Classifiez automatiquement des biens de consommation

OPENCLASSROOMS

ENJEUX

Contexte : Automatisation de l'attribution de catégories de produits pour une marketplace e-commerce.

Objectif : Développer un moteur de classification des articles en utilisant les descriptions textuelles et les images.

Importance :

- Améliorer l'expérience utilisateur pour les vendeurs et acheteurs.
- Passer à l'échelle pour gérer un volume croissant d'articles.

Approches Utilisées :

- Prétraitement et extraction de features pour le texte et les images.
- Réduction de dimension pour visualisation et validation des clusters.
- Classification supervisée pour affiner le modèle.
- Exploration de l'API pour élargir l'offre de produits.

SOMMAIRE

Méthodologie

Textes :

Etape 1 : Prétraitement des Données

- Nettoyage/Prétraitement du texte
- Descriptions textuelles → représentations numériques
- Word Embeddings

Etape 2 : Extraction de Caractéristiques

Texte prétraité → caractéristiques exploitables pour l'apprentissage automatique.

Etape 3 : Modélisation

Application de modèles de classification basiques ou basés sur des réseaux de neurones.

Etape 4 : Evaluation

Regression Logistique → Accuracy Score
TSNE / kmeans → Adjusted Rand Index (ARI) Score

Images :

Etape 1 : Prétraitement des Images

- Normalisation des images
- Extraction des descripteurs d'images à l'aide d'algorithmes

Etape 2 : Clustering des Descripteurs

Application de kmeans pour regrouper les descripteurs extraits
→ Réduit la dimensionalité
→ Crée des histogrammes de caractéristiques pour chaque image.

Etape 3 : Extraction de Caractéristiques

Modèles de réseaux de neurones pré-entraînés → extrait des caractéristiques

Etape 4 : Modélisation

Etape 5 : Evaluation

Prétraitement du Texte

Texte original :	Delicious Organic Almond Butter - Better for Spreads and Baking!
Après conversion en minuscules :	delicious organic almond butter - better for spreads and baking!
Après tokenization :	['delicious', 'organic', 'almond', 'butter', 'better', 'for', 'spreads', 'and', 'baking']
Après suppression des stopwords :	['delicious', 'organic', 'almond', 'butter', 'better', 'spreads', 'baking']
Après stemming :	['delici', 'organ', 'almond', 'butter', 'better', 'spread', 'bake']
Après lemmatization :	['delici', 'organ', 'almond', 'butter', ‘good', 'spread', 'bake']
Texte final après prétraitement :	delici organ almond butter good spread bake

Vectorisations Basiques des Données

Bag of Words (BoW)

- Convertit le texte en un format numérique en comptant la fréquence des mots dans chaque document
- Utilisation d'un vecteur de mots qui crée une matrice :
  → ligne = document
  → colonne = mot
- Chaque cellule contient le nombre d'occurrences du mot dans le document

TF-IDF (Term Frequency - Inverse Document Frequency)

- Texte → représentation numérique
- Mots pondérés selon leur fréquence dans le document et leur rareté dans le corpus

TF : Fréquence d'un mot dans un document.

IDF : Mesure l'importance d'un mot dans l'ensemble des documents (moins fréquent = plus important).

- Atténue l'importance des mots courants, en favorisant ceux plus spécifiques au document.

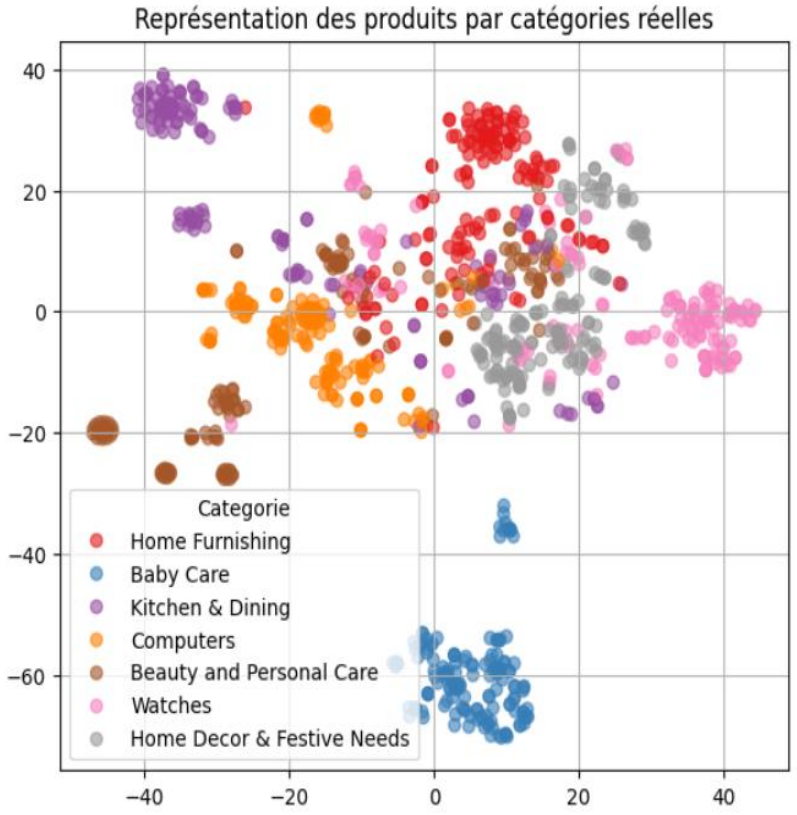
Bag of Words (BoW)

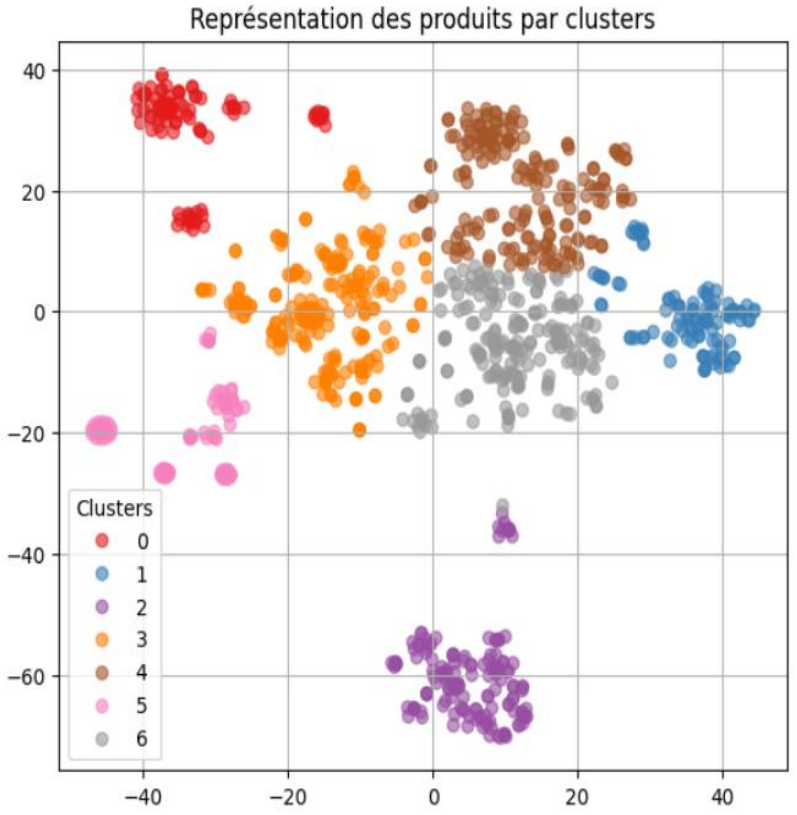
Source = Description

Récurrence max. = 95%
Récurrence min. = 1%

Régression Logistique
→ Accuracy = 93,33%

Clusterisation k-means
→ ARI Score = 42,84%





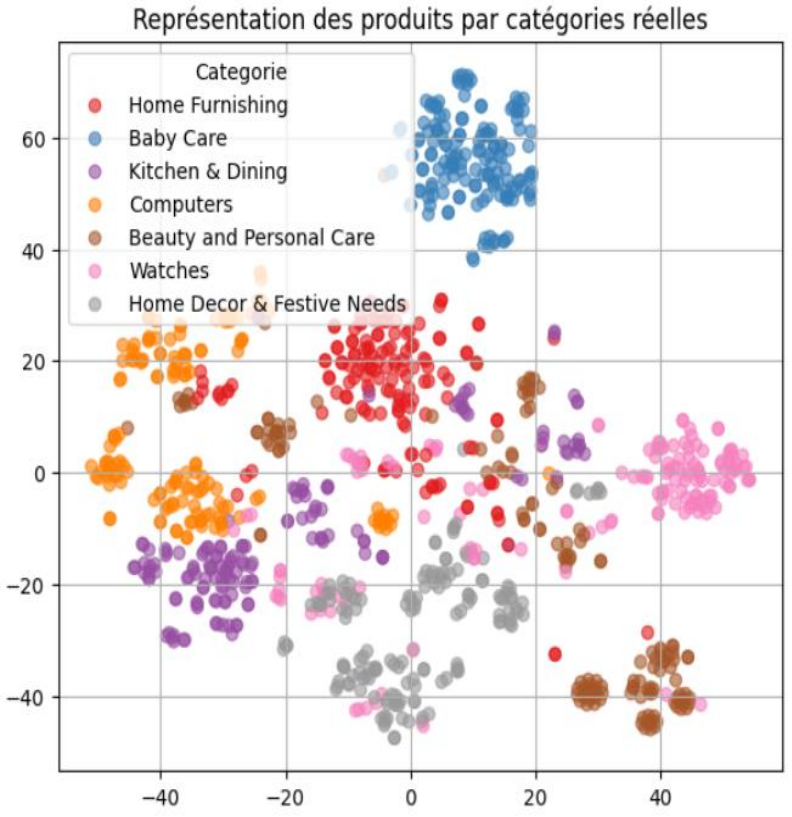
Term Frequency-Inverse Document Frequency (TF-IDF)

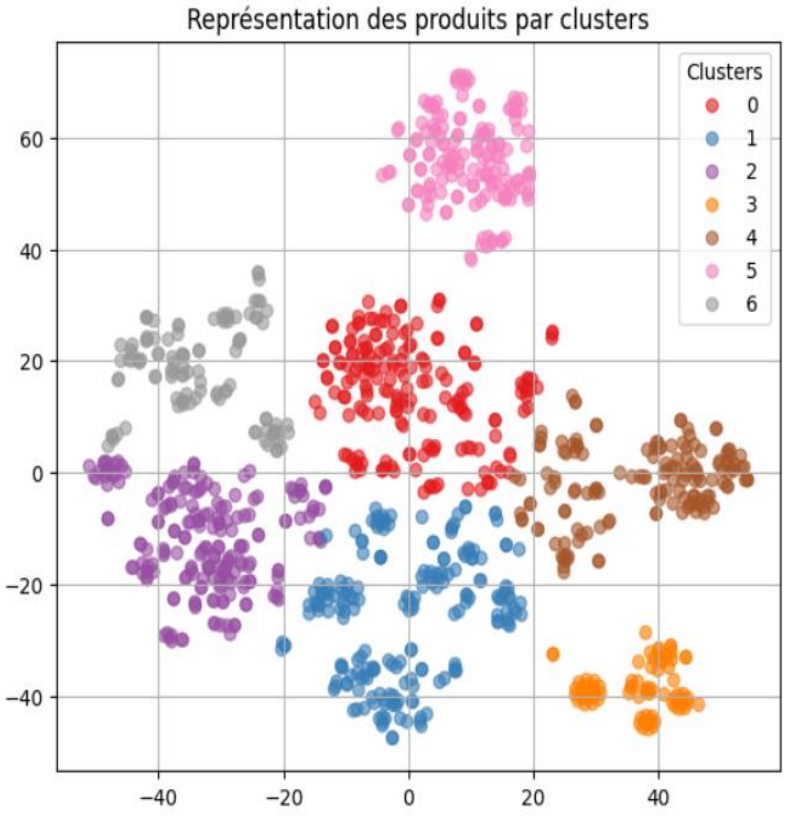
Source = Description

Récurrence max. = 99%
Récurrence min. = 2

Régression Logistique
→ Accuracy = 94,76%

Clusterisation k-means
→ ARI Score = 53,08%





Vectorisations Avancées des Données

Word2Vec

- Crée des vecteurs de mots basés sur leur contexte dans les phrases, capturant des relations sémantiques.
- Phrases tokenisées & vecteur d'embedding moyen calculé pour chaque document à partir des vecteurs des mots qu'il contient.

Universal Sentence Encoder (USE)

- Encode les phrases entières en vecteurs de manière → capture la signification globale des phrases
- Phrases traitées par lots & embeddings générés concaténés → vecteur représentant chaque document

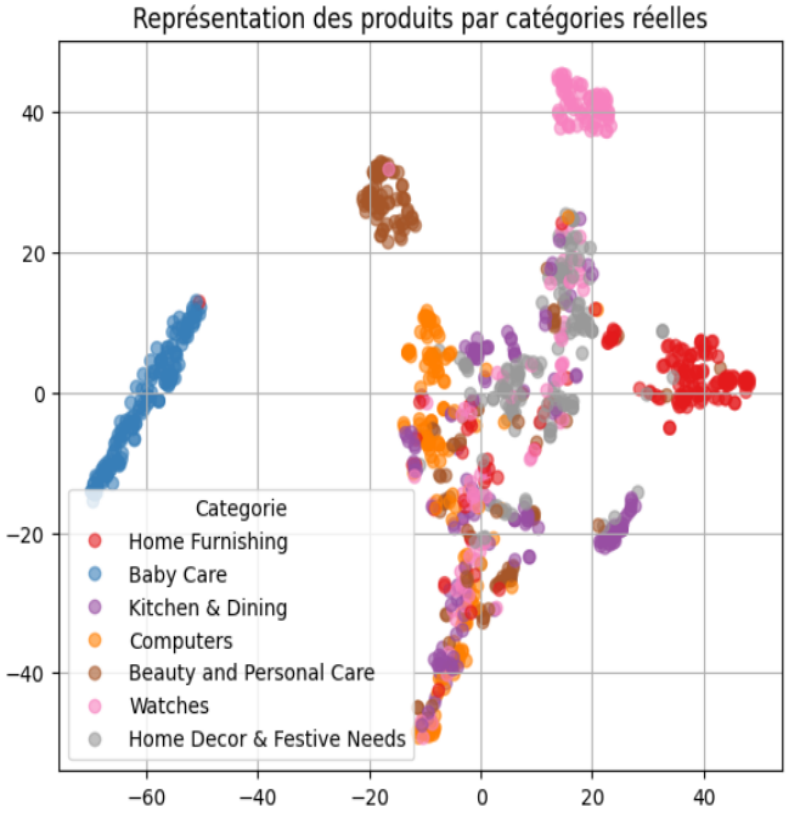
Word2Vec

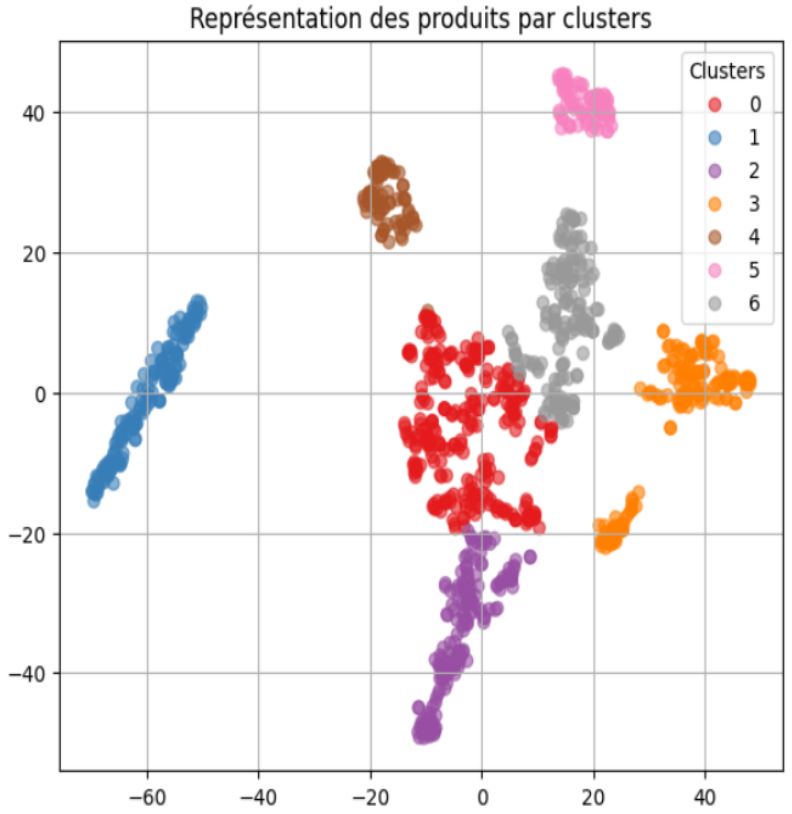
Source = product_name

Récurrence min. = 1
Mots combinés = 5
Dimension vecteurs = 100

Régression Logistique
→ Accuracy = 73,81%

Clusterisation k-means
→ ARI Score = 36,39%





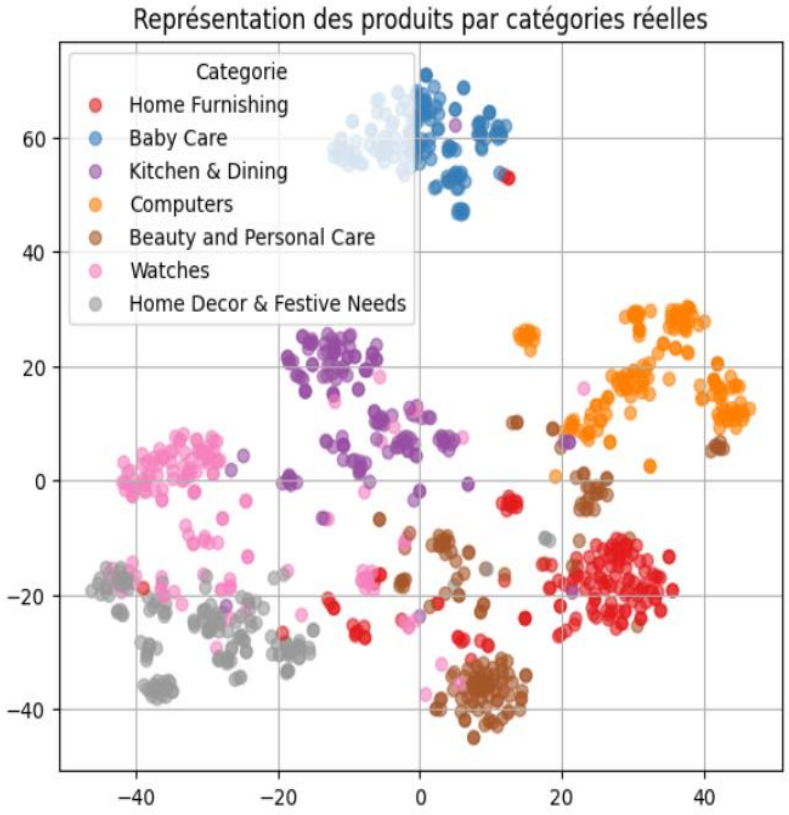
Universal Sentence Encoder (USE)

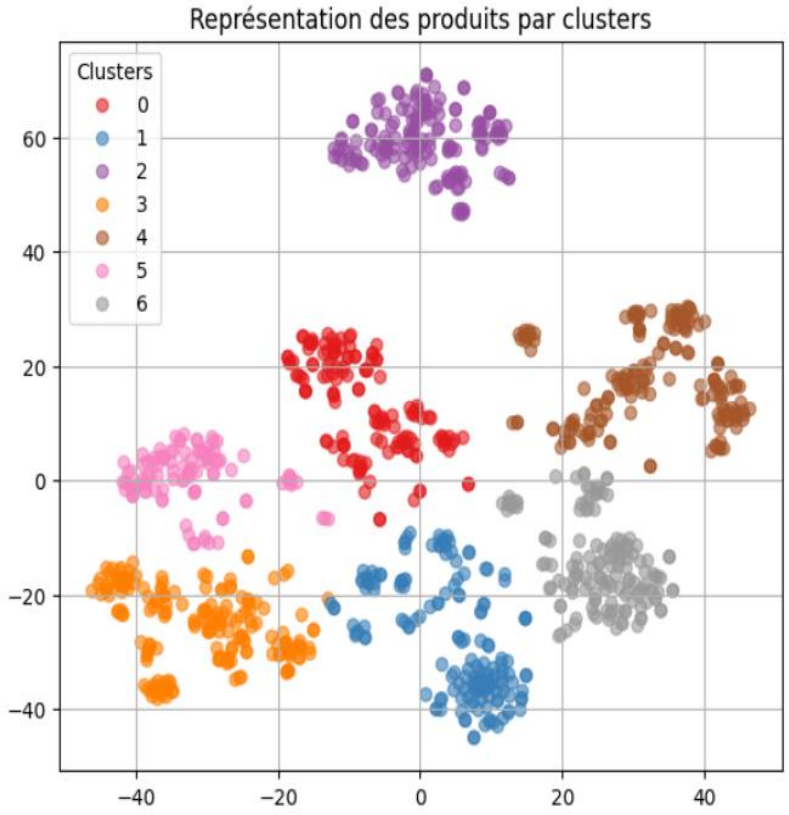
Source = product_name

Application du modèle par lots de 10 phrases

Régression Logistique
→ Accuracy = 89,05%

Clusterisation k-means
→ ARI Score = 70,44%





Vectorisations Avancées des Données

BERT (Hugging Face)

- Modèle pré-entraîné capable de comprendre le contexte des mots dans une phrase
- Modèle de langage basé sur des Transformers. Entraîné sur des tâches de remplissage de masque et de prédiction de phrase
- Représentations contextuelles bidirectionnelles des mots, adaptées via la bibliothèque Hugging Face pour une personnalisation simple et rapide.

BERT (TensorFlow)

- Version de BERT utilisant TensorFlow
- Conçu pour des applications similaires à celles du modèle Hugging Face, avec un focus sur l'intégration et la prédiction via TensorFlow.
- Méthode d'extraction des caractéristiques identique + intégration de TensorFlow pour des opérations de prétraitement et de traitement en lot.

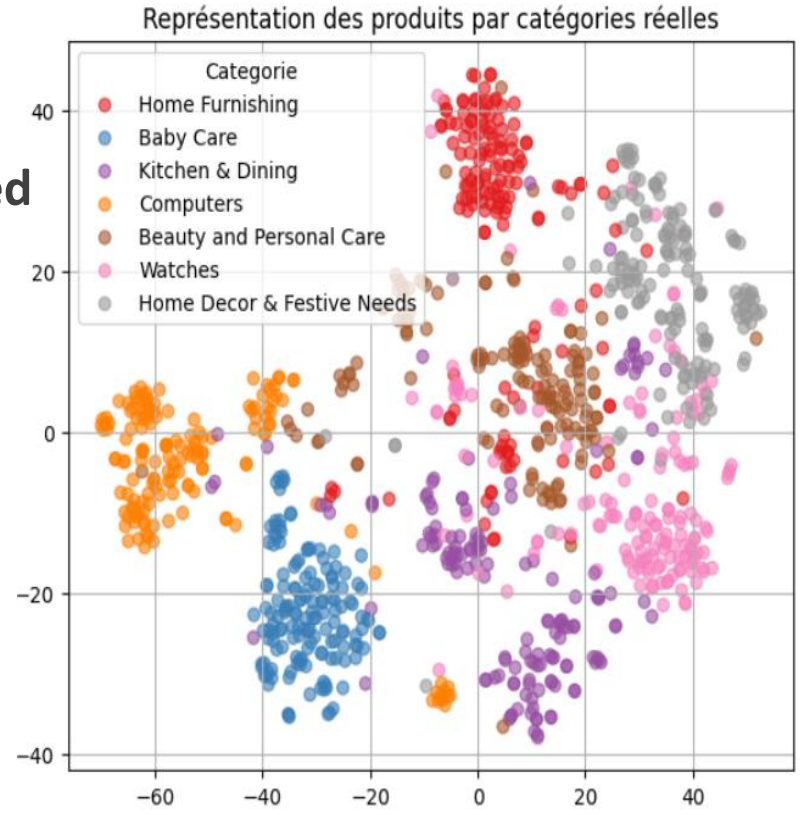
BERT (Hugging Face)

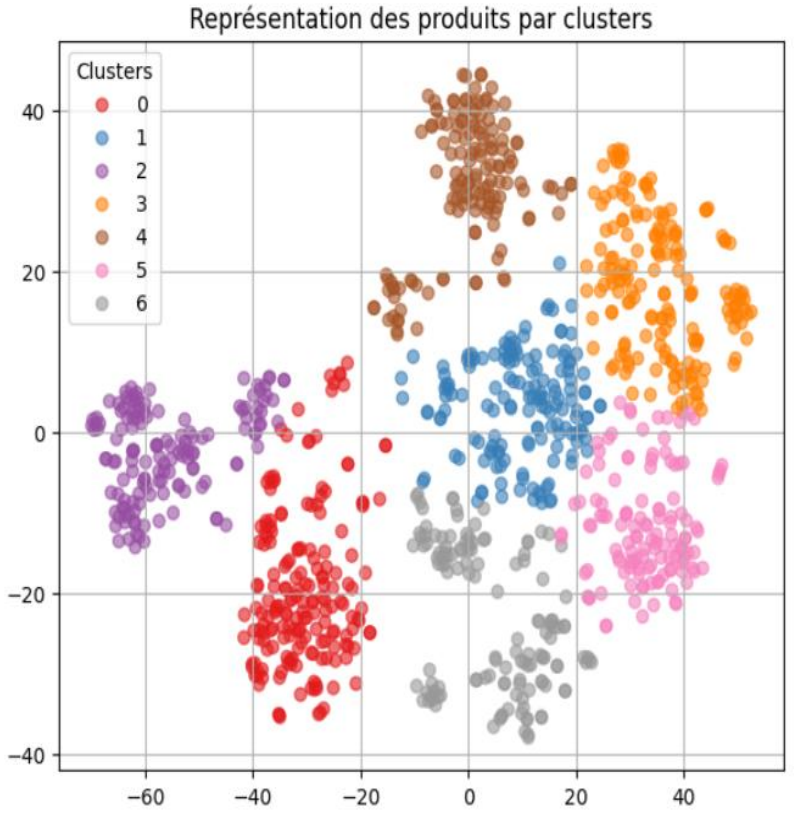
Source = product_name

Modèle : bert-base-uncased
Max tokens/phrase = 160

Régression Logistique
→ Accuracy = 92,86%

Clusterisation k-means
→ ARI Score = 60,49%





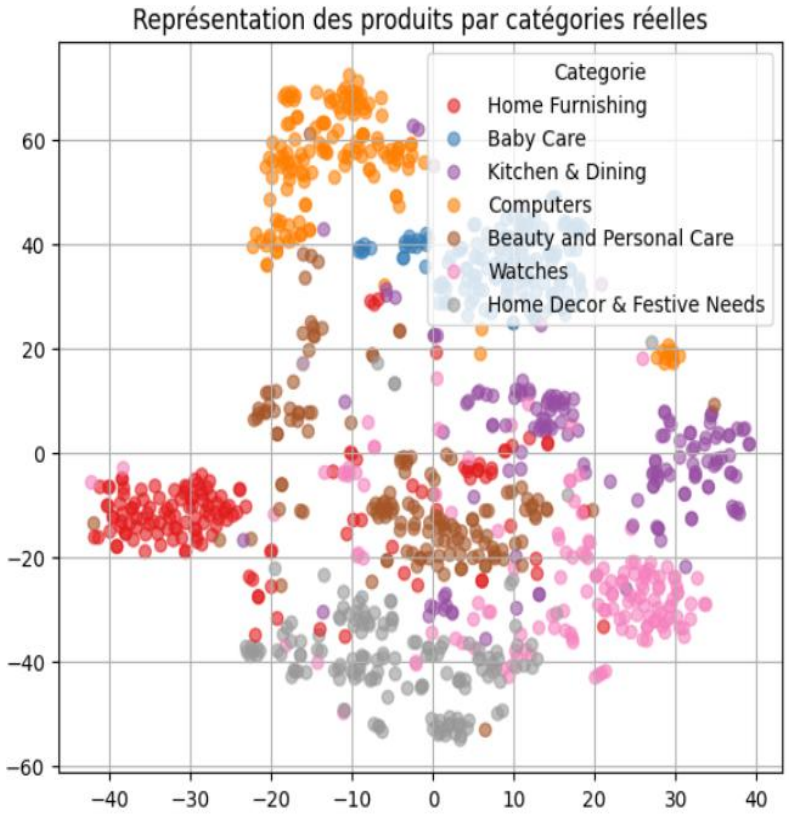
BERT (Tensor Flow)

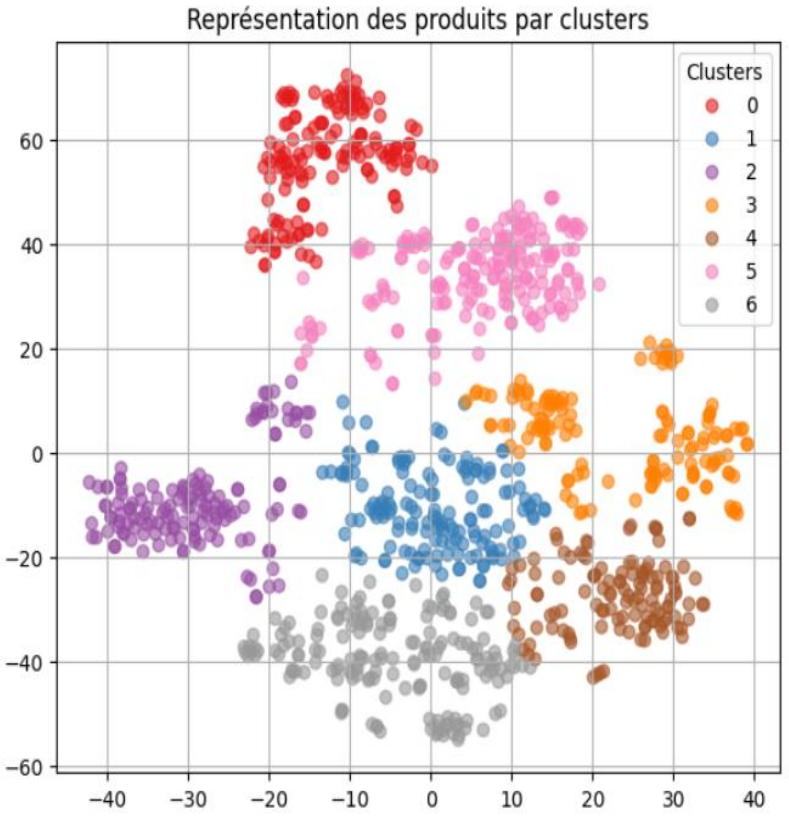
Source = product_name

TensorFlow Hub : bert_en_uncased_L-12_H-768_A-12
Même schéma que Hugging Face

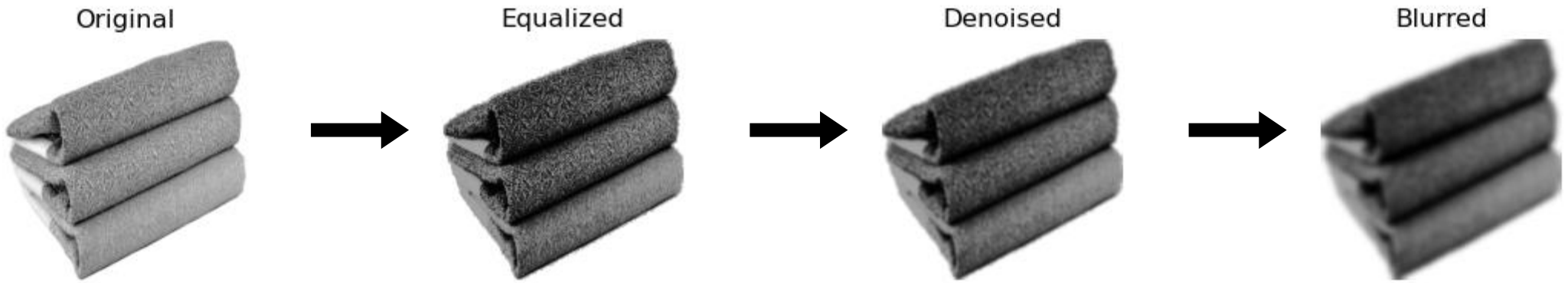
Régression Logistique
→ Accuracy = 92,86%

Clusterisation k-means
→ ARI Score = 60,09%





Prétraitement de l’Image



- Conversion en niveaux de gris
- Redimensionnement

Égalisation :

+ contraste global

Uniformisation des niveaux d'intensité

Filtrage & Réduction du bruit :

- Détails inutiles
- Irrégularités visuelles

Méthodes d’Extraction Basiques

Scale-Invariant Feature Transform (SIFT)

- Identifie des points d’intérêt dans une image, invariants aux changements d’échelle, de rotation, et d’éclairage.
- Robuste pour la reconnaissance d'objets sous différentes perspectives et conditions visuelles, mais coûteux en calcul.

Oriented FAST and Rotated BRIEF (ORB)

- Détecte des points d’intérêt et génère des descripteurs rapidement.
- Efficace pour les systèmes en temps réel, avec invariance à la rotation, mais sensible aux changements d'échelle.

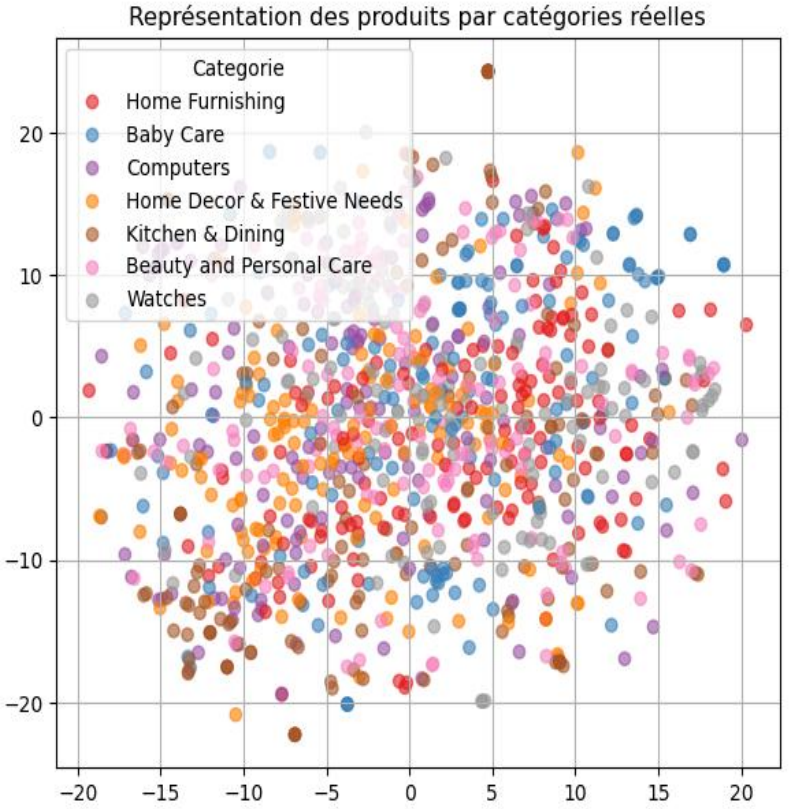
Scale-Invariant Feature Transform (SIFT)

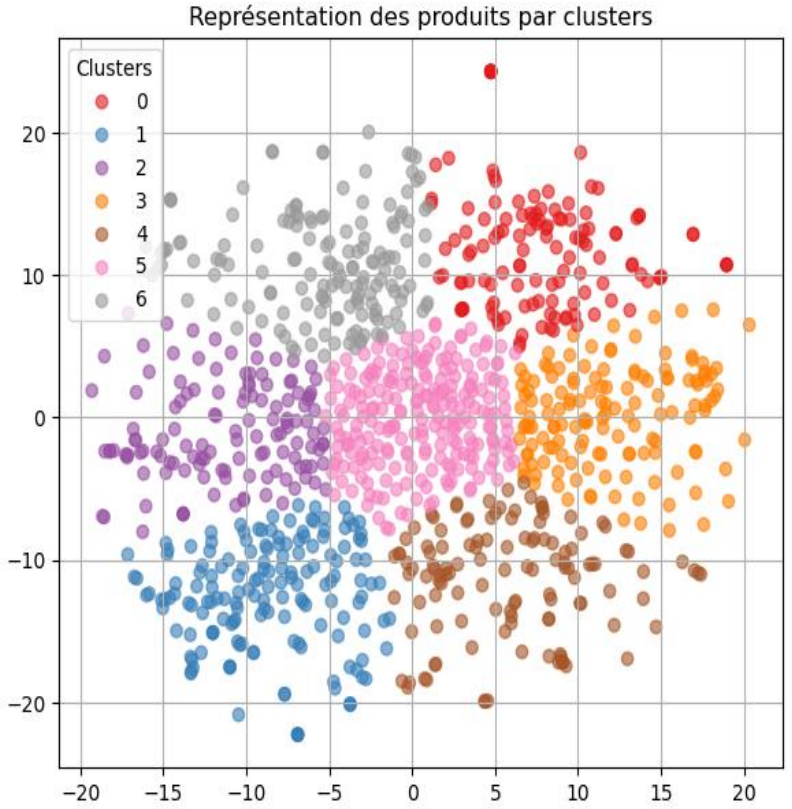
Nombre de descripteurs = 285114

Nombre de clusters estimés = 534

Régression Logistique
→ Accuracy = 19,52%

Clusterisation k-means
→ ARI Score = 4,19%





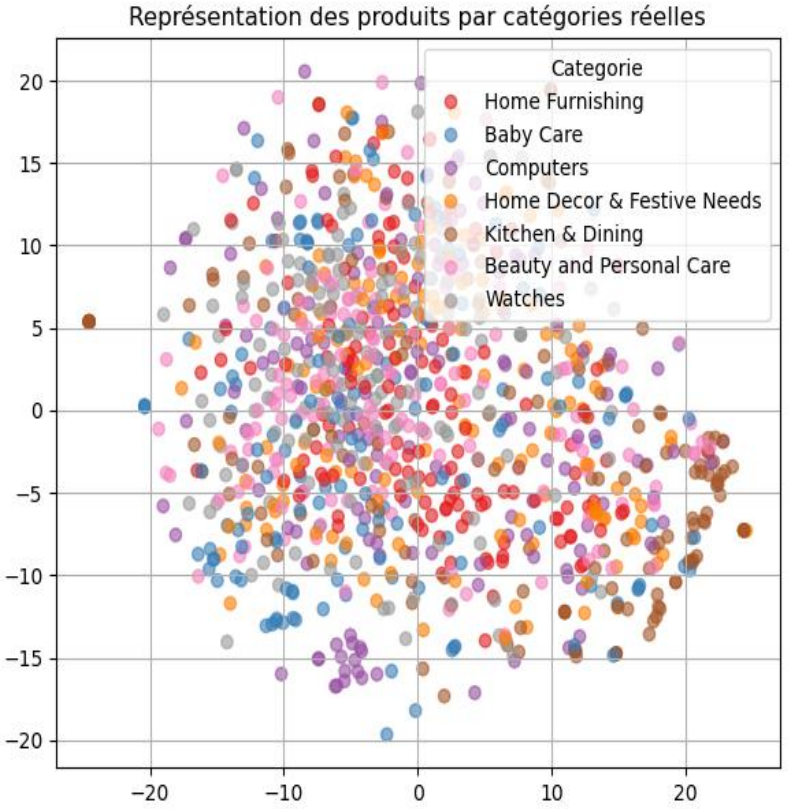
Oriented FAST and Rotated BRIEF (ORB)

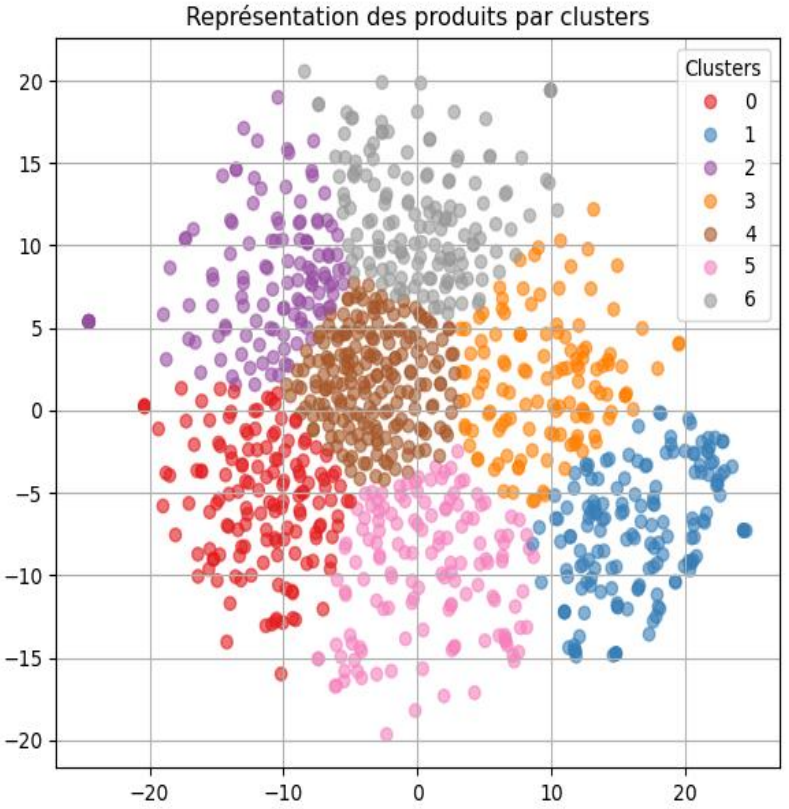
Nombre de descripteurs = 329971

Nombre de clusters estimés = 574

Régression Logistique
→ Accuracy = 18,57%

Clusterisation k-means
→ ARI Score = 2,87%





Réseaux de Neurones Pré-Entraînés (CNN)

VGG16

Réseau de neurones convolutifs profond constitué de 16 couches, utilisé pour la classification d'images.

Utilise des filtres de petite taille (3x3), offrant une grande précision mais avec un temps d'entraînement et une consommation de mémoire élevés.

ResNet50

Réseau de neurones convolutifs à 50 couches intégrant des connexions résiduelles pour prévenir la dégradation des performances avec l'augmentation de la profondeur.

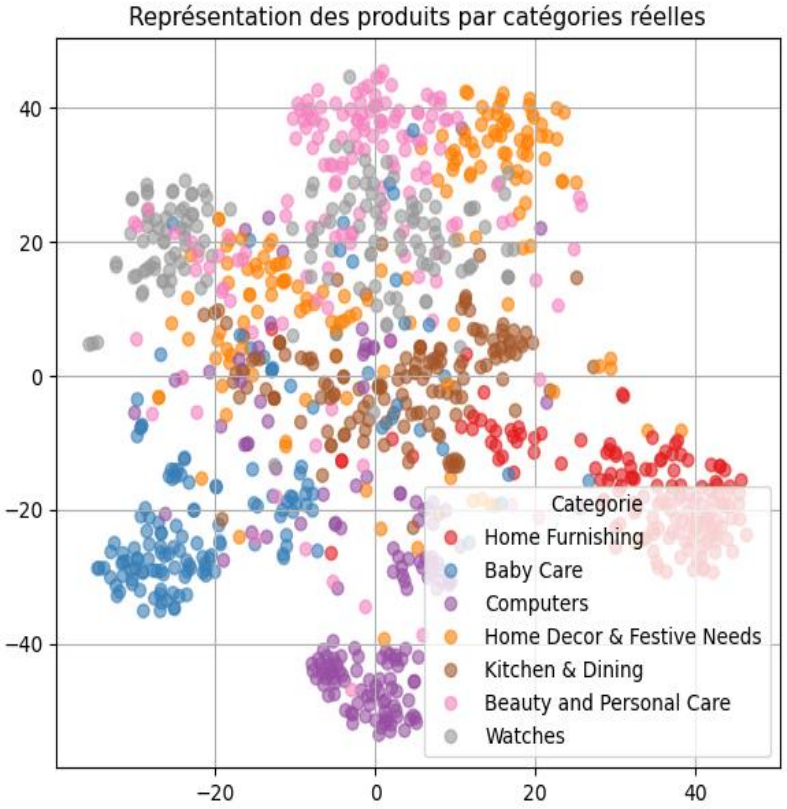
Permet un entraînement de réseaux très profonds, tout en conservant des performances élevées sur des tâches complexes comme la classification d'images.

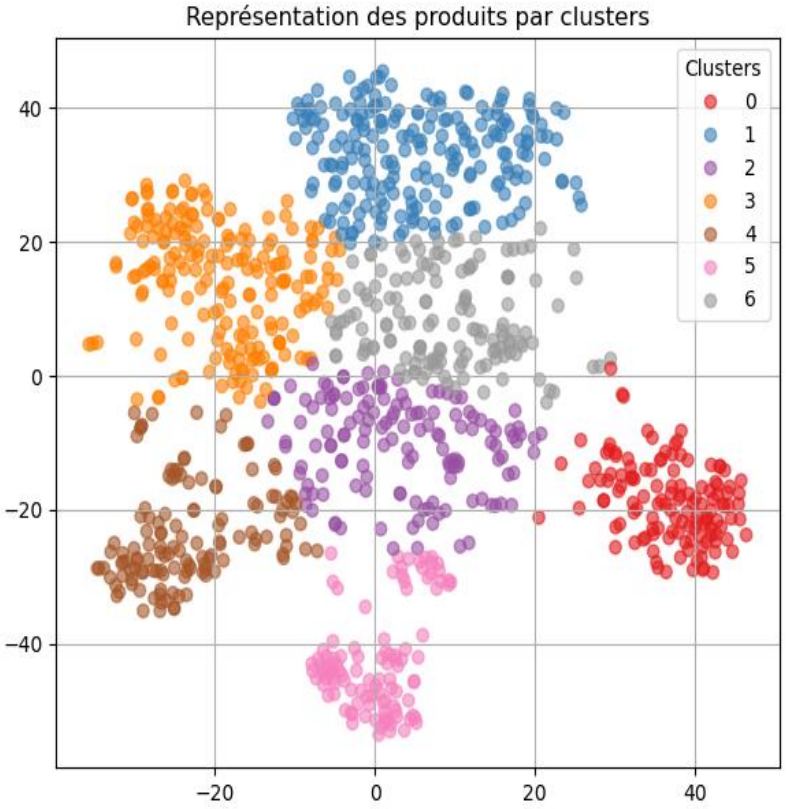
VGG16

Nombre de descripteurs = 4096

Régression Logistique
→ Accuracy = 81,90%

Clusterisation k-means
→ ARI Score = 35,47%



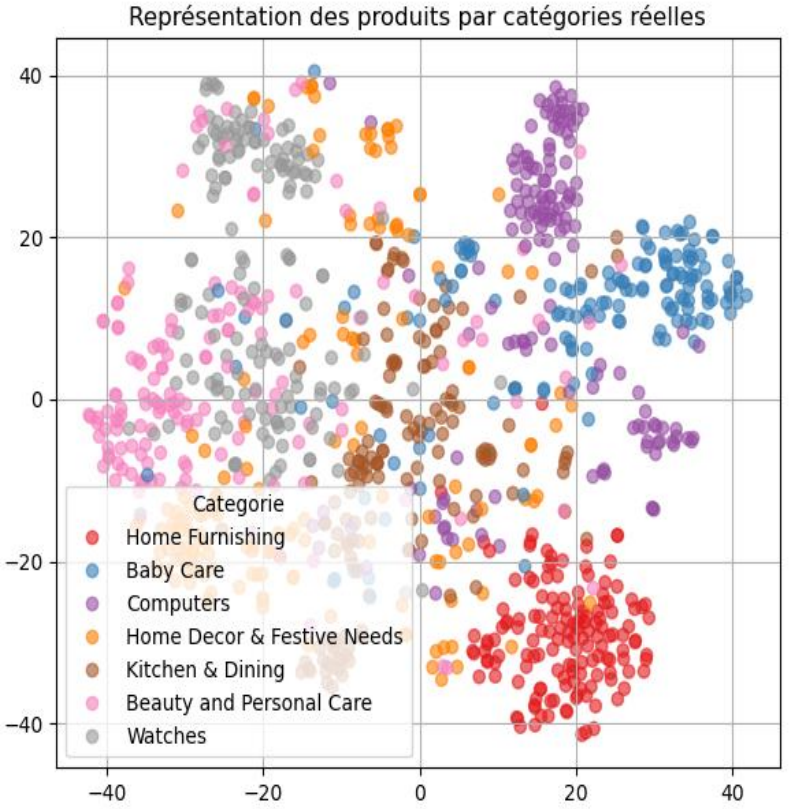


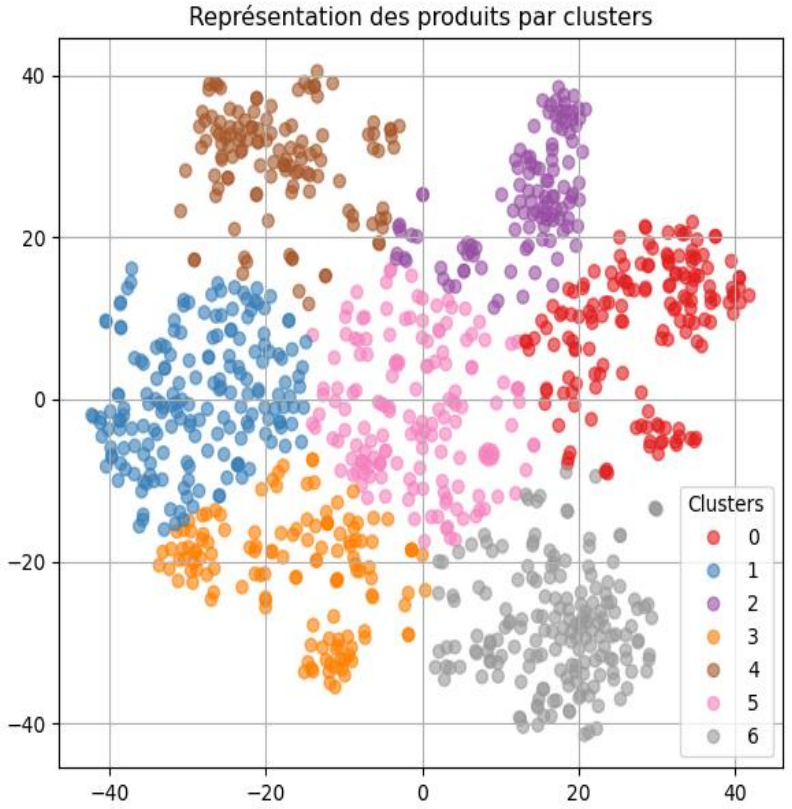
RestNet50

Nombre de descripteurs = 2048

Régression Logistique
→ Accuracy = 82,86%

Clusterisation k-means
→ ARI Score = 38,77%





Etude de Faisabilité – Bilan

Textes

1. USE :

Accuracy = 89,05%
ARI Score = 70,44%

2. BERT (HF) :

Accuracy = 92,86%
ARI Score = 60,49%

3. BERT (TF) :

Accuracy = 92,86%
ARI Score = 69,09%

Images

1. RestNet50 :

Accuracy = 82,86%
ARI Score = 38,05%

2. VGG16 :

Accuracy = 83,33%
ARI Score = 35,01%

Textes

- Techniques modernes surpassent les modèles traditionnels
- Importance des représentations contextuelles → Amélioration significative des performances en traitement du langage naturel.

Images

- Méthodes traditionnelles largement dépassées par réseaux de neurones convolutifs
- Précision supérieure dans le traitement des images de produits, facilitant la séparation des classes.

Utilisation de méthodes avancées cruciale pour améliorer les performances de classification et de recommandation.

Synergie des approches texte et image : vers des systèmes plus robustes et intuitifs, améliorant l'expérience utilisateur.

Classification Supervisée

Approche Préparation Initiale des Images :

- Prétraitement des images (redimensionnement, normalisation) avant l'entraînement.
- Pas de transformations dynamiques, uniquement des modifications statiques.

Approche ImageDataGenerator avec Data Augmentation :

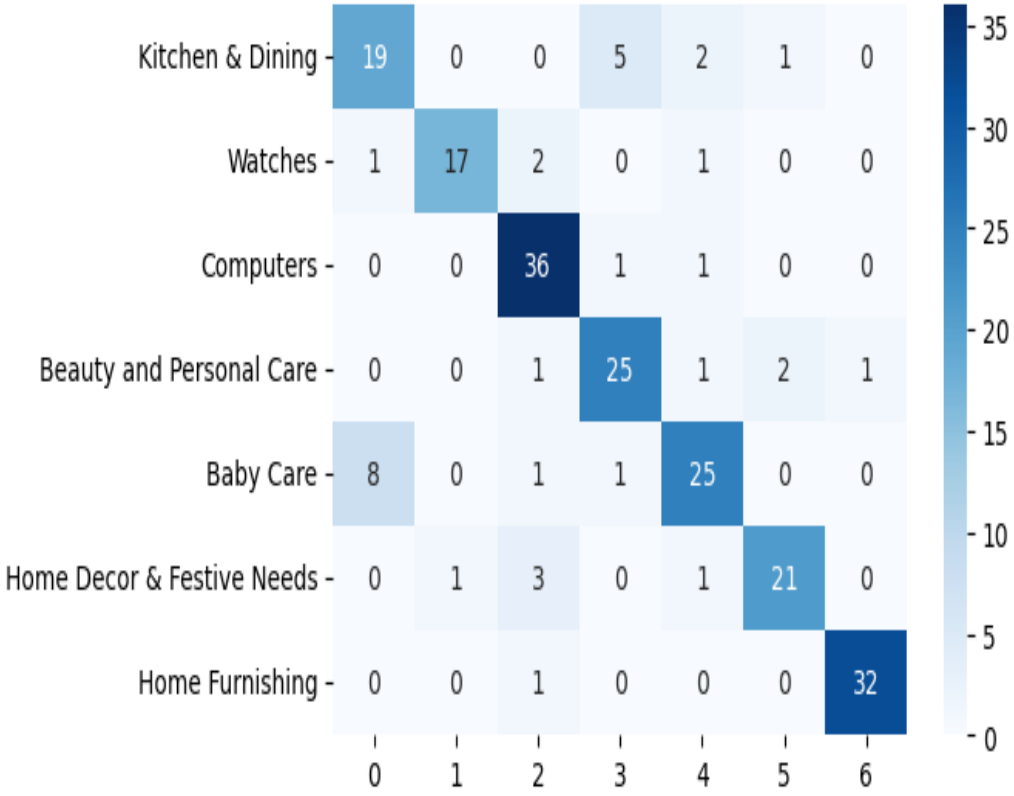
- Applique des transformations en temps réel (rotation, zoom, etc.) pour augmenter la diversité du dataset.
- Génère des images augmentées dynamiquement pour améliorer la robustesse du modèle.

Approche Sans Data Augmentation :

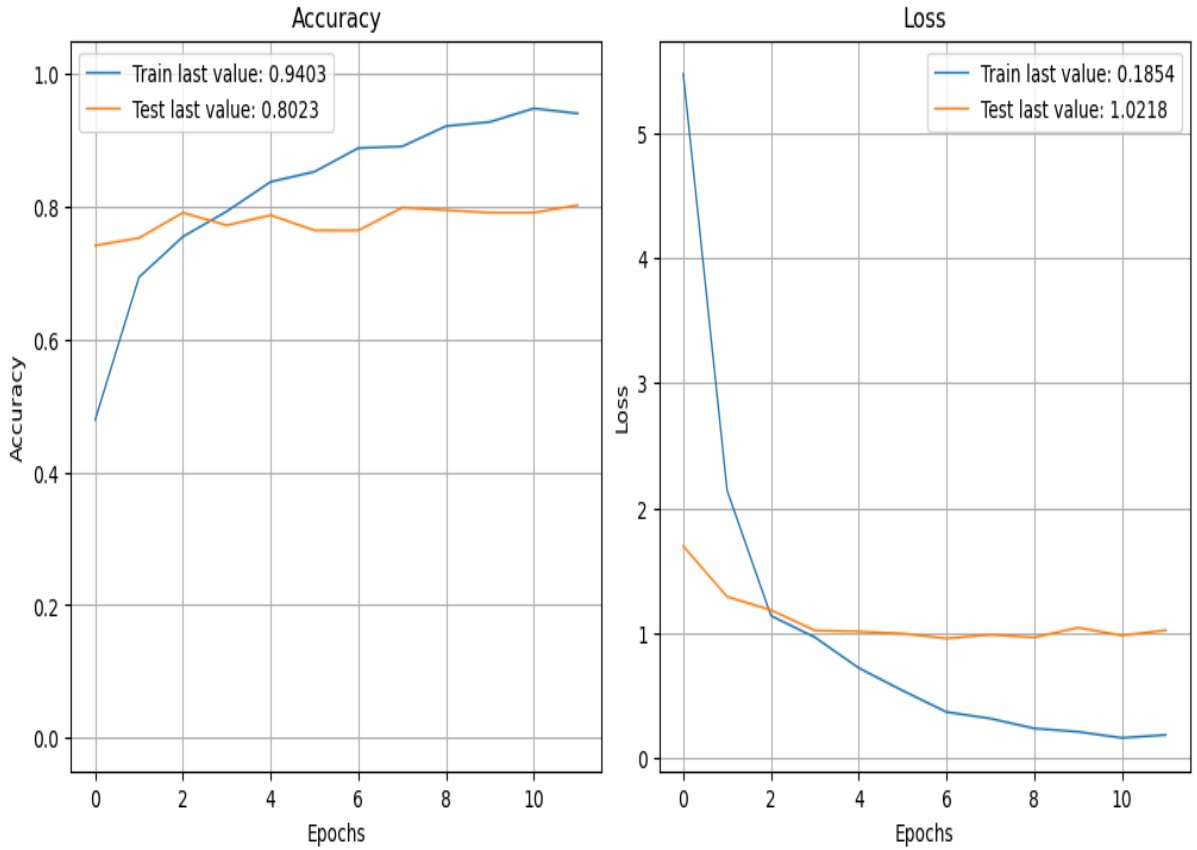
- Le modèle s'entraîne sur les images d'origine sans transformations.
- Aucune augmentation ou modification des données.

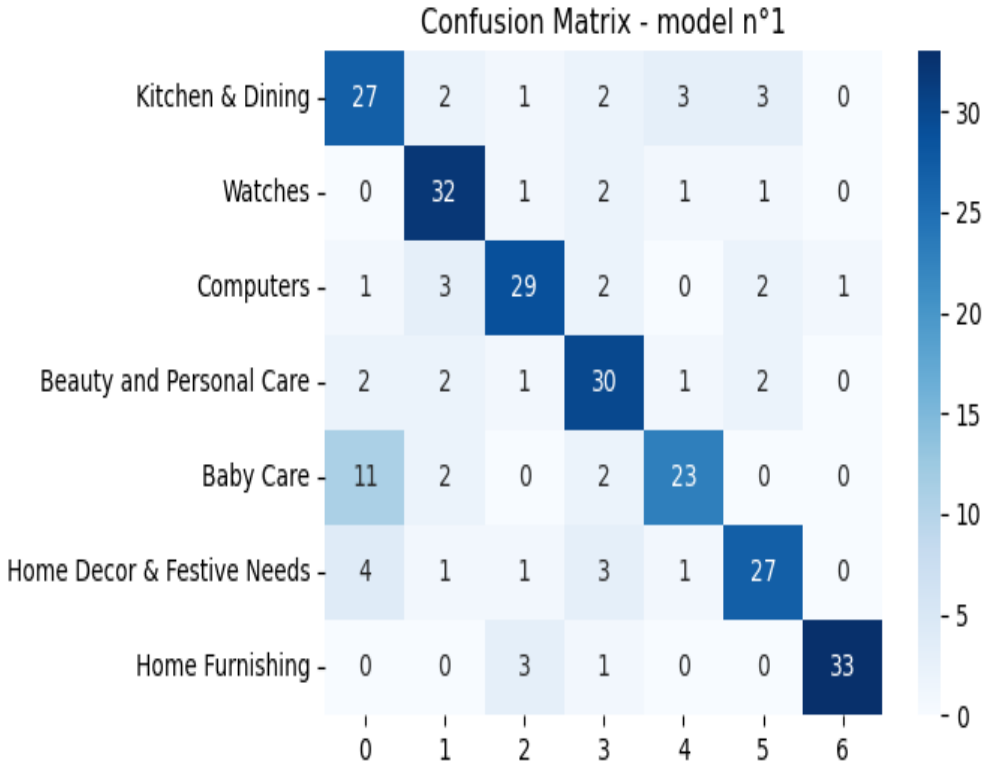
Approche avec Data Augmentation Intégrée :

- Les transformations sont intégrées directement dans le modèle pendant l'entraînement.
- Améliore la performance en limitant l'overfitting.

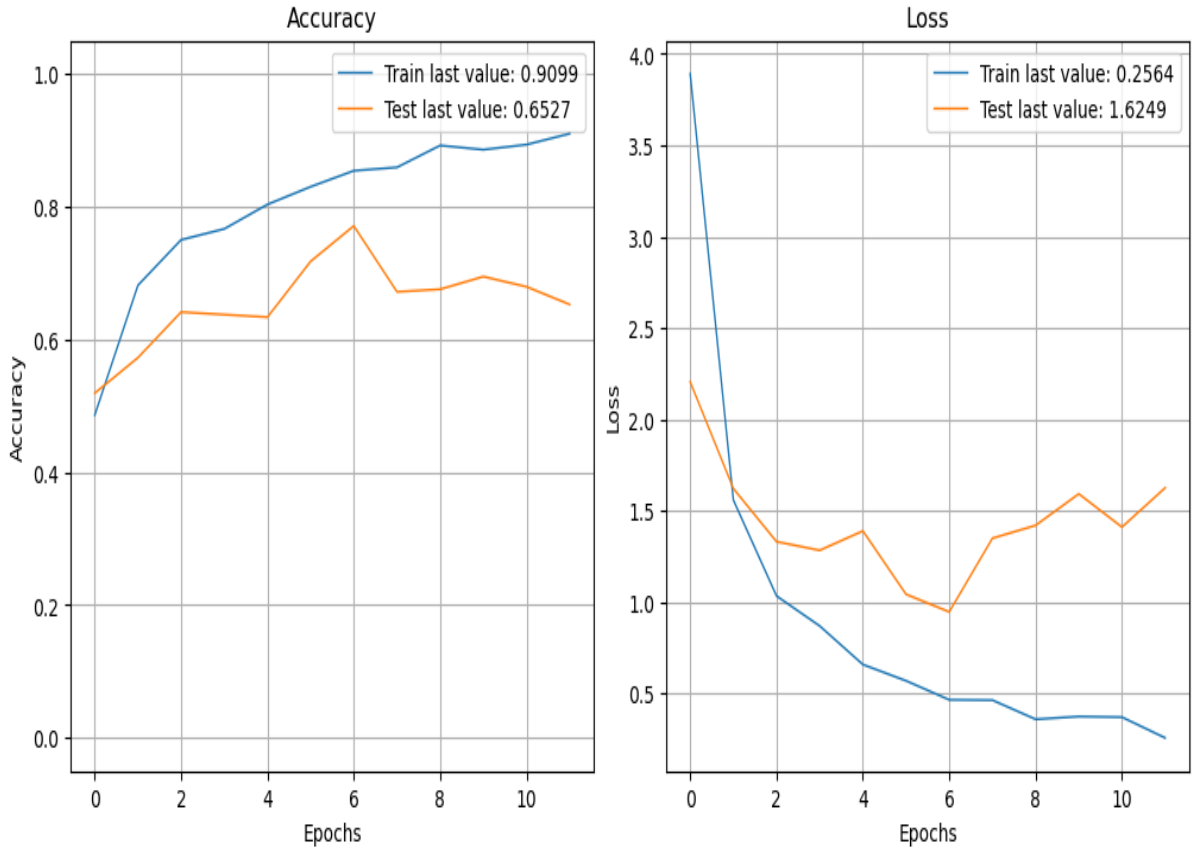


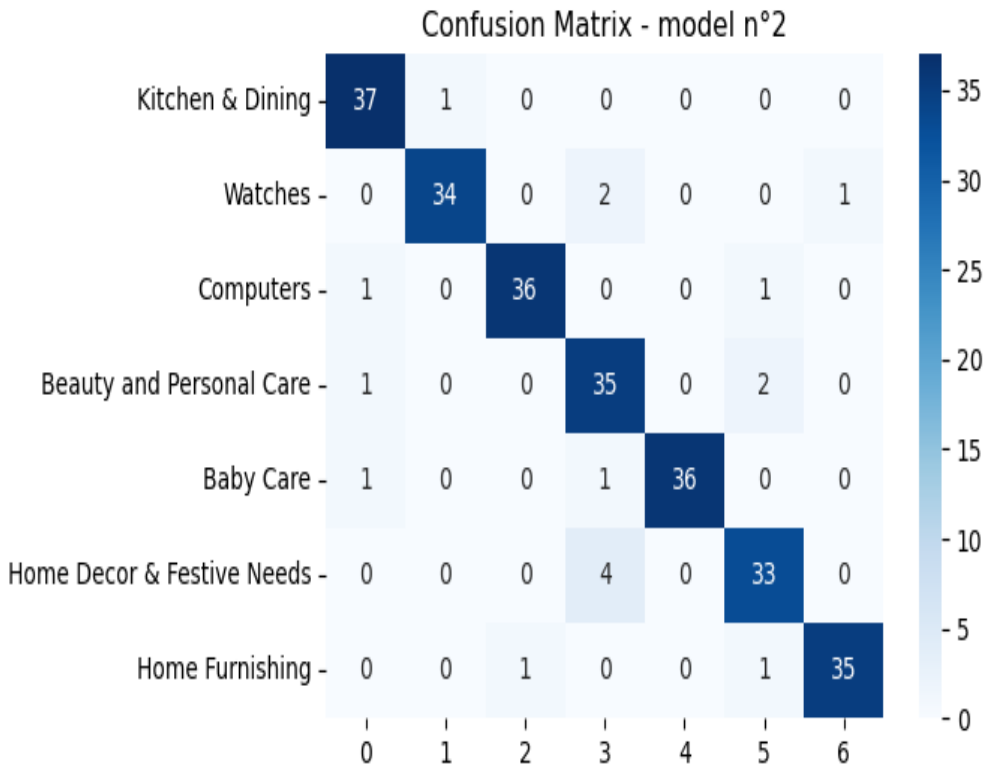
Modélisation Initiale des Images



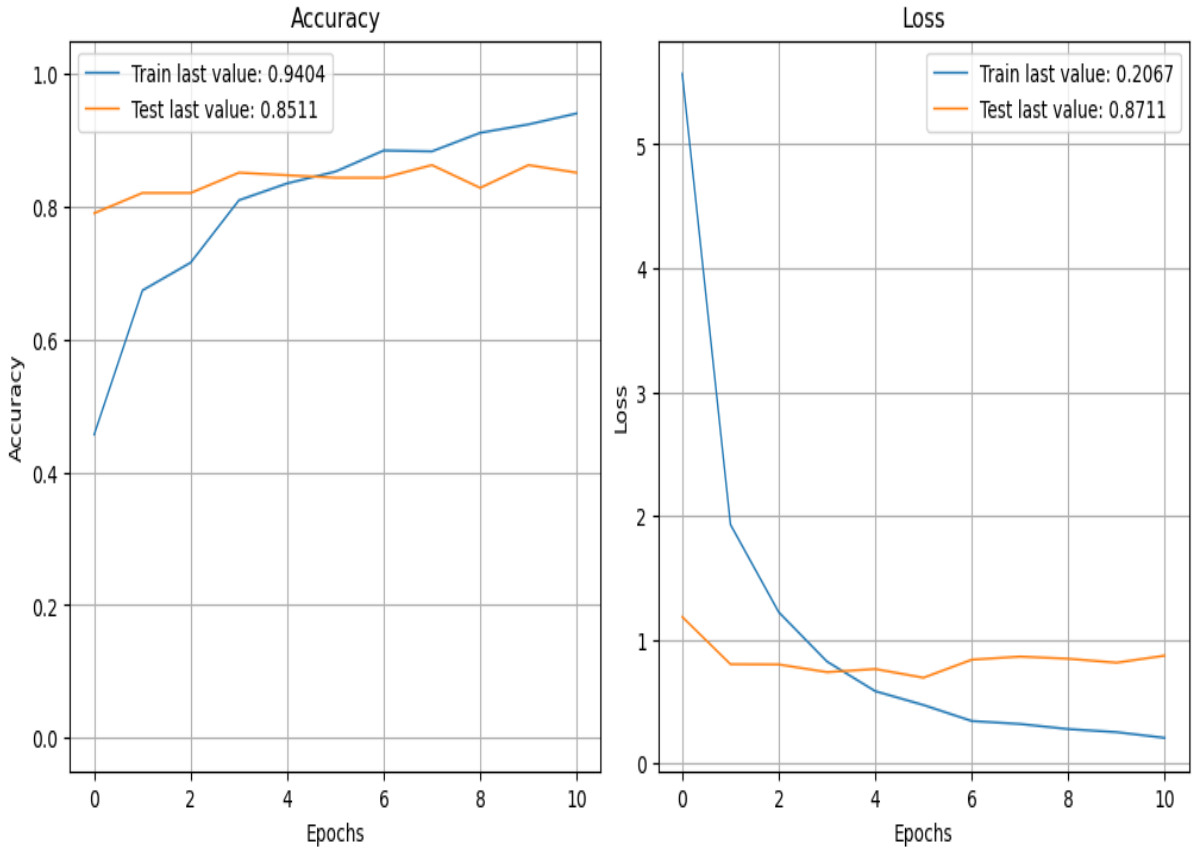


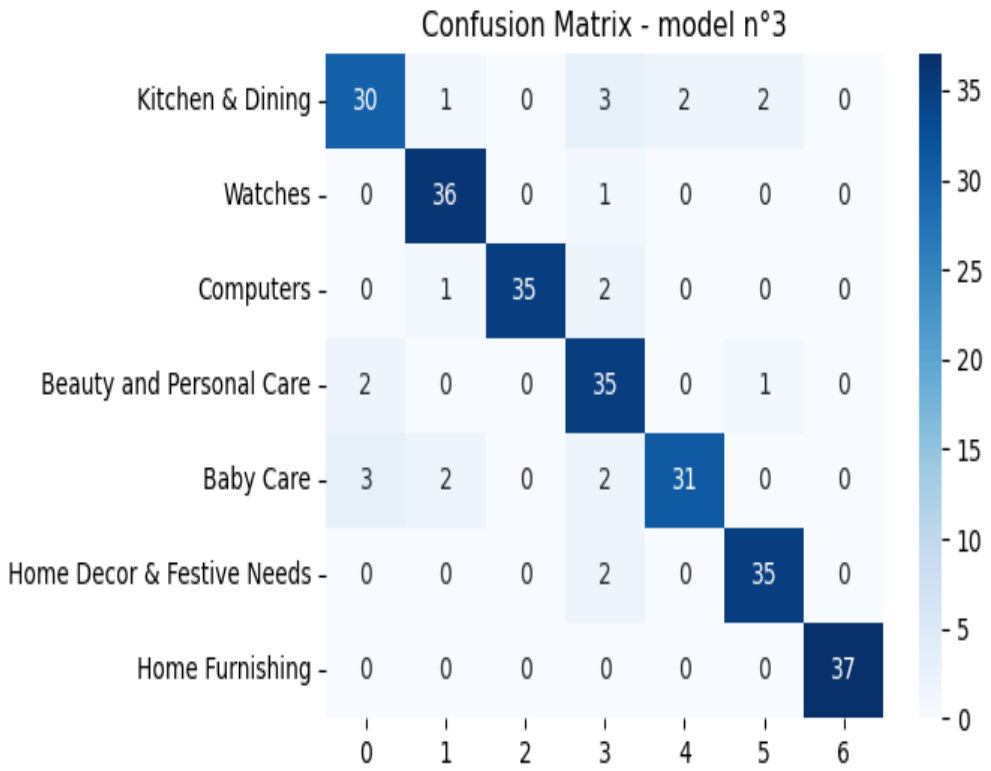
Modélisation avec Data Augmentation



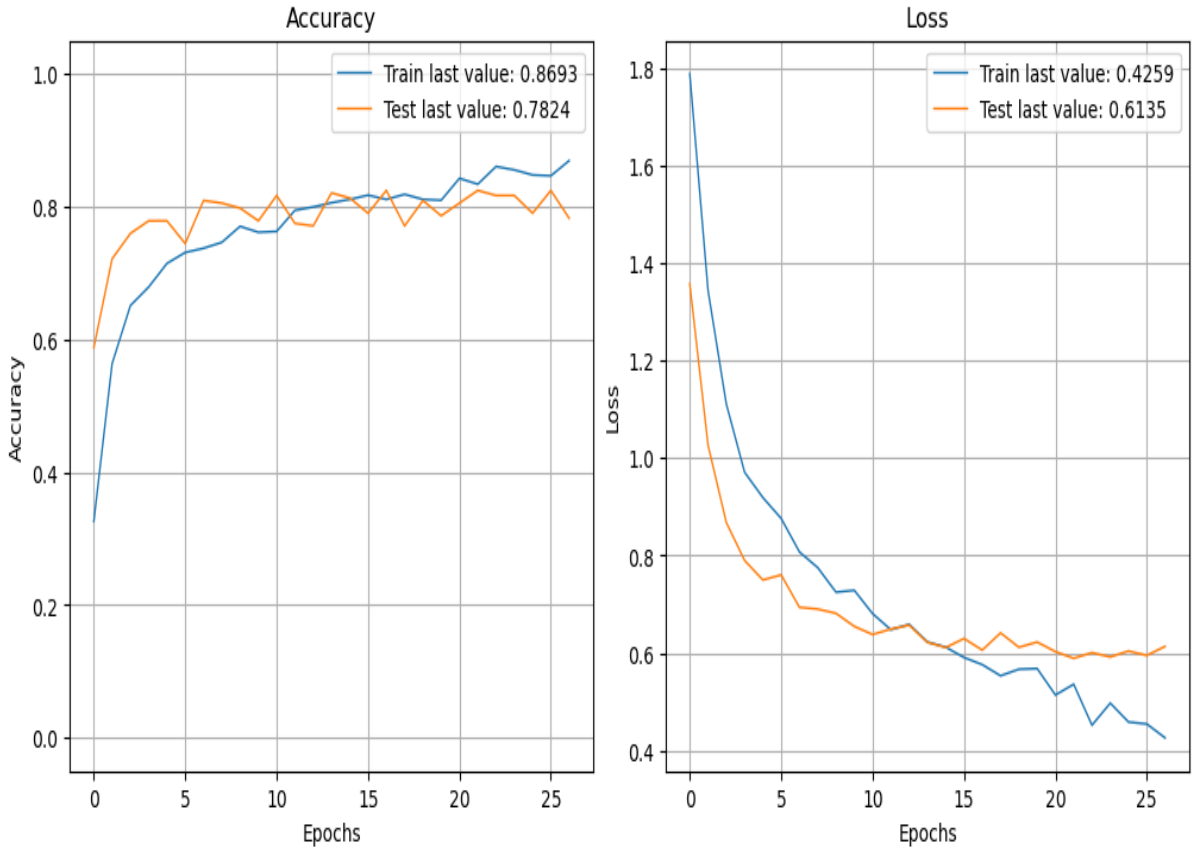


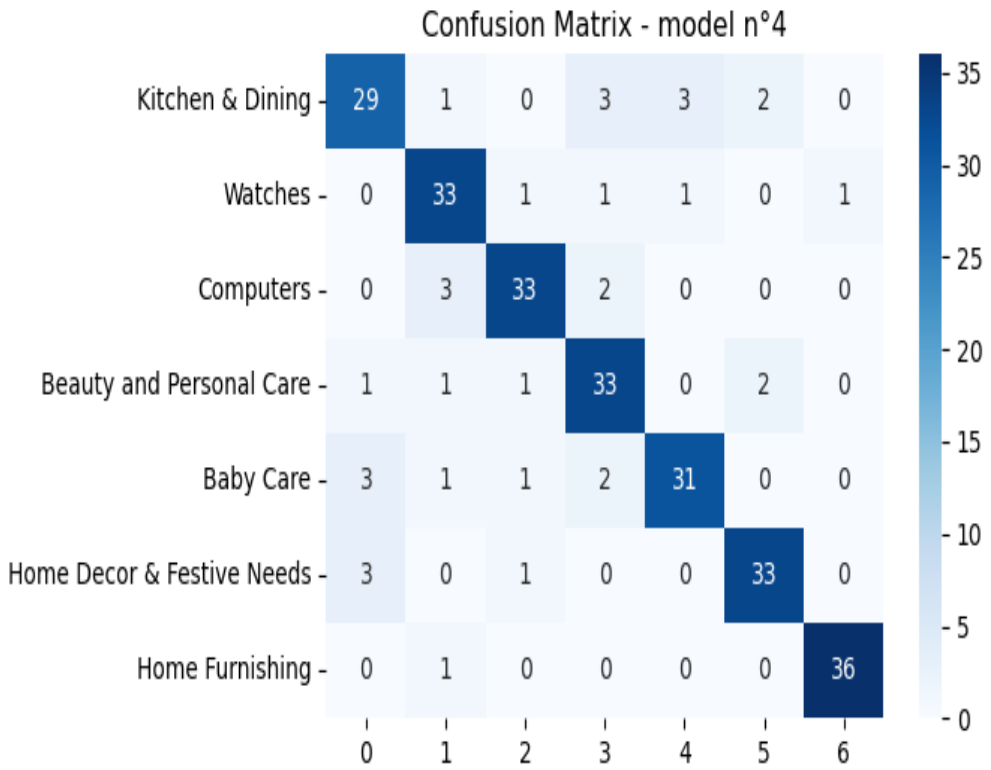
Modélisation sans Data Augmentation





Modélisation avec Data Augmentation Intégrée au Modèle





Classification Supervisée – Bilan

Modèle 1 : Préparation initiale efficace, mais faible capacité de généralisation.

Modèle 2 : Précision de validation insuffisant

Modèle 3 : Meilleurs résultats → Importance d'une structuration adéquate des données.

Modèle 4 : Performances prometteuses malgré une légère baisse en entraînement.

→ Optimiser l'apprentissage des modèles pour s'adapter aux nouvelles données, crucial pour améliorer la classification d'images.

Extension de Gamme : Collecte d'informations sur des produits à base de champagne via API, script Python pour extraire les données clés, organisées dans un fichier CSV.

Test API

foodId	label	category	foodContentsLabel	image
food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
food_b753ithamdb8psbt0w2k9aquo06	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	NaN
food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN
food_a79xmnya6togreaeukbroa0thhh0	Champagne Chicken	Generic meals	Flour; Salt; Pepper; Boneless, Skinless Chicke...	NaN

Conclusion

Méthodes textuelles : méthodes basiques → performance fiable, surpassent les méthodes avancées

Méthodes basées sur l'image : CNN → meilleure robustesse

Approches de classification supervisée : Approches privilégiées = Sans augmentation & avec augmentation intégrée au modèle

Traitement de texte basiques + Traitement d'image en Deep Learning

Merci pour votre attention