UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

GUIFT DIMITRI **2024-2025**
12214381

MASTER 2 MoSEF DATA SCIENCE

RAPPORT DE FIN D'ÉTUDES

CREDIT AGRICOLE ASSURANCES

ASSURANCES

Interprétabilité des clauses contractuelles en assurances dommages et modélisation du sinistre tempête

Tuteur d'alternance : M. Mustapha BENARBIA

Unité d'accueil - lieu d'alternance
Ville : Paris Pays : France

Tuteur pédagogique : M. Marc-Arthur DIAYE

Remerciements

Avant d'entamer la présentation de ce rapport, je souhaite exprimer toute ma gratitude envers celles et ceux qui ont contribué à la réussite de cette année d'alternance. Leur soutien, leurs conseils et leur disponibilité ont largement enrichi cette expérience et m'ont permis de progresser tant sur le plan professionnel que personnel.

Je tiens à remercier Monsieur Mustapha BENARBIA, mon tuteur d'entreprise, ainsi que Monsieur Marc-Arthur DIAYE, mon tuteur académique, pour leur accompagnement, leur écoute et leur pédagogie tout au long de cette année. Leur expertise et leur patience ont été pour moi un véritable appui et une source constante d'apprentissage.

J'adresse également mes remerciements à l'ensemble de mes collègues de la Direction de l'Audit des Assurances de Crédit Agricole Assurances. Leur disponibilité, leur esprit d'équipe et leur bienveillance ont favorisé mon intégration et contribué à rendre cette expérience particulièrement formatrice.

Je souhaite aussi remercier mes enseignants et encadrants du Master MoSEF pour la qualité de leur enseignement et leur investissement, qui m'ont permis d'acquérir des connaissances solides et utiles à mon avenir professionnel. J'exprime une reconnaissance particulière à Madame Rania KAFFEL, directrice du Master 2 MoSEF, pour m'avoir accueillie dans ce programme et accompagnée avec bienveillance dans ce parcours exigeant.

Je tiens également à exprimer ma gratitude à mes camarades de classe du Master, avec qui j'ai partagé des moments d'échanges, de collaboration et de solidarité. Leur présence et leur soutien ont rendu cette formation plus enrichissante et stimulante.

Enfin, je remercie ma famille et mes proches pour leur soutien indéfectible et leurs encouragements constants, qui m'ont accompagnée dans chacune des étapes de mon parcours.

Mes sincères remerciements vont à toutes ces personnes qui ont rendu cette expérience d'alternance à la fois unique, enrichissante et inoubliable.

Table des matières

Introduction

Depuis le 30 septembre 2024, j'ai l'opportunité de réaliser mon alternance au sein de Crédit Agricole Assurances, situé à Paris Montparnasse, au sein de la Direction de l'Audit des Assurances (DAA). Sous la responsabilité de Monsieur Mustapha BENARBIA, j'occupe la fonction de data scientist au sein du pôle Data, une équipe dédiée à l'exploitation et à la valorisation des données pour appuyer les missions d'audit.

La DAA représente la troisième ligne de défense du groupe Crédit Agricole Assurances. Elle conduit des missions d'audit visant à évaluer le degré de maîtrise des risques techniques, financiers, opérationnels, informatiques ou encore de conformité. Dans ce contexte, le pôle Data joue un rôle transversal en développant des analyses massives et des outils automatisés, permettant d'améliorer la fiabilité et l'efficacité des travaux d'audit.

Ce rapport rend compte de cette expérience. Il débute par une présentation du groupe Crédit Agricole et de son environnement, puis décrit les missions qui m'ont été confiées. Parmi celles-ci figurent une étude sur l'interprétabilité des clauses contractuelles en assurance dommages à l'aide de techniques de traitement automatique du langage, ainsi qu'une analyse du risque climatique lié au vent appliqué à l'assurance agricole, construite à partir de données météorologiques ERA5 couplées au portefeuille Multirisque Agricole. Enfin, une réflexion critique est proposée sur les compétences acquises et les enseignements tirés de cette alternance au regard de mon projet professionnel.

En définitive, cette introduction vise à situer le cadre de mon alternance, à en préciser les enjeux et à présenter la logique du rapport qui suit. Elle témoigne également de l'importance d'intégrer la data science dans des métiers traditionnellement ancrés dans l'audit et le contrôle, et souligne la valeur ajoutée que représente cette démarche dans un contexte de transformation digitale et de complexité croissante des risques.

Chapitre 1

Présentation de l'environnement professionnel et du contexte

1.1 Secteurs d'activité du groupe Crédit Agricole

Le Groupe Crédit Agricole (CA) est l'un des principaux groupes bancaires européens et le premier assureur en France. Il repose sur un modèle coopératif et mutualiste, structuré autour des caisses locales et régionales, et coordonné par sa holding Crédit Agricole S.A. Présent en France et à l'international, le Groupe propose une gamme complète de services comprenant la banque de détail, l'assurance (via Crédit Agricole Assurances), la gestion d'actifs, les services financiers spécialisés et la banque de financement et d'investissement. Fortement ancré dans les territoires, il s'appuie sur des valeurs de proximité, de responsabilité et de solidarité pour accompagner les transitions économiques, écologiques et sociétales.

1.2 Objectifs, projets et engagements du Groupe Crédit Agricole

Le Crédit Agricole a défini sa raison d'être : *« Agir chaque jour dans l'intérêt de nos clients et de la société »*. Cette ambition repose sur trois piliers stratégiques : l'excellence relationnelle, la responsabilité en proximité et l'engagement sociétal.

La stratégie du Groupe s'articule autour de trois projets majeurs :

— **Projet client** : devenir la banque de référence en matière de satisfaction auprès des particuliers, des entrepreneurs et des institutions.

— **Projet humain** : soutenir les collaborateurs dans la transformation digitale, tout en favorisant la proximité humaine et le développement de compétences adaptées aux

enjeux actuels.

— **Projet sociétal** : agir comme un acteur clé de la transition écologique, solidaire et responsable, en contribuant au développement durable des territoires et de la société.

1.3 Chiffres clés du Groupe Crédit Agricole en 2024

En 2024, le Groupe Crédit Agricole s'impose comme un acteur majeur de la finance mondiale. Présent dans **46 pays**, il compte **54 millions de clients** et s'appuie sur un réseau de **8 200 agences**, dont **6 660 en France**. Avec **12,1 millions de sociétaires**, il confirme son modèle coopératif et mutualiste. Il occupe la première place en tant que **financeur de l'économie française**, **gestionnaire d'actifs européen** et **assureur en France**. Il est également la **première banque de proximité de l'Union européenne** et la **première banque coopérative au monde**. Enfin, il se classe au **9e rang des banques mondiales**.

FIGURE 1.1 – Chiffres clés 2024 du Groupe Crédit Agricole – synthèse graphique.

1.4 Filiale Crédit Agricole Assurances (CAA)

Crédit Agricole Assurances est la filiale d'assurance du Groupe Crédit Agricole et le premier bancassureur en France. Il regroupe des entités comme Predica (épargne et retraite), Pacifica (assurance dommages) et CACI (prévoyance et assurance emprunteur). Ses produits sont distribués principalement par les réseaux Crédit Agricole et LCL. En 2024, Crédit Agricole Assurances a confirmé sa solidité financière et son rôle de leader, tout en plaçant la proximité, la digitalisation et la responsabilité sociétale au cœur de sa stratégie.

1.4.1 Objectifs et activités

Les objectifs de Crédit Agricole Assurances s'inscrivent dans la stratégie globale du Groupe Crédit Agricole. Ils visent à renforcer la proximité avec les clients, à poursuivre la digitalisation des services et à accompagner les grandes transitions économiques et sociétales. L'entreprise met également l'accent sur la responsabilité sociale et environnementale, en soutenant le financement d'initiatives durables. Les activités de CAA se structurent autour de trois grands pôles complémentaires :

— **Épargne et retraite**, principalement portée par Predica et Spirica ;
— **Assurance dommages**, développée par Pacifica, couvrant l'habitation, l'automobile et la responsabilité civile ;
— **Prévoyance et assurance emprunteur**, assurées par CACI et d'autres entités spécialisées.

La stratégie de CAA repose également sur cinq piliers définis dans le plan Horizon 2025 :

— Accélérer sur la **protection des biens et des personnes** ;
— Proposer de nouvelles solutions d'**épargne responsables et accessibles** ;
— Renforcer le modèle de **bancassurance universel pour les entreprises** ;
— Développer les activités **à l'international** ;
— Devenir l'**assureur digital de référence**.

1.4.2 Chiffres clés de Crédit Agricole Assurances (CAA)

En 2024, Crédit Agricole Assurances a réalisé un chiffre d'affaires de **43,6 milliards d'euros**, confirmant sa place de **premier bancassureur en France** et d'acteur majeur en Europe. L'épargne et la retraite représentent la plus grande part de l'activité avec **32,1 milliards d'euros**, plaçant CAA parmi les leaders européens du secteur. Les encours d'assurance vie atteignent **347,3 milliards d'euros**, dont **104,1 milliards en unités de compte**, renforçant son rôle de **premier assureur vie en France**. L'assurance dommages poursuit sa croissance avec **6,2 milliards d'euros** et plus de **16,7 millions de contrats**, ce qui en fait un acteur de référence en France. Enfin, la protection des personnes génère **5,3 milliards d'euros**, tandis que le résultat net s'établit à **1,96 milliard d'euros**, confortant la solidité financière du groupe, classé parmi les **premiers assureurs européens**.

1.4.3 Organigramme de Crédit Agricole Assurances

L'organigramme ci-dessous présente la structure de Crédit Agricole Assurances (CAA) et met en évidence ses principales filiales et partenariats.

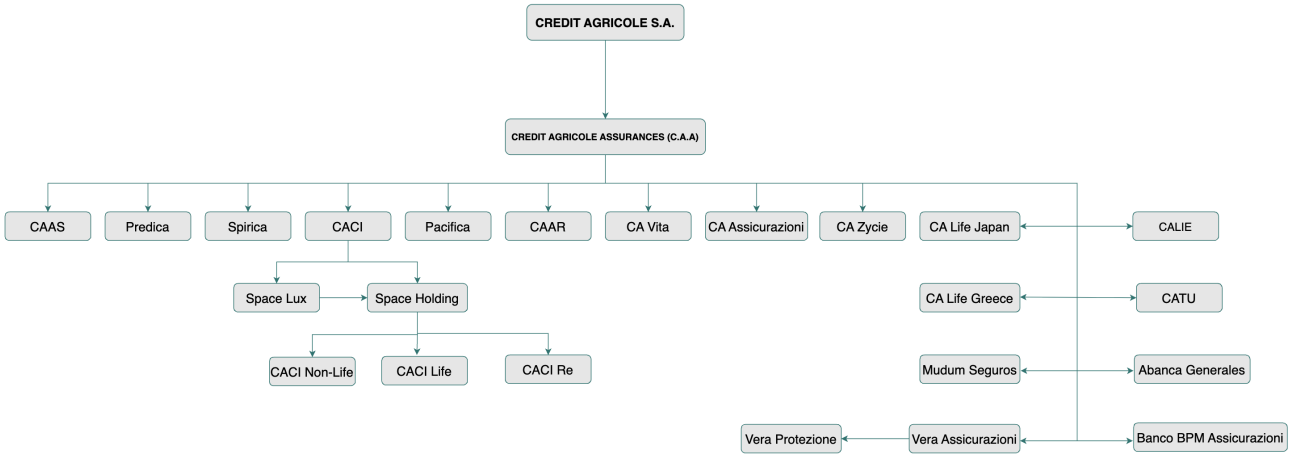

FIGURE 1.2 – Structure de Crédit Agricole Assurances (CAA)

1.4.4 Direction de l'Audit des Assurances (DAA)

Mon alternance s'est déroulée au sein de la **Direction de l'Audit de Crédit Agricole Assurances (DAA)**. Cette direction assure la « fonction d'audit interne » du Groupe CAA, conformément aux exigences de la directive *Solvabilité II*, ainsi que le rôle de « contrôle périodique ». La SU Audit interne des Assurances conduit des investigations à la fois sur site et sur documents, afin d'évaluer le degré de maîtrise des risques liés à l'ensemble des activités du groupe : risques techniques, financiers, opérationnels, informatiques ou encore de non-conformité.

La direction regroupe une équipe d'environ **36 collaborateurs** (hors alternants er stagiaires), habilitée à accéder à l'ensemble des données de l'entreprise, ce qui permet d'analyser un volume croissant d'informations. Le Directeur de l'Audit Interne exerce la fonction clé d'audit interne pour le Groupe CAA et certaines de ses filiales, sur décision du Conseil d'administration et sous réserve d'approbation par les autorités de supervision compétentes. Son rôle et son positionnement dans le dispositif de gestion des risques sont définis dans la politique de gouvernance du Groupe. La coordination avec les autres fonctions clés repose sur des échanges réguliers, la transmission de rapports et de tableaux de suivi, ainsi que la participation active à certaines instances de gouvernance.

1.4.5 Fonctions à la DAA

Au sein de la direction d'audit interne de Crédit Agricole Assurances, nous retrouvons des fonctions bien précises assurées par chaque les collaborateurs des différents paliers hiérarchiques :

Poste	Fonctions
Auditeurs	— Contribuer à des missions d'audit interne au sein de l'ensemble des compagnies et métiers du Groupe Crédit Agricole Assurances (France, international, prestataires essentiels). — Participer à des missions menées par leur direction ou par d'autres unités d'audit de la Ligne Métier Audit Inspection du Groupe Crédit Agricole.
Chefs de mission	— Conduire des missions d'audit interne au sein des compagnies et métiers du Groupe CAA, en France et à l'international. — Contribuer à des missions portant sur d'autres entités du Groupe Crédit Agricole. — Apporter leur expertise, participer à la prise de décision et à la mise en œuvre du plan d'actions de leur entité. — Suivre les évolutions techniques, en analyser les impacts et proposer/mettre en œuvre les adaptations nécessaires.
Superviseurs	— Contribuer à la définition et à la mise en œuvre du plan d'audit de la Direction de l'Audit du Groupe CAA. — Anticiper les évolutions et assurer la préparation et la supervision des missions d'audit confiées. — Prendre en charge des dossiers spécifiques et les actions relevant de leur domaine d'expertise. — Représenter l'Audit auprès des instances de gouvernance du Groupe CAA, en relais du Directeur de l'Audit, et contribuer aux travaux de la Ligne Métier Audit Inspection du Crédit Agricole. — Manager une équipe de chefs de mission et d'auditeurs.

TABLE 1.1 – Tableau récapitulatif des postes et fonctions au sein de la Direction de l'Audit des Assurances

1.5 Pôle Data au sein de la DAA

Au sein de la **Direction de l'Audit des Assurances**, les collaborateurs s'organisent autour d'un rythme de **trois vagues de missions par an**. Les missions d'audit de chaque vague durent généralement entre trois et quatre mois, et chaque équipe de direction se voit confier des travaux spécifiques pendant cette période. Une équipe est composée d'un **chef de mission**, d'**auditeurs** et, en fonction des thématiques abordées, d'**experts IT** ou d'**actuaires**.

Le **pôle Data** occupe un rôle transversal puisqu'il met à disposition des auditeurs des outils et des analyses permettant d'exploiter efficacement les données et de renforcer la qualité des missions. Mon positionnement au sein de cette équipe m'a conduit à collaborer directement avec les auditeurs, tout en bénéficiant de l'accompagnement de mon tuteur d'alternance. J'ai ainsi pu contribuer aux travaux de la DAA en apportant un appui technique et méthodologique sur différentes problématiques liées à la **modélisation statistique** et à l'**analyse de données**.

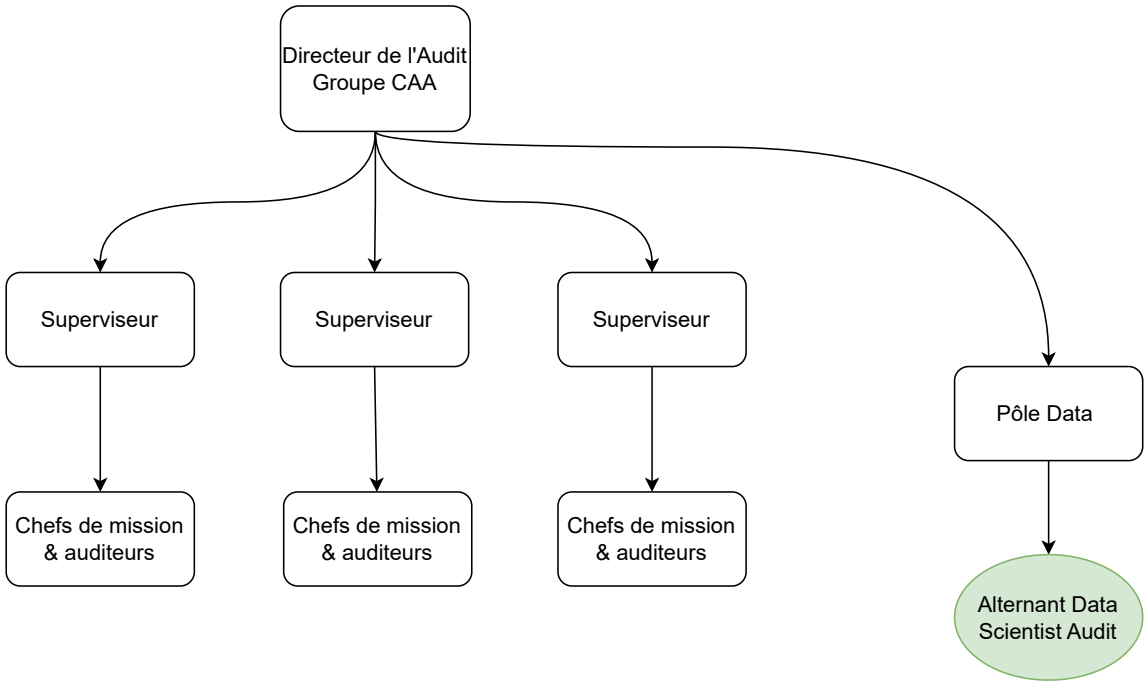

FIGURE 1.3 – Organigramme hiérarchique synthétique de la DAA

Chapitre 2

Interprétabilité des clauses contractuelles en assurance dommages

Introduction

Dans les contrats d'assurance, et en particulier en assurance dommages, les clauses contractuelles jouent un rôle central : elles déterminent les conditions de couverture, les garanties, mais aussi les exclusions qui limitent le champ de l'assurance. Or, ces clauses ne sont pas toujours rédigées de manière claire et transparente. Certaines formulations laissent place à des incertitudes, ce qui pose un problème d'interprétabilité.

La jurisprudence et les régulateurs ont souligné à plusieurs reprises cette difficulté. La **Cour de cassation** rappelle que l'assurance, en tant que mécanisme de protection contre les aléas, doit être aussi prévisible que possible. Une clause obscure ou ambiguë peut fragiliser la sécurité juridique du contrat, en laissant l'assuré dans l'incertitude quant à l'étendue de sa couverture. Dans le même sens, l'**Autorité de Contrôle Prudentiel et de Résolution (ACPR)** a récemment mis en garde les assureurs contre la présence, dans certains contrats d'habitation ou d'automobile, de clauses d'exclusion dites « non formelles et limitées », comme celles relatives au « défaut d'entretien », au « non-respect des règles de l'art » ou encore à la « négligence de l'assuré ». Ces expressions, trop générales, ne permettent pas à l'assuré de comprendre clairement dans quels cas la garantie s'applique, et sont régulièrement censurées par la jurisprudence.

Ces constats traduisent un enjeu majeur : l'interprétabilité des clauses contractuelles n'est pas seulement une question de rédaction juridique, mais aussi une condition essentielle de transparence et de confiance entre l'assureur et l'assuré. Elle touche à la protection du consommateur, à la qualité de la gouvernance des contrats et, plus largement, à la stabilité du marché de l'assurance.

C'est dans ce contexte que s'inscrit notre travail. L'objectif est de développer un **modèle de classification binaire** capable de distinguer automatiquement deux types de clauses

contractuelles dans les contrats d'assurance dommages :

— les clauses **interprétables**, c'est-à-dire celles qui laissent place à la confusion ;

— et les clauses **non interprétables**, rédigées de manière claire et précise.

Pour atteindre cet objectif, nous avons constitué un jeu de données de clauses issues de différentes sources, puis nous avons appliqué un prétraitement linguistique (nettoyage, normalisation) et une vectorisation adaptée. Enfin, plusieurs modèles d'apprentissage supervisé ont été entraînés et comparés afin d'évaluer leur capacité à prédire la classe d'une clause. Ce travail associe ainsi une problématique juridique et opérationnelle avec des outils issus du traitement automatique du langage (NLP) et du machine learning, dans la perspective de renforcer la lisibilité et la qualité des contrats d'assurance.

2.1 Contexte et enjeux de l'interprétabilité

L'interprétabilité des clauses contractuelles est un enjeu central dans le domaine de l'assurance dommages. Elle ne se limite pas à un aspect juridique : elle touche directement à la compréhension des contrats, à la confiance entre les parties et à la stabilité du marché de l'assurance.

Un contrat clair permet à l'assuré de savoir exactement dans quelles situations il est protégé. À l'inverse, un contrat mal rédigé ou comportant des clauses ambiguës peut générer des litiges et alimenter un sentiment de méfiance.

On peut distinguer plusieurs niveaux d'enjeux :

— **Pour l'assuré** : la compréhension du contrat est essentielle pour éviter les mauvaises surprises au moment d'un sinistre. Un manque de clarté peut aboutir à un refus d'indemnisation difficilement anticipé.

— **Pour l'assureur** : des clauses interprétables augmentent le risque de contentieux et peuvent fragiliser la relation commerciale avec les clients. La clarté des contrats contribue à la satisfaction et à la fidélisation.

— **Pour le régulateur** : garantir que les contrats respectent les principes de transparence et de protection du consommateur est un objectif de gouvernance. L'interprétabilité devient alors une exigence de conformité, en plus d'un enjeu de marché.

2.2 Définition opérationnelle

Dans ce travail, nous adoptons une définition pragmatique de l'interprétabilité des clauses contractuelles. Une clause est dite **interprétable** lorsqu'elle laisse place à la confusion, soit

parce qu'elle utilise une formulation vague, soit parce qu'elle manque de précision et ouvre la porte à plusieurs interprétations possibles. À l'inverse, une clause est considérée comme **non interprétable** lorsqu'elle est rédigée de manière claire, précise et sans ambiguïté pour l'assuré.

Exemples illustratifs

Pour rendre cette distinction plus concrète, on peut comparer deux formulations :

— **Clause interprétable** : « l'assuré doit prendre toutes les mesures nécessaires pour prévenir le risque ». Cette formulation est problématique car elle ne précise pas quelles mesures sont attendues, ni selon quels critères elles doivent être évaluées.

— **Clause non interprétable** : « l'assurance couvre les dommages causés par un incendie à condition que l'assuré ait installé un détecteur de fumée conforme à la norme NF-EN 14604 ». Ici, la condition est claire, vérifiable et ne laisse pas de place à des interprétations divergentes.

Cette définition opérationnelle sert de base à la construction de notre jeu de données et à la mise en place de la classification binaire. Chaque clause contractuelle étudiée appartient donc à l'une de ces deux catégories, ce qui permet de transformer une problématique juridique et qualitative en un problème de *machine learning* supervisé.

Ainsi, l'interprétabilité s'impose comme un sujet transversal : elle conditionne à la fois la sécurité juridique, la qualité de la relation client et le respect des obligations réglementaires. C'est pourquoi son étude dépasse le champ théorique et doit s'appuyer sur des méthodes d'analyse concrètes et reproductibles, comme celles issues du traitement automatique du langage.

2.3 Constitution et préparation des données

La mise en place d'un modèle de classification nécessite la constitution d'un jeu de données représentatif, regroupant à la fois des clauses interprétables et des clauses non interprétables. Cette étape a permis de transformer des documents contractuels hétérogènes en une base exploitable par des algorithmes d'apprentissage supervisé.

2.3.1 Sources des clauses

Les clauses utilisées dans ce travail proviennent de plusieurs sources complémentaires :

— des clauses jugées **interprétables**, identifiées par le *Médiateur de l'assurance* et enrichies par des exemples générés via Copilot ;
— des clauses considérées comme **non interprétables**, extraites des contrats d'assurance dommages de *Pacifica*, réputés pour la rigueur de leur rédaction.

2.3.2 Construction du jeu de données

Après extraction des textes bruts (via le package `Fitz`), chaque clause a été intégrée dans une base structurée comprenant :

— le texte brut de la clause,
— une variable cible binaire (interpretabilite) indiquant son niveau d'interprétabilité (1 = interprétable, 0 = non interprétable),
— une version vectorisée pour la phase de modélisation.

Le jeu de données final comprend **596 clauses**, dont 79% sont non interprétables et 21% interprétables dans le jeu d'entraînement de 417 clauses.

2.3.3 Prétraitement du texte

Un prétraitement linguistique a été appliqué afin de rendre les données exploitables par les modèles de machine learning. Cette étape comprend plusieurs opérations successives, chacune ayant un rôle précis :

— **Suppression des balises HTML et caractères parasites** : permet de nettoyer les résidus liés à l'extraction des textes depuis les fichiers PDF, afin de ne conserver que le contenu utile.
— **Conversion en minuscules** : harmonise l'ensemble du texte pour éviter que les majuscules et minuscules soient considérées comme des mots différents (exemple : « Assurance » et « assurance »).
— **Tokenisation en mots** : décompose chaque clause en une suite de mots (ou tokens), ce qui facilite leur traitement par les algorithmes.
— **Suppression des *stopwords*** : élimine les mots très fréquents (articles, conjonctions…) qui n'apportent pas d'information discriminante pour la classification.
— **Lemmatisation** : réduit chaque mot à sa forme canonique (par exemple « couvertures » devient « couverture »), afin de regrouper les variantes d'un même terme.
— **Reconstitution du texte normalisé** : reconstruit une version homogène des clauses, composée uniquement des mots significatifs et lemmatisés.

Ce nettoyage a permis d'obtenir une base textuelle cohérente, prête à être transformée

en vecteurs numériques. Ces représentations sont indispensables pour l'entraînement des modèles de classification supervisée.

2.4 Approche méthodologique en NLP

Une fois le jeu de données constitué et les textes prétraités, l'étape suivante a consisté à transformer les clauses contractuelles en représentations numériques exploitables par des modèles de classification, puis à tester différents algorithmes supervisés.

2.4.1 Vectorisation des clauses

La transformation du texte en vecteurs a été réalisée à l'aide du modèle `Word2Vec`. Ce modèle permet de représenter chaque mot sous forme de vecteur dans un espace de dimension réduite, de manière à capturer les relations sémantiques entre les termes. Les principaux paramètres retenus sont :

- `vector_size = 100` : dimension des vecteurs de mots,
- `window = 5` : taille de la fenêtre contextuelle, soit cinq mots avant et cinq mots après le mot cible,
- `min_count = 1` : nombre minimal d'occurrences d'un mot pour qu'il soit pris en compte.

Chaque clause est ainsi représentée par un vecteur agrégé, permettant son intégration dans les modèles de classification supervisée.

2.4.2 Modèles de classification testés

Trois familles de modèles ont été évaluées dans ce travail. Chaque approche présente des avantages et des limites, ce qui permet d'obtenir une comparaison pertinente dans le cadre de la classification binaire des clauses contractuelles.

- **Régression logistique** : ce modèle linéaire sert de point de départ (*baseline*). Il apprend une relation entre les variables explicatives (vecteurs de clauses) et la variable cible (interprétable / non interprétable). Bien que simple et rapide, il peut manquer de flexibilité pour capturer des relations complexes. Son principal atout est sa lisibilité et sa capacité à donner une première référence de performance.

— **Random Forest** : ce modèle repose sur la combinaison d'un grand nombre d'arbres de décision. Chaque arbre est entraîné sur un sous-échantillon du jeu de données (technique du *bootstrap*), et à chaque nœud, une sélection aléatoire de variables est utilisée pour décider des séparations. Cette double part de hasard permet de réduire le risque de surapprentissage présent dans un arbre unique (qui aurait tendance à mémoriser les données d'entraînement). Le résultat final est obtenu par un vote majoritaire entre tous les arbres. Le Random Forest est robuste au bruit, gère bien les données déséquilibrées et capte des relations non linéaires, mais il peut être coûteux en calcul et moins interprétable.

— **Gradient Boosting** : à la différence du Random Forest qui entraîne plusieurs arbres en parallèle, le Gradient Boosting construit ses arbres de manière séquentielle. Chaque nouvel arbre vient corriger les erreurs commises par les précédents (*weak learners*). Ce processus permet d'obtenir des modèles très performants, notamment lorsque les données sont complexes. Deux hyperparamètres clés influencent fortement ses performances : le *learning rate*, qui contrôle la vitesse d'apprentissage et le risque de surapprentissage, et le *subsample*, qui introduit de la régularisation en limitant la part des données utilisées à chaque itération. Le Gradient Boosting est reconnu pour donner d'excellents résultats, mais il est sensible aux réglages et plus long à entraîner.

Ainsi, la comparaison de ces trois modèles permet d'opposer : une méthode de référence simple et interprétable (régression logistique), un modèle robuste basé sur l'agrégation aléatoire d'arbres (Random Forest), et un modèle séquentiel plus sophistiqué visant la performance maximale (Gradient Boosting).

2.4.3 Gestion du déséquilibre des classes

Le jeu de données étant déséquilibré (moins de clauses interprétables que de clauses non interprétables), la méthode `SMOTE` (*Synthetic Minority Oversampling Technique*) a été utilisée. Elle consiste à générer artificiellement de nouvelles instances de la classe minoritaire (on parle de suréchantillonnage de la classe minoritaire) afin d'équilibrer le jeu d'entraînement et d'améliorer la capacité de généralisation des modèles.Un sous-échantillonnage de la classe majoritaire était aussi une option, mais moins envisageable dans notre cas car nous aurions perdu des données utiles.

2.5 Résultats et interprétation

Après entraînement et optimisation, les performances des différents modèles ont été comparées sur le jeu de test afin d'évaluer leur capacité à distinguer les clauses interprétables des clauses non interprétables.

2.5.1 Comparaison des modèles

Trois modèles ont été testés :

— la **régression logistique**, utilisée comme modèle de base ;
— le **Random Forest**, basé sur une agrégation d'arbres de décision ;
— le **Gradient Boosting**, qui apprend de manière séquentielle en corrigeant les erreurs successives.

Pour chaque modèle, plusieurs hyperparamètres ont été explorés.

— **Régression logistique** : aucun terme de régularisation n'a été ajouté, la variable `clauses_vect` étant l'unique prédicteur.
— **Random Forest** : nous avons testé différentes valeurs de `n_estimators` (100, 200, 300), de profondeur maximale (`max_depth` = 0, 10, 20, 30), ainsi que de `min_samples_split` (2, 5, 10) et `min_samples_leaf` (1, 2, 4).
— **Gradient Boosting** : les hyperparamètres évalués incluent `n_estimators` (50, 100, 200), `learning_rate` (0.01, 0.1, 0.2), `max_depth` (3, 5, 10), `min_samples_split` (2, 10), `min_samples_leaf` (1, 5), ainsi que le `subsample` (0.7, 1.0).

2.5.2 Résultats obtenus

Modèles de classification	Accuracy	AUC	Hyperparamètres
Régression logistique	0,44	0,78	C = 1 (pas de régularisation possible)
Random Forest	0,81	0,82	max_depth = None min_samples_leaf = 1 min_samples_split = 2 n_estimators = 200
Gradient Boosting	0,83	0,82	learning_rate = 0,2 max_depth = 5 min_samples_leaf = 1 min_samples_split = 2 n_estimators = 200 subsample = 0,7

Le tableau des métriques montre que le **Gradient Boosting** obtient les meilleures performances sur le jeu de test. Ce modèle parvient mieux à équilibrer précision et rappel, et se démarque particulièrement par ses résultats sur les clauses interprétables.

Au-delà des métriques numériques, il est utile de représenter graphiquement les performances du modèle retenu. Deux visualisations complémentaires ont été produites pour le Gradient Boosting : la matrice de confusion et la courbe ROC.

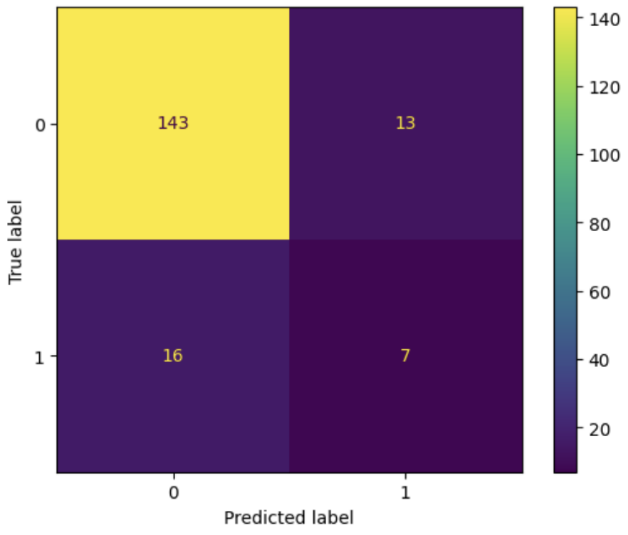

FIGURE 2.1 – Matrice de confusion du jeu de test avec Gradient Boosting

La matrice de confusion met en évidence que le modèle parvient à bien classer les clauses non interprétables, mais reste en difficulté sur la classe interprétable. On observe ainsi un nombre plus élevé de faux négatifs (clauses interprétables prédites comme non interprétables), ce qui confirme l'analyse des métriques.

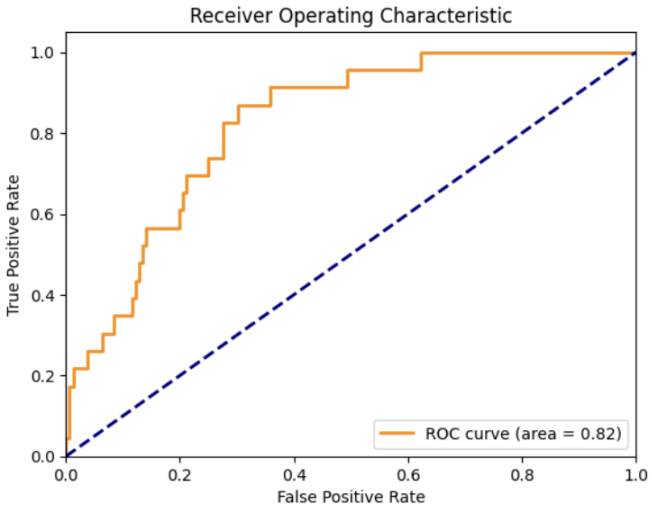

FIGURE 2.2 – Courbe ROC du modèle Gradient Boosting

La courbe ROC montre une aire sous la courbe (AUC) de l'ordre de 0,82, indiquant une bonne capacité de discrimination globale du modèle. On remarque également que l'aspect en « escalier » de la courbe s'est atténué après l'ajout des clauses Pacifica, ce qui suggère une meilleure robustesse et une couverture plus homogène du jeu de données.

2.5.3 Interprétation

En résumé : Les résultats obtenus permettent de tirer plusieurs enseignements. Tout d'abord, on constate que les modèles parviennent plus facilement à classer correctement les clauses non interprétables, qui sont plus fréquentes et souvent rédigées de manière explicite. À l'inverse, les clauses interprétables restent difficiles à identifier, car elles reposent sur des formulations vagues ou ambigües, parfois proches de clauses claires, ce qui complique la tâche de la classification automatique.

La comparaison des modèles montre également une hiérarchie claire. La régression logistique, utilisée comme baseline, obtient des performances modestes (accuracy de 0,44, AUC de 0,78). Le Random Forest améliore nettement les résultats (accuracy de 0,81, AUC de 0,82) grâce à sa capacité à capter des relations non linéaires. Enfin, le Gradient Boosting se distingue légèrement avec une accuracy de 0,83 et une AUC de 0,82, confirmant qu'il

s'agit du modèle le plus adapté à notre problématique.

Cependant, malgré ces performances, les **F1-scores restent inférieurs à 0,5**, ce qui traduit une difficulté persistante à bien détecter les clauses interprétables. Cela confirme que la complexité du langage juridique et le déséquilibre des données limitent la performance actuelle des modèles.

En résumé, le Gradient Boosting constitue une première solution prometteuse, mais il reste nécessaire d'enrichir les données et d'améliorer la représentation des textes pour progresser sur la détection des clauses véritablement ambigües.

2.6 Limites et perspectives

Malgré les résultats encourageants obtenus, plusieurs limites doivent être soulignées :

— **Taille du jeu de données limitée** : seulement 596 clauses, ce qui restreint la capacité des modèles à généraliser.

— **Biais introduit par l'augmentation des données** : la génération artificielle de nouvelles données peut améliorer l'équilibre mais aussi dégrader la qualité de l'apprentissage.

— **Complexité du langage juridique** : certaines expressions comme « défaut d'entretien » ou « négligence de l'assuré » sont volontairement générales et difficiles à traiter automatiquement.

— **Performances encore insuffisantes** : les modèles détectent mieux les clauses non interprétables, alors que les F1-scores restent inférieurs à 0,5 pour la classe interprétable.

— **Subtilités linguistiques non prises en compte** : Notre modèle ne capture pas la structure syntaxique ni les nuances contextuelles

— Il nous a été compliqué voire impossible d'explorer la large gammes de modèles de Hugging Face et les implémenter sur nos données.

Plusieurs pistes peuvent être envisagées pour améliorer ces résultats :

— **Enrichir le jeu de données** en intégrant davantage de contrats et de décisions du Médiateur, afin d'élargir la couverture des cas.

— **Explorer des modèles plus avancés**, comme BERT ou CamemBERT, mieux adaptés aux subtilités linguistiques et au contexte.

— **Optimiser les hyperparamètres** et tester d'autres méthodes de rééquilibrage pour améliorer la détection des clauses interprétables.

— **Développer des approches hybrides** combinant machine learning et expertise humaine, en utilisant le modèle comme un outil d'aide à la décision pour les juristes.

En résumé, ce travail constitue une première étape vers l'automatisation de la détection

des clauses interprétables. Il montre que les modèles de machine learning peuvent fournir une aide précieuse, mais que des efforts supplémentaires sont nécessaires pour capturer toute la complexité du langage contractuel.

Chapitre 3

Le risque climatique et les enjeux pour l'assurance agricole

Introduction

Le changement climatique accentue la fréquence et l'intensité des phénomènes extrêmes tels que tempêtes, sécheresses, inondations ou gelées, affectant fortement le secteur agricole. Ce contexte rend la gestion du risque climatique incontournable pour les assureurs agricoles, confrontés à une sinistralité croissante.

En France, le système d'assurance multirisque agricole (MA) repose sur un partenariat public-privé dans lequel Pacifica, filiale du Crédit Agricole Assurances, joue un rôle central. Cependant, la soutenabilité économique de ce modèle est remise en question, appelant à une modernisation des approches actuarielles.

Plusieurs études confirment ce besoin. Möhring et al. (2020) montrent que l'assurance influence les décisions agricoles et nécessite des modèles adaptés. Capitanio (2022) appelle à un meilleur encadrement de l'incertitude. Santeramo et al. (2024) soulignent le potentiel des outils indiciels face aux événements extrêmes. Enfin, le rapport FI-Compass (2023) recommande l'amélioration des données météorologiques pour mieux cibler les indemnisations.

Les régulateurs partagent cette préoccupation. En juillet 2025, la Banque Centrale Européenne a envisagé une sanction de 7 millions d'euros contre le Crédit Agricole pour une prise en compte insuffisante du risque climatique dans ses outils de pilotage (La BCE, 2025 – Agefi). Cette pression réglementaire incite à une meilleure maîtrise du risque environnemental.

L'objectif de cette partie est d'estimer la probabilité d'un sinistre tempête à l'échelle contrat–trimestre, en combinant les données assurantielles issues du portefeuille de Pacifica avec les données climatiques ERA5 (vent moyen, rafales…), à l'aide de modèles de machine learning.

Ce chapitre présente : un cadrage du risque climatique en agriculture, la méthode de construction de la base sinistres–vent, et la logique prédictive adoptée au regard de la littérature.

3.1 Construction de la base de données

La modélisation du risque climatique à l'échelle *contrat–trimestre* repose sur la constitution préalable d'une base de données robuste, propre et exploitable. Dans ce travail, deux sources principales de données ont été mobilisées : d'un côté, les données internes issues du portefeuille de contrats d'assurance Multirisque Agricole (MA) géré par Pacifica ; de l'autre, les données climatiques ouvertes issues de la base ERA5, produite par le programme Copernicus, portant sur les vitesses de vent et les rafales.

Ces deux sources répondent à des logiques différentes : les données assurantielles permettent d'identifier les expositions assurées et les sinistres déclarés, tandis que les données météorologiques permettent de quantifier l'intensité des aléas climatiques subis localement. Pour être croisées de manière pertinente, ces données ont dû faire l'objet d'un traitement rigoureux.

L'ensemble de ces opérations a conduit à la création d'une base finale structurée, dans laquelle chaque ligne représente un contrat actif pour un trimestre donné, enrichi par ses caractéristiques assurantielles et les mesures météorologiques correspondantes.

Cette section présente la logique de construction de cette base, en détaillant successivement :

— la structuration des données assurantielles du portefeuille,
— la préparation et l'agrégation des données climatiques ERA5,
— et enfin, la fusion des deux sources dans un jeu de données consolidé.

3.1.1 Données assurantielles issues du portefeuille Multirisque Agricole

Les données assurantielles utilisées dans ce travail proviennent du portail interne de **Pacifica**, hébergé sur l'environnement **SAS Enterprise Guide**. Ce portail structure les données en bibliothèques mensuelles appelées `DateYYMM`, chacune contenant un ensemble de tables métiers. Dans ce contexte, l'exploitation des données s'est concentrée sur trois sources principales : les **contrats**, les **sinistres**, et les **garanties**.

— **Identification des contrats actifs au 06/2025**

La première étape a consisté à identifier les contrats du produit *Multirisque Agricole (MA)* encore en vigueur au mois de juin 2025. Pour cela, la bibliothèque `Date0625` a été mobilisée, en se focalisant sur la table `MVTPRMA`, qui contient l'historique de chaque contrat, représenté par plusieurs « images » successives. Chaque image correspond à un état du contrat à une date donnée. Seules les dernières images dispo-

nibles ont été conservées, représentant l'état le plus à jour des contrats au 06/2025.

— **Filtrage des contrats avec un historique suffisant**

L'analyse ne s'est pas limitée à une photographie du portefeuille à l'instant T. Pour assurer une profondeur temporelle suffisante en vue de la modélisation, seuls les contrats disposant d'au moins cinq années d'historique ont été retenus. Ce critère garantit la disponibilité d'un suivi trimestriel complet depuis 2020, en cohérence avec la période d'observation retenue pour les sinistres.

— **Constitution de la table des sinistres**

Les données relatives aux sinistres ont été extraites des différentes versions mensuelles de la table `SIN_DGL`, disponibles dans les bibliothèques `DateYYMM` de juin 2020 à juin 2025. Toutes ces tables ont été concaténées pour former un historique complet des sinistres sur la période. Pour chaque sinistre, seule la dernière image disponible a été conservée afin de ne pas dupliquer les informations.

— **Ajout des informations sur les garanties déclenchées**

En parallèle, les données de la table `SIN_GTIE`, qui précisent les garanties activées par chaque sinistre, ont également été concaténées sur la même période. Cette étape est essentielle pour qualifier la nature du sinistre subi par un contrat.

NB : Les tables `SIN_DGL` permettent d'identifier les contrats ayant subi un sinistre, tandis que les tables `SIN_GTIE` renseignent sur la nature des sinistres. Cela permet de filtrer et ne retenir que les sinistres climatiques spécifiquement liés au vent ou aux tempêtes, en cohérence avec l'objectif de cette étude.

3.1.2 Données climatiques issues d'ERA5

Les données climatiques mobilisées dans ce travail proviennent de la base **ERA5**, produite par le programme européen *Copernicus Climate Change Service (C3S)* et gérée par le *Centre Européen pour les Prévisions Météorologiques à Moyen Terme (CEPMMT)*. ERA5 constitue une référence en matière de réanalyses climatiques, offrant des données de haute résolution spatiale (0,25° x 0,25°) et temporelle (horaire), couvrant l'ensemble du globe.

— **Variables sélectionnées**

Dans le cadre de ce travail, seules les variables liées au vent ont été extraites depuis l'API *Climate Data Store*, à l'aide du package `cdsapi`, à partir d'un script automatisé. Les variables concernées sont :

— la vitesse moyenne du vent à 10 mètres (en m/s),
— la vitesse moyenne du vent à 100 mètres (en m/s),
— la vitesse maximale des rafales de vent (gusts, en m/s).

— **Période et étendue géographique**

  La période couverte s'étend de **2012 à 2025**, mais pour rester cohérent avec la plage temporelle des données assurantielles, seules les données de vent entre **juin 2020 et juin 2025** ont été utilisées. Les extractions ont été limitées à la **France métropolitaine**, en sélectionnant les coordonnées géographiques correspondant aux points ERA5 situés sur ce territoire.

— **Agrégation spatiale par commune**

  Une jointure spatiale a été effectuée entre les **points ERA5** et les **communes INSEE**, à l'aide d'un shapefile contenant les limites géographiques des communes françaises (traitement effectué via la bibliothèque `geopandas` en Python). Chaque point de mesure a été associé à une commune en fonction de sa position géographique, permettant de regrouper les mesures horaires à la maille *commune–jour*, puis *commune–mois*.

— **Agrégation temporelle à la maille trimestre**

  Les données climatiques ont ensuite été agrégées à la maille *commune–trimestre*, en cohérence avec la structure des contrats d'assurance. Pour chaque commune et chaque trimestre, les indicateurs suivants ont été calculés pour chaque variable de vent (10m, 100m, rafales) :
  — Moyenne,
  — Maximum,
  — Minimum,
  — Écart-type.

— **Construction de la base météorologique finale**

  Les données trimestrielles obtenues ont été concaténées dans une base unique, nommée `df_vent_final`, dans laquelle chaque ligne correspond à une *commune* pour un *trimestre donné*, enrichie de 12 variables climatiques (moyenne, max, min, écart-type pour les trois types de vent).

3.1.3 Fusion des bases de données climatiques et assurantielles

Cette dernière étape consiste à fusionner les deux sources de données précédemment construites — la base assurantielle à la maille *contrat–trimestre* et la base climatique à la

maille *commune–trimestre* — afin d'obtenir une base finale prête à être exploitée pour la modélisation.

— **Création de la table de correspondance contrat–commune**

  Chaque contrat identifié comme actif au 06/2025 dans la base assurantielle a été rattaché à une commune INSEE, en se basant sur les variables d'adresse géographique présentes dans les tables métier de Pacifica. Cette opération a permis de construire une table de correspondance entre chaque identifiant contrat et le code INSEE de sa commune d'implantation.

— **Appariement temporel des trimestres**

  Pour chaque contrat disposant d'un historique d'au moins cinq ans, un enregistrement a été généré pour chaque trimestre de la période `2020-T2 - 2025-T2`, avec comme fréquence une ligne par *contrat–trimestre*.

— **Jointure avec la base des données sur le vent**

  Une jointure a ensuite été effectuée entre cette base `contrat-trimestre` et la base climatique de vent, construite précédemment à la maille `commune-trimestre`. La correspondance a été réalisée sur deux clés : le code INSEE de la commune et le trimestre de référence. Chaque contrat s'est ainsi vu enrichi de l'ensemble des mesures climatiques liées au vent observées dans sa commune, sur le trimestre concerné.

— **Base finale consolidée**

  Le résultat de cette fusion est une base consolidée dans laquelle chaque ligne représente un contrat actif sur un trimestre donné, enrichi :
  — de ses caractéristiques assurantielles,
  — de l'occurrence éventuelle d'un sinistre pour ce trimestre,
  — et des indicateurs climatiques correspondants (vent 10m, vent 100m, rafales, avec moyenne, min, max et écart-type).

Cette base finale servira de fondation à la phase de modélisation du risque climatique lié aux tempêtes, détaillée dans la section suivante.

3.2 Modélisation des sinistres

Après avoir consolidé les données assurantielles et climatiques à la maille *contrat–trimestre*, cette section est consacrée à la modélisation du risque de sinistre lié au vent. L'objectif est de construire des modèles de classification capables d'identifier les périodes et contrats les plus exposés, à partir de variables explicatives historiques.

Cette phase mobilise des techniques de machine learning, en tenant compte des spécificités du problème : fort déséquilibre de classes, dépendance temporelle des événements, et hétérogénéité des contrats. Plusieurs approches seront explorées, depuis des modèles de base jusqu'à des algorithmes plus avancés en intégrant des méthodes de rééchantillonnage et d'optimisation des seuils de décision.

Les sections suivantes détaillent successivement l'objectif de la modélisation, les enjeux méthodologiques, les modèles testés ainsi que les résultats obtenus et les limites rencontrées.

3.2.1 Objectif de la modélisation

L'objectif de cette modélisation est de prédire la survenue d'un sinistre lié au vent pour un contrat donné, à une échéance trimestrielle. Il s'agit d'un problème de classification binaire, où la variable cible indique si un contrat a subi ou non un sinistre au cours d'un trimestre donné.

Le choix d'une granularité *contrat–trimestre* s'explique avant tout par des considérations pratiques : d'une part, il permet de réduire la complexité computationnelle en limitant le volume total de données, et d'autre part, il atténue le fort déséquilibre de classes inhérent au problème. En effet, une maille mensuelle aurait conduit à une proportion extrêmement faible d'occurrences positives (sinistres), rendant l'entraînement des modèles beaucoup plus instable.

La variable cible, notée `SINISTRE`, prend la valeur 1 lorsqu'un sinistre climatique lié au vent est enregistré sur le contrat au trimestre considéré, et 0 sinon. L'objectif est donc d'estimer, pour chaque contrat actif et chaque trimestre, la probabilité d'occurrence d'un sinistre afin d'identifier de manière précoce les profils ou périodes à risque.

À travers cette approche, il s'agit d'exploiter conjointement les données assurantielles historiques et les conditions climatiques observées pour renforcer les outils de prévision du risque climatique dans un contexte de sinistralité croissante.

3.2.2 Enjeux méthodologiques

La modélisation du risque de sinistre climatique à l'échelle *contrat–trimestre* soulève plusieurs enjeux méthodologiques majeurs. Le premier concerne le fort déséquilibre entre les classes, avec une proportion très faible de sinistres par rapport aux observations sans sinistre, rendant nécessaire l'utilisation de techniques de rééchantillonnage telles que le sur-échantillonnage (SMOTE) ou l'undersampling. Le second défi réside dans la stratification temporelle et contractuelle : il est indispensable de s'assurer que les observations issues d'un même contrat ne soient pas réparties entre l'entraînement et le test afin d'éviter toute fuite d'information. Par ailleurs, la richesse des données climatiques et assurantielles rend cruciale une sélection pertinente des variables, afin d'éviter le sur-apprentissage et de réduire la complexité computationnelle. Enfin, la définition du bon seuil de décision pour la classification s'avère stratégique, notamment pour optimiser des métriques adaptées aux situations déséquilibrées comme le F1-score de la classe minoritaire.

3.2.3 Régression logistique

La régression logistique a été choisie comme modèle de référence en raison de sa simplicité, de sa robustesse et de son caractère interprétable. Elle permet d'établir une première baseline dans un contexte de classification binaire, ici la survenue ou non d'un sinistre sur une période donnée. Deux versions du modèle ont été évaluées : une version basique sans prétraitement avancé, puis une version enrichie intégrant plusieurs techniques d'optimisation.

Dans un premier temps, la version simple a consisté à entraîner un modèle de régression logistique avec pénalisation L2 (ridge), en prenant en compte le déséquilibre des classes via le paramètre `class_weight="balanced"`. Cette configuration permet de ne pas biaiser le modèle en faveur de la classe majoritaire (absence de sinistre), tout en conservant une architecture de base minimaliste. Sur le jeu de test, le modèle atteint un AUC de **0.8176** et un F1-score de **0.1225**, traduisant une faible capacité de détection des sinistres.

Dans un second temps, une version optimisée a été proposée afin d'améliorer la détection de la classe minoritaire. Cette version inclut : une sélection de variables par pénalisation L1 (lasso), un pipeline de rééquilibrage combinant SMOTE et undersampling, ainsi qu'une optimisation du seuil de décision. Toutefois, malgré ces ajustements, les performances n'ont pas significativement progressé. L'AUC obtenu diminue (**0.7764**), tandis que le F1-score demeure faible (**0.1137**). Cela suggère que ces techniques, bien que théoriquement pertinentes, n'ont pas permis une meilleure séparation des classes dans ce cas précis.

Ce constat met en lumière les limites des approches linéaires dans des contextes de forte

imbalance et de non-linéarité potentielle, ce qui justifie l'expérimentation de modèles de type arbre ou ensemble abordés dans la section suivante.

3.2.4 Modèles d'ensemble

Après avoir constaté les limites de la régression logistique dans un contexte de données fortement déséquilibrées, l'attention s'est portée sur les modèles d'ensemble, reconnus pour leur capacité à capturer des relations non linéaires et à mieux gérer les situations d'*imbalance*. Deux grandes approches ont été explorées : le **bagging**, avec la Random Forest, et le **boosting**, avec LightGBM et CatBoost.

— **Le Bagging (Bootstrap Aggregating)**, incarné ici par la Random Forest, repose sur l'entraînement parallèle de multiples arbres de décision, chacun sur un sous-échantillon bootstrapé des données. Les prédictions individuelles sont ensuite agrégées (par vote ou moyenne), ce qui permet de réduire la variance et d'améliorer la stabilité du modèle.

— **Le Boosting**, quant à lui, procède de manière séquentielle : chaque nouvel arbre est entraîné pour corriger les erreurs commises par les arbres précédents. Cette méthode permet de réduire le biais et d'affiner les prédictions, notamment sur les exemples mal classés. Dans ce travail, le **LightGBM**, une implémentation rapide et optimisée du Gradient Boosting, ainsi que **CatBoost**, plus robuste aux variables catégorielles, ont été mobilisés.

Pour chaque méthode, plusieurs optimisations ont été systématiquement mises en œuvre :

— Une **stratification par contrat**, afin d'éviter les fuites d'information entre les ensembles d'apprentissage, de validation et de test ;

— Une **sélection des variables** basée sur leur importance calculée par les modèles eux-mêmes, avec des seuils ajustés pour ne retenir que les plus discriminantes ;

— L'utilisation combinée de **SMOTE** (sur-échantillonnage de la classe minoritaire) et de **RandomUnderSampler** (sous-échantillonnage de la classe majoritaire), afin de rééquilibrer les classes dans le pipeline d'entraînement ;

— Et enfin, une **optimisation du seuil de décision**, souvent plus efficace que le seuil par défaut (0.5), permettant de maximiser le F1-score sur les données de validation.

3.2.4.1 Random Forest

La Random Forest a constitué la première approche de modèle d'ensemble testée dans ce travail. Grâce à sa robustesse naturelle et à sa capacité à gérer les déséquilibres de classes, elle s'est révélée pertinente pour une première montée en complexité après la régression

logistique.

La modélisation s'est appuyée sur un découpage rigoureux du jeu de données via une *Stratified Group K-Fold* par contrat, garantissant l'absence de fuites d'information entre les phases d'apprentissage, de validation et de test. Pour chaque itération, un modèle de forêt aléatoire a été entraîné, et les variables importantes ont été extraites à partir des importances de Gini.

Une boucle d'optimisation a été mise en place pour explorer différents seuils d'importance des variables, avec comme critère la maximisation du **F1-score** sur le jeu de validation. Les variables les plus discriminantes ont ensuite servi à réentraîner un modèle final, intégré dans un pipeline combinant :

— le suréchantillonnage de la classe minoritaire via SMOTE ;

— le sous-échantillonnage de la classe majoritaire via RandomUnderSampler ;

— et l'optimisation du seuil de décision, testant des valeurs entre 0.01 et 1.

Les résultats montrent une nette amélioration du F1-score(**0.2357**) comparé aux versions précédentes, traduisant une meilleure capacité du modèle à mieux identifier les sinistres, malgré leur rareté. La courbe ROC associée témoigne également d'un bon pouvoir discriminant, avec une AUC de **0.2357** Cette étape valide l'intérêt du bagging dans notre contexte, notamment grâce à sa gestion naturelle de la variance et sa capacité à capter des interactions non linéaires entre les variables explicatives.

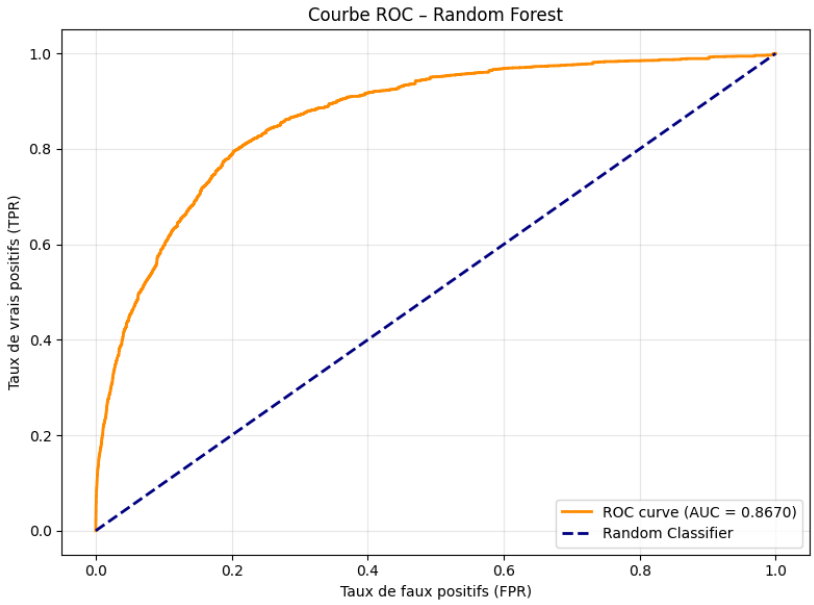

FIGURE 3.1 – Courbe ROC du modèle Random Forest sur le jeu de test.

3.2.4.2 Modèle LightGBM

Le modèle **LightGBM** (Light Gradient Boosting Machine) a été mobilisé dans le cadre de cette étude en tant qu'algorithme de type *boosting*, reconnu pour sa rapidité, son efficacité sur de grands jeux de données, ainsi que pour sa capacité à traiter efficacement les déséquilibres de classes. Contrairement à la Random Forest qui agrège des arbres construits indépendamment, le boosting fonctionne de manière **séquentielle** : chaque nouvel arbre corrige les erreurs commises par les arbres précédents, ce qui en fait une méthode particulièrement adaptée aux contextes où les observations positives (sinistres) sont rares et difficiles à isoler.

Dans le cadre de cette modélisation, le pipeline suivant a été mis en œuvre :

— **Stratification par contrat** afin de garantir l'absence de fuite d'information entre les jeux d'entraînement, de validation et de test ;

— **Sélection des variables** par importance calculée à partir du modèle lui-même, avec exploration de différents seuils pour ne retenir que les plus contributives ;

— **Rééquilibrage des classes** à l'aide d'une combinaison de SMOTE (suréchantillonnage) et de RandomUnderSampler (sous-échantillonnage) ;

— **Optimisation du seuil de décision** en maximisant le F1-score sur la validation, plutôt que d'utiliser le seuil classique de 0,5.

L'entraînement a été réalisé sur les trois premiers folds d'une cross-validation stratifiée par groupe (contrat), la validation sur le quatrième, et l'évaluation finale sur un jeu de test indépendant.

Les résultats obtenus montrent une **valeur d'AUC** satisfaisante de **0.8670**, traduisant une bonne capacité de discrimination. Le **F1-score**, quant à lui, a été légèrement amélioré en passant à 0.2506, renforçant la pertinence de cette stratégie pour des données déséquilibrées.

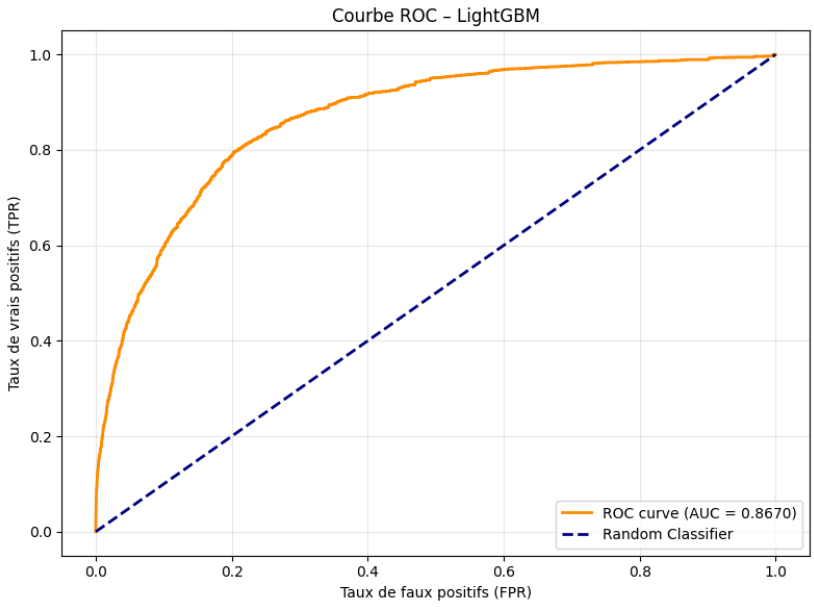

FIGURE 3.2 – Courbe ROC du modèle LightGBM sur le jeu de test

3.2.4.3 CatBoost

Le dernier modèle testé dans cette série d'expérimentations est **CatBoost**, une variante de gradient boosting développée par Yandex, particulièrement réputée pour sa robustesse face aux variables catégorielles et sa capacité à capturer des relations complexes sans nécessiter de prétraitement approfondi. Bien que les variables catégorielles aient été transformées en variables numériques dans ce travail, l'intérêt de CatBoost réside ici dans sa capacité à apprendre efficacement à partir de données déséquilibrées, tout en maintenant de bonnes performances générales.

Le pipeline utilisé pour CatBoost est identique à celui mis en place pour LightGBM : il inclut une **stratification par contrat**, une **sélection de variables** basée sur l'importance des features, un **rééquilibrage des classes** par SMOTE et RandomUnderSampler, ainsi qu'une **optimisation du seuil de décision** pour maximiser le F1-score sur l'ensemble de validation.

Les performances obtenues sur l'ensemble de test sont encourageantes, avec une **AUC de 0.8670** et un **F1-score de 0.2632**. Ces résultats suggèrent que CatBoost parvient à détecter un nombre plus important de sinistres que les modèles précédents, tout en conservant un bon niveau de précision globale. Cela en fait un candidat particulièrement pertinent pour des tâches de prédiction dans des contextes fortement déséquilibrés comme celui de ce projet.

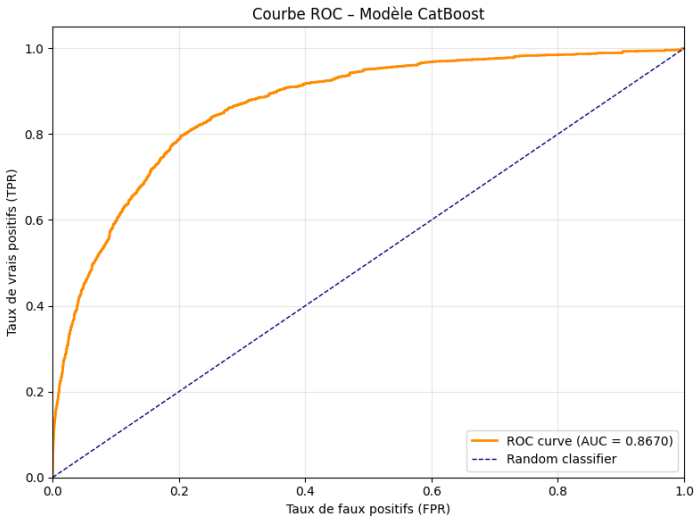

FIGURE 3.3 – Courbe ROC – Modèle CatBoost

3.2.4.4 Optimisation des hyperparamètres du modèle CatBoost

Dans le but de rendre meilleur notre modélisation, nous avons procédé à l'optimisation des hyperparamètres du modèle CatBoost afin d'améliorer sa capacité prédictive. Pour cela, une recherche exhaustive a été menée sur une grille de paramètres jugée « riche mais

raisonnable », comprenant : la profondeur des arbres ($\texttt{depth} \in \{4, 6, 8\}$), le taux d'apprentissage ($\texttt{learning\_rate} \in \{0.01, 0.05, 0.1\}$), le nombre d'itérations ($\texttt{iterations} \in \{100, 200, 300\}$) et le coefficient de régularisation des feuilles ($\texttt{l2\_leaf\_reg} \in \{1, 3, 5\}$). Cette combinaison représente un total de 81 modèles testés.

Chaque configuration a été évaluée au sein du même pipeline. L'optimisation a été réalisée en maximisant le **F1-score** sur le jeu de validation, tout en ajustant dynamiquement le seuil de décision.

À l'issue de cette recherche, les meilleurs hyperparamètres retenus sont les suivants :

$$\texttt{depth} = 8, \quad \texttt{learning\_rate} = 0.1, \quad \texttt{iterations} = 200, \quad \texttt{l2\_leaf\_reg} = 3.$$

En réentraînant le modèle CatBoost avec ces paramètres optimaux sur l'ensemble des données d'entraînement et de validation, puis en l'évaluant sur le jeu de test, nous obtenons un **AUC de 0.8081** et un **F1-score de 0.2658** (au seuil optimal fixé à 0.79). Ces résultats, bien qu'imparfaits, marquent quand même une amélioration par rapport aux versions précédentes et valident l'intérêt de l'optimisation des hyperparamètres dans notre contexte.

3.2.4.5 Importance des variables

L'interprétation du modèle est une étape clé afin de comprendre les mécanismes sous-jacents aux prédictions. Dans notre cas, nous nous concentrons sur le modèle CatBoost, retenu comme modèle final après la phase d'optimisation des hyperparamètres. L'algorithme fournit un indicateur d'*importance des variables*, mesuré par la contribution relative de chaque variable à la réduction de l'erreur de prédiction au sein des arbres de décision. La Figure 3.4 présente les 25 variables les plus influentes selon l'importance interne calculée par CatBoost. On observe que l'**ancienneté de l'exploitation**, la **note du risque globale**, et le **nombre pondéré de sinistre sur 1 an** occupent les premières positions, traduisant leur rôle déterminant dans la prédiction de la probabilité de sinistre. Les mésures du vent comme la **mésure maximale de rafales atteinte sur un trimestre** et la **vitesse du vent à 100 m** contribue à aussi à discriminer les contrats les plus suceptibles de subir un sinistre. Ces résultats mettent en évidence le poids considérable des variables liées aux carastéristiques du bien d'exploitation et aux mésures du vent. À l'inverse, certaines variables comme la **souscription à une garantie pour les évènments de base** ou **pour les évènements de base et les options** présentent une importance marginale, confirmant leur faible contribution au processus décisionnel.

L'analyse de ces importances permet non seulement de renforcer la confiance dans le modèle, mais également d'apporter un éclairage métier : elle souligne que les prédictions s'appuient principalement sur des variables cohérentes avec la logique assurantielle et les déterminants identifiés dans la littérature. Cela constitue une étape essentielle pour relier la

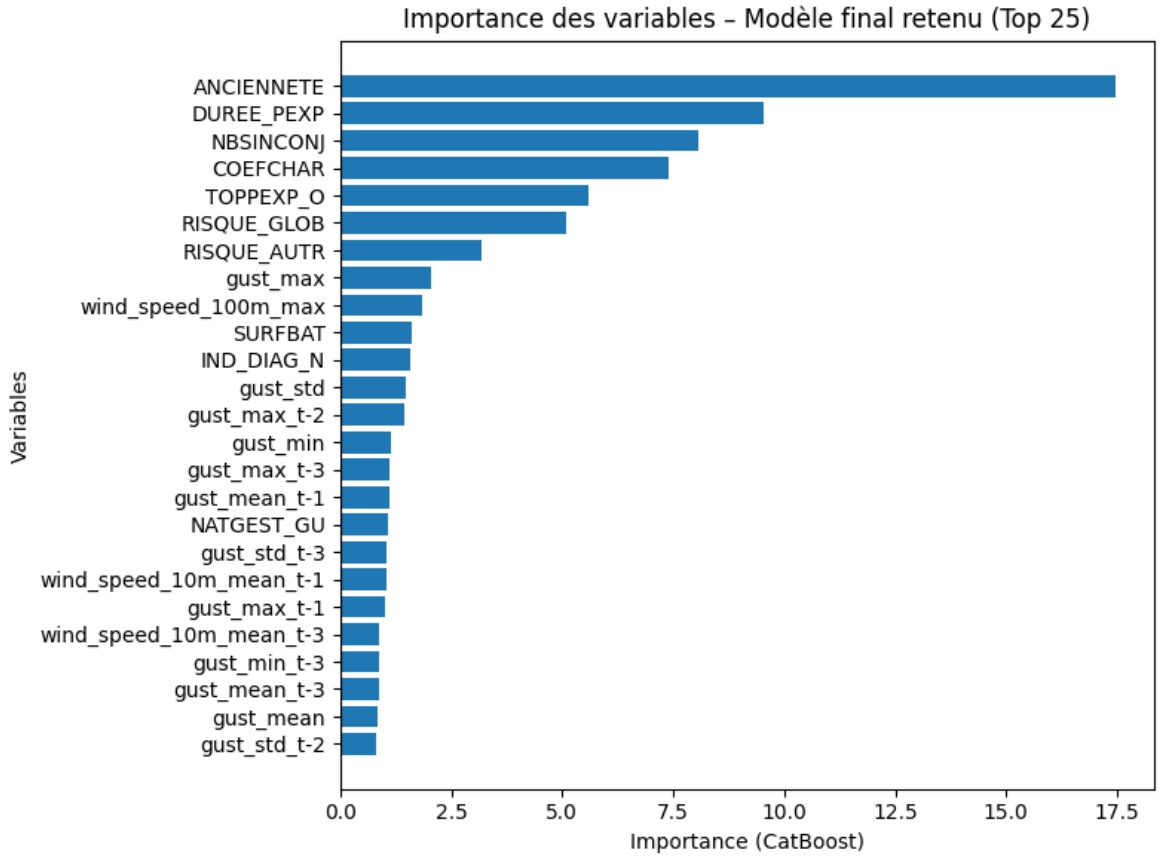

FIGURE 3.4 – Importance des variables selon CatBoost (Top 25).

performance du modèle à une compréhension opérationnelle exploitable.

3.2.5 Défis rencontrés et perspectives d'améliorations

Malgré l'intérêt des résultats obtenus, ce travail comporte plusieurs limites méthodologiques et pratiques qu'il convient de souligner. Ces limites ouvrent également la voie à des pistes d'amélioration et de recherche future.

3.2.5.1 Limites

— **Qualité et disponibilité des données** : Les données assurantielles issues du portefeuille Multirisque Agricole (MA) présentent certaines limites : Certaines années manquaient de tables utiles ce qui nous a forcé à ne pas inclure certaines variables.

— **Appariement spatial et temporel** : L'appariement entre les grilles ERA5 et les communes INSEE repose sur une jointure spatiale qui introduit nécessairement une approximation. De plus, l'agrégation des données à une échelle mensuelle ou trimestrielle atténue la capacité du modèle à détecter les phénomènes extrêmes (rafales violentes, tempêtes localisées) qui sont pourtant les principaux déclencheurs de sinistres.

— **Simplifications méthodologiques** : l'utilisation d'une approche de séries temporelles statique peut ne pas permettre de capter des dépendances temporelles voire des patterns(si elles existent). Par ailleurs, l'étude n'a pas intégré de dimensions économiques complètes (coûts de transaction, gestion du risque de portefeuille, tarification des primes), car les données concernant les montants sur les contrats étaient donnés par image et nous avons retenu les dernières images.

— **Une meilleure intégration d'une dimension économique** : Etant à une maille contrat, il nous était difficile d'utiliser certaines variables avec une dimension économique lié au sinistre.

3.2.5.2 Améliorations possibles et pistes à exploiter

— **Amélioration de la qualité des données** : L'utilisation directe des tables "archives" qui contiennent l'historique des contrats faciliterait la tâche dans le sens où nous n'aurions pas à faire des jointures et agrégations qui étaient insupportables pour la RAM de l'ordinateur.

— **Granularité plus fine des données météo** : L'exploitation de données climatiques à une échelle journalière, voire horaire, permettrait de mieux capter l'occurrence de phénomènes extrêmes. L'intégration de modèles à plus haute résolution (ex. AROME) constituerait également une piste pertinente pour affiner les appariements (Toutefois cela devrait être accompagné d'initiatives pour une meilleure puissance computationnelle des ordinateurs)

— **Méthodes de modélisation avancées** : L'utilisation de modèles dynamiques (par exemple des modèles à mémoire longue comme LSTM, ou des approches bayésiennes hiérarchiques) pourrait permettre de mieux capter les dépendances temporelles et la variabilité intra-annuelle. Mais toutefois nous ne sommes pas sûr qu'il soit possible de détecter des patterns car à priori la météo d'aujourd'hui n'est pas sensé conditionner celles de demain.

— **Intégration de la dimension économique** : Les variables relatives à la dimension économique concerne les garanties, donc il faudrait utiliser une maille garantie-mois (par exemple) pour faciliter l'usage de telles variables.

— **Ouverture vers d'autres risques climatiques et d'autres pays** : Enfin, la méthodologie développée pour le vent pourrait être étendue à d'autres aléas climatiques pertinents pour l'agriculture (sécheresse, grêle, inondations). Nous pourrions utiliser cette même approche pour des filiales de Crédit Agricoles Assurances implentées

dans d’autres pays (Portugal, Pologne, Italie)

Chapitre 4

Autre tâche réalisée

En plus des deux sujets qui nous ont été confiés, nous avons eu l'occasion de travailler sur une mission d'audit intitulé : « Groupe CAA : lutte contre la fraude ». Il s'agit d'une mission sur la fonction conformité au sein du Groupe CAA. Cette mission a été pilotée par un chef de mission accompagné de quatre auditeurs généralistes. Pour cette mission, le périmètre de cette mission concerne le Groupe CAA, PACIFICA, BU CAAPE et la BU ERI. Il couvrira Le dispositif de lutte contre la fraude et analysera l'organisation, la gouvernance ainsi que le pilotage de la thématique par la Direction de la Conformité Groupe CAA, le dispositif de prévention et de dissuasion de la fraude (animation, actions de sensibilisation, veille), les dispositifs de lutte contre la fraude externe (outils de détection, traitement de la fraude) et de la fraude interne (processus de notes de frais, processus d'alertes), les risques de fraude liés au SI (gestion des habilitations, qualité des données, contrôles) et enfin, le dispositif de maitrise des risques liés à la fraude (contrôles permanents, cartographie des risques, collecte des incidents opérationnels, suivi du cadre d'appétence). La tâche que nous avons pu réaliser concernait la fraude interne en l'occurrence le processus de notes de frais. Nous avions pour objectif d'identifier les attestations sur l'honneur parmi les notes de frais et vérifier si les montants mentionnés sur les documents coïncident avec ceux remboursés. Nous avions à notre disposition des pièces jointes de plusieurs types (.pdf, .jpg, .docx, .png) qui représentaient les notes de frais. A l'aide d'un code python qui utilise l'OCR dont la bibliothèque Python « Pytesseract », nous avons pu parcourir les documents afin d'identifier les possibles attestations sur l'honneur. Cette recherche s'est faite en plusieurs étapes avec plusieurs fonctions Python :

-

1. **Extraction d'Images (extract_images_from_pdf)** : Cette étape est essentielle pour récupérer tout contenu visuel des fichiers PDF. Les images extraites seront ensuite soumises à l'OCR pour convertir le texte qu'elles contiennent.

2. **Traitement OCR (perform_ocr)** : Après avoir extrait les images, cette fonction prend en charge à la fois les fichiers PDF et les images. Elle commence par extraire le

texte directement des pages PDF. Ensuite, pour chaque image extraite, elle applique l'OCR pour obtenir le texte contenu dans ces images. Cela permet de s'assurer que tout le texte, qu'il soit dans des images ou dans le texte brut du PDF, est capturé.

3. **Traitement de Dossier (process_folder)** : Cette fonction est le point d'entrée qui gère l'ensemble du processus pour tous les fichiers d'un dossier. Elle liste les fichiers, détermine leur type (PDF ou image), et appelle la fonction perform_ocr pour chaque fichier. Cela crée un flux de travail automatisé, où chaque fichier est traité de manière uniforme.

4. **Identification de Mots Clés** : Une fois que le texte a été extrait de chaque fichier, cette étape utilise une liste de mots clés pour déterminer si le texte contient des références à des attestations sur l'honneur. Cela permet de filtrer et d'identifier rapidement les documents pertinents, ajoutant une valeur analytique au processus d'extraction.

Nous avons identifié par la suite 9 attestations sur l'honneur et rien à signaler comme possible fraude interne car les montants remboursés correspondaient à ceux renseignés sur les notes de frais.

Conclusion

Ce rapport d'alternance retrace une année riche au sein du pôle Data de la Direction de l'Audit des Assurances de Crédit Agricole Assurances. Cette expérience m'a permis de comprendre concrètement le rôle de l'audit interne dans la maîtrise des risques et de mesurer la place croissante que prend la donnée dans ce métier.

Mon intégration dans l'entreprise s'est déroulée dans de très bonnes conditions. Dès mon arrivée, j'ai bénéficié d'un accueil chaleureux de la part des équipes, ce qui m'a permis de trouver rapidement mes repères. L'accompagnement attentif et la disponibilité de mon tuteur, Monsieur Mustapha Benarbia, ont constitué un véritable soutien tout au long de l'année. Ses conseils et son suivi régulier ont grandement facilité mon évolution et m'ont permis de progresser avec confiance dans un environnement exigeant.

Les projets menés m'ont permis de mettre en pratique les compétences acquises au Master MoSEF, mais aussi de développer des aptitudes essentielles pour le monde professionnel : rigueur, autonomie, capacité à collaborer avec des profils variés et à restituer des résultats de manière claire et utile pour le métier.

Enfin, cette alternance m'a confirmé l'importance stratégique de la data science dans le secteur de l'assurance. Les outils et méthodes mobilisés ne sont pas seulement techniques : ils apportent une réelle valeur ajoutée pour la compréhension des risques et la prise de décision.