|  |  |
| --- | --- |
| **Assignments Case** |  |
| COMP6140001  Data Mining |
| **Computer Science** | **O223-COMP6140-JE07-01** |
| ***Valid on*** *Odd Semester Year 2021/2022* | **Revision 00** |

1. Seluruh mahasiswa tidak diperkenankan untuk:

*All students are not allowed to:*

* + - Berdiskusi dan/atau bekerja sama dengan mahasiswa lainnya

*Discuss and/or work together with other student participants*

* + - Melihat sebagian atau seluruh jawaban mahasiswa lain

*Seeing a part or the whole answer from another student*

* + - Membuka dan menyalin dari **BUKU** atau **CATATAN**, **VIDEO** dari pengajar (recording kelas, VBL, Youtube, dsb) dan **REFERENSI** lainnya

*Open and copy from any resources such as notes, videos (class recording, VBL, Youtube, etc) and other references*

* + - Membuka dan menyalin jawaban dari internet (google, stackoverflow, dsb)

*Open and copy answer from the internet (google, stackoverflow, etc)*

* + - Mengerjakan soal yang tidak sesuai dengan tema yang ada di soal,

*Working with another theme which is not in accordance with the existing theme in the matter of the case,*

* + - Melakukan tindakan kecurangan lainnya,

*Committing other dishonest actions,*

* + - Secara sengaja maupun tidak sengaja melakukan segala tindakan kelalaian yang menyebabkan hasil karyanya berhasil dicontek oleh orang lain / kelompok lain.

*Accidentally or intentionally conduct any failure action that cause the results of the project was copied by someone else / other groups.*

1. Jika mahasiswa terbukti melakukan tindakan seperti yang dijelaskan butir 1 di atas, maka **nilai mahasiswa** yang melakukan kecurangan (menyontek maupun dicontek) akan di – **NOL** – kan.

*If the student is proved to the actions described in point 1 above, the score of the student which committed dishonest acts (cheating or being cheated) will be “Zero”*

1. Perhatikan jadwal pengumpulan jawaban, segala jenis pengumpulan jawaban di luar jadwal tidak dilayani.

*Pay attention to the submission schedule, all kinds of submission outside the schedule will not be accepted*

1. 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*

1. Persentase penilaiaan untuk matakuliah ini adalah sebagai berikut:

*Marking percentage for this subject is described as follows:*

|  |  |
| --- | --- |
| **Tugas Mandiri**  *Assignment* | **Proyek**  *Project* |
| 40% | 60% |

1. Software yang digunakan pada matakuliah ini adalah sebagai berikut:

*Software will be used in this subject are described as follows:*

|  |
| --- |
| **Software**  *Software* |
| R 4.0.2  RStudio 1.3.9.959  RapidMiner Studio 9.7 |

## Ekstensi file yang harus disertakan dalam pengumpulan tugas mandiri, dan uap untuk matakuliah ini adalah sebagai berikut:

*File extensions should be included in assignment and project collection for this subject are described as follows:*

|  |  |
| --- | --- |
| **Tugas Mandiri**  *Assignment* | **UAP**  *Final Exam* |
| Folder Project (RDATA, RHISTORY, RPROJ, R, RPROJ.USER) | RMP, PROPERTIES |

## Soal

*Case*

**Favorite Food**

**Jett** is a volunteer researcher to improve the food consumption quality based on the students’ favorite food type. **Jett** gathered all the necessary data including the student data and their food of choice.

**Jett** has created forms asked manually and goes to school personally to collect data necessary for his research. Below are the provided data in the csv format as following:

* **favorite\_food\_data**.**csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| Student Name | Character | The name of the corresponding Student. |
| Favorite Food | Character | The favorite food type of the corresponding Student. |

* **Student\_data**.**csv**

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Data Type** | **Description** |
| id | Integer | The unique identifier of the corresponding Student. |
| Student | Character | The name of the corresponding Student. |
| Age | Integer | The age of the corresponding Student. |
| Height | Integer | The height of the corresponding Student. |
| Education Grade | Integer | The education grade of the corresponding Student. |
| Education Year | Integer | Education year of the corresponding Student. |

You as an expert in **Data Mining** is hired by **Jett** to help him to **analyze** and **visualize** the data. The requirements are as below:

1. **Data Visualization**

To help **Jett** to understand the data easier, you are asked to **visualize** the data in **graph** **form**. Some data that needed to be visualized are:

* 1. Show the **frequency of** **favourite food types** by students**.**

Chart, pie chart

Description automatically generated

Figure 1. Frequency of Favourite Food Types

* 1. Show the **number of students by their education grade and education year**.

Chart, bar chart

Description automatically generated

Figure 2. Student Distribution by Grade Chart

* 1. Show the **average height of all primary school (“SD”) students by their education year** and the **students’ age** must **less than equal to 12.**

Chart, line chart

Description automatically generated

Figure 3. Average Height of Primary School Students by Their Education Year Chart.

1. **Frequent Pattern Analysis**

**Jett** wants to know which **Food Type** is **popular nowadays with the** **students**. You are asked to do **frequent pattern analysis** to search for **frequent Food Type** that is liked by junior and senior high school. To get the **frequent Food Type** data, use the “**favorite\_food\_type**.**csv**” and “**student\_data**.**csv**” files and do some steps below:

* 1. **Data Pre-processing**

In **Data Pre-processing** phase, there are some data that can’t be used for the further analysis. Do the following task to **cleanse** the data:

* **Remove** all **Students** with **empty name**.
* **Remove** all **Students** with an **Education Grade of Primary** (**“SD”**)
* **Remove** all **Duplicated data.**
  1. **Data Transformation**

In this phase, you need to **change** **the** **data**, so it is suitable to be used in the **Apriori** analysis. Prepare the data in term of the **Favourite Food Type.**

* 1. **Data Mining**
* Show **frequent** **Favourite Food Type** using **Apriori** algorithm with **minimum support**: **0**.**25** based on the data that have already pre-processed.

Text

Description automatically generated with medium confidence

Figure 4. Frequent Food Types using Apriori.

* Show the **association** **rules** using **minimum confidence**: **0**.**6** based on the frequent **Favourite Food Type** that resulted from step above.

Text

Description automatically generated

Figure 5. Association Rules.

**Files to be collected**:

* R Studio Project Folder consist of:
  + - .RData
    - .Rhistory
    - .Rproj
    - .r