**MACHINE LEARNING POC TEST (ANOMALY DETECTION)**

The test cases are based on a real problem in which you must create a predictive machine learning model for supervised and unsupervised.

Your first step is to read the data description below and create an observation data based on ID of order.

**ORDER DATASET**

|  |  |
| --- | --- |
| column\_name | description |
| id | ID of order |
| wholesaler\_id | ID of wholesaler |
| retailer\_id | ID of retailer |
| buyer\_type | retailer = registered, guest = unregistered |
| shipping\_type | method of shipping |
| order\_type | app = online order via app, otc = order via wholesaler cashier |
| book\_time | purchase time |
| last\_status | recent status of order |
| last\_status\_time | time of last status |
| anomaly\_label | 1= yes, 0 = no (this column is used only for supervised case) |

**ORDER PRODUCT DATASET**

|  |  |
| --- | --- |
| column\_name | description |
| order\_id | ID of order |
| id | ID of order\_products |
| sku\_id | ID of sku |
| brand | name of brand |
| packaging | unit of measurement |
| packaging\_amount | quantity of corresponding unit of measurement |
| amount | product quantity |
| price | product price |
| book\_time | purchase time |
| last\_status | recent status of order product |
| last\_status\_time | time of last status |

1. **Supervised Machine Learning**

Using the dataset of orders and order\_products, you must predict the anomaly\_label and complete the following tasks

**a. Explanation**

- Please explain thoroughly which features that you are going to use and please provide the reasons on why you choose those features.

- How you deal with missing data and/or outlier data

- What kind of algorithms you use. Please provide the reason on why do you choose it and the result of those algorithms as well as the evaluation metrics and why you choose it.

**b. Output**

Your output will be **anomaly label** in which you must provide the label (1/0) to the **test dataset**

1. **Unsupervised Machine Learning**

Using the dataset of orders and order\_products, use only for **label = 0** for the train data and then build a unsupervised model that use clustering/outlier detection/pattern recognition to establish 3 pattern types

**a. Explanation**

- Please explain thoroughly which features that you are going to use and please provide the reasons on why you choose those features.

- How you deal with missing data and/or outlier data

- What kind of algorithms you use. Please provide the reason on why do you choose it and the result of those algorithms.

**b. Output**

Your output will be **anomaly label** (0/1/2) and **patter type** (2 columns). Your pattern type should consist of 3 type of clusters/outlier types/patterns and give the description about it.