



# CHALLENGE RAKUTEN: PRÉDICTION MULTIMODALES DES CODES DE TYPE DE PRODUIT

Réalisé par : Salma Benmoussa, Charlotte Cegarra  
Dans le cadre du Master 2 Modélisations, Statistiques, Economiques et Financières  
Année 2024-2025



キーワード検索



# CONTEXTE ET OBJECTIF

*Challenge organisé par le Rakuten Institute of Technology (RIT)*

- L'objectif principal est d'améliorer la classification des codes de type de produit sur la plateforme Rakuten France.
- Prédire le code de type de produit pour chaque article du catalogue Rakuten France en utilisant des données multimodales, comprenant des images, des titres et des descriptions des produits.



キーワード検索



# DIFFICULTES

- **Données bruitées** : Les descriptions et les images peuvent être incomplètes ou dupliquées.
- **Déséquilibre des classes** : Certaines catégories de produits sont sous-représentées dans le catalogue.
- **Grande échelle des données** : Près de 100 000 produits à classer avec des informations multimodales complexes.

# Sommaire

- 01 Préparation des données
- 02 Modélisations et Performances
- 03 Conclusions et perspectives

# PARTIE 1

## PREPARATION DES DONNEES

1. Analyse exploratoire des données (EDA)
2. Nettoyage et traitement des données

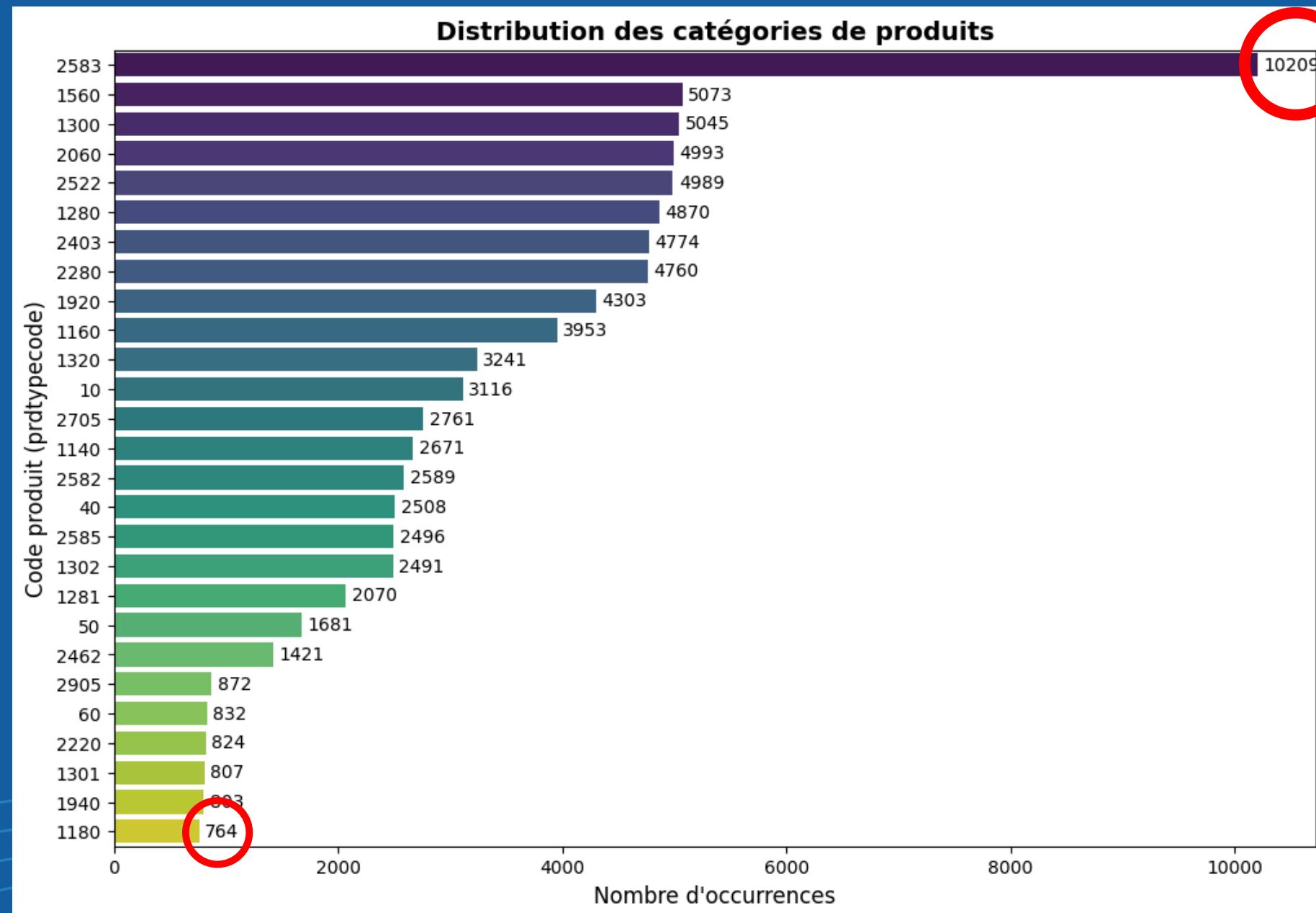
## 1. Analyse exploratoire des données (EDA)

### Metadonnées

Dataset	Variable	Description
X_train/ X_test (84916, 4) / (13812, 4)	designation	Titre du produit
	description	Description détaillée du produit
	productid	Identifiant unique du produit. Utilisé pour l'association des images et des informations du produit
	imageid	Identifiant unique de l'image associée au produit
y_train (84916, 1)	prdtypecode	Code de type de produit, la variable cible à prédire

## 1. Analyse exploratoire des données : Texte

### Visualisation de la variable cible “prdtypecode”



On observe un déséquilibre des classes significatif entre les catégories de produits

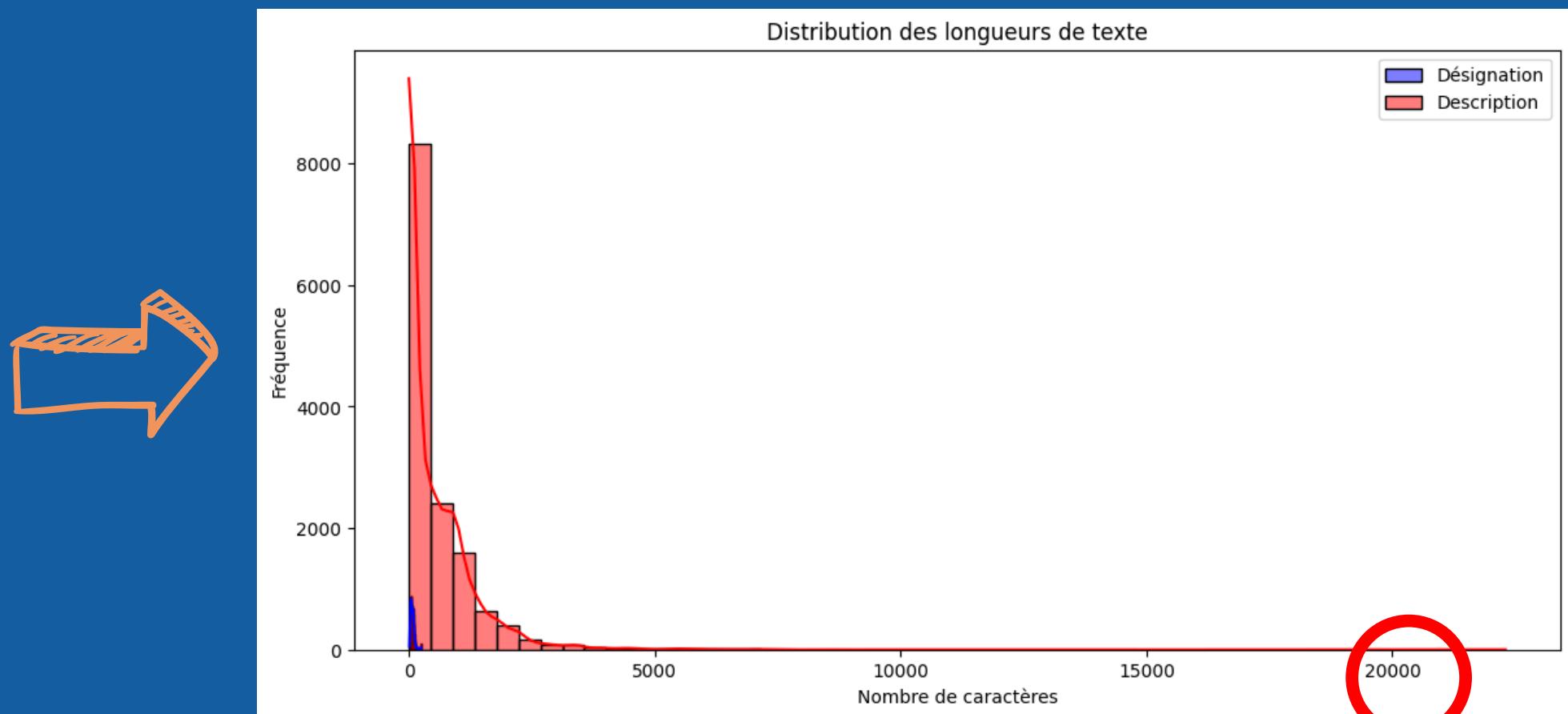
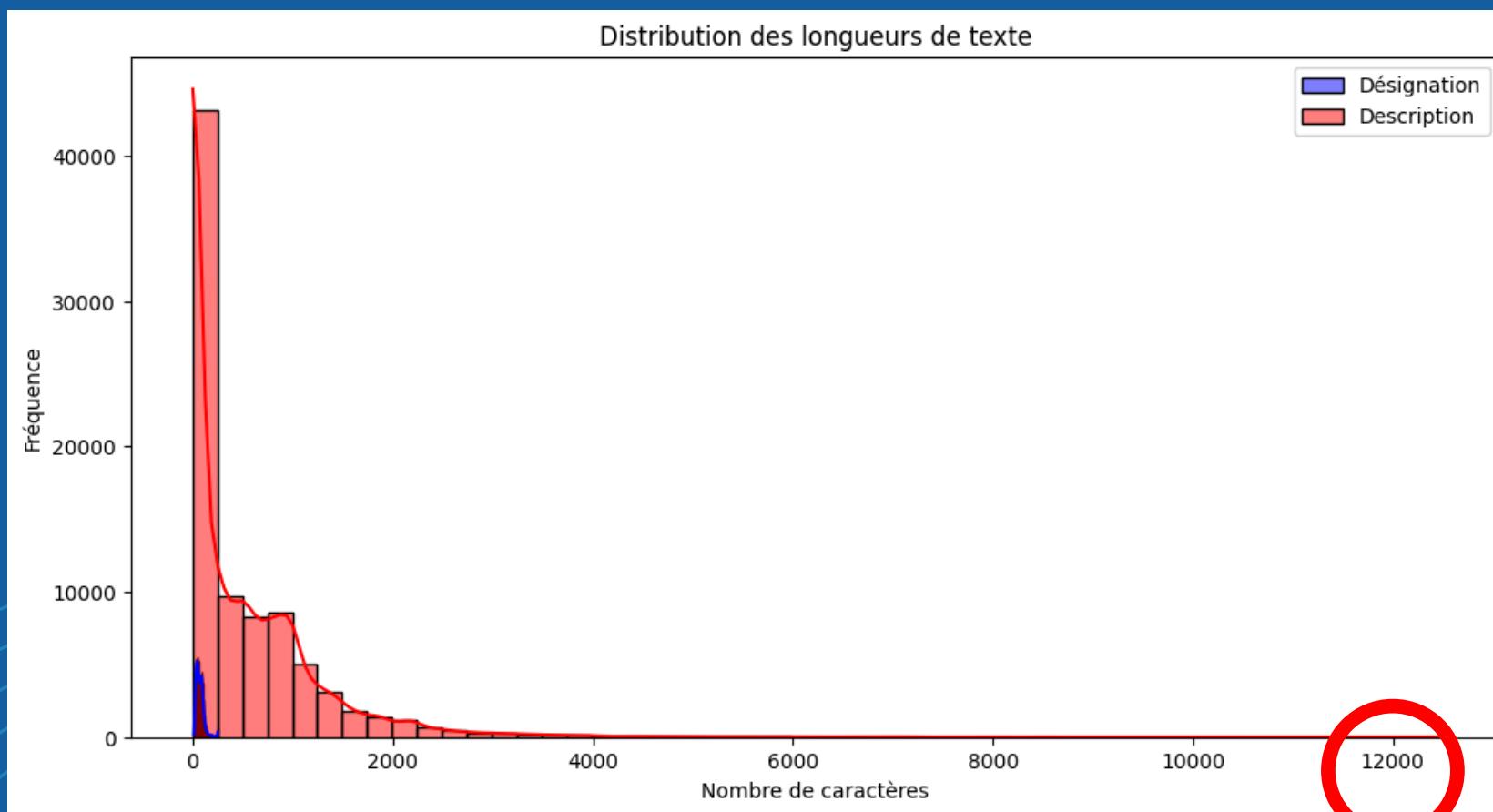
- **Code produit 2583** : largement représenté, avec **10209 occurrences**
- **Code produit 1180** : sous représenté avec **764 occurrences**

Impact : entraîne un **biais dans les prédictions** du modèle, avec une préférence pour les classes majoritaires  
=> Solution : rééquilibrer les classes

## 1. Analyse exploratoire des données : Texte

### Description des variables “designations” et “description”

Variable	X_train	X_test
Désignation Min	11 caractères	11 caractères
Désignation Max	250 caractères	250 caractères
Description Min	0 caractères	0 caractères
Description Max	12451 caractères	22299 caractères

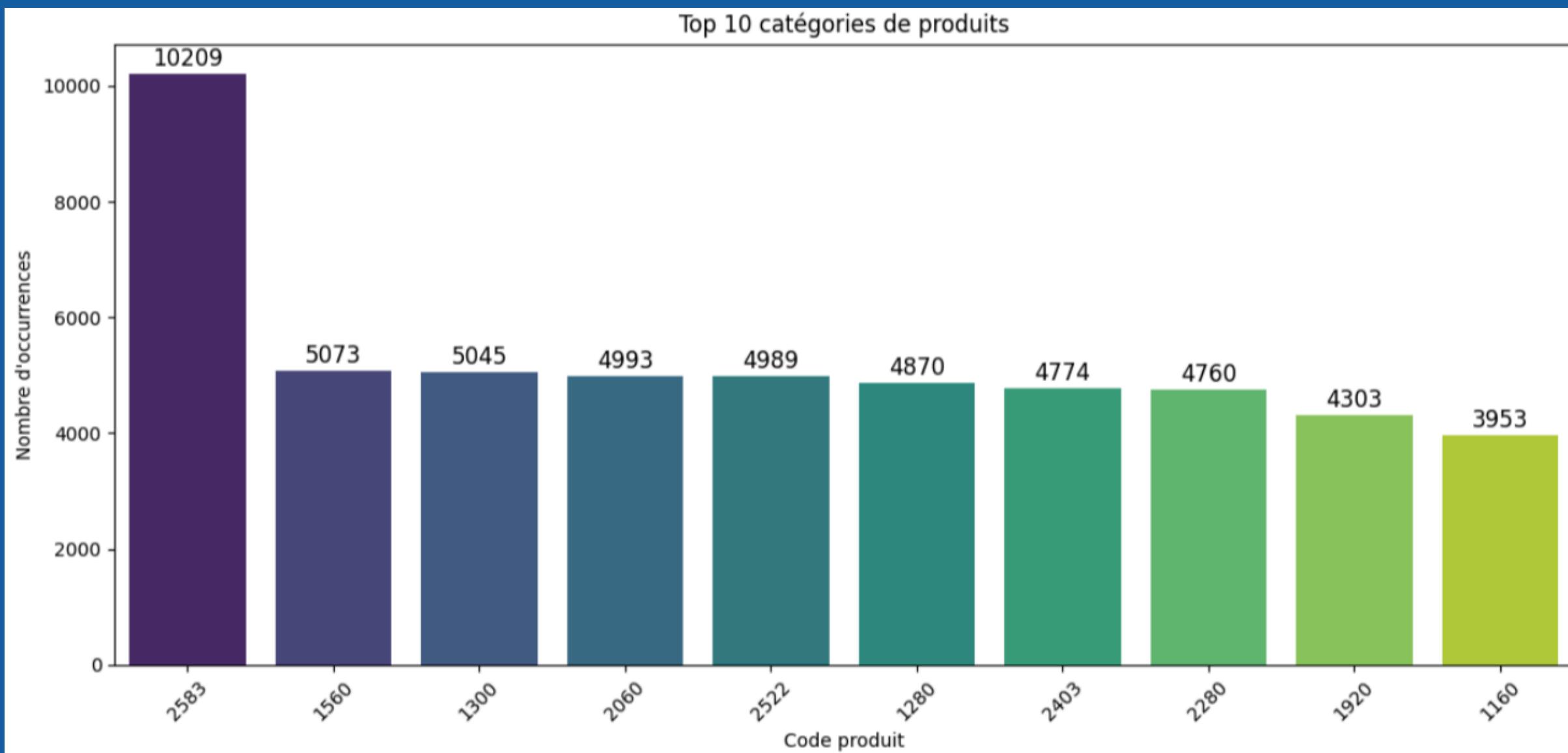


X\_train

X\_test

## 1. Analyse exploratoire des données : **Texte**

### Top 10 des catégories de produits



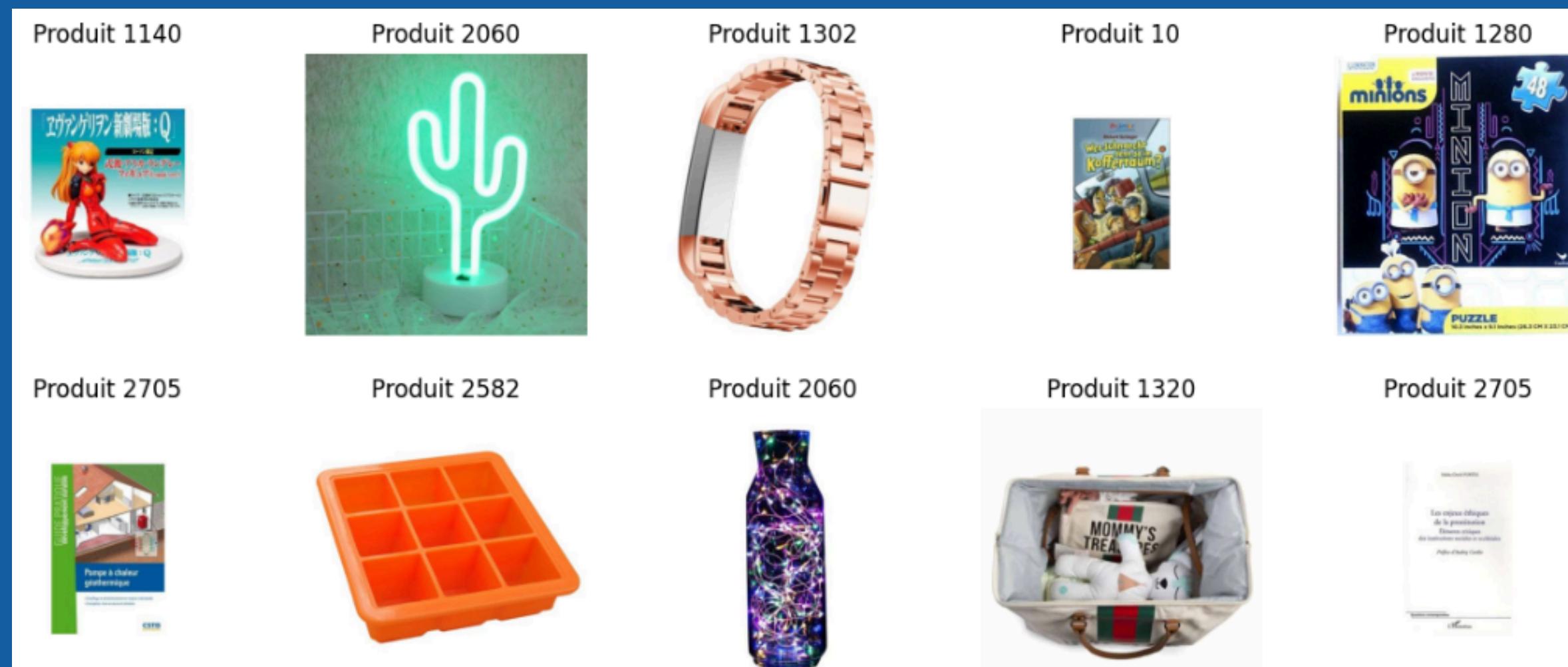
Hypothèses du top 3 des catégories de code produit :

- Code produit 2583 : équipements extérieurs (abris de jardin, arrosoir, ...)
- Code produit 1560 : meubles (canapé, fateuils, ...)
- Code produit 1300 : produit électroniques (drônes, appareils électroniques)

# 1. Analyse exploratoire des données : **Image**

## Les images

- Toutes les images sont au format **jpg**
- Nombre d'images d'entraînement : **84916**
- Nombre d'images de test : **13812**
- Les images ont les mêmes dimensions : **500 pixels** de hauteur et de largeur.



*Exemple d'images dans l'ensemble d'entraînement*

# 1. Analyse exploratoire des données : **Image**

**Produit: 2583**

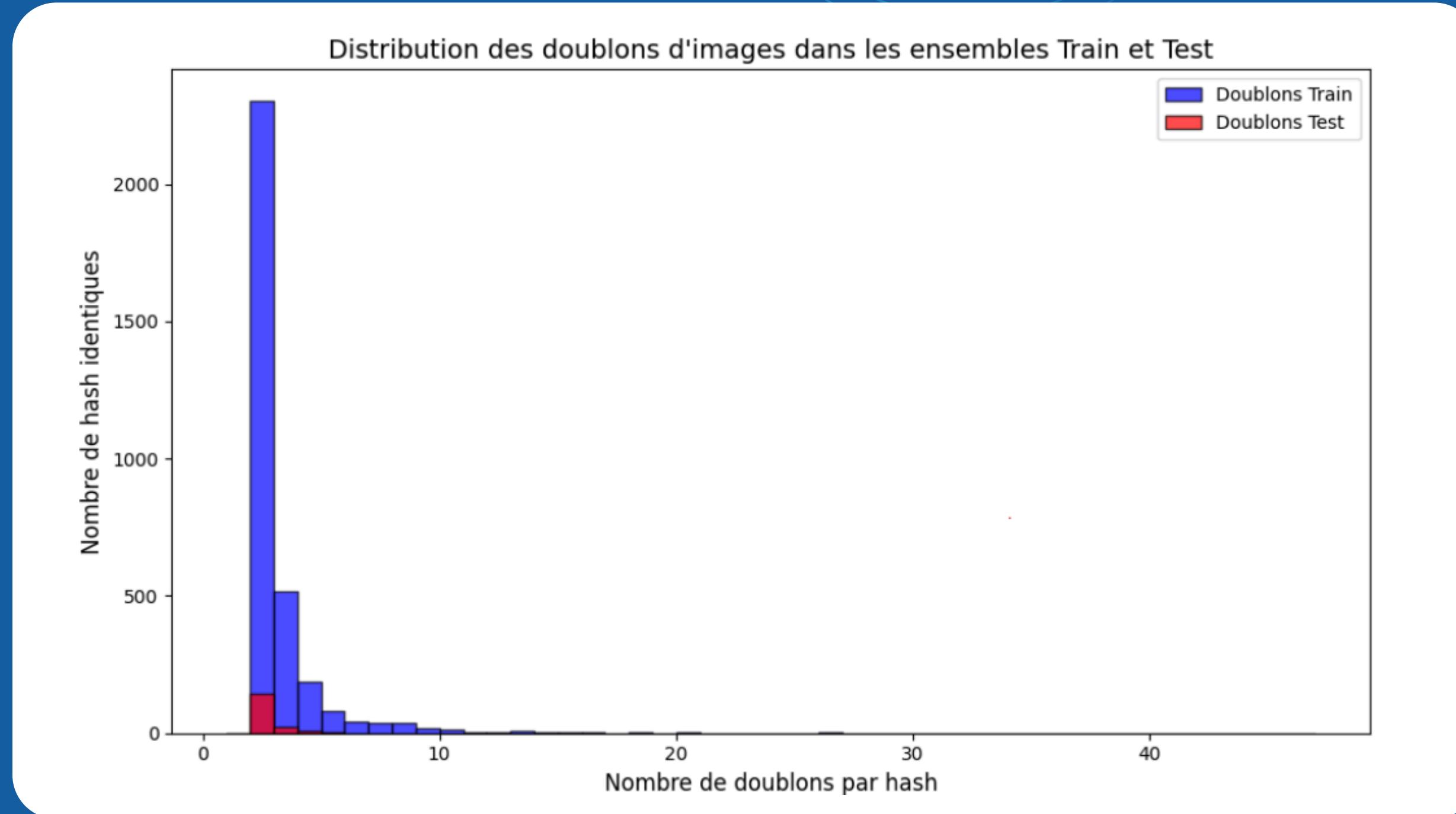
**Titre: Nature Housse pour barbecue au gaz PE 63x165x90 cm 6031605**

Cette housse de Nature pour le barbecue au gaz protège votre barbecue contre les conditions météorologiques défavorables !



correspond bien aux équipements extérieurs

# 1. Analyse exploratoire des données : Image



- **Doublons nombreux dans l'entraînement :** Plus de 2000 occurrences de doublons partageant le même hash
- **Doublons plus rares dans le test :** Moins fréquents et dispersés

## 2. Nettoyage et Traitement des données : Texte

### Comment avons nous prétraiter nos variables textuelles ?

Voici un extrait de “description” :

*Tente pliante V3S5 Pro PVC 500 gr/m<sup>2</sup> – 3 x 4m50.Que vous soyez **un** particulier pour votre jardin **ou un** professionnel pour stand commercial **ou** pour **vos** réceptions le barnum V3S5 Pro de 135 m<sup>2</sup> sera vous combler.Imaginez un **<strong>stand</strong>** robuste léger adaptable à chacun pliable **et** peu encombrant... Le déploiement ultra-rapide et le réglage de la hauteur se font maintenant via des poignées **d’indexage.***

#### Nettoyage

- Suppression des balises HTML
- Conversion des entités HTML
- Conversion en minuscules
- Suppression des accents
- Suppression des caractères spéciaux
- Suppression des mots vides (stopwords)
- Suppression des mots trop courts (1-2 lettres) et trop fréquents

#### Traitement

- Tokenisation : Diviser le texte en mots
- Application du stemming : réduction des mots à leur racine

#### Vectorisation

- TF IDF
- Word2vec

## 2. Nettoyage et Traitement des données : Texte

## Nuage de mots des désignations



# Nuage de mots des descrip

## ► Nuage de mots des titres

## Nuage de mots des descriptions



## 2. Nettoyage et Traitement des données : **Image**

### Comment avons nous prétraiter nos images ?

- **Suppression des doublons** : chaque image a été soumise à un processus de déduplication basé sur le calcul de son empreinte MD5. Toutes les occurrences d'un même contenu visuel ont été détectées et un seul exemplaire a été conservé par groupe de doublons.
- **Réduction de dimension** : image standardisé à  $224 \times 224$  pixels et application d'une normalisation conforme aux statistiques d'ImageNet
  - > garantit une convergence plus rapide et une meilleure stabilité d'entraînement sur les architectures de vision utilisées par la suite

# PARTIE 2

# MODELISATIONS ET PERFORMANCES

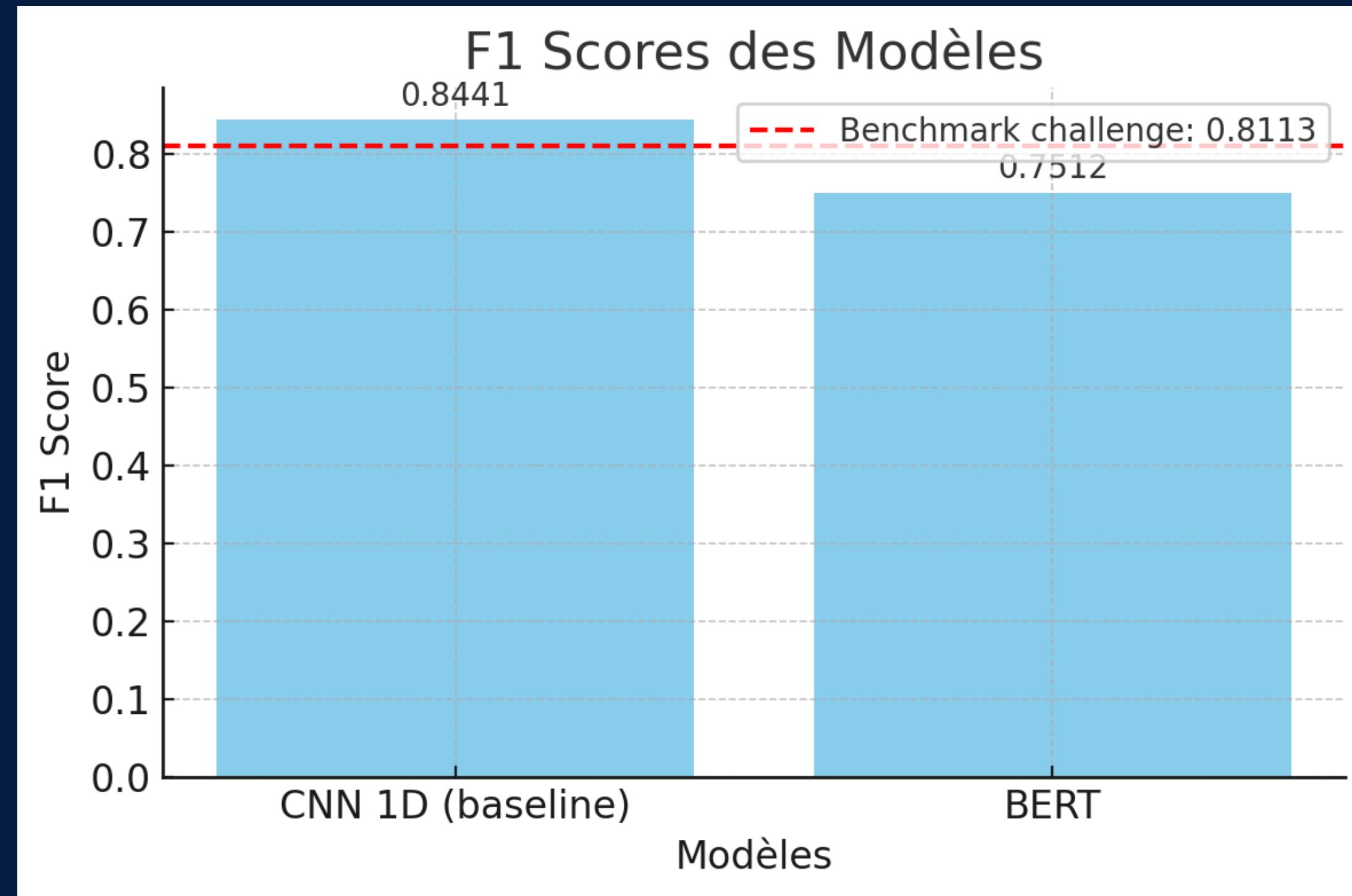
1. Modélisation basée sur le texte
2. Modélisation basée sur les images
3. Fusion Multimodale
4. Benchmark final

## 1. Modélisation basée sur le texte

2 types de modèles déployés :

- **CNN 1D (baseline)** : choisi comme modèle de référence en raison de sa simplicité et de son efficacité pour traiter des séquences de texte, en exploitant les relations locales entre les mots et en extrayant des caractéristiques pertinentes pour la classification.
- **Bert (approche avancée)** : modèle de pointe pour le traitement du langage naturel, reposant sur des mécanismes d'attention qui permettent de capturer des relations globales et complexes entre les mots, en prenant en compte le contexte complet des phrases.

## 1. Modélisation basée sur le texte : Résultat



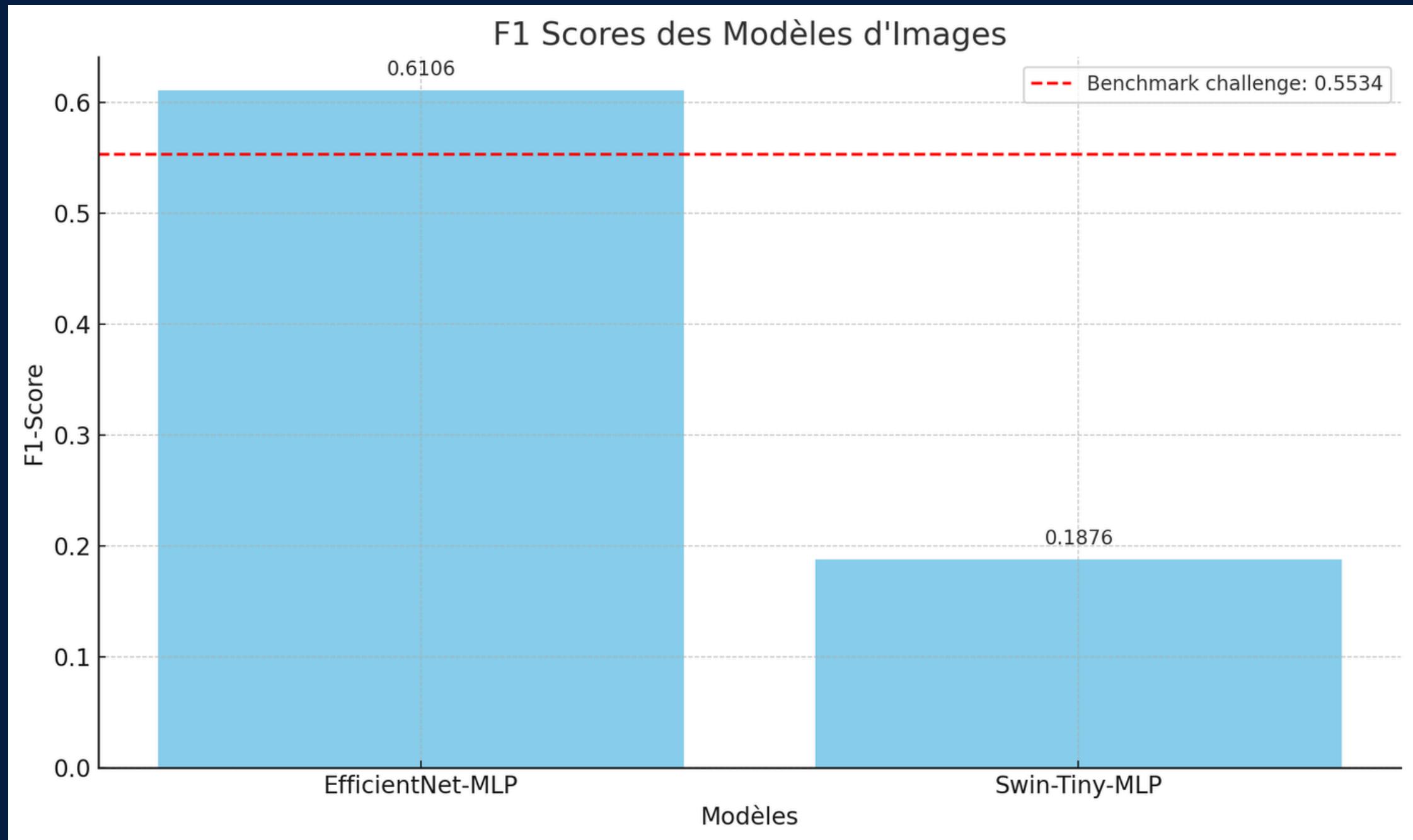
Le modèle CNN 1D a surpassé BERT ainsi que le benchmark du challenge en F1 score, montrant qu'une approche plus simple peut être efficace pour cette tâche.

## 2. Modélisation basée sur les images

2 types de modèles déployés :

- **EfficientNet-BO** : a été utilisé pour extraire des descripteurs visuels des images de produits en raison de son efficacité et de ses bonnes performances en classification d'images. Les caractéristiques extraites ont été moyennées spatialement pour réduire la dimensionnalité à un vecteur de 1280, facilitant ainsi le traitement. Ce vecteur a ensuite été traité par un perceptron multicouche (MLP) pour réaliser la classification des produits.
- **Swin-Tiny** : a été choisi pour sa capacité à capturer des relations globales grâce à ses mécanismes d'attention, permettant ainsi une meilleure modélisation des images complexes. Nous avons réduit les cartes de caractéristiques extraites par le modèle à un vecteur global pour le rendre compatible avec la classification. Ce vecteur a été ensuite envoyé dans un perceptron multicouche (MLP) pour effectuer la classification des produits.

## 2. Modélisation basée sur les images : Résultat



EfficientNet-BO a surpassé le benchmark avec un F1-score de 0.6106, prouvant son efficacité pour capturer les caractéristiques visuelles des produits.

### 3. Fusion Multimodale

3 approches

#### 1. Concaténation simple (MLP)

Score F1 : 0.8521

- Concaténation du vecteur 1280-D d'EfficientNet-BO avec le vecteur 300-D de Word2Vec sur les descriptions de produits.
- Le vecteur concaténé a été traité par un MLP à trois couches pour la classification.
- Cette approche a montré une complémentarité puissante entre les informations visuelles et textuelles.

#### 2. Fusion avec attention cross-modal

Score F1 : 0.8593

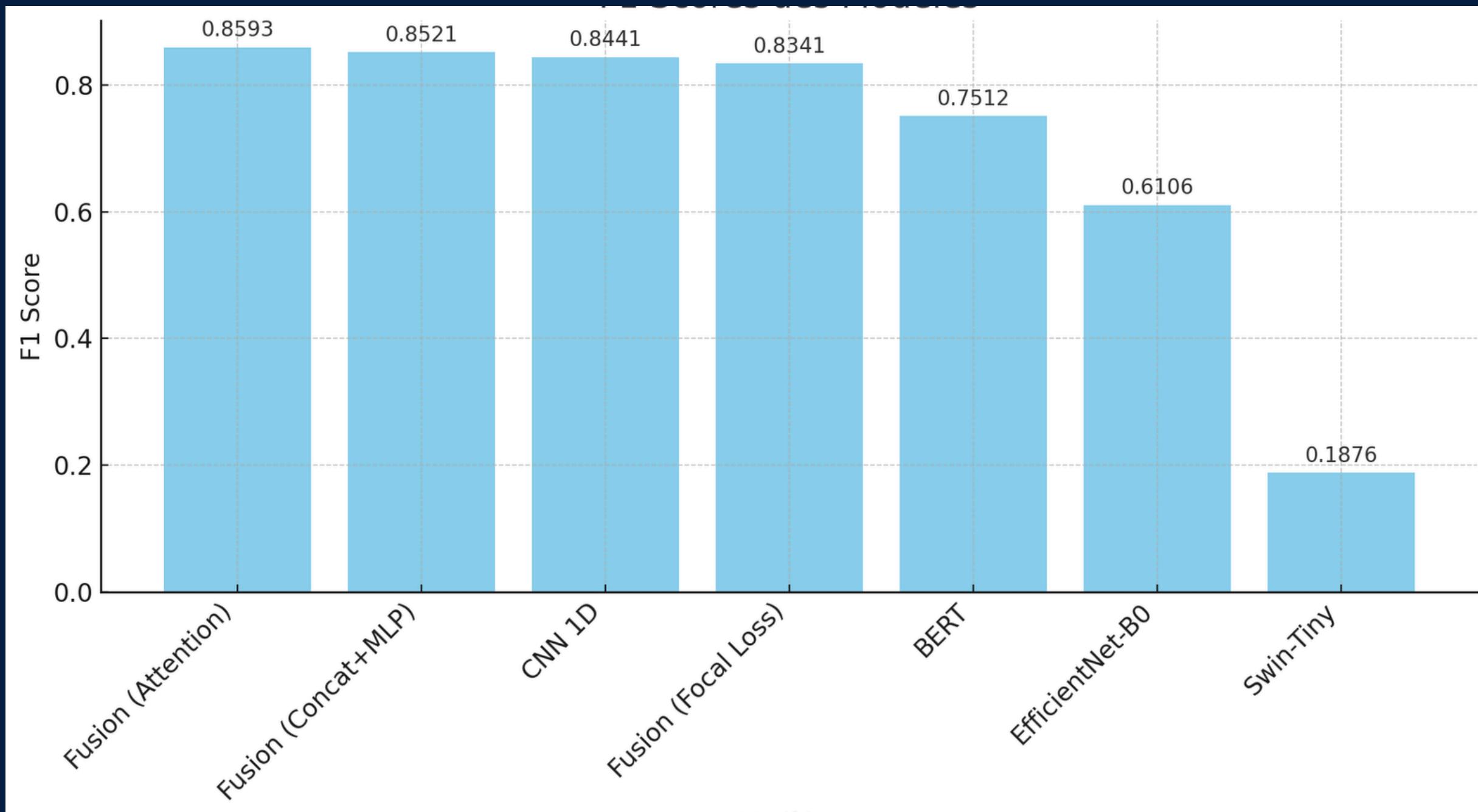
- Les vecteurs d'image et de texte ont été projetés en deux "tokens" de même dimension.
- Ces tokens ont été traités par un bloc MultiHeadAttention, puis agrégés pour la classification.
- Cette approche a permis de modéliser plus finement les interactions entre les modalités, améliorant légèrement le F1-score.

#### 3. Fusion avec Focal Loss

Score F1 : 0.8341

- Utilisation de la Focal Loss ( $\gamma = 2$ ) pour accentuer l'apprentissage des classes rares. Malgré une légère amélioration sur les classes rares, cette approche a diminué la performance globale par rapport à l'MLP simple. Cela suggère que le déséquilibre des classes n'était pas le principal facteur limitant et que se concentrer sur les classes rares a nui à la précision des classes majoritaires.

### 3. Benchmark final



Le modèle le + performant est un modèle de fusion multimodal avec  
un mécanisme d'attention cross-modal

---

## PARTIE 4

# CONCLUSION ET PERSPECTIVES

## Conclusion :

- Les informations textuelles surpassent nettement les informations visuelles seules, même avec des architectures complexes.
- La fusion multimodale, combinant les représentations visuelles et textuelles avec un mécanisme d'attention cross-modal, est la plus performante.
- La complémentarité entre texte et image montre sa force dans l'amélioration des performances.
- Certaines catégories demeurent difficiles à discerner, selon l'analyse de la matrice de confusion.

## Perspectives :

- Explorer des modèles multimodaux avancés comme ViLBERT ou MMBT pour renforcer la synergie image–texte.

# NOUS VOUS REMERCIONS POUR VOTRE ECOUTE !



Charlotte  
CEGARRA  
Data Scientist



Salma  
BENMOUSSA  
Data Scientist