# PROJET 1

* **Enoncé :**  Recommandations de produits personnalisées dans le commerce électronique

L'objectif des recommandations de produits personnalisées en e-commerce est de proposer des recommandations aux clients en fonction de leurs habitudes de navigation et d'achat. L'analyse des descriptions et des avis produits permet d'atteindre cet objectif grâce à des algorithmes de filtrage collaboratif et des techniques de traitement du langage naturel (TALN).

* Titre du projet : Recommandations de produits personnalisées dans le commerce électronique
* Ensemble de données utilisées : données des clients de cartes de crédit
* Niveau de difficulté : 2
* Concepts impliqués : Prétraitement, nettoyage des données, suppression du bruit
* Code source : <https://github.com/alanramponi/recommEngine.git>
* **Consigne :**  Instructions de Présentation des projets :
* Présentation du Data Pipeline, sous forme d'un schéma
* Présentation du DataSet
* Présentation des traitements effectués sur la data
* Pourquoi le choix d’un tel algo
* Une bonne interprétation des résultats
* Une synthèse sur le projet

# Plan du projet

## 1. Présentation du Data Pipeline (Schéma)

### 1️. Analyse des besoins (Business Understanding)

CRISP-DM : Cross-Industry Standard Process for Data Mining

CRISP-DM est un modèle de processus utilisé pour structurer les projets de Data Mining et de Machine Learning. Il divise le projet en six étapes clés, de la compréhension du problème métier jusqu’au déploiement.

**Problème métier** : Comment proposer des recommandations personnalisées aux clients en fonction de leurs habitudes d’achat ?

**Besoins des utilisateurs** :

✅ Offrir une meilleure expérience d’achat en suggérant des produits pertinents.

✅ Augmenter les ventes en incitant à des achats complémentaires.

✅ Optimiser les stratégies marketing grâce aux insights clients.

**Indicateurs de performance (KPI) :**

✅ Taux de clics sur les recommandations.

✅ Taux de conversion après recommandation.

✅ Satisfaction des utilisateurs (avis, notes).

### 2️. Cycle de vie du projet Data

💡 Ce cycle décrit les grandes étapes du projet et comment les données sont exploitées.

1️) Collecte des données

Importation des données des clients (historique d’achats, navigation, avis).

2️) Préparation et nettoyage

Gestion des valeurs manquantes, suppression des doublons, normalisation.

3️) Exploration et analyse

Visualisation et étude des patterns d’achat.

4️) Modélisation et algorithmes

Mise en place du système de recommandations (collaboratif, basé sur le contenu).

5️) Évaluation et optimisation

Tests sur la qualité des recommandations et amélioration du modèle.

6️) Déploiement et suivi

Mise en production (si applicable) et surveillance des performances.

Une image contenant texte, capture d’écran, multimédia, logiciel

Le contenu généré par l’IA peut être incorrect.

## 2. Présentation du Dataset (Data understanding)

### 1. Source des données

* Données collectées à partir des transactions de cartes de crédit des clients.

### 2. Type de données

* Historique d’achats : Produits achetés, montants dépensés.
* Navigation : Produits consultés, interactions sur le site.
* Avis clients : Notes et commentaires laissés sur les produits.
* Catégories de produits : Types de produits achetés par chaque client.

### 3. Variables clés

* ID Client : Identifiant unique du client.
* Achats précédents : Liste des produits achetés.
* Notes attribuées : Évaluation donnée aux produits.
* Fréquence d’achat : Périodicité des achats.

### 4. Exploration des données

* Statistiques descriptives : Moyenne, médiane, écart-type des achats.
* Distribution des données : Répartition des notes attribuées.
* Visualisation des tendances : Évolution des achats dans le temps, catégories les plus populaires.

### 5. Problèmes possibles

* Données manquantes : Clients n’ayant pas laissé d’avis.
* Bruit : Erreurs de saisie, achats inhabituels.
* Valeurs aberrantes : Achats anormalement élevés ou faibles.

### 6. Objectif

* Explorer et analyser les données disponibles pour comprendre les comportements d’achat et identifier des tendances.

#### Illustration

* **Dataset 1000\_Reviews (avis de clients)**

|  |  |  |
| --- | --- | --- |
| Colonnes | Description | Exemple |
| reviewerID | Identifiant unique de l'auteur de l'avis. | A1ZQZ8RJS1XVTX |
| asin | Identifiant du produit. | 0657745316 |
| reviewerName | Nom du client ayant laissé l'avis. | gsxrgirl |
| helpful | Indicateur de la pertinence de l'avis sous forme de liste [positifs, totaux]. | [0, 0] |
| reviewText | Texte de l'avis laissé par le client. | "No sugar, no GMO garbage, no fillers that come with store bought extracts. This stuff is just amazing..." |
| overall | Note globale donnée par le client (sur 5). | 5 |
| summary | Résumé de l'avis. | "Best vanilla I've ever had" |
| unixReviewTime | Date de publication de l'avis en format Unix timestamp (secondes depuis 1970). | 1381449600 |
| reviewTime | Date de publication de l'avis au format "jour mois, année". | 10 11, 2013 |

* **Dataset 500 Products (informations sur les produits)**

|  |  |  |
| --- | --- | --- |
| Colonnes | Description | Exemple |
| Asin | Identifiant unique du produit. | 0657745316 |
| Title | Titre du produit. | 100 Percent All Natural Vanilla Extract |
| imUrl | URL de l'image du produit. | http://ecx.images-amazon.com/images/I/41gFi5h0jYL.\_SY300\_.jpg |
| Related  (Relations avec d'autres produits) | |  | | --- | | Produits similaires associés à celui-ci. |  |  | | --- | |  | | {'also\_viewed': ['B001GE8N4Y']}  (Contient des dictionnaires avec des clés comme "also\_viewed" et des listes de produits associés.) |
| salesRank | Classement du produit dans une catégorie spécifique. | {'Grocery & Gourmet Food': 374004} |
| categories | Catégories du produit, souvent sous forme de liste. | [['Grocery & Gourmet Food']] |
| price | Prix du produit. | nan |
| brand | Marque du produit. | nan |

**Le dataset Reviews** contient des informations détaillées sur les avis des utilisateurs sur les produits, y compris la note, le texte de l'avis, et la date de publication.

**Le dataset Products** donne des informations sur les produits eux-mêmes, comme le titre, l'image, les catégories, le prix et la marque.

## 3. Prétraitement et Nettoyage des Données (Data preparation)

### Vérification des valeurs manquantes et outliers.

* Identification des valeurs manquantes dans le dataset.
* Détection des valeurs aberrantes (ex : achats extrêmement élevés).

### Normalisation des données si nécessaire. (ex : mise à l’échelle des prix)

* Ajustement des échelles (ex : normalisation des prix pour éviter les écarts trop grands).
* Standardisation si nécessaire (ex : transformation des notes de 1 à 5 en un score de 0 à 1).

### Suppression des valeurs manquantes et des doublons.

* Suppression ou imputation des valeurs manquantes (ex : remplacement par la moyenne/médiane).
* Élimination des doublons pour éviter les biais dans l’analyse.

### Création de nouvelles variables

Génération de nouvelles informations exploitables comme :

* **Fréquence d’achat** : nombre d’achats par mois.
* **Catégorie préférée** : identification des préférences du client.

### Transformation des avis en données exploitables

* Analyse des avis clients avec le **Traitement Automatique du Langage Naturel (TALN)**.
* Extraction des **sentiments** (positif, neutre, négatif).
* Création d’un score d’avis basé sur les commentaires et notes.

### Objectif

* Nettoyer et structurer les données pour garantir une analyse fiable et précise.

## 4. Choix de l’algorithme (Modeling)

### a) Filtrage collaboratif : basé sur les interactions utilisateurs.

**Principe :**  
Cet algorithme repose sur l’analyse des interactions entre utilisateurs et produits (ou articles). Il identifie des utilisateurs similaires en fonction de leur comportement passé (achats, évaluations, clics) et recommande des produits populaires auprès de ces utilisateurs similaires.

**Avantages :**

* Utilise des données réelles (interactions passées) pour faire des recommandations.
* Efficace pour les plateformes avec un grand nombre d'utilisateurs et une variété de produits.

**Inconvénients :**

* Nécessite une grande quantité de données sur les utilisateurs pour être efficace (le "cold start" est un problème pour les nouveaux utilisateurs ou nouveaux produits).
* Peut ne pas fonctionner de manière optimale si les utilisateurs ont un comportement très unique ou si les produits sont très spécifiques.

### b) Filtrage basé sur le contenu : basé sur la similarité des produits.

**Principe :**  
Cet algorithme recommande des produits similaires à ceux qu'un utilisateur a déjà aimés, basé sur des caractéristiques spécifiques des produits (ex : genre, description, catégorie, etc.).

**Avantages :**

* Ne dépend pas des interactions avec d'autres utilisateurs (pas de problème de "cold start" pour les utilisateurs ou produits).
* Permet de personnaliser les recommandations en fonction des préférences précises d'un utilisateur.

**Inconvénients :**

* Peut manquer de diversité dans les recommandations (si l'utilisateur aime un certain type de produit, l'algorithme risque de recommander uniquement des produits similaires).
* Nécessite une description détaillée des produits, ce qui peut être coûteux à maintenir.

### ****c) Quel algorithme choisir ?****

#### **Pour ce cas spécifique, l'algorithme le plus adapté serait le filtrage collaboratif.**

Cela est particulièrement pertinent si tu as beaucoup de données d’interactions (achats, évaluations, etc.) provenant de nombreux utilisateurs, car il permet de capturer des patterns comportementaux qui ne seraient pas évidents à identifier avec un filtrage basé sur le contenu.

Cependant, si tu manques de données sur les interactions des utilisateurs ou si tu veux éviter le problème du "cold start", le filtrage basé sur le contenu pourrait être une bonne alternative, surtout si les produits ont des caractéristiques distinctes et bien définies.

**Objectif :** Appliquer des algorithmes pour générer des recommandations.

L’objectif est de choisir l'algorithme qui maximisera la pertinence des recommandations, en fonction de la quantité de données disponibles et du type de produits. Pour un grand volume d'interactions utilisateurs, le **filtrage collaboratif** est généralement plus performant, tandis que le **filtrage basé sur le contenu** est utile lorsque les interactions sont limitées.

## 5. Interprétation des Résultats (Evaluation)

* Visualisation des recommandations faites aux clients.
* Évaluation de la pertinence des suggestions.
* **Objectif :** Vérifier la performance des recommandations.

## 6. Synthèse et Conclusion (Deployment & suivi)

* Résumé du projet.
* Difficultés rencontrées et solutions trouvées.
* Propositions d'amélioration.
* **Objectif :** Intégrer le modèle dans un environnement de production et surveiller son efficacité.