|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ujian Akhir Praktikum - Semester Ganjil 2020/2021**  *Practicum Final Exam –Odd Semester Year 2020/2021* | | | | | | |
| **Matakuliah**  *Subject* | | | **COMP6579 – Big Data Processing** | | |  |
| **Kelas**  *Class* | **:** | **BA08** | | **Tanggal Mulai**  *Start Date* | **: 22 December 2020** |
| **Waktu Mulai**  *Start Time* | **: 15:20** |
| **Dosen**  *Lecturer* | **:** | **D5403 - Dr. Sani Muhamad Isa, S.Si., M.Kom.** | | **Tanggal Selesai**  *End Date* | **: 22 December 2020** |
| **Waktu Selesai**  *End Time* | **: 17:20** |

**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 untuk membuka dan menyalin dari **BUKU** atau **CATATAN**, **VIDEO** dari pengajar (recording kelas, VBL, Youtube, dsb) dan **REFERENSI** lainnya

*Student isn't allowed to open and copy from any resources such as notes, videos (class recording, VBL, Youtube, etc) and other references*

* Mahasiswa tidak diperbolehkan membuka dan menyalin jawaban dari internet (google, stackoverflow, dsb)

*Student isn't allowed to open and copy answer from the internet (google, stackoverflow, etc)*

* 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, apabila terlambat mengumpulkan maka jawaban tidak akan dikoreksi dan nilai mahasiswa adalah 0

*Submit the answer on time, if not, then the answer will not be checked, and the students will receive 0 (ZERO)*

* 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:*

* VM Cloudera
* Jupyter Notebook

**FILE YANG DIKUMPULKAN:**

*File must be collected:*

* 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*

**TokocAmzAne**

**TokocAmzAne** is a camera store run by a businessman called Ucok. As the sales went high,Ucok hires you to make some analysis about the data they have. Unfortunately, the database that is used to store the sales data is deleted by Ucok’s best assistant, Butet.However, Butet still has the backup file stored in **CSV file**. Below is the **TokocAmzAne ERD** that Ucok wants you to analyze:

A screenshot of a social media post

Description automatically generated

**Figure 1. TokocAmzAne ERD**

Below is the task you must do to analyze the data:

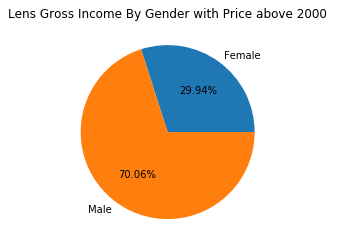
1. **Load Data from CSV to Spark**

Using **SparkSession**, **read** the following files (“camera\_transactions.csv”, “cameras.csv”, “lens\_transactions.csv”, “lenses.csv”, “members.csv”, “staffs.csv”, “transaction\_headers.csv”).

1. **Query Analysis and Visualization**

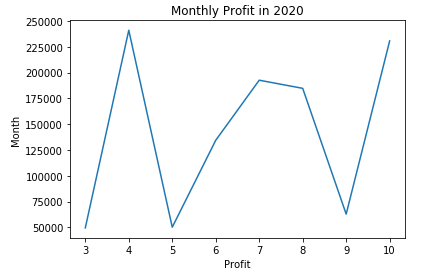
You are asked to gain some sales insight about the data. Below are some statements you need to answer. Use **SparkSQL** to answer the question and **pyplot** package to **visualize** the answer.

1. Show the **percentage of lens gross income based on gender** for lens with **price above 2000** using **pie plot**. Don’t forget to add **title**, **labels,** and **percentage** for the plot.

****

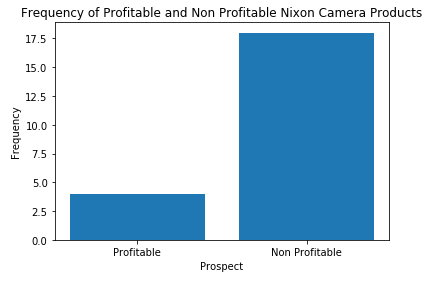
**Figure 2. Lens Gross Income By Gender Figure**

1. Show **monthly** **profit** from **lens** and **cameras** **per month in 2020** using **line plot**. Don’t forget to add **title**, **x-label**, and **y-label** for the plot.



**Figure 3. Monthly Profit in 2010 Figure**

1. Showthe **frequency** **of** **profitable cameras** where the camera is **manufactured by** **Nikon** using **bar plot**. A camera is called a **profitable camera** if it makes **more than 10.000** in all transactions. Don’t forget to add **title**, **x-label**, and **y-label** for the plot.



**Figure 4. Nixon Profitable Camera Frequency Figure**

**TokocAmzAne Customer’s Satisfaction**

**Ucok** wants to give the best service to **TokocAmzAne**. Since then, **Ucok** has collected so many **customer’s behaviors** which is stored in **Customer\_Training.csv**. Your task is to **make classification model** and **test** it to **Customer\_Testing.csv** to proof whether your model is **accurate** or **not**, so that the model can be used in the future. Here is the description of the columns in **Customer\_Training.csv** and **Customer\_Testing.csv**:

|  |  |
| --- | --- |
| Column Name | Description |
| Age | All customers’ age |
| Gender | All customers’ gender |
| Voice | Customers’ voice when interacting with customer service |
| Rating | Customer service rating given by customers when finishing a transaction |
| Walking Behavior | Customers’ walking behavior when leaving TokocAmzAne |
| Tipping | Tipping amount given by customers to customer service |
| Is Happy | Showing whether the customer is happy or not. |

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

1. **Load Data**

Given the file “**Customer\_Training.csv**” and “**Customer\_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.

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 ‘**Voice**’ 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 ‘**20**’ as the max iteration.

1. **Model Testing and Evaluation**

After the model is generated, you can **test** the model to predict whether the person is happy or not. Use **BinaryClassificationEvaluator** package to print the accuracy of your model. Getthe **model** with **minimum accuracy 85% or higher**.

**Good Luck 😊**