Formation Data Scientist

Soutenance Projet 6 :
Classifiez automatiquement des biens de consommation

-- Mélanie WARY --

Novembre 2021

PROBLÉMATIQUE

L'entreprise "Place de marché", qui lance une marketplace e-commerce et souhaite **automatiser l'attribution de la catégorie de chaque article**.

MISSIONS

(1) Réaliser une première **étude de faisabilité** d'un moteur de **classification** d'articles en différentes catégories, **basée sur l'image et la description** fournies pour chaque article, avec un **niveau de précision suffisant**.

(2) Présenter les résultats de la classification sous la forme d'une **représentation en 2D**

CONTRAINTES & INTERPRÉTATION :

- **Pré-traitement et extraction de features** des données
  - **textuelles (description)**
  - **visuelles (photo)** – SIFT / ORB / SURF
- **Réduction(s) dimensionnelle(s)** – 2D
- **Clustering non supervisé** – nombre de clusters = nombre de « catégories vraies »
- **Optimisation** et mesure du niveau de **précision** par comparaison à la « catégorisation vraie ».

JEU DE DONNEES

➔ ***1050 produits***

➔ ***3 variables d'intérêt:***

CATEGORIES

Arbre de catégories

Min. 2 et max. 7 catégories

Format =

["Categ1 >> Categ2"]

["Categ1 >> Categ2 >> Categ3 >> … >> Categ7 "]

IMAGES

1 par produit

En couleur

Différentes tailles

Généralement sur fond blanc

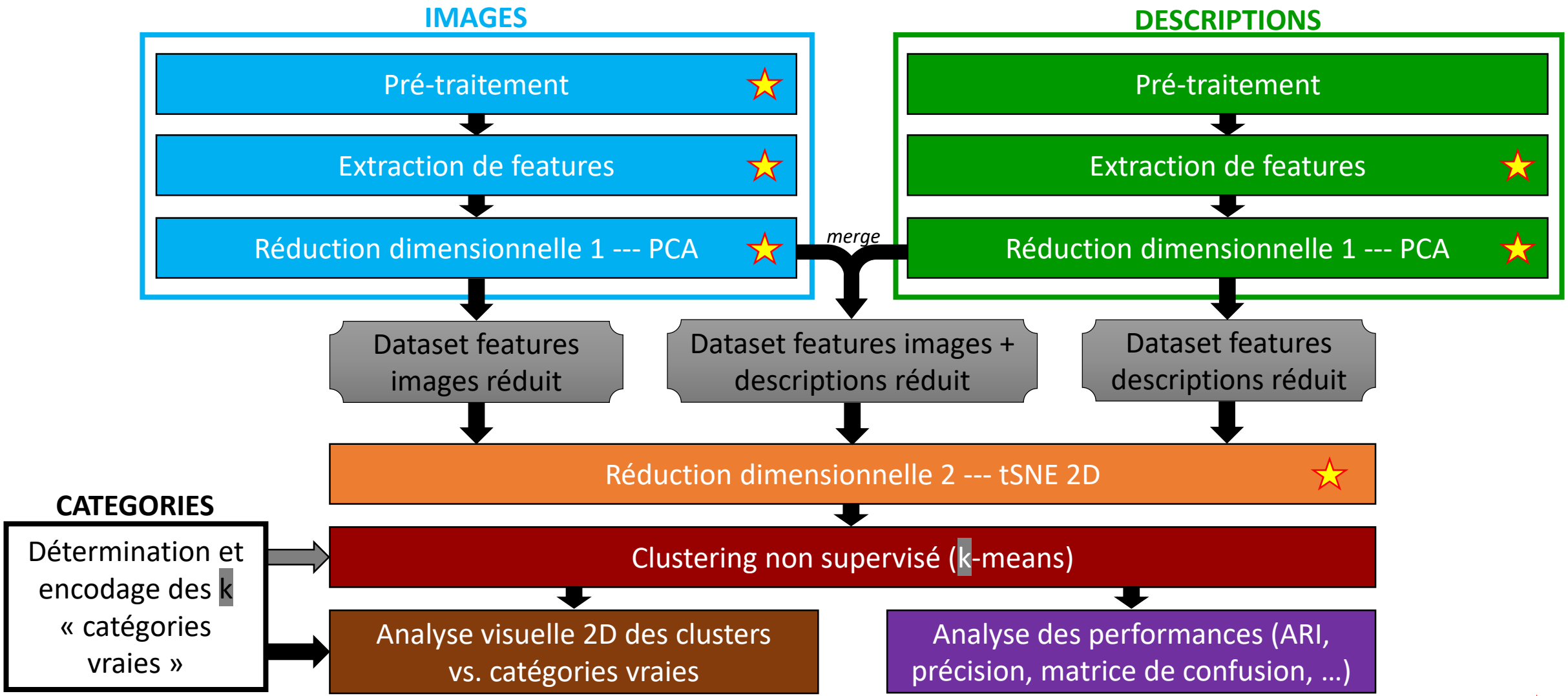
DESCRIPTIONS

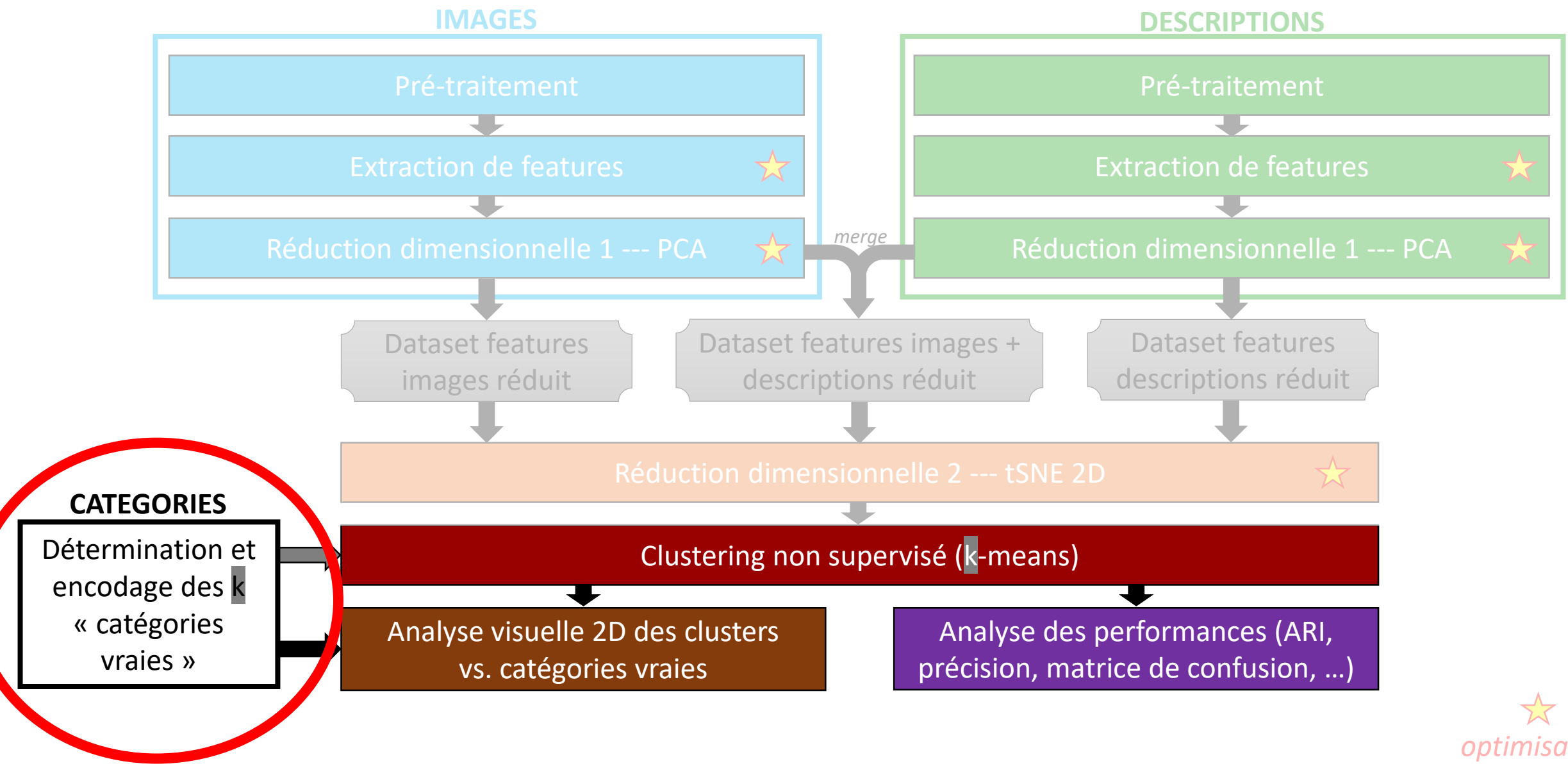
1 par produit

En anglais

De 13 à 572 mots

MÉTHODOLOGIE GÉNÉRALE




IMAGES
Pré-traitement
Extraction de features
Réduction dimensionnelle 1 --- PCA
merge
DESCRIPTIONS
Pré-traitement
Extraction de features
Réduction dimensionnelle 1 --- PCA
Dataset features images réduit
Dataset features images + descriptions réduit
Dataset features descriptions réduit
Réduction dimensionnelle 2 --- tSNE 2D
CATEGORIES
Détermination et encodage des k « catégories vraies »
Clustering non supervisé (k-means)
Analyse visuelle 2D des clusters vs. catégories vraies
Analyse des performances (ARI, précision, matrice de confusion, …)
optimisation

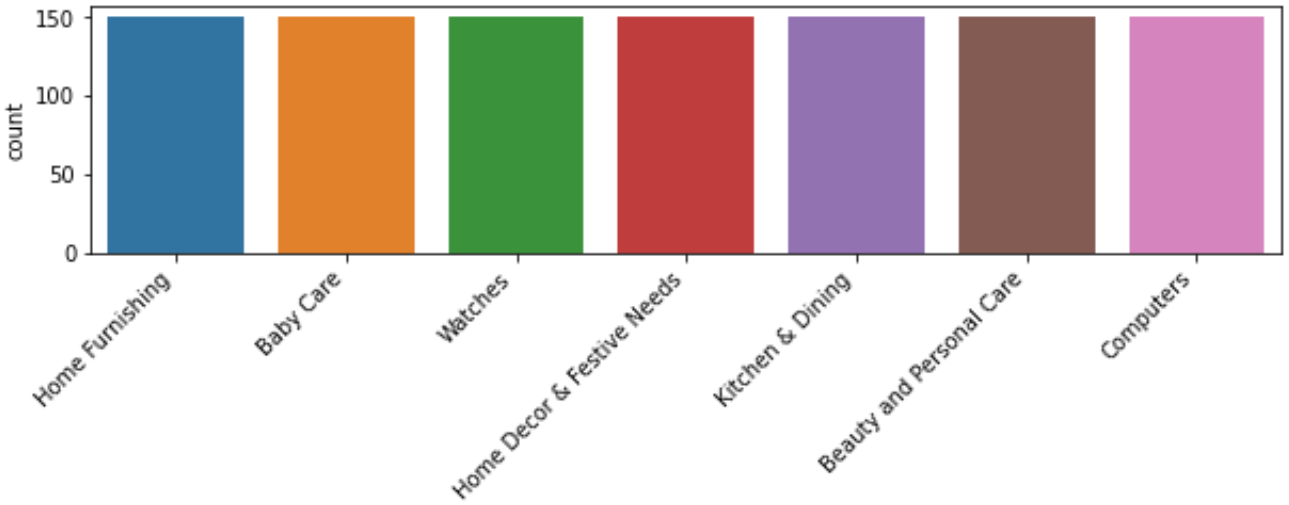
CATEGORIES

Détermination et encodage des k « catégories vraies »

Min. 2 catégories

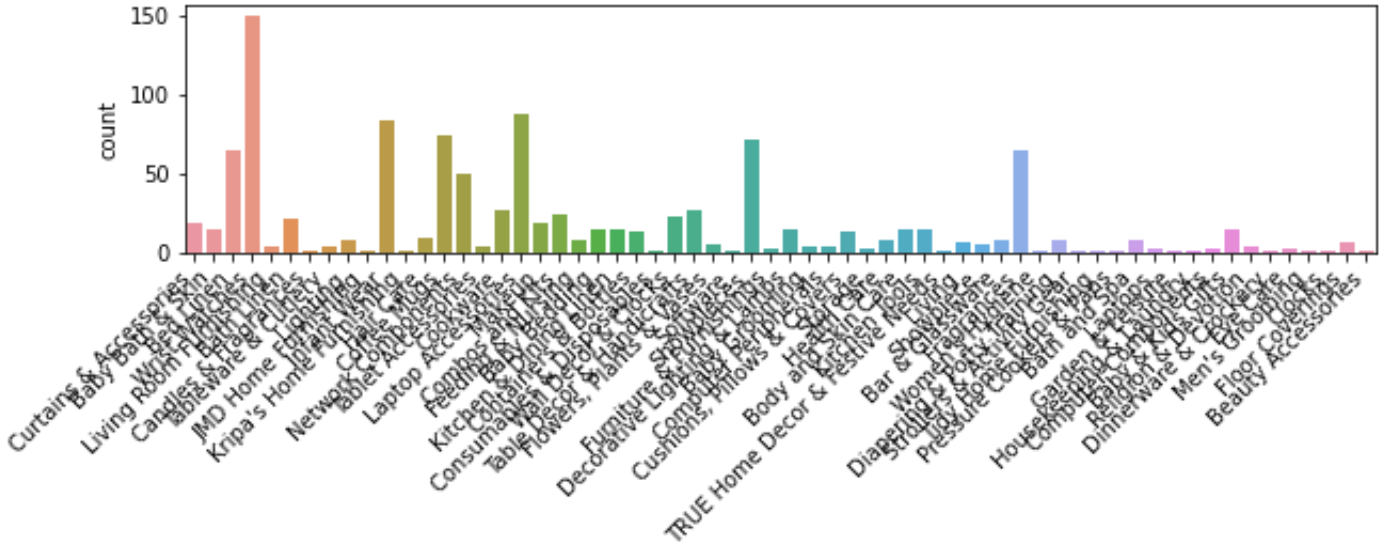
Perte d'individus avec l'utilisation des niveaux de catégorie supérieurs à 2

Catégories de niveau 1



Sélection des k=7 catégories de niveau 1 comme « catégories vraies » car équilibrées

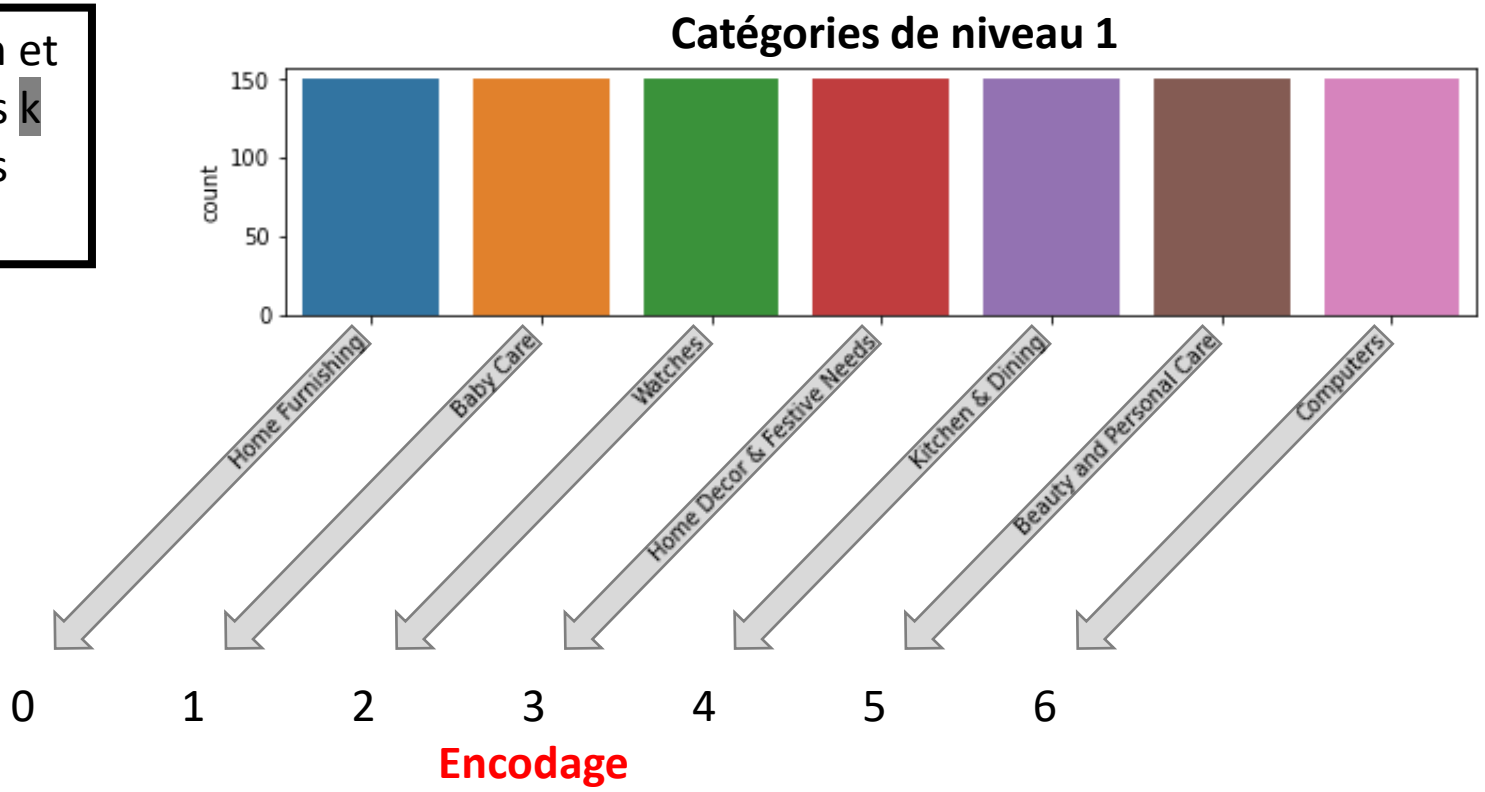
Catégories de niveau 2



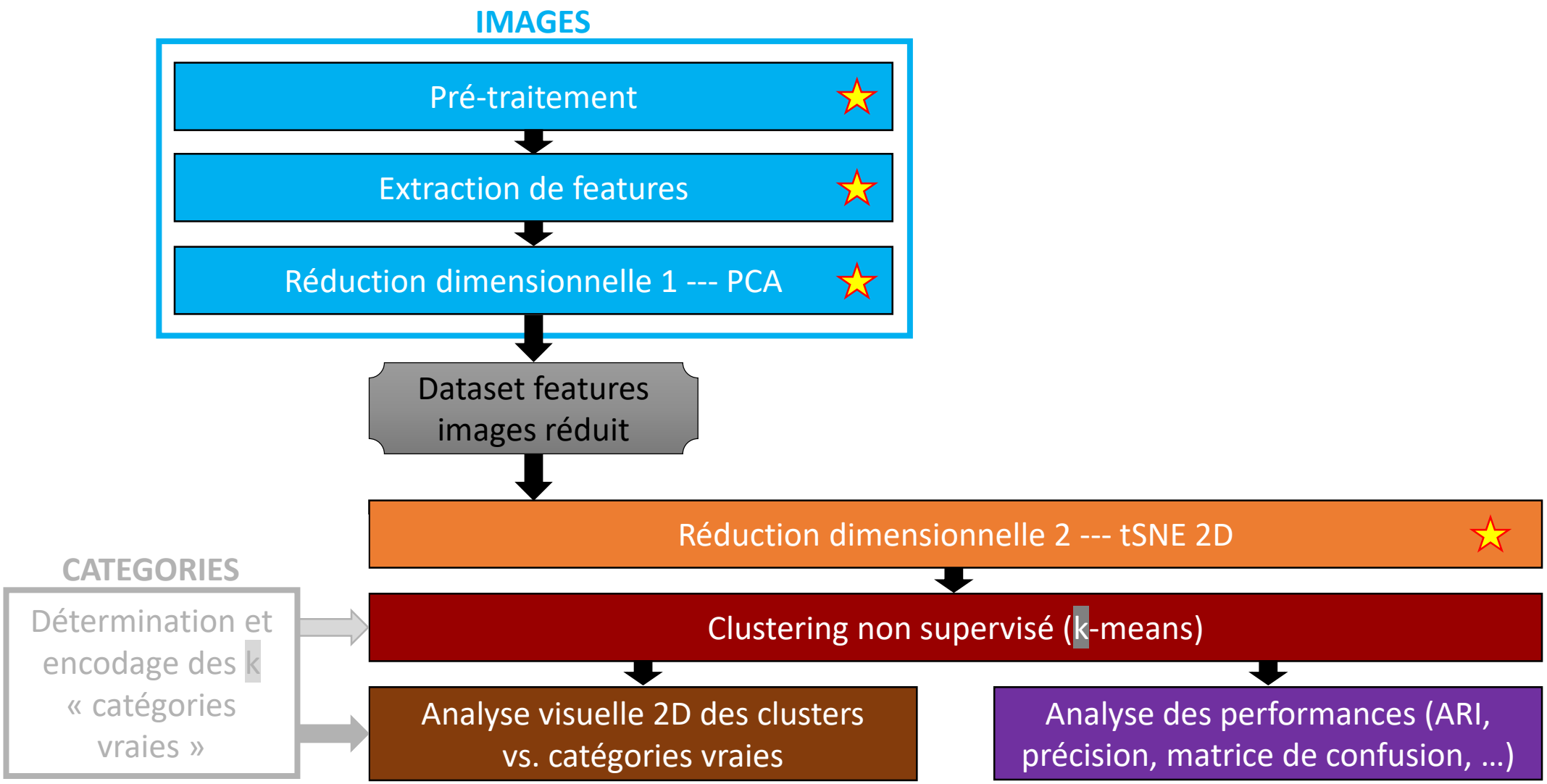
Non équilibrées

CATEGORIES

Détermination et encodage des k « catégories vraies »



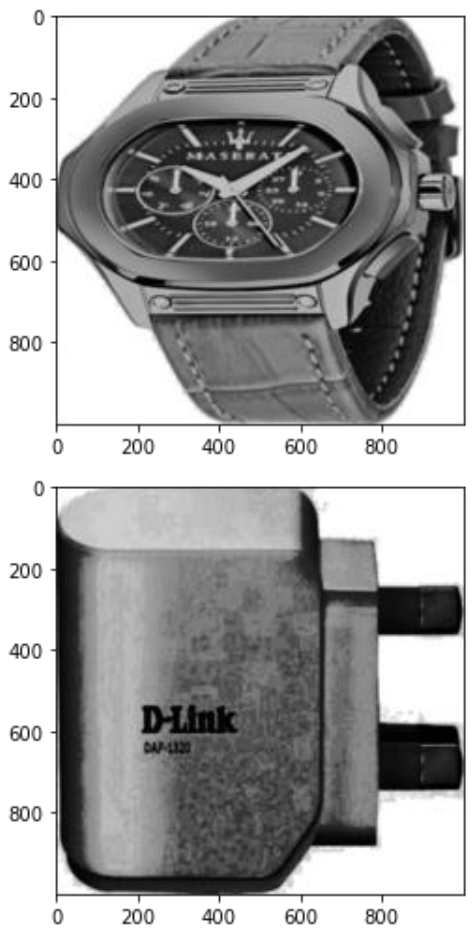
Sélection des k=7 catégories de niveau 1 comme « catégories vraies » car équilibrées


IMAGES
Pré-traitement
Extraction de features
Réduction dimensionnelle 1 --- PCA
Dataset features images réduit
Réduction dimensionnelle 2 --- tSNE 2D
CATEGORIES
Détermination et encodage des k « catégories vraies »
Clustering non supervisé (k-means)
Analyse visuelle 2D des clusters vs. catégories vraies
Analyse des performances (ARI, précision, matrice de confusion, …)
optimisation

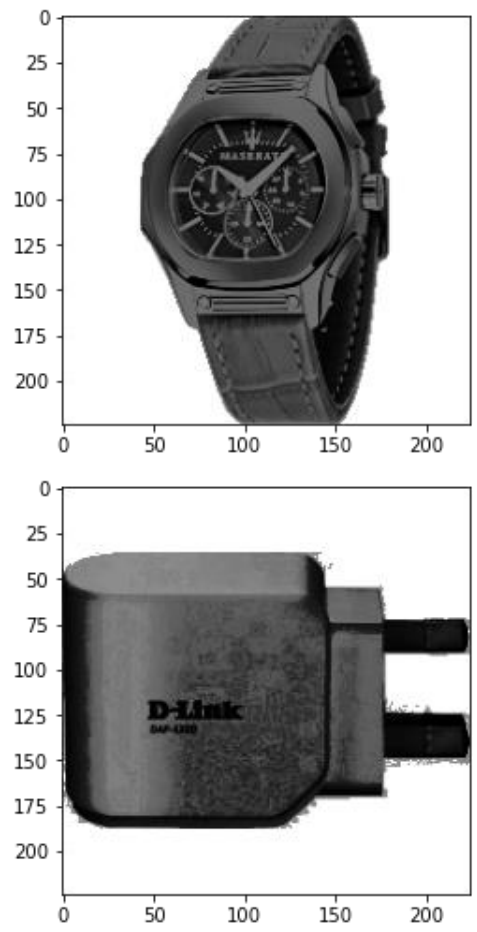
Pré-traitement

Images originales

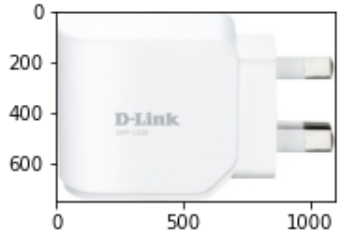
Redimensionnement (1000,1000)
Conversion en niveaux de gris
Egalisation histogramme (contraste)
Bruit filtré



Redim. (224,224) + padding blanc
Conversion en niveaux de gris
Egalisation histogramme (contraste)
Bruit NON filtré



Images originales



Pré-traitement

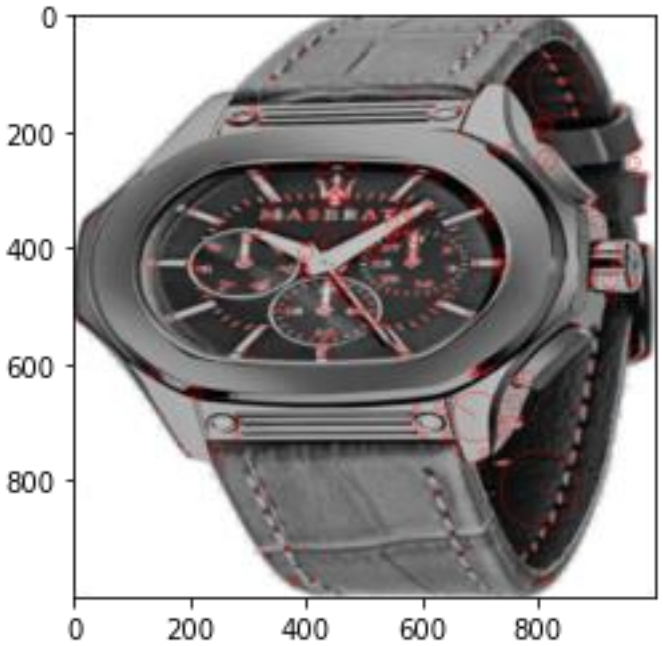
Size1000 + noise | *Size224 + padding*

SIFT

Extraction de features

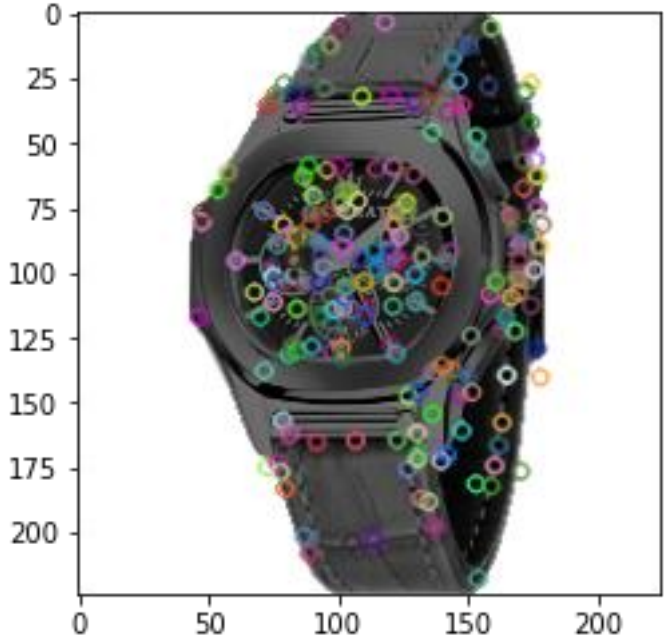
Nb max descripteurs : [25,100,500,1000,5000,10000]

Ex: nb max descripteurs = 500



Size1000 + noise

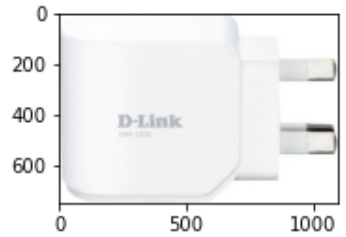
500 descripteurs



Size224 + padding

244 descripteurs

Images originales



Pré-traitement

Size1000 + noise | *Size224 + padding*

SIFT

Extraction de features

Nb max descripteurs : [25,100,500,1000,5000,10000]

Nb visual words: [k_{min}, k_{mean}, k_{max}]

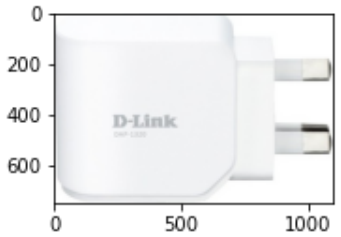
MiniBatchKMeans

$$k_{min} = 10 \times nb\ catégories = 70$$

$$k_{max} = \lfloor \sqrt{nb\ total\ de\ descripteurs\ de\ l'ensemble\ des\ images} \rfloor$$

$$k_{mean} = \left\lfloor \frac{k_{min} + k_{max}}{2} \right\rfloor$$

Images originales



Pré-traitement

Size1000 + noise | *Size224 + padding*

SIFT

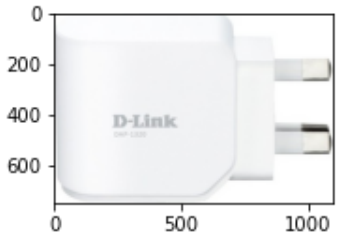
Extraction de features

Nb max descripteurs : [25,100,500,1000,5000,10000]

MiniBatchKMeans ← *Nb visual words: [k_{min}, k_{mean}, k_{max}]*

Vectorisation – Bag of visual words

Images originales



Pré-traitement

Size1000 + noise | *Size224 + padding*

SIFT

Extraction de features

Nb max descripteurs : [25,100,500,1000,5000,10000]

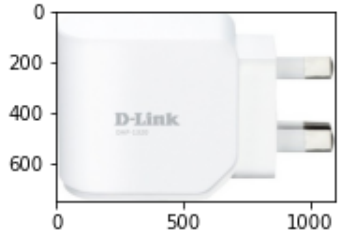
MiniBatchKMeans ← *Nb visual words: [k_{min}, k_{mean}, k_{max}]*

Vectorisation – Bag of visual words

Réduction dimensionnelle 1 --- PCA

Variance expliquée: [80%, 90%, 99%]

Images originales



Pré-traitement

Size1000 + noise | *Size224 + padding*

SIFT

Extraction de features

Nb max descripteurs : [25,100,500,1000,5000,10000]

MiniBatchKMeans ← *Nb visual words: [k_{min}, k_{mean}, k_{max}]*

Vectorisation – Bag of visual words

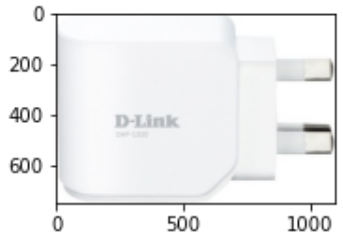
Réduction dimensionnelle 1 --- PCA

Variance expliquée: [80%, 90%, 99%]

Réduction dimensionnelle 2 --- tSNE 2D

Perplexity: [2,5,10,25,50,75,100,500]

Images originales



Pré-traitement

Size1000 + noise | *Size224 + padding*

SIFT

Extraction de features

Nb max descripteurs : [25,100,500,1000,5000,10000]

MiniBatchKMeans ← *Nb visual words: [k_{min}, k_{mean}, k_{max}]*

Vectorisation – Bag of visual words

Réduction dimensionnelle 1 --- PCA

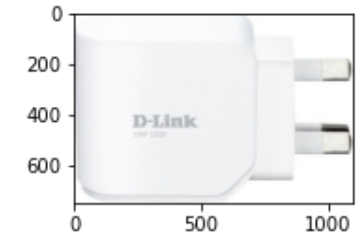
Variance expliquée: [80%, 90%, 99%]

Réduction dimensionnelle 2 --- tSNE 2D

Perplexity: [2,5,10,25,50,75,100,500]

Clustering non supervisé (k-means à k=7)

Images originales



Pré-traitement

Size1000 + noise | *Size224 + padding*

SIFT

Extraction de features

Nb max descripteurs : [25,100,500,1000,5000,10000]

MiniBatchKMeans ← *Nb visual words: [k_{min}, k_{mean}, k_{max}]* =610

Vectorisation – Bag of visual words

Réduction dimensionnelle 1 --- PCA

Variance expliquée: [80%, 90%, 99%]

Réduction dimensionnelle 2 --- tSNE 2D

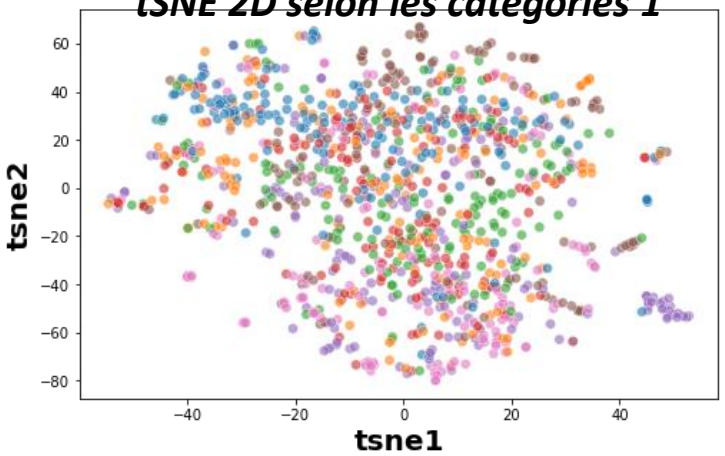
Perplexity: [2,5,10,25,50,75,100,500]

Clustering non supervisé (k-means à k=7)

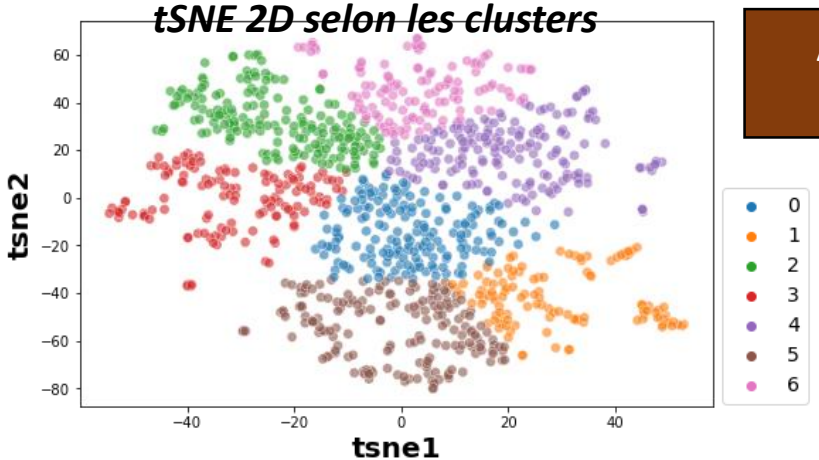
Analyse des performances (ARI)

Meilleur ARI ≈ 0.08

tSNE 2D selon les catégories 1



tSNE 2D selon les clusters



Pré-traitement

Size1000 + noise | *Size224 + padding*

SIFT

MiniBatchKMeans

Extraction de features

Nb max descripteurs : [25,100,500,1000,5000,10000]

Nb visual words: [k_{min}, k_{mean}, k_{max}] =610

Vectorisation – Bag of visual words

Réduction dimensionnelle 1 --- PCA

Variance expliquée: [80%, 90%, 99%]

Réduction dimensionnelle 2 --- tSNE 2D

Perplexity: [2,5,10,25,50,75,100,500]

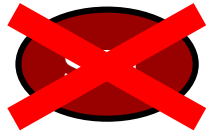
Clustering non supervisé (k-means à k=7)

Analyse visuelle 2D des clusters vs. catégories vraies

Analyse des performances (ARI)

Meilleur ARI ≈ 0.08

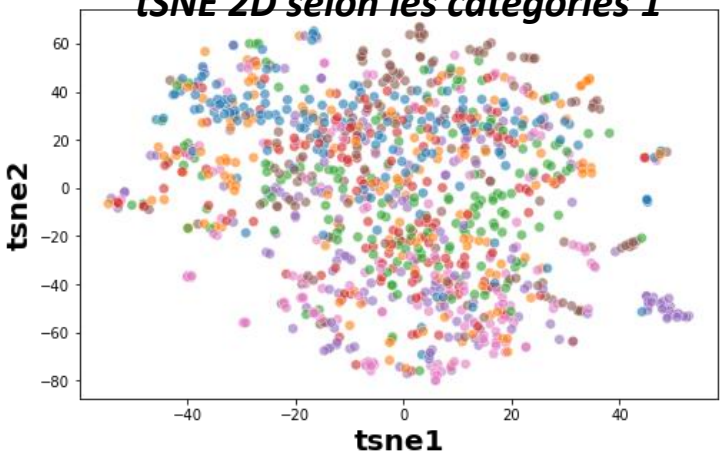
CNN (VGG16) – Transfer Learning



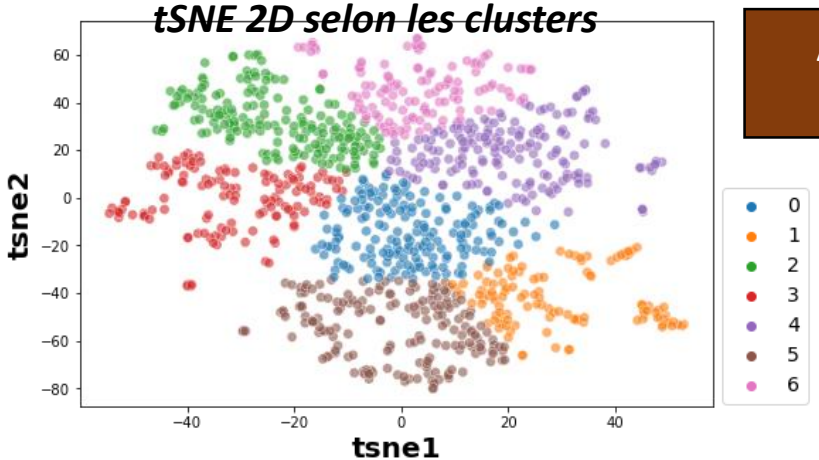
MiniBatchKMeans

- Pré-traitement
  - *Size1000 + noise*
  - *Size224 + padding*
- Extraction de features
  - *Nb max descripteurs : [25,100,500,1000,5000,10000]*
  - *Nb visual words: [k_{min}, k_{mean}, k_{max}] =610*
  - *Vectorisation – Bag of visual words*
- Réduction dimensionnelle 1 --- PCA
  - *Variance expliquée: [80%, 90%, 99%]*
- Réduction dimensionnelle 2 --- tSNE 2D
  - *Perplexity: [2,5,10,25,50,75,100,500]*
- Clustering non supervisé (k-means à k=7)
- Analyse visuelle 2D des clusters vs. catégories vraies
- Analyse des performances (ARI)

tSNE 2D selon les catégories 1

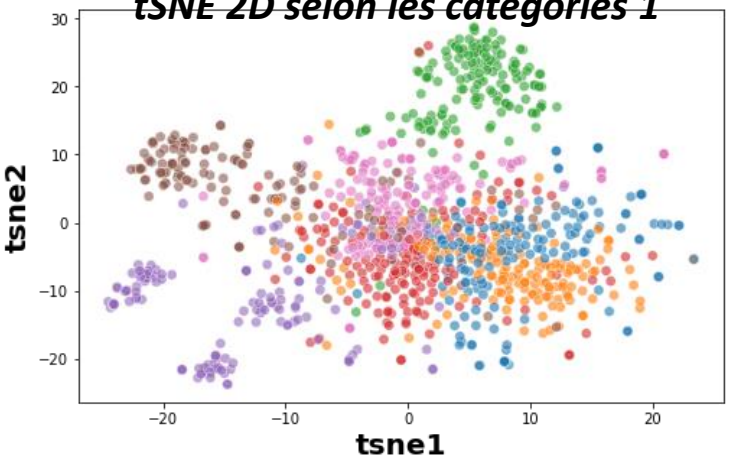


tSNE 2D selon les clusters

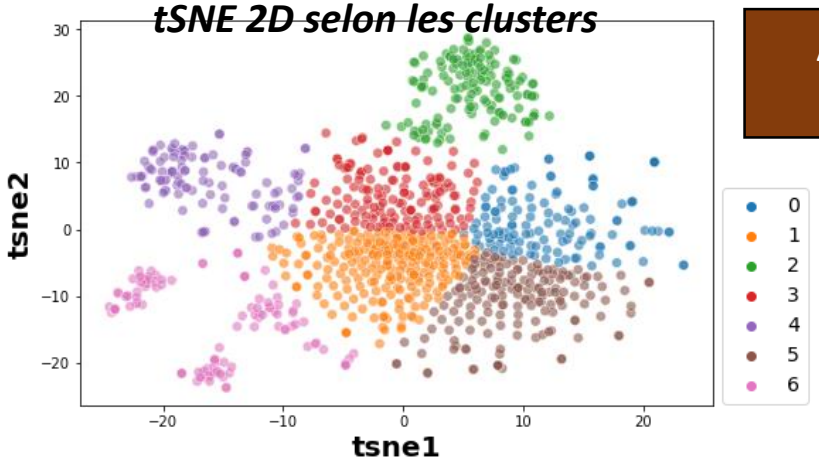


Meilleur ARI ≈ 0.08

tSNE 2D selon les catégories 1



tSNE 2D selon les clusters



VGG16

Pré-traitement

Prétraitement VGG16

Extraction de features

Base VGG16 + Flatten	*Base VGG16 + Pooling + Flatten*
Vecteur output de taille 25088	*Vecteur output de taille 4608*

Réduction dimensionnelle 1 --- PCA

Variance expliquée: [80%, 90%, 99%]

Réduction dimensionnelle 2 --- tSNE 2D

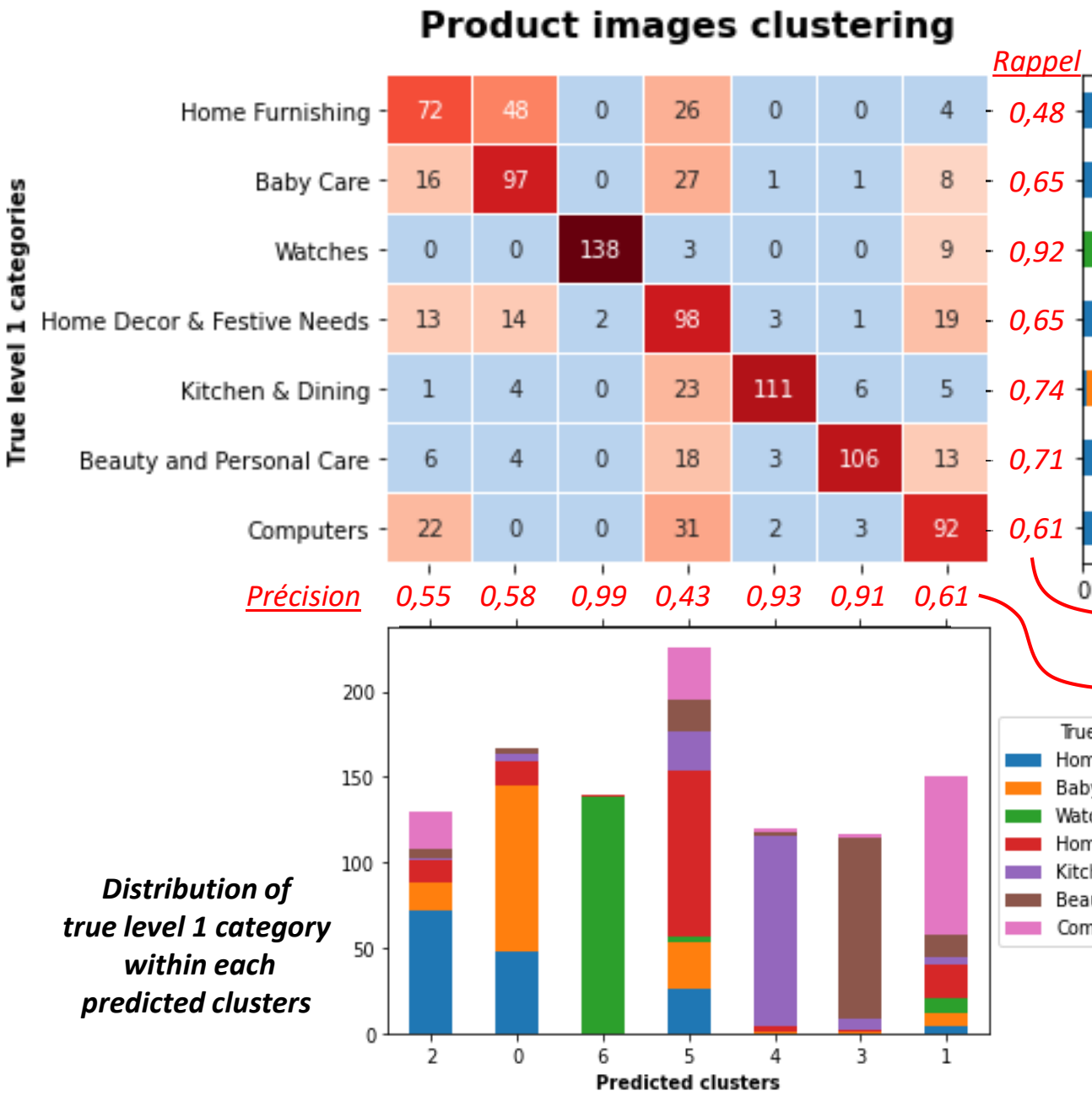
Perplexity: [2,5,10,15,20,25,30,35,40,45,50,75,100,500,750] *+ essai 3D*

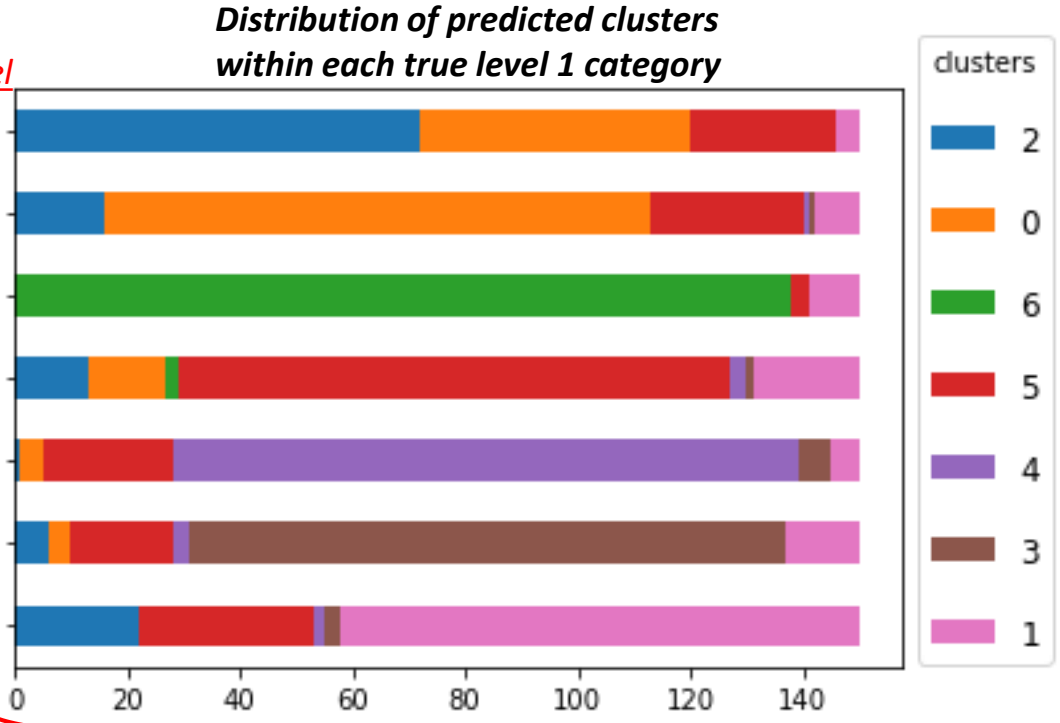
Clustering non supervisé (k-means à k=7)

Analyse visuelle 2D des clusters vs. catégories vraies

Analyse des performances (ARI, précision, matrice de confusion, …)

Meilleur ARI ≈ 0.42





Distribution of true level 1 category within each predicted clusters

Rappel moyen = 0,68
➔ 68% des produits de chaque « catégorie vraie » correctement identifiés

Précision moyenne = 0,71
➔ 71% des produits de chaque cluster correctement classifiés

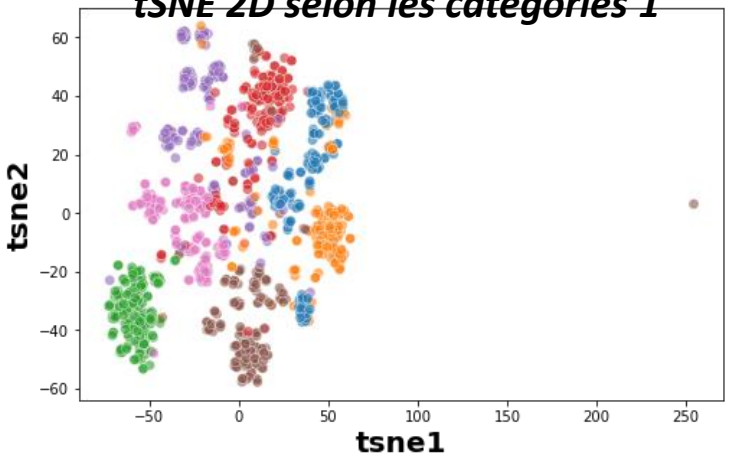
→ Bilan intermédiaire 1 (IMAGES):

- Extraction features via SIFT moins performante (et moins rapide) que via VGG16-Transfer Learning
- Meilleures performances : ARI = 0.42, rappel = 0.68, précision = 0.71

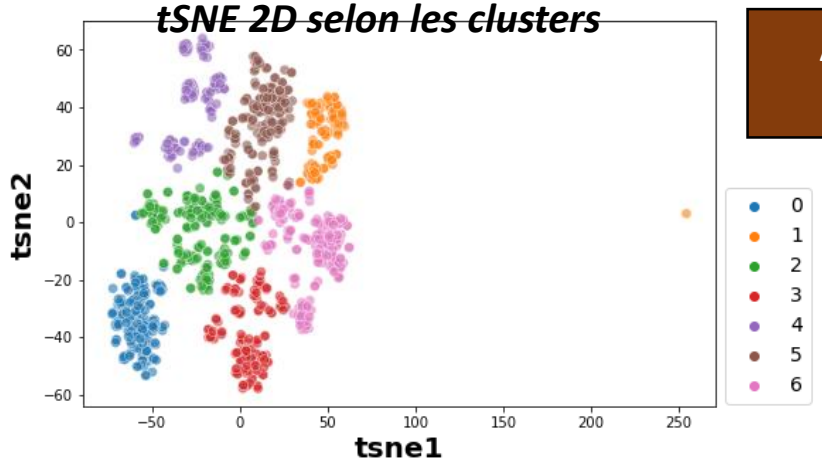
Améliorations possibles:

- Niveau dataset :
  - Regarder les produits mal classifiés pour identifier les raisons sous-jacentes et si possible les corriger avant le passage à l'échelle de la marketplace ?
- Niveau pré-traitement et extraction de features :
  - Tester ORB / SURF ?
  - Tester autre modèle CNN (+ Transfer Learning) avec images plus similaires ?
  - Ajouter une couche de pooling pour réduire encore la dimension des vecteurs features en sortie du CNN ?
- Niveau classification :
  - Autre approche Transfer Learning = ajouter un classifieur 7 classes (couche Dense, activation softmax) en fin de modèle → entrainement seulement du classifieur → prédiction des clusters ?

tSNE 2D selon les catégories 1



tSNE 2D selon les clusters



Pré-traitement

Tokenisation, suppr ponctuation, suppr caractères numériques, mise en minuscules, suppr English stopwords + mots fréquents, lemmatisation

Extraction de features

BoW | *BoW tf-idf* | *bigrams tf-idf*

Word2Vec
Nb vecteurs = [50,100, 200,300,500]

Réduction dimensionnelle 1 --- PCA

Variance expliquée: [80%, 90%, 99%]

Réduction dimensionnelle 2 --- tSNE 2D

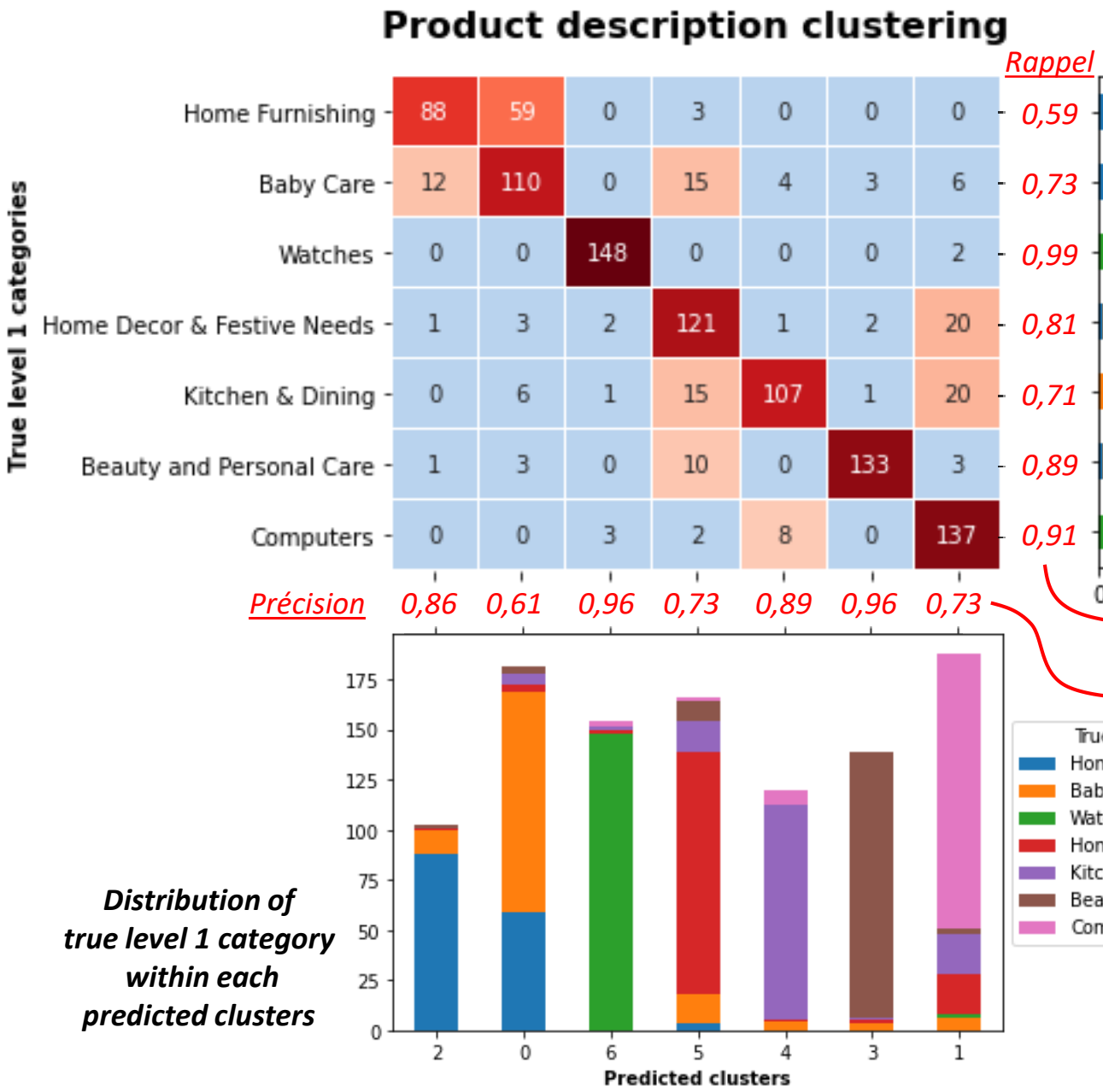
Perplexity: [2,5,10,15,20,25,30,35,40,45,50,75,100,500] *(+ init. PCA pour W2V)*

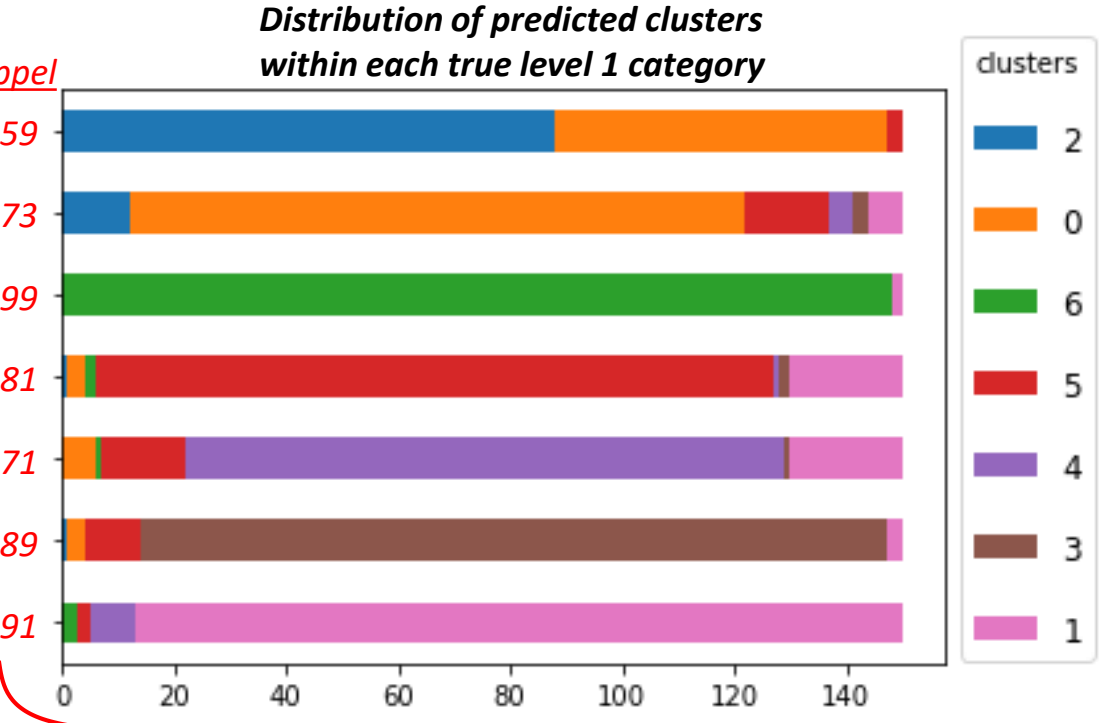
Clustering non supervisé (k-means à k=7)

Analyse visuelle 2D des clusters vs. catégories vraies

Analyse des performances (ARI, précision, matrice de confusion, …)

Meilleur ARI ≈ 0.63





Rappel moyen = 0,80
➔ 80% des produits de chaque « catégorie vraie » correctement identifiés

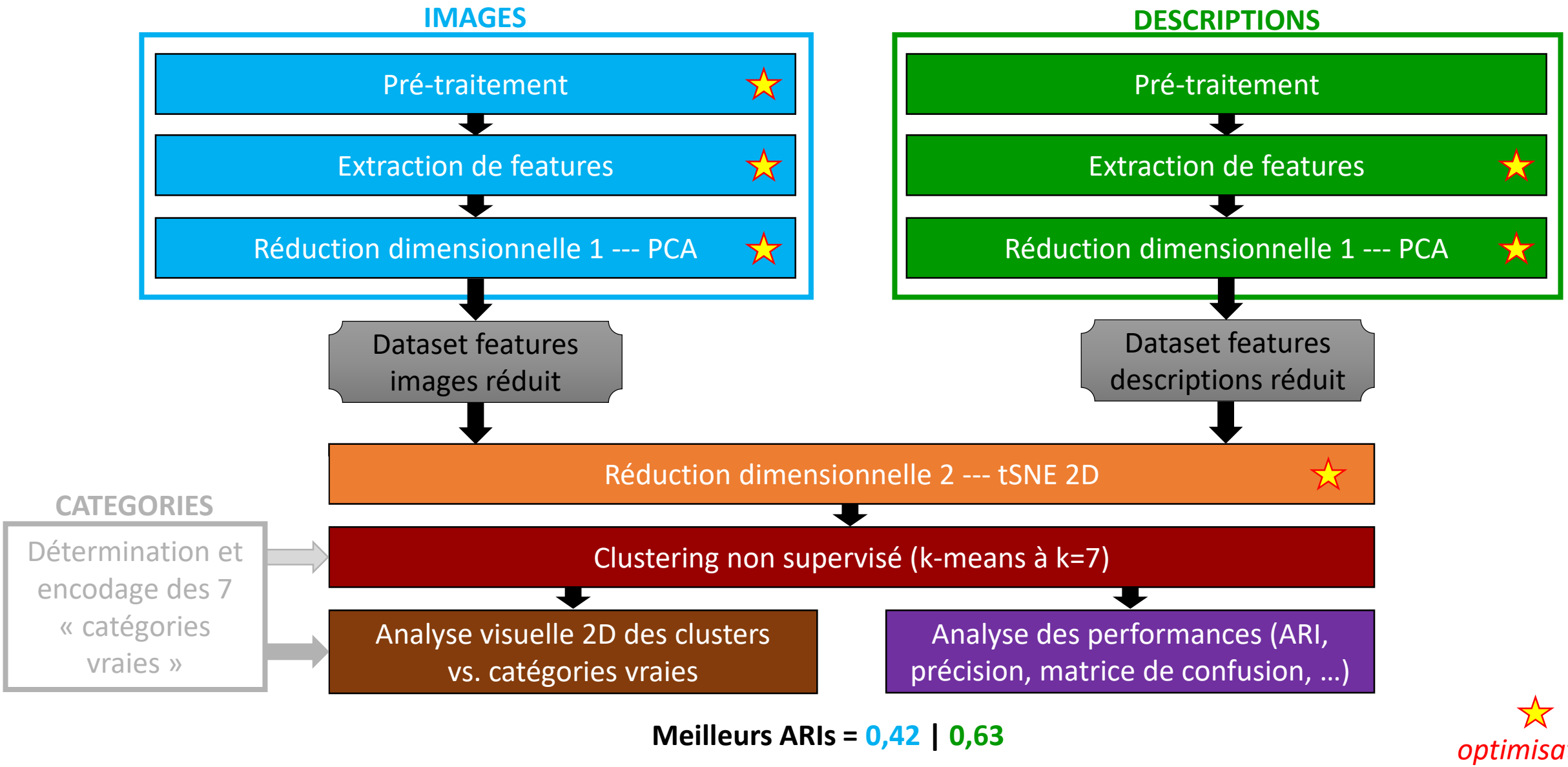
Précision moyenne = 0,82
➔ 82% des produits de chaque cluster correctement classifiés

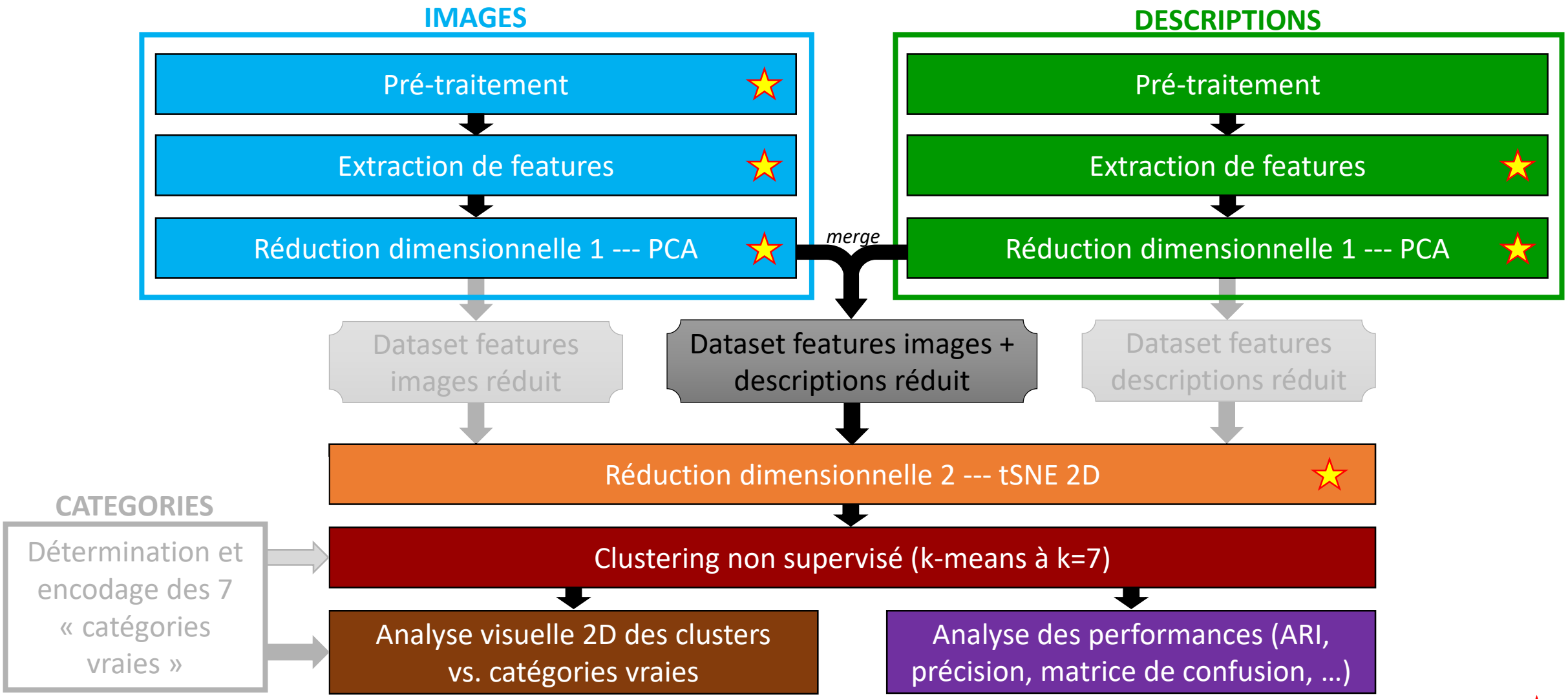
➔ Bilan intermédiaire 2 (DESCRIPTIONS):

- extraction features via Bag-of-words tf-idf plus performante que BoW classique, Bag-of-bigrams tf-idf et word embedding via Word2Vec
- Meilleures performances : ARI = 0.63, rappel = 0.80, précision = 0.82

Améliorations possibles:

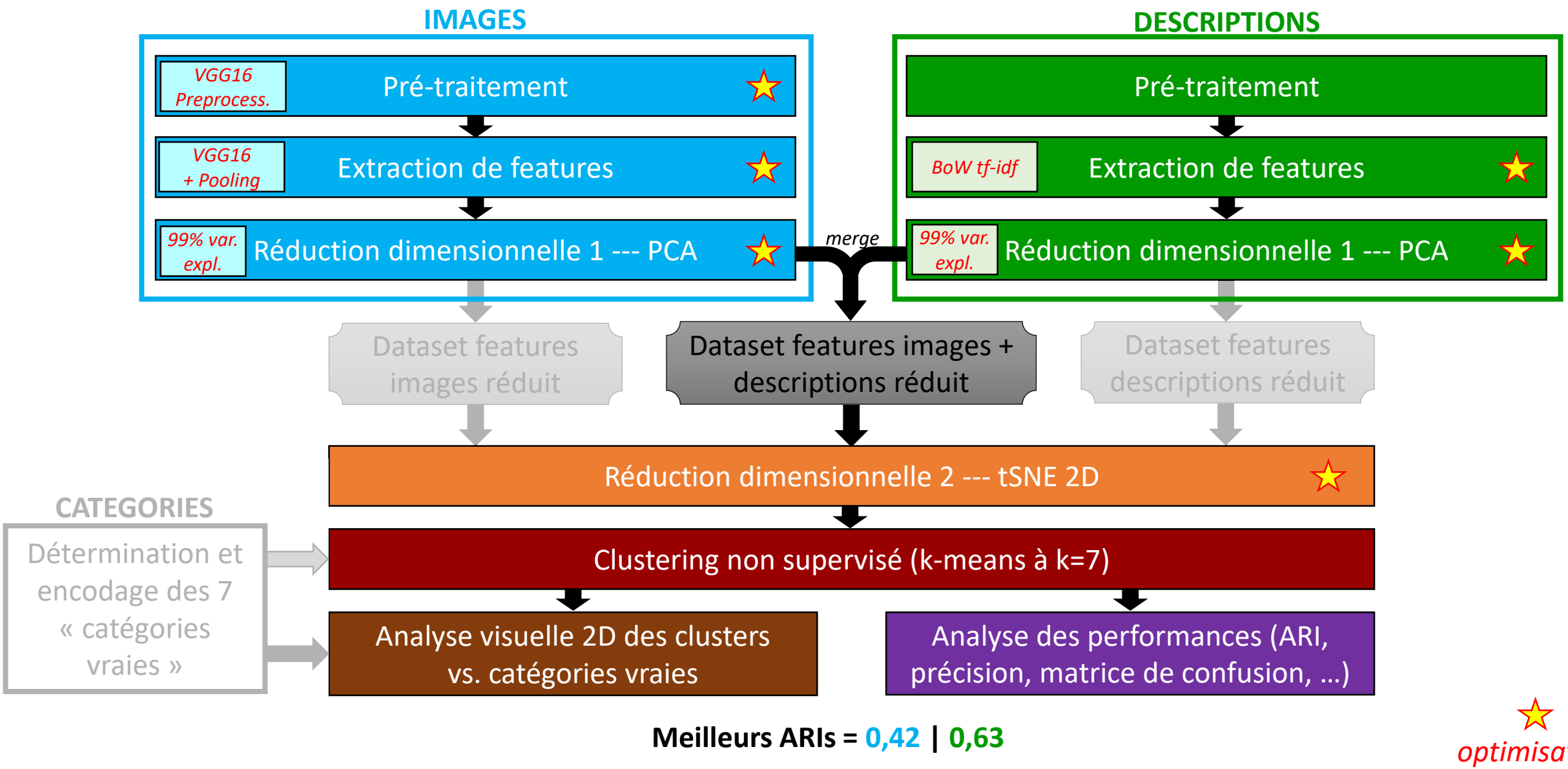
- Niveau dataset :
  - Regarder les produits mal classifiés pour identifier les raisons sous-jacentes et si possible les corriger avant le passage à l'échelle de la marketplace ?
- Niveau pré-traitement et extraction de features :
  - Optimiser le pré-traitement de texte (e.g. lemmatisation vs racinisations)
  - Tester un autre algorithme de plongement de mots ?
  - Tester un modèle RNN (+ Transfer Learning) ?
- Niveau modélisation:
  - Tester clustering sur données réduites en 3D ?
  - Tester l'utilisation d'un RNN (+ Transfer Learning) pour la classification ?


IMAGES
Pré-traitement
Extraction de features
Réduction dimensionnelle 1 --- PCA
DESCRIPTIONS
Pré-traitement
Extraction de features
Réduction dimensionnelle 1 --- PCA
Dataset features images réduit
Dataset features descriptions réduit
Réduction dimensionnelle 2 --- tSNE 2D
CATEGORIES
Détermination et encodage des 7 « catégories vraies »
Clustering non supervisé (k-means à k=7)
Analyse visuelle 2D des clusters vs. catégories vraies
Analyse des performances (ARI, précision, matrice de confusion, …)
Meilleurs ARIs = 0,42 | 0,63
optimisation



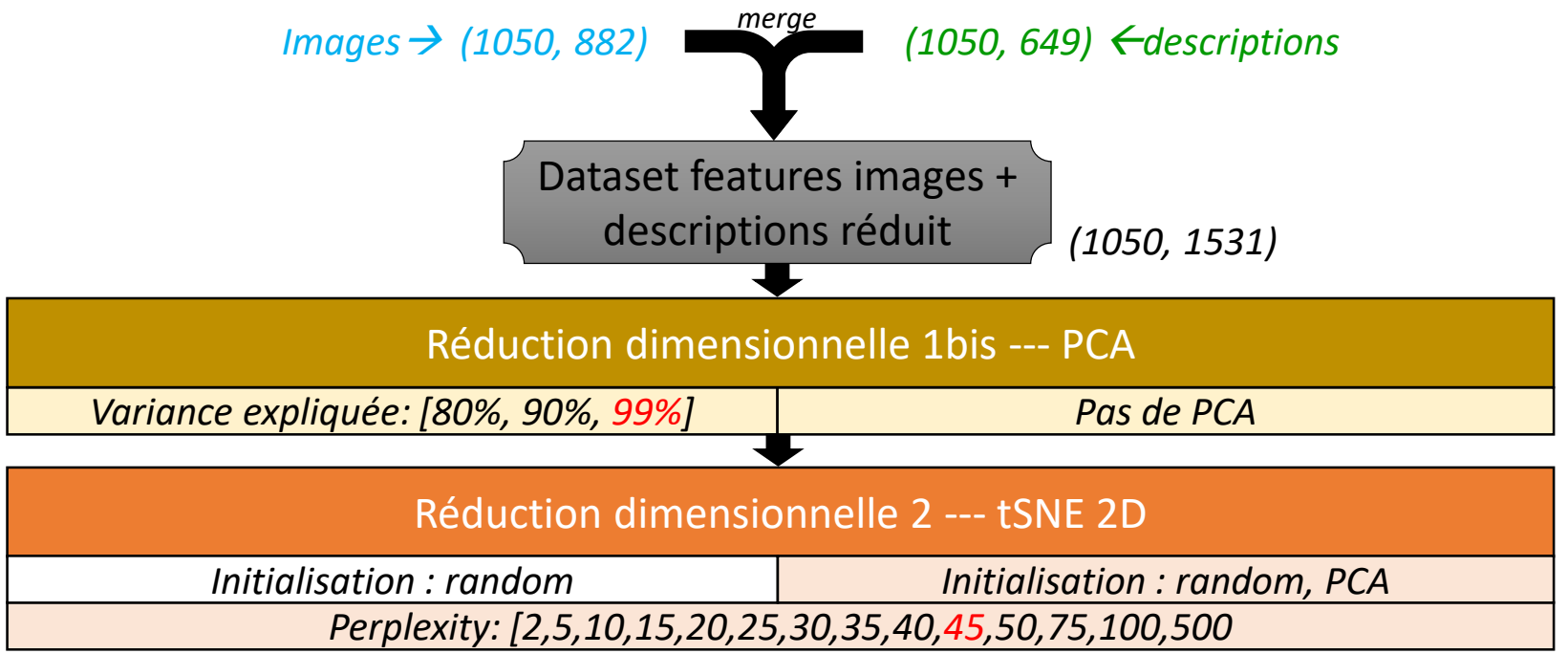
Meilleurs ARIs = 0,42 | 0,63

optimisation

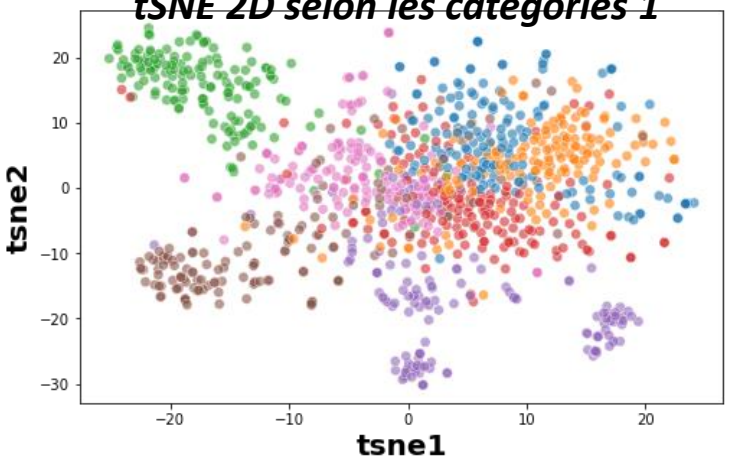


Meilleurs ARIs = 0,42 | 0,63

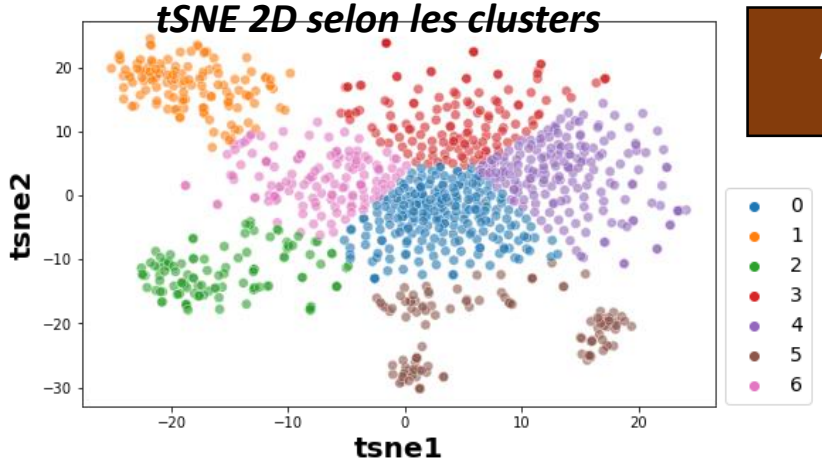
optimisation


Images → (1050, 882)
merge
(1050, 649) ← descriptions
Dataset features images + descriptions réduit
(1050, 1531)
Réduction dimensionnelle 1bis --- PCA
Variance expliquée: [80%, 90%, 99%]
Pas de PCA
Réduction dimensionnelle 2 --- tSNE 2D
Initialisation : random
Initialisation : random, PCA
Perplexity: [2,5,10,15,20,25,30,35,40,45,50,75,100,500

tSNE 2D selon les catégories 1



tSNE 2D selon les clusters



Images → (1050, 882) — *merge* — *(1050, 649) ← descriptions*

Dataset features images + descriptions réduit *(1050, 1531)*

Réduction dimensionnelle 1bis --- PCA

Variance expliquée: [80%, 90%, 99%] | *Pas de PCA*

Réduction dimensionnelle 2 --- tSNE 2D

Initialisation : random | *Initialisation : random, PCA*

Perplexity: [2,5,10,15,20,25,30,35,40,45,50,75,100,500

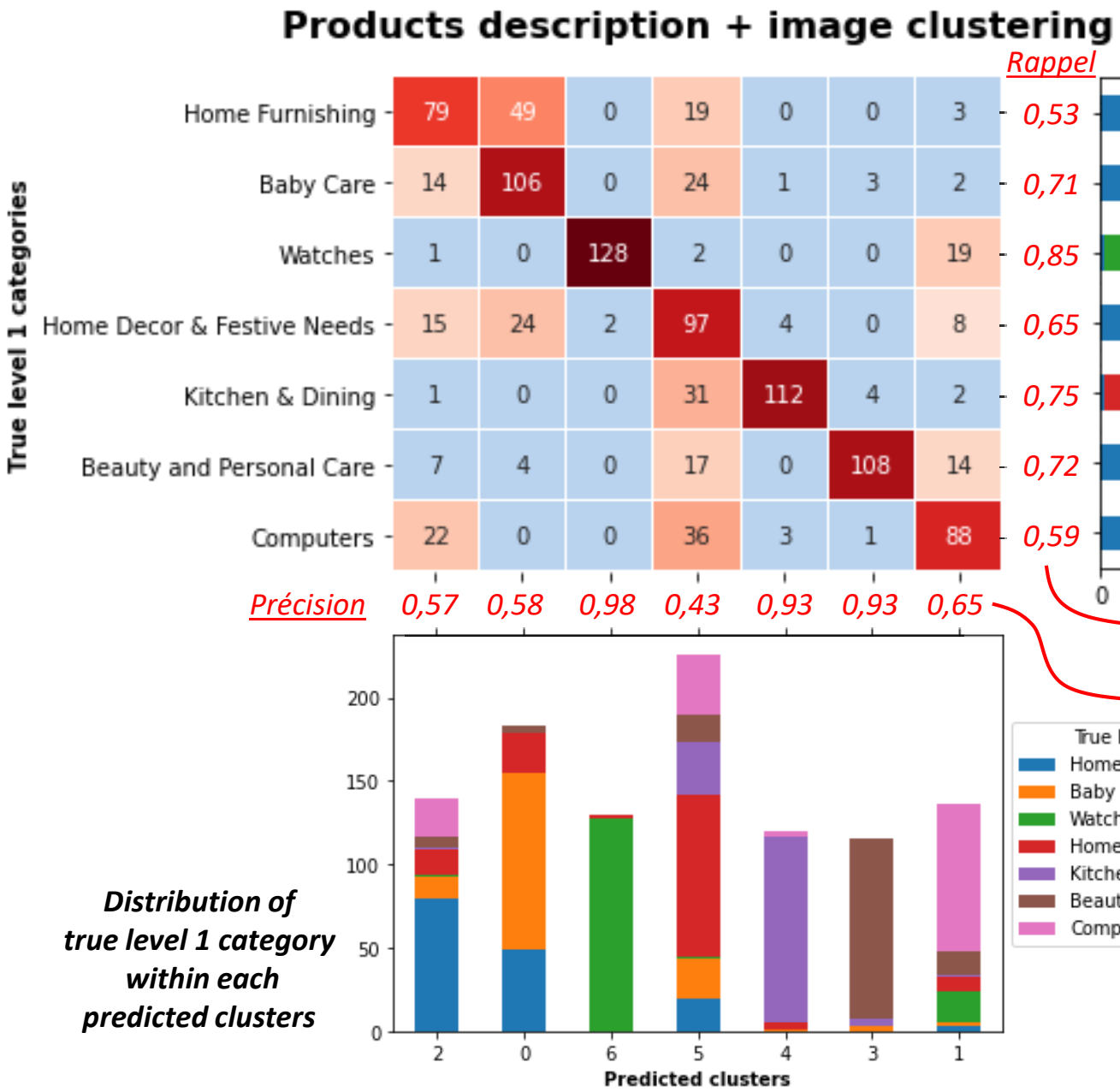
Clustering non supervisé (k-means à k=7)

Analyse visuelle 2D des clusters vs. catégories vraies

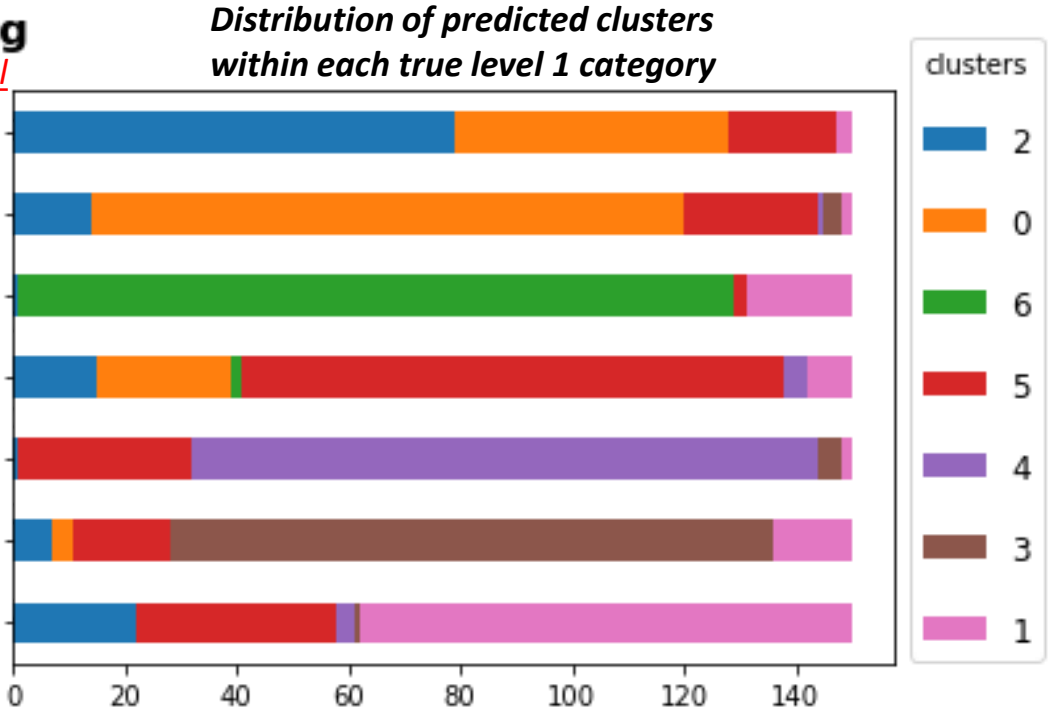
Analyse des performances (ARI, précision, matrice de confusion, …)

Meilleur ARI ≈ 0.43

Meilleurs ARIs = 0,42 | 0,63



Distribution of true level 1 category within each predicted clusters



vs. 0,68 | 0,80

Rappel moyen = 0,68
➔ 68% des produits de chaque « catégorie vraie » correctement identifiés

Précision moyenne = 0,72 **vs. 0,71 | 0,82**
➔ 72% des produits de chaque cluster correctement classifiés

➔ Bilan intermédiaire 3 (IMAGES + DESCRIPTIONS):

- Performances non améliorées, équivalentes à celles obtenues sur la base des images seules

Données	Meilleur ARI	Meilleur rappel	Meilleure précision
Images	0,42	0,68	0,71
Descriptions	0,63	0,80	0,82
Images + descriptions	0,43	0,68	0,72

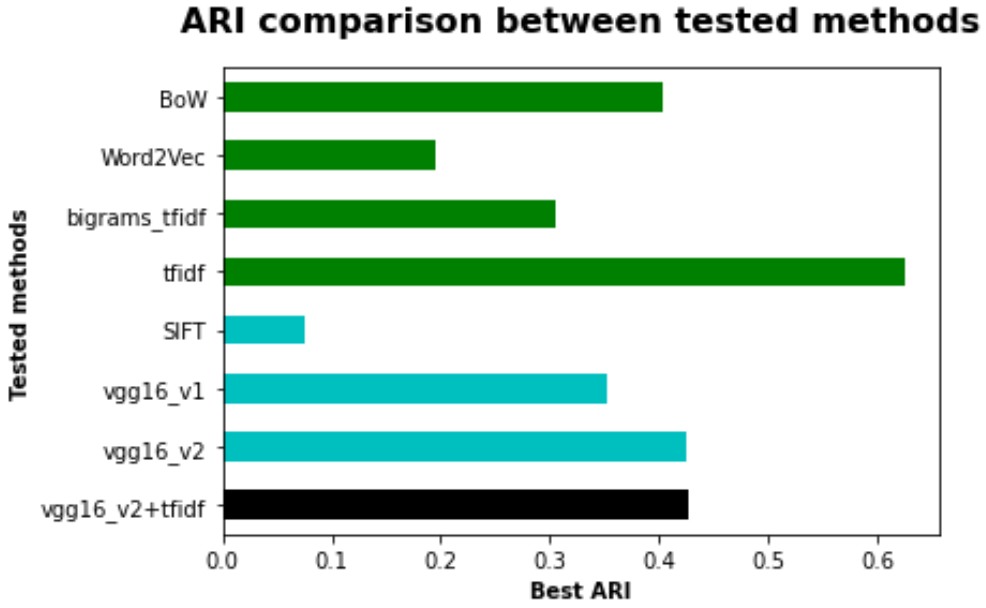
Améliorations possibles, en plus de celles spécifiques aux images et aux descriptions déjà mentionnées:

- Niveau pré-traitement et extraction de features :
  - Tester d'autres combinaisons de méthodes (pas uniquement les plus performantes séparément) ?
- Niveau modélisation:
  - Tester clustering sur données réduites en 3D ?

➔ Conclusions générales

- Meilleures performances obtenues sur classification basée sur descriptions seules

Données	Meilleur ARI	Meilleur rappel	Meilleure précision
Images	0,42	0,68	0,71
Descriptions	0,63	0,80	0,82
Images + descriptions	0,43	0,68	0,72



- Suffisamment bonnes pour attester de la faisabilité (avec une bonne précision) :
  - d'une automatisation non supervisée (après tests supplémentaires et améliorations)
  - d'une automatisation **supervisée**, avec un **jeu de données plus conséquent** (après passage à l'échelle de la marketplace), de l'attribution de catégories aux produits.

Améliorations possibles, en plus de celles déjà mentionnées:

- Vérifier la validité des « catégories vraies » actuelles / détecter l'occurrence éventuelle de catégories plus adaptées ?
- Classification « manuelle » des descriptions via l'identification de mots-clés spécifiques pour chaque catégorie?

Baby Care





Home Furnishing





Watches





Home Decor & Festive Needs





Kitchen & Dining





Beauty and Personal Care





Computers