International Journal of Informatics and Information Systems ISSN 2579-7069
Vol. 5, No. 4, December 2022, pp. 175-180

# Comparison of Data Normalization for Wine Classification Using K-NN Algorithm

Rohitash Chandra [1,*], Kaylash Chaudhary [2], Akshay Kumar [3]

School of Science and Technology, The University of Fiji, Lautoka, Fiji
[1] rohitashc1@unifiji.ac.fj*; [2] kaylashc@unifiji.ac.fj; [3] akshay@unifiji.ac.fj
* corresponding author

(Received: September 13, 2022 Revised: October 15, 2022 Accepted: December 10, 2022, Available online: December 22, 2022)

**Abstract**

The range of values that are not balanced on each attribute can affect the quality of data mining results. For this reason, it is necessary to pre-process the data. This preprocessing is expected to increase the accuracy of the results from the wine dataset classification. The preprocessing method used is data transformation with normalization. There are three ways to do data transformation with normalization, namely min-max normalization, z-score normalization, and decimal scaling. Data that has been processed from each normalization method will be compared to see the results of the best classification accuracy using the K-NN algorithm. The K used in the comparisons were 1, 3, 5, 7, 9, 11. Before classifying the normalized wine dataset, it was divided into test data and training data with k-fold cross validation. The division of the data using k is equal to 10. The results of the classification test with the K-NN algorithm show that the best accuracy lies in the wine dataset which has been normalized using the min-max normalization method with K = 1 of 65.92%. The average obtained is 59.68%.

*Keywords:* Normalization, K-fold cross validation, K-NN

## 1. Introduction

Wine is the result of anaerobic fermentation (without the presence of O2) of grape juice in the form of an alcoholic beverage, by yeast. Wine is a popular drink that is very much in demand, especially abroad. Not only as connoisseurs, but some people who frequently consume various types of wine develop into wine experts. The wine expert is responsible for labeling the types of wine. Therefore, classification can be done on wine data to reduce the role of experts in labeling [1].

Preprocessing is an initial stage that must be carried out in data mining. The purpose of pre-processing in data mining is to prepare raw data before other processes are carried out. Data preprocessing is done by eliminating inappropriate data or changing data into a form that is easier for the system to process. Preprocessing is also carried out to obtain more accurate results, reduce calculation time for large scale problems, and make data values smaller without changing the information contained in them. Data preprocessing can be in the form of data cleaning, data integration, data reduction, and data transformation [2].

In some datasets there are different ranges of values for each attribute. The difference in the range of values for each attribute causes the attribute that has a much smaller value to function than the other attributes. Therefore, it is necessary to transform the data with normalization to equate the range of values for each attribute with a certain scale. In order to produce better data mining. Data transformation with normalization can be done in several ways, namely min-max normalization, z-score normalization, decimal scaling, sigmoid, and softmax.

Classification is one of the important stages in data mining. Classification is grouping new data or objects into classes or labels based on certain attributes [9]. The technique of classification is to look at variables from existing data groups. Classification aims to predict the class of an object that is not known beforehand. Classification consists of three stages, namely model building, model application, and evaluation. Model building is building a model using training data that already has attributes and classes. Then, these data are applied to determine the class of the new data or object. After that, the data is evaluated to see the level of accuracy of the development and application of the

model to the new data [10]. The classification process consists of two phases, namely the training phase and the testing phase. The training phase is the phase where the data is used to build a model while the testing phase is testing the model that has been made with other data to determine the accuracy of the model [3].

Research related to the classification of wine has been done before, the first obtained results of an accuracy of 68.75% [8]. Then the second classification uses the k-Nearest Neighbor (K-NN) algorithm by applying the k-fold cross validation method (k = 3) in dividing data to produce better accuracy, which is equal to 72.97% [4]. Based on the indicators above, we will conduct research on the effect of data transformation using the normalization method for accuracy results in classification using the K-NN algorithm

## 2. Research Methodology

Several stages of the research are outlined in the diagram as follows:

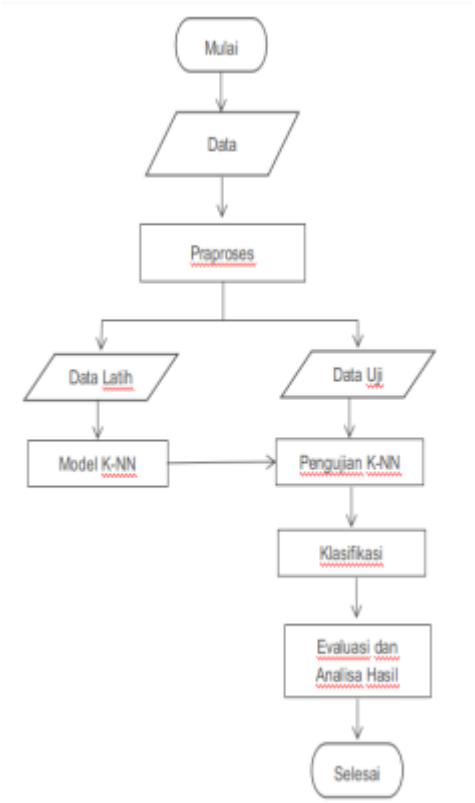


**Figure. 1.** Flowchart

### 2.1. Research Datasets

The dataset used is the wine dataset taken from UCI Machine Learning. The wine dataset is 1599 data. The dataset has eleven attributes and one output attribute in the form of a class with a range of 0-10. The ranges are displayed in letter form, namely z=0, a=1, b=2, c=3, d=4, e=5, f=6, g=7, h=8, i=9, j= 10.

### 2.2. Preprocess

The pre-processing stage is carried out as an initial stage and an important stage in research. The method used in this study is data transformation using normalization. Normalization is the process of scaling attribute values from data so that they can lie in a certain range [5]. The following are the stages of normalization that were carried out:

1) Min-Max Normalization: Min-Max normalization is a normalization method by carrying out a linear transformation of the original data so as to produce a balance of comparative values between the data before and after processing [6]. This method can use the following formula:

$$normalized\,|x| = \frac{minRange + |x - minValue|\,|maxRange - minRange|}{maxValue - minValue}$$

**Figure. 2.** Equality 1

2) Z-score Normalization: Z-score normalization is a normalization method based on the mean and standard deviation of the data. This method is very useful if you do not know the actual minimum and maximum values of the data. The formula used is as follows:

$$nilaibaru = \frac{nilailama - mean}{stdev}$$

**Figure. 3.** Equality 2

3) Decimal Scaling Normalization: Decimal scaling is a normalization method by moving the decimal value of the data in the desired direction. The formula used is as follows:

$$nilaibaru = \frac{nilailama}{10^{i}}$$

**Figure. 4.** Equality 3

## 2.3. Classification

The classification stage is the stage for classifying the quality of the wine dataset. Classification stage as follows:

1) K-fold Cross Validation: Cross validation is a model validation technique to assess the accuracy of analysis results. Data that has been pre-processed is cross-validated by dividing the data into training data and test data for the classification process [7]. Data division was carried out using k-fold cross validation with a value of k equal to 10.

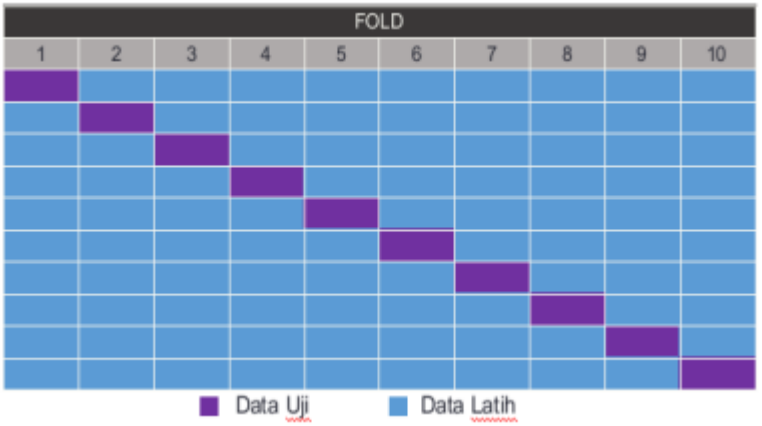


**Figure. 5.** Illustration of 10-fold cross validation

2) K-Nearest Neighbor: After dividing the test data and training data, the classification process is continued using K-NN. The basic concept of K-NN is to find the shortest distance between the data to be evaluated and its k nearest neighbours. The value of the distance between the test data and the training data is sorted from the lowest value. The sorting process is carried out to choose a minimum distance of K fruit. The k values used in this study are 3, 5, 7, and 11. The calculation is done with the following equation:

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

**Figure. 6.** Equality 4

## 2.4. Evaluation

The evaluation stage is carried out by analyzing the accuracy of the wine dataset. The accuracy calculation is done by dividing the number of correct test data (true positive and true negative) by the total number of test data then multiplied by 100%.

$$akurasi = \frac{true\ positive + true\ negative}{true\ positive + true\ negative + false\ positive + false\ negative}$$

**Figure. 6.** Equality 5

## 3. Result and Discussion

This stage is the decomposition of the research results obtained along with their explanations.

## 3.1. Preprocess

In this stage, pre-processing of the wine dataset is carried out. There are several preprocessing stages that are carried out, including min-max normalization, z-score normalization, decimal scaling.

**Table. 1.** Wine Dataset

| No | Fixed acidity | volatile acidity | Citric acid | …… | quality |
|---|---|---|---|---|---|
| 1 | 7.4 | 0.7 | 0 | …….. | e |
| 2 | 7.8 | 0.88 | 0 | …….. | e |
| 3 | 7.8 | 0.76 | 0.04 | ……. | e |
| 4 | 11.2 | 0.28 | 0.56 | …….. | f |
| …. | ….. | ….. | ……. | …… | …… |
| 1599 | 6 | 0.31 | 0.47 | ……. | f |

1) Min-Max Normalization: In table I you can see the original value of the wine dataset before pre-processing. Many data have different ranges that must be normalized. The dataset is transformed using the min-max normalization method by processing the minimum and maximum values of each attribute. The range used in this method is 0-1. The formula in equation 1 can be used to normalize using the min-max normalization method. The results of changing the value with this method can be seen in table 2. The resulting value after processing has a balanced range of values.

**Table. 2.** Wine dataset after Min-Max Normalization

| No | Fixed acidity | volatile acidity | Citric acid | …… | quality |
|---|---|---|---|---|---|
| 1 | 0.247788 | 0.39726 | 0 | …….. | e |
| 2 | 0.283186 | 0.520548 | 0 | …….. | e |
| 3 | 0.283186 | 0.438356 | 0.04 | ……. | e |
| 4 | 0.584071 | 0.109589 | 0.56 | …….. | f |

| …. | ….. | ….. | ……. | …… | …… |
|---|---|---|---|---|---|
| 1599 | 0.283186 | 0.130137 | 0.47 | ……. | f |

2) Z-score Normalization: The original wine dataset which can be seen in table I will be re-transformed with a different method. The next method used is z-score normalization. The formula used in this method can be seen in equation 2. Z-score normalization is done by processing the mean and standard deviation of the attribute values. The results of the transformation of this method can be seen in table III.

**Table. 3.** Wine dataset after Z-Score Normalization

| No | Fixed acidity | volatile acidity | Citric acid | …… | quality |
|---|---|---|---|---|---|
| 1 | -0.528194 | 0.961576 | -1.391037 | …….. | e |
| 2 | -0.298454 | 1.966827 | -1.391037 | …….. | e |
| 3 | -0.298454 | 1.29666 | -1.185699 | ……. | e |
| 4 | -1.332285 | -1.384011 | 1.483689 | …….. | f |
| …. | ….. | ….. | ……. | …… | …… |
| 1599 | -1.332285 | -1.216469 | 1.02168 | ……. | f |

3) Decimal Scaling: For further comparisons, the wine data will be pre-processed using a data transformation method, namely decimal scaling. Table IV is the result of normalization with the decimal scaling method.

**Table. 4.** Wine dataset after Decimal Scaling

| No | Fixed acidity | volatile acidity | Citric acid | …… | quality |
|---|---|---|---|---|---|
| 1 | 0.074 | 0.07 | 0 | …….. | e |
| 2 | 0.078 | 0.088 | 0 | …….. | e |
| 3 | 0.078 | 0.076 | 0.04 | ……. | e |
| 4 | 0.112 | 0.028 | 0.56 | …….. | f |
| …. | ….. | ….. | ……. | …… | …… |
| 1599 | 0.06 | 0.031 | 0.47 | ……. | f |

In calculating the new value, the formula in equation 3 is used. It can be seen that the changes that have occurred, the values generated for each attribute have a range that is not too far away.

## 3.2. Classification

After preprocessing the data and before classifying, the data is divided into test data and training data first. Data sharing is done by k fold cross validation. K used in k-fold is equal to 10. The next step is to classify the wine dataset that has been pre-processed using the K-NN algorithm. K used, namely 3, 5, 7, 11. Calculation of the distance in this algorithm can be seen in equation 4. Following are the results of the research table.

**Table. 4.** Accuracy results with the normalization method using K-NN

| K-NN | Normalization Method | | |
|---|---|---|---|
| | Decimal scaling | Min-max normalization | Z-score normali zation |
| K=1 | 63,10% | 65,92% | 65,85% |
| K=3 | 52,47% | 59,35% | 59,22% |
| K=5 | 53,22% | 57,41% | 56,60% |
| K=7 | 50,47% | 58,03% | 57,54% |
| K-9 | 51,66% | 58,66% | 57,60% |
| K=11 | 51,47% | 58,72% | 57,85% |
| Mean | 53,73% | 59,68% | 59,11% |

The final stage in this process is calculating accuracy using the formula in equation 5. The accuracy results can be seen in table V. Comparison between the accuracy of the results of the min-max normalization, z-score normalization, decimal scaling methods shows that the highest accuracy lies in the data processed using the method min-max normalization with an average accuracy of 59.68%.

## 4. Conclusion

From the results of the research that has been done, it can be concluded that:

- The highest accuracy lies in the wine dataset using the min-max normalization method in the pre-processing stage with K = 1 of 65.92%.
- Accuracy with the highest average is 59.68% using the min-max normalization method.
- The lowest accuracy is found in the dataset using the decimal scaling method, with an average of 53.73%..
- The choice of data preprocessing method in data mining affects the accuracy of the data classification results.
- This study found that the accuracy of using pre-processed data with the normalization method was no better than the accuracy of previous studies conducted by Arandika et al. 2014 with an accuracy rate of 68.75% and research by Saputra & Siahaan. 2007 amounted to 72.97%.

## References

[1] K. Saputra and A. P. U. Siahaan, “Klasifikasi Data Minuman Wine Menggunakan Algoritma K-Nearest Neighbor,” pp. 2– 4, 2007.

[2] S. Mujilahwati, “Pre-Processing Text Mining Pada Data Twitter,” Semin. Nas. Teknol. Inf. dan Komun., vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[3] A. C. Imanda, N. Hidayat, and M. T. Furqon, “Klasifikasi Kelompok Varietas Unggul Padi Menggunakan Modified KNearest Neighbor,” vol. 2, no. 8, pp. 2392–2399, 2018.

[4] P. Studi, T. Informatika, J. T. Informatika, F. Sains, D. A. N. Teknologi, and U. S. Dharma, “Deteksi Outlier Pada Data Campuran Numerik Dan Kategorikal Menggunakan Algoritma Enhanced Class Outlier Distance Based ( Ecodb ) Algoritma Enhanced Class Outlier Distance Based ( Ecodb ).”

[5] T. T. Hanifa, S. Al-faraby, F. Informatika, and U. Telkom,“Analisis Churn Prediction pada Data Pelanggan PT . Telekomunikasi dengan Logistic Regression dan Underbagging,” vol. 4, no. 2, pp. 3210–3225, 2017.

[6] R. E. Putri, Suparti, and R. Rahmawati, “Perbandingan Metode Klasifikasi Naïve Bayes Dan K-Nearest Neighbor Pada Analisis Data Status Kerja Di Kabupaten Demak Tahun 2012,” J. Gaussian, vol. 3, pp. 831–838, 2014.

[7] Arandika A, Mardji, Cholisson I. Implementasi Algoritma KNeaerest Neighbor (K-NN) Untuk Klasifikasi Data Wine. Jurnal Mahasiswa PTIIK UB.Volume 4, Number 12. 2014.

[8] Septianto, Ryan Hendy. 2015. Diagnosa Penyakit Tanaman Kopi Arabika dengan Metode Modified K-Nearest Neighbor (MK-NN). Skripsi. Universitas Brawijaya, Malang.

[9] Kumalasari, Noviana Ayu. 2014. Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Menentukan Tingkat Resiko Penyakit Lemak Darah (Profil Lipid). Skripsi. Universitas Brawijaya, Malang