|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ujian Akhir Praktikum - Semester Genap 2019/2020**  *Practicum Final Exam - Even Semester Year 2019/2020* | | | | | | |
| **Matakuliah**  *Subject* | | | **COMP6579 - Big Data Processing** | | |  |
| **Kelas**  *Class* | **:** | **BA01 / BB01 / BC01 / BD01 / BE01 / BF01 / BG01 / BH01 / BI01 / BJ01 / BK01 / BL01 / BM01 / BN01** | | **Tanggal Mulai**  *Start Date* | **: 15 June 2020** |
| **Waktu Mulai**  *Start Time* | **: 13.20** |
| **Dosen**  *Lecturer* | **:** | **D3338 - Dr. Ir. Edy Irwansyah, S.T., M.Si., IPM, ASEAN Eng.**  **D5403 - Dr. Sani Muhamad Isa, S.Si., M.Kom.**  **D5627 - I Gede Putra Kusuma Negara, B.Eng., PhD**  **D6198 - Fepri Putra Panghurian, S.Kom., M.T.I.** | | **Tanggal Selesai**  *End Date* | **: 16 June 2020** |
| **Waktu Selesai**  *End Time* | **: 13.20** |

**PERATURAN UJIAN:**

*Exam Regulations:*

* Mahasiswa tidak diperbolehkan berdiskusi dan/atau bekerja sama dengan peserta ujian lainnya

*Student is not allowed to discuss and/or work together with other exam participants*

* Mahasiswa tidak diperbolehkan menyalin jawaban dari intenet

*Student is not allowed to copy answer from the intenet*

* Asisten **BERHAK** memberi nilai 0 **(NOL)** bagi peserta ujian yang melakukan segala bentuk kecurangan

*Assistant is able to give 0 (ZERO) score for exam participant who does any cheating actions*

* Kumpulkan jawaban tepat pada waktunya di <https://laboratory.binus.ac.id/lab>

*Submit the answer on time at* [*https://laboratory.binus.ac.id/lab*](https://laboratory.binus.ac.id/lab)

* Bila Anda tidak membaca peraturan ini, maka Anda dianggap telah membaca dan menyetujuinya

*If you have missed to read these regulations, so you are considered to have read and agreed on it*

**SOFTWARE YANG DIGUNAKAN:**

*Software will be used:*

* VMware - Cloudera
* Jupyter Notebook

**FILE YANG DIKUMPULKAN:**

*File must be collected:*

* Case 1 (\*.ipynb)
* Case 2 (\*.ipynb)
* Case 3 (\*.ipynb)

**PERHATIAN!**

*Attention!*

* Bagi yang mengerjakan tidak sesuai dengan soal, maka akan diberikan nilai **NOL (0)**

*For those who do not work in accordance with the exam case will be marked as* ***ZERO (0)***

* Bagi yang mengerjakan tidak sesuai dengan software dan versi yang telah ditetapkan, maka akan tetap dikoreksi dengan software dan versi yang telah ditetapkan

*For those who do not work in accordance with the software and specific version will be corrected by the predefined software and version*

* Kompres semua jawaban yang akan diunggah. Pastikan format pengumpulan nama file dan ekstensi sesuai dengan format berikut: **[NIM]-[NAMA].zip**

*Compress all file that will be uploaded. Make sure the format for collecting file name and extension according to the following format:* ***[NIM]-[NAME].zip***

**Soal**

*Case*

**Case 1 (Query + Visualization)**

**Ramen Shop**

**Ramen Shop** is a store located in Jakarta. Since the sales are going high, they need to improve the store sales more quickly. To do that they intended to do some **analysis** from **different** **kinds** of data they have.

From the sales business process, the data can be analyzed to gain sales insight. The data is stored in **Comma-Separated** **Values (CSV)** file and the data schema is drawn using **Entity Relationship Diagram (ERD)** below:

*A screenshot of a social media post

Description automatically generated*

**Figure 1. Ramen Shop ERD**

You were given the task to gain some insight from the sales data. Below is the task you must do:

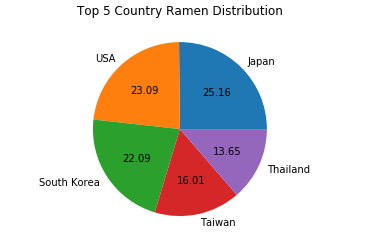
1. **Load Data from CSV to Spark**

Using **SparkSession**, **read** the following files (“MsBrand.csv”, “MsCountry.csv”, “MsCustomer.csv”, “MsRamen.csv”, “MsRamenStyle.csv”, “TransactionHeader.csv”, “TransactionDetail.csv”).

1. **Query Analysis and Visualization**

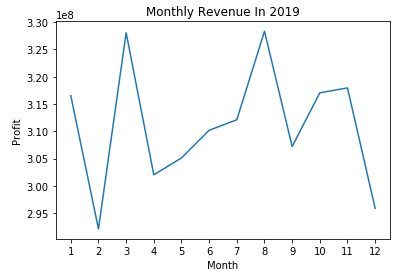
You are asked to gain some sales insight about the data. Below are some statements you need to answer. Use **SparkSQL** to answer the question and **pyplot** package to **visualize** the answer.

1. Show **top 5 country** which **distributed** **the most number of ramen** using **pie plot**. Don’t forget to add **title**, **labels,** and **percentage** for the plot.



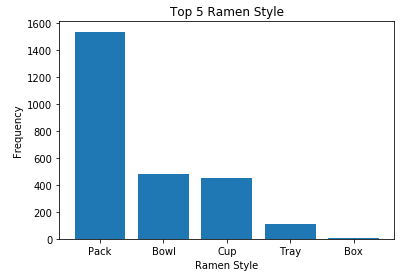
**Figure 2. Top 5 Country Ramen Distribution Figure**

1. Show the **amount of revenue** for **each month** within **year 2019** using **line plot**. Don’t forget to add **title**, **x-label**, and **y-label** for the plot.



**Figure 3. Monthly Revenue in 2019 Figure**

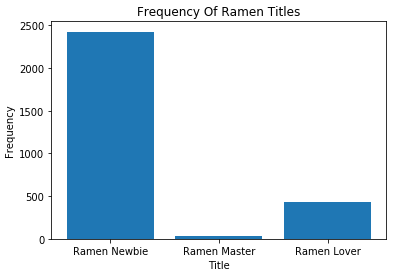
1. Show **top 5 ramen style** that is **used in making the ramen** using **bar plot**. Don’t forget to add **title**, **x-label**, and **y-label** for the plot.



**Figure 4. Top 5 Ramen Style Figure**

1. Show the **frequency of customer** who earned “**Ramen Master**”, “**Ramen Lover**”, and “**Ramen Newbie**” **title** using **bar plot**. Don’t forget to add **title**, **x-label**, and **y-label** for the plot. The **title** is achieved based on the following condition:

|  |  |
| --- | --- |
| **Ramen Ordered** | **Title** |
| > 125 | Ramen Master |
| 76 - 125 | Ramen Lover |
| 1 - 75 | Ramen Newbie |



**Figure 5. Frequency of Ramen Titles Figure**

**Case 2 (Classification)**

**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.**

**Case 3 (Clustering)**

**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.**

**Good Luck 😊**