DATA SCIENCE & IA - 3I026

Analyses de données de films - Movie Lens

REDJEL Skander - HAZOURLI Ahmed Rachid

Introduction

Dans ce projet nous avons utilisé la base de données Movie lens ainsi que d'autres databases concernant des films. Notre but a été de pouvoir établir certaines corrélations entre les données afin de pouvoir prédire par exemple la note d'un film , la catégorie d'un film.

PROBLEMATIQUE 1 : Régression Supervisé

Première partie :

Au cours de cette analyse nous avons voulu prédire la note qui peut être attribué à un film grâce à l'utilisation des catégories d'un film, sachant que les films sont notés sur 5.

- **Algorithme utilisé :** Moindres carrés
- **Remarque :** On pouvait aussi appliquer les algos suivant:
  KNN / Non Stochastique / Stochastique

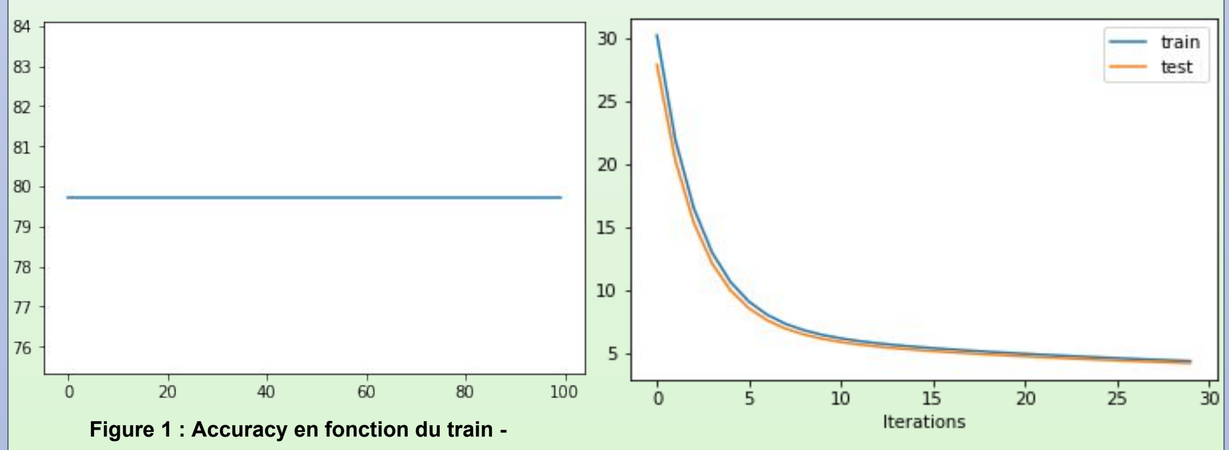
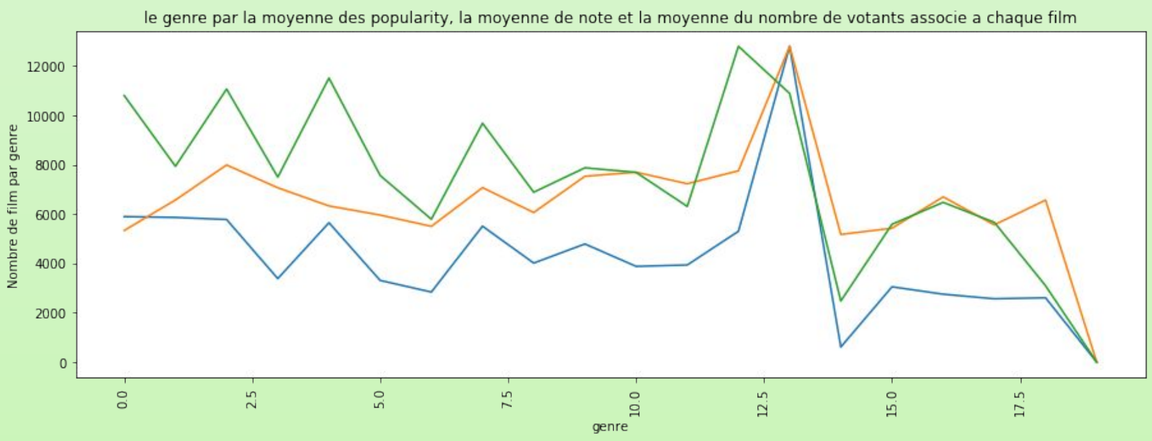


Figure 1 : Accuracy en fonction du train -

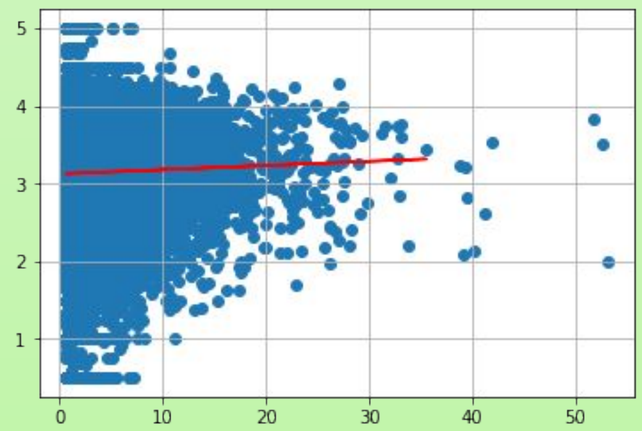
Deuxieme partie :

Nous avons voulu prédire la note moyenne à partir de la valeur de la popularité et le nombre des votants de chaque film.

- **Algorithme utilisé :** Moindres carré
- **Remarque :** On pouvait aussi appliquer les algos suivant:
  KNN / Non Stochastique / Stochastique



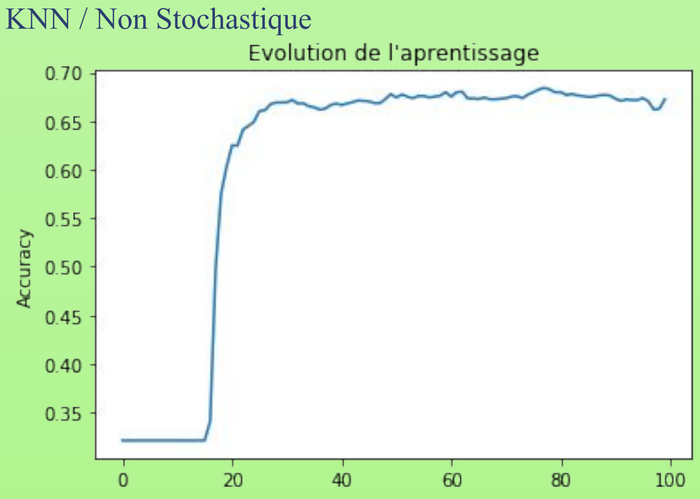
Resultat : Taux de precision : 90%



PROBLEMATIQUE 2 : Classification supervisé

Dans cette section nous nous intéressons à un problème de classification supervisé. afin d'essayer de prédire le sexe d'un acteur, si c'est un homme ou une femme.

- **Algorithme utilisé :** Gradient stochastique
- **Remarque :** On pouvait aussi appliquer les algos suivant:
  KNN / Non Stochastique



- Figure 2 : Evolution de l'appresntissage -

Accuracy

Voici un tableau récapitulatif des différentes accuracy que nous avons obtenu lorsque nous considérons divers paramètres avec divers algorithmes :Figure : Clusters en fonction des categories drame et comedie

		Id Act/ Id Prod	Popularité /Note	Id prod/ Note	IdA /IdP /pop
KNN	Accuracy Entrainement :	62.52%	61.94%	52.4%	64.10 %
	Accuracy Test:	68.23%	65.5 %	78.92%	80.5 %
Perceptron	Accuracy Entrainement :	56.55%	52.84 %	58.97 %	50.36 %
	Accuracy Test:	61.10%	58.47 %	63.00 %	58.00%
Arbre	Accuracy Entrainement :	68.02 %	70.42 %	71.97 %	64.20 %
	Accuracy Test:	73.85 %	70.01 %	72.04 %	83.51 %

PROBLEMATIQUE 3 : Classification Non Supervisé

Dans cette section, nous avons voulu catégoriser les acteurs en fonction de ce qu'ils jouent et la note moyenne des films en plus l'année moyenne.

- **Algorithme utilisé :** Clustering a l'aide de l'algorithme des K-Moyennes.
- **Résultat :** Nous avons utilisé la librairie SKlearn afin de comparer l'efficacité de notre code

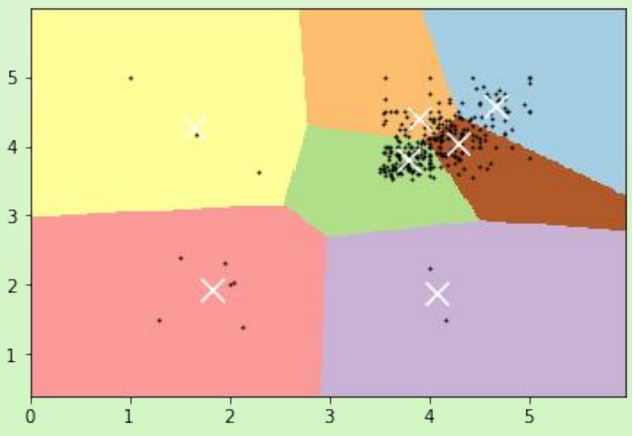


Figure : Clusters en fonction des categories drame et comedie

Nous avons voulu rajouter une dimension a notre cluster, en ajoutant une troisième catégorie qui est : Romance.

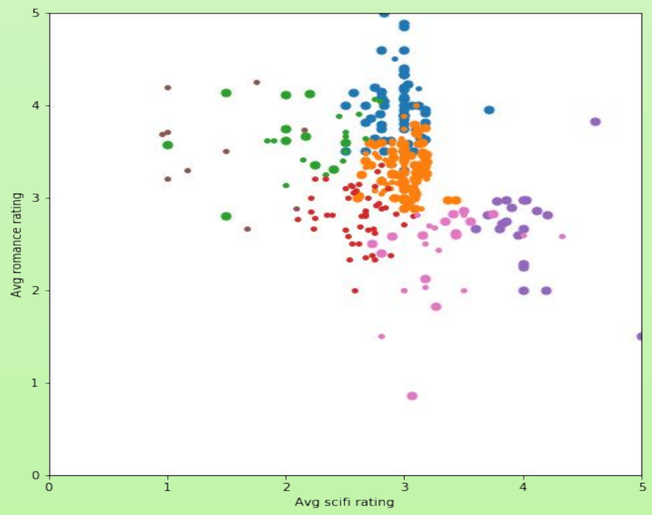


Figure : Clusters 3 Dimensions

Afin de bien choisir notre K pour le cluster on a pris le k = 7 ce qui donne un bon Silhouette score

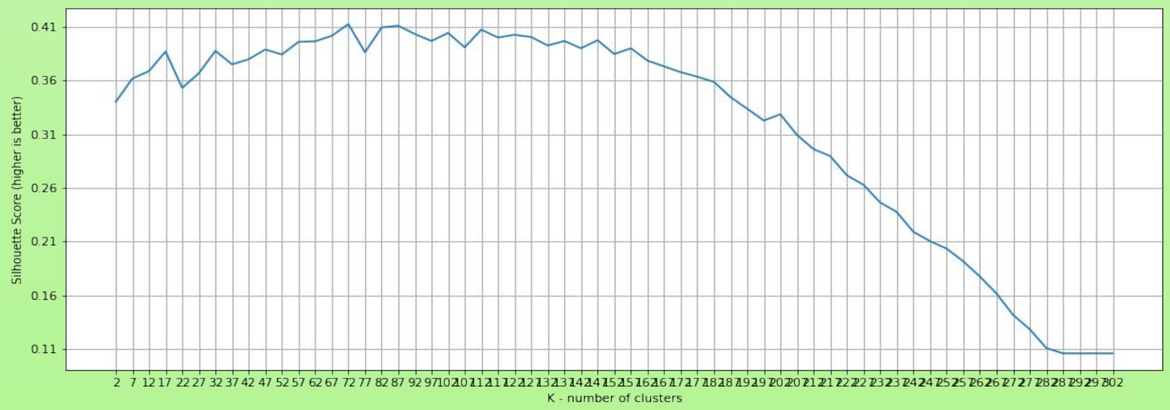


Figure : Silhouette score en fonction de K