Classifiez automatiquement des biens de consommations

Projet 6 – Parcours Data Scientist

Sommaire


Contexte & données
Etude de faisabilité (texte et image)
Classification supervisée
Test API
Conclusion & Perspectives

place de marché

Contexte

Place de marché est un site de e-commerce où les vendeurs proposent des articles à des acheteurs en postant une photo et une description.

Attribution de la catégorie

Manuelle

Actuellement :
- Peu fiable
- Volume très petit

Automatique

Futur :
- Plus fiable, expérience plus fluide
- Passage à l'échelle

Jeu de données d’articles

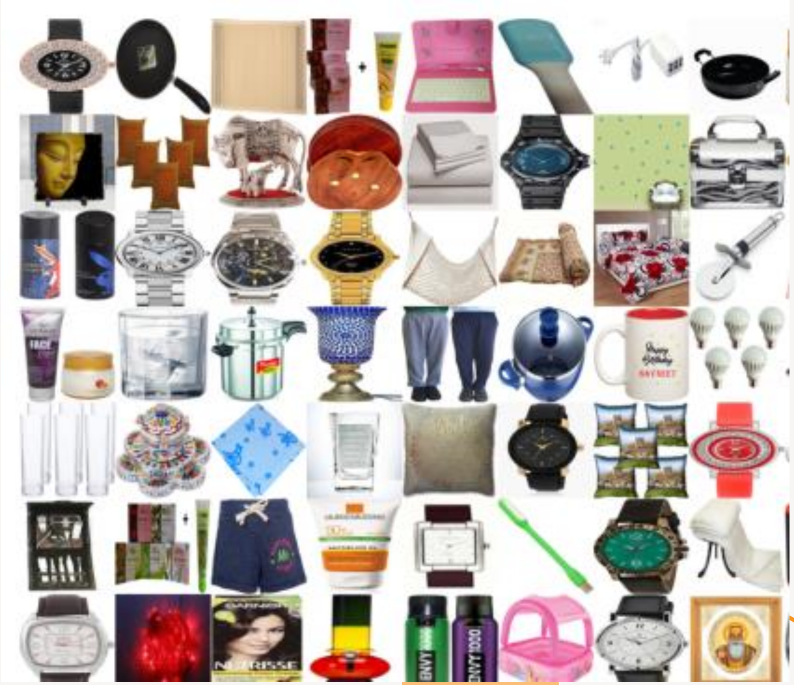
1050 articles

Description
15 variables

7 catégories principales

main_category	count
Home Furnishing	150
Baby Care	150
Watches	150
Home Decor & Festive Needs	150
Kitchen & Dining	150
Beauty and Personal Care	150
Computers	150

Images



Etude de faisabilité d'un moteur de classification automatique

Text pre-processing

```
Multicolor Abstract Eyelet Door Curtain (213 cm in
Height, Pack of 2) Price: Rs. 899 This curtain
enhances the look of the interiors.
```

```
'Multicolor', 'Abstract', 'Eyelet', 'Door', 'Curtain',
'213', 'cm', 'in', 'Height', 'Pack', 'of', '2',
'Price', 'Rs', '899', 'This', 'curtain', 'enhances',
'the', 'look', 'of', 'the', 'interiors'
```

```
'multicolor', 'abstract', 'eyelet', 'door', 'curtain',
'height', 'pack', 'price', 'this', 'curtain',
'enhances', 'the', 'look', 'the', 'interiors'
```

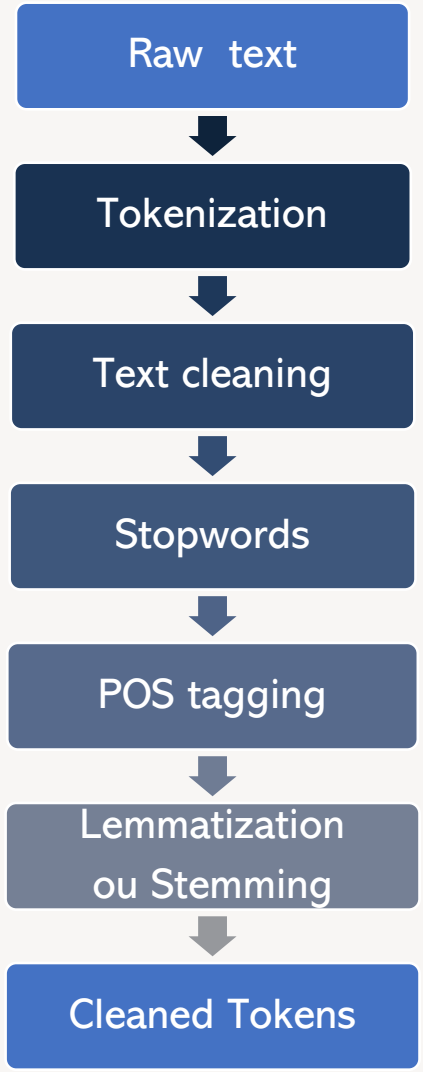
```
'multicolor', 'abstract', 'eyelet', 'door', 'curtain',
'height', 'pack', 'price', 'curtain', 'enhances',
'look', 'interiors'
```

- POS tagging + Lemmatization

```
'multicolor', 'abstract', 'eyelet', 'door', 'curtain',
'height', 'pack', 'price', 'curtain', 'enhances',
'look', 'interior'
```

- POS tagging + Stemming

```
'multicolor', 'abstract', 'eyelet', 'door', 'curtain',
'height', 'pack', 'price', 'curtain', 'enhanc',
'look', 'interior'
```



- caractères spéciaux
- minuscules
- len(token) >2

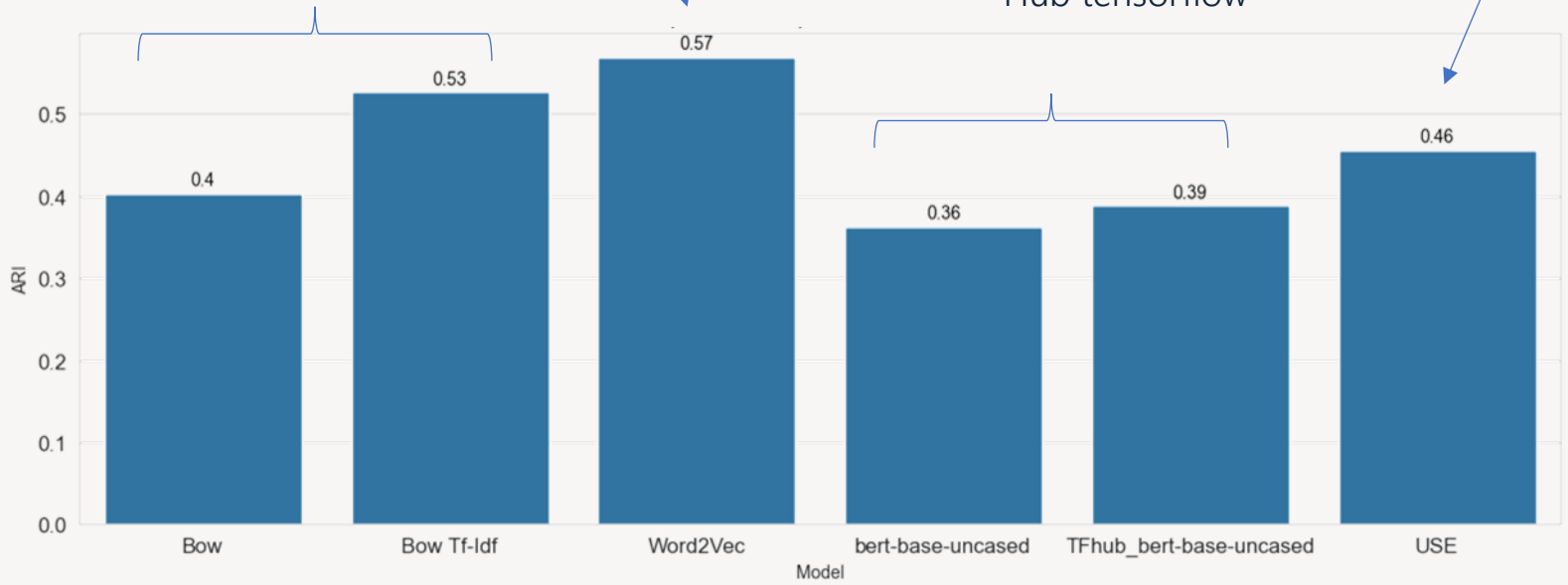
catégorie grammaticale : adjectif, adverbe, …

Features Extractions : modèles testés

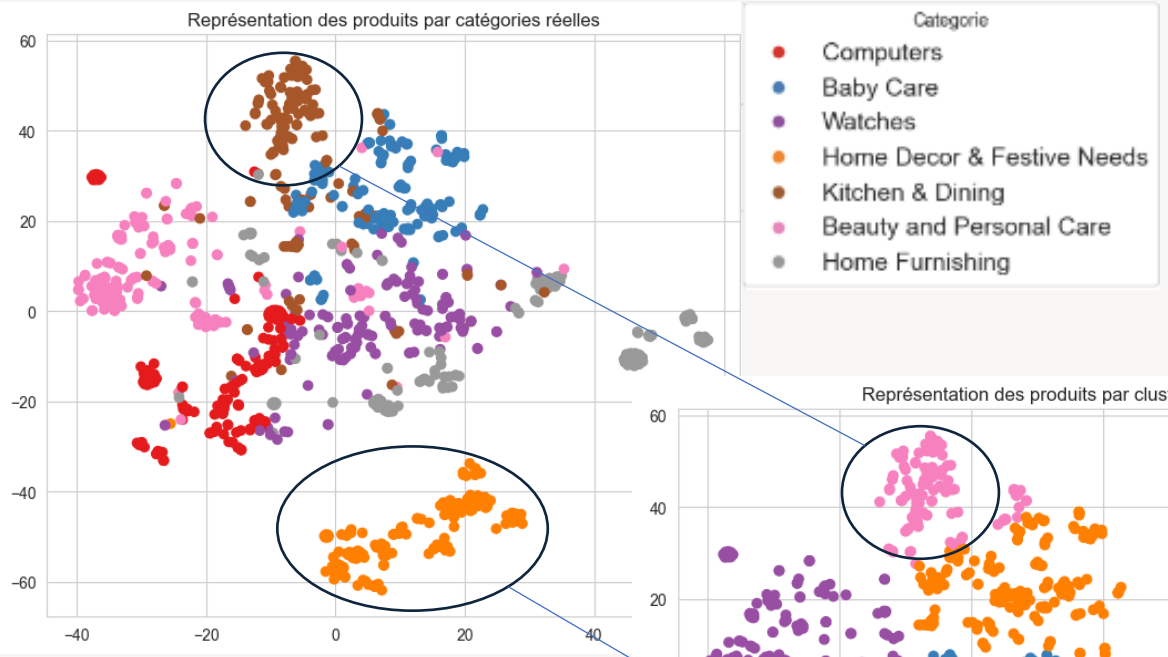
- Bag-of-words
- TF-Idf

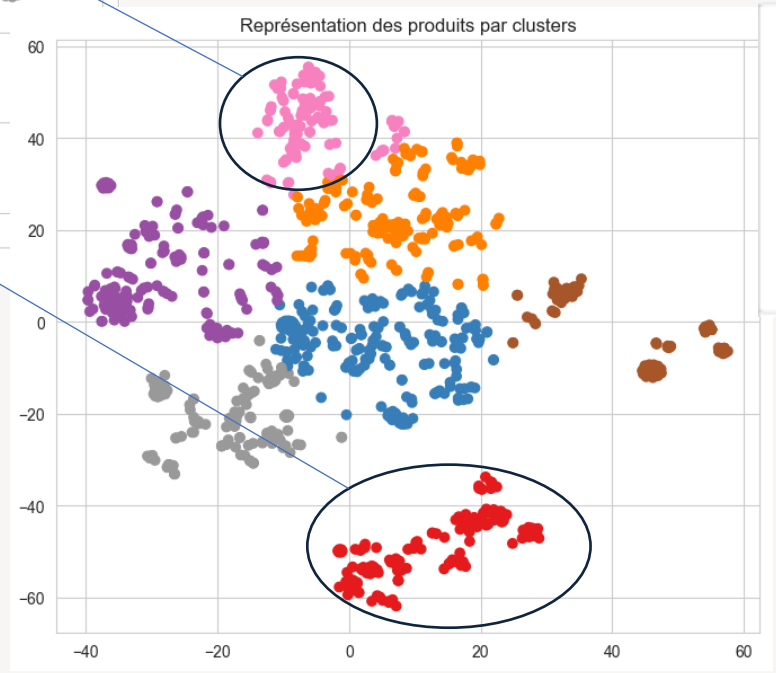
bert-base-uncased

- Huggingface
- Hub tensorflow



Word2Vec – ARI 0,57





Système de vote en utilisant les prédictions de plusieurs modèles?

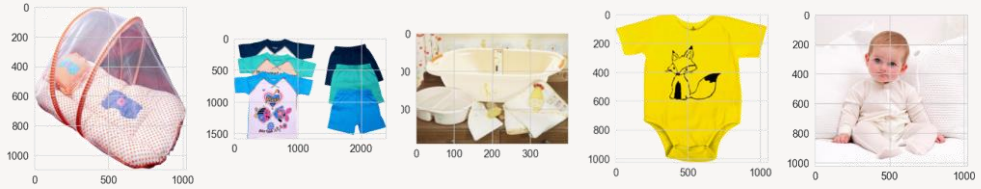
Model	ARI	silhouette_score	Precision / recall - Computers	Precision / recall - Home Furnishing	Precision / recall - Home Decor & Festive Needs	Precision / recall - Watches	Precision / recall - Baby Care	Precision / recall - Beauty and Personal Care	Precision / recall - Kitchen & Dining
Bow	0.4033	0.45	0.27 / 0.38	0.70 / 0.70	0.42 / 0.33	1.00 / 1.00	0.66 / 0.70	0.42 / 0.53	0.97 / 0.49
Bow Tf-Idf	0.5263	0.45	0.59 / 0.64	0.76 / 0.75	0.65 / 0.78	0.88 / 1.00	0.75 / 0.59	0.95 / 0.83	0.53 / 0.49
Word2Vec	0.5417	0.48	0.79 / 0.72	0.70 / 0.84	0.51 / 0.75	1.00 / 0.99	0.87 / 0.65	0.77 / 0.87	0.91 / 0.51
bert-base-uncased	0.3627	0.44	0.89 / 0.68	0.65 / 0.48	0.46 / 0.49	0.83 / 0.90	0.55 / 0.63	0.82 / 0.55	0.34 / 0.52
TFhub_bert-base-uncased	0.3739	0.44	0.88 / 0.78	0.61 / 0.46	0.46 / 0.49	0.83 / 0.90	0.60 / 0.63	0.76 / 0.53	0.35 / 0.51
USE	0.5406	0.47	0.89 / 0.88	0.59 / 0.63	0.67 / 0.73	0.95 / 1.00	0.62 / 0.67	0.66 / 0.58	0.81 / 0.69

Précision : Proportion de vrais positifs (TP) parmi les prédictions positives (TP+FP)
Recall : Proportion des vrais positifs (TP) parmi tous les éléments réellement positifs (TP+FN)

Etude de faisabilité d'un moteur de classification automatique

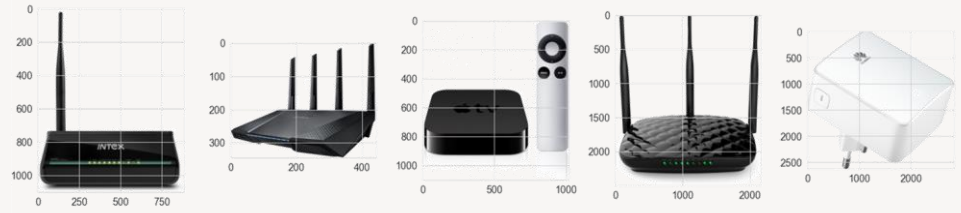
En image : des articles très différents au sein d'une même catégorie

Baby Care

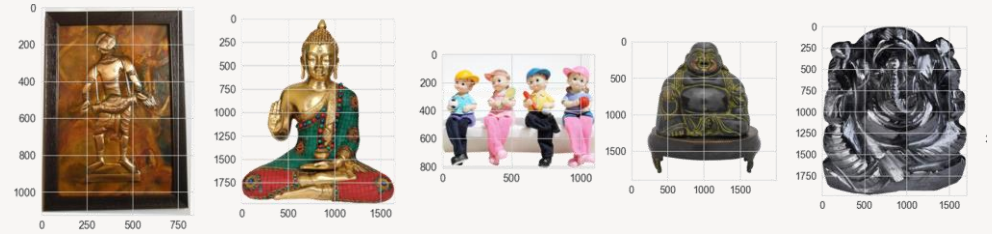


Beauty and Personal Care

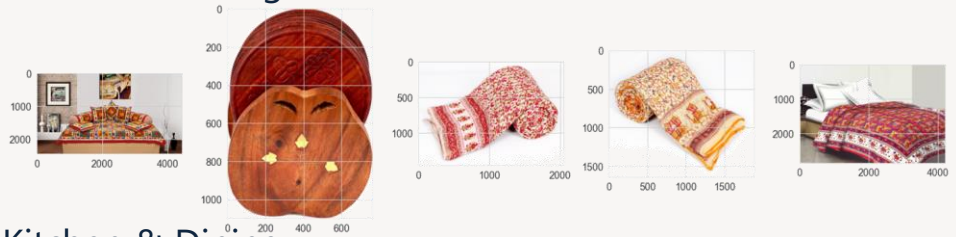
Computers



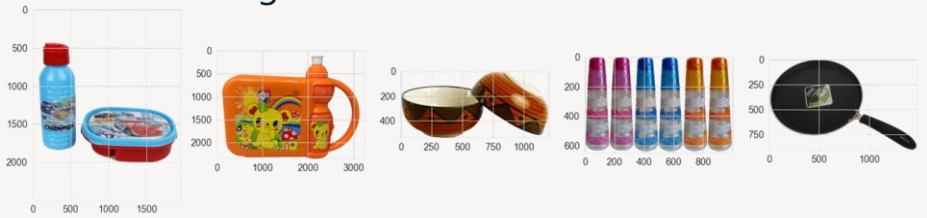
Home Decor & Festive Needs



Home Furnishing



Kitchen & Dining



Watches

Des images de dimensions différentes

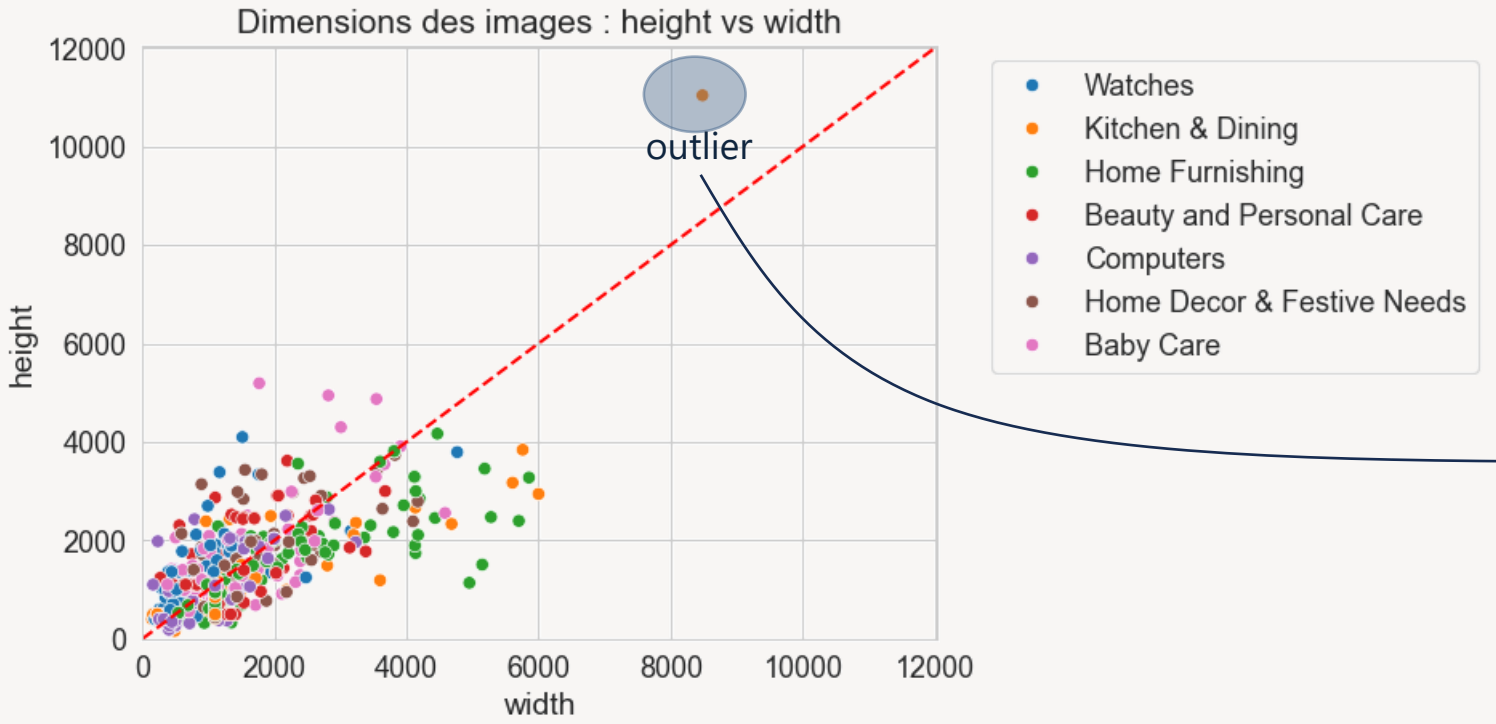
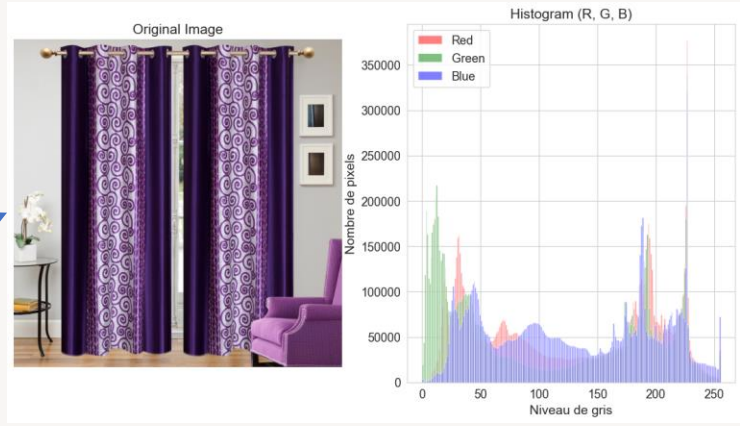


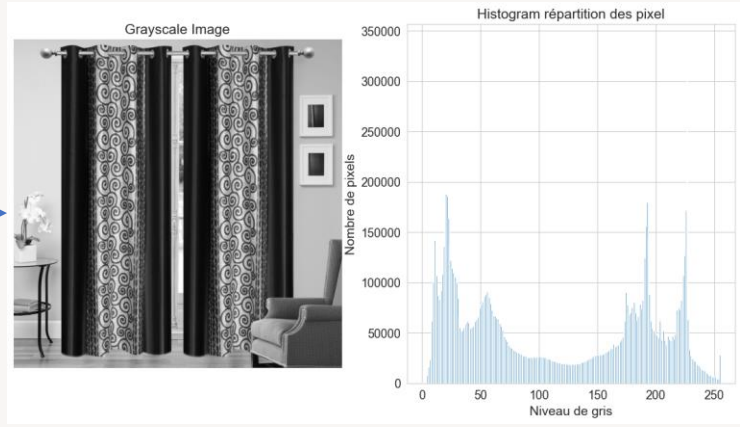
Image pre-processing

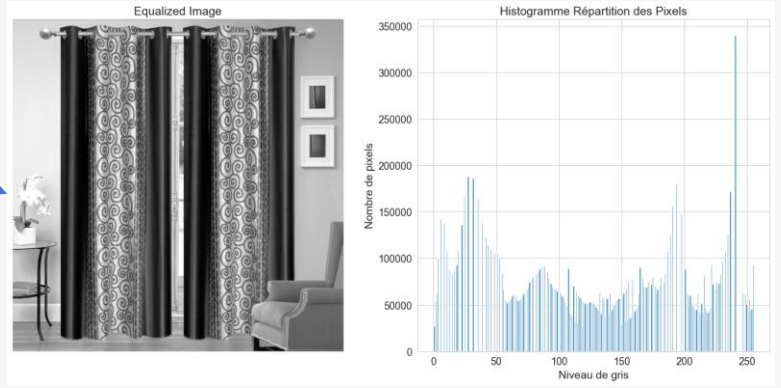
Raw image

Gray scale

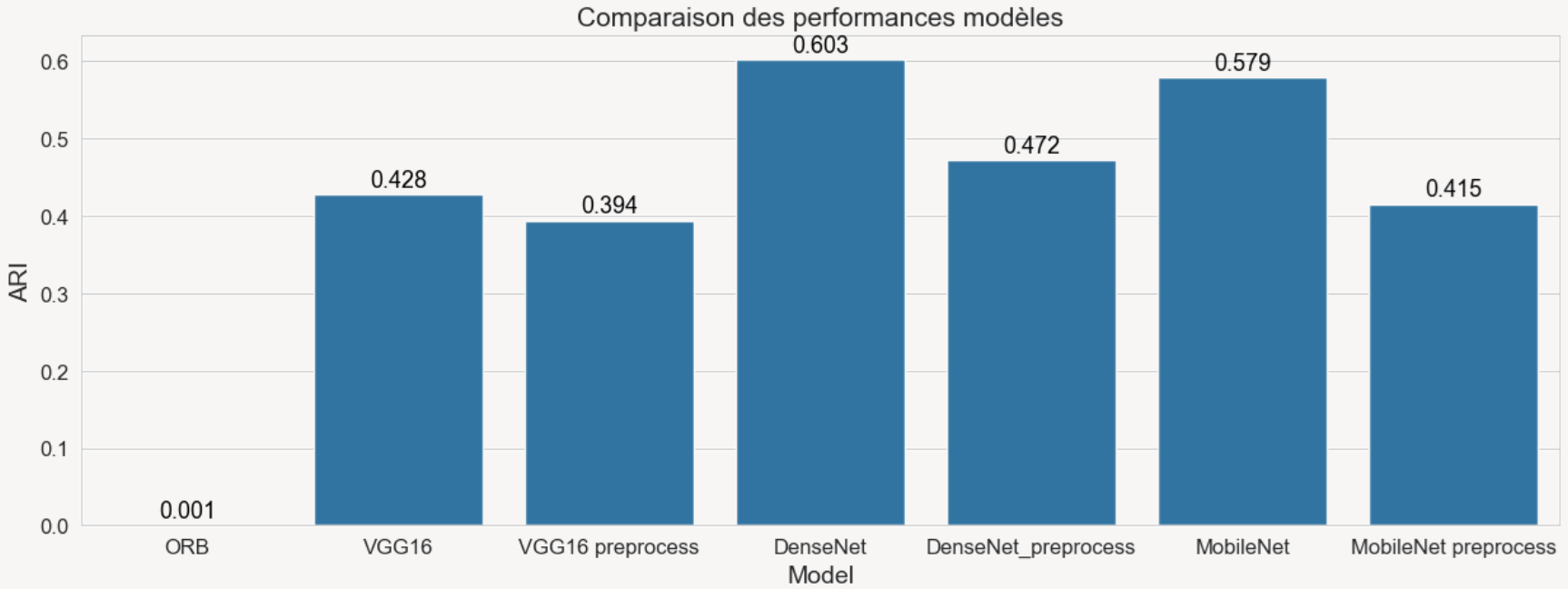
Equalized contrast



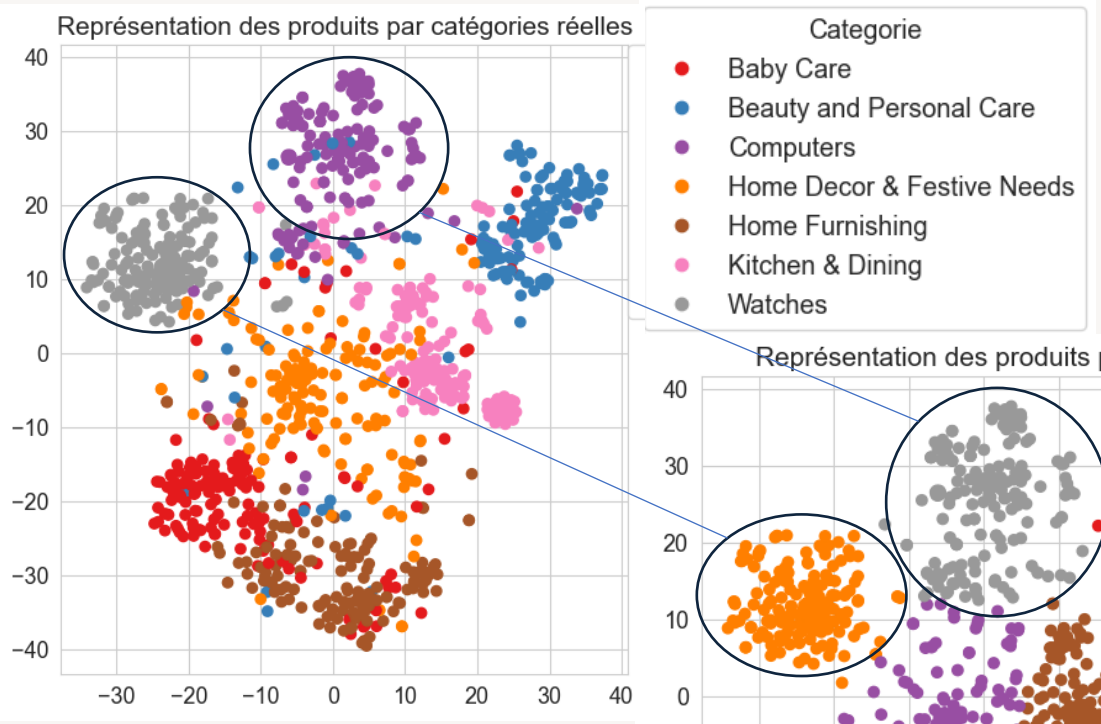


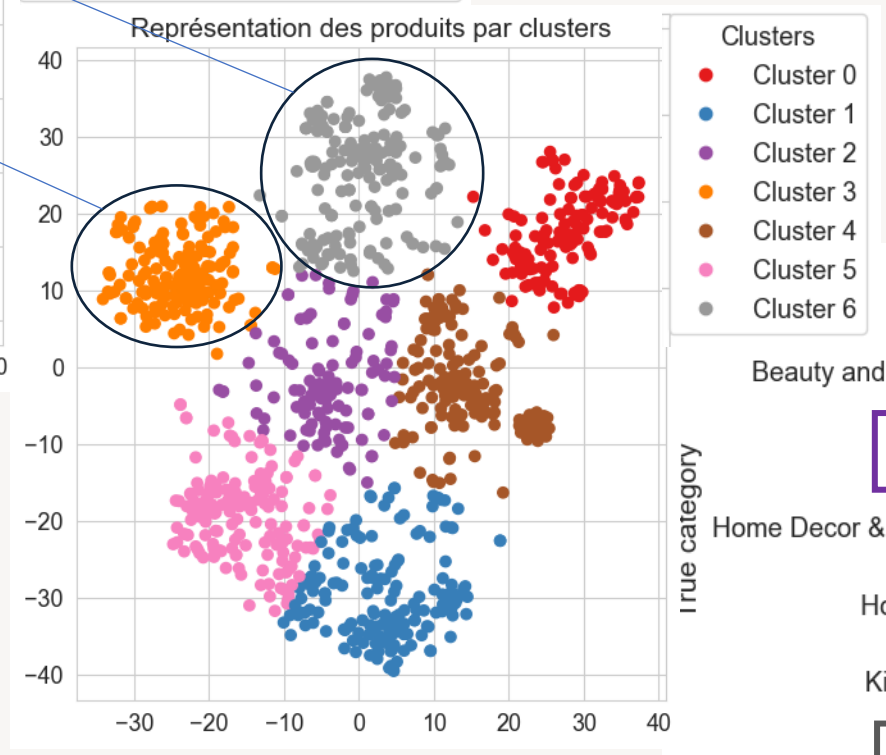


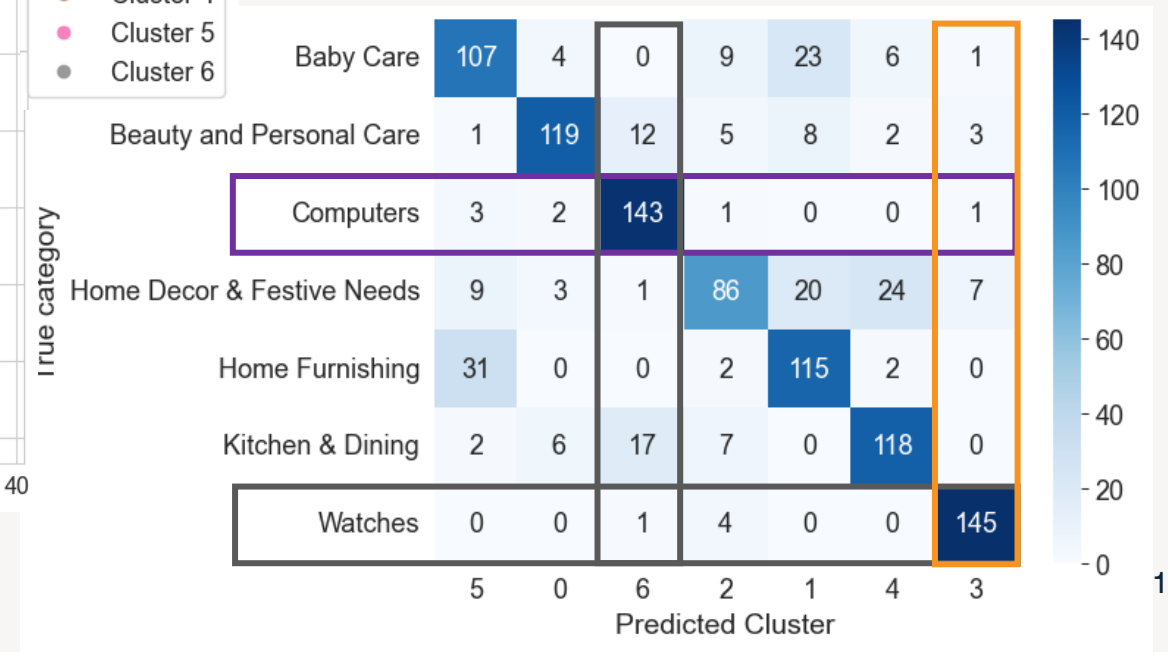
Features Extractions : modèles testés



DenseNet – ARI 0,6







Zoom sur la catégorie Watches

True Positive

False Positive

False Negative

Baby care, Home Decor & Festive Needs, Home Furnishing : des catégories difficiles à détecter

	Model	ARI	silhouette_score	Precision / recall - Baby Care	Precision / recall - Beauty and Personal Care	Precision / recall - Computers	Precision / recall - Home Decor & Festive Needs	Precision / recall - Home Furnishing	Precision / recall - Kitchen & Dining	Precision / recall - Watches
0	ORB	0.0012	0.40	0.15 / 0.19	0.20 / 0.25	0.24 / 0.10	0.17 / 0.23	0.19 / 0.15	0.19 / 0.17	0.19 / 0.19
1	VGG16	0.4284	0.44	0.61 / 0.71	0.95 / 0.73	0.53 / 0.64	0.47 / 0.62	0.53 / 0.52	0.93 / 0.54	0.97 / 0.93
2	VGG16 preprocess	0.3942	0.47	0.54 / 0.62	0.90 / 0.67	0.50 / 0.72	0.55 / 0.41	0.52 / 0.48	0.81 / 0.75	0.83 / 0.87
3	DenseNet	0.6026	0.50	0.70 / 0.71	0.89 / 0.79	0.82 / 0.95	0.75 / 0.57	0.69 / 0.77	0.78 / 0.79	0.92 / 0.97
4	DenseNet_preprocess	0.4721	0.48	0.60 / 0.73	0.84 / 0.82	0.66 / 0.63	0.63 / 0.51	0.70 / 0.57	0.65 / 0.72	0.85 / 0.97
5	MobileNet	0.5790	0.49	0.69 / 0.56	0.76 / 0.81	0.77 / 0.90	0.77 / 0.74	0.61 / 0.73	0.96 / 0.71	0.94 / 0.99
6	MobileNet preprocess	0.4154	0.47	0.52 / 0.63	0.85 / 0.75	0.49 / 0.55	0.39 / 0.32	0.54 / 0.57	0.90 / 0.75	0.82 / 0.91

Précision : Proportion de vrais positifs (TP) parmi les prédictions positives (TP+FP)
Recall : Proportion des vrais positifs (TP) parmi tous les éléments réellement positifs (TP+FN)

Classification supervisée

Méthodologie

Create Model → **Split dataset** → **Pre-processing & Data augmentation** → **Train Model** → **Evaluate Model**

Models
(VGG16, DenseNet)

Main dataset

Stratification : *Main category*

Train dataset (64%)

Validation dataset (16%)

Test dataset (20%)

Data augmentation :
Yes / No

Metrics : *Train / val accuracy, Fit time*
Loss : Categorical Crossentropy

Metrics :
Test accuracy, Prediction / Recall per category

Data augmentation

Transformations des images

- Rotation
- Flip
- Zoom
- Crop
- Brightness
- Contrast
- Saturation
- Noise addition

Utilisée *Non utilisée*

Collecte de nouvelles images

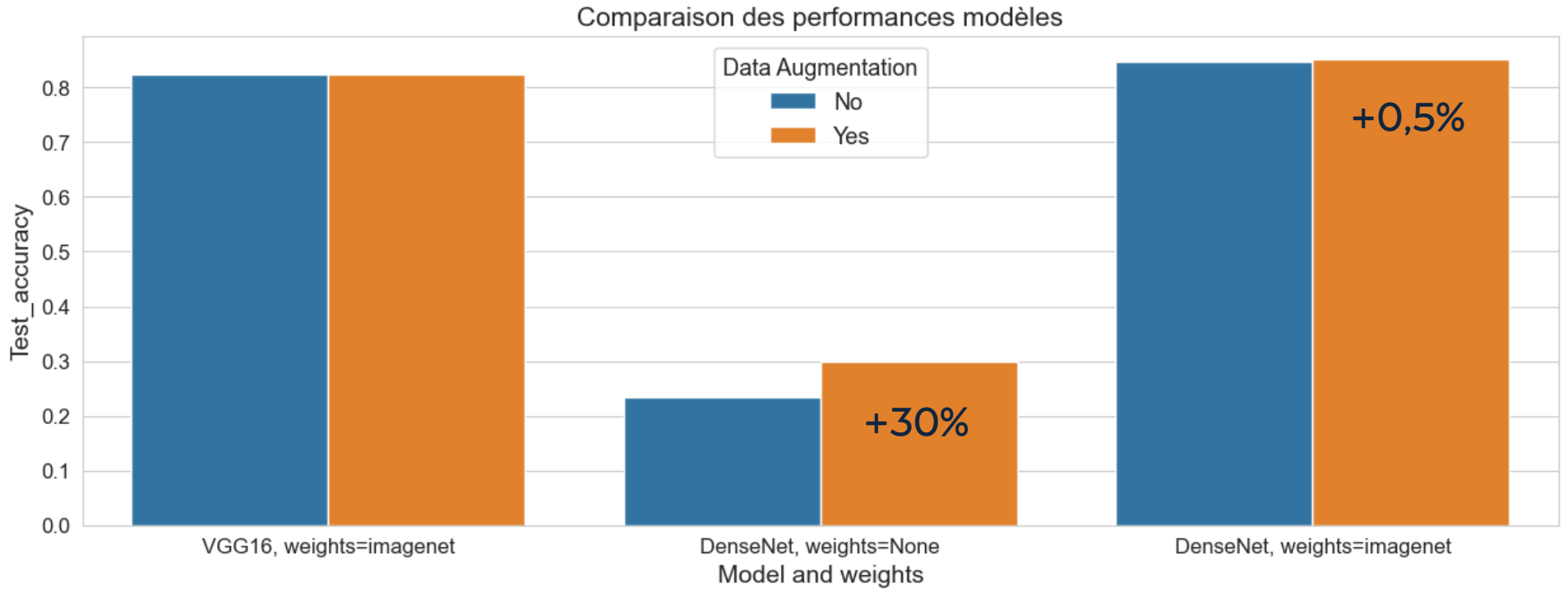
- Base de données d'images, API
  - Exemples : Pexels, Shutterstock

Propriété intellectuelle

- Photos des shooting non retenues du photographe

Accuracy améliorée avec la Data Augmentation

	Model and weights	Data Augmentation	Weights_nb	Epoch	Train_accuracy	Validation_accuracy	Fit Time (s)	Test_accuracy
0	VGG16, weights=imagenet	No	14980935	10	0.972	0.821	646.45	0.824
1	VGG16, weights=imagenet	Yes	14980935	10	0.930	0.786	430.53	0.824
2	DenseNet, weights=None	No	19309127	10	0.196	0.268	566.59	0.233
3	DenseNet, weights=None	Yes	19309127	10	0.205	0.268	576.52	0.300
4	DenseNet, weights=imagenet	No	19309127	9	0.982	0.869	517.71	0.848
5	DenseNet, weights=imagenet	Yes	19309127	10	0.952	0.839	579.43	0.852



Meilleures performances avec le transfer learning

	Model	Precision / recall - Baby Care	Precision / recall - Beauty and Personal Care	Precision / recall - Computers	Precision / recall - Home Decor & Festive Needs	Precision / recall - Home Furnishing	Precision / recall - Kitchen & Dining	Precision / recall - Watches
0	DenseNet, weights=imagenet	0.70 / 0.71	0.89 / 0.79	0.82 / 0.95	0.75 / 0.57	0.69 / 0.77	0.78 / 0.79	0.92 / 0.97

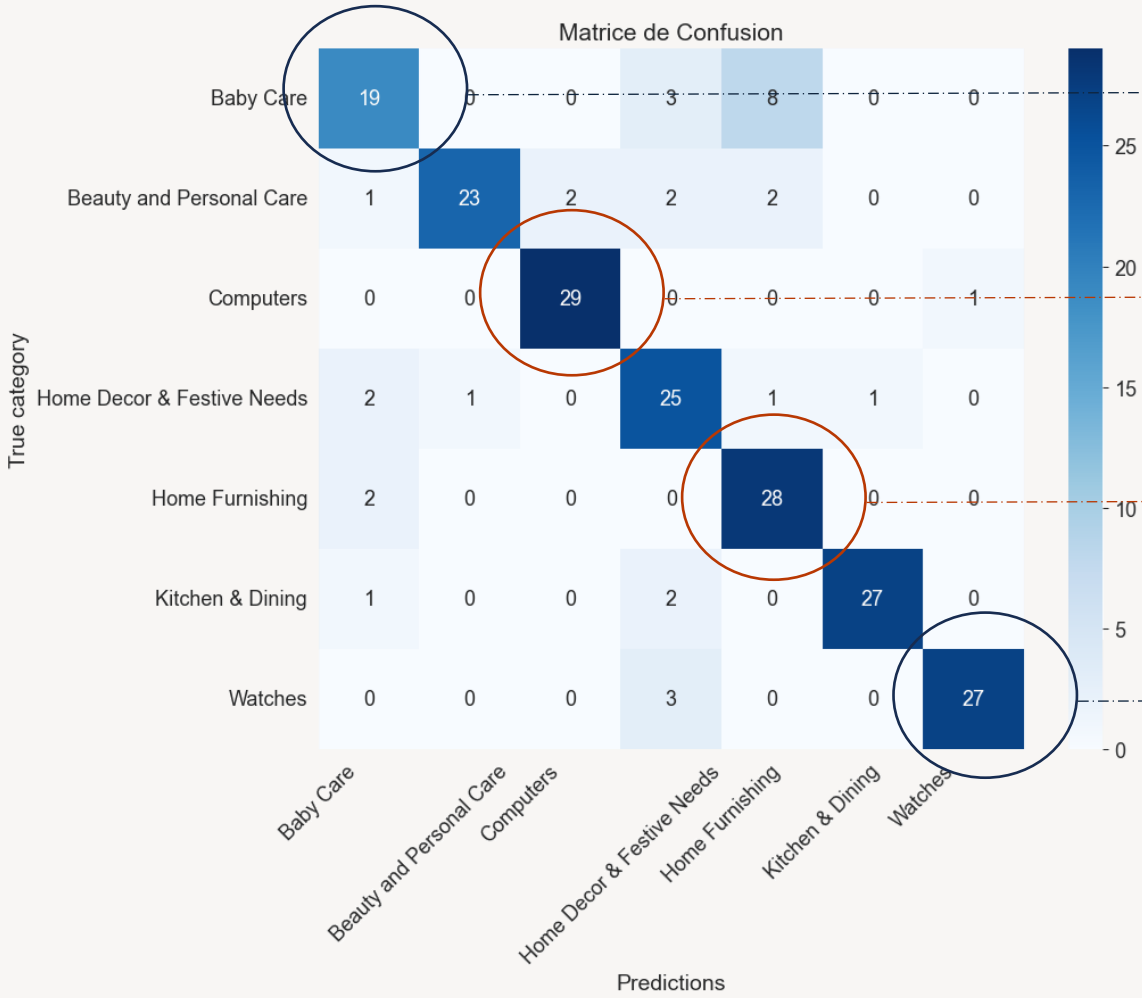
Transfer learning

	Model	Precision / recall - Baby Care	Precision / recall - Beauty and Personal Care	Precision / recall - Computers	Precision / recall - Home Decor & Festive Needs	Precision / recall - Home Furnishing	Precision / recall - Kitchen & Dining	Precision / recall - Watches
4	DenseNet, weights=imagenet	0.76 / 0.63	0.96 / 0.77	0.94 / 0.97	0.71 / 0.83	0.72 / 0.93	0.96 / 0.90	0.96 / 0.90
5	DenseNet, weights=imagenet, with data augment	0.75 / 0.80	0.96 / 0.73	0.90 / 0.93	0.82 / 0.77	0.80 / 0.93	0.85 / 0.93	0.96 / 0.90

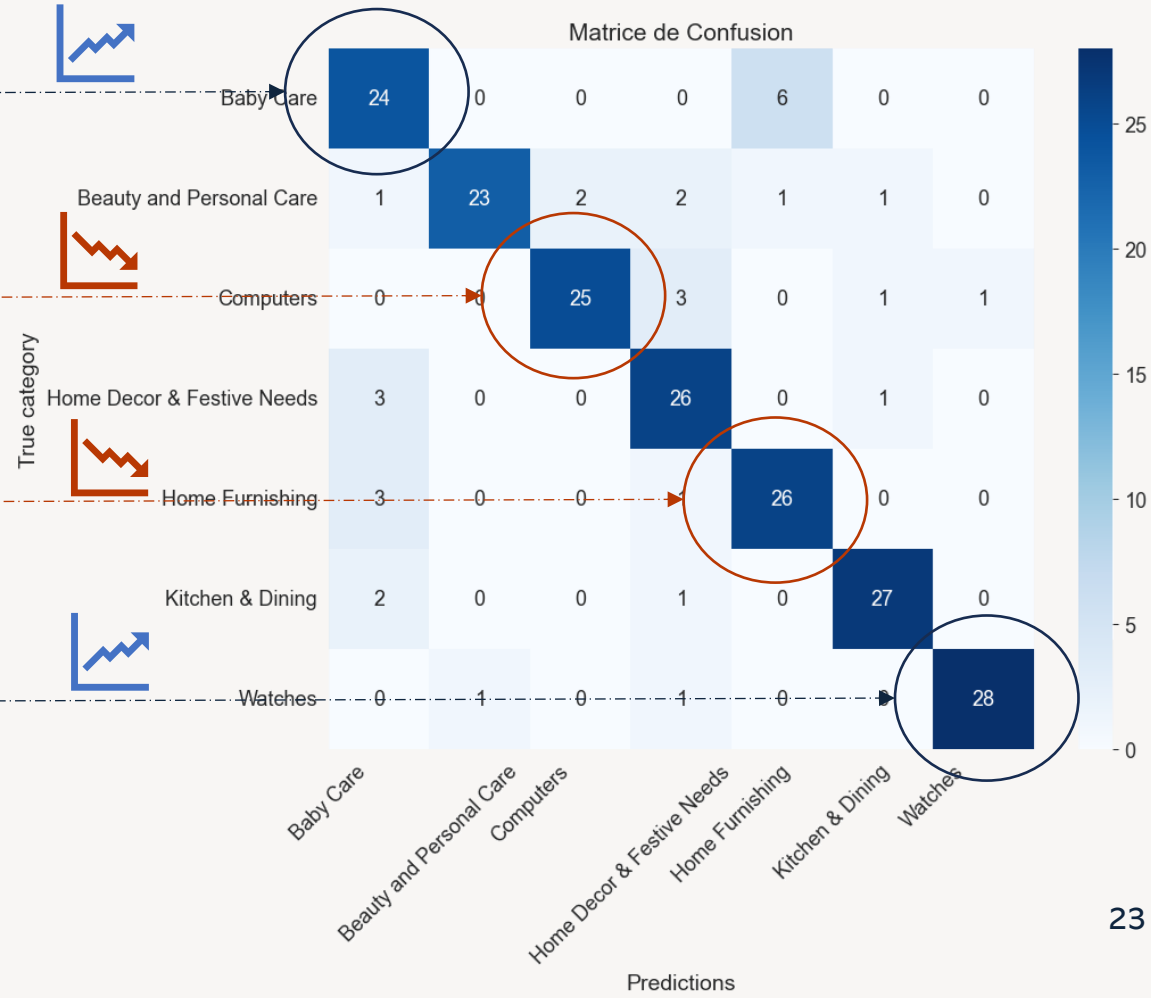
Baby Care, Home Decor & Festive Needs, Home Furnishing : toujours difficiles à classifier

Impact de la Data augmentation par catégorie

Sans data augmentation



Avec data augmentation



Test collecte de données via API

Nouvelle gamme de produits : Epicerie Fine

Objectif : Tester la collecte de données (texte + image) pour enrichir la base de données Epicerie fine

Données à extraires : foodId, label, category, foodContentsLabel, image

Filtre : {"ingr": "champagne"}

Requête sur l'API : https://edamam-food-and-grocery-database.p.rapidapi.com/api/food-database/v2/parser

RGPD et propriété intellectuelle

Produits à base de « champagne »

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	N/A	https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	N/A
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...	N/A
4	food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	N/A
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;...	N/A
7	food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	N/A
8	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	N/A
9	food_a79xmnya6togreaeukbroa0thhh0	Champagne Chicken	Generic meals	Flour; Salt; Pepper; Boneless, Skinless Chicke...	N/A

Peu d'images disponibles, images de mauvaises qualité, peu représentatives du produit « Champagne »

Conclusion

Performances

- Meilleurs modèles : **Word2Vec** (textes)
  DenseNet (images)
- **Certaines catégories difficiles à classifier**

Données

- Produits similaires dans plusieurs catégories
- Produits en situation ou non, portés ou non
- Erreur de catégorisation

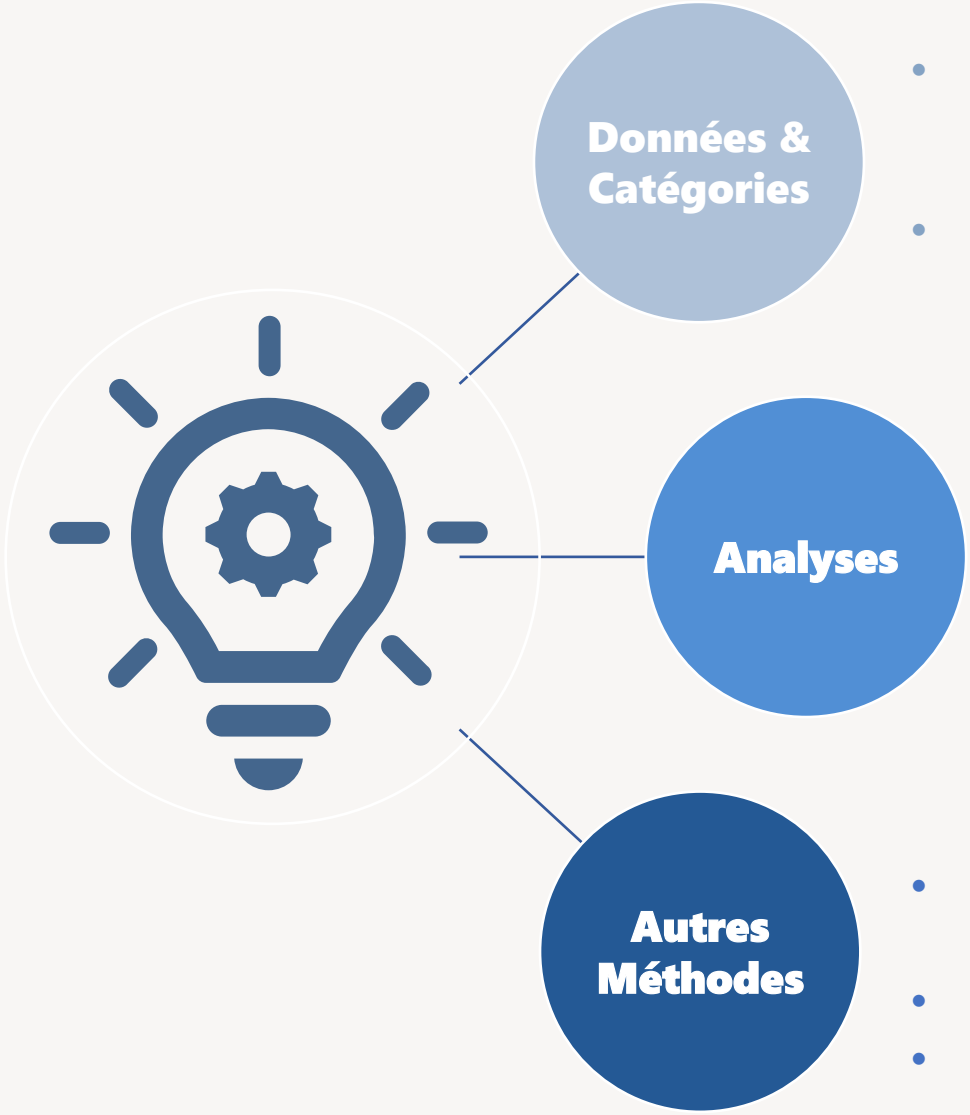
Data Augmentation

- **Améliore l'accuracy globale** du modèle
- **Peut diminuer la performance locale** par catégorie

Test API

- **Bonne solution pour la Data Augmentation**
- **Vigilance :** Pertinence des données, qualité des données, propriété intellectuelle

Perspectives



- **Données :** Tester d'autres pre-processing des images, Data Augmentation spécifique à chaque catégorie, utiliser des modèles entraînés des jeux autres que ImageNet
- **Revoir les catégories /sous-catégories** pour celles qui ne fonctionnent pas (ex. Vêtement portés ou non)

- **Class Activation Maps :** Visualiser où le modèle concentre son attention
- **Analyse des filtres des couches convolutionnelles :** Comprendre comment le modèle apprend

- **One Shot / Few shot Learning** (1 à n produit(s) par sous-catégorie)
  Ensemble Learning (systèmes de votes)
- **Multimodal Learning** (combiner Texte + Image)
- **Curriculum Learning** (apprentissage progressif,+ simple -> + complexe)

Back-up slides

5 grands principes du Règlement Général sur la Protection des Données (RGPD)

FINALITÉ

Informations enregistrées et utilisées dans un but bien précis, légal et légitime

PROPORTIONNALITÉ ET PERTINENCE

Informations enregistrées pertinentes et strictement nécessaires

DURÉE DE CONSERVATION LIMITÉE

Durée de conservation précise fixée en fonction du type d'informations et de la finalité du fichier

SÉCURITÉ ET CONFIDENTIALITÉ

Restrictions d'accès aux informations enregistrées aux seules personnes autorisées

DROITS DES PERSONNES

Droit à l'information, recueil du consentement, droit d'opposition, droits d'accès et rectification, droit à la portabilité

Propriété intellectuelle : Les textes et images de la base de données ne relèvent pas d'une propriété intellectuelle.