding the problem to be studied is carried out, which is the reason for this research, by conducting literature studies and literature studies as well as reviewing related research by looking for other source data that will be used as references related to data mining, predictions, the K-Nearest Neighbors method, and so on.

2. *Shallot Data Collection*

This stage study aims to collect shallot data, which will be used as training and testing data. The data used is obtained through shallot agents from Karo district. This dataset consists of 303 data starting from January 2021 to October 2022 based on the rupiah price per kilogram and has 5 features, namely H1 (farmer price), H2 (seller price), H3 (retail price), H4 (seed price), and H5 (crop yield). Here is an example of the data that will be used.

Table 2.1 Shallot price data

Date	H1	H2	H3	H4	H5
July 22, 2022	IDR 31.000	IDR 32.500	IDR 35.500	IDR 76.000	2169/kg
July 23, 2022	IDR 18500	IDR 20.000	IDR 23.000	IDR 51.000	2432/kg
July 26, 2022	IDR 18500	IDR 20.000	IDR 23.000	IDR 51.000	2340/kg
July 27, 2022	IDR 23000	IDR 24.500	IDR 27.500	IDR 60/000	2082/kg

3. *Design of Prediction System with KNN Algorithm*

This prediction system planning is done by analyzing the system first to find out what processes will be made. Then, a design is made that describes the results of this system later. The design encompasses various components, namely input design, output design, database design, and process design.

4. *Implementation of Shallot Price Prediction System*

The implementation of the shallot price prediction system is carried out by transforming the system design into a programming language to be compiled into an application. This implementation process is based on the results of the previous system design.

5. *Testing of Shallot Price Prediction System*

Testing is done to test system performance and find system defects, which are then carried out in the improvement process. In this process, an evaluation of the results of predicting the price of merag onions using the K-Nearest Neighbors algorithm is also carried out.

6. *Evaluation and Conclusion of Research Result*

Evaluation is done by calculating the Mean Squared Error (MSE) value and the Mean Absolute Error (MAE) value based on the calculation of the distance of the k value and the division between training data and testing data that has been determined. Research results will be concluded by paying attention to the results of the calculation of MSE and MAE values with the use of different k values and split data.

2.2 Proposed System

1. Ongoing System Architecture

Figure 2.2 is a description of the current conditions where farmers plant shallots without being able to predict the price of shallots in the future; after the shallots are harvested, the farmers will sell them to the shallot sales agent, then the farmers do not expect that the price of shallots has decreased drastically and often fluctuates suddenly so that the farmers experience losses that make them afraid and reluctant to plant shallots again in the future.

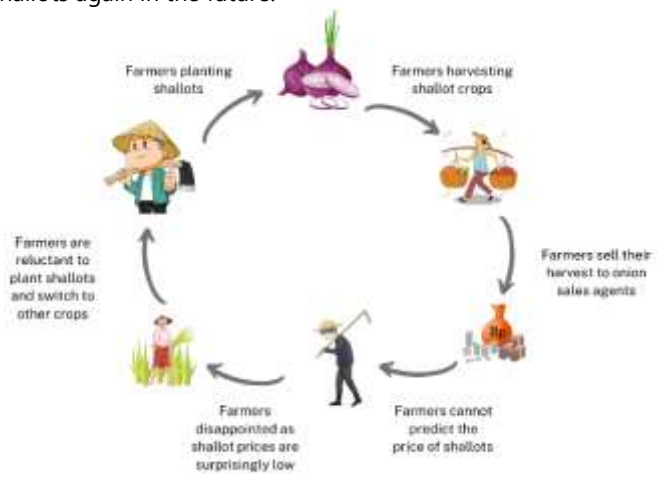

Figure 2.2 Current system architecture

2. Proposed System Architecture

The proposed system for predicting the price of shallots using the K-Nearest Neighbors algorithm can be seen in Figure 2.3.

Figure 2.3 Architecture of the proposed system

Figure 2.3 is a description of the proposed system where the user opens the system that has been made previously using the internet; then, the user will be asked to input shallot price parameters such as shallot prices from farmers to agents, retail shallot prices, harvest areas, and crop yields. Then, the system will perform calculations using the K-Nearest Neighbor algorithm, and then the results of the shallot price prediction calculation will be submitted to the user.

3. System Flowchart

a. System Flowchart Testing Process Page: The flowchart at the testing process stage explains the testing stage carried out by the system; the initial display at this stage is that the system will display the testing data form, the user can see the data from the training stage that has been done before, after that the system will perform the pattern determination calculation process, then the user will be asked to input testing data according to the attributes and the data will be processed in the prediction calculation using the K-Nearest Neighbor method as shown in Figure 2.4 below:

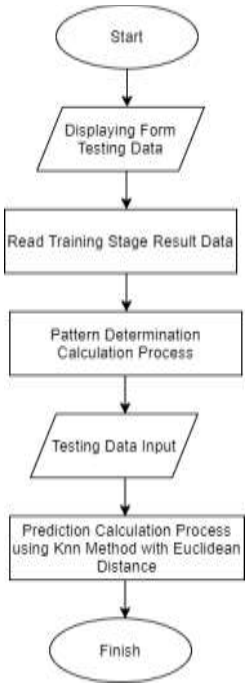

Figure 2.4 Testing process stage system flow

b. K-Nearest Neighbor Algorithm Flowchart: Flowchart at the testing process stage explains how the K-nearest neighbor algorithm works.

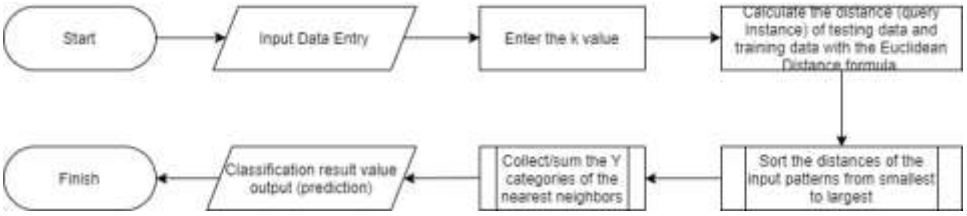

Figure 2.5 K-Nearest Neighbor algorithm system flow

The following is an explanation of the steps in the K-Nearest Neighbors flowchart :

- Start the process
- Input test data in this step, data input will be carried out by users who will test the system the test data is in the form of data on price_seller, price_retail, price_bibit, and yield_harvest.
- Determining the K value

- Then, the distance calculation (query instance) is carried out on the testing data and training data using the Euclidean distance calculation formula.

  Here is the formula for the Euclidean Distance equation

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

  Description :
  d(x,y) = distance between data x to data y
  i = data variable, (i = 1, 2, 3, ..., n)
  n = data dimension
  xi = X value in training data (x1, x2, x3, ..., xn)
  yi = Y value in testing data (y1, y2, y3, ..., yn)

- Next, sort the results of the distance calculation from the smallest to the largest (ascending)
- Sum the Y category values of the nearest neighbors
- The final result of this process is the output of shallot price prediction results.
- Process completed

c. System Interface Design: The interface design is used as a guide for creating the system interface. Figure 2.6 is an example of the interface design of one of the system views.

Figure 2.6 Home page interface design

3. Results and Discussion

3.1 System Implementation

The following is an analysis of the supporting needs of the system that will be created later, such as software for development, devices for testing, and so on. These needs include hardware and software**.**

1. Software requirements
   The following are the software requirements needed, namely:
   a) Operating System Windows 11
   b) Code Editor Visual Studio Code
   c) Draw.io

2. Hardware requirements
   The following are the hardware requirements needed, namely:
   a) Processor : AMD Ryzen 3 4300U with Radeon Graphics
   b) RAM : 8GB DDR4
   c) Storage : 512GB SSD M.2 2242 PCIe NVMe 3.0x2

After the previous interface design process, the results of the system implementation can be seen in Figure 3.1 below.

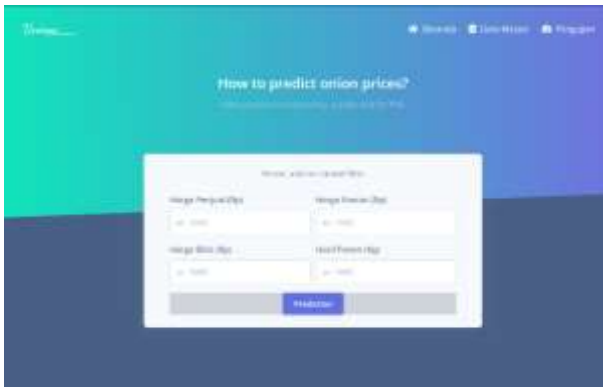

Figure 3.1 Home page implementation

Figure 3.1 Is a dashboard page in the form of a main page when the user enters the system, the page displays several menu bars and a form that must be filled in first before making predictions.

Figure 3.2 Implementation of the master data page

Figure 3.2 This is a master data page that serves to display information from the dataset used. This dataset is presented in the form of a column containing onion price information.

Figure 3.3 Implementation of prediction results

Figure 3.3 is a picture of the prediction result form of shallots by inputting several inputted attributes such as shallot prices (sellers), shallot prices (retail), shallot seed prices, shallot harvests that occur at that time, and the results of shallot price predictions.

Figure 3.4 Testing page implementation

It is a page that displays information about the test model used, such as the amount of test data used, the value of k, the value of the Mean Squared Error (MSE) calculation results and the Mean Absolute Error (MAE) value, and information about the shallot price data distribution graph. Then, on this page, we can also see the history of test results that have been done before.

3.2 Prediction Result

1. First Test with 7 : 3 Ratio

Table 3.1 presents the initial test outcomes for the computation of the mean absolute error (MAE) and mean squared error (MSE) metrics, employing various k values.

Table 3.1 First test result table

No	Number of K	MAE Value	MSE Value
1	K=1	97701	126
2	K=2	96264	137
3	K=3	120255	177
4	K=4	150887	212
5	K=5	159818	231
6	K=6	369383	329
7	K=7	627671	442

As seen in the table, the test results using 70% training data and 30% testing data, the best MAE and MAE values are at k = 1 with an MAE value of 97701 and an MSE value of 126.

2. First Test with 8 : 2 Ratio

Table 3.2 presents the initial test outcomes pertaining to the computation of the mean absolute error (MAE) and mean squared error (MSE) values, employing various k values.

Table 3.2 Second test result table

No	Number of K	MAE Value	MSE Value
1	K=1	37241	75
2	K=2	27586	72
3	K=3	33371	108

4	K=4	34710	113
5	K=5	64743	159
6	K=6	134266	223
7	K=7	368292	321

As seen in the table, the test results using 80% training data and 20% testing data, the best MAE and MAE values are at k = 2 with an MAE value of 27586 and an MSE value of 72.

3. First Test with 9 : 1 Ratio

Table 3.3 presents the initial test outcomes pertaining to the computation of the mean absolute error (MAE) and mean squared error (MSE) values, employing various k values.

Table 3.3 Third test result table

No	Number of K	MAE Value	MSE Value
1	K=1	52068	72
2	K=2	99482	124
3	K=3	124647	128
4	K=4	168965	186
5	K=5	140900	217
6	K=6	191907	338
7	K=7	576805	417

As seen in Table 4 of the test results using 90% training data and 10% testing data, the best MAE and MSE values are at k = 1 with an MAE value of 52068 and an MSE value of 72.

4. Prediction result evaluation model

Furthermore, this research analyzes shallot price prediction by utilizing The utilization of the K-Nearest Neighbor approach for machine learning modeling is employed to test the data. The following are the results that have been done in this study in predicting the price of shallots from farmers to agents.

Table 3.4 Original data and prediction data table

No	Original Data (IDR)	Predicted Data (IDR)
1	24.500	14.750
2	14.000	14.000
3	12.000	12.000
4	20.500	20.750
5	33.000	32.750
6	24.000	23.500
7	20.000	20.000
8	19.500	19.750

9	20.000	20.000
10	18.000	17.750

Table 3.4 is an image of the original data and data prediction of shallot prices from farmers to agents using a data division ratio of 8 : 2 and with the number k = 2; the observed disparity in prices between the original data and the forecasted data is rather small, suggesting that the utilization of the K-Nearest Neighbors algorithm yields favorable outcomes in predictive modeling.

4. Conclusion And Future Scope

Based on the shallot price prediction system using the K-Nearest Neighbors algorithm with several parameter inputs, it can be concluded that the model created is able to predict shallot prices in the future well. The optimal test results are obtained by employing 80% training data and 20% testing data split, with a value of k = 2. This configuration yields a Mean Absolute Error (MAE) value of 25786 and a Mean Squared Error (MSE) value of 72. Consequently, it can be inferred that this system has the potential to assist farmers in forecasting future shallot prices prior to selling their produce to sales agents.

This research certainly still has shortcomings where this system only uses a few features as factors that affect the fluctuations in shallot prices so that for development, it can add other factors that affect shallot prices. This system can also be developed for platforms other than the web so that the system can be easily accessed anywhere and from many platforms.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Anggrasari H and Mulyo, J.H. (2019). The Trade Of Indonesian Spice Commodities In International Market, *Agro Ekonomi*, vol. 30, doi: 10.22146/ae.41665.
[2] Basuki, S., Wulanjari, M. E., Komalawati, V and Sahara, D. (2021). The Performance of Production, Price and Marketing System of Shallot in Central Java, in *E3S Web of Conferences*, EDP Sciences, Nov. doi: 10.1051/e3sconf/202131602004.
[3] Khaleel, A.H., Al-Suhail, G.A and Hussan, B.M. (2017). *International Journal of Computer Science and Mobile Computing* A Weighted Voting of K-Nearest Neighbor Algorithm for Diabetes Mellitus. [Online]. Available: www.ijcsmc.com
[4] Kalyani M. (2022). A Study on Crop Yield Prediction using Machine Learning Techniques. [Online]. Available: www.imd.gov.in
[5] Larose, D.T. (2005). *Discovering Knowledge in Data*, John Wiley & Sons, United States of America
[6] Larose, D.T. (2005). *Discovering Knowledge in Data*, John Wiley & Sons, United States of America.
[7] Naik R and Deepika, N. (2016). Data Mining System and Applications: A Study, *International Journal of Computer Science and Mobile Computing (IJCSMC)*.103 - 110. Dec.
[8] Pane T.C and Supriana, T. (2020). The supply chain of North Sumatera shallot, in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics Publishing, Apr. doi: 10.1088/1755-1315/454/1/012037.
[9] Pane T. C and Supriana, T. (2020). The supply chain of North Sumatera shallot," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics Publishing, Apr. doi: 10.1088/1755-1315/454/1/012037.
[10] Pulungan, A.F., Zarlis, M and Suwilo, S. (2019). Analysis of Braycurtis, Canberra and Euclidean Distance in KNN Algorithm, *SinkrOn*, doi: 10.33395/sinkron.v4i1.10207.
[11] Rahmawati, A., Fariyanti, A and Rifin, A. (2018). Spatial Market Integration of Shallot in Indonesia, *Jurnal Manajemen dan Agribisnis*, Nov. doi: 10.17358/jma.15.3.258.
[12] Rahmawati, A., Fariyanti, A and Rifin, A. (2018). Spatial Market Integration of Shallot in Indonesia, *Jurnal Manajemen dan Agribisnis*, Nov. 2018, doi: 10.17358/jma.15.3.258.
[13] Yu W and Zhengguo, W.A. (2007). *Fast KNN algorithm for text categorization, in Proc. of 6thInternational Conference on Machine Learning and Cybernetics*, Hong Kong,