Film Data Analysis

Hakim CHEKIROU

Introduction:
To analyse and extract meaningful information from the IMDB data. We preprocess the data. For the sake of deciding what dataset to use , We statistically analyse the data before applying the machine learning algorithms.

PRE-PROCESSING:

> Reduction of Ratings' size.

> Deleting the meaningless features on every table.

> **Dropping missing values**.

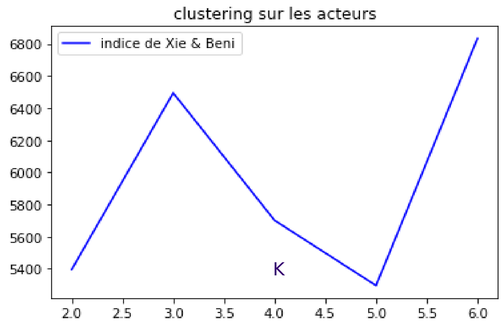
> **Adding new features**.

> **Link all the tables together**.

Non Supervised Learning : K-means

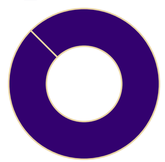
Our goal in this section is to find clusters of actor based on the types of movies they play in and there performance in them. The roles have been grouped by actor and new features have been added. We limited this analysis on only the first 10000 actors

Looking for the optimal cluster number



Analysis on the resulting clusters

K = 2



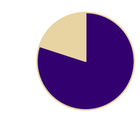
Cluster 1 male actors

Cluster 2 only composed of actresses

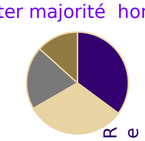
K = 5

cluster 1 hommes

cluster 2 hommes

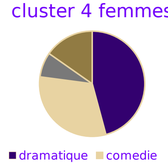


Cluster majorité hommes



cluster 3 mixte

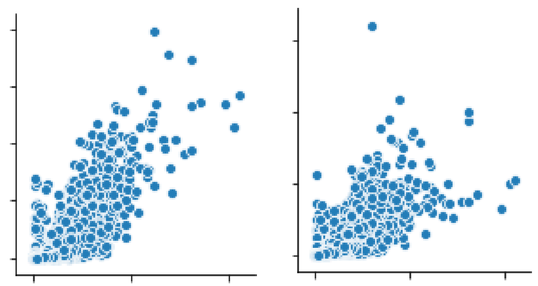
cluster 4 femmes



Regression

In this section, we are aiming at prediction the rating of a movie.

Correlation



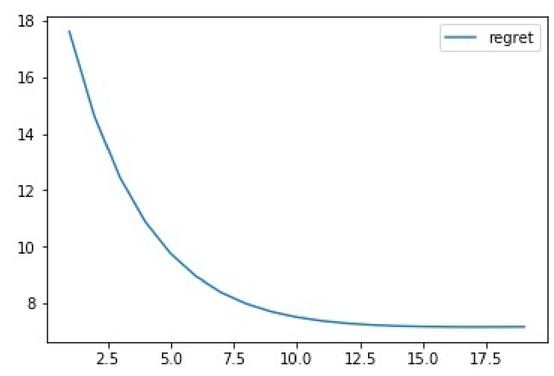
Movie Ratings

Movie Ratings

Perceptron batch

We have trained the algorithm on 25000 films, with a 0,002 learning rate.

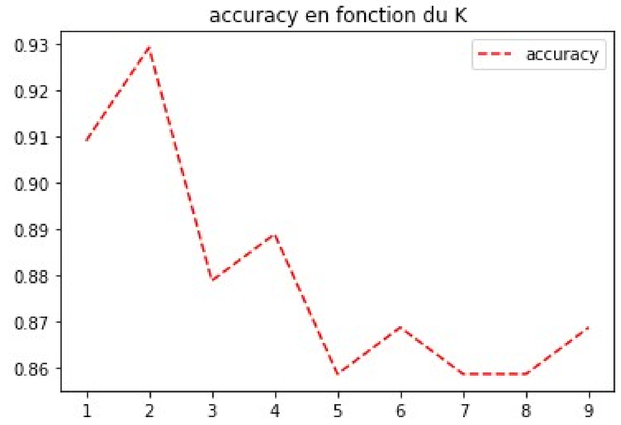
Regret



iteration

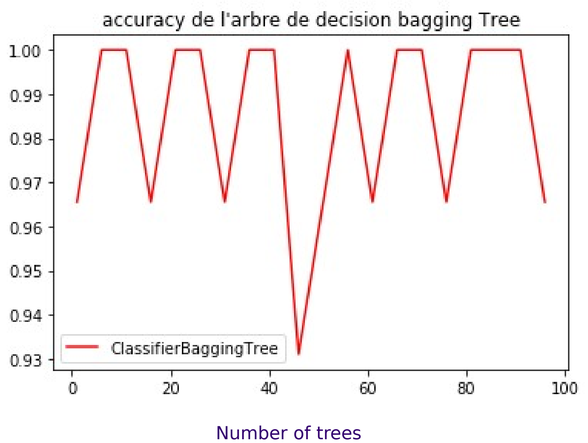
Classification

Our goal here is to predict the type of the movie.



Decision Tree

Prediction if a person is a director.
Epsilon = 0.01.



Number of trees