INTERNATIONAL EDUCATIONAL JOURNAL OF SCIENCE AND ENGINEERING (IEJSE)

Volume: 3, Issue: 1, Jan-Feb 2020 | E-ISSN NO: 2581-6195

IMPLEMENTATION K-NEAREST NEIGHBOR AND ~~LINEAR DISCRIMINANT~~ ANALYSIS METHOD FOR CLASSIFICATION TYPE OF WORKS IN JAKARTA

Yunita Sartika Sari

Faculty of Computer Science, Mercu Buana University, Indonesia.

Submission date: 13 Jan 2020, Acceptance date: 6 Feb 2020, Publication date: 15 Feb 2020

ABSTRACT: *In this study, explaining how to implement the classification of types of work in DKI Jakarta using the K-Nearest Neighbor and Linear Discriminant Analysis methods. Types of work are grouped by occupation. After testing the classification of work types, the results of testing accuracy are 50% for the Linear Discrimant Analysis Method and 32% for K-Nearest Neighbor.*

KEYWORDS: *k-nearest neighbor, linear discriminant analysis, classification*

1. INTRODUCTION:

In the digital era, the data is one of the components that are important in decision making. Data must be processed first so that it can be understood by the recipient data. The results of data processing is called information. Later this information to be used as a benchmark by a person, institution or company in decision making.

The increasing number of population and the existing technology, the amount of data is also growing and the information can be obtained from such data is becoming more diverse. Indonesia is one country with the highest number of inhabitants in the world. Most of the Indonesian population lives on the island of Java and Jakarta as the capital of Indonesia has a high population density.

Jakarta is one of the provinces in Indonesia which is the business center so that residents have a very diverse professions. The diversity of professions people living in Jakarta as well as administration areas were divided into six regions, create jobs data people living in Jakarta must be treated to information about the distribution of occupations by region of residence can be obtained.

Based on these descriptions, will be the classification of areas in Jakarta based on the type of work with comparative several methods of classification.

2. PLATFORM THEORY:

2.1 Definition of K-Nearest Neighbor Method:

K-nearest neighbors or KNN is an algorithm that functions to classify a data based on learning data (train data sets), which are taken from k nearest neighbors (nearest neighbors). With k is the number of nearest neighbors.

Following are some formulas used in the KNN algorithm:

a. *Euclidean Distance:* Euclidean distance is a formula for finding the distance between 2 points in two-dimensional space.

b. *Hamming Distance:* Hamming distance is a way to find the distance between two points calculated by the length of the binary vector formed by these two points in the binary code block.

c. *Manhattan Distance:* Manhattan Distance or Taxicab Geometry is a formula for finding the distance d between 2 vectors p, q in the n dimensional space.

d. *Minkowski Distance:* Minkowski distance is a formula for measuring between 2 points in a normal vector space which is a hybridization that generalizes euclidean distance and mahattan distance.

2.2. Definition of Linear Discriminant Analysis Method:

Linear Discriminant Analysis (LDA) is one of the methods used to group data into several classes. Determination of grouping is based on the boundary line (straight line) obtained from the linear equation.

3. RESULTS AND DISCUSSION:

a. *Data:* Data were obtained from Jakarta Open Data, Jakarta Open Data is a website that provides the dataset associated with the data contents in Jakarta. The data obtained is the data in 2014. The data has a format (.csv) with a total of 267 rows and 95 columns, and the size of 86 kb. Here is a view of the data to be used for processing.

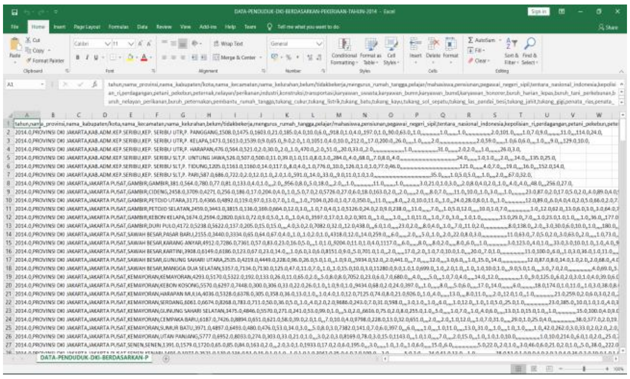


Figure 1: Dataset

b. *Device:* Processing is done using the programming language Python 3.7.3 (based on anaconda) and the text editor of Visual Studio Code with python interpreter, as already mentioned.

Figure 2: Device

c. *Modeling:* Data processing method used is the classification. Classification method has several algorithms for processing. Data processing was performed using K-Nearest Neighbor and Linear Discriminant Analysist. Many modeling performed to determine the ratio between the model and get the best data accuracy of some types of modeling. The accuracy of the data that will either maximize data processing is done.

d. *Data visualization:*

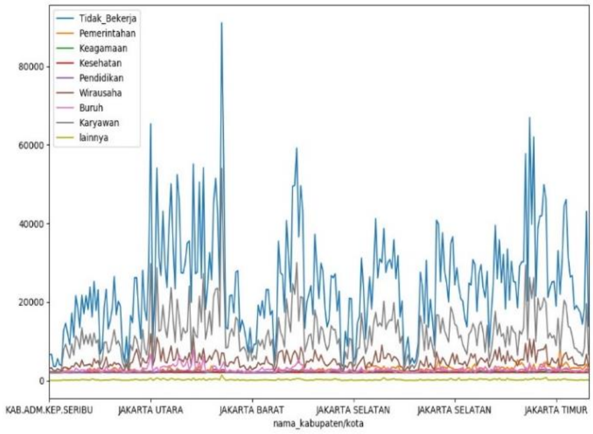


Figure 3: Data Visualization

3.1. Implementation Method:

The method used are K-Nearest Neighbor and Linear Discriminant Analysis.

a. *K-Nearest Neighbor:* Supervised learning algorithm in which the results of the new instance are classified based on the majority of the k-neighbor closest categories. The model used uses the default k-nearest neighboors parameter in python. Next, the fitting process is carried out on the model using the x train and y train parameters. The results of the modeling process using k-nearest neighbors produce an accuracy value of 59% for the training set and 32% for the testing set with the value of the confusion matrix as follows:

```
[[ 9  4  5  2  0  0]
 [ 6  2  5  3  1  0]
 [ 4  5  8  6  0  0]
 [ 7  2  4 15  0  0]
 [ 6  0  5  4  0  0]
 [ 0  0  4  0  0  0]]
```

Confusion matrix can be used to produce the following classification evaluations:

```
                    precision    recall  f1-score   support

      JAKARTA BARAT      0.28      0.45      0.35        20
      JAKARTA PUSAT      0.15      0.12      0.13        17
    JAKARTA SELATAN      0.26      0.35      0.30        23
      JAKARTA TIMUR      0.50      0.54      0.52        28
      JAKARTA UTARA      0.00      0.00      0.00        15
 KAB.ADM.KEP.SERIBU      0.00      0.00      0.00         4

           accuracy                          0.32       107
          macro avg      0.20      0.24      0.22       107
       weighted avg      0.26      0.32      0.28       107
```

Visualization of the comparison of actual and predictive labels is shown in the following pictures:

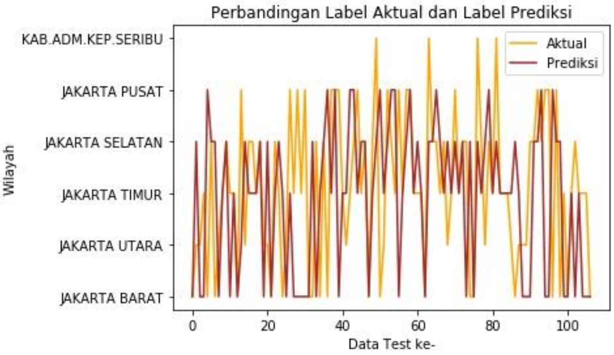
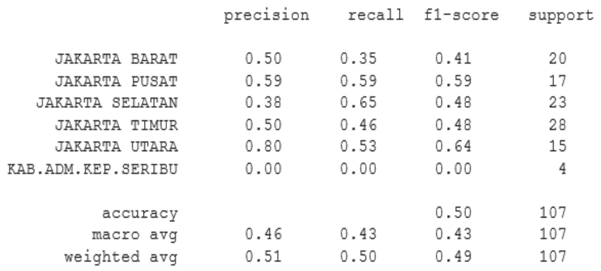


Figure 5: Visualization of Comparation K-Nearest Neighbor Methods

b. *Linear Discriminant Analysis:* Linear Discriminant Analysis (LDA) is general pattern recognition to find linear combinations of features that characterize or separate two or more classes of objects or events. The resulting combination can be used as a linear classification. The model used uses the default linear discriminat analysis parameter in python. Next, the fitting process is carried out on the model using the x train and y train parameters. The results of the modeling process using k-nearest neighbors produce an accuracy value of 59% for the training set and 32% for the testing set with the value of the confusion matrix as follows:

```
[[ 7  3  8  2  0  0]
 [ 0 10  4  2  1  0]
 [ 1  0 15  7  0  0]
 [ 2  3  9 13  1  0]
 [ 4  1  0  2  8  0]
 [ 0  0  4  0  0  0]]
```

Confusion matrix can be used to generate the evaluation of the following classifications:



```
                    precision    recall  f1-score   support

      JAKARTA BARAT      0.50      0.35      0.41        20
      JAKARTA PUSAT      0.59      0.59      0.59        17
    JAKARTA SELATAN      0.38      0.65      0.48        23
      JAKARTA TIMUR      0.50      0.46      0.48        28
      JAKARTA UTARA      0.80      0.53      0.64        15
KAB.ADM.KEP.SERIBU      0.00      0.00      0.00         4

           accuracy                          0.50       107
          macro avg      0.46      0.43      0.43       107
       weighted avg      0.51      0.50      0.49       107
```

Visualization label comparison of actual and predicted label shown in the following:

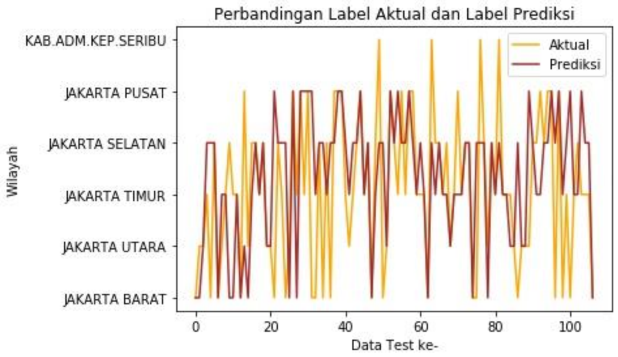


Figure 6: Visualization of Comparation Linear Discriminant Analysis Method

Comparison of K-Nearest Neighbor and Linear Discriminant Analysis methods that have been done can be seen in the following table:

Table 2: Comparison of Accuracy of Method

No.	Method	accuracy	
		training	testing
1	K-Nearest Neighbor	59%	32%
2	Linear Discriminant Analysis	57%	50%

4. CONCLUSION:

Based on the results of several methods can be concluded that:

1. The method is good enough Linear Discriminant Analysis testing which resulted in an accuracy of 50% means that a predictive model has been pretty good.

2. Distribution of the work that most of the field work is not work (yet / Not Working, Taking Care of Household, Student / Students and Pensioners) and is located in West Jakarta.

REFERENCES:

1. Nia Rahma Kurnianda & Yunita Sartika Sari. Analysis and Design of Information System for Self-Journal on Food Based Dietary Assessment Record for Diabetes Patients. International Research Journal of Computer Science (IRJCS). Volume 06 Issue 5. 2018
2. Ranggadara, Indra & Suhendra "Naive Bayes Algorithm with Chi Square and NGram Feature for Reviewing Laptop Product on Amazon Site.".IRJCS, Vol 7, issue 2.2018
3. William H. Kruskal and Judith M. Tanur, ed. (1982), "Linear Hypotheses," International Encyclopedia of Statistics. Free Press, v. 1,
4. Lindley, D.V. (1987). "Regression and correlation analysis," New Palgrave: A Dictionary of Economics, v. 4, pp. 120–23.
5. Birkes, David and Yadolah Dodge, Alternative Methods of Regression. ISBN 0-471-56881-3
6. Corder, G.W. and Foreman, D.I. (2009).Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach Wiley, ISBN 978-0-470-45461-9
7. Draper, N.R. and Smith, H. (1998).Applied Regression Analysis Wiley Series in Probability and Statistics
8. Fox, J. (2017). Applied Regression Analysis, Linear Models and Related Methods. Sage
9. Hardle, W., Applied Nonparametric Regression (1990), ISBN 0-521-42950-1
10. Meade, N. and T. Islam (1995) "Prediction Intervals for Growth Curve Forecasts," Journal of Forecasting, 14, pp. 413–430.

11. Witten, H.Ian, Eibe, F. and Mark, H., 2011. Data Mining Practical Mechine Learning Tools and Techniques Third Edition (3rd Ed.). Elsevier Inc