Analyse de données de films

Merrouche Aymen & Sidhoum Imad

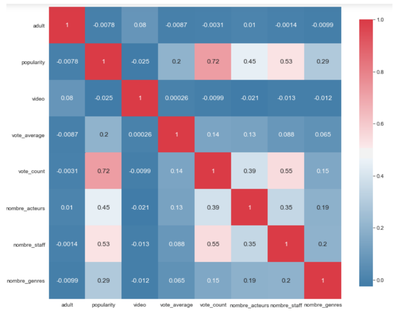
Faculté de sciences et ingénierie, Sorbonne Université

Introduction

Dans ce projet nous allons analyser des données décrivant des films.
Les problématiques traitées sont:

- Clustering des acteurs selon le genre des films dans lesquels ils tournent.
- Clustering des acteurs selon la qualité des films dans lesquels ils tournent.
- La prédiction de la note moyenne donnée à un film par les internautes
- La prédiction de la popularité d'un film
- Classification des films selon la note moyenne donnée à un film par les internautes

Corrélations entre les attributs :



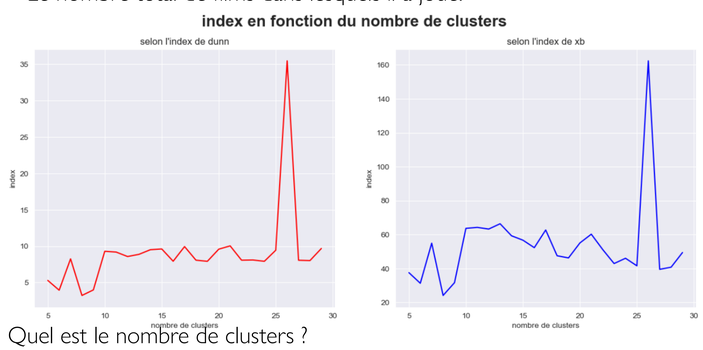
- On remarque que le nombre de vote est extrêmement corrélé avec la popularité, c'est logique puisque si le film est populaire plus de personnes vont donner leurs avis sur sa qualité.
- Le « vote average » est très légèrement corrélé avec la popularité, le nombre de votes.
- La popularité et le nombre de vote sont très corrélés avec le nombre de membres du staff et le nombre d'acteurs. Ils sont aussi légèrement corrélé avec nombre de genres dans lesquels le film est catégorisée

Catégorisation non supervisée :

Regroupement des acteurs selon les films -kmoyennes- :

Selon le genre des films dans lesquels ils ont joué, chaque acteur est caractérisé par

- Pour chaque genre : le nombre de films auxquels il a participé.
- Le nombre total de films dans lesquels il a joué.

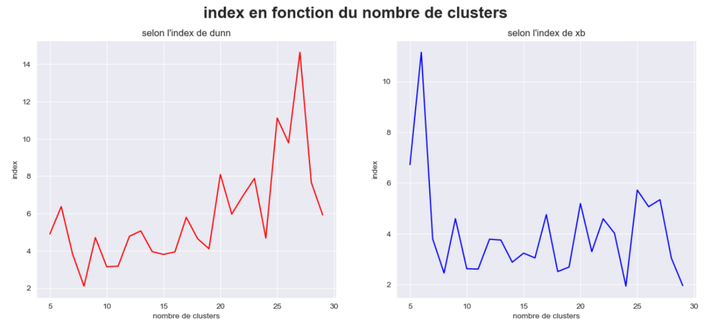


Quel est le nombre de clusters ?

- Nombre de clusters minimisant l'index de «Dunn» : 8.
- Nombre de clusters minimisant l'index de «Xb» : 8.

Selon le qualité des films dans lesquels ils ont joué, chaque acteur est caractérisé par :

- La popularité des films dans lesquels ils tournent.
- Leurs note moyenne.
- Le nombre de personnes qui ont participé au vote des films dans lesquels ils ont tourné.

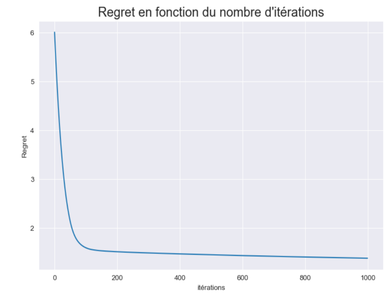


Quel est le nombre de clusters ?

- Nombre de clusters minimisant l'index de «Dunn» : 8.
- Nombre de clusters minimisant l'index de «Xb» : 24.

Régression supervisée

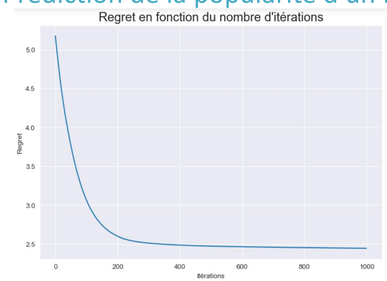
Prédiction de la note attribuée par les internautes -Moindres Carrés- :



On va prédire l'attribut "vote average" a partir de :

- Du genre, i.e. utilisation du Dumby coding:.
- De la langue, i.e. utilisation du Dumby coding:.
- De la popularité.
- Du nombre de personne qui ont voté.
- Pour adultes ou pas.

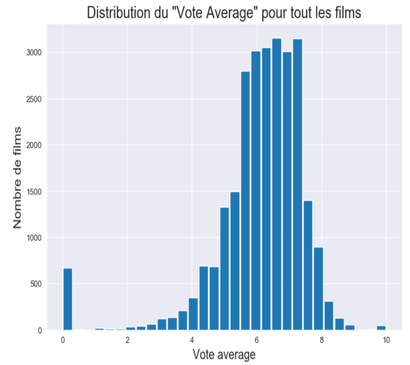
Prédiction de la popularité d'un film -Moindres Carrés- : :



Comme on a vu précédemment dans la matrice de corrélations, la popularité d'un film est corrélée avec le nombre d'acteur et le nombre de membres du staff ainsi qu'avec le nombre de votes sur le film, on va utiliser ces trois attributs pour prédire la popularité d'un film

Classification supervisée -One Against All- :

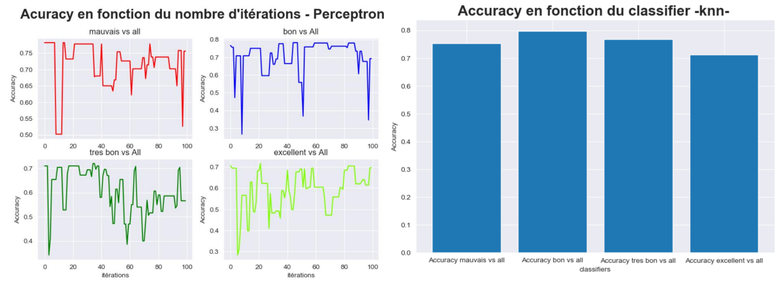
Distribution du vote average :



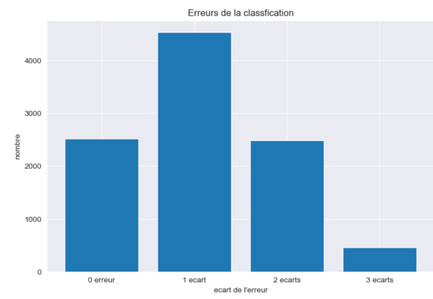
On va classifier sur l'attribut "vote average" a partir de :

- Du genre, i.e. utilisation du Dumby coding.
- De la popularité.
- Du nombre de personne qui ont voté.
- On va discrétiser l'attribut "vote average" en utilisant les quartiles.

Résultat des la classification -One Against All-:



Combinaison des résultats :



Pour estimer la qualité de la classification on définit quatre types d'erreurs selon l'écart :

- 0 : aucune erreur.
- 1 : erreur d'un écart.
- 2 : erreur de deux écarts.
- 3 : erreur de trois écarts.

email@university.edu