Olivier RAYMOND
Olivier.raymond.17@eigsi.fr

Mentor: Khalid Moustapha Askia
Jury: Zied Jemai

Formation Data Scientist

PROJET N°6 – Classifier automatiquement des biens de consommation

SOMMAIRE

Contexte

I. Etude de faisabilité

II. Regression Supervisée

Conclusion

Contexte

Vendeurs proposent articles en postant photo et description (MANUEL)

Problématique: Catégorisation peu fiable ET augmentation du volume des articles

Proposition: Automatiser la tâche d'attribution de la catégorie.

Objectifs:

1. Réaliser une étude de faisabilité d'un moteur de classification automatique d'articles via : 1. des descriptions et 2. des images.
2. Réaliser une classification supervisée pour les textes et les images.
3. Utilisation d'un API pour retrouver des informations sur des produits contenant du Champagne.

Contexte Dataset

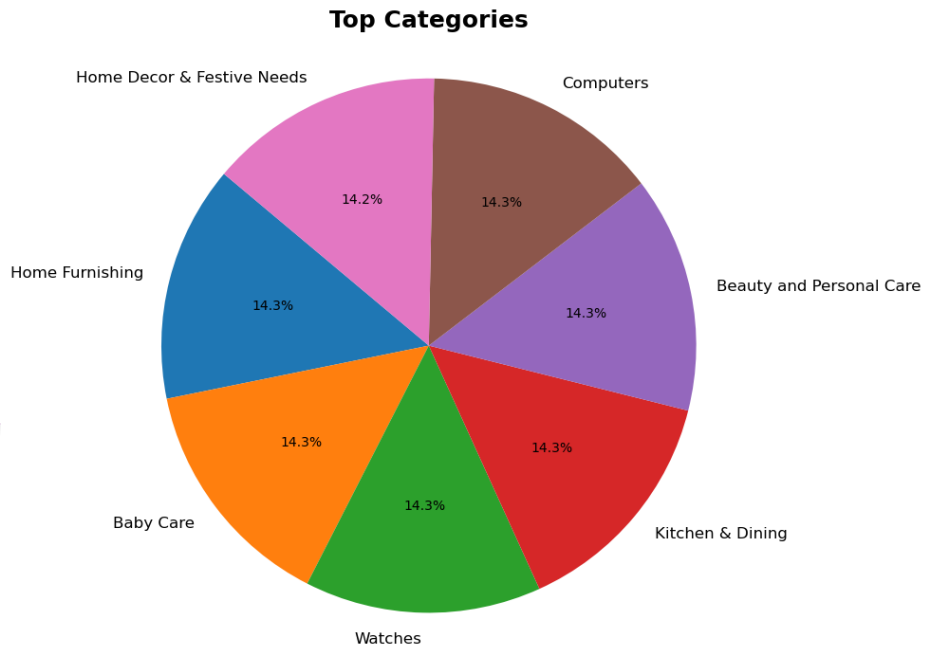
place de marché

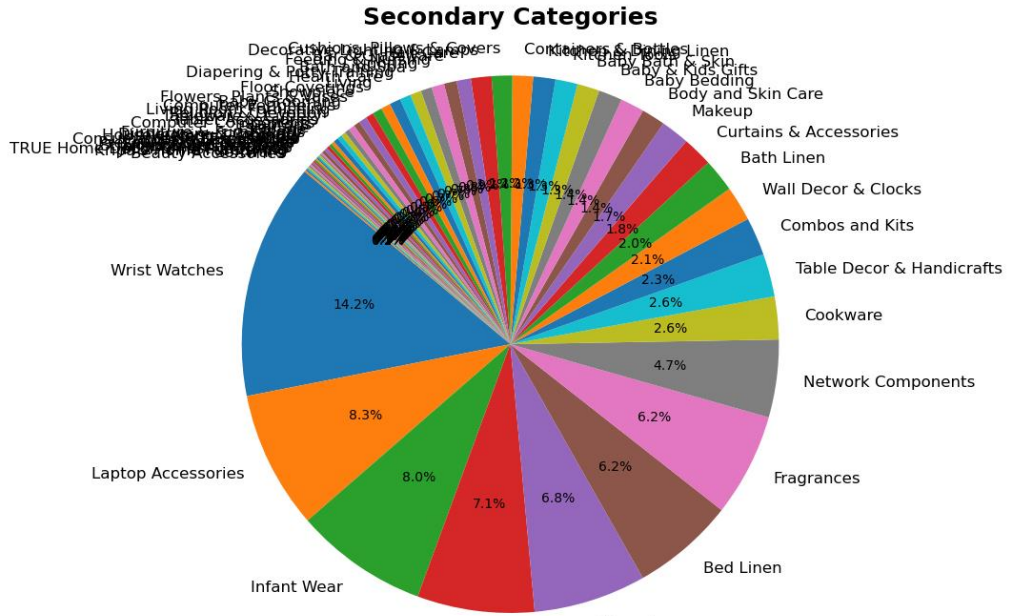
uniq_id	302c95f6eae5f4ce217fcedc4ef91262
crawl_timestamp	2015-12-01 12:40:44 +0000
product_url	http://www.flipkart.com/rastogi-handicrafts-showpiece-20-cm/p/itme5u6n9tgrjbf2?pid=SHIE5U6NBCUHQZTS
product_name	Rastogi Handicrafts Showpiece - 20 cm
product_category_tree	["Home Decor & Festive Needs >> Showpieces >> Rastogi Handicrafts Showpieces"]
pid	SHIE5U6NBCUHQZTS
retail_price	999.0
discounted_price	450.0
image	302c95f6eae5f4ce217fcedc4ef91262.jpg
is_FK_Advantage_product	False
description	Buy Rastogi Handicrafts Showpiece - 20 cm for Rs.450 online. Rastogi Handicrafts Showpiece - 20 cm at best prices with FREE shipping & cash on delivery. Only Genuine Products. 30 Day Replacement Guarantee.
product_rating	No rating available
overall_rating	No rating available
brand	Rastogi Handicrafts
product_specifications	{"product_specification"=>[{"key"=>"Brand", "value"=>"Rastogi Handicrafts"}, {"key"=>"Model Number", "value"=>"GS-AME-S"}, {"key"=>"Type", "value"=>"Nature"}, {"key"=>"Material", "value"=>"Crystal"}, {"key"=>"Color", "value"=>"Purple"}, {"key"=>"Height", "value"=>"20 cm"}, {"key"=>"Width", "value"=>"15 cm"}, {"key"=>"Sales Package", "value"=>"1 Showpiece Figurine"}, {"key"=>"Pack of", "value"=>"1"}]}

Contexte Catégories

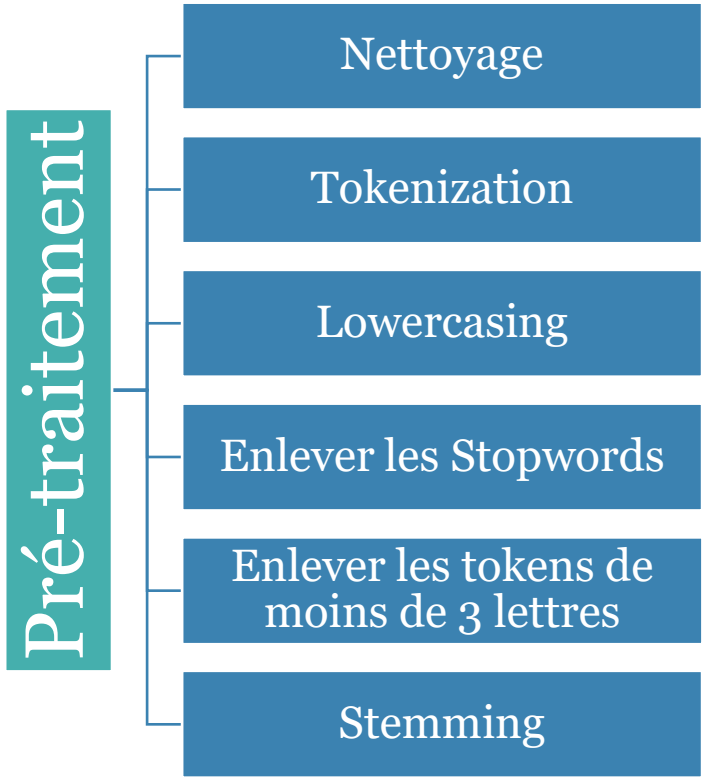
Extraction des catégories

1. TOP catégories (7)
2. Catégories secondaires (62)
3. Catégories tertiaires (240)





II. Faisabilité de classification
Pré-traitement du TEXTE



Phrase d'origine:
"The quick brown fox jumps over the lazy dog."

Nettoyage: The quick brown fox jumps over the lazy dog
Tokenization: ['The'. 'quick'. 'brown'. 'fox'. 'jumps'. 'over'. 'the'. 'lazy'. 'dog']
Lowercasing: ['the'. 'quick'. 'brown'. 'fox'. 'jumps'. 'over'. 'the'. 'lazy'. 'dog']
Enlever stopwords: ['quick'. 'brown'. 'fox'. 'jumps'. 'lazy'. 'dog']
Enlever les tokens de moins de 3 : ['quick'. 'brown'. 'fox'. 'jumps'. 'lazy'. 'dog']
Stemming: ['quick'. 'brown'. 'fox'. 'jump'. 'lazi'. 'dog']

Phrase nettoyée:
quick brown fox jump lazi dog

II. Faisabilité de classification
Méthode utilisée

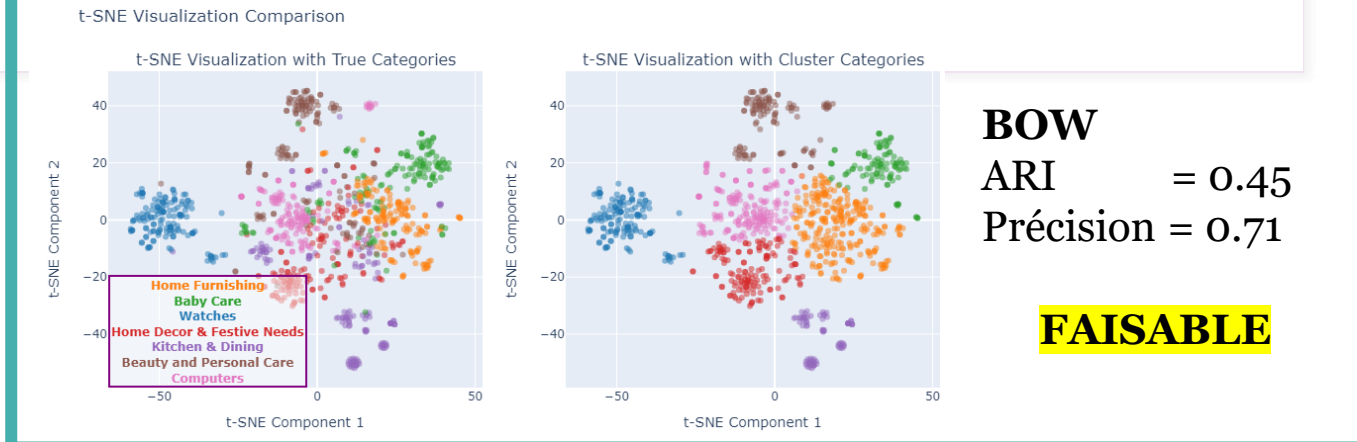
Méthode :

1. **Encodage du texte** (BOW, Bag of N-grams, TF-IDF, Word2Vec, BERT, USE)
2. **Réduction de la dimensionnalité** des caractéristiques textuelles à 2D avec T-SNE pour la visualisation.
3. **Entraînement d'un classificateur KMEANS** sur l'ensemble du dataset.
4. **Évaluation du classificateur (ARI, précision)**.
5. **Smart mapping**
6. Création **d'une visualisation par nuage de points** comparant les étiquettes réelles et prédites.

II. Faisabilité de classification TEXTE

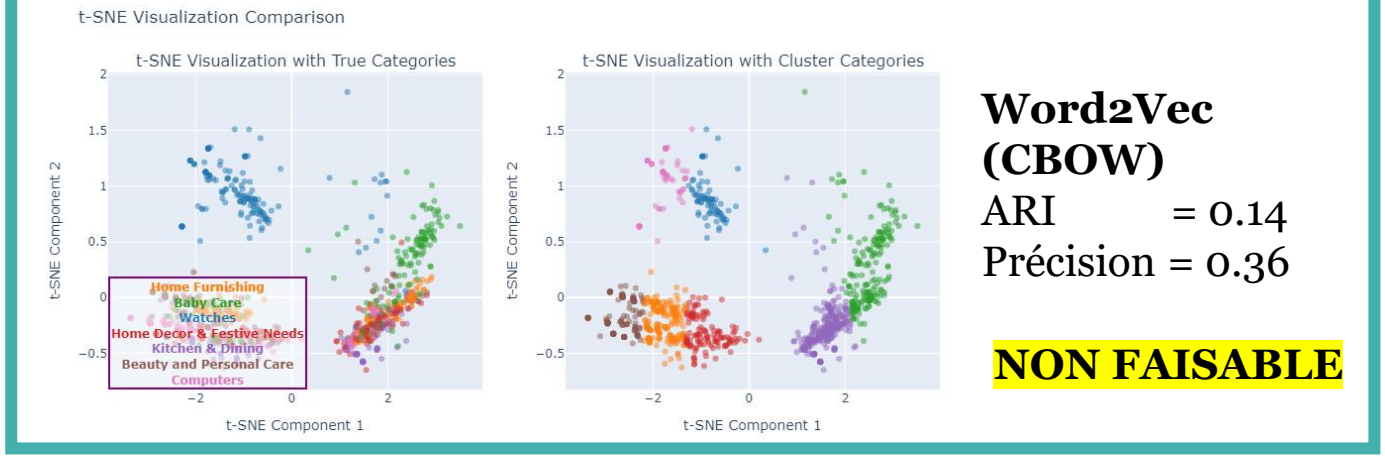
Deux possibilités:

1. **<u>Analyse visuelle</u>** du dataset réduit à 2 dimensions (PCA ou T-SNE).
2. **<u>Clustering sur la sortie du T-SNE</u>** et évaluation de l'ARI et PRECISION.



BOW
ARI = 0.45
Précision = 0.71

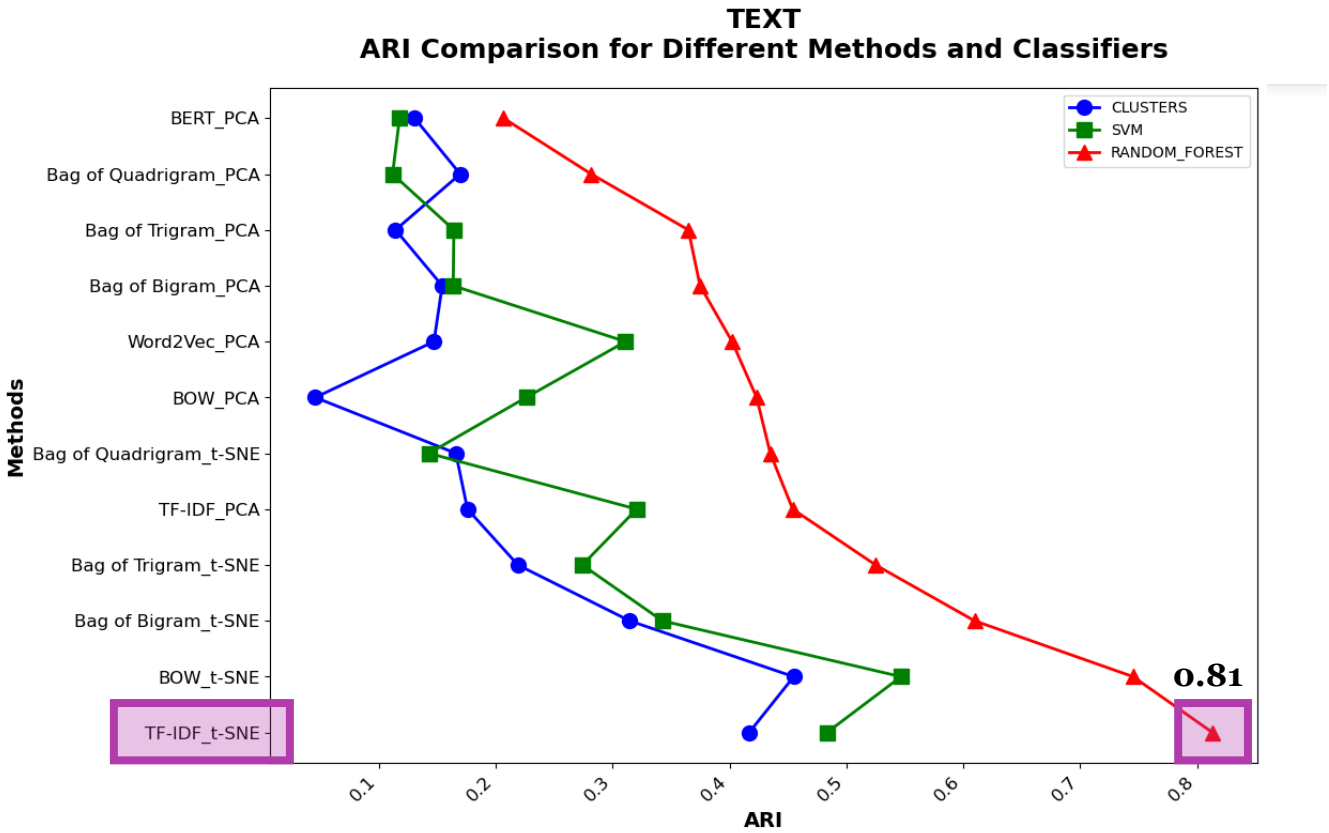
FAISABLE

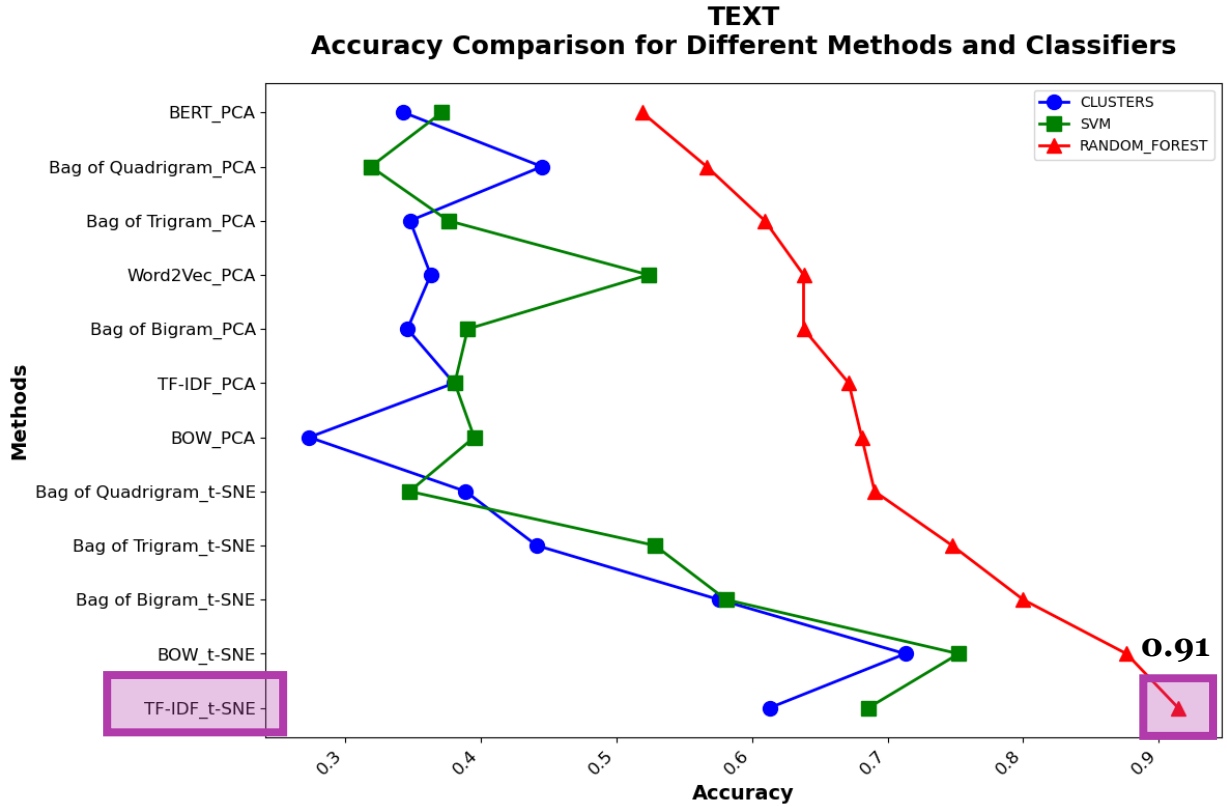


Word2Vec (CBOW)
ARI = 0.14
Précision = 0.36

NON FAISABLE

II. Faisabilité de classification Régression supervisée

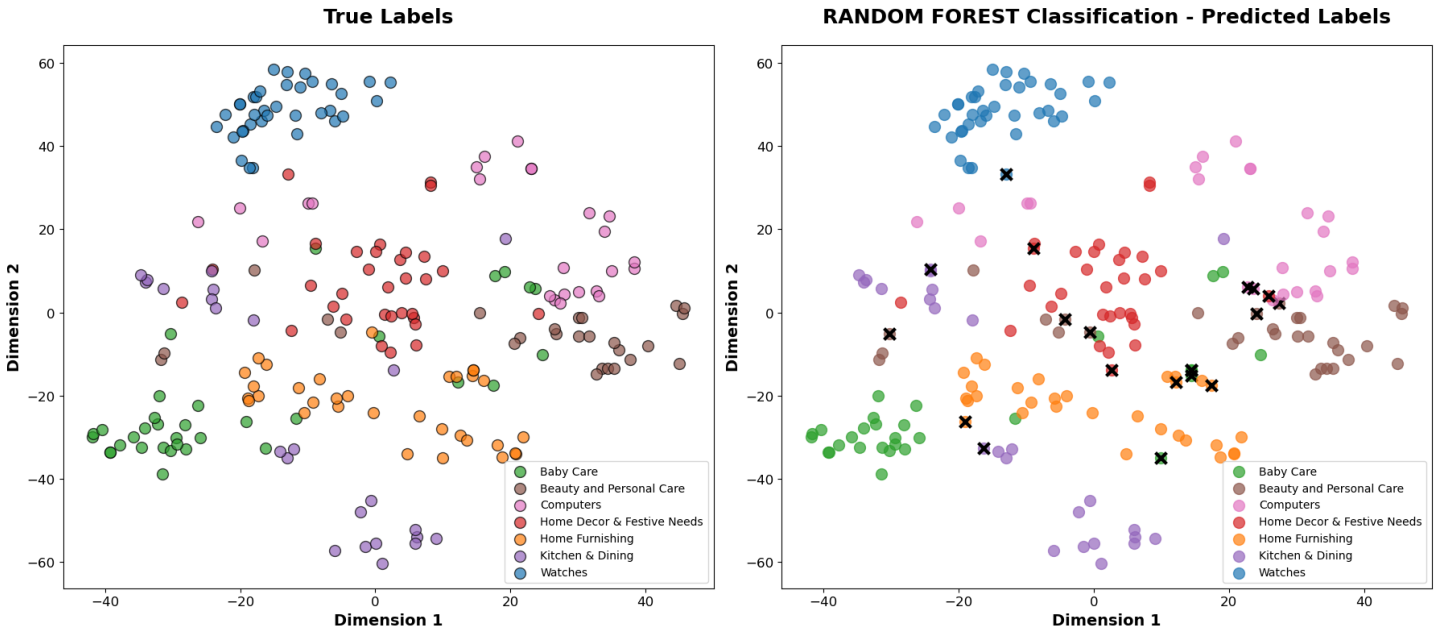


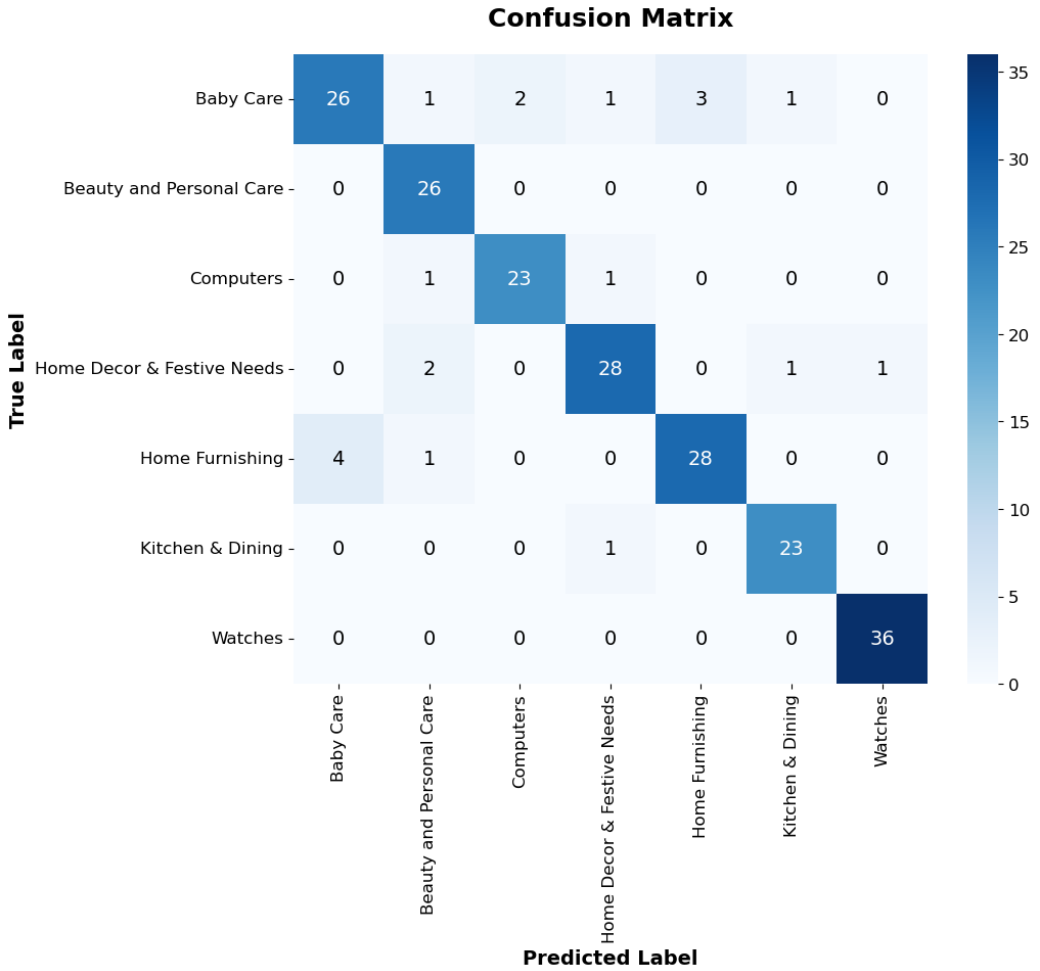


/!\ Résultats SVM & RANDOM FOREST sur les data tests.

Choix : RANDOM FOREST avec TF-IDF et représentation T-SNE

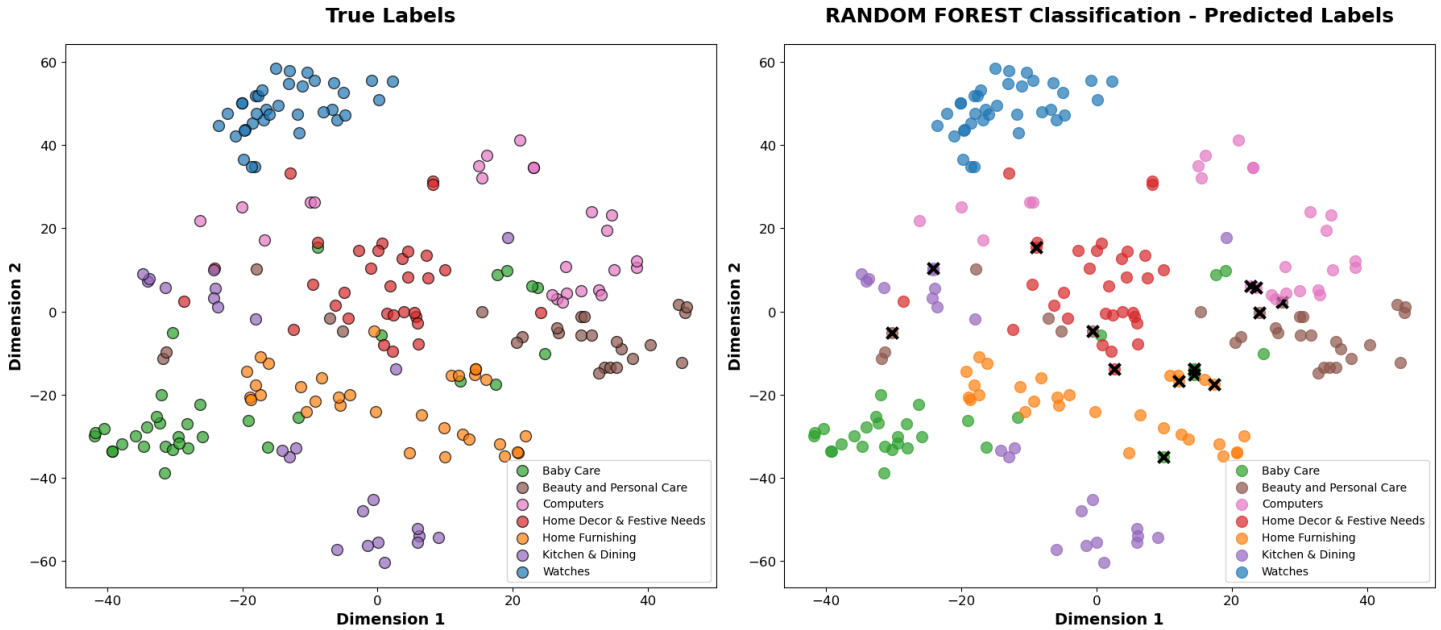
II. Faisabilité de classification Régression supervisée - Résultats





RANDOM FOREST (val par défaut) avec TF-IDF et représentation T-SNE

II. Faisabilité de classification
Régression supervisée – Résultats
Optimisation des hyperparamètres

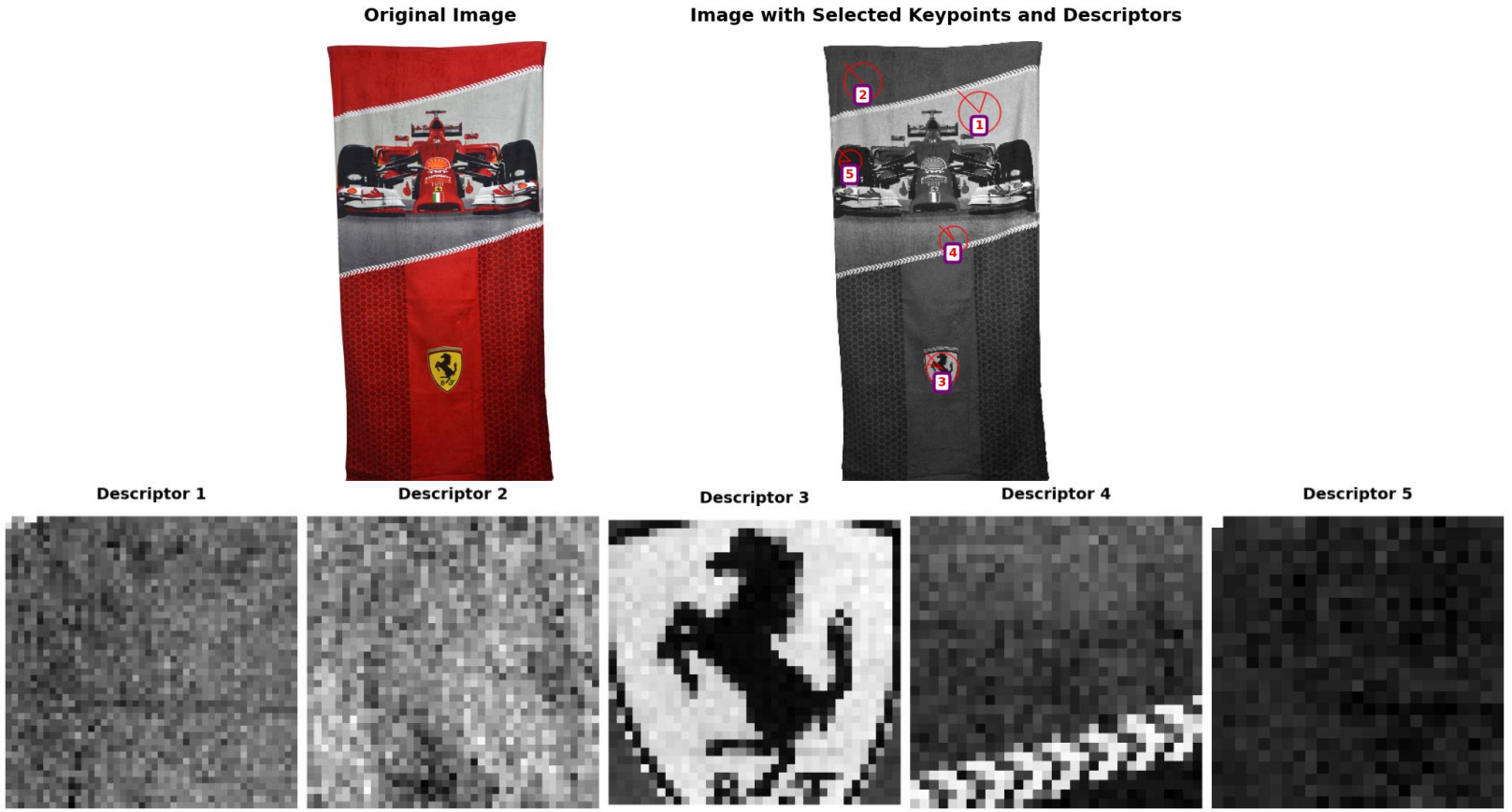


RANDOM FOREST avec TF-IDF et représentation T-SNE

```
param_grid = {
    'n_estimators': [50, 100, 150],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2'],
    'criterion' : ['gini', 'entropy', 'log_loss']
}
```

Précision : 0.91 → 0.93
ARI : 0.81 → 0.84

II. Faisabilité de classification SIFT – Preprocessing & Extraction de features



Représentation de 5 descripteurs uniquement

Méthode de pré-traitement :

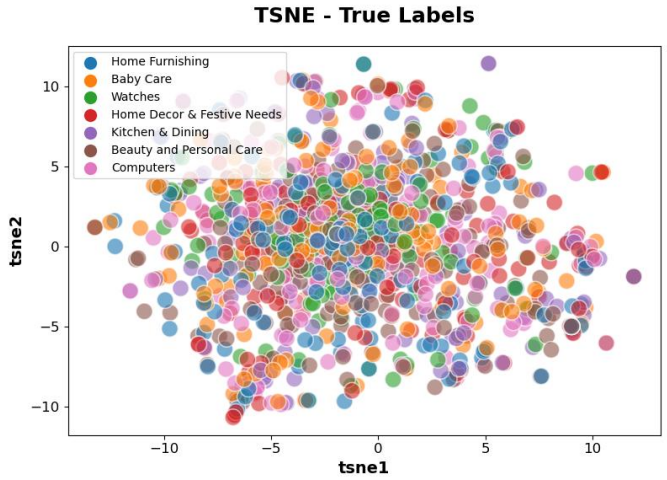
1. Resize
2. Passage en nuance de gris
3. Egalisation de l'histogramme
4. ~~Flou Gaussien~~
5. ~~Tresholding (OTSU ou Adaptive ou Canny)~~

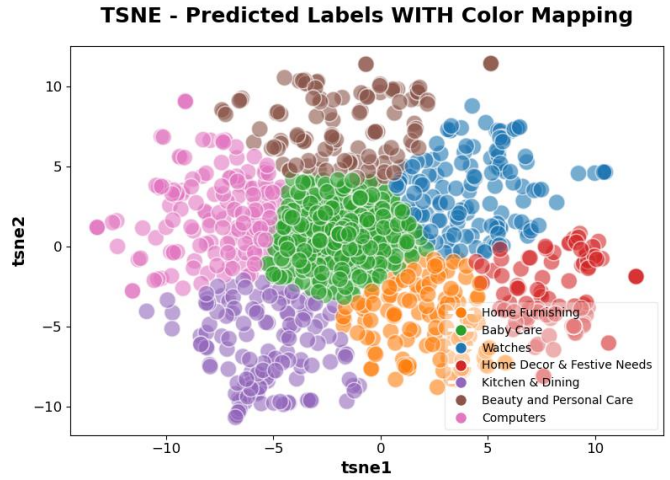
II. Faisabilité de classification
Méthode utilisée

Méthode :

1. **Pré-traitement d'images**
2. **Extraction de descripteurs** (SIFT, ORB, AKAZE, BRISK, FAST)
3. Création de **clusters de descriptors**
4. Création de features (histogram) <-> **Bag of Visual Word (BoVW)**
5. Représentation en 2D (**PCA ou T-SNE**)
6. **Analyse visuelle puis évaluation** de l'ARI et PRECISION

II. Faisabilité de classification (BASIQUE)





Non faisable

→ Passage à un autre type de classification :
Transfert Learning via CNN

Extractor Method	ARI	Accuracy	Temps de traitement	Nb Descripteurs	Nb Clusters
SIFT	0.01	0.21	34.50	(350279, 128)	592
ORB	0.01	0.20	20.63	(436277, 32)	661
AKAZE	0.00	0.19	26.69	(307139, 61)	554
BRISK	0.00	0.19	48.81	(1157843, 64)	1076
FAST	0.00	0.19	47.33	(1848474, 64)	1360

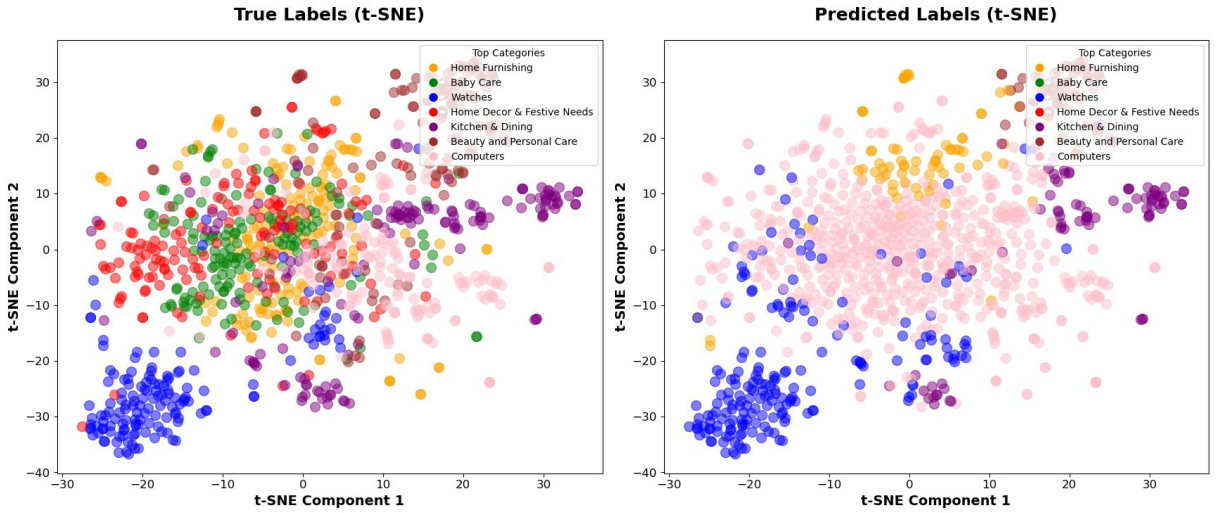
II. Faisabilité de classification (CNN + Transfert Learning ImageNet)

Faisable

→ Amélioration de performances ?

→ **Deux Stratégies**

→ **Data Augmentation**



ARI = 0.20
Précision = 0.46

II. Faisabilité de classification (CNN) STRATEGIES

STRATEGIE 2 : Extraction des features

Adaptée aux petites collections d'images **similaires** à celles de l'entraînement initial, réduisant ainsi le risque de surapprentissage.

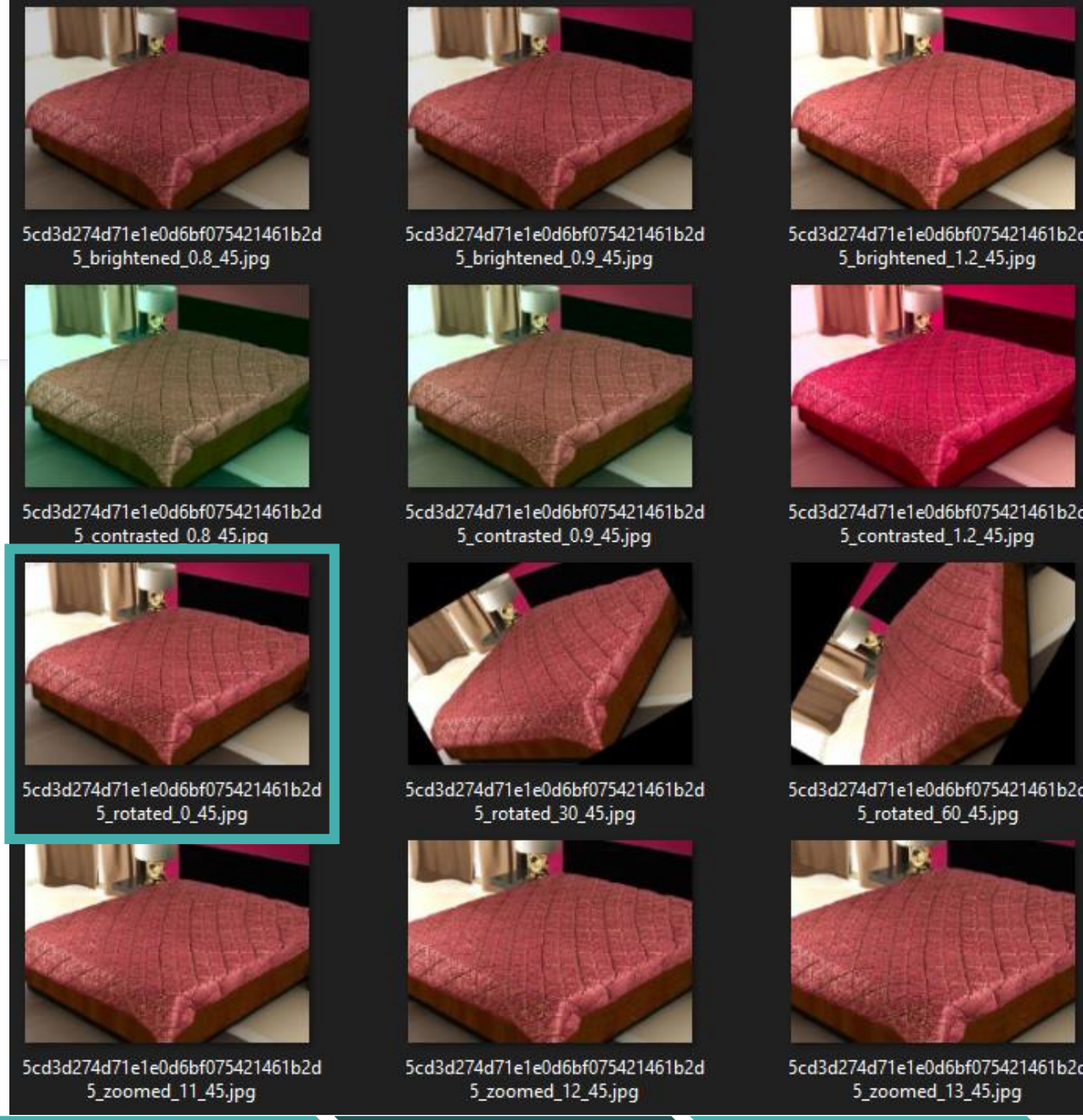
STRATEGIE 3 : Fine-tuning partiel

Idéale pour de petites collections d'images **très différentes** de celles de l'entraînement initial, minimisant ainsi le risque de surapprentissage.

Peu de similarité entre ImageNet et mon dataset
→ **STRATEGIE 3 à privilégier.**

Data Augmentation

- **Augmentation du dataset via l'ajout de :**
  - 3 brightened (0.8 – 0.9 – 1.2)
  - 3 contrasted (0.8 – 0.9 – 1.2)
  - 3 rotations (0° - 30° - 60°)
  - 3 zooms (10% - 20% - 30%)
- **<u>Objectif</u>** :
  - Améliorer la capacité du modèle à généraliser à de nouvelles données tout en évitant le surapprentissage.



II. Faisabilité de classification Régression supervisée

	Classificateur	ARI	Précision
STRATEGY 2	Random Forest – Data Augmented 1 & 2	0.73	0.88
STRATEGY 2	**SVM – Data Augmented 1 & 2**	**0.78**	**0.90**
STRATEGY 2	Random Forest	0.55	0.77
STRATEGY 2	SVM	0.59	0.79
STRATEGY 2	KMEANS	0.18	0.45
STRATEGY 3	CNN – Transfert Learning – Data Augmented 1	0.79	0.9
STRATEGY 3	**CNN – Transfert Learning – Data Augmented 2**	**0.85**	**0.93**
STRATEGY 3	CNN – Transfert Learning	0.5	0.77

Data Split :

- Train (70%), test (15%) et **validation** (15%).

Risque de Data Leakage :

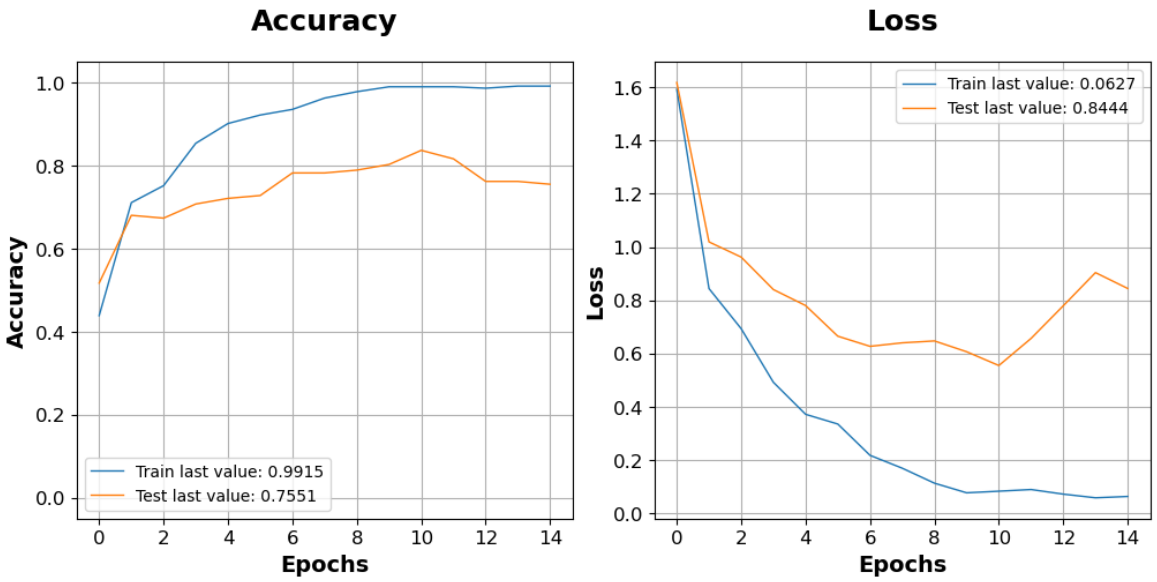
Utilisation de stratify

Data Augmented :

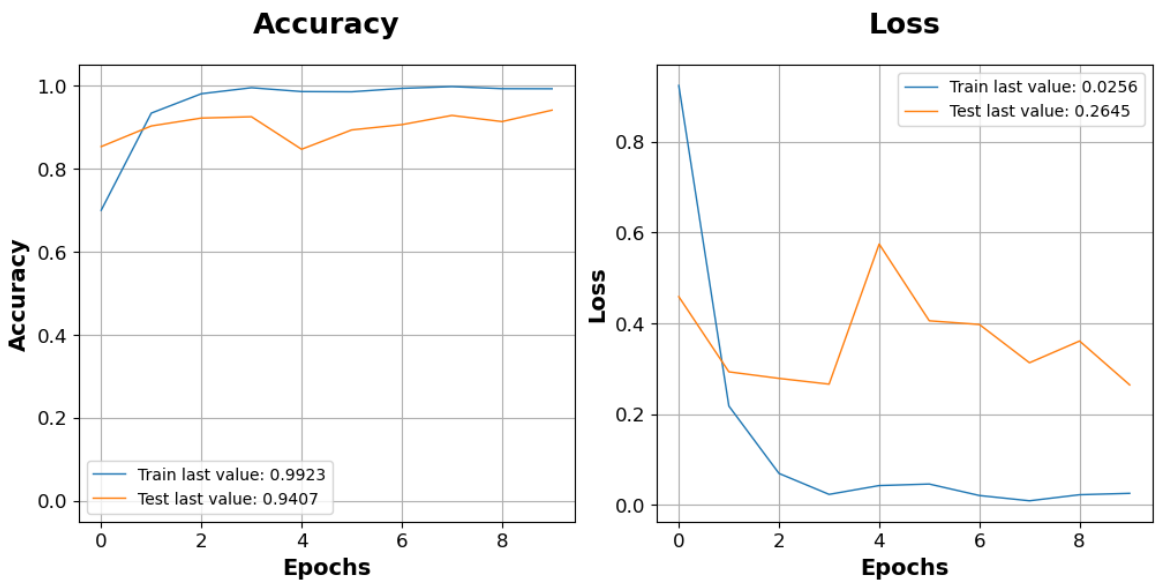
- 1 : ZOOM + ROTATION
- 2 : ZOOM + ROTATION + LUMINOSITE + CONTRASTE

Stratégie 3 : Risque d'overfitting

Sans Data Augmentation



Avec Data Augmentation



Réduction de l'erreur entre train et test : 0.24 à 0.9
Réduction du risque d'overfitting

Test de l'API

- Edanam Food and Grocery Database API

- **Recherche d'un aliment par mot-clé, nom d'aliment ou code UPC/Barcode**

Méthode:

- Création d'un compte
- Obtention de X-RapidAPI-**Key** et X-RapidAPI-**Host**.
- Fetch Data
- Extraction des informations importantes
- Sauvegarde (CSV)

- **19 produits trouvées qui contiennent du champagne**
- **3 produits ont une image**

API

foodId	label	category	foodContentsLabel	image
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
food_a79xmnya6togreaeukbroa0thhh0	Champagne Chicken	Generic meals	Flour, Salt, Pepper, Boneless, Skinless Chicke...	NaN

CONCLUSION

TEXTE

- **Faisabilité ?**
  - Oui : Méthodes basiques (TF-IDF ou BOW)
- **Classification supervisée**
  - Choix : TF-IDF avec représentation T-SNE
  - **ARI =0.84 et Précision = 0.93**

IMAGES

- **Faisabilité ?**
  - Oui : Méthodes avancées (CNN + transfert learning)
- **Classification supervisée**
  - Choix : CNN – Transfert Learning – Data Augmented (Rotation & Zoom & Brightness & Contrast)
  - **ARI =0.85 et Précision = 0.93**

Merci pour votre écoute. avez vous des questions?