CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

Parcours Data sciences | ABBOUD Marwa | 10 Mars 2021

Encadrant : Bertrand Beaufils

Evaluateur : Pierre-Antoine Ganaye

Présentation de la problématique

- L'entreprise « Place de la marche » lancer un marketplace d'e-commerce en proposant des produits à la vente, souhaite **automatiser l'attribution des catégories aux articles**.

Objectifs de l'étude

- Etudier la **faisabilité de classification** des produits en différentes catégories à partir de la:
  - **Description** du produit
  - **Image** du produit

Source de données

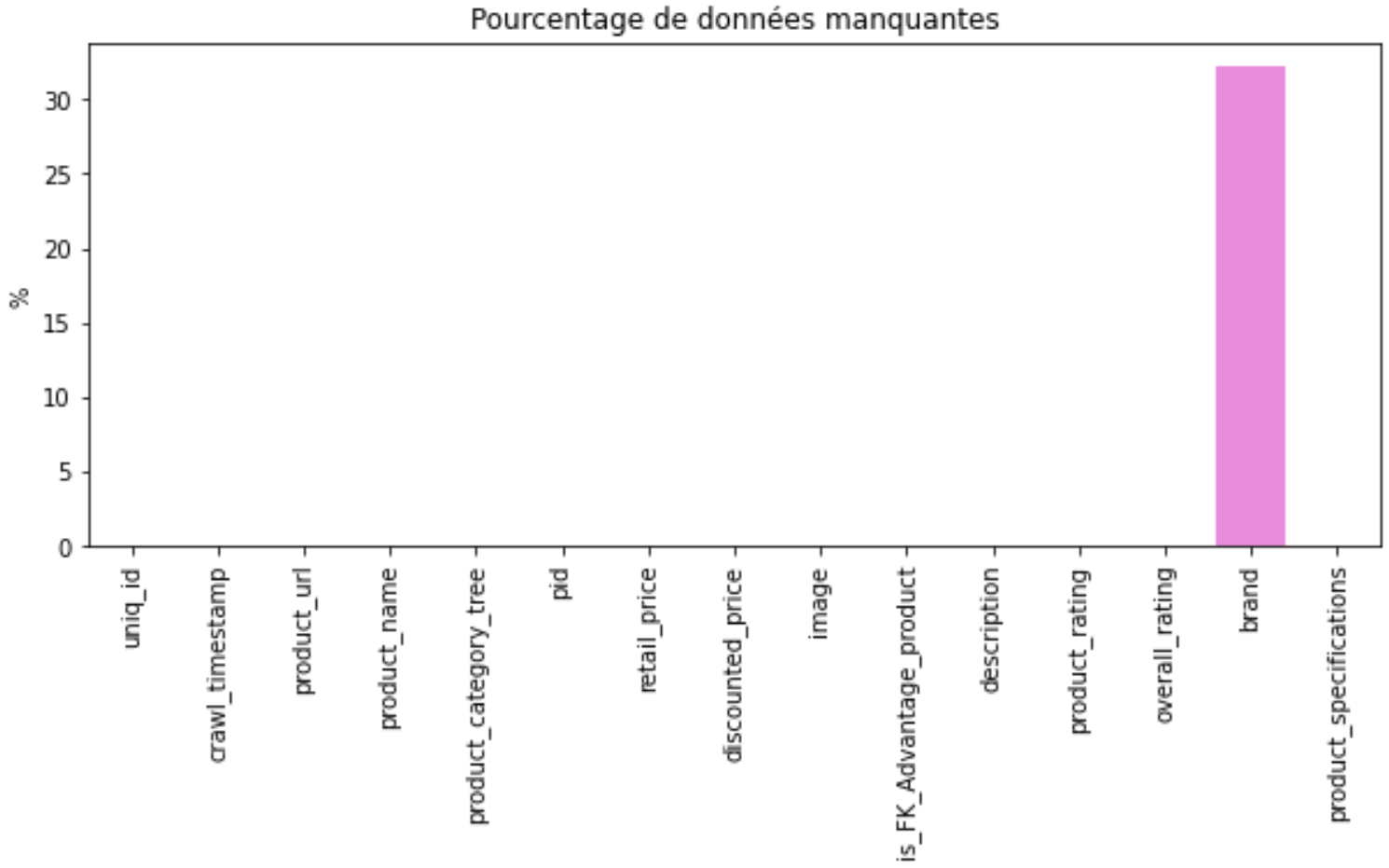
- https://openclassrooms.com/fr/paths/164/projects/631/assignment

SOMMAIRE

PRESENTATION DES DONNEES

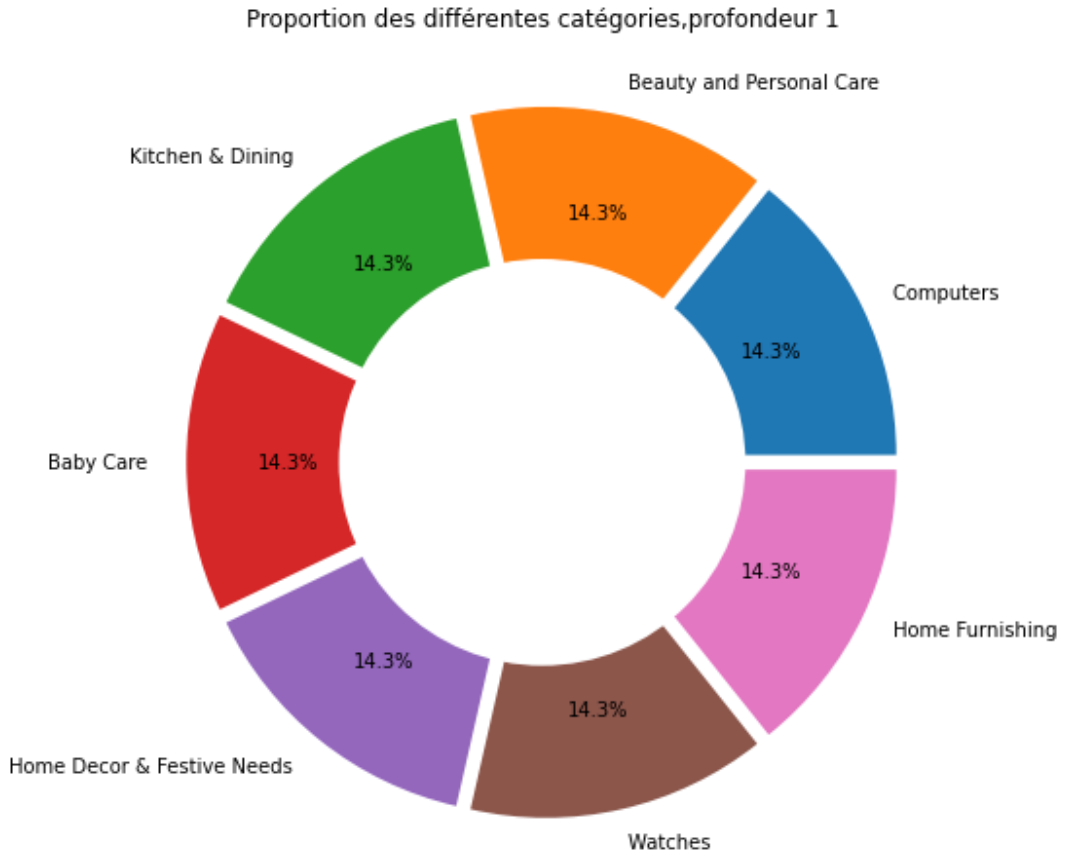
- Base de données composée de 1050 produits

Chaque produit contient une image, une description associée et chaque produit est catégorisée suivant un système d'arbres



DONNEES TEXTUELLES

Target



Première profondeur:
- 7 catégories uniques de produits
- 150 produits par catégories

Possibilité d'avoir des catégories plus fines

DONNEES TEXTUELLES

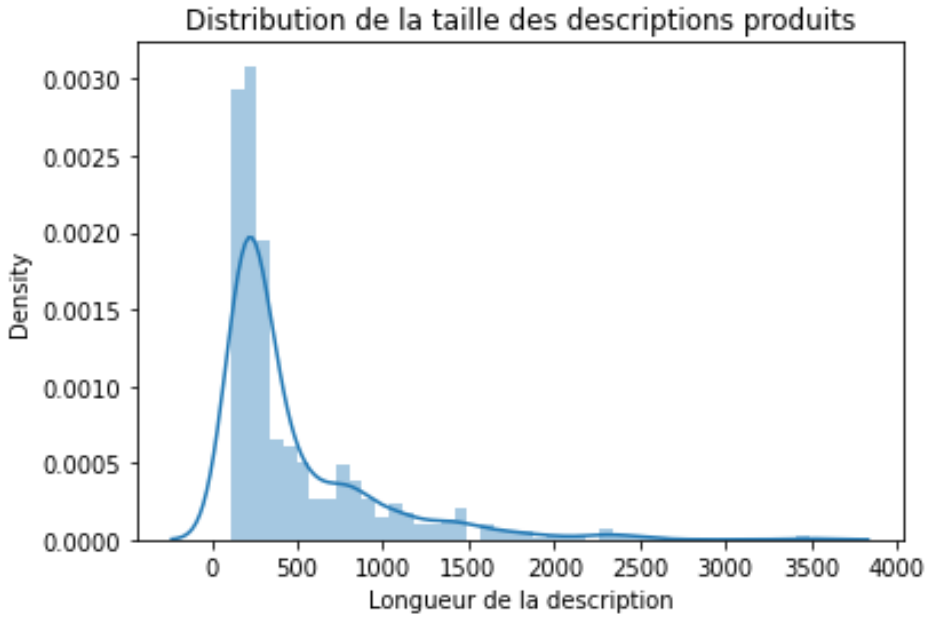
Features	Représentations des descriptions

```
1 df["description"].iloc[6]
```

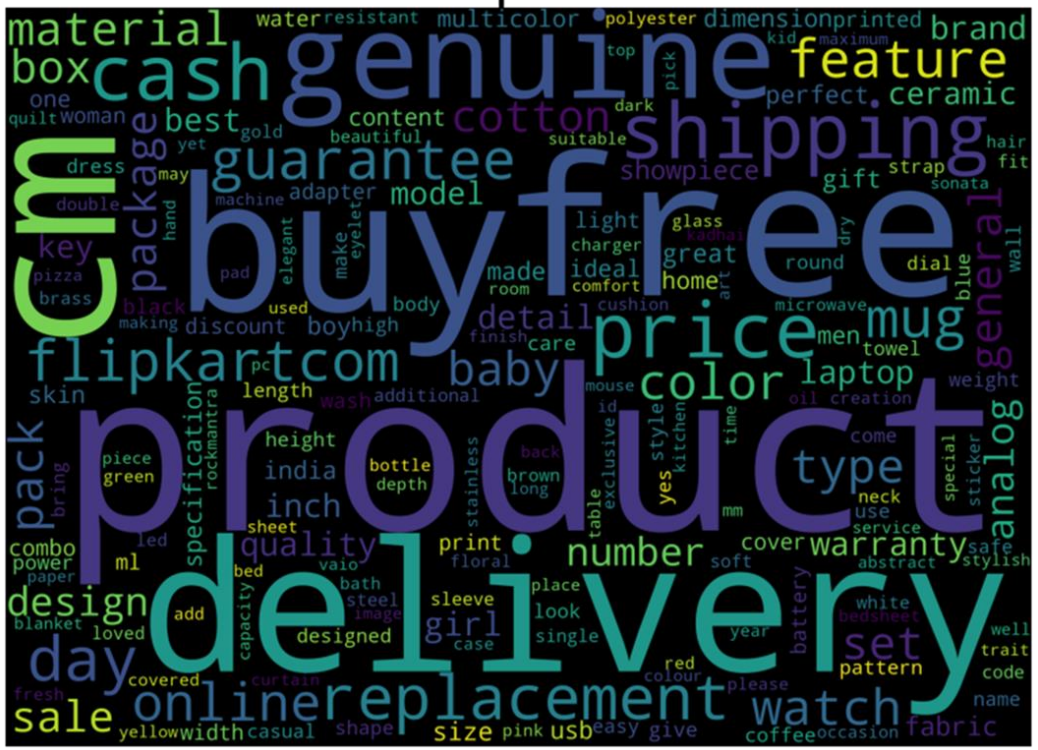
```
'Camerii WM64 Elegance Analog Watch  - For Men, Boys - Buy Camerii WM64 Elegance Analog Watch
eplacement Guarantee, Free Shipping. Cash On Delivery!'
```

```
1 df["description"].iloc[14]
```

```
'Srushti Art Jewelry Megnet_Led_Sport_BlackRed1 Digital Watch  - For Men, Women, Boys, Girls -
ed1 Online at Rs.200 in India Only at Flipkart.com. Led Watch, Sports Led, Megnet watch, Fresh
0 Day Replacement Guarantee, Free Shipping. Cash On Delivery!'
```



Termes les plus utilisées



DONNEES TEXTUELLES

Etapes de preprocessing

Lower Case

Ponctuation

Tokenizer

Stop Word

Lemmatizer

Vectorisation

Transformation BOW et TF-IDF

```
PREPROCESSING
Camerii WM64 Elegance Analog Watch  - For Men, Boys - Buy Camerii WM64 Elegance Analog Watch  - For Men, Boys  WM64 Online at Rs.449 in India Only at Flipkart.com.

LOWERCASE
camerii wm64 elegance analog watch  - for men, boys - buy camerii wm64 elegance analog watch  - for men, boys  wm64 online at rs.449 in india only at flipkart.com.

REMOVE NUMBER
camerii wm elegance analog watch  - for men, boys - buy camerii wm elegance analog watch  - for men, boys  wm online at rs. in india only at flipkart.com. - great

 TOKENIZER
['camerii', 'wm', 'elegance', 'analog', 'watch', 'for', 'men', 'boys', 'buy', 'camerii', 'wm', 'elegance', 'analog', 'watch', 'for', 'men', 'boys', 'wm', 'online',

STOPWORDS
['camerii', 'wm', 'elegance', 'analog', 'watch', 'men', 'boys', 'buy', 'camerii', 'wm', 'elegance', 'analog', 'watch', 'men', 'boys', 'wm', 'online', 'rs', 'india',

LEMMATISATION
['camerii', 'wm', 'elegance', 'analog', 'watch', 'men', 'boy', 'buy', 'camerii', 'wm', 'elegance', 'analog', 'watch', 'men', 'boy', 'wm', 'online', 'r', 'india',
```

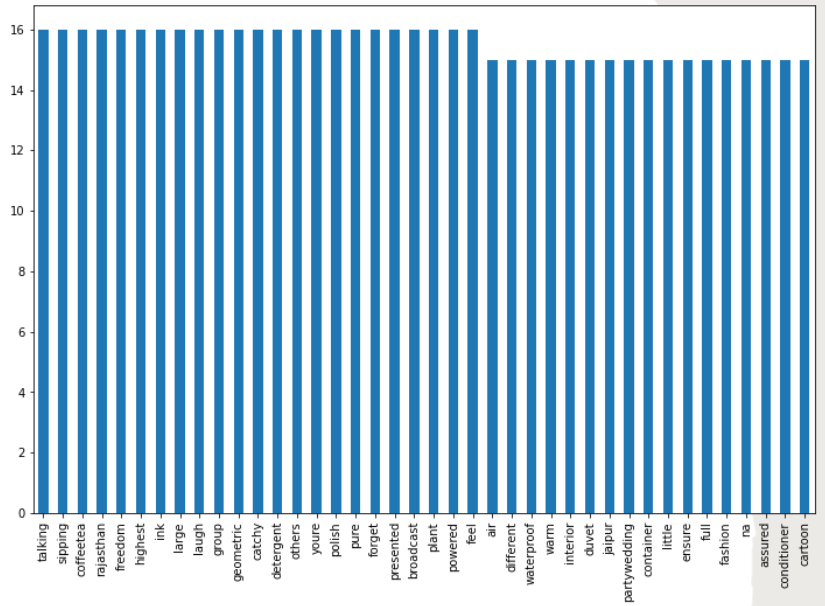
DONNEES TEXTUELLES

Bag of Words **BOW**

- Création d'un Bag Of Words (TF) avec :

```
CountVectorizer(max_df=0.8, min_df=0.2,max_features=200 , stop_words=stopwords.words('english'))
```

Fréquence d'occurrence des mots



Term Frequency Inverse Document Frequency **TFIDF**

- Création d'un TF-IDF avec:

```
TfidfVectorizer(max_features=200,min_df=1,max_df=0.8,stop_words=stopwords.words('english'),ngram_range=(1,2))
```

mot	tfidf
ceramic mug	0.470415
mug	0.466275
ceramic	0.460302
delivery genuine	0.177603
price free	0.177603
best price	0.176312
best	0.166811
online	0.129228

Ignorer les termes trop présents (max_df) et peu présents (min_df).

Garder les mots qui apparaissent plus que deux fois

→ BOW et TF IDF avec une Classification
Supervisée : Gradient Boosting
Non supervisée : NMF et LDA

DONNEES TEXTUELLES

Topic modeling : Latent Dirichlet Allocation **LDA** et Non-Negative Matrix Factorization **NMF** → Modèles non supervisés

Affichage des 10 mots les plus importants de chaque topic

LDA

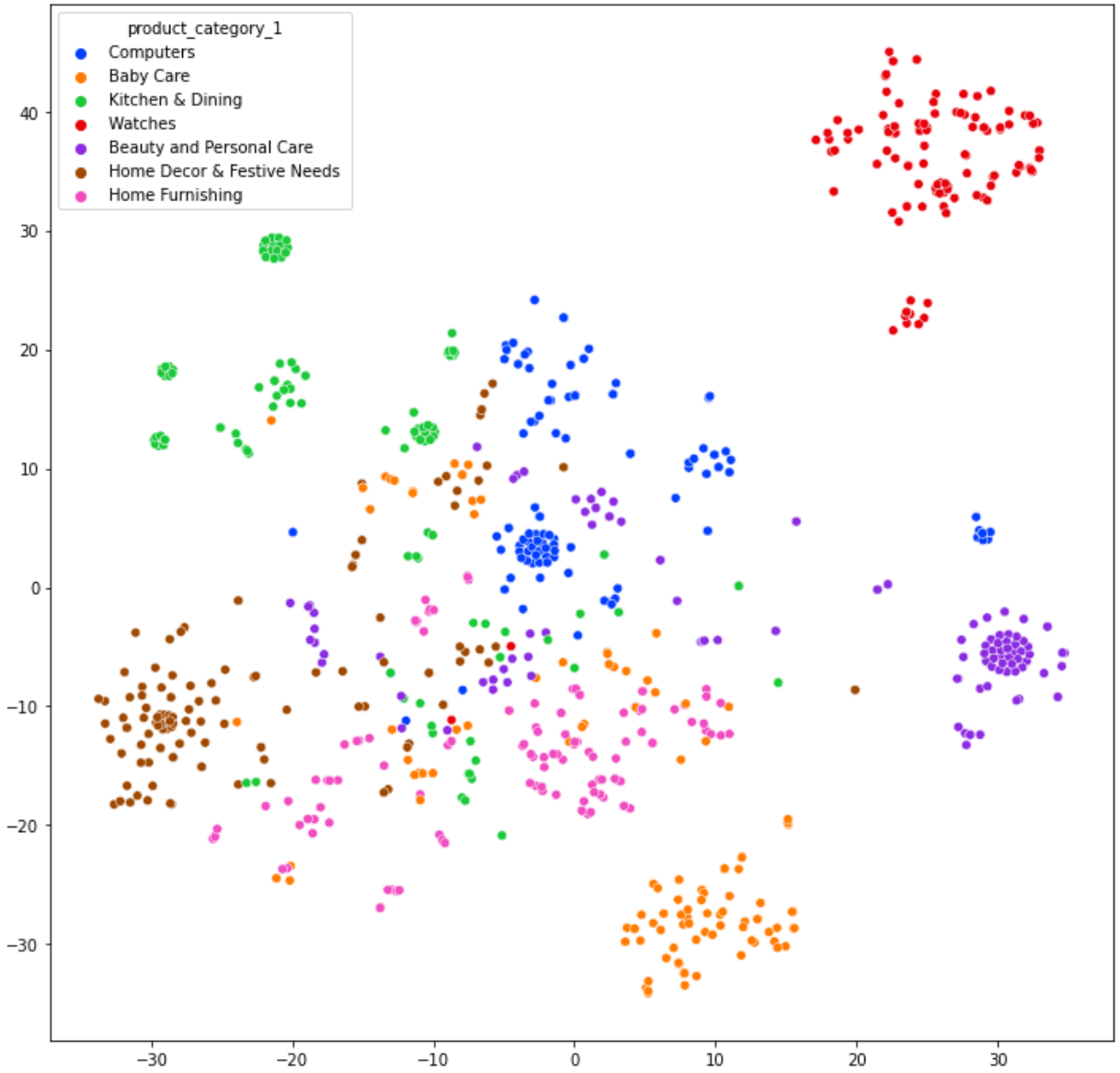
```
Topic 0:
mug ceramic gift home one coffee perfect price art add

Topic 1:
cm color pack package cotton feature box sale number general

Topic 2:
mug design usb paper bring perfect eyelet coffee curtain designed

Topic 3:
baby detail girl set boy dress usb ceramic sleeve mug

Topic 4:
skin set laptop towel type light color ml feature price

Topic 5:
free delivery cash genuine shipping buy product day guarantee replacement

Topic 6:
product warranty laptop adapter battery quality power replacement charger please
```

NMF

```
Topic 0:
flipkartcom guarantee replacement day combo set online lowest towel range

Topic 1:
watch analog men discount india great woman online flipkartcom guarantee

Topic 2:
cm pack inch box color model feature cover number material

Topic 3:
showpiece cm best online guarantee replacement day handicraft statue buddha

Topic 4:
mug ceramic perfect coffee rockmantra one gift loved prithish safe

Topic 5:
baby girl detail fabric cotton dress boy sleeve neck pattern

Topic 6:
usb laptop battery cell led light hp power lapguard replacement
```

DONNEES TEXTUELLES

Représentation via TNSE | BOW



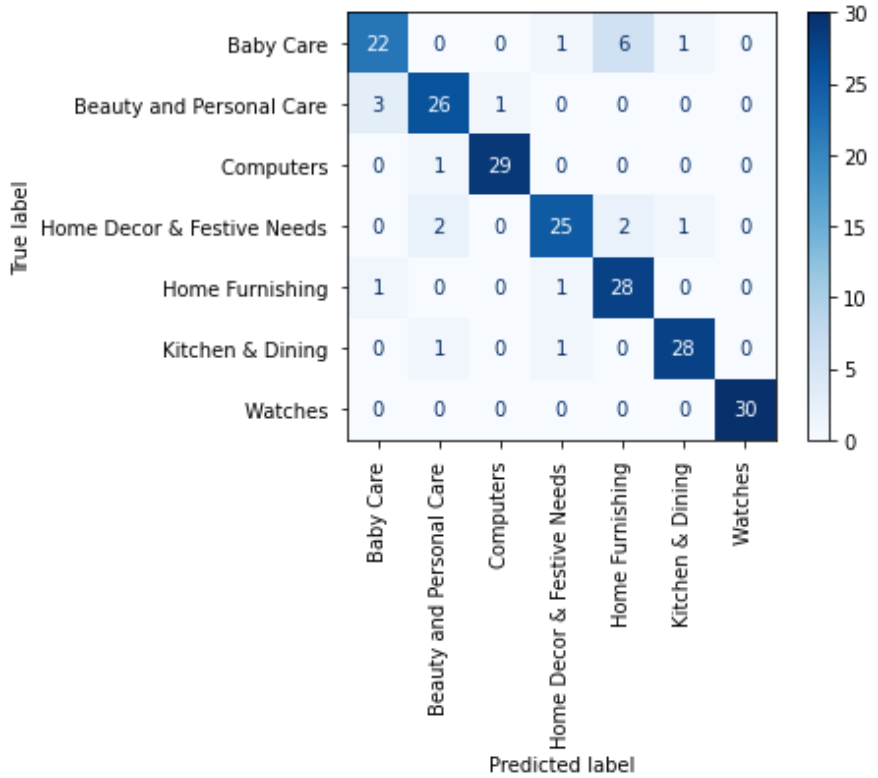
DONNEES TEXTUELLES

Classification Supervisée → XGBoosting

- Recherche sur grille des paramètres
  - Hyperparamètres de preprocessing: max_df,min_df et max_features
  - Hyperparamètres de XGBoosting: n_estimators , learning-rate,
- Mesure la qualité de Classification : **accuracy score**

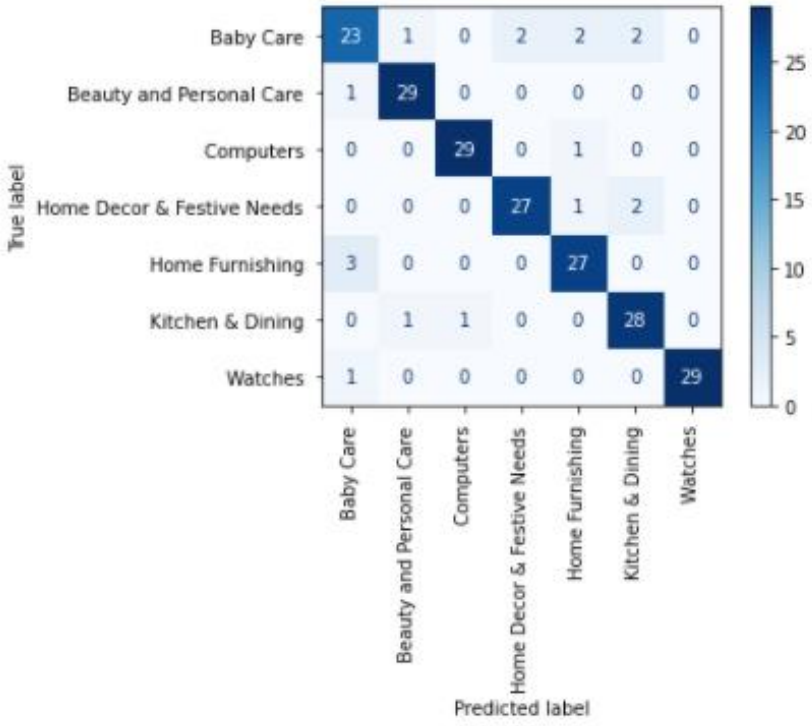
BOW

accuracy sur jeu test : 0.8952380952380953



TFIDF

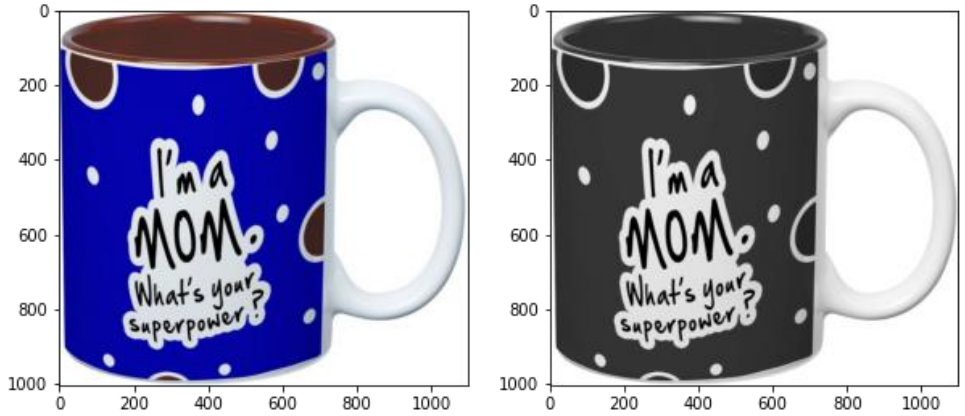
accuracy sur jeu test : 0.9142857142857143



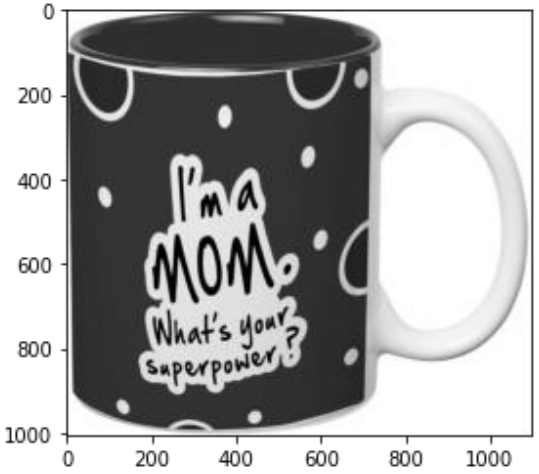
DONNEES VISUELLES

Etape de preprocessing

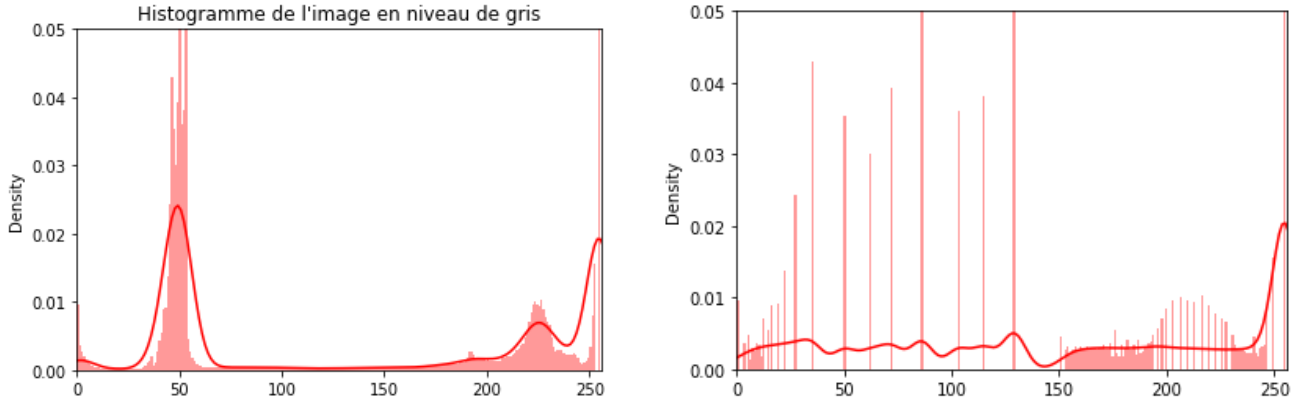
Chargement des images en niveau de gris



Application d'un filtre Gaussien



Contraste en normalisant l'histogramme de chaque image

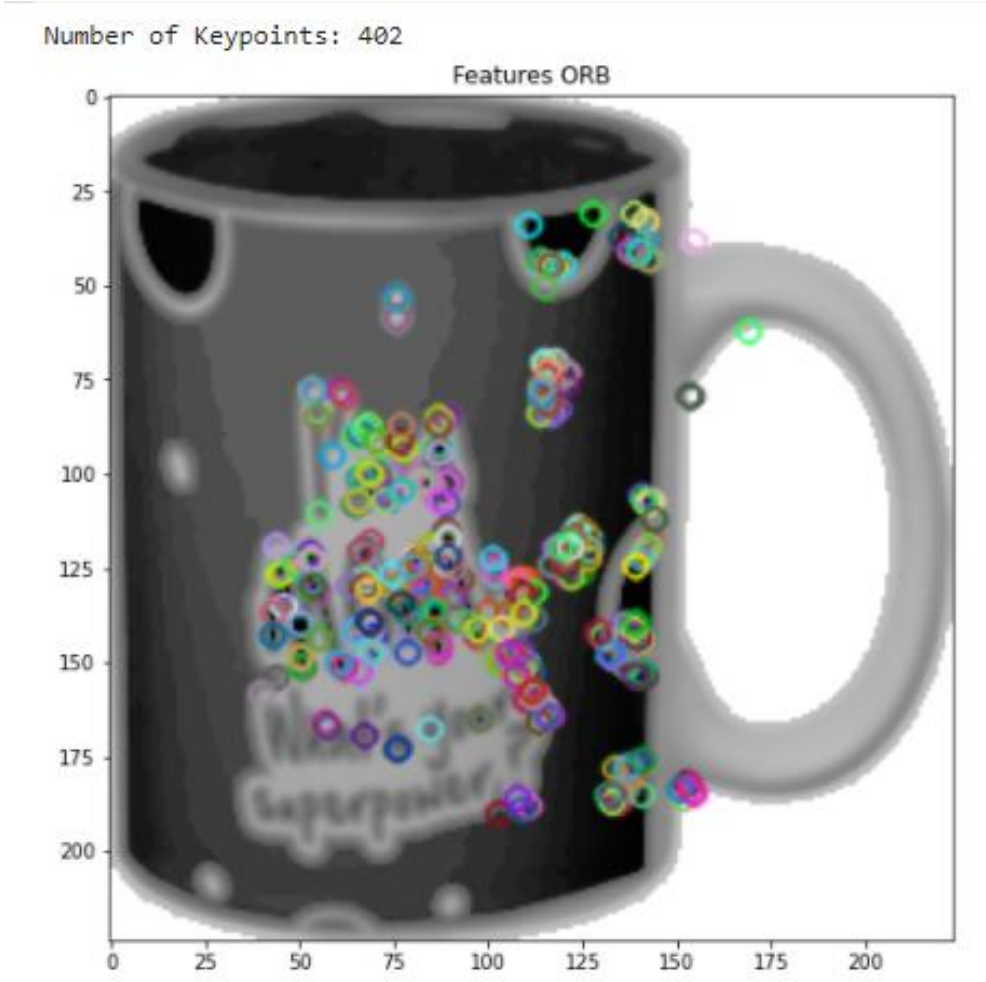


Redimensionné l'image (224,224)

DONNEES VISUELLES

Extraction de features de l'image par la méthode ORB

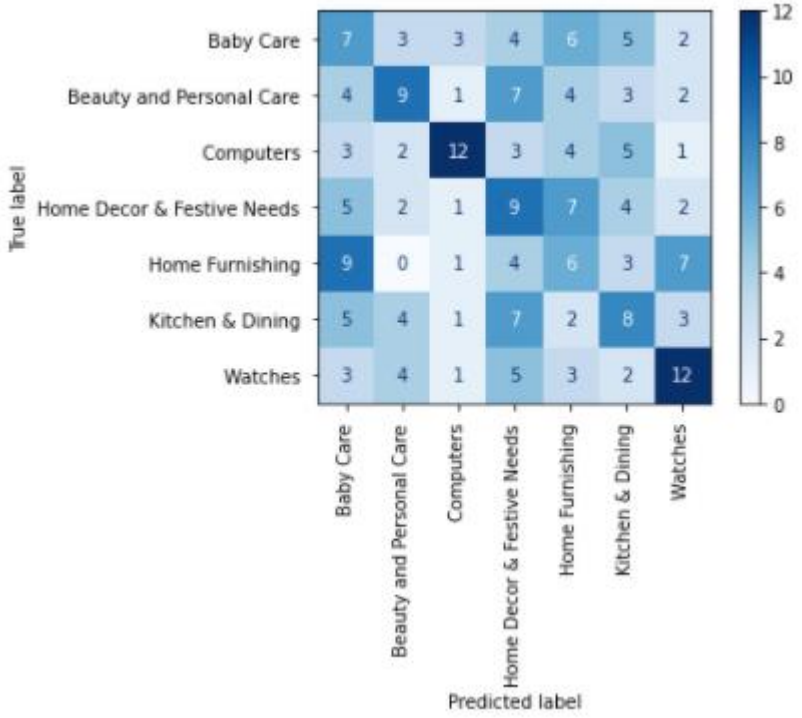
- Calcul des descripteurs sur l'ensemble des images
- Kmeans sur l'ensemble de descripteurs (pour créer des catégories de descripteurs)
- Création de Bag of Visual Words pour chaque image
- Classification de BoVW avec un Gradient Boosting



DONNEES VISUELLES

Classification Supervisée

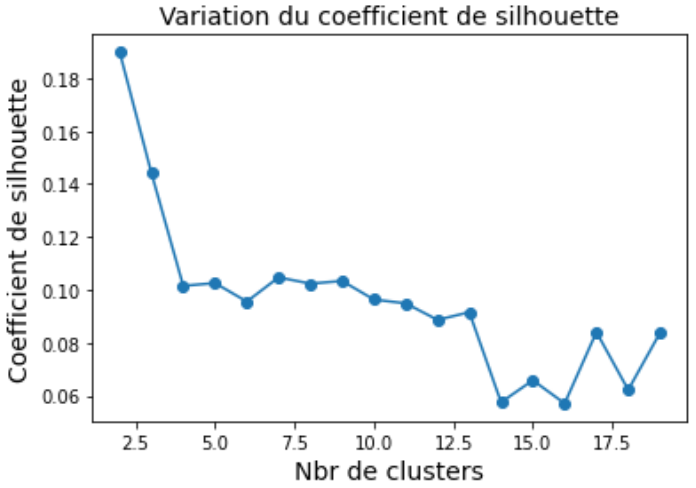
accuracy sur jeu test : 0.3



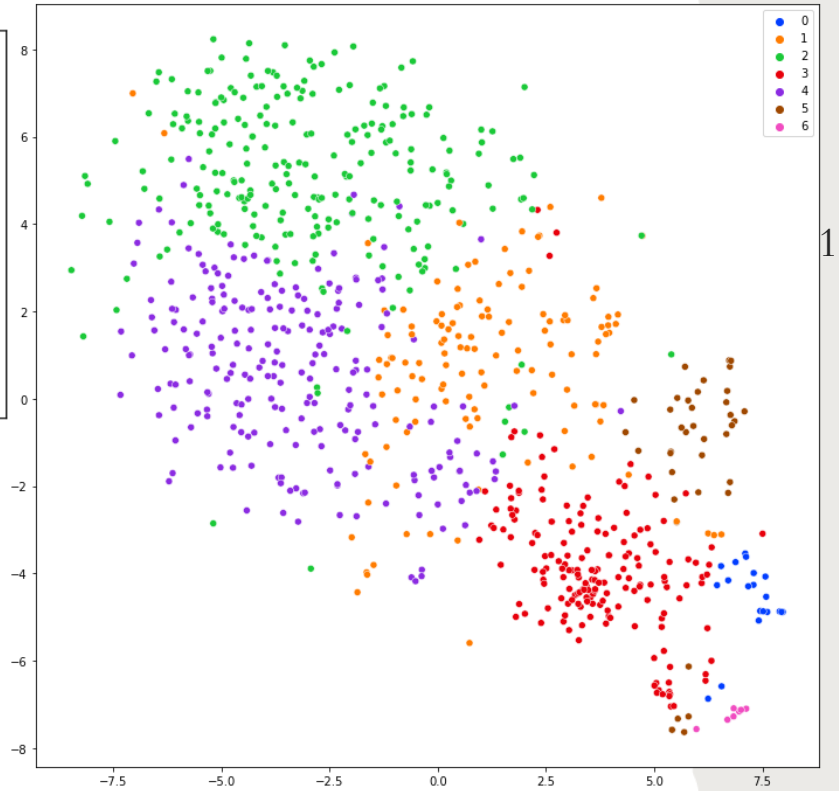
Classification Non supervisée

Mesure de qualité :

Coef silhouette : 0.11
Score Adjusted Rand index : 0.03



	labels	nbr_individus
0	0	21
1	1	127
2	2	241
3	3	179
4	4	226
5	5	39
6	6	7



DONNEES VISUELLES

Convolutional Neural Network (CNN) | VGG16

- Modèle CNN entraîné sur 14 millions d'images provenant du site ImageNet pour une classification à 1000 classes

```
model = Sequential()
vgg= VGG16(weights="imagenet",include_top=True)
```

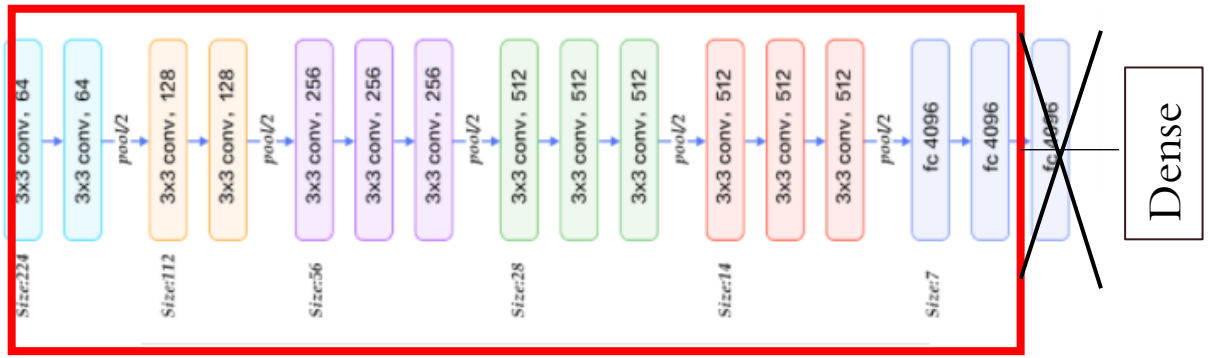
- Pour du Transfer learning, on supprime uniquement la dernière couche propre à la problématique initiale pour en recréer une adaptée à la nôtre

```
for i in range(len(vgg.layers)-1) :
    vgg.layers[i].trainable = False
    model.add(vgg.layers[i])
```


- Ajout de la couche Dense de prédiction adaptée à notre problème de classifications (7 classes)

```
model.add(Dense(7, activation='softmax'))
```

- Réentraîner la dernière couche avec les images preprocessing



Architecture de VGG-16

→ Effectuer une classification supervisée

```
Train Score (accuracy):  0.9964285714285714
```

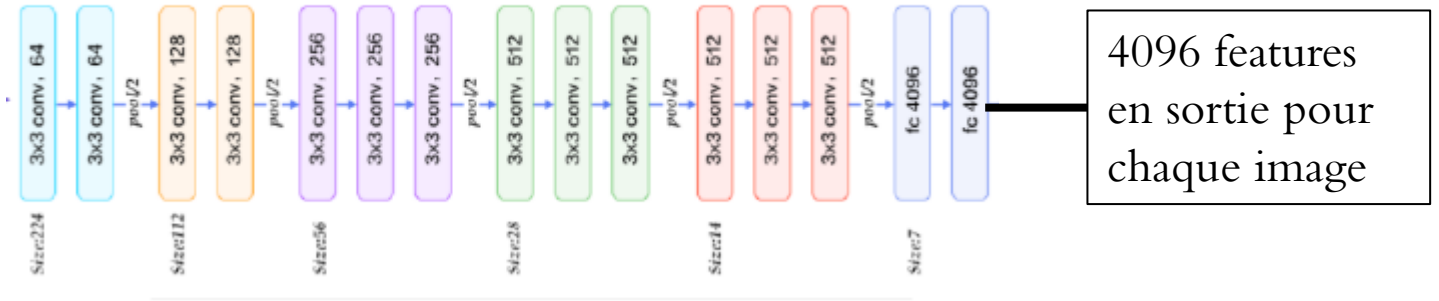
DONNEES VISUELLES

Convolutional Neural Network (CNN) | VGG16

- Pour l'extraction des features

  - Supprime la dernière couche

  - Prédire les 4096 features de sorties associées à chaque image de notre base de données



4096 features en sortie pour chaque image

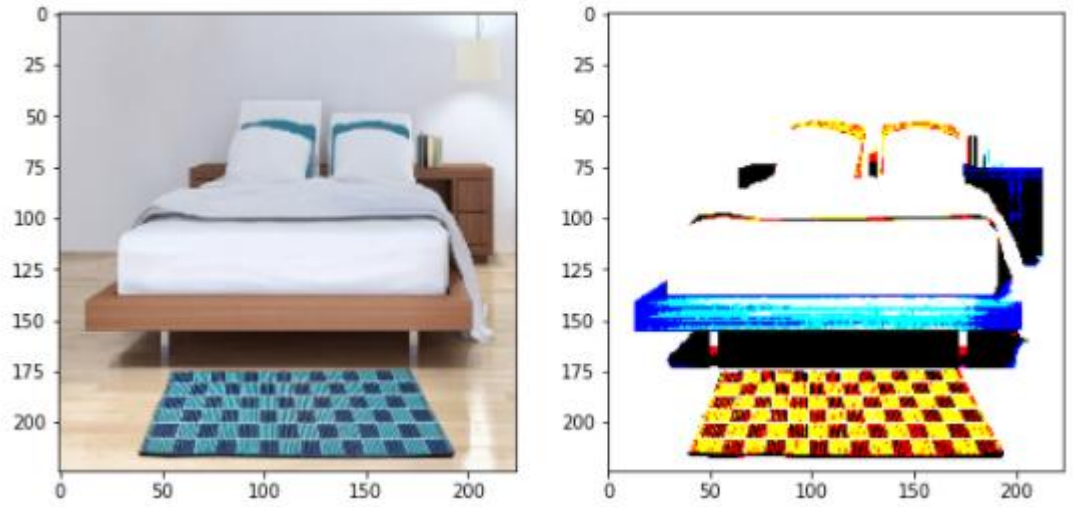
DONNEES VISUELLES

Convolutional Neural Network (CNN) | ResNet50

- Extraction des features par réseau de neurone convolutionnel ResNet50 pré-entraîné sur Image Net

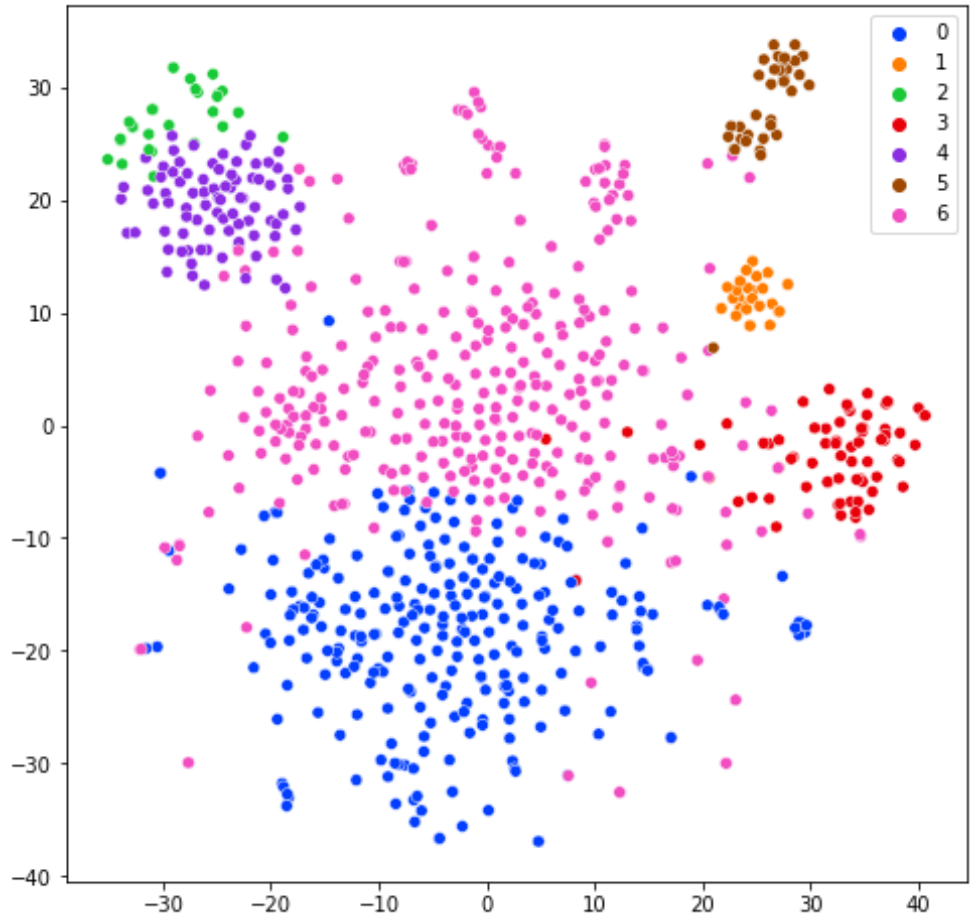
```
model = ResNet50(weights='imagenet', include_top=False)# include_top : remove the last layer
```

Avant et après le preprocessing de Resnet50

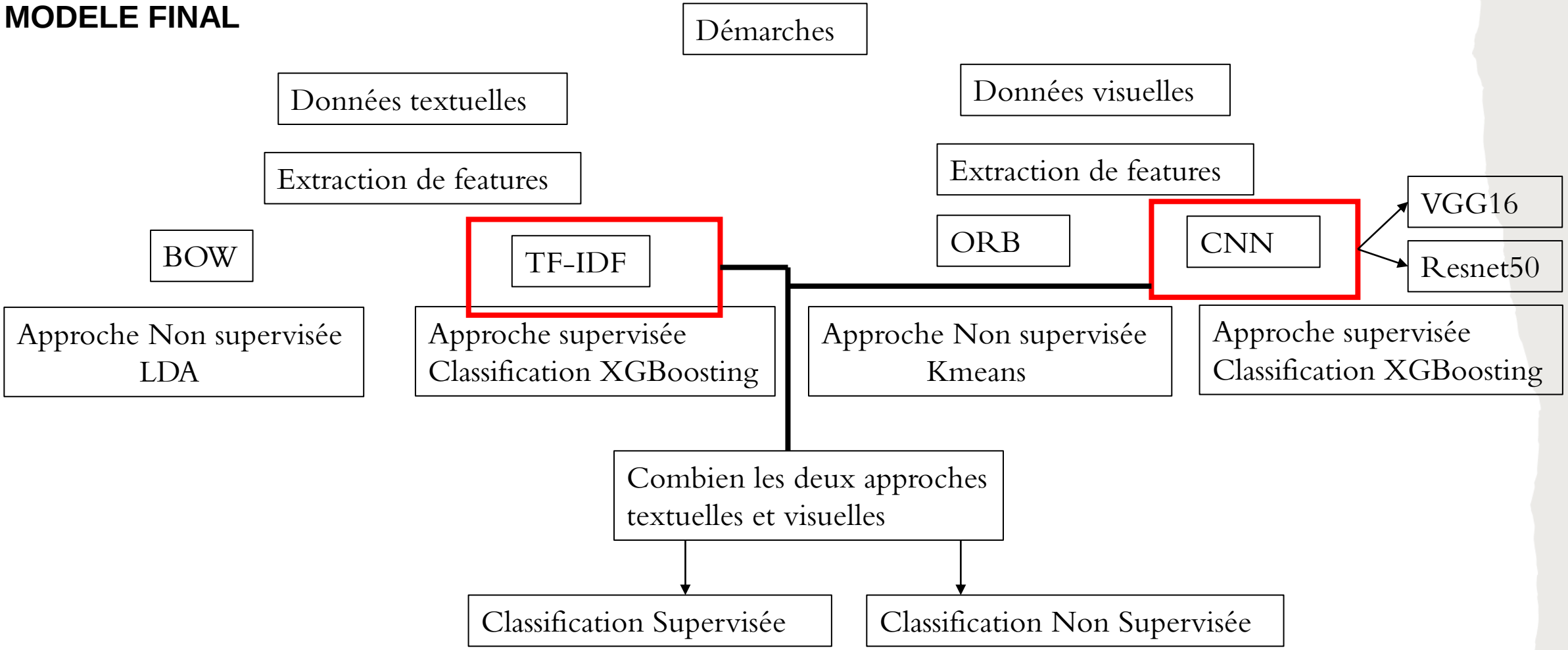


- Effectuer une classification non supervisée

```
Score Adjusted Rand index : 0.3262405727759149
```



MODELE FINAL

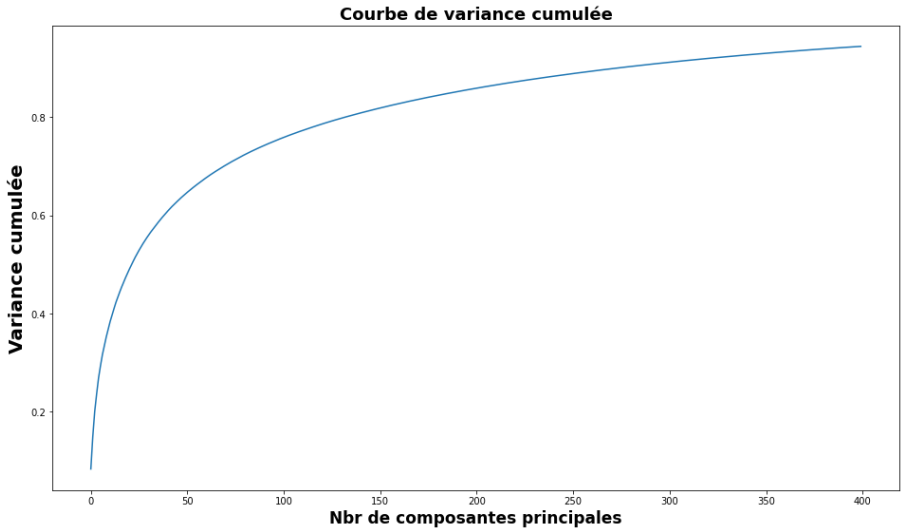


MODELE FINAL

- Regroupement des features issue de TF-IDF et VGG16
- Dimensions des données à concaténer :
  - Features textuelles: (840, 400)
  - Réseau de neurone Imagenet : (840, 4096)

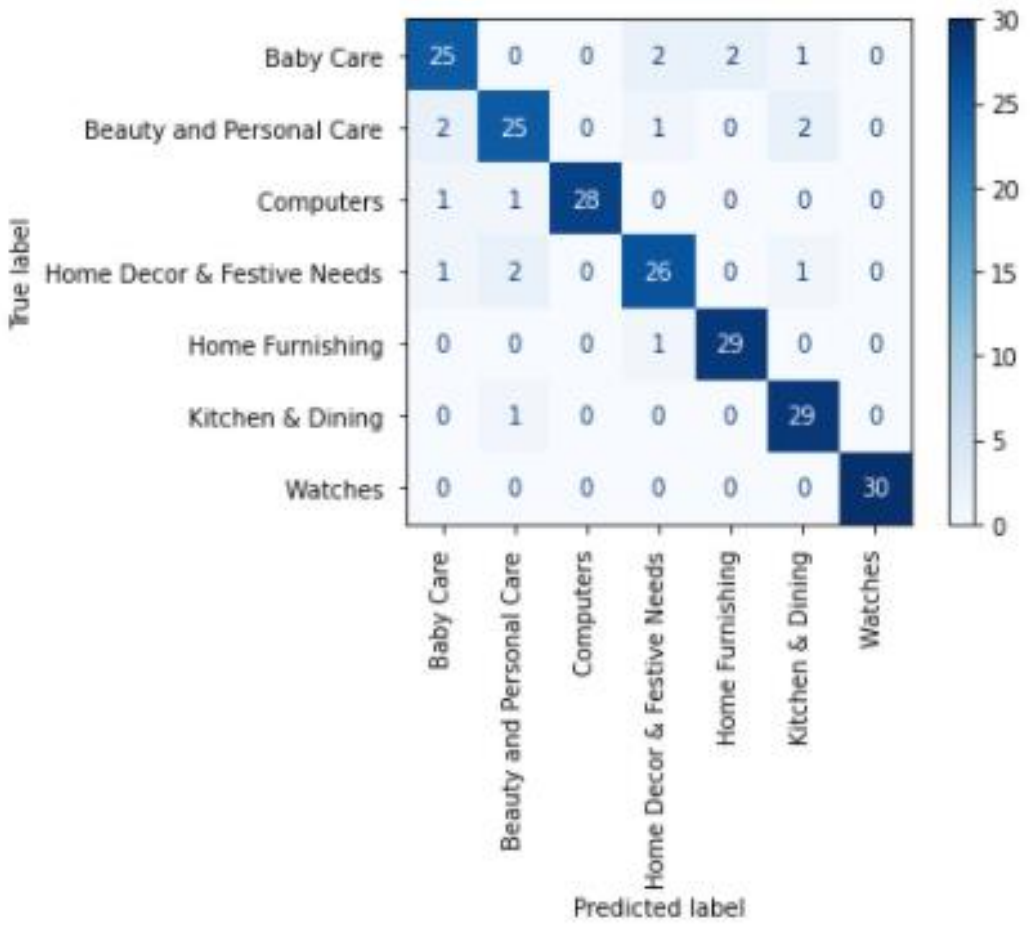
Fort déséquilibre entre dimensions des données textuelles et visuelles

Réalisation d'une ACP afin de faire une réduction dimensionnelle



Classification supervisée

accuracy sur jeu test : 0.9142857142857143



CONCLUSION

- Analyse des données textuelles et visuelles
- Extraction de features en utilisant des techniques adaptées
- Identification de produits plus difficiles à catégoriser
- La concaténation des features du meilleur modèle image avec le meilleur modèle texte permet d'améliorer la performance du VGG16.

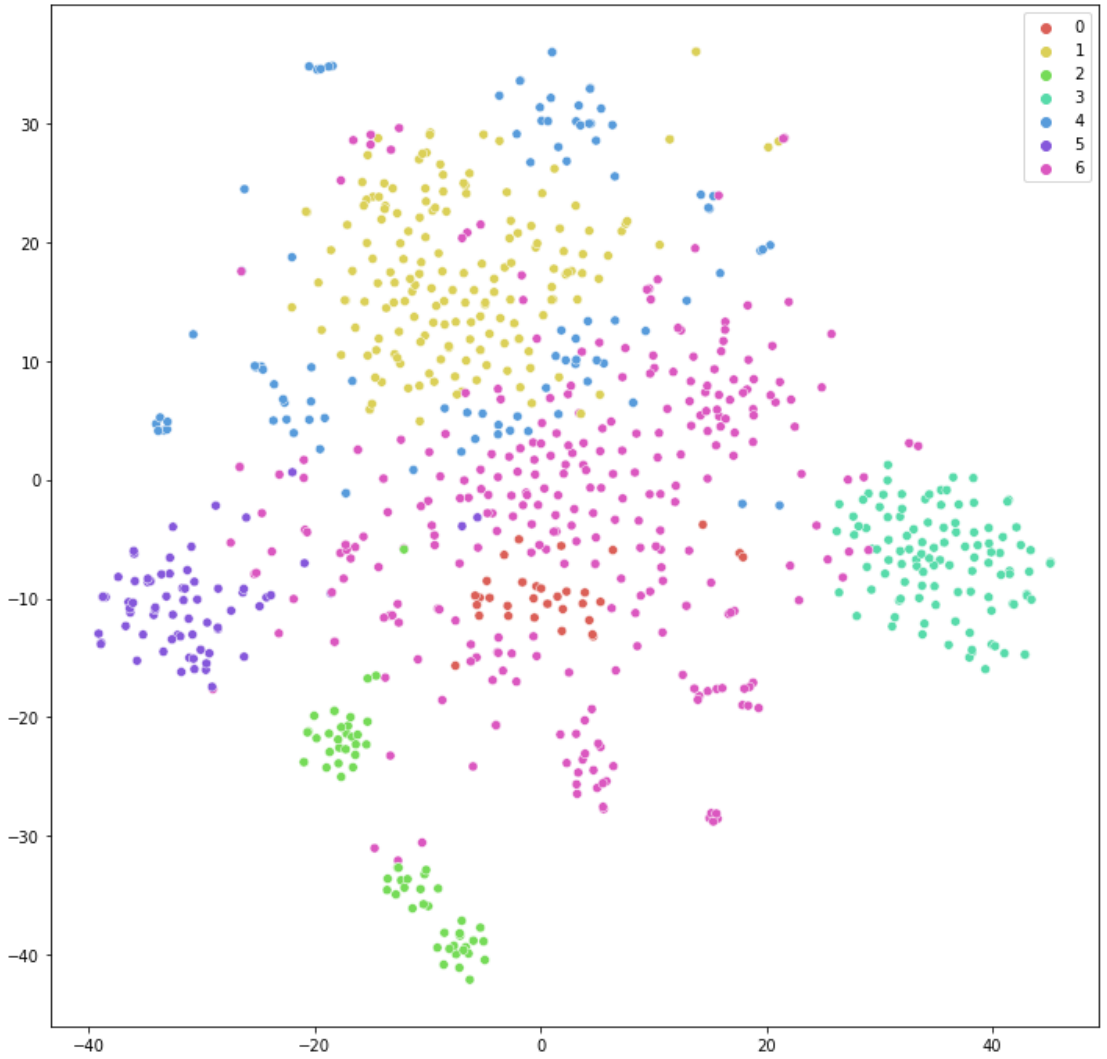
MODELE FINAL

- Regroupement des features issue de TF-IDF et resnet50
- Dimensions des données à concaténer :

  - Features textuelles: (840, 500)
  - Réseau de neurone Imagenet : (840, 100352)

Classification Non supervisée

Coef silhouette : 0.24
Score Adjusted Rand index : 0.32



Merci pour votre attention