

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Collecting Data

In this project data was accessed using the API through <https://api.darksky.net/forecast> (Asia /data Jakarta) with access intervals every 2 days. The collected data are weathers parameter data which includes Time, Summary, Precip Intensity, Precip Probability, Temperature, Apparent Temperature, Dew Point, Humidity, Pressure, Wind Speed, Wind Gust, Wind Bearing, Cloud Cover, UV Index, Visibility, and Ozone. There are 16 data column. Each column contains a value that changes at anytime and has its own interval. Then 20 rows data are used as trials.

3.2 Processing Data

The first step is to determine the number of K values according to user input and the date intervals. Then from the input results, a new centroid was formed in the form of X and Y randomly as many as the number of previous K values. The coordinate value of X is Time and the value of Y is adjusted to the entire attributes used (for example, one attribute used is temperature as Y value of 16 data columns).

The second step is to read the entire CSV data, then calculate the distance of each data with the each predetermined centroid using Euclidean Distance formula, as follows:

$$D(i, j) = \sqrt{(x_{1i} - y_{1j})^2 + (x_{2i} - y_{2j})^2 + \dots + (x_{ni} - y_{nj})^2}$$

Dimana :

$D(i, j)$ = Distance between data at points x and y

x = Initial data point

y = Final data point

n = Number of attributes / data characteristics

If the distance calculation process and data grouping have not been completed, the data will be read again until the overall data is successfully calculated. The third step, calculates the average of cluster members data, this averages will become the center point for the next iteration.

For the last step, If there is a change in cluster members, then the process of forming a new centroid value will be repeated again. It will be done until the average values doesn't change anymore. Then the results of the conditions were determined from the data formed from each cluster.

3.3 Final Results Data

Each attribute has its own cluster results. Each cluster has its own weather conditions. Subsequently, determine the frequency of the most weather condition occurrence from each cluster. Then proceed with determining the most conditions of the entire cluster of each attribute. Furthermore, for the last step, determine the frequency of the most occurring weather condition from the overall attributes. Last, the results of that merger will be the final results of determining the weather conditions.

3.4 Report Writing

Write conclusion about the classification using K-Means algorithm, including how the process determines the final results of weather conditions and the time intervals effect from available data. After the conclusion, the author also add suggestion for the next study.