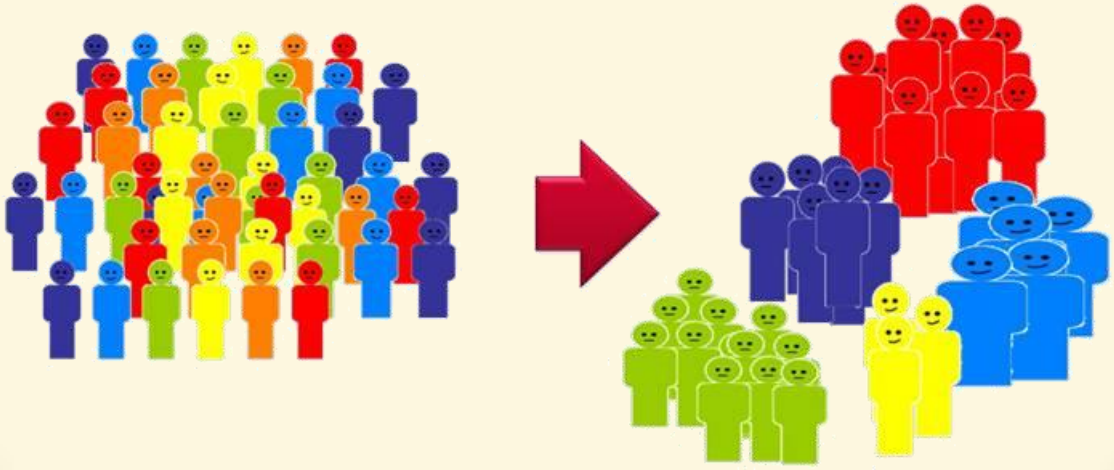
CentraleSupélec

OPENCLASSROOMS

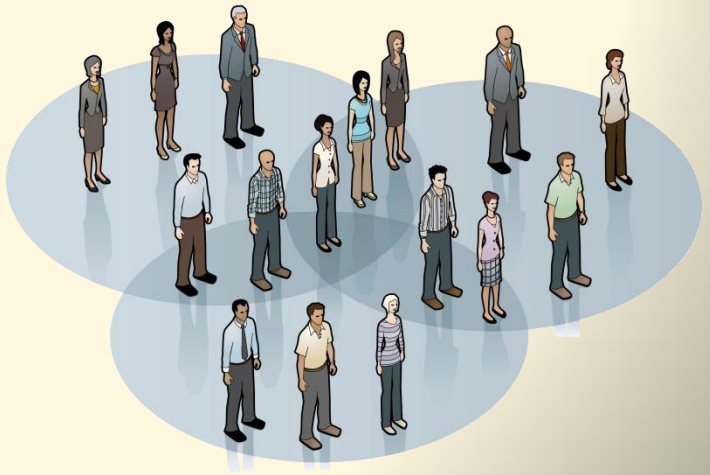
Parcours Data Scientist

Projet 5 :
Segmentation des clients



Sommaire

- Présentation et Objectifs
- Nettoyage / Exploration
- Clustering des Clients
- Interprétation des Clusters
- Classification
- API
- Pistes d'évolutions
- Conclusion



Présentation et Objectifs

- Dataset issue d'un magasin
  - Clients (ID + pays)
  - Articles achetés et qté
  - Date d'achat
  - Numéro de facture
- Objectif
  - Segmenter la clientèle par comportement
  - Prédire au plus tôt dans quel groupe sera une personne

Nettoyage et Exploration

- Achat annulés
  - Nouvelle features
  - Gardé pour la fin
    - Suppression des achats faits puis annulés
    - Explication des outliers

Cas 1 :

Achat Objet A – qte N
Annulation Objet A – qte M
avec (M < N)

➢ Difficile à gérer
❖ InvoiceNo achat manquant

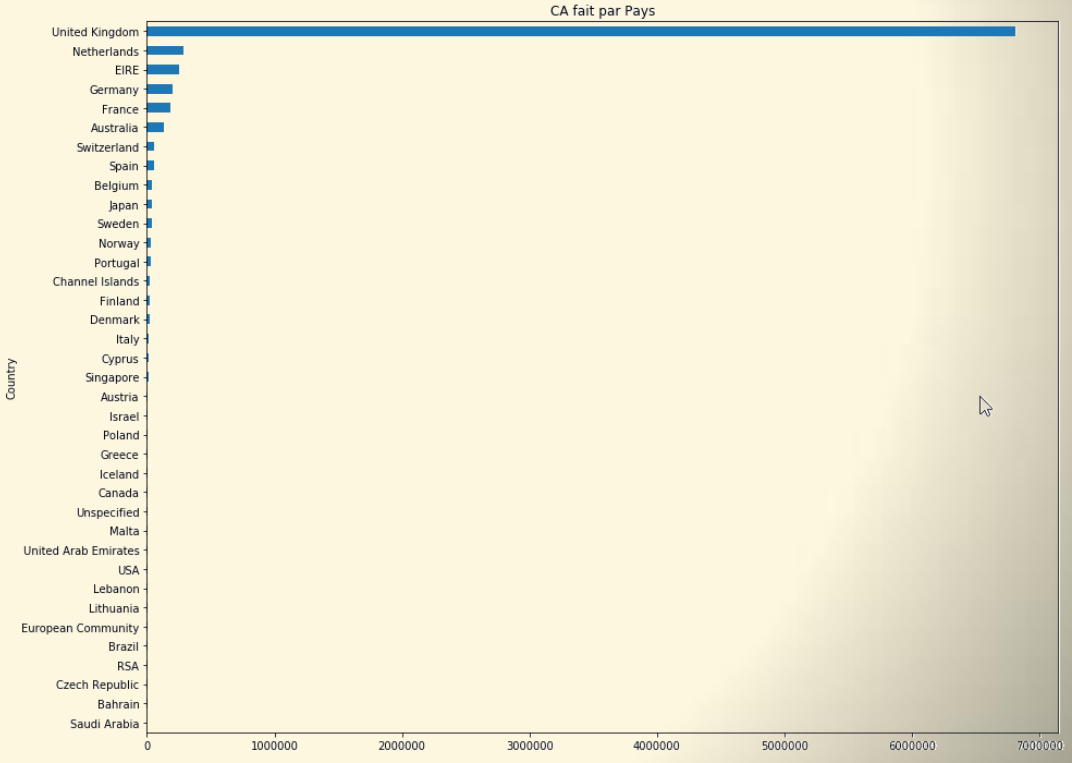
Cas 2 :

Achat Objet A – qte N
Annulation Objet A – qte N

➢ Tentative de suppression

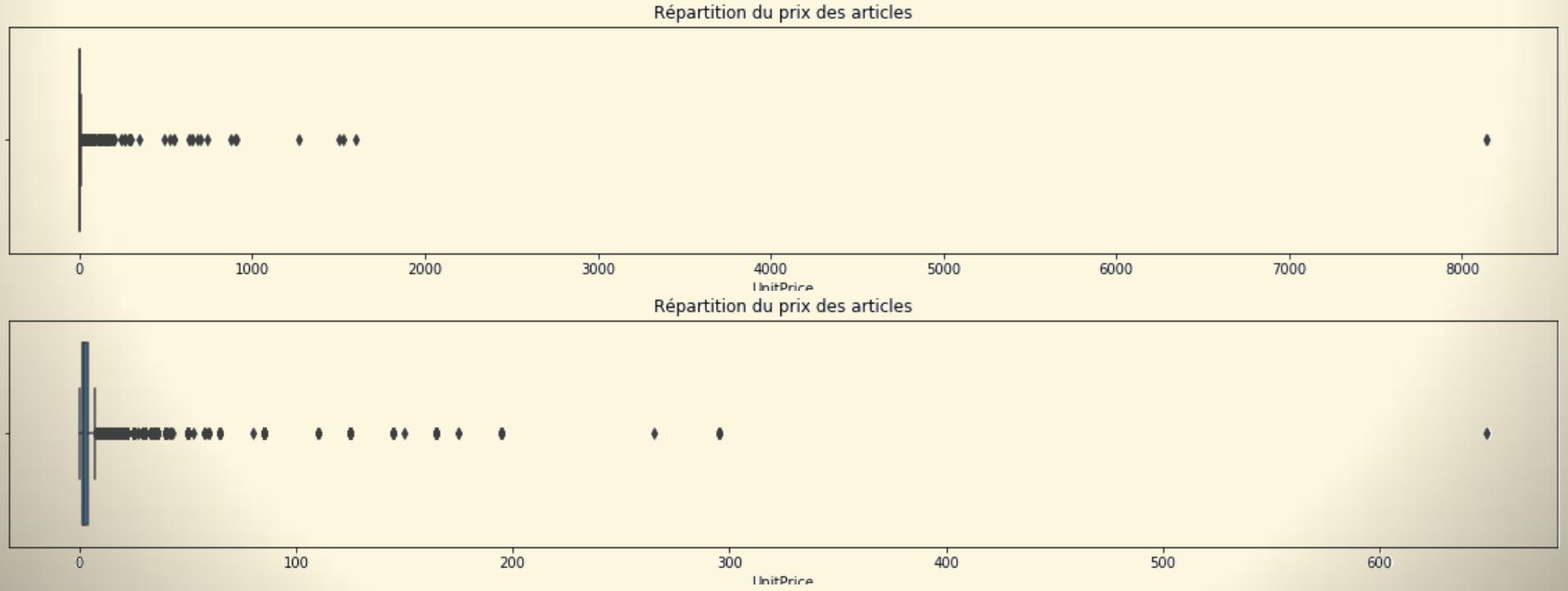
Nettoyage et Exploration

- Pays d'origine
  - 11 % des clients sont étrangers
  - 18 % du CA
  - ➢A conserver
  - Labélisation par Classement



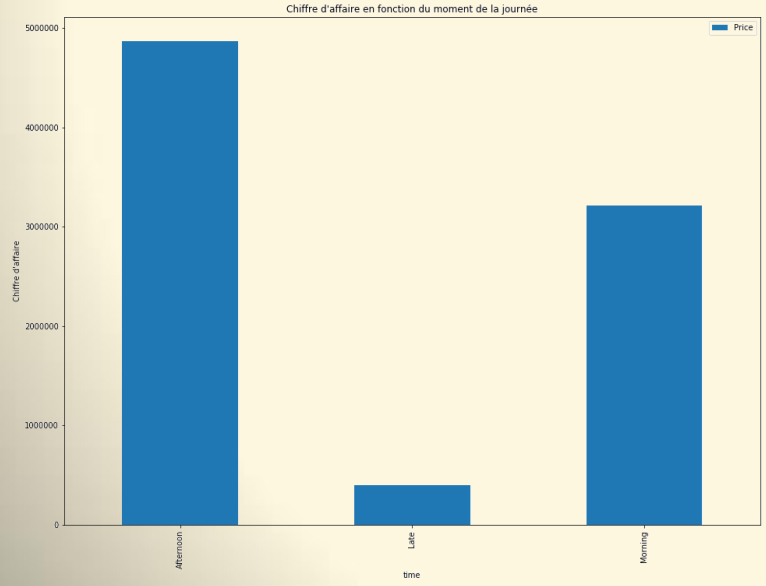
Nettoyage et Exploration

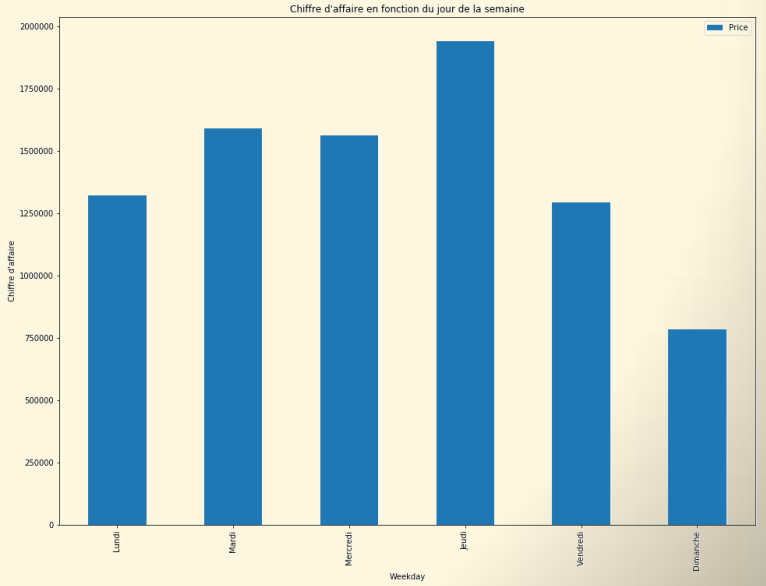
- Quantity & Prix Unitaire
  - Boxplot et exploration des outliers
    - Peu de vrais outliers (certains articles chers ou grosse qté)



Nettoyage et Exploration

- Date
  - Découpage
    - jour de la semaine
    - Moment de la journée





Nettoyage et Exploration

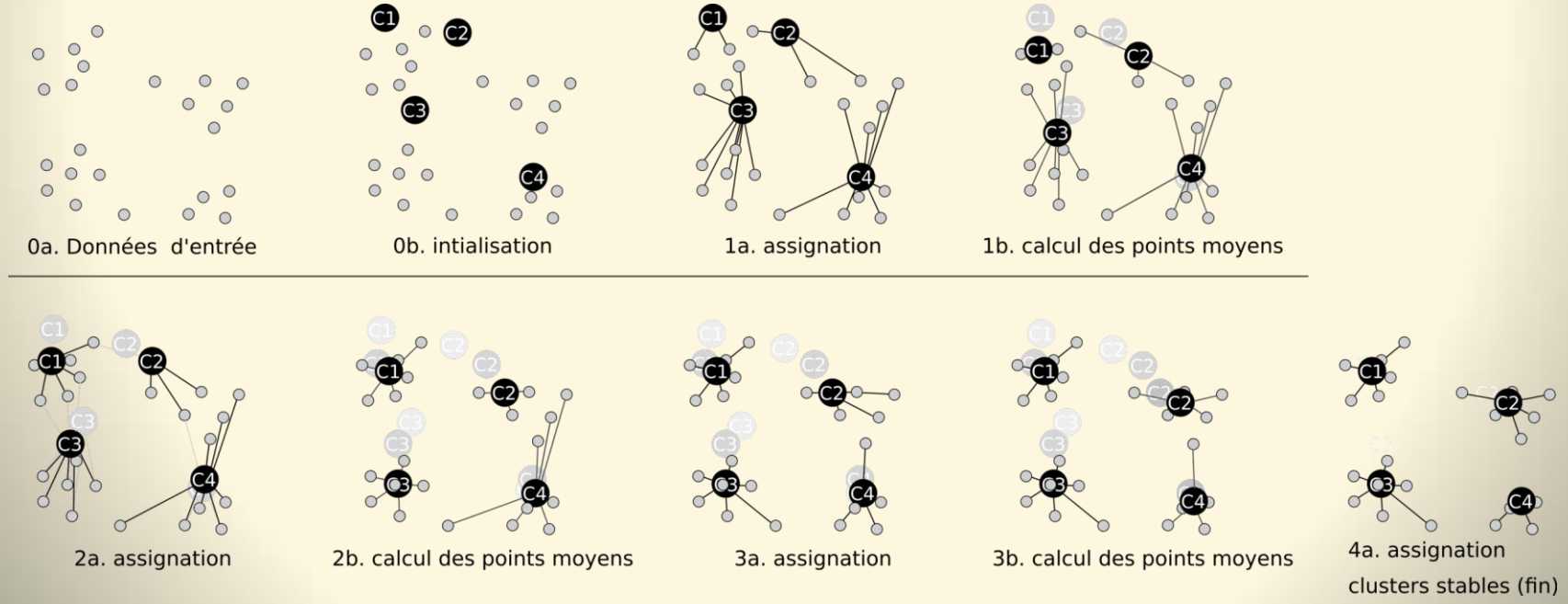
- Description
  - Suppression articles en minuscule (faux articles)
  - Suppression StockCode (POST / DOT)
    - Frais de ports
  - Suppression client Nan (Magasin)

Nettoyage et Exploration

- Description
  - Nouveau Dataset
    - Objet Initial=> Objet cleané
      - **BLUE** RETRO KITCHEN WALL CLOCK => RETRO KITCHEN WALL CLOCK
      - **SET OF 6** RIBB~~ON~~S COUNTRY STYLE => RIBBS COUNTRY STYLE
    - Objet Cleané => keyword
      - 95 % d'apparitions
    - Matrice Objet Initial x Keywords (4000 x 1300)
    - Discount Factor (0,99)
    - Clustering Kmeans

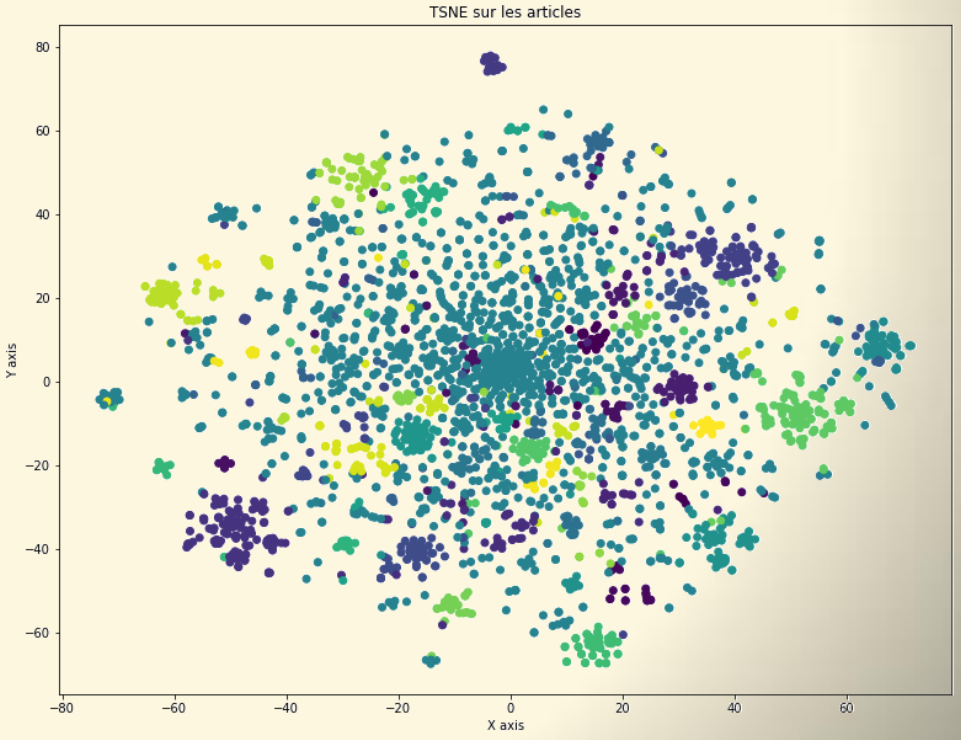
Nettoyage et Exploration

- Présentation du Kmeans (aparté)
  - Minimisation ∑ (distances)²
  - Glisser les moyenne n fois



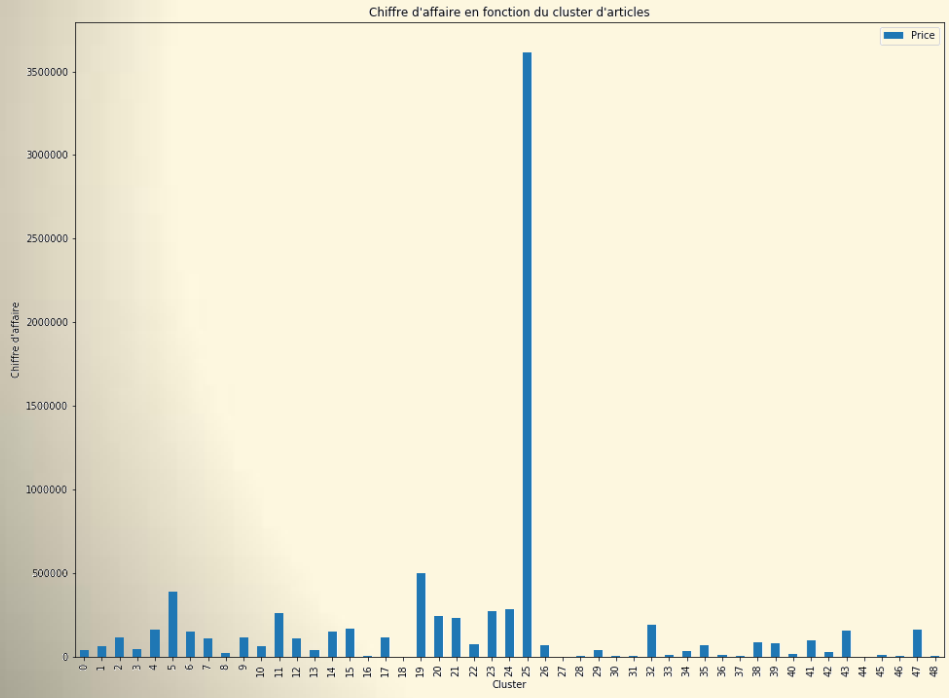
Nettoyage et Exploration

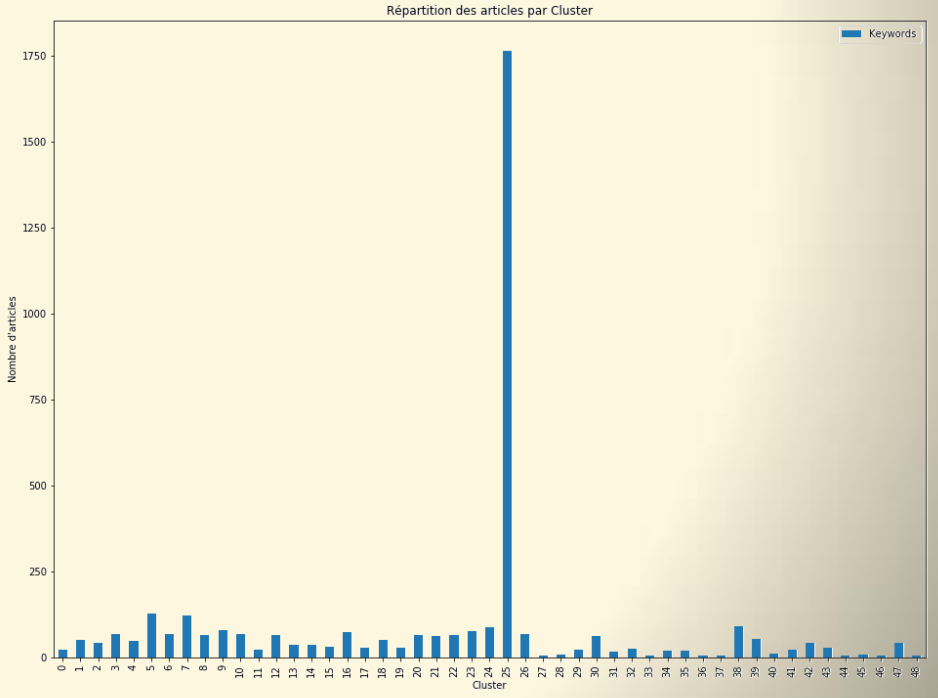
- Description
  - N-cluster basé sur Score de Silhouette
  - Plusieurs essais
    - Avec et sans top keywords
    - Avec et sans Discount factor
  - Faible Score
    - Superposition



Nettoyage et Exploration

- Description





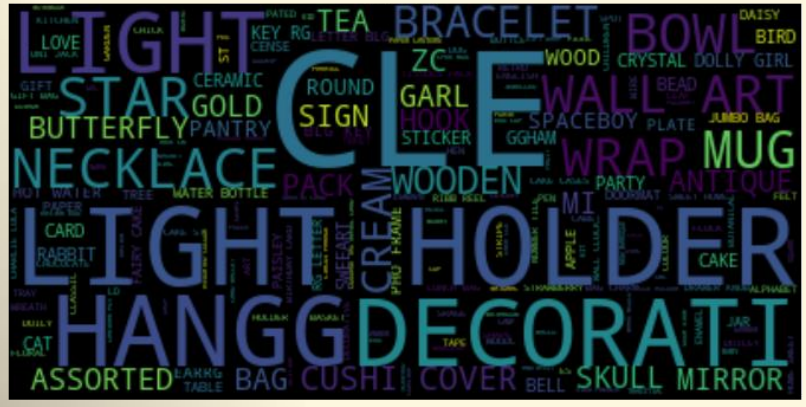
Nettoyage et Exploration

- Description

Cluster 19

Cluster 25

Cluster 27



Cluster 28

Nettoyage et Exploration

- Description
  - Encodage des articles par cluster
  - OneHotEncoding de ces clusters

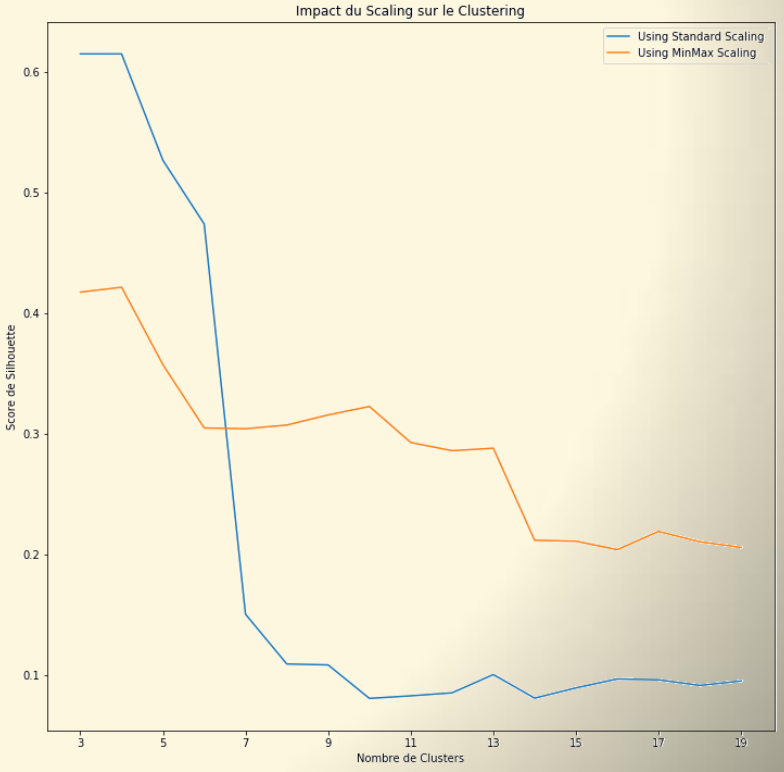
- Fin du nettoyage / Exploration
  - Sauvegarde du dataset
  - Sauvegarde du converter Pays => Label
  - Sauvegarde du converter Objet => Cluster

Clustering des Clients

- Préparation du dataset Clients
  - Phase 1 :
    - Dataset groupé par invoice
    - Ajout de Features
      - Jour depuis achat
      - Prix Panier
      - Nb articles / clusters
  - Phase 2 :
    - Dataset groupé par clients
    - Ajout de features
      - 1ere visite/dernière visite
      - Nb visites
      - Temps moyen entre visites

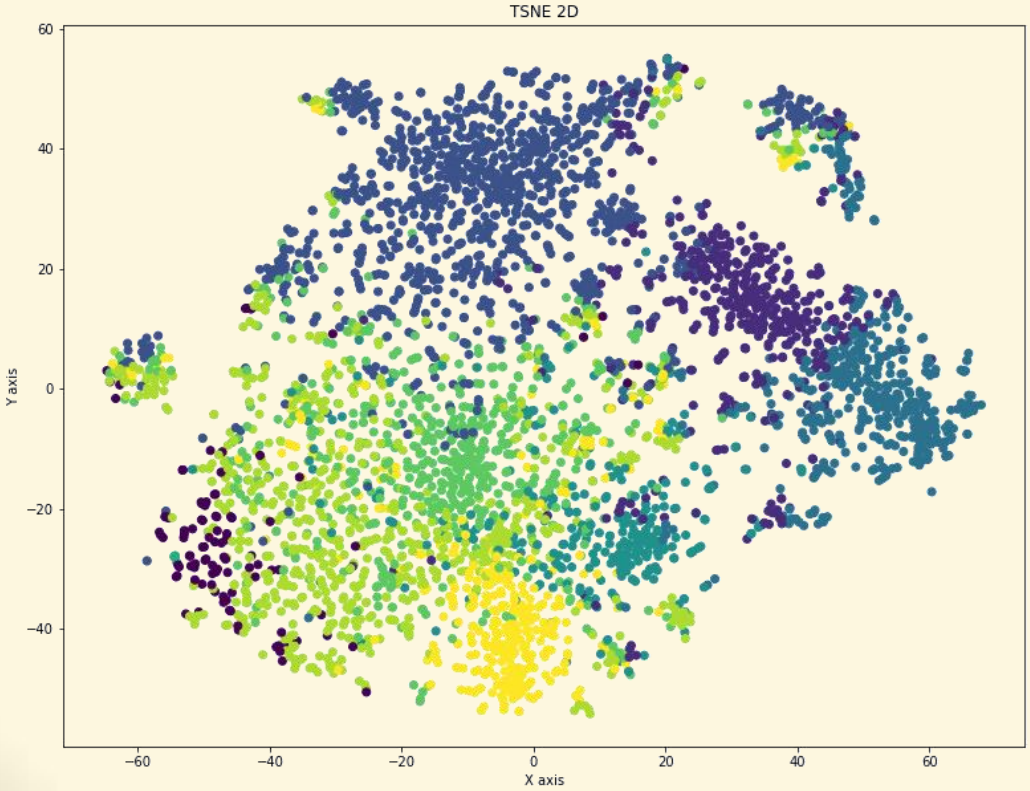
Clustering des Clients

- 2nd Clustering Kmeans
  - N-cluster basé sur Score de Silhouette
  - Scaling Standard et MinMax
    - MinMax meilleur si N cluster > 7
  - Std Scaling trouve les outliers (1 seul gros cluster env 3500 pers.)
  - Garder suffisamment de pers. / cluster

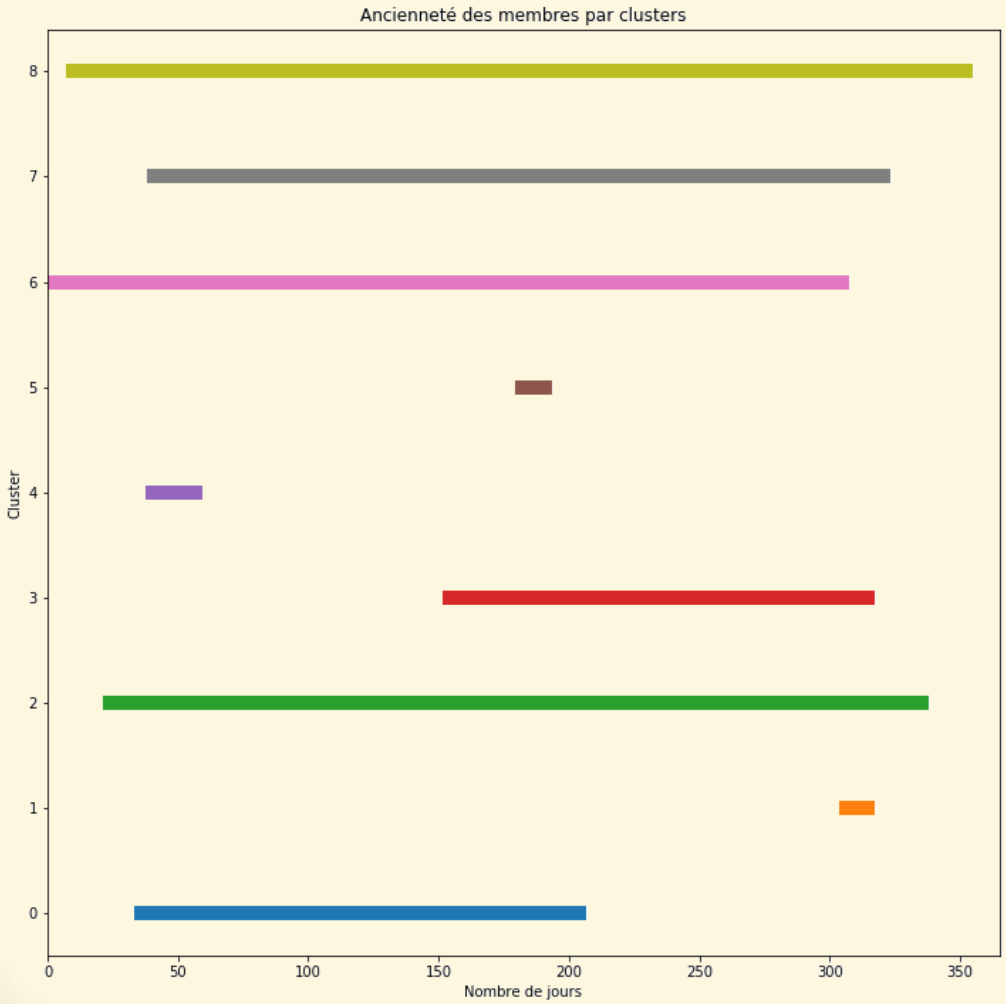


Clustering des Clients

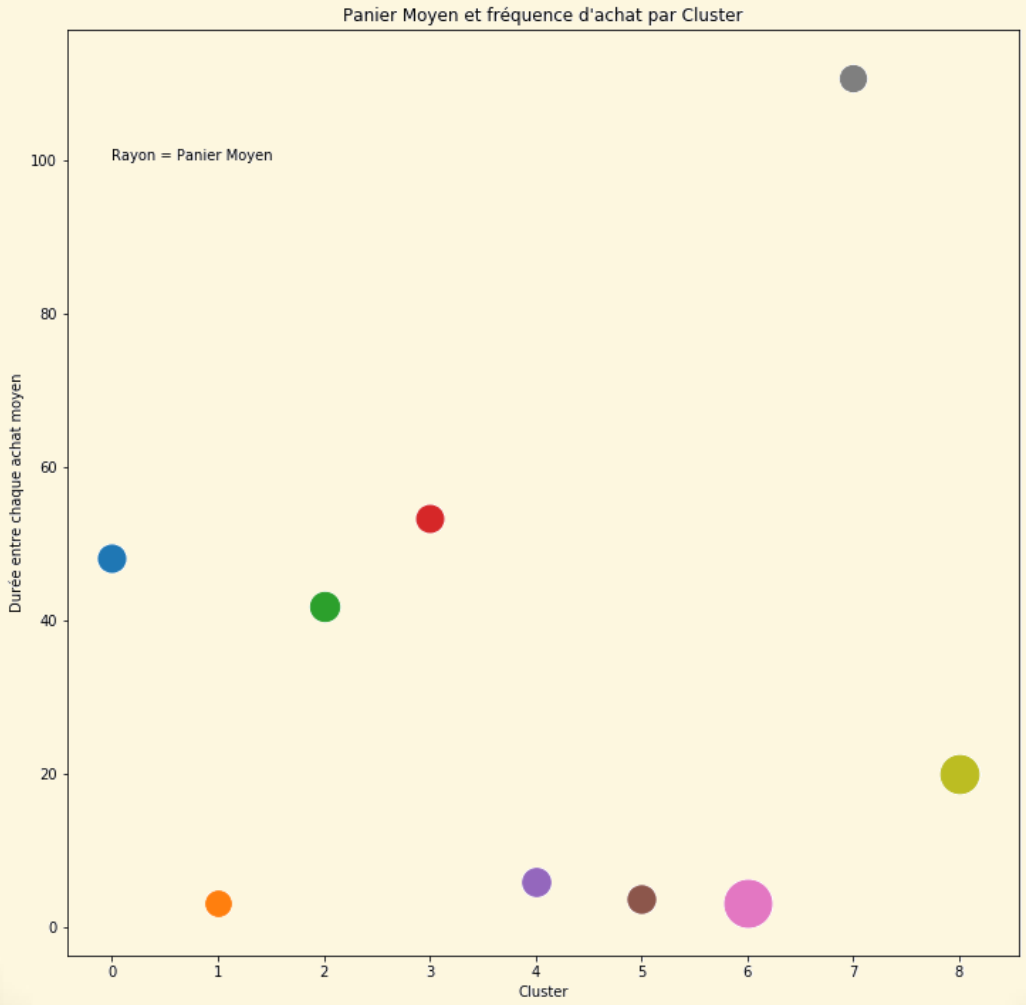
- 2nd Clustering Kmeans
  - Visualisation sur TSNE



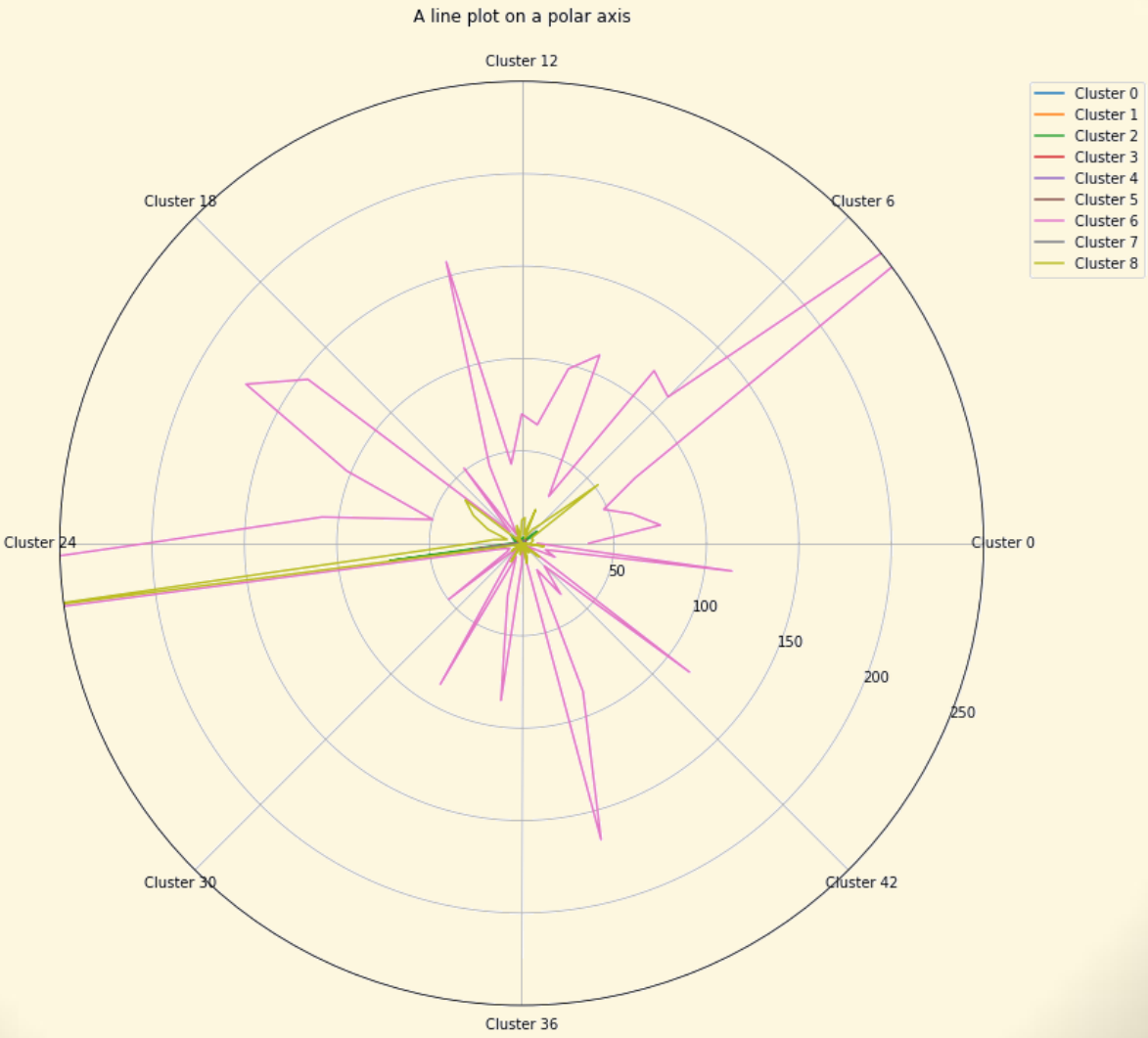
Interprétation des Clusters



Interprétation des Clusters



Interprétation des Clusters

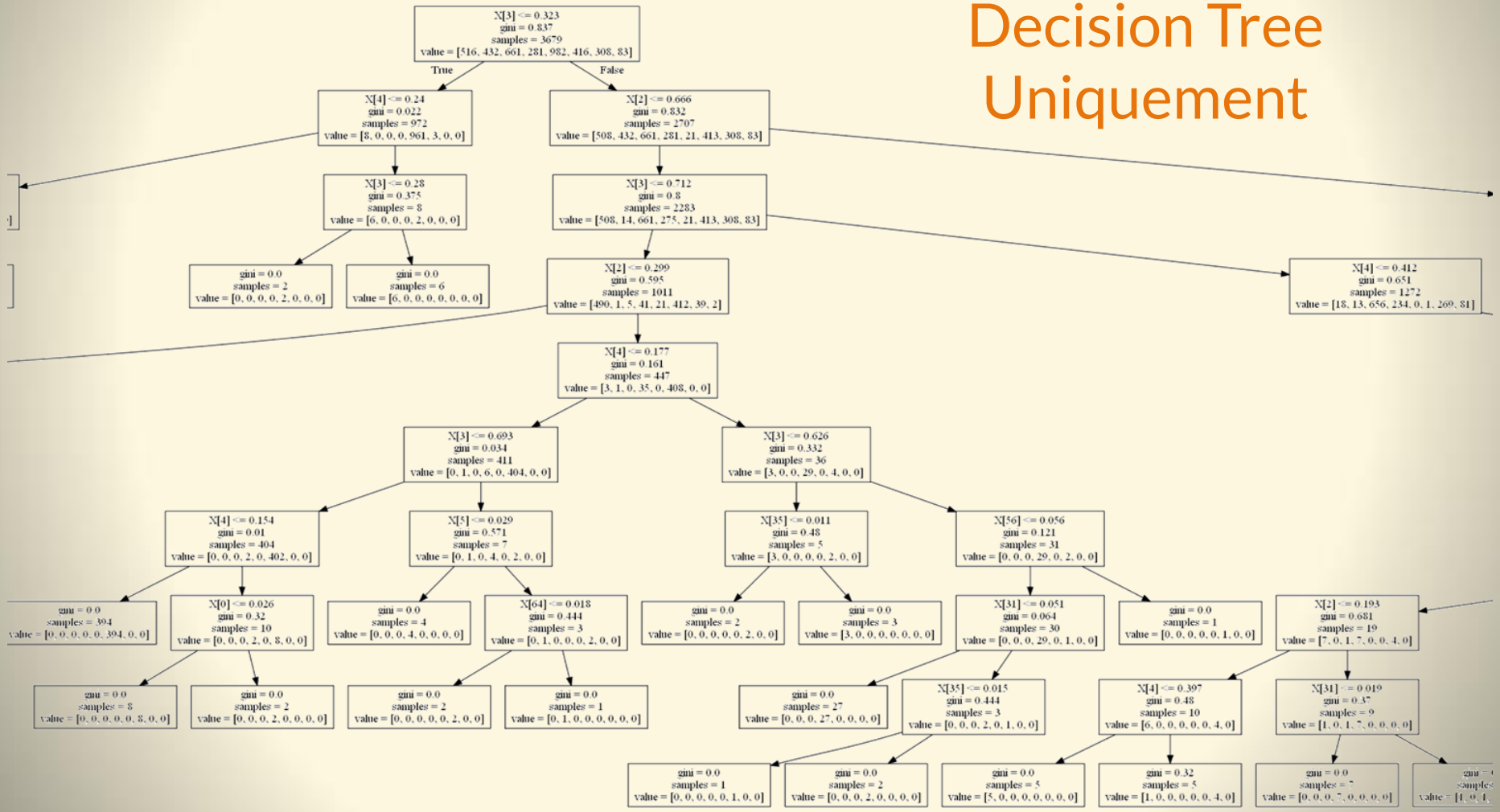


Classification

- Suppression Cluster 6 (4 clients – VIP)
- ~~Standard~~ & MinMax Scaling
- Multiple Modèles testés
  - KNN (Accuracy 93,6 %)
  - SVC (Accuracy 83,6 %)
  - Naive Bayes (Accuracy 46,3%)
  - Decision Tree(Accuracy 94 %)
  - ✓Random Forest(Accuracy 95,6 %)
- Grid Search + export top model fitted

Interprétation des Clusters

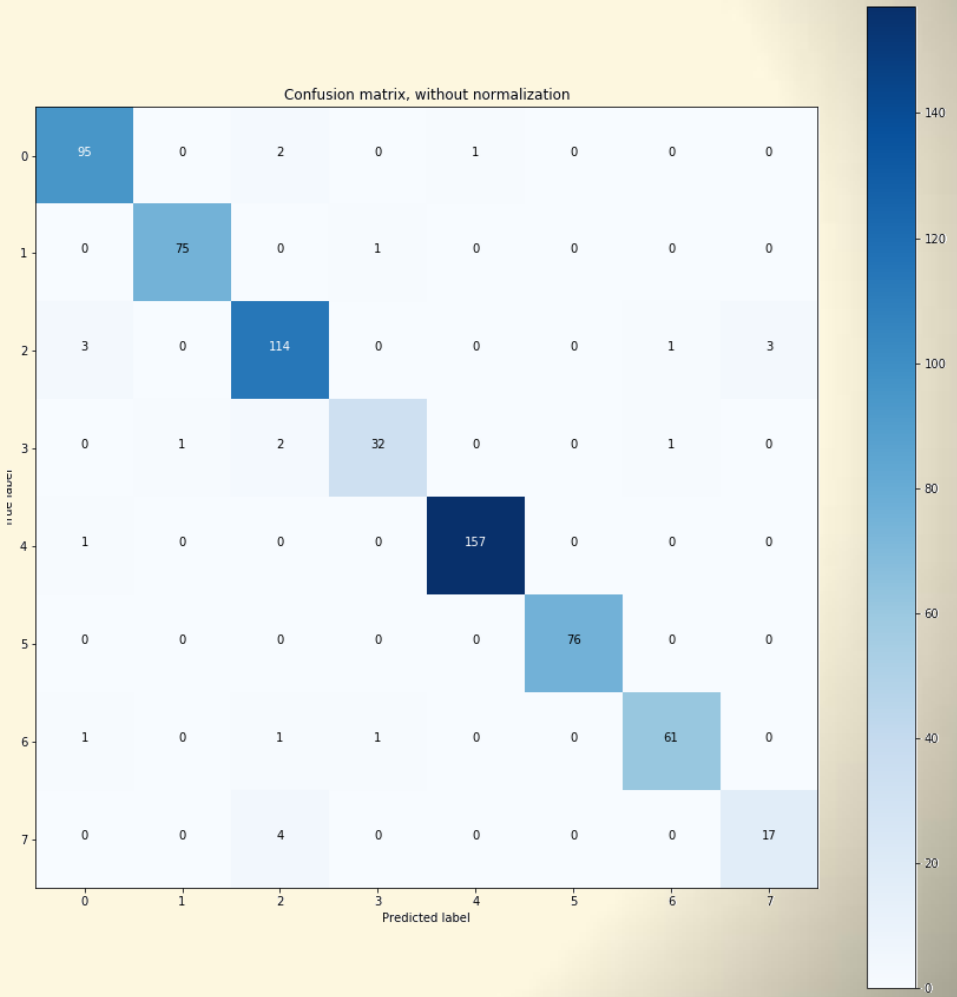
Decision Tree Uniquement



Classification

- Résultats
  - Accuracy : 96,46%
  - Recall : 94,29%
  - Precision : 95,21%

  - FP > FN



API

- Test
  - Extraction du dataset initial
    - 1 Client par Cluster (sauf 6)
    - Sauvegarde sous testN.csv (N=cluster)
  - Code
    - Ouvre chaque dataset
    - Reforme le dataset du training (shape)
    - Charge scaler + modèle fitté => prédiction
  - Runtest.bat
    - Appelle test.py
    - ❖python test.py client.csv …
    - Certains dataset = 2 invoices
    - Peut-être du train set

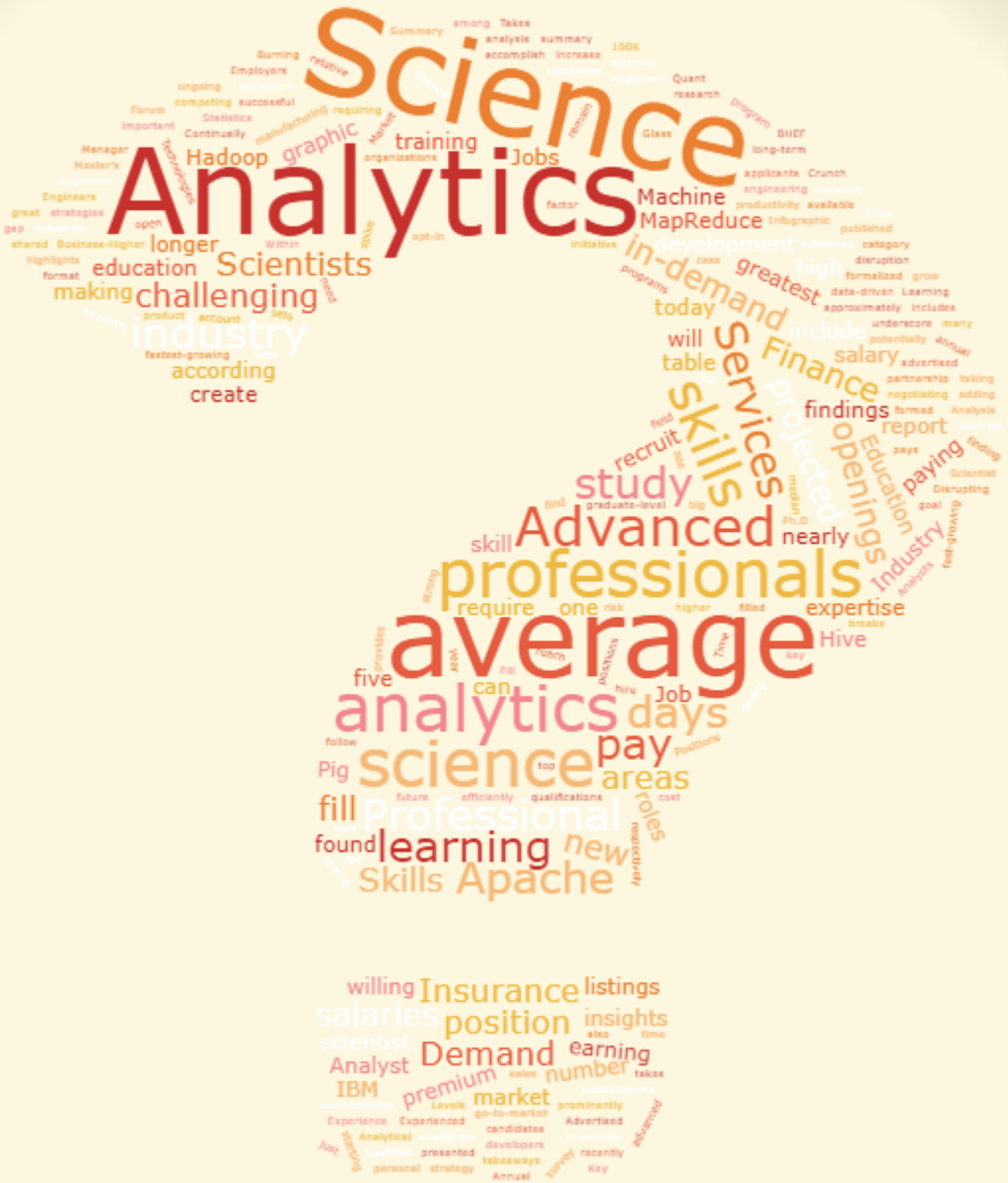
```
Prediction sur test1.csv
Ce client est predit pour appartenir au Groupe 1
Prediction sur test2.csv
Ce client est predit pour appartenir au Groupe 2
Prediction sur test3.csv
Ce client est predit pour appartenir au Groupe 3
Prediction sur test4.csv
Ce client est predit pour appartenir au Groupe 4
Prediction sur test5.csv
Ce client est predit pour appartenir au Groupe 5
Prediction sur test7.csv
Ce client est predit pour appartenir au Groupe 7
Prediction sur test8.csv
Ce client est predit pour appartenir au Groupe 8
```

Pistes d'évolutions

- Le Clustering des objets impacte le Clustering des Clients
  - Meilleur Extraction des articles ?
    - Jonction avec fournisseur par exemple
    - Type de Produits (jouets, soins, décorations, bijoux, …)
  - Moins de Mots-clés pour les articles ?
  - 2nd Clustering par couleur ?
- Convertir comme le RFM (perte d'info mais aide au Clustering)

Conclusion

- On a groupé les Clients dans 8 groupes
- Permet de savoir :
  - Qui ne reviendra potentiellement pas
  - Qui revient fréquemment
  - Qui a tendance à être bon acheteur
- Aide à la décision :
  - Cibler des clients pour des offres
    - Faire revenir des clients perdu avec des promotions ?
  - Cibler les produits/catégories phares
- Risque d'erreur assez faible même avec peu d'achats
- Evolution des clients de certains clusters dans le temps ?
  - Fidélisation ou perte du client


Science
Analytics
Hadoop
graphic
training
Jobs
Machine
MapReduce
longer
education
Scientists
making
challenging
industry
according
create
development
in-demand
greatest
high
today
include
Finance
Services
will
table
salary
skills
findings
report
recruit
study
Advanced
skill
nearly
professionals
Education
Industry
paying
openings
require
one
expertise
Hive
average
five
can
Job
analytics
days
pay
Pig
science
areas
fill
Professional
roles
found
learning
new
Skills
Apache
willing
Insurance
listings
salaries
position
insights
scientist
Demand
earning
Analyst
number
premium
IBM
market