Classifiez automatiquement des biens de consommation

Projet 6 - Julien Agneray

Classification automatique pour la marketplace

“Place de marché” est une marketplace anglophone où des vendeurs proposent des articles à des acheteurs en postant une photo et une description.

- Pour l'instant, l'attribution de la catégorie d'un article est effectuée manuellement par les vendeurs, et est donc peu fiable. De plus, le volume des articles est pour l’instant très petit.
- Pour rendre l’expérience utilisateur des vendeurs et des acheteurs la plus fluide possible, et dans l'optique d'un passage à l'échelle, il devient nécessaire d'automatiser cette tâche d‘attribution de la catégorie.

Objectif : Etudier la faisabilité d’un moteur de classification des articles en différentes catégories à partir du texte et de l’image.

Sommaire

1. Prétraitement des données textuelles
2. Méthodes de vectorisation et clustering
3. Classification basée sur les images
4. Conclusions générales
5. Requête API pour l’extraction de produits

Prétraitement des données textuelles

Nous disposons d'un jeu de données textuel contenant les descriptions d'articles, que nous commençons par nettoyer afin de le rendre exploitable pour l'analyse.

product_name	description
Elegance Polyester Multicolor Abstract Eyelet ...	Key Features of Elegance Polyester Multicolor ...
Sathiyas Cotton Bath Towel	Specifications of Sathiyas Cotton Bath Towel (...
Eurospa Cotton Terry Face Towel Set	Key Features of Eurospa Cotton Terry Face Towe...
SANTOSH ROYAL FASHION Cotton Printed King size...	Key Features of SANTOSH ROYAL FASHION Cotton P...
Jaipur Print Cotton Floral King sized Double B...	Key Features of Jaipur Print Cotton Floral Kin...

Prétraitement des données textuelles

Nettoyage des descriptions textuelles avec les étapes suivantes :

1. Suppression de la ponctuation
   - Exemple : "La qualité du produit !" ⟶ "La qualité du produit"
2. Mise en minuscules
   - Exemple : "LA QUALITÉ DU PRODUIT" ⟶ "La qualité du produit"
3. Suppression des mots superflus (stopwords)
   - Exemple : "la qualité du produit est excellente" ⟶ "qualité produit excellente"
4. Lemmatisation
   - Exemple : "les meilleures qualités des produits" ⟶ "meilleur qualité produit"
5. Stemming
   - Exemple : "qualité" et "qualifier" ⟶ "qualit"

Exemple final après nettoyage : « La qualité exceptionnelle du produit est garantie ! » ⟶ « qualit exceptionnel produit garanti »

Prétraitement des données textuelles

Exemple de transformation du texte brut suite au nettoyage : réduction des données à l'essentiel pour une meilleure analyse.

product_name	description	cleaned_product_name	cleaned_description
Elegance Polyester Multicolor Abstract Eyelet ...	Key Features of Elegance Polyester Multicolor ...	eleg polyest multicolor abstract eyelet door c...	key featur eleg polyest multicolor abstract ey...
Sathiyas Cotton Bath Towel	Specifications of Sathiyas Cotton Bath Towel (...	sathiya cotton bath towel	specif sathiya cotton bath towel 3 bath towel ...
Eurospa Cotton Terry Face Towel Set	Key Features of Eurospa Cotton Terry Face Towe...	eurospa cotton terri face towel set	key featur eurospa cotton terri face towel set...
SANTOSH ROYAL FASHION Cotton Printed King size...	Key Features of SANTOSH ROYAL FASHION Cotton P...	santosh royal fashion cotton print king size d...	key featur santosh royal fashion cotton print ...
Jaipur Print Cotton Floral King sized Double B...	Key Features of Jaipur Print Cotton Floral Kin...	jaipur print cotton floral king size doubl bed...	key featur jaipur print cotton floral king siz...

Méthodologie

- Exploration des méthodes de vectorisation

  CountVectorizer, TF-IDF, Word2Vec, BERT, USE – des techniques pour transformer les descriptions en vecteurs exploitables.

- Réduction de dimension

  PCA et t-SNE pour simplifier et visualiser les données.

- Clustering pour regrouper les produits

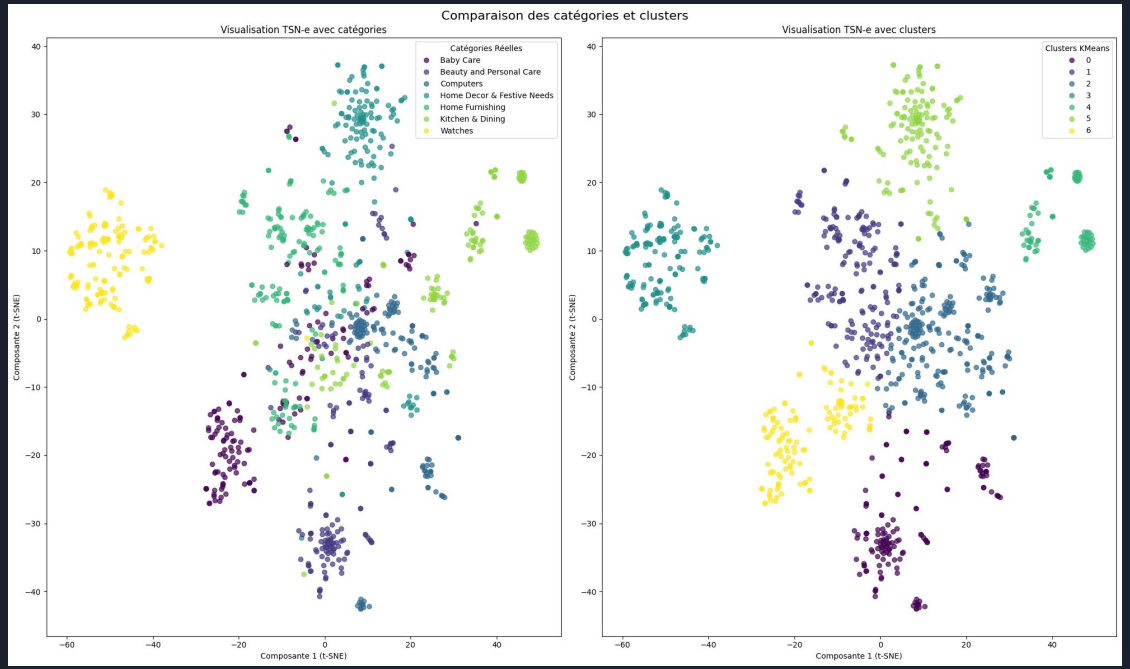
  KMeans pour créer des clusters et découvrir des groupes de produits similaires.

- Évaluation de la qualité des regroupements

  Indice ARI pour mesurer la cohérence avec les catégories réelles.

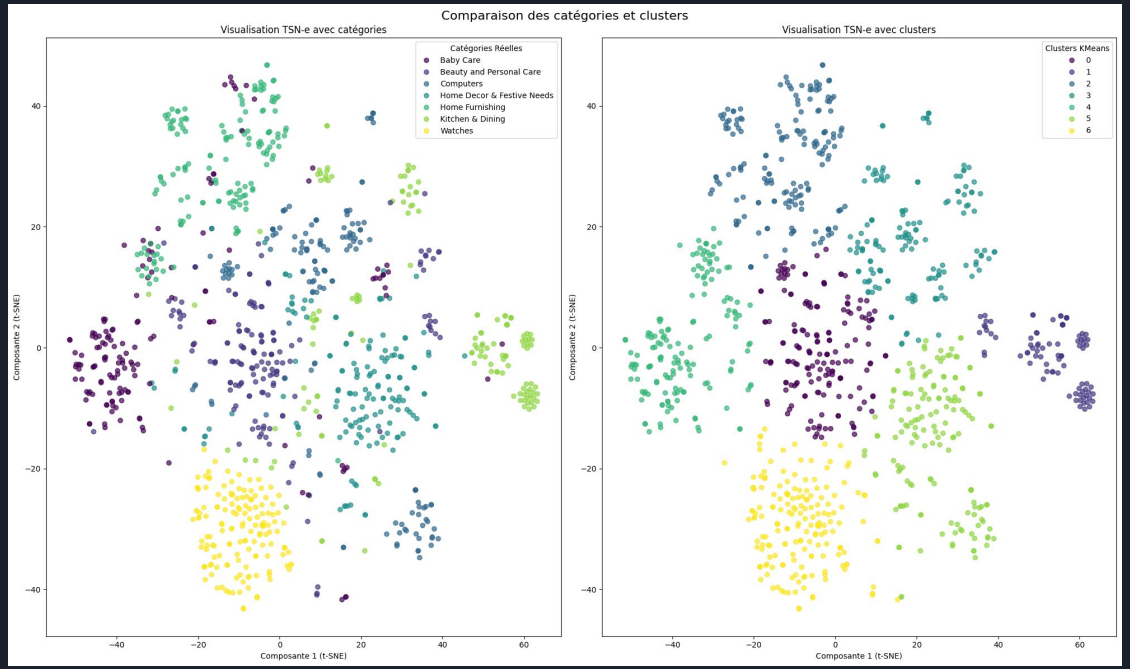
CountVectorizer

ARI (Adjusted Rand Index) : 0.46



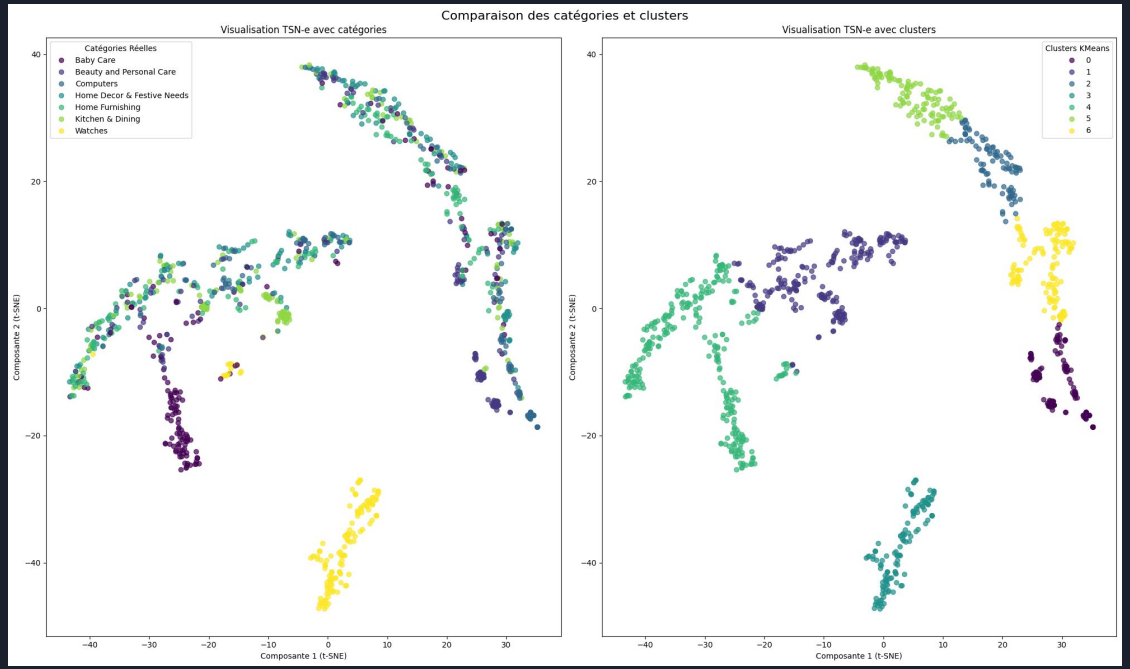
Tf-idf

ARI (Adjusted Rand Index) : 0.46



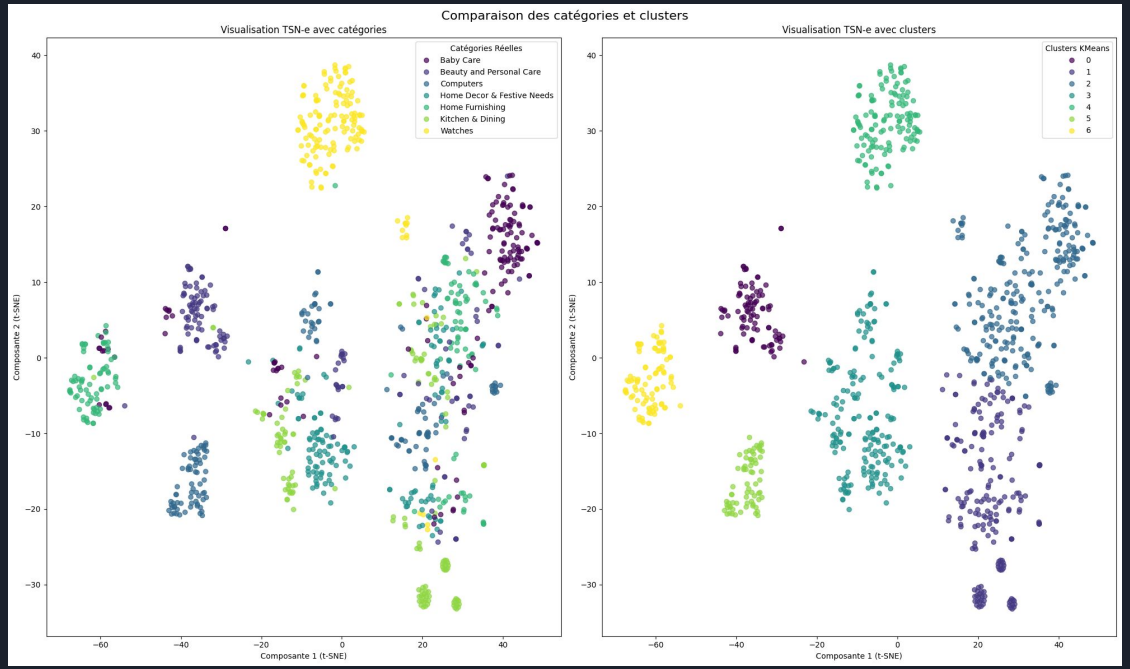
Word2Vec

ARI (Adjusted Rand Index) : 0.2



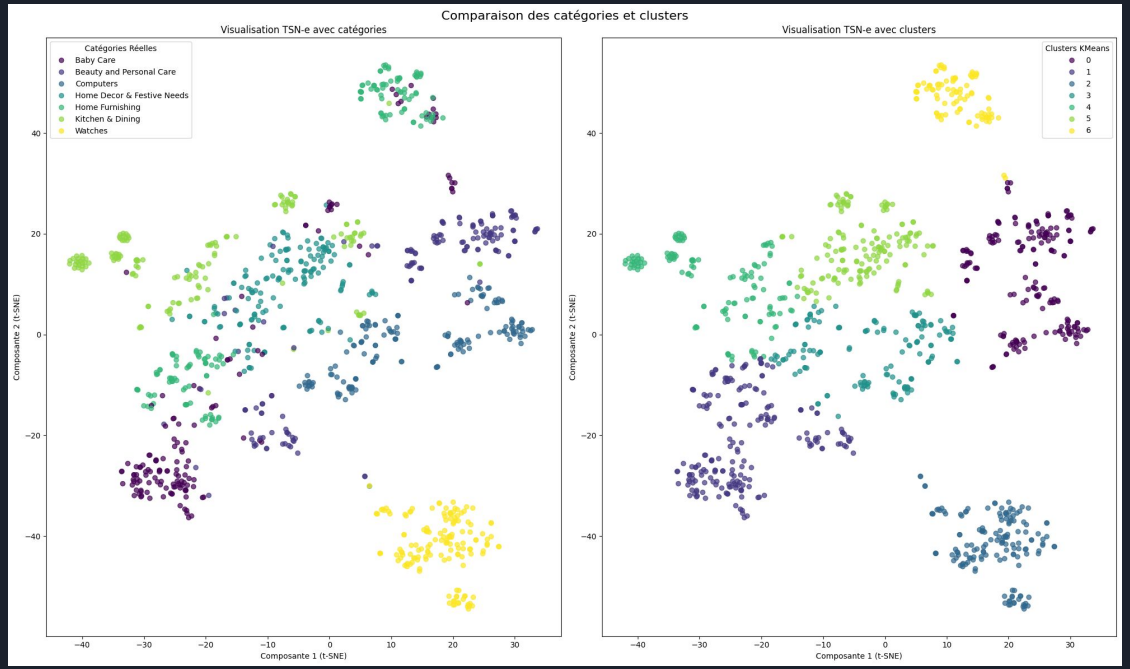
BERT

ARI (Adjusted Rand Index) : 0.29



USE

ARI (Adjusted Rand Index) : 0.46



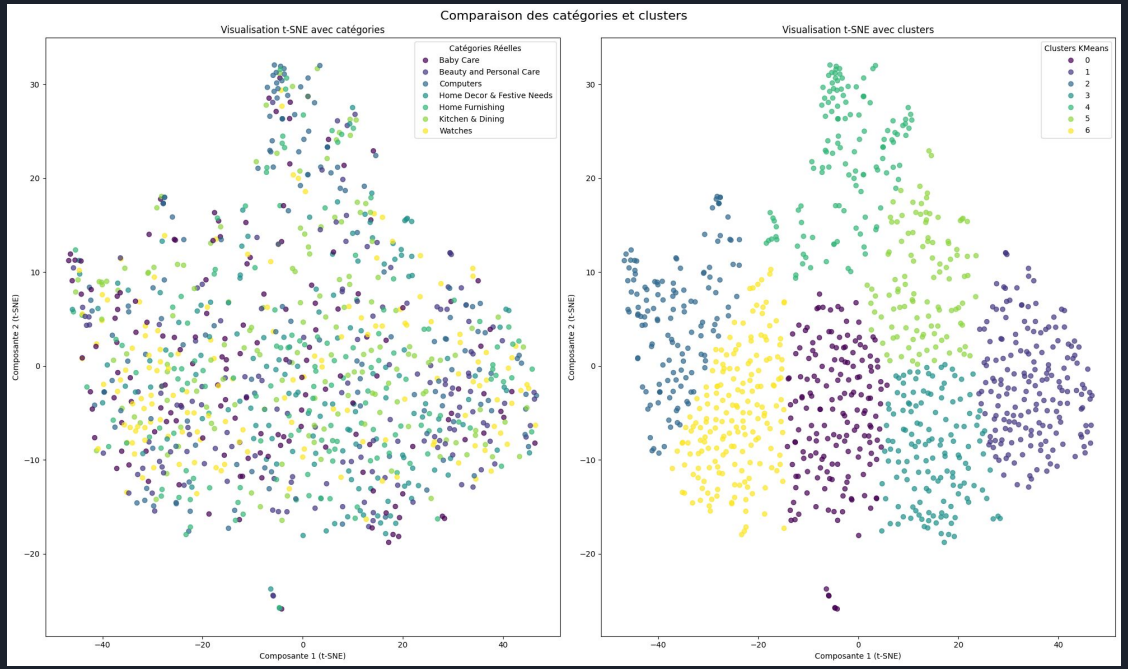
Récapitulatif des scores ARI

Les méthodes TF-IDF, CountVectorizer et USE ont donné les meilleurs résultats, mais des limites subsistent pour différencier certaines catégories.

	Méthode	ARI
0	CountVectorizer	0.46
1	Tf-idf	0.46
2	Word2Vec	0.20
3	BERT	0.29
4	USE	0.46

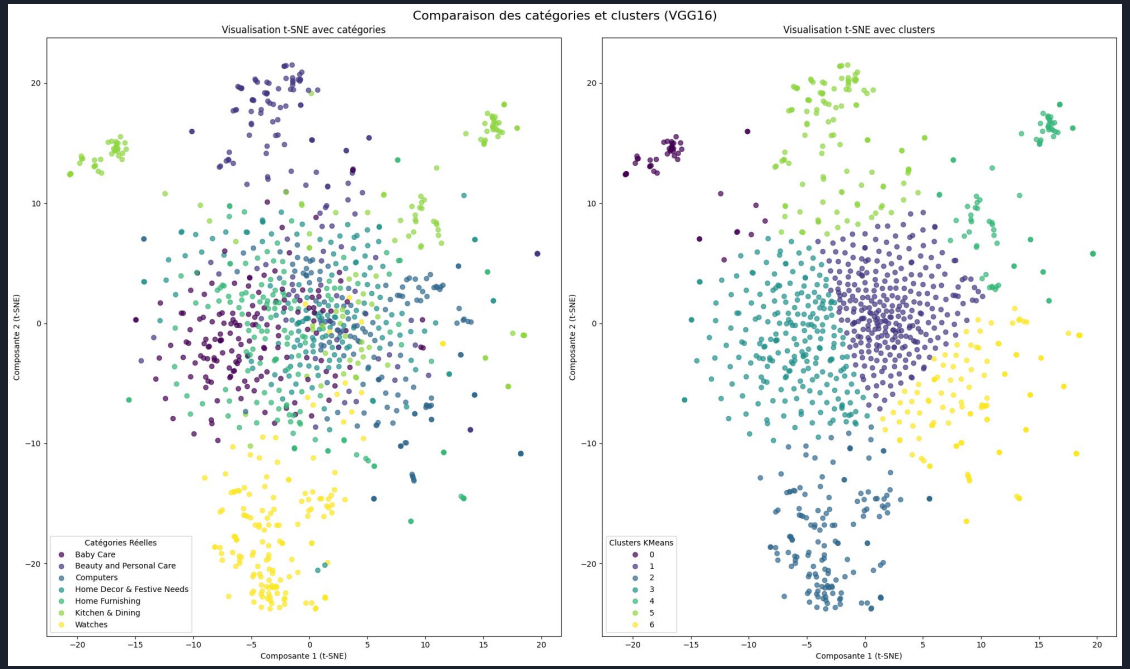
ORB

ARI (Adjusted Rand Index) : 0.02



VGG16

ARI (Adjusted Rand Index) : 0.3



Conclusion de l’étude de faisabilité

À travers diverses méthodes de vectorisation de texte et d’extraction de features d’image, certaines approches, comme TF-IDF et USE pour le texte, ou VGG16 pour les images, montrent des performances satisfaisantes pour distinguer certaines catégories. Cependant, les catégories aux descriptions ou caractéristiques visuelles similaires posent encore des défis.

Bien que ces méthodes capturent des différences entre produits, la séparation des catégories n’est pas encore optimale pour une automatisation fiable, en raison du chevauchement de vocabulaire et de caractéristiques visuelles.

La prochaine étape consiste à tester une classification supervisée sur les images, avec data augmentation, pour optimiser les modèles et améliorer les performances sur les catégories complexes.

Classification des Images

Objectif : Évaluer la faisabilité de la classification automatique des produits en analysant leurs images.

Approches Testées :

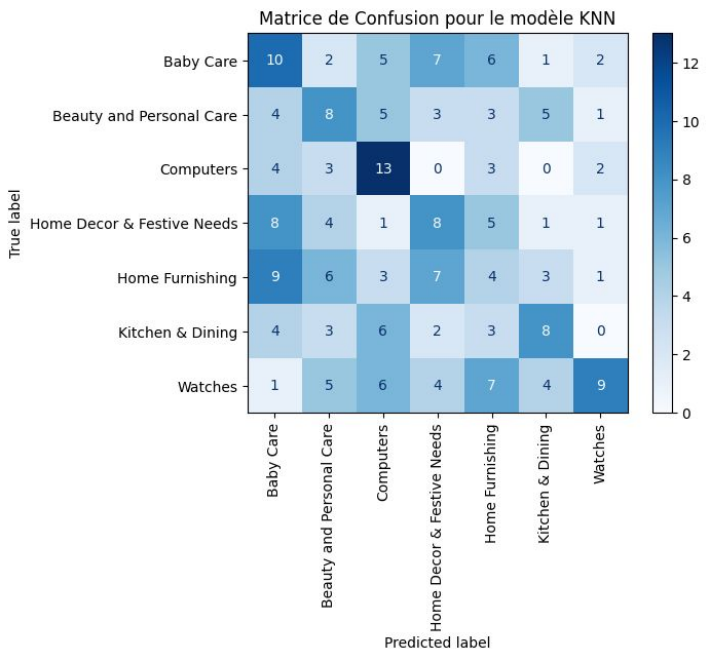
- ORB et KNN pour une approche basée sur les caractéristiques locales (points clés).
- VGG16 et ResNet50 pour une approche basée sur des réseaux de neurones convolutionnels profonds.

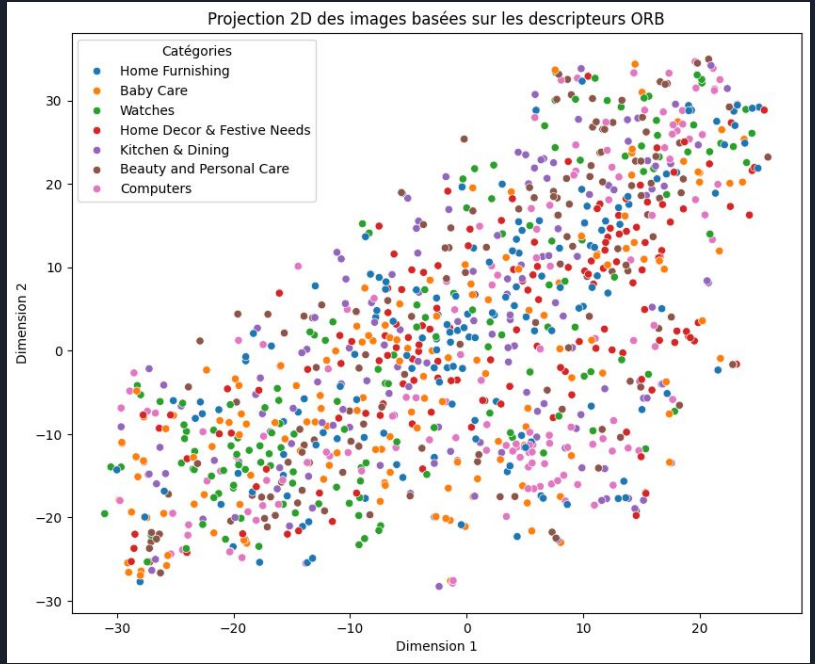
Défis Anticipés :

- Similitudes visuelles : Certaines catégories partagent des caractéristiques visuelles proches, comme 'Home Decor' et 'Home Furnishing'.
- Variabilité des images : Les différences de qualité et d’angle des images ajoutent une complexité à la tâche de classification.

ORB et KNN

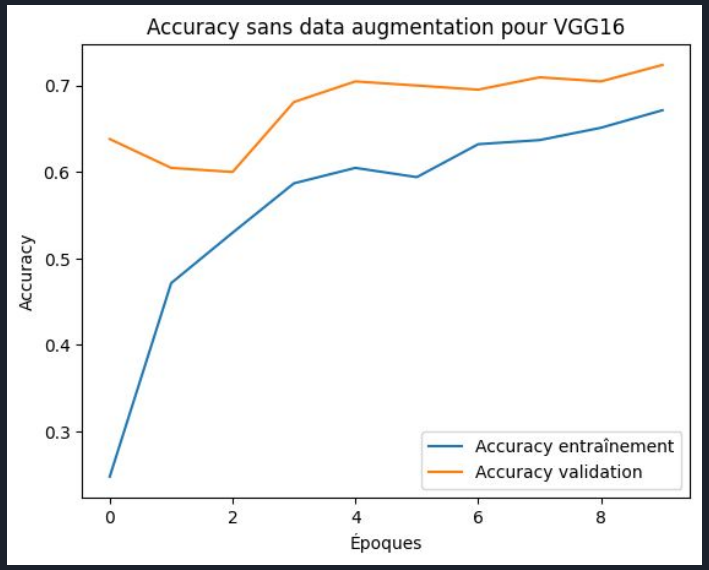
Précision du modèle KNN sur les features ORB : 0.29

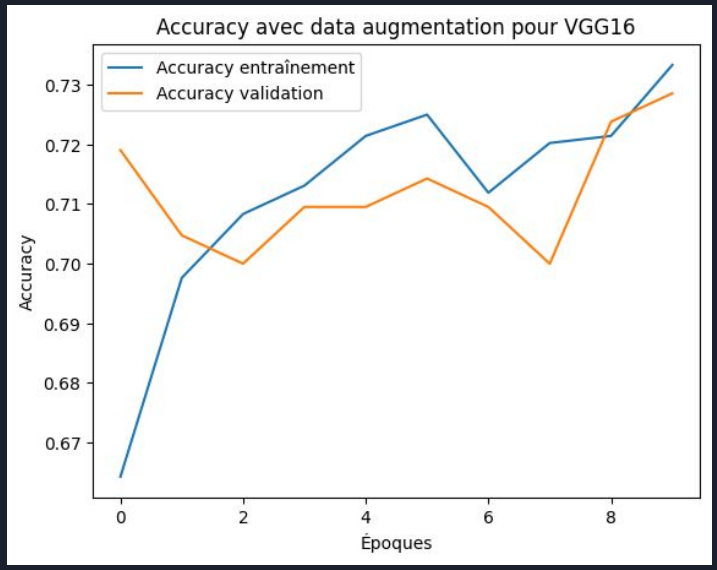




Classification avec VGG16

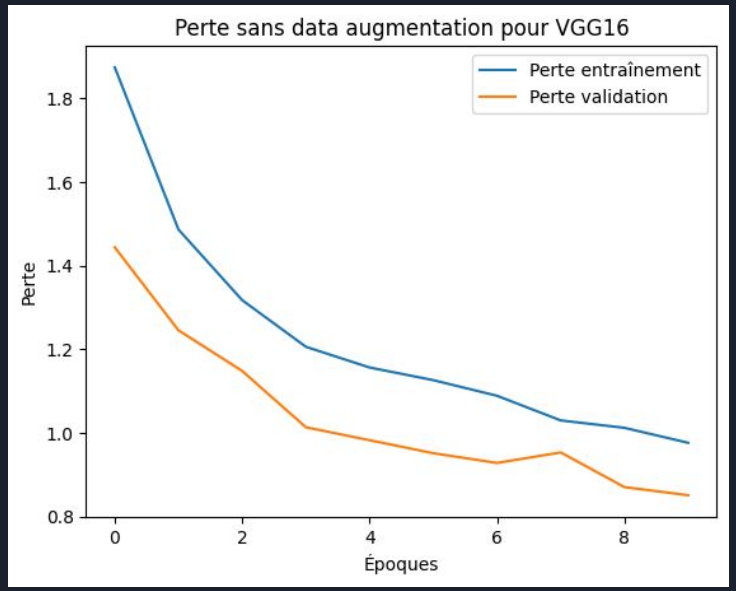
Comparaison de l'évolution de la précision d'entraînement et de validation de VGG16, avec et sans data augmentation.

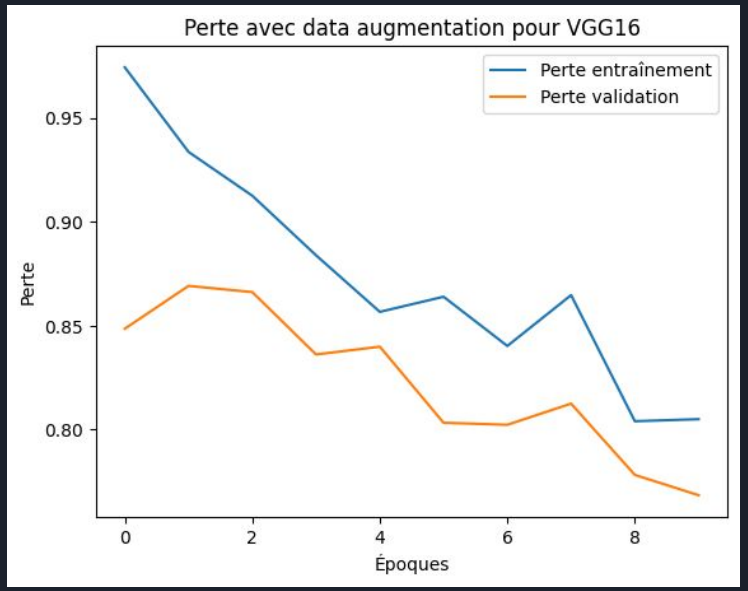




Classification avec VGG16

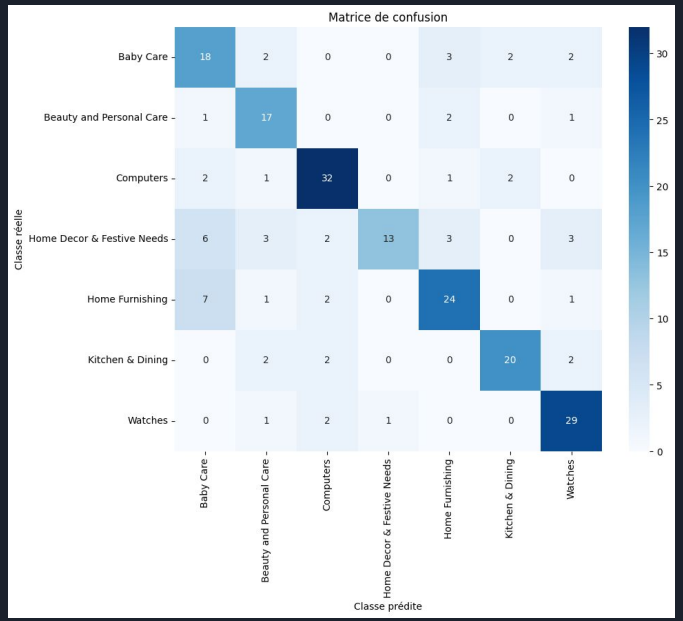
Comparaison de l'évolution de la perte d'entraînement et de validation de VGG16, avec et sans data augmentation.





Évaluation Finale des Performances de VGG16

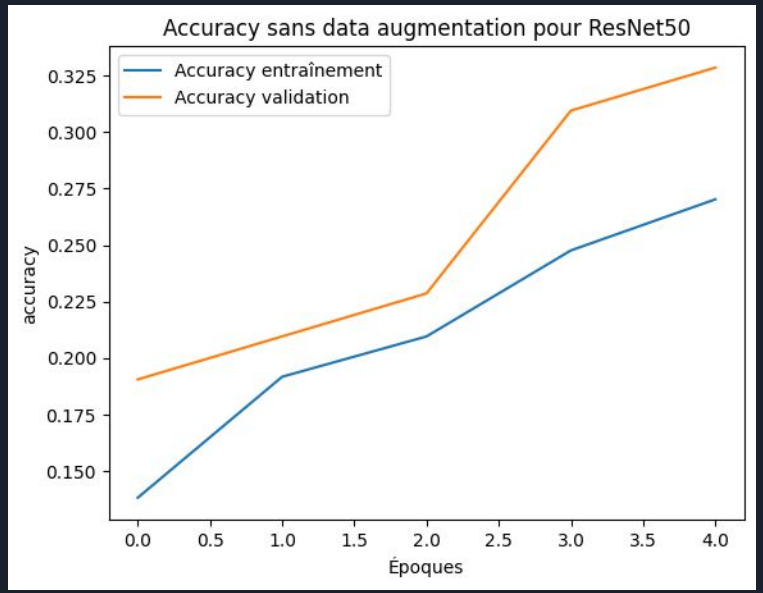
Résumé des performances de VGG16 par catégorie, avec la matrice de confusion et les mesures de précision, rappel et F1-score.

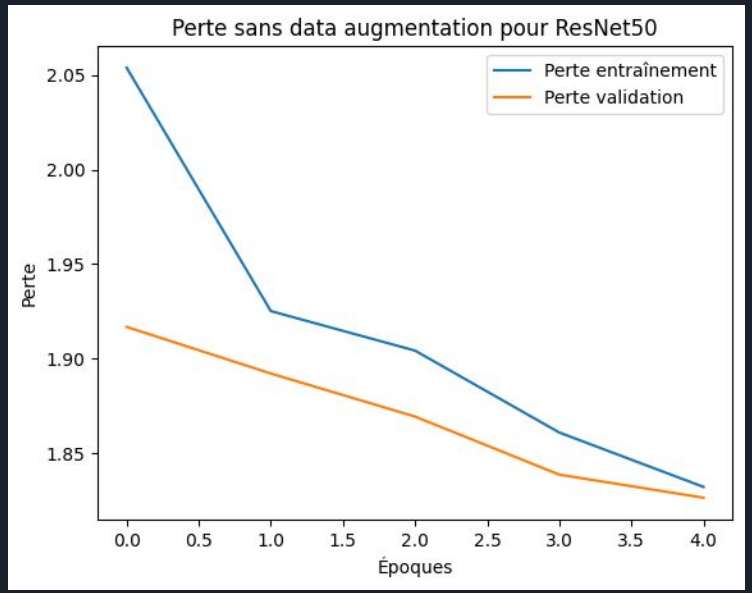


	precision	recall	f1-score	support
Baby Care	0.53	0.67	0.59	27
Beauty and Personal Care	0.63	0.81	0.71	21
Computers	0.80	0.84	0.82	38
Home Decor & Festive Needs	0.93	0.43	0.59	30
Home Furnishing	0.73	0.69	0.71	35
Kitchen & Dining	0.83	0.77	0.80	26
Watches	0.76	0.88	0.82	33
accuracy			0.73	210
macro avg	0.74	0.73	0.72	210
weighted avg	0.75	0.73	0.72	210

ResNet50

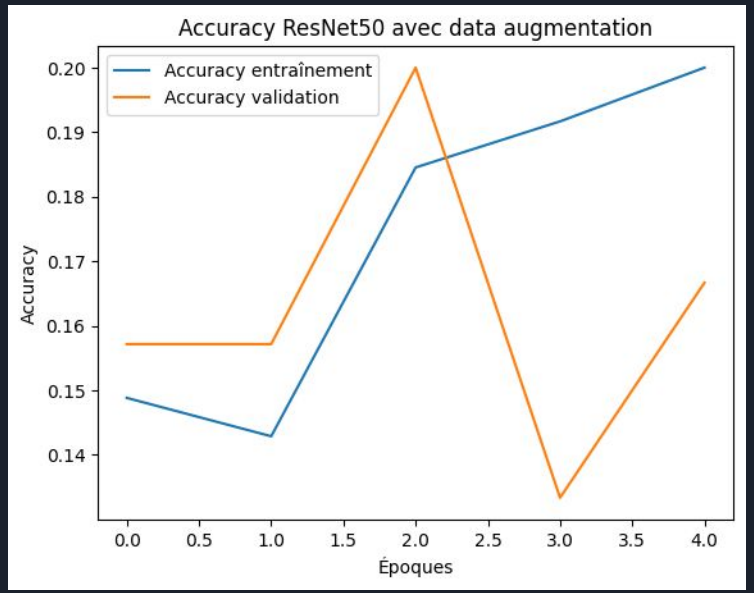
Évolution de la précision et de la perte de ResNet50 sans data augmentation.

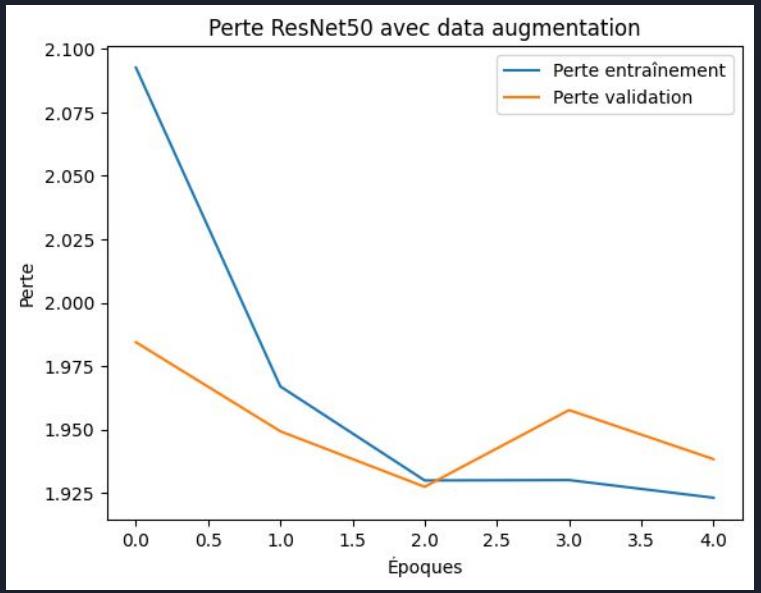




ResNet50

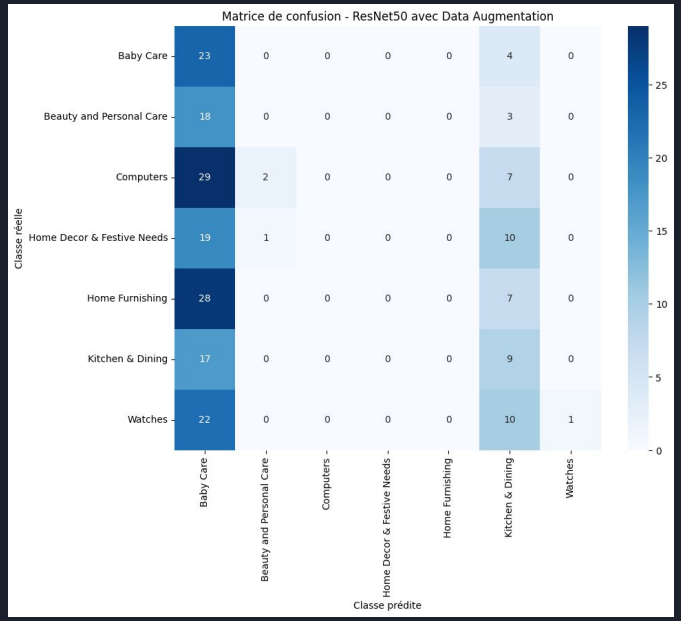
Évolution de la précision et de la perte de ResNet50 avec data augmentation.





Évaluation Finale des Performances de ResNet50 par Catégorie

Résumé des performances de ResNet50 par catégorie, avec matrice de confusion et mesures de précision, rappel et F1-score.



	precision	recall	f1-score	support
Baby Care	0.15	0.85	0.25	27
Beauty and Personal Care	0.00	0.00	0.00	21
Computers	0.00	0.00	0.00	38
Home Decor & Festive Needs	0.00	0.00	0.00	30
Home Furnishing	0.00	0.00	0.00	35
Kitchen & Dining	0.18	0.35	0.24	26
Watches	1.00	0.03	0.06	33
accuracy			0.16	210
macro avg	0.19	0.18	0.08	210
weighted avg	0.20	0.16	0.07	210

Bilan et Perspectives

Vectorisation textuelle

	Méthode	ARI
0	CountVectorizer	0.46
1	Tf-idf	0.46
2	Word2Vec	0.20
3	BERT	0.29
4	USE	0.46

Méthodes visuelles de base

	Méthode	ARI
0	ORB	0.02
1	VGG16	0.30

Résultats de VGG16

	precision	recall	f1-score	support
Baby Care	0.53	0.67	0.59	27
Beauty and Personal Care	0.63	0.81	0.71	21
Computers	0.80	0.84	0.82	38
Home Decor & Festive Needs	0.93	0.43	0.59	30
Home Furnishing	0.73	0.69	0.71	35
Kitchen & Dining	0.83	0.77	0.80	26
Watches	0.76	0.88	0.82	33
accuracy			0.73	210
macro avg	0.74	0.73	0.72	210
weighted avg	0.75	0.73	0.72	210

Résultats de ResNet50

	precision	recall	f1-score	support
Baby Care	0.15	0.85	0.25	27
Beauty and Personal Care	0.00	0.00	0.00	21
Computers	0.00	0.00	0.00	38
Home Decor & Festive Needs	0.00	0.00	0.00	30
Home Furnishing	0.00	0.00	0.00	35
Kitchen & Dining	0.18	0.35	0.24	26
Watches	1.00	0.03	0.06	33
accuracy			0.16	210
macro avg	0.19	0.18	0.08	210
weighted avg	0.20	0.16	0.07	210

API Edamam

Résultats de l'intégration de l'API Edamam

	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_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
8	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN
9	food_a79xmnya6togreaeukbroa0thhh0	Champagne Chicken	Generic meals	Flour; Salt; Pepper; Boneless, Skinless Chicke...	NaN

Avez-vous des questions?