|  |  |
| --- | --- |
| **Practicum Case** |  |
| COMP6579001  Big Data Processing |
| **Computer Science** | **E201-COMP6579001-DD01-10** |
| ***Valid on*** *Even Semester Year 2019/2020* | **Revision 00** |

## Learning Outcomes

* Understand Big Data Analytics and Visualizations

## Topic

* Session 10 – Clustering and Model Evaluation

## Subtopics

* Clustering using Spark
* Model Evaluation using Spark

## Soal

*Case*

**Marine Pollution**

You are working in a science laboratory and are currently doing some research about **marine pollution**. **Marine pollution** is a condition where marine waters are getting polluted from industrial, agricultural, residential waste, and some invasive organisms. Based on the data you have collected, you decided to make a **clustering** model that will divide your data into **2 cluster**, **polluted water cluster** and **non-polluted water cluster**. You will be given “**Training.csv”** and “**Testing.csv”** and here is the description of the columns:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Trash Pollution | The level of manufactured products such as plastic  that end up in the ocean. |
| Oil Concentration | The amount of oil residue in waters. Oil cannot  dissolve in water and forms a thick sludge in the water.  The amount of harm caused depends on how an  organism is exposed and to how much oil. |
| Bacteria Level | The level of bacteria in Colony Forming Units (CFU)  per 100 ml of water.  Bacteria is a microorganism that live in diverse environment and can in soil, ocean, and human guts.  The presence of bacteria can be beneficial while some  that is pathogenic can be harmful. |
| Algae Concentration | The concentration of algae living in the waters.  Algae are a diverse group of aquatic organisms that  have the ability to conduct photosynthesis. Most algae  are harmless and an important part of the natural ecosystem, but under certain conditions, it can harm humans, fish, and other animals. |
| Humidity | The concentration of water vapour present.. |
| Wind Direction | The surrounding wind direction in degree. |
| Air Temperature | The surrounding air temperature in Celcius. An  increase in the air temperature will cause water temperatures to increase as well. |

**Figure 1. Training.csv**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Trash Pollution | The level of manufactured products such as plastic  that end up in the ocean. |
| Oil Concentration | The amount of oil residue in waters. Oil cannot  dissolve in water and forms a thick sludge in the water.  The amount of harm caused depends on how an  organism is exposed and to how much oil. |
| Bacteria Level | The level of bacteria in Colony Forming Units (CFU)  per 100 ml of water.  Bacteria is a microorganism that live in diverse environment and can in soil, ocean, and human guts.  The presence of bacteria can be beneficial while some  that is pathogenic can be harmful. |
| Algae Concentration | The concentration of algae living in the waters.  Algae are a diverse group of aquatic organisms that  have the ability to conduct photosynthesis. Most algae  are harmless and an important part of the natural ecosystem, but under certain conditions, it can harm humans, fish, and other animals. |
| Humidity | The concentration of water vapour present.. |
| Wind Direction | The surrounding wind direction in degree. |
| Air Temperature | The surrounding air temperature in Celcius. An  increase in the air temperature will cause water temperatures to increase as well. |
| Polluted | Whether the water is polluted or not (Yes, No). |

**Figure 2. Testing.csv**

Below are the steps you are required to do to generate the model:

1. **Load Data**

Given the file “**Training.csv**” and “**Testing.csv**”, you are asked to load the data using **SparkSession**.

1. **Select Features**

After you load the data, you need to **select important features** that will be used for training. Pick **three** **important features**.

1. **Data Preprocessing**

In this step, please remove any **missing values** in the data.

1. **Transform Data**

In this step, transform the raw data so that it is suitable for training. For example, **recode** the ‘**Trash Pollution**’ column value to be either **0**, **1**, or **2**.

1. **Normalization**

After data preprocessing, you are required to **normalize** the data. Use the **StandardScaler**

package to normalize the data.

1. **Generate Model**

Next, you are required to **generate** a **model** from the data. Use the **KMeans** package to generate the model into **2 cluster**.

1. **Visualization**

After the model is generated, you can **visualize** the model using the **pyplot** package. Don’t forget to add **x-label**, **y-label**, and **title** for your plot.

1. **Model Testing and Evaluation**

Then, you can **test** the model to check predict whether the data will be in **polluted water cluster** or in the **non-polluted water cluster**. Print the accuracy of your model and get the **model** with **minimum accuracy 80% or higher.**

**Please ask your teaching assistant if there are any related questions.**