PROJET INTER PROMO 2022

TRAVAIL PRÉPARATOIRE SUR LES DONNÉES

INVENTAIRE
MISE EN FORME
ECHANTILLONNAGE

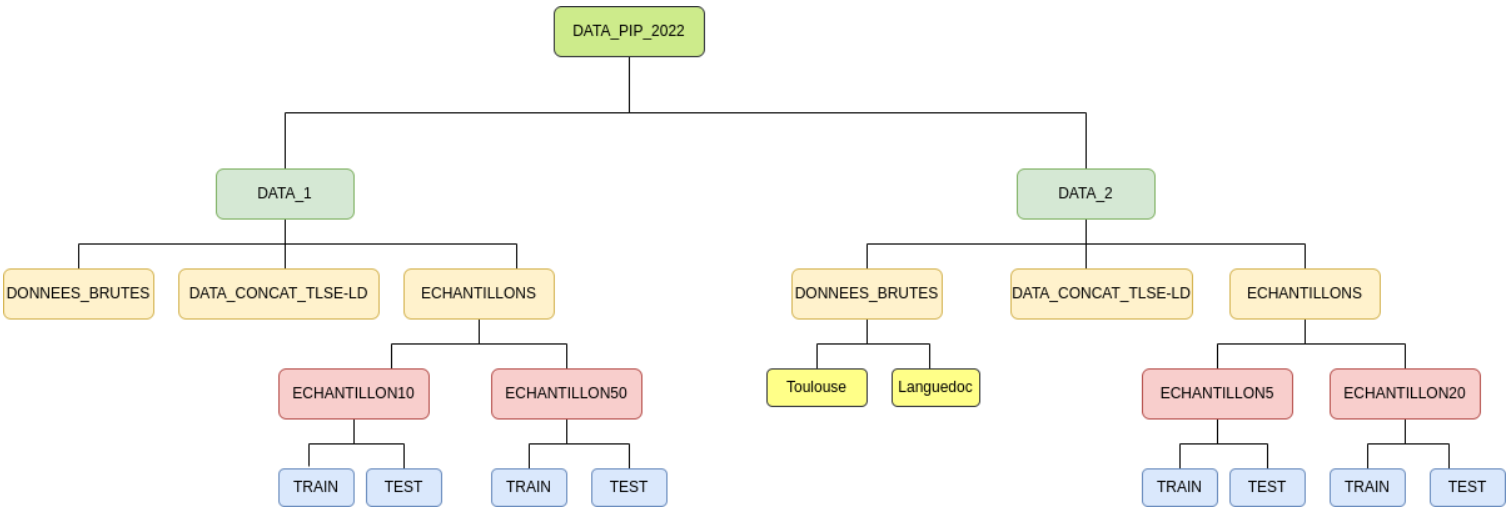
RÉALISÉ PAR :
AMAL Ghita
CHALIGNÉ Damien
AFESSA Emanuel
HADDA Hassan

ENCADRÉ PAR:
CHOUQUET Cecile
SERRURIER Mathieu
GRANDGIRARD Emma

Les Notebook :

Vous trouverez joint à ce document les notebooks Python détaillants les procédures de mise en forme des données . Ainsi qu'un fichier zip organisé selon l'arborescence décrite ci dessous .

- IMPORT_DATA_1.ipynb : Importation des jeux de données 1
- IMPORT_DATA_2.ipynb : Importation des jeux de données 2
- INFO_DATA_2.ipynb : Information sur les variables des jeux de données 2
- JEUX_DONNEES_1.ipynb : Informations sur les jeux de données 1 + echantillonnage
- JEUX_DONNEES_2_TOULOUSE.ipynb : Manipulation des jeux de données 2 pour Toulouse (jointures entre dataframes)
- PREPARATION_DATA_2VF.ipynb: Manipulation des jeux de données 2 pour Languedoc et exemples d'échantillonnage
- fonction.py : fonctions utiles pour importer les jeux de données 2 et afficher les informations
- fonction_import_data.py : fonctions utiles pour importer les fichiers zip



JEUX DE DONNÉES 1 - Détection de clients sensibles à la fraude via phishing

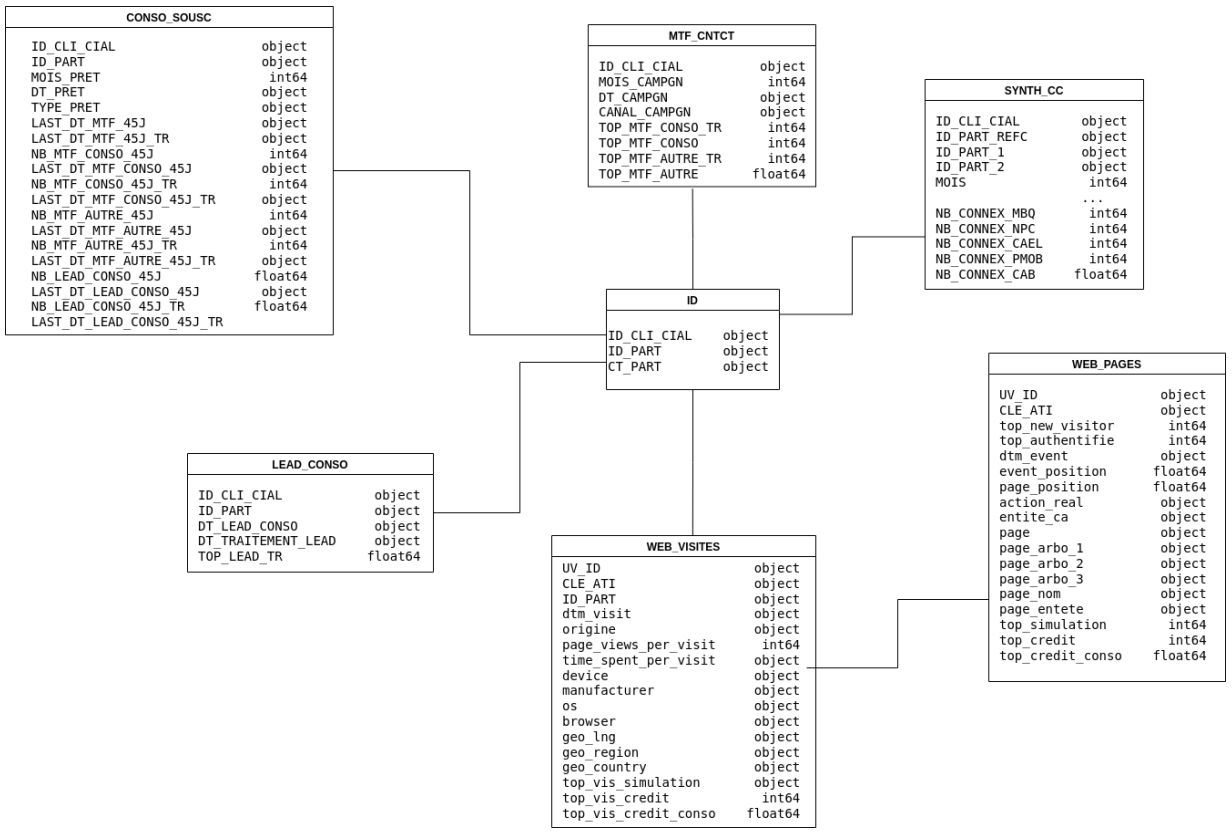
Après avoir récupérer les données du jeu 1 , nous avons réalisé les tâches suivantes :

1. Description des données
    a. Nombre de lignes et de colonnes.
    b. Type de chaque colonne .
    c. Affichage et comptage des valeurs distinctes pour chaque colonne.
    d. Pourcentage des valeurs nulles pour chaque variable.

2. Concaténation des jeux de données de Toulouse et Languedoc

3. Échantillonnage stratifié de 10% et 50%
    Séparation des échantillons en jeux de Test (50%) et jeux d'entraînement (50%).

JEUX DE DONNÉES 2 - Contribution des parcours digitaux sur la souscription d'un crédit consommation:



Afin de faire apparaitre la variable à expliquer dans toutes les tables et aussi pourvoir tracer le comportement des clients à travers les pages Web, tout en récupérant leur infos nous avons suivi les étapes ci dessous :

1. Import des données Toulouse + Languedoc
2. Informations sur les les dataFrames WEB_PAGES et WEB_VISITES de LangueDoc (très volumineux)
3. LANGUEDOC - Extraction de la variable à expliquer TOP_CONSO¶
4. TOULOUSE - Extraction de la variable à expliquer TOP_CONSO
5. Concaténation LANGUEDOC+TOULOUSE¶
6. Exportation des tables concaténées en fichier csv
7. Echantillonnage de 5%¶
8. Jointures les tables lead_conso, conso_sousc, synth_cc, mtf_cntct avec l'echantillon de 5%¶
9. Echantillonage de 20%
10. Jointures les tables lead_conso, conso_sousc, synth_cc, mtf_cntct avec l'echantillon de 20%
11. Jeu d'entrainement et Jeu de Test (pour l'echantillon de 5% et de 20%)
12.
13. Jointures des tables des tables web_pages_L et web_visites_L et synth_cc_L avec l'echantillon de 5%