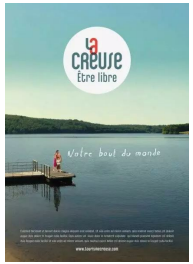
Projet 2 :
Création d’un moteur de recommandations de films




Les étapes de notre démarche



Étape 1 : Étude de marché sur la consommation de films au cinéma sur la Creuse

Étape 2 : Appropriation, exploration des données et nettoyage

Étape 3 : Machine learning et recommandations

Étape 4 : Affinage, interface et présentation

Insee
Mesurer pour comprendre

pandas

scikit learn
Machine Learning with Scikit-Learn

Streamlit

Étape 1: Étude de marché sur la consommation de cinéma sur la Creuse

La majorité de la population se situe entre **45 ans et 75 ans**.

Top 3 des genres : Films **Français**, Films **Arts et essais**, Films **US**

6 cinémas: Le **Sénéchal** sort du lot !

Et plus généralement en France ?
- Les **femmes** vont plus au cinéma que les hommes
- Les **retraités** : 25 %
- Les **élèves/étudiants** : 25 %

Source: INSEE, CNC, Google

Étape 2: Appropriation, exploration des données et nettoyage

1- Téléchargement des tables

Nous avons commencé par :

-> Comprendre chaque table avec la documentation disponible

-> Déterminer ce qui va nous intéresser pour l'analyse et le traitement des données

-> Déterminer les relations entre les tables et les données qu'elle contiennent

2- Visualisation des données mise sous format HTML

-> *Ydata-profiling*

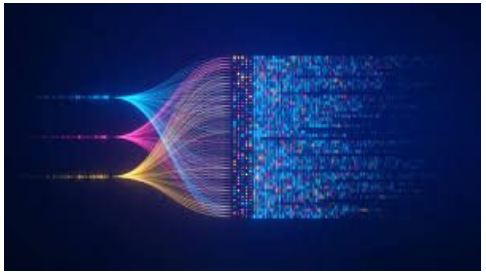
Étape 2: Appropriation, exploration des données et nettoyage

1- Tri et sélection des données

- Choix de ne **conserver que les films de la région "FR"**
- Conservation uniquement des types **"*movies*"** (films)
- Conservation uniquement des **“is original title”** (pour éviter doublons)
- Exclusion : films **n'ayant pas de date de sortie** et genre **"*adult*"**
- Exclusion de la base des films qui ne sont **pas encore sortis**

→ **Une seule table pour le système de recommandation**

<u>Étape 3:</u> Machine learning et recommandations



Mise en place dans notre DataFrame d'une série contenant

les informations permettant d'entraîner l'algorithme:

- ***Cleaning*** des données, vérifier la présence et l'absence des valeurs, mettre au bon format, création d'une nouvelle colonne ;
- Retrait des ***Stopwords,*** préparation du texte ;
- ***Stemmatisation*** des données, rapide et intéressant pour le système de recommandation ;
- Test pour le ***Lemmatizer***, mais abandonné car trop long et peu pertinent pour les résultats ;
- Mis en place de ***TF-IDF*** (vecteurs) et application du modèle ***Nearest Neighbors*** (cosinus) ;
- Suggestions avec des titres proches: ***get_close_matches.***

Étape 4: Affinage, interface et présentation

Réalisation du site avec Streamlit:

1ère page: Notre système de recommandation de films +

ajout avis/favoris/note de l'utilisateur

2nde page: Présentation des Films à l'affiche dans la Creuse

via webscrapping (Allociné)

3ème page: Quizz de recommandation en fonction des genres, des années

et de la note (grâce au ML)

Outils utilisés

Traitement des données : Jupyter - Bibliothèque Pandas - DuckDB - Natural Language Toolkit (NLTK) - Regex - Power BI

Site internet : Streamlit

Livrables : Google slide - Google Collab

Communication de groupe : Slack - Trello

Difficultés rencontrées et pistes d’amélioration

Choix et tri dans les données

Amélioration de l’algorithme du machine learning afin d’optimiser la suggestion

Enregistrer le login-Id de l’utilisateur

Mettre powerBI sur Streamlit pour un plus bel aspect esthétique

Ajout d’une page personnalisé avec favoris/liste d’envies

Envoi des notifications/emails à l’utilisateur

Ajout des suggestions des nouvelles sorties films.

Optimisation de nos ensembles de codes

Présentation des KPI - via Power bi

Présentation de notre site