




PLACE DE MARCHÉ

Moteur de classification
&
Classification supervisée



OBJECTIFS DE L'ÉTUDE

- 
- Elargissement gamme de produits grâce à une API
 - Faisabilité moteur classification:
 - À partir des textes
 - Puis des images
 - Réalisation classification supervisée d'images
 - RGPD & Propriété intellectuelle



COLLECTE DE DONNÉES VIA UNE API



Création compte sur:
« <https://developer.edamam.com/food-database-api-docs> »

Collecte des données depuis API 'Food search' disponible ici:
(<https://api.edamam.com/api/food-database/v2/parser>)

Recherche de l'ingrédient 'Champagne'

Extraction du retour json des Champs:
'foodid', 'label', 'category', 'foodContentsLabel', 'image'

Limitation aux dix premières occurrences

	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_am5egz6aq3fpjlaf8xpkcdbc2asis	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

Sauvegarde du DataFrame au format csv



JEU DE DONNÉES

- Fichier csv associé à une archive de photos
- 1050 produits / 15 features / 1050 photos

```
RangeIndex: 1050 entries, 0 to 1049
Data columns (total 15 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   uniq_id                               1050 non-null   object
1   crawl_timestamp                       1050 non-null   object
2   product_url                           1050 non-null   object
3   product_name                           1050 non-null   object
4   product_category_tree                 1050 non-null   object
5   pid                                   1050 non-null   object
6   retail_price                          1049 non-null   float64
7   discounted_price                      1049 non-null   float64
8   image                                1050 non-null   object
9   is_FK_Advantage_product              1050 non-null   bool
10  description                           1050 non-null   object
11  product_rating                        1050 non-null   object
12  overall_rating                        1050 non-null   object
13  brand                                 712 non-null    object
14  product_specifications                1049 non-null   object
dtypes: bool(1), float64(2), object(12)
memory usage: 116.0+ KB
```


DONNÉES SOURCES

Texte original (484 mots): Key Features of Mom and Kid Baby Girl's Printed Blue, Grey Top & Pyjama Set Fabric: Cotton Brand Color: Blue, Grey, Mom and Kid Baby Girl's Printed Blue, Grey Top & Pyjama Set Price: Rs. 309 Girls Pyjama Set, Specifications of Mom and Kid Baby Girl's Printed Blue, Grey Top & Pyjama Set General Details Pattern Printed Ideal For Baby Girl's Night Suit Details Number of Contents in Sales Package Pack of 1 Fabric Cotton Type Top & Pyjama Set Neck Round Neck In the Box 1 Top & Pyjama Set



DONNÉES CIBLES

Distribués uniformément dans 7 catégories

```
category
Home Furnishing      150
Baby Care            150
Watches              150
Home Decor & Festive Needs 150
Kitchen & Dining      150
Beauty and Personal Care 150
Computers            150
Name: count, dtype: int64
```



MOTEUR DE CLASSIFICATION TEXTUELLE



FONCTION NETTOYAGE

Fonction unique

- Passage en minuscule
- Tokenisation
- Suppression :
 - Stopwords (si fournis)
 - Mots rares (si fournis)
 - Mots en dessous d'un nombre de caractères (si fournis)
 - Caractères non alphabétiques (si activé)
- Application d'un Stemmer ou d'un Lemmatizer
- Filtre des mots non anglais (si fournis)
- Filtre de mots additionnels (si fournis)

Test fonction nettoyage

Texte original (93 caractères): Color: Blue, Grey,Mom and Kid Baby Girl's Printed Blue, Grey Top & Pyjama Set Price: Rs. 309

Passage minuscule (92 caractères): color: blue, grey,mom and kid baby girl's printed blue, grey top & pyjama set price: rs. 309

Tokenisation (18 mots): ['color', 'blue', 'grey', 'mom', 'and', 'kid', 'baby', 'girl', 's', 'printed', 'blue', 'grey', 'top', 'pyjama', 'set', 'price', 'rs', '309']

Suppression stopwords (16 mots): ['color', 'blue', 'grey', 'mom', 'kid', 'baby', 'girl', 'printed', 'blue', 'grey', 'top', 'pyjama', 'set', 'price', 'rs', '309']

Suppression mots rares (16 mots): ['color', 'blue', 'grey', 'mom', 'kid', 'baby', 'girl', 'printed', 'blue', 'grey', 'top', 'pyjama', 'set', 'price', 'rs', '309']

Suppression mots au dessous de 3 lettres (15 mots): ['color', 'blue', 'grey', 'mom', 'kid', 'baby', 'girl', 'printed', 'blue', 'grey', 'top', 'pyjama', 'set', 'price', '309']

Caractères alpha (14 mots): ['color', 'blue', 'grey', 'mom', 'kid', 'baby', 'girl', 'printed', 'blue', 'grey', 'top', 'pyjama', 'set', 'price']

Application de lemm (14 mots): ['color', 'blue', 'grey', 'mom', 'kid', 'babi', 'girl', 'print', 'blue', 'grey', 'top', 'pyjama', 'set', 'price']

Filtre dictionnaire anglais (13 mots): ['color', 'blue', 'grey', 'kid', 'babi', 'girl', 'print', 'blue', 'grey', 'top', 'pyjama', 'set', 'price']

Suppression d'extra-words (11 mots): ['blue', 'grey', 'kid', 'babi', 'girl', 'print', 'blue', 'grey', 'top', 'pyjama', 'set']

Mots du corpus d'origine: 497512
Mots après traitement : 63371 dont 4066 uniques



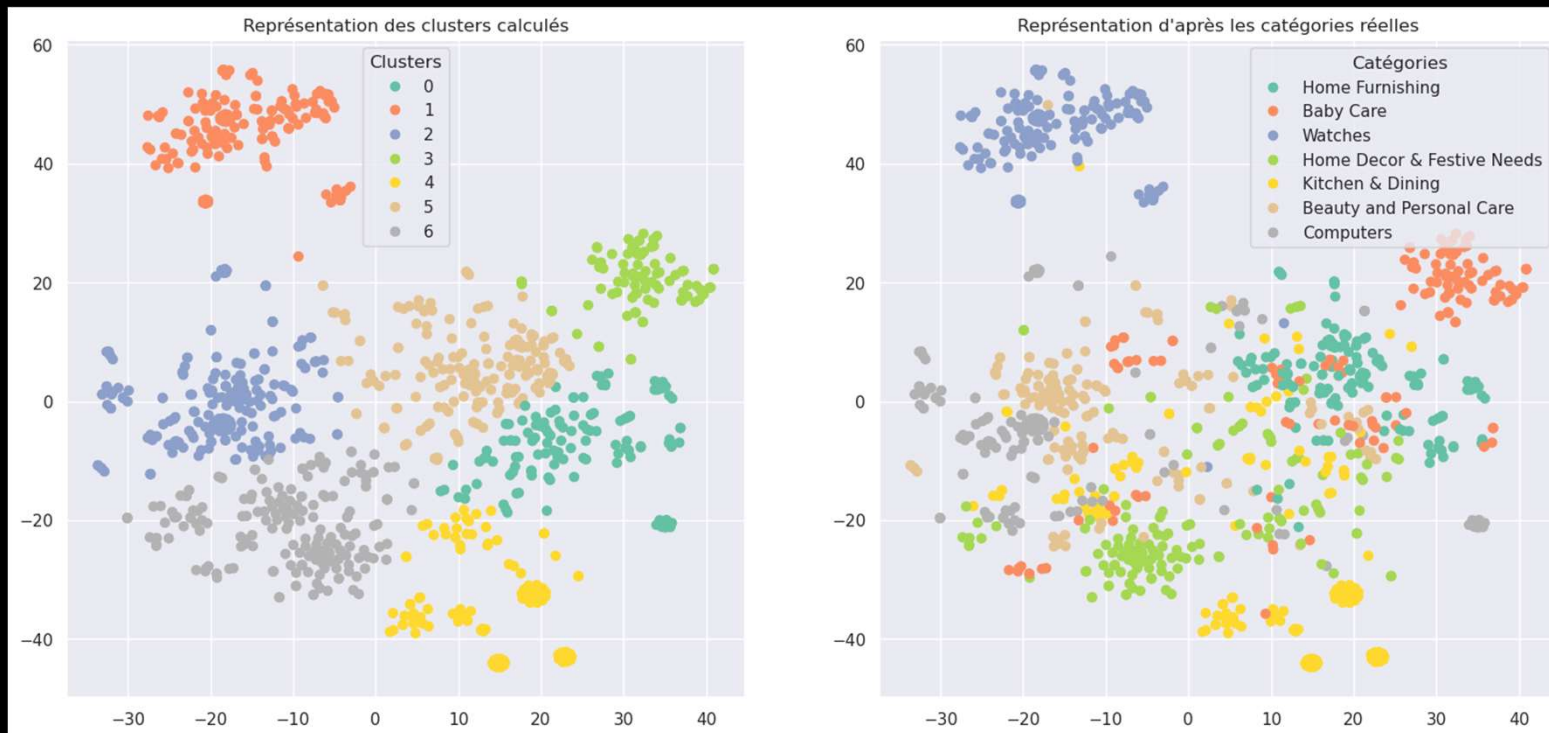
Fonction de calcul et sauvegarde des résultats

Appliquée a tous les modèles pour établir la faisabilité

- Mesure durée
- Réduction en deux dimensions (t-SNE)
- Création de 7 clusters (Kmeans)
- Calcul de l'ARI
- Graphique du cluster calculé
- Graphique des catégories réelles

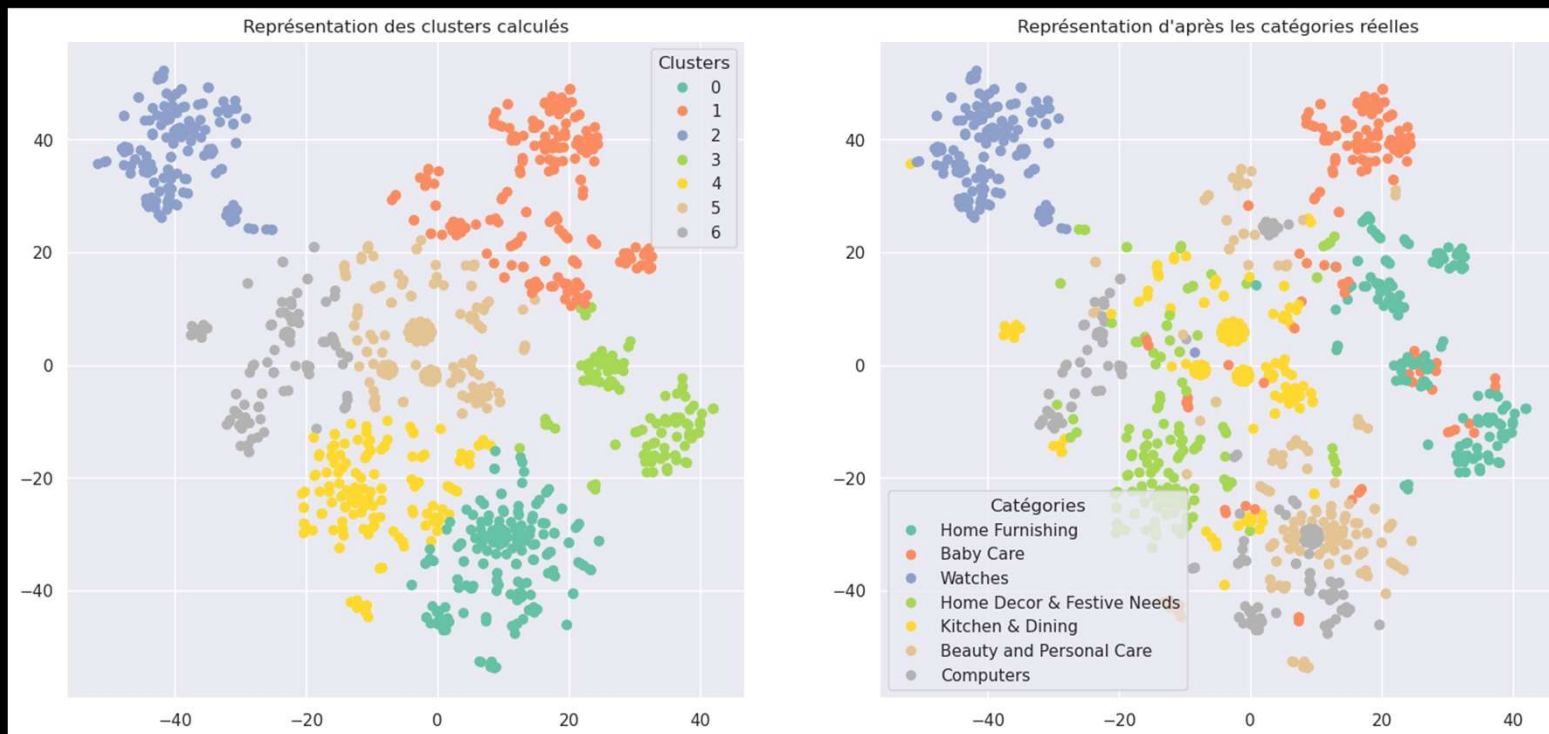
Modèle : BagOfWords CountVectorizer

durée transformation: 4.93 ARI: 0.3663



Modèle : BagOfWords Tfidf

durée transformation: 4.46 ARI: 0.4351



Modèle : Doc2Vec

durée transformation: 12.32 ARI: 0.1745



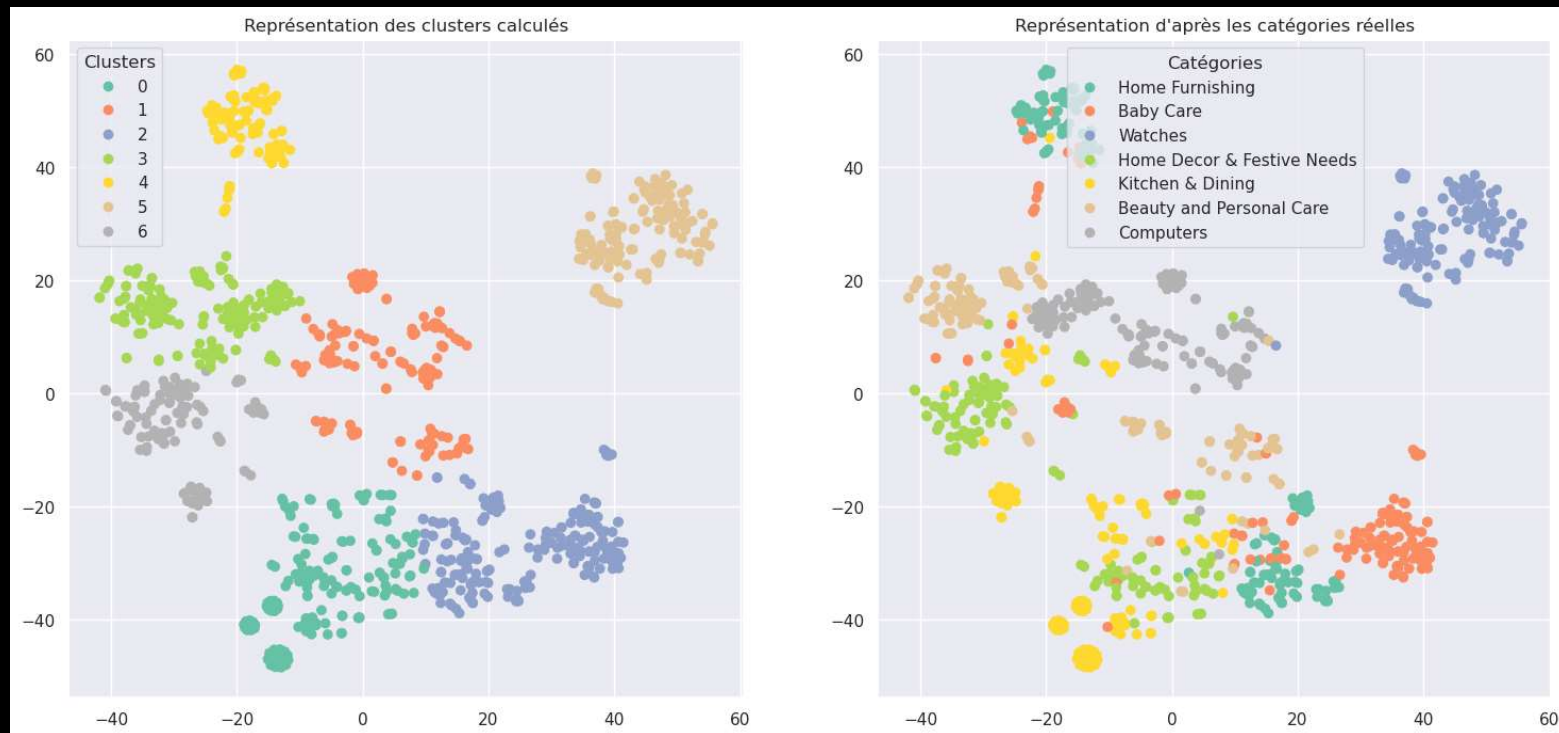
Modèle : Bert

durée transformation: 81.21 ARI: 0.3009



Modèle : Use

durée transformation: 10.74 ARI: 0.4504



Comparatif classification textuelle

Modèle	Temps	ARI
Cleaning Corpus	77.6014	
BagOfWords + CountVectorizer	4.9313	0.3663
BagOfWords + Tfidf	4.4615	0.4351
Doc2Vec	12.3258	0.1745
Bert	81.2102	0.3009
Use	10.7446	0.4504

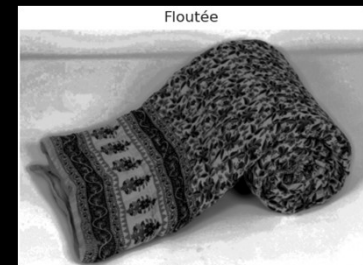


MOTEUR DE CLASSIFICATION VISUELLE

FONCTION PRÉTRAITEMENT IMAGES

Fonction unique

- Passage en niveaux de gris
- Débruitage
- Égalisation
- Floutage





Modèle : SIFT

durée transformation: 385.50 ARI: 0.0405



Modèle : Vgg16

durée transformation: 211.83 ARI: 0.4683



Comparatif classification textuelle & visuelle

Modèle	Temps	ARI
Cleaning Corpus	77.6014	
BagOfWords + CountVectorizer	4.9313	0.3663
BagOfWords + Tfidf	4.4615	0.4351
Doc2Vec	12.3258	0.1745
Bert	81.2102	0.3009
Use	10.7446	0.4504
Sift	385.5095	0.0405
Vgg16	211.8345	0.4683



CLASSIFICATION SUPERVISÉE D'IMAGES



Cibles : 7 catégories de produits

Séparation des données avec stratification:
Train 80% Valid 10% Test 10%

Utilisation de modèles pré-entraînés

3 approches :

- Préparation initiale des images avant classification supervisée
- DataSet, sans data augmentation
- DataSet, avec data augmentation intégrée au modèle



Réglages des modèles

Paramètres :

Loss = 'categorical_crossentropy'

Optimizer = 'rmsprop'

EarlyStopping = 'val_loss'

Hyperparamètre

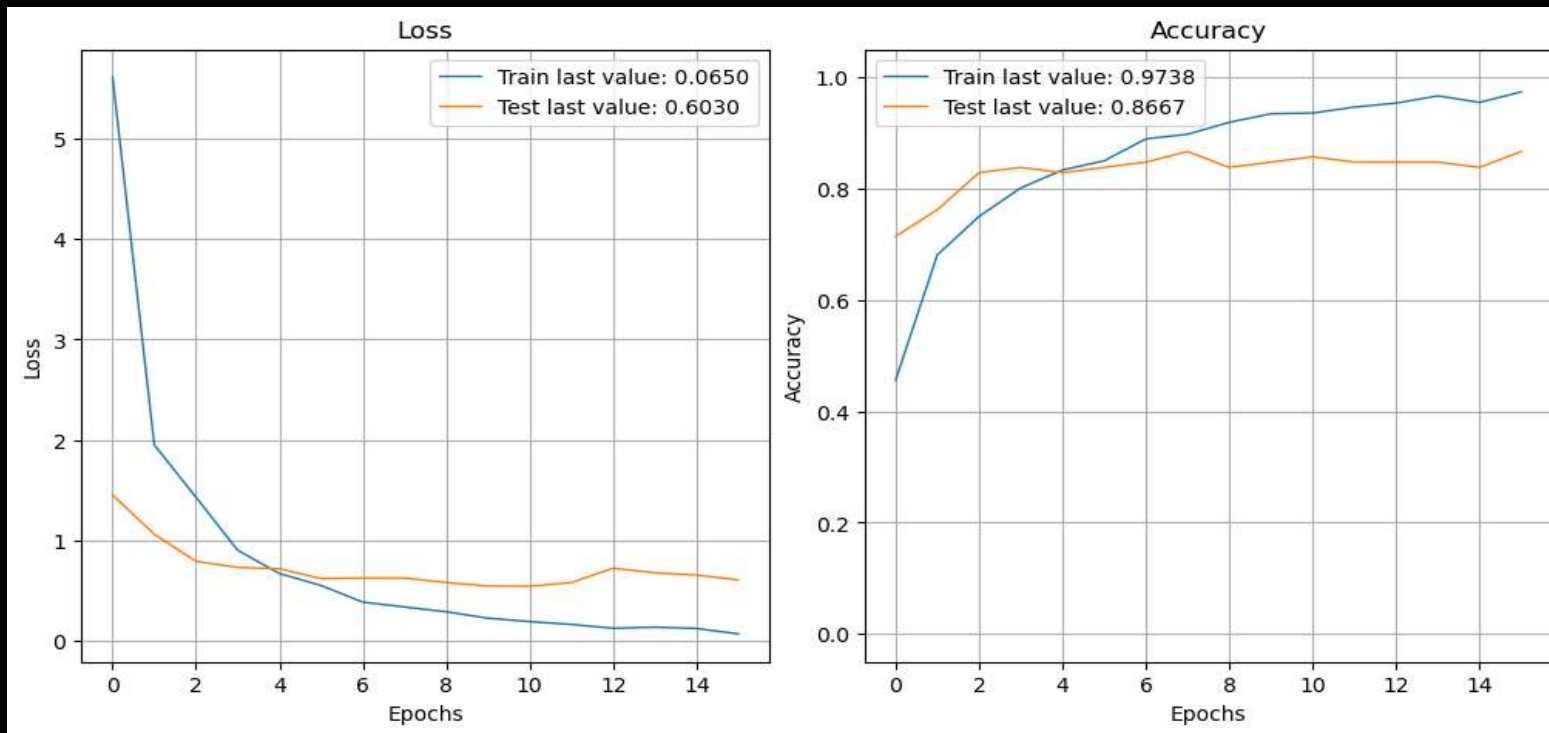
Nombre d'epochs

Metric de Mesure

Accuracy (Exactitude)

Modèle : VGG16 / préparation images

Durée: 1071s Accuracy valid: 0,85 Accuracy test: 0,76

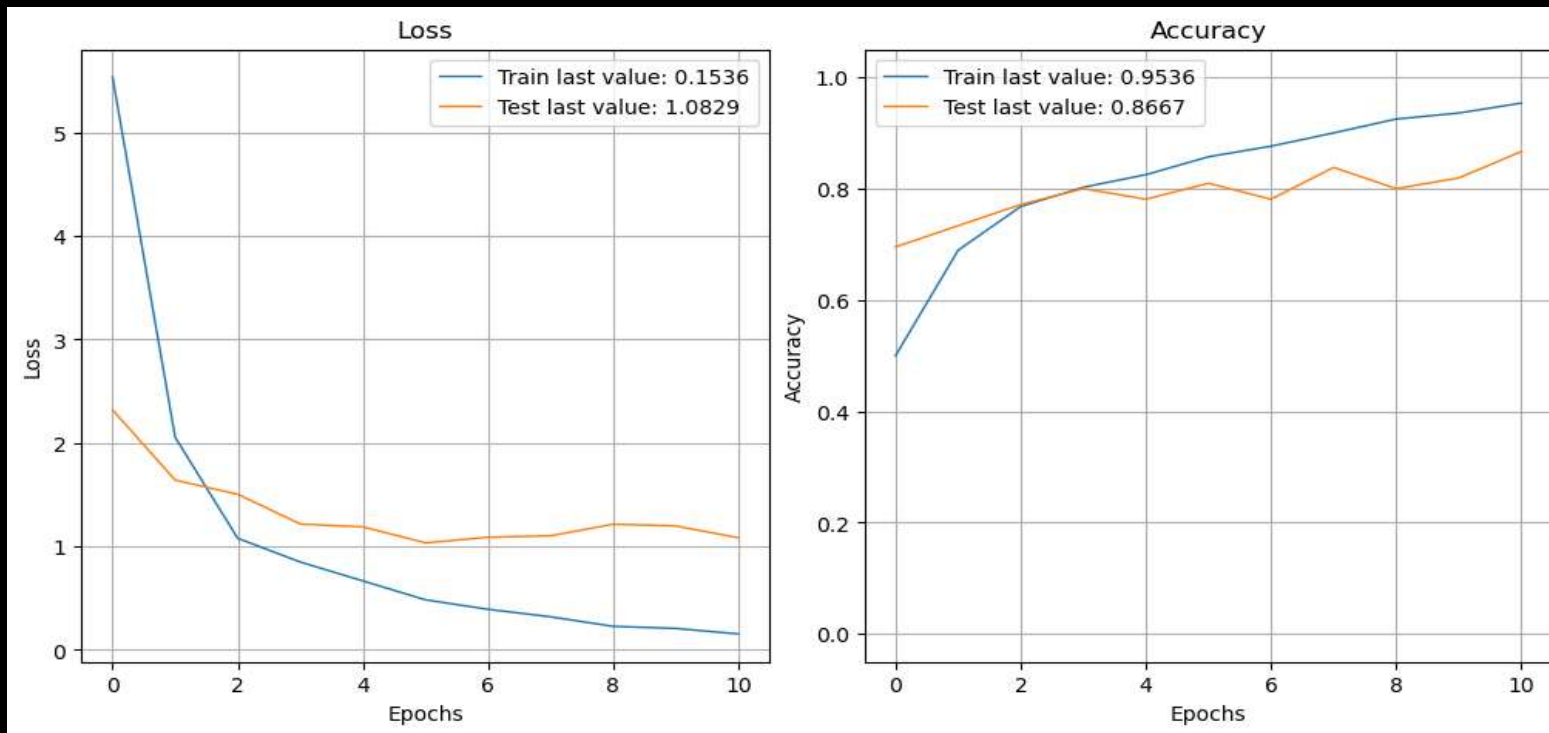


Modèle : Dataset / sans data augmentation

Durée: 1006s

Accuracy valid: 0,84

Accuracy test: 0,86





Augmentation de données:

Intégrée au modèle

Séquentielle

RandomFlip

RandomRotation

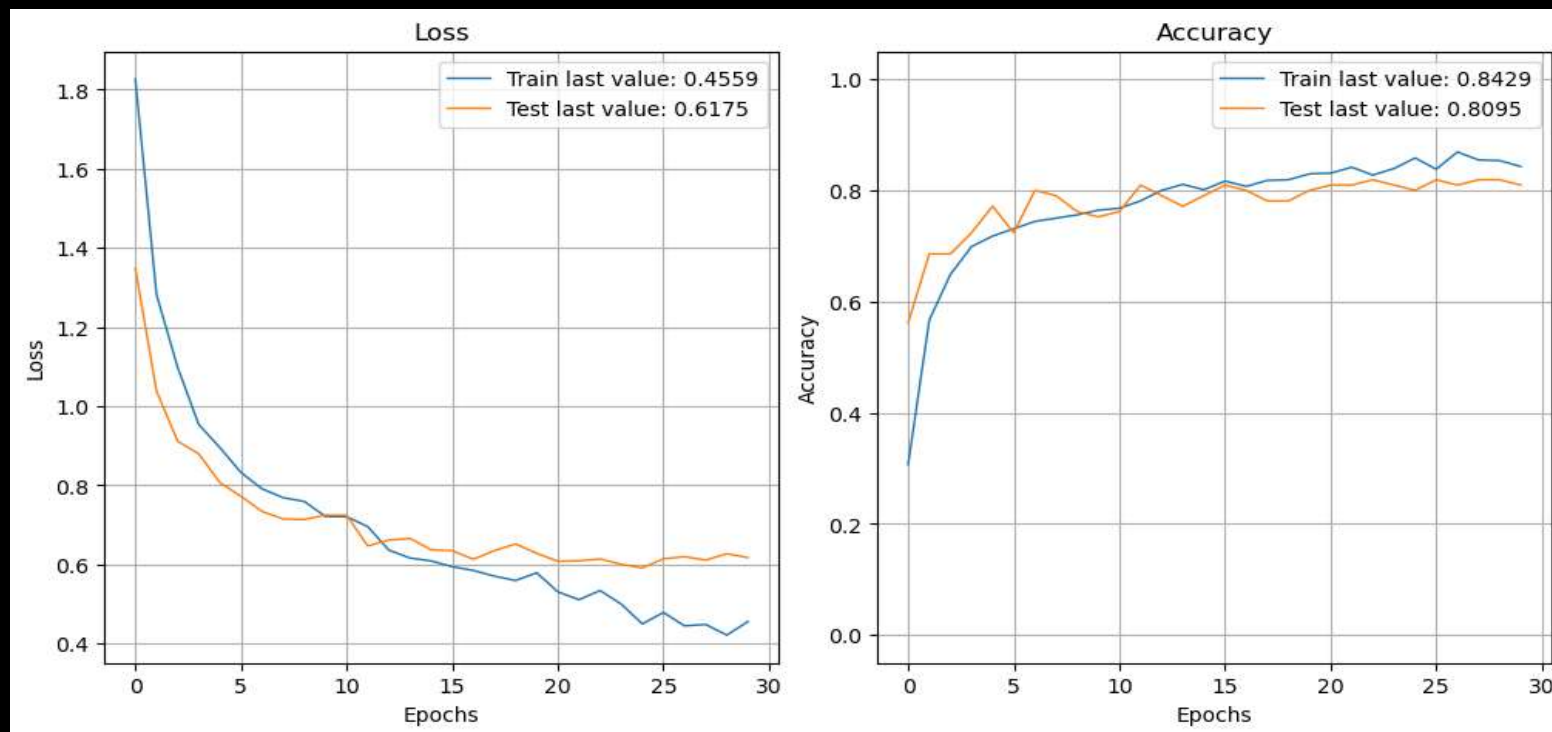
RandomZoom

Modèle : Dataset / avec data augmentation

Durée: 2368s

Accuracy valid: 0,80

Accuracy test: 0,87



Comparatif classification visuelle

	Modèle	Epoch	Loss test	Accuracy valid	Accuracy test	Temps
Préparation initiale images		15	1.232836	0.857143	0.761905	1071.54
Dataset sans data augmentation		12	0.549108	0.847619	0.866667	1006.20
Dataset avec data augmentation		29	0.419771	0.800000	0.876190	2368.69



RGPD



CINQ GRANDS PRINCIPES DU RGPD

- Licéité, loyauté et transparence
- Limitation des finalités
- Minimisation des données
- Exactitude
- Conservation

Données extraites API

Champs: 'foodid', 'label', 'category', 'foodContentsLabel', 'image'
Produit: 'Champagne'
Rétention: durée du projet

Pas de données extraites à caractère personnel





PROPRIÉTÉ INTELLECTUELLE

Aucune contrainte de propriété sur les
données et images d'après le mail de Linda

Merci pour votre attention

