Rapport projet Data Mining

Sujet : Le datamining pour la prévision de succès des films

1) Compréhension du besoin métier

Problème Métier : La prévision de succès des films

Objectif du datamining dans ce cas : Utiliser le datamining pour ressortir des donnes descriptives pour mieux comprendre si un film sera à succès ou non **(A revoir à la fin)**

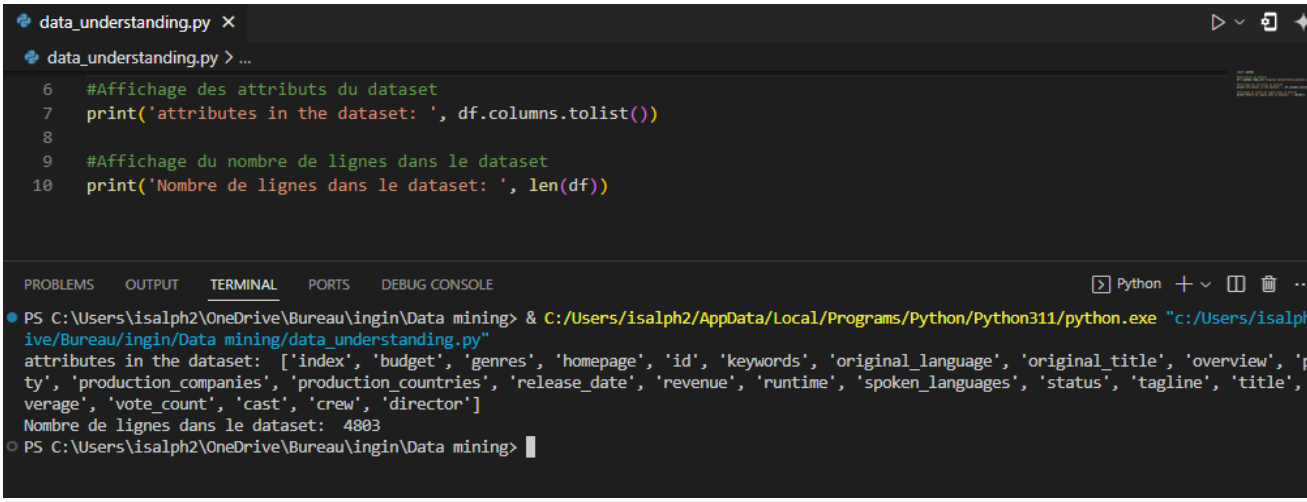
2) Compréhension des donnees

- **Collecte des données** : Ici, on se servira du dataset **movie_dataset.csv** de chez Kaggle via le lien : https://www.kaggle.com/datasets/utkarshx27/movies-dataset?resource=download
- **Description des du dataset** :
- **Description des colonnes** :
  - **Index** : Les numéros de chaque instance
  - **Budget** : Le budget de production du film
  - **Genre** : genre du film (ex : Action, Drame etc…)
  - **Homepage** : Le site web du film
  - **Id** : identifiant unique du film dans **The movie Database**
  - **Keywords** : Mots clés associes au film
  - **Original language** : Langue originale du film
  - **Original title** : Titre original
  - **Overview** : Description du film
  - **Popularity** : indicateur TMDB pour mesurer la popularité
  - **Production-company** : Compagnie de production du film
  - **Production country** : Pays de production
  - **Release date** : date de sortie du film
  - **Revenue** : Recette générer par le film (en $)
  - **Runtime** : Duree du film en minutes
  - **Spoken-language** : Liste des langues disponibles pour le film

- **Status** : Etat du film (disponible, en production etc..)
- **Tagline** : Slogan du film
- **Title** : Titre du film
- **Vote average** : Note moyenne note par les utilisateurs
- **Vote count** : Nombre de votes par les utilisateurs
- **Cast** : Liste des acteurs principaux du film
- **Crew** : Membres de l'équipe technique
- **Director** : Nom du réalisateur du film

- **<u>Donnees à analyser</u>** :
  - Succès Commercial (revenue)
  - Succès critique (Vote)

  <u>Nb</u> : On pourra etudier ces succès en fonction d'autres variables explicatives comme le revenu en fonction du genre etc…

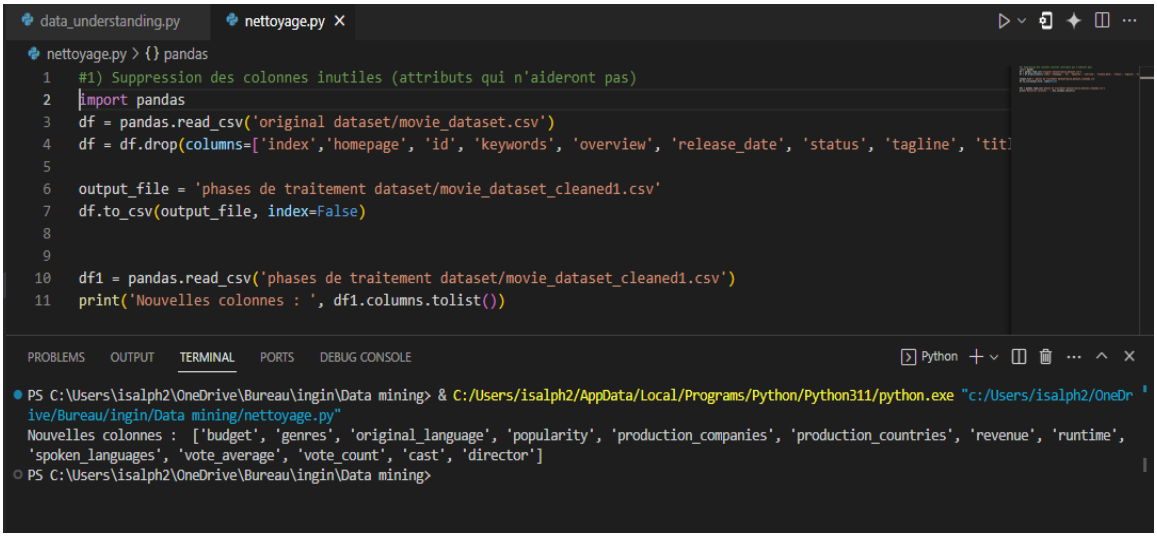


3) <u>Construction du Data Hub</u>

Ici, on constate avec amertume que les donnees ne sont pas prêtes à l'utilisation.

- Suppression des colonnes contenant des attributs inutiles à savoir :
  - Index
  - Homepage
  - Id
  - Keywords
  - Overview
  - Release date

- Status
- Tagline
- Title
- Original Title
- Crew



4) Modélisation

Voir le fichier **process.ipynb**

5) Evaluation

Métrique	K-Means	K-Medoids	Apriori
Silhouette (Commercial)	0.45–0.55	0.42–0.52	N/A
Silhouette (Critique)	0.40–0.50	0.38–0.48	N/A
Itemsets fréquents	N/A	N/A	106
Règles générées	N/A	N/A	356
Lift moyen	N/A	N/A	1.728
Lift maximal	N/A	N/A	2.274s
Classement global	**1er**	**2e**	_

On peut conclure ici que le meilleure modèle ici est le modèle **K-means** à cause des meilleurs résultats qu'il Offre ; En outre, avec la méthode **Apriori**, On obtient un lift moyen >1 ce qui signifie qu'il y'a forte dépendance entre les variables explicatives ;

6) Déploiement

Voir déploiement.py

On remarque par exemple ici qu'avec la création d'une nouvelle instance avec les coordonnées suivantes, on obtient un succès commercial dans le cluster des films à succès mais avec un revenue bas probable **(mais est dans le meilleur cluster)**

Budget (en dollars): 150000000
Popularité (0-100): 40
Runtime (en minutes): 65
Vote Average (0-10): 8
Vote Count (nombre de votes): 10000