

Projet 6 :

Classifiez automatiquement des biens de consommation

Lecerf Defer Amandine

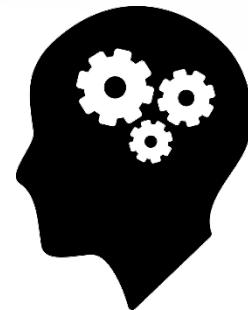
Plan

- I. Contexte + Données disponibles
- II. Traitement des Descriptions
 - a. Prétraitement
 - b. Clustering
 - c. Classification
- III. Traitement des Images
 - a. Prétraitement
 - b. Clustering
 - c. Classification
- IV. Faisabilité d'un moteur de recherche : Descriptions + Images



Compétences évaluées

- Prétraiter des données image pour obtenir un jeu de données exploitable
- Mettre en œuvre des techniques de réduction de dimension
- Représenter graphiquement des données à grandes dimensions
- Prétraiter des données texte pour obtenir un jeu de données exploitable





Problématique + données

Contexte

- Data Scientist pour « Place de marché »
 - mise en place marketplace en ligne
 - Produit = Description + Image
 - Actuellement peu de produits => catégorisation des produits manuelle par les vendeurs = **Peu fiable**
 - Souhaits :
 - faciliter la mise en ligne de nouveaux articles
 - faciliter la recherche de produits
 - Augmenter le nombre de produits disponibles
 - Inconvénient : catégorisation manuelle impossible
- Automatiser la catégorisation grâce à la description et à l'image du produit



Objectifs

- **Objectifs :**
 - Analyse un jeu de données :
 - Traitement de Descriptions
 - Traitement d'Images
 - Réduction de Dimensions
 - Clustering
 - Faisabilité d'un moteur de catégorisation à partir d'une Description et d'une Image pour un produit

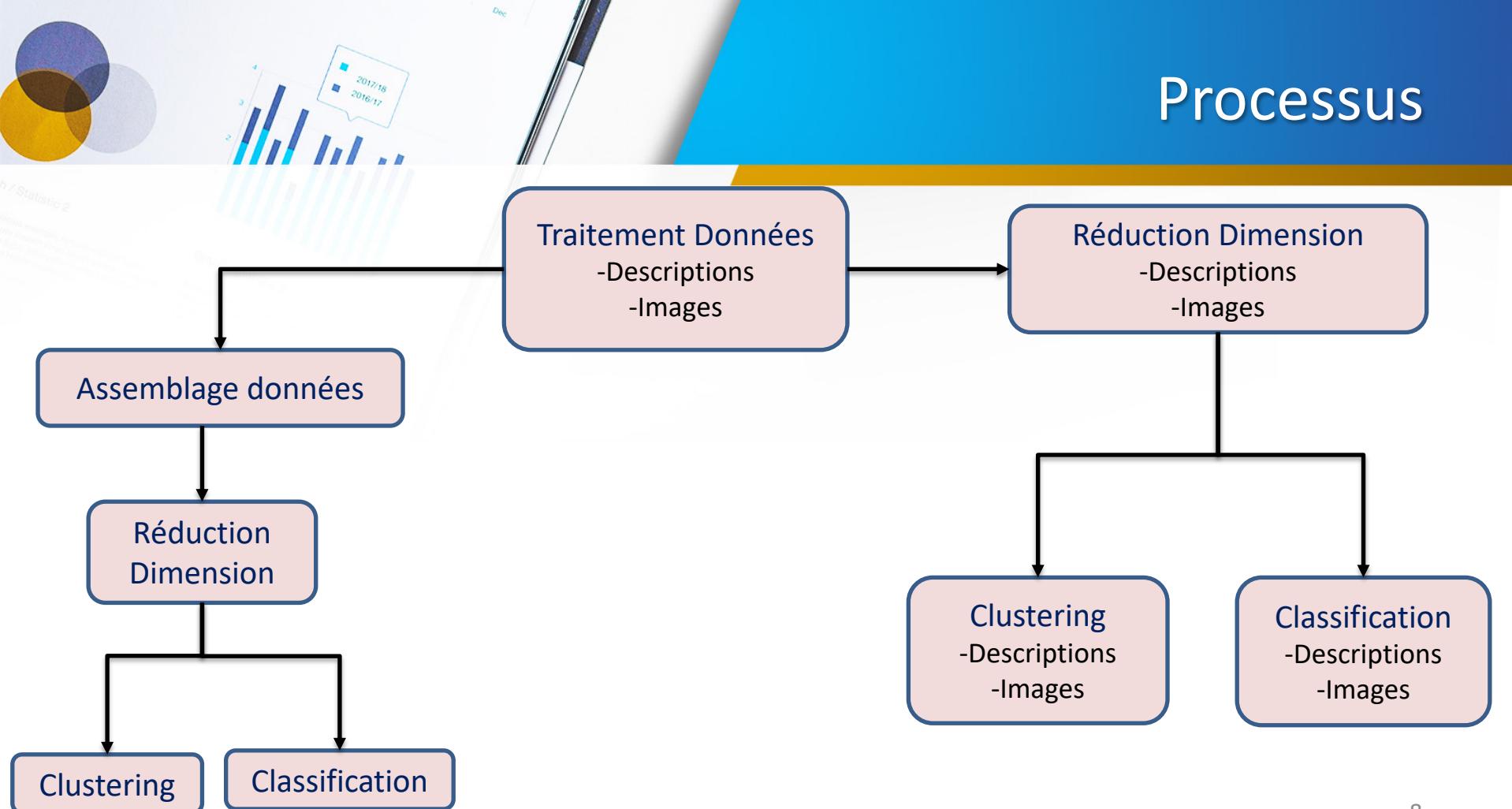


Données Disponibles

- 1050 articles
- 1 csv avec 15 colonnes :
 - Id, nom, catégorie, marque, description produit
 - Prix, prix soldé,
 - Note du produit ...
- 1 dossier avec une image pour chaque article
- Genre d'article :
 - Montres
 - Ordinateurs
 - Décoration
 - Produits d'hygiène
 - Articles pour bébé
 - ...



Processus

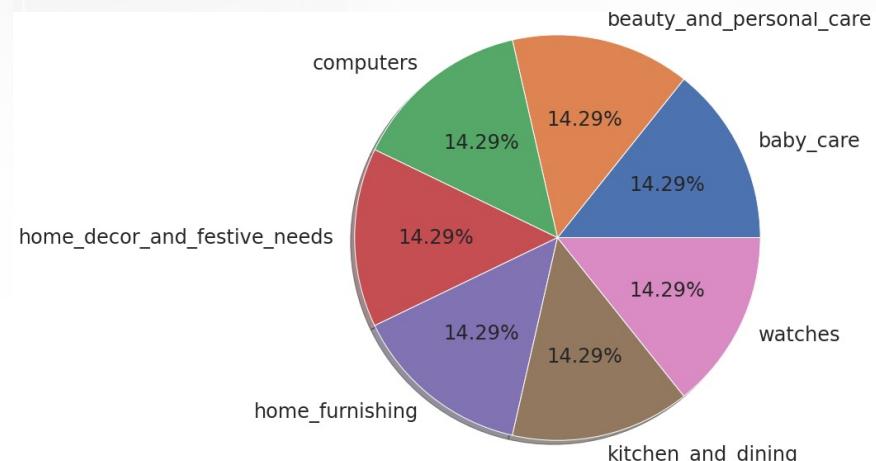




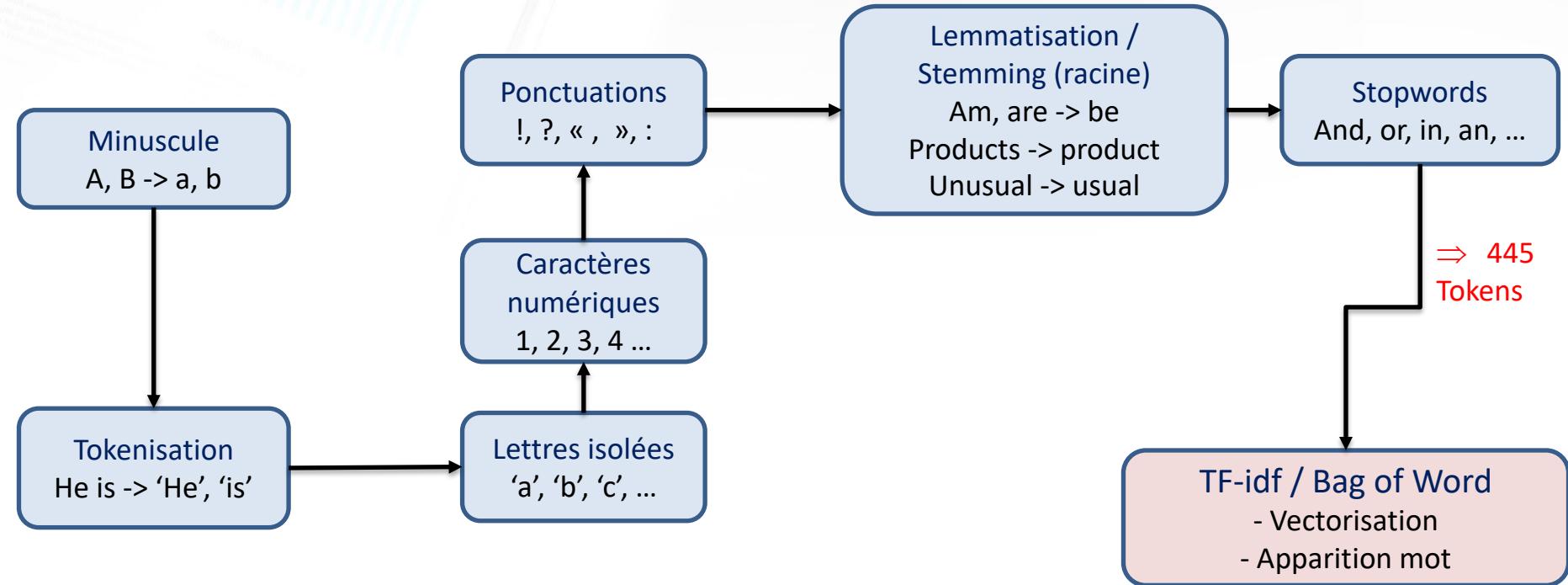
Traitement des Descriptions

Nettoyage CSV

- 2,17% NaN: marques produits
- Pas de doublon
- Colonnes gardées :
 - image,
 - product_name,
 - product_category_tree,
 - description
- Catégorisation manuelle
 - arbre avec plusieurs profondeurs
 - Racine de l'arbre = 7 catégories (150 produits/catégorie (14,29%))
 - Niveau 1 arbre = 62 catégories => trop de catégories
 - Nouvelle colonne = catégorie générale



Prétraitement Description



Description

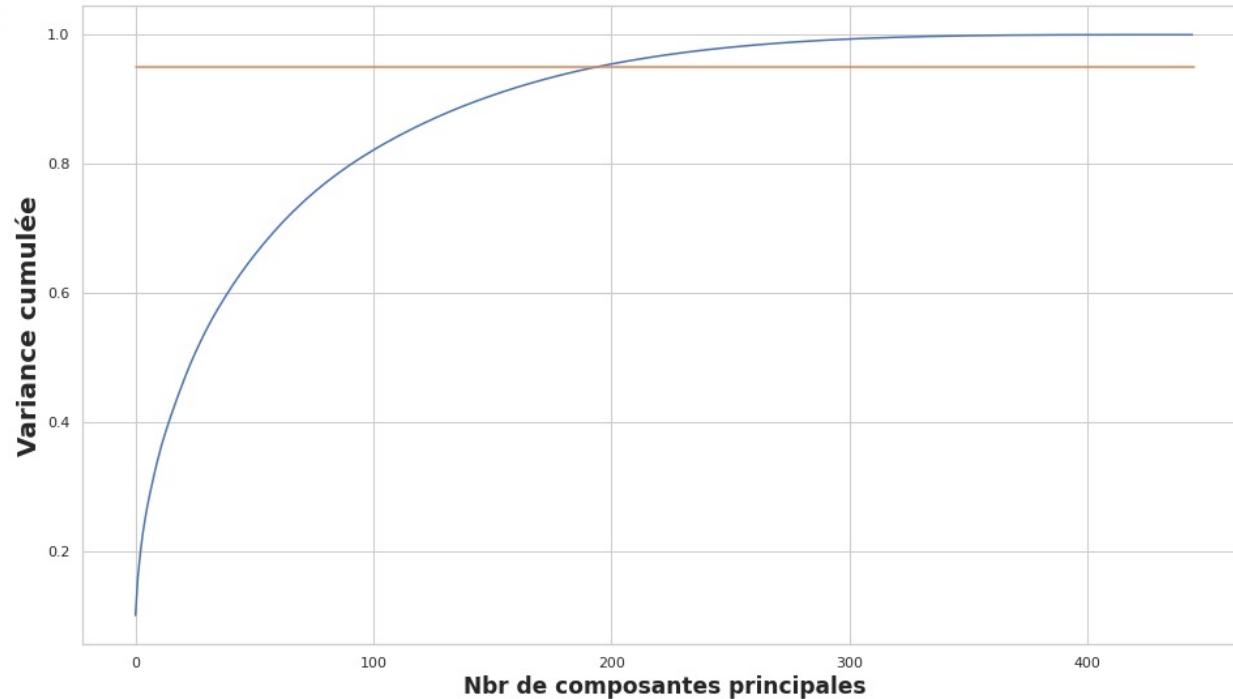
- Je privilégie le Tf-id

| | image | category | abstract | wash | washable | watch | water | waterproof | wear | wedding | weight | well | white | wide | width |
|---|--------------------------------------|-----------------|----------|----------|----------|-------|----------|------------|------|----------|----------|------|----------|------|----------|
| 0 | 55b85ea15a1536d46b7190ad6fff8ce7.jpg | home_furnishing | 0.202646 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.000000 |
| 1 | 7b72c92c2f6c40268628ec5f14c6d590.jpg | baby_care | 0.000000 | 0.000000 | 0.099876 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.072941 |
| 2 | 64d5d4a258243731dc7bbb1eef49ad74.jpg | baby_care | 0.000000 | 0.136487 | 0.000000 | 0.0 | 0.042381 | 0.0 | 0.0 | 0.000000 | 0.044419 | 0.0 | 0.000000 | 0.0 | 0.042507 |
| 3 | d4684dcdc759dd9cdf41504698d737d8.jpg | home_furnishing | 0.000000 | 0.048721 | 0.062329 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.139163 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.136561 |
| 4 | 6325b6870c54cd47be6ebfbfa620ec7.jpg | home_furnishing | 0.000000 | 0.000000 | 0.051890 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.039601 | 0.0 | 0.140024 | 0.0 | 0.113689 |

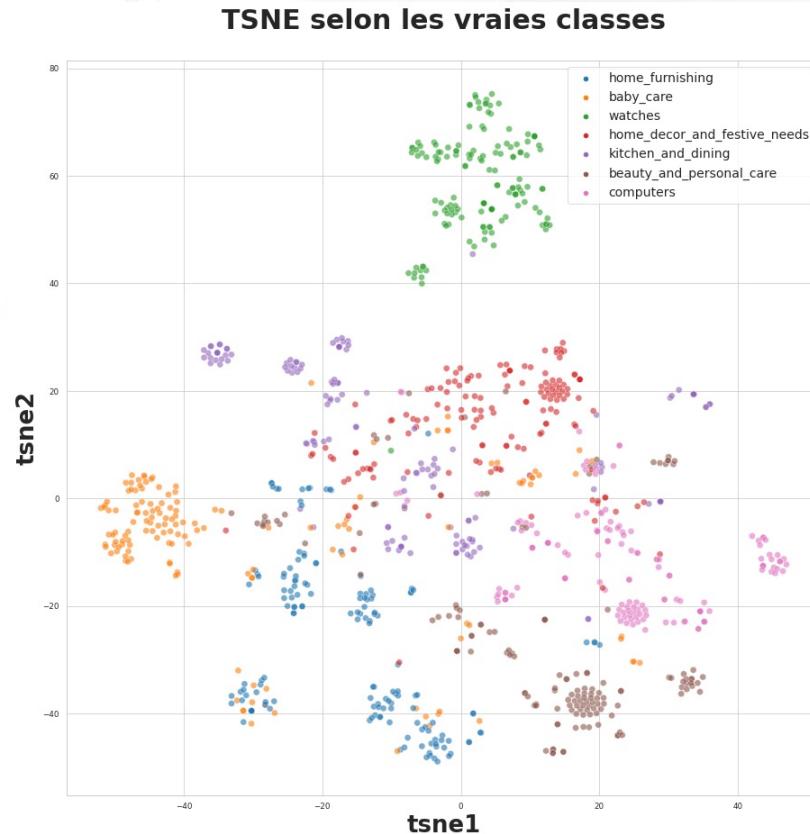
Description

=> 195 descripteurs textuels pour expliquer 95% de la variance cumulée du dataset

Courbe de variance cumulée



Description

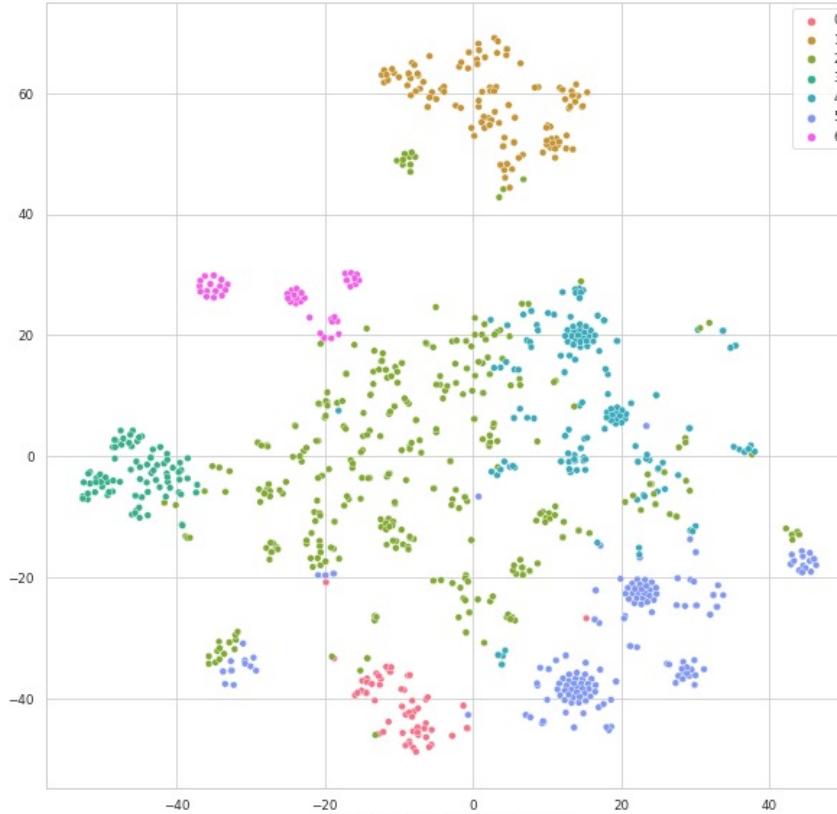


Catégorisation non supervisée

Clustering

| | Algorithm | silhouette | ARI |
|---|-------------------------|------------|----------|
| 0 | Kmeans | 0.122183 | 0.256783 |
| 1 | clustering hiérarchique | 0.109586 | 0.227130 |

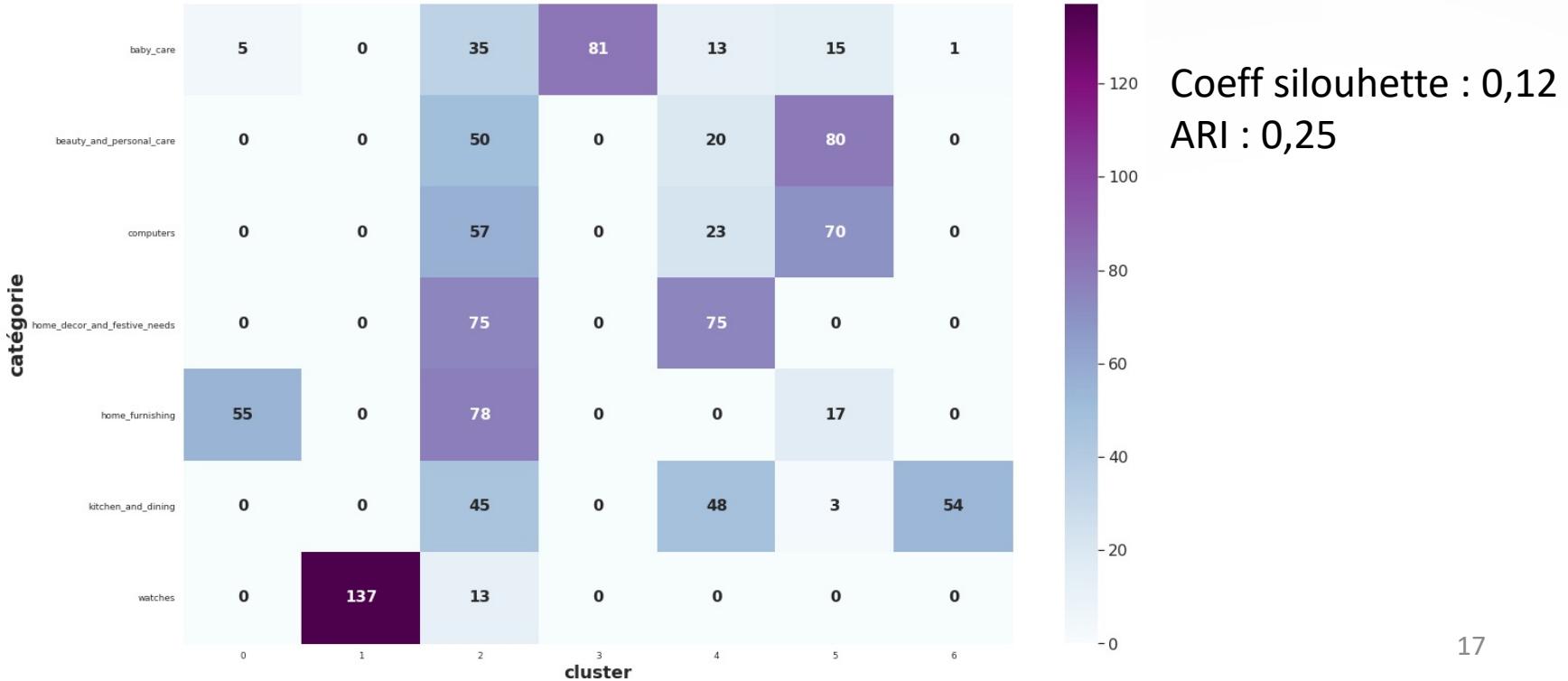
Catégorisation non supervisée



Coeff silouhette : 0,12
ARI : 0,25

Catégorisation non supervisée

Répartition des clusters



Catégorisation non supervisée

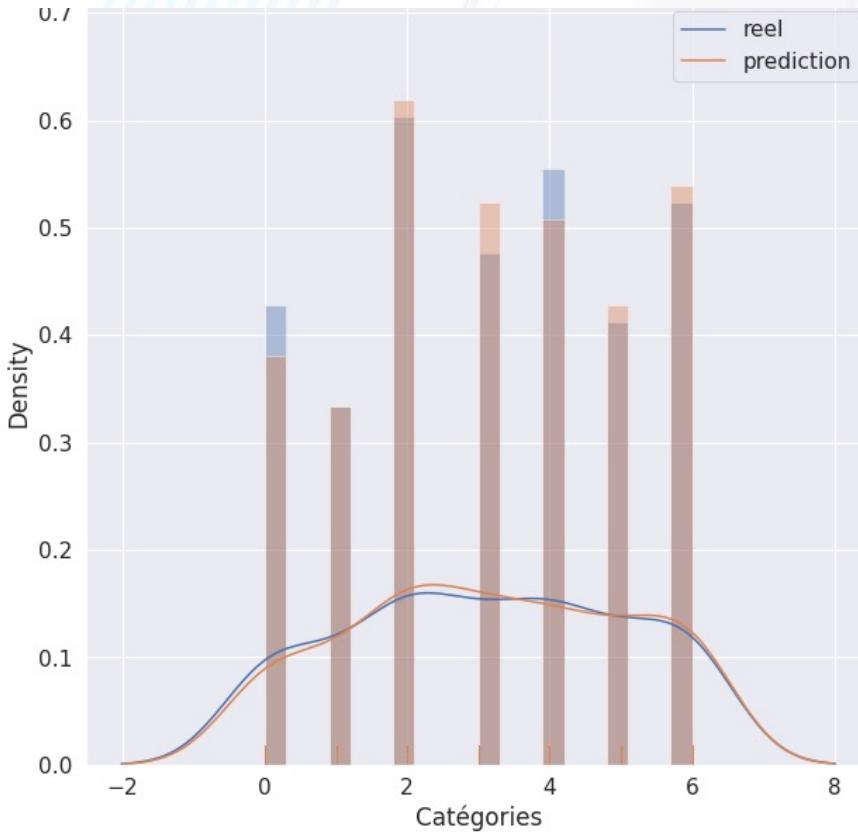
=> Quid d'une catégorisation supervisée ?

Catégorisation supervisée

Classification

| | Algorithmme | accuracy | precision | recall | f1_score |
|----------|---------------|----------|-----------|----------|----------|
| 0 | knn | 0.861905 | 0.863177 | 0.861905 | 0.861288 |
| 1 | Random Forest | 0.909524 | 0.910222 | 0.909524 | 0.908815 |
| 2 | MultinomialNB | 0.895238 | 0.905212 | 0.895238 | 0.894759 |

Catégorisation supervisée



0 = baby_care

1 = beauty_and_personal_care

2 = computers

3 = home_decor_and_festive_needs

4 = home_furnishing

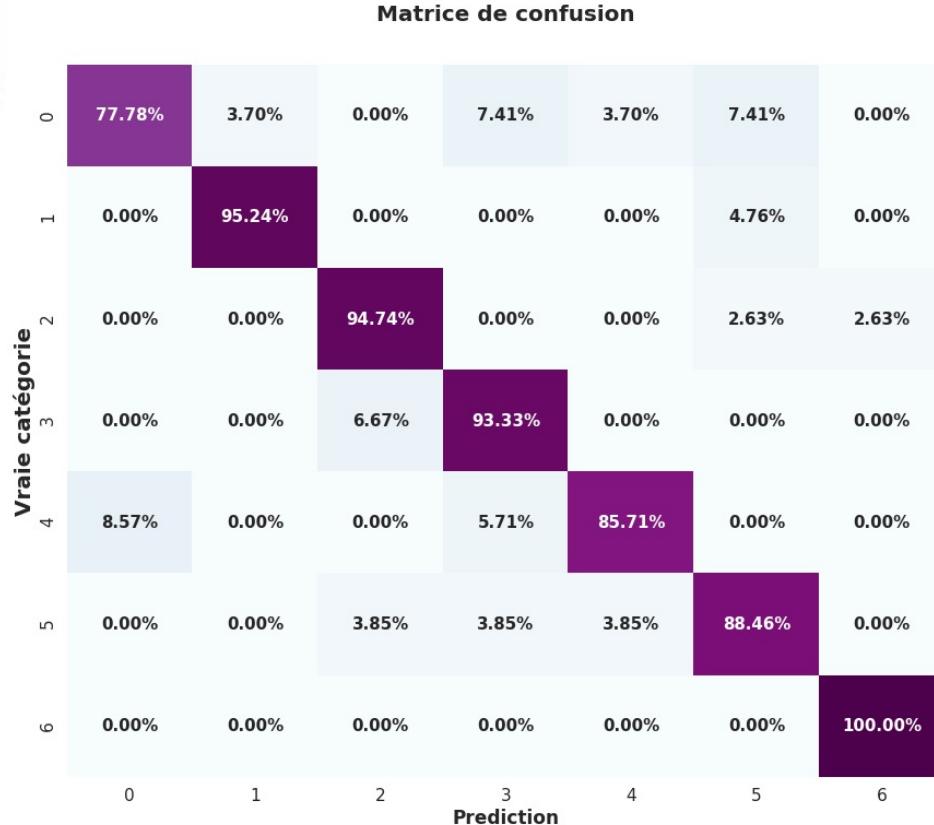
5 = kitchen_and_dining

6 = watches

Accuracy : 0,909

Précision : 0,910

Catégorisation supervisée



0 = baby_care

1 = beauty_and_personal_care

2 = computers

3 = home_decor_and_festive_needs

4 = home_furnishing

5 = kitchen_and_dining

6 = watches

Accuracy : 0,909

Précision : 0,910



Traitement des Images

Prétraitement Images



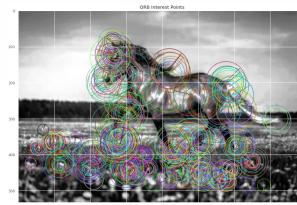
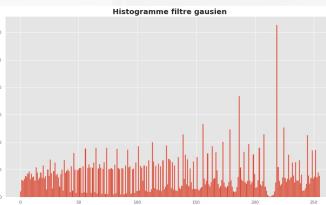
Conversion en gris



Redimensionner
image
-width : 400



Amélioration
du contraste



Débruitage



Filtre gausien



Vectorisation

691
descripteurs

Clustering Descripteurs

477 138
descripteurs

Recherche Descripteurs
- ORB

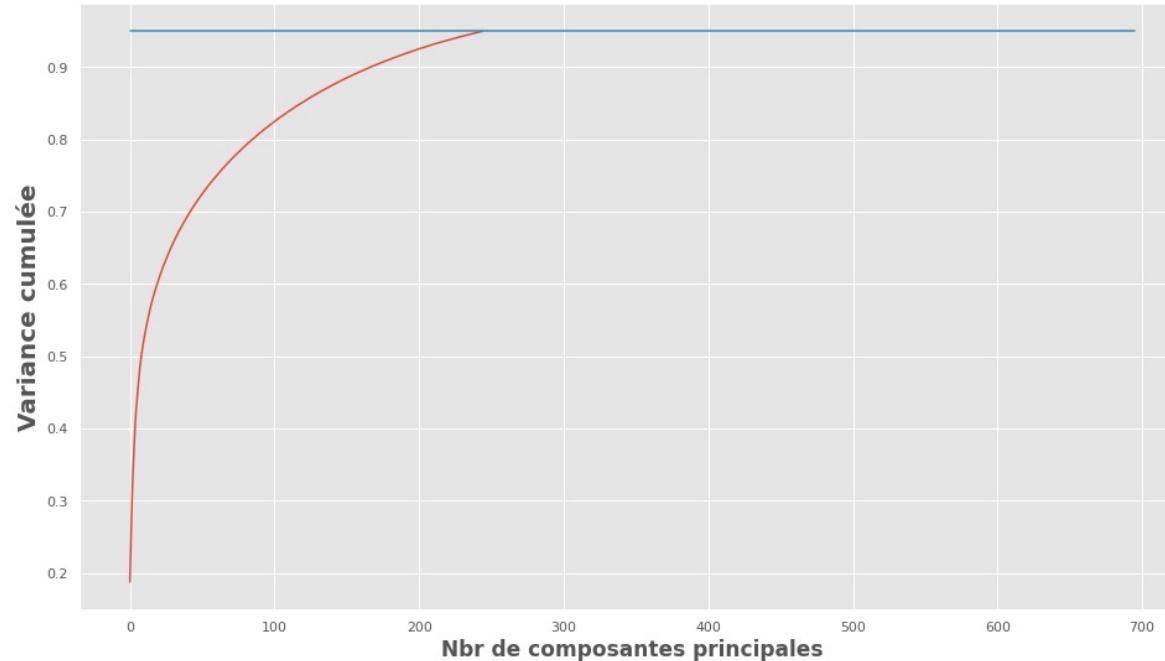


Descripteurs visuels

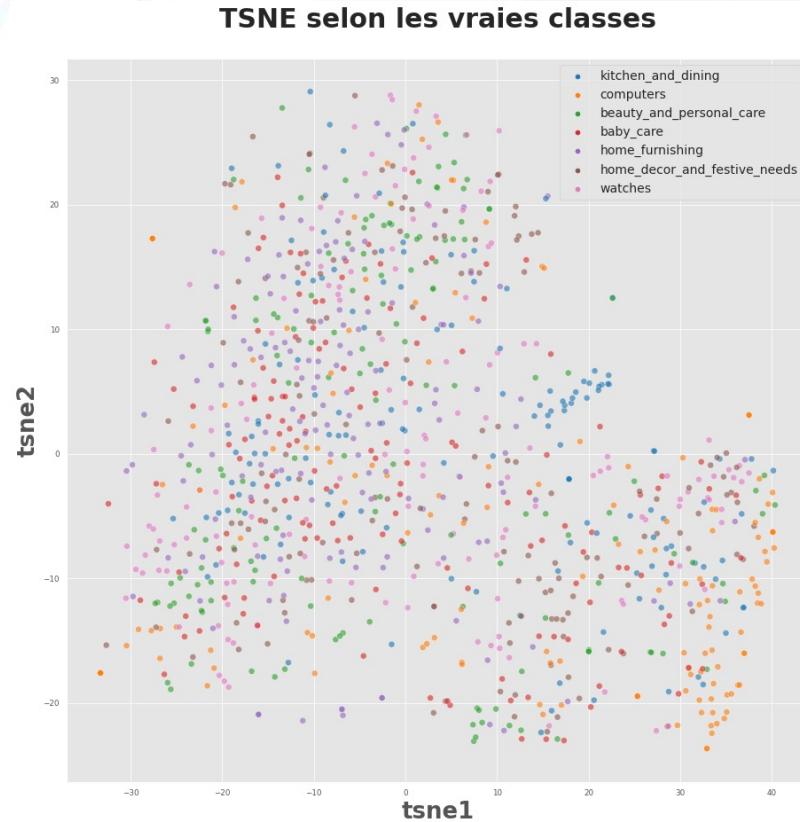
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 0 | 0.002033 | 0.004065 | 0.002033 | 0.000000 | 0.002033 | 0.000000 | 0.002033 | 0.004065 | 0.000000 | 0.000000 | 0.002033 | 0.002033 | 0.004065 |
| 1 | 0.000000 | 0.005102 | 0.000000 | 0.000000 | 0.000000 | 0.005102 | 0.002551 | 0.002551 | 0.005102 | 0.002551 | 0.000000 | 0.005102 | 0.002551 |
| 2 | 0.000000 | 0.006000 | 0.000000 | 0.000000 | 0.002000 | 0.002000 | 0.002000 | 0.012000 | 0.004000 | 0.002000 | 0.002000 | 0.000000 | 0.000000 |
| 3 | 0.005814 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.005814 | 0.002907 | 0.000000 |
| 4 | 0.005814 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.005814 | 0.002907 | 0.000000 |

=> 245 descripteurs visuels pour expliquer 95% de la variance cumulée du dataset

Courbe de variances cumulées



Images

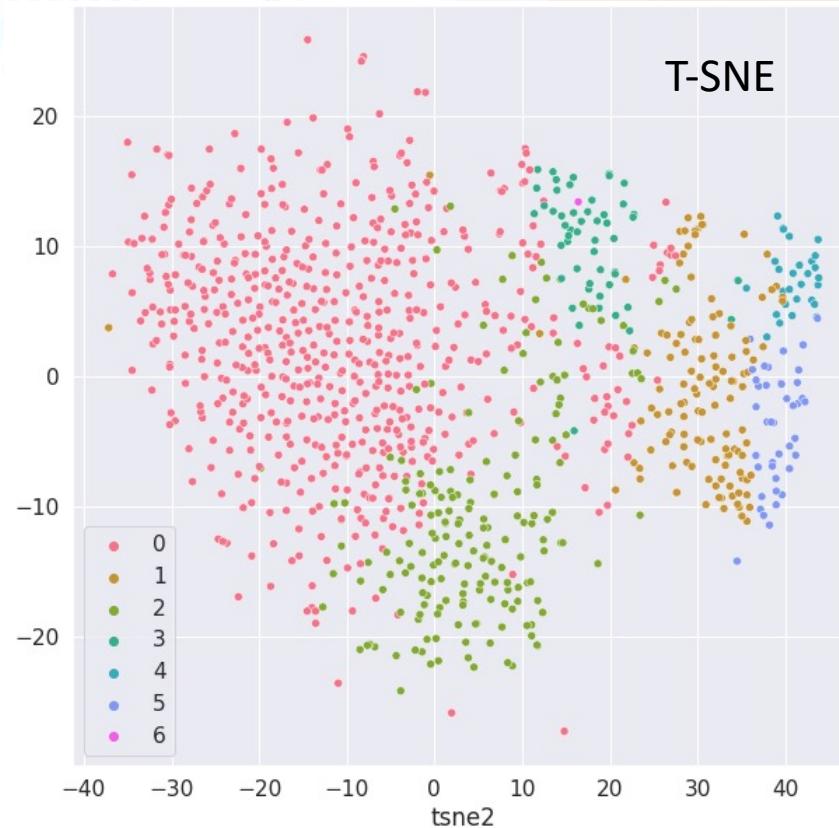


Catégorisation non supervisée

Clustering

| | Algorithm | silhouette | ARI |
|---|-------------------------|------------|----------|
| 0 | k-means | 0.079861 | 0.018805 |
| 1 | clustering hiérarchique | 0.192383 | 0.011107 |

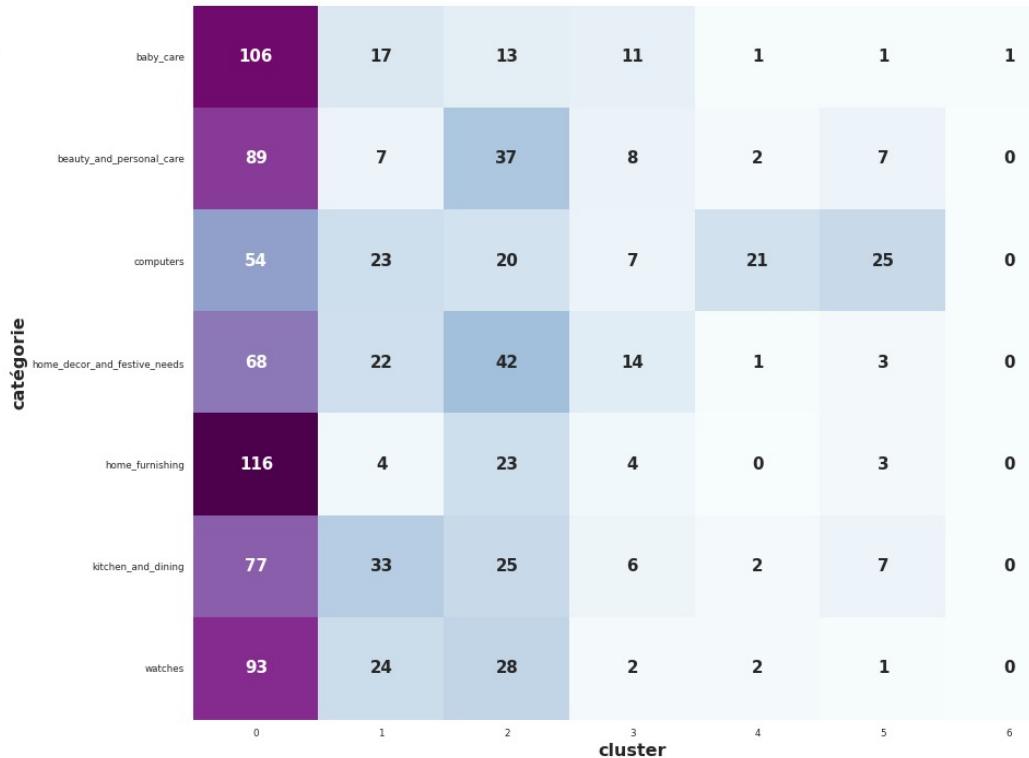
Catégorisation non supervisée



Coeff silouhette : 0,079
ARI : 0,018

Catégorisation non supervisée

Répartition des clusters



Coeff silhouette : 0,079
ARI : 0,018

Catégorisation non supervisée

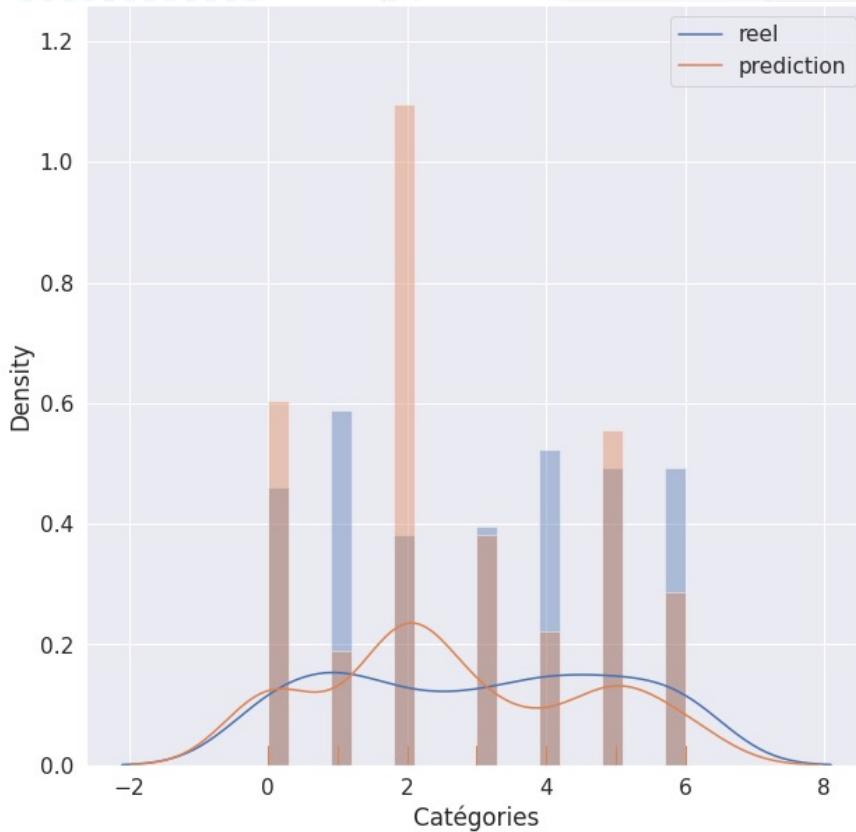
- => Catégorisation non supervisée donne de très mauvais résultats
- => Quid d'une catégorisation supervisée ?

Catégorisation supervisée

Classification :Train et test + traitement + ACP

| | Algorithme | accuracy | precision | recall | f1_score |
|---|---------------|----------|-----------|----------|----------|
| 0 | knn | 0.109524 | 0.112537 | 0.109524 | 0.103226 |
| 1 | Random Forest | 0.157143 | 0.138722 | 0.157143 | 0.134244 |

Catégorisation supervisée



0 = baby_care

1 = beauty_and_personal_care

2 = computers

3 = home_decor_and_festive_needs

4 = home_furnishing

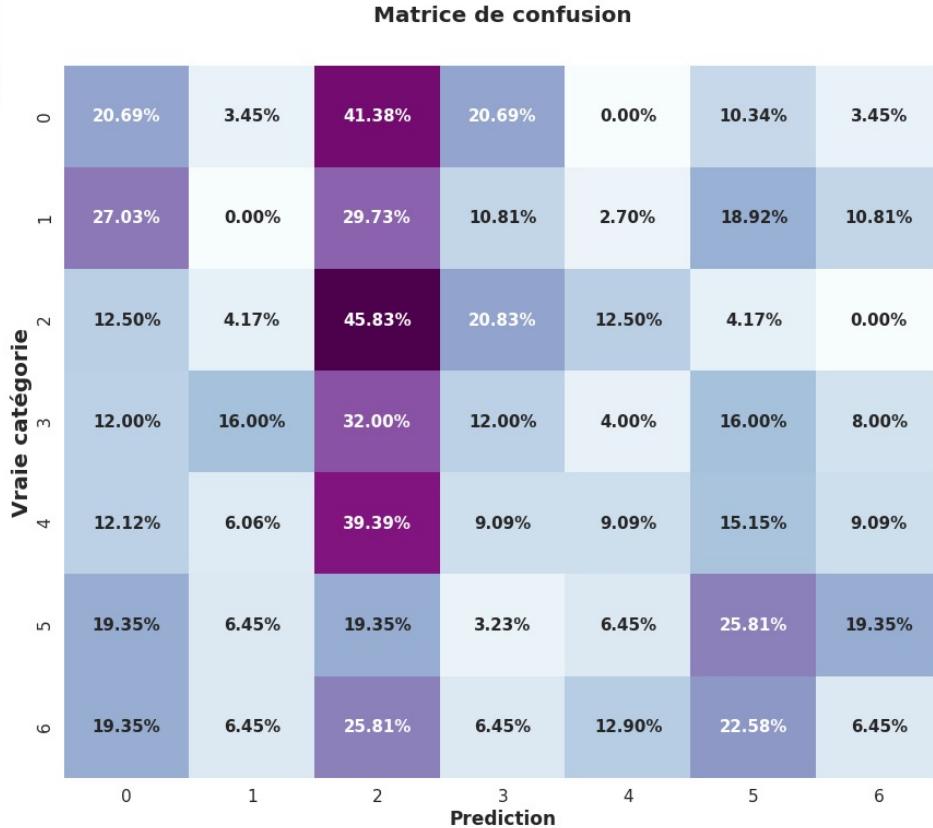
5 = kitchen_and_dining

6 = watches

Accuracy : 0,15

Précision : 0,13

Catégorisation supervisée



0 = baby_care

1 = beauty_and_personal_care

2 = computers

3 = home_decor_and_festive_needs

4 = home_furnishing

5 = kitchen_and_dining

6 = watches

Accuracy : 0,15

Précision : 0,13

CNN & Transfert Learning : Modèle

CNN = empilement de couches de neurones

Transfert Learning = utilisation modèle pré-entraîné

- VGG16 (ImageNet) : extraction descripteur image
 - sans les couches fully_connected
 - Entrainement dernier classifieur pas toutes les couches
- CNN :
 - couche Flatten : aplatissement image : matrice -> vecteur
 - couche Dense : affiner classification à 7 catégories

```
1 new_model = Sequential()
2
3 # Charger VGG-16 pré-entraîné sur ImageNet et sans les couches fully-connected (couches hautes).
4 model = VGG16(weights="imagenet", include_top=False, input_shape=(224, 224, 3))
5
6 # On entraîne seulement le nouveau classifieur et on ne ré-entraîne pas les autres couches :
7 for layer in model.layers:
8     layer.trainable = False
9
10 new_model.add(model)
11
12 # Ajout d'une couche Flatten pour réduire les dimensions de l'input.
13 new_model.add(Flatten())
14
15 # Ajout de la couche Dense de prédiction adaptée à notre problème de classifications (7 classes).
16 new_model.add(Dense(7, activation='softmax'))
```

```
1 # Compilation du modèle
2 new_model.compile(loss="categorical_crossentropy", optimizer=optimizers.SGD(
3     lr=0.0001, momentum=0.9), metrics=["accuracy"])
4
5 batch_size = 32
```

CNN & Transfert Learning : Traitement Image



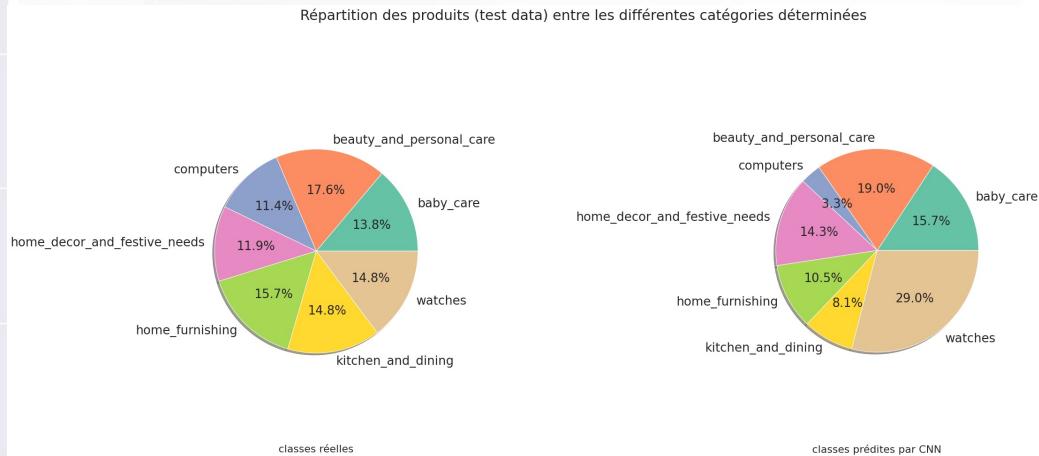
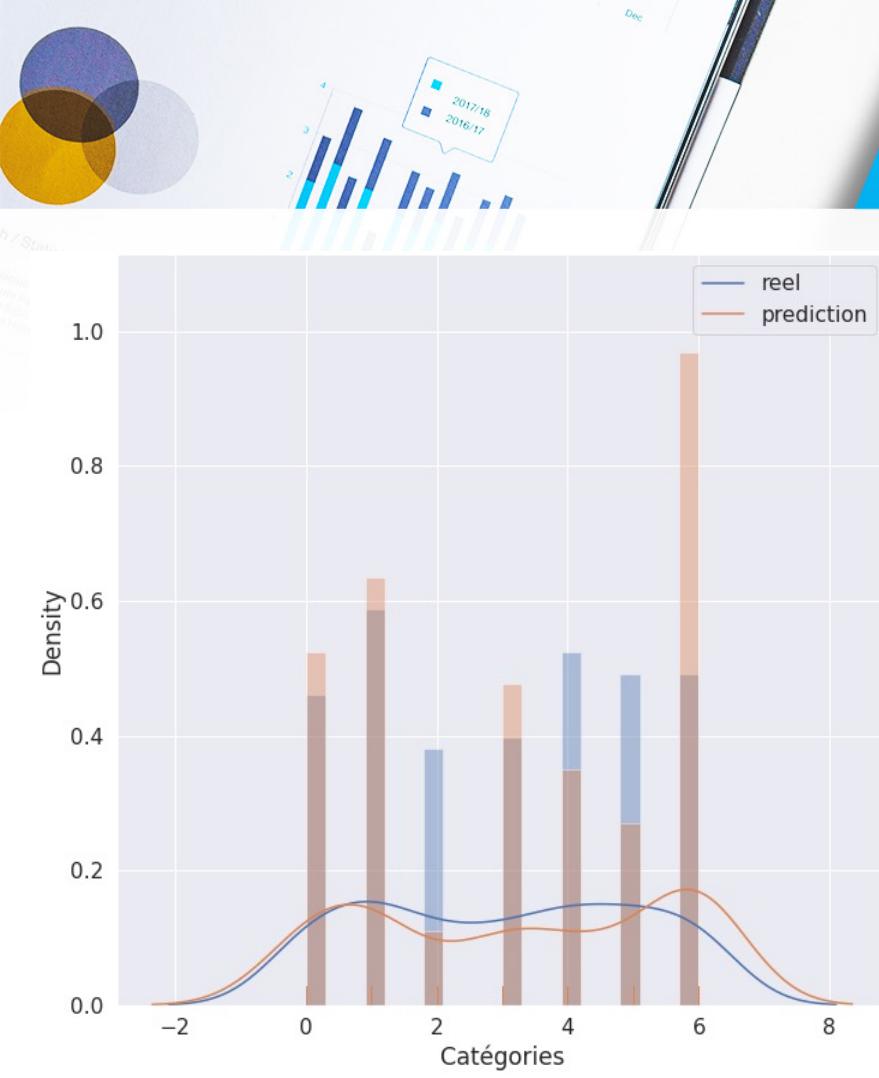
CNN supervisé

- Séparation images en un dossier train et test
- VGG-16 reçoit des images de taille (224, 224, 3) : redimensionner image
- Keras traite les images comme des tableaux numpy : transformation en tableau
- Appliquer les mêmes pré-traitements pour les données d'entraînement et de prédiction

```
10 for j in os.listdir(folder):
11     folder_courant = folder + '/' + j
12
13     for filename in os.listdir(folder_courant):
14         y_test_category.append(j)
15
16         # Path du fichier.
17         path = folder_courant + '/' + filename
18
19         # Charger l'image à la bonne dimension.
20         img_resize = load_img(path, target_size=(224, 224))
21
22         # Transformation de l'image en tableau numpy.
23         img = img_to_array(img_resize)
24         img = img.reshape([1, img.shape[0], img.shape[1], img.shape[2]])
25
26         # Preprocessing.
27         img = preprocess_input(img)
28
29         # Ajout de l'image à la liste de stockage.
30         if (i == 0):
31             xl = img
32             i = 1
33         else:
34             xl = np.concatenate((xl, img), axis=0)
```

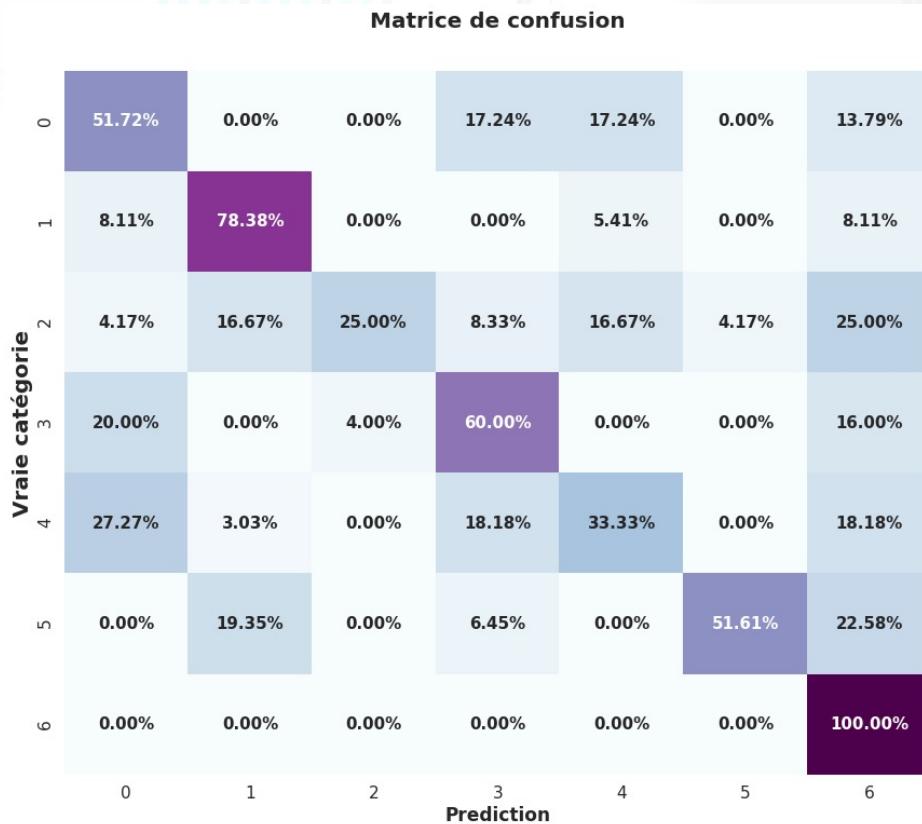
```
5 train_dir = os.path.join(base_dir, 'Training Set')
6 test_dir = os.path.join(base_dir, 'Testing Set')
7
8
9 train_datagen = ImageDataGenerator(preprocessing_function=preprocess_input)
10
11 # Testing Augmentation
12 test_datagen = ImageDataGenerator(preprocessing_function=preprocess_input)
13
14 # Generates batches of Augmented Image data
15 train_generator = train_datagen.flow_from_directory(train_dir, target_size=(224, 224),
16                                                       batch_size=batch_size,
17                                                       class_mode='categorical')
18
19 # Generator for validation data
20 validation_generator = test_datagen.flow_from_directory(test_dir,
21                                                       target_size=(224, 224),
22                                                       batch_size=batch_size,
23                                                       class_mode='categorical')
```

CNN résultat



Accuracy : 0,58
Precision : 0,64

CNN résultat



0 = baby_care

1 = beauty_and_personal_care

2 = computers

3 = home_decor_and_festive_needs

4 = home_furnishing

5 = kitchen_and_dining

6 = watches

Accuracy : 0,58

Precision : 0,64



Association Description + Image

Préparation

Descriptions

- Tf-IDF + ACP
- size : (1050, 197)

| | image | product_category | Feature BoW 0 | Feature BoW 1 | Feature BoW 2 | Feature BoW 3 | Feature BoW 4 | Feature BoW 5 | Feature BoW 6 | Feature BoW 7 | Feature BoW 8 | Feature BoW 9 | Feature BoW 10 | Feature BoW 11 | Feature BoW 12 | Feature BoW 13 | Feature BoW 14 | Feature BoW 15 | Feature BoW 16 | Feature BoW 17 | Feature BoW 18 | Feature BoW 19 | Feature BoW 20 | Feature BoW 21 | Feature BoW 22 | Feature BoW 23 | Feature BoW 24 | Feature BoW 25 | Feature BoW 26 | Feature BoW 27 | Feature BoW 28 | Feature BoW 29 | Feature BoW 30 | Feature BoW 31 | Feature BoW 32 | Feature BoW 33 | Feature BoW 34 | Feature BoW 35 | Feature BoW 36 | Feature BoW 37 |
|---|--------------------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| 0 | 55b85ea15a1536d46b7190ad6fffc8e7.jpg | home_furnishing | -0.290536 | -0.055178 | -0.021944 | -0.137678 | -0.118319 | -0. | -0.055178 | -0.021944 | -0.137678 | -0.118319 | -0. | -0.055178 | -0.021944 | -0.137678 | -0.118319 | -0. | -0.055178 | -0.021944 | -0.137678 | -0.118319 | -0. | -0.055178 | -0.021944 | -0.137678 | -0.118319 | -0. | -0.055178 | -0.021944 | -0.137678 | -0.118319 | -0. | -0.055178 | -0.021944 | -0.137678 | -0.118319 | -0. | | |
| 1 | 7b72c92c2f6c40268628ec5f14c6d590.jpg | baby_care | -0.298032 | -0.042844 | -0.127226 | 0.014708 | -0.127724 | -0. | -0.042844 | -0.127226 | 0.014708 | -0.127724 | -0. | -0.042844 | -0.127226 | 0.014708 | -0.127724 | -0. | -0.042844 | -0.127226 | 0.014708 | -0.127724 | -0. | -0.042844 | -0.127226 | 0.014708 | -0.127724 | -0. | -0.042844 | -0.127226 | 0.014708 | -0.127724 | -0. | -0.042844 | -0.127226 | 0.014708 | -0.127724 | -0. | | |
| 2 | 64d5d4a258243731dc7bbbe1ef49ad74.jpg | baby_care | -0.284868 | -0.053634 | -0.095143 | -0.006220 | -0.105319 | -0. | -0.053634 | -0.095143 | -0.006220 | -0.105319 | -0. | -0.053634 | -0.095143 | -0.006220 | -0.105319 | -0. | -0.053634 | -0.095143 | -0.006220 | -0.105319 | -0. | -0.053634 | -0.095143 | -0.006220 | -0.105319 | -0. | -0.053634 | -0.095143 | -0.006220 | -0.105319 | -0. | -0.053634 | -0.095143 | -0.006220 | -0.105319 | -0. | | |
| 3 | d4684ddcc759dd9cdf41504698d737d8.jpg | home_furnishing | -0.304484 | -0.048035 | -0.065740 | -0.032090 | -0.147239 | -0. | -0.048035 | -0.065740 | -0.032090 | -0.147239 | -0. | -0.048035 | -0.065740 | -0.032090 | -0.147239 | -0. | -0.048035 | -0.065740 | -0.032090 | -0.147239 | -0. | -0.048035 | -0.065740 | -0.032090 | -0.147239 | -0. | -0.048035 | -0.065740 | -0.032090 | -0.147239 | -0. | -0.048035 | -0.065740 | -0.032090 | -0.147239 | -0. | | |
| 4 | 6325b6870c54cd7be6ebfbfa620ec7.jpg | home_furnishing | -0.300111 | -0.059830 | -0.058106 | -0.053589 | -0.148128 | -0. | -0.059830 | -0.058106 | -0.053589 | -0.148128 | -0. | -0.059830 | -0.058106 | -0.053589 | -0.148128 | -0. | -0.059830 | -0.058106 | -0.053589 | -0.148128 | -0. | -0.059830 | -0.058106 | -0.053589 | -0.148128 | -0. | -0.059830 | -0.058106 | -0.053589 | -0.148128 | -0. | -0.059830 | -0.058106 | -0.053589 | -0.148128 | -0. | | |



Ordre des produits
(id_image)

| | image | Principal Component 0 | Principal Component 1 | Principal Component 2 | Principal Component 3 | Principal Component 4 | Principal Component 5 | Principal Component 6 | Principal Component 7 | Principal Component 8 | Principal Component 9 | Principal Component 10 | Principal Component 11 | Principal Component 12 | Principal Component 13 | Principal Component 14 | Principal Component 15 | Principal Component 16 | Principal Component 17 | Principal Component 18 | Principal Component 19 | Principal Component 20 | Principal Component 21 | Principal Component 22 | Principal Component 23 | Principal Component 24 | Principal Component 25 | Principal Component 26 | Principal Component 27 | Principal Component 28 | Principal Component 29 | Principal Component 30 | Principal Component 31 | Principal Component 32 | Principal Component 33 | Principal Component 34 | Principal Component 35 | Principal Component 36 | Principal Component 37 |
|---|--------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| 0 | f01140c02fd09e5c56312a64fa8db3da.jpg | 47.501534 | 78.069267 | 278.580475 | 88.402855 | 226.873260 | -24 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 1 | b4d35b6536b623c83b293eb3a1c02bba.jpg | -158.536484 | -161.657761 | -24.215393 | -68.864708 | 91.076775 | -16 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| 2 | 69446af66066146718cd56978e99fff.jpg | -190.977737 | 370.004974 | 486.494110 | 723.511597 | -87.442993 | 12 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 3 | 6329ddb46087cb8c09ced95471dd9643.jpg | -214.660141 | -126.670815 | 67.009918 | -16.149027 | 146.044571 | -25 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| 4 | 7e6ae1b372e8e0b509b83ba3658b8107.jpg | -214.660019 | -126.671494 | 67.009209 | -16.149225 | 146.044785 | -25 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |

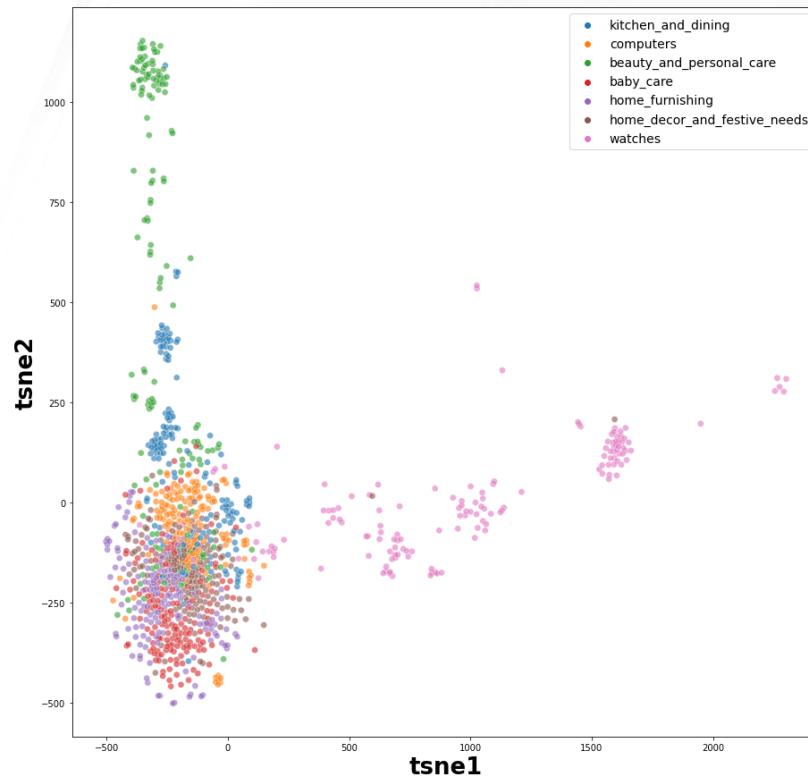


Size :
(1050, 942)

| | image | product_category | Feature BoW 0 | Feature BoW 1 | Feature BoW 2 | Feature BoW 37 | ... | Principal Component 900 | Principal Component 901 | Principal Component 902 | Principal Component 903 | Principal Component 904 | Principal Component 905 | Principal Component 906 | Principal Component 907 | Principal Component 908 |
|---|--------------------------------------|--------------------|---------------|---------------|---------------|----------------|-----|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| 0 | f01140c02fd09e5c56312a64fa8db3da.jpg | kitchen_and_dining | -0.339980 | -0.025793 | 0.005679 | -0.033565 | ... | 7.619818 | -6.366675 | -3.012838 | 0.889892 | -9.145618 | 6.133828 | 6.141134 | -8.115696 | -4.474566 |
| 1 | b4d35b6536b623c83b293eb3a1c02bba.jpg | computers | -0.256665 | -0.025599 | 0.028749 | 0.057838 | ... | -17.493641 | 2.857285 | 3.176433 | -11.563011 | 1.904783 | 17.578993 | 2.459184 | -0.416151 | 7.820480 |
| 2 | 69446af66066146718cd56978e99fff.jpg | kitchen_and_dining | -0.328396 | -0.071964 | 0.067223 | -0.0111543 | ... | -12.689688 | -23.040579 | 11.368026 | 4.525949 | -14.030891 | 1.327769 | 6.503119 | -12.271140 | 27.528654 |
| 3 | 6329ddb46087cb8c09ced95471dd9643.jpg | computers | -0.260871 | -0.033179 | 0.004478 | -0.111121 | ... | 0.051887 | 0.940511 | 0.878449 | -1.411968 | -1.886258 | 3.731393 | -0.873620 | -1.681192 | 2.796400 |
| 4 | 7e6ae1b372e8e0b509b83ba3658b8107.jpg | computers | -0.260871 | -0.033179 | 0.004478 | -0.111121 | ... | 0.051787 | 0.940556 | 0.878507 | -1.412057 | -1.886347 | 3.731398 | -0.873591 | -1.681193 | 2.796565 |

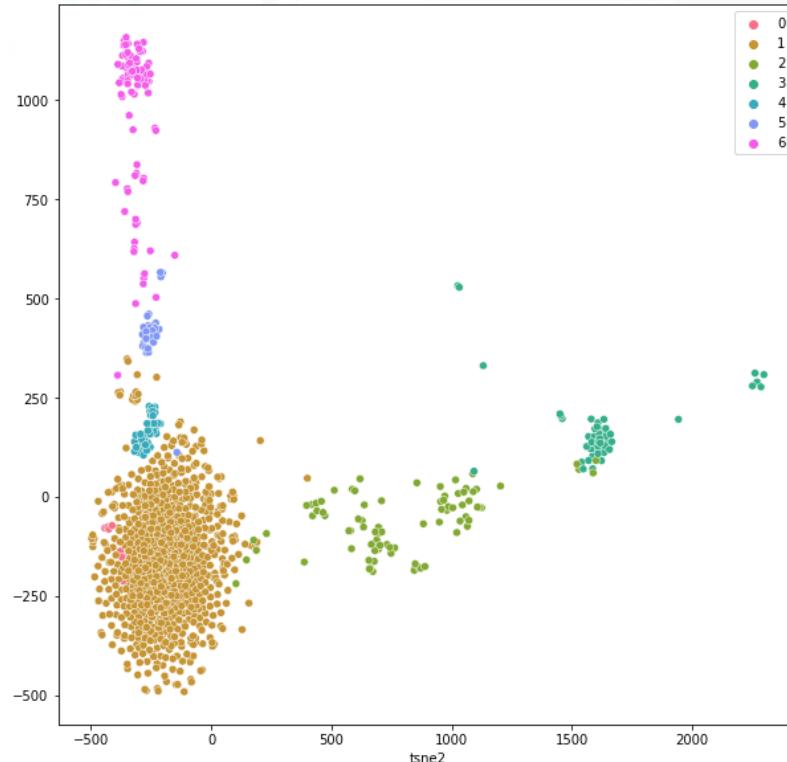
Représentation

TSNE selon les vraies classes



Catégorisation non supervisée

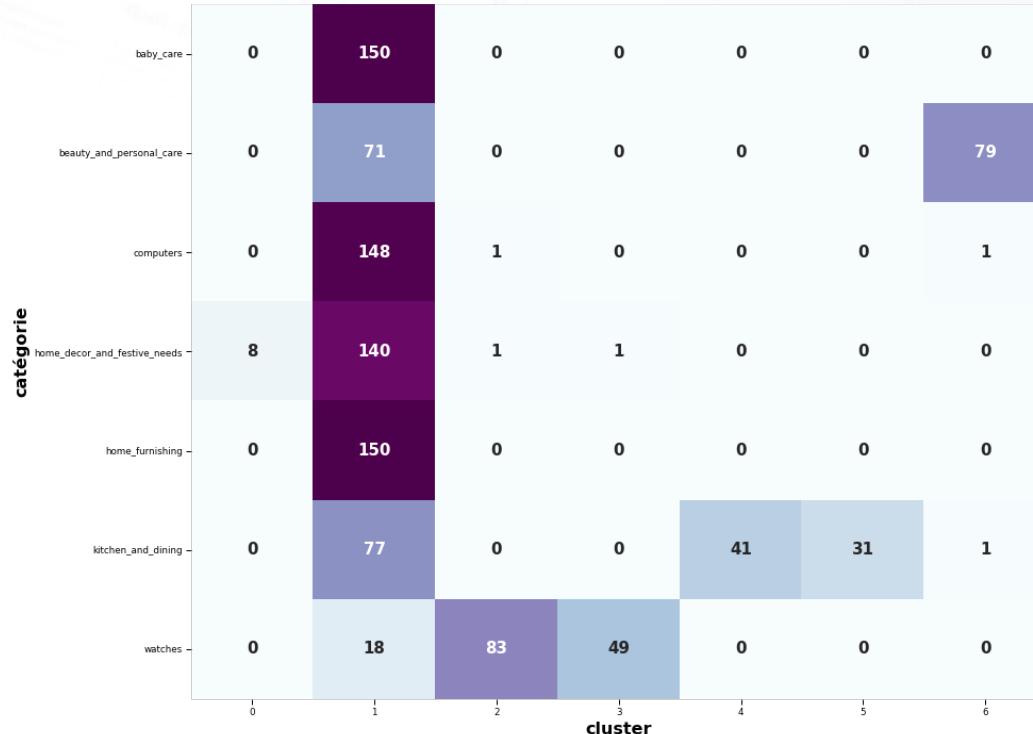
- Clustering : Algorithme k-means



Coeff silouhette : 0,035
ARI : 0,10

Catégorisation non supervisée

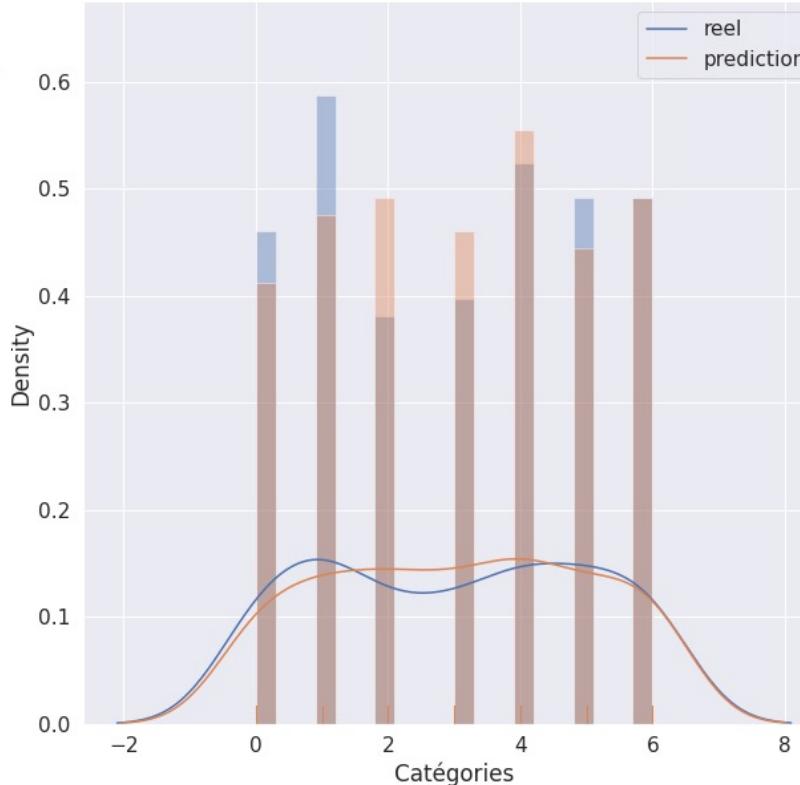
Répartition des clusters



Coeff silouhette : 0,035
ARI : 0,10

Catégorisation supervisée

- Classification : Algorithme Random Forest



0 = baby_care

1 = beauty_and_personal_care

2 = computers

3 = home_decor_and_festive_needs

4 = home_furnishing

5 = kitchen_and_dining

6 = watches

Accuracy : 0,87

Précision : 0,88

Catégorisation supervisée

Matrice de confusion

| Vraie catégorie | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|-----------------|--------|--------|--------|--------|--------|--------|---------|
| Prediction | 72.41% | 0.00% | 0.00% | 6.90% | 17.24% | 3.45% | 0.00% |
| 0 | 72.41% | 0.00% | 0.00% | 6.90% | 17.24% | 3.45% | 0.00% |
| 1 | 0.00% | 81.08% | 10.81% | 8.11% | 0.00% | 0.00% | 0.00% |
| 2 | 0.00% | 0.00% | 95.83% | 4.17% | 0.00% | 0.00% | 0.00% |
| 3 | 0.00% | 0.00% | 12.00% | 88.00% | 0.00% | 0.00% | 0.00% |
| 4 | 9.09% | 0.00% | 0.00% | 0.00% | 90.91% | 0.00% | 0.00% |
| 5 | 0.00% | 3.23% | 9.68% | 3.23% | 0.00% | 83.87% | 0.00% |
| 6 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 100.00% |

0 = baby_care

1 = beauty_and_personal_care

2 = computers

3 = home_decor_and_festive_needs

4 = home_furnishing

5 = kitchen_and_dining

6 = watches

Accuracy : 0,87

Precision : 0,88

Importance Description / Image

Image

- CNN supervisé
- Accuracy : 0,58
- Précision : 0,64
- Probabilité classification

Description

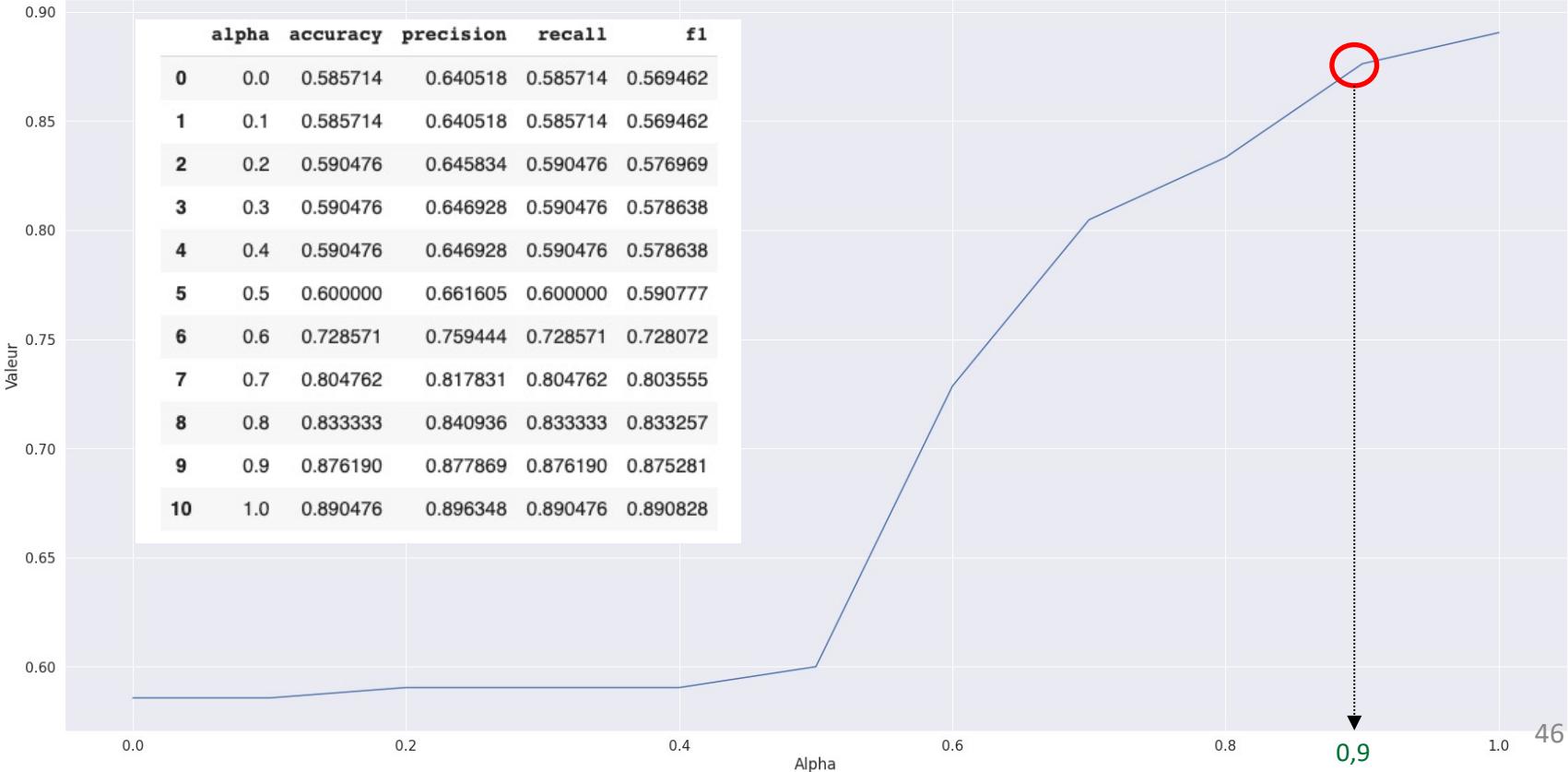
- Random Forest + ACP
- Accuracy : 0,890
- Précision : 0,896
- Probabilité classification

$\text{prob_combinée} = \alpha * \text{proba_description} + (1-\alpha) * \text{proba_image}$

Pour α allant de 0 à 1 :

- accuracy
- score précision

Importance Description / Image



Importance Description / Image

$\text{prob_img} = 0,9 * \text{proba_description_rf} + 0,1 * \text{proba_image_cnn}$

Conclusion

- Le Clustering permet bien de rassembler les articles entre eux (grâce à l'association description et image) mais les résultats et l'exactitude de celui-ci est trop peu précis.
- La Classification quant à elle donne de meilleurs résultats.
=> privilégier une catégorisation supervisée
- Catégorisation par descriptions : plus fiable
- Axe d'amélioration :
 - Photos de bonne qualité : améliorer détection descripteurs visuels
 - Enrichir base de données
 - Palier de l'accuracy non atteint : augmenter nombre d'epoch pour CNN

Fin de la présentation



Merci pour
votre attention

