CLASSIFICATION AUTOMATIQUE DES BIENS DE CONSOMMATION

Formation Data Scientist – Projet 6

Octave POUILLOT

Juillet 2023

SOMMAIRE

- Mission & Dataset
- Etude de faisabilité
- Classification supervisée
- API
- Conclusions

MISSION & DATASET

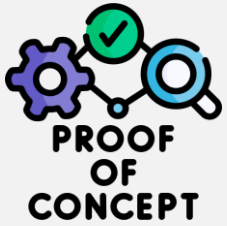
Place de marché - Marketplace e-commerce.

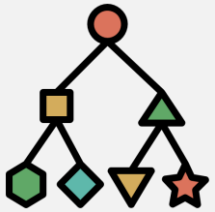
place de marché

MISSION

- Catégorisation automatique des produits (image & description)

- Etude de faisabilité d'un moteur de classification d'articles
  - Texte
  - Image
- Classification supervisée à partir des images
- Tester la collecte de produits via une API

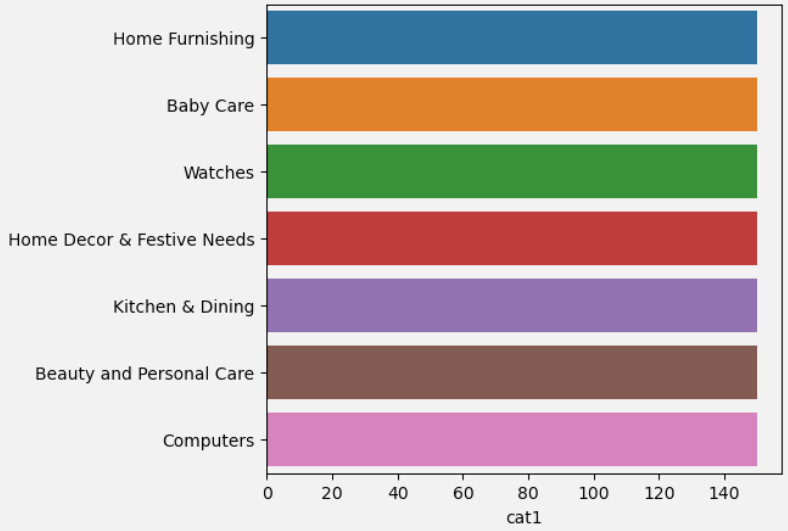




DATASET

- 1050 produits, 15 features
- 1050 images (une par produit)

- 7 catégories principales
(62 secondaires)



RAPPEL RGPD

- **Principe de finalité** : enregistrer que dans un but précis, légal et légitime
- **Principe de proportionnalité et de pertinence** : informations pertinentes et strictement nécessaires
- **Principe de durée de conservation limitée** : fonction du type d'information et de la finalité
- **Principe de sécurité et de confidentialité** : garantir la sécurité et l'accès limité aux informations
- **Droits des personnes** : Information, portabilité, opposition, …

➔ Pas de données personnelles

➔ Linda : « J'ai bien vérifié qu'il n'y avait aucune contrainte de propriété intellectuelle sur les données et les images. »

ETUDE DE FAISABILITÉ

Descriptions textuelles

Images

ETUDE DE FAISABILITÉ - GÉNÉRALE

Description	Image
5 modèles (Count, TF-IDF, Word2Vec, BERT, USE)	2 modèles (SIFT, CNN)
Prétraitement	
Extraction de features	
Réduction en 2 dimensions - T-SNE	
Classification non supervisée - KMeans	
Score ARI	

ETUDE DE FAISABILITÉ - DESCRIPTION PREPROCESSING

Tokenizer	• RegexpTokenizer • lower
Stop words	• Liste « english » • Possible ajout de mots
Lemmatizer Stemmer	• WordNetLemmatizer • PorterStemmer

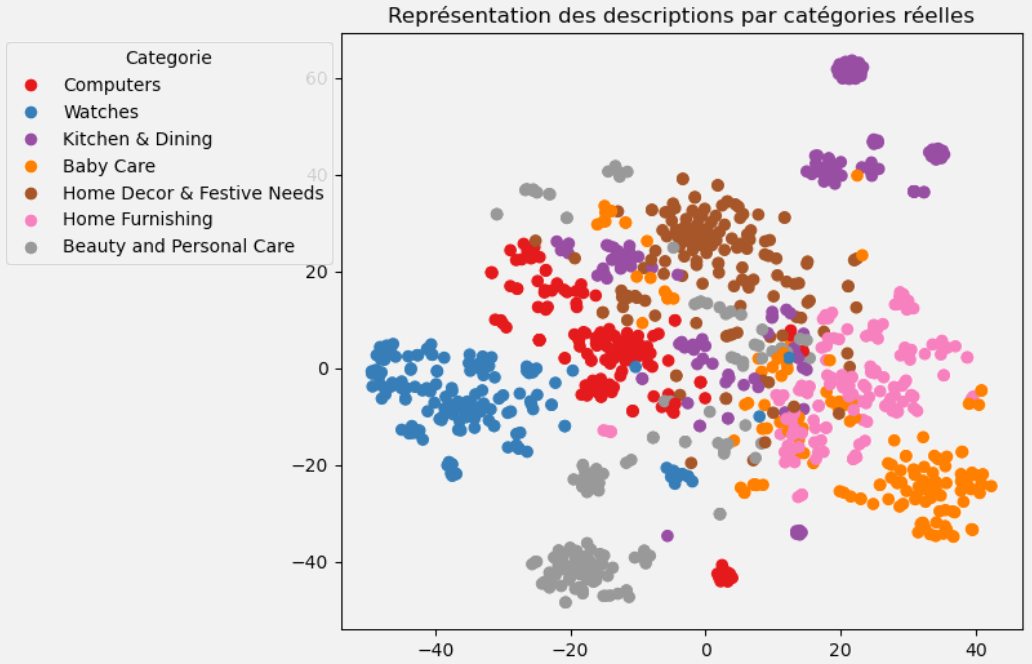
- 3 colonnes à partir de « description »
  - Bag-of-word AVEC lemmatisation

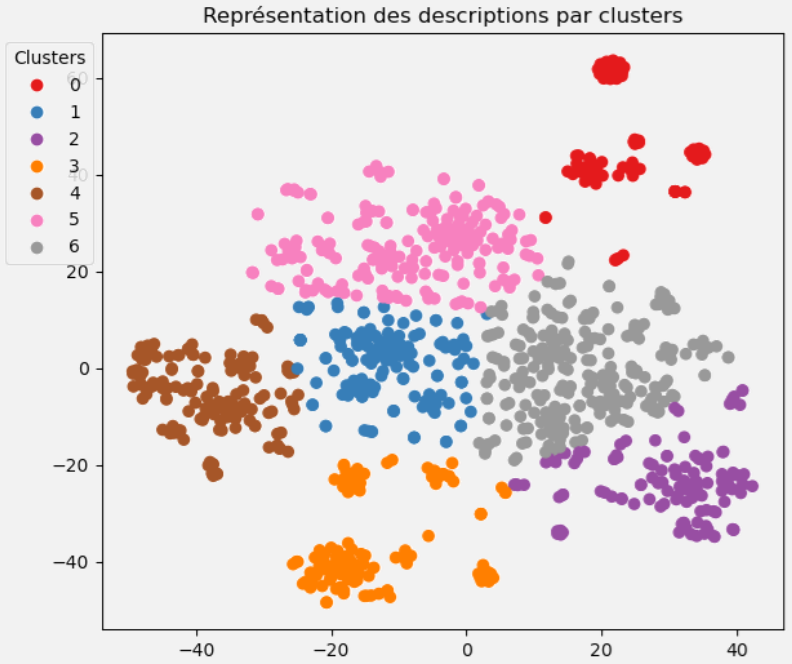
  - Bag-of-word SANS lemmatisation

  - Deep Learning

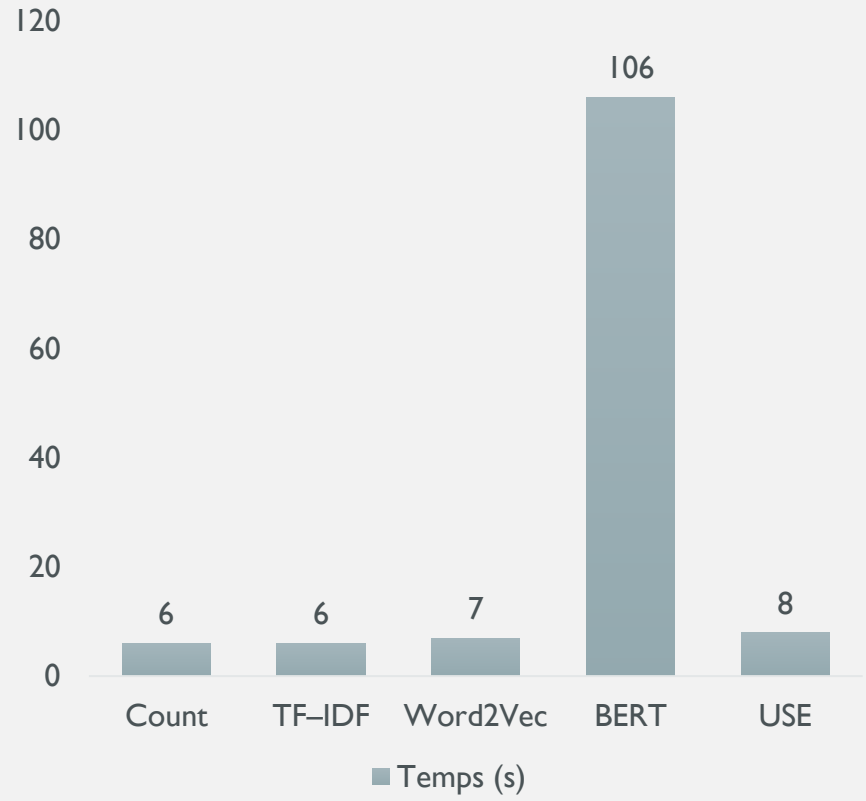
ETUDE DE FAISABILITÉ – DESCRIPTION EVALUATION

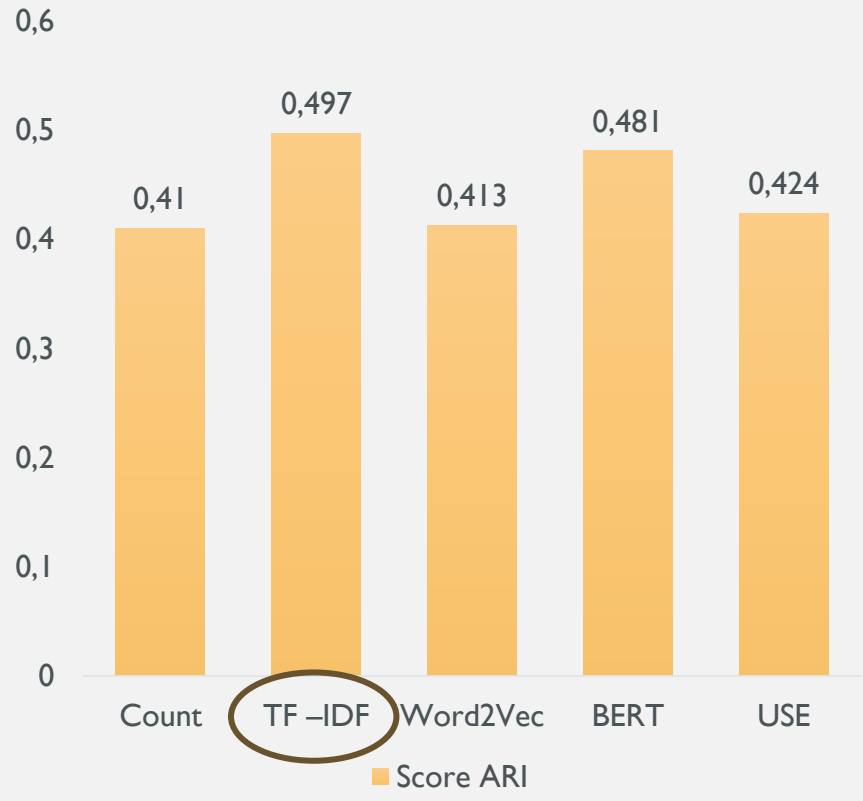
- Fonction « TSNE – KMEANS – ARI » :
  - Réduction en 2 dimensions
  - Kmeans avec k=7
  - ARI catégories vs clusters



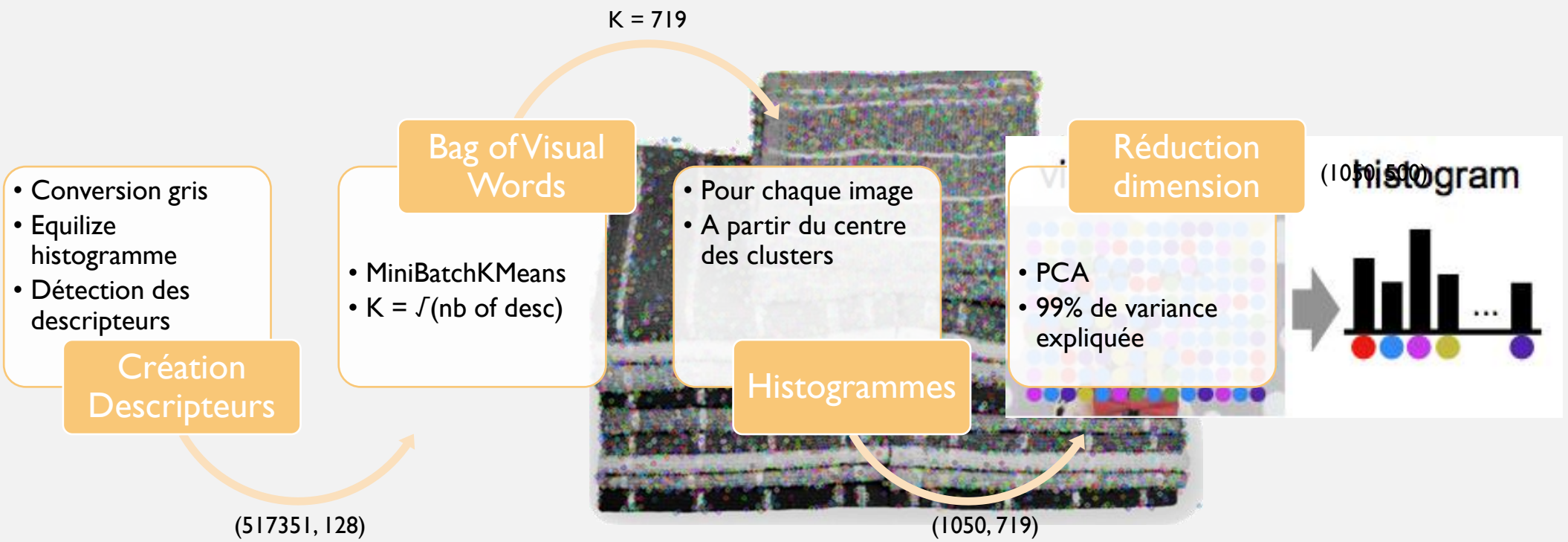


ETUDE DE FAISABILITÉ – DESCRIPTION RÉSULTATS



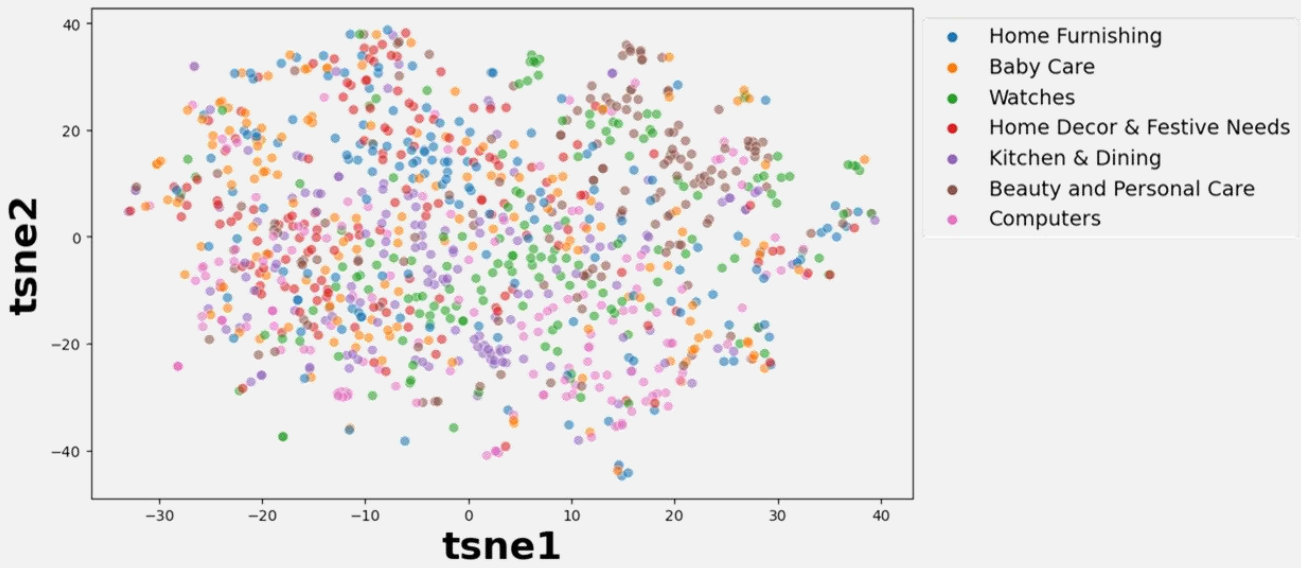


ETUDE DE FAISABILITÉ – IMAGE SIFT

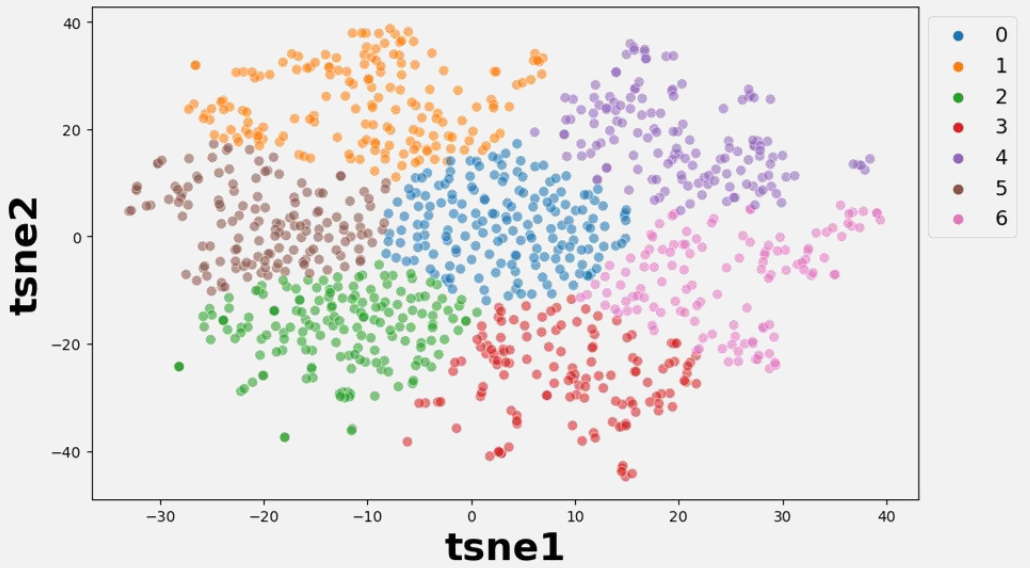


ETUDE DE FAISABILITÉ – IMAGE SIFT

TSNE SELON LES VRAIES CLASSES

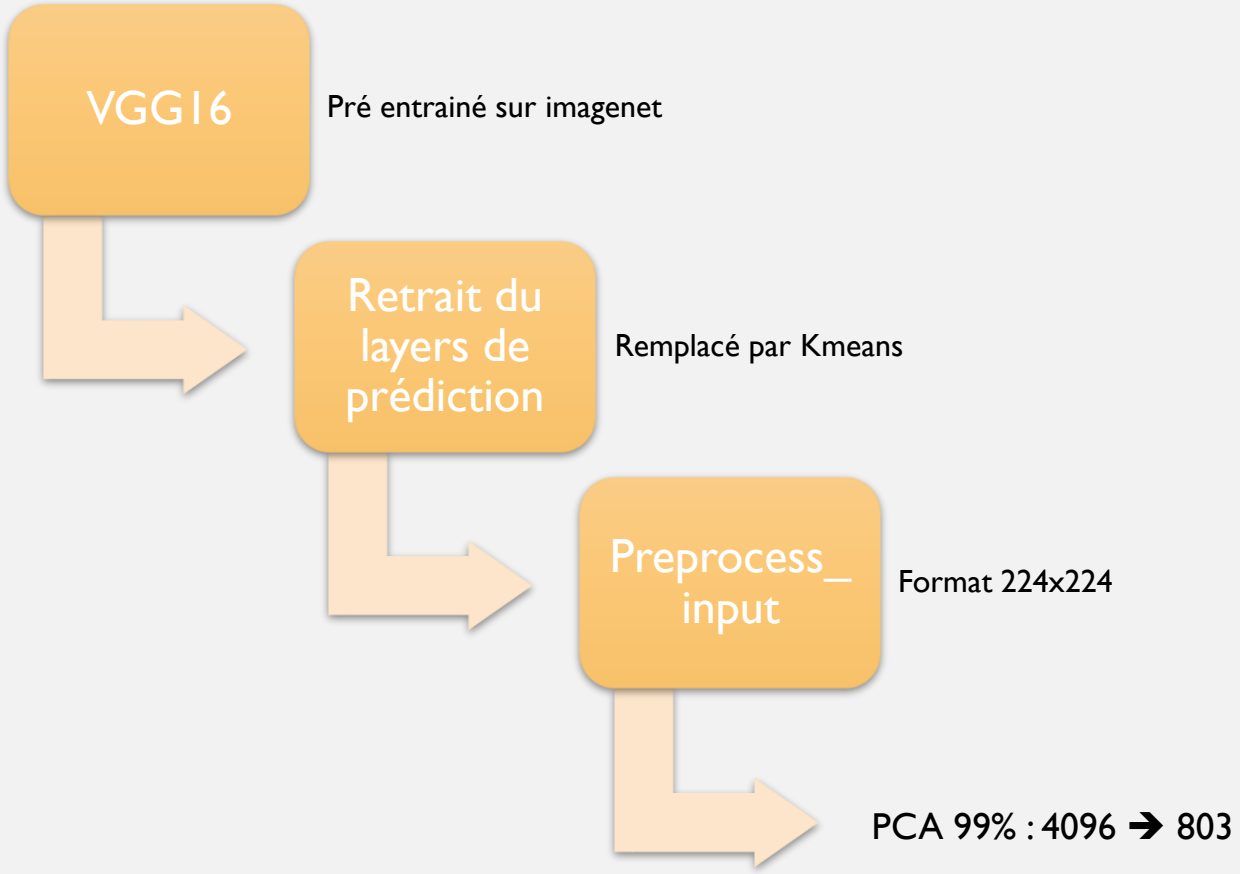


TSNE SELON LES CLUSTERS



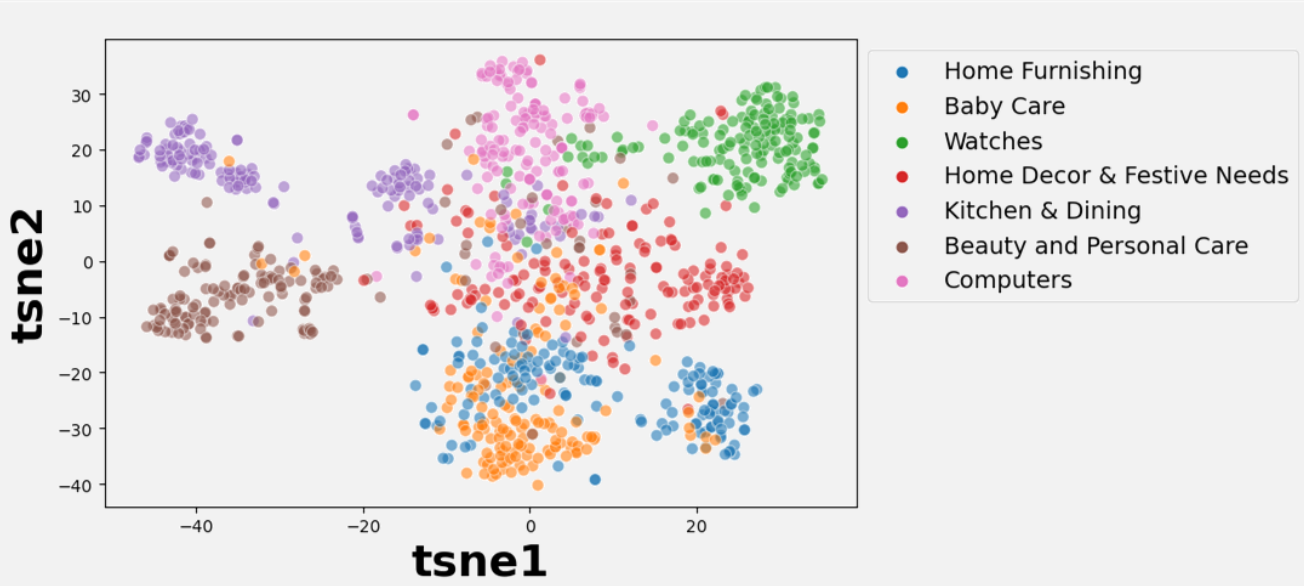
Score ARI : 0.0599

ETUDE DE FAISABILITÉ – IMAGE CNN

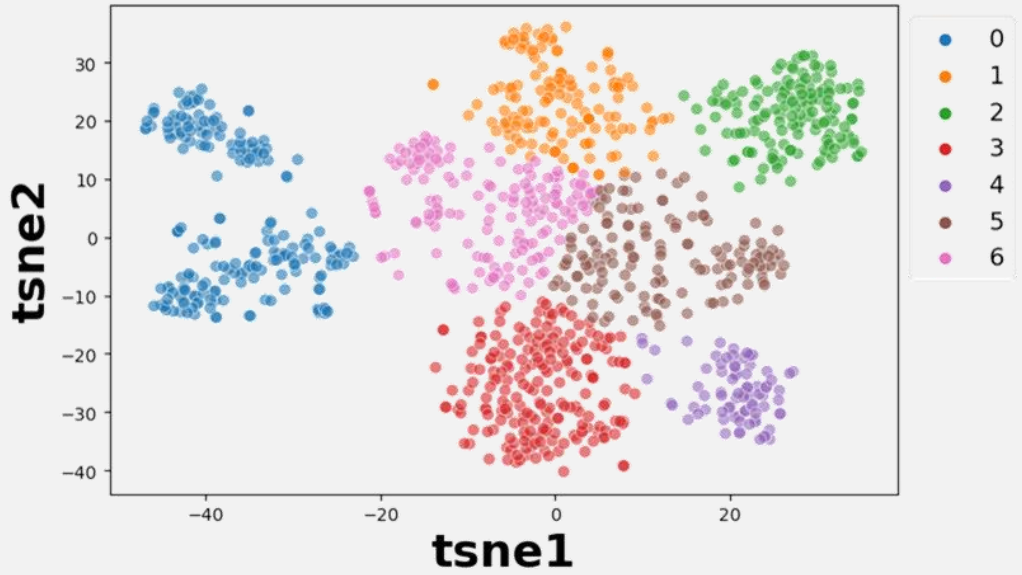


ETUDE DE FAISABILITÉ – IMAGE CNN

TSNE SELON LES VRAIES CLASSES

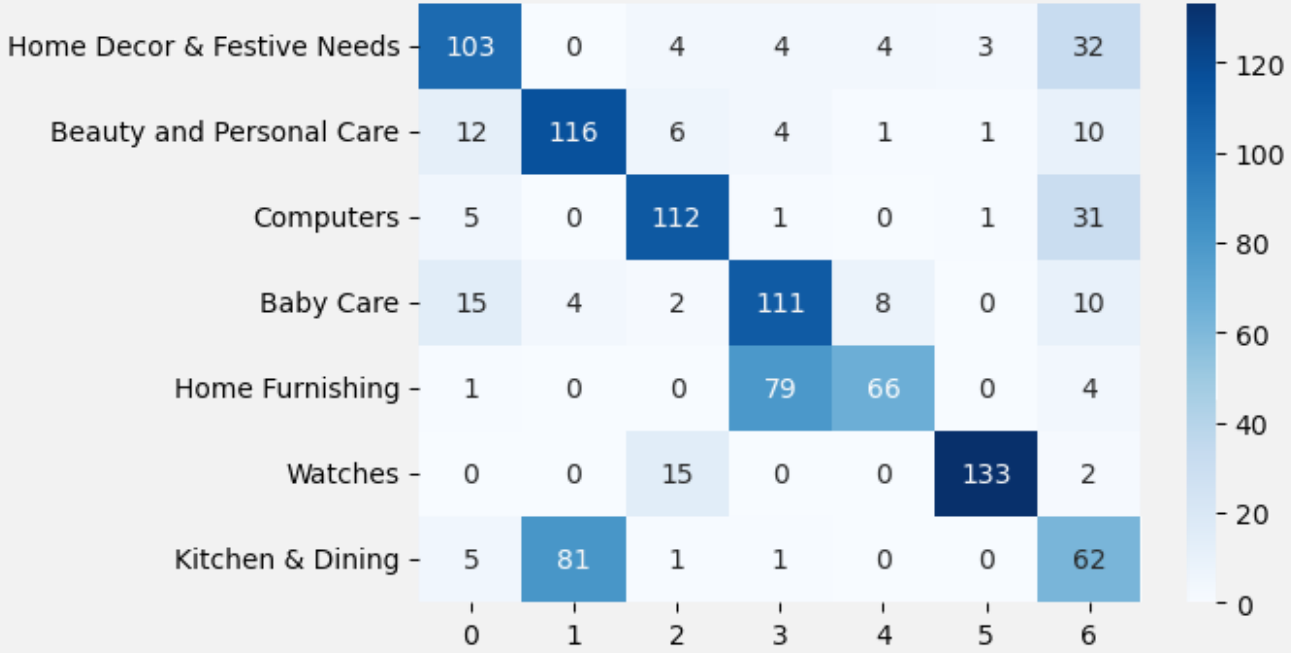


TSNE SELON LES CLUSTERS



Score ARI : 0.4777

ETUDE DE FAISABILITÉ – IMAGE CNN



	precision	recall	f1-score
0	0.73	0.69	0.71
1	0.58	0.77	0.66
2	0.80	0.75	0.77
3	0.56	0.74	0.63
4	0.84	0.44	0.58
5	0.96	0.89	0.92
6	0.41	0.41	0.41

Accuracy : 0.67

CLASSIFICATION SUPERVISÉE

Images

CNN Transfer Learning

Data Augmentation

CLASSIFICATION SUPERVISÉE
STRATÉGIE

4 approches :

- Simple : préparation initiale de l'ensemble des images avant classification supervisée
- Par data generator : Même approche avec Data Augmentation
- Par DataSet : Préparation initiale par image_dataset_from_directory
- Par DataSet avec Data Augmentation (intégrée au modèle)

CLASSIFICATION SUPERVISÉE STRATÉGIE

- Séparation du dataset :
- Hyperparamètres :
  - Batch size
  - Epoch
  - Train / val / test size

CLASSIFICATION SUPERVISÉE
MODÈLE

VGG16 avec poids « imagenet »

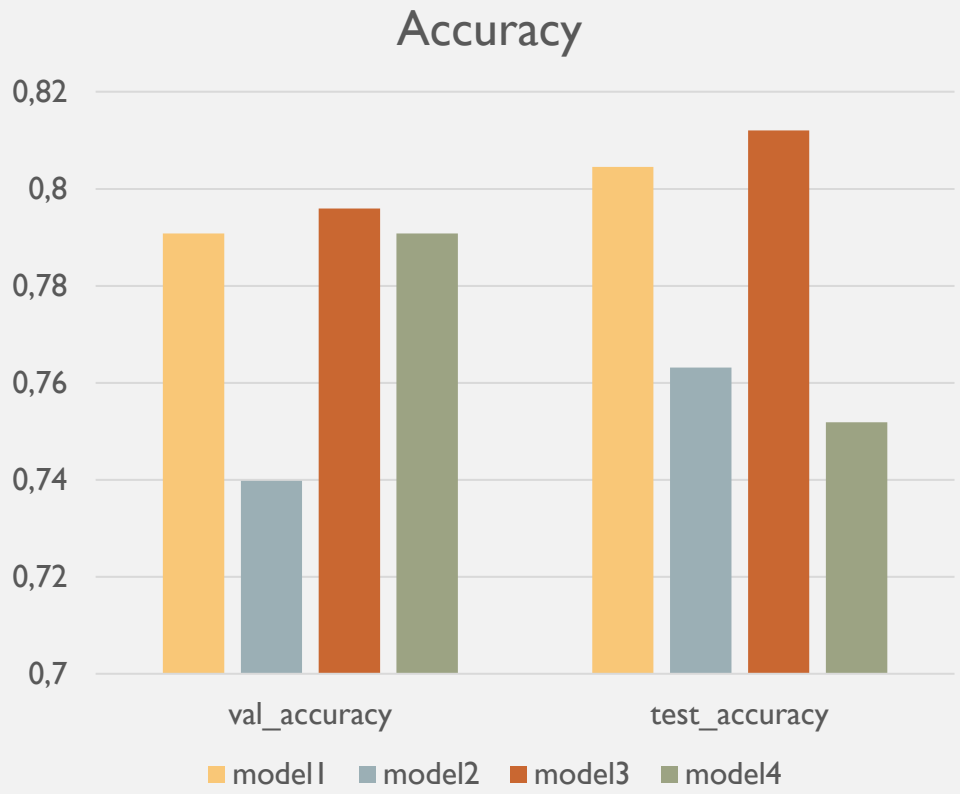
Retrait des layers fully-connected

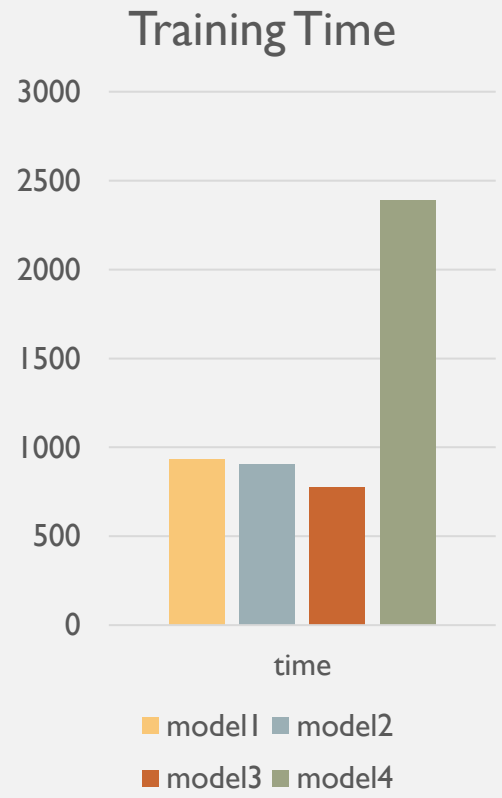
Ajout :

- GlobalAveragePooling2D()(x)
- Dense(256, activation='relu')(x)
- Dropout(0.5)(x)
- Dense(7, activation='softmax')(x)

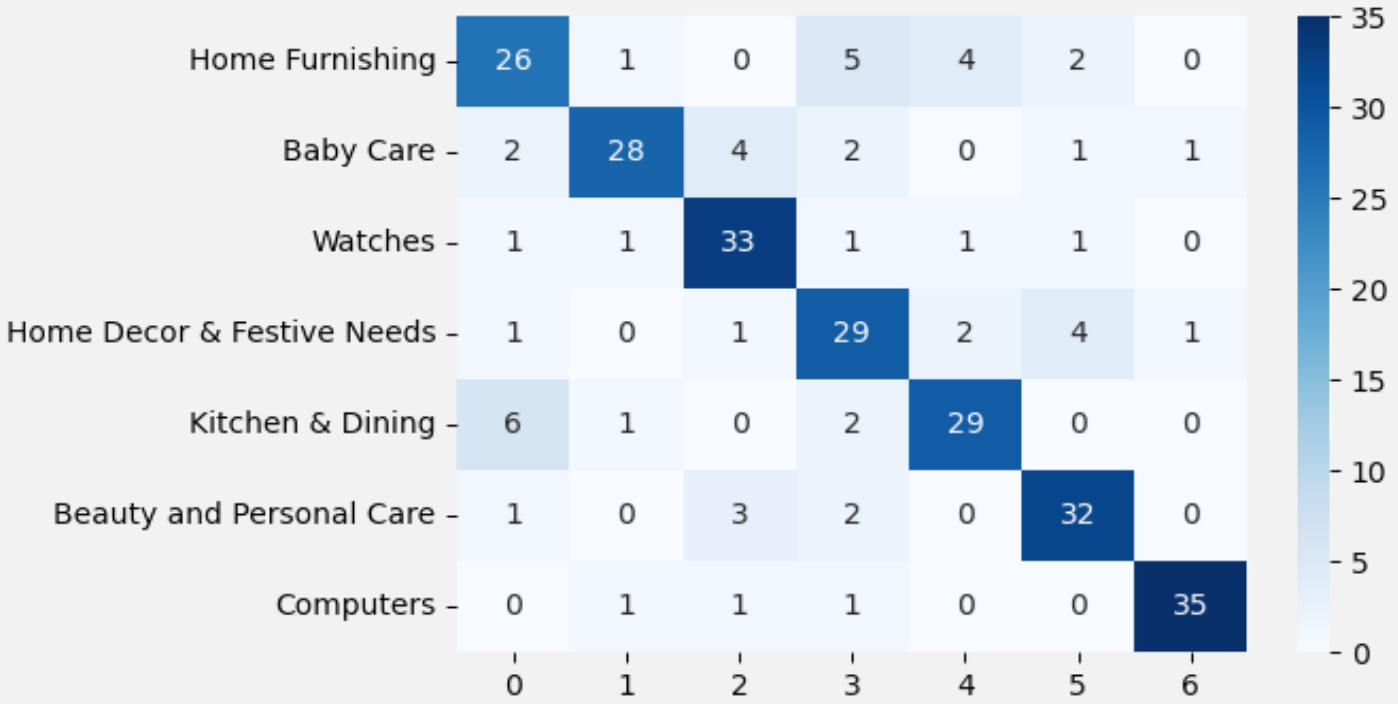
- Métriques :
  - Accuracy & Loss
  - Validation & Test
  - Training Time

CLASSIFICATION SUPERVISÉE
RÉSULTAT





ETUDE DE FAISABILITÉ – IMAGE CNN



	precision	recall	f1-score
0	0.70	0.68	0.69
1	0.88	0.74	0.80
2	0.79	0.87	0.82
3	0.69	0.76	0.72
4	0.81	0.76	0.78
5	0.80	0.84	0.82
6	0.95	0.92	0.93

Accuracy : 0.80

Rapid

place de marché

TEST API

Collecte de produits à base de “champagne”

TEST API

- Requet vers API avec query « ingr":"['champagne’] »
- Récupération Json
- Focus sur « hints / food »
- Filtre sur les colonnes demandées : 'foodId', 'label', 'category', 'foodContentsLabel', 'image’
- Enregistrement en csv avec tabulation comme séparateur

CONCLUSION

CONCLUSION

- Faisabilité
- Classification supervisée :
  - Modèle : CNN avec image_dataset_from_directory
  - ~~Data augmentation~~
- Test de l'API

MERCI DE VOTRE ATTENTION !

Des questions ?