# LAPORAN TUGAS BESAR

diajukan untuk memenuhi tugas mata kuliah (CII4I3) Penambangan Data

oleh: Otniel Abiezer (NIM 1301180469)



PROGRAM STUDI S1 INFORMATIKA
FAKULTAS INFORMATIKA
UNIVERSITAS TELKOM
BANDUNG
2022

## 1. HUBUNGAN ANTARA FILE

Items.csv, category\_hierarchy.csv, dan order.csv sebagai train, dan submission.csv sebagai test. Pada items.csv memiliki itemID sebagai PK (Primary Key) dan category sebagai FK (Foreign Key) yang berisi keterangan tentang item, sementara pada category\_hierarchy dengan category sebagai PK, dan order.csv dengan menggunakan FK dari items.

#### 2. PREPROCESSING DAN ANALISIS

Semua proses preprocessing yang dilakukan adalah

1) Imputasi missing value

Karena hampir semua data tidak memiliki missing value kecuali pada kolom categories di items

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32776 entries, 0 to 32775
Data columns (total 8 columns):
    Column Non-Null Count Dtype
    itemID
              32776 non-null int64
             32776 non-null int64
1
    brand
    feature 1 32776 non-null int64
2
3 feature 2 32776 non-null int64
4 feature 3 32776 non-null int64
    feature 4 32776 non-null int64
    feature 5 32776 non-null int64
    categories 25988 non-null float64
dtypes: float64(1), int64(7)
memory usage: 2.0 MB
```

Banyaknya data yang missing adalah 20% dan melakukan imputasi missing value dengan nilai -1 karena data bernilai kategori dan -1 untuk menambah kategori baru.

# 2) Drop Duplikasi

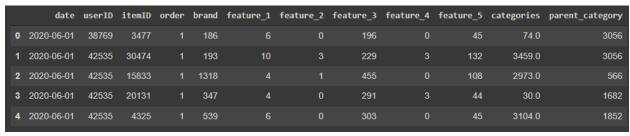
```
[19] category['category'].nunique()
4301

/ [20] category.shape[0]
4333

/ [21] category = category.drop_duplicates(subset=['category'], keep='first')
category.shape[0]
4301
```

Pada kolom category di category\_hierarchy.csv memiliki duplikasi karena seharusnya seluruh category bernilai unik, tetapi total datanya lebih banyak daripada nilai unik tersebut (4333 > 4301). Untuk itu, dilakukan drop duplikat dengan mengambil nilai pertama saja yang diakui.

3) Penggabungan ketiga dataframe

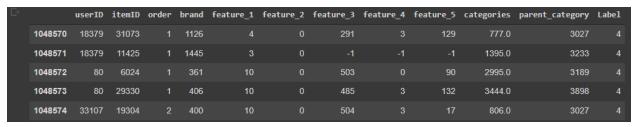


Seluruh tiga file csv terpisah disatukan dengan menggunakan Join berdasarkan hubungan PK-FK

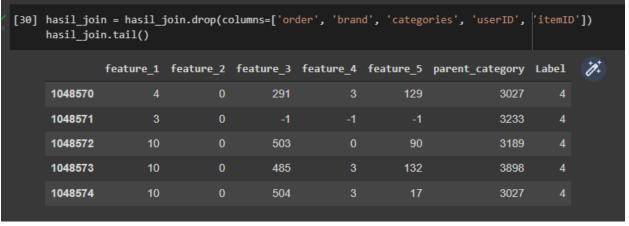
# 4) Membuat Label berdasarkan Date

Karena memerlukan prediksi tentang minggu ke berapa (ada 4 minggu dalam 1 bulan), maka dibuat label baru dengan melihat day pada date lalu melihat kondisinya sebagai berikut

- a) Tanggal  $1 \le x \le 7$  menjadi 1
- b) Tanggal  $8 \le x \le 14$  menjadi 2
- c) Tanggal  $15 \le x \le 21$  menjadi 3
- d) Sisanya menjadi 4



5) Drop kolom tidak perlu



6) Melakukan pergeseran nilai +1 dengan kolom bernilai minimal -1

Hal ini dilakukan agar dapat dilakukan algoritma klasifikasi berikutnya

# 3. METODE

```
[37] cnb = CategoricalNB()
cnb.fit(X_train, y_train.values.ravel())

CategoricalNB()
```

Metode yang digunakan adalah Categorical Naïve Bayes karena semua hasil akhir memiliki nilai kategorikal, sehingga dengan menggunakan categorical naïve bayes merupakan pendekatan yang tepat dan juga merupakan algoritma klasifikasi yang cukup cepat untuk data yang banyak.

Untuk pembagian data test dan data train adalah 75% dan 25%.

# 4. EVALUASI

Evaluasi dengan menggunakan 2 metrik, yaitu akurasi dan Confusion Matrix seperti yang terlihat

Untuk klasifikasi 4 kelas, sudah termasuk baik.

## 5. VIDEO PRESENTASI

https://youtu.be/TDGDgI0xtLk