Projet 6
-
Classifiez automatiquement des biens de consommation

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 –> avril 2024

Problématique

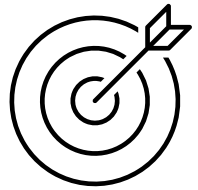
Flipkart

L’entreprise de e-commerce indien « Flipkart » permet à des vendeurs de proposer des articles à des acheteurs en postant une photo et une description.

=> automatiser l’attribution de la catégorie des articles mis en ligne par les utilisateurs pour améliorer l’expérience client (acheteur et vendeur).

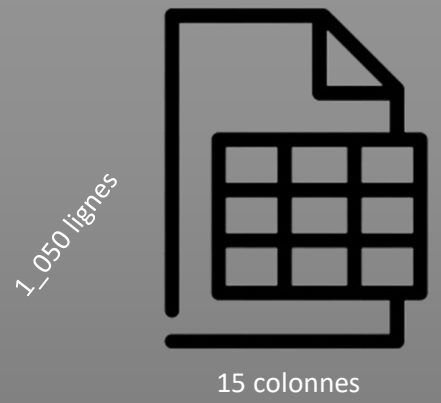
Missions :

1. Faire une étude de la faisabilité d’un moteur de classification automatique
2. Réaliser une classification supervisée à partir des images
3. Collecte de produits « champagne » sur une API



Présentation du jeu de données

flipkart_com-ecommerce_sample_1050



➢ <u>Descriptif des produits :</u>

- Des informations **textuelles**: description, marque, etc.
- Des informations de **visuelles** : image, url, etc.
- Des informations **financières** : prix, promotion, etc.
- Des informations **d'appréciation**: note produit, note globale, etc.
- Des informations **temporelles**: récupération des données

➢ <u>Valeurs manquantes :</u>

- 2% de Nan (30% des marques)
- tous les prix d'1 seul produit

➢ <u>Doublons</u>

Pas de doublons sur la colonne de l'identifiant unique des produits.

Nettoyage et EDA

1. Sélection des features

	image		description	product_category_tree
0	55b85ea15a1536d46b7190ad6fff8ce7.jpg	0	Key Features of Elegance Polyester Multicolor ...	["Home Furnishing >> Curtains & Accessories >>...
1	7b72c92c2f6c40268628ec5f14c6d590.jpg	1	Specifications of Sathiyas Cotton Bath Towel (...	["Baby Care >> Baby Bath & Skin >> Baby Bath T...
2	64d5d4a258243731dc7bbb1eef49ad74.jpg	2	Key Features of Eurospa Cotton Terry Face Towe...	["Baby Care >> Baby Bath & Skin >> Baby Bath T...

2. Détermination catégories

product_name
Home Furnishing
Baby Care
Baby Care

3. LabelEncoder

product_name	label
Home Furnishing	4
Baby Care	0
Baby Care	0

Nettoyage et EDA

X 1_050

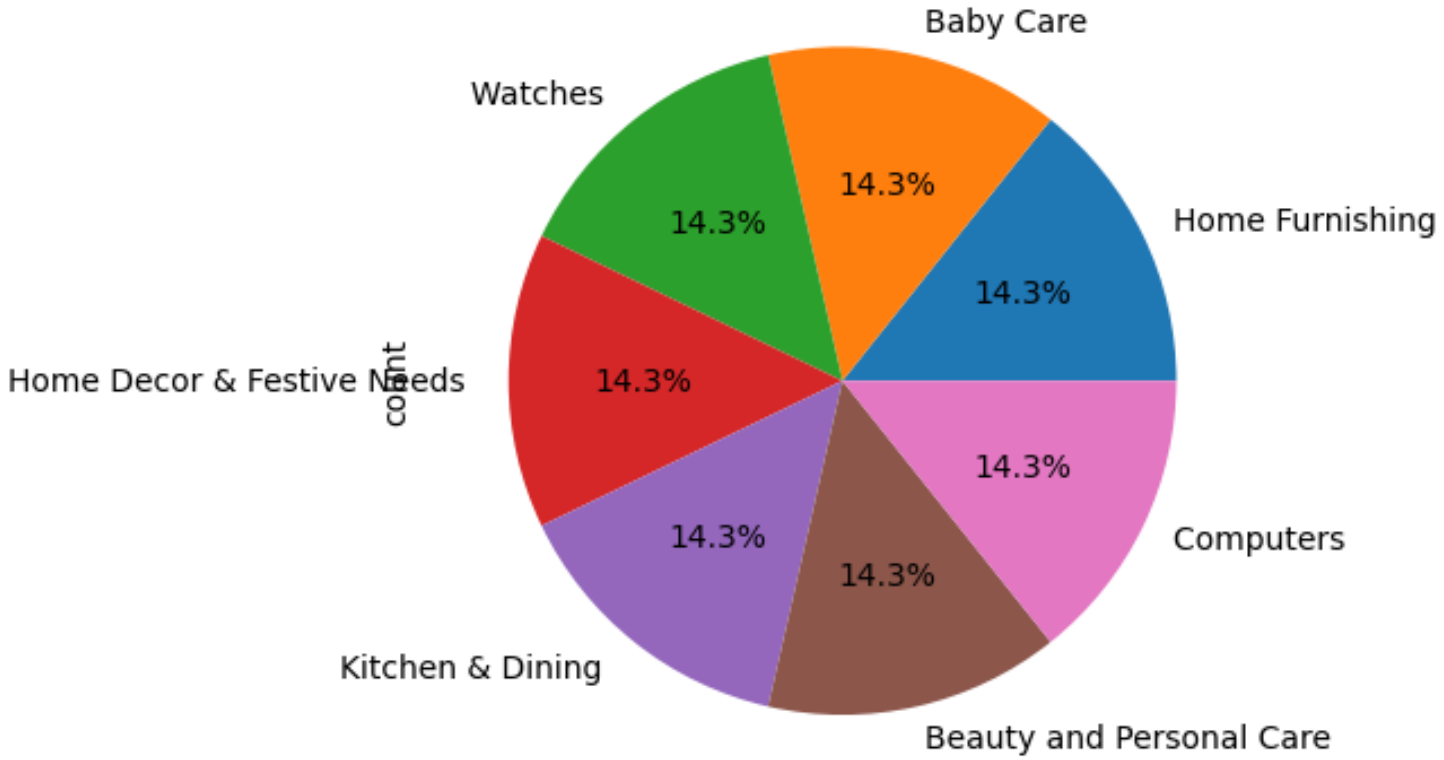


	image
0	55b85ea15a1536d46b7190ad6fff8ce7.jpg
1	7b72c92c2f6c40268628ec5f14c6d590.jpg
2	64d5d4a258243731dc7bbb1eef49ad74.jpg

description	product_name	label
Key Features of Elegance Polyester Multicolor ...	Home Furnishing	4
Specifications of Sathiyas Cotton Bath Towel (...	Baby Care	0
Key Features of Eurospa Cotton Terry Face Towe...	Baby Care	0

7 catégories

Etude de faisabilité

Missions :

1. Faire une étude de la faisabilité d'un moteur de classification automatique

Etude de faisabilité

Document
X 1_050

	description	product_name	label
0	Key Features of Elegance Polyester Multicolor ...	Home Furnishing	4
1	Specifications of Sathiyas Cotton Bath Towel (...	Baby Care	0
2	Key Features of Eurospa Cotton Terry Face Towe...	Baby Care	0

1. **Nettoyage de base** (casse, ponctuation, espace)

X 6_676

2. **Tokenisation**

X 6_560

3. **Stopwords**

X 6_486

4. **Token omniprésent**

X 6_075

5. **Longueur de mot**

X 5_702

6. **Lemmatisation**

X 3_070

7. **Vérification mots anglais**

NLTK

PRE PROCESSING

"Analog Watch 15 cm - For Boys - Buy Maserati "

"analog watch 15 cm for boys buy maserati"

'analog', 'watch', '15', 'cm', 'for', 'boys', 'buy', 'maserati'

'analog', 'watch', '15', 'cm', 'boys', 'buy', 'maserati'

'analog', 'watch', '15', 'cm', 'boys', 'maserati'

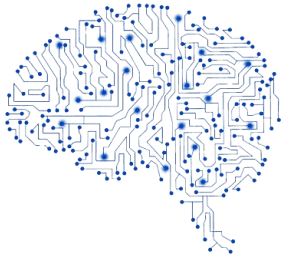
'analog', 'watch', 'boys', 'maserati'

'analog', 'watch', 'boy', 'maserati'

'watch', 'boy'

Etude de faisabilité

VS



Deep Learning

1% < Seuil de fréquence < 95% + minuscule

1. Fit Vectorizer
(Count ou tf-idf)

1. Modélisation
Deep Learning

2. ACP
(99% variance)

3. t-SNE

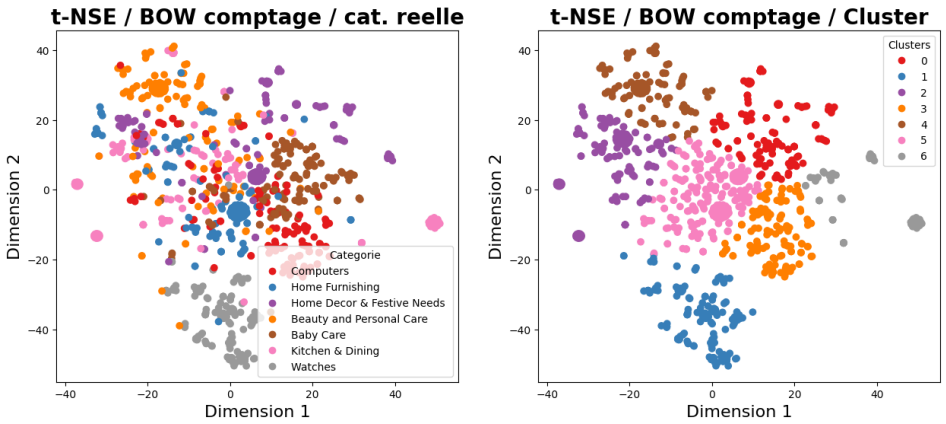
4. Clustering

5. ARI

Etude de faisabilité

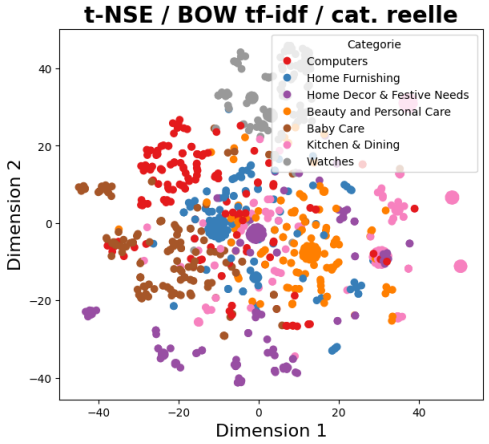
Comptage simple

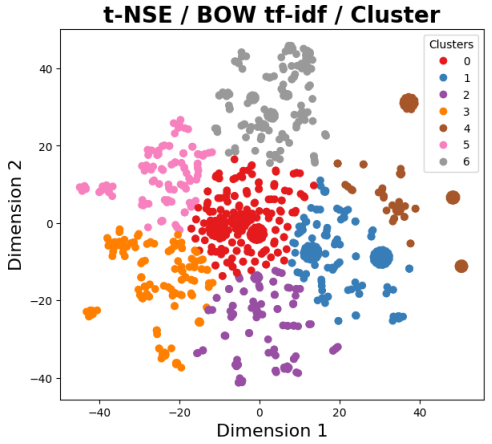
ARI = 0,31



TF-IDF

ARI = 0,30





Etude de faisabilité

Word2Vec

GENSIM topic modelling for humans

K

ARI = 0,24

Google BERT

HuggingFace

K

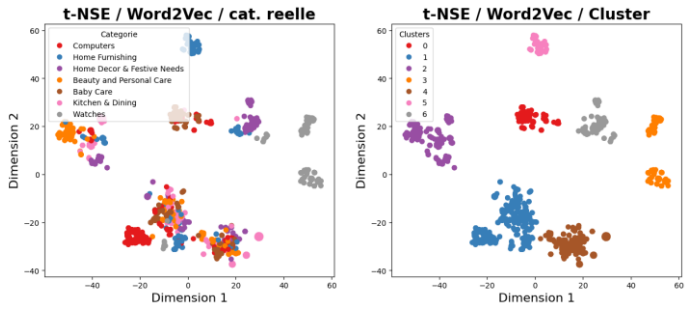
TensorFlow Hub

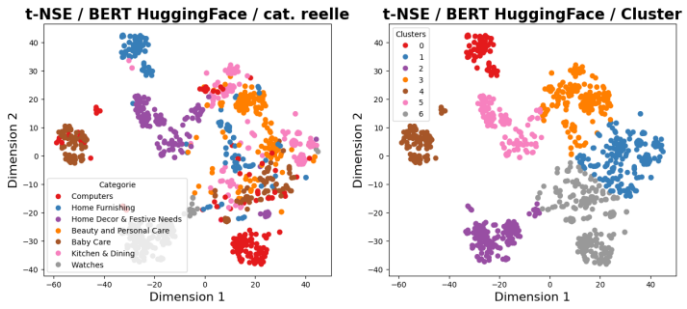
ARI = 0,39

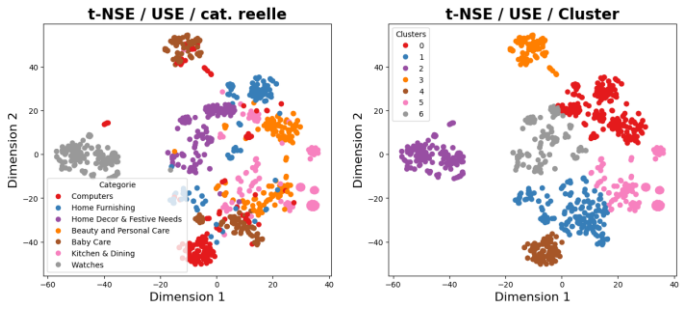
Google USE

K

ARI = 0,42

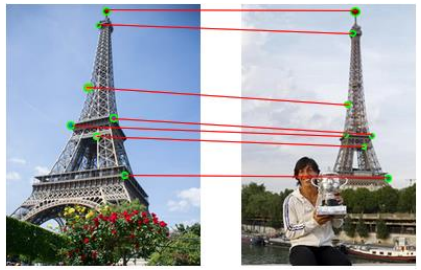




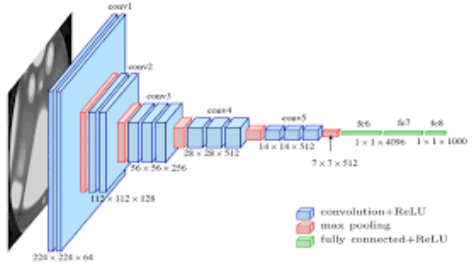


Etude de faisabilité

SIFT
Scale-Invariant Feature Transform



VS



Vgg16

PRE PROCESSING

1. Pré-traitement SIFT

2. ACP
(99% variance)

3. t-SNE

4. Clustering

5. ARI

PRE PROCESSING

1. Prétraitement
CNN avec Transfer learning
VGG16

Etude de faisabilité

X 1_050

	image	product_name	label
0	55b85ea15a1536d46b7190ad6fff8ce7.jpg	Home Furnishing	4
1	7b72c92c2f6c40268628ec5f14c6d590.jpg	Baby Care	0
2	64d5d4a258243731dc7bbb1eef49ad74.jpg	Baby Care	0

1. Lecture de l'image

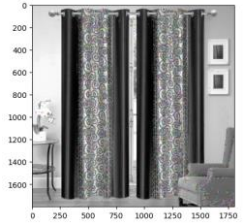
2. Conversion en gris

3. Filtre bruit « poivre et sel »

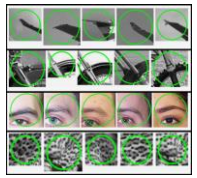
4. Flou (réduction bruit HF)

5. Egalisation histo. (amélioration contraste)

6. Réduction taille /2



7. Extraction Keypoints & Calcul des descripteurs SIFT



8. Création des visuals words

9/ Bag-visual-words

SIFT
Scale-Invariant Feature Transform

OpenCV

PRE PROCESSING

Etude de faisabilité

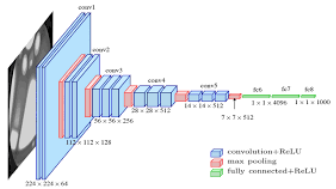
X 1_050

	image	product_name	label
0	55b85ea15a1536d46b7190ad6fff8ce7.jpg	Home Furnishing	4
1	7b72c92c2f6c40268628ec5f14c6d590.jpg	Baby Care	0
2	64d5d4a258243731dc7bbb1eef49ad74.jpg	Baby Care	0

1. Redimensionnement
224 x 224 pixels

2. Normalisation

3. RGB to BGR



4. Création modèle
TL VGG16

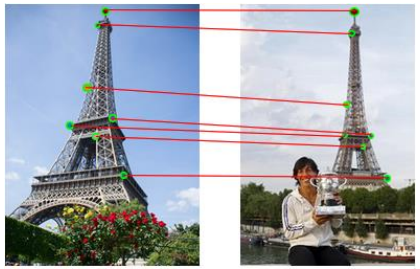
Sauf les 2 dernières couches fully connected

5. Prédiction du modèle

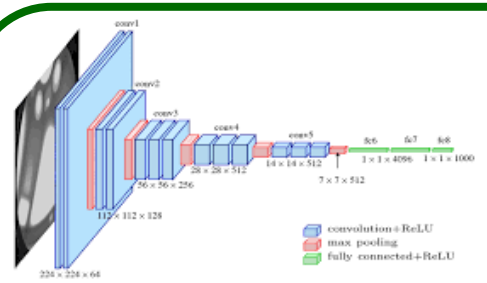
PRE PROCESSING

Etude de faisabilité

SIFT
Scale-Invariant Feature Transform

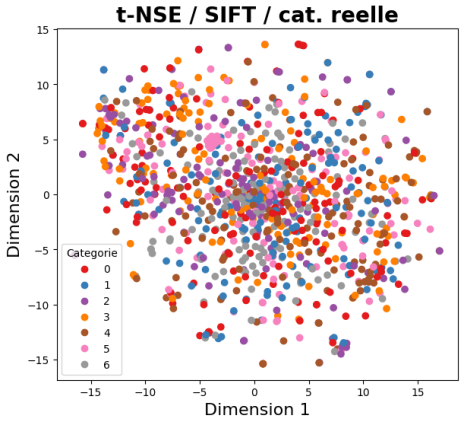


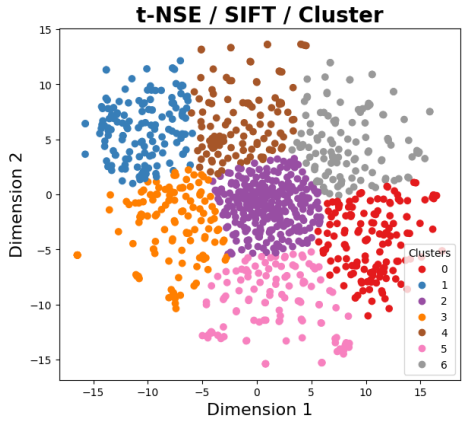
ARI = 0,01

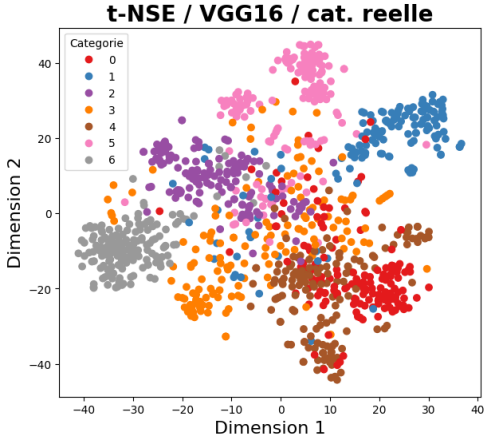


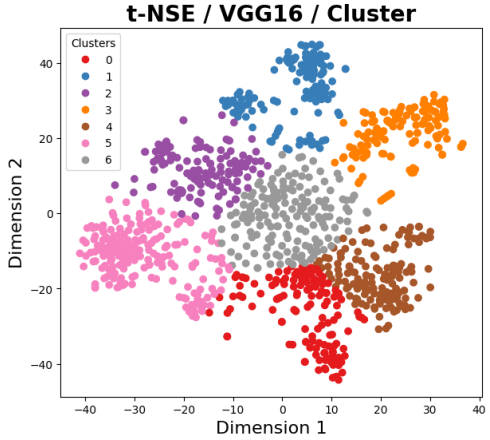
Vgg16

ARI = 0,48



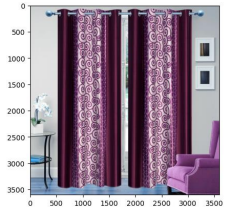






Classification supervisée

2. Réaliser une classification supervisée à partir des images



X 1_050

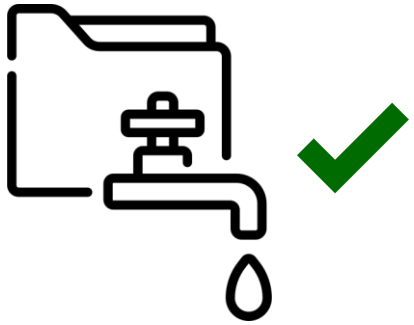
	image	product_name	label
0	55b85ea15a1536d46b7190ad6fff8ce7.jpg	Home Furnishing	4
1	7b72c92c2f6c40268628ec5f14c6d590.jpg	Baby Care	0
2	64d5d4a258243731dc7bbb1eef49ad74.jpg	Baby Care	0

1. Séparation
Features / **Target**

2. Train test split

75% train *=> pour l'entrainement*
15% validation *=> pour l'optimisation des hyperparamètres*
15% test *=> pour l'évaluation*

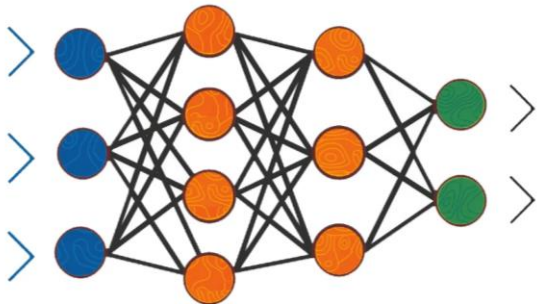
3. Test de différents modèles



Classification supervisée

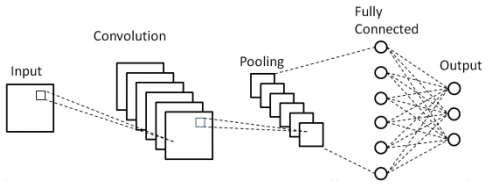
3. Test de différents modèles

?



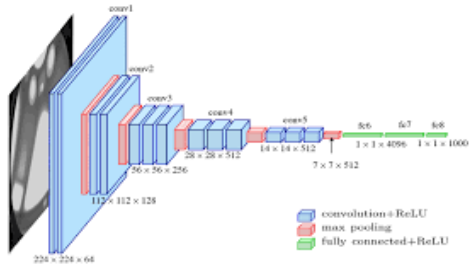
MultiLayer Perceptron (MLP)

VS



Convolutional Neural Network (CNN)

VS



CNN + 

Classification supervisée

3. Test de différents modèles

1. Prétraitement

2. Création modèle

```
fonction_perte = 'categorical_crossentropy'
optimiseur = 'rmsprop'
metric_eval = 'accuracy'
```

3. Optimisation hyperparamètres

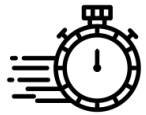
```
# Définir les batch_sizes à tester
batch_sizes = [32, 64, 128, 256]
```

4. Entrainement modèle optimal

=> 15 epochs

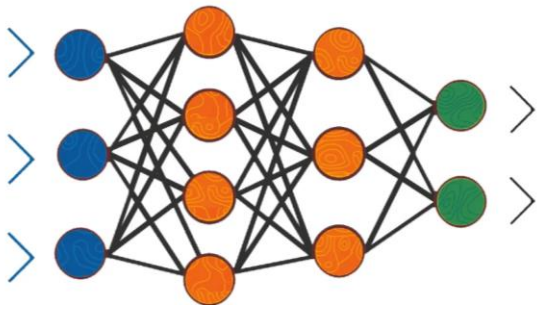
2 Callbacks :
- ModelCheckpoint
- EarlyStopping (patience =3)

5. Evaluation du modèle

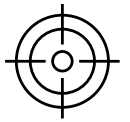


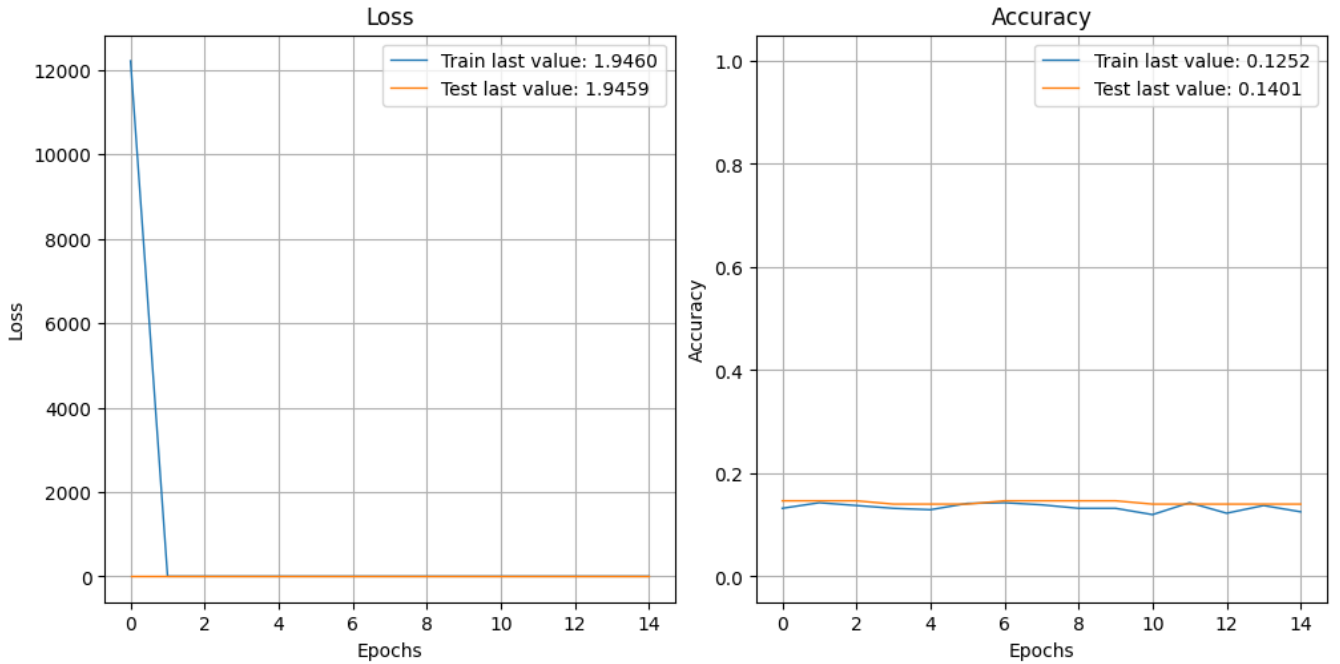
Classification supervisée

Validation Accuracy	
Dummy	0.146497



MultiLayer Perceptron (MLP)

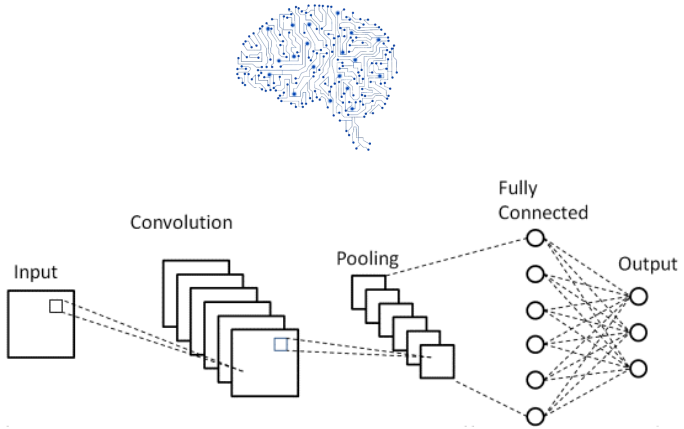




5 couches (27 neurones) :
- Input
- Flatten
- Dense (10 neurones) avec fonction ReLu
- Dense (10 neurones) avec fonction ReLu
- Dense (7 neurones) avec fonction Softmax

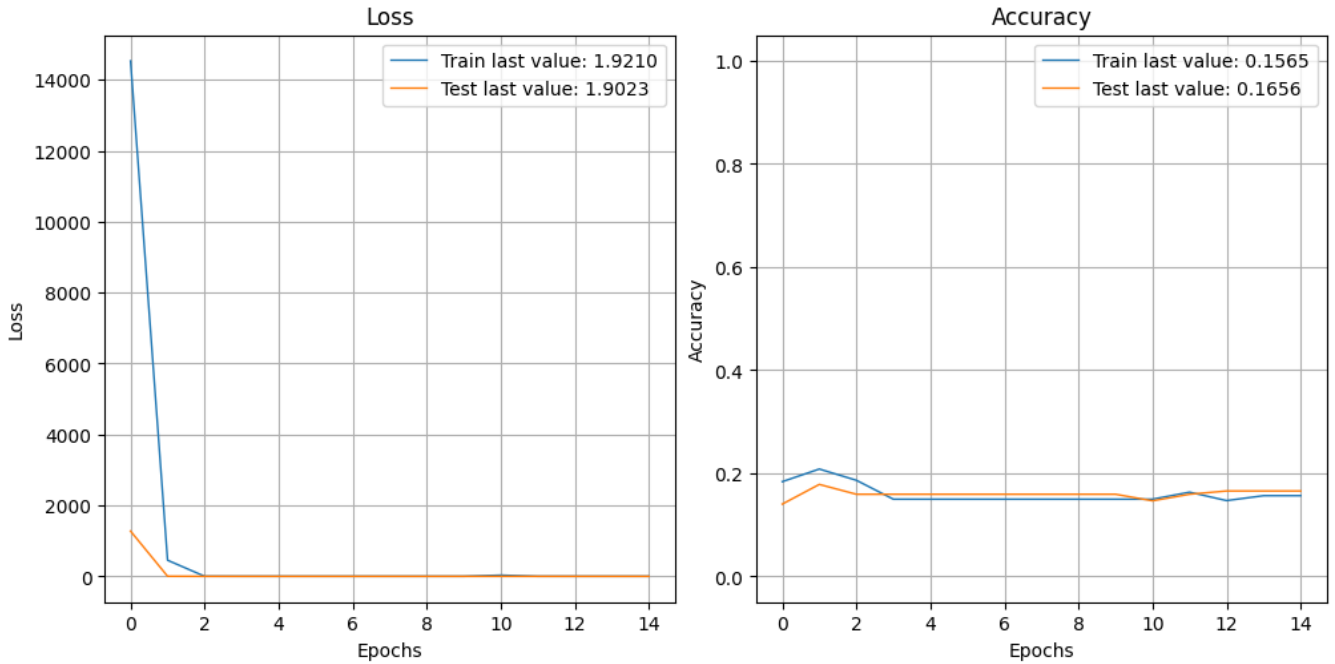
Classification supervisée

Validation Accuracy	
Dummy	0.146497



Convolutional Neural Network (CNN)

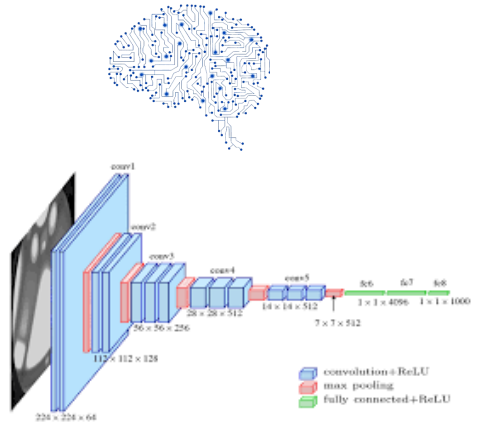
5 couches (63 neurones) :
- Convolution 2D
- MaxPooling 2D
- Flatten
- Dense (56 neurones) avec fonction ReLu
- Dense (7 neurones) avec fonction Softmax



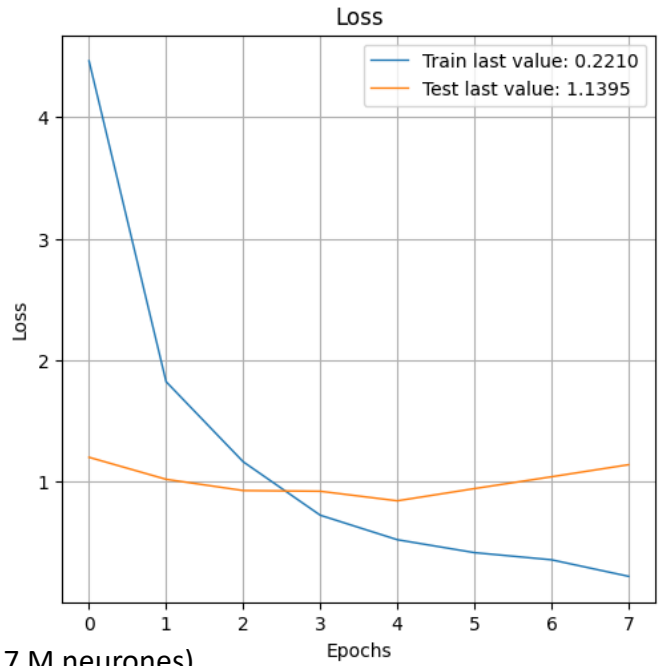
64

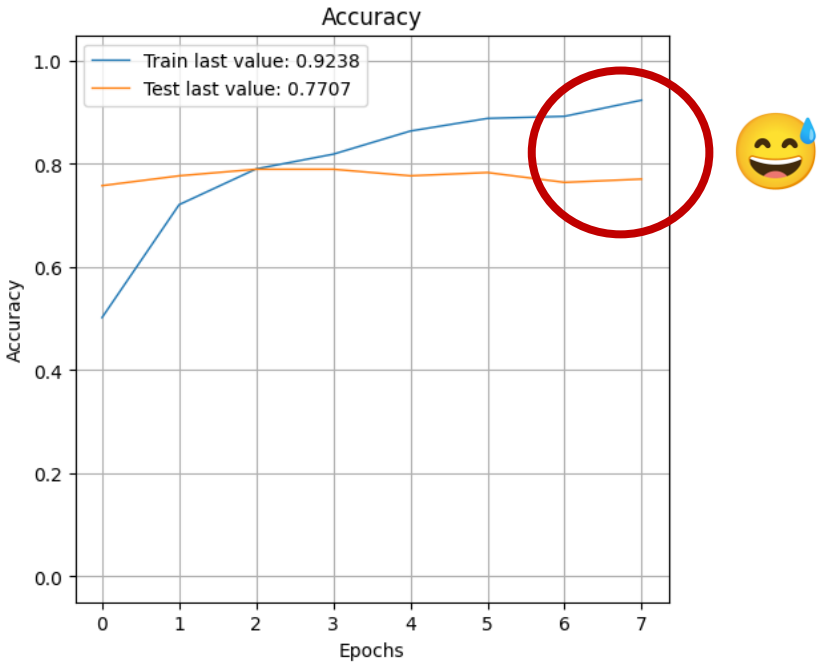
Classification supervisée

Validation Accuracy	
Dummy	0.146497



CNN + Vgg16



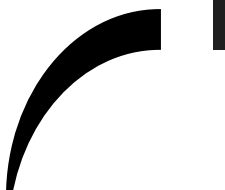


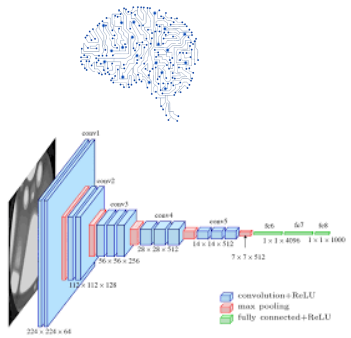
22 couches (≈14,7M neurones) :
- Modèle VGG16 : 5 blocs de convolution/Max pooling (14,7 M neurones)
- GlobalAveragePooling 2D (≈ 25_000 neurones)
- Dense (256 neurones) avec fonction ReLu
- Dropout (0,5)
- Dense (7 neurones) avec fonction Softmax

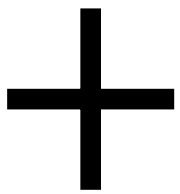
32

Classification supervisée

	Validation Accuracy	Temps (sec)	Epochs	batch_size
mlp	0.140127	10.536837	15	128
cnn	0.165605	62.145483	15	64
tf_vgg16_noDA	0.770701	80.939456	8	32



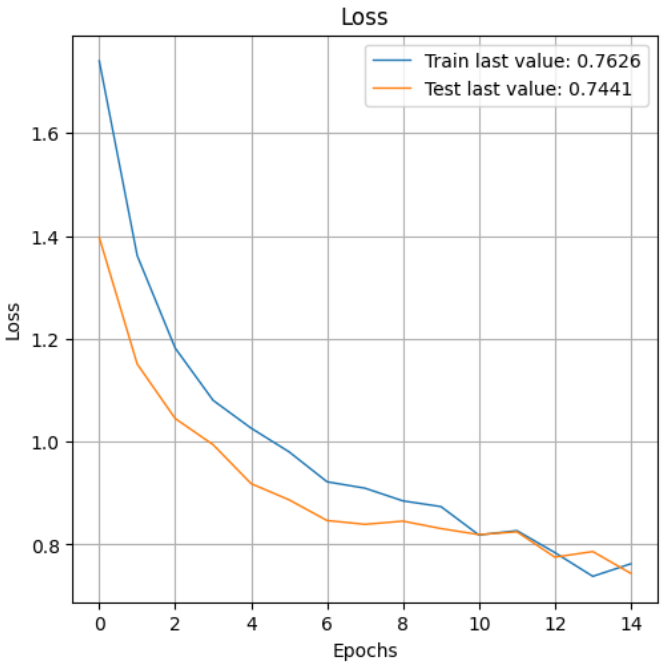


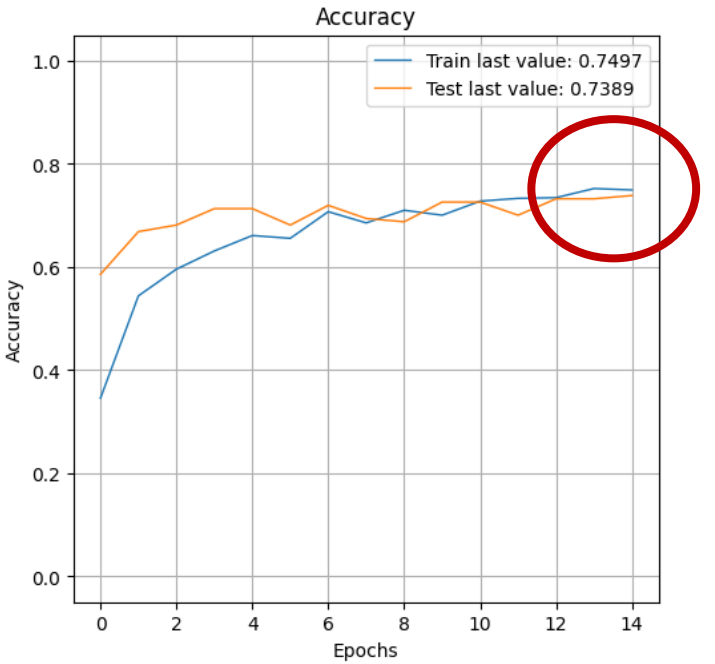


Classification supervisée

DATA AUGMENTATION

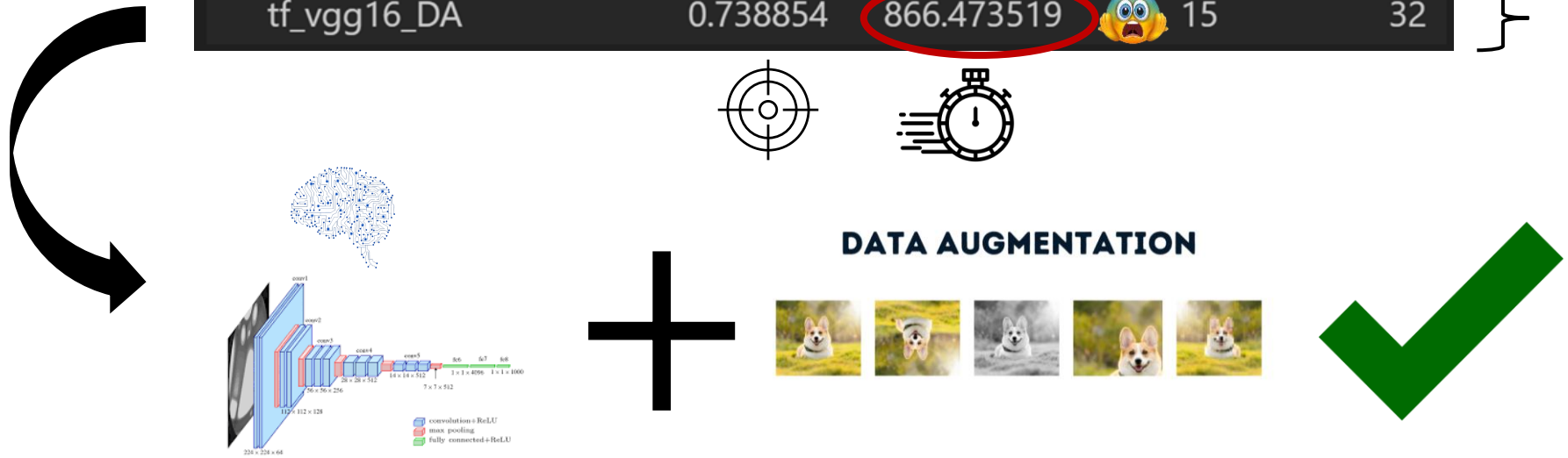
```python
# Créer un générateur de données avec Data Augmentation
datagen = ImageDataGenerator(
    featurewise_center=True,
    featurewise_std_normalization=True,
    rotation_range=20,
    width_shift_range=0.2,
    height_shift_range=0.2,
    horizontal_flip=True,
    brightness_range=(0.5, 1.),
    zoom_range=(0.3, 1.5)
)
```





Classification supervisée

	Validation Accuracy	Temps (sec)	Epochs	batch_size
mlp	0.140127	10.536837	15	128
cnn	0.165605	62.145483	15	64
tf_vgg16_noDA	0.770701	80.939456	8	32
tf_vgg16_DA	0.738854	866.473519	15	32



Test API

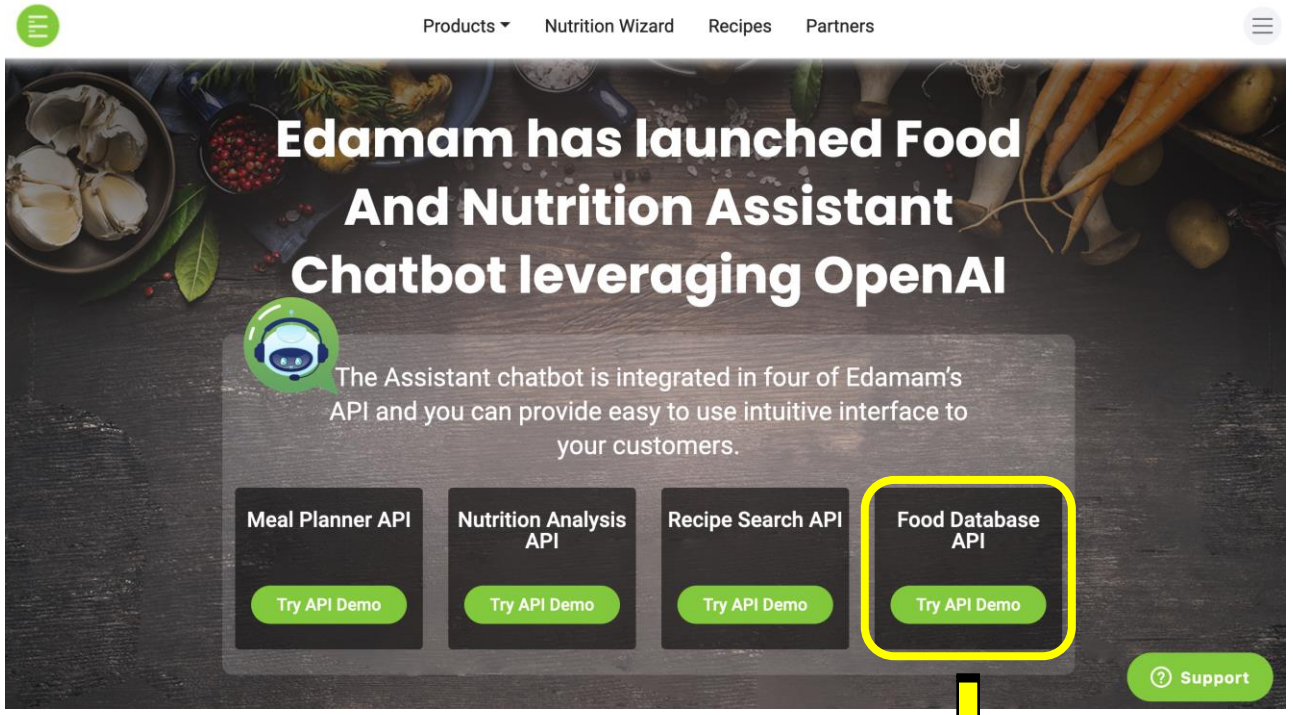
Missions :

3. Collecte de produits « champagne » sur une API.

Voici les 5 grand principe RGPD (Règlement sur la Protection des Données Personnelles) :

1. Finalité
2. Pertinence
3. Durée limitée de conservation
4. Sécurité
5. Droits des personnes

Test API



EDAMAM

```python
# URL de l'API
url = "https://api.edamam.com/api/food-database/v2/parser"

# Clé et ID de l'API
API_KEY = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
API_ID = 'xxxxxxxxxx'
# (volontairement effacés après utilisation pour des raisons de sécurité
# de ce notebook)

# Paramètres de la requête (ingrédient : champagne)
query_string = {"app_key": API_KEY, "app_id": API_ID, "ingr": "champagne"}

# Requête GET (réponse HTTP en format JSON) et transformation en dictionnaire
response_dict = requests.request("GET", url, params=query_string).json()
```

python™

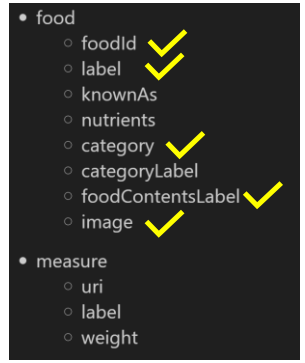
Requests

Test API

EDAMAM


```
{
    "text": "champagne",
    "parsed": [
        {
            "food": {
                "foodId": "food_a656mk2a5dmqb2adiamu6beihduu",
                "label": "Champagne",
                "knownAs": "dry white wine",
                "nutrients": {
                    "ENERC_KCAL": 82.0,
                    "PROCNT": 0.07,
                    "FAT": 0.0,
                    "CHOCDF": 2.6,
                    "FIBTG": 0.0
                },
                "category": "Generic foods",
                "categoryLabel": "food",
                "image": "https://www.edamam.com/food-img/a71/a718cf3c52add522128929f1f324d2ab.jpg"
            }
        }
    ],
"hints": [
    {
        "food": {
            "foodId": "food_a656mk2a5dmqb2adiamu6beihduu",
            "label": "Champagne",
            "knownAs": "dry white wine",
            "nutrients": {
                "ENERC_KCAL": 82.0,
                "PROCNT": 0.07,
                "FAT": 0.0,
                "CHOCDF": 2.6,
                "FIBTG": 0.0
            },
    "_links": {
        "next": {
            "title": "Next page",
            "href": "https://api.edamam.com/api/food-database/v2/parser?session=40&app_key=b52ee0b5d5f28834f1cbedd57ca49692&app_id=f367f20d&ingr=champagne"
        }
    }
}
```

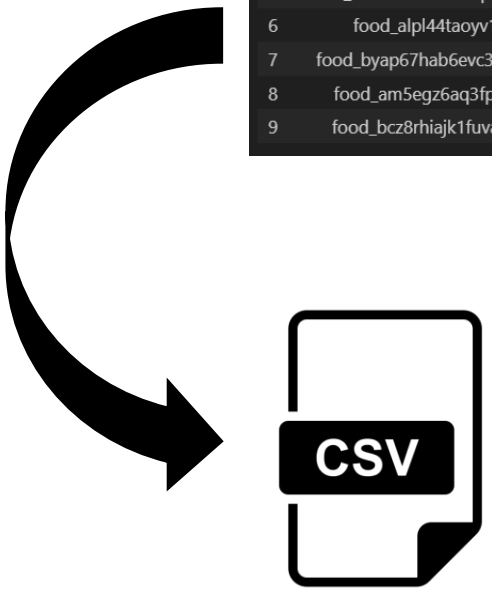

Chaque produit



Test API

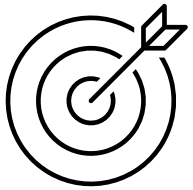
EDAMAM

	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_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	NaN
8	food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
9	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN




```
# Nom du fichier
export_nom_fichier = 'berthe_pierrick_5_export_api_122023.csv'
```

Conclusion



Missions :

1. Faire une étude de la faisabilité d'un moteur de classification automatique. ✓

2. Réaliser une classification supervisée à partir des images ✓

3. Collecte de produits « champagne » sur une API ✓

Limites :

- Problèmes de puissance de calcul pour tester correctement la data augmentation
- Coupler l'analyse du texte et des images combinés peut-être plus performant

Merci pour votre attention

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 –> avril 2024