**TP Individuel**

**Formation** : Data stewards

**Cours** : Système d’information décisionnelle

**Soumissionnaire** : Hubert Azonvidé DOSSA

**2- Explication de la donnée récupérer sur le site de Kaggle:** [**https://www.kaggle.com/**](https://www.kaggle.com/)

La base exploitée provient du site de Kaggle et est dénommée " **Credit Card Transactions Fraud Detection Dataset**". Elle est disponible directement à travers ce lien : <https://www.kaggle.com/datasets/kartik2112/fraud-detection> .

L'ensemble de données "Credit Card Transactions Fraud Detection Dataset" disponible sur Kaggle est une base de données simulée contenant des transactions par carte de crédit, avec à la fois des transactions légitimes et frauduleuses. Les jeux de données sont de deux catégories : Table des transactions de test (Test Dataset) et Table des transactions de train (Train Dataset). Ces données peuvent être utilisées pour entraîner et tester des modèles de détection de fraude de cartes de crédit.

Voici une description détaillée de cette base de données :

* **Durée de la période :**

Les données couvrent la période du 1er janvier 2019 au 31 décembre 2020.

* **Clients :**

La base de données inclut des transactions effectuées par 1000 clients.

* **Marchands :**

Les transactions sont effectuées avec un pool de 800 marchands.

* **Source de la simulation :**

Les données ont été générées à l'aide de l'outil Sparkov Data Generation développé par Brandon Harris. Cet outil utilise la bibliothèque Python "faker" pour générer des transactions en se basant sur des listes prédéfinies de marchands, clients et catégories de transactions. Il simule des transactions en fonction de différents profils tels que l'âge, le sexe et la localisation, ainsi que des paramètres comme le nombre de transactions par jour, la distribution sur les jours de la semaine et la distribution des montants pour créer des transactions réalistes.

* **Agrégation des données :**

Les transactions de tous les profils sont générées et fusionnées pour créer une représentation globale des transactions simulées.

**3- Quelques pistes d’analyse possible de la donnée**

* **Analyse exploratoire des données (EDA) :**

Comprendre la distribution des transactions, examiner les montants des transactions, identifier les tendances dans les transactions légitimes par rapport aux transactions frauduleuses.

* **Modélisation de détection de fraude :**

Développer des modèles d'apprentissage automatique pour détecter les transactions frauduleuses en se basant sur des caractéristiques telles que le montant, le marchand, l'heure de la transaction, etc.

* **Segmentation des clients :**

Segmenter les clients en fonction de leur comportement de transaction (fréquence, montant), identifier les clients à forte valeur, etc.

* **Analyse des marchands :**

Analyser les volumes et les montants des transactions pour différents marchands, identifier les marchands populaires, etc.

* **Analyse temporelle :**

Examiner les tendances et les schémas dans l'activité des transactions au cours de la période de deux ans.

**La base Mysql**

|  |
| --- |
|  |

**Gui Bash**

|  |
| --- |
|  |