SUPMTI
CISI4
TP : MACHINE LEARNING

Rapport de Mini Projet

Une analyse des données de films et de leurs évaluations sur IMDb

Préparé par:
ABID Safae et CHIGUEUR Manal
Encadré par :
ZAKARIA HAJA

Remerciement

Nous tenons à dédier ce projet à ZAKARIA HAJA, dont le soutien et les conseils précieux ont été une source d'inspiration tout au long de ce semestre. Votre expertise et votre engagement envers notre réussite ont été inestimables, et nous vous en sommes profondément reconnaissants.

Ce projet est le fruit de votre mentorat attentif et de votre dévouement à l'excellence académique. Merci pour votre guidance et votre encouragement constant.

Sommaire

Introduction

L'essor des technologies de l'apprentissage automatique et du traitement de données a révolutionné la façon dont nous abordons et analysons les ensembles de données complexes. Dans le cadre de ce projet, nous nous concentrons sur le développement d'un programme complet reposant sur ces techniques, avec pour objectif la création et l'application de pipelines d'apprentissage automatique pour la classification ou la régression.

Ce projet repose sur plusieurs piliers essentiels : la collecte et le montage d'un ensemble de données pertinent, la définition claire d'un objectif à atteindre, et enfin la mise en place de pipelines d'apprentissage automatique pour réaliser cet objectif de manière efficace et précise.

Tout d'abord, nous nous attacherons à la construction d'un dataset solide, un processus crucial qui nécessite la sélection, l'agrégation et la préparation de données provenant de sources diverses et variées. Cette étape est essentielle pour garantir la qualité et la fiabilité des données sur lesquelles nous allons travailler.

Ensuite, nous définirons clairement notre objectif principal, que ce soit une tâche de classification ou de régression. Il est crucial de bien comprendre le problème à résoudre et les métriques à optimiser pour évaluer la performance de nos modèles.

Enfin, nous mettrons en œuvre plusieurs pipelines d'apprentissage automatique, composés de différentes étapes telles que le prétraitement des données, la sélection des caractéristiques, le choix et la configuration des modèles, et l'évaluation des performances. Ces pipelines seront conçus de manière à maximiser l'efficacité, la précision et la généralisation de nos modèles.

Dans l'ensemble, ce projet vise à explorer et à exploiter les techniques avancées de l'apprentissage automatique et du traitement de données pour résoudre des problèmes complexes et réels, et à fournir des solutions robustes et performantes dans divers domaines d'application.

Aperçu des données

1 Description du Jeu de Données

- **Source des Données:** Le jeu de données provient d'un fichier CSV contenant des informations sur les films, y compris leur titre, année de sortie, note IMDb, etc.
- **Colonnes et Types de Données:** Après le prétraitement, le jeu de données contient les colonnes suivantes :
  - Title : Le titre du film (chaîne de caractères).
  - Year : L'année de sortie du film (numérique).
  - Rating : La note IMDb du film (numérique).

2 Aperçu des Lignes

- data.head(): une vue d'ensemble rapide des valeurs présentes dans chaque colonne.

	Name	Rating
0	Inception\n(2010)	8.8
1	The Matrix\n(1999)	8.7
2	The Lord of the Rings: The Return of the King\...	9.0
3	The Lord of the Rings: The Two Towers\n(2002)	8.8
4	The Departed\n(2006)	8.5

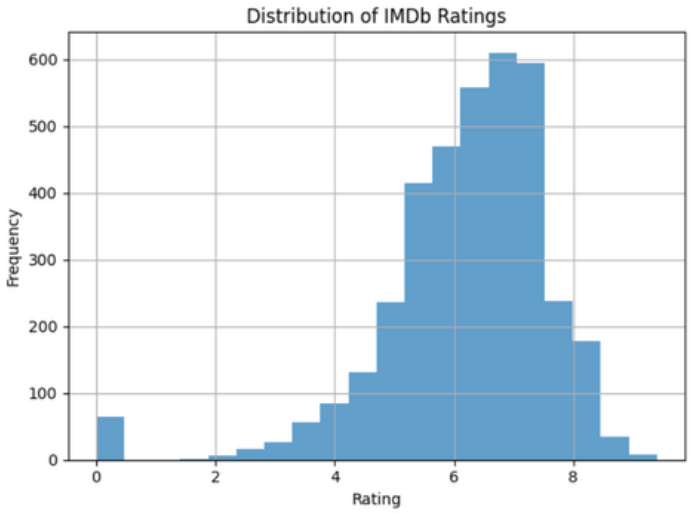
- data.info(): Vérifier s'il y a des valeurs manquantes et de s'assurer que les types de données sont corrects.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3740 entries, 0 to 3739
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Rating  3728 non-null   float64
 1   Title   3008 non-null   object
 2   Year    3008 non-null   object
dtypes: float64(1), object(2)
memory usage: 87.8+ KB
```

3 Visualisation Initiale des Données

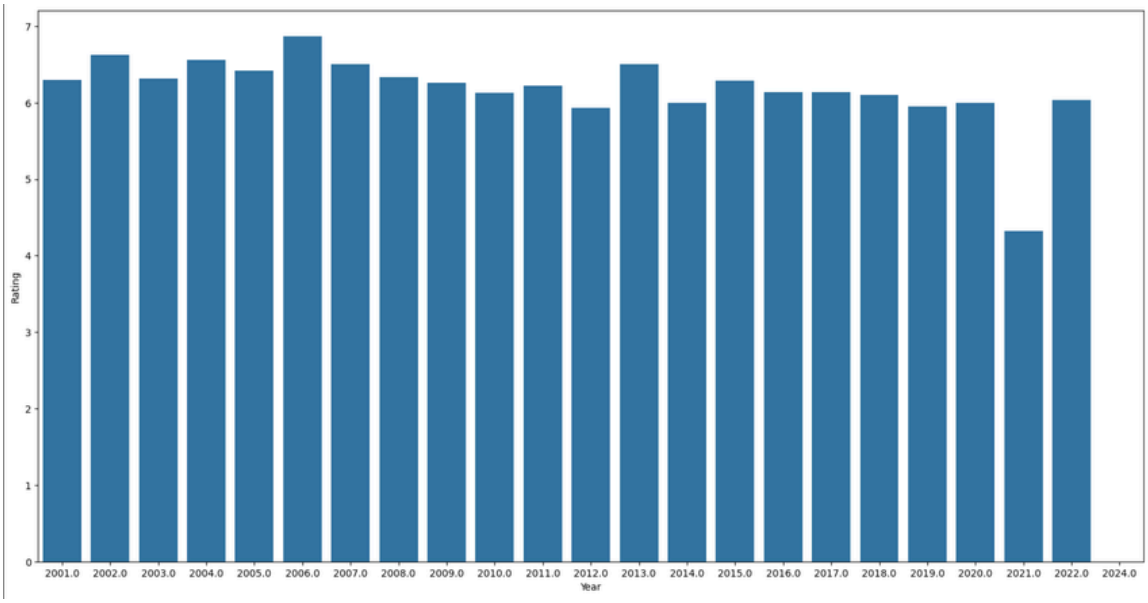
- plt.hist(data['Rating'], bins=20, alpha=0.7)
- plt.title('Distribution of IMDb Ratings')
- plt.xlabel('Rating')
- plt.ylabel('Frequency')
- plt.grid(True)
- plt.tight_layout()

Cette histogramme montre la répartition des notes IMDb, permettant de visualiser comment les notes sont distribuées parmi les films.



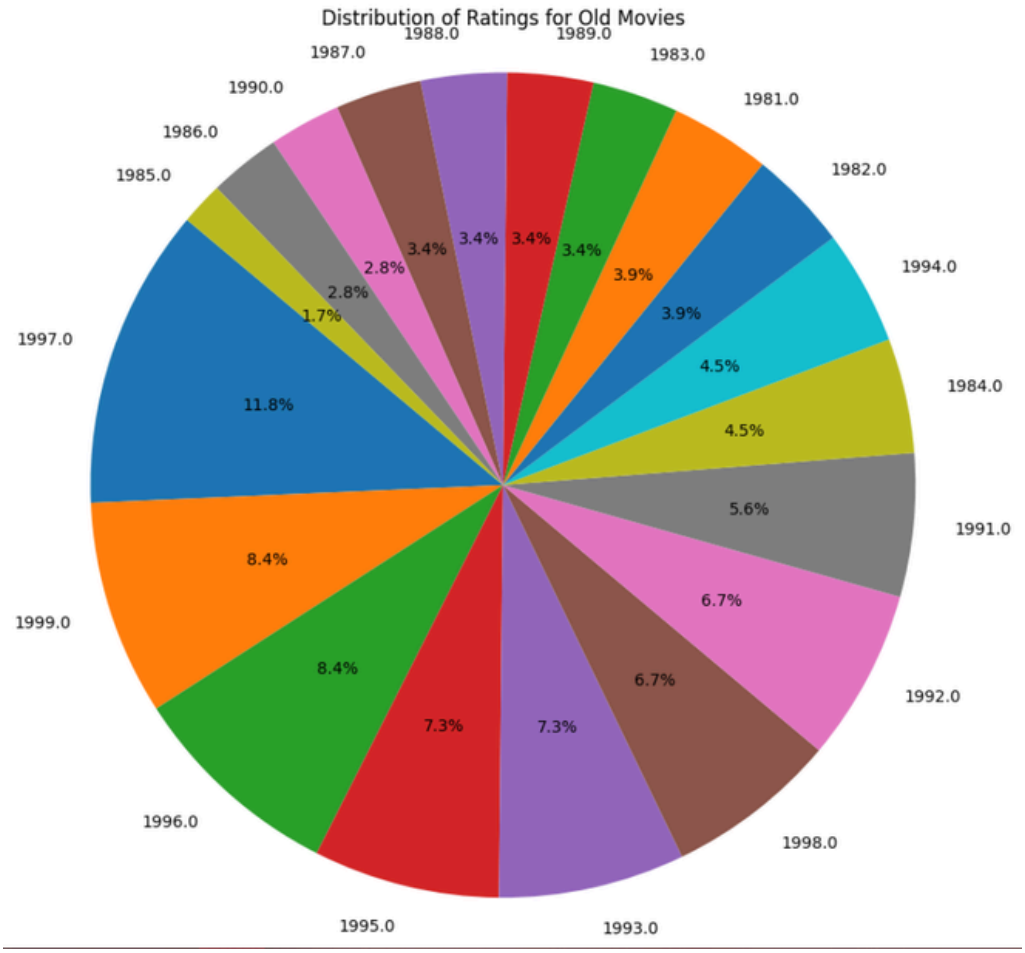
- plt.figure(figsize=(20,10))
- sn.barplot(data , x = "Year" , y = "Rating" , errorbar=None)

Ce graphique en barres montre la répartition des notes IMDb par année pour les films sortis après 2000, permettant d'identifier les tendances au fil du temps.



- plt.figure(figsize=(10, 10))
- plt.pie(old_movie["Year"].value_counts(), labels=old_movie["Year"].value_counts().index,autopct='%1.1f%%', startangle=140)
- plt.title('Distribution of Ratings for Old Movies')
- plt.axis('equal')

Ce graphique en secteurs montre la répartition des films par année pour les films sortis entre 1980 et 2000, offrant une vue d'ensemble de la production cinématographique de cette période.



Prétraitement des données

Etape de Prétraitement	Code	Justification
Lecture du Jeu de Données:	`data = pd.read_csv('/content/imdb_data1.csv')` `data = pd.read_csv(path)`	• le jeu de données est lu à partir d'un fichier CSV dans un DataFrame pandas
Extraction du Titre et de l'Année:	`data[['Title', 'Year']] = data['Name'].str.extract(r'(.+)\s\((\d{4})\)')`	• La colonne 'Name' contient à la fois le titre du film et l'année de sortie dans une seule chaîne de caractères. Pour une meilleure manipulation et analyse des données, il est important de les séparer en colonnes distinctes.
Suppression de la Colonne 'Name' d'Origine:	`data = data.drop("Name", axis=1)`	• suppression de la colonne 'Name' pour réduire l'utilisation de la mémoire et d'éviter toute confusion ou redondance dans le jeu de données.
Conversion de 'Year' en Type Numérique:	`data['Year'] = pd.to_numeric(data['Year'])`	• La colonne 'Year', initialement extraite comme une chaîne de caractères, doit être convertie en type numérique pour permettre les opérations et comparaisons numériques.
Filtrage des Données en Fonction de l'Année:	`old_movie = data[(data["Year"] > 1980) & (data["Year"] < 2000)]` `data = data[data["Year"] > 2000]`	• les films sortis entre 1980 et 2000 sont séparés dans le DataFrame old_movie, et les films sortis après 2000 restent dans le DataFrame data.

Développement de l'Interface Utilisateur

Dans cette section, nous décrivons la création d'une interface utilisateur pour visualiser et interagir avec le dataset des films. L'interface offre les fonctionnalités suivantes :

- Afficher une liste des films avec leurs notes IMDb.
- Permettre la recherche de films par titre.
- Ajouter de nouveaux films avec leurs notes.

Outils et Bibliothèques Utilisés

- CSV : Pour la lecture et l'écriture du dataset à partir d'un fichier CSV.
- Tkinter : Pour la création de l'interface graphique.
- Messagebox : Pour afficher des messages d'information et d'erreur.

1 Fonctionnalités Implémentées

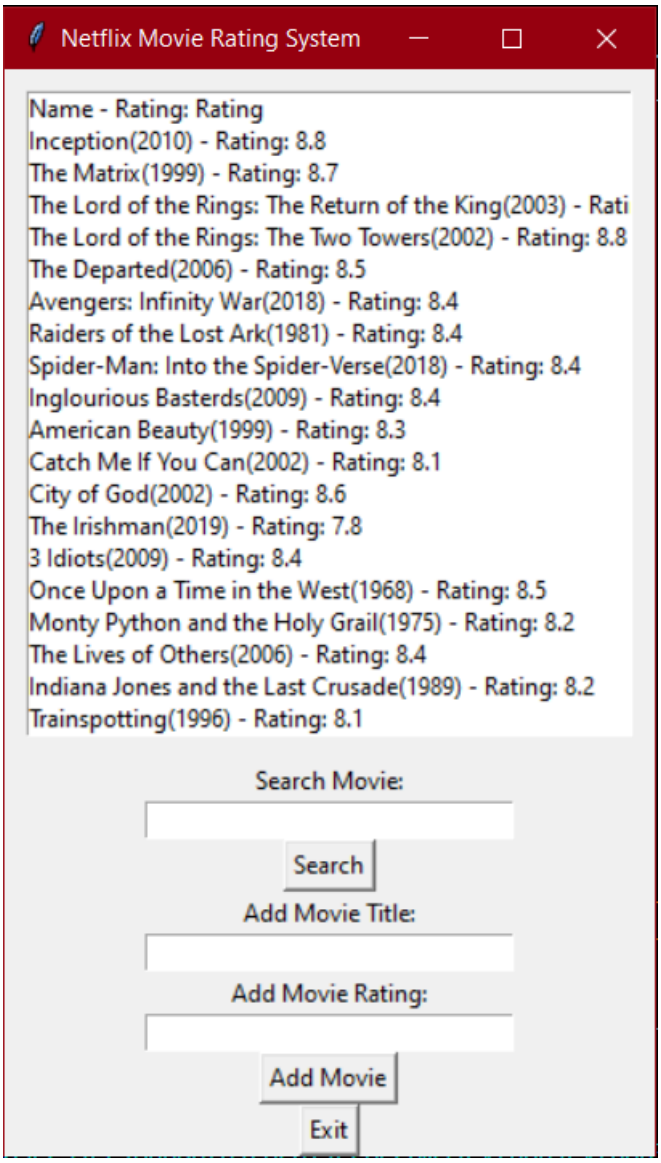
1. Chargement des Données:
   - La fonction load_dataset lit le dataset à partir d'un fichier CSV et le charge dans une liste en mémoire.
2. Affichage des Films:
   - La fonction view_all_movies affiche tous les films et leurs notes dans une liste déroulante Tkinter.
3. Recherche de Films:
   - La fonction search_movie permet de rechercher des films par titre en utilisant un champ de saisie et un bouton de recherche. Les résultats de la recherche sont affichés dans une boîte de dialogue.
4. Ajout de Films:
   - La fonction add_movie permet d'ajouter de nouveaux films avec leur titre et note IMDb. Les nouvelles entrées sont ajoutées à la liste et affichées dans la liste déroulante.
5. Interface Graphique:
   - Des champs de saisie et des boutons sont fournis pour rechercher et ajouter des films. Un bouton "Exit" permet de quitter l'application.

2 Interface Graphique

L'interface graphique est construite en utilisant Tkinter, offrant une expérience utilisateur simple et intuitive :

- Liste déroulante pour afficher les films.
- Champs de saisie et boutons pour rechercher et ajouter des films.
- Messagebox pour afficher les résultats de la recherche et les messages d'information.

3 Visualisation d'interface



Conclusion

Ce projet visait à analyser et interagir avec un dataset de films IMDb, en mettant en place une interface utilisateur intuitive. Nous avons effectué plusieurs étapes clés pour atteindre cet objectif, comprenant le prétraitement des données, l'analyse exploratoire, et le développement d'une interface graphique.

Annexe

DATASET:
https://www.kaggle.com/datasets/sukhmandeepsinghbrar/netflix-all-movie-ratings-by-imdb/data

Google Collab:
https://colab.research.google.com/drive/1J-y5YUX2J_h0M7y519Av5rL79tUXZYfy?usp=sharing

Visual Studio Code:
Pour la réalisation d'interface en utilisant python.