Analysis of Accuracy Improvement in K-Nearest Neighbor using Principal Component Analysis (PCA)

A H Lubis1*, P Sihombing1* and E B Nababan1*

1 Faculty of Computer Science and Information Technology, University of Sumatera Utara, Medan, Indonesia

*Email: auliahadi17@gmail.com, poltak@usu.ac.id, ernabr@usu.ac.id

Abstract. This research conducted to simplify and eliminate attributes or features that are less relevant without reducing the intent and purpose of the original data using Principal Component Analysis (PCA) to improve the performance (accuracy) of the K-NN (K-Nearest Neighbour) classification method. Modifying the method yields an average accuracy of 88%, where the K value is at K = 3 to K = 9. While the lowest accuracy value generated from K-Nearest Neighbors Conventional with PCA + K-NN classification models for the Pekanbaru Air Quality dataset has an average accuracy of 87% where when the K value is at K = 2, K = 4, K = 6 and K = 8. The highest average error is 0.11% where when the value of K is at K = 2, K = 4, K = 6 and K = 9. While the lowest average error value generated from K-Nearest Neighbors Conventional with PCA + K-NN classification model of the Pekanbaru Air Quality dataset is 0.08% where when the K value was at K = 5.

1. Introduction
Based on research by Raikwal in 2012, the results obtained using the process of cross-validation and Euclidean Distance to find accuracy and other performance parameters [8]. There are deficiencies in the K-NN implementation stage to get proper classification results, only not in textual data; the occurrence of performance parameters varies according to the size of the dataset. If the size of the data increases, the results obtained on K-NN are not right.

Meanwhile, K-nearest neighbor produces 159 prediction data that are close to the same difference only between the first choice and second choice and 16 data that have the same prediction between the first and second choice due to the many attributes used so that the prediction results are less accurate.

Hussain et al. used the Principal Component Analysis (PCA) approach as a method of feature selection to reduce indicators related to the prediction of survival rates of patients infected with breast cancer [3]. The data used were from the SEER dataset of 684,394 patient medical records. With the proposed approach, the accuracy is 92%.
2. Research methods

2.1. K-Nearest Neighbor
K-Nearest Neighbor (KNN) is a group of instances-based learning. K-Nearest Neighbor is also one of the lazy learning methods [2]. KNN finds the k group of objects in the training data closest to the object in the new data or testing data [8]. K-Nearest Neighbor algorithm is a method that functions to classify objects based on learning data that is the closest distance to the tested object [5]. Nearest Neighbor is an approach to look for cases by calculating the closeness between a new problem with an old problem in matching weights from the number of existing features [6].

In the training phase, this algorithm only stores feature vectors and classifies training sample data. In this phase, the same features are calculated for testing data whose classification is unknown. The distance of the new vector to the entire training sample vector is calculated, and the closest number of k is taken.

2.2. Principal component analysis
PCA is a linear combination of primary variables that geometrically this linear combination is a new coordinate system obtained from the rotation of the original system. The PCA method is advantageous to use if the data available has a large number of variables and has a correlation between the variables. The calculation of principal component analysis is based on the calculation of Eigenvalues and eigenvectors that represent the spread of data from a dataset. By using PCA, variables that were as many as n variables will be selected into k new variables called principal components, with the number of k less than n. Using only the k-principal component will produce the same value using n variables. The outcome variable of the selection is called the principal component.

Principal component analysis as a method for obtaining coefficient values or weights from the linear combination of the forming variables is as follows [7]:

• There are as many p principal components as there are variables observed and each principal component is a linear combination of these variables
• Each principal component is orthogonal to one another (perpendicular) and free from each other.
• Principal components are formed based on the order of variance from the largest to the smallest.

3. Identification problems
One of the weaknesses of the classification technique in the K-NN method is the level of accuracy that is still not optimal. It is influenced by many attributes that are less relevant so that it affects the accuracy of the classification. Therefore, a method is needed to reduce the attributes that are less relevant to distance calculation is performed so that it is expected to be able to improve the classification accuracy of the K-NN method.

4. Result and discussion
This section describes the implementation of research with the support of the Python version programming language 3.7.3 and Visual Studio code editor code. This research also aims to improve the accuracy of the K-Nearest Neighbors (K-NN) classification using the Principal Component Analysis (PCA) method and see the comparison. Both methods use the conventional K-Nearest Neighbors classification method based on the resulting accuracy.
4.1. Result of initial data preparation (data preprocessing)

Table 1. Normalization results of Pekanbaru city air quality data [1].

| No | PM10 | SO2  | CO   | O3   | NO2  |
|----|------|------|------|------|------|
| 1  | -0.14| 2.80 | -0.20| 1.04 | -0.86|
| 2  | -0.13| 2.80 | -0.12| -0.06| -0.86|
| 3  | -0.23| 2.80 | -0.12| -0.46| -0.86|
| 4  | -0.36| 2.72 | -0.65| 0.46 | -0.92|
| 5  | -0.35| 2.72 | -0.58| -0.09| -0.92|
| 6  | -0.41| 2.72 | -0.58| 0.53 | -0.86|
| 7  | -0.44| 2.72 | -0.35| 0.82 | -0.92|
| 8  | -0.39| 2.72 | -0.50| -0.09| -0.92|
| 9  | -0.33| 2.72 | -0.27| -0.09| -0.86|
| 10 | -0.26| 2.72 | -0.43| -0.09| -0.86|
| ⋮  | ⋮    | ⋮    | ⋮    | ⋮    | ⋮    |
| 1080| -0.23| -0.48| -0.27| -0.57| -0.59|

Z-Score for PM10 attribute from data to -1 as equation 1:

\[ z = \frac{x - \mu}{\sigma} \]  
\[ z = \frac{47 - 61.66}{105.50} = -0.14 \]  

4.2. Results of PCA Analysis

4.2.1. Correlation calculation results. The next process is to calculate the correlation value between attributes using the covariance equation. Covariance measures the magnitude of the relationship between two attributes. The following is the calculation of Pekanbaru City Air Quality values based on normalization results for PM10 (X1) attributes with SO2 (X2) attributes, namely:

Table 2. Correlation Calculation (Covariance) Air Quality dataset in Pekanbaru City.

| No | X1  | X2  | X1-Xavg | X2-Xavg | Product |
|----|-----|-----|---------|---------|---------|
| 1  | -0.14| 2.80| -0.14   | 2.80    | -0.39   |
| 2  | -0.13| 2.80| -0.13   | 2.80    | -0.36   |
| 3  | -0.23| 2.80| -0.23   | 2.80    | -0.64   |
| 4  | -0.36| 2.72| -0.36   | 2.72    | -0.98   |
| 5  | -0.35| 2.72| -0.35   | 2.72    | -0.95   |
| ⋮  | ⋮    | ⋮    | ⋮       | ⋮       | ⋮       |
| 1080| -0.23| -0.48| -0.23   | -0.48   | 0.11    |

\[ \text{Cov}(X_1, X_2) = \frac{\sum(-0.14-0)(2.80-0)+(-0.13-0)(2.80-0)+(-0.23-0)(2.72-0)+(-0.35-0)(-0.48-0)}{1080-1} = 0.297 \]

Then proceed with calculating the percentage value of the variance proportion of covariance for the Pekanbaru City Air Quality dataset as follows:
Proporsi PC 1 (%) = \frac{\text{Eigen PC} 1}{\text{Varians kovarians}} \times 100\% = \frac{284.5}{5} \times 100\% = 56.90\%

Proporsi PC 2 (%) = \frac{\text{Eigen PC} 2}{\text{Varians kovarians}} \times 100\% = \frac{85.33}{5} \times 100\% = 17.06\%

Proporsi PC 3 (%) = \frac{\text{Eigen PC} 3}{\text{Varians kovarians}} \times 100\% = \frac{59.18}{5} \times 100\% = 11.83\%

Proporsi PC 4 (%) = \frac{\text{Eigen PC} 4}{\text{Varians kovarians}} \times 100\% = \frac{45.47}{5} \times 100\% = 9.09\%

Proporsi PC 5 (%) = \frac{\text{Eigen PC} 5}{\text{Varians kovarians}} \times 100\% = \frac{25.70}{5} \times 100\% = 5.14\%

Eigen value is obtained by multiplying the value of the proportion of the variance with the sum of the variance of the overall attributes of the covariance. Can be seen as follows:

\[
\text{Eigen value} = \frac{\text{Proportion of variance}}{\text{Covariance variance}} \times \text{Overall variance of attributes} \tag{2}
\]

so on in the same way also applies to PC 2 to PC 5.

4.2.2. Orthogonal principal component (PC) results. In this study, the number of principal components chosen is the maximum proportion of variance-covariance that can explain the variance of covariance from the original attribute 1 and 2.

Table 3. Orthogonal Principal Components of Air Quality in Pekanbaru City.

| PC  | Eigen | Variance proposition (%) | Cumulative |
|-----|-------|--------------------------|------------|
| 1   | 284.5 | 56.90                    | 56.90      |
| 2   | 85.33 | 17.06                    | 73.95      |

Variance Threshold = 73.95%

The percentage of the variance of PC 1 variance fulfills 56.90% of the total variance of original data covariance so that the required percentage of variance for PC2 variance is 17.06%. Thus, PC1 and PC2 can represent the maximum proportion of covariance variance from the original data of 73.95% of the original data.

Table 4. Dominant factors of air quality in Pekanbaru City.

| No | Factor | Attribute | Eigenvalue | Load value | Variance % |
|----|--------|-----------|------------|------------|------------|
| 1  | PC 1   | PM10      | 284.5      | 0.489      | 56.90%     |
| 2  | PC 1   | CO        | 284.5      | 0.510      | 56.90%     |
| 3  | PC 2   | SO2       | 85.33      | 0.819      | 17.06%     |
| 4  | PC 2   | NO2       | 85.33      | 0.496      | 17.06%     |
4.3. Accuracy results of conventional K-Nearest Neighbors classification

![Figure 1. Conventional Accuracy-NON Score (K-Fold = 1 to 10).](image1)

![Figure 2. Conventional K-NN Error-Rate Graph with K Value (K = 1 to 10).](image2)

Table 5. Configuration Matrix Data Air Quality Pekanbaru City (K-NN Conventional).

| Actual Class       | Good | Dangerous | Very unhealthy | Middle | No data | Unhealthy |
|--------------------|------|-----------|----------------|--------|---------|-----------|
| Actual. Good       | 177  | 0         | 0              | 5      | 0       | 0         |
| Actual. Dangerous  | 0    | 7         | 0              | 0      | 0       | 0         |
| Actual. Very unhealthy | 0  | 1         | 1              | 0      | 0       | 0         |
| Actual. Middle     | 16   | 0         | 0              | 48     | 0       | 0         |
| Actual. No data    | 1    | 0         | 0              | 0      | 0       | 0         |
| Actual. Unhealthy  | 0    | 0         | 0              | 3      | 0       | 11        |

Next proceed to calculate the Accuracy value and classification error rate (Classification_error) of the Conventional K-Nearest Neighbors Classification Model against the Pekanbaru Air Quality dataset, the following are the results of the calculation:

- **Accuracy** = \( \frac{177+7+1+48+0+11}{177+7+1+48+0+11+16+1+1+5+3} = \frac{244}{270} = 0.9037 \times 100\% = 90.37\% \)
- **Error classification** = \( \frac{16+1+1+5+3}{177+7+1+48+0+11+16+1+1+5+3} = \frac{26}{270} = 0.096 \times 100\% = 9.63\% \)
4.4. Testing the Accuracy of the PCA + K-NN Classification Model

![Figure 3. Accuracy PCA+ K-NN (K-Fold=1 to 10).](image)

![Figure 4. Error-rate PCA+ K-NN to K Value (K=1 to 10).](image)

Table 6. Configuration Matrix Data Air Quality Pekanbaru City (PCA + K-NN).

| Actual Class           | Predicted Well | Predicted Dangerous | Predicted Very unhealthy | Predicted Almost Healthy | Predicted There is no data | Predicted Not healthy |
|------------------------|----------------|---------------------|--------------------------|--------------------------|----------------------------|-----------------------|
| Actual. Well           | 175            | 0                   | 0                        | 7                        | 0                          | 0                     |
| Actual. Dangerous      | 0              | 7                   | 0                        | 0                        | 0                          | 0                     |
| Actual. Very unhealthy | 0              | 0                   | 2                        | 0                        | 0                          | 0                     |
| Actual. Almost Healthy | 15             | 0                   | 0                        | 48                       | 0                          | 0                     |
| Actual. There is no data | 0            | 0                   | 0                        | 1                        | 0                          | 0                     |
| Actual. Not healthy    | 0              | 0                   | 0                        | 0                        | 0                          | 14                    |

Next proceed to calculate the Accuracy value and the classification error level (Classification_error) of the PCA + K-NN Classification Model against the Pekanbaru Air Quality dataset, the following are the results of the calculation:

- **Accuracy** = \( \frac{175 + 7 + 2 + 48 + 0 + 14}{270} = \frac{246}{270} = 0.911 \times 100\% = 91\% \)
- **Error classification** = \( \frac{15 + 7 + 1 + 1}{270} = \frac{24}{270} = 0.088 \times 100\% = 8.8\% \)

4.5. Test results of classification models

Following are the test results and accuracy comparison based on the values of K = 1 to K = 10 K-Nearest Neighbors Conventional classification model using PCA + K-NN classification model using the Cross-Validation method, the number of folds as much as K = 1 to K = 10 in the Pekanbaru Air Quality dataset, can be seen in the following chart:
The following results of the evaluation and comparison of the Error-average for the values of K = 1 to K = 9 of the classification model of the conventional K-Nearest Neighbors classification model with the PCA + K-NN classification model on the Pekanbaru Air Quality dataset, can be seen in the following graph:

5. Conclusion
Average accuracy of 88% where when the value of K is at K = 3 to K = 9. While the lowest accuracy value generated from K-Nearest Neighbors Conventional with PCA + K-NN classification models for the Pekanbaru Air Quality dataset has an average accuracy of 87% where when the K value is at K = 2, K = 4, K = 6 and K = 8.

The highest average error is 0.11% where when the value of K is at K = 2, K = 4, K = 6 and K = 9. While the lowest average error value generated from K-Nearest Neighbors Conventional with PCA + K-NN classification model of the Pekanbaru Air Quality dataset is 0.08% where when the K value was at K = 5.
References

[1] Danades, A., Pratama, D., Anggraini, D., Anggriani, D. 2016. Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status. *International Conference on System Engineering and Technology*, pp. 137-141.

[2] Gou, J., Yi, Z., Du, L. & Xiong, T. January 2012. A Local Mean-Based k-Nearest Centroid Neighbor Classifier. *The Computer Journal* 55(6): pp. 1058-1071.

[3] Hussain, H., Quazilbash, N.Z., Bai, S. & Khoja, S. 2015. Reduction of Variables for Predicting Breast Cancer Survivability Using Principal Component Analysis. *International Conference on Computer-Based Medical Systems*, pp. 131-134.

[4] Iswarya, P. & Radha, V. March 2015. Ensemble learning approach in Improved K Nearest Neighbor Algorithm for Text Categorization. *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIECS)*, pp. 1-5.

[5] Jo, T. March 2017. Using K Nearest Neighbors for Text Segmentation with Feature Similarity. *International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, pp. 1-5.

[6] Kotu, V., & Desphande, B. 2015. Predictive Analytics and Data Mining. Waltham, USA: Morgan Kaufmann Publishers.

[7] Raikwal, J.S & Saxena, Kanak. 2012. *Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set*. International Journal of Computer Applications (0975-8887). Volume.50. pp. 59-61.

[8] Purwanti, E. and Calista, E.. 2017. Detection of Acute Lymphocyte Leukemia Using K-Nearest Neighbor Algorithm based on Shape and Histogram Features. *Journal of Physics: Conference Series* Vol. 853, No. 1, p. 012011.