International Journal of Multidisciplinary Research and Publications
ISSN (Online): 2581-6187

KNN on Iris Data with Python Programming

Ifan Prihandi
Computer Science, Mercubuana University, Kembangan, Jakarta Barat, 11650
Email address: ifan[DOT]prihandi[AT]mercubuana[DOT]ac[DOT]id

Abstract**—** *K-Nearest Neighbor (K-NN) is a classification technique that makes explicit predictions on test data based on a comparison of K nearest neighbors. In the process of data mining will extract valuable information by analyzing the existence of certain patterns or relationships of large data. Data mining is related to other fields of science, such as Database Systems, Data Warehousing, Statistics, Machine Learning, Information Retrieval, and High-Level Computing. In addition, data mining is supported by other sciences such as Neural Network, Pattern Recognition, Spatial Data Analysis, Image Database, Signal Processing. Several surveys of the modeling process and methodology state that, "Data mining is used as a guide, where data mining presents the essence of history, description and as a standard guide regarding the future of a data mining model process.*

Keywords**—** *K-Nearest Neighbor, classification, Data mining, model.*

I. INTRODUCTION

(Kamil, Kemas, Eng, W, & Kom, 2015) Dataset is the embodiment of data in memory that provides a consistent relational program model regardless of the origin of the data source. Used to set the query itself to be run by using DataAdapter in using parameters in report generation.

(Yulianton, 2014) The formal definition of data mining is the process of extracting valid, useful, unknown, and understandable information from data and using it to make business decisions. Data mining is also commonly referred to as "Data or knowledge discovery" or discovering hidden patterns in data. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is defined as the process of extracting or mining knowledge needed from large amounts of data.

In the process of data mining will extract valuable information by analyzing the existence of certain patterns or relationships of large data. Data mining is related to other fields of science, such as Database Systems, Data Warehousing, Statistics, Machine Learning, Information Retrieval, and High-Level Computing. In addition, data mining is supported by other sciences such as Neural Network, Pattern Recognition, Spatial Data Analysis, Image Database, Signal Processing. Several surveys of the modeling process and methodology state that, "Data mining is used as a guide, where data mining presents the essence of history, description and as a standard guide regarding the future of a data mining model process.

(Dwi Retnosari, 2014) Machine Learning is an area in artificial intelligence that is related to the development of techniques that can be programmed and learned from past data. Pattern recognition, data mining and machine learning are often used to refer to the same thing. This field deals with probability and statistics, sometimes optimization. Machine learning becomes an analytical tool in data mining.

1.1 Problem

Based on the background described above, then in broad outline, the formulation of the problem is:

1. Find the group in the data, with the number represented by the K value. Variable K itself is the number of clusters desired?
2. What is the Train to Splits Process?

1.2 Writing

The purpose of this paper is:

1. The centroid of cluster K, which can be used to label new data.
2. Label training data and testing data.

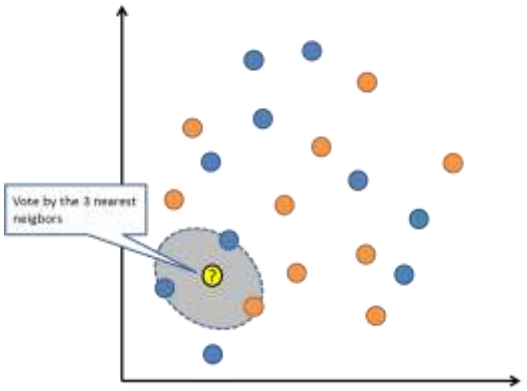
1.3 Scope / Limitation of Problems

Based on the identification of the problems above, the authors limit the problems discussed in this study, namely the application of the K-NN algorithm only to training data and testing data.

II. THEORETICAL BASIS

2.1 K-Nearest Neighbor

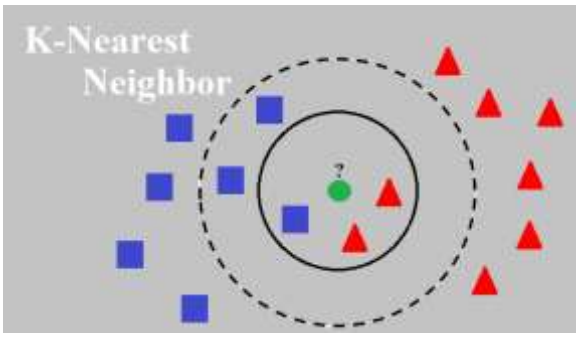
(Prasetyo, 2015) K-Nearest Neighbor (K-NN is a classification technique that makes explicit predictions on test data based on a comparison of K nearest neighbors),



KNN is one of the nonparametric machine learning algorithms (models). The discussion on parametric models and nonparametric models can be their article, but in short, the definition of nonparametric models is a model that does not assume anything about the distribution of instances in the dataset. Nonparametric models are usually more difficult to interpret, but one of the advantages is that the class decision

Ifan Prihandi, "KNN on Iris Data with Python Programming," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 2, Issue 7, pp. 6-8, 2019.

lines produced by these models can be very flexible and nonlinear.



The k-NN method algorithm is very simple, working based on the shortest distance from the test sample to the training sample to determine its k-NN. After collecting k-NN, the majority of k-NN is then taken as a prediction from the test sample. Data for the k-NN algorithm consists of several multi-variate Xi attributes that will be used to classify Y. Data from k-NN can be of any size scale, from ordinal to nominal.

2.2 Phyton

(Habibi, Setiawidayat, & Mukhsim, 2017) Python is a multipurpose interpretive programming language with a design philosophy that focuses on the level of code readability. Python is claimed to be a language that combines capabilities, capabilities with very clear code syntax and is complemented by large and comprehensive standard library functionality. At present Python code can be run on a variety of operating systems, some of which are Linux / Unix, Windows, Mac OSX, Java Virtual Machines, OS / 2, Amiga, Palm, and Symbian (for Nokia products).

The term high-level programming language refers to the level of closeness of a programming language to the electrical codes used by computers. A programming language is called a high-level programming language (high-level programming language) because the command or program code used is similar to human language (English).

III. ANALYSIS

(Anugerah & Cipta, 2014) an analysis is a set of activities and processes. One of the forms of analysis activity is to compile raw data into interpreted information. To answer and describe the problems that arise from the background and formulation of the problem, the research method can be seen in the figure below.

Based on the method framework drawings above, it can be explained as follows:

1. Determination of research objects
   In determining the object of research the authors conducted an analysis to see the results of observations. To complete the results of observation.
2. Analysis
   The formulation of the design consists of among other backgrounds.
3. Testing
   Develop a research model that has been set in the form of an application so that it can be developed into the Android versionEquations Style.

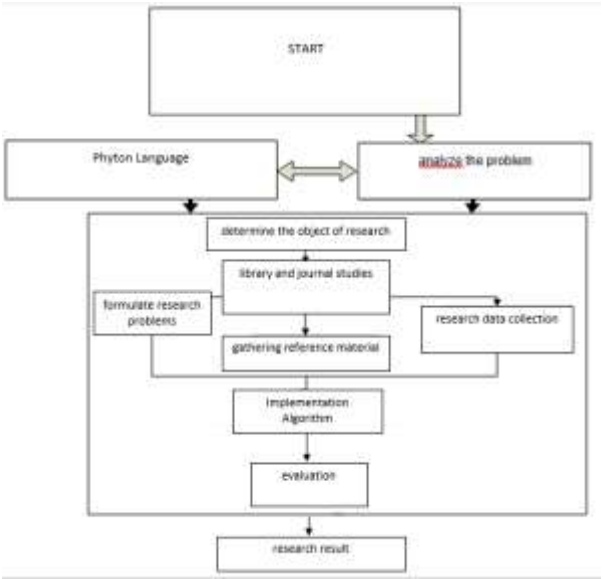


Fig. 3.1. Framework for solving problems

IV. TESTING DATA

System implementation and testing phases are done after the analysis phase is completed. At this stage, the implementation and testing of the system that has been made will be explained. The implementation phase has two scopes, namely the system requirements specification which includes hardware and software and the implementation of supporting application systems which include the coding process and the application of the interface design process (user interface) following the design. [5]

How the KNN algorithm works with the sample dataset from the dataset using Iris. This is available by default from Sklearn. This dataset contains 3 types of flower species along with petal and sepal sizes. This data is CSV, which is used as a table, each row shows different types of flower species, while the columns show data features, namely: sepal length, sepal width, petal length, and petal width respectively. There are 3 types of flowers, namely Setosa, Versicolor and Virginia. There are 50 data samples for each type of interest. So if in total there are 50x3 data samples = 150 data samples.

4.1 Train Test Data

Train Test Data by dividing all 150 data into 2 parts, training data and testing data. The ratio will be automatically 80:20 percent. There is an x value for training and testing [8], as well as 'y' there will be y for training and testing for prediction [9].

Ifan Prihandi, "KNN on Iris Data with Python Programming," *International Journal of Multidisciplinary Research and Publications (IJMRAP)*, Volume 2, Issue 7, pp. 6-8, 2019.

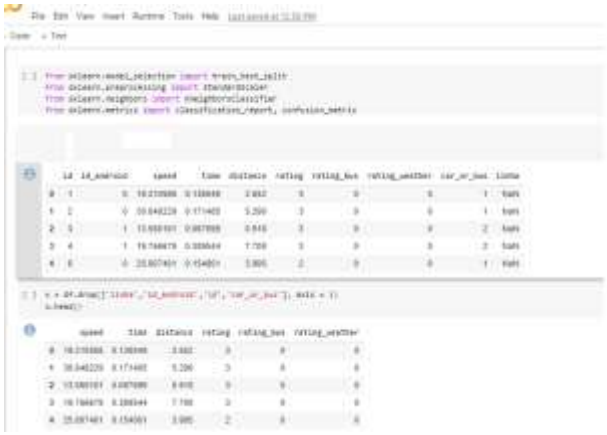


Fig. 4.1. Library and dataset

4.2 K-Fold Cross-Validation

(Lidya, Sitompul, & Efendi, 2015) K-fold cross-validation divides documents into k sections, one of the popular Cross-Validation methods is by folding K as much data and repeating (iterating) the experiment as much as K.

Fig. 4.2. Testing data

Fig. 4.3. Testing data

4.3 Tunning Parameter

Accurate classification of data classification. KNN has several parameters that determine the level of classification accuracy, namely the number of K in KNN. parameters with high accuracy, Tuning Parameters or Finding Hyperparameter. Here 1–40 K. All scores on K are different and stored in the k_score variable.

Fig. 4.4. Tunning parameter

4.4 Plotting

The final result is in plotting KNN score

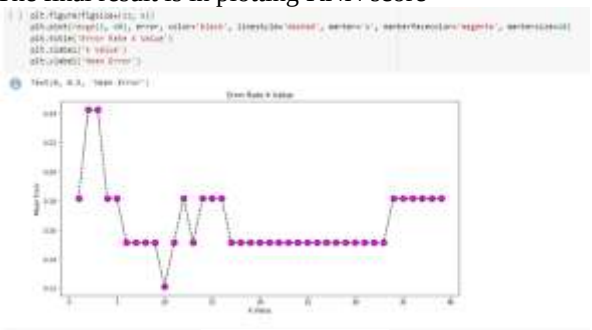


Fig. 4.5. Visualization

V. CLOSING

5.1 Conclusion

1. K-Nearest Neighbor (K-NN) is a classification method for a collection of data based on learning data that has been classified
2. This line chart illustrates that K which produces the highest score is 2 and 4. After that the score tends to go down, then we penalty that 4 is the highest K and produces the highest score.

REFERENCES

[1] Anugerah, P. T., & Cipta, B. (2014). Analisa Dan Perancangan Sistem Informasi Pemasaran. *Universitas Mercu Buana*, (18), 1–10. Retrieved from fasilkom.mercubuana.ac.id/wp-content/uploads/2017/10/Modul-Analisa-Perancangan-Sistem-Informasi.pdf%0D

[2] Dwi Retnosari. (2014). Sistem Aplikasi Data Mining Untuk Menampilkan. *Jurnal Integrasi Sistem Industri UMJ*, *1*(2), 13–20.

[3] Habibi, F. N., Setiawidayat, S., & Mukhsim, M. (2017). *Alat Monitoring Pemakaian Energi Listrik Berbasis Android Menggunakan Modul PZEM-004T*. *01*(01), 157–162.

[4] Kamil, I., Kemas, N., Eng, M., W, G. A. A., & Kom, S. (2015). *Analytical Processing pada Graph Dataset dengan Metode Graph OLAP Framework Analytical Processing on Graph Dataset with Graph OLAP framework Methodology*. (1).

[5] Lidya, S. K., Sitompul, O. S., & Efendi, S. (2015). Sentiment Analysis Pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM). *Seminar Nasional Teknologi Dan Komunikasi 2015*, *2015*(Sentika), 1–8. https://doi.org/10.1016/j.eswa.2013.08.047

[6] Prasetyo, E. (2015). Fuzzy K-Nearest Neighbor in Every Class Untuk Klasifikasi Data Fuzzy K-Nearest Neighbor in Every Class. *Seminar Nasional Teknik Informatika (SANTIKA 2012)*, (November), 1–5.

[7] Yulianton, H. (2014). Data Mining untuk Dunia Bisnis. *Teknologi Informasi DINAMIK*, *XIII*(1), 9–15.

[8] I. Ranggadara, G. Wang, and E. R. Kaburuan, "Applying Customer Loyalty Classification with RFM and Naïve Bayes for Better Decision Making," in *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2019, pp. 564–568.

[9] I. N. Budi, I. Ranggadara, I. Prihandi, N. R. Kurnianda, and Suhendra, "Prediction using C4.5 Method and RFM Method for Selling Furniture," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 535–541, 2019.