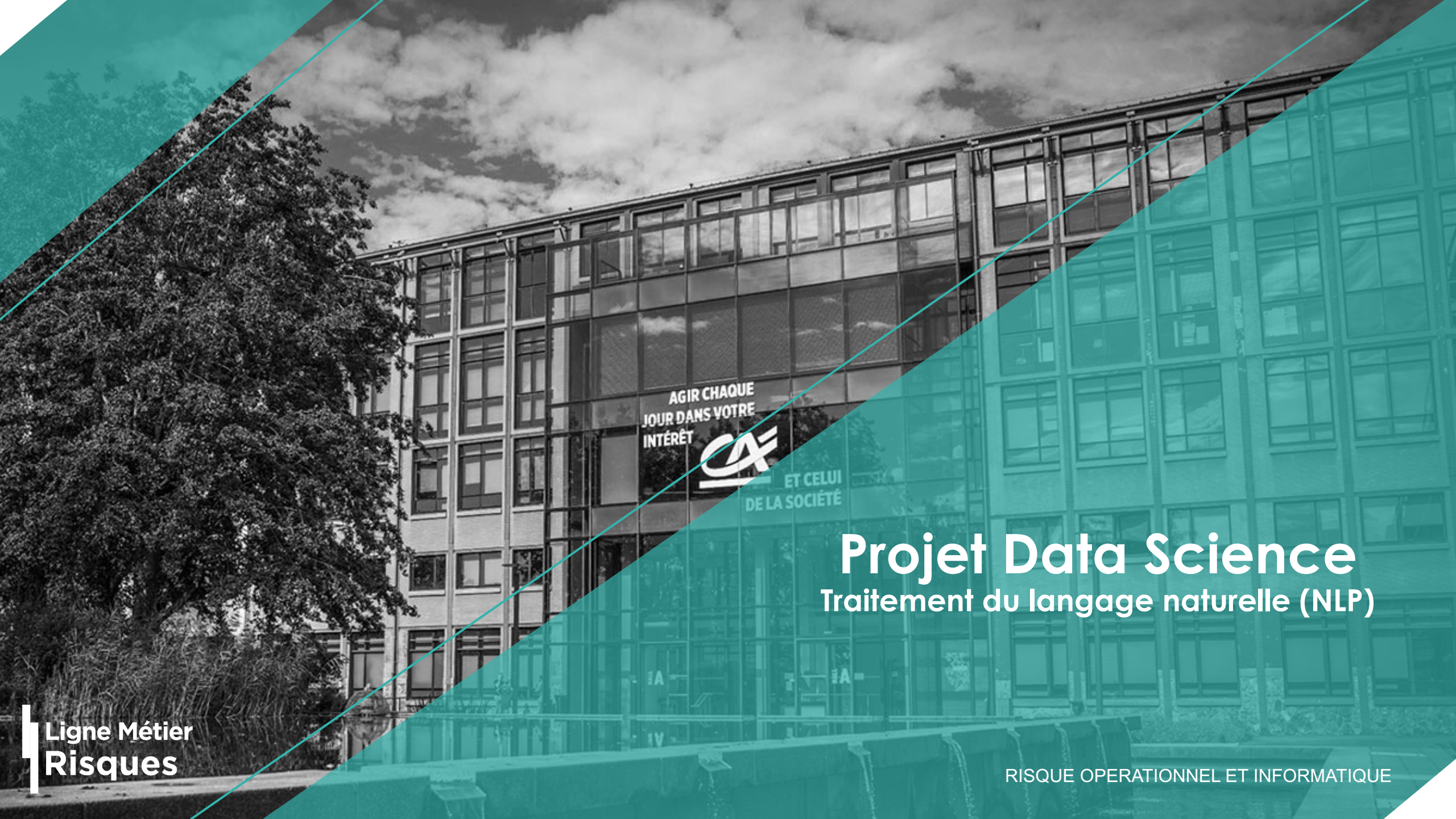


Projet Data Science

Traitement du langage naturelle (NLP)

Ligne Métier
Risques

RISQUE OPERATIONNEL ET INFORMATIQUE

Sommaire

I. Présentation du Groupe Crédit Agricole

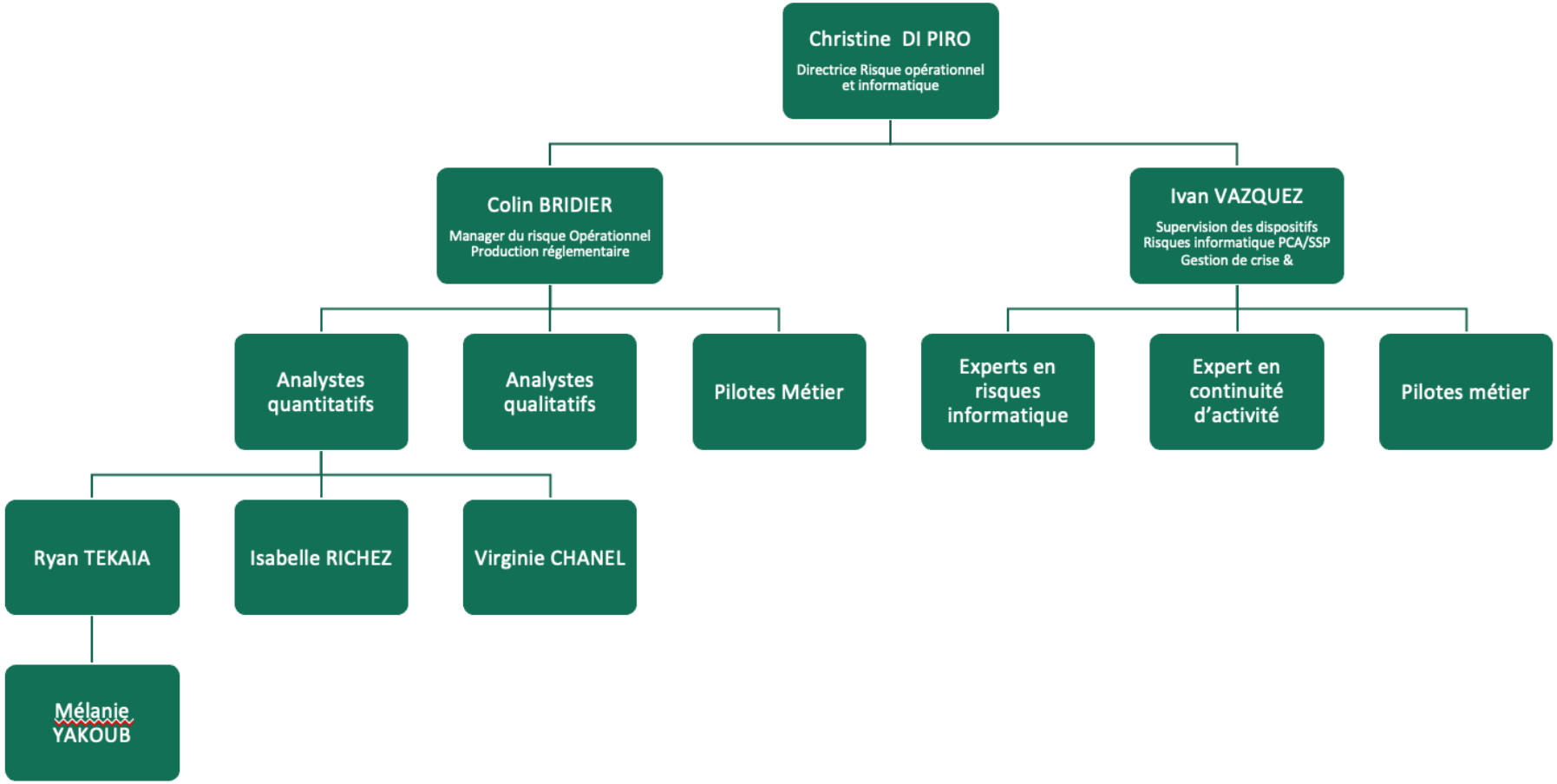
Organe Central

- **145 000** collaborateurs dans le monde
- **11,5 millions** de sociétaires
- **53 millions** de clients

- **10ème** banque mondiale
- **1er** gestionnaire d'actifs en Europe
- **1er** financeur de l'économie en France
- **1er** assureur en France

I. Présentation du Groupe Crédit Agricole

Équipe



II. Présentation des missions

Collecte des données trimestrielles

Incidents opérationnels → Déclaration de la perte financière → Renseignements
- Date
- Montant
- Description courte
- Catégorie de risque (ET3)
- …

Prédiction de la catégorie d'un événement de risque ET3

Objectif

- Prédire la catégorie de risque ET3 d'une description
- Challenger les entités du Groupe sur la qualité de la collecte
- Généraliser les prédictions à d'autres catégories
- Nettoyer les descriptions à d'autres fins tel que la recherche par mots clés

Entrées
- Les nouveaux incidents de chaque trimestre
- Entraînement du modèle sur des incidents classifiés par des experts
- Sélection de la catégorie de risque 1 et émetteur

→

Sortie
- Les incidents avec leur catégorie prédite

Cartographie (Clustering) des événements de risques

Objectif

- Formation de cluster (groupe de mots) en fonction de la description de chaque incident
- Identification de catégorie d'incidents émergents

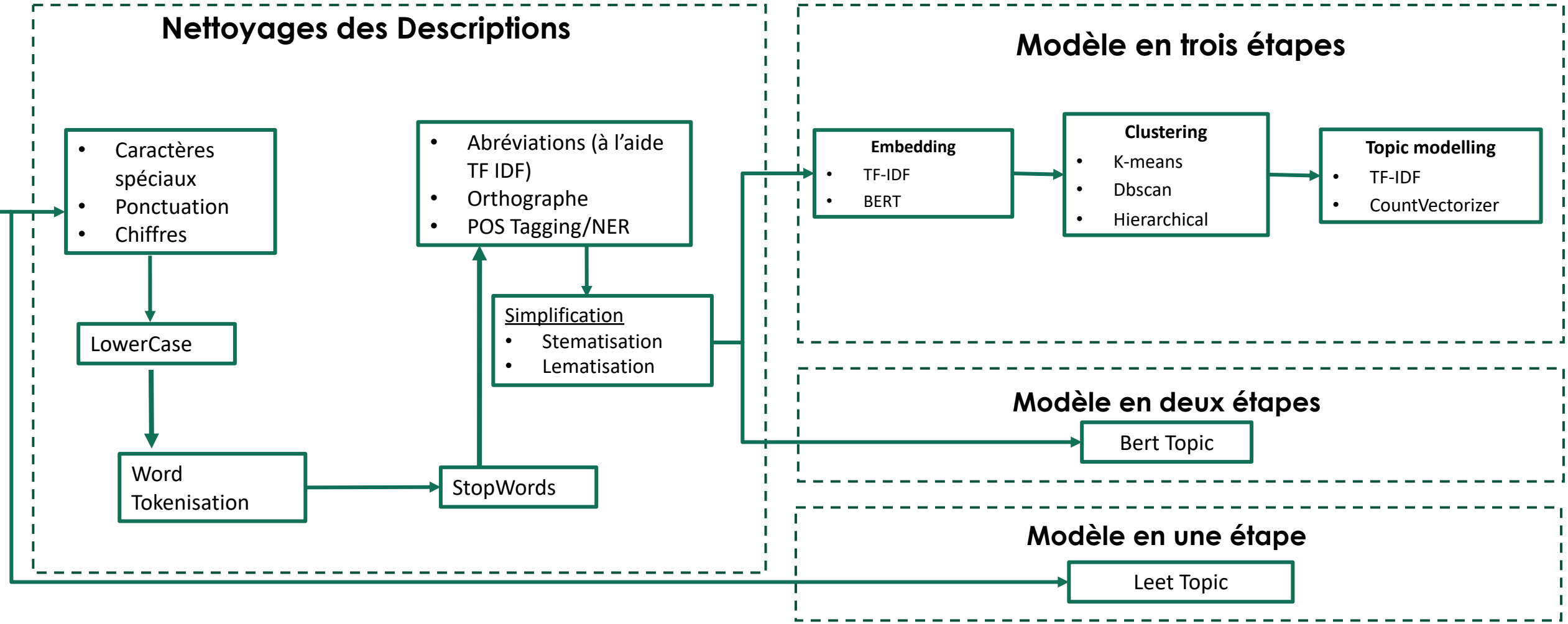
Entrées
- Sélection de la période et de l'émetteur

→

Sortie
- Cluster et nuages de mots
- Analyses et Statistiques sur les proportions des catégories de risques au sein de chaque cluster
- Comparaison avec les années, trimestres, mois

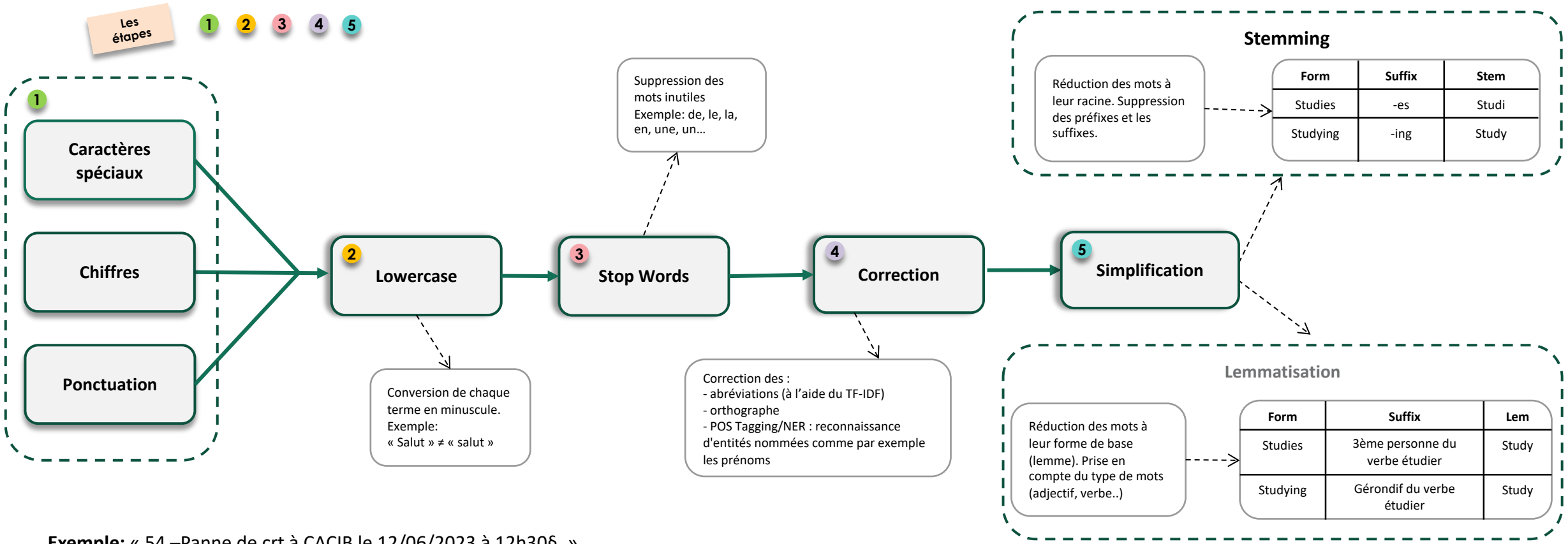
III. Projet 1 : Cartographie des risques

Vision globale



III. Projet 1 : Cartographie des risques

Nettoyages



Exemple: « 54 –Panne de crt à CACIB le 12/06/2023 à 12h30§. »

1. **Panne de crt à CACIB le à**
2. **panne de crt à cacib le à**
3. **panne crt cacib**
4. **Panne courant cacib**
5. **Panne courant cacib**

III. Projet 1 : Cartographie des risques

Vectorisation des descriptions : TF IDF

Corpus: METTRE EXEMPLE DE LA BANQUE AVC PEU DE MOTS DIFFERENTS

- Document 1: This is the first document
- Document 2: This document is the second document
- Document 3: And this is the third one

Formule:

$$TF - IDF = TF(mot, document) \times IDF$$
$$= \text{Nombre d'occurrence du mot dans le document} \times \log\left(\frac{nombre\ de\ document}{idf}\right)$$

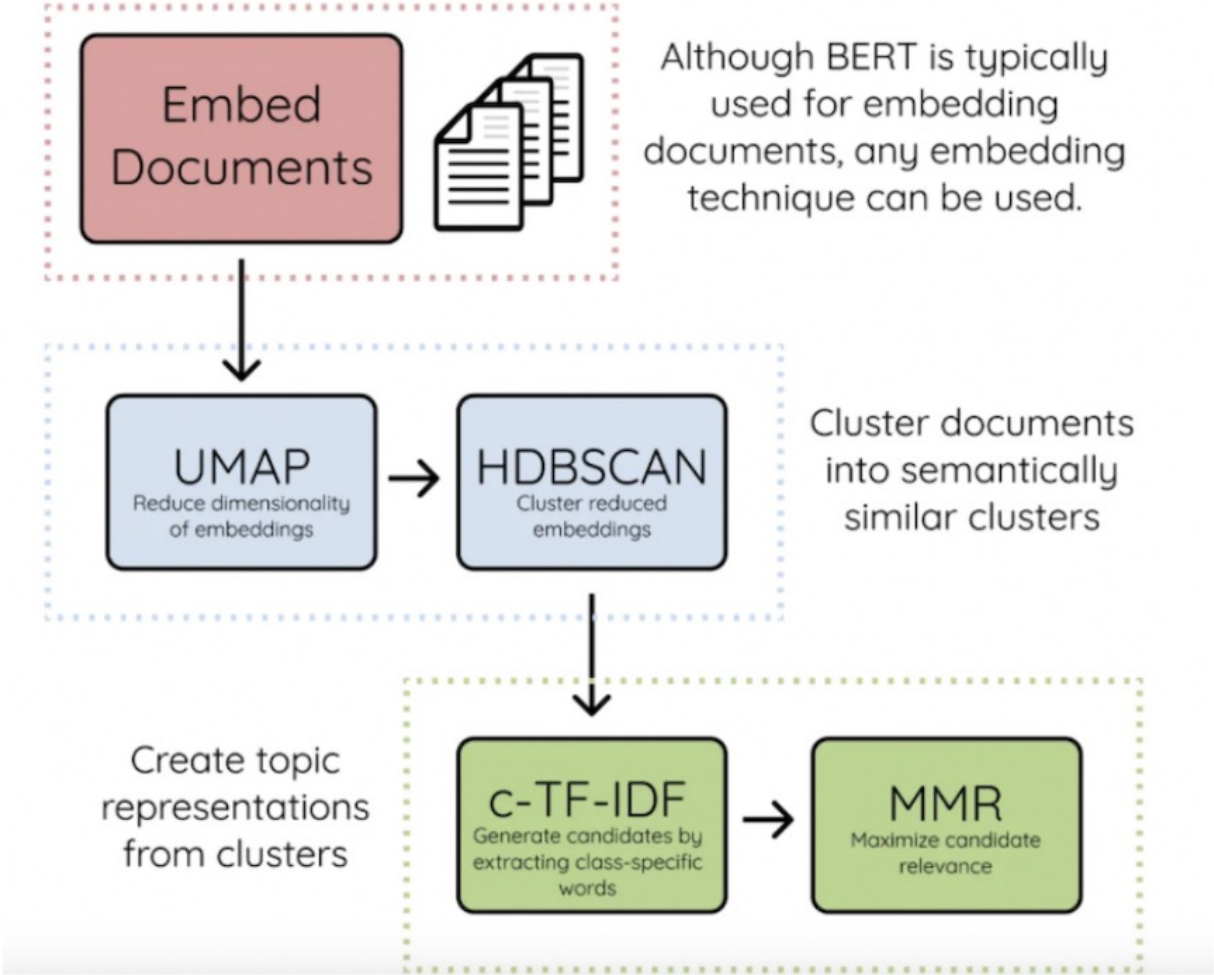
Les mots clés:

- First
- Second
- Third
- And
- One

	This	Is	The	First	Document	Second	And	Third	One
Nombre d'occurrence du mot dans le document phrase 1 (TF)	1	1	1	1	1	0	0	0	0
Nombre d'apparition du mot dans le corpus (IDF)	3	3	3	1	2	1	1	1	1
TF-IDF phrase 1	$\frac{1}{5}\times\log(\frac{3}{3})$	$\frac{1}{5}\times\log(\frac{3}{3})$	$\frac{1}{5}\times\log(\frac{3}{3})$	$\frac{1}{5}\times\log(\frac{3}{1})$	$\frac{1}{5}\times\log(\frac{3}{2})$	$\frac{0}{5}\times\log(\frac{3}{1})$	$\frac{0}{5}\times\log(\frac{3}{1})$	$\frac{0}{5}\times\log(\frac{3}{1})$	$\frac{0}{5}\times\log(\frac{3}{1})$
Résultats	0	0	0	0,1	0	0	0	0	0

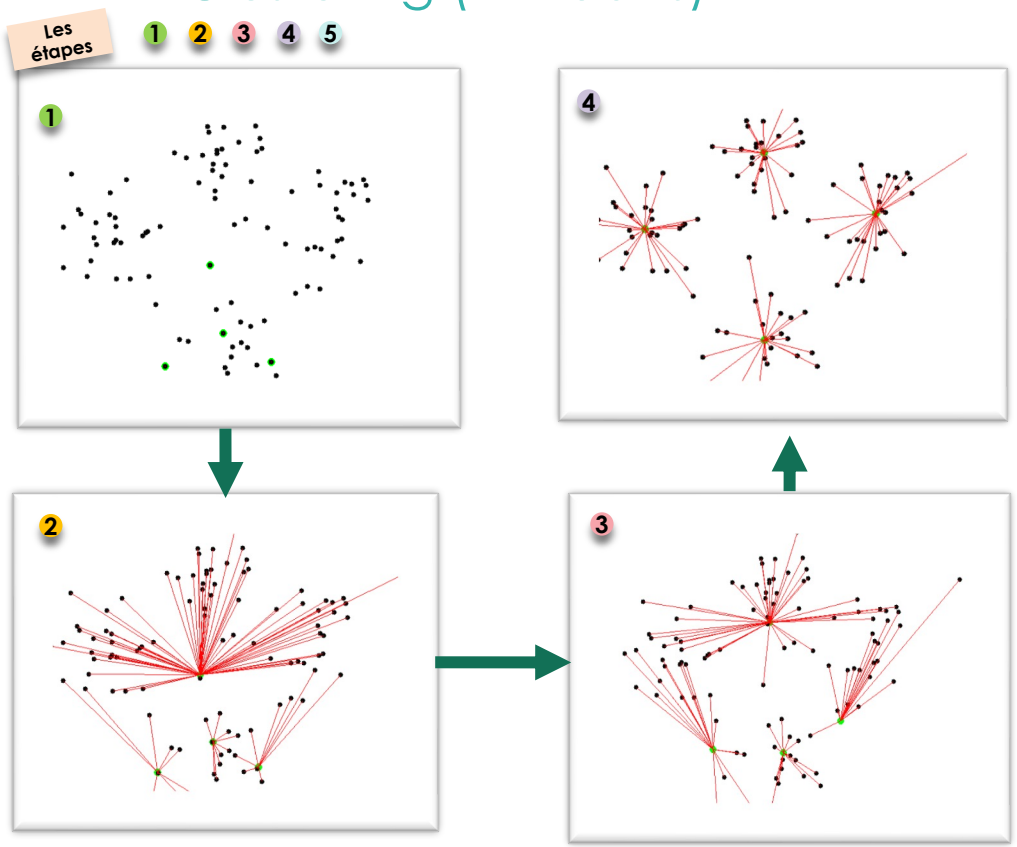
III. Projet 1 : Cartographie des risques

Vectorisation des descriptions : BERTopic, une méthode avancé

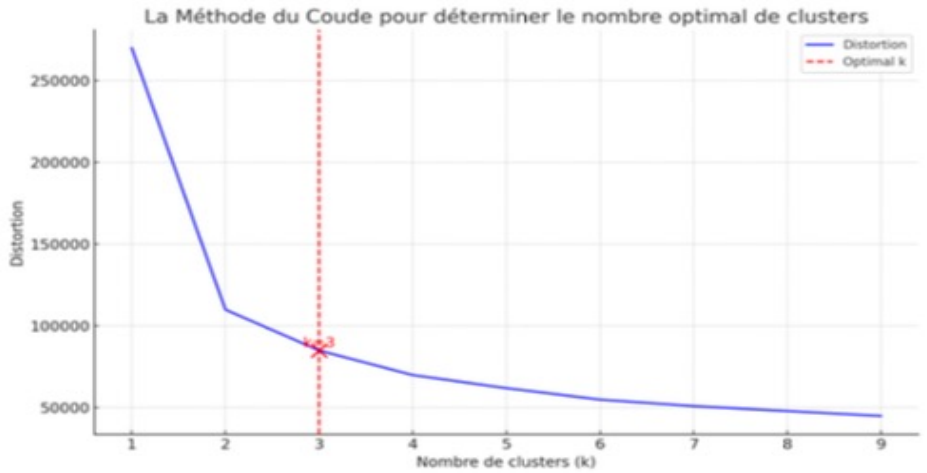


III. Projet 1 : Cartographie des risques

Clustering (k-means)



1. **Initialiser les centroïdes :** Définir le nombre de clusters (ici 4) et placer ces 4 points (centroïde) de façon aléatoire dans l'espace
2. **Assignation des points** : Affecter chaque point à son centroïde le plus proche
3. **Recalcul des centres de cluste**r: Définir les nouvelles positions des centroïdes en utilisant la population associée
4. **Itération :** Répéter l'étape 3 et 4 jusqu'à ce que les centroïdes soient stables



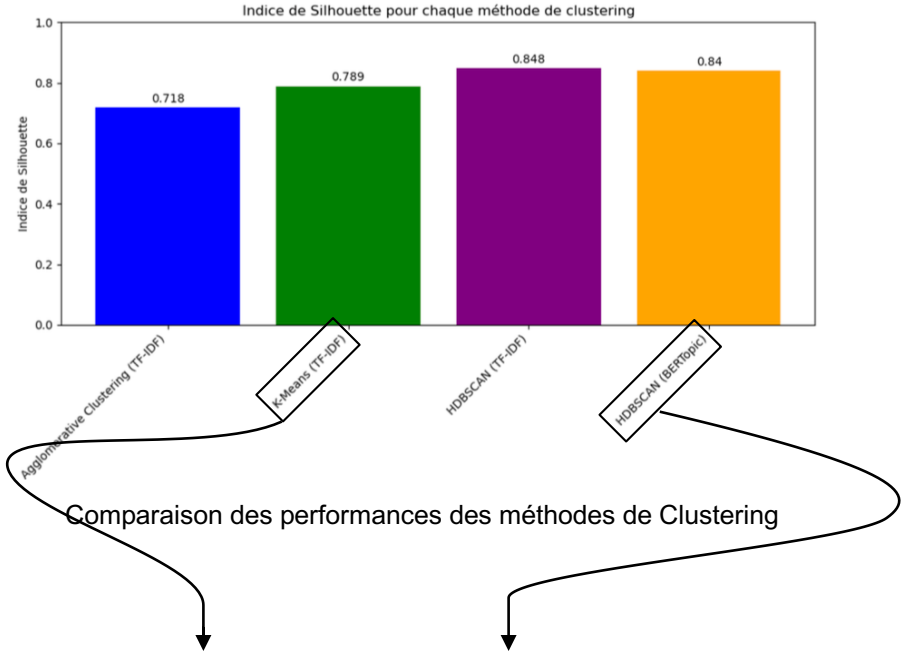
$$\text{Inertie} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

où k est le nombre de clusters, C_i représente le i-ème cluster, x est un point de données dans le cluster C_i, et μ_i est le centroïde du cluster C_i.

- Le nombre optimal de clusters estimé initialement était 3, mais ce résultat a été pris avec prudence en raison des chevauchements et de la complexité des données.
- La variable de catégorie de risque a été prise en compte, avec 7 catégories pour le risque 1 et 175 pour le risque 3, rendant 3 clusters insuffisants.
- Après plusieurs tests et analyses des résultats, nous avons décidé de fixer le nombre de clusters à 6 pour le k-means, pour mieux refléter la structure des données.

III. Projet 1 : Cartographie des risques

Performances des modèles et interprétations des clusters



Sur la base des résultats et des analyses, deux méthodes complémentaires ont été retenues :

- **TF-IDF avec k-means**
- **BERTopic avec HDBSCAN** comme modèle de clustering.

Ces approches permettent une meilleure segmentation et interprétation des données textuelles.

Figure 11: Interpretation des Clusters

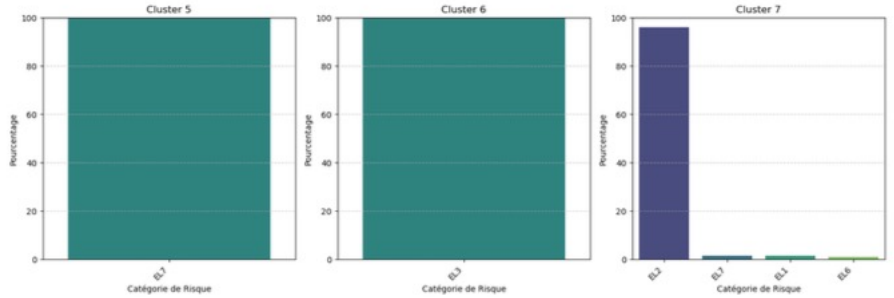
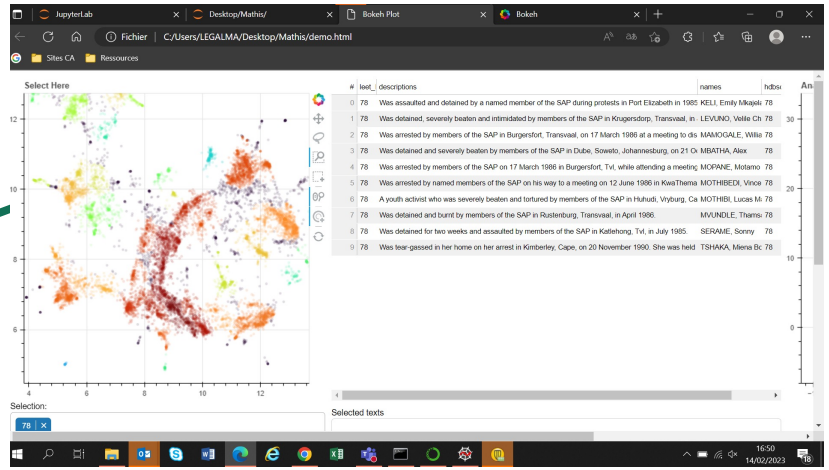
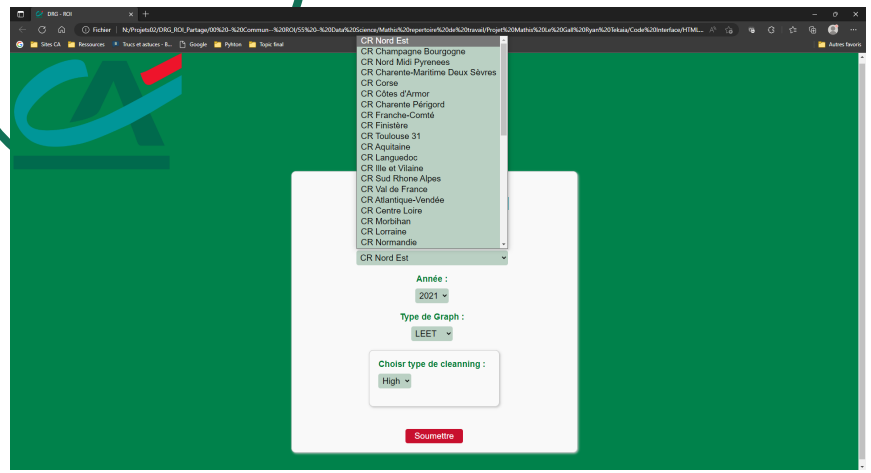


Figure 12: Proportion des catégories de risque bâloises au sein des clusters

III. Projet 1 : Cartographie des risques

Résultats et application :





Les méthodes **TF-IDF avec K-Means** et **BERTopic** ont été intégrées dans l'application, offrant à l'utilisateur la possibilité de choisir le modèle souhaité. L'application affiche d'abord des résultats globaux sur les clusters de pertes par entité. Ensuite, la méthode **Leet-topic** permet de visualiser les clusters en détail, avec la description de chaque observation, les montants concernés, et des statistiques sur les mots-clés associés à chaque cluster.

IV. Projet 2 : Prédiction de la catégorie de risque ET3

Visualisation globale des étapes

Les étapes 1 2 3 4

Nettoyages des Descriptions (1)

- Caractères spéciaux
- Ponctuation
- Chiffres

LowerCase

Word Tokenisation

StopWords

Simplification
- Stemmatisation
- Lemmatisation

- Abréviations (à l'aide TF IDF)
- Orthographe
- PoS Tagging/NER

Variables et Déséquilibres (2)

Sélections
- Catégorie bâloise
- Caisse

Déséquilibres
- Catégorie Autres/suppression classe dominante
- Équilibrages effectifs par classe
- Détection et suppression des doublons

Vectorisation (3)

Text Embedding
- TF-IDF

Sub-word Tokenisation

Word Embedding
- Word2Vect
- LSTM/GRU
- Transformers : Bert (plusieurs versions)

Classification (Entraînement, Test, Inférence) (4)

Algorithmes de Machine Learning

- Bayésien Classifier
- Régression logistique / SVC
- SVM
- Bagging : Random forest
- Boosting : xgboost, Catboost, lithgbm, adaboost
- Stacking
- Finetunning : Réseaux de neurones Bert (avec une étape de data collator)

Optimisation	Ensemble de prédiction	Application web

IV. Projet 2 : Prédiction de la catégorie de risque ET3

Nettoyages

Chaque niveau de nettoyage inclut les opérations des niveaux précédents, tout en ajoutant des étapes plus avancées pour un traitement plus rigoureux :

No :

- Conversion en minuscules, correction d'orthographe et gestion des abréviations.
- Aucun autre nettoyage spécifique : les caractères spéciaux et les mots courants (stop words) ne sont pas supprimés.

Light :

- Suppression des caractères non alphabétiques et des mots courants (stop words).
- Maintien de l'essentiel du contenu.

Medium :

- Lemmatisation pour ramener les mots à leur forme de base.
- Suppression des mots très courts (moins de 4 lettres).
- Simplification du texte tout en préservant les informations clés.

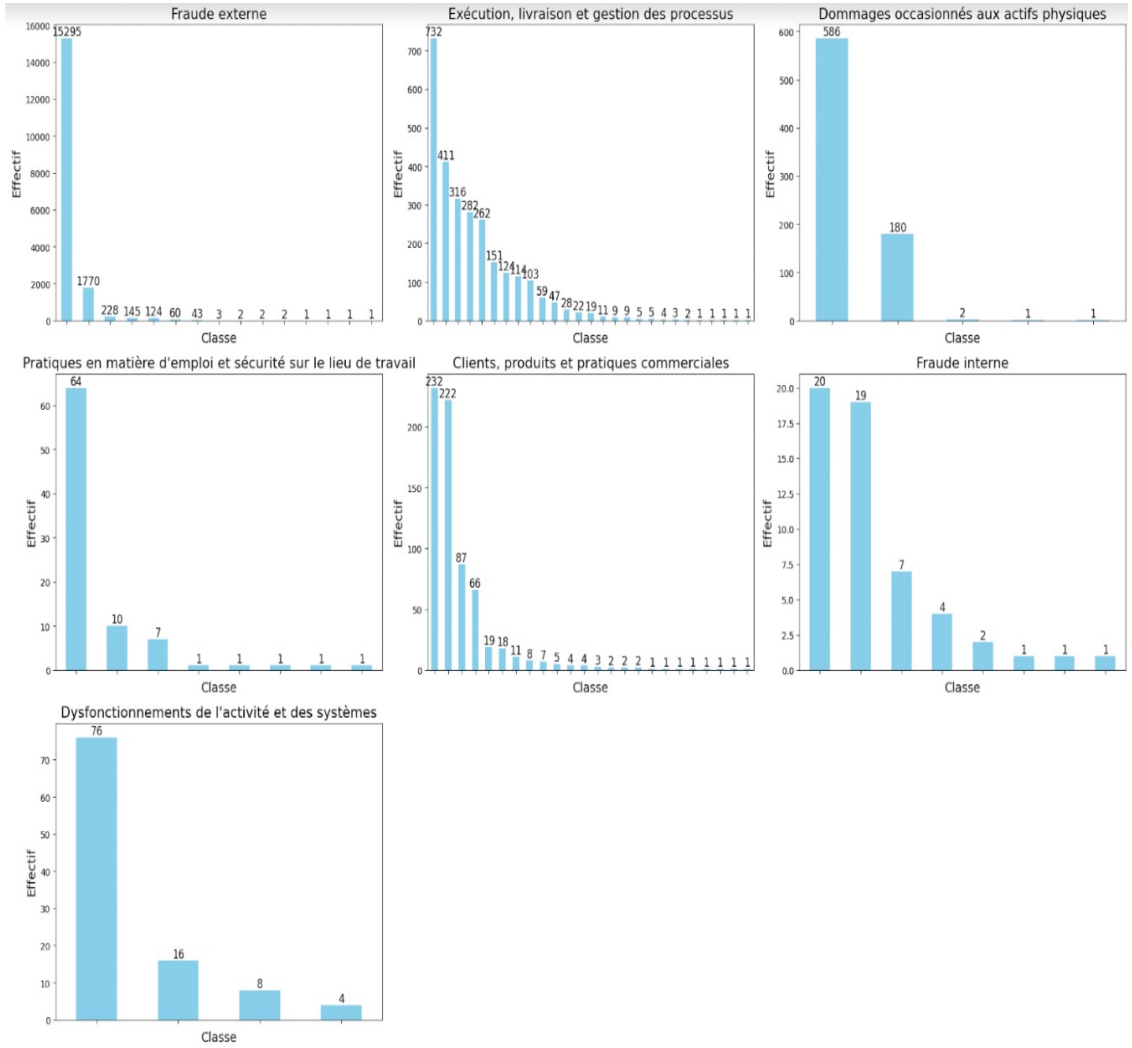
Advanced :

- Suppression complète des abréviations et des accents.
- Le texte est entièrement débarrassé des stop words et autres éléments non essentiels.
- Utilisé pour une normalisation maximale du texte.

IV. Projet 2 : Prédiction de la catégorie de risque ET3

Variables et déséquilibres

- Le projet couvre 175 classes ET3 et 2 297 731 observations, avec un modèle général par caisse et un focus spécifique sur la caisse d'Île-de-France (22 116 observations).
- Cette image illustre visuellement le déséquilibre entre les différentes classes, en particulier la surreprésentation de la catégorie "fraude externe".
- Un zoom particulier est fait sur la catégorie "Exécution, livraison et gestion des processus". Pour pallier le déséquilibre des classes, notamment celles avec un effectif inférieur à 5, une catégorie "Autres" a été créée, réduisant 91 classes à 36.
- Biais potentiel vers la classe majoritaire et des mauvaises performances globales.



Effectif des catégories ET3 au sein de chaque catégorie de risque bâloise

IV. Projet 2 : Prédiction de la catégorie de risque ET3

Variables et déséquilibres

Approches principales pour corriger le déséquilibre des classes :

- **Méthodes Algorithm-Level** : Ajustement des modèles en introduisant des pondérations spécifiques pour chaque classe à travers des métriques telles que la Balanced Accuracy et le F1 Score, permettant au modèle de traiter les classes déséquilibrées de manière plus équitable et robuste.
- **Rééchantillonnage (Data-Level)** : Utilisation de techniques comme le sous-échantillonnage aléatoire pour réduire la classe majoritaire et SMOTE pour générer de nouveaux échantillons dans les classes minoritaires.

SMOTE génère des échantillons synthétiques pour les classes minoritaires en interpolant entre des exemples existants. Cela introduit de la variabilité, réduit le risque de sur-apprentissage et améliore l'équilibre global des classes.

Résultats observés et impact des méthodes :

- Grâce à l'application de ces techniques, notamment SMOTE, la précision d'une classe est passée de 0 % à 40 %.
- Certaines limitations, liées aux abréviations dans les données textuelles, ont parfois réduit l'efficacité de SMOTE, mais l'amélioration globale des performances reste notable.

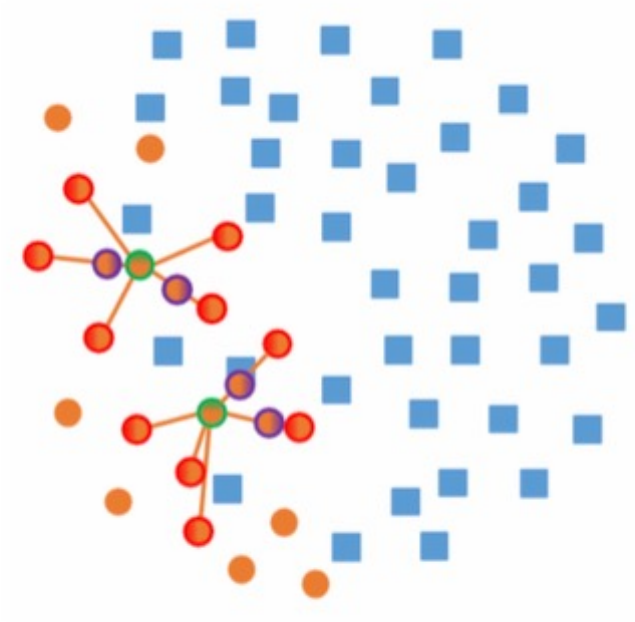
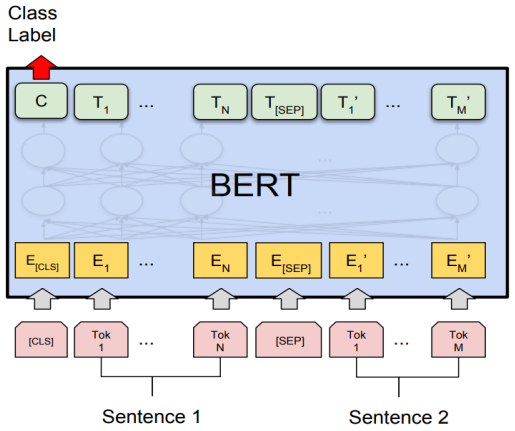


Figure 17: SMOTE

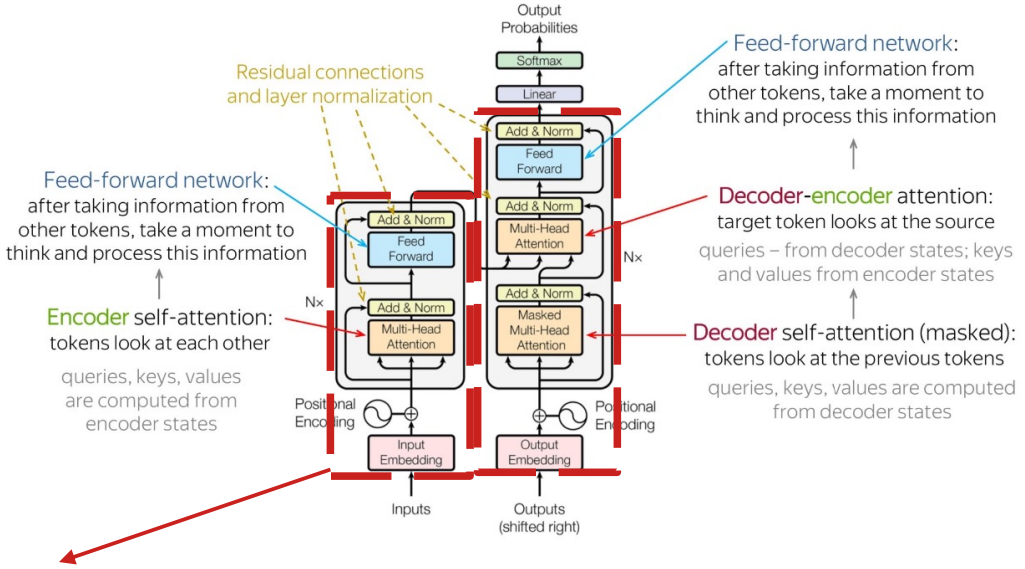
IV. Projet 2 : Prédiction de la catégorie de risque ET3

Modèles de vectorisation (BERT)



Autres méthodes testées
TF IDF / Word2Vec

Encodeur
BERT, CamemBERT (large corpus français),
DistilBERT (version allégée de CamemBERT)



Encodeur

- 1 bloc d'attention (self-attention)
- 1 bloc feed-forward
- Connexions résiduelles

Utilisation du token spécial CLS dans BERT
Le CLS token est ajouté au début de chaque séquence pour capturer une représentation globale du texte, utile pour les tâches de classification :

- BERT utilise le token [CLS] avec des modèles ML pour la tâche de classification
- BERT est fine-tuné, ajustant le token [CLS] pour une classification plus précise et adaptée à la tâche spécifique.

Transformer : Type de réseau de neurones qui traite les mots en parallèle pour mieux comprendre le contexte.

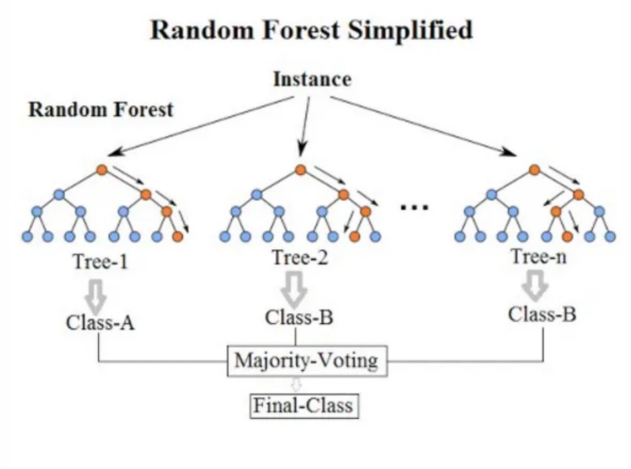
Entraînement Bidirectionnel : Comprend le contexte complet d'une phrase en tenant compte des mots avant et après.

Puissance de Calcul : Exige une capacité élevée pour l'entraînement et l'inférence.

IV. Projet 2 : Prédiction de la catégorie de risque ET3

Modèles de classification avec optimisation bayésienne

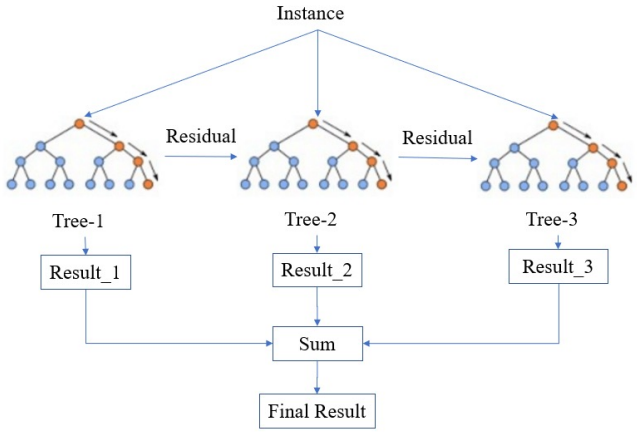
- Random Forest



Avantages : Robuste aux données bruitées et aux classes déséquilibrées.

Inconvénients : Moins performant sur des données textuelles complexes par rapport à XGBoost ou réseaux de neurones.

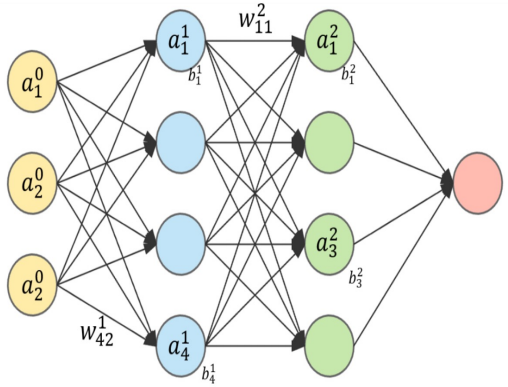
- XGBoost



Avantages : Très performant, notamment sur des données de grande dimension et déséquilibrées.

Inconvénients : Plus complexe à ajuster et à interpréter.

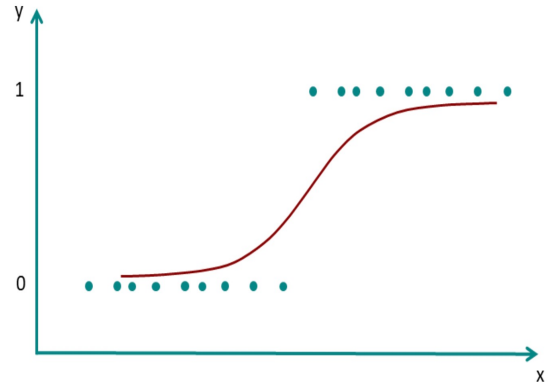
- Réseaux de neurones



Avantages : Capture des relations complexes dans les données textuelles.

Inconvénients : Nécessite beaucoup de données et de ressources, avec un temps de calcul plus long.

- Régression logistique

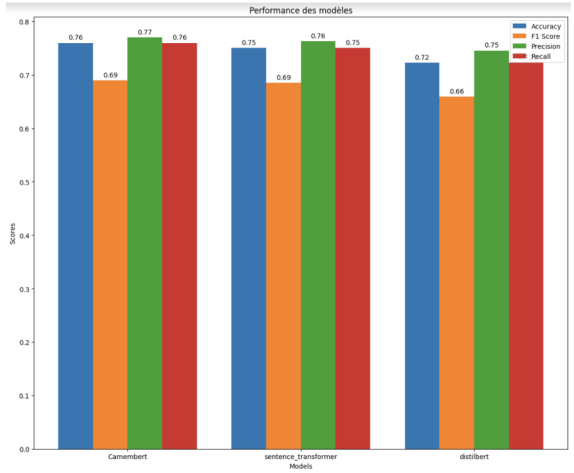


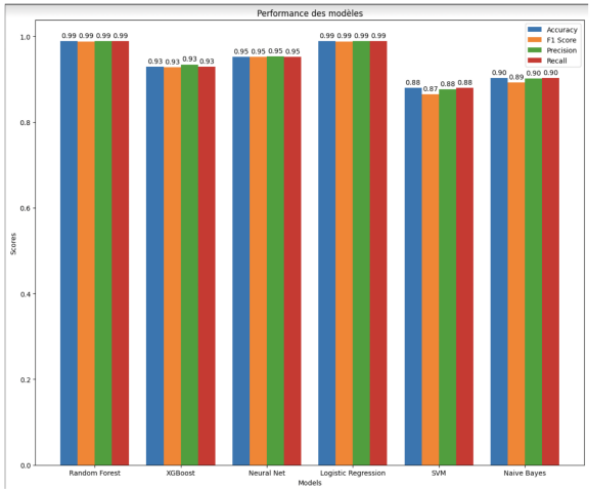
Avantages : Simple à mettre en œuvre et rapide à entraîner.

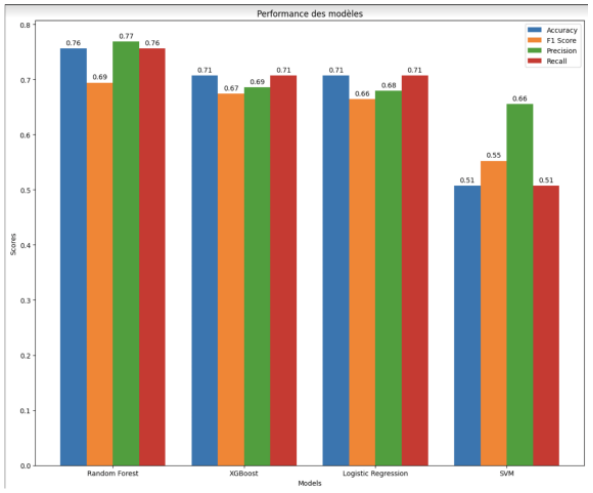
Inconvénients : Limité pour capturer des relations complexes entre les données.

IV. Projet 2 : Prédiction de la catégorie de risque ET3

Choix du modèle







Bert | Camembert | DistilBert | TF IDF

Choix du modèle ?

finetuning

Camembert | TF IDF

+

RF | XGBoost | NN | SVM

Meilleures performances ?

Bert finetuné | Bert RF | TF IDF RF

Projet 1 : Prédiction de la catégorie de risque ET3

Résultats et Application web

Benchmark :

Vectorisation	Modèle ML	Nettoyage	Sous selection de variables	Caisse et période	Paramètres optimisés	Accuracy
TF IDF	Random Forest	Avancé	Non	Ile de France	Oui	92%
Bert	Random Forest	Légé	Non	Ile de France/ 2018 à 2022	Oui	88%
Bert	Finetunning	Légé	Non	Ile de France	Oui	87%
Word2vect	Random Forest	Avancé	Non	Ile de France	Oui	87%
TF IDF	Random Forest	Avancé	Oui	Ile de France	Oui	99%
Bert	Random Forest	Légé	Oui	Ile de France	Oui	76%
Bert	Finetunning	Légé	Oui	Ile de France	Oui	72%

Interface application :

V. Conclusion

- Maîtrise de Python et développement d'une expertise en NLP, avec utilisation de Transformers et fine-tuning via TensorFlow et PyTorch.
- Utilisation de MLflow pour le suivi des modèles et dockerisation du projet pour faciliter son déploiement.
- Développement de compétences en Business Intelligence à travers la création de visualisations claires avec Power BI.
- Renforcement de la confiance en soi grâce aux présentations régulières, malgré un manque de collaboration approfondie avec l'équipe.

Des questions ?