Classification automatique des biens de consommation

OPENCLASSROOMS

Sommaire

Place de marché est une entreprise qui souhaite lancer une place de marché e-commerce.
La classification manuelle des articles par les vendeurs est peu fiable et limitée.

En tant que Data Scientist, je dois évaluer la faisabilité d'un système de classification automatique des produits, à l'aide de photos et de descriptions.

Dans cette présentation, nous aborderons :

1. **L'étude de faisabilité d'un moteur de catégorisation automatique des articles (à partir de leurs descriptions puis à partir de leurs images)**

2. **Les différentes approches de modélisation pour la classification automatique des articles à partir de leurs images**

3. **Les résultats des tests de collecte d'articles à base de champagne via l'API *Edamam Food and Grocery Database***

<u>NB</u> : Les images et descriptions ainsi que toutes les données utilisées ici sont des données publiques, non soumises à des droits de propriété intellectuelle.

Classification automatique des articles à partir de leurs descriptions

—

Faisabilité

Description & nettoyage des données textuelles

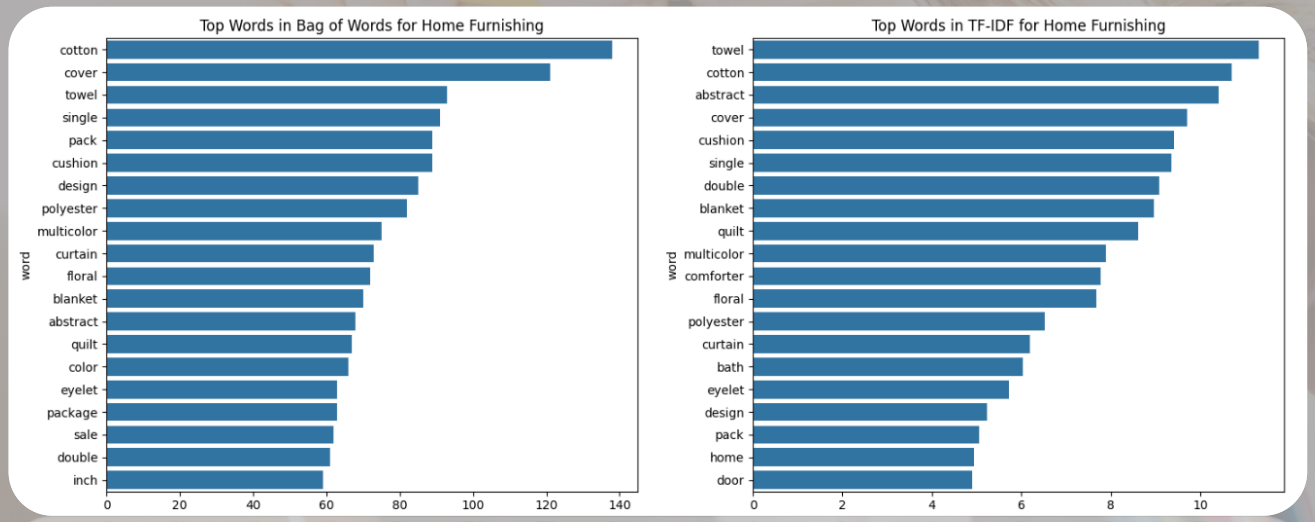
- Nous avons réuni les valeurs des variables product_name et description pour faciliter le nettoyage et la vectorisation des données textuelles.
- Par ailleurs, la variable product_category_tree a été nettoyée afin de ne conserver que la catégorie générale du produit : **nous dénombrons 7 catégories uniques de produits.**
- Nous avons procédé à la tokenisation des données, suivi par un nettoyage des tokens obtenus, afin de pouvoir évaluer par la suite la performance de la lemmatisation par rapport à celle de la stemmisation : **la lemmatisation surpasse la stemmisation en termes de précision grâce à son analyse contextuelle. Dans le domaine du traitement du langage naturel, où le contexte et la précision sont cruciaux, la lemmatisation est généralement privilégiée, tandis que la stemmisation peut convenir à des applications de recherche et de filtrage plus élémentaires.**
- Nous avons recensé puis supprimé le top 10 des lemmes les plus présents sur l'ensemble des catégories (ex : *« shipping »*, *« delivery »*, *« cash »*, etc.)
- Nous constatons un écart type relativement important de 44 lemmes sur la taille des descriptions. **Par ailleurs, l'écart entre la médiane de 25 lemmes et la moyenne de 43 à 44 lemmes relève une distribution asymétrique du nombre de lemmes par description.**
- En ce qui concerne la diversité des lemmes, une moyenne de 0.51 suggère que **les descriptions de produits utilisent une bonne gamme de vocabulaire, sans trop de répétitions excessives.**

  Par ailleurs, une faible variabilité (écart type de 0.09) indique une cohérence dans l'utilisation des mots entre les produits.

Comparaison des approches BoW et TF-IDF

Analyse de texte : Bag of words vs TF-IDF

Le Bag of Words capture les fréquences élevées de termes clés, tandis que le TF-IDF offre des scores plus modérés pour ces mêmes termes, illustrant la différence d'approche dans la pondération de l'importance des mots.



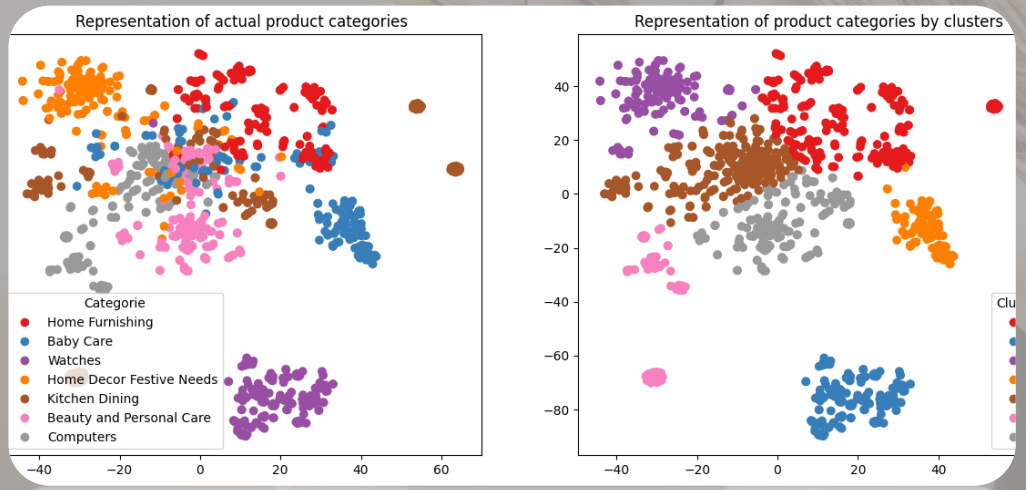
En BoW, « *cotton* » et « *cover* » dominent dans la catégorie « *Home Furnishing* », tandis que leur importance relative diminue dans TF-IDF à cause de leur présence répandue, à l'inverse de « abstract ».

L'approche BoW, en combinaison avec la lemmatisation, s'avère plus performante pour notre jeu de données, où la fréquence des mots est un indicateur clé de la catégorisation.

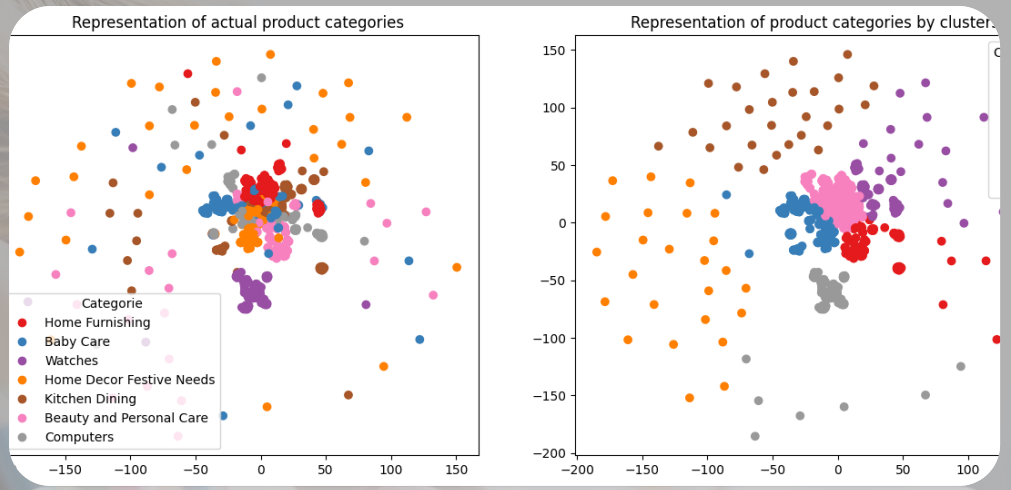
Visualisation de la segmentation

Clustering des données textuelles

BoW clustering



TF-IDF clustering



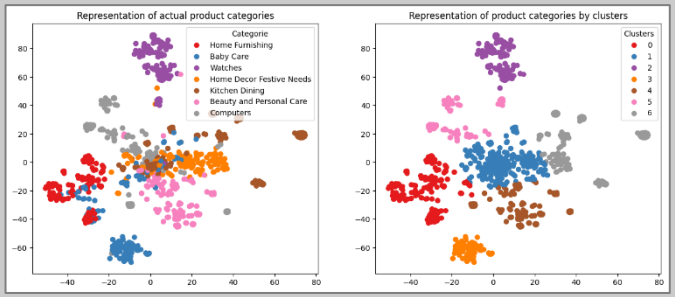
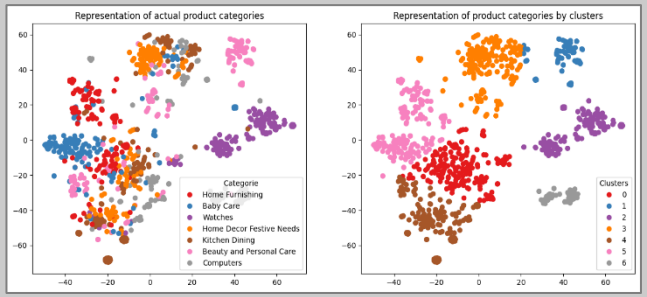
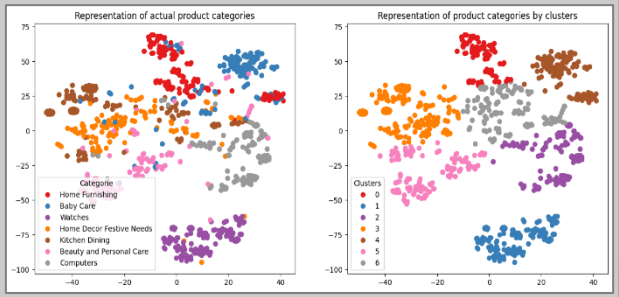
Les clusters formés à partir des caractéristiques BoW correspondent étroitement aux catégories réelles, montrant une séparation distincte, tandis que la représentation TF-IDF présente un chevauchement plus bien important des clusters.

Le modèle BoW aligné avec la lemmatisation est préférable pour des catégories bien définies, offrant une visualisation plus nette et des clusters mieux délimités.

Vectorisation de textes pour la catégorisation automatique

Comparaison des modèles d'embedding FastText, BERT et USE

A partir des descriptions semi nettoyées de nos produits, nous avons mis en œuvre FastText pour capturer les nuances morphologiques, BERT pour sa compréhension contextuelle des mots, et USE pour la signification globale des phrases.

	FastText	**BERT**	**USE**
Modèle	*Skip-gram model*	*BERT-base model*	*Universal Sentence Encoder*
ARI score	0.4922	0.2676	0.5496
Clustering Visuel (t-SNE)			
Similarité Intra-catégorie	Haute (proche de 1.0)	Haute (proche de 1.0)	Haute (proche de 1.0)
Similarité Inter-catégories	Bonne séparation (0.89 à 0.97)	Bonne séparation (0.94 à 0.99)	Distinction claire (0.37 à 0.89)
Optimisation et réglages	Hyperparamètres via Optuna	Paramètres par défaut BERT	Paramètres par défaut USE
Conclusion et application potentielle	**Catégorisation assez précise**	**Bon dans la distinction contextuelle**	**Distinction catégorielle supérieure**

Vectorisation de textes pour la catégorisation automatique

Résultats et recommandations

Nous recommandons l'utilisation de USE pour la catégorisation automatique chez Place de Marché.

Sa capacité à distinguer clairement entre différentes catégories de produits et à traiter des descriptions textuelles complètes le rend particulièrement adapté pour votre plateforme.

Cette approche est susceptible de minimiser les erreurs de catégorisation et d'améliorer l'expérience utilisateur.

Un POC (Proof of Concept) ou une phase pilote avec USE serait utile dans une prochaine étape pour valider l'approche avant un déploiement à grande échelle.

Classification automatique des articles à partir de leurs images

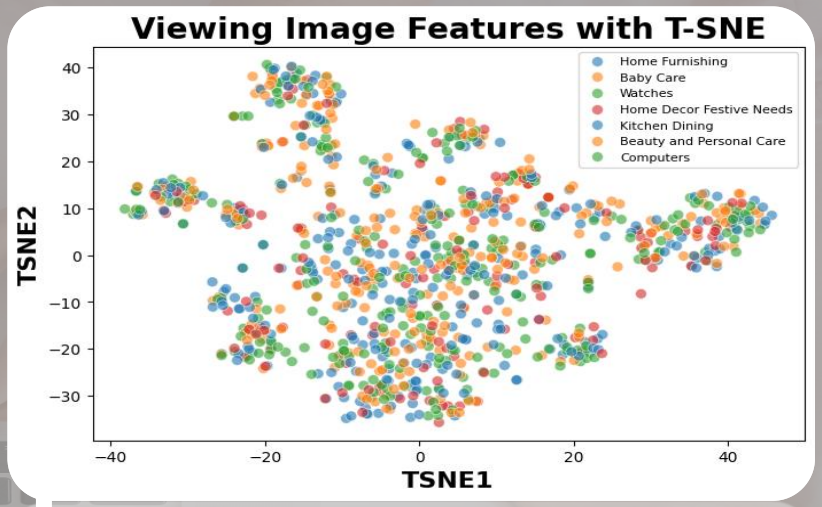
—

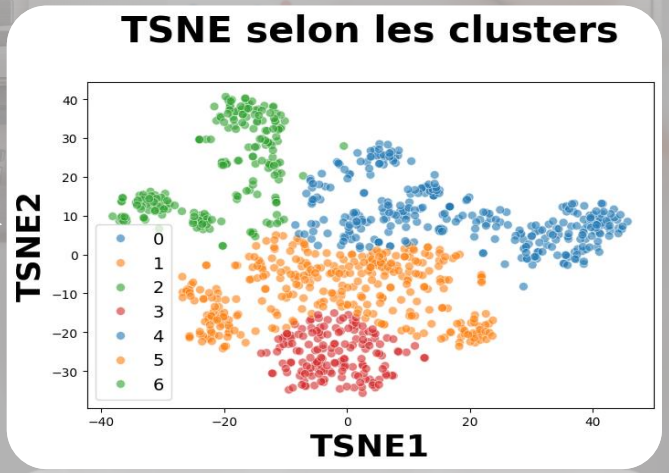
Faisabilité

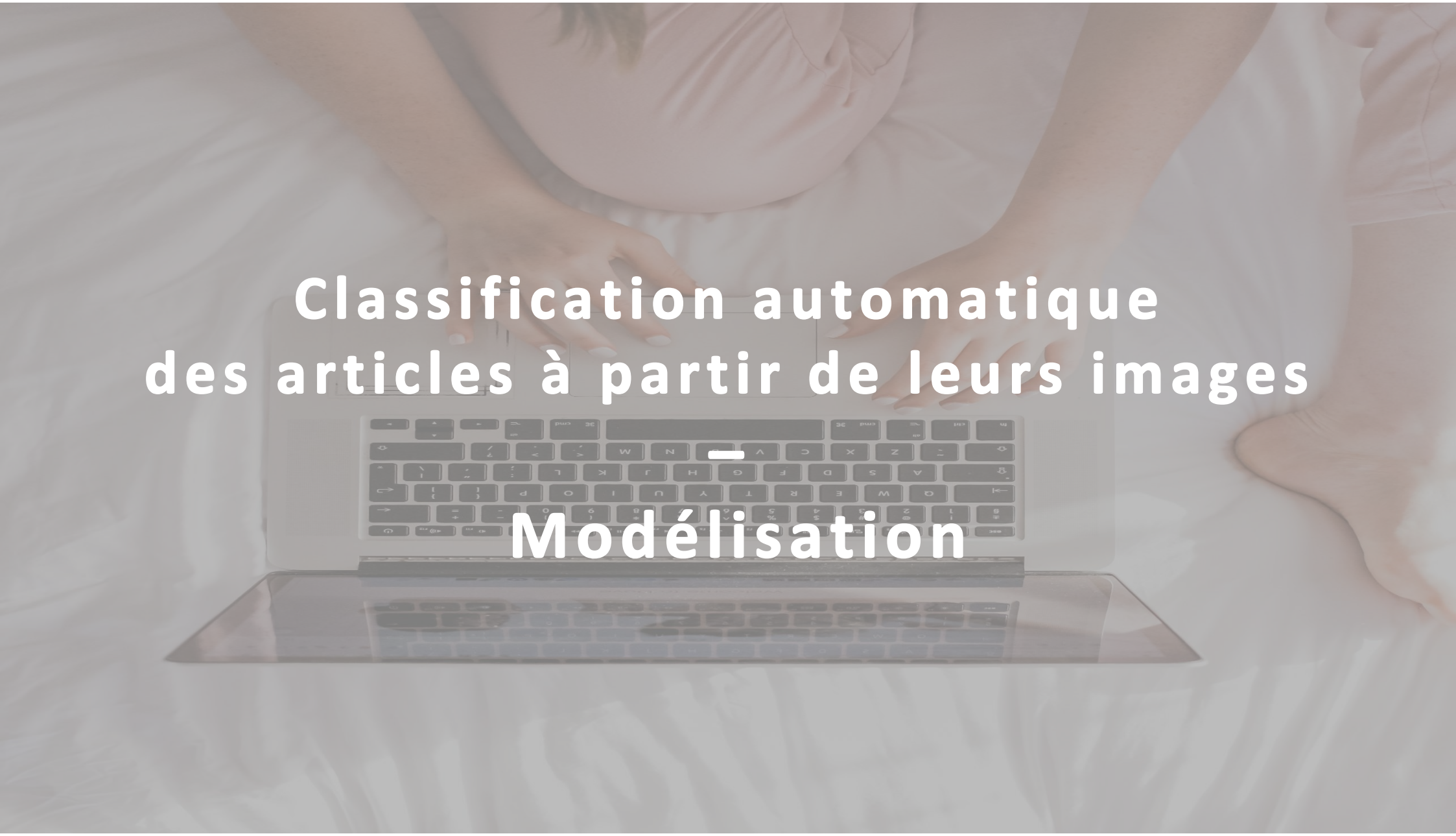
Des images aux features: prétraitement et analyse

Extraction des caractéristiques et clustering

- **Traitement avancé des images** : redimensionnement, conversion en array numérique, utilisation de la méthode de traitement VGG16 pour l'amélioration du contraste et la normalisation.
- **Feature engineering avec ORB**: création d'un 'bag-of-images' par extraction des points clés avec ORB, formant un premier ensemble de caractéristiques discriminantes.
- **Extraction profonde via CNN**: mise en œuvre du Transfer Learning avec un modèle CNN (ici VGG16) pour une extraction plus fine et pertinente des caractéristiques.
- **Analyse t-SNE (features CNN)**: réduction dimensionnelle avec PCA avant t-SNE afin de conserver les fonctionnalités les plus informatives améliore la capacité du t-SNE à distinguer visuellement les clusters dans un espace bidimensionnel.
- **Résultats et interprétation**: Un ARI négatif proche de zéro suggère que l'attribution des clusters est aléatoire, sans pour autant contrevenir à une structure intrinsèque des données. Un score de silhouette de -0.049 révèle une séparation imparfaite des clusters, soulignant le besoin d'un ajustement plus fin du modèle pré-entraîné sur nos données spécifiques étiquetées.







Classification automatique des articles à partir de leurs images

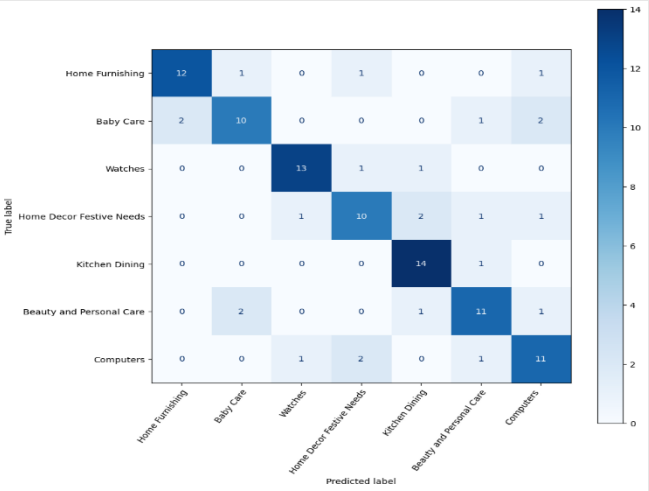
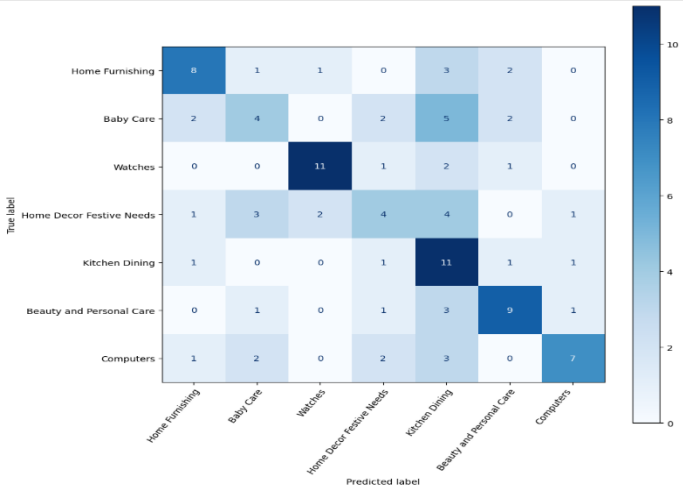
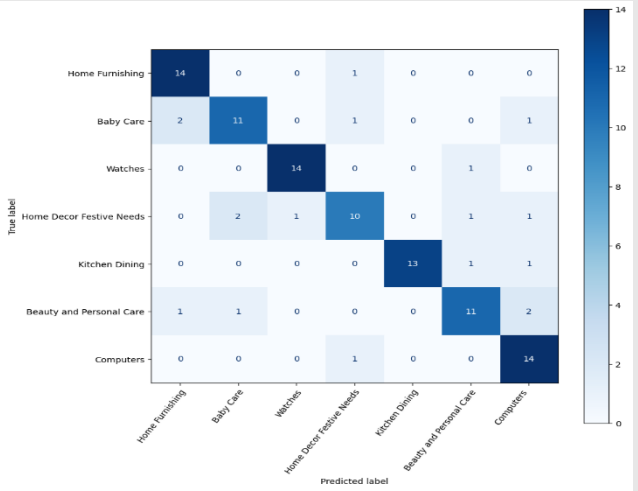
—

Modélisation

Évaluation de modèles CNN en apprentissage par transfert

Comparaison des modèles de Deep Learning CNN : VGG16, ResNet et Inception

Ces modèles utilise l'apprentissage par transfert pour affiner un modèle général pré-entraîné sur nos données prétraitées.

	VGG16	ResNet	Inception
Optimisation et Convergence	*EarlyStopping* après 13 époques	*EarlyStopping* après 29 époques	*EarlyStopping* après 15 époques
Matrice de confusion			
ARI score	0.8149	0.6203	0.976
Précision et Perte	Test de precision: ~77% - Test loss: 0.761	Test de precision: ~51.4% - Test loss: 1.469	Test de precision: ~82.6% - Test loss: 0.692
Performance globale	**Bonne capacité de généralisation avec marge d'amélioration**	**Nécessite un raffinement supplémentaire pour obtenir de meilleures performances**	**Performance globale supérieure avec de fortes capacités de généralisation**

Évaluation de modèles CNN en apprentissage par transfert

Augmentation de données basic et avancée

Avec déjà la meilleure performance globale parmi les trois modèles, l'augmentation des données sur le modèle Inception pourrait soit solidifier sa robustesse, soit à l'inverse ne pas apporter une amélioration aussi marquée que pour les autres modèles moins performants. Nous saurons néanmoins si les capacités de généralisation d'Inception sont déjà maximisées ou si elles peuvent être encore améliorées.

	Basic data augmentation	Advanced data augmentation
Optimisation et Convergence	*EarlyStopping* après 17 époques	*EarlyStopping* après 26 époques
ARI score	0.8149	0.804
Précision et Perte	Test de precision: ~77.14% - Test loss: 0.646	Test de precision: ~79% - Test loss: 0.612
Performance globale	- Meilleure gestion de l'erreur malgré une baisse de la précision - Grande diversité dans les images augmentées qui complique la tâche de clustering (introduction de défis supplémentaires)	- Une perte réduite par rapport à la data augmentation basique - Complexité accrue qui peut nécessiter un entraînement plus long et un ajustement fin des hyperparamètres.
Conclusion	**L'utilisation de l'augmentation de données, qu'elle soit basique ou avancée, doit être soigneusement évaluée en fonction des performances de base du modèle et des caractéristiques spécifiques de l'ensemble de données. Pour les ensembles de données déjà divers et bien représentatifs, une augmentation de données avancée pourrait ne pas apporter d'avantages significatifs en termes de précision, bien qu'elle puisse contribuer à réduire la perte et à augmenter la robustesse du modèle.**	


Test de collecte de données
via API

Test de collecte d’articles à base de champagne via l’API *Edamam Food and Grocery Database*

- **Requête API** : Une requête a été écrite et testée pour interroger l'API d'Edamam en se concentrant sur les critères de recherche spécifiques, notamment pour des aliments contenant l'ingrédient "champagne".

- **Récupération des champs nécessaires** : Les données récupérées incluent les champs spécifiés : foodId, label, category, foodContentsLabel, image.

- **Application d’un filtre** : Un filtre a été appliqué sur l'ingrédient ("ingr") pour ne collecter que les données associées au "champagne".

- **Stockage des données collectées** :Les données collectées ont été stockées dans un fichier CSV, utilisable pour des analyses ultérieures ou l'importation dans d'autres systèmes.

En examinant les données collectées, il semble que le script a respecté la minimisation des données en ne récupérant que les informations pertinentes et nécessaires pour le projet.

Il est important de noter que toute collecte de données doit obtenir le consentement des personnes concernées si les données personnelles sont impliquées, bien que, dans ce cas, les données semblent non personnelles et se concentrent sur des produits.

Conclusion du projet

Conclusion

Synthèse du projet

- Nous avons évalué avec succès la faisabilité d'un système de classification automatique des produits pour la société "Place de marché".
- Les tests ont démontré la capacité du modèle à classifier précisément les articles à partir de leurs images et descriptions textuelles.

Résultats clés

- Le modèle basé sur l'Universal Sentence Encoder (USE) a démontré sa capacité à classifier précisément les articles à partir de leurs descriptions textuelles.
- Le modèle Inception a surpassé les autres modèles CNN en termes de précision et de perte pour la classification à partir des images.

Respect de la RGPD

- Nous avons veillé au respect des principes du RGPD, en assurant la minimisation et la sécurité des données collectées.
- La collecte des données via l'API Edamam a été conforme aux exigences légales et éthiques.

Etapes futures

- Continuer à peaufiner les modèles en fonction des retours utilisateurs et des performances en conditions réelles.
- Explorer de nouvelles approches et technologies émergentes pour maintenir l'efficacité du système de classification.