Human Development Index Classification and Clustering

Kholed Langsari & Gervais Hatungimana

Department of Informatics Faculty of Information Technology, Institut Teknologi Sepuluh Nopember (ITS)

Email: langsaree@gmial.com

**Abstract.** Human Development Index (HDI) classification or clustering refers to application of machine learning techniques to study trends between variables of collected statistical data. In this paper we explore different classification and clustering methods to the dataset originally from survey to categorize geographical area (cities) based on four level of development. We use data from Badan Pusat statistiks of Indonesia 2004-2012; we use sample of 342 out of 22,366 full data. If you are looking about how machine learning techniques is used in social sciences then you are in the right track; we hope through this paper you will get to know the way out.

**Keywords:** Human Development Index; HDI; Indek Pembangunan Manusia; IPM; KDD; Classification; Clustering; Data mining.

# Introduction

The Human Development Index (HDI) is a composite statistic of life expectancy, education, and per capita income indicators, which are used to rank countries into four tiers of human development [1]. *Indek Pembangunan Manusia* (IPM) is HDI acronym in Indonesia language.

Mostly organization like UNDP program conduct the survey in every country about life expectancy at birth, Mean year of schooling, expected year of schooling and GNI per capita in order to estimate HDI (Human Development Index) and IHDI (Inequality-adjusted HDI),

This paper studies the human development index (HDI), which has become one of the most widely used measures to communicate a country’s development status. Compared to the gross domestic product (GDP), the HDI is a broader measure of development, as it captures not only the level of income, but also incorporates measures of health and education. The United Nations Development Programme, which releases the HDI statistics, classifies each country into one of three categories: ‘low human development’ for HDI scores between 0.0 and 0.5, ‘medium human development’ for scores between 0.5 and 0.8 and ‘high human development’ for scores between 0.8 and 1.0 [2]. And this paper is basically on statistics that Badan Pusat Statistik has released for the country of Indonesia the year of 2004-2012.

Although UNDP standards has three level or categories of HDI, Badan Pusat Statistik of Indonesia decided to categorize HDI into four levels.

Four tiers of human development are very high, high, medium, and low of human development [1].

Although these development categories were not originally designed to determine international relations, development aid, nor should they imply any other legal consequences, today these three mutually exclusive categories are used in politics, academia and the corporate world. In business relations, the categories have been used for international pricing purposes [2].

Data mining techniques [3] have always been a powerful tool for any type of categorization. Data mining techniques include supervised and unsupervised learning. Supervised learning requires training data with the label of the classes while unsupervised learning does not require the training patterns to be labeled.

# Research Methodology

KDD is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

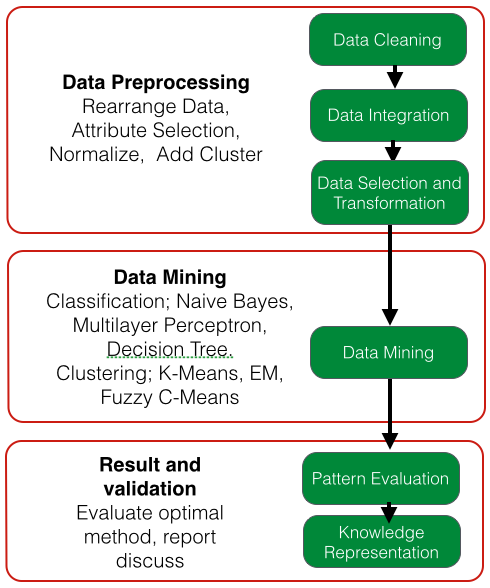
In our work we use KDD (knowledge discovery in databases) as well know methodology process. KDD are consist of big three steps process include; data preprocessing, data mining, and result and validation as illustrated in Figure 1.

## Data Preprocessing

Because data in the real world is dirty, incomplete and noisy. Incomplete in lacking attributes values and lacking attributes of interest or containing only aggregate value noisy in terms of containing errors or outliers and inconsistent containing discrepancies in names or codes [5].

Data is not clean, Duplicity of data and the no quality data and the most important is no quality result so data preprocessing is important. Quality decisions must be based on the quality data [5].

Data cleaning and preprocessing: includes basic operations, such as noise removal and handling of missing data. Data from real-world sources are often erroneous, incomplete, and inconsistent, perhaps due to operation error or system implementation flaws. Such low quality data needs to be cleaned prior to data mining [7].



1. KDD (knowledge discovery in databases) process.

During the data cleaning and preprocessing stage noise is removed from the data [4].

Some data can be available but not related ,or its impact is does not give good results depending on the purpose of the project. So at this stage we look the contribution each variable brings to the objectives then remove unnecessary and keep those one which are important. If value of attributes have big standard deviation which means the dependency is normal we use WEKA to clean the data by normalization module.

Data integration includes integrating multiple, heterogeneous data sources. In data integration stage we do integrate same kind of raw data from difference resources.

Most of raw data are available in excel format. In order to be read in machine learning application such as WEKA or Matlab application, we need to convert it to appropriate format as to .ARFF, .CSV or .DAT format. For example WEKA application does not read .DAT while Matlab application does not read .ARFF format.

Also raw data arrangement might not be suitable to implement in machine learning process. Process like transpose matrix might take place before do training the dataset.

## Data Mining

Data mining is an interdisciplinary field with a general goal of predicting outcomes and uncovering relationships in data. It uses automated tools employing sophisticated algorithms to discover hidden patterns, associations, anomalies and/or structure from large amounts of data stored in data warehouses or other information repositories [3].

Data mining involves fitting models to determining patterns from observed data. A data mining algorithm constitutes some combination of the model, search algorithm and preference criteria [7].

### Classification

Data mining refers to extracting or mining knowledge from large amounts of data. Classification is one functionality of data mining who’s task is generalizing known structure to apply to new data. Classification is to identify objects into predefined groups or classes based on some features or characteristics. Classification is a supervised machine learning

technique of finding patterns in the data, i.e., these algorithms work with a class attributes. In this paper we implement and compare three of classification methods, Naïve Bayes, Decision Tree and Multilayer Perceptron.

### Clustering

Clustering is the task of grouping a set of data in such a way that data in the same group are more similar to each other than to those in other groups. Different clustering algorithms attempt to group the data in a different way. In this article we applied clustering techniques include, K-Means, Expectation Maximization and Fuzzy C-Means clustering.

K-Means clustering [6] finds the cluster centers and assigns the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. It is an optimization problem.

Expectation Maximization (EM) [6] algorithm is used to classify each point into the most likely Gaussian and estimate the parameters of each distribution.

The fuzzy c-means (FCM) clustering algorithm is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. It provides a method that shows how to group data points that populate some multidimensional space into a specific number of different clusters. FCM has a wide domain of applications such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis, and target recognition [8].

# Experimentation

We carry experimentation using WEKA and MatLab. WEKA is one of the user friendly and an open source software runs on any platform. WEKA tool was developed by the University of Waikato in New Zealand. In beginning WEKA tool was written in C language, Later the application has been rewritten in java language. WEKA provides the implementation of algorithm which can be applied to data set. It includes many algorithms for clustering, association rule mining, attribute selection and regression. WEKA has two file format ARFF (attribute relation file format) and CSV (comma separated values). WEKA helps us to learn more about the data from analyzing the output result. It contains a collection of visualization tools and algorithms available for data; analysis, preprocess, classification, modeling and evaluation [3].

MATLAB is a high-level programming language used in scientific numerical computation. It has many rich features [9] including user-friendly interface for technical computing, flexible development environment for managing programming code, interactive tools for iterative exploration, design and problem solving and functions for integrating MATLAB with external applications and languages. It exploits high-level programming features including easy-to- write syntax, rich library and simple programming paradigm which makes it popular even for the people who do not have much programming expertise. In our case we use MATLAB the Fuzzy logic tool to simulate C-means method with our dataset.

## Data set

The data set is HDI of Indonesia country from year 2004 to 2012. The sample of data set we select for Jawa Timur province. In this data set its consist of 5 attributes and 1 class with 342 instances after data preprocessing process. Bandan Pusat Statistik of Indonesia reported on Indonesia Development Report where they distinguished four levels of Human Development Index (HDI) according to data in 2009.

In this project we also look for distinguishing four levels of HDI based on data from 2004- 2012. The levels of HDI are assigned as follow; below 54– 64.99 is class LOW, 63.00 – 67.99 is class MEDIUM, 68.00 – 71.99 is class HIGH, 72.00 – above is class VERY HIGH.

# Results

After comparing outputs from various methods, it seems that in general classification methods performs better than those of clustering on our dataset.

Classification with Multilayer Peceptron is able to classify data into appropriate classes at 99.7076% with relative absolute error of 2.2775% and root relative square error is 10.9161%. Then Classification with NaiveBayes is able to classify data into appropriate classes at 95.9864% with relative absolute error of 8.8907% and root relative square error is 30.373%. Finally classification with Decision Tree is able to classify data into appropriate classes at satisfactory level of 99.4152% with absolute error of 1.1864% and root relative square error is 10.8979%. We can conclude that Multilayer Perceptron is the best classification method to be used for this kind of dataset as illustrated in Table 2.

Three methods of clustering are evaluated, Simple K-Means, Simple EM (Expectation Maximization) and Fuzzy C-Mean; all methods show improvement after the dataset has been normalized; Simple K-Mean reduced square error (from 32.904 to 17.771), Simple EM increased the likelihood (from -11.70632 to 3.93363) and Fuzzy C-Mean fcn has improved also to 8.6518. However looking at the visualization results both K-Mean and EM can be used alternatively but Fuzzy C-Mean has a lot of misclassified data as illustrated in Table 1.

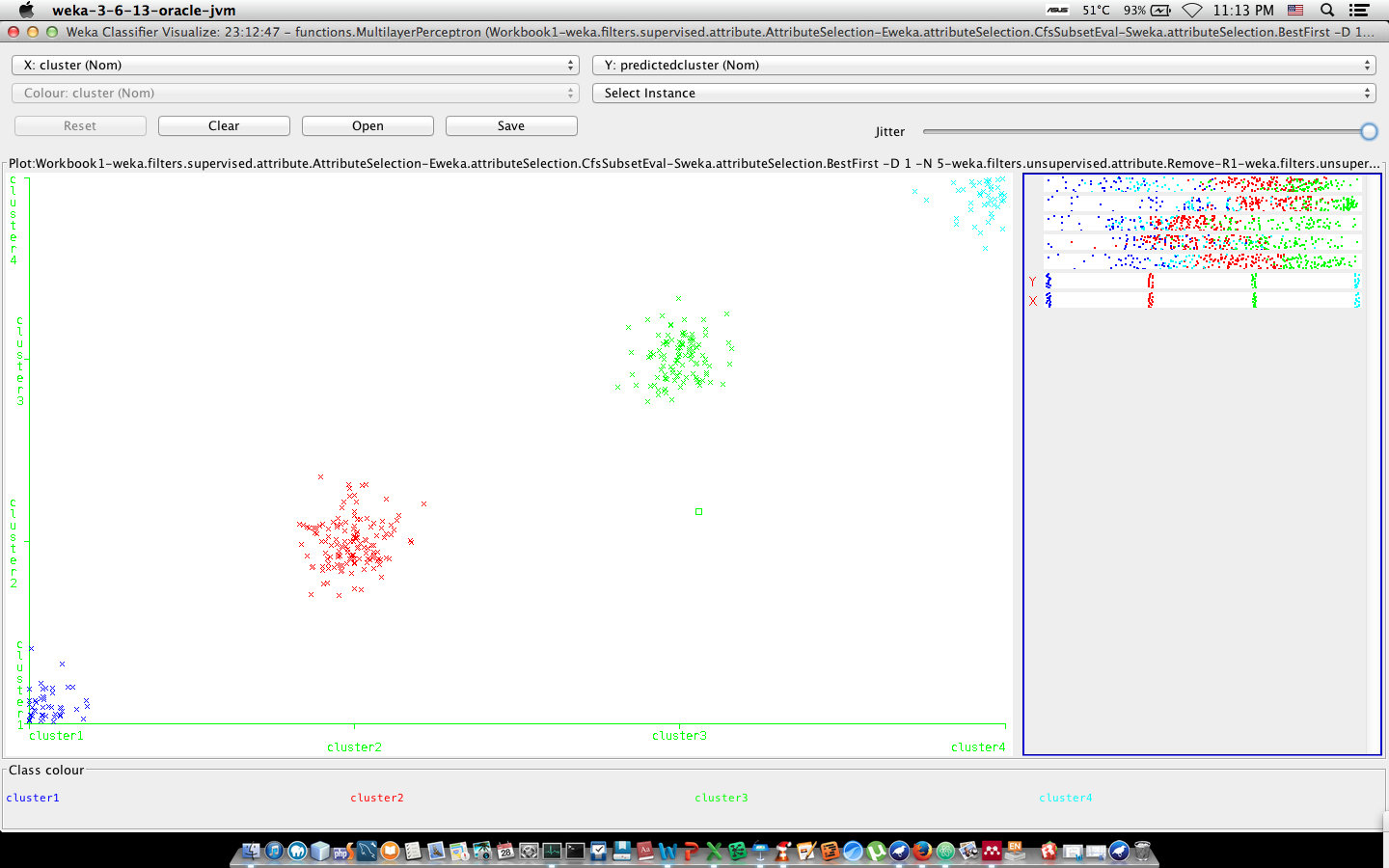
1. Clustering results.

|  |  |  |
| --- | --- | --- |
| **No** | **Clusterer** | **accuracy** |
| 1 | Simple K-Mean | 17.771 square error |
| 2 | Simple EM | 3.93363 likelihood |
| 3 | Fuzzy | 8.6518 fcn |
|  |  |  |

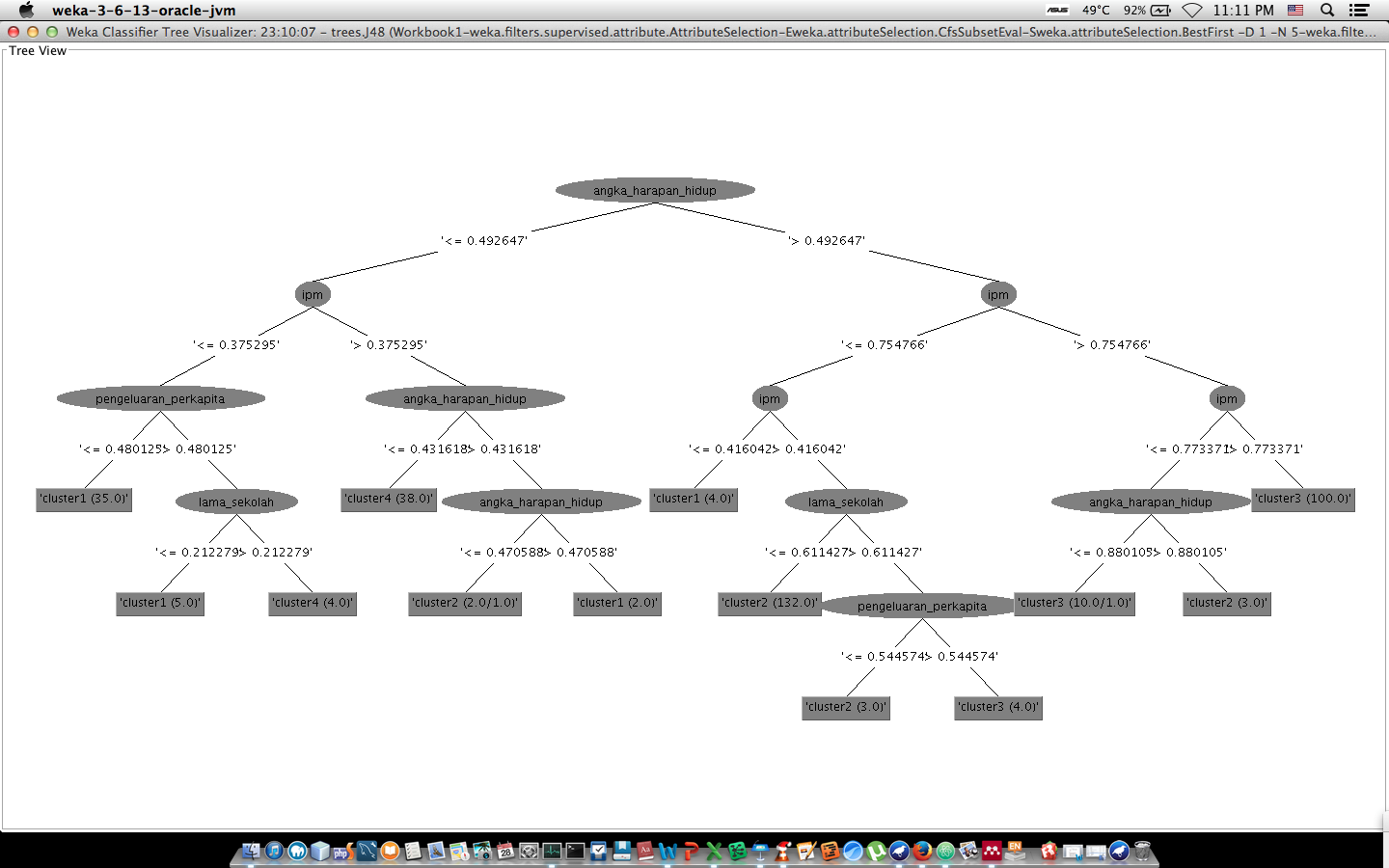
1. Classification results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Classifier** | **Relative absolute error** | **Root relative squared error** | **Correctly Classified Instances** |
| 1 | Mutilayer Peceptron | 2.2775% | 10.9161% | 99.7076% |
| 2 | Decision Tree (J48) | 1.1864% | 10.8979% | 99.4152% |
| 3 | NaïveBayes | 8.8907% | 30.373% | 95.9864% |

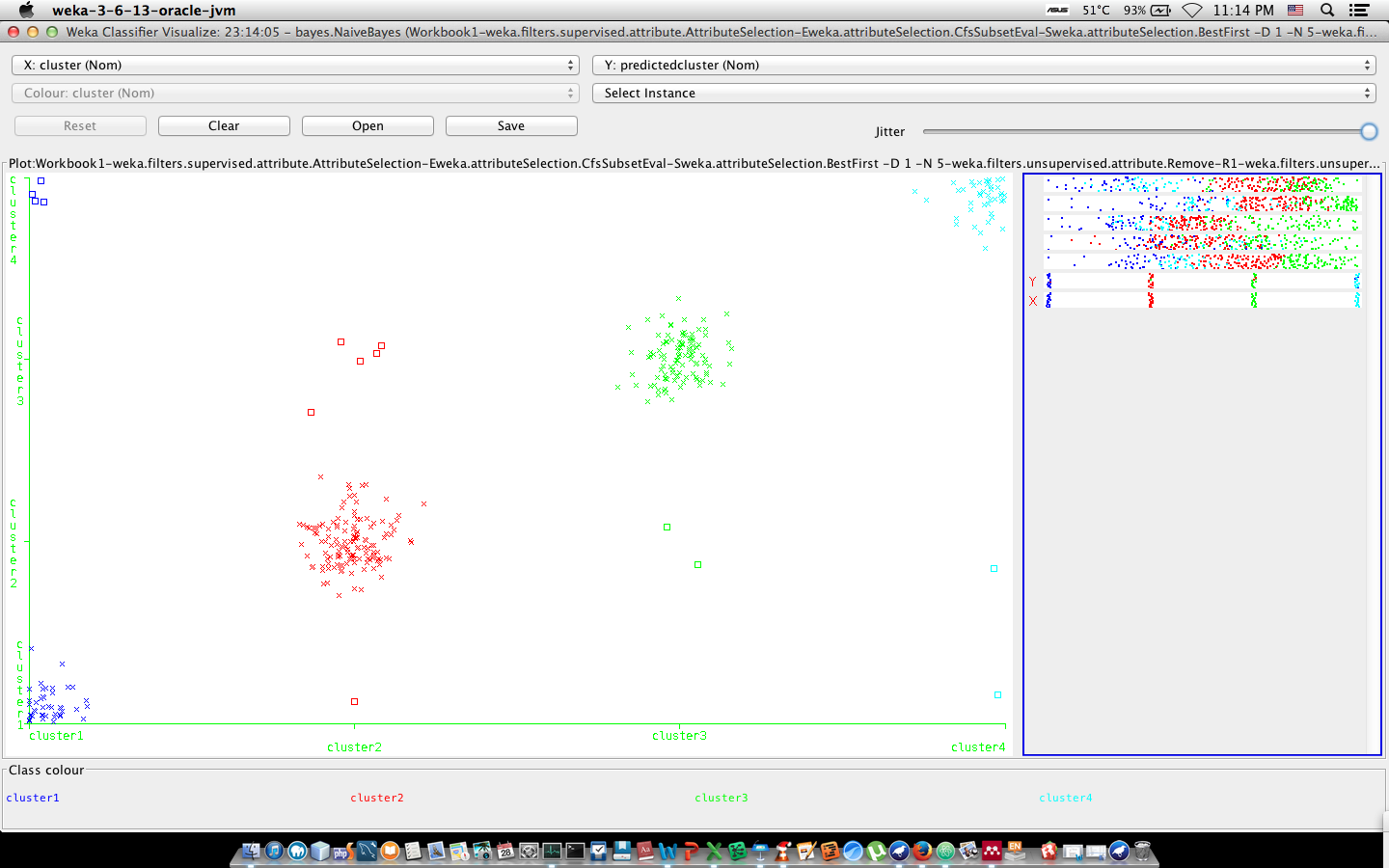
Visualized results of classification can be found in Figure 1, Figure 2 and Figure 3 and visualized results of clustering can be found in Figure 4, Figure 5 and Figure 6.



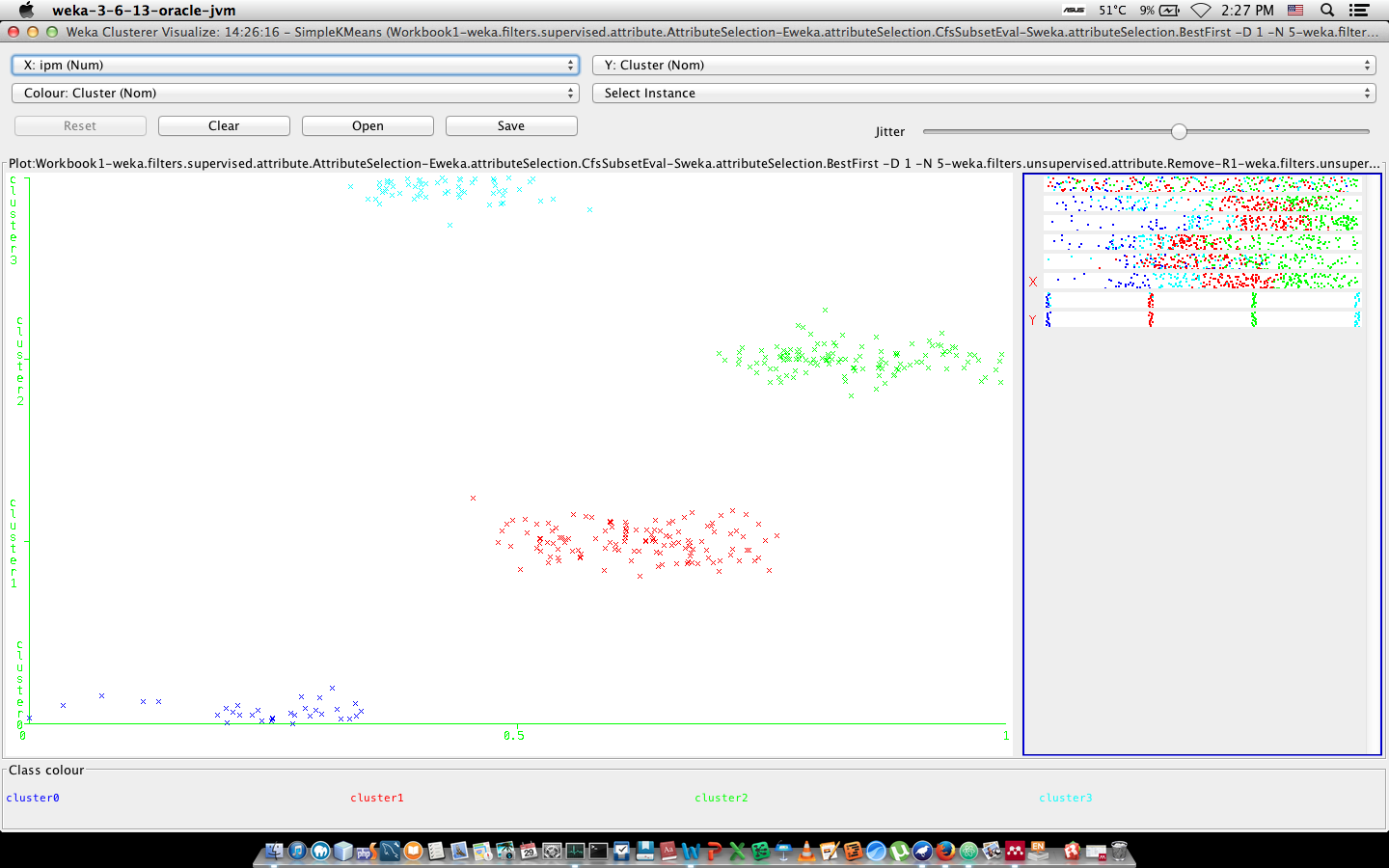
1. Visualize classifier of Multilayer Perceptron



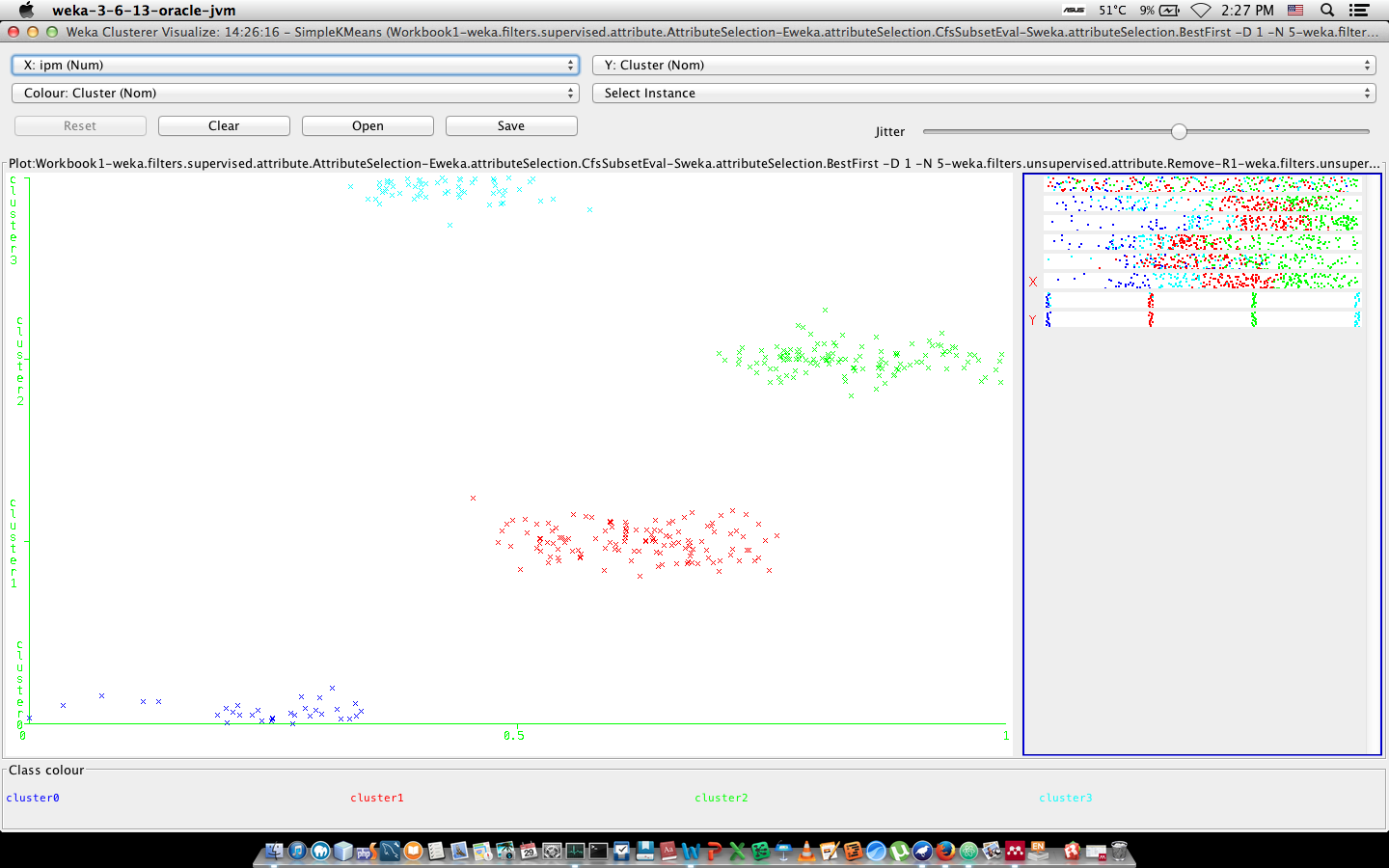
1. Visualize tree of Decision Tree



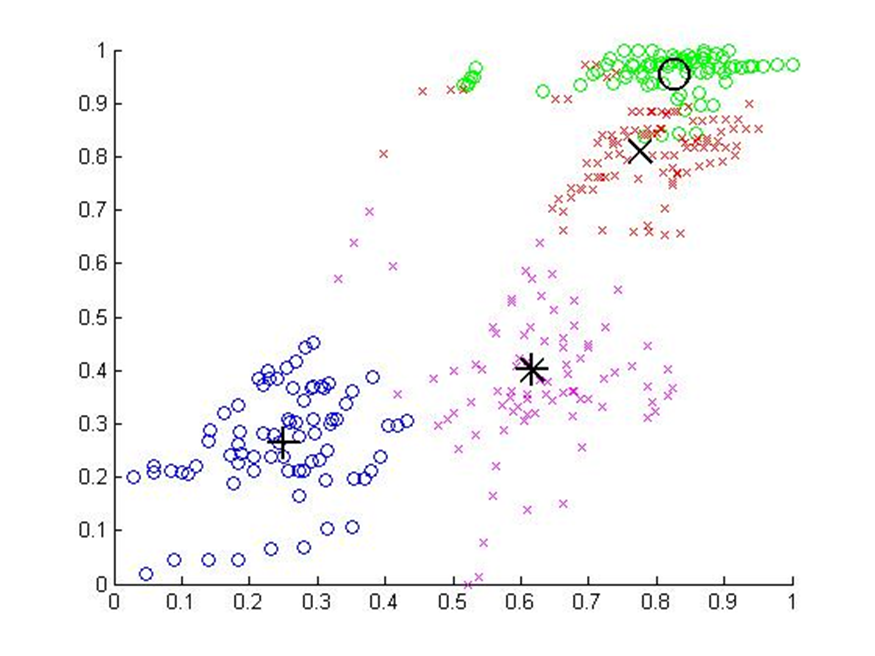
1. Visualize classifier of Naïve Bayes method



1. Virtualize cluster assignment of Simple K-Means



1. irtualize cluster assignment of Simple EM (Expectation Maximization)
2. Virtualize of fuzzy C-Means



# Conclusion

In this paper we have presented raw data statistically arranged ;our purpose was to take make use of the machine learning techniques to group different areas of Indonesia into four level of Human Development Indicators (HDI) basically on the report of Badan Pusat Statistiks of Indonesia in 2009. The course of work involved rearrange raw data according to the requirements of machine learning methods (Data Processing) and apply different techniques.

Looking the task as supervised machine learning techniques we first labeled our instances by adding class attribute then experiment using Naïve Bayes ,Multilayer Perceptron and Decision Tree classification. Looking the task as unsupervised machine learning techniques class label was removed then experiment using Simple K-Means ,EM, and Fuzzy C-Means clustering methods.

Considering the results from both statistics and visual ones we conclude that the suitable machine learning techniques for kind of dataset like this should be the classification with Multilayer Perceptron. The concentration in this paper was to group area based on HDI ;HDI is a results of many other attributes as per UNDP formula; In case someone else is motivated to research with different purpose some other methods may yield better results.

**References**

1. <https://en.wikipedia.org/wiki/Human_Development_Index>, 2015
2. Wolff, H. Chong, and M. Auffhammer, *“Classification, detection and consequences of data error: evidence from the human development index,”* The economic Journal, vol. 121, pp. 843–870, 2011.
3. I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)*, Morgan Kaufmann, 2011.
4. S. B. Kotsiantis and D. Kanellopoulos, *“Data preprocessing for supervised leaning,”* International Journal of …, vol. 1, no. 2, pp. 1–7, 2006.
5. S. Singhal and M. Jena, “*A Study on WEKA Tool for Data Preprocessing , Classification and Clustering,”* International Journal of Innovative Technology and Exploring Engineering, vol. 2, no. 6, pp. 250–253, 2013.
6. Nancy.P and R. G. Ramani, *“Discovery of Patterns and evaluation of Clustering Algorithms in SocialNetwork Data (Face book 100 Universities) through Data Mining Techniques and Methods,”* International Journal of Data Mining & Knowledge Management Process, vol. 2, no. 5, pp. 71–82, 2012.
7. R. Kalavathy, R. M. Suresh, and R. Akhila, *“KDD and data mining,”* IET-UK International Conference on Information and Communication Technology in Electrical Sciences (ICTES 2007), no. Ictes, pp. 1105–1110, 2007.
8. Hopner, K, R., Runkler, *Fuzzy Cluster Analysis,* John Wily & sons. 1999
9. http://www.mathworks.com/products/matlab/description1.html, retrieved on 10 March, 2011.