**Cahier des charges simplifié pour un mini-projet de prédit des notes de films.**

### 1. Compréhension des Affaires (Business Understanding)

* **Objectif principal** : Prédire la **note des films** sur la base de leurs caractéristiques, notamment les **genres**, la **durée**, l'**année de sortie**, et les **synopsis**. En parallèle, explorer des **tendances** dans les genres et notes pour fournir des informations pertinentes pour des cinéphiles, critiques, ou plateformes de streaming.
* **Question d'affaires** : Peut-on prédire avec précision la note d'un film en analysant ses caractéristiques descriptives ? Y a-t-il des tendances importantes ou des relations intéressantes dans les données (ex. : genres populaires ou années de haute qualité) ?
* **Valeur ajoutée** : Ce modèle pourrait aider les entreprises de divertissement à anticiper la performance de nouveaux films et à optimiser leur contenu en fonction des genres populaires ou des caractéristiques de films bien notés.

**2. Compréhension des Données (Data Understanding)**

* **Exploration du Dataset IMDB** : Analyser les colonnes disponibles comme le **titre**, le **genre**, la **durée**, l'**année de sortie**, la **note**, et les **synopsis**.
* **Analyse exploratoire des données** :
  + Vérifier la répartition des notes et genres pour comprendre les biais potentiels.
  + Étudier les distributions des caractéristiques (ex. : distribution des genres, durée moyenne).
  + Examiner des exemples de synopsis pour évaluer la complexité du texte et le vocabulaire.
* **Visualisation initiale** : Utiliser des graphiques de répartition et de tendances pour avoir un aperçu des relations entre les genres, les années de sortie et les notes.

**3. Préparation des Données (Data Preparation)**

* **Nettoyage des Données** :
  + Traiter les valeurs manquantes (ex. : synopsis manquants ou durées non renseignées).
  + Formater les variables temporelles et catégorielles pour faciliter leur utilisation.
* **Prétraitement Textuel** :
  + Utiliser un modèle **BERT** (via Hugging Face) pour transformer les synopsis en vecteurs de caractéristiques.
  + Tokenisation et normalisation du texte pour l'adapter au modèle BERT.
* **Transformation des Caractéristiques** :
  + Encoder les variables catégorielles comme le genre en valeurs numériques.
  + Normaliser ou standardiser les variables numériques (ex. : durée des films).

**4. Modélisation (Modeling)**

* **Modèle Hybride Deep Learning** :
  + Utiliser le modèle **BERT** pour transformer les synopsis en vecteurs de caractéristiques, afin de capturer la signification sémantique des synopsis.
  + Construire un **réseau dense (DNN)** pour traiter les autres caractéristiques (genre, durée, année de sortie).
  + Fusionner les sorties du modèle BERT et du réseau dense dans une couche dense finale pour prédire la note.
* **Sélection et Entraînement des Modèles** :
  + Entraîner le modèle hybride en utilisant des techniques comme la régularisation et le dropout pour améliorer la généralisation.
  + Tester des alternatives (Random Forest, régression linéaire) comme baselines pour comparer les performances.
* **Validation** :
  + Diviser le dataset en ensembles d’entraînement et de test, puis utiliser la **validation croisée** pour évaluer la stabilité du modèle.
  + Utiliser des métriques comme l’erreur quadratique moyenne (RMSE) pour mesurer la précision des prédictions.

**5. Évaluation (Evaluation)**

* **Évaluer la Précision** :
  + Utiliser les métriques (RMSE, R²) pour évaluer les performances du modèle sur les données de test.
* **Analyse des Performances** :
  + Comparer les performances entre le modèle deep learning et les modèles plus simples (Random Forest, régression).
* **Interprétation** :
  + Analyser les caractéristiques les plus influentes dans la prédiction des notes (ex. : genres, durée, tendances temporelles).
  + Examiner les erreurs du modèle pour identifier les types de films ou genres où les prédictions sont moins précises.

**6. Déploiement (Deployment)**

* **Utilisation pratique** : Mettre le modèle en production dans un cadre de recommandation ou de prédiction de notes sur une plateforme de streaming ou de critique.
* **Rapport final** : Rédiger un rapport contenant les principales observations, les tendances identifiées, et les recommandations basées sur les résultats de l’analyse.
* **Améliorations futures** : Envisager d’intégrer de nouvelles données (comme les critiques utilisateurs) pour améliorer les prédictions et explorer d’autres modèles NLP pour un meilleur traitement des synopsis.

**Résumé des Étapes CRISP-DM pour ce Projet**

* **Objectif d'affaires** : Prédire les notes des films et analyser les tendances.
* **Données** : Analyse des colonnes IMDB (genres, synopsis, durée, année de sortie).
* **Préparation** : Nettoyage, BERT pour le texte, encodage des genres, normalisation des durées.
* **Modélisation** : Modèle hybride BERT + DNN pour prédire la note.
* **Évaluation** : Mesurer la précision et la performance (RMSE), interpréter les résultats.
* **Déploiement** : Utilisation des prédictions dans un système de recommandation ou de consultation pour les utilisateurs.