



FORMATION DATA SCIENTIST : PROJET 6



Classifiez automatiquement des des biens de consommation



Sommaire

- ▶ Introduction
- ▶ Analyse exploratoire
- ▶ Etude de faisabilité sur les descriptions
- ▶ Etude de faisabilité sur les images
- ▶ Classification supervisée des images
- ▶ Test d'une API
- ▶ Conclusion et Recommandations



1. Introduction

- **Contexte du Projet:**

- **Entreprise :** Place de marché, une marketplace e-commerce en pleine expansion.
- **Utilisateurs :** Vendeurs postant articles via photos et descriptions;
Acheteurs recherchant des produits.
- **Défi Actuel :** Catégorisation manuelle des articles par les vendeurs, entraînant des erreurs et une inefficacité croissante avec l'augmentation du volume d'articles.



- **Problématique:**

- **Fiabilité :** Classification manuelle actuelle peu fiable et subjective.
- **Scalabilité :** Avec l'augmentation prévue du volume de produits, il est crucial de simplifier et d'automatiser le processus de mise en ligne et de recherche de produits.
- **Objectif :** Étudier la faisabilité d'un moteur de classification automatique des articles en utilisant les descriptions textuelles et les images des produits pour améliorer l'expérience utilisateur.



2. Analyse exploratoire

- Les Données

Un fichier csv + un dossier d'images

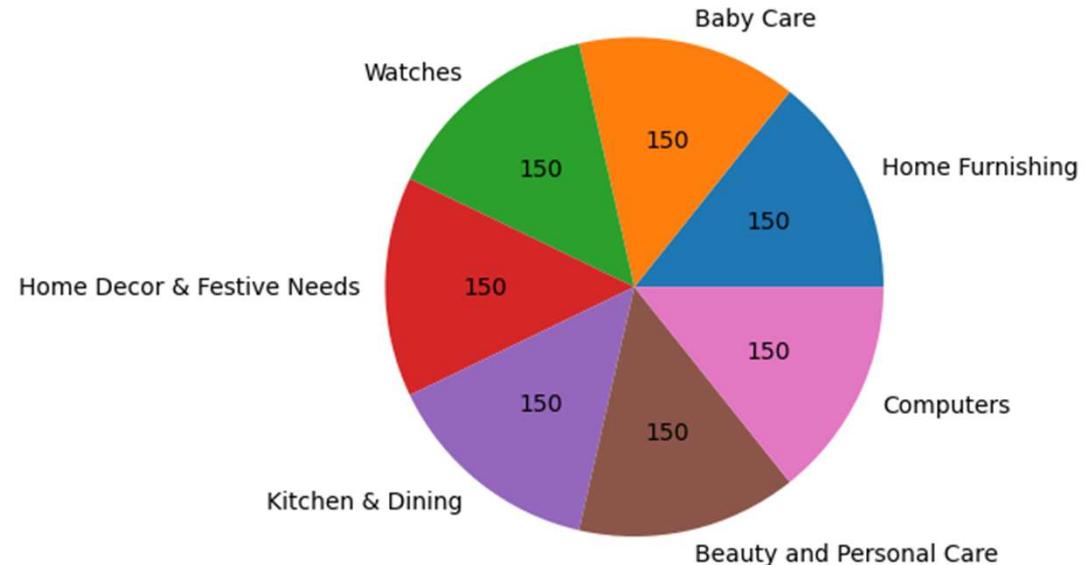
1050 lignes et images avec la colonne uniq_id qui correspond au nom des fichiers images [uniq_id].jpg

On conserve uniq_id et description

On transforme product_category_tree en category

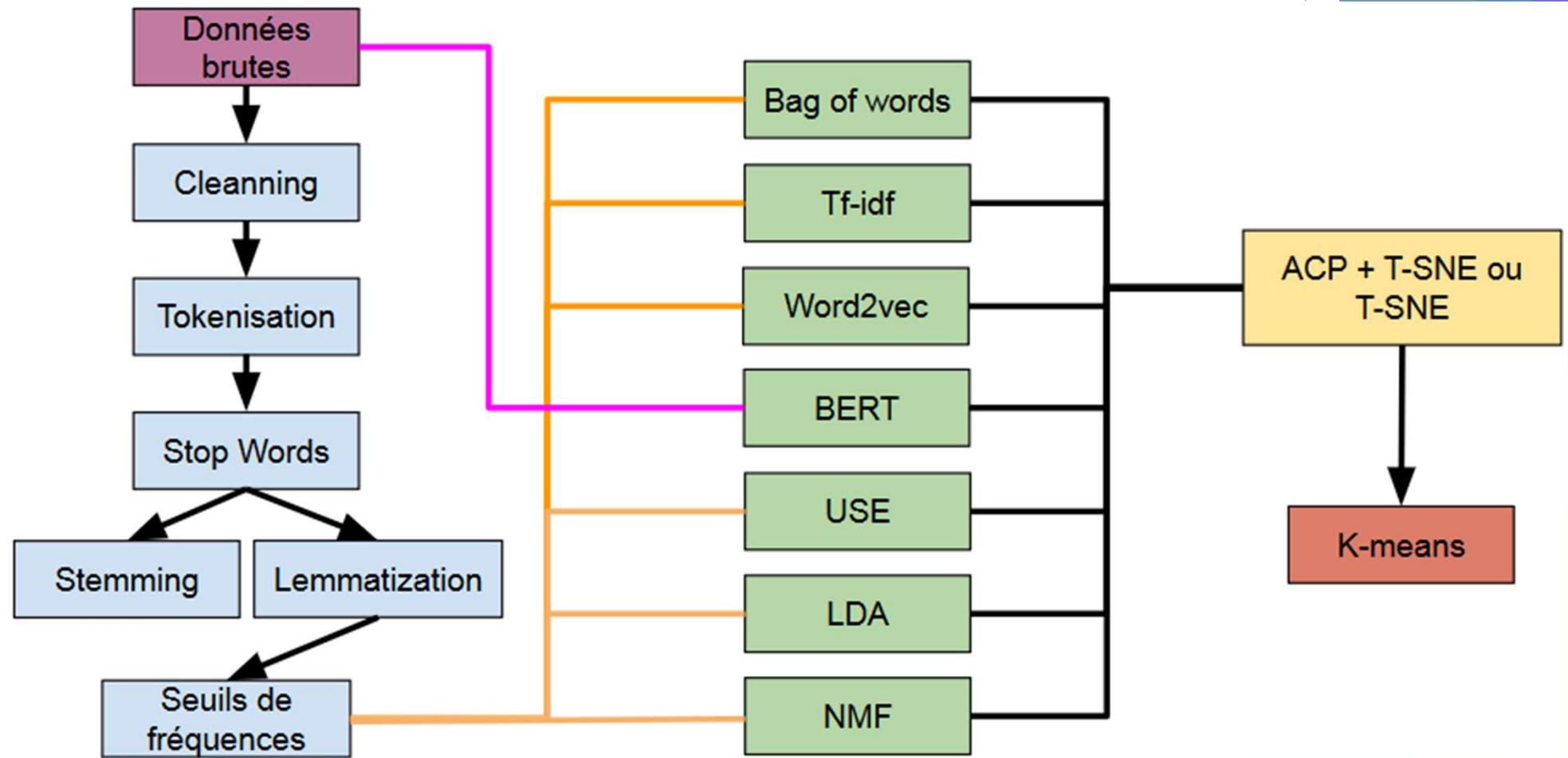
Pas de données ou d'images manquantes sur les variables choisies

- Les catégories



3. Étude de faisabilité sur les descriptions

- La méthodologie



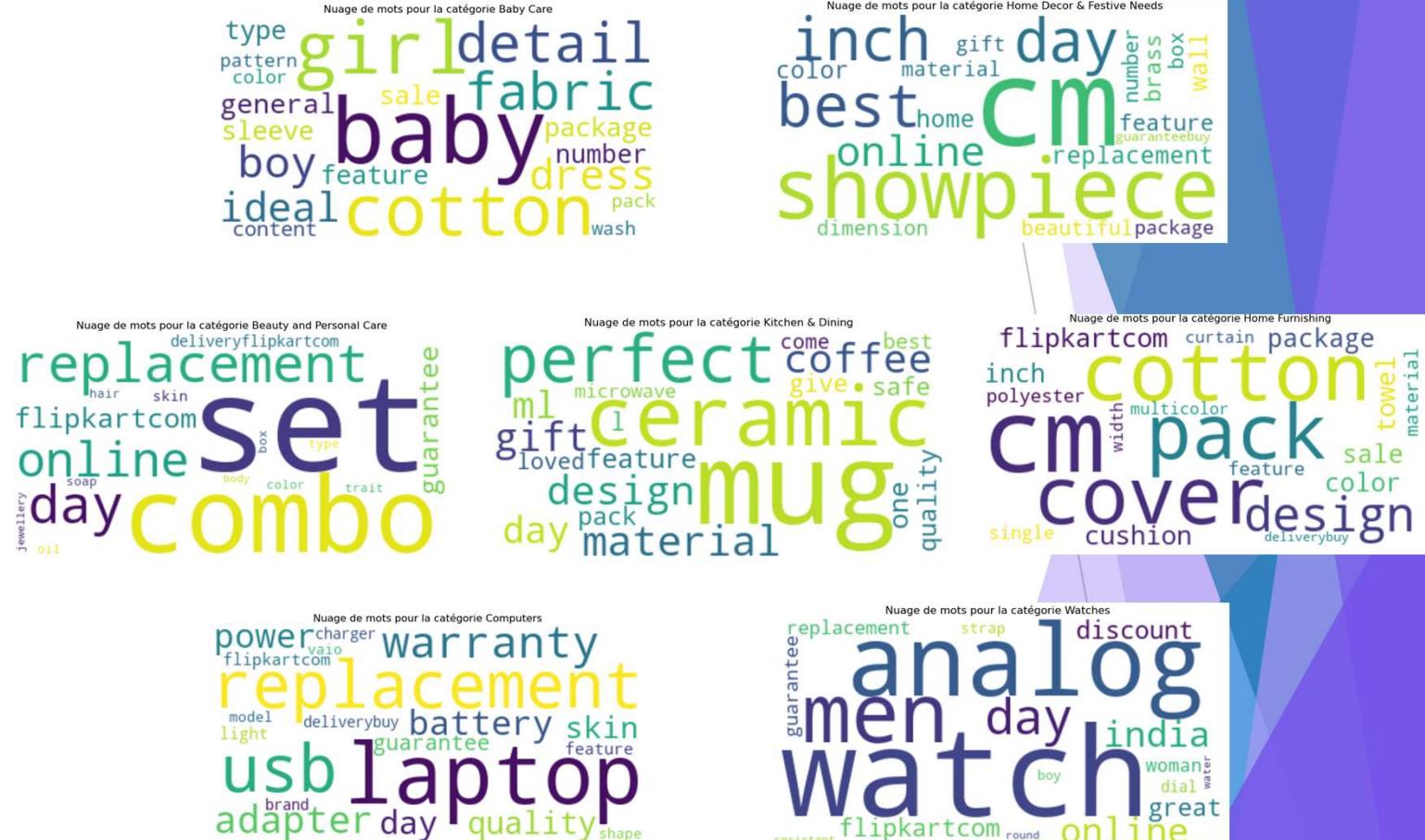
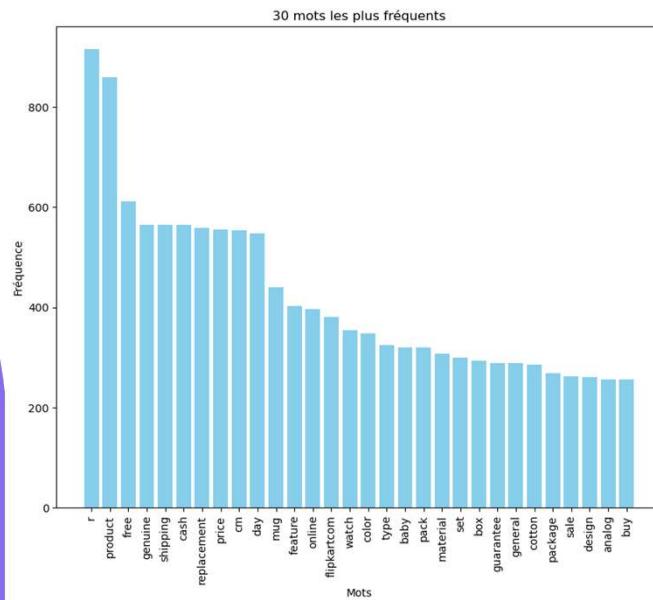
3. Étude de faisabilité sur les descriptions

- Exemple de traitement :

Donnée brute	Buy Lapguard HP Pavilion dv5-1014tx 6 Cell Laptop Battery only for Rs. 0.0 from Flipkart.com. Only Genuine Products. 30 Day Replacement Guarantee. Free Shipping. Cash On Delivery!
Nettoyage	buy lapguard hp pavilion dvtx cell laptop battery only for rs from flipkartcom only genuine products day replacement guarantee free shipping cash on delivery
Tokenisation	'buy', 'lapguard', 'hp', 'pavilion', 'dvtx', 'cell', 'laptop', 'battery', 'only', 'for', 'rs', 'from', 'flipkartcom', 'only', 'genuine', 'products', 'day', 'replacement', 'guarantee', 'free', 'shipping', 'cash', 'on', 'delivery'
Stop words	'buy', 'lapguard', 'hp', 'pavilion', 'dvtx', 'cell', 'laptop', 'battery', 'rs', 'flipkartcom', 'genuine', 'products', 'day', 'replacement', 'guarantee', 'free', 'shipping', 'cash', 'delivery'
Stemming	'buy', 'lapguard', 'hp', 'pavilion', 'dvtx', 'cell', 'laptop', 'batteri', 'rs', 'flipkartcom', 'genuin', 'product', 'day', 'replac', 'guarante', 'free', 'ship', 'cash', 'deliveri'
Lemmatization	'buy', 'lapguard', 'hp', 'pavilion', 'dvtx', 'cell', 'laptop', 'battery', 'r', 'flipkartcom', 'genuine', 'product', 'day', 'replacement', 'guarantee', 'free', 'shipping', 'cash', 'delivery'

3. Étude de faisabilité sur les descriptions

- Seuils de fréquence

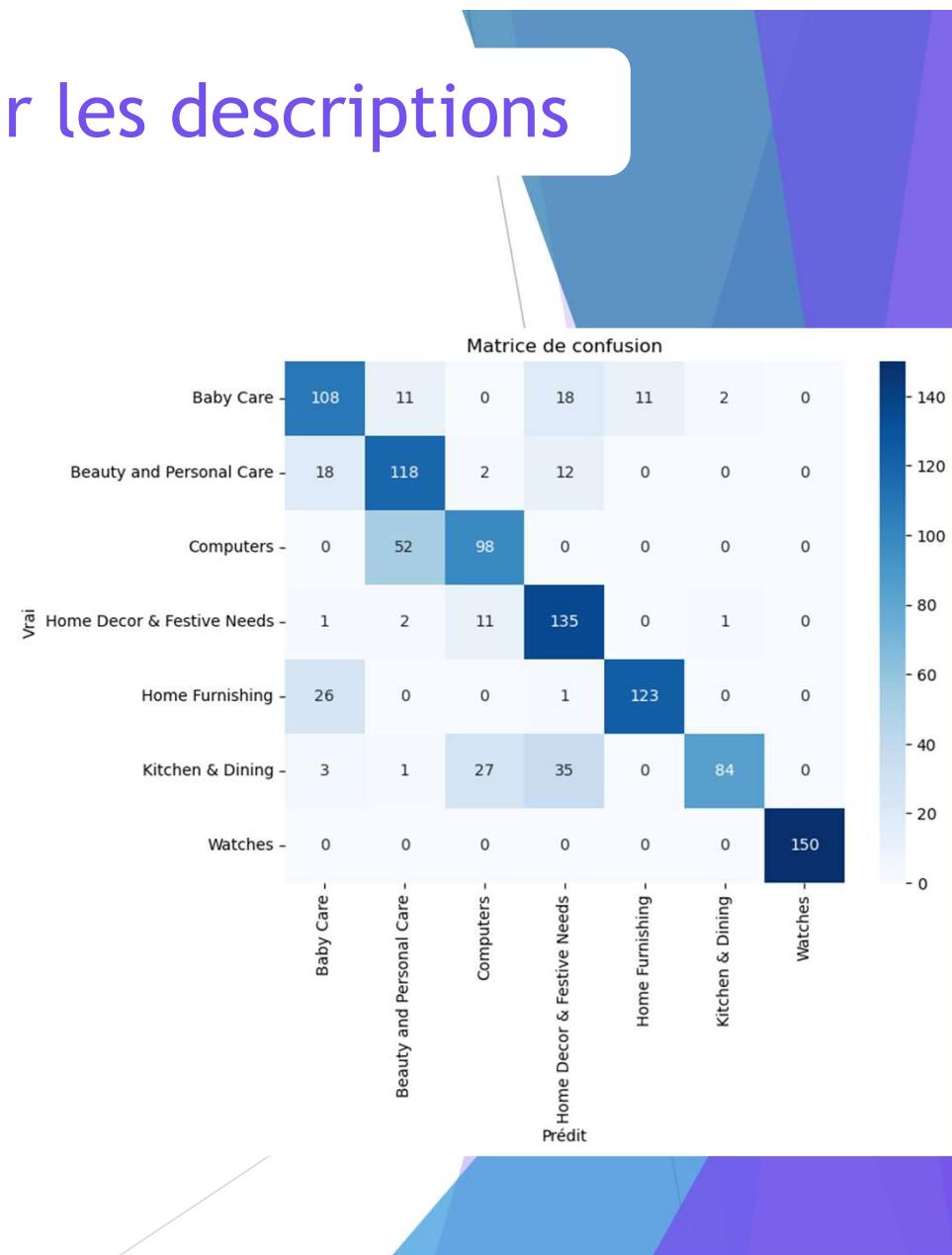
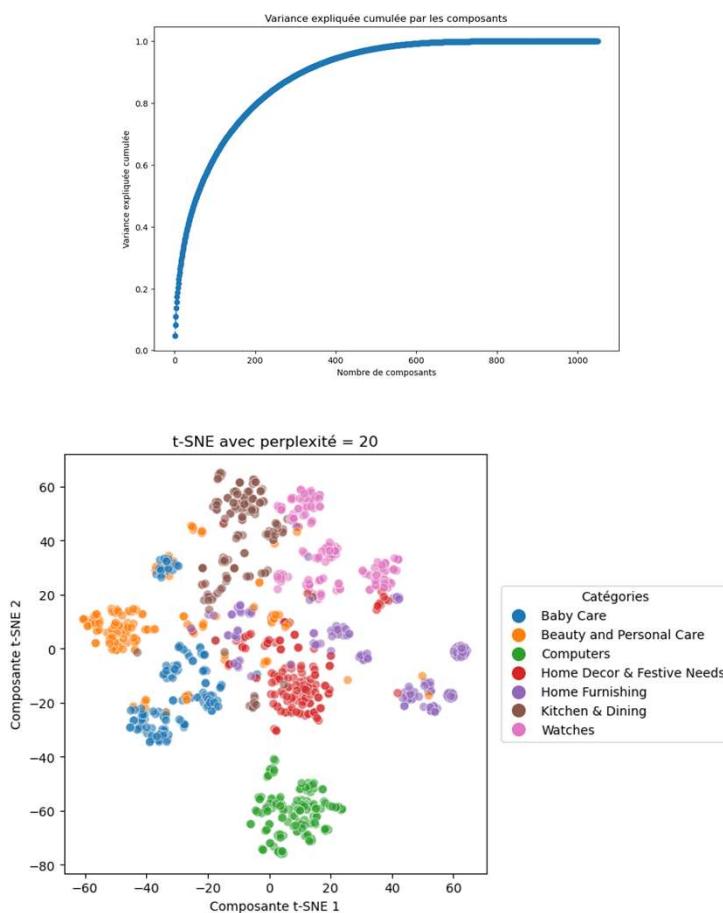


Mots avec une fréquence documentaire supérieure à 0.5 : ['price', 'r', 'product', 'free', 'shipping', 'delivery', 'genuine', 'buy', 'cash']
 Mots avec une fréquence documentaire inférieur à 0.002 : ['given', 'shrinkage', 'welcome', 'slide', 'stitch', 'apart', 'sathiyas', 'absorbency', 'softness', 'eurospa', 'feeling', 'shuvampcftsetassorted', 'exporting', 'santosh', 'gifted', 'returne', 'origional', 'waranty', 'machinewash', 'wm']

3. Étude de faisabilité sur les descriptions

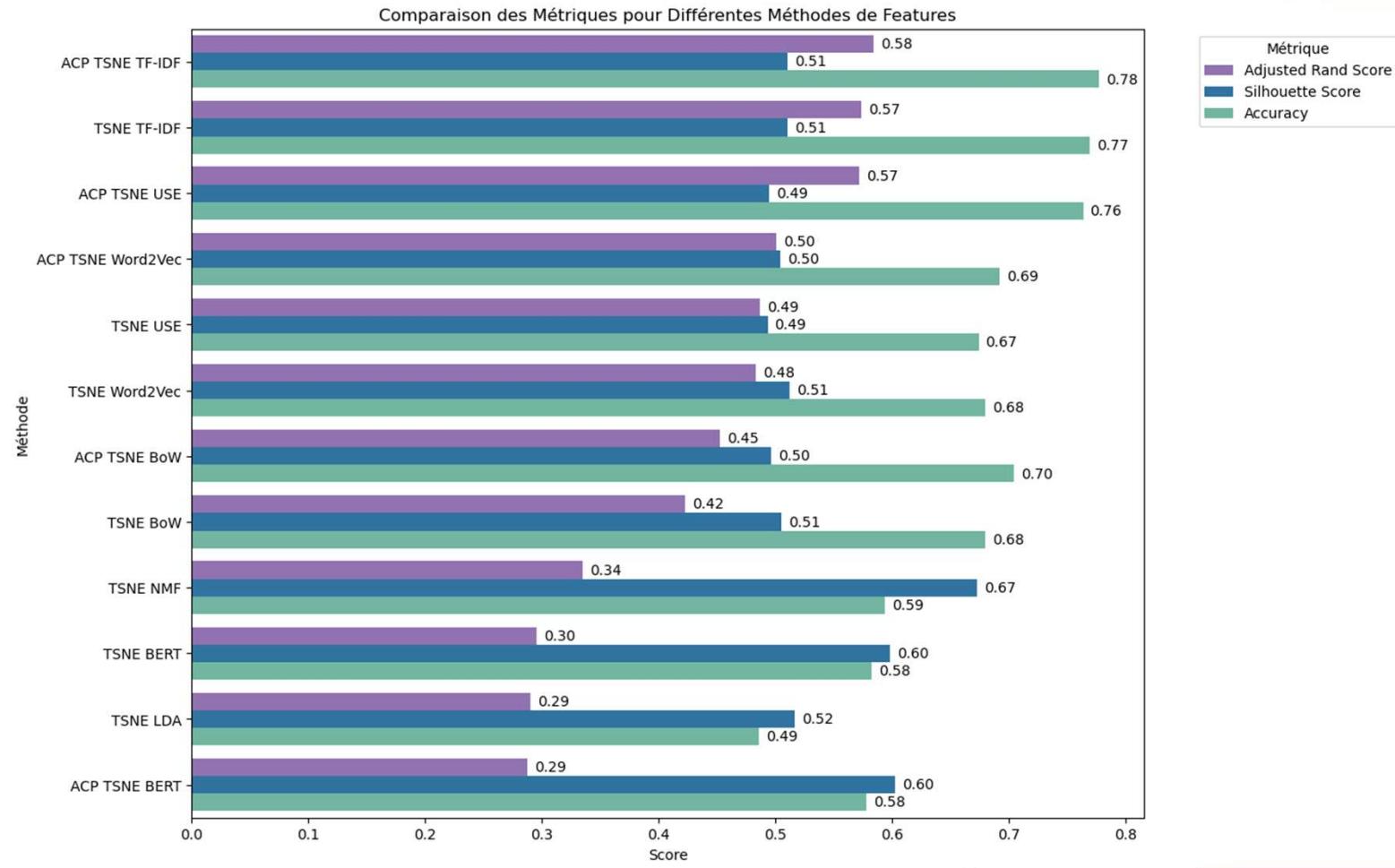
- Transformation des données

exemple avec TF-IDF



3. Étude de faisabilité sur les descriptions

- Comparaison des résultats



4. Étude de faisabilité sur les images

- Exemple d'images



4. Étude de faisabilité sur les images

- Pré-traitement avec SIFT

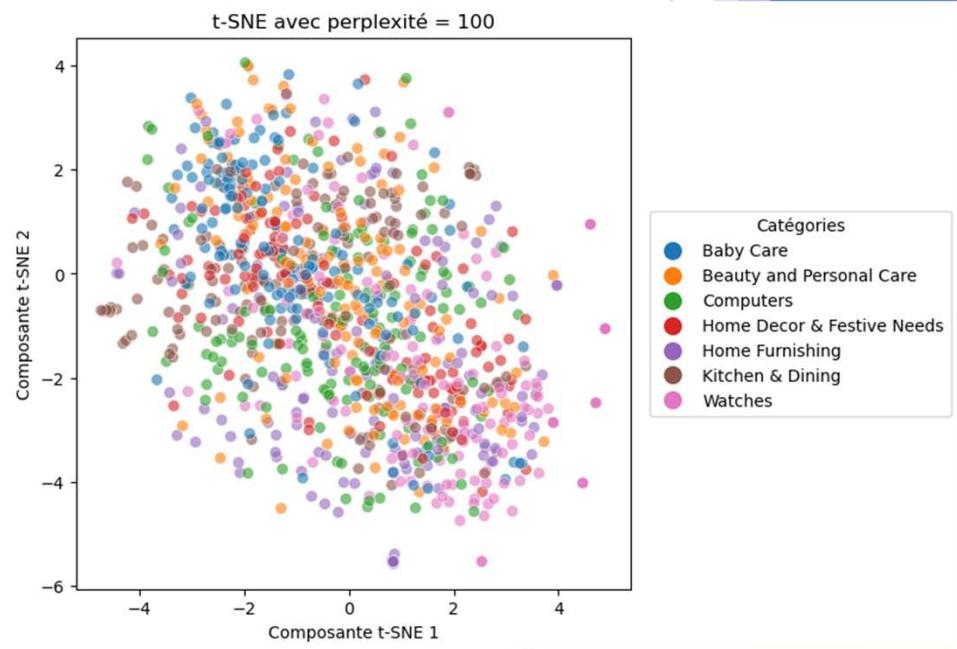
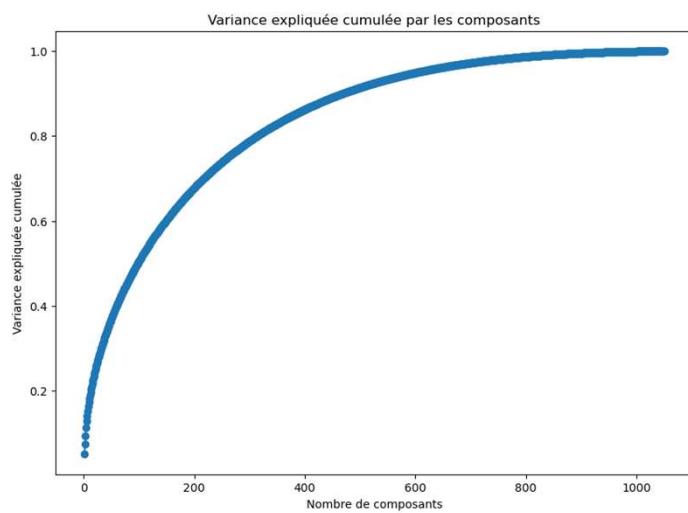
- Détermination des descripteurs

- Création des clusters de descripteurs

- Création des features

- Réduction de dimension

- Affichage 2D



4. Étude de faisabilité sur les images

- Pré-traitement avec CNN

- Basique** : Input -> Convolution 2D -> MaxPooling2D -> Convolution 2D -> MaxPooling2D -> GlobalAveragePooling2D -> Couche Dense (fully connected) -> Couche de sortie (Dense)
- Intermédiaire** : Input -> Convolution 2D -> MaxPooling2D -> Convolution 2D -> MaxPooling2D -> Convolution 2D -> MaxPooling2D -> GlobalAveragePooling2D -> Couche Dense (fully connected) -> Dropout -> Couche de sortie (Dense)

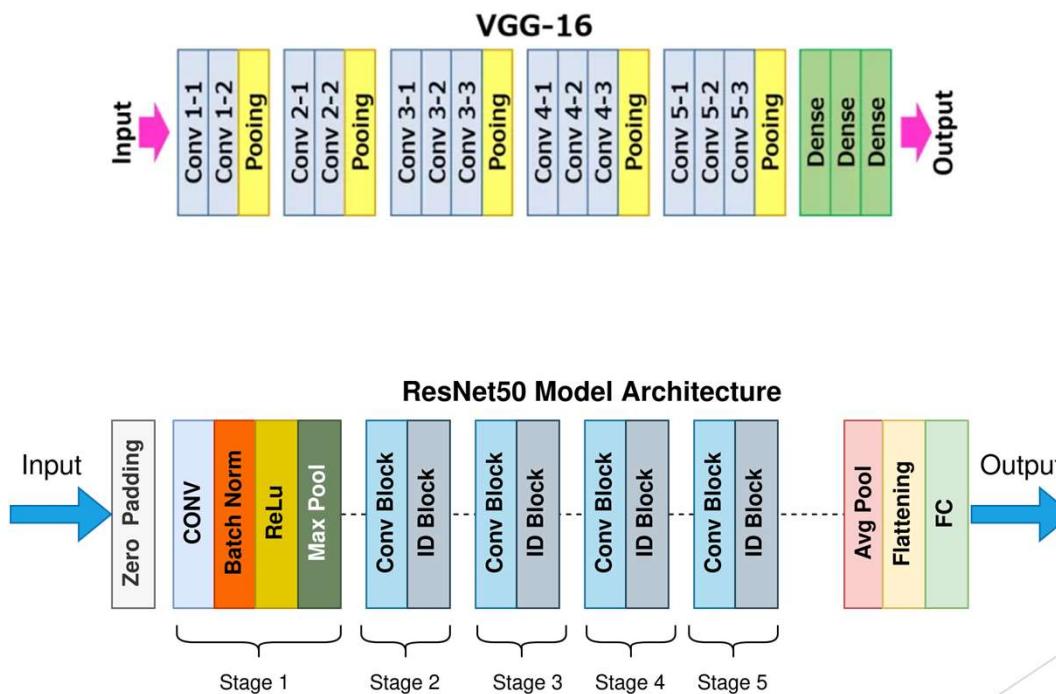
- VGG16**

- ResNet50**

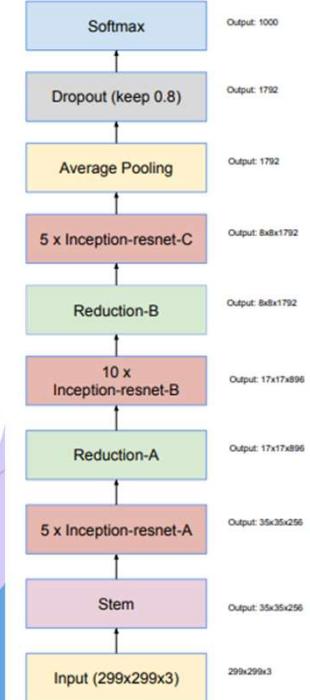
- InceptionV3**

- InceptionResNetV2**

- DenseNet201**

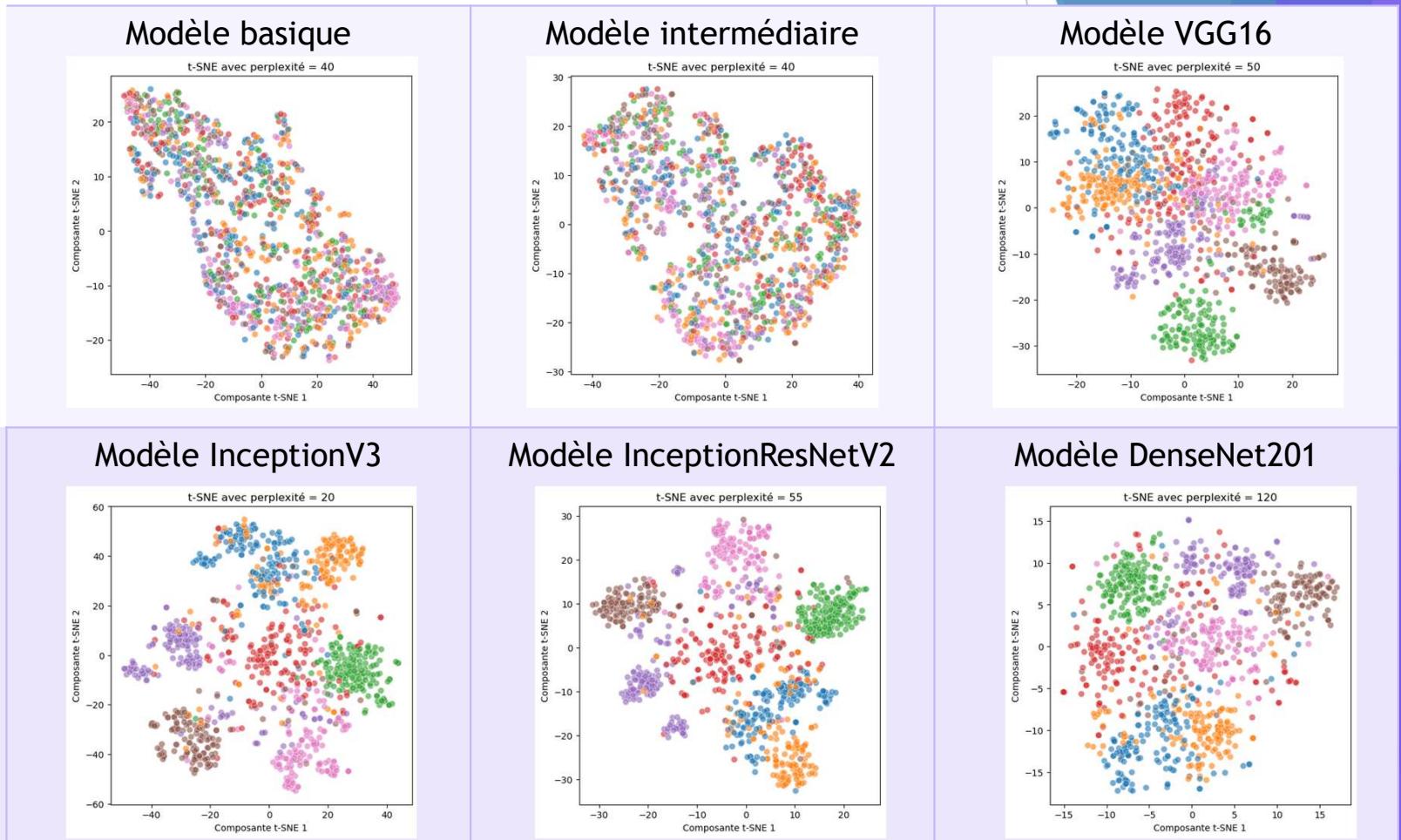


InceptionResNetV2



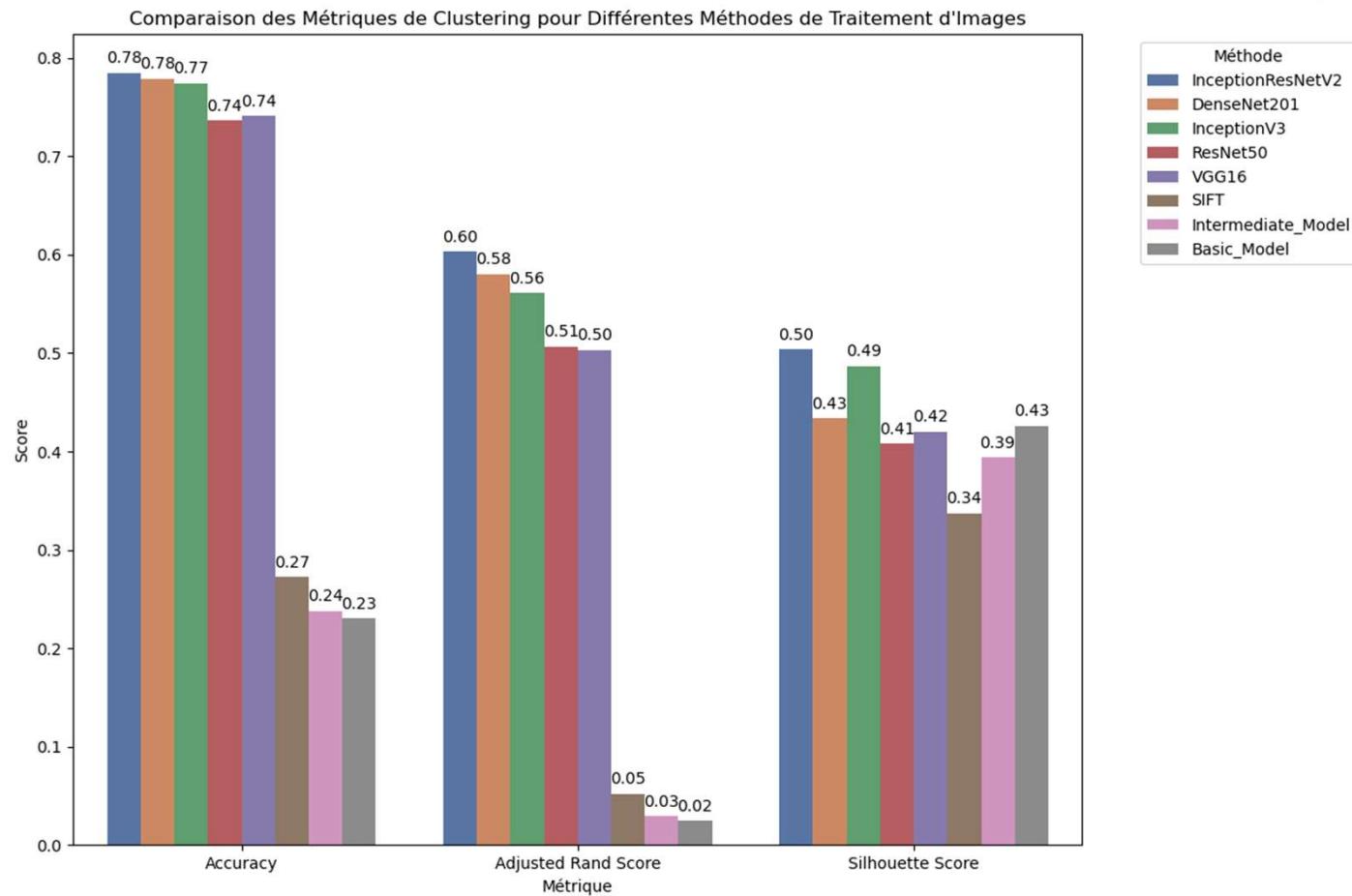
4. Étude de faisabilité sur les images

- Visualisation avec CNN



4. Étude de faisabilité sur les images

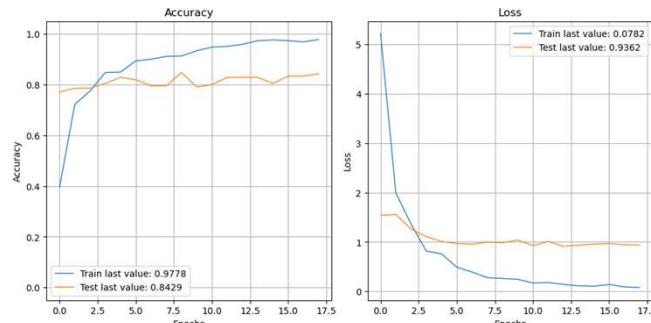
- Comparaison des résultats



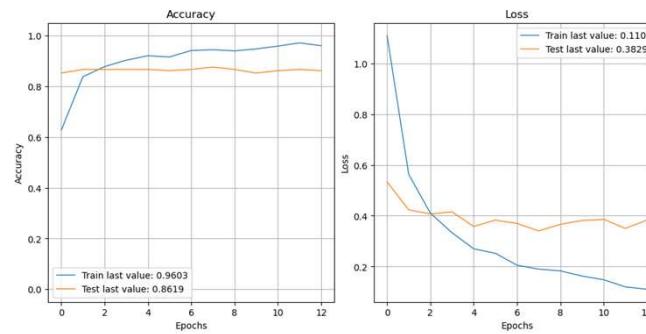
5. Classification supervisée des images

- Classification avec Transfert Learning

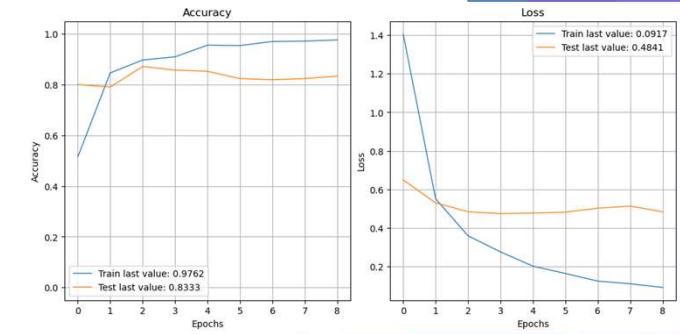
VGG16



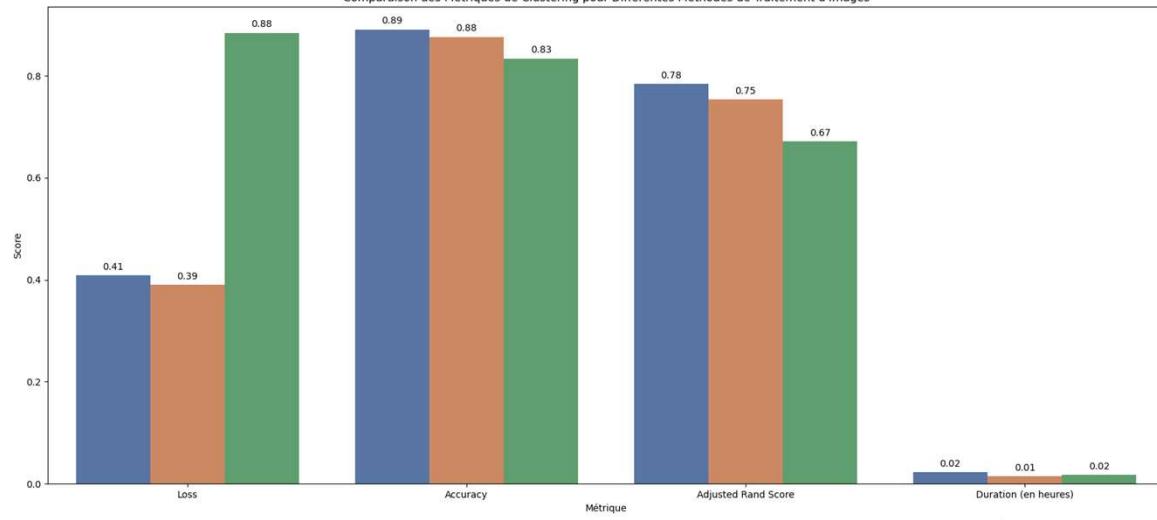
InceptionResNetV2



DenseNet201



Comparaison des Métriques de Clustering pour Différentes Méthodes de Traitement d'Images



5. Classification supervisée des images

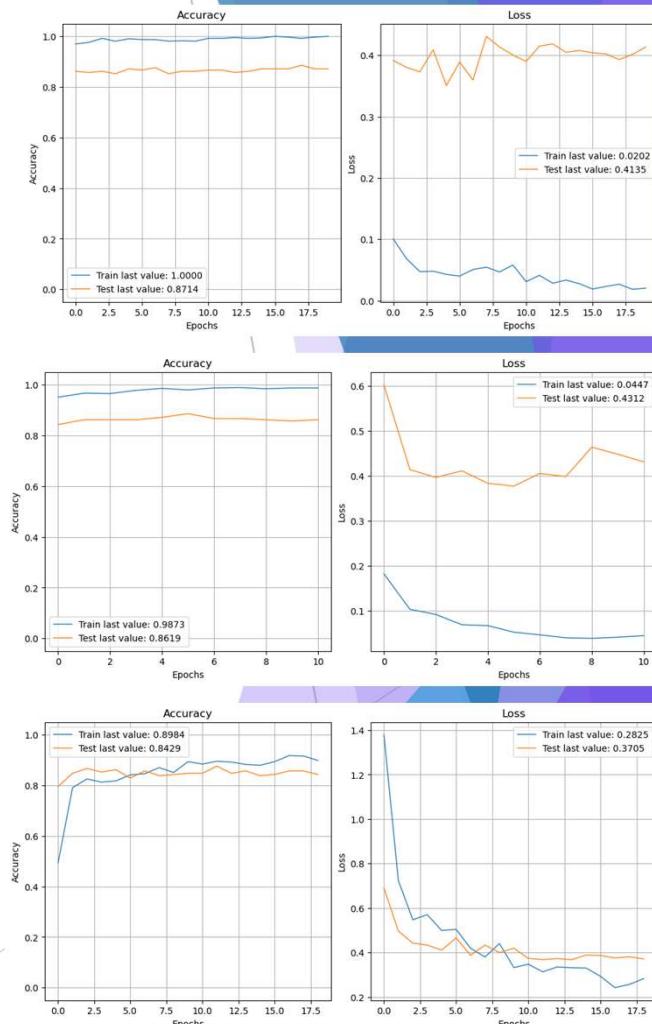
- Recherche des Hyperparamètres

Sur le modèle de base InceptionResNetV2

Hyperparamètres testés : 3 taux d'apprentissage (0.00075, 0.0005, 0.0001) et deux tailles de lot (32, 64)

Durée d'entraînement fixée à 50 époques mais avec EarlyStopping à 15

Reduction du taux d'apprentissage déclenché à 5



- Data augmentation

- **ImageDataGenerator** : rotation, décalage horizontal et vertical, zoom, inclinaison et normalisation des valeurs de pixels + preprocess_inceptionresnetv2 sur le jeu d'entraînement

- **Sequential** : Sur le modèle de base InceptionResNetV2 avec ajout d'une séquence de transformation des données flip vertical, rotation aléatoire, zoom aléatoire

5. Classification supervisée des images

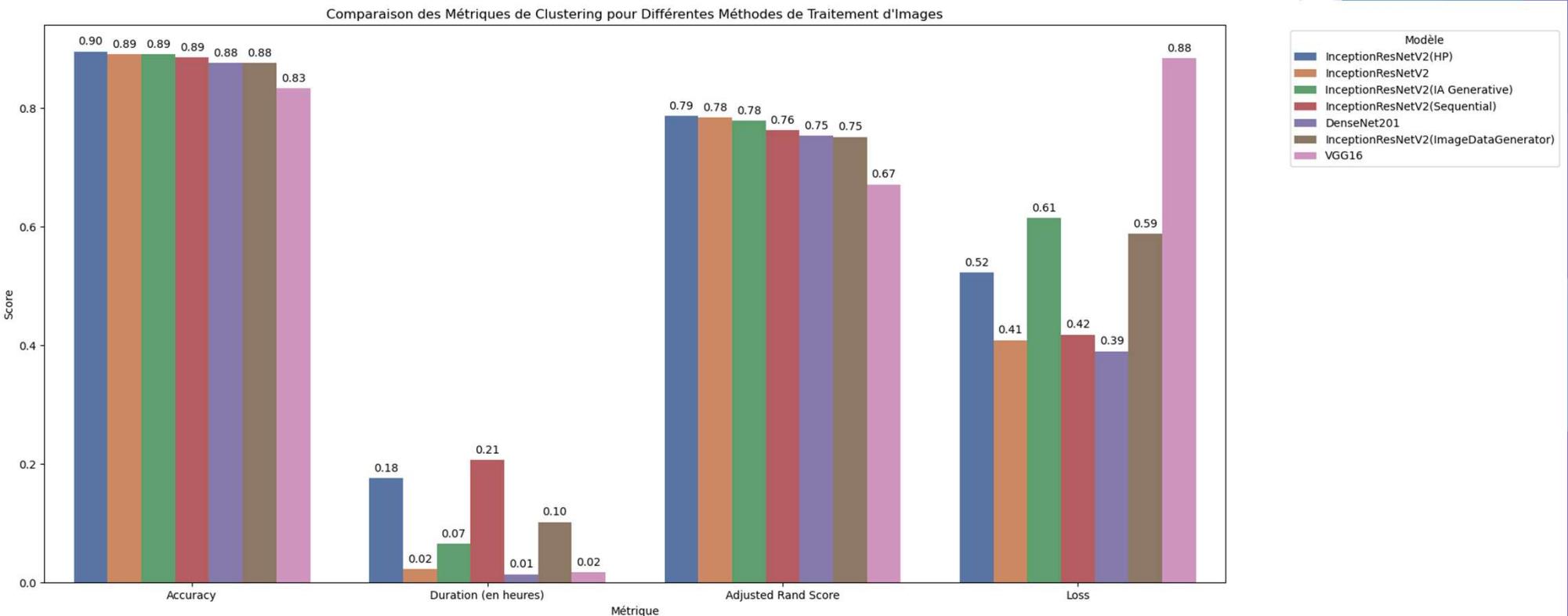
- IA Generative

Génération de 18 x 7 images avec CoPilot
Modèle InceptionResNetV2



5. Classification supervisée des images

- Comparaison des résultats



6. Test de l'API

- Collecte de données via API

- **Objectif** : Élargissement de notre gamme à l'épicerie fine, spécifiquement les produits à base de champagne.
- **API Utilisée** : Edamam Food and Grocery Database.
- **Méthode** : Requête API pour filtrer et collecter des informations sur les produits à base de champagne.
- **Filtre Appliqué** : Recherche par ingrédient spécifique "champagne".
- **Données Collectées** : foodId, label, category, foodContentsLabel, image.

```
def fetch_champagne_products(api_key):  
    url = "https://edamam-food-and-grocery-database.p.rapidapi.com/api/food-database/v2/parser"  
    query = {  
        "ingr": "champagne"  
    }  
    headers = {  
        'x-rapidapi-host': "edamam-food-and-grocery-database.p.rapidapi.com",  
        'x-rapidapi-key': api_key  
    }  
  
    response = requests.get(url, headers=headers, params=query)  
    if response.status_code == 200:  
        return response.json()  
    else:  
        return None
```

6. Test de l'API

- Test sur les produits à base de champagne

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods		https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	
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...	
4	food_an4jueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	
5	food_bmu5dmkazwvpaa5prh1daa8jxs0	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;...	
7	food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	
8	food_am5egz6aq3fpjlaf8xpdkbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	
9	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	

Les données ont été enregistrées dans un fichier CSV, prêtes pour analyses ultérieures.

6. Test de l'API

• Principes Clés du RGPD

Il s'applique spécifiquement à la collecte, au stockage, au traitement et à la gestion des données personnelles.

- **Licéité, Loyauté et Transparence**
Traitement légal, honnête et transparent des données personnelles.
 - **Limitation des Finalités**
Collecte des données uniquement pour des finalités explicites et légitimes.
 - **Minimisation des Données**
Seules les données nécessaires à la finalité prévue sont collectées et traitées.
 - **Exactitude**
Les données doivent être exactes, à jour et corrigées si nécessaire.
 - **Limitation de la Conservation**
Conservation des données pour une durée limitée, spécifique à la finalité de la collecte.
 - **Intégrité et Confidentialité**
Protection des données contre l'accès non autorisé et les risques de dommage, perte ou destruction.
-
- **Dans notre cas :**
Collecte limitée aux informations nécessaires, aucune donnée personnelle sensible n'est traitée.



Conclusion et Recommendations

- **Synthèse des résultats**

- **Classification Textuelle** : Les techniques TF-IDF et USE ont montré un excellent potentiel pour la classification.
- **Classification d'images** : Modèles CNN et notamment InceptionResNetV2 a montré une haute précision qui n'a pas été nettement amélioré ni par la recherche des hyperparamètres ni par les différentes techniques de data augmentation testées.
- **Test de l'API** : Intégration réussie

- **Recommendations**

- **Amélioration Continue** : Poursuivre le raffinement des modèles de classification textuelle et visuelle.
- **Suivi des Performances**: Mettre en place un monitoring régulier des KPIs pour ajuster les stratégies en temps réel..

- **Prochaines Étapes**

- **Combiner les classifications textuelle et image** pour améliorer les performances
- **Tests en Conditions Réelles** : Déployer les modèles sur un échantillon plus large de la plateforme pour tester l'efficacité en production.
- **Analyse de l'Impact sur les Ventes et l'Engagement** : Évaluer comment les améliorations impactent les ventes et l'engagement des utilisateurs.
- **Révision Basée sur le Feedback** : Intégrer les retours des utilisateurs pour optimiser les modèles et les processus.



Merci pour votre attention

Des questions ?

