|  |  |
| --- | --- |
| **Practicum Case** |  |
| COMP6140 | COMP6140001 | COMP6140049  Data Mining |
| **Computer Science** | **E241-COMP6140-RE01-11** |
| ***Valid on*** *Even Semester Year 2023/2024* | **Revision 00** |

## Learning Outcomes

* LO 2 – apply various data mining techniques

## Topic

* Session 11 – Outlier detection

## Sub Topics

* Distance-based outlier detection (nested loop)
* Density-based outlier detection (LOF)

## Soal

*Case*

**Sales Outlier**

The datasets given consists of sales in a pharmacy stores. From the datasets given, identify which of the data is outlier using the outlier detection algorithm.

You are given the data to be identified named **sales.csv**. Do the outlier detection using **Distance Based** and **LOF (Local Outlier Factor)** outlier detection. The **details** of the data are:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| datum | Date | The age of the participants |
| M01AB | Integer | Anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives and related substances |
| M01AE | Integer | Anti-inflammatory and antirheumatic products, non-steroids, Propionic acid derivatives |
| N02BA | Integer | Other analgesics and antipyretics, Salicylic acid and derivatives |
| R03 | Integer | Drugs for obstructive airway diseases |

Your tasks for the **outlier detection** are:

* To achieve **good results** on the outlier detection, you must do some **tasks** below:
  + **Set** the **role** of the **attributes** that doesn’t needed to be counted for the **outlier** **detection**.
  + **Tuning** the **parameters** of the **LOF algorithm**.
  + **Tuning** the **parameters** of the **distance based algorithm**.
* **Filter** the **results** to the outlier data.