DevFest
Perros-Guirec

Eléa Petton

A QUEL POINT DEVONS-NOUS OPTIMISER NOS MODÈLES D'IA ?

Testons, évaluons, comparons !

6 OCTOBRE 2023

ÉLÉA PETTON

Machine Learning Engineer

OVHcloud

AI Solutions Team

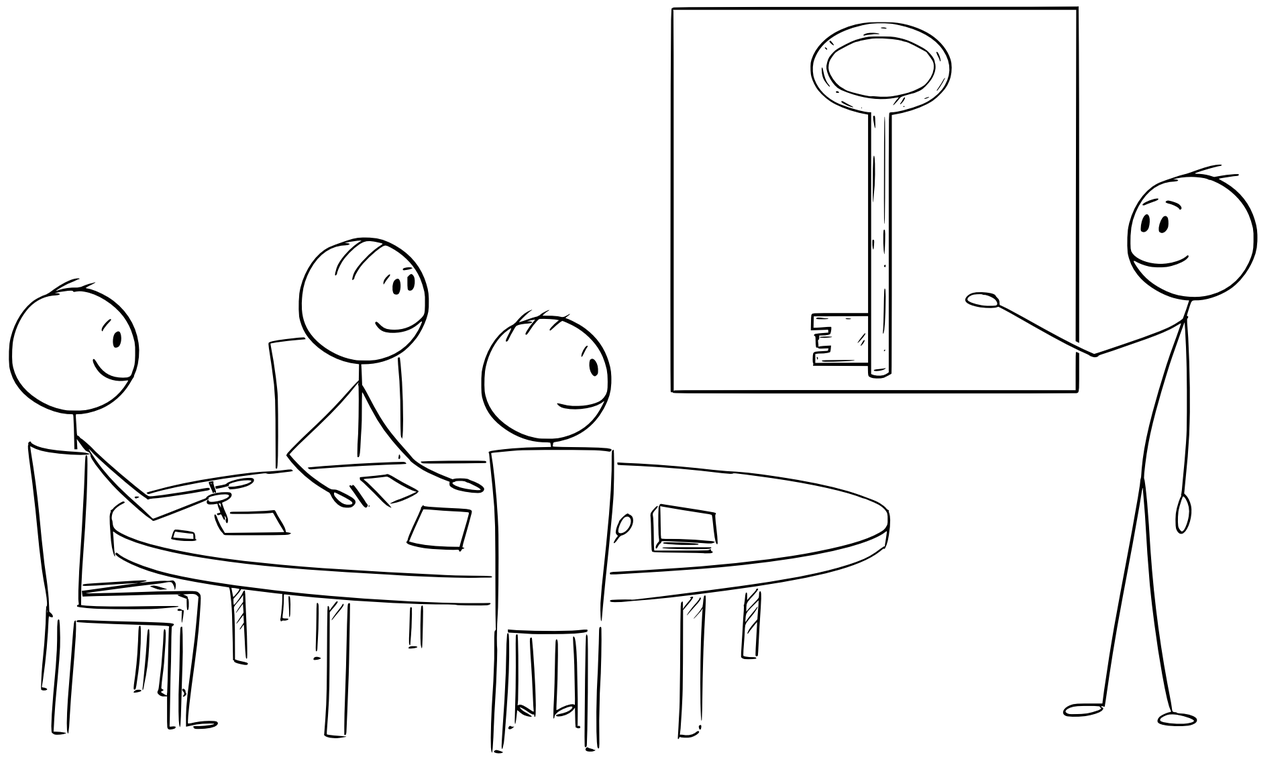
@EleaPetton

eleapttn

Eléa PETTON

OBJECTIFS

L'OPTIMISATION, C'EST LA CLÉ !



DÉFINITIONS

OPTIMISATION

Performance

A quel point le modèle est précis, efficace et pertinent ?

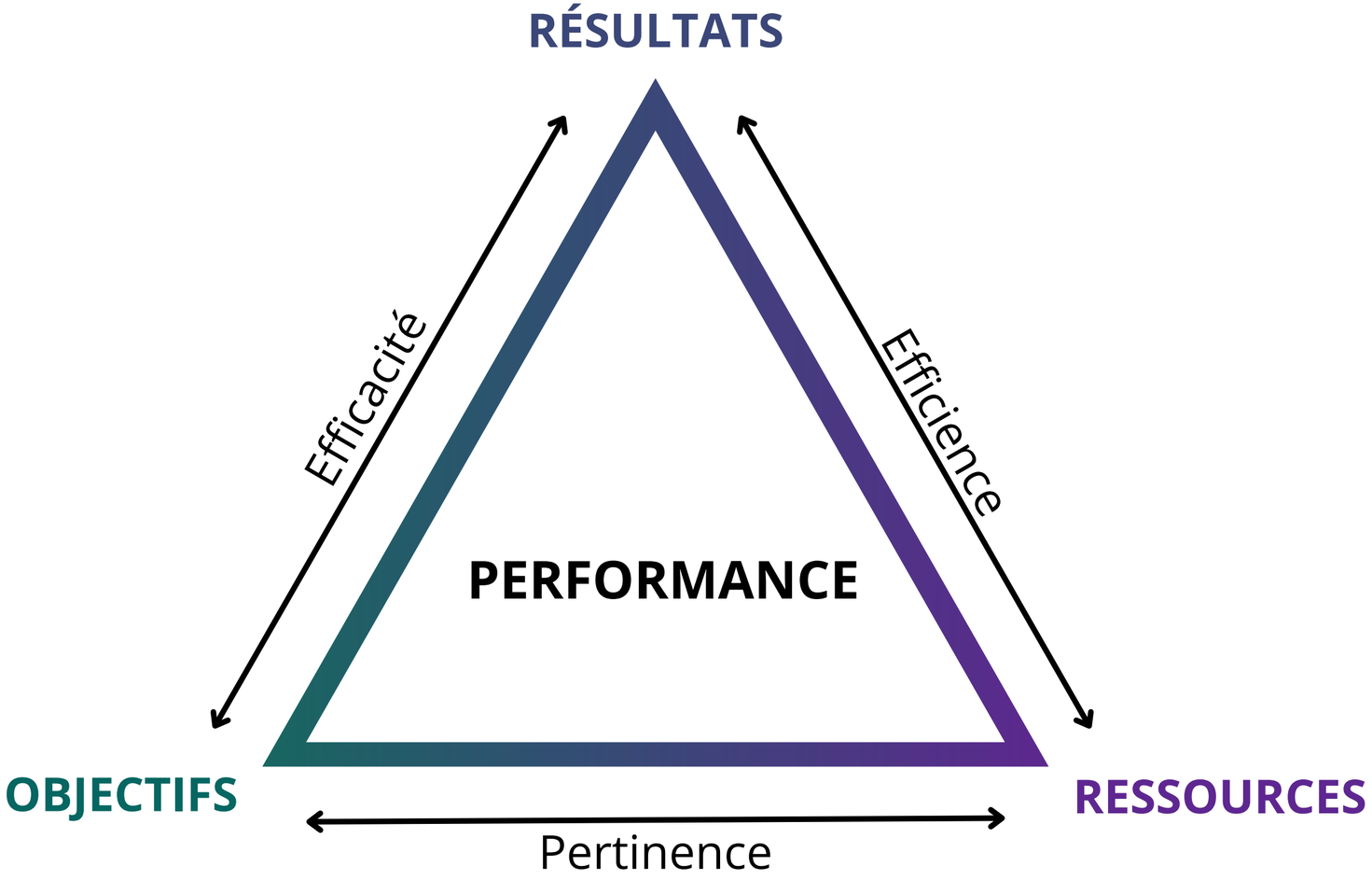
Explicabilité

Le fonctionnement et les résultats du modèle sont-ils intelligibles et transparents ?

Durabilité

Est-ce que les coûts et la quantité d'énergie utilisée permettent de rendre cette IA durable ?

PERFORMANCE



EXPLICABILITÉ

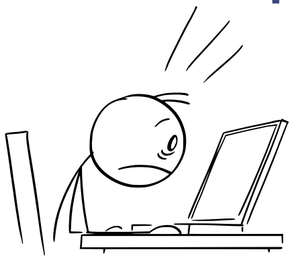
Comment le modèle fonctionne ?

Compréhension du modèle

- Le modèle est **adapté** à l'objectif
- Les résultats sont **intelligibles**
- Le modèle est facilement **debuggable**
- Le modèle est **améliorable** et **maintenable** dans le temps
- Le déploiement en production est **viable**

Qu'est-ce qui me permet de prendre une décision ?

Evaluation des risques



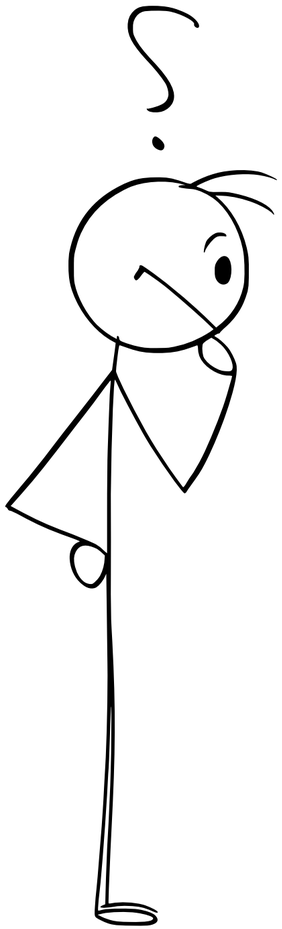
- Le modèle est **robuste**
- Le modèle **respecte** la réglementation
- Le modèle comporte peu de biais **éthiques** et **moraux**
- Le modèle n'impacte pas **négativement** l'utilisateur

Est-ce que je peux croire le modèle ?

Fiabilité du modèle

- L'utilisateur est conscient de l'**impact** que peut avoir le modèle
- L'utilisateur a connaissance des **biais** du modèle
- L'utilisateur peut **interpréter les résultats** et les utiliser

DURABILITÉ

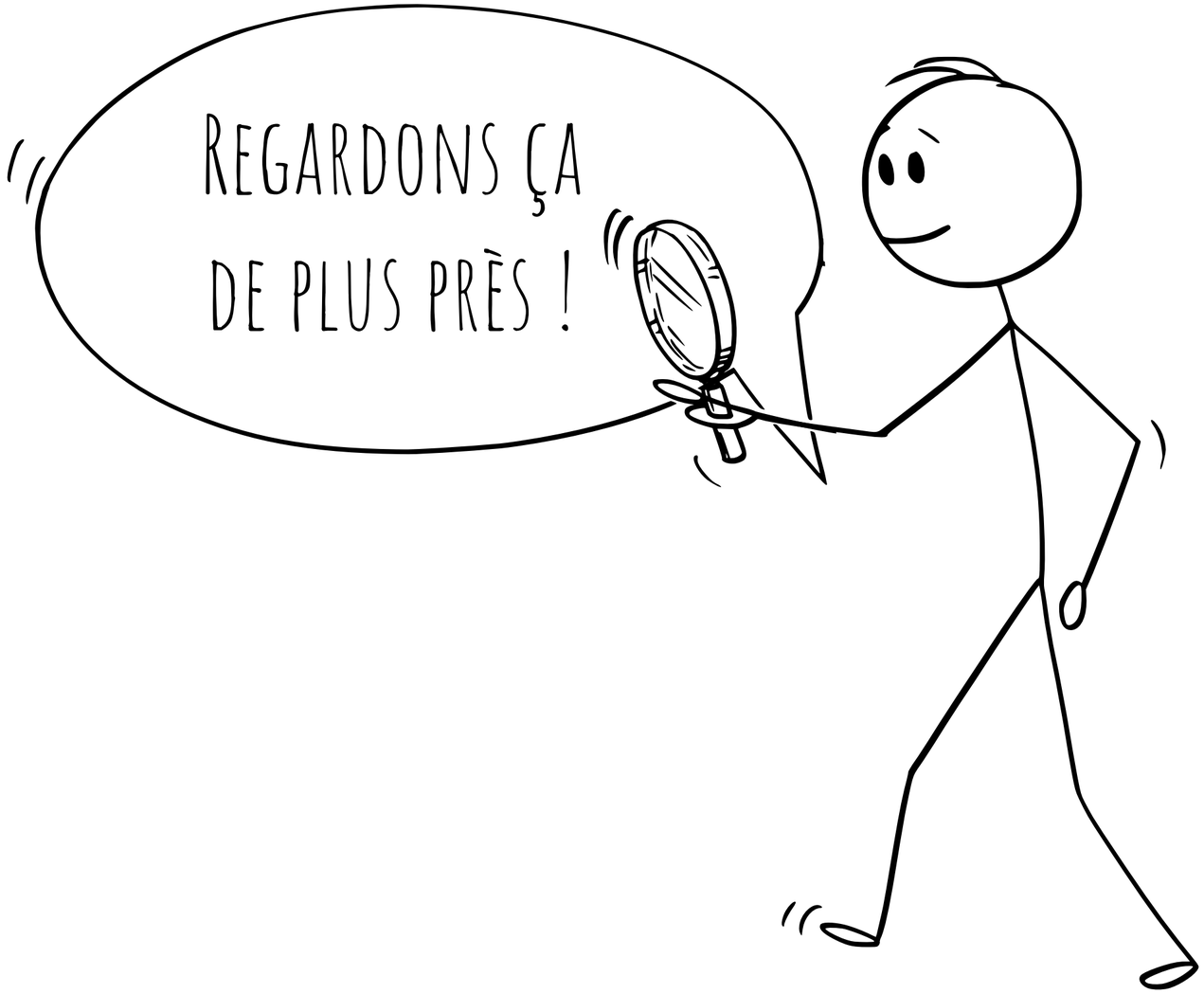


Le modèle d'IA
Opter pour un petit modèle, optimisé

La ressource de calcul
Choisir une alternative aux ressources coûteuse

La consommation énergétique
Être vigilant sur la consommation énergétique et l'empreinte du modèle

OPTIMISATION D'UN MODÈLE D'IA



LES ÉLÉMENTS-CLÉS

Le besoin
Définir la cible, le cas d'usage métier.

La donnée
Récolter, nettoyer, traiter et extraire la donnée.

Le modèle
Construire, entraîner, tester le modèle d'IA.

Les métriques
Evaluer, améliorer, optimiser le modèle.

La décision
Comparer, interpréter, décider de la meilleure solution.

12 ÉTAPES DE L'OPTIMISATION

1- Définir le cas d'usage

2- Récolter la donnée

3- Explorer la donnée

4- Nettoyer la donnée

5- Feature Engineering

6- Sélectionner les caractéristiques

7- Construire le modèle

8- Entraîner le modèle

9- Evaluer le modèle

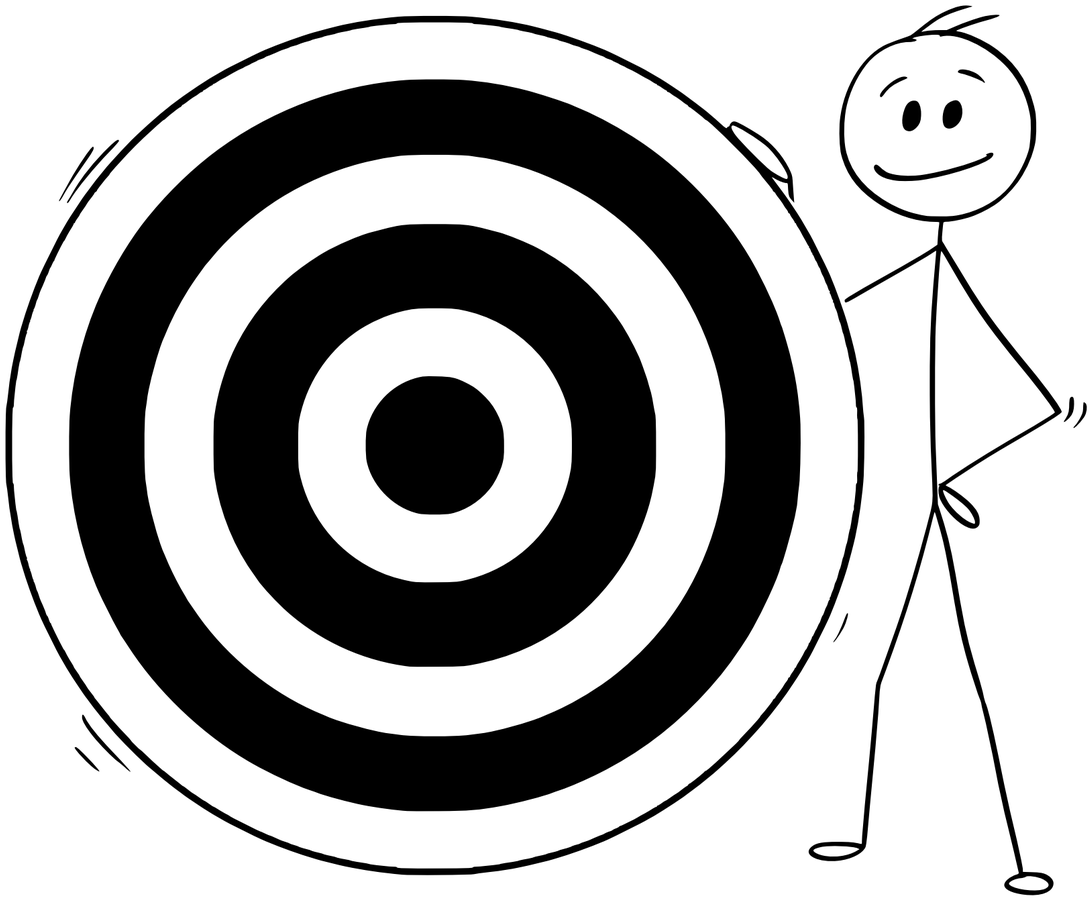
10- Optimiser le modèle

11- Comparer les résultats

12- Décider du modèle à déployer

DÉFINIR LE CAS D'USAGE

Ça, c'est notre objectif !



DÉFINIR LE CAS D'USAGE

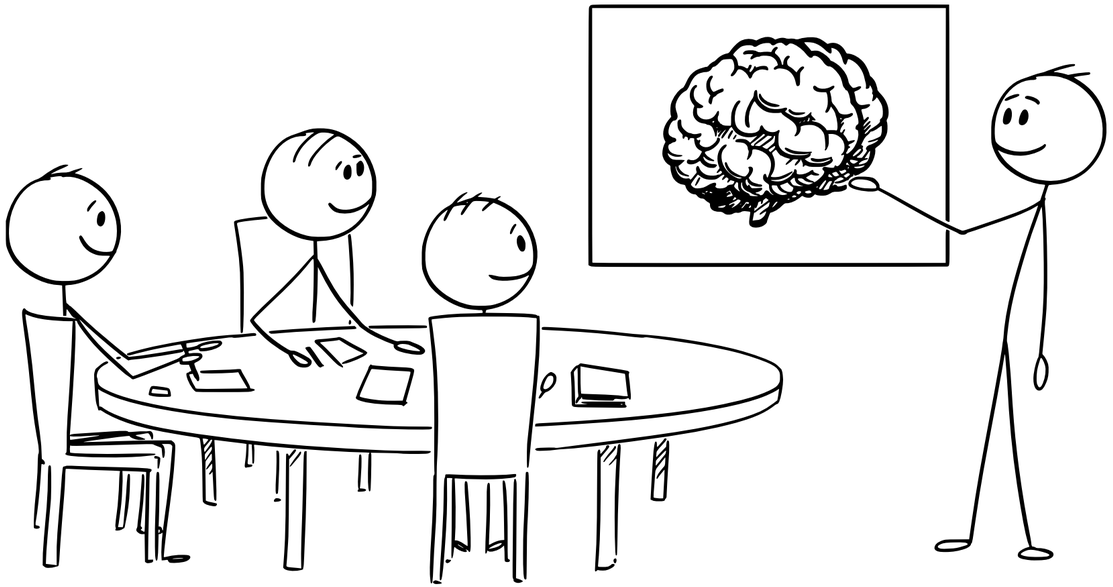
Produit : site de e-commerce de vêtements

Objectif : avoir le sentiment moyen des consommateurs pour pouvoir améliorer les produits et l'expérience client

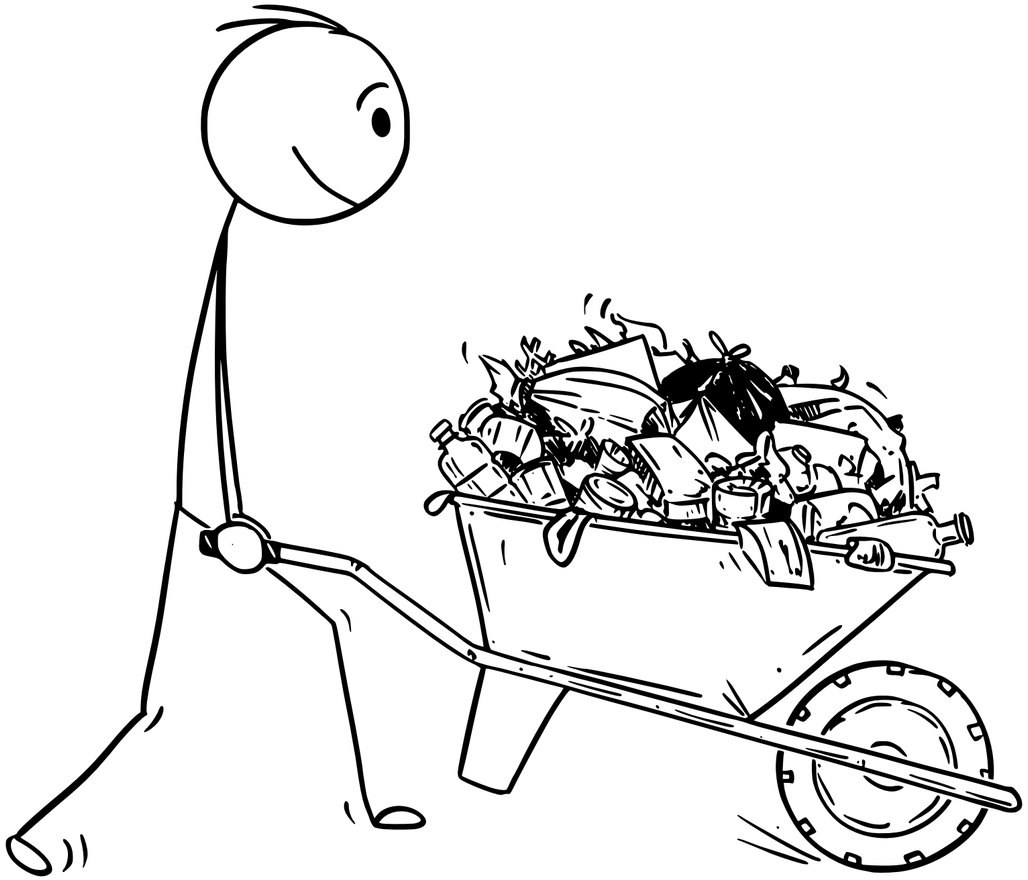
Solution : déployer un modèle d'IA permettant de classifier les avis clients laissés sur les différents produits

Contraintes : budget restreint, utilisation quotidienne

Une IA nous permettrait d'améliorer l'expérience client en se basant sur leurs avis...



Ça fait beaucoup de données...



RÉCOLTER LA DONNÉE

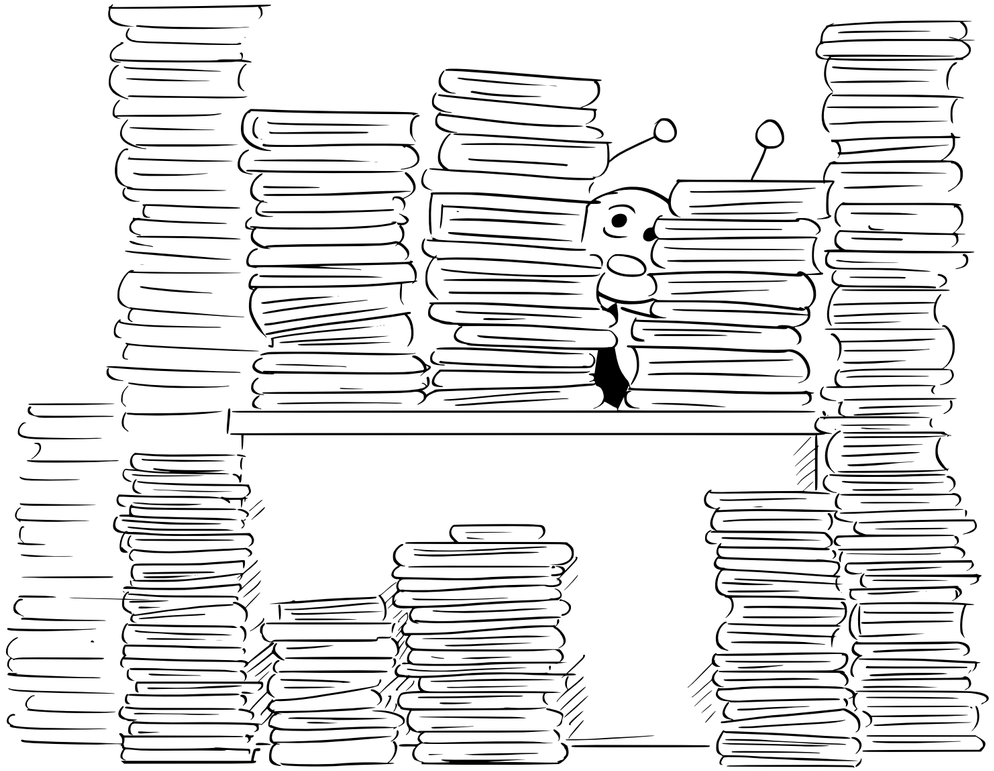
RÉCOLTER LA DONNÉE

Dataset : Women Clothing e-commerce reviews

Description : le jeu de données contient plusieurs informations de natures différentes

- *review_text* : contenu du commentaire
- *age* : âge du client
- *rating* : notation de 1 à 5 étoiles
- *positive_feedback_count* : nombre de retours positifs
- *division_name* : catégorie de taille du produit
- *department_name* : catégorie du produit concerné
- *class_name* : produit concerné
- *recommended_ind* : label pour de la classification binaire

Je crois que j'ai trouvé les données qu'il nous faut !



EXPLORER LA DONNÉE

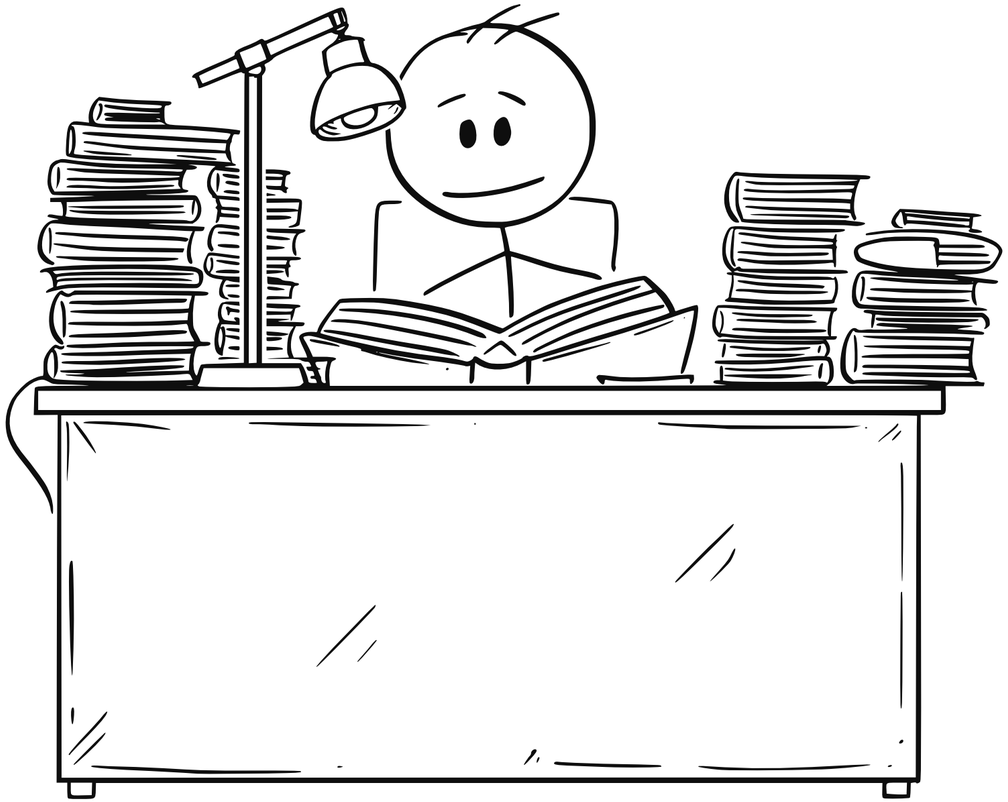
Partons en exploration !

EXPLORER LA DONNÉE

Les éléments à vérifier : vérification de la qualité et de la pertinence du jeu de données

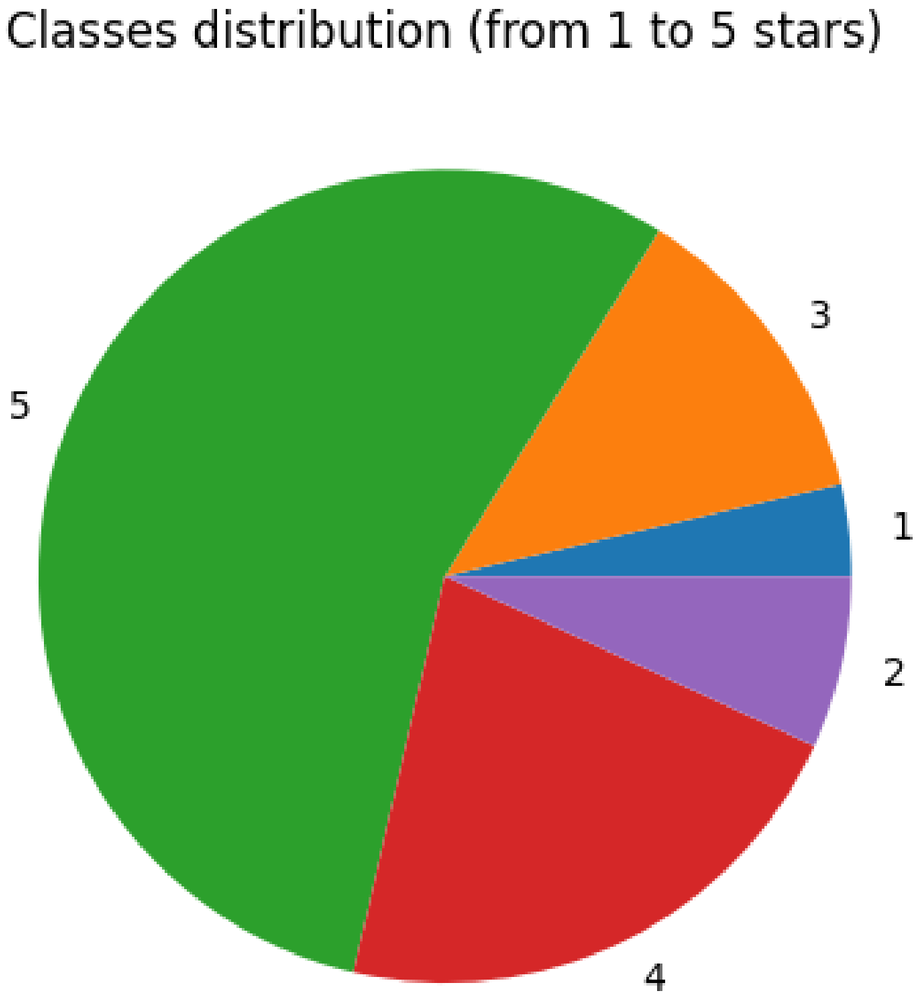
- *Equilibre des classes* : proportions respectées entre les classes
- *Corrélation entre les données* : similarités entre les types de données
- *Longueur des commentaires :* nombre de caractères dans chaque commentaire
- *Pertinence des données :* utilité de l'information
- *Fréquence des mots :* nombre de fois où les mots sont utilisés dans les commentaires

On y voit déjà un petit peu plus clair !



EXPLORER LA DONNÉE

- Equilibre des classes



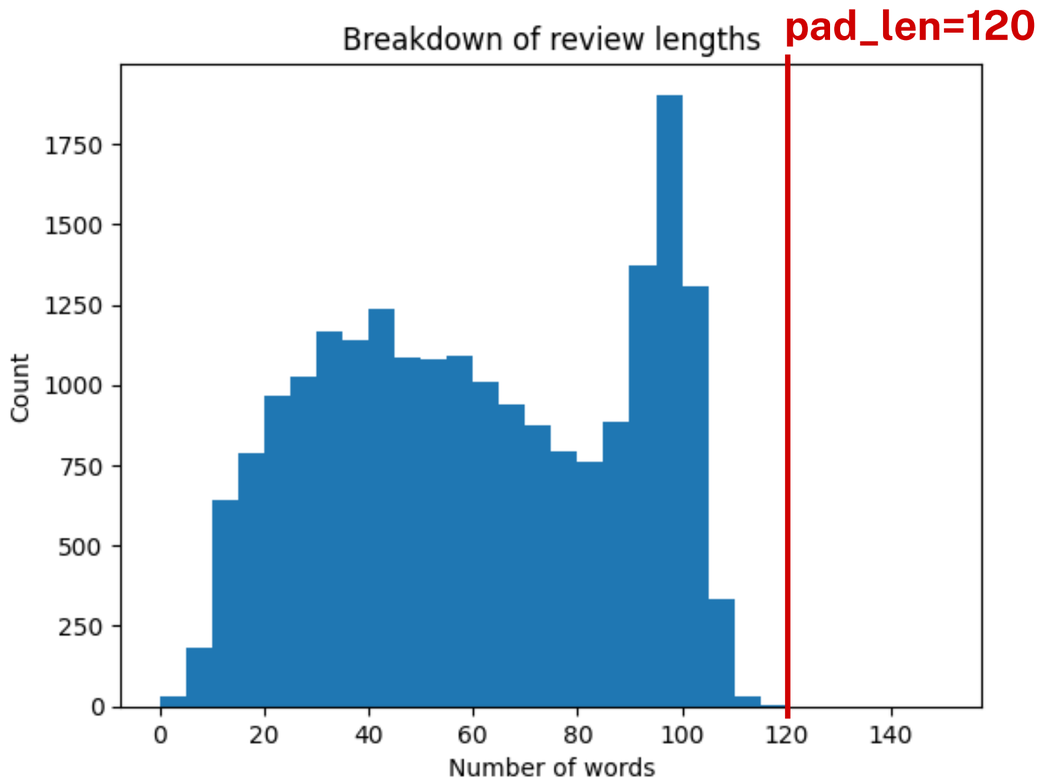
EXPLORER LA DONNÉE

- Equilibre des classes
- Corrélation entre les données

	review_text	age	rating	positive_feedback_count	division_name	department_name	class_name	recommended_ind
0	I loved this shirt until the first time i washed it. it shrunk so much it became unwearable ...	39	1	0	General	Tops	Knits	0

EXPLORER LA DONNÉE

- Equilibre des classes
- Corrélation entre les données
- Longueur des commentaires



EXPLORER LA DONNÉE

- Equilibre des classes
- Corrélation entre les données
- Longueur des commentaires
- Pertinence des données

	review_text	age	rating	positive_feedback_count	division_name	department_name	class_name	recommended_ind
0	I loved this shirt until the first time...	39	1	0	General	Tops	Knits	0

EXPLORER LA DONNÉE

- Equilibre des classes
- Corrélation entre les données
- Longueur des commentaires
- Pertinence des données
- Fréquence des mots

term	frequence
00	302
000	55
0000	1
00020b	2
000mah	3
...	...
zymox	1
zyrtec	1
zz	1
zzzzzzz	1
ítem	3

[13 594 rows x 2 columns]

suppression des mots qui ne sont présents qu'une seule fois

term	frequence>1
00	302
000	55
00020b	2
000mah	3
01	8
...	...
zte	7
zucchini	20
zumba	5
zwave	4
ítem	3

[7 972 rows x 2 columns]

Il est temps de faire du ménage !

NETTOYER LA DONNÉE

NETTOYER LA DONNÉE

Il faut vraiment tout garder ?

Les éléments à éliminer : suppression de certaines informations inutiles pour la compréhension du modèle d'IA

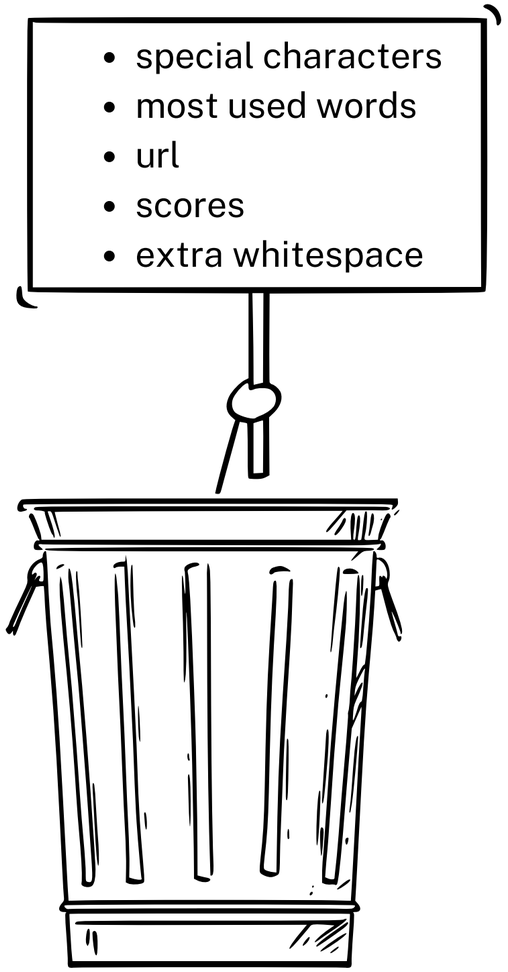
- *Stop Words :* mots les plus communs dans une langue
- *URL :* mots commençant par HTTP
- *Emoji :* ce qui n'est pas textuel
- *Term Frequency :* mots qui apparaissent presque jamais

Le texte à normaliser : rendre les commentaires pertinents et utilisables pour le NLP

La langue à sélectionner : garder uniquement la langue qui nous intéresse (**english**)

NETTOYER LA DONNÉE

- Application du “Data cleaning” sur les reviews

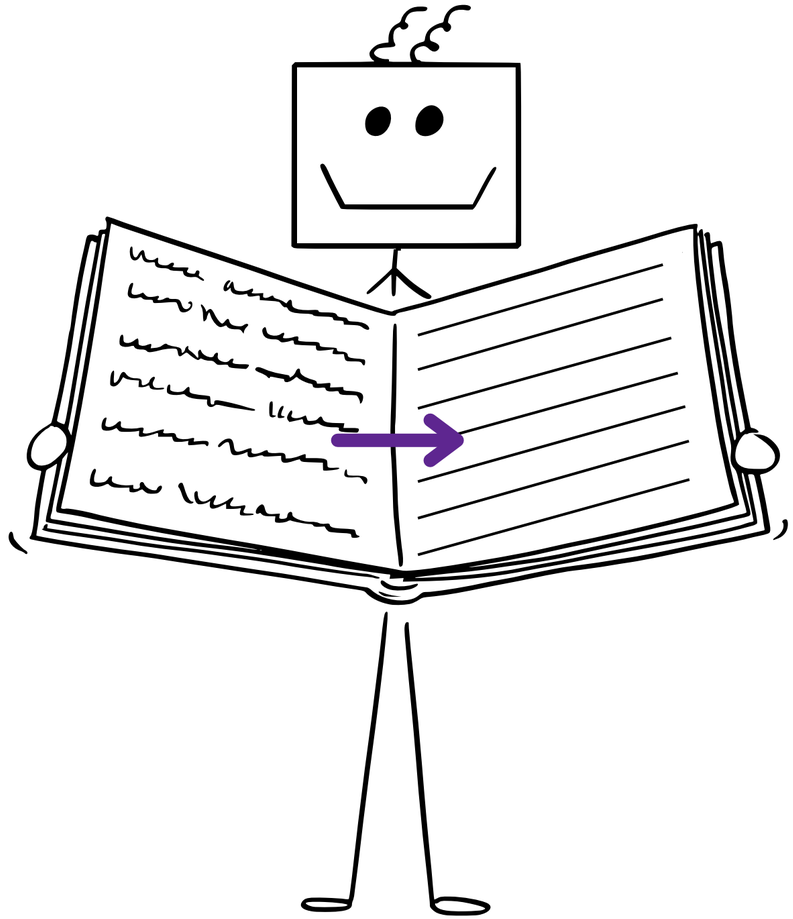


NETTOYER LA DONNÉE

- Application du “Data cleaning” sur les reviews
- Suppression des “Stop Words”

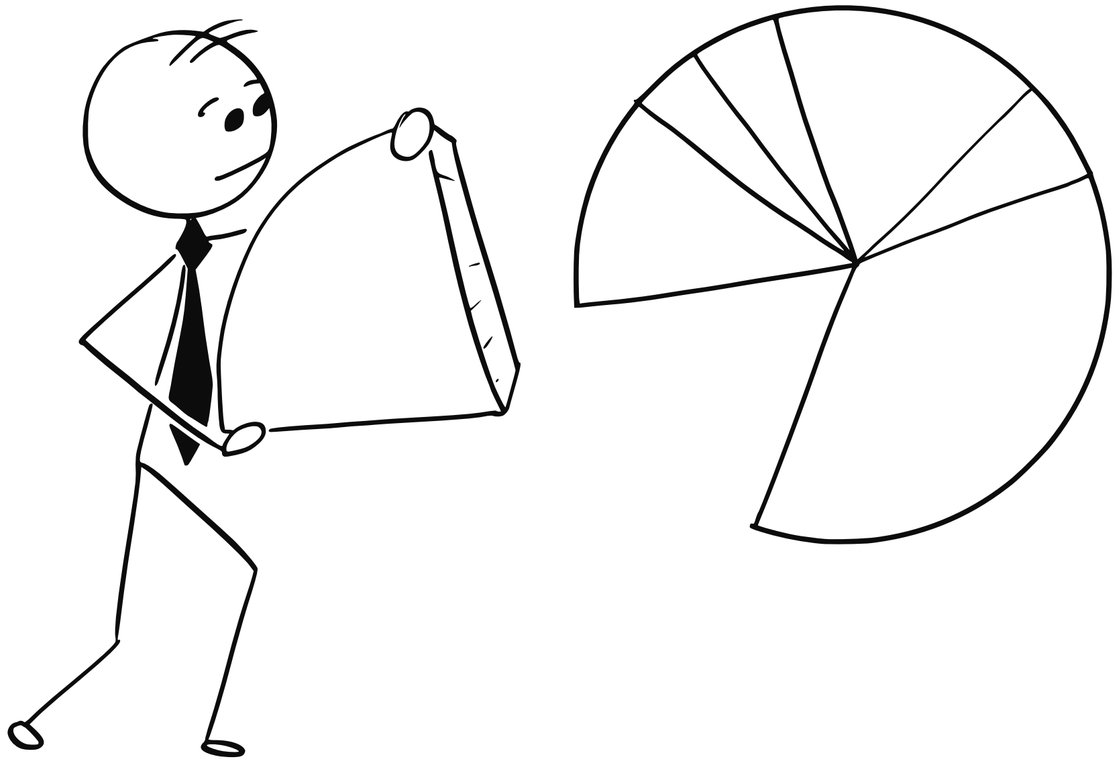
NETTOYER LA DONNÉE

- Application du “Data cleaning” sur les reviews
- Suppression des “Stop Words”
- Standardisation du text



SÉLECTIONNER LES CARACTÉRISTIQUES

C'EST LE MOMENT DE FAIRE UN CHOIX...

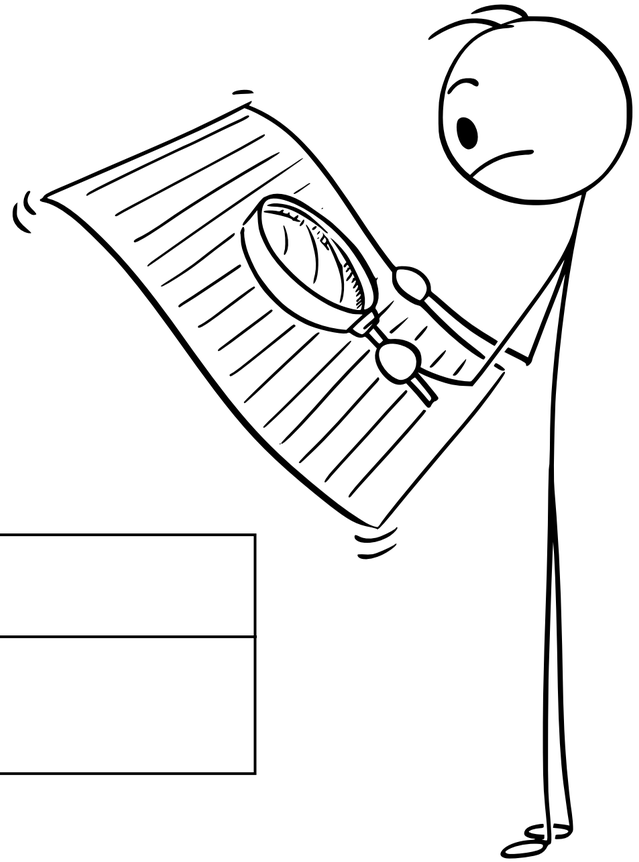


SÉLECTIONNER LES CARACTÉRISTIQUES

ET SI ON N'EN GARDAIT QUE 2 ?

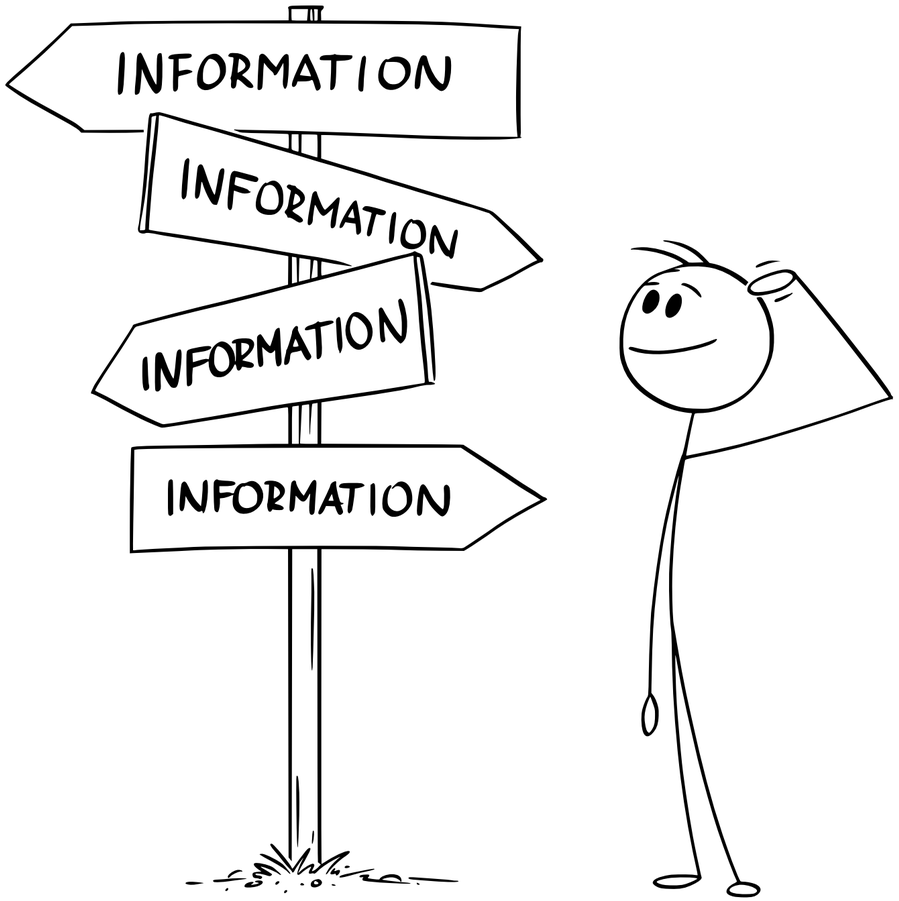
Garder uniquement l'information utile :
sélectionner parmi les types de données lesquelles sont les plus pertinentes

- *rating : notation de 1 à 5 étoiles*
- *review_text : contenu du commentaire*



	rating	review_text
0	1	I loved this shirt until the first time...

Comment utiliser ces informations ?



FEATURE ENGINEERING

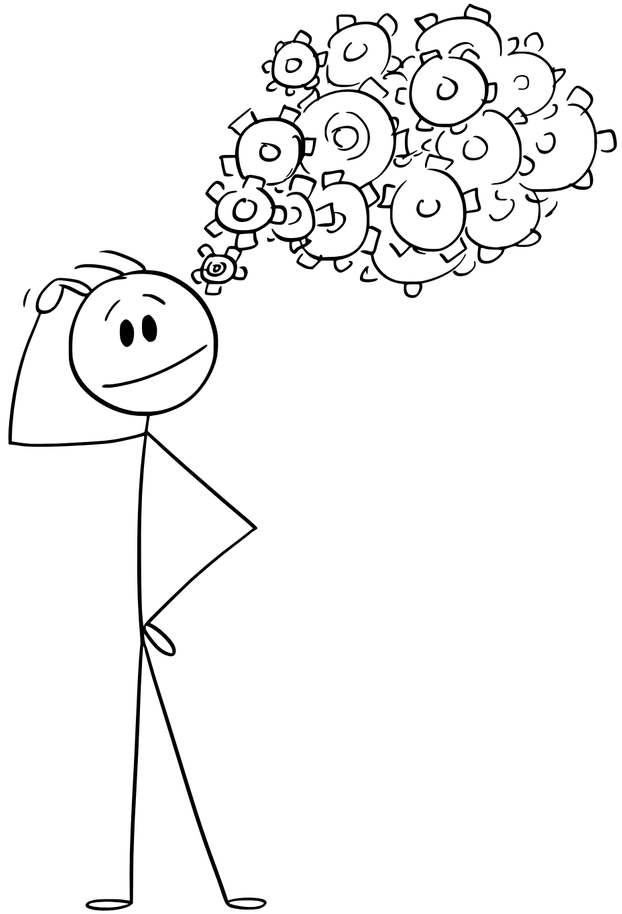
FEATURE ENGINEERING

Comment faire pour que mon IA comprenne ces informations ?

Traiter l'information : harmoniser et rendre intelligible le jeu de données

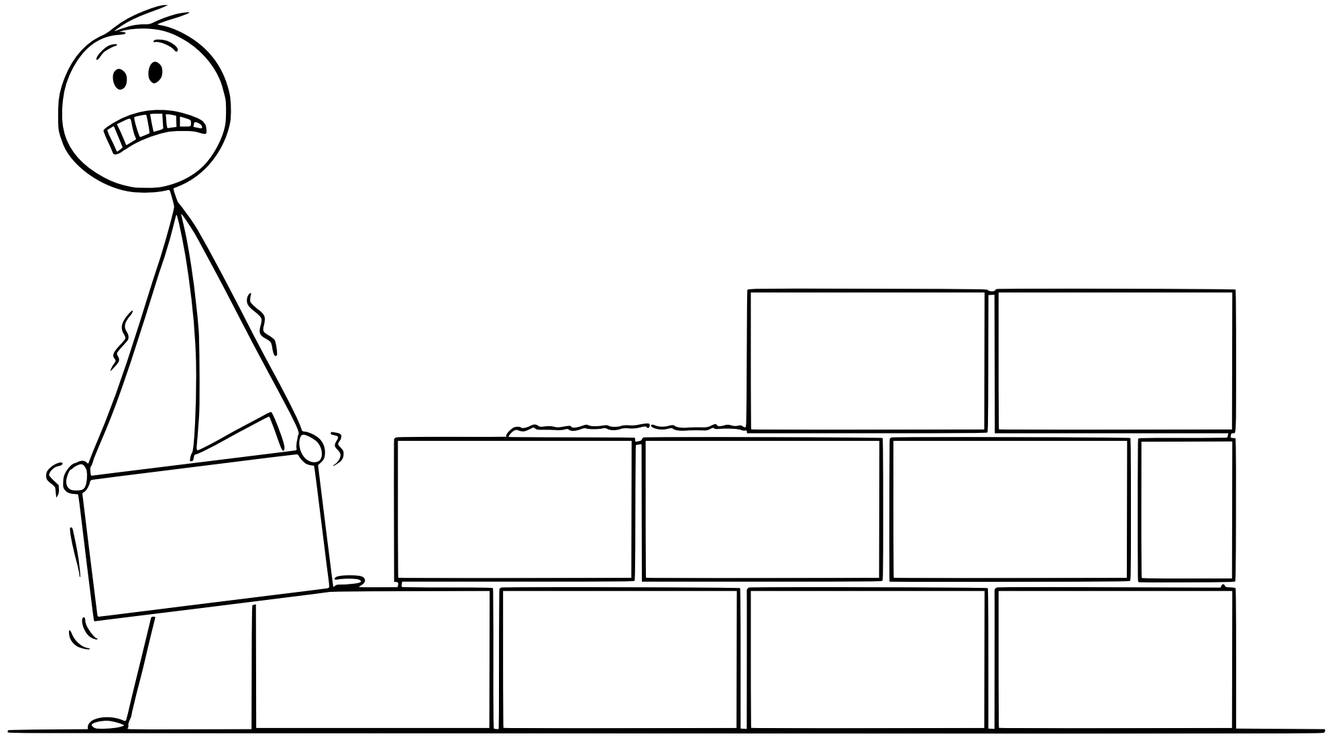
- *Valeurs manquantes* : gestion des cases vides dans le jeu de données
- *Encodage* : gestion des variables catégorielles
- *Standardisation* : égalisation du poids de chaque dimension

rating	*encodage*	*label_id*
1	→	0
2	→	1
...		...
5	→	4



CHOISIR / CONSTRUIRE LE MODÈLE

Quand faut y aller, faut y aller !

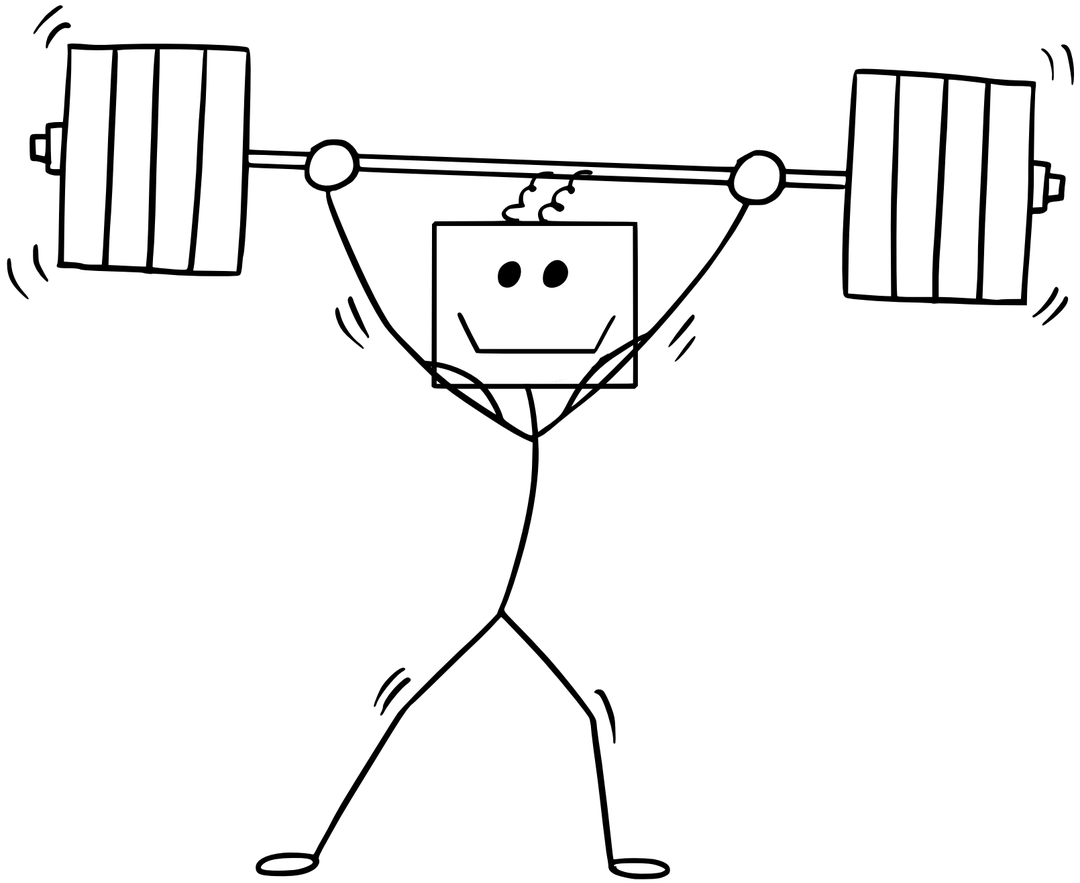


CHOISIR / CONSTRUIRE LE MODÈLE

- **Fine-Tune** un modèle de type “BERT”
- **Fine-Tune** un modèle de type “LSTM”
- Utiliser un modèle existant disponible **“on-shelf”**

Regardons le deuxième modèle

Un jour je serai la meilleure IA...



ENTRAÎNER LES MODÈLES

ENTRAÎNER LES MODÈLES

Entraînement des 2 modèles BERT et LSTM :

- **Précision** des modèles
- **Durée du training** des modèles
- **Consommation** des ressources (1 GPU - Tesla V100S)
- **Coût** de l'entraînement (prix)

ENTRAÎNER LES MODÈLES

Précision des modèles

		BERT	LSTM
Training set	Accuracy	0.852	0.711
	Loss	0.537	0.703
Validation set	Accuracy	0.652	0.653
	Loss	0.877	0.888

EVALUER LES MODÈLES

Durée du training / consommation des ressources

	BERT	LSTM
Temps de l'entraînement (**min**)	11	24
Consommation GPU (%)	99	25

EVALUER LES MODÈLES

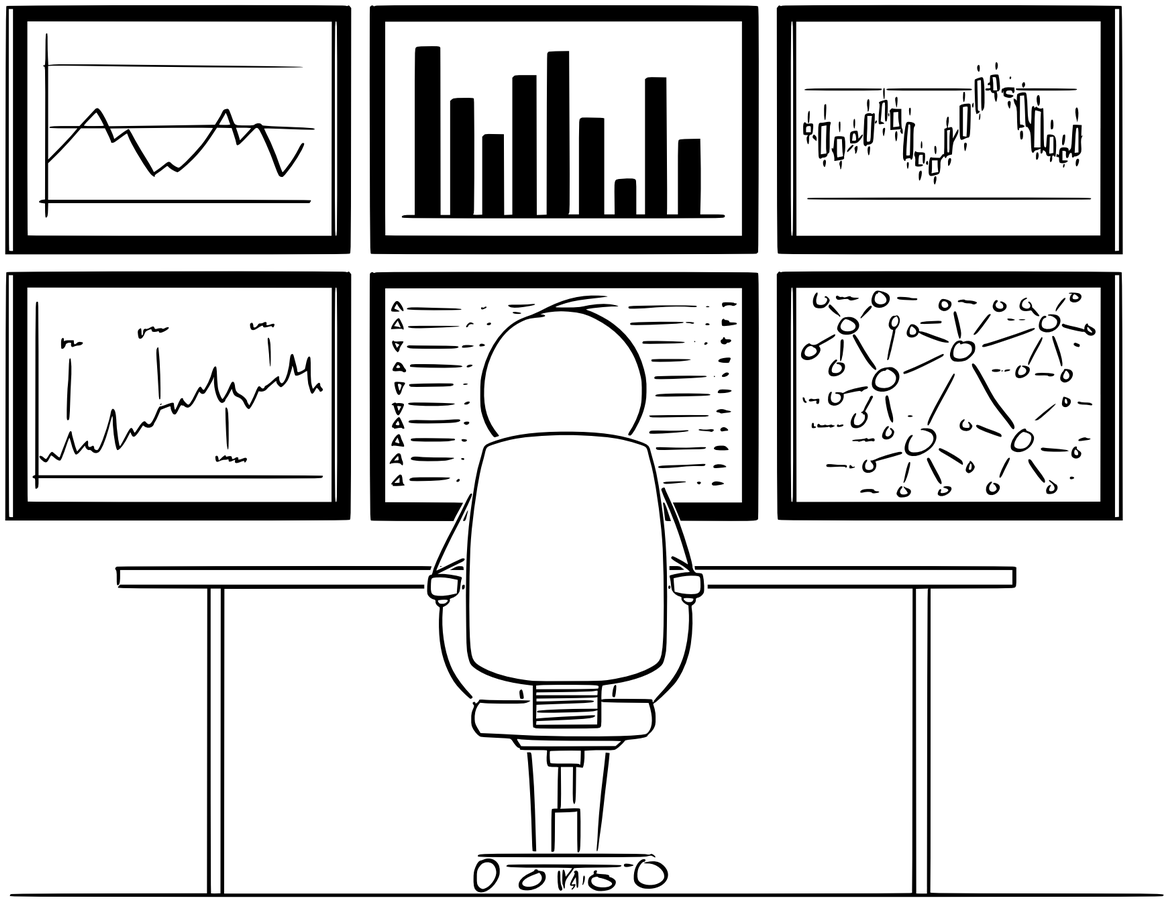
Coût de l’entraînement (prix)

	BERT	LSTM
Temps d’inférence (**sec**)	11	24
Nombre de GPU (**Tesla V100S**)	1	1
Coût total HT (€)*	**0.33**	**0.72**

**GPU Tesla V100S => 1.93€/heure HT*

EVALUER LES MODÈLES

ÇA FAIT BEAUCOUP D'INFORMATIONS...



EVALUER LES MODÈLES

Evaluation des 3 modèles sur le dataset de test :

- **Précision** des modèles
- **Latence** des modèles
- **Consommation** des ressources (GPU)
- **Coût** de l'inférence (prix)

EVALUER LES MODÈLES

Précision des modèles

	BERT	LSTM	LETTRIA
accuracy	0.63	0.63	0.63
precision	0.47	0.44	0.50
recall	0.46	0.39	0.59
f1 score	0.46	0.40	0.53

EVALUER LES MODÈLES

Latence et consommation des ressources

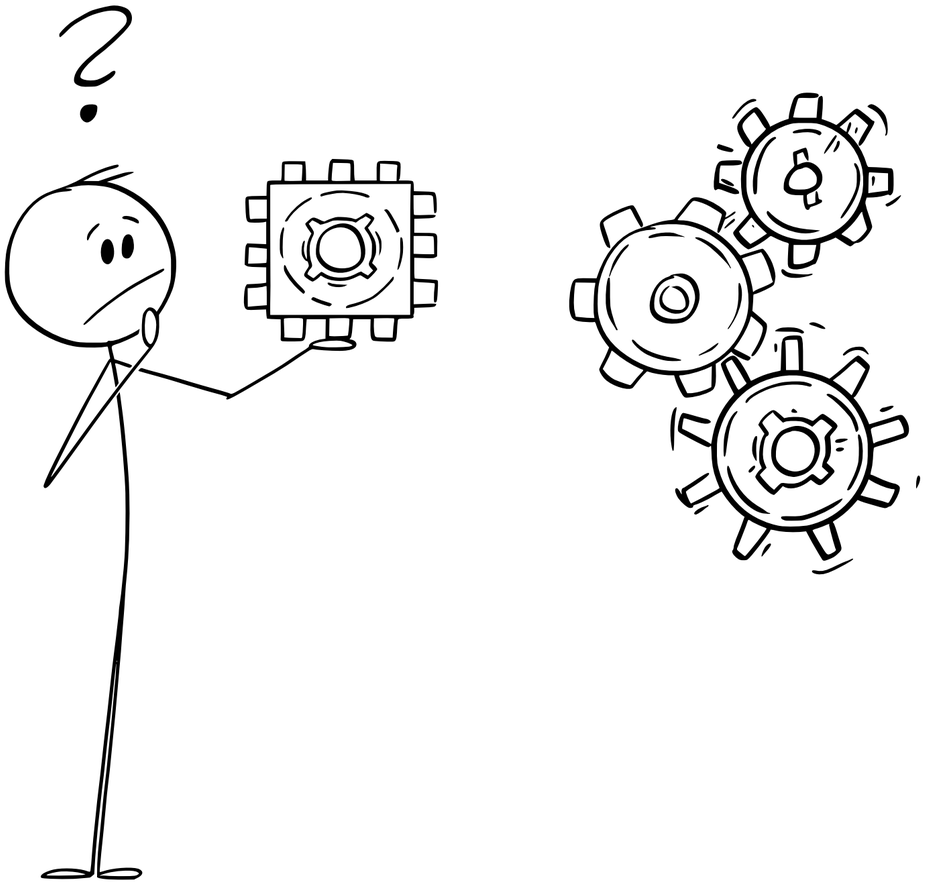
	BERT	LSTM	LETTRIA
Temps d'inférence (**sec**)	1.8	1.9	65
Consommation GPU (%)	5	9	18

EVALUER LES MODÈLES

Coût de l’inférence (prix)

	BERT	LSTM	LETTRIA
Temps d’inférence (**sec**)	1.8	1.9	65
Nombre de GPU (**Tesla V100S**)	1	1	1
Coût total HT (€)	**0.03**	**0.03**	**0.14**

Là ça devient compliqué...



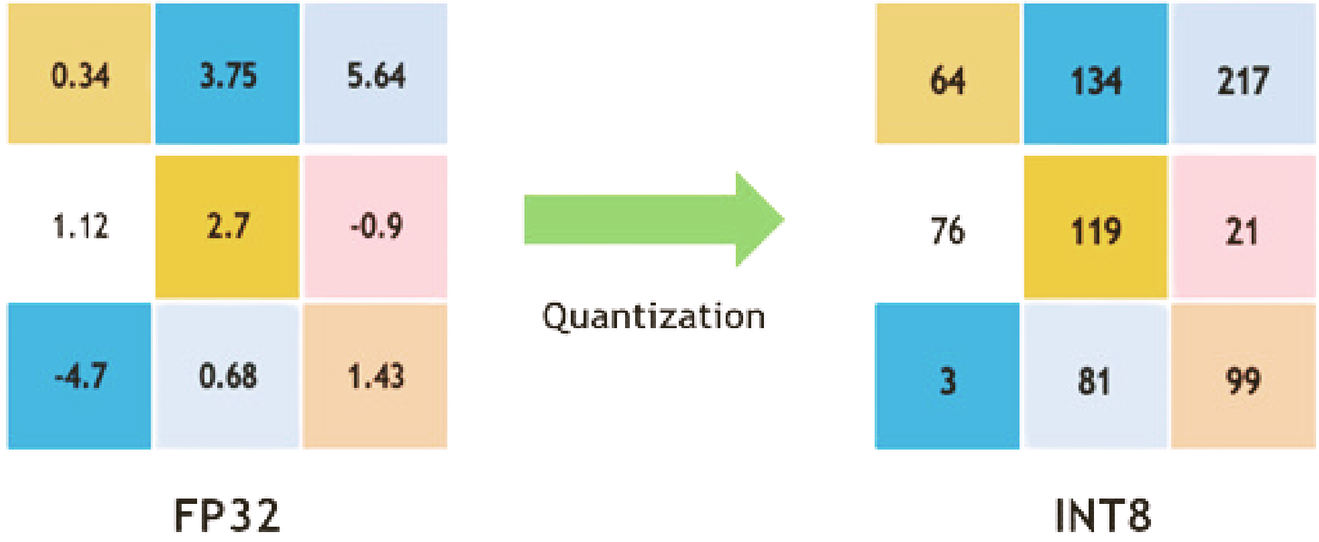
OPTIMISER LE MODÈLE

OPTIMISER LE MODÈLE

→ Optimisation du modèle

Quantization

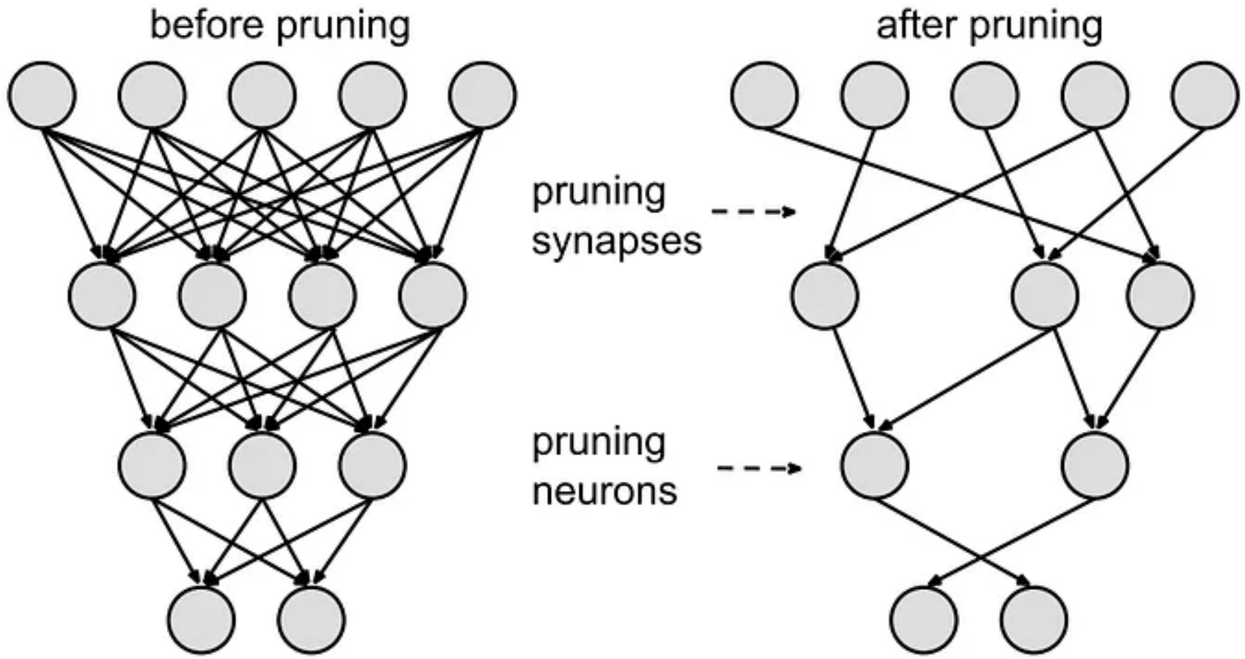
- aware training
- post training



OPTIMISER LE MODÈLE

→ Optimisation du modèle

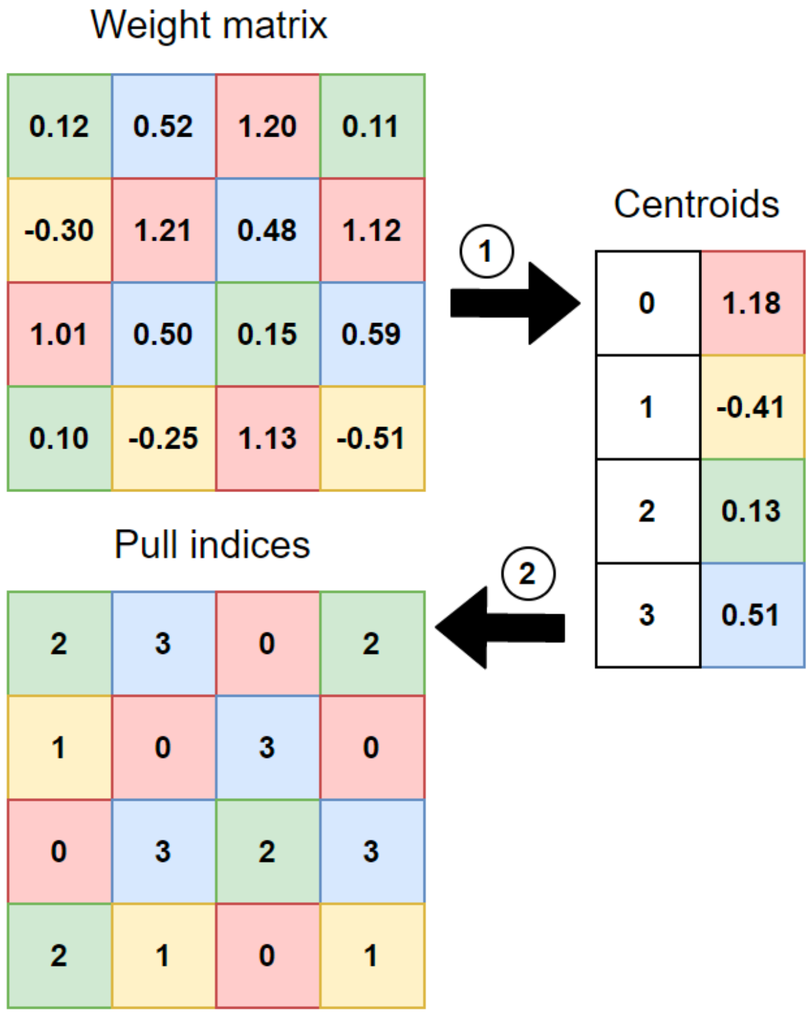
- Quantization
  - aware training
  - post training
- **Pruning**



OPTIMISER LE MODÈLE

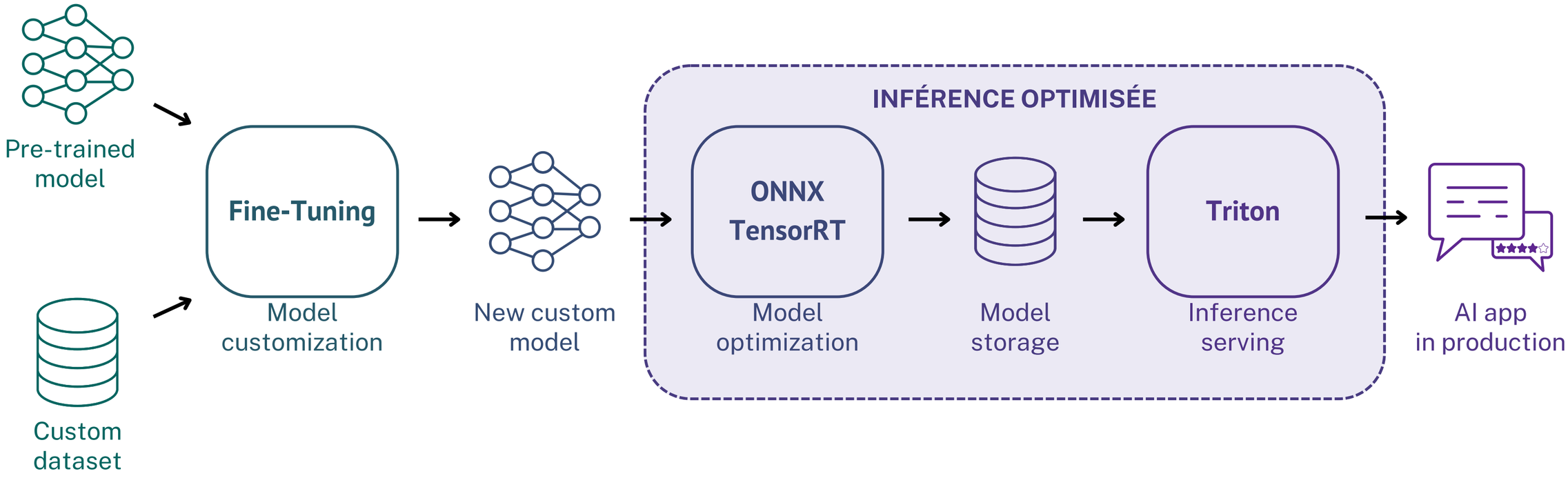
→ Optimisation du modèle

- Quantization
  - aware training
  - post training
- Pruning
- **Clustering des poids**



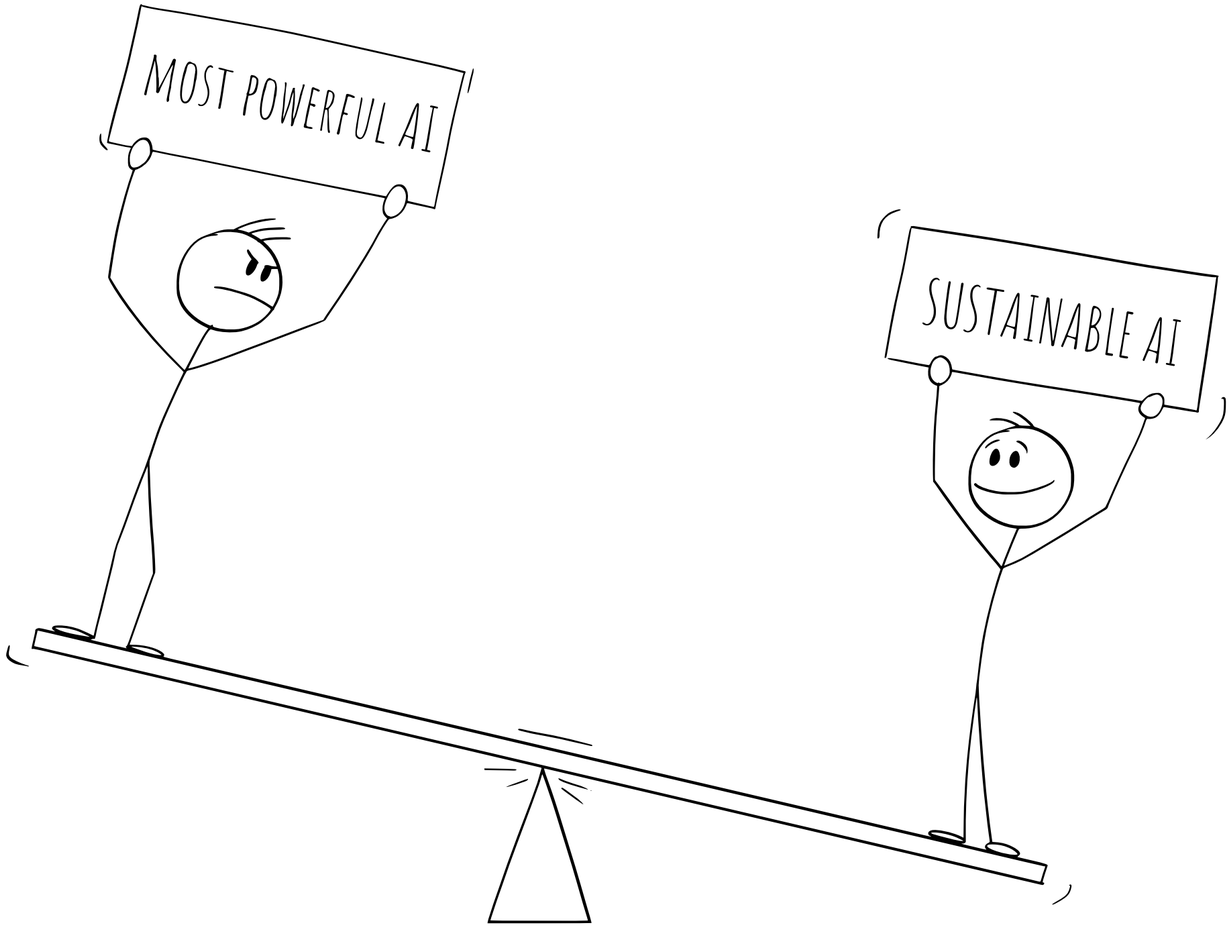
OPTIMISER LE MODÈLE

→ Optimisation de l'inférence



COMPARER LES RÉSULTATS

De quoi avons nous finalement besoin ?



RAPPEL - LE CAS D'USAGE

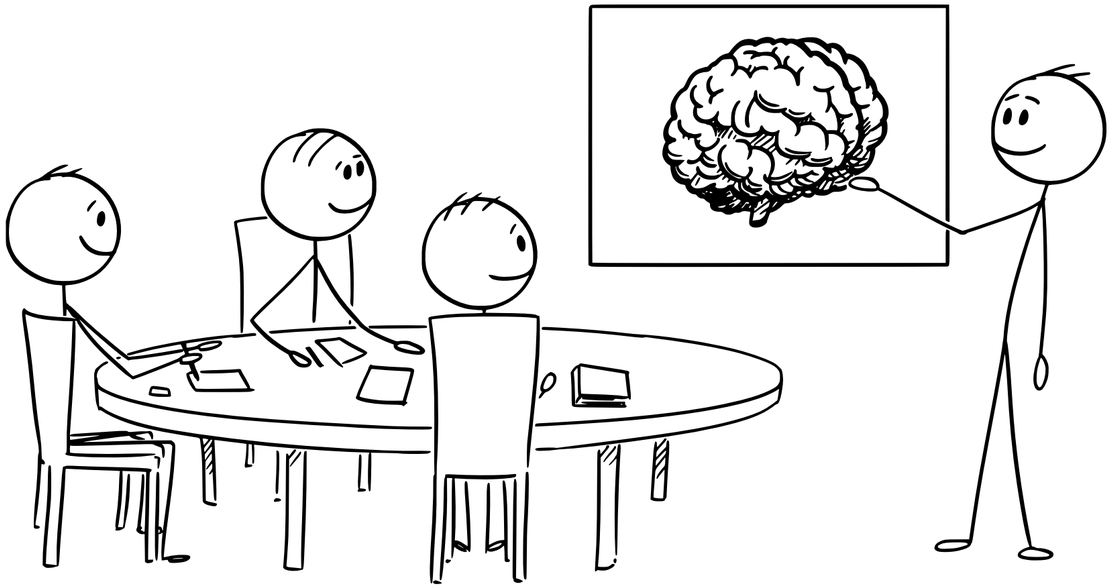
Produit : site de e-commerce de vêtements

Objectif : avoir le sentiment moyen des consommateurs pour pouvoir améliorer les produits et l'expérience client

Solution : déployer un modèle d'IA permettant de classifier les avis clients laissés sur les différents produits

Contraintes : budget restreint, utilisation quotidienne

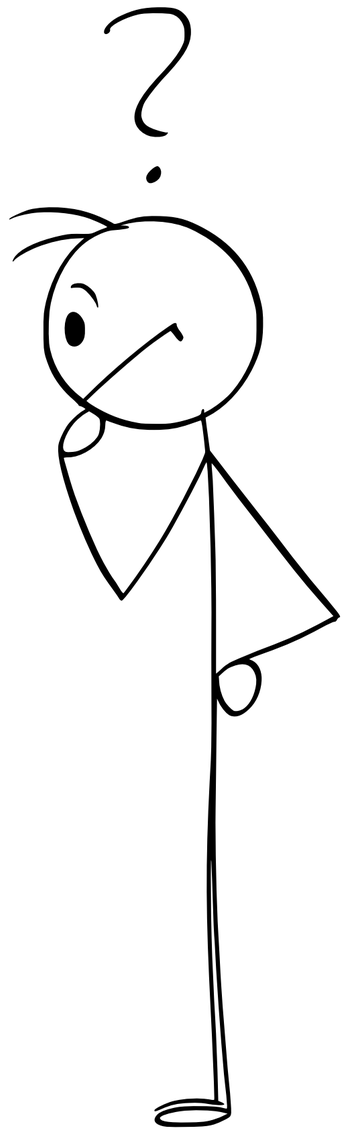
Une IA nous permettrait d'améliorer l'expérience client en se basant sur leurs avis...



COMPARER LES RÉSULTATS

Qu'est-ce qui correspond le plus à mon besoin ?

	BERT	LSTM	LETTRIA
Précision			✅
Latence	✅	✅	
Consommation	✅		
Prix	✅	✅	



En route vers la mise en production !

DÉCIDER DU MODÈLE À DÉPLOYER

DÉCIDER DU MODÈLE À DÉPLOYER

Besoin

- **Objectif** => extraire le sentiment global des clients
- **Occurence** => 1 fois par jour
- **Temps de réponse** => pas besoin de réponse en temps réel
- **Type d'usage** => call API

↦

Réponse

- **Solution** => sentiment analysis app
- **Ressources** => 1 GPU
- **Scaling** => static - 1 réplica
- **Start/stop** => à la demande - démarrage une fois par jour, stop à la fin de l'analyse

CONCLUSION


CUSTOM
dataset
data
PROCESSING
model
BUILDING
model
TRAINING
model
EVALUATION
model
OPTIMIZATION
model
EXPORTATION
model
DEPLOYMENT
optimized model
IN PRODUCTION

DevFest
Perros-Guirec

Avez-vous des **QUESTIONS ?**

RÉFÉRENCES

- **Repo GitHub :** *https://github.com/eleapttn/project-model-optimization-sentiment-analysis.git*
- **OVHcloud AI documentations :** *https://help.ovhcloud.com/csm/worldeuro-documentation-public-cloud-ai-and-machine-learning?id=kb_browse_cat&kb_id=574a8325551974502d4c6e78b7421938&kb_category=1f34d555f49801102d4ca4d466a7fd7d*
- **Women e-commerce clothing reviews dataset :** *https://github.com/ya-stack/Women-s-Ecommerce-Clothing-Reviews*
- **BERT VS. LSTM: Performances in Sentiment Classification** - *https://medium.com/@cd_24/bert-vs-lstm-performances-in-sentiment-classification-b82075184d60*
- **10 steps to build and optimize a ML model** - *https://dev.to/mage_ai/10-steps-to-build-and-optimize-a-ml-model-4a3h*
- **Fine-tuning BERT model for Sentiment Analysis** - *https://www.geeksforgeeks.org/fine-tuning-bert-model-for-sentiment-analysis/*
- **Sentiment Analysis using LSTM** - *https://jagathprasad0.medium.com/sentiment-analysis-using-lstm-b3efee46c956#:~:text=Long%20short%2Dterm%20memory%20is,short%2Dterm%20memory%20of%20data.*
- **Sentiment Analysis with LSTM** - *https://www.analyticsvidhya.com/blog/2022/01/sentiment-analysis-with-lstm/*
- **Model optmization techniques** - *https://medium.com/analytics-vidhya/model-optimization-techniques-79a3a96b6427*


A QUEL POINT DEVONS-NOUS
OPTIMISER NOS MODÈLES D'IA ?