

# **Classifiez automatiquement des biens de consommation**

**Joyce Kuoh Moukouri,  
P6, Soutenance du 08/07/2023**

# Ordre du jour

## **Classifiez automatiquement des biens de consommation**

1. La mission
  2. Présentation du jeu de données
  3. Étude de faisabilité : classification automatique des données textuelles
  4. Étude de faisabilité : classification automatique des données visuelles
  5. Résultats de la classification supervisée d'images
  6. Data augmentation
  7. Test API
- Conclusion

# 1. La mission

# La mission

Rappel des objectifs fixés par Linda de l'équipe de « Place de marché »

- Étudier la faisabilité d'un moteur de **classification automatique d'articles**, basé sur une image et une description, pour l'attribution de la catégorie de l'article.
- Implémentation d'un **modèle supervisé de classification d'images**
- Optimisation du modèle par **data augmentation**

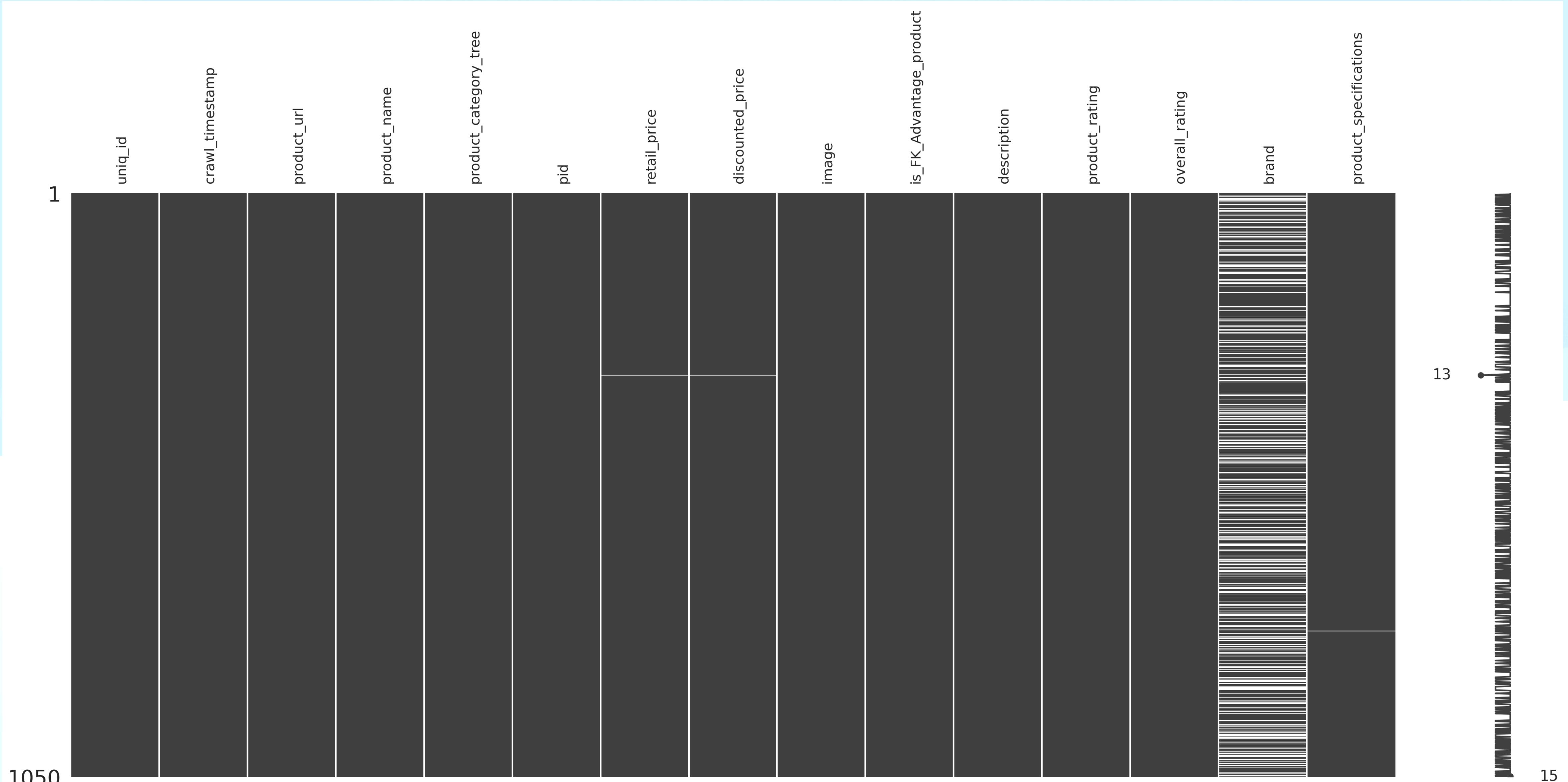
## **2. La base de données**

## 2. La base de données

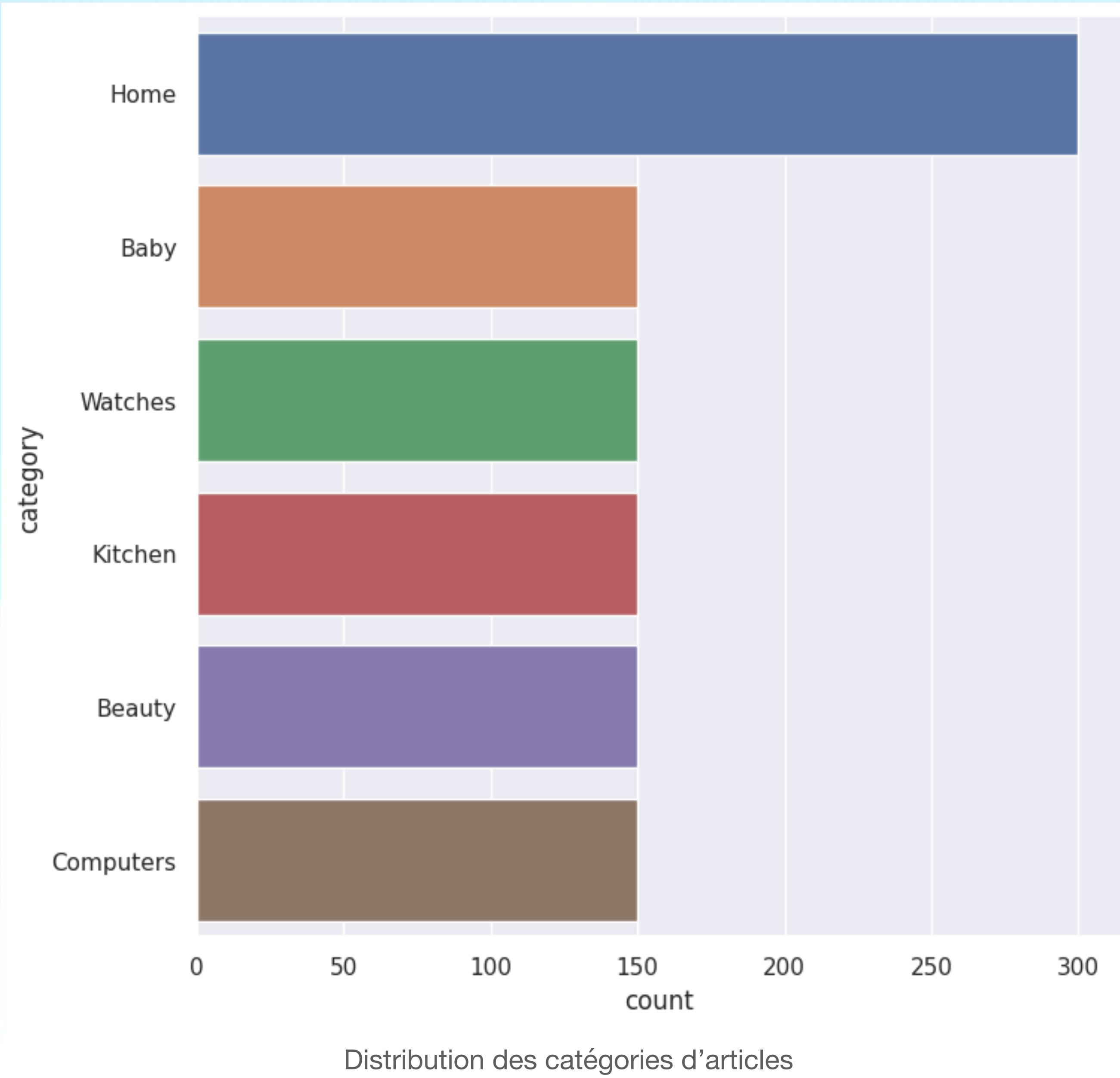
- Jeu de données extrait du site de e-commerce **flipkart.com**
- Le **site n'autorise pas de web-scraping** mais les données sont disponibles sur Kaggle
- (H1) : Il est supposé que les images sont libres de droits



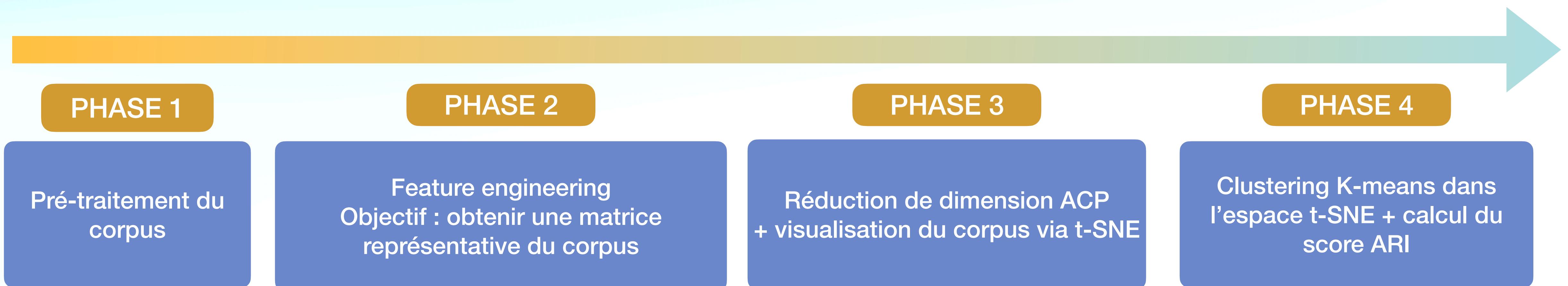
- Une table listant 1050 articles de consommations, décrits en **15 features**
- Un dossier comportant les 1050 images correspondantes à chaque articles
- Absence de doublons et pas de valeurs manquantes (hormis pour la feature « brand » , 32%)



Aperçu de la base de données et des valeurs manquantes

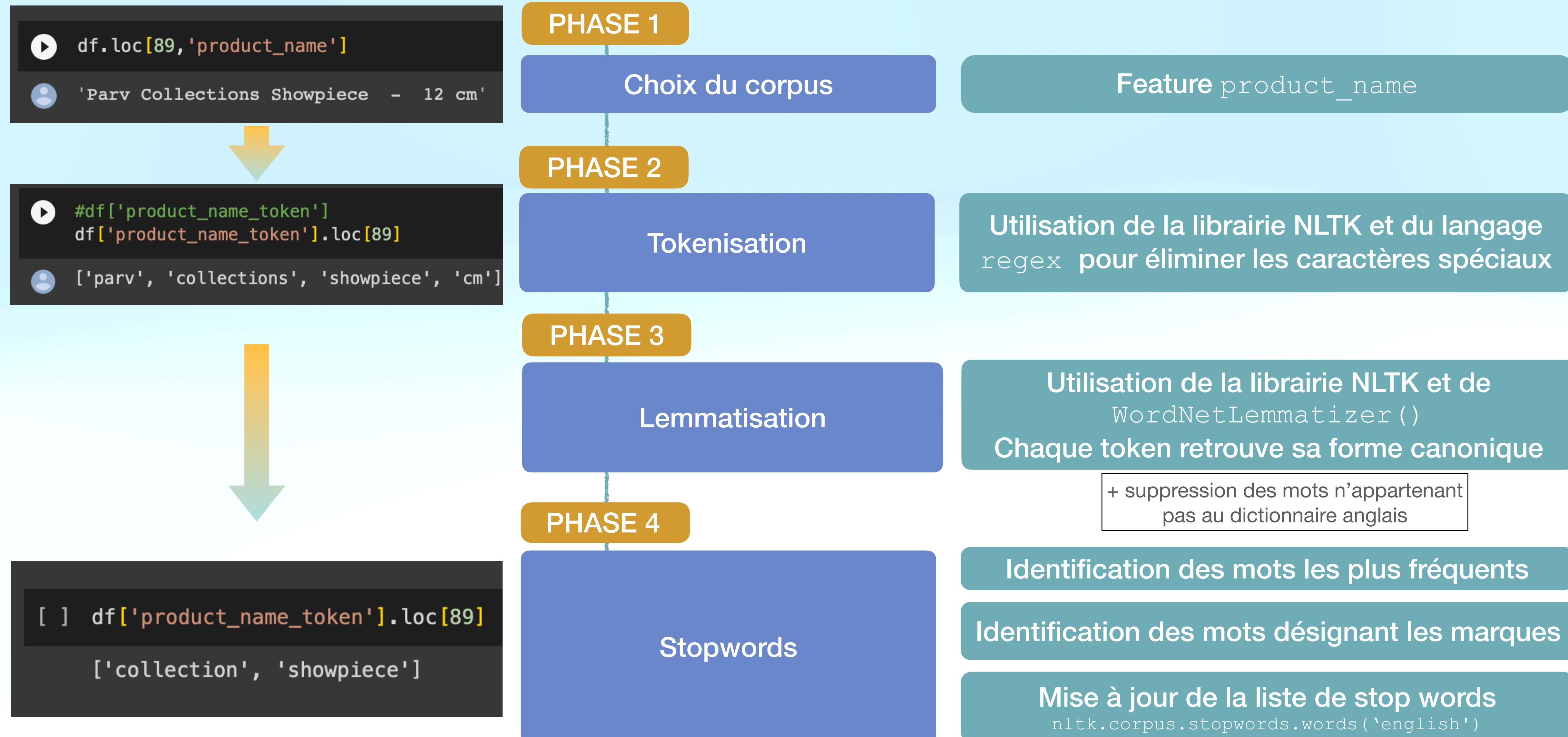


# 3. Étude de faisabilité : classification automatique des données textuelles



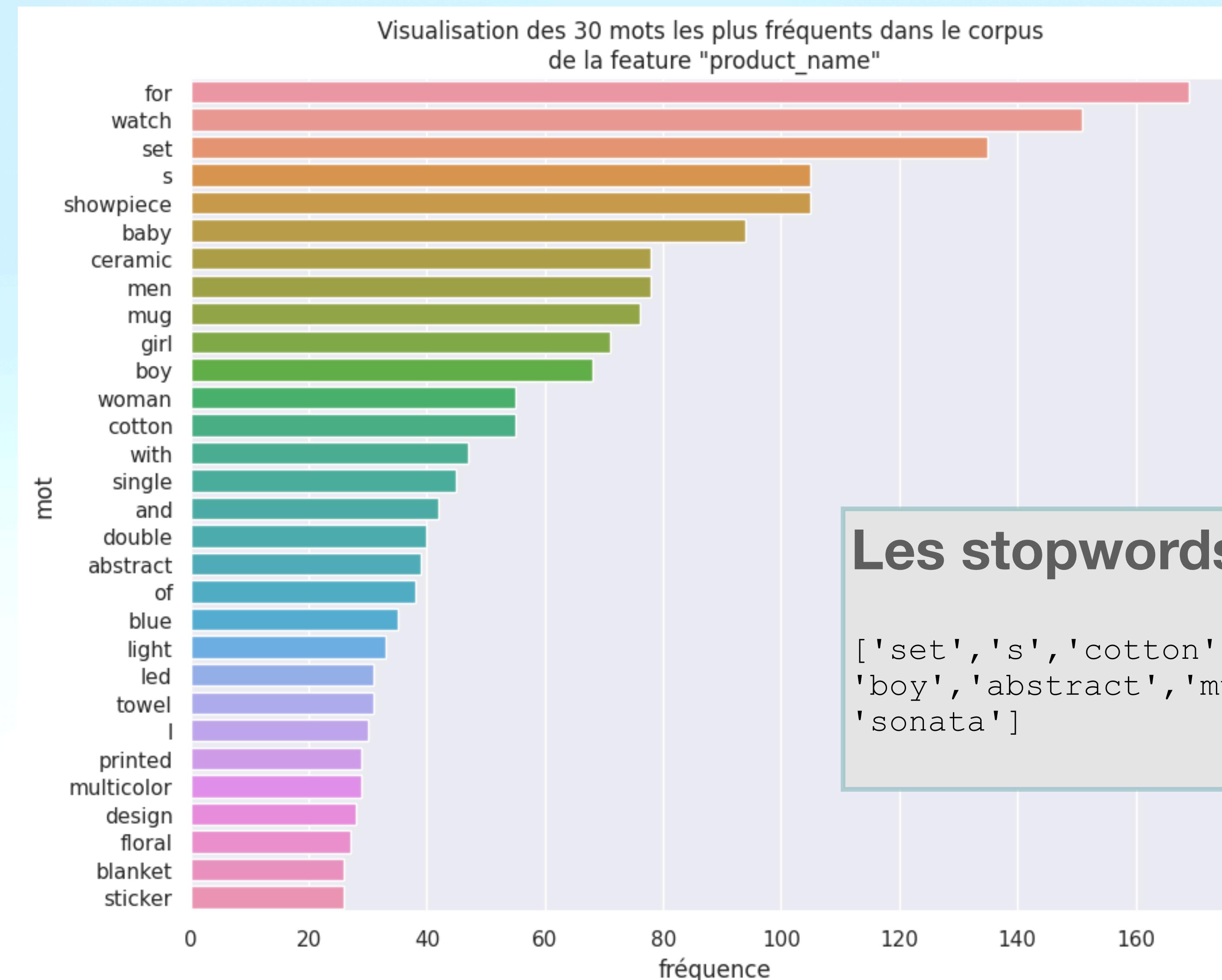
### 3. Étude de faisabilité : classification automatique des données textuelles

#### A. Pré-traitement : cas des données textuelles



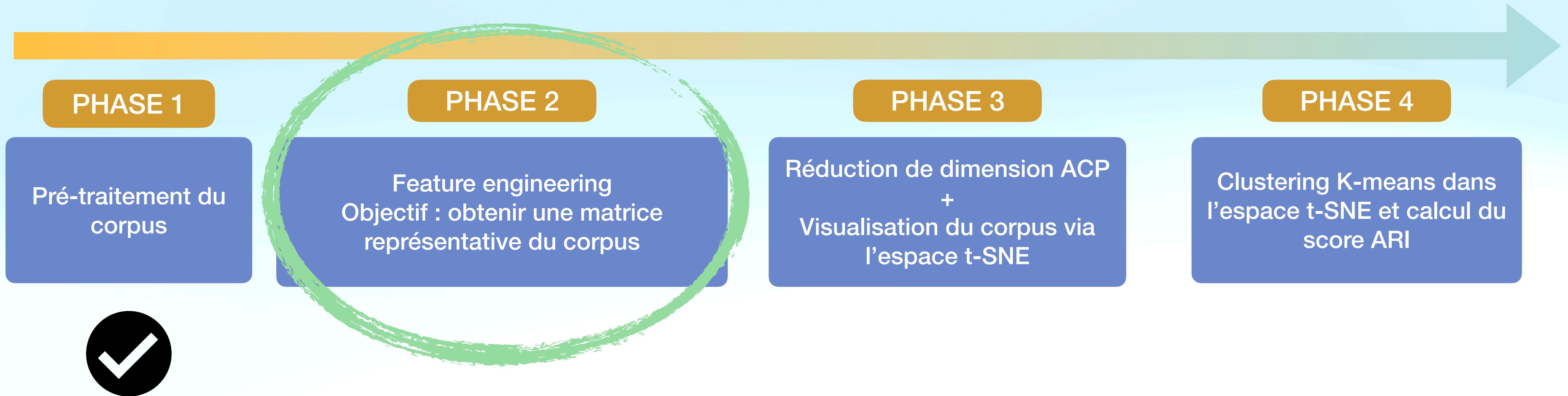
### 3. Étude de faisabilité : classification automatique des données textuelles

#### A. Pré-traitement : cas des données textuelles



### 3. Étude de faisabilité : classification automatique des données textuelles

#### B. Feature engineering



### 3. Étude de faisabilité : classification automatique des données textuelles

#### B. Feature engineering : vectorization du corpus

Méthodes testées	
Bag of word	Matrices éparses de grandes dimensions vectorization unique au corpus.
TF-IDF	
Word2vec + modèle de sentence embedding	
BERT par Transfer Learning	Méthodes de word/ sentence embedding, matrice dense
USE par Transfer Learning	

# Classification automatique des produits en fonction de leur désignation



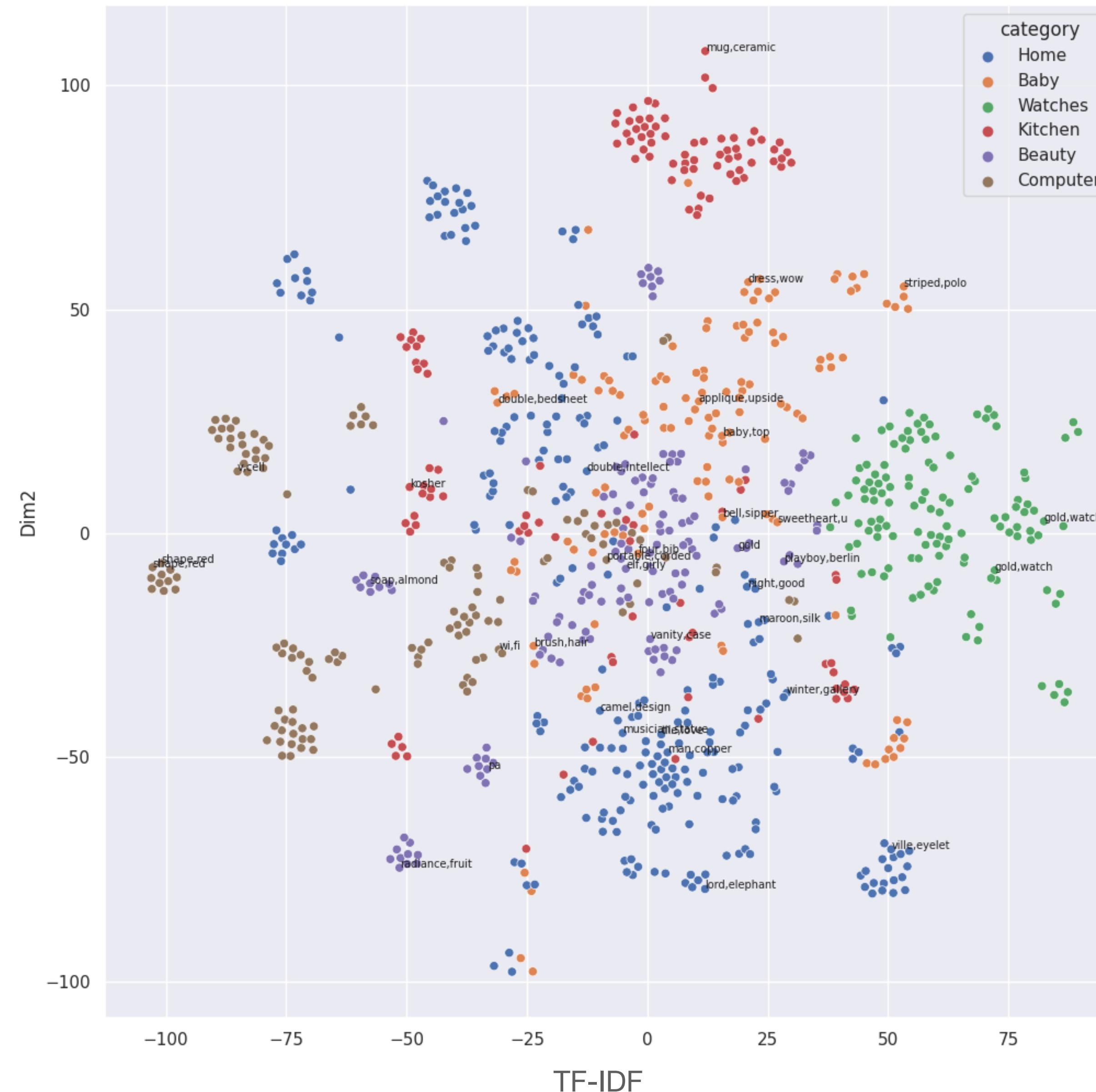
Bag of word

# Classification automatique des produits en fonction de leur désignation



## Bag of Word, ARI = 0.

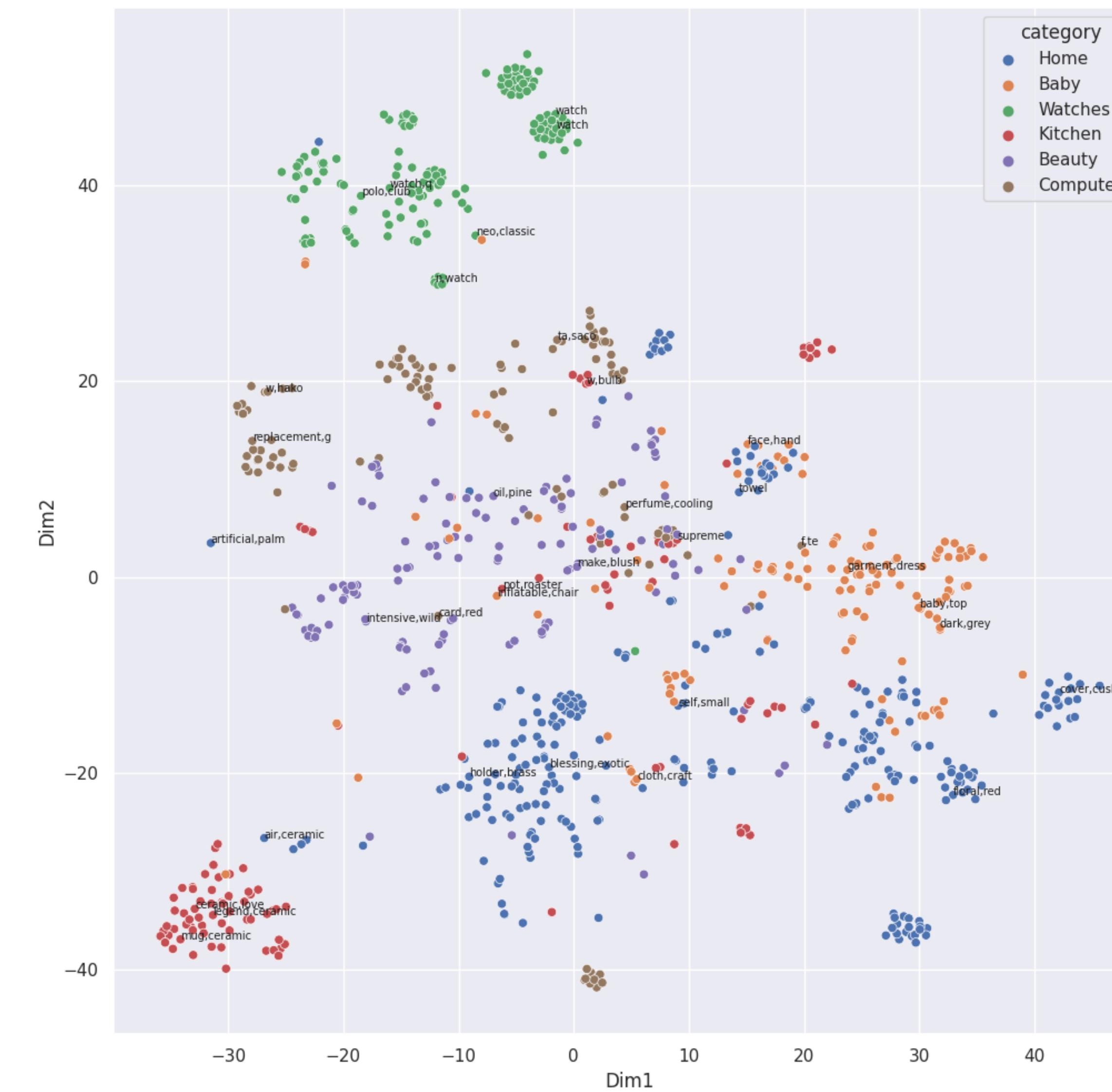
# Classification automatique des produits en fonction de leur désignation



# Classification automatique des produits en fonction de leur désignation



# Classification automatique des produits en fonction de leur désignation



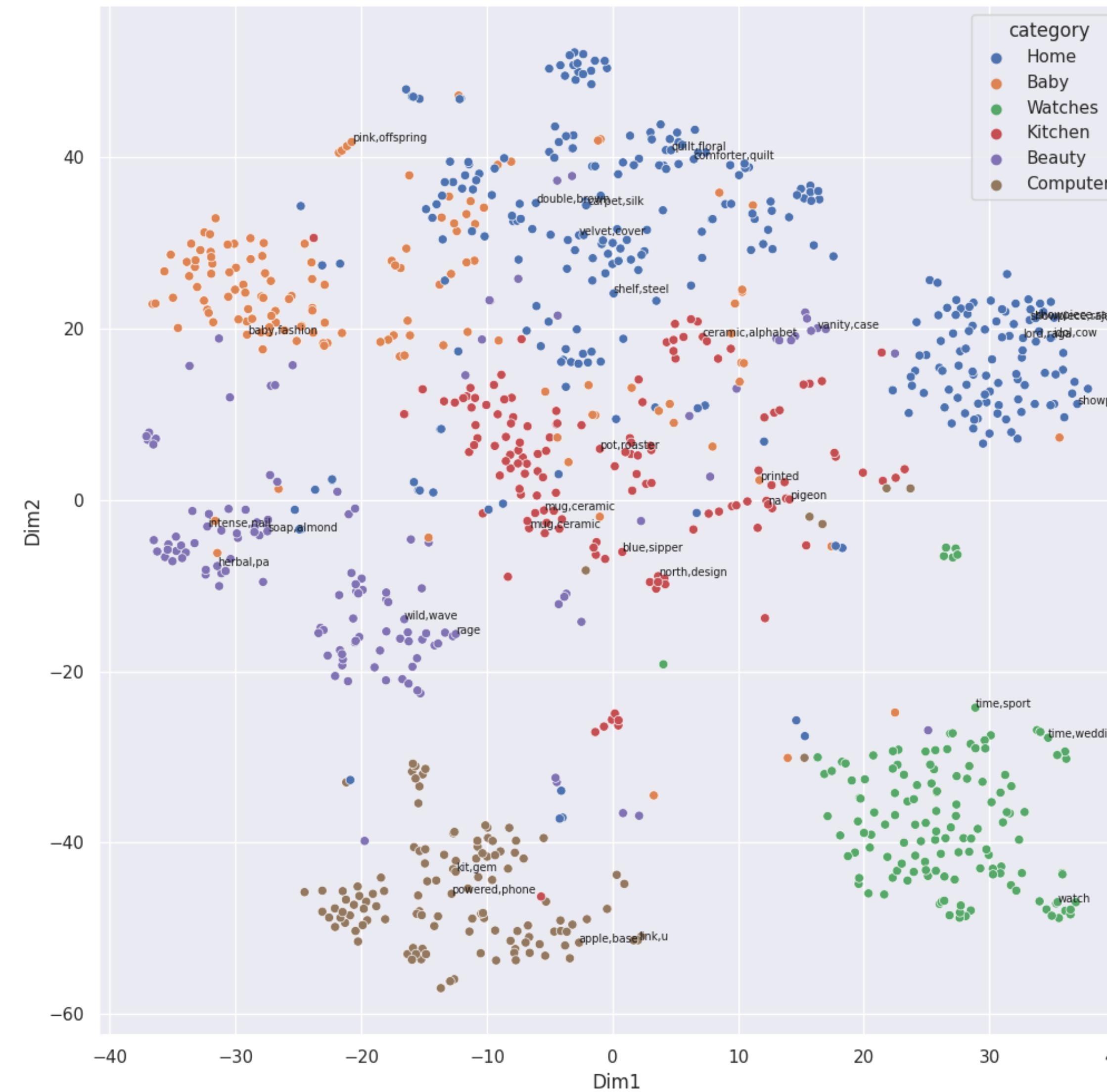
# Word2vec

# Classification automatique des produits en fonction de leur désignation



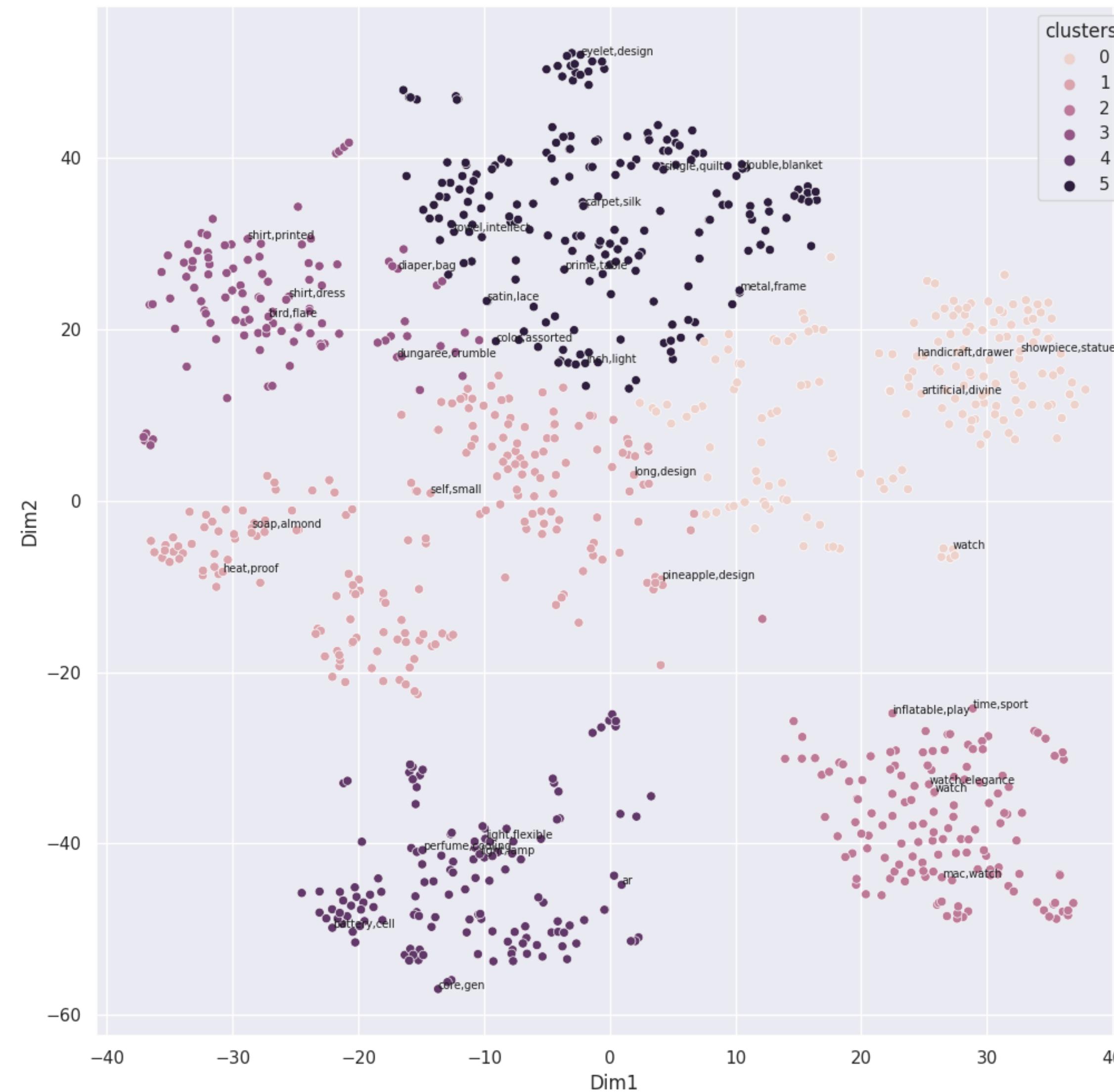
Word2Vec, ARI : 0.33

# Classification automatique des produits en fonction de leur désignation



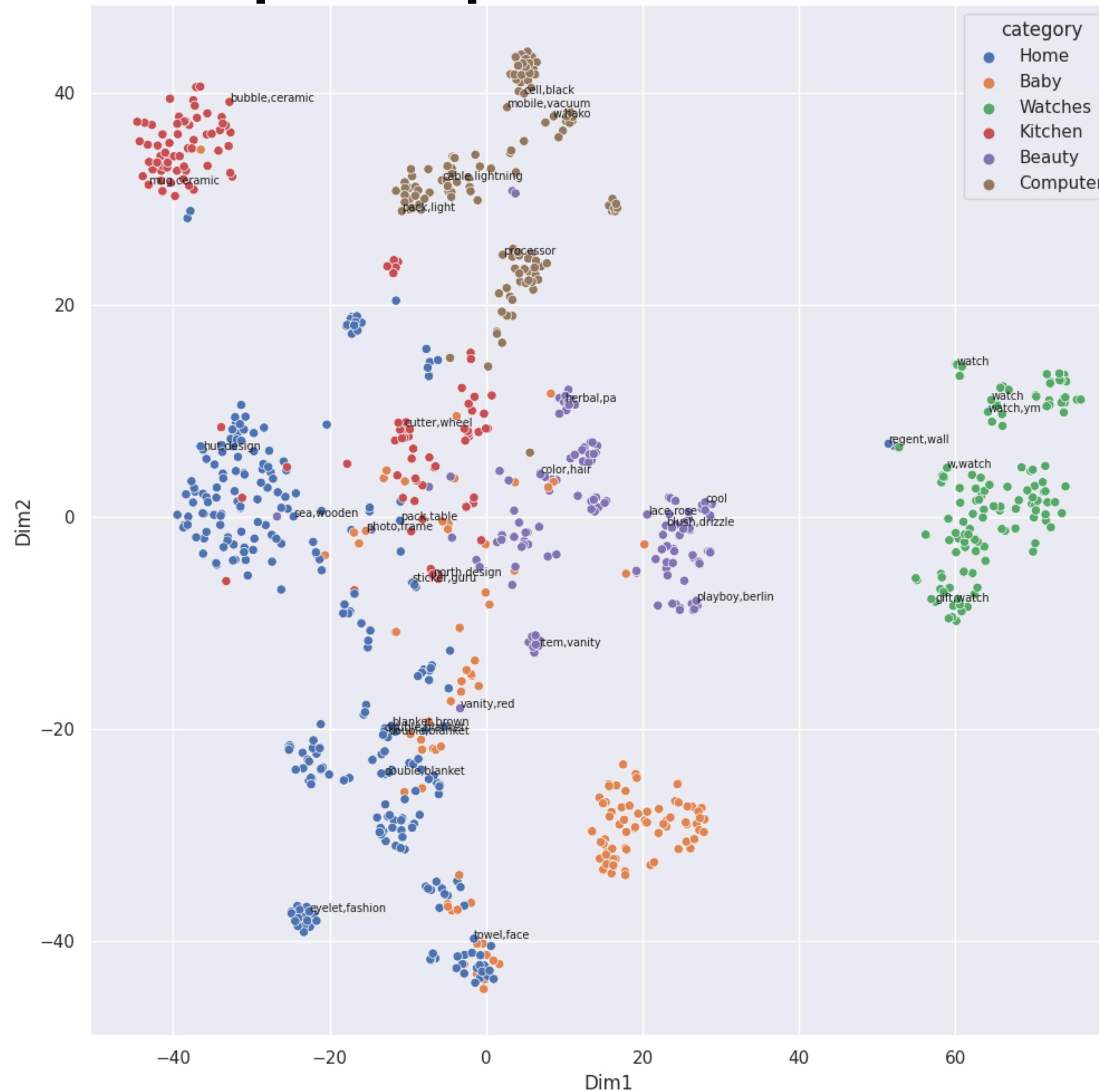
BER

# Classification automatique des produits en fonction de leur désignation



BERT, ARI : 0.46

# Classification automatique des produits en fonction de leur désignation

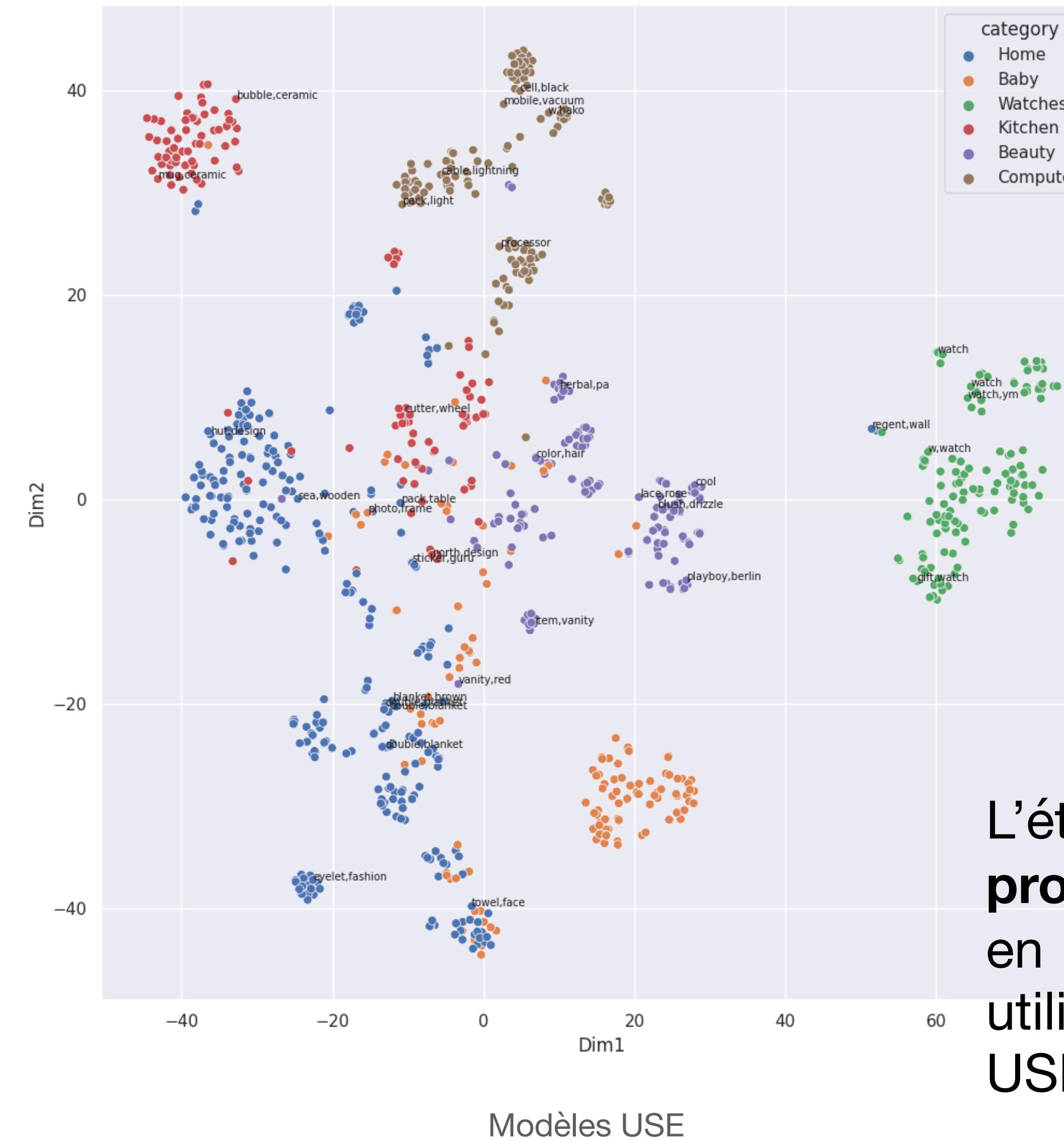


# Classification automatique des produits en fonction de leur désignation



USE, ARI = 0.5

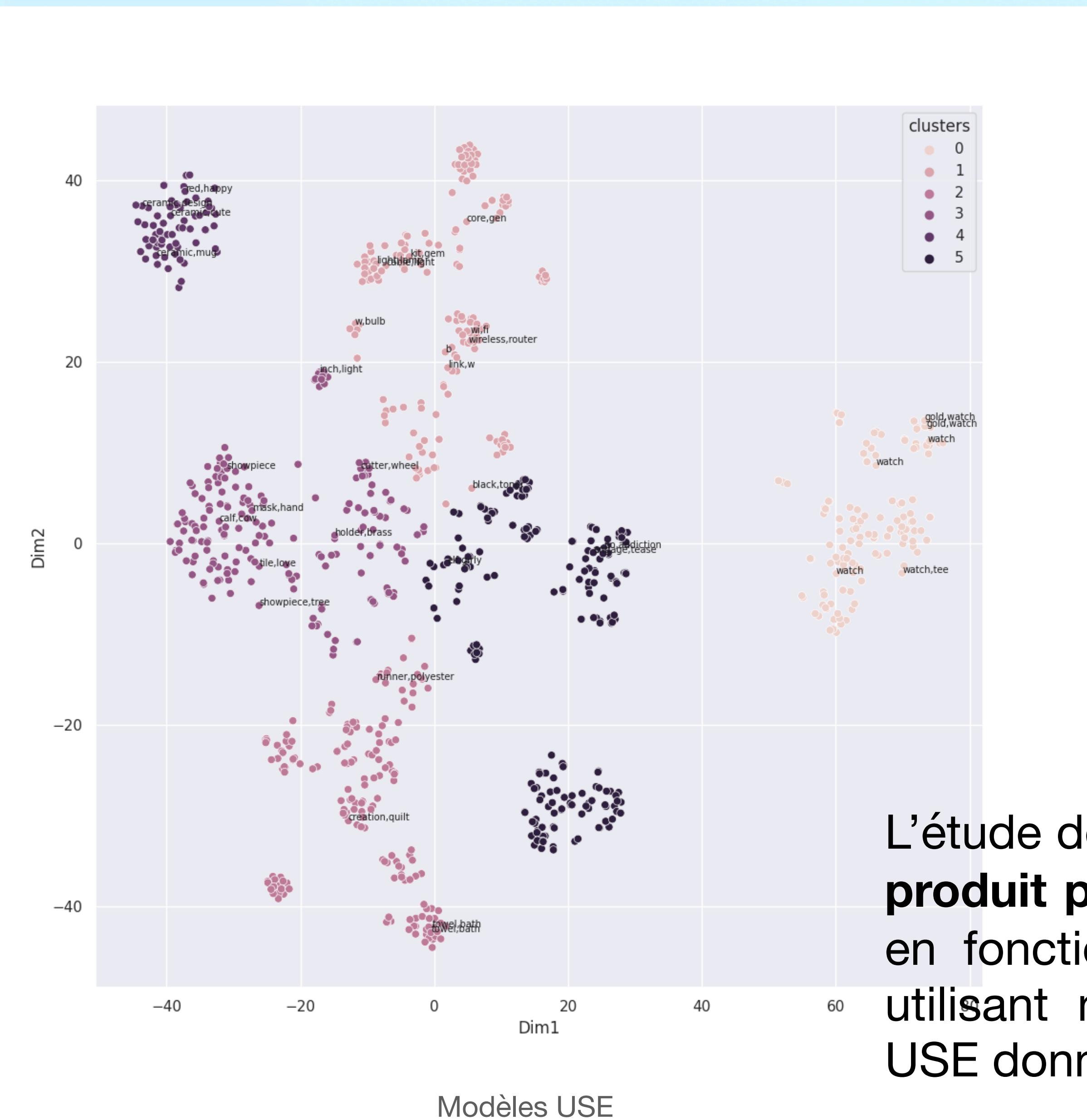
# Classification automatique des produits en fonction de leur désignation



## Conclusion

L'étude de faisabilité démontre que **6 catégories de produit peuvent se distinguer automatiquement** en fonction uniquement de leur désignation, en utilisant notamment des transformers. Le modèle USE donne les meilleurs résultats.

# Classification automatique des produits en fonction de leur désignation



# Conclusion

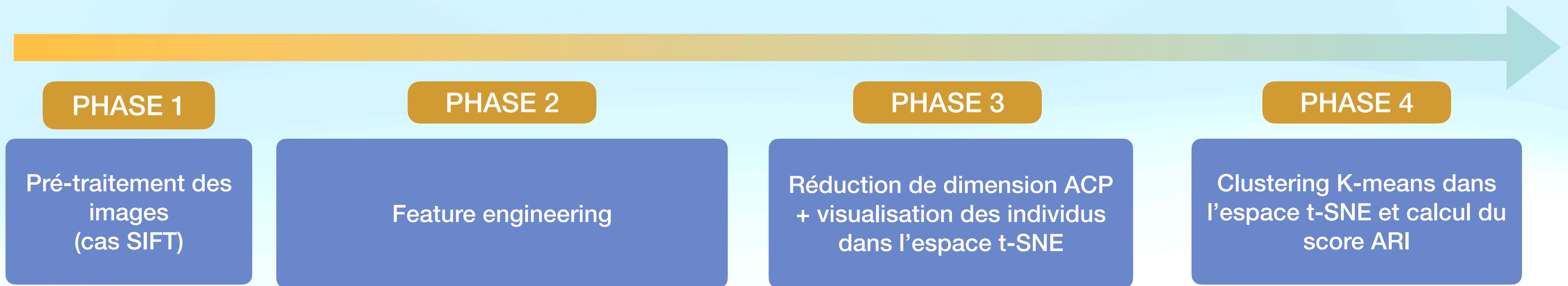
L'étude de faisabilité démontre que **6 catégories de produit peuvent se distinguer automatiquement** en fonction uniquement de leur désignation, en utilisant notamment des transformers. Le modèle USE donne les meilleurs résultats.

# 4. Étude de faisabilité : classification automatique des données visuelles

# 4. Étude de faisabilité : classification automatique des données visuelles

Méthode testées
SIFT Scale-invariant feature transform
TransferLearning à partir du modèle VGG16

## 4. Étude de faisabilité : classification automatique des données visuelles

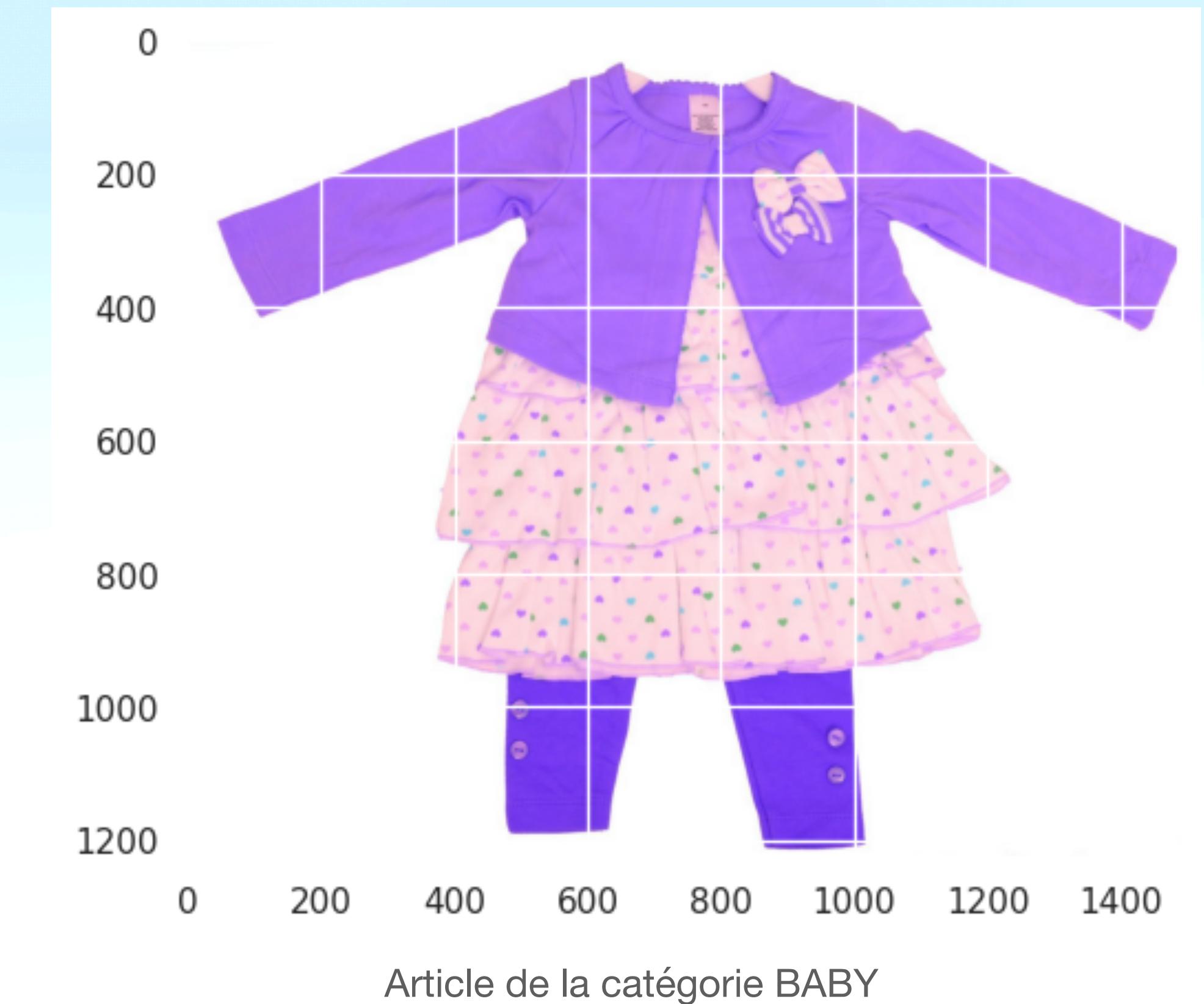


## 4. Étude de faisabilité : classification automatique des données visuelles

### A. Pré-traitement des images - modèle SIFT

Pour mieux détecter les points d'intérêts :

- On passe chacune des images au gris
- On égalise l'histogramme de l'image pour une meilleure répartition du contraste de l'image

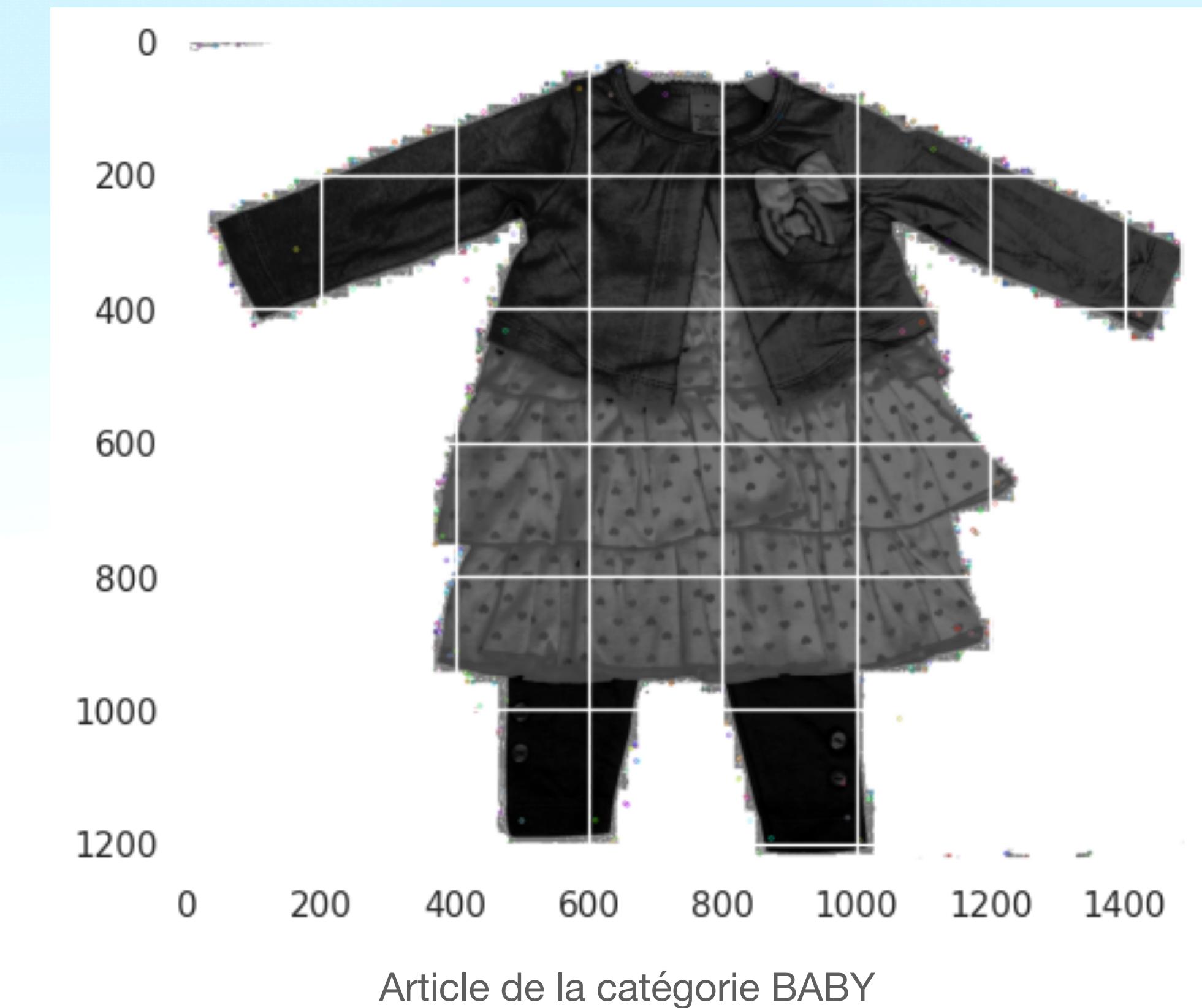


## 4. Étude de faisabilité : classification automatique des données visuelles

### A. Pré-traitement des images - modèle SIFT

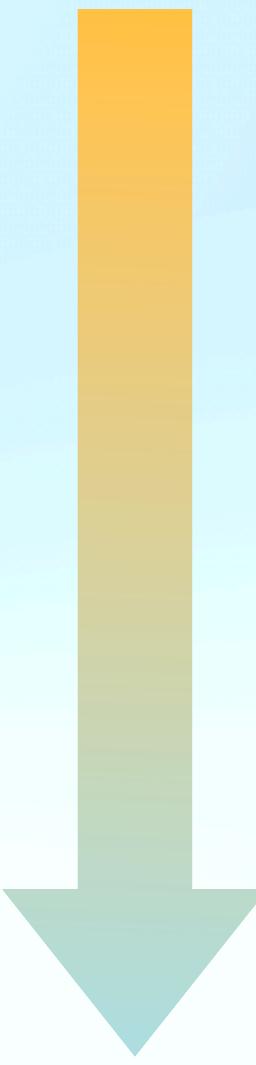
Pour mieux détecter les points d'intérêts :

- On passe chacune des images au gris
- On égalise l'histogramme de l'image pour une meilleure répartition du contraste de l'image



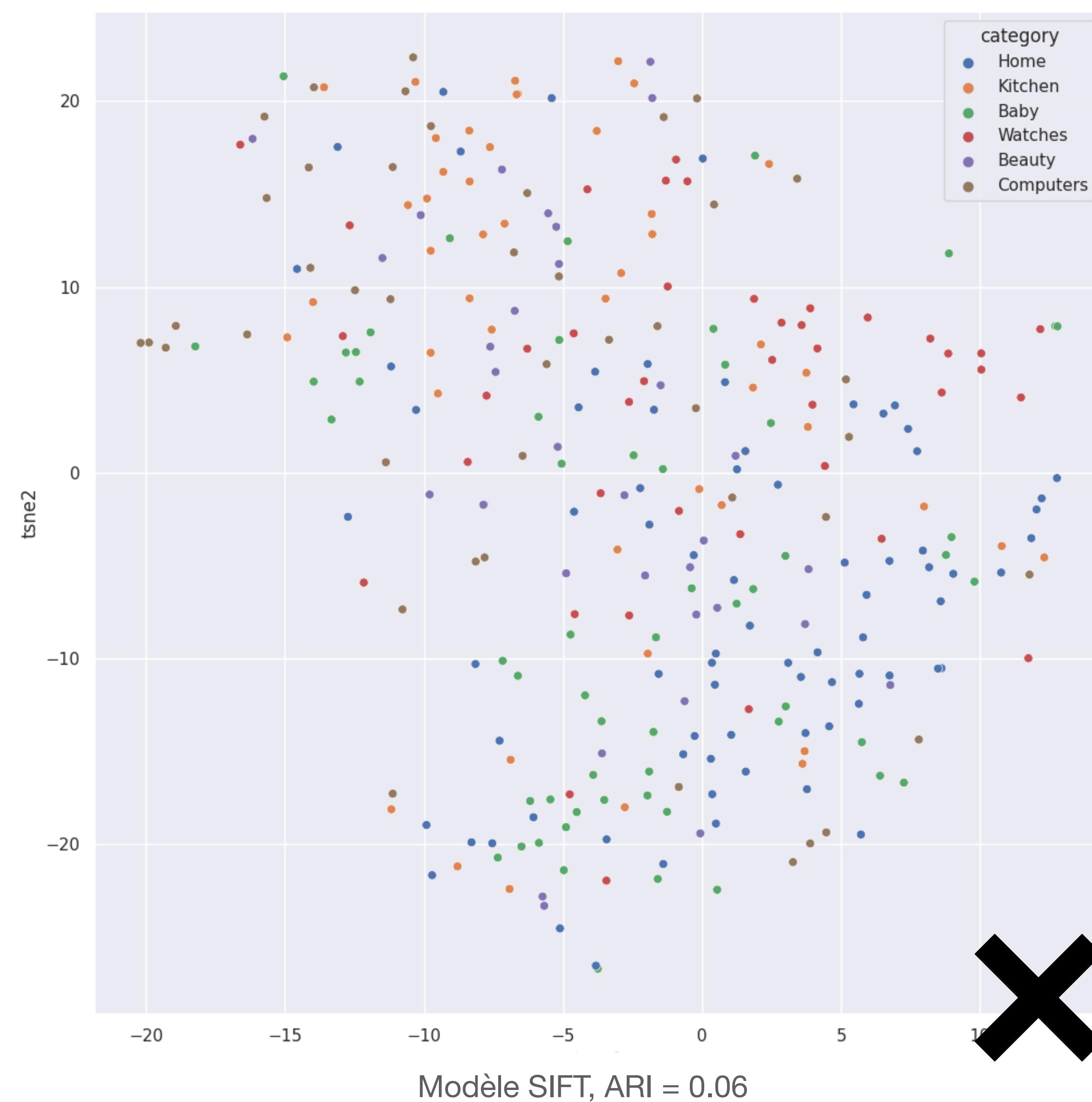
### 3. Étude de faisabilité : classification automatique des données visuelles

#### B. Feature engineering - modèle SIFT

- 
1. On détecte les points d'intérêt
  2. On génère les descripteurs de chaque points pour chaque image
  3. On regroupe chaque descripteurs dans un des  $k$  clusters
  4. On crée une matrice de type ***bag of features***

# Classification automatique des produits en fonction de leur images

## Modèle SIFT

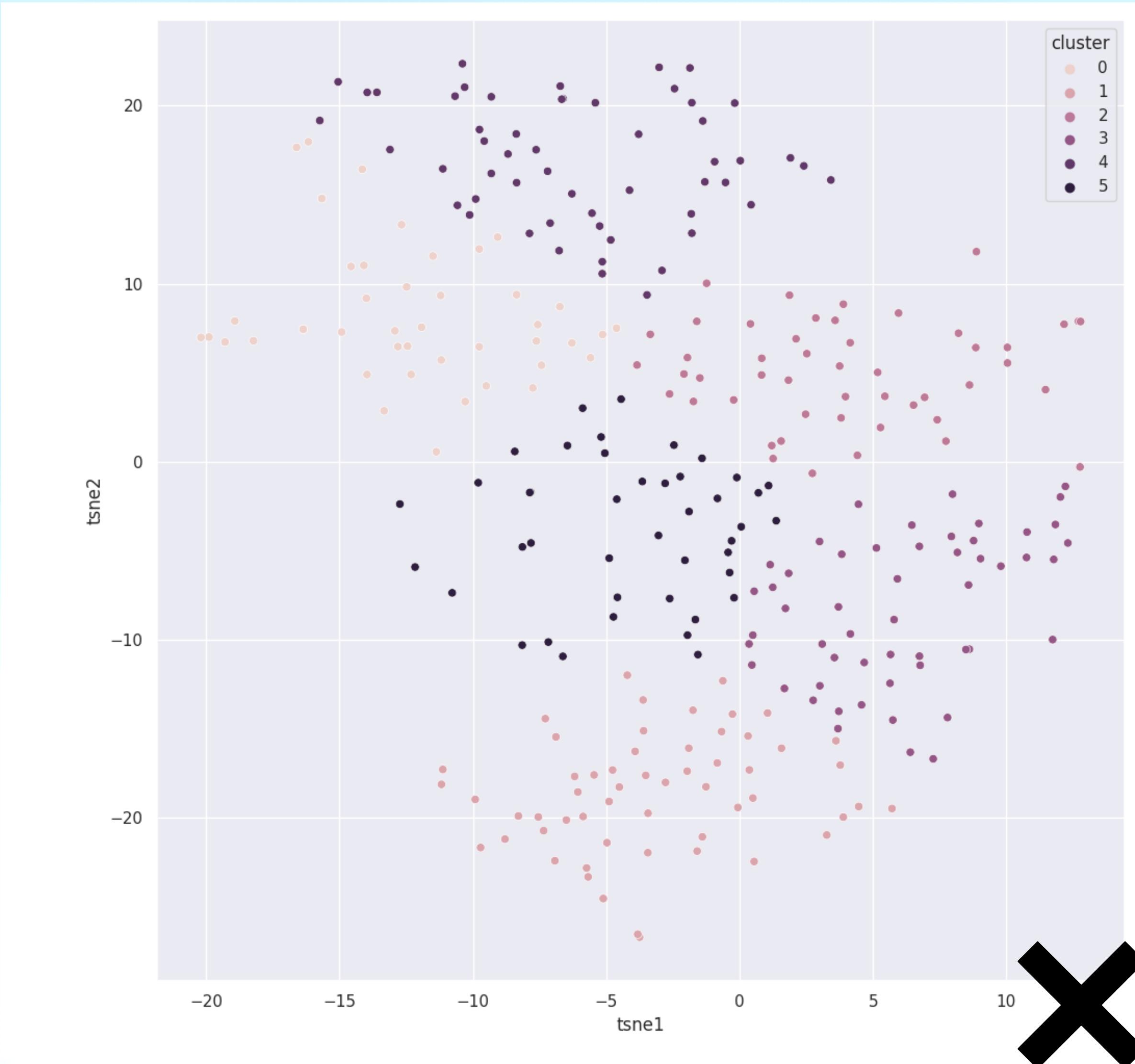


### Conclusion

- Le modèle SIFT ne permet d'admettre que la classification des produits en fonction de leur image est faisable.
- Les scores ARI obtenus varient entre 0.05 et 0.09
- Deux nombreux paramètres font varier le résultats de l'étude : le nombre de cluster de descripteur k, les paramètres du t-sne
- L'étude est coûteuse en terme de calcul.  
Échantillonage à 300 images pour aboutir.

# Classification automatique des produits en fonction de leur images

## Modèle SIFT



Modèle SIFT, ARI = 0.06

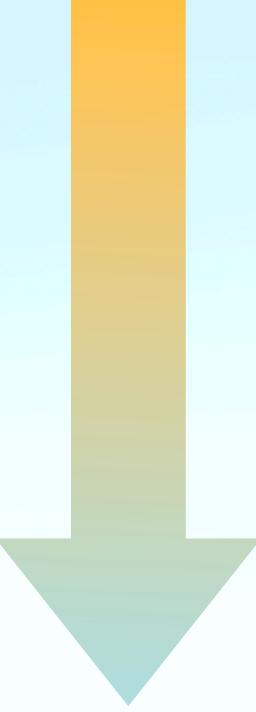
### Conclusion

- Le modèle SIFT ne permet d'admettre que la classification des produits en fonction de leur image est faisable.
- Les scores ARI obtenus varient entre 0.05 et 0.09
- Deux nombreux paramètres font varier le résultats de l'étude : le nombre de cluster de descripteur k, les paramètres du t-sne
- L'étude est coûteuse en terme de calcul.  
Échantillonage à 300 images pour aboutir.

### 3. Étude de faisabilité : classification automatique des données visuelles

#### C. Feature engineering - Modèle VGG16 par Transfer learning

Modèle de CNN (Convolutional Neural Network) qui a révolutionné le domaine de l' « image recognition » en 2014.

- 
1. On charge le modèle pré-entraîné sur plus de 15 millions d'image et 20 000 catégories.
  2. On extrait les features pour chacune des images avec un .predict

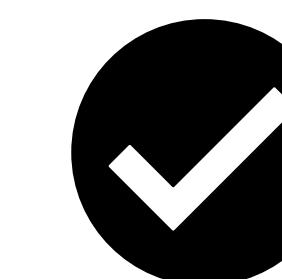
# Classification automatique des produits en fonction de leurs images

## Modèle VGG16



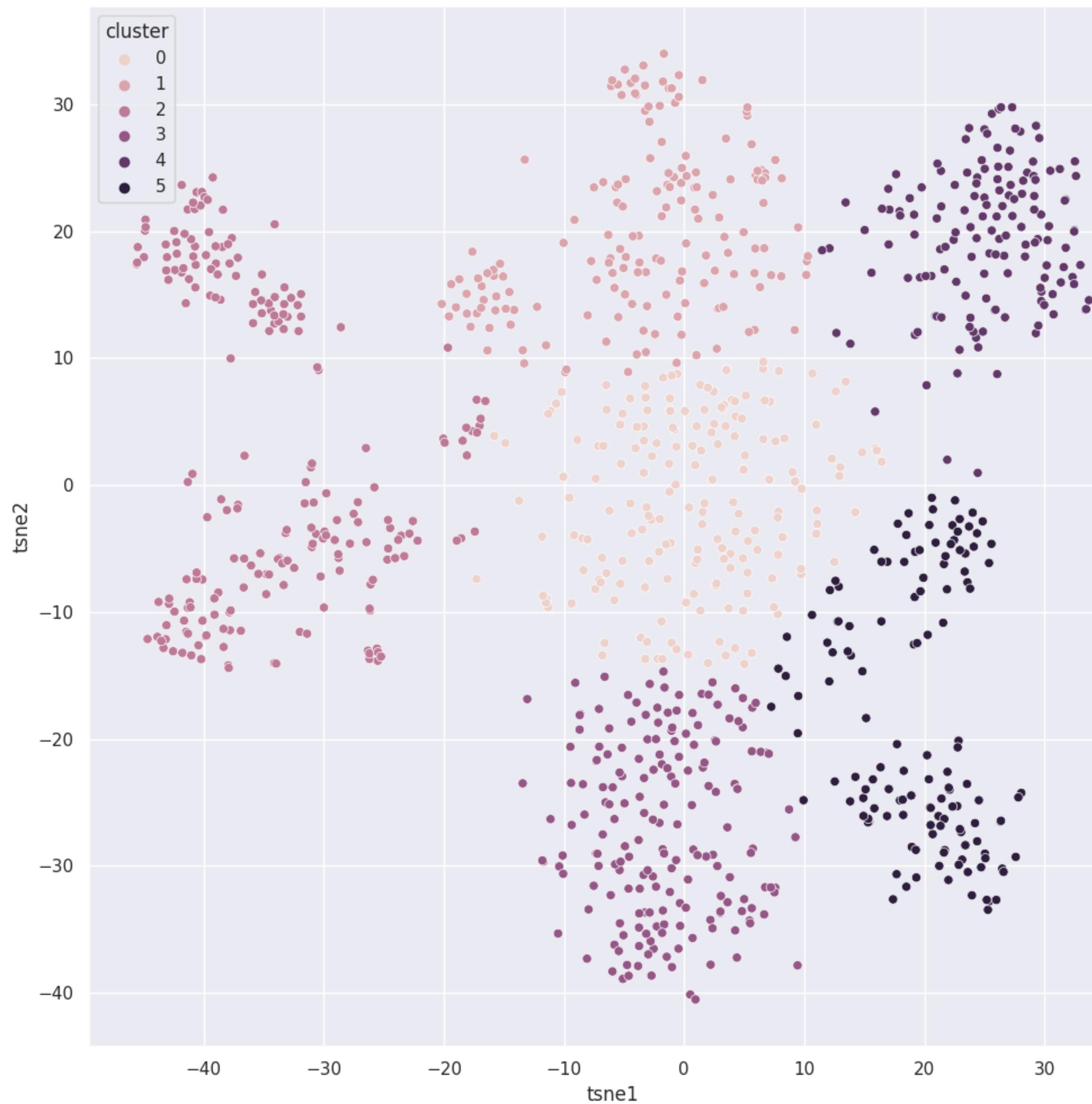
## Conclusion

L'étude de faisabilité démontre que **6 catégories de produit peuvent se distinguer automatiquement** en fonction uniquement de leurs images en utilisant le modèle pré-entraîné VGG16

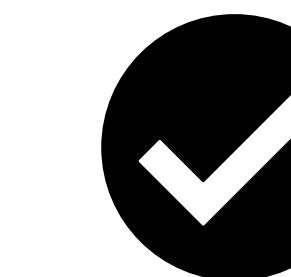


# Classification automatique des produits en fonction de leurs images

## Modèle VGG16



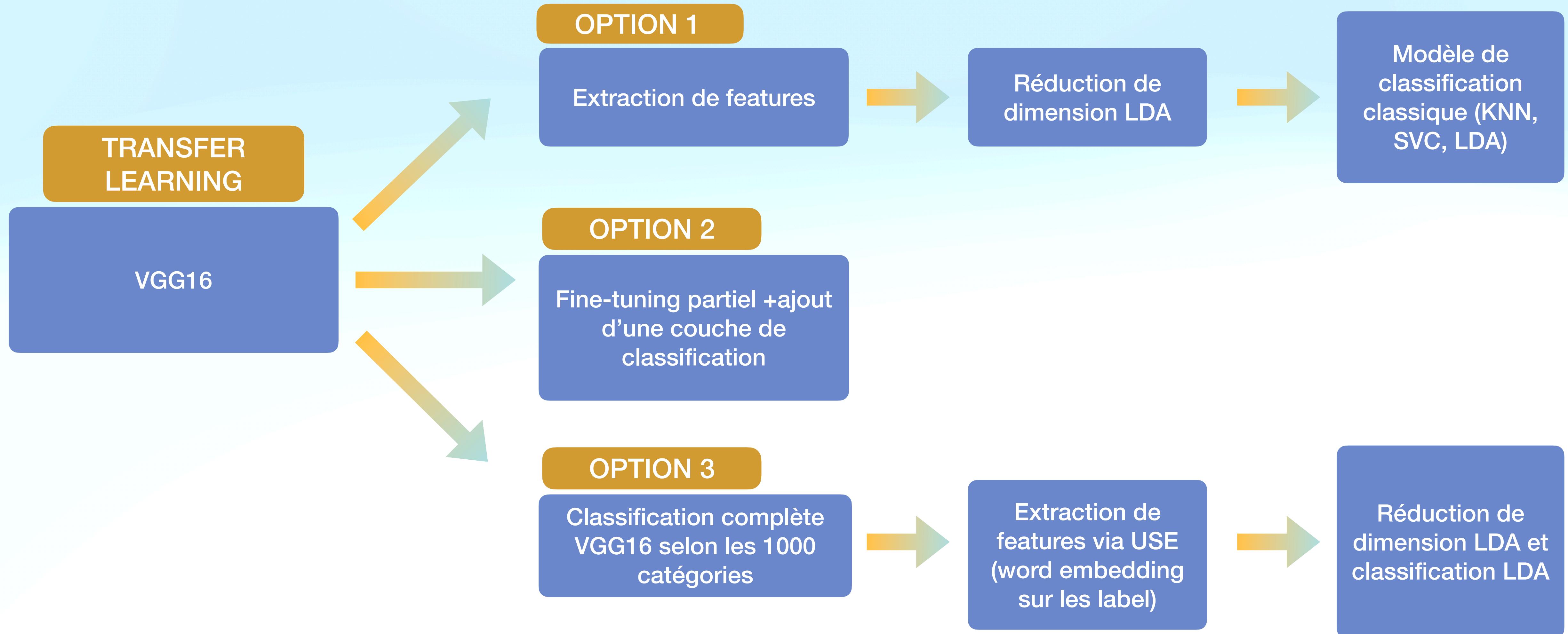
Modèle VGG16, ARI = 0.38



## Conclusion

L'étude de faisabilité démontre que **6 catégories de produit peuvent se distinguer automatiquement** en fonction uniquement de leurs images en utilisant le modèle pré-entraîné VGG16

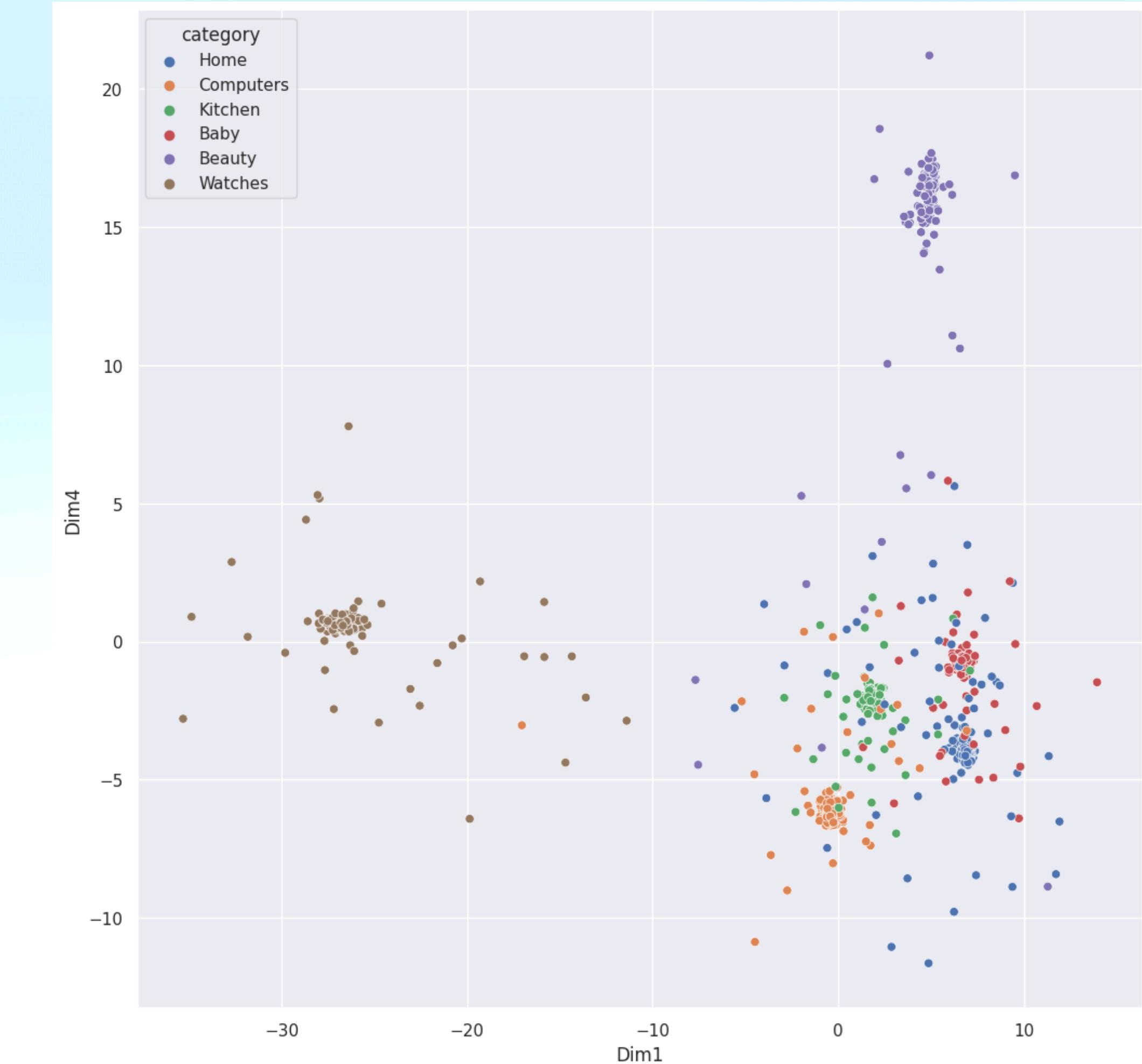
# 5. Résultats de la classification supervisée d'images



# 5. Résultats de la classification supervisée d'images

## Linear Discriminant Analysis (LDA)

Le choix de la LDA comme réducteur de dimension est motivé par sa faculté à maximiser la variance inter-classe

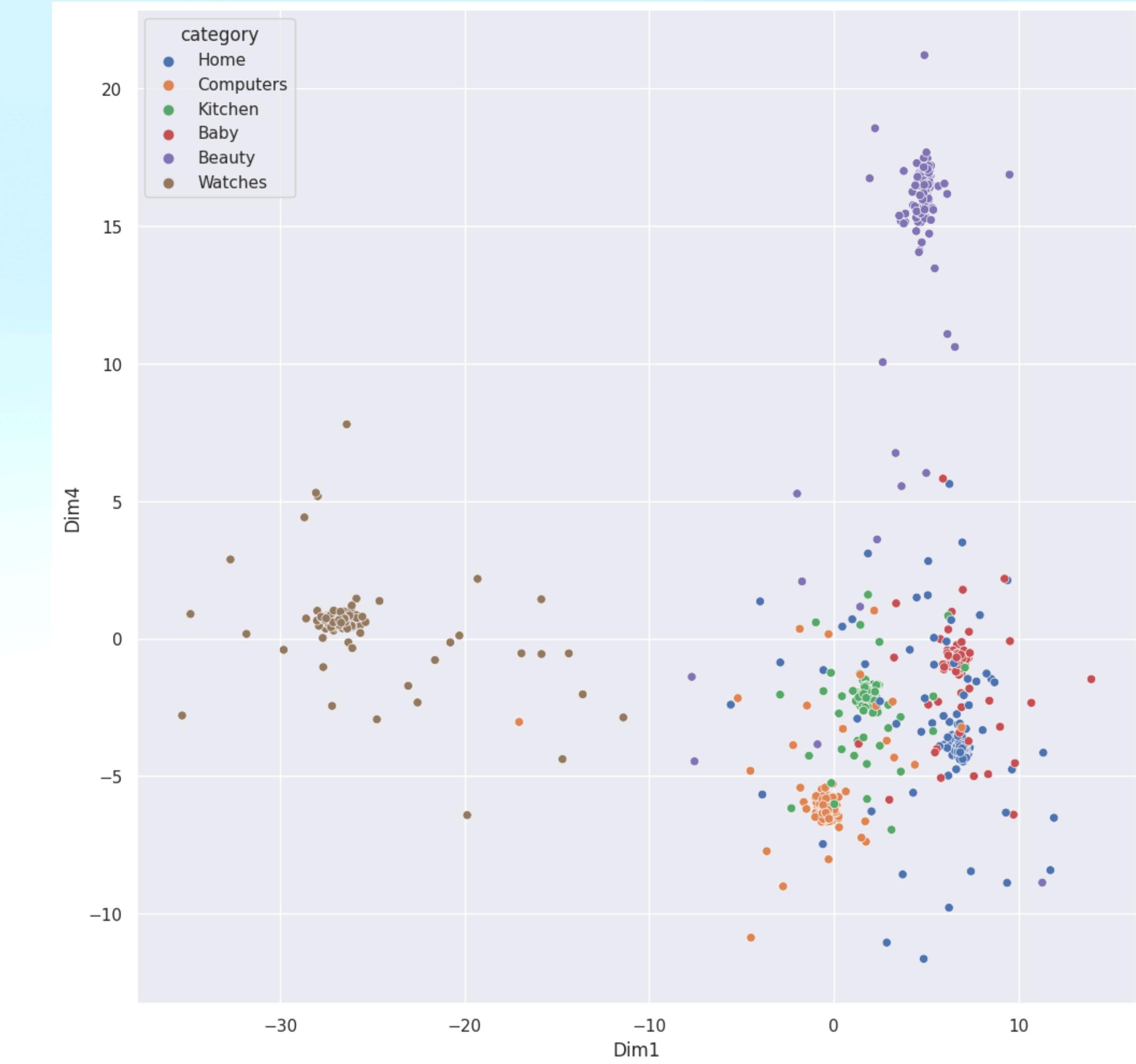


Linear Discriminant Analysis (LDA) - OPTION 1  
Extraction de feature VGG16 puis réduction de dimension via LDA

# 5. Résultats de la classification supervisée d'images

## Linear Discriminant Analysis (LDA)

Le choix de la LDA comme réducteur de dimension est motivé par sa faculté à maximiser la variance inter-classe

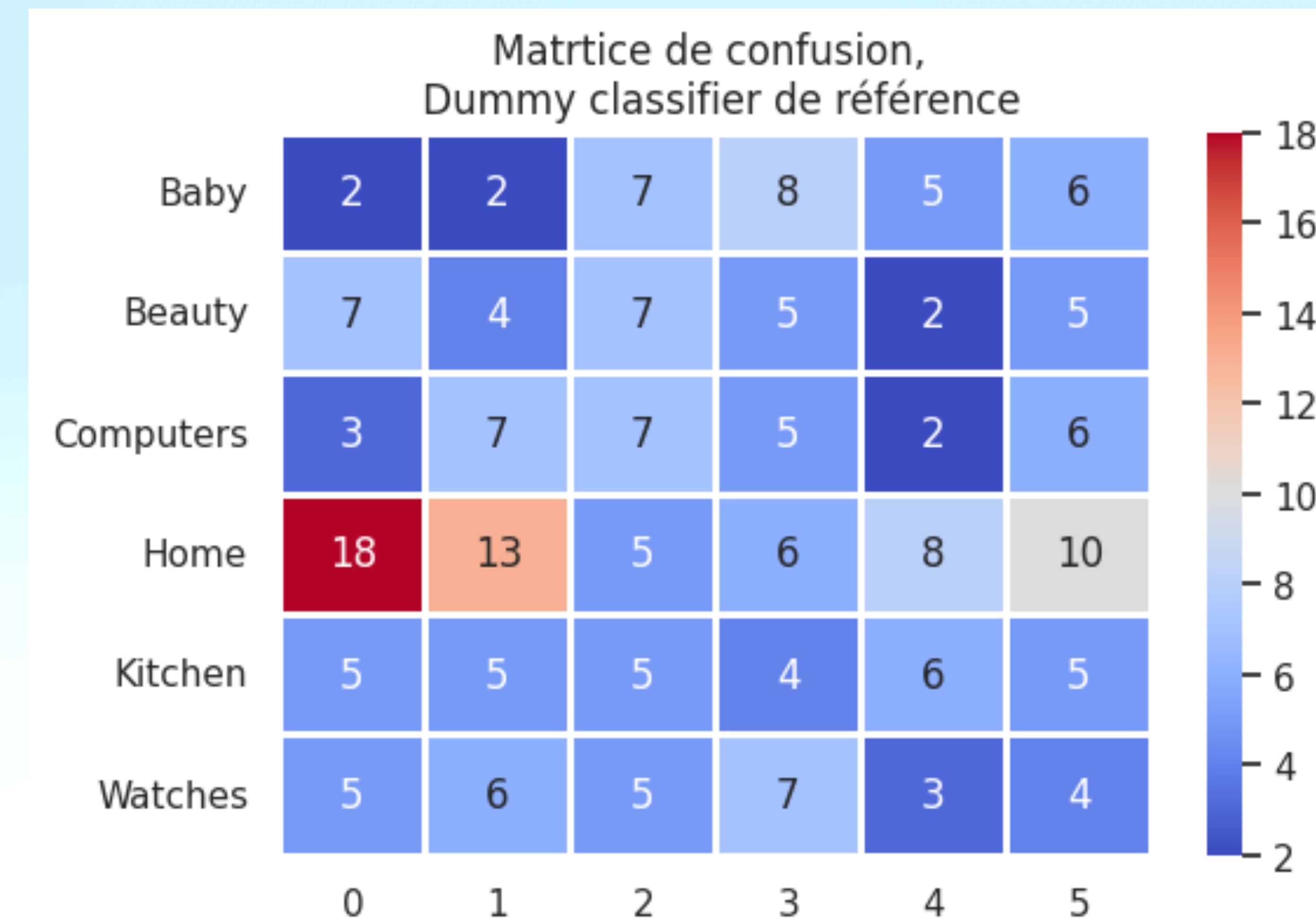


Linear Discriminant Analysis (LDA) - OPTION 1  
Extraction de feature VGG16 puis réduction de dimension via LDA

# 5. Résultats de la classification supervisée d'images

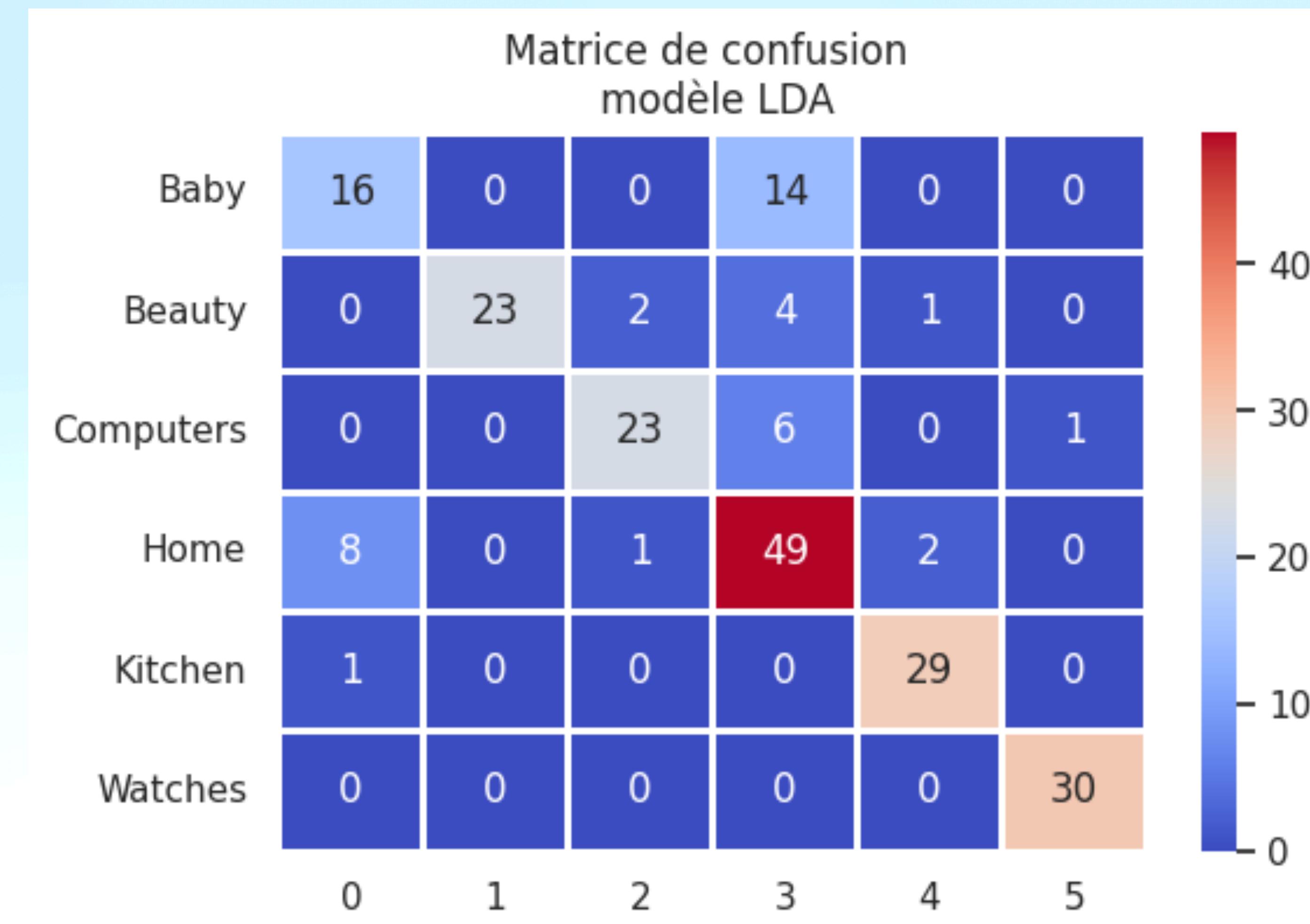
	#	Modèles testés	Temps d'entraînement	Accuracy moyenne (validation)	Accuracy Moyenne (test)	Précision moyenne (test)	Sensitivité moyenne (Test)
<b>OPTION 1</b> Extraction de feature via VGG16	1.0	Dummy classifier	-	-	0.14	0.15	0.14
	1A	Classification par Linear Discriminant Analysis (LDA)	31 sec	0.86	0.81	0.84	0.81
	1B	Classification par KNN	30 sec	0.85	0.81	0.85	0.81
	1C	Classification par SVC kernel linear	30 sec	0.85	0.81	0.84	0.81
<b>OPTION 2</b> Fine-tuning partiel	2A	Modèle VGG16 + Classifieur   entraînement des 10 dernières couches   epoch = 50   batch_size = 64   patience = 5	4h10	0.33	0.25	-	-
	2B	Modèle VGG16 sans les couches fully-connected + Couche de pooling + Classifieur (toutes les couches du VGG basiques sont fixes)   epoch = 50   batch_size = 64   patience = 5	1h38	0.82	0.79	0.81	0.78
<b>OPTION 3</b> Classification VGG16 complète	3	Modèle VGG16 total + USE en transfer learning sur les labels + Classification par Linear Discriminant Analysis	30 sec	0.76	0.73	0.75	0.72

## 5. Résultats de la classification supervisée d'images

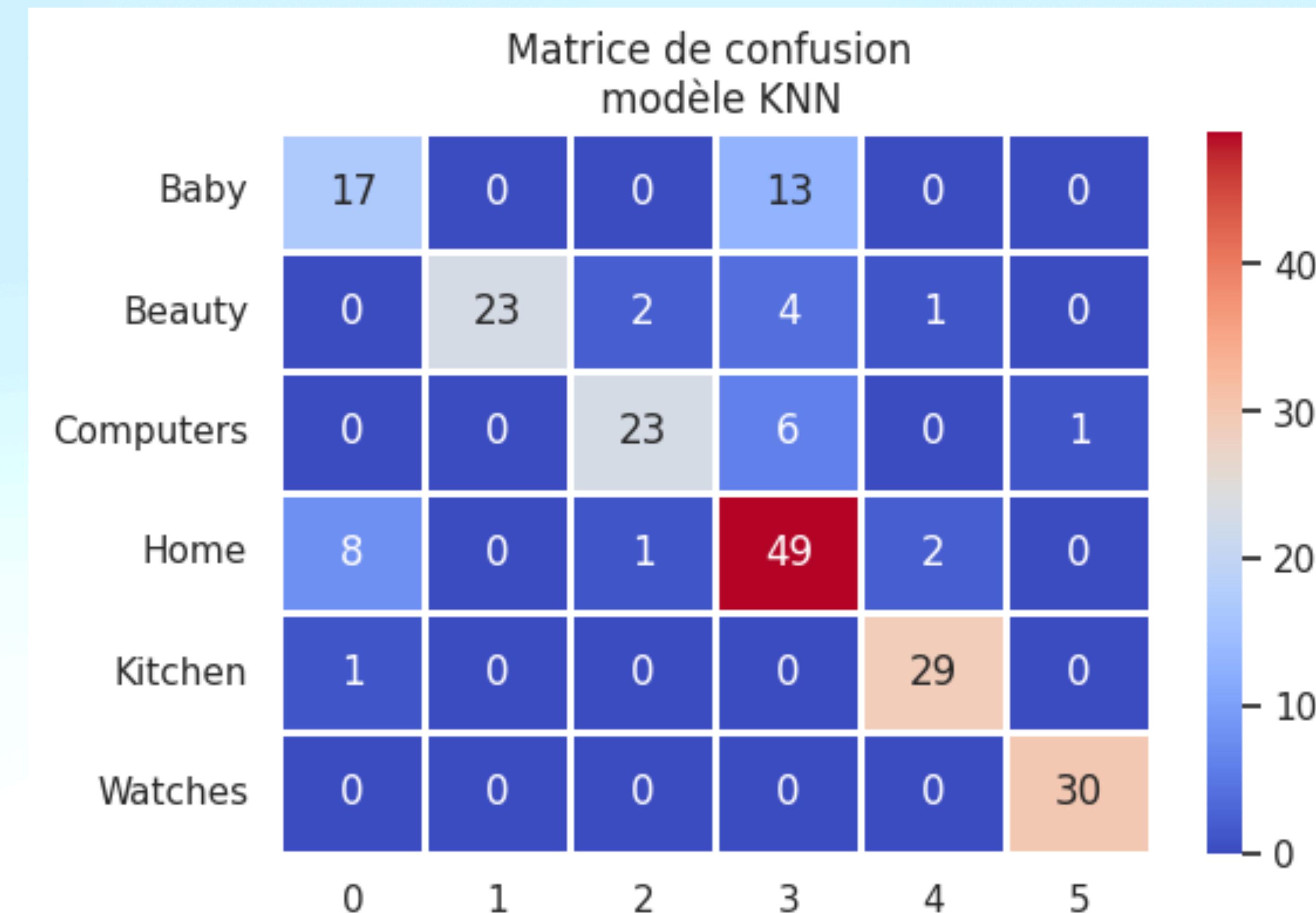


Extraction de features VGG16 + Dummy Classifier dans l'espace LDA

## 5. Résultats de la classification supervisée d'images

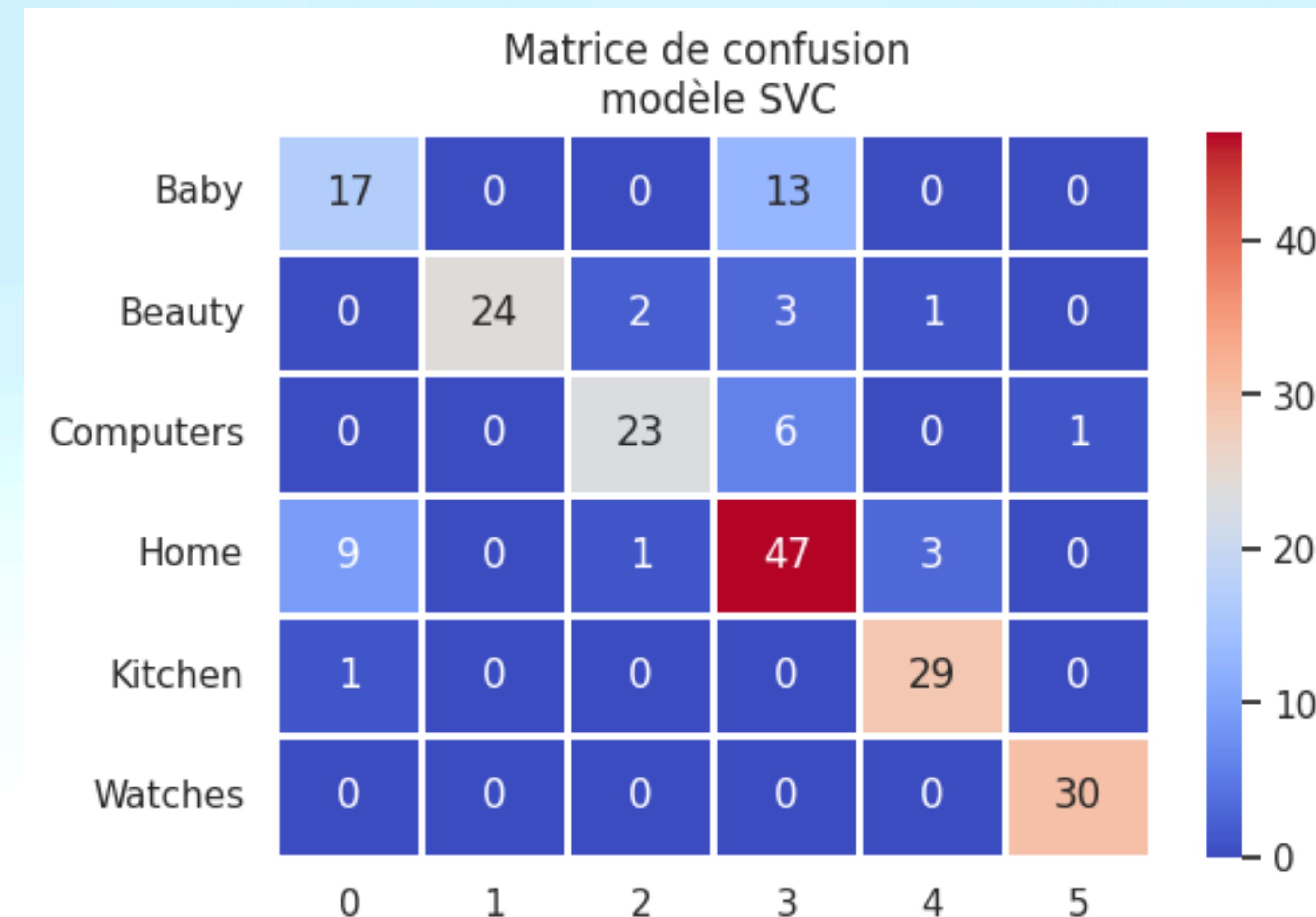


## 5. Résultats de la classification supervisée d'images



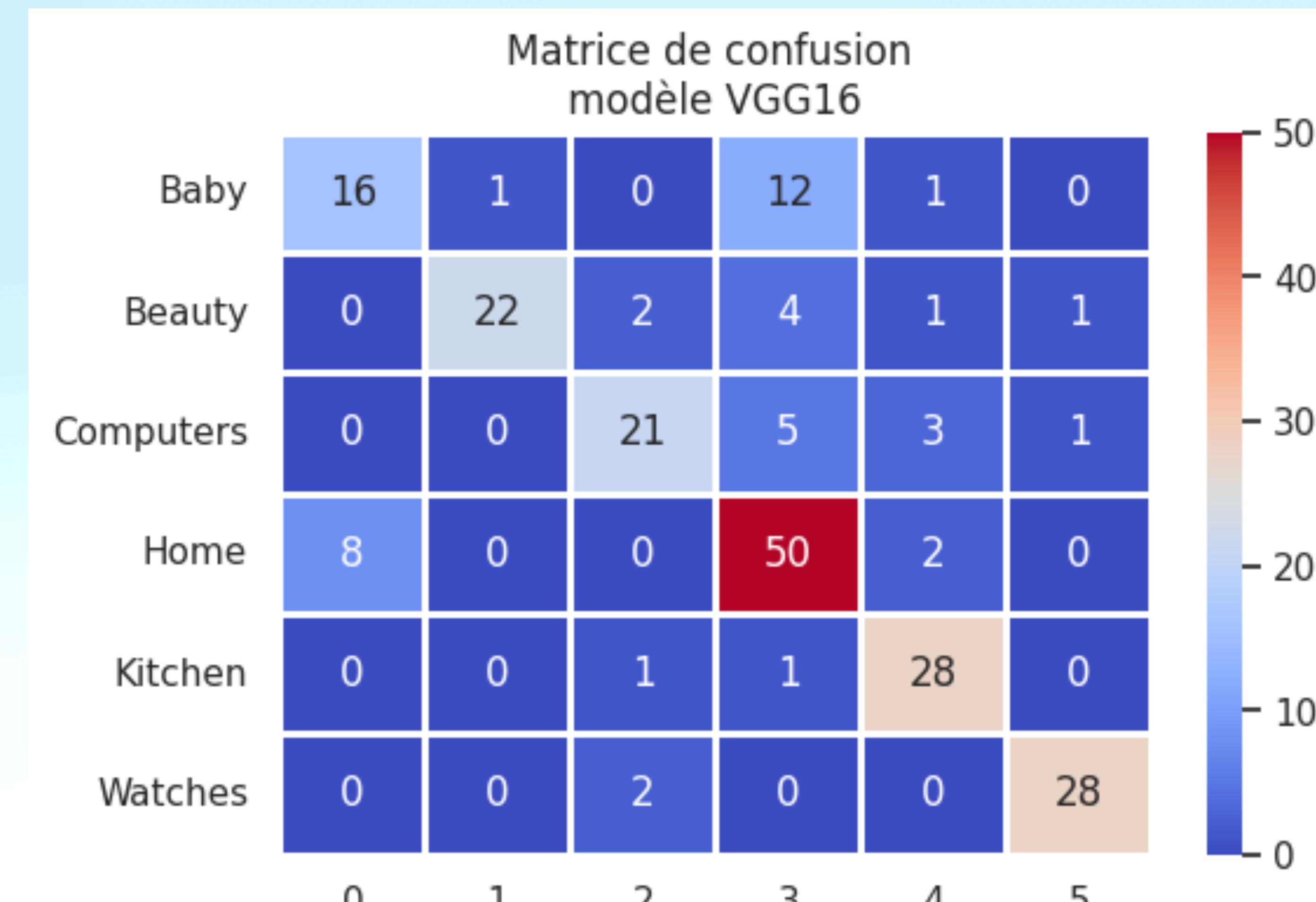
Extraction de features VGG16 + entraînement du modèle KNN avec  $k = 4$  dans l'espace LDA.

## 5. Résultats de la classification supervisée d'images



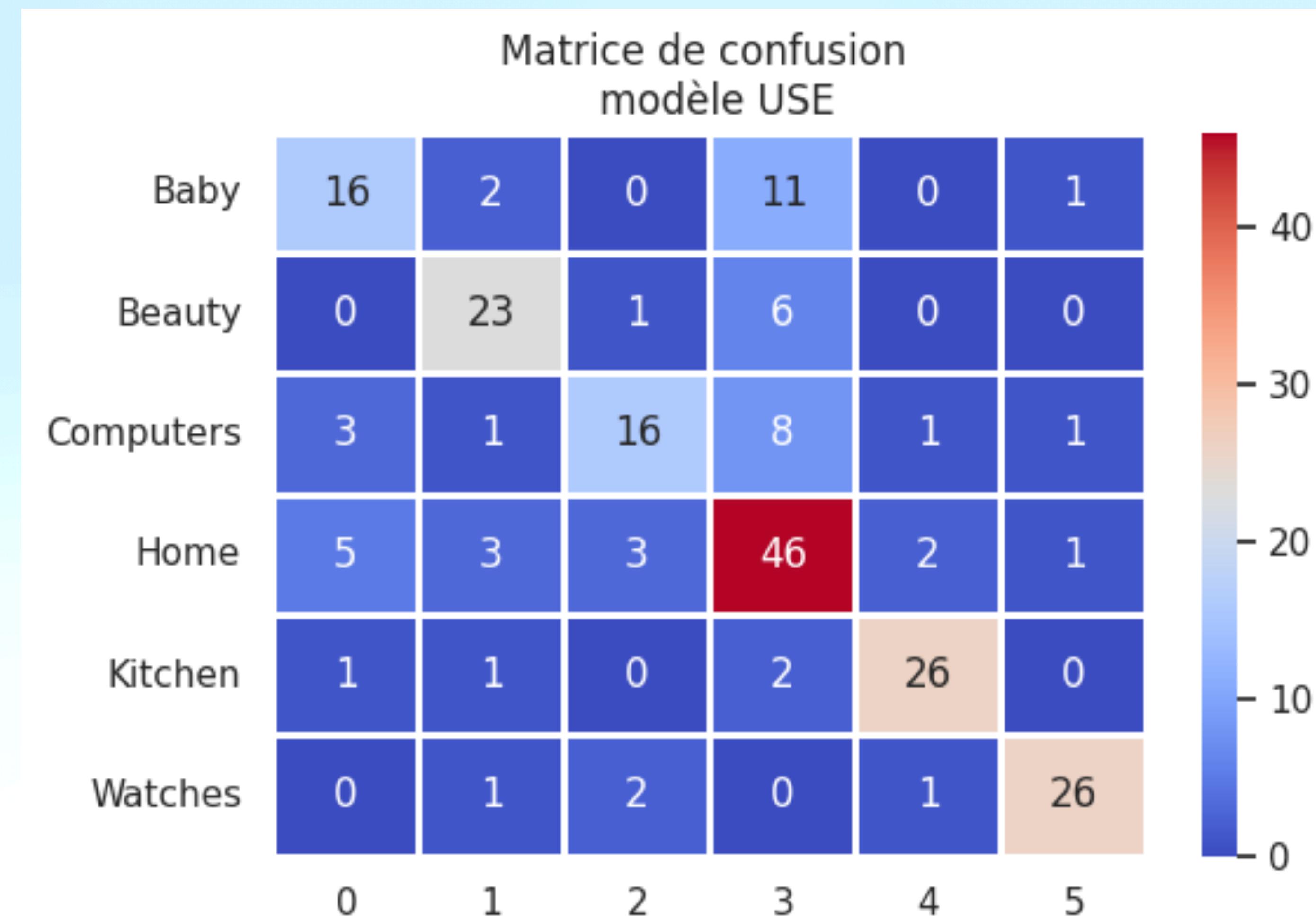
Extraction de features VGG16 + entraînement du modèle SVC avec k = 4 dans l'espace LDA.

## 5. Résultats de la classification supervisée d'images



Fine-tuning partiel. Entraînement du classifieur VGG16

## 5. Résultats de la classification supervisée d'images



# 5. Résultats de la classification supervisée d'images

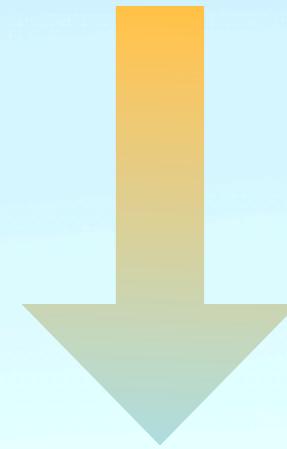
	#	Modèles testés	Temps d'entraînement	Accuracy moyenne (validation)	Accuracy Moyenne (test)	Précision moyenne (test)	Sensitivité moyenne (Test)
<b>OPTION 1</b> <b>Extraction de feature via VGG16</b>	1A	Classification par Linear Discriminant Analysis (LDA)	31 sec	0.86	0.81	0.84	0.81
	1B	Classification par KNN	30 sec	0.85	0.81	0.85	0.81
	1C	Classification par SVC kernel linear	30 sec	0.85	0.81	0.84	0.81

Remarque : Les 15 millions d'images et les 20 000 catégories de la base de données ImageNet, sur laquelle est pré-entraîné le modèle VGG16, comprennent des catégories proches des catégories FlipKart -> -> -> **Le modèle pré-entraîné est une base convenable !**

# **6. Data augmentation**

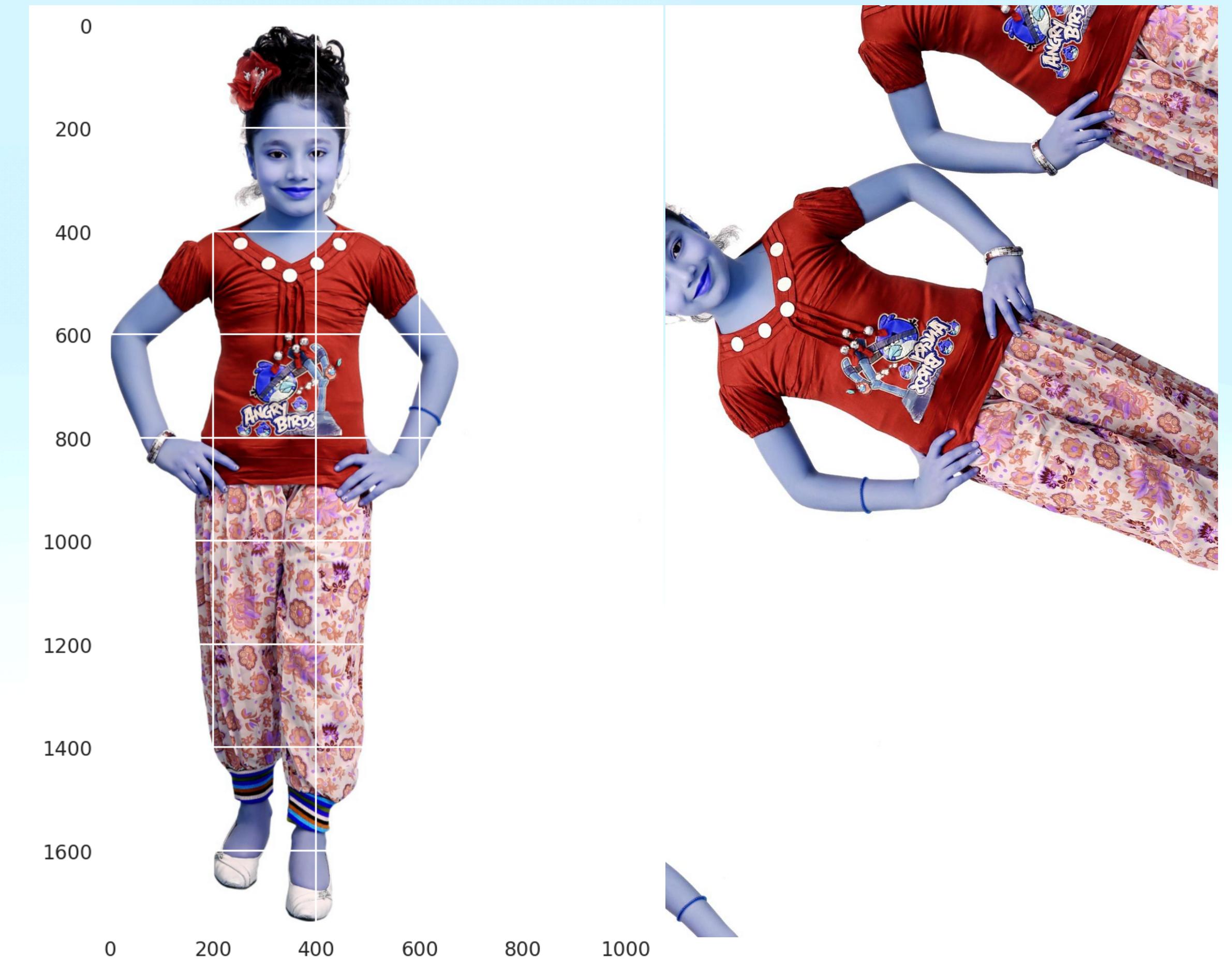
# 6. Data augmentation

Objectif : Améliorer les performances de mes modèles en augmentant les données d'entraînement.



## Démarche

1. Sélection de 120 images du jeu d'entraînement parmi chacune des deux catégories 'BABY' et 'HOME'.
2. Sélection aléatoire de 200 images du jeu d'entraînement toute catégorie confondues.
3. Traitement d'image aléatoire



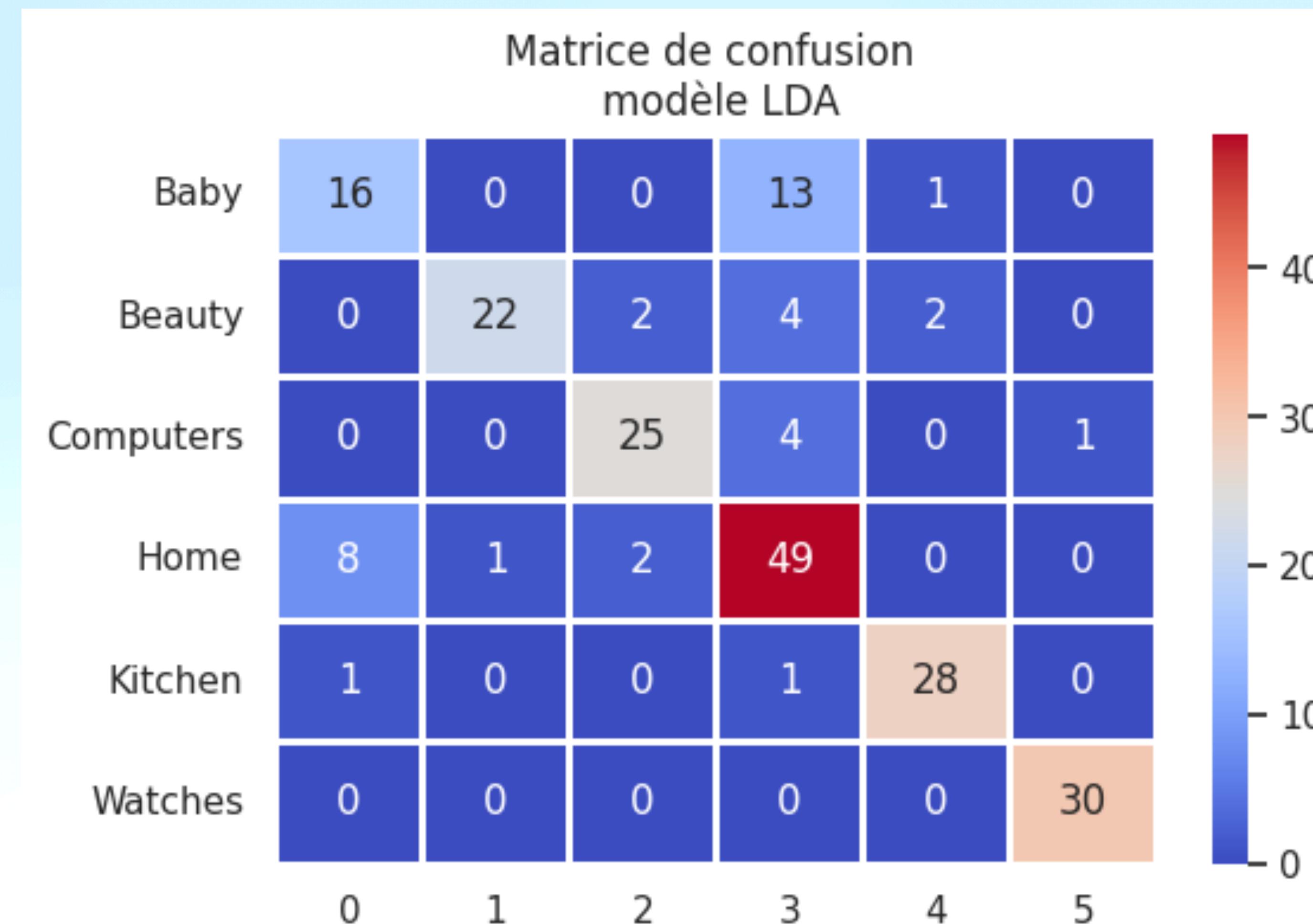
# 6. Data augmentation

## Résultats

	#	Modèles testés	Temps d'entraînement	Accuracy moyenne (validation)	Accuracy Moyenne (test)	Précision moyenne (test)	Sensitivité moyenne (Test)
<b>OPTION 1</b> Extraction de feature via VGG16	1A	Classification par Linear Discriminant Analysis (LDA)	31 sec	0.79	0.82	0.84	0.82
	1B	Classification par KNN	30 sec	0.79	0.82	0.84	0.82
	1C	Classification par SVC kernel linear	30 sec	0.8	0.82	0.84	0.83
<b>OPTION 2</b> Fine-tuning partiel	2B	Modèle VGG16 sans les couches fully-connected + Couche de pooling + Classifieur (toutes les couches du VGG basiques sont fixes)   epoch = 50   batch_size = 64   patience = 5	3h30	0.85	0.79	0.82	0.79

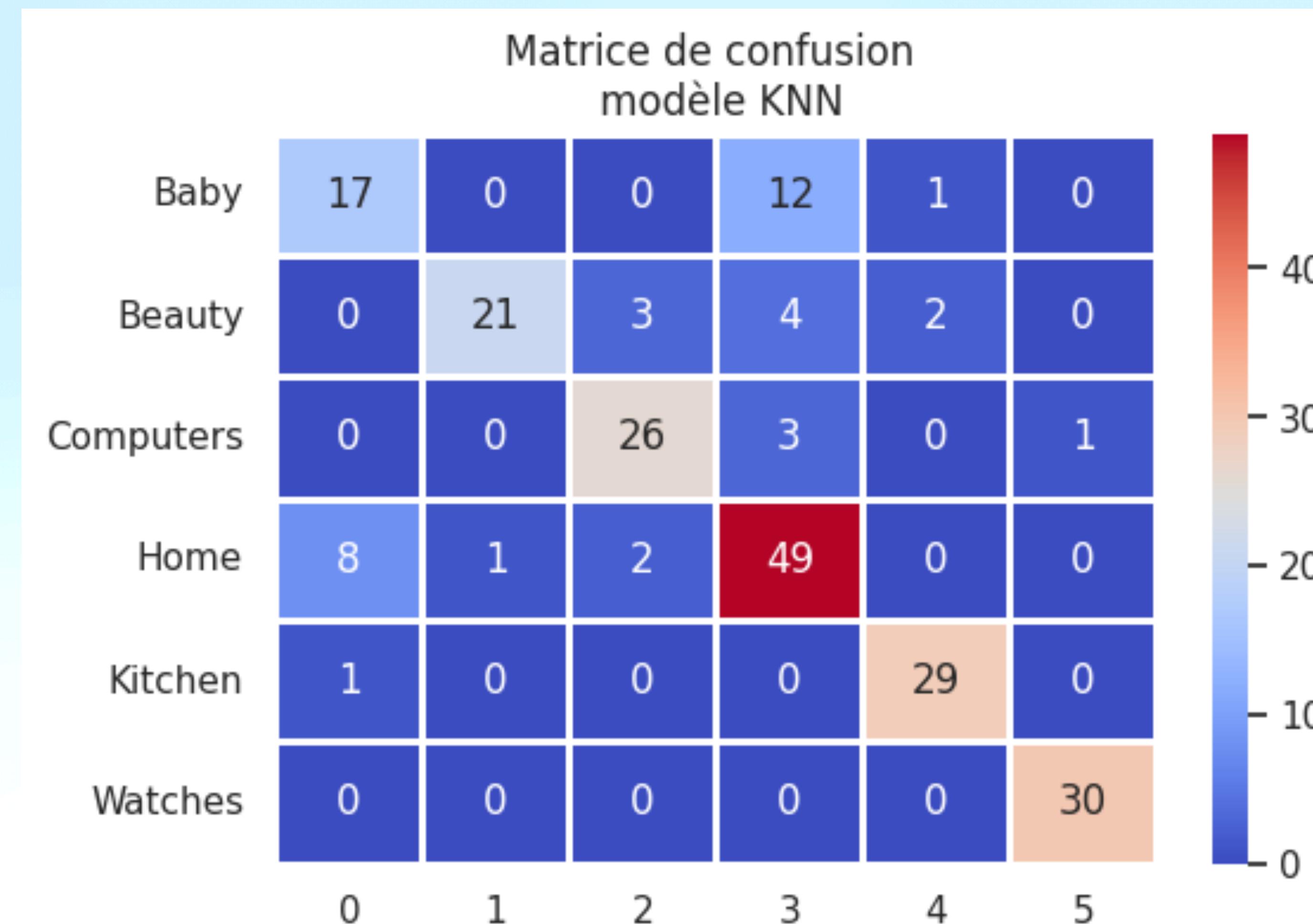
- On observe une légère augmentation de la sensibilité moyenne (ou recall)

# 6. Data augmentation



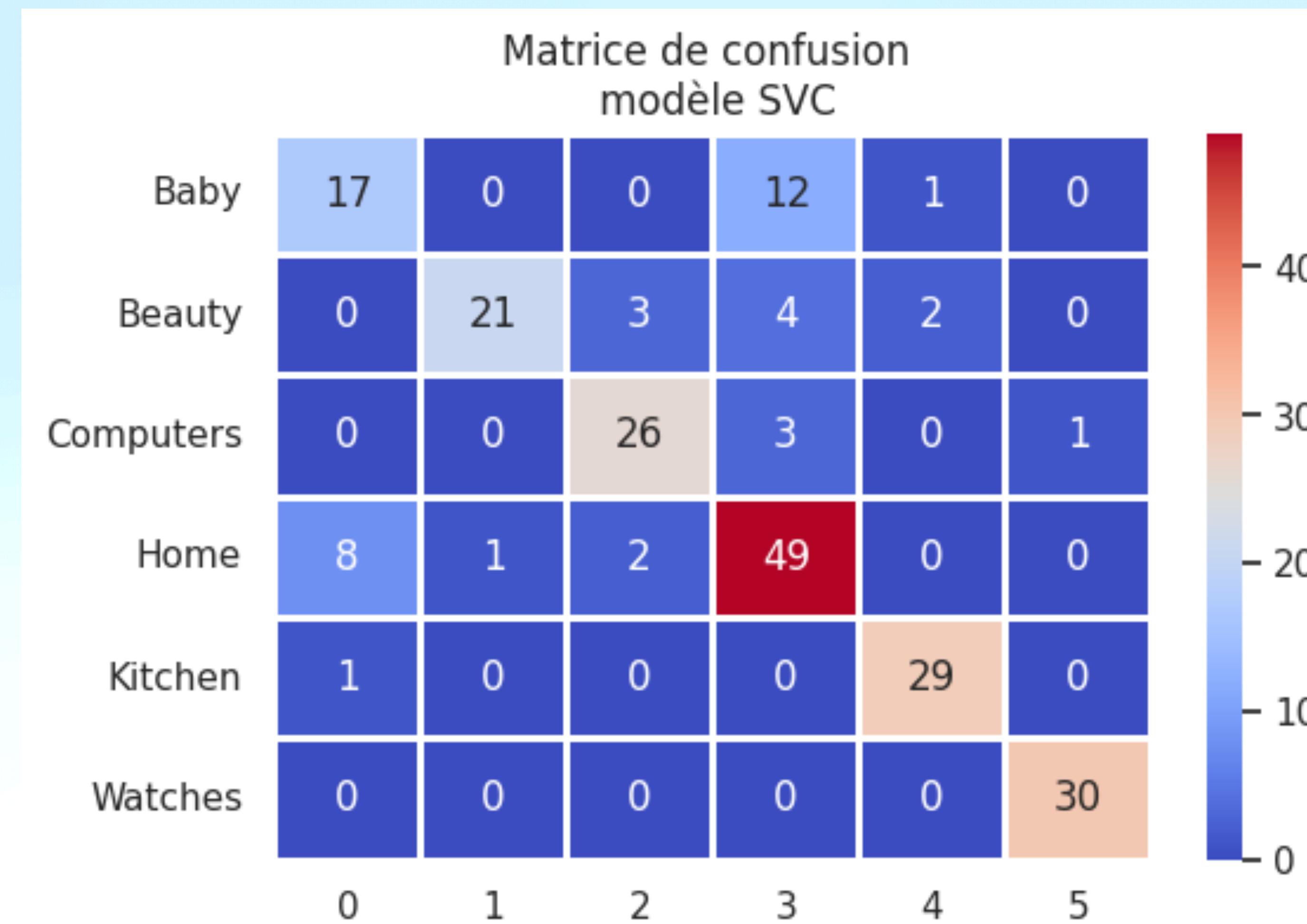
Extraction de features VGG16 + classification via LDA

# 6. Data augmentation



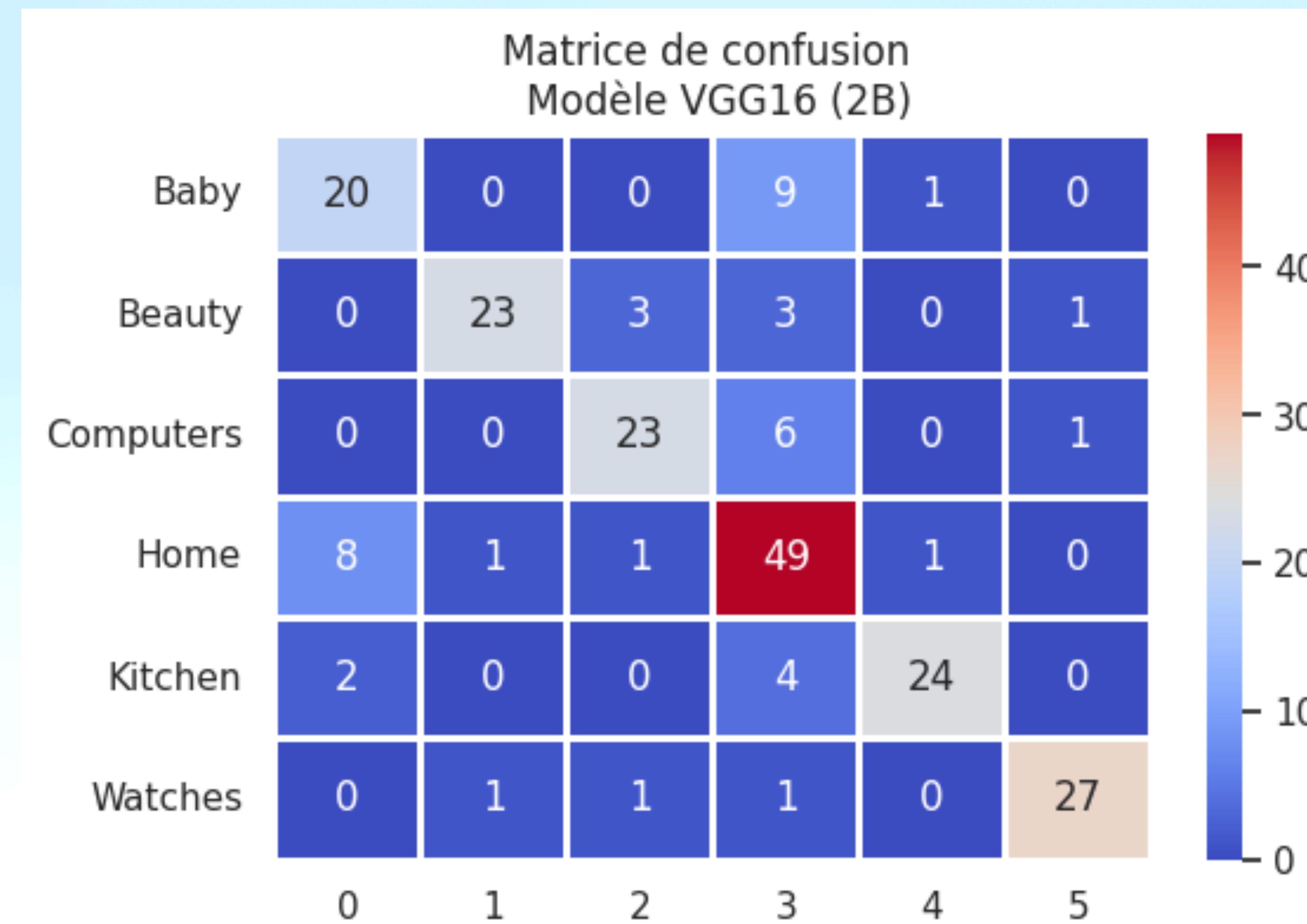
Extraction de features VGG16 + classification via KNN

# 6. Data augmentation



Extraction de features VGG16 + classification via SVC

# 6. Data augmentation



Fine-tuning partiel. Entraînement du classifieur VGG16

# 7. Test API

```
▶ import requests

url = "https://edamam-food-and-grocery-database.p.rapidapi.com/api/food-database/v2/parser"

querystring = {"ingr":"champagne"}

headers = {
    "X-RapidAPI-Key": "9f903a4ce5msh846c986d15894f5p1a6b3ajsn0198db85cf24",
    "X-RapidAPI-Host": "edamam-food-and-grocery-database.p.rapidapi.com"
}

response = requests.get(url, headers=headers, params=querystring)

print(response)
```

👤 <Response [200]>

Nouvelle compétence ! 🚀 🙌

- J'ai pu collecter des données à partir de l'API edamam-food.
- Utilisation de la librairie `requests`
- **La base de données n'étaient pas exploitables pour la data augmentation mais le test de collecte a fonctionné**

# 7. Test API

foodId	label	category	foodContentsLabel	image
0 food_bu12urpb tuo9v6b4jpvk2a1fh4hh	Champagne Simply Dressed Vinaigrette, Champagne	Packaged foods	FILTERED WATER; CANOLA OIL; CHAMPAGNE AND WHIT...	<a href="https://www.edamam.com/food-img/736/736a3e27a6...">https://www.edamam.com/food-img/736/736a3e27a6...</a>
1 food_a656mk2a5dmqb2adiamu6bei hduu	Champagne	Generic foods		NaN <a href="https://www.edamam.com/food-img/a71/a718cf3c52...">https://www.edamam.com/food-img/a71/a718cf3c52...</a>
2 food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	<a href="https://www.edamam.com/food-img/ab2/ab2459fc2a...">https://www.edamam.com/food-img/ab2/ab2459fc2a...</a>
3 food_b3dyababjo54xobm6r8jzbghjqqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	<a href="https://www.edamam.com/food-img/d88/d88b64d973...">https://www.edamam.com/food-img/d88/d88b64d973...</a>
4 food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
5 food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
6 food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
7 food_alp44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	SUGAR; BUTTER; SHORTENING; VANILLA; CHAMPAGNE;...	NaN
8 food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	SUGAR; LEMON JUICE; BRANDY; CHAMPAGNE; PEACH	NaN
9 food_am5egz6aq3fpjlaf8xp kdbc2asis	Champagne Truffles	Generic meals	BUTTER; COCOA; SWEETENED CONDENSED MILK; VANIL...	NaN
10 food_bcz8rhiajk1fuva0vkfm eakbouc0	Champagne Vinaigrette	Generic meals	CHAMPAGNE VINEGAR; OLIVE OIL; DIJON MUSTARD; S...	NaN
11 food_a79xmny a6t ogre aeu kbroa0thhh0	Champagne Chicken	Generic meals	FLOUR; SALT; PEPPER; BONELESS, SKINLESS CHICKE...	NaN
12 food_aoxaf73b3o0igebj6wjga6kqhco	Strawberry Champagne	Generic meals	CHAMPAGNE; STRAWBERRIES	NaN
13 food_ax1n26waalpd9cbc64bjob7pw6hg	Champagne Jelly	Generic meals	CHAMPAGNE; GELATINE; CASTER SUGAR; BLUEBERRIES	NaN
14 food_b4va8u0bb6pf74akh2rtcb3llna9	Champagne Punch	Generic meals	CHAMPAGNE; SIMPLE SYRUP; ORANGE JUICE; BLUEBER...	NaN
15 food_a4j8wm8ayflf13b45t3c3bk9w4ek	Champagne Sangria	Generic meals	MINT LEAVES; CHAMPAGNE; ORANGE JUICE; LEMON; L...	NaN
16 food_bw7gtgx bnn7nbwa62ppwpar9ljc1	Champagne Cotton Candy, Champagne	Packaged foods	SUGAR; ARTIFICIAL & NATURAL FLAVOR.	NaN
17 food_bba727vaimolf0b8stgoibx7ujei	Champagne Cake	Generic meals	FLOUR; BAKING POWDER; SALT; BUTTER; SUGAR; EGG...	NaN
18 food_a6mj2obbqy38soat01vrxaqn vvet	Champagne Cupcakes	Generic meals	BUTTER; SUGAR; EGGS; CHAMPAGNE; PLAIN YOGURT; ...	NaN
19 food_aj3tbbpb l068bhagn76uubtzyzyv	Champagne Vinegar	Packaged foods	CALIFORNIA CHAMPAGNE WINE VINEGAR; FRESH TARRA...	NaN

Nouvelle compétence !  

- J'ai pu collecter des données à partir de l'API edamam-food.
- Utilisation de la librairie requests
- **La base de données n'étaient pas exploitables pour la data augmentation mais le test de collecte a fonctionné**

# Conclusion

- Étudier la faisabilité d'un moteur de classification automatique d'articles, basé sur une image et une description, pour l'attribution de la catégorie de l'article. **OK ----- J'ai eu recours au Transfer Learning et aux modèles faisait intervenir les réseaux de neurones pour obtenir les résultats les plus convaincants**
- Implémentation d'un modèle supervisé de classification d'images. **OK ----- La solution la moins coûteuse consistait à utiliser l'OPTION 1 d'extraction de features à partir du modèle pré-entraîné VGG16. Modèle star de la compétition ImageNet.**
- Optimisation du modèle par data augmentation **OK ----- L'augmentation du jeu d'entraînement par traitement successif d'image a conduit à une légère amélioration des critères de classification.**
- Le test de l'API proposé a fonctionné bien que les données n'était pas disponible pour la data augmentation
- De nouvelles compétences acquise : NLP, Regex, RNN, CNN, collecte via API, Linear Discriminant analysis