|  |  |
| --- | --- |
| **Example Case** |  |
| COMP6140 | COMP6140001 | COMP6140016 | COMP6140049  Data Mining |
| **Computer Science** | **Example Case** |
| ***Valid on*** *-* | **Revision 00** |

## Soal

*Case*

* + - 1. **Classification**

**COVID19 Detector**

In the year 2020, there are COVID19 outbreaks which happen in almost all the countries. These datasets provide the prediction whether the user’s risk of getting COVID19 is **high** or **low**, depending on the data.

You are given the data for **training** named “**covid-train.csv**” and testing data named “**covid-test.csv**” in separate CSV files. Do classification with **K-NN**, **Naïve Bayes**, or **Decision Tree** to classify that the patient is **COVID19 risk** as **high** or **low** for the **testing data** that has not been classified yet, based on the model generated by the **training data**. You must make sure that the algorithm already achieves the **highest** performance of the model. The details of the data are:

***Table 1. covid-train.csv***

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| ID | Polynomial | The id of the datasets where the id has the form of an UUID |
| Region | Polynomial | The region where the person related is currently living |
| Age | Integer | The age of the person related |
| Body\_Temperature | Polynomial | The body temperature of the person related |
| Blood\_Pressure | Polynomial | The blood pressure of the person related |
| Asphyxiation | Polynomial | Indicating whether the person related feels asphyxiation upon examination |
| Weight | Integer | The weight of the person related (Kg) |
| Height | Integer | The height of the person related (cm) |
| Last\_Contact | Polynomial | Indicating where is the last contact with another people of the person related (overseas inflow, public place, contact with patient, none) |
| Risk | Polynomial | The risk of the person related of getting COVID19 (High, Low) |

***Table 2. covid\_test.csv***

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| ID | Polynomial | The id of the datasets where the id has the form of an UUID |
| Region | Polynomial | The region where the person related is currently living |
| Age | Integer | The age of the person related |
| Body\_Temperature | Polynomial | The body temperature of the person related |
| Blood\_Pressure | Polynomial | The blood pressure of the person related |
| Asphyxiation | Polynomial | Indicating whether the person related feels asphyxiation upon examination |
| Weight | Integer | The weight of the person related (Kg) |
| Height | Integer | The height of the person related (cm) |
| Last\_Contact | Polynomial | Indicating where is the last contact with another people of the person related (overseas inflow, public place, contact with patient, none) |

Your tasks for the **classification** are:

1. **Part 1**

* **Preprocess** the given **training data** based on data **requirements** that described on the table above.
* To achieve the **highest** **performance**, you must do some **tasks** below:
  + **Select** the **attributes** that want to be used for classification using **correlation matrix**.
  + Apply the classification algorithm **K-NN, Naïve bayes** or **Decision Tree**.
  + Ensure that the selected algorithm must achieve the **highest** **accuracy**.

**Documentation**

* Write down each **algorithm** you have tried, along with the **accuracy**.
* Write down the **attributes** you have choose to get the **highest accuracy** along with the **reason** **why**.

1. **Part 2**

* **Preprocess** the given **training data** based on data **requirements** that described on the table above.
* **Create** a **model** from the **training data** using the **algorithm** with the **highest accuracy** that you have find out in **Part 1**
* **Preprocess** the given **testing data** based on data **requirements** that described on the table above.
* **Apply** the classification **model** from the training data into the testing data to **classify** or **label** the **testing** **data**.
* **Output** the **labeling** result to **\*.csv file**.
  + - 1. **Clustering**

**Weather Grouping**

The datasets given consists of weather-related attributes. Using this dataset, we are going to **cluster** them into each of the **weather group**.

You are given the weather data to be clustered named “**weather\_clustering.csv**”. Do the clustering with **K-Means** to cluster the **label** of the **weather**. The details of the data are:

***Table 3. weather\_clustering.csv***

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| ID | Polynomial | The id of the datasets where the id has the form of an UUID |
| Location | Polynomial | Location where data is taken |
| Temperature | Polynomial | Average upper bound temperature taken in degree Celsius |
| Air\_Quality\_Index | Polynomial | Index of the pollution rate, higher index means cleaner air |
| Wind\_Speed | Polynomial | Average wind speed in kilometers/hour |
| Humidity | Polynomial | Whether or not the air is above the humidity threshold of 40% |
| Precipitation | Polynomial | Quantity of water deposit in milimetres (volume/area) |

Your tasks for the **clustering** are:

* **Preprocess** the given **data** based on data **requirements** that described on the table above.
* To achieve the **highest** **accuracy**, you must do some tasks below:
  + **Select** the **attributes** that want to be used for classification using **correlation matrix**.
  + Apply the clustering algorithm **K-Means**
* Change every **value** of the **clustering result** based on the following detail:
  + **Sunny**, for every weather which is in the cluster that have the following criteria:
* **High** temperature,
* **Dry** air,
* **Normal** wind speed,
* **Low** precipitation, and
* **Good** air quality index.
  + **Cloudy**, for every weather which is in the cluster that have the following criteria:
* **Normal** temperature,
* **Normal to Humid** air,
* **Low** wind speed,
* **Slight** precipitation, and
* **Good to Moderate** air quality index.
  + **Rainy**, for every weather which is in the cluster that have the following criteria:
* **Low** temperature,
* **Humid** air,
* **Low** wind speed,
* **High** precipitation, and
* **Normal** air quality index.

**Documentation**

* Write down the **attributes** you have chosen to cluster the data, along with the **reasons** (you might as well write the correlation result to support your reasoning).
* Write down the **cluster centroid table** and **analyze** why the cluster is mapped as **Sunny, Cloudy,** or **Rainy**

**Files to be collected**

* **Classification**
  + Process & Data (**.rmp**, **.ioo**, **.md**, **.properties**) to determine the highest performance of model using algorithms for classification.
  + Process & Data (**.rmp**, **.ioo**, **.md**, **.properties**) to label the testing data.
  + **\*.csv** as the labeling result
* **Clustering**
  + Process & Data (**.rmp**, **.ioo**, **.md**, **.properties**) to cluster the data.
* **Documentation**
  + Documentation described in each classification and clustering (**.docx**).

**If there’s something you don’t understand, feel free to ask your teaching assistants.**