



CLASSIFIER AUTOMATIQUEMENT DES BIENS DE CONSOMMATION



Place de Marché



PROBLÉMATIQUE

Lancement d'une marketplace e-commerce

Enjeu :

Catégoriser automatiquement des biens de consommation selon leur description et/ou leur image, afin d'améliorer l'expérience des vendeurs et des clients, et de gagner en temps et en efficacité.



JEU DE DONNÉES

Fichier de données :

[flipkart_com-ecommerce_sample_1050.csv](#)

Dataset comportant l'ensemble des informations relatives aux 1050 images (nom, URL, catégorie, description, nom de l'image correspondante).

Images :

1050 images au format .jpg, dont les noms correspondent sont ceux inclus dans le dataset Flipkart. Les images ne sont pas classées par catégories.



PRÉTRAITEMENTS TEXTE

ETUDE DE FAISABILITÉ

Caractéristiques du fichier Flipkart :

- Ni valeurs nulles, ni doublons
- Identification des catégories principales ('main_cat') au nombre de 7, équilibrées car comprenant chacune 150 produits :

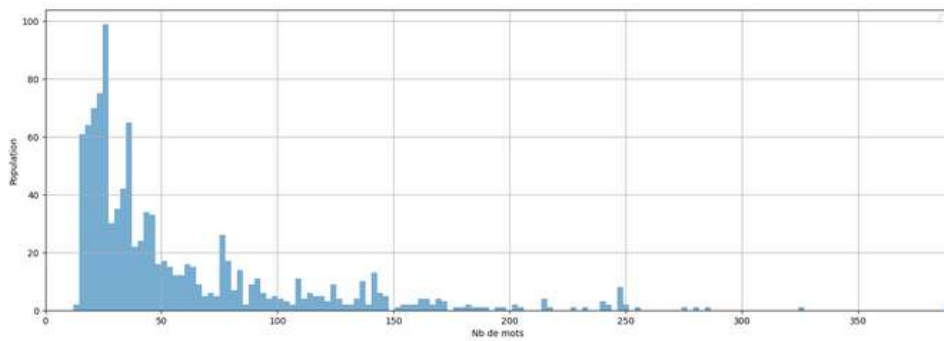
Home Furnishing	150
Baby Care	150
Watches	150
Home Decor & Festive Needs	150
Kitchen & Dining	150
Beauty and Personal Care	150
Computers	150

ETUDE DE LA FEATURE 'DESCRIPTION'

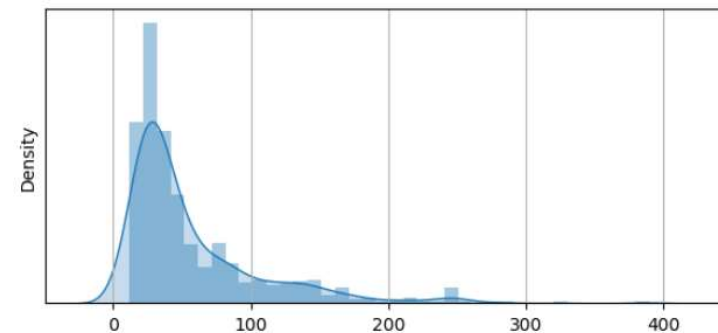
Feature comprenant une description textuelle de chaque produit.

Analyse de la feature :

Longueur des descriptions



Longueur max. : 3 588 mots





MANIPULATION DU TEXTE DE DESCRIPTION

- Cleaning général :

minuscules, suppression des URL, suppression des ponctuations etc.

- Suppression des contractions :

ex : « you're » devient « you are », « yall » devient « you all »

- Suppression des Stop words

ex : « this », « do », « more »

- Tokenisation des phrases :

ex : [« Projet 6. Classification textes et images »] devient ['Projet 6.', 'Classification textes images']

- Stemmatisation :

ex : programming / programmer / programs deviennent : 'program'

- Lemmatisation :

ex : meeting devient 'meet', 'was' devient 'be'

Application au texte descriptif :

'This curtain enhances the look of the interiors' ➡ ['curtain', 'enhanc', 'look', 'interior']



ESSAIS DE MODELISATION

Nous appliquerons un Kmeans ($n_clusters = 7$) couplé à nos différentes méthodes de réduction de dimension et aux méthodes de traitement du texte.

Modèles testés :

- Sentence Embedding :
 - USE
 - Sentence BERT
- Bag of Words (BoW) :
 - CounterVectorizer
 - TF-IDF
- Word Embedding :
 - Word2Vec
 - FastText
 - BERT (bert-base-uncased)
 - BERT (TensorFlow Hub)

Réduction de dimension :

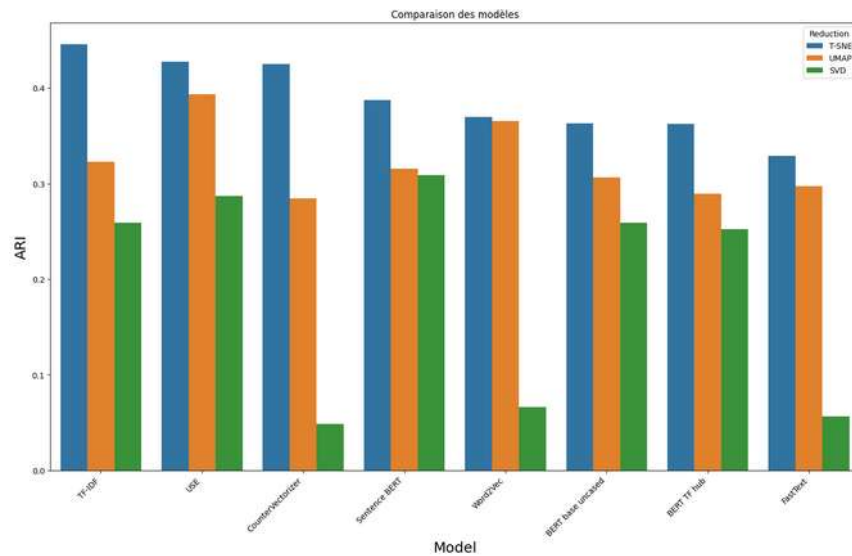
- T-SNE
- UMAP
- SVD

Métrique d'évaluation :

- ARI
- Matrice de confusion
- Precision
- Recall
- F-1 score

BILAN DE LA MODELISATION

Comparaison des différents essais :



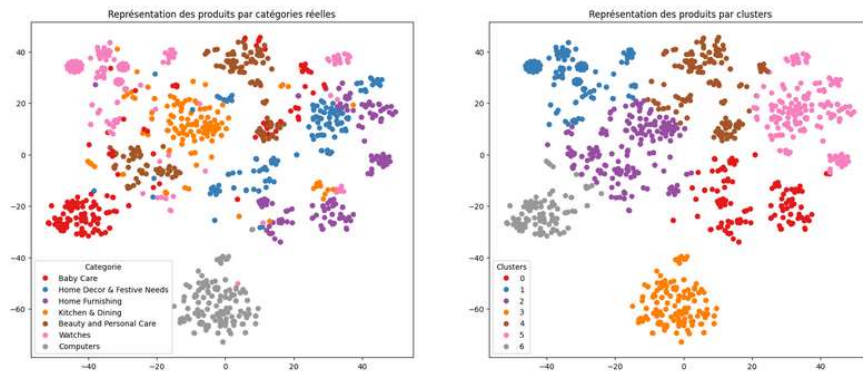
Les scores ARI sont globalement bas ($< 0,5$).

La modélisation la plus pertinente est la combinaison : TF-IDF avec réduction de dimension par T-SNE (ARI = 0,446)

BILAN DE LA MODELISATION

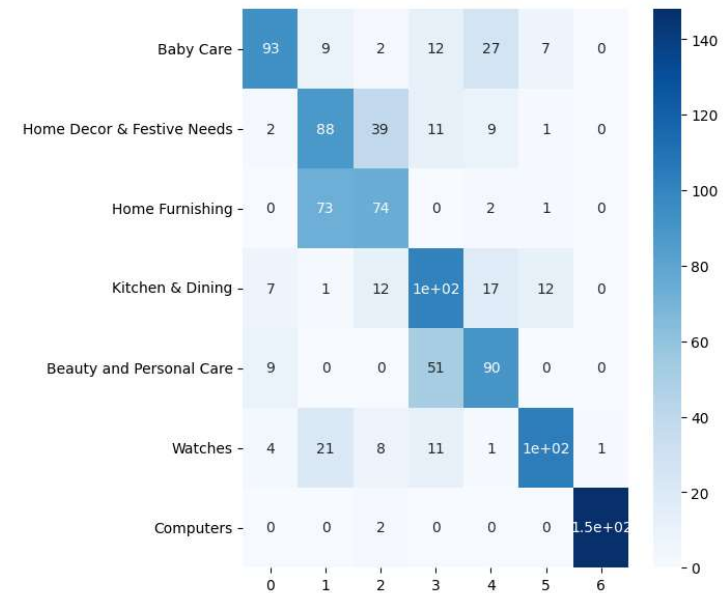
Evaluation de la meilleure modélisation (TF-IDF) :

ARI : 0.4459



La qualité de la prédiction dépend des catégories. La catégorie 'Computers' est ici bien mieux prédite que la catégorie 'Baby Care'.

	precision	recall	f1-score	support
0	0.81	0.62	0.70	150
1	0.46	0.59	0.51	150
2	0.54	0.49	0.52	150
3	0.54	0.67	0.60	150
4	0.62	0.60	0.61	150
5	0.83	0.69	0.76	150
6	0.99	0.99	0.99	150
accuracy			0.66	1050
macro avg	0.68	0.66	0.67	1050
weighted avg	0.68	0.66	0.67	1050



PRÉTRAITEMENT IMAGES

ETUDE DE FAISABILITÉ

Création d'une feature 'img' combinant le chemin d'accès au dossier contenant les images, et le nom de l'image.

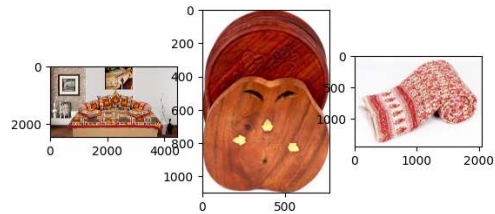
Exemple d'image :



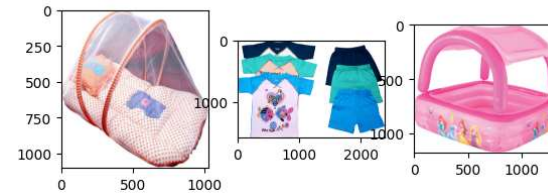
LISTING DES CATEGORIES REELLES

3 exemples d'images par catégorie :

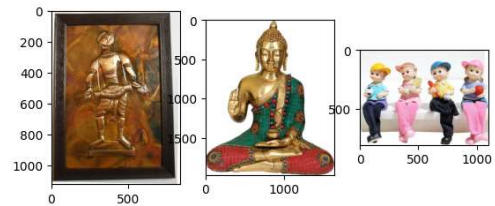
Home Furnishing



Baby Care



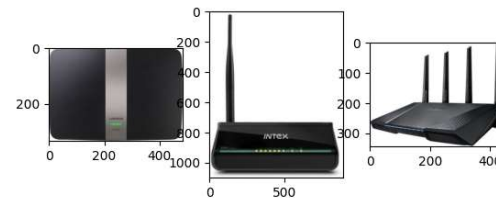
Home Decor & Festive Needs



Kitchen & Dining



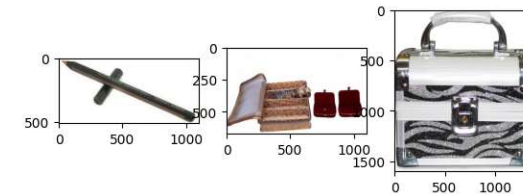
Computers



Watches



Beauty and Personal Care



MANIPULATION DES IMAGES

Redimensionnement :

Formats acceptés par les méthodes qui seront utilisées : (224 x224) et (299 x 299)

Flou gaussien : uniformiser les parties d'une image en les floutant et en harmonisant ses détails.

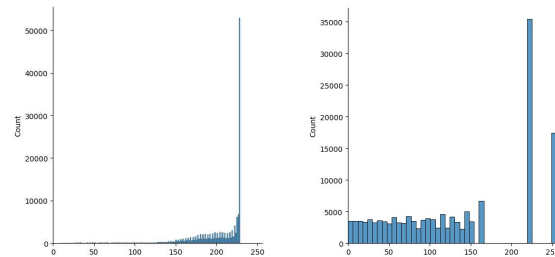
▪ Ex :



MANIPULATION DES IMAGES

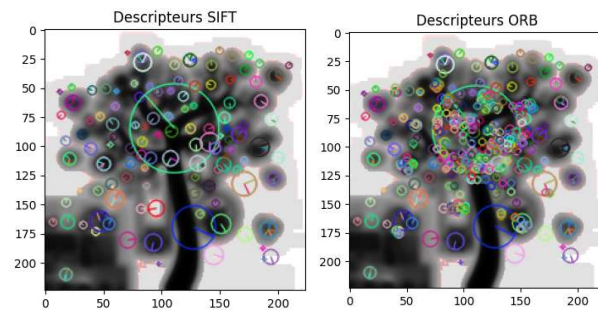
Egalisation des couleurs : augmenter le contraste, rétablir les points les plus foncés et les plus clairs, et répartir uniformément les valeurs entre ces deux points.

Ex :



Descripteurs (SIFT et ORB) : calcul de caractéristiques visuelles dans un ensemble de sous régions de l'image.

Ex :





ESSAIS DE MODELISATION

Nous appliquerons un Kmeans ($n_clusters = 7$) couplé à nos différentes méthodes de réduction de dimension et à nos méthodes de traitement d'images.

Méthodes testées :

- Features SIFT
- ResNet50
- VGG16
- VGG19
- Xception
- InceptionV3

Réduction de dimension :

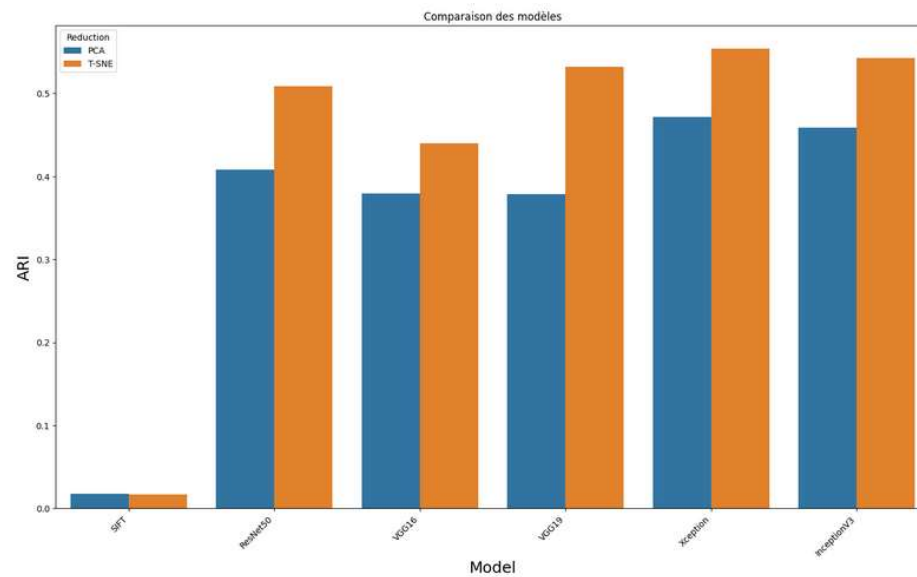
- T-SNE
- PCA

Métrique d'évaluation :

- ARI
- Temps de traitement
- Matrice de confusion
- Precision
- Recall
- F-1 score

BILAN DE LA MODELISATION

Comparaison des différents essais :

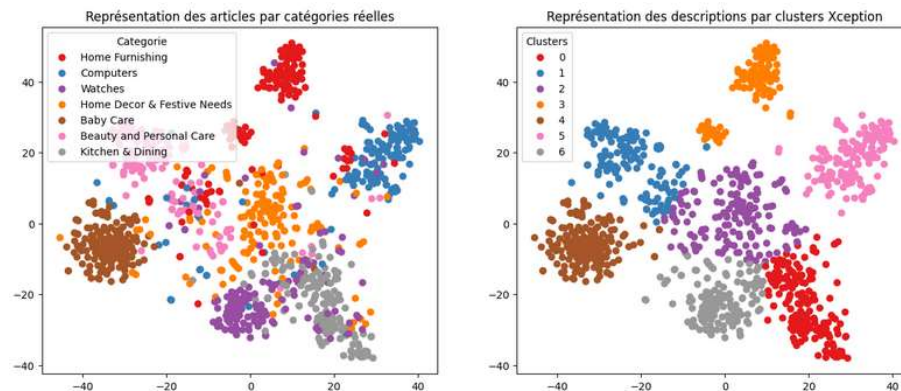


Les scores ARI sont globalement bas ($< 0,6$).

La modélisation la plus pertinente est la combinaison :
Xception avec réduction de dimension par T-SNE (ARI = 0,554)

BILAN DE LA MODELISATION

Evaluation de la meilleure modélisation (Xception) :



La qualité de la prédiction dépend des catégories. La catégorie 'Kitchen & Dining' est ici bien mieux prédite que la catégorie 'Beauty and Personal Care'.

La segmentation confirme la faisabilité d'une classification des images selon une méthode supervisée.

Score ARI : 0.5537

```
[[ 93  6  3  7 35  3  3]
 [  9 121  6  5  3  1  5]
 [  3  3 118 24  1  0  1]
 [ 18  4  3 105 12  0  8]
 [ 21  0  0  12 117  0  0]
 [  1 15 15 13  0 104  2]
 [  0  0  0  1  1  0 148]]
```

	precision	recall	f1-score	support
0	0.64	0.62	0.63	150
1	0.81	0.81	0.81	150
2	0.81	0.79	0.80	150
3	0.63	0.70	0.66	150
4	0.69	0.78	0.73	150
5	0.96	0.69	0.81	150
6	0.89	0.99	0.93	150
accuracy			0.77	1050
macro avg	0.78	0.77	0.77	1050
weighted avg	0.78	0.77	0.77	1050





CLASSIFICATION SUPERVISÉE D'IMAGES

APPROCHE

Modèle retenu :

- Xception :
 - weights = 'imagenet'
 - input_shape = (299, 299, 3)

Méthodes testées :

- Préparation initiale des images - sans data augmentation
- ImageDataGenerator (TensorFlow Keras) –data augmentation
- Image_dataset_from_directory (TensorFlow Keras) - data augmentation

Paramètres retenus :

- Batch_size : 32
- Random_state : 42
- Encodeur : LabelEncoder()

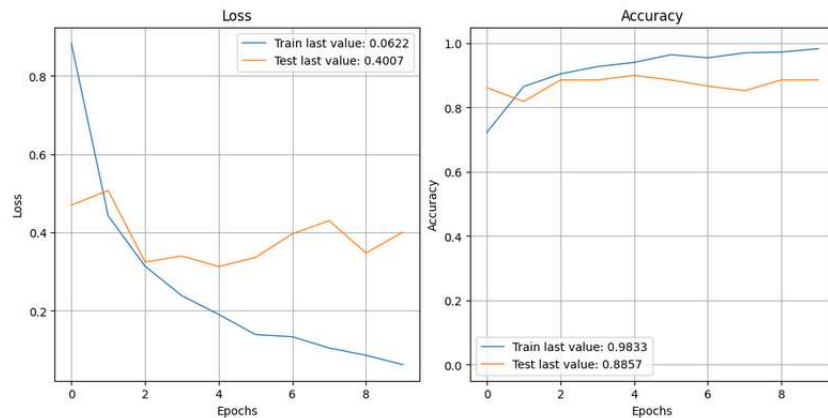
Métriques d'évaluation :

- Accuracy / loss
- Temps d'entraînement

Split du dataset :

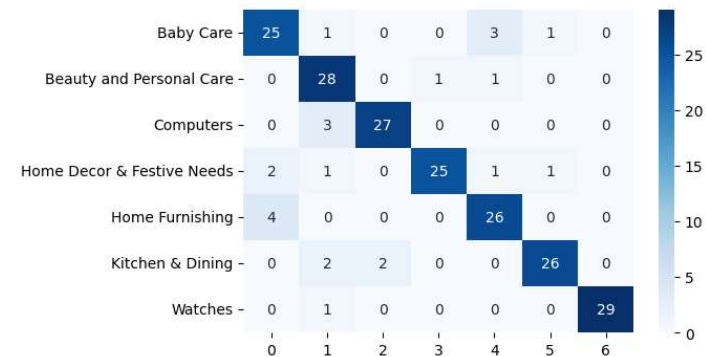
- Train / Validation / Test

PRÉPARATION INITIALE DES IMAGES - SANS DATA AUGMENTATION

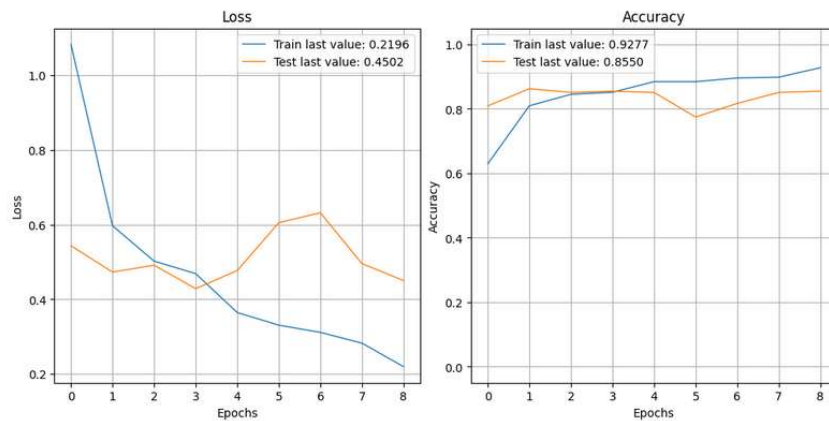


Validation accuracy : 0.8857
Validation loss : 0.3351
Test accuracy : 0.959
Test loss : 0.1347

Malgré l'absence de data augmentation dans cette modélisation, les résultats d'accuracy obtenus sont très acceptables.

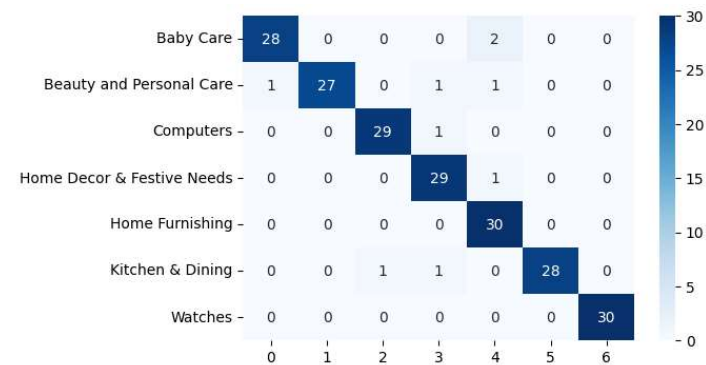


IMAGEDATAGENERATOR (TENSORFLOW KERAS) – AVEC DATA AUGMENTATION

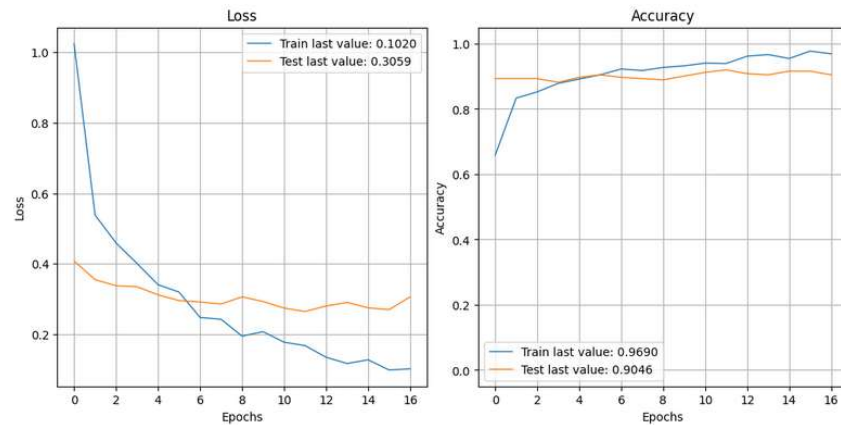


Validation accuracy : 0.8397
Validation loss : 0.53
Test accuracy : 0.9333
Test loss : 0.2141

Les résultats apparaissent ici inférieurs à ceux de notre méthode précédente n'utilisant pas de data augmentation (accuracy sur les dataset de validation et de test)



IMAGE_DATASET_FROM_DIRECTORY (TENSORFLOW KERAS) - AVEC DATA AUGMENTATION



Validation accuracy : 0.9046
 Validation loss : 0.2883
 Test accuracy : 0.9657
 Test loss : 0.1185

La méthode utilisant `Image_dataset_from_directory`, couplée à une data augmentation, permet d'obtenir les meilleurs résultats d'accuracy sur nos dataset de validation et de test.



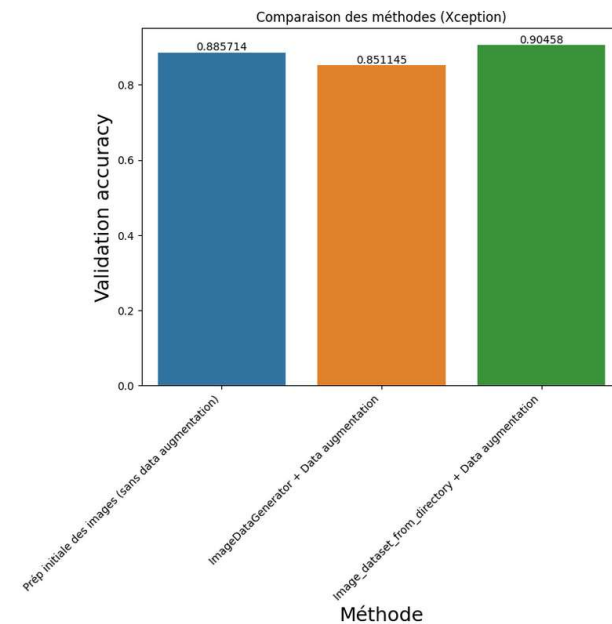
BILAN DE LA CLASSIFICATION SUPERVISÉE D'IMAGES

Modèle retenu : Xception
Encoder : LabelEncoder
Batch_size : 32
Epochs : 50 (with early stopping)
Random_state : 42

	Method	Training time	Train acc	Train loss	Val. acc	Val. loss	Test acc	Test loss
0	Prép initiale des images (sans data augmentation)	526.51	0.992857	0.028407	0.885714	0.400726	0.971429	0.102871
1	ImageDataGenerator + Data augmentation	371.80	0.956853	0.139504	0.851145	0.499971	0.940000	0.186125
2	Image_dataset_from_directory + Data augmentation	995.23	0.977381	0.071859	0.904580	0.305940	0.958095	0.131765

Conclusion :

Notre dernière méthode (Image_dataset_from_directory) donne les meilleurs résultats d'accuracy, et démontre que la classification automatique d'images est réalisable de manière efficace.





TEST D'UNE API

Enjeu :


Elargir la gamme de produits disponibles sur la market place, particulièrement concernant l'épicerie fine.

Approche :

Nous utiliserons l'API fournie (RapidAPI – Edanam Food and Grocery Database) et essaierons d'en extraire 10 produits contenant le mot 'Champagne', en ne gardant que les features nécessaires et en enregistrant l'extrait au format .csv afin qu'il soit exploitable.



EXTRACTION VIA L'API



Edamam Food and Grocery Database

By Edamam | Updated 17 days ago | Food

Popularity

9.7 / 10

Latency

711ms

Service Level

100%

[Endpoints](#) [About](#) [Tutorials](#) [Discussions](#) [Pricing](#) [Subscribed](#)

Edamam Food and Grocery Database API Documentation

This API provides you with tools to find nutrition and diet data for generic foods, packaged foods and restaurant meals. In addition it employs NLP (Natural Language Processing) which allows for extraction of food entities from unstructured text.

Covered Use Cases:

- Search for a food by keyword, food name or UPC/Barcode
- Sourcing of nutrition facts for a given food, including: macro and micro nutrients, allergen labels, lifestyle and health labels
- Search for a food by given nutrient quantity for 28 nutrients
- Search for foods within a given brand
- Built in food-logging context it allows for NLP requests for chat bots and natural language calorie counters

Données extraites enregistrées sous :
'champagne_extract.csv'

Données extraites :

	food.foodid	food.label	food.category	food.foodContentsLabel	food.image
0	food_a556mk2a5dmqb2adlamu6belhduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718d3c52...
1	food_b753lthamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
2	food_b3dyababjo54xobm6r8zbgghjqge	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_alpl44taoyv11ra0lct1qa8xculli	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne...	NaN
7	food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	NaN
8	food_am5egz6aq3fpja8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
9	food_bcz8thiajk1fuva0vdfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN



ce de A

MERCI