3 Janvier 2022

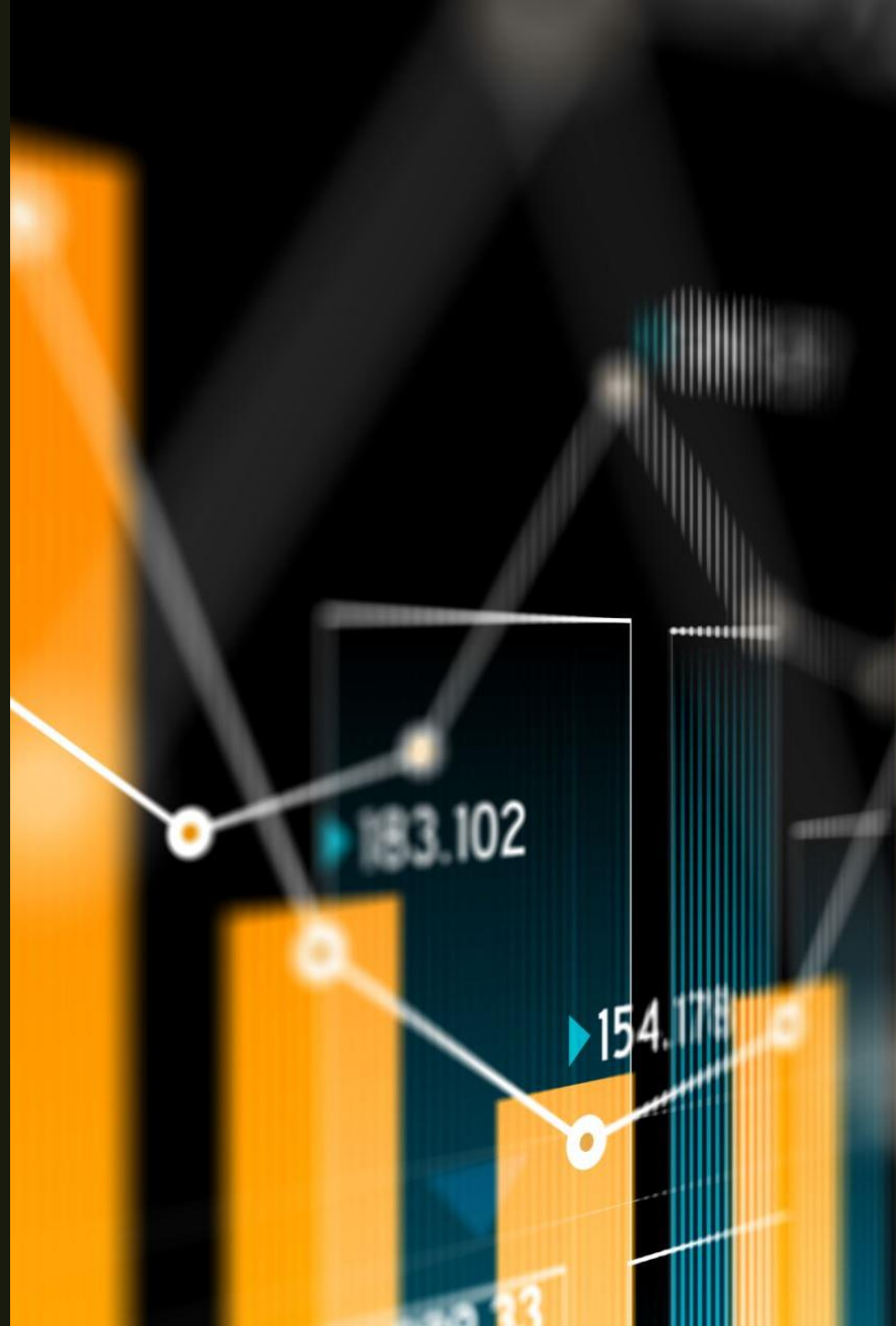
CLASSIFIEZ AUTOMATIQUEMENT DES BIENS DE CONSOMMATION

Soutenance Projet n°6

Présenter par Louisy-Louis Matthieu

OPENCLASSROOMS

SOMMAIRE



I. PRÉSENTATION DU PROJET ET DE LA PROBLÉMATIQUE

Notre entreprise **Place de marché** souhaite lancer une marketplace e-commerce.

Problématique

On veut automatiser la catégorisation d'un article pour rendre l'expérience utilisateur plus fluide.

Objectif

Etudier la faisabilité d'un moteur de classification des articles en différentes catégories.

Outils utilisés

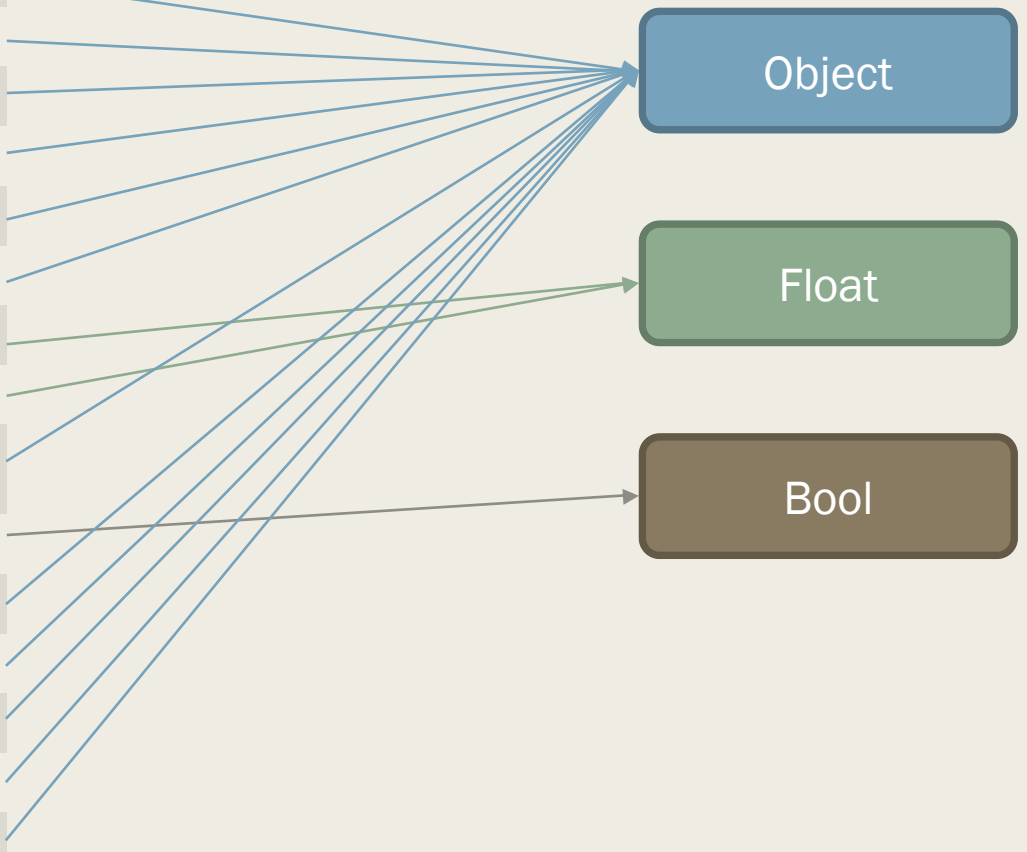
- Anaconda
- JupiterLab
- Python
  - *Pandas*
  - *ScikitLearn*
  - *NLTK*
  - *CV2*
  - *PIL*
  - *Keras*

II. PRÉSENTATION DES DONNÉES

II. Présentation des données

Dataset : 1050 entrées, 15 colonnes

Uniq_id	Identifiant unique
Crawl_timestamp	Date de création
Product_url	Adresse url du produit
Product_name	Nom du produit
Product_category_tree	Arbre de catégorisation
Pid	Code (?)
Retail_price	Prix
Discounted_price	Prix après remise
Image	Nom de l'image associée
Is_FK_Advantage_product	(?)
Description	Description du produit
Product_rating	Note du produit
Overall_rating	Note générale
Brand	Marque du produit
Product_specifications	Spécifications du produit



Peu de données manquantes

```
data.isnull().sum()
```

```
uniq_id                    0
crawl_timestamp            0
product_url                0
product_name               0
product_category_tree      0
pid                        0
retail_price               1
discounted_price           1
image                      0
is_FK_Advantage_product    0
description                0
product_rating             0
overall_rating             0
brand                    338
product_specifications     1
dtype: int64
```

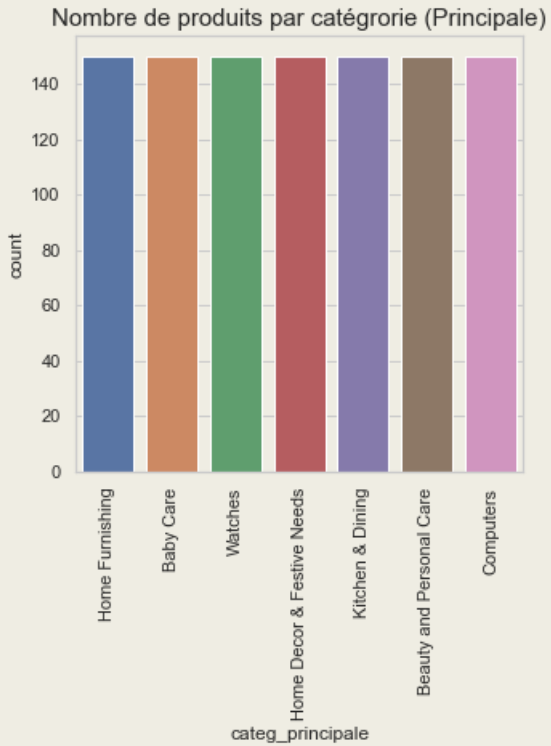
7 Catégories principales

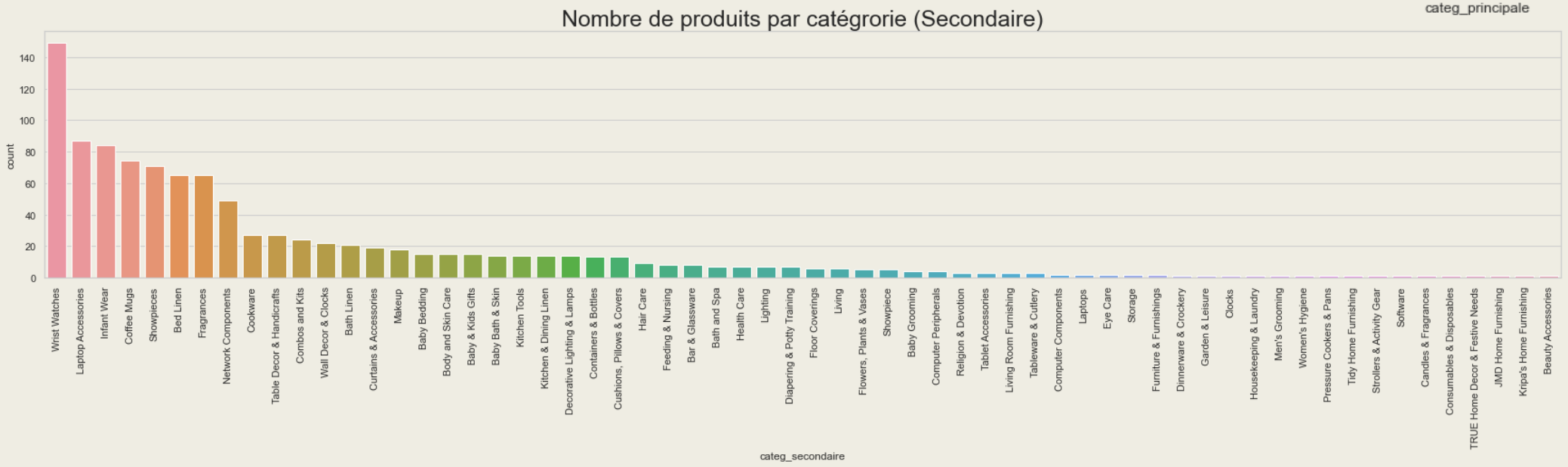
- Home Furnishing
- Baby Care
- Watches
- Home Decor & Festive Needs
- Kitchen & Dining
- Beauty & Personal Care
- Computers

62 Catégories secondaires

Répartition des produits par catégories

150 Produits par catégorie principale





III. TRAITEMENT DES DONNÉES TEXTUELLES

Pré-traitement

Exemple sur les 10 premiers mots d'une description d'un produit

```
data["description"].iloc[215]
```

```
'King International Ergonomic Design with Long Gripped Handle Rolling Pizza
```

Tokenizer	King	International	Ergonomic	Design	with	Long	Gripped	Handle	Rolling	Pizza
Stemmer	king	intern	ergonom	design	with	long	grip	handl	roll	pizza
Lemmatizer	king	international	ergonomic	design	with	long	grip	handle	roll	pizza

```
#Exemple avec une phrase simple
print([lemmatizer.lemmatize(w, get_wordnet_pos(w)) for w in tokenizer.tokenize("Rico's child is playing soccer with his feet ;) ")])
```

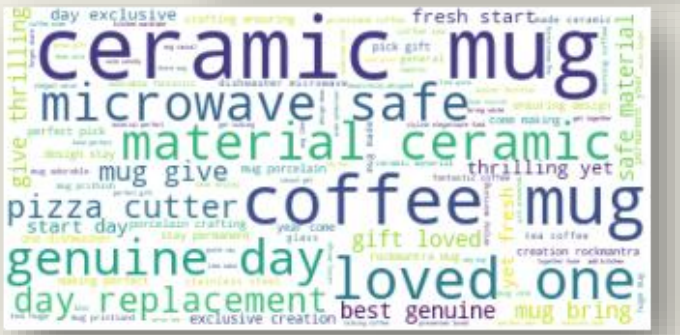
```
['Rico', 's', 'child', 'be', 'play', 'soccer', 'with', 'his', 'foot']
```

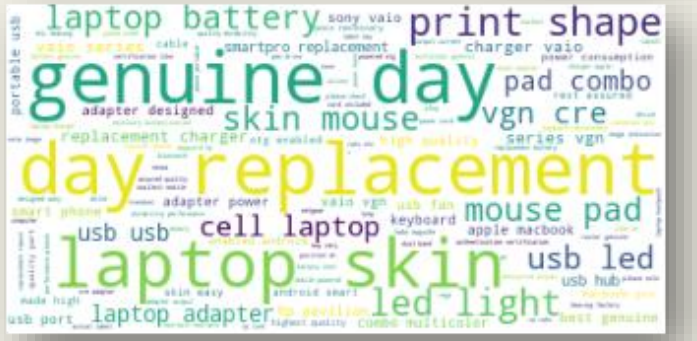
Word Cloud


fabric cotton
baby boy
baby girl
girl gathered
general pattern
fabric care
ideal baby
dress
mini klub
cotton
round neck
box
half sleeve
occasion casual


genuine day
day replacement
best genuine
fruit massage
box
combo
mxofere combo
massage cream
hair
cream vitamin
combo online
soap kit
online genuine
almond honey
lowest genuine
jewellery vanity


cushion cover
printed
bath-towel
quilt
comforter
box
aroma
comfort
cotton
brown
polyester
genuine
towel
home
soft
door curtain
single quilt
double blanket

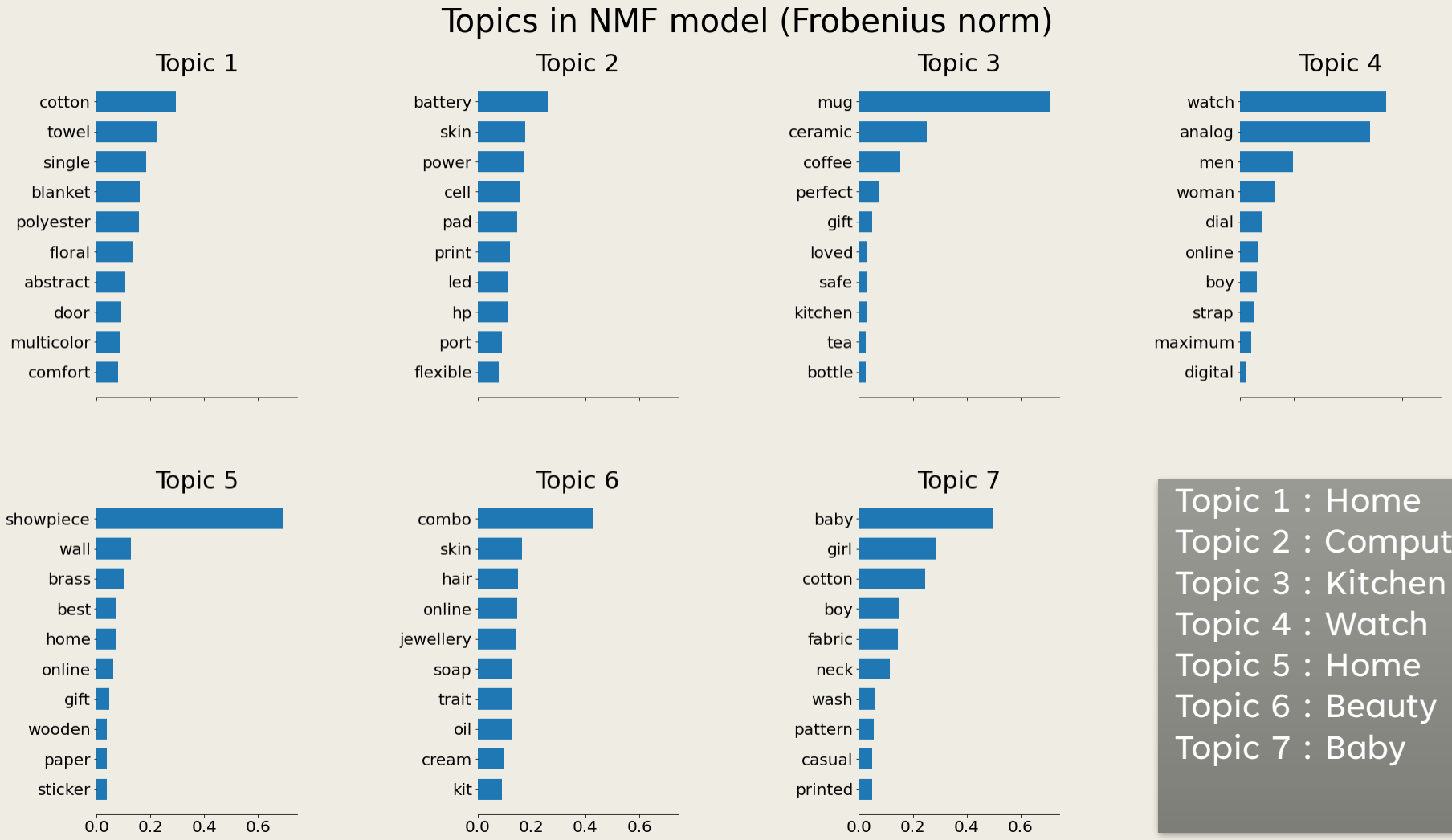

ceramic mug
microwave safe
material ceramic
mug give
coffee mug
pizza cutter
genuine day
loved one
day replacement
best genuine
fresh start


laptop battery
print shape
genuine day
pad combo
skin mouse
vgn cre
day replacement
mouse pad
usb usb
cell laptop
laptop skin
usb led
led light
laptop adapter


day replacement
showpiece
showpiece online
showpiece best
best genuine
rice light
genuine day
decorative
beautiful
lord ganesha
lal haveli
design


discount genuine
watch men
watch woman
online india
analog watch
great discount
day replacement
genuine day

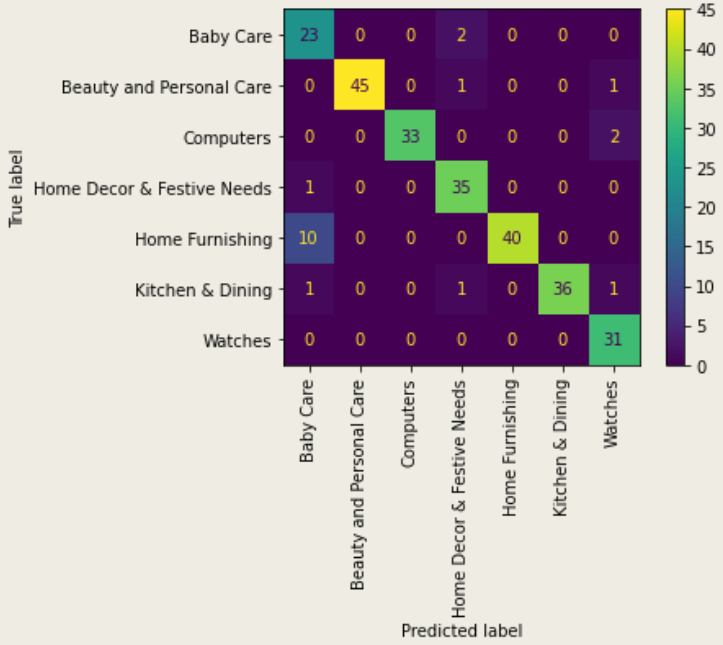
Algorithme non-supervisé



Topic 1 : Home
Topic 2 : Computers
Topic 3 : Kitchen
Topic 4 : Watch
Topic 5 : Home
Topic 6 : Beauty
Topic 7 : Baby

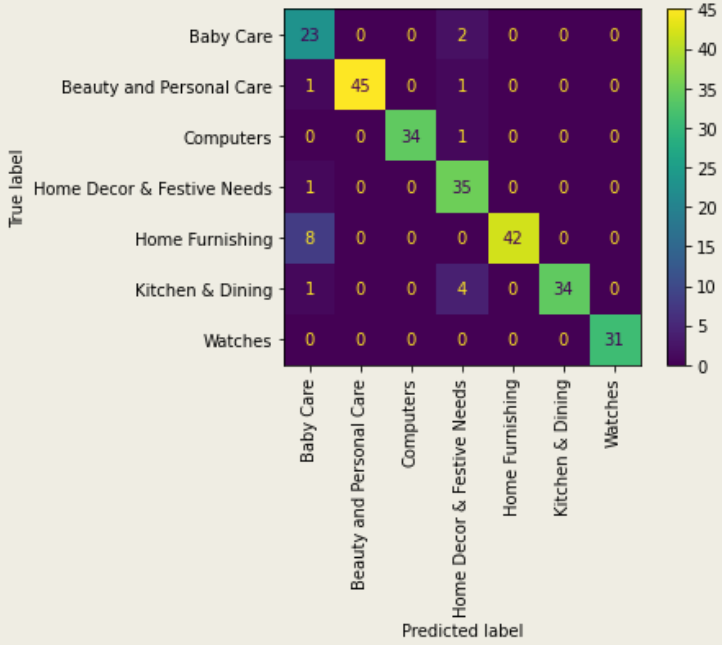
Algorithme supervisé

MNB



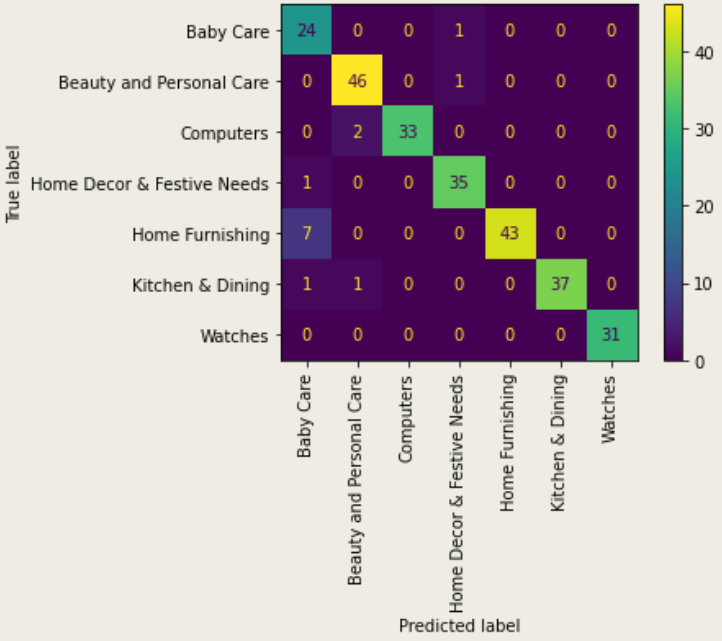
Train accuracy : 0,983
Test accuracy : 0,924

SVC



Train accuracy : 0,999
Test accuracy : 0,923

MLP



Train accuracy : 1,0
Test accuracy : 0,947

IV. TRAITEMENT DES DONNÉES VISUELLES

Aperçu des images

Echantillon en fonction de chaque catégorie

Computers

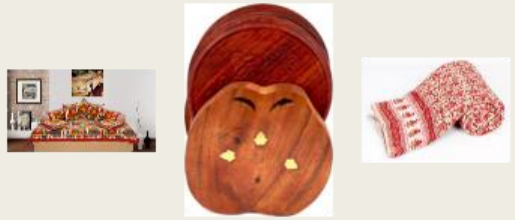
Beauty & Personnal Care

Home Deco & Festive Needs

Baby

Watches

Kitchen & Dining

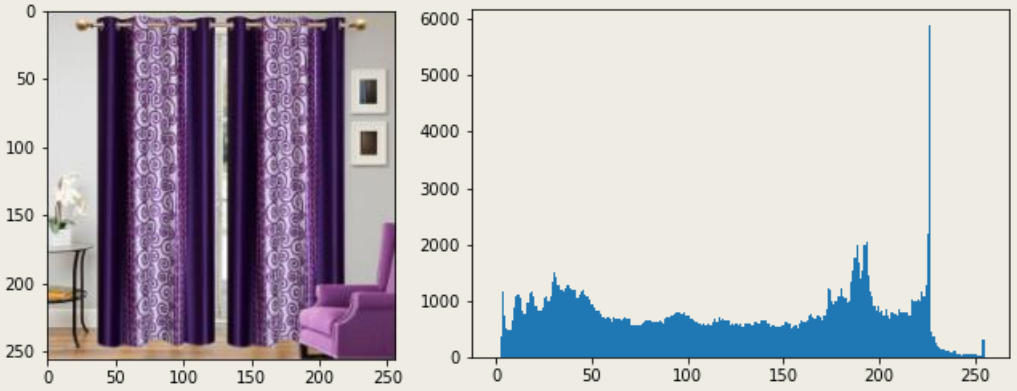


Home Furnishing

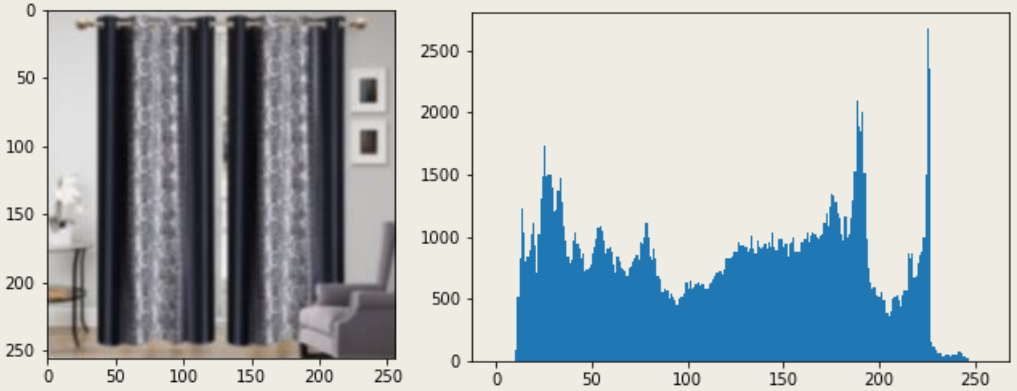
Pré-traitement de l'image

Exemple sur une image du dataset

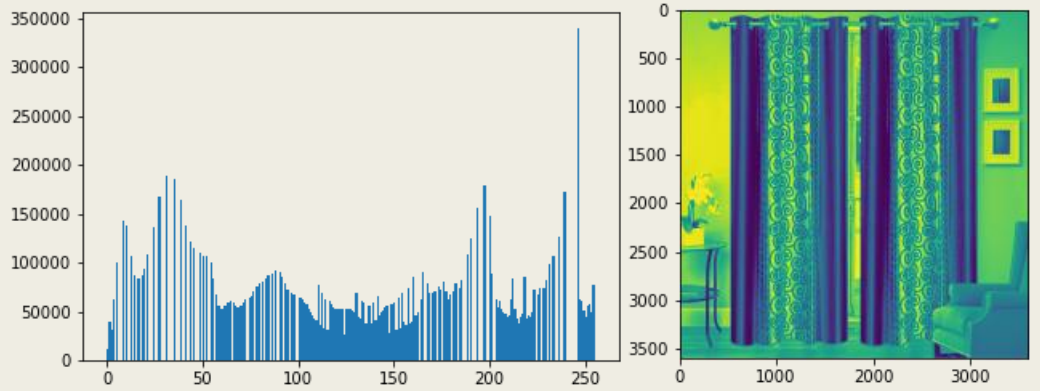
Image originale (rétréci)



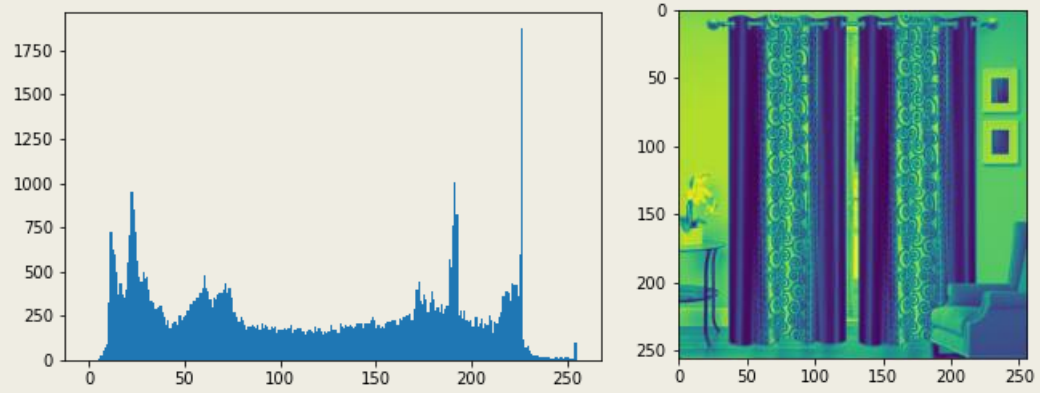
Gaussian Blur



Equalize Histogram



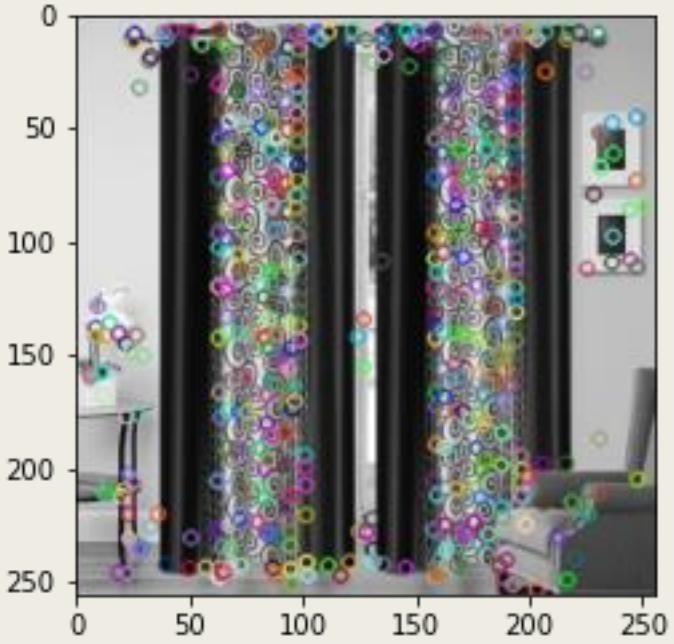
Gray



Exraction des features

Pré-traitement de l'image:

- Passage en gris
- Redimensionnement
- Ajout de flou
- Réglage du contraste



Observation des keypoints detectés par SIFT

Création de clusters des descripteurs

Création des histogrammes

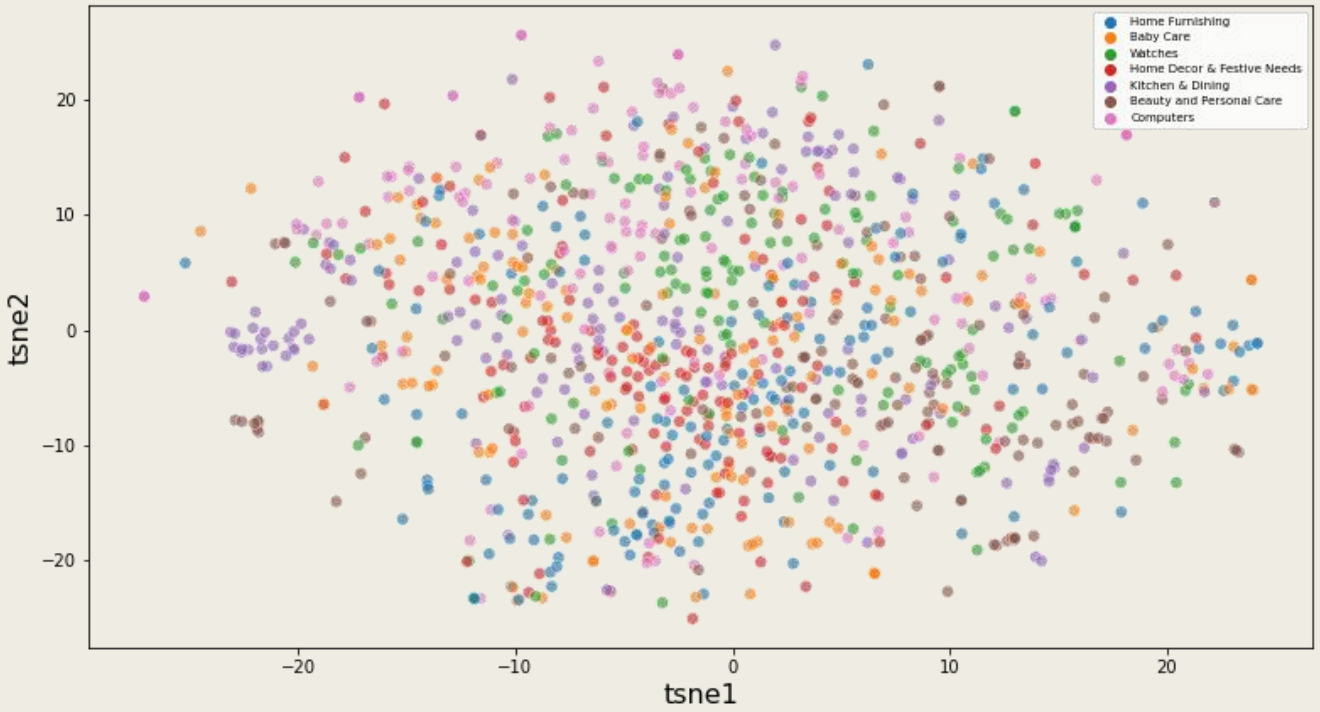
Réduction de dimensions

Application PCA sur image features (1050,608)

- Maintien d'un niveau de variance expliquée élevé (99%)
- Création de features décorrelées entre elles
- Diminution de la dimension

```
Dimensions dataset avant réduction PCA :  (1050, 608)
Dimensions dataset après réduction PCA :  (1050, 479)
```

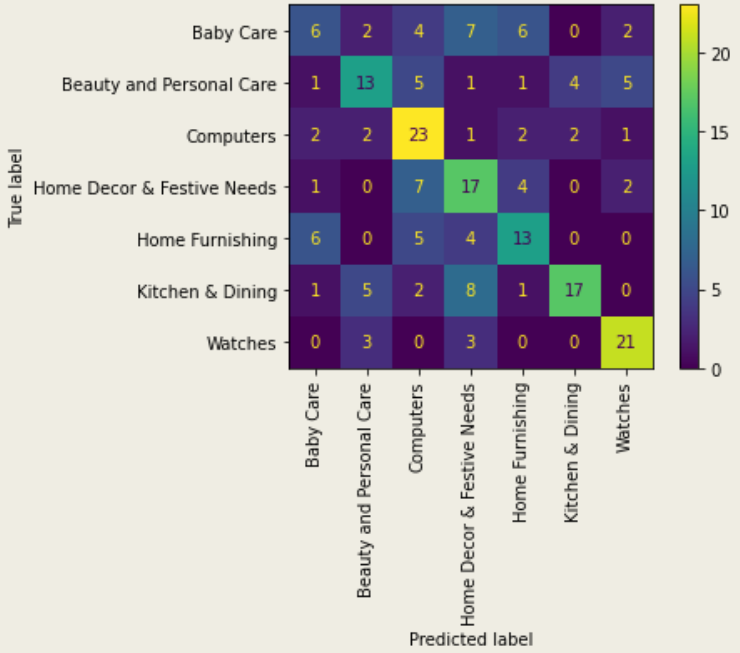
TSNE selon les vraies classes



Algorithme supervisé sur les features d'image

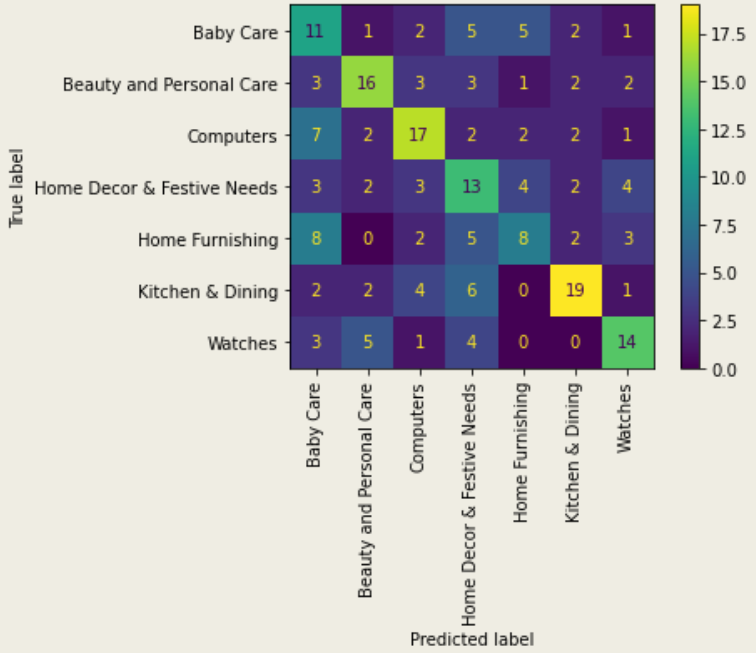
Division train/test

SVC



	precision	recall	f1-score	support
Baby Care	0.35	0.22	0.27	27
Beauty and Personal Care	0.52	0.43	0.47	30
Computers	0.50	0.70	0.58	33
Home Decor & Festive Needs	0.41	0.55	0.47	31
Home Furnishing	0.48	0.46	0.47	28
Kitchen & Dining	0.74	0.50	0.60	34
Watches	0.68	0.78	0.72	27
accuracy			0.52	210
macro avg	0.53	0.52	0.51	210
weighted avg	0.53	0.52	0.52	210

MLP



	precision	recall	f1-score	support
Baby Care	0.30	0.41	0.34	27
Beauty and Personal Care	0.57	0.53	0.55	30
Computers	0.53	0.52	0.52	33
Home Decor & Festive Needs	0.34	0.42	0.38	31
Home Furnishing	0.40	0.29	0.33	28
Kitchen & Dining	0.66	0.56	0.60	34
Watches	0.54	0.52	0.53	27
accuracy			0.47	210
macro avg	0.48	0.46	0.47	210
weighted avg	0.48	0.47	0.47	210

Construction d'un réseau de neurone convolutif

Input : 224x224x3

Application d'algorithme supervisé

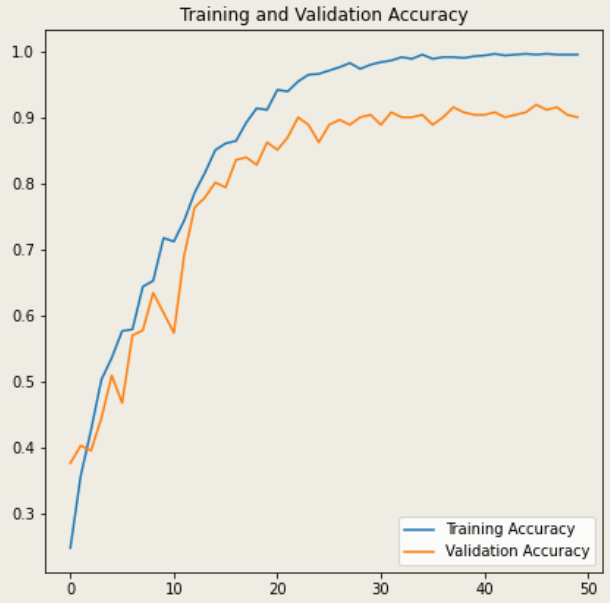
```
opt = tf.keras.optimizers.Adam(learning_rate=0.0001)
model.compile(optimizer = opt ,loss = tf.keras.losses.SparseCategoricalCrossentropy() , metrics = ['accuracy'])
```

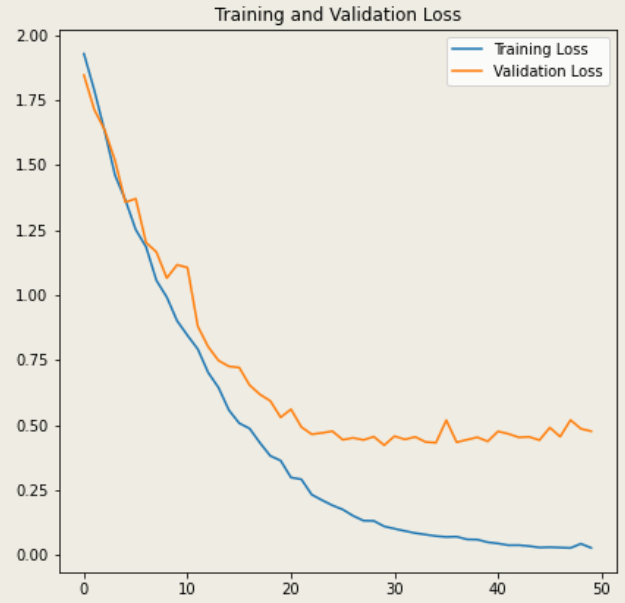
```
history = model.fit(X_train,y_train, epochs = 50 , validation_data = (X_test, y_test))
```

```
test_loss, test_acc = model.evaluate(X_test, y_test)
test_acc
```

```
9/9 [==============================] - 2s 204ms/step - loss: 0.4762 - accuracy: 0.9011
0.9011406898498535
```

ACCURACY = 90,1%





Convolution 2D

224x224x32

Maxpooling

112x112x32

Convolution 2D

112x112x32

Maxpooling

56x56x32

Convolution 2D

56x56x64

Maxpooling

28x28x64

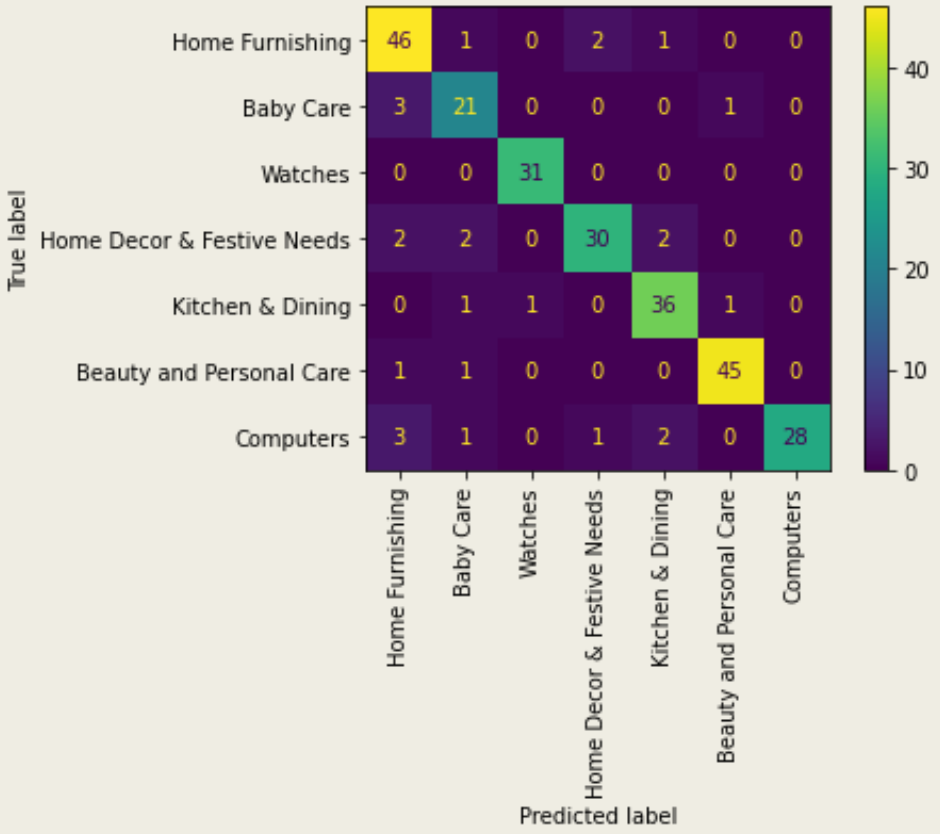
Flatten

50176

Dense

7

Details des résultats du CNN



	precision	recall	f1-score	support
Home Furnishing	0.84	0.92	0.88	50
Baby Care	0.78	0.84	0.81	25
Watches	0.97	1.00	0.98	31
Home Decor & Festive Needs	0.91	0.83	0.87	36
Kitchen & Dining	0.88	0.92	0.90	39
Beauty and Personal Care	0.96	0.96	0.96	47
Computers	1.00	0.80	0.89	35
accuracy			0.90	263
macro avg	0.90	0.90	0.90	263
weighted avg	0.91	0.90	0.90	263

Essai sur des images inédites

Traitement image → Prédiction du modèle → Catégorie : %

Catégorie : Kitchen & Dining
% de confiance : 64,75

Catégorie : Home Decor & Festive Need
% de confiance : 98,16

Catégorie : Watch
% de confiance : 99,51

Catégorie : Baby Care
% de confiance : 55,74

Catégorie : Computers
% de confiance : 81,81

V. BILAN

Des résultats satisfaisants

Avec les algorithmes supervisés

- Accuracy de 95% sur les données textes
- Accuracy de 90% sur les données visuelles

	precision	recall	f1-score	support
Baby Care	0.97	0.78	0.86	45
Beauty and Personal Care	0.89	0.97	0.93	34
Computers	0.95	0.95	0.95	41
Home Decor & Festive Needs	0.91	0.98	0.94	41
Home Furnishing	0.79	0.96	0.87	28
Kitchen & Dining	1.00	0.92	0.96	37
Watches	1.00	1.00	1.00	37
accuracy			0.93	263
macro avg	0.93	0.94	0.93	263
weighted avg	0.94	0.93	0.93	263

Données texte

	precision	recall	f1-score	support
Home Furnishing	0.84	0.92	0.88	50
Baby Care	0.78	0.84	0.81	25
Watches	0.97	1.00	0.98	31
Home Decor & Festive Needs	0.91	0.83	0.87	36
Kitchen & Dining	0.88	0.92	0.90	39
Beauty and Personal Care	0.96	0.96	0.96	47
Computers	1.00	0.80	0.89	35
accuracy			0.90	263
macro avg	0.90	0.90	0.90	263
weighted avg	0.91	0.90	0.90	263

Données image

✓ Possibilité de classifier automatiquement les produits selon les catégories

Des pistes pour aller plus loin

- Taille du jeu de données
- Stopwords spécifique
- Transfer Learning

Etudier les données textes et images ensemble pour optimiser le machine learning

FIN DE LA SOUTENANCE

MERCI POUR VOTRE ATTENTION