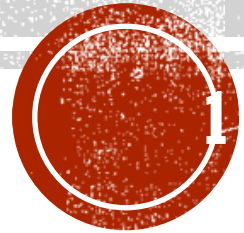
CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

Etude de la faisabilité d'un moteur de classification



Formation Data Scientist Projet 6
Openclassrooms / Res Entreprise
Tulomba Divine
Octobre 2023

SOMMAIRE

1. **Introduction, contexte & présentation de la problématique**
2. **Analyse & préparation du jeu de données**
3. **Faisabilité classification Texte (Bags of Words et Text Embedding)**
4. **Faisabilité classification Image (Sift et Transfert Learning)**
5. **Classification supervisée avec ou sans Data Augmentation**
6. **Présentation test API**
7. **Conclusion**

INTRODUCTION

- Objectif:
  - une étude de faisabilité d'un moteur de classification
  - classification supervisée et non supervisée à partir des images
  - tester la collecte de produits à base de “champagne” via API
- Données:
  - informations sur les produits disponibles sur le site e-commerce
  - API
- Approche:
  - **Analyse**
  - **Nettoyage**
  - **Modèle**
  - **Teste**

Flipkart

DESCRIPTION DU JEU DE DONNÉES

- **Contexte** : informations sur les produits disponibles sur le site e-commerce Flipkart

- **Nombre de produits** : 1050

- **Caractéristiques des produits** : des produits un identifiant unique, le nom du produit, la catégorie, la marque, le prix de détail, le prix réduit, etc.

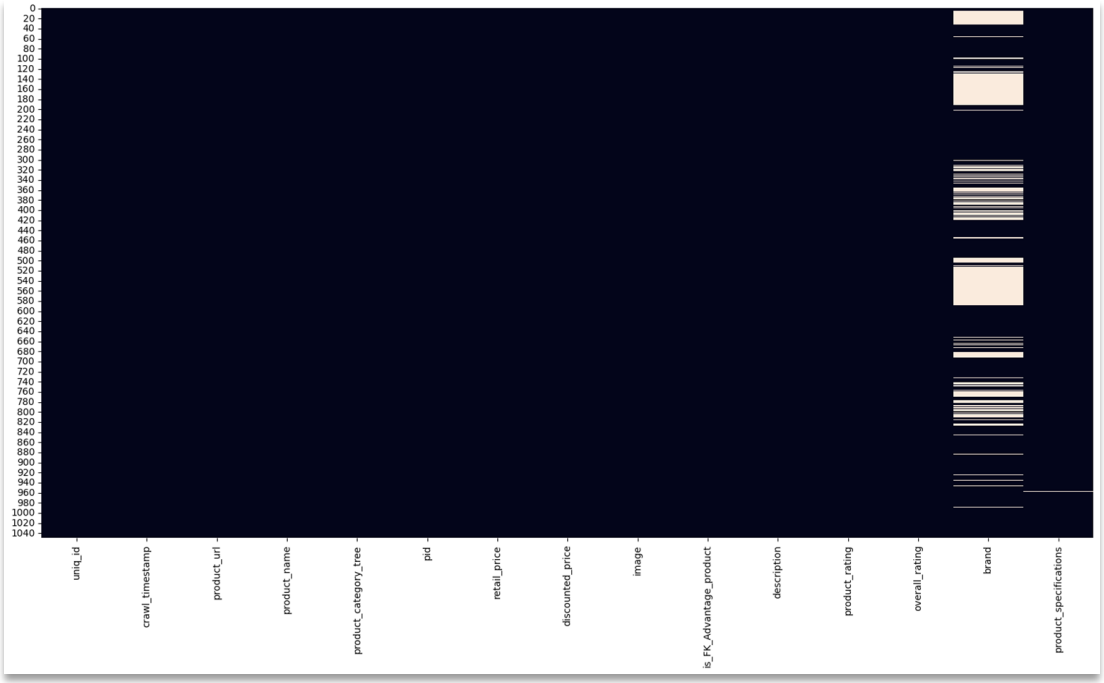
- **Descriptif des colonnes** : 15 colonnes

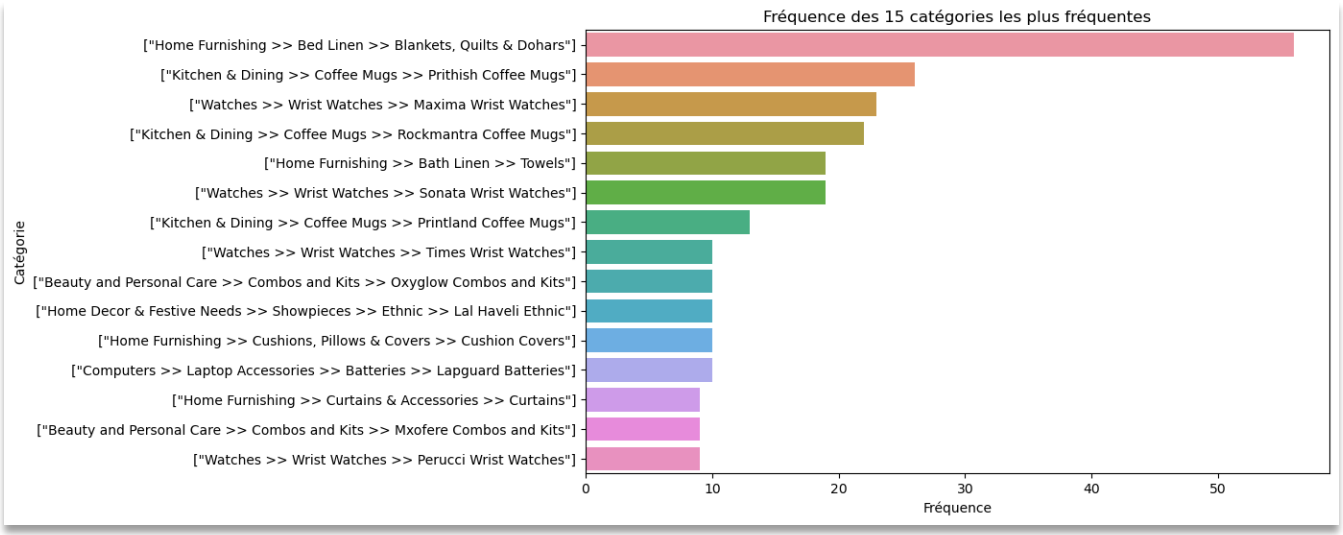
- **Échantillon de données :**

	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid	retail_pric
0	55b85ea15a1536d46b7190ad6fff8ce7	2016-04-30 03:22:56 +0000	http://www.flipkart.com/elegance-polyester-mul...	Elegance Polyester Multicolor Abstract Eyelet ...	["Home Furnishing >> Curtains & Accessories >>...	CRNEG7BKMFFYHQ8Z	1899
1	7b72c92c2f6c40268628ec5f14c6d590	2016-04-30 03:22:56 +0000	http://www.flipkart.com/sathiyas-cotton-bath-t...	Sathiyas Cotton Bath Towel	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEGFZHGBXPHZUH	600
2	64d5d4a258243731dc7bbb1eef49ad74	2016-04-30 03:22:56 +0000	http://www.flipkart.com/eurospa-cotton-terry-f...	Eurospa Cotton Terry Face Towel Set	["Baby Care >> Baby Bath & Skin >> Baby Bath T...	BTWEG6SHXTDB2A2Y	Na
3	d4684dcdc759dd9cdf41504698d737d8	2016-06-20 08:49:52 +0000	http://www.flipkart.com/santosh-royal-fashion-...	SANTOSH ROYAL FASHION Cotton Printed King size...	["Home Furnishing >> Bed Linen >> Bedsheets >>...	BDSEJT9UQWHDUBH4	2699
4	6325b6870c54cd47be6ebfbffa620ec7	2016-06-20 08:49:52 +0000	http://www.flipkart.com/jaipur-print-cotton-fl...	Jaipur Print Cotton Floral King sized Double B...	["Home Furnishing >> Bed Linen >> Bedsheets >>...	BDSEJTHNGWVGWWQU	2599

ANALYSE RAPIDE DU JEU DE DONNÉES

	colonne	valeurs manquantes	pourcentage
13	brand	338	32.190476
6	retail_price	1	0.095238
7	discounted_price	1	0.095238
14	product_specifications	1	0.095238
0	uniq_id	0	0.000000
1	crawl_timestamp	0	0.000000
2	product_url	0	0.000000
3	product_name	0	0.000000
4	product_category_tree	0	0.000000
5	pid	0	0.000000
8	image	0	0.000000
9	is_FK_Advantage_product	0	0.000000
10	description	0	0.000000
11	product_rating	0	0.000000
12	overall_rating	0	0.000000





Base de données au niveau de texte :

- Beaucoup de données manquantes au niveau de la variable « Brand » et peu sur product_specifications
- Les catégories les plus présentes dans notre base sont les produits de maison et literie.

Home Furnishing

Baby Care

Watches

Home Decor & Festive Needs

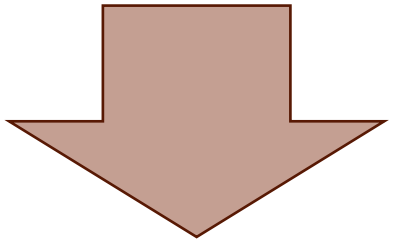
Kitchen & Dining

Beauty and Personal Care

Computers

CLASSIFICATION DESCRIPTION : FEATURES EXTRACTION

TF-IDF	Word2Vec	BERT	USE
▪ Ajoute un poids pour les mots les plus utilisés	▪ Prend en compte la sémantique des mots	▪ Pré-entrainé ▪ Utilise le texte dans les 2 sens	▪ Pré-entrainé ▪ Fonctionne sur les phrases



▪ Réduction 1 PCA → ▪ Réduction 2 T-SNE → ▪ Clustering K-means → ▪ Calcul Score ARI
▪ Visualisation TSNE

3. FAISABILITÉ DE CLASSIFICATION – TEXTE

NETTOYAGE DES FEATURES TEXTUELLES

❑ **Tokenisation** :

```
# Exemple d'utilisation
example_sentence = "Le chat dort."
tokenized_sentence = tokenize_sentence(example_sentence)
tokenized_sentence
```

```
['Le', 'chat', 'dort', '.']
```

❑ **Stemming / Lemmatisaion:**

```
# Exemple d'utilisation
example_sentence = "The cats are running and jumping."
stemmed_sentence = stem_sentence(example_sentence)
stemmed_sentence
```

```
'the cat are run and jump .'
```

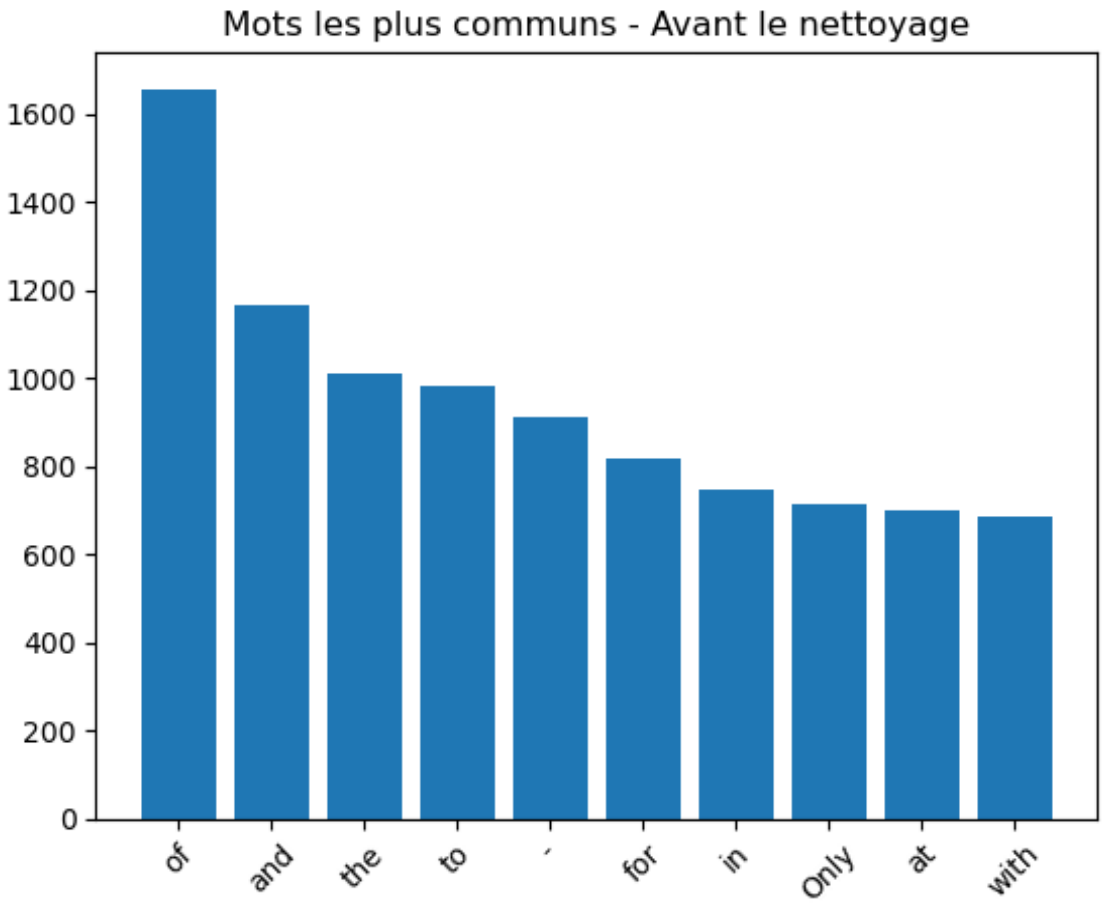
❑ **Suppression ponctuation / conversion minuscule :**

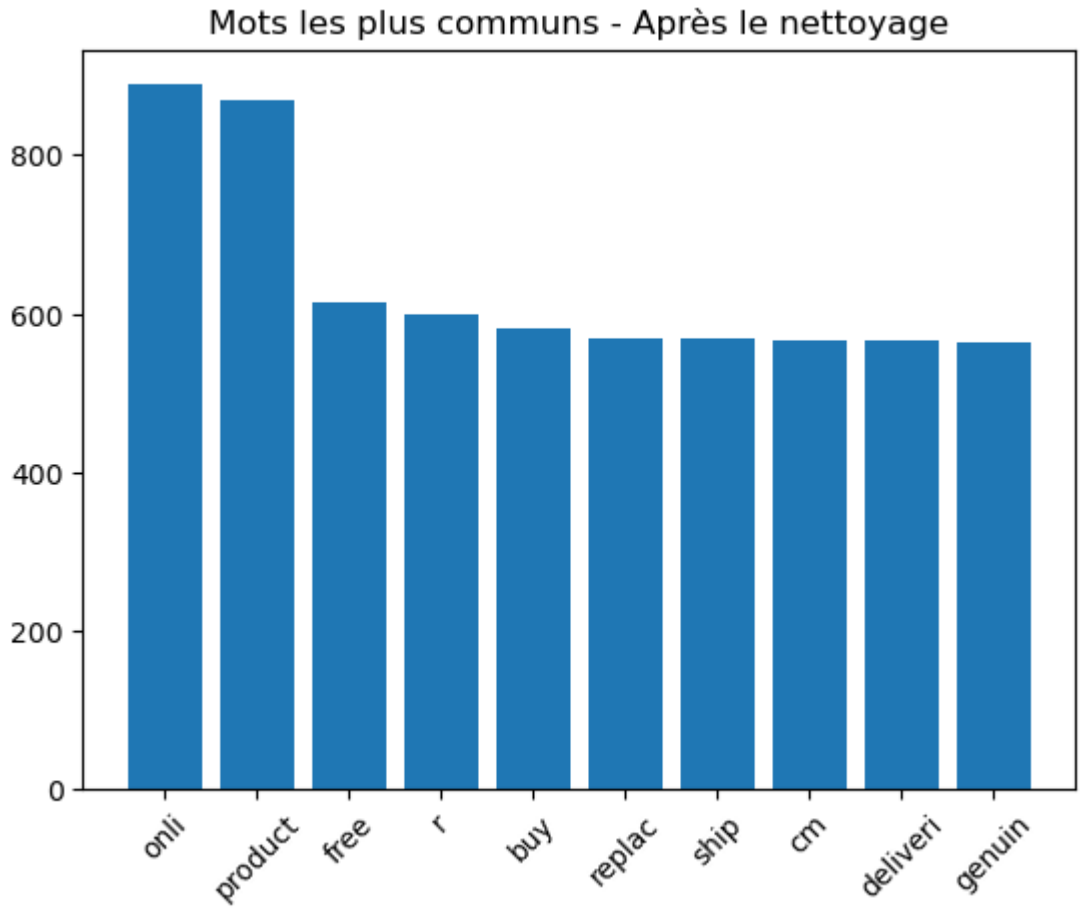
```
# Application de la fonction de nettoyage à la colonne de description
df['cleaned_description'] = df['description'].apply(clean_text)
# Application de la fonction de nettoyage à la colonne de nom du produit
df['cleaned_product_name'] = df['product_name'].apply(clean_text)

# Affichage des résultats
df[['description', 'cleaned_description','product_name','cleaned_product_name']].head()
```

	description	cleaned_description
0	Key Features of Elegance Polyester Multicolor ...	key featur eleg polyest multicolor abstract ey...
1	Specifications of Sathiyas Cotton Bath Towel (...	specif sathiya cotton bath towel 3 bath towel...
2	Key Features of Eurospa Cotton Terry Face Towe...	key featur eurospa cotton terri face towel set...
3	Key Features of SANTOSH ROYAL FASHION Cotton P...	key featur santosh royal fashion cotton print ...
4	Key Features of Jaipur Print Cotton Floral Kin...	key featur jaipur print cotton floral king siz...

AVANT / APRÈS NETTOYAGE

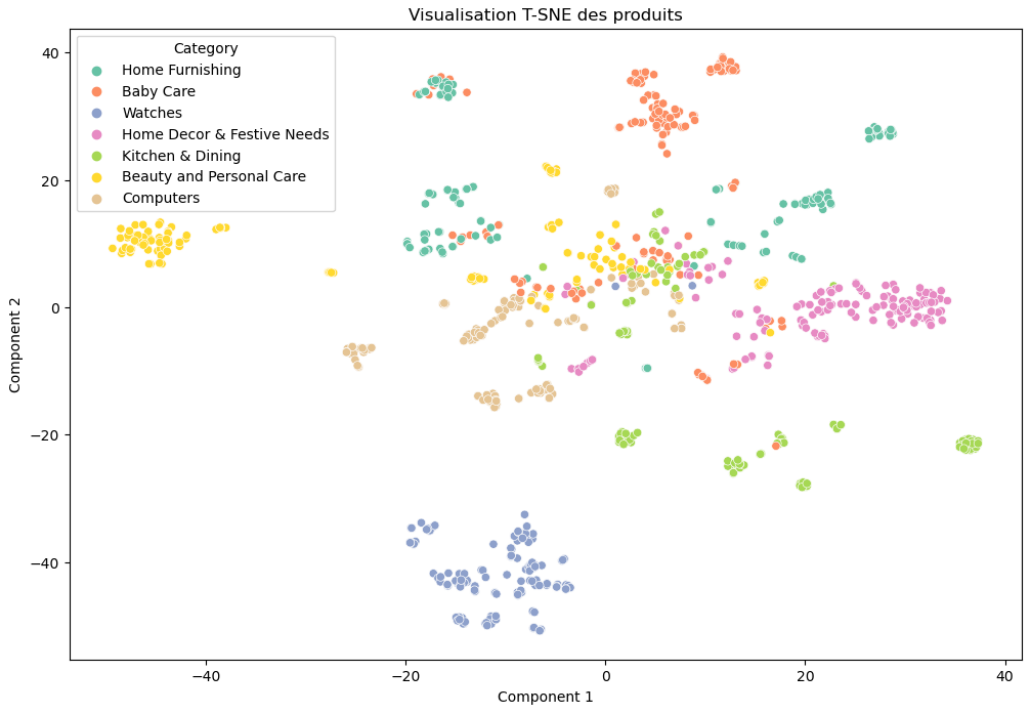
Il n'y a plus de stopwords, mais uniquement des mots liés aux descriptions des biens

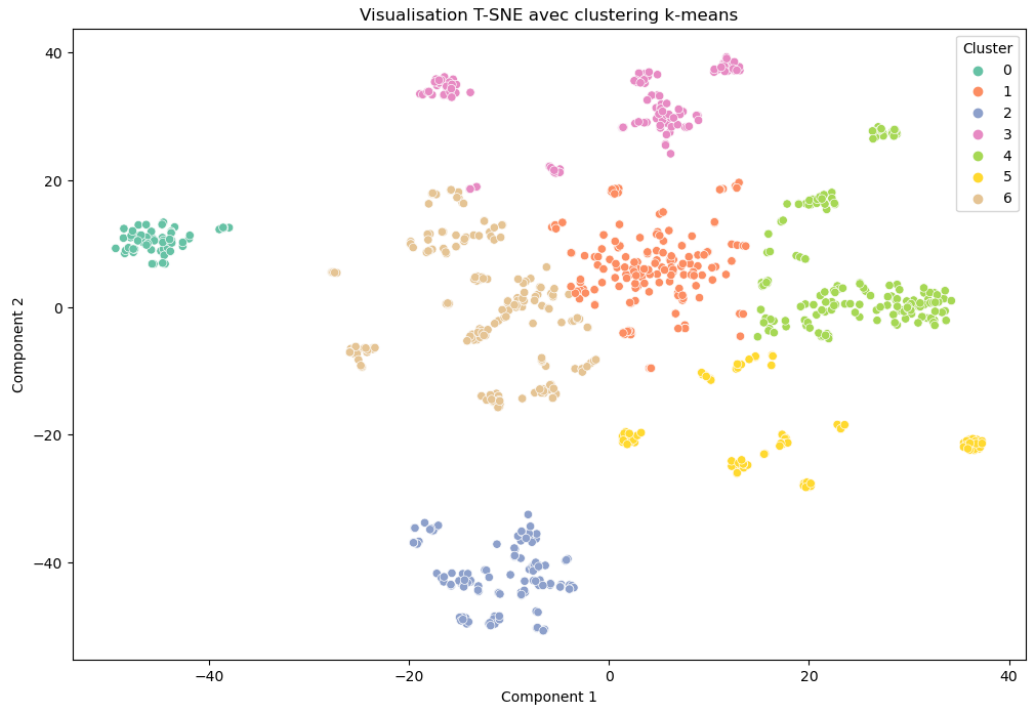




Faisabilité de classification – Texte
Bag-of-Words

- Fit et Transform sur « product_name » + « description » TfidVectorizer






```
Adjusted Rand Score: 0.3988821944927195
```

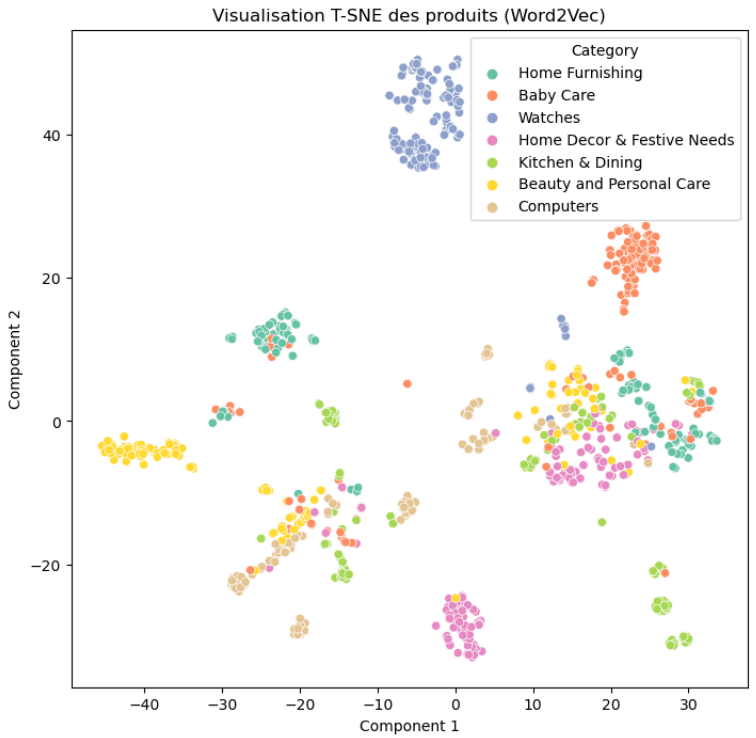
MÉTHODOLOGIE ADOPTÉE POUR LA CLASSIFICATION NON SUPERVISÉE DE DONNÉES TEXTUELLES

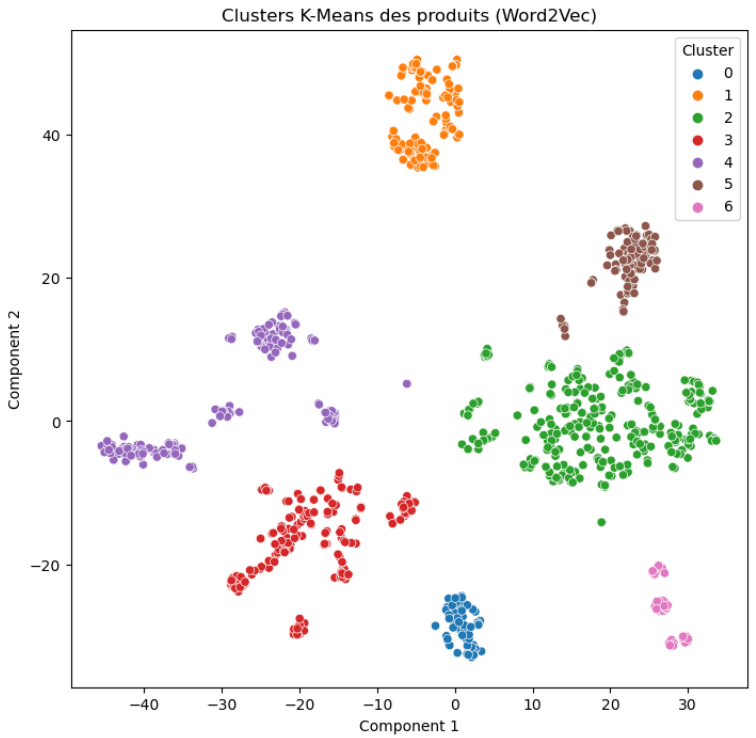
TEXT EMBEDDING

transformer du text brut non structuré et difficile à traiter par les modeles d'IA, en représentations structurées

3 outils : Word2Vec / BERT / USE

Word2Vec

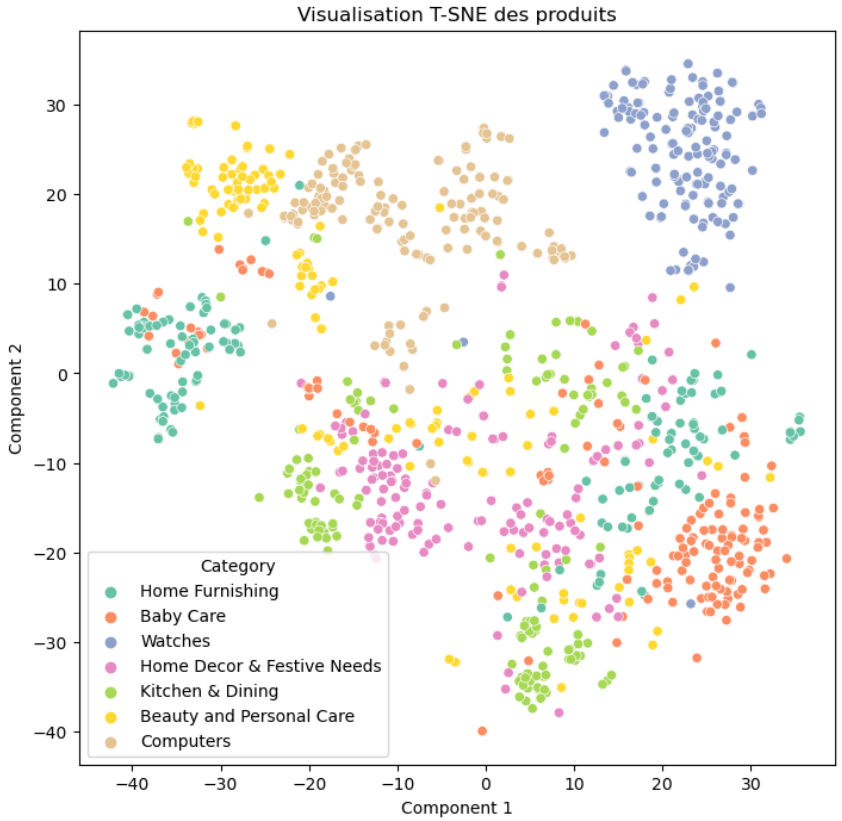


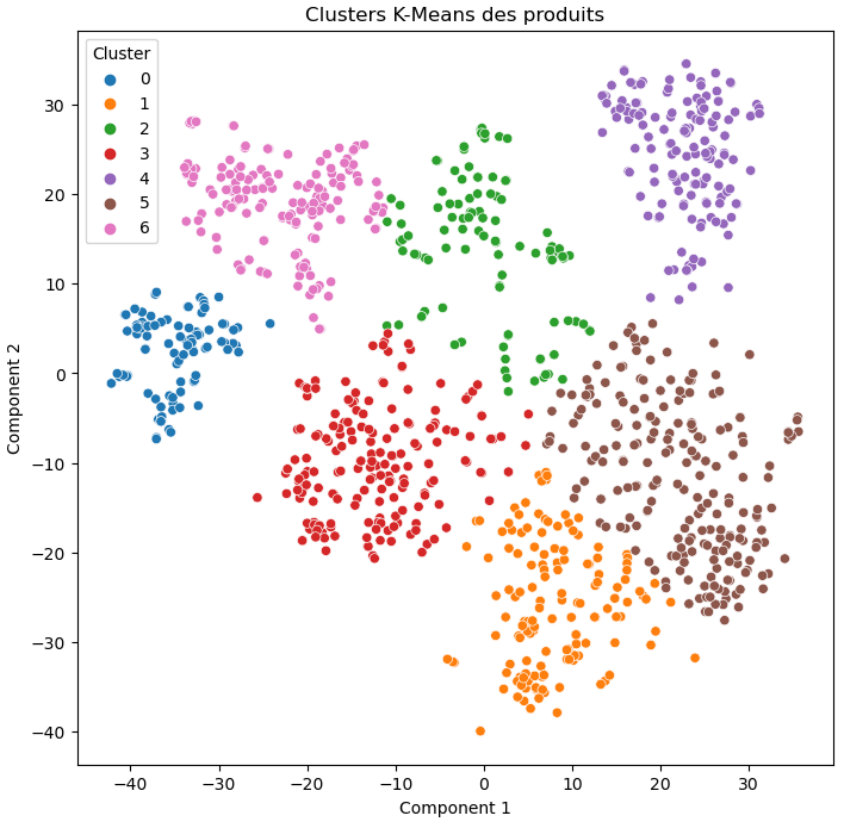



```
Adjusted Rand Score (Word2Vec): 0.2708816939510262
```

Ordre de grandeur de 0.3, la classification est faisable.

BERT

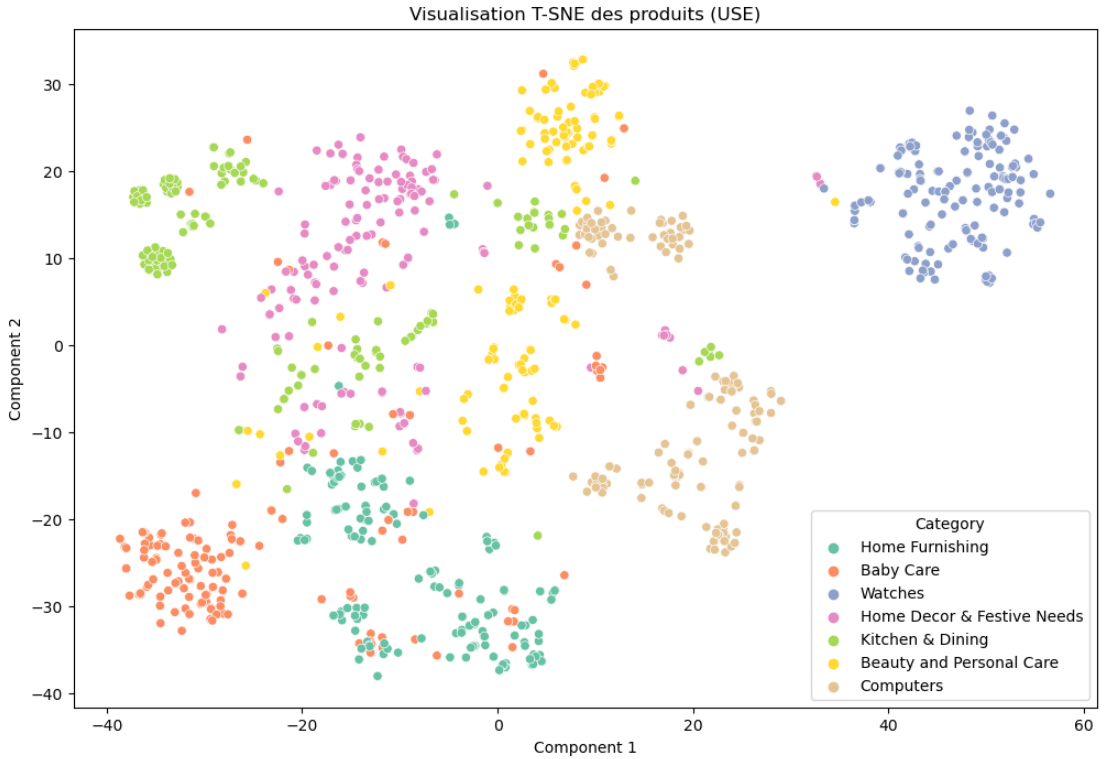


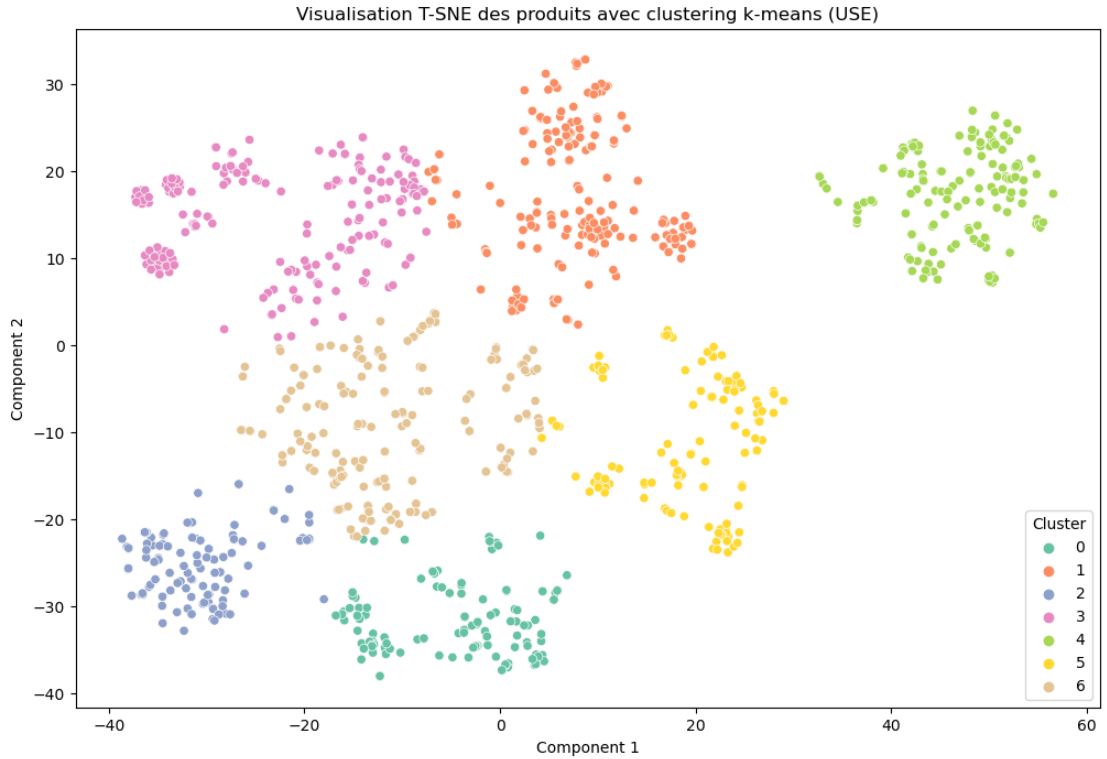



```
Adjusted Rand Score: 0.36237013040045846
```

Ordre de grandeur de 0.3, la classification est faisable.

USE






```
Adjusted Rand Score (USE): 0.43914335331065046
```

Ordre de grandeur de 0.4, la classification est faisable.

4.MÉTHODOLOGIE ADOPTÉE POUR LA CLASSIFICATION NON SUPERVISÉE DE DONNÉES IMAGES

ANALYSE D'UNE IMAGE ET DIFFÉRENTES APPROCHES DE TRANSFORMATION

Image Originale

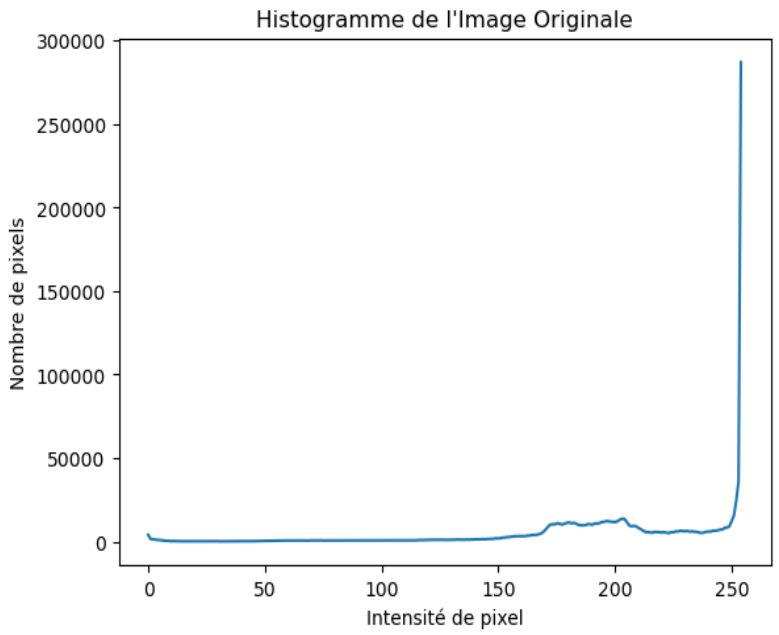
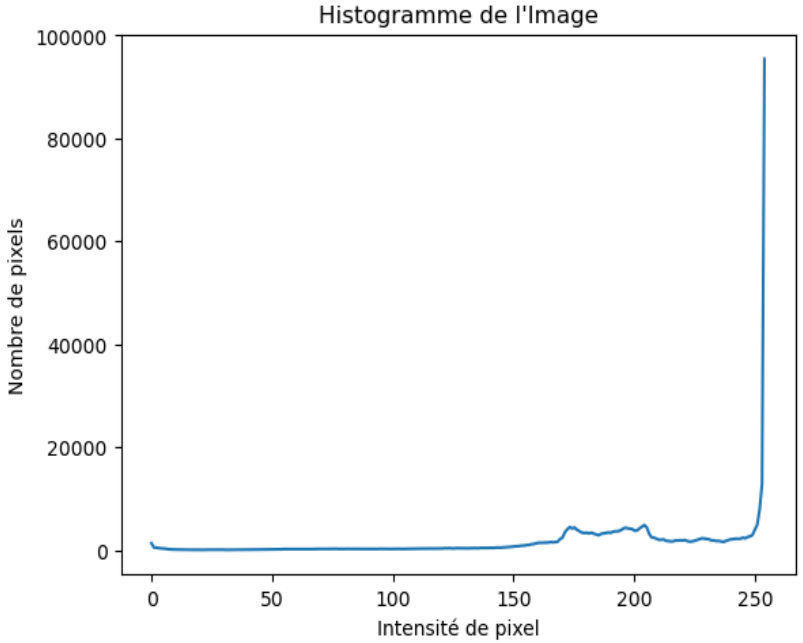


Image en Niveaux de Gris



FAISABILITÉ DE CLASSIFICATION AUTOMATIQUE D'IMAGES

ANALYSE D'UNE IMAGE ET DIFFÉRENTES APPROCHES DE TRANSFORMATION

Image avec Égalisation d'Histogramme

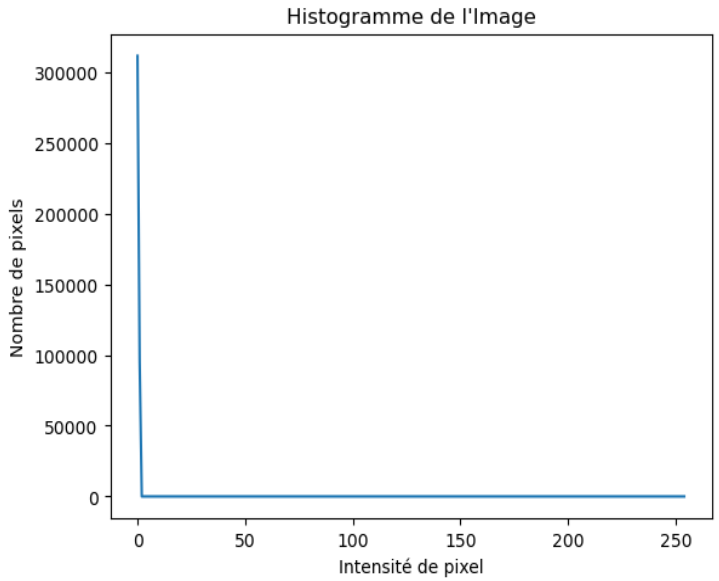
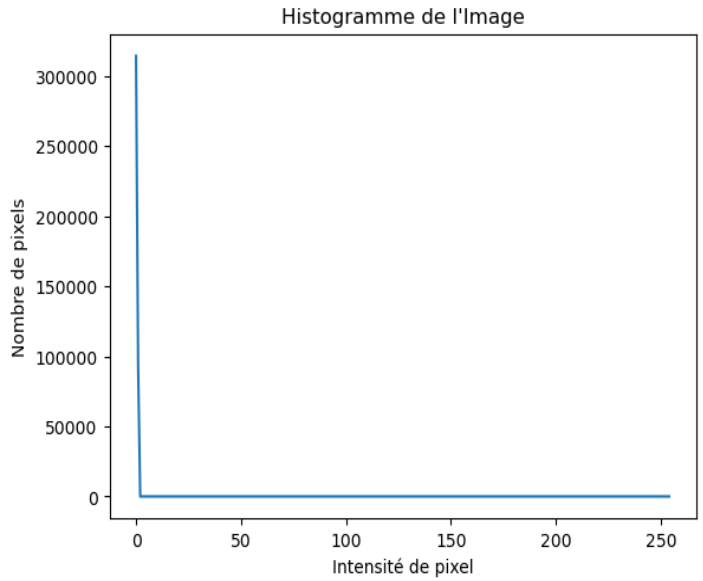


Image avec Filtrage du Bruit



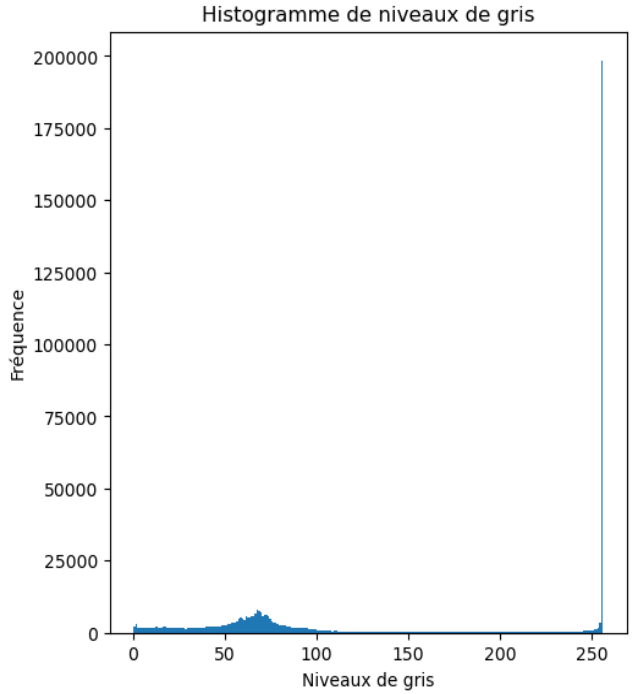
FAISABILITÉ DE CLASSIFICATION AUTOMATIQUE D'IMAGES

ANALYSE D'UNE IMAGE ET DIFFÉRENTES APPROCHES DE TRANSFORMATION

Image Originale

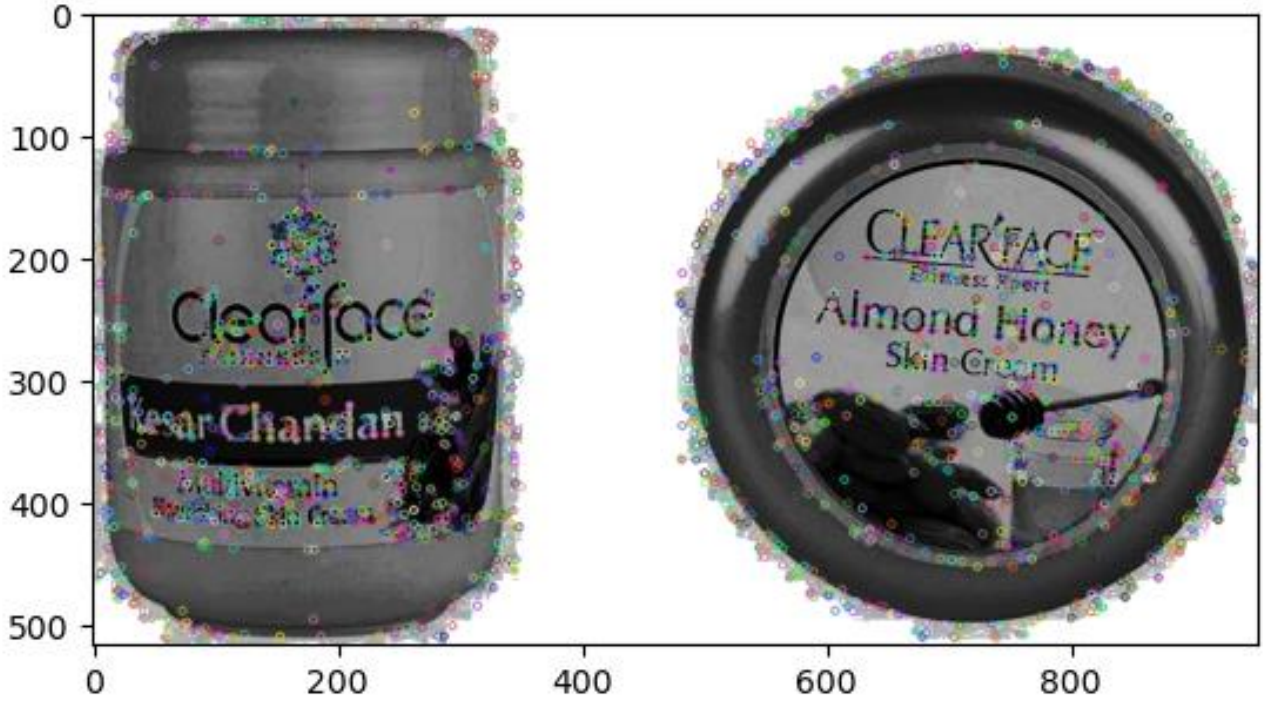
Image avec Contraste Amélioré

Image originale



FAISABILITÉ DE CLASSIFICATION AUTOMATIQUE D'IMAGES

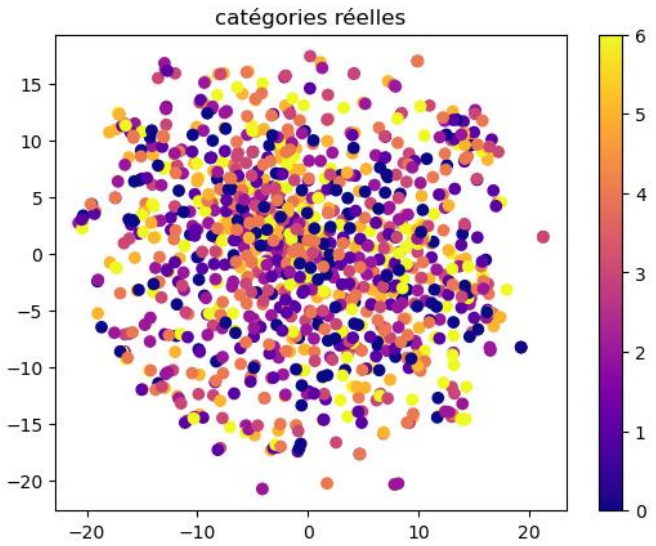
Le Sift détecte des points d'intérêt, construis de l'échelle de l'image, détecte de l'orientation, décris des caractéristiques.

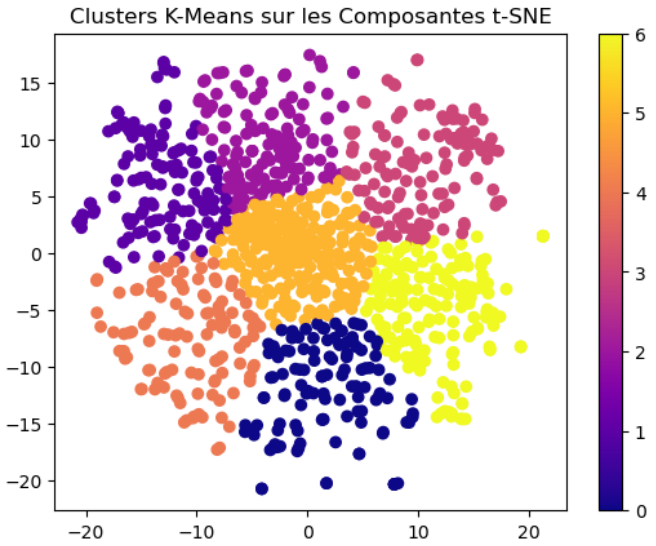



```
Descripteurs :  (2387, 128)

[[124.  47.  26. ...   0.   0.   0.]
 [  1.   2.   0. ...   0.   0.   3.]
 [  0.   0.   0. ...  20.   1.   7.]
 ...
 [  0.   0.   0. ...   2.   1.   2.]
 [  0.   0.   2. ...  51.   2.   0.]
 [  0.   0.   0. ...  70.   3.   1.]]
```

Création descripteur des images → Créations des clusters de descripteurs → Extraction des features

Pca 95%de variance → T-sne → ARI

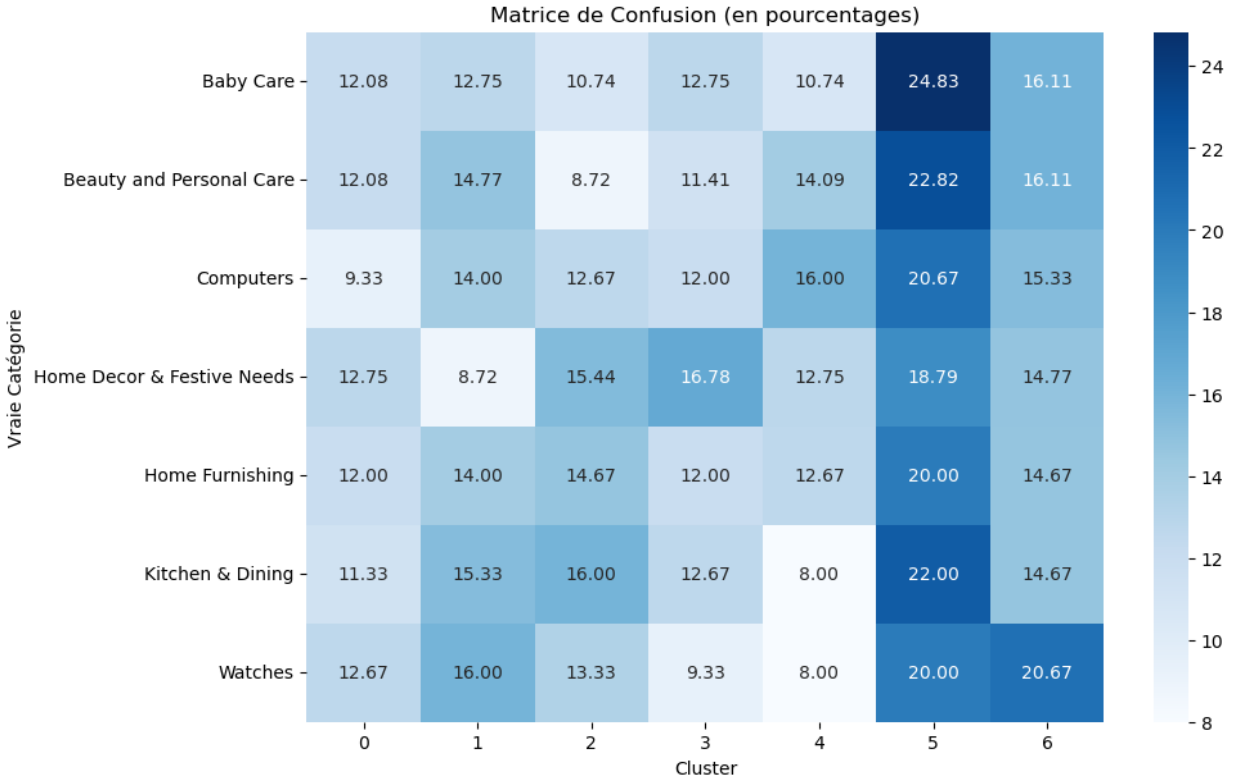





```
Adjusted Rand Score (ARI): -0.00197064706630822
```

La valeur, de l'ordre de 0.00 à 0.1, confirme le visuel.

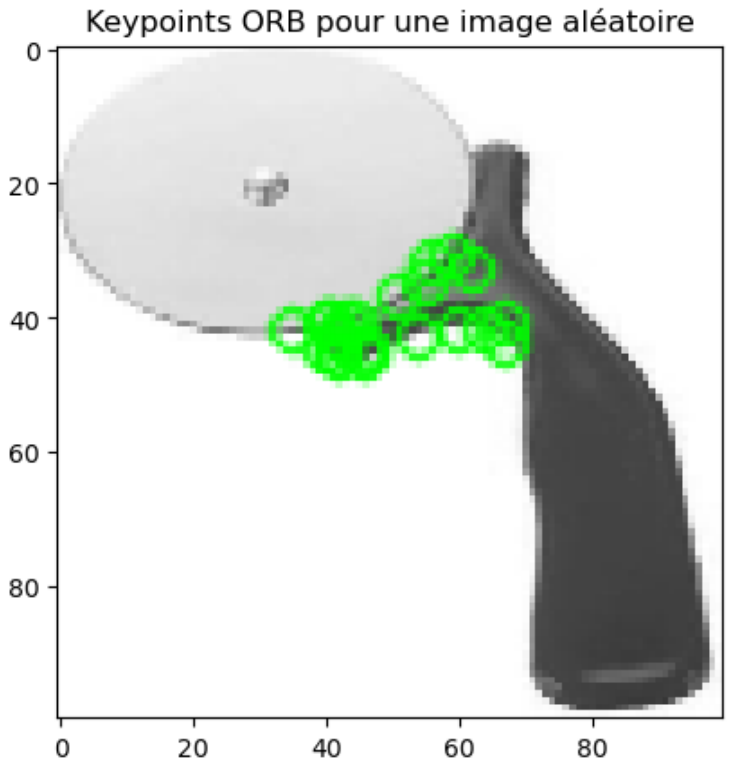
ANALYSE PAR CLASSE



	precision	recall	f1-score	support
0	0.15	0.12	0.13	149
1	0.15	0.15	0.15	149
2	0.14	0.13	0.13	150
3	0.19	0.17	0.18	149
4	0.15	0.13	0.14	150
5	0.15	0.22	0.18	150
6	0.18	0.21	0.19	150
accuracy			0.16	1047
macro avg	0.16	0.16	0.16	1047
weighted avg	0.16	0.16	0.16	1047

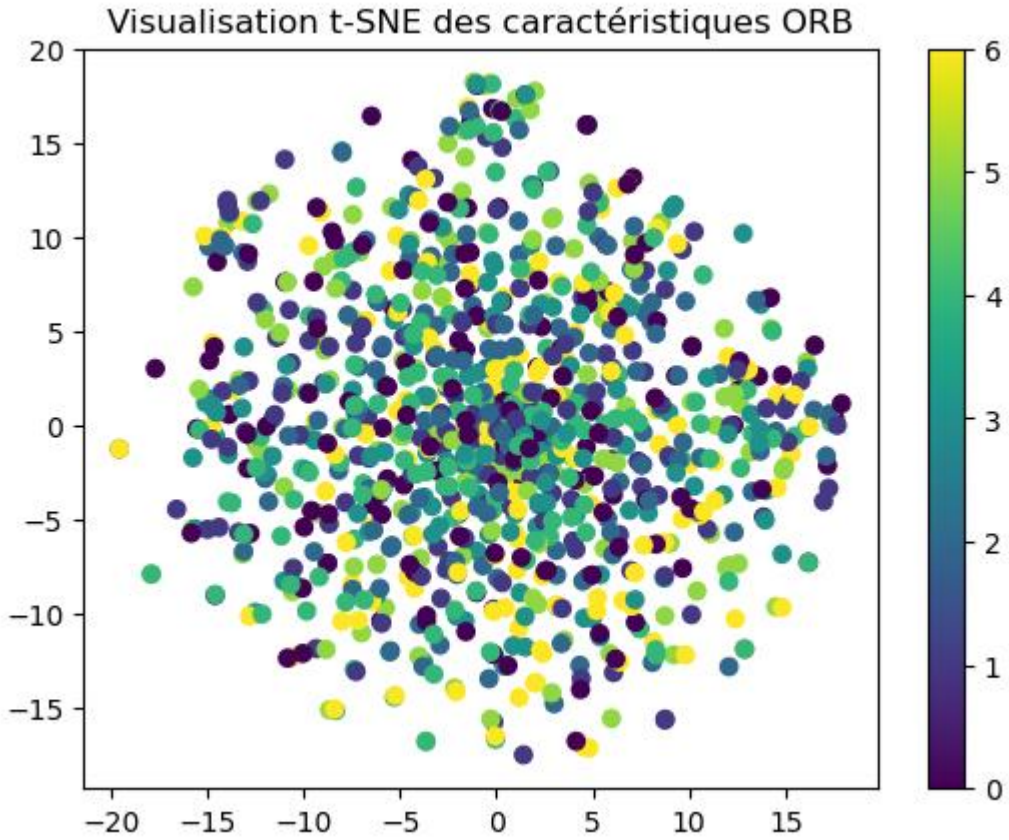
FAISABILITÉ DE CLASSIFICATION AUTOMATIQUE D'IMAGES

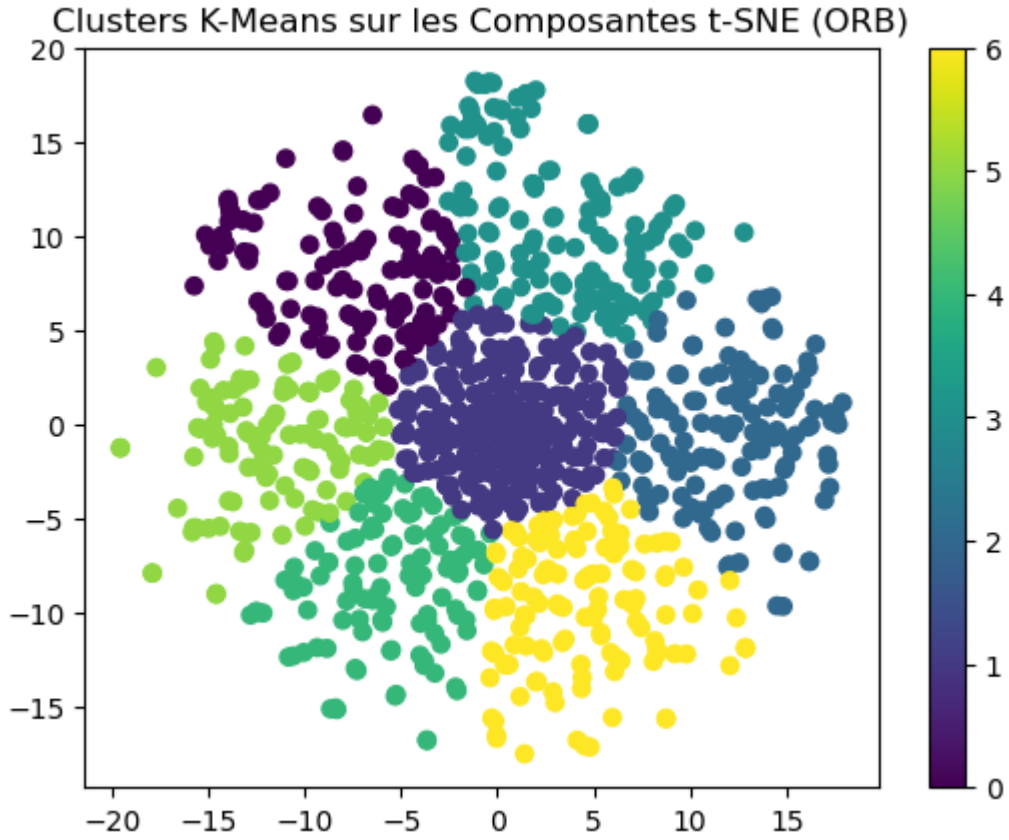
ORB



Extraction descripteurs → Créations des clusters de descripteurs → Extraction des features

Pca 95%de variance → T-sne → ARI



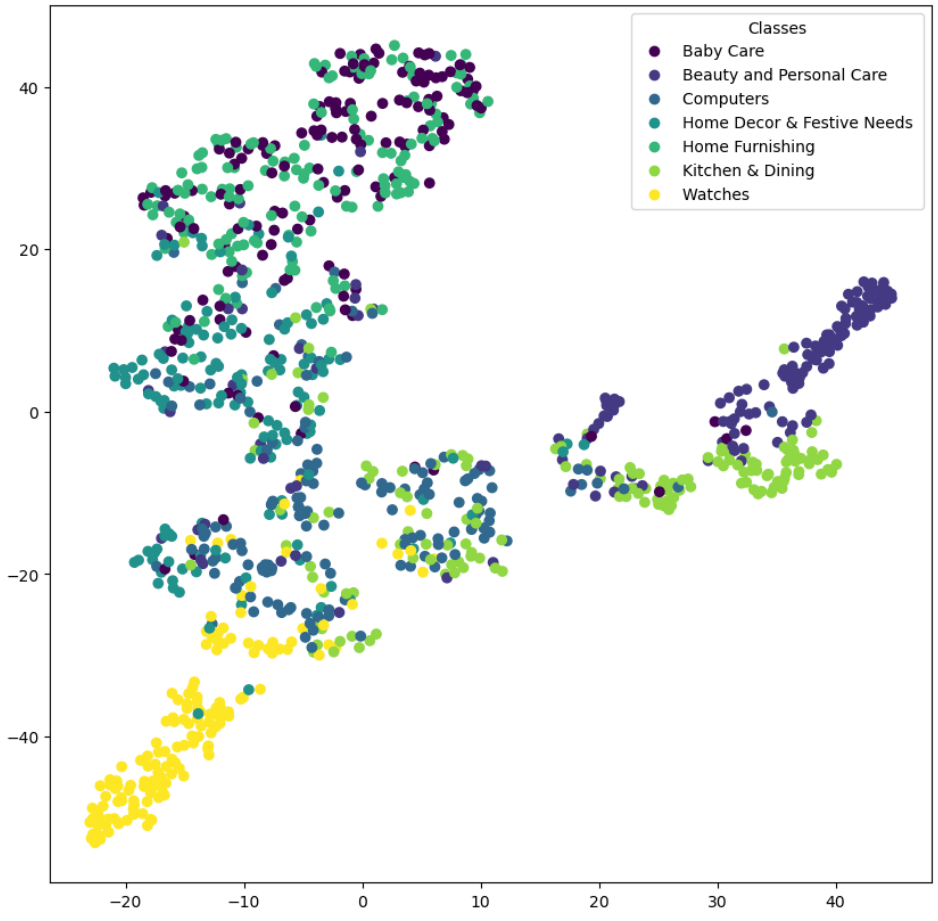


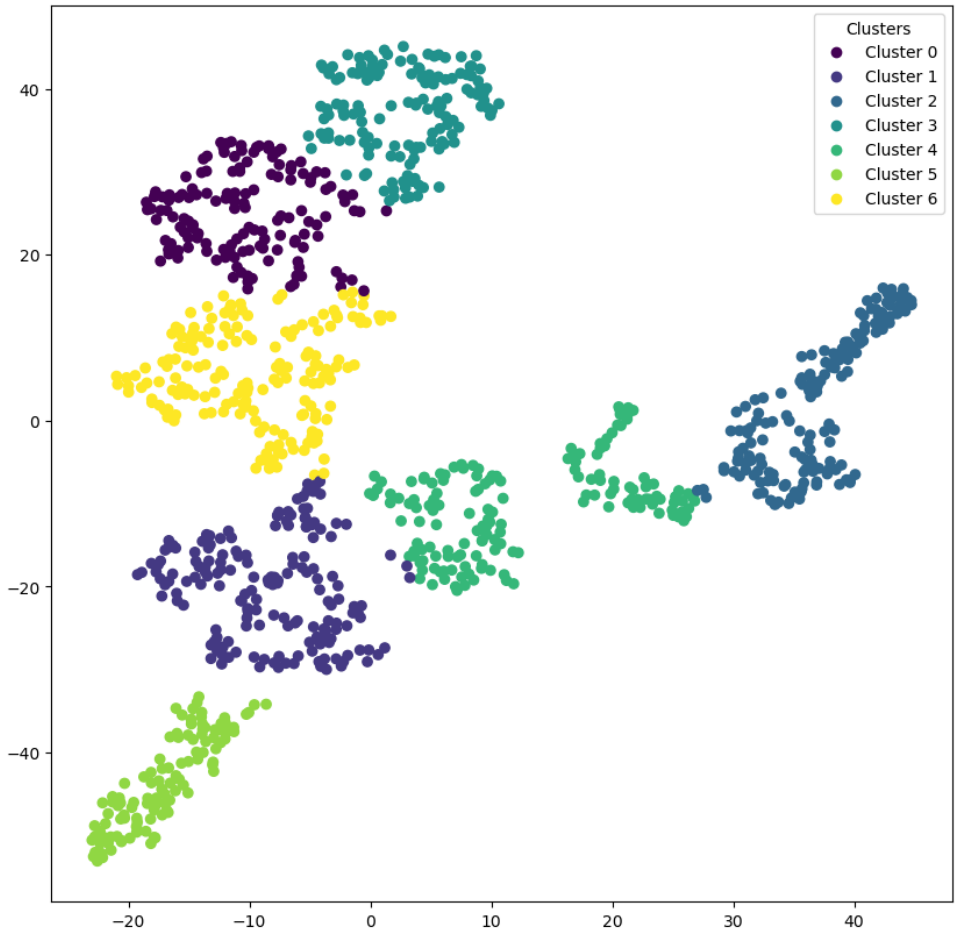

```
Adjusted Rand Score (ARI) avec ORB: 0.00039670233175964935
```

Le résultat est similaire au sift

```
Adjusted Rand Score (ARI): -0.0012080496299577332
```

FAISABILITÉ DE CLASSIFICATION AUTOMATIQUE D'IMAGES VIA CNN TRANSFERLEARNING





Score Ari :

```
ARI entre t-SNE et K-Means: 0.30636696012242487
```

5. CLASSIFICATION SUPERVISÉ

- **Classification Supervisé**

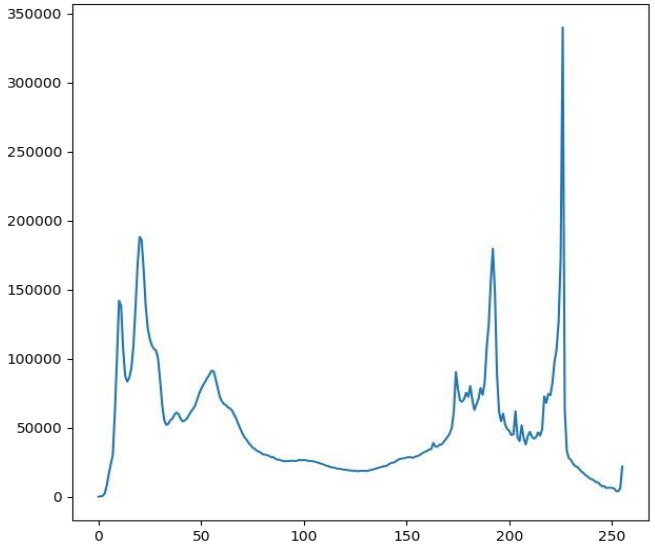
- **Classification Supervisé avec Data Augmentation avant le modèle**

CLASSIFICATION IMAGE : PRE-PROCESSING

Image en niveaux de gris avec PIL

Image après équilibrage avec PIL

Histogramme avant équilibrage avec PIL



Histogramme après équilibrage avec PIL

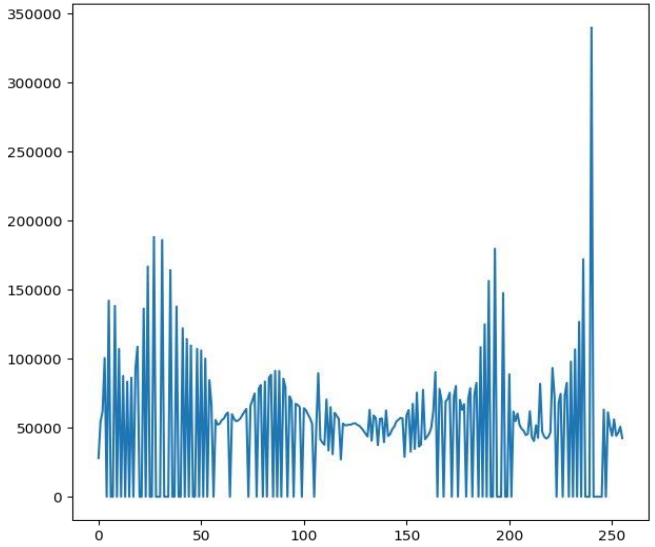
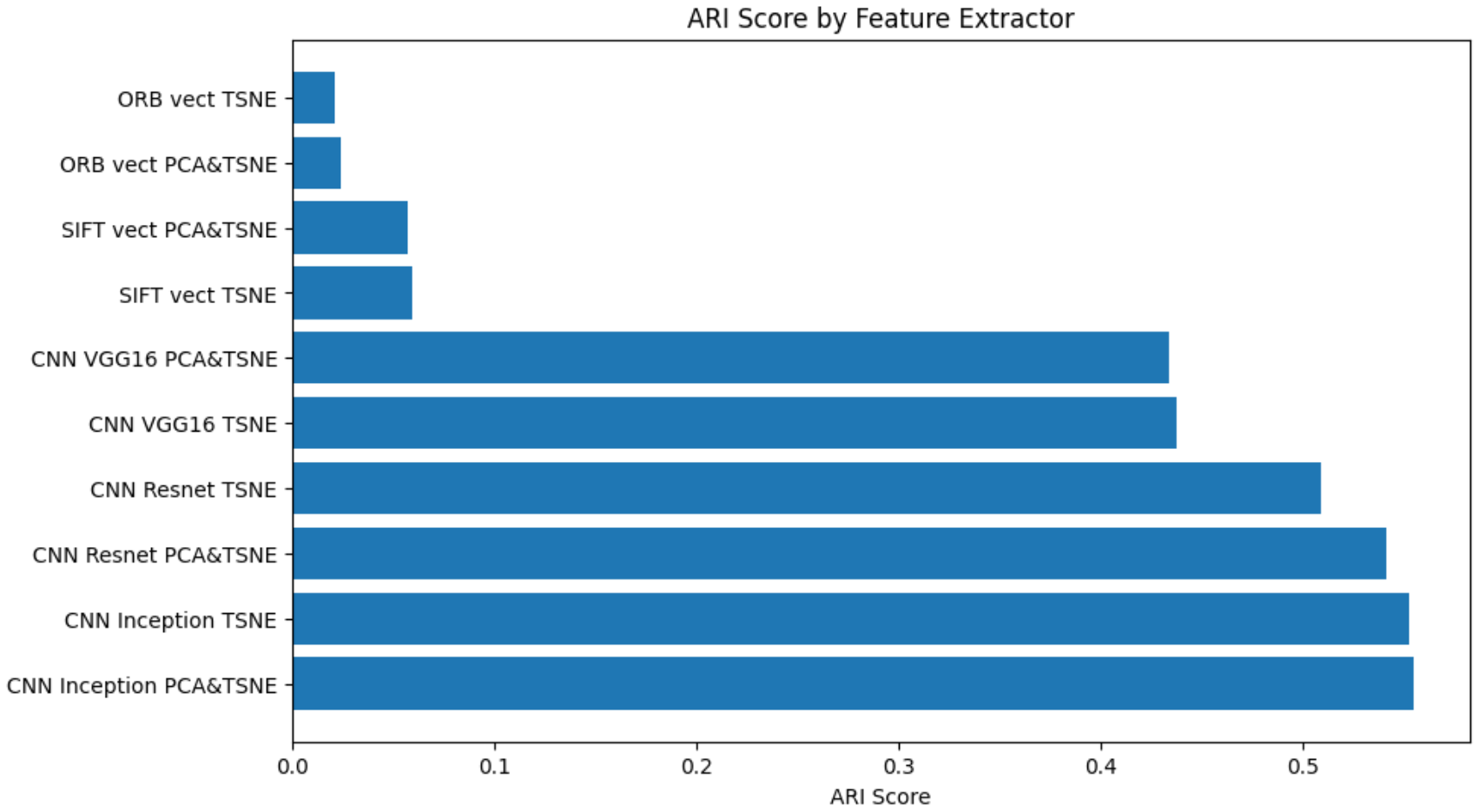
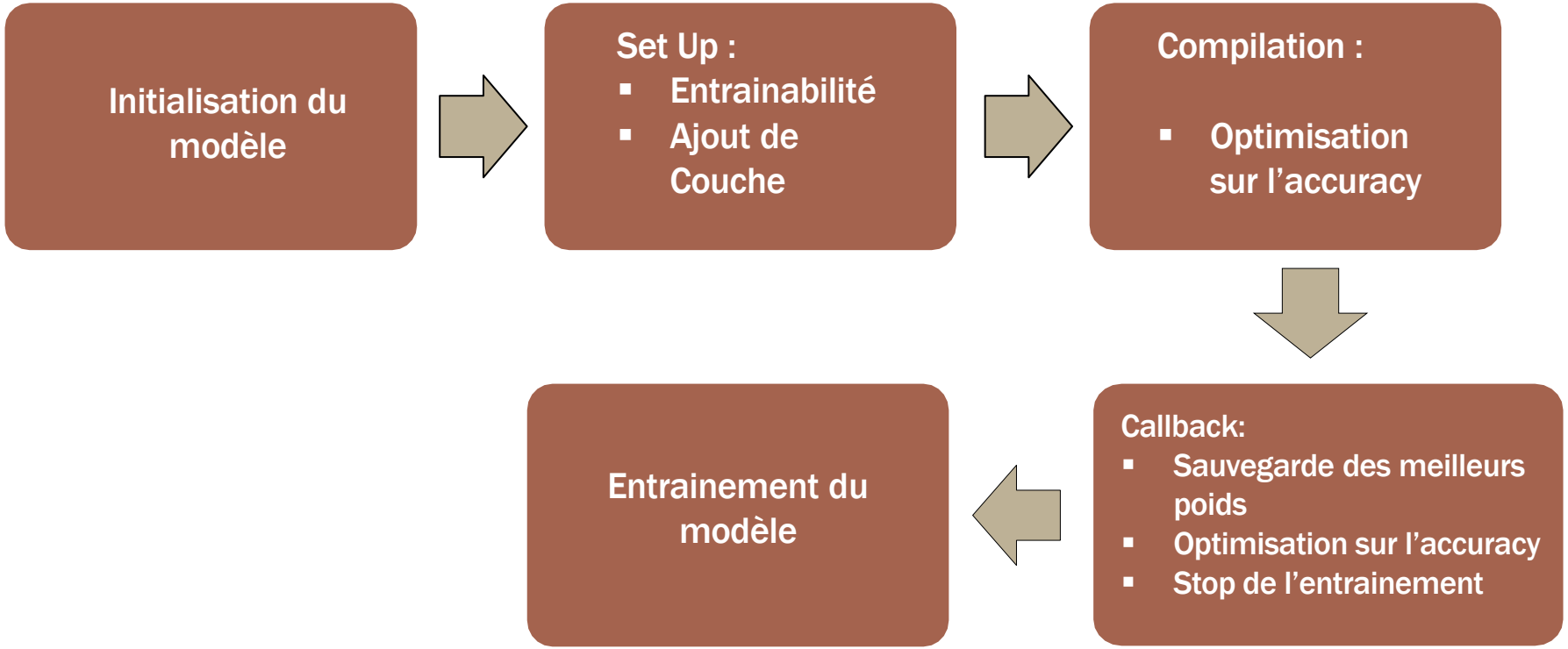


Image après filtrage médian

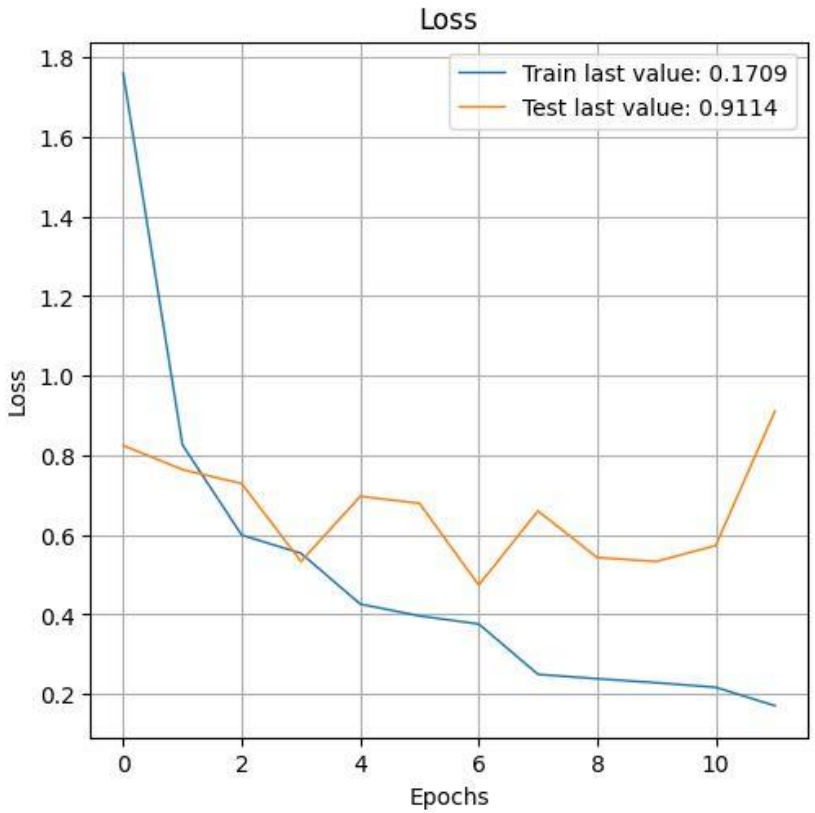
CLASSIFICATION IMAGE: RÉDUCTION DE DIMENSION ET EFFICACITÉ

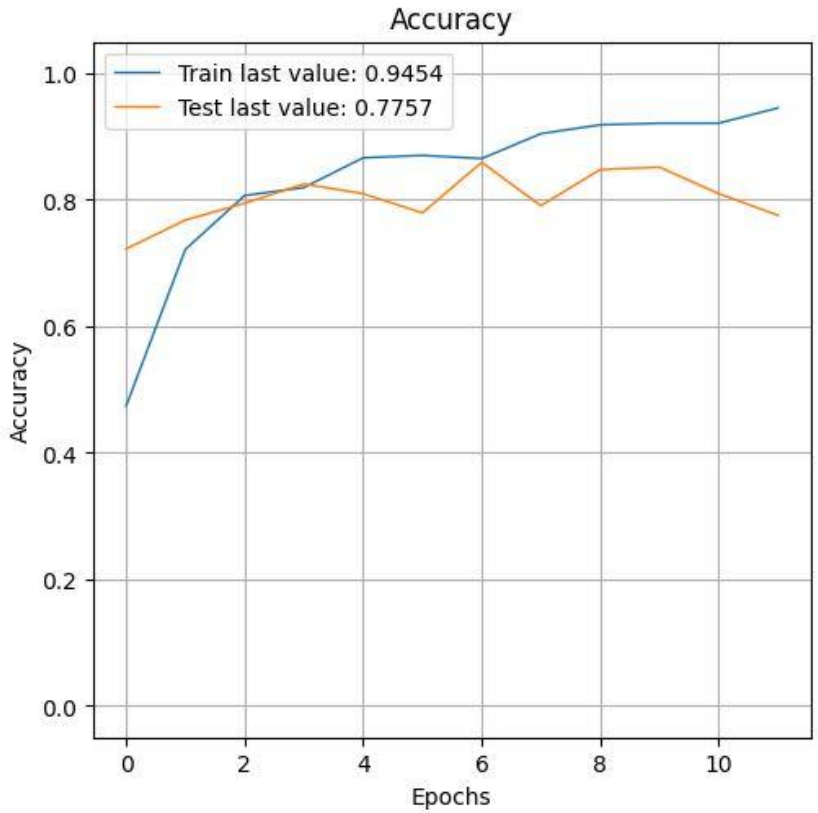


CLASSIFICATION SUPERVISÉE D'IMAGES



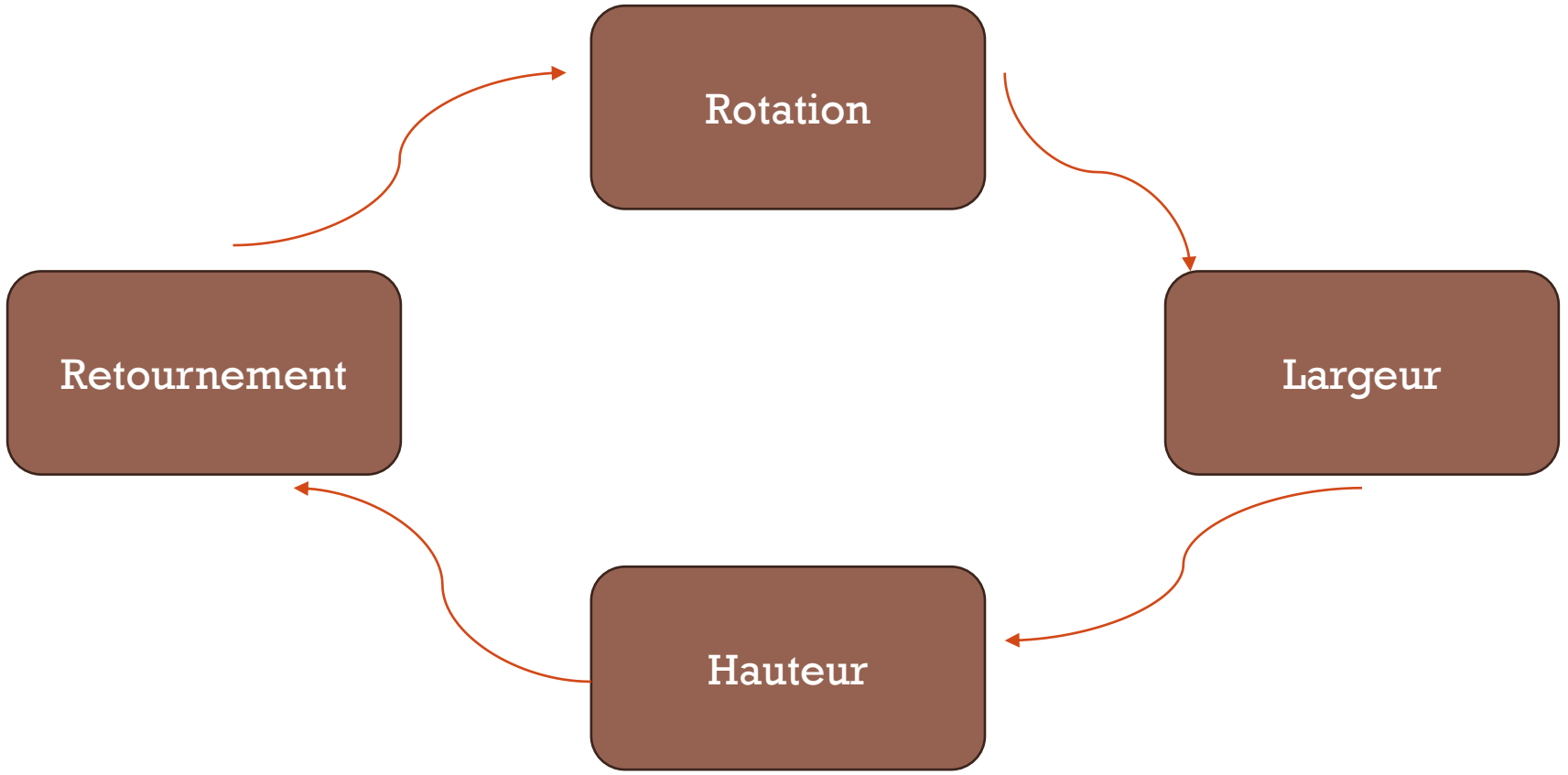
CLASSIFICATION IMAGE SUPERVISÉE





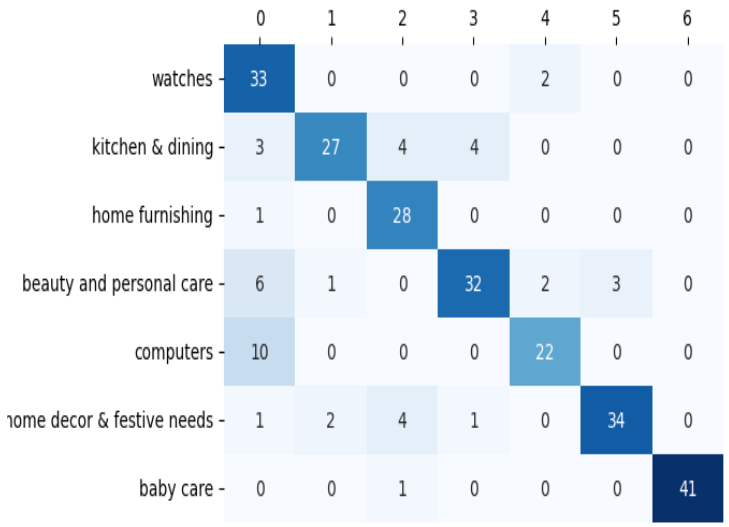
CLASSIFICATION IMAGE : SUPERVISÉE AVEC DATA AUGMENTATION AVANT LE MODÈLE

❖ DATA AUGMENTATION AVANT LE MODEL



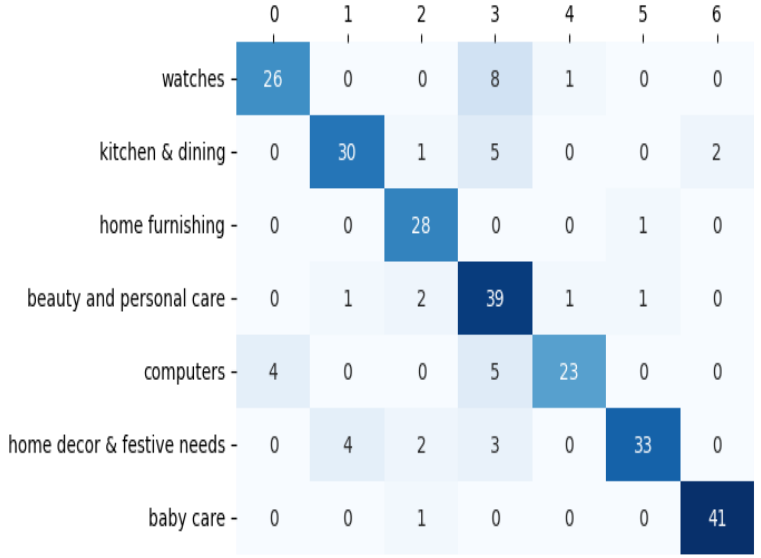
Classification Image: Supervisee, Accuracy

Sans data augmentation

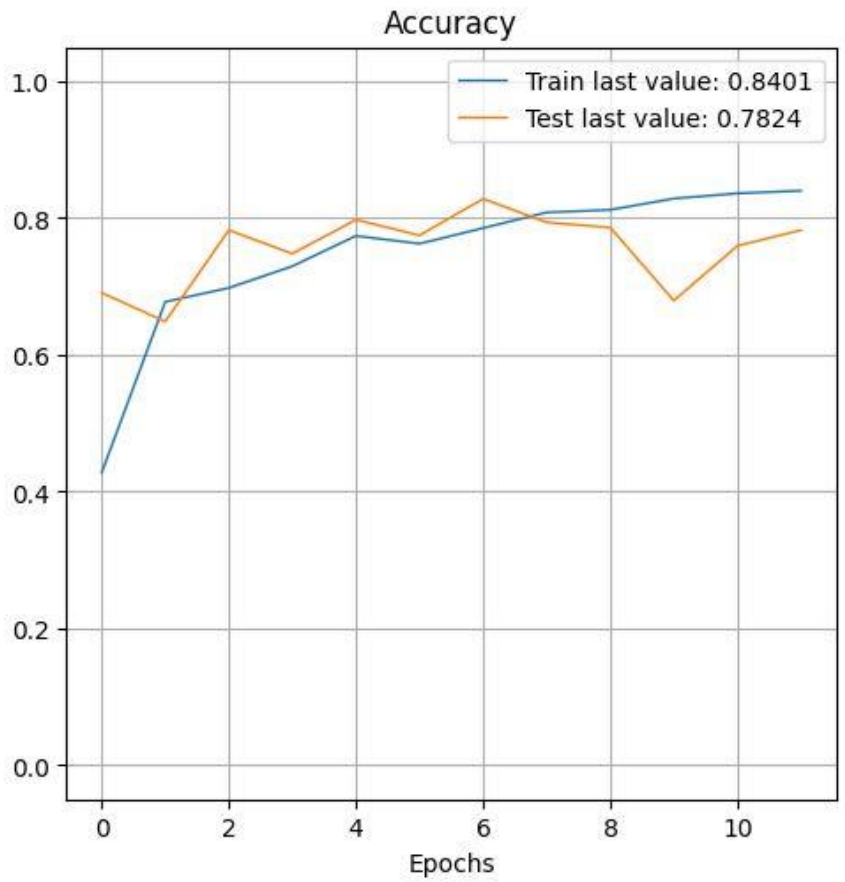


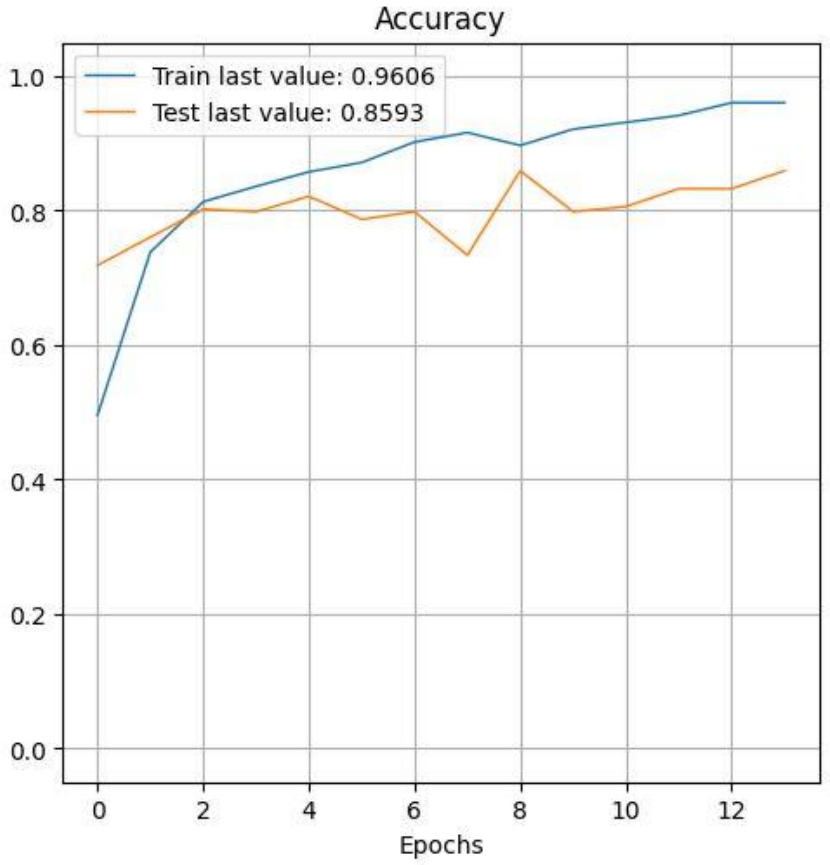
Accuracy		
Watches	1,00	0,95
Kitchen	0,92	0,94
Home furnish	0,85	0,92
Beauty	0,90	0,86
Computer	0,76	0,82
Home decor	0,86	0,65
Baby care	0,61	0,87
Avg	0,84	0,86

Avec data augmentation



Classification Image : Supervisée, sur-apprentissage





6 -PRÉSENTATION DU TEST DE L'API

Utilisation API Edamam food → Extraction 10 premiers produits ingrédient champagne → Création du fichier csv

RESPECT DES NORMES RGPD

```
querystring = {"ingr":"champagne"}
```

```
['foodId', 'label', 'category', 'foodContentsLabel', 'image']
```

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
2	food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
3	food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
4	food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
5	food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
6	food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	NaN
7	food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	NaN
8	food_am5egz8aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
9	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN

CONCLUSION

- Possibilité de Classification Automatique

- S’orienter vers les images

MERCI POUR VOTRE ÉCOUTE