Projet 6
Classifiez automatiquement des biens de consommation

Guille Anaïs – Parcours Data Scientist

Mentor : Ahmed Tidiane Balde

Sommaire

I- Problématique

- ***Place de marché*** : Entreprise anglophone qui souhaite lancer une marketplace e-commerce

Etudier la faisabilité d'un moteur de classification d'articles en utilisant leur image et leur description

Réaliser une classification supervisée à partir des images

Tester l'API de collecte de produit à base de 'Champagne'

II- Présentation du jeu de données

#	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

- 1 fichier csv de 1050 articles
- 1 dossier de 1050 images
- Aucune donnée dupliquée
- Aucune contrainte de propriété intellectuelle sur les données et les images

III- Etude de faisabilité sur le texte

3.1) Nettoyage des données

1) Sélection des colonnes d'intérêts :

   'product_category_tree' et 'description'

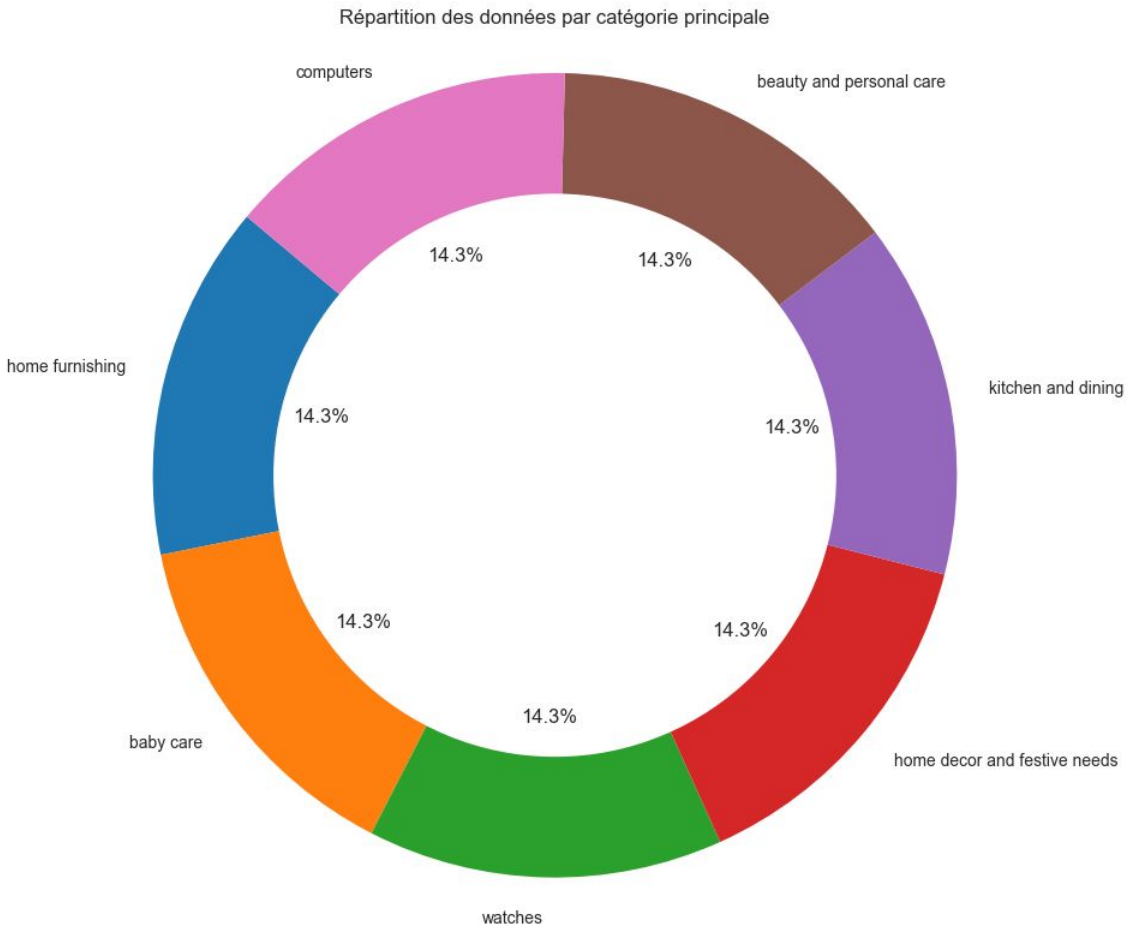
2) Création de la variable 'main_category'

3) Nettoyage de la variable 'main_category' :

   a) Lower()

   b) Remplacer '&' par 'and'

   c) Supprimer les espaces de début et de fin



III- Etude de faisabilité sur le texte

3.1) Nettoyage des données

4) Nettoyage de la variable ‘description’

```
# Lower()
df['final_description'] = df['description'].apply(lambda x : x.lower())

# Elimination des adresses URL // balises HTML // caractères non-ASCII
df['final_description'] = df['final_description'].apply(lambda x : fc.clean_text(x))

#Tokenization
tokenizer = RegexpTokenizer(r'\w+')
df['final_description'] = df['final_description'].apply(tokenizer.tokenize)

# Élimination des stopwords
stop_words = set(stopwords.words('english'))
df['final_description'] = df['final_description'].apply(lambda x: [w for w in x if w not in stop_words])

# Elimination des stopwords (communs aux 7 catégories)
df["final_description"] = df["final_description"].apply(lambda x : [w for w in x if w not in tokens_communs])

# Elimination des mots non anglais
english_words = set(nltk.corpus.words.words())
df['final_description'] = df['final_description'].apply(lambda x: [w for w in x if w in english_words])

# Suppression des mots n'apparaissant qu'une fois
df['final_description'] = df['final_description'].apply(lambda x: [w for w in x if w not in liste_unique_words])

# Elimination des mots possédant un mélange de chiffres et de lettres
df["final_description"] = df["final_description"].apply(lambda x : [w for w in x if w.isalpha()])

# Lemmatization
lemmatizer = WordNetLemmatizer()
df['final_description'] = df['final_description'].apply(lambda x: [lemmatizer.lemmatize(w) for w in x])

# Jointure des tokens
df['final_description'] = df['final_description'].apply(lambda x: " ".join(x))
```

III- Etude de faisabilité sur le texte

3.1) Nettoyage des données

4) Nettoyage de la variable ‘description’

Description d'un produit avant le pré-nettoyage

Key Features of Elegance Polyester Multicolor Abstract Eyelet Door Curtain Floral Curtain,Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) Price: Rs. 899 This curtain enhances the look of the interiors.This curtain is made from 100% high quality polyester fabric.It features an eyelet style stitch with Metal Ring.It makes the room environment romantic and loving.This curtain is ant- wrinkle and anti shrinkage and have elegant apparance.Give your home a bright and modernistic appeal with these designs. The surreal attention is sure to steal hearts. These contemporary eyelet and valance curtains slide smoothly so when you draw them apart first thing in the morning to welcome the bright sun rays you want to wish good morning to the whole world and when you draw them close in the evening, you create the most special moments of joyous beauty given by the soothing prints. Bring home the elegant curtain that softly filters light in your room so that you get the right amount of sunlight.,Specifications of Elegance Polyester Multicolor Abstract Eyelet Door Curtain (213 cm in Height, Pack of 2) General Brand Elegance Designed For Door Type Eyelet Model Name Abstract Polyester Door Curtain Set Of 2 Model ID Duster25 Color Multicolor Dimensions Length 213 cm In the Box Number of Contents in Sales Package Pack of 2 Sales Package 2 Curtains Body & Design Material Polyester

Description du même produit après nettoyage final

elegance polyester multicolor abstract eyelet door curtain floral curtain elegance polyester multicolor abstract eyelet door curtain curtain curtain polyester fabric eyelet stitch metal ring room environment curtain ant wrinkle anti shrinkage elegant give home bright appeal attention sure heart contemporary eyelet slide smoothly draw apart first thing morning welcome bright sun want wish good morning whole draw evening create special beauty given soothing bring home elegant curtain room right amount sunlight elegance polyester multicolor abstract eyelet door curtain elegance door eyelet model name abstract polyester door curtain set model id multicolor length content polyester

III- Etude de faisabilité sur le texte

3.2) Etude de la faisaibilité

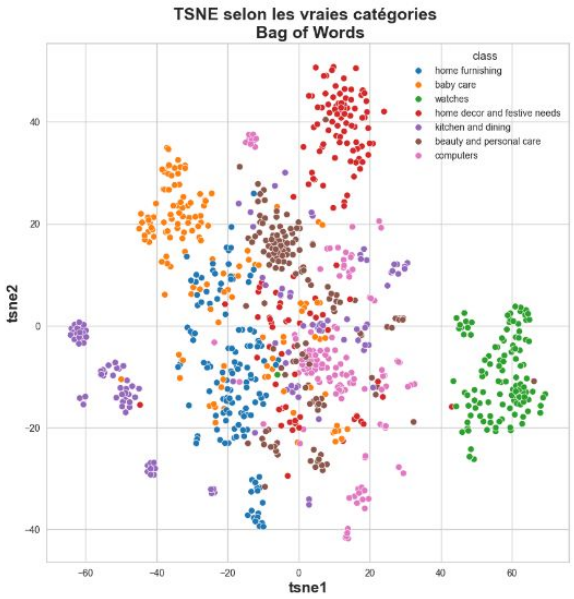
1) **Extraction des features et encodage par diverses méthodes**
   a) Bag of Words : CountVectorizer() et TdidfVectorizer()
   b) Word Embedding : Word2Vec, BERT et USE

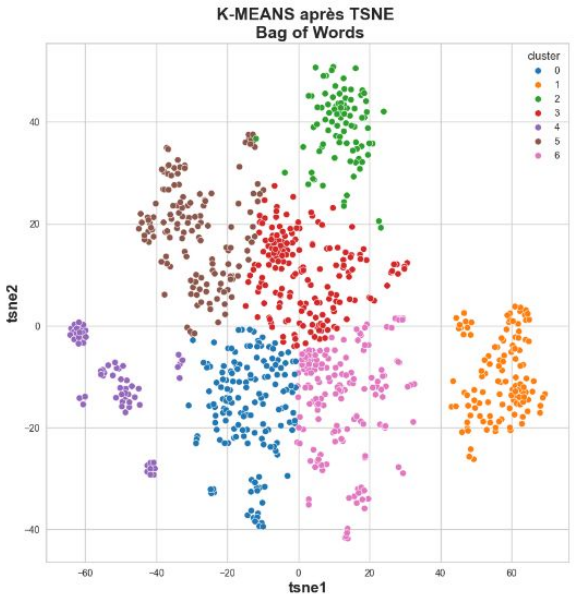
2) **Réduction en 2 dimensions**
   a) PCA
   b) T-SNE

3) **Analyse de la faisabilité**
   a) Visualisation selon les vraies catégories
   b) K-means avec k=7
   c) Calcul du score ARI

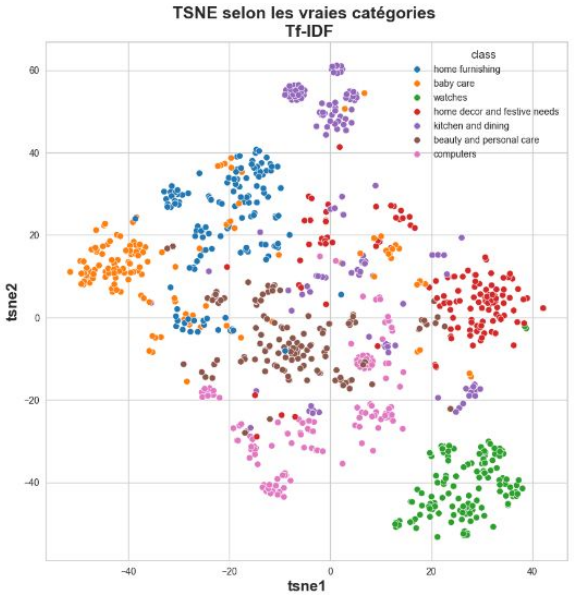
III- Etude de faisabilité sur le texte

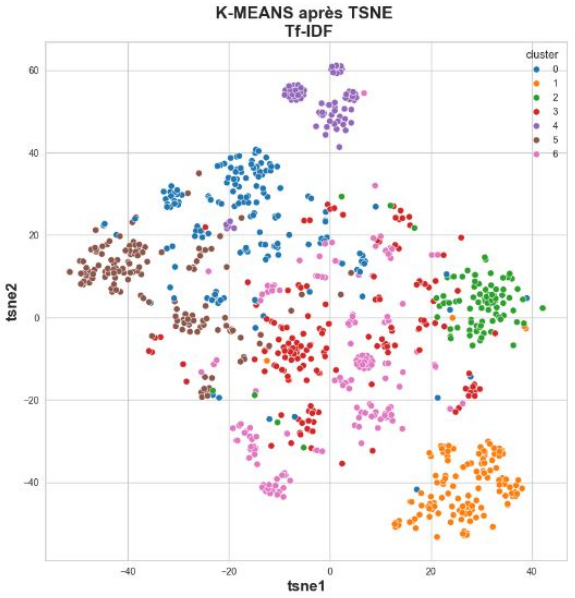
3.2) Etude de la faisaibilité





ARI Score : 0.43

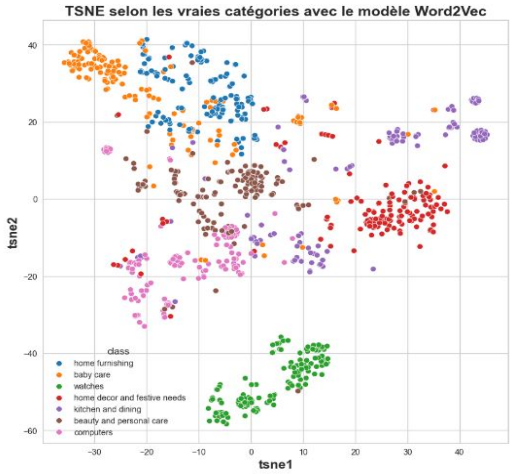


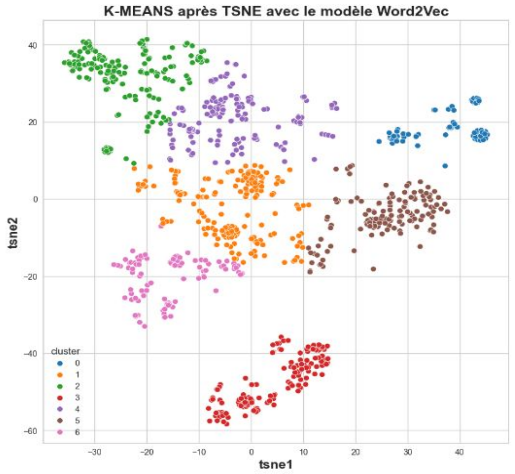


ARI Score : 0.43

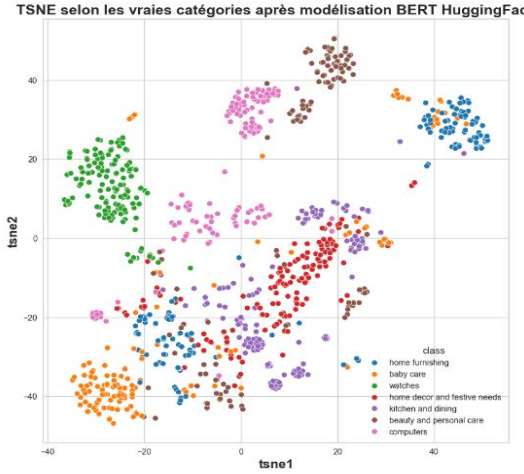
III- Etude de faisabilité sur le texte

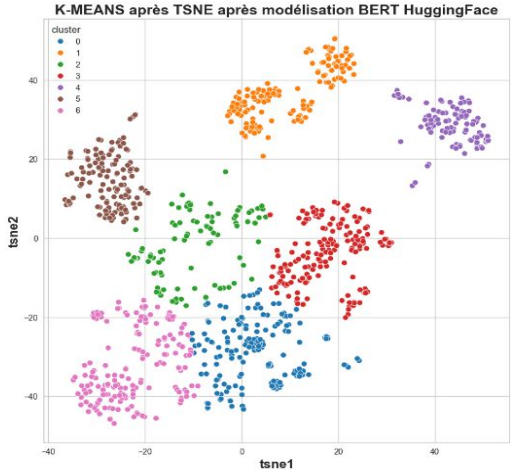
3.2) Etude de la faisaibilité



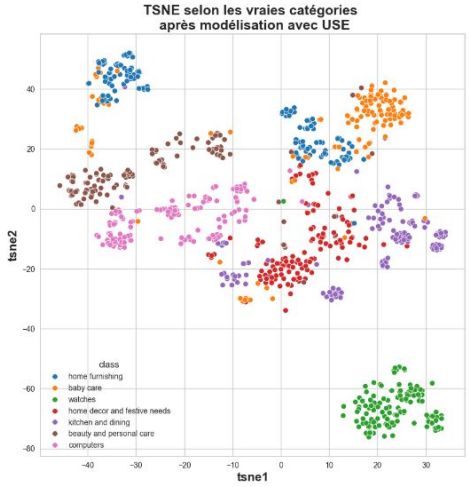


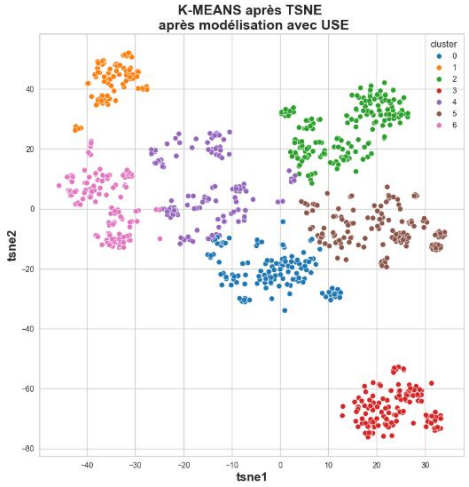
ARI Score : 0.51





ARI Score : 0.37





ARI Score : 0.47

III- Etude de faisabilité sur le texte

3.3) Conclusion

- Faisabilité de la classification des produits selon leur description :

OUI

- Possibilité de réaliser une meilleure classification avec un travail plus poussé

- Les catégories avec le plus de ressemblance sont celles qui posent le plus de problème.

IV- Etude de faisabilité sur les images

4.1) Test de prétraitement des images avec Pillow

Original →

Nuance de gris →

Contraste →

Filtre médian

IV- Etude de faisabilité sur les images

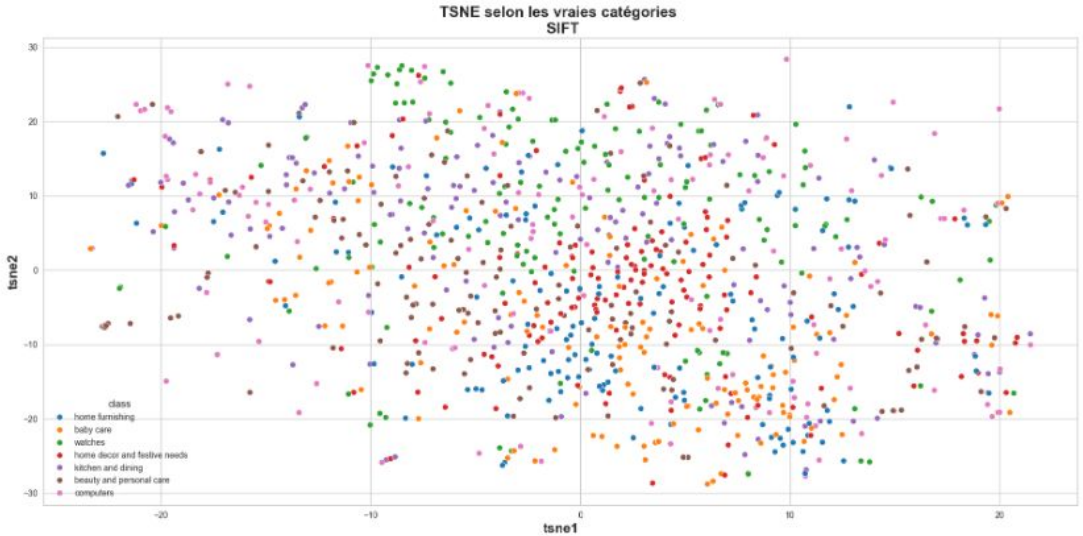
4.2) SIFT

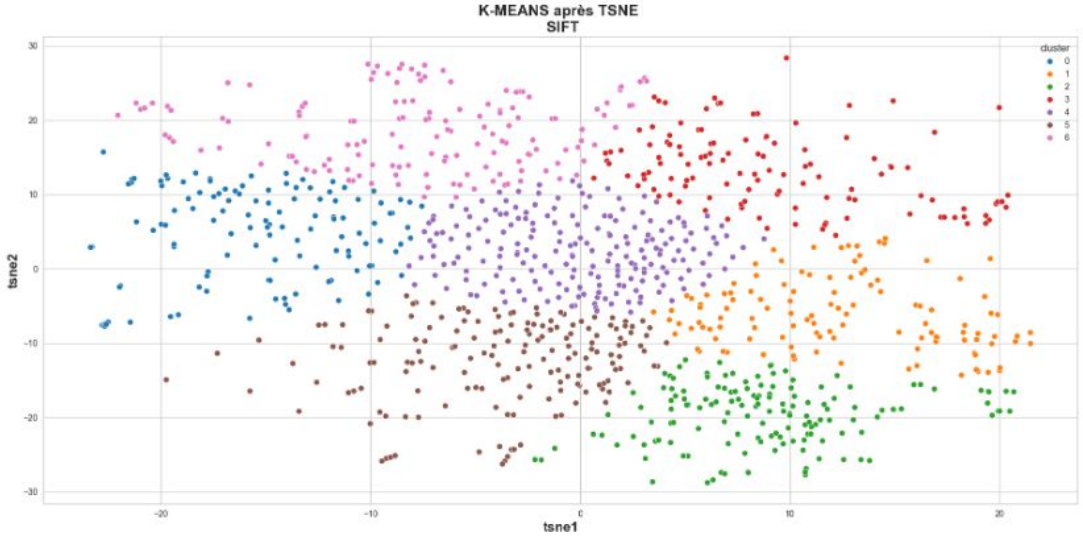
1) **Traitement des images avec Pillow**

2) **Extraction des features**
   a) Création des descripteurs
   b) Création des clusters de descripteurs
   c) Extraction des features

3) **Réduction en 2 dimensions**
   a) PCA
   b) T-SNE

4) **Analyse de la faisabilité**
   a) Visualisation selon les vraies catégories
   b) K-means avec k=7
   c) Calcul du score ARI



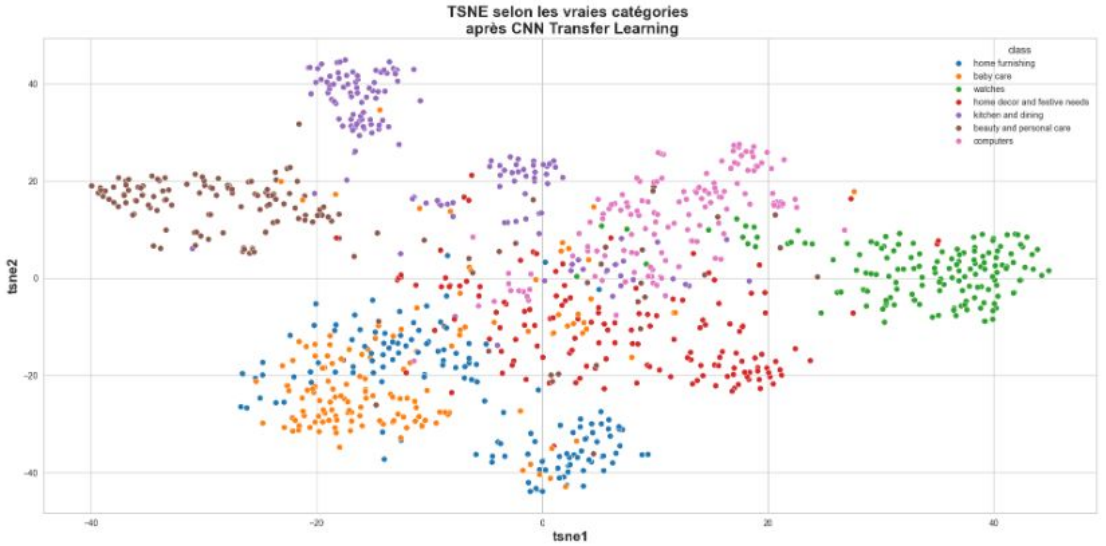


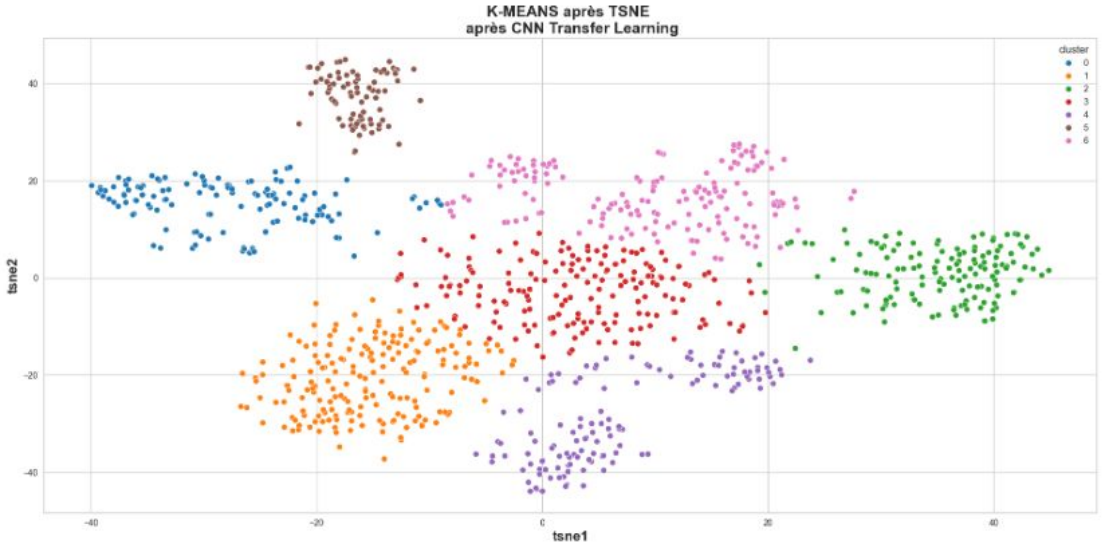
ARI Score : 0.04

IV- Etude de faisabilité sur les images

4.2) CNN Transfer Learning (VGG16)

1) **Création d'un modèle pré-entraîné**
2) **Création des features des images**
3) **Réduction en 2 dimensions**
   a) PCA
   b) T-SNE
4) **Analyse de la faisabilité**
   a) Visualisation selon les vraies catégories
   b) K-means avec k=7
   c) Calcul du score ARI





ARI Score : 0.45

IV- Etude de faisabilité sur les images

4.3) Conclusion

- Faisabilité de la classification des produits selon leur image:

OUI

- Possibilité de réaliser une classification supervisée pour déterminer automatiquement les classes des images

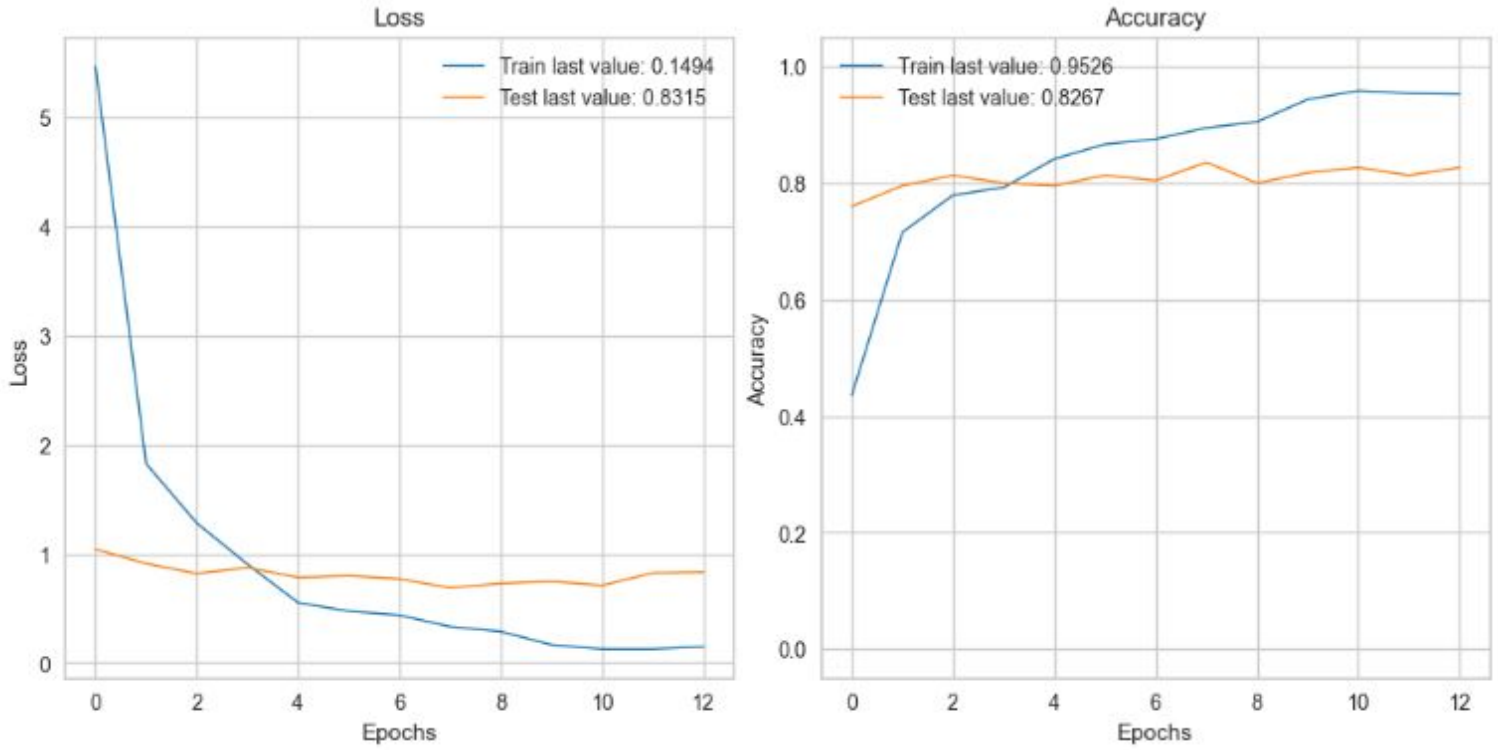
V- Classification supervisée

5.1) Les approches

- **Approche 1 :** Une approche simple par préparation initiale de l'ensemble des images avant classification supervisée
- **Approche 2 :** Une approche par Data Generator, permettant facilement la data augmentation. Les images sont directement récupérées à la volée dans le répertoire des images
- **Approche 3 :** Une approche récente par DataSet, avec data augmentation intégrée au modèle (layer en début de modèle)

V- Classification supervisée

5.2) Approche simple

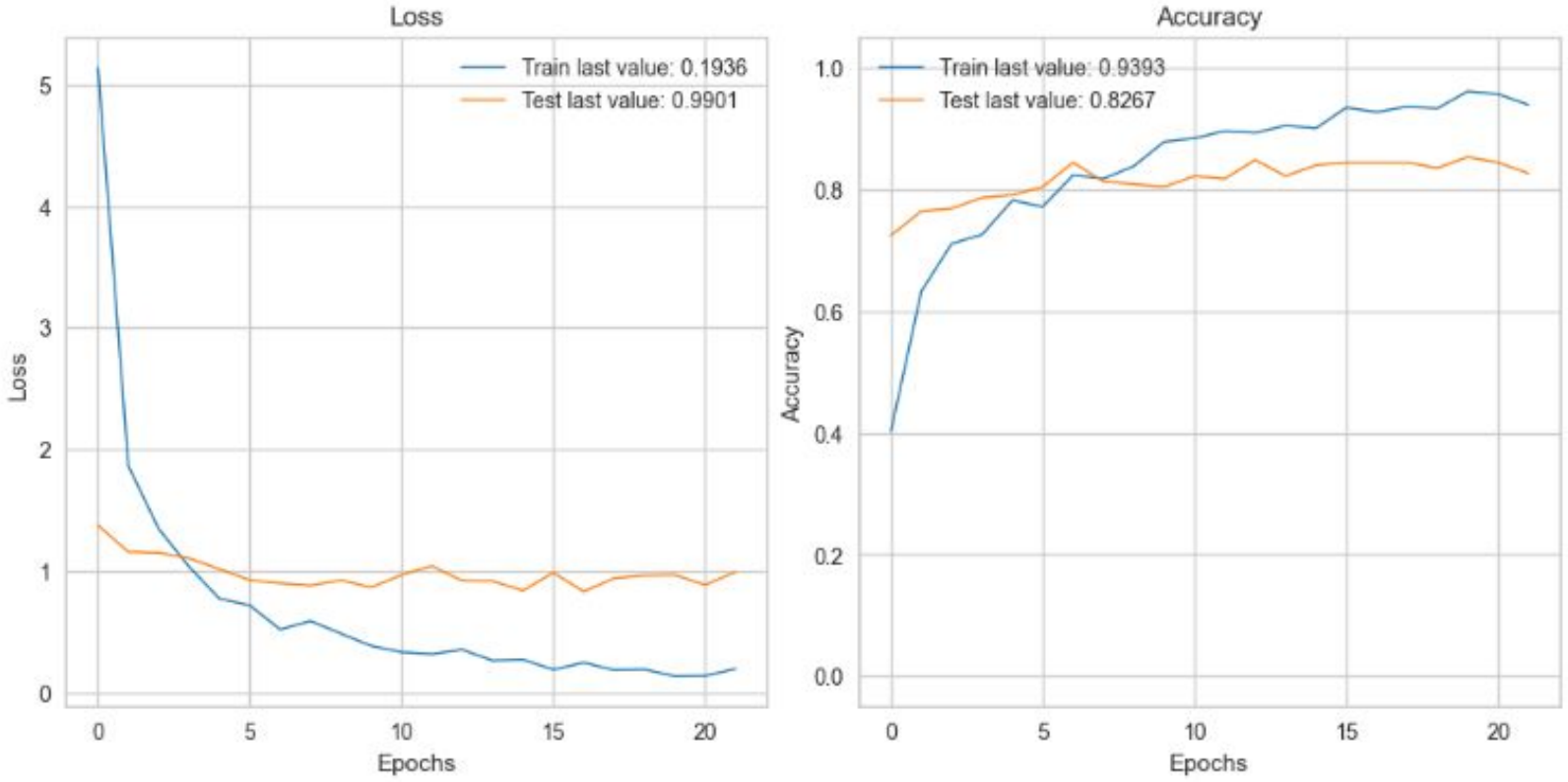


<u>Accuracy</u>

- Training : 0.997
- Validation : 0.83
- Test : 0.82

V- Classification supervisée

5.3) Approche par ImageDataGenerator (data augmentation)

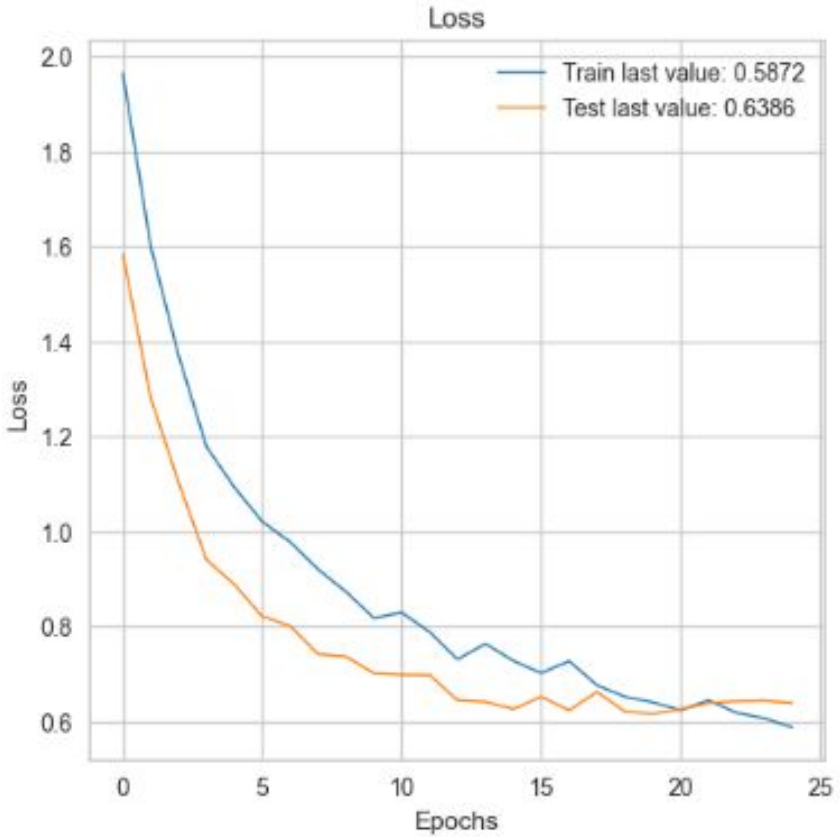


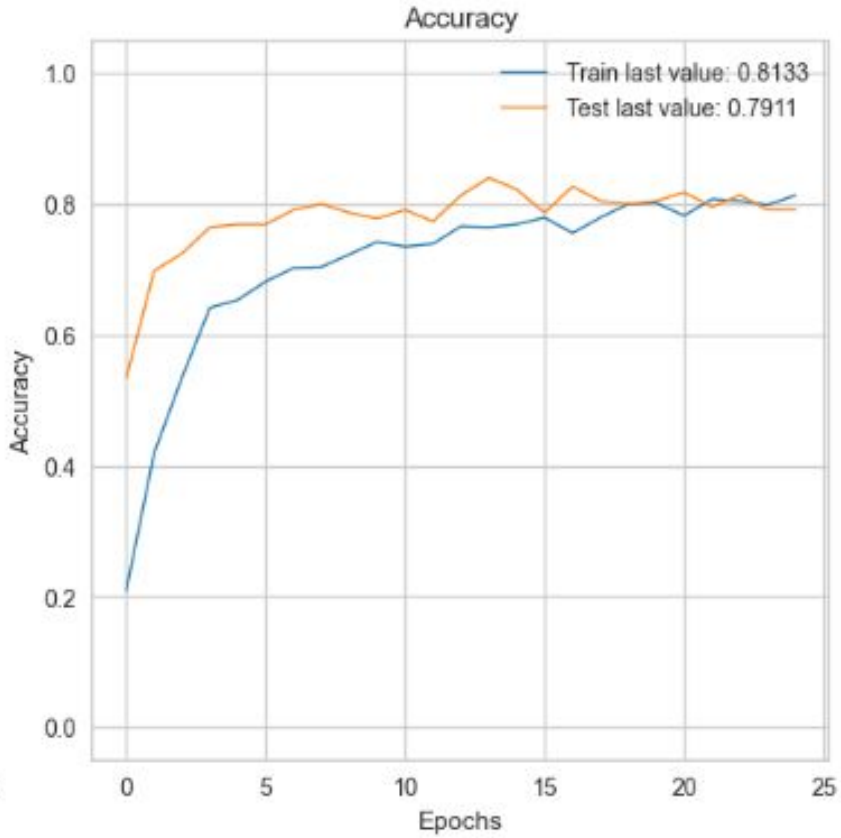
<u>Accuracy</u>

- Training : 0.99
- Validation : 0.84
- Test : 0.78

V- Classification supervisée

5.4) Approche par DataSet (data augmentation)





<u>Accuracy</u>

- Training : 0.86
- Validation : 0.83
- Test : 0.73

V- Classification supervisée

5.5) Conclusion

- **Modèle retenu** : Approche par DataSet avec data augmentation

- Data augmentation permet d'éviter le sur-apprentissage

VI - Test de l'API

- Collecte de produits à base de 'Champagne' via l'API
- Extraction des 10 premiers produits dans un fichier .csv

```
url = "https://edamam-food-and-grocery-database.p.rapidapi.com/api/food-database/v2/parser"

#Filtrer L'ingrédient champagne
querystring = {"ingr": "champagne"}

headers = {
    "X-RapidAPI-Key": "3bbff61965msh1c7c69097fe7bf8p143182jsn5fe6caec4f33",
    "X-RapidAPI-Host": "edamam-food-and-grocery-database.p.rapidapi.com"
}

response = requests.get(url, headers=headers, params=querystring)
data = response.json()
```

On récupère les champs nécessaires, à savoir : foodId, label, category, foodContentsLabel, imag

```
filtered_data = []

if "hints" in data:
    for item in data["hints"]:
        food_data = item.get('food', {})
        relevant_info = {
            'foodId': food_data.get('foodId'),
            'label': food_data.get('label'),
            'category': food_data.get('category'),
            'foodContentsLabel': food_data.get('foodContentsLabel'),
            'image': food_data.get('image')
        }
        filtered_data.append(relevant_info)
```

On récupère les 10 premiers produits

```
with open('champagne_data.csv', 'w', newline='') as csvfile:
    fieldnames = ["foodId", "label", "category", "foodContentsLabel", "image"]
    writer = csv.DictWriter(csvfile, fieldnames=fieldnames)

    writer.writeheader()
    # Prendre seulement les 10 premiers éléments de la liste
    for row in filtered_data[:10]:
        writer.writerow(row)
```

foodId	label	category	foodContentsLabel	image
food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	NaN
food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	NaN
food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN

VI - Test de l'API

Les normes RGPD

1. Ne collecter que les données nécessaire pour atteindre l'objectif
2. Être transparent
3. Organiser et faciliter l'exercice des droits des personnes
4. Fixer des durées de conservation
5. Sécurisez les données et identifiez les risques

VII- Conclusion

- Classification automatiquement d'articles en utilisant image et description est faisable
- Classification supervisée avec VGG16, data augmentation, rmsprop et batch de 64
- Data augmentation permet d'éviter le sur-apprentissage
- Collecte des 10 premiers produits à base de champagne via l'API

Merci pour votre attention