|  |  |
| --- | --- |
| **Practicum Case** |  |
| COMP6579  Big Data Processing |
| **Computer Science** | **E201-COMP6579-DD01-09** |
| ***Valid on*** *Even Semester Year 2019/2020* | **Revision 00** |

## Learning Outcomes

* Understand Big Data Analytics and Visualizations

## Topic

* Session 09 – Data Exploration and Classification

## Subtopics

* Data Exploration using Spark
* Handling Missing Value in Spark
* Classification using Spark

## Soal

*Case*

**Bluejek Hospital**

**Bluejek Hospital** is a hospital located in Jakarta which is known for its psychology. As more and more people coming every day to **Bluejek Hospital**, the hospital found out that most people who come consulted about depression. As a way to improve their performance, they intended to create a **predictive model** that will **classify whether a person is likely to be depressed or not depressed**.

You will be given **Classification\_Train.csv** and **Classification\_Test.csv** and here is the description of the columns:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| **Name** | The person’s name. |
| **Gender** | The person’s gender (Male, Female). |
| **Height** | The person’s height in cm. |
| **Education Level** | The person’s level of education (Low, Intermediate,  High). |
| **Eye Color** | The person’s eye color (Blue, Black, Brown, Gray). |
| **Married** | Whether the person is married or not (No, Yes). |
| **Salary Income** | The person’s income per year. |
| **Depressed** | Whether the person is depressed or not (No, Yes). |

**Figure 7. Classification\_Train.csv and Classification\_Test.csv**

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

1. **Load Data**

Given the file “**Classification\_Train.csv**” and “**Classification\_Test.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 ‘**Married**’ column value to be either 0 or 1.

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 **LogisticRegression** package to generate the model with ‘**10**’ as the max iteration.

1. **Model Testing and Evaluation**

After the model is generated, you can **test** the model to predict whether the person chance of depression. Use **BinaryClassificationEvaluator** package to print the accuracy of your model. Getthe **model** with **minimum accuracy 85% or higher.**

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