Introduction

« Ta maison, Tongilianus, t'avait couté deux cent mille sesterces ; mais un de ces incendies, malheureusement trop communs dans Rome, te l'a fait perdre. Une souscription vient de te rétablir dans un million de sesterces, ne peut tu pas, Tongilianus, voyons, passer pour avoir mis toi-même le feu à ta maison ? »

« Marcus Valerius Martialis»

Cette épigramme du poète romain Martial fournit une preuve évidente que le phénomène de la fraude à l'assurance était déjà connu dans l'Empire romain au cours du premier siècle après J.-C. il ne s'agit pas d'un phénomène récent, elle est aussi ancienne que l'assurance elle-même. Quoique l'histoire des assurances soit jalonnée des manœuvres frauduleuses qui devenues de plus en plus sophistiquées, le risque de fraude n'a été mis en lumière que depuis quelques années suite aux cumuls des grosses pertes opérationnelles liées à la fraude. En effet, des incidents particulièrement étonnants ont impliqué d'énormes indemnisations indues de la part des compagnies d'assurance. Ces manœuvres, cependant, ne sont que la partie émergée de l'iceberg.
Ce qui fait que l'ampleur de ce risque est minimisée en raison de sa nature dissimulée.

En Tunisie, l'étude menée par la banque mondiale en collaboration avec le comité général des assurances a révélé que seulement 25% des manœuvres frauduleuses ont été détectées. Pour cette raison, il est difficile de déterminer avec un degré de certitude la valeur globale des pertes liées à la fraude dans l'assurance.

Autre fait remarquable, c'est que la fraude est incrustée dans l'ensemble des branches de l'assurance particulièrement la branche automobile. D'ailleurs, la sinistralité élevée de cette branche pèse lourdement sur sa rentabilité ce qui fait que l'un des défis posés aux assureurs est de pouvoir repérer les assurés fragilisant leurs résultats. À ce propos, la fraude liée aux dommages matériels représente, à elle seule, un enjeu financier redoutable pour les assureurs.

Dès lors que la directive européenne a intégré le risque de fraude comme un risque opérationnel, elle allait de pair avec une attention particulière portée sur des méthodes innovantes telle que la Machine Learning.

La complexité de ce phénomène justifie, bien entendu, l'engouement des assureurs pour les techniques d'apprentissage automatique en raison des avantages qui en découlent à l'instar de sa capacité à repérer des comportements frauduleux cachés et non identifiés par le gestionnaire et de bloquer l'indemnisation.

Dans ce cadre, ce présent mémoire va s'attacher à apporter des solutions opérationnelles à travers l'identification des déterminants qui permettraient de tirer des conclusions sur la probabilité d'une demande d'indemnisation frauduleuse. Précisément, nous souhaitons développer des modèles d'apprentissage statistique qui vont permettre in fine d'affecter les assurés à deux classes différentes celle des fraudeurs et l'autre des non-fraudeurs.

Nous tenterons dès lors d'appliquer deux méthodes d'apprentissage supervisé, nous les verrons, permettant d'apporter une aide à la décision. Cette méthode sera fondée sur l'extraction d'informations à partir de l'historique des données comprenant les caractéristiques de l'assuré luimême, du véhicule assuré, de la police souscrite et du sinistre.

Le premier chapitre de ce mémoire présente un aperçu général sur la notion de la fraude en assurance en faisant la lumière sur les éléments constitutifs de la fraude, ses typologies et les principaux enjeux de lutte contre celle-ci. Puis, le reste du chapitre sera réservé à cerner, au plus près, la fraude dans la branche automobile en abordant ses formes les plus répandues. Nous nous attarderons également sur la méthodologie de construction des modèles de classification. Dans le troisième chapitre, nous allons donner un aperçu théorique des méthodes traditionnellement utilisées dans la littérature pour des problèmes de classification à l'instar des arbres de décision et des forêts aléatoires.

Finalement, le dernier chapitre sera consacré à mettre en application les méthodes abordées dans le chapitre précèdent. Nous présenterons également quelques recommandations.

Chapitre I : Le risque de fraude dans l’assurance

La vision financière classique évoquait les risques du marché, les risques stratégiques ou les risques de réputation comme les facteurs essentiels de défaillance. Toutefois, les incidents découverts au début du 21e siècle au sein de certaines des plus prestigieuses sociétés multinationales démontrent bien la présence d’une autre source de pertes importantes qui pourrait provenir des risques opérationnels afférents aux éléments internes à l’entreprise, comme ses procédures, son personnel ou ses systèmes d’information ; ou des éléments extérieurs.

Certes, ces risques opérationnels recouvrent particulièrement la fraude qui s'étend bien au-delà des institutions financières d’autant que plusieurs d’autres organismes sont fortement exposés à ce risque dont les répercussions, loin d'être négligeables, sont susceptibles de les ébranler fortement.

Aujourd'hui, la fraude est considérée comme l'un des risques les plus importants auxquels une organisation est exposée, ayant un lien étroit avec les risques de marché ou de réputation.

Ce premier chapitre sera consacré donc à définir ce risque de fraude et ses éléments constitutifs au travers de la littérature en insistant notamment sur les apports théoriques sur les principales typologies du risque de fraude et le profil des fraudeurs. Nous présenterons ensuite les effets de la fraude sur l’assurance et les enjeux de la lutte contre celle-ci. Nous nous attacherons enfin aux principales sanctions encourues par les fraudeurs.

Section 1 : Le contexte de la fraude dans l'assurance

1.1. Les éléments constitutifs de la fraude à l'assurance

Tout le monde peut être une victime de fraude, qu'il s'agisse d'un individu ou d'une organisation, et les victimes peuvent être ciblées par des individus ou des groupes organisés d'individus. Il est difficile de définir la fraude en raison de l'éventail des comportements qui peuvent impliquer la malhonnêteté ou la tromperie. D'ailleurs, l'absence d'une définition opérationnelle commune de la fraude est l'une des limites persistantes à une quantification efficace de l'ampleur de ce phénomène.

Pour cette raison, plusieurs définitions du risque de fraude ont été proposées par quelques organismes à l'instar de l'IAA (Institute of Internal Auditors) et ACFE (Association of Certified Fraud Examiners) qui ont suggéré la définition suivante : « la fraude consiste à tromper délibérément autrui pour obtenir un bénéfice illégitime, ou pour contourner des obligations légales ou des règles de l'organisation. Un comportement frauduleux suppose donc un élément factuel et intentionnel ainsi qu'un procédé de dissimilation de l'agissement non autorisé »[1]

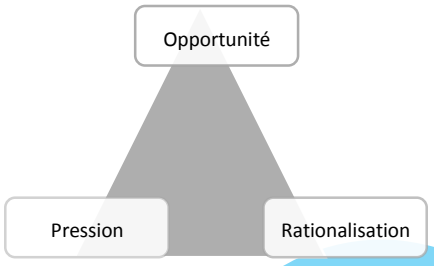
Le risque de fraude au sens large du terme est un acte visant à obtenir un bénéfice ou un avantage matériel ou moral indu par le biais des moyens illicites ou illégaux, il comprend notamment tout acte intentionnel ou volontaire visant à priver une autre personne de biens ou d'argent par la tromperie ou d'autres moyens illicites ou illégaux.

Dans le but d'expliquer la propension naturelle d'une personne à la fraude, le sociologue Donald Cressey a mis au point un modèle, en 1986, à partir des entretiens avec des personnes condamnées pour fraude, en essayant de tirer des points communs à chaque cas. Le modèle montre que la commission d'une fraude est le résultat d'une combinaison de trois éléments : la pression, l'opportunité et la rationalisation.

---

[1] Cahier de recherche IFACI, La fraude Comment mettre en place et renforcer un dispositif de lutte anti-fraude ? Définition issue du glossaire, cadre de référence international des pratiques professionnelles de l'audit interne, IIA, IFACI, janvier 2009

Figure 1:Triangle de la fraude de Cressey (1950)

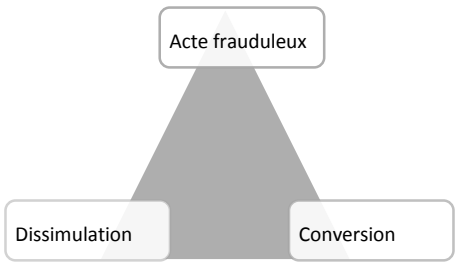


Tout d'abord, le fraudeur doit être soumis à des pressions financières (train de vie élevé, endettement à éponger...) ou à des vices (toxicomanie, addiction aux jeux…) qu'il ne peut pas divulguer à son entourage et pour lesquels il ne peut pas demander d'aide. Deuxièmement, il doit repérer une occasion pour résoudre ses difficultés financières sans pour autant que ces pratiques soient décelées.

Enfin, le fraudeur rationalise l'acte qu'il a commis parce qu'il ne se voit pas comme un criminel. De cette façon il développe une justification qui rend son acte frauduleux légitime tout en restant dans sa zone de confort moral.

Le modèle mis par Cressey (1950) est retenu dans des nombreuses études pour comprendre les éléments de fraude ou pour justifier les actes frauduleux. Cependant, un nouveau modèle a été récemment développé pour comprendre les actes de fraudeurs, il s'agit de triangle de l'acte frauduleux (Kranacher et al., 2011 ; Dorminey et al., 2012) créé dans le but de se concentrer non pas aux individus mais aux processus de la fraude. De ce fait, les trois facettes de l'acte frauduleux désormais les suivantes :

Figure 2: Triangle de l'acte frauduleux



- L’acte frauduleux, un élément factuel, qui implique le développement d'une certaine méthodologie et de la mettre en œuvre tels que le détournement de biens, vol.
- La dissimulation où le fraudeur essaye de cacher la fraude
- La conversion qui représente un processus servant à dissimuler la provenance illicite des bénéfices

L'intérêt majeur de ce modèle consiste à mieux appréhender les outils de lutte contre la fraude tandis que le modèle traditionnel permet de repérer les fraudeurs.

Ce risque de fraude touche un bon nombre d’entreprises et de secteurs d’activité dont celui de l’assurance qui n’échappe pas à ce problème tant en interne qu’en externe. Cependant, l'assurance est un secteur assez particulier du point de vue du mode de fonctionnement, ce qui implique une exposition importante en termes de volume, de fréquence et de variété de sources de fraude.

1.2. Les principales typologies du risque de fraude et le profil des fraudeurs

(Dedene, Stijn Viaene et Guido, 2004) ont analysé l'essence et les types de la fraude à l’assurance. Ils ont divisé les types de fraude en trois grandes catégories, à savoir la souscription par rapport aux sinistres, interne et externe et « soft fraud » par rapport au « Hard fraud ». Comme il ressort des lignes précédentes, étant donné que les cas de fraude sont tellement nombreux et variés nous avons opté pour une distinction basée sur la nature de l’opération, sur le type des fraudeurs et sur le mode opératoire :

Le risque de fraude peut provenir de sources tant internes qu'externes à l'organisation ; la fraude externe est un acte frauduleux commis à l'encontre de la compagnie d’assurance par des personnes ou des entités à l'extérieure de l'organisme et qui sont aussi diverses que les suivantes : assurés , intermédiaires d'assurance (les courtiers, bancassurance, poste assurance…) , prestataires de soins médicaux (médecins, opticiens, pharmaciens…), fournisseurs, prestataires de services ou tout autre tiers. Pour sa part, la fraude interne, par définition, correspond à un fait illicite délibéré volontairement par les dirigeants ou les membres du personnel au sens large dans le but d'obtenir un avantage particulier ou personnel.

Au-delà de la distinction entre fraude interne et fraude externe, l'on peut distinguer entre les différents types de fraudeurs qui peuvent être présents dans la plupart des milieux professionnels et sociaux. Il est toutefois impossible de dresser un tableau objectif du fraudeur, comme le note Michel Wright,2005) : « si la question du pourquoi frauder est souvent claire et se formalise avant tout par des motivations économiques à court ou moyen terme, il est nettement plus complexe pour les organisations de déterminer « comment » les fraudeurs s'y prennent ». Donc, pour simplifier, nous allons classer les fraudeurs en trois catégories selon leur intention et leur degré d'organisation. En effet, les types d'acteurs qui sont susceptibles de réaliser des fraudes sont classés par (Clarke,1989) comme suit : "l'opportuniste", "l'amateur" et "le professionnel" ;

L'opportuniste profite d'une perte réelle et couverte par la police d'assurance pour commettre une fraude, par exemple, en réclamant à côté des pertes réelles des objets non volés lors d'un vol ou en demandant à un fournisseur d'établir un devis d'un montant supérieur aux dégâts réellement subis ou d'établir une fausse facture afin de récupérer la franchise. L'amateur peut commencer par commettre une fraude opportuniste, puis franchir une étape supplémentaire... Cependant, le caractère « amateur » de ces tentatives facilite leur détection par l'assureur. Des dizaines d'autres cas de fraude peuvent être inclus dans cette liste, car ce type de fraudeurs est le plus fréquent.

Le professionnel, sans doute le type de fraudeur le plus dangereux, se livre à des fraudes systématiques, tant à titre individuel que dans le cadre d'un réseau organisé. Il possède ainsi des connaissances pointues sur le domaine assurantiel et qu'il les utilise pour améliorer sa situation financière d'une façon abusive. De nombreuses preuves empiriques indiquent que ce type de fraudeurs sont présents fréquemment dans la branche de l'assurance automobile.

Il est toutefois difficile de distinguer un fraudeur d'un véritable demandeur sur la base des seules caractéristiques personnelles et sociales. Plusieurs recherches menées pour classer un échantillon de demandes frauduleuses ont montré que les fraudeurs présentent des caractéristiques qui les rendent pour la plupart impossibles à distinguer du véritable demandeur (Dodd, 1998). Le fait que les fraudeurs professionnels n'utilisent généralement pas d'identités authentiques aggrave cette difficulté.

Outres ces distinctions citées dans les lignes précédentes, la variété des pratiques de fraude résulte aussi de la diversité des modus operandi des fraudeurs qui s'explique notamment par le

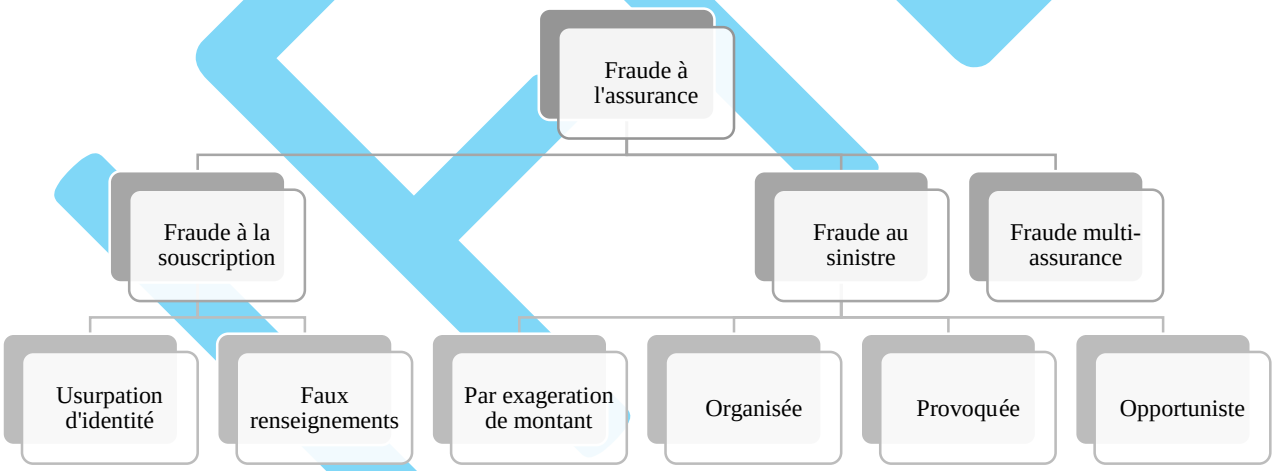
fait que le phénomène de fraude est désormais polymorphe et peut surgir à tout moment et à toutes les étapes d'exécution du contrat.

Le lien contractuel, qui unit une compagnie d'assurance à un souscripteur et qui les engage envers certaines obligations, se résume à des échanges monétaires : une prestation généralement pécuniaire en cas de réalisation d'un sinistre, moyennant le paiement d'une prime ou cotisation.

Les montants de ces échanges dépendent clairement, des déclarations de l'assuré, que ce soit à la signature du contrat pour fixer le montant de la prime ou durant la vie du contrat. Clairement, le secteur de l'assurance compte un nombre important de fraudeurs qui tentent de réaliser des gains illicites pour manipuler les montants de ces échanges, ce qui entraine l'apparition de plusieurs types de fraude.

Pour illustrer cette notion de « polymorphe », nous pourrions distinguer principalement trois phases: la fraude à la souscription, la fraude au sinistre et la fraude de multi-assurance :

Figure 3:Les types de fraudes à l'assurance



Source : Élaboré par moi-même.

Dans le processus de souscription sont réunies toutes les tentatives pour payer des primes moins chères. En effet, un fraudeur donnera délibérément de fausses informations afin de faire baisser la prime tels que l'usurpation d'identité, la fausse déclaration du risque, capitaux sous-évalués, situation du risque erronée, fausse déclaration de la sinistralité passée, mauvaise description du risque, minoration de la gravité du risque. Ce type de fraude est relativement facile à commettre et

à détecter grâce à l'intervention de l'expertise pour évaluer de visu les dommages subis suite à un sinistre, ce qui peut aider à détecter la fraude.

La fraude aux sinistres est souvent regroupée en deux grandes catégories : la fraude organisée (ou planifiée) et la fraude opportuniste [Dionne & Gagne, 2001]. La fraude organisée est préméditée et peut impliquer la mise en scène d'un accident ou d'un événement dans le but de réclamer une indemnité d'assurance. La fraude opportuniste se produit après un incident légitime et le niveau des dommages ou des blessures est délibérément exagéré afin d'avoir une indemnité plus élevée, ou réclamer pour quelque chose qui n'est pas couverte par la police (Weisberg et Derrig, 1993). Par ailleurs, la détection de la fraude est rendue plus ardue car les fraudeurs savent désormais s'adapter. La fraude n'est plus un phénomène statique, les tendances changent au cours du temps (au rythme des cycles saisonniers et économiques) et de nouveaux types de fraudes inconnues auparavant émergent particulièrement dans le domaine de la fraude organisée.

Quant à la fraude de la multi-assurance, elle consiste à souscrire plusieurs polices d'assurance pour couvrir un bien déjà assuré contre le même risque afin de toucher des prestations multiples. Ce type de fraude est moins fréquent et peu détectable par les assureurs.

1.3. Les effets de la fraude sur l'assurance

L'IFACI soulignait que « La fraude crée un préjudice financier, matériel, moral ou d'image à l'organisation ou à ses clients, partenaires ou autres interlocuteurs », elle peut entraîner notamment des répercussions néfastes, pour l'assureur comme pour l'assuré.

Effectivement, pour l'assureur, les fraudes commises à son encontre peuvent altérer directement sa situation financière en lui causant un manque à gagner considérable. Afin d'apprécier le poids de ce risque, une étude commandée par la fédération des sociétés d'assurance européenne « Insurance Europe » a évalué les fraudes détectées et non détectées à 10 % du coût des sinistres liés aux demandes de remboursement en Europe. Ce chiffre varie selon les pays et les catégories d'assurance en raison d'un certain nombre de facteurs, tels que le fonctionnement du marché ou la prévalence locale d'un type d'assurance.

Tableau 1 : L'effet de la fraude par pays et par branche

Pays	branches	Montant
Allemagne	Toutes les branches	10% du coût des sinistres
Australie	Toutes les branches	10% du coût des sinistres
Canada	Toutes les branches	10% à 15% du coût des sinistres
Espagne	Automobile	22% du coût des sinistres
Grande Bretagne	Branche des particuliers	7% du coût des sinistres
Scandinavie	Toutes les branches	5% à 10% du coût des sinistres
États-Unis	Automobile	11% à 15% du coût des sinistres
États-Unis	Toutes les branches	10% du coût des sinistres

Source : Insurance Europe,2013

En ce qui concerne la Tunisie, « l'impact de fraude est considéré par tous comme extrêmement important, même si, par construction, les statistiques disponibles sont rares. » (La banque mondiale, 2015). En effet, selon les statistiques publiées par la Fédération Tunisienne des Sociétés d'Assurance (FTUSA), la fraude à l'assurance coûte aux assureurs 150 millions de dinars par an, soit presque 10% des primes du secteur.

Les conséquences directes de la fraude se situent également dans les primes payées par les assurés qui devront payer plus (Goel, 2014). Ceci s'explique par le fait que l'assurance est basée sur le mécanisme de partage des risques qui suppose que les coûts des sinistres soient partagés entre les assurés, de manière à ce qu'ils se compensent mutuellement. Tout acte frauduleux peut donc perturber gravement l'équilibre du principe de mutualisation des risques d'une manière à ce qu'une bonne partie des assurés honnêtes payent la malhonnêteté des fraudeurs par le biais d'une augmentation de la tarification des produits d'assurance. La compétitivité des compagnies d'assurance sera compromise en l'occurrence.

Les effets vont au-delà des pertes directes, ils nuisent également à la réputation de l'entreprise en générant une mauvaise publicité pour l'image de l'entreprise ce qui peut éventuellement occasionner des manques à gagner indirects

1.4. Les principaux enjeux de la lutte contre la fraude

Pour les entreprises d'assurance, les enjeux financiers, commerciaux et marketing de la lutte contre la fraude sont sans doute considérables et nécessitent un engagement au plus haut niveau. Elles devraient, par conséquent, tout mettre en œuvre pour prévenir les effets significatifs de la fraude étudiés aux lignes précédentes.

La bonne réputation de l'organisme d'assurance, le maintien de confiance accordée par toutes les parties prenantes, l'amélioration de la qualité du service et l'accélération de la cadence des règlements pour les dossiers non douteux, sont tous des enjeux commerciaux majeurs de la lutte contre la fraude. L'évaluation adéquate de ce risque contribue à l'atteinte de tels objectifs et plus généralement à la protection de l'image et de l'intégrité de l'organisme.

les enjeux financiers sont également conséquents, comme le souligne (Éric Lamarque[2],2009) : « Si les autorités de régulation internationale se sont saisies du problème, c'est que le coût financier est apparu de plus en plus important et est de nature à affecter significativement la rentabilité et les fonds propres des établissements ». Ceci s'illustre par les dépenses énormes qui sont, pour une grande part, répercutées sur les assurés, via les primes payées. Pour remédier à ce problème, les assureurs doivent songer à limiter ce risque pour optimiser la tarification et baisser les charges de sinistres.

Alors, face à tous les préjudices dévastateurs portés au secteur assurantiel, il se pose l'importante question de savoir si les organismes d'assurance et les autorités ont engagé des actions civiles et pénales pour sanctionner les comportements frauduleux.

1.5. Les sanctions encourus par les fraudeurs

Eu égard au fléau de la fraude à l'assurance, les initiatives de lutte contre la fraude ont multiplié avec un foisonnement législatif qui reflète la prise de conscience de l'étendu de ce risque. Cette prise de conscience a débouché sur un consensus quant aux différents sanctions encourues par les fraudeurs. Il s'agit des textes de loi prélevés notamment du code civil, code des obligations et des contrats, code des assurances, code pénal, …

[2] Éric Lamarque; Dans Revue française de gestion 2009

Le tableau ci-après résume les différentes sanctions prévues par le code des assurances encourues par les fraudeurs, selon la réglementation tunisienne :

Comportements frauduleux	Les sanctions encourues	Textes applicables
Réticence d'information ou fausse déclaration avant et après sinistre	Nullité du contrat ou réduction de sinistre	Article 8 du code des assurances
Souscription d'assurances multiples cumulative	Nullité /plafonnement de l'indemnisation à la chose assuré et application du principe de proportionnalité entre les polices d'assurance	Article 18 du code des assurances
Surévaluation du bien assuré	Application de la règle proportionnelle si elle est prévue dans le contrat	Article 17 du code des assurances
Sinistres déclarés trop tardivement	Déchéance pour déclaration tardive	Article 7 du code des assurances
Adhèrent fictif dans le cadre d'assurance groupe	Exclusion de l'adhèrent du contrat	Article 32 du code des assurances
Suicide inconscient de l'assuré	Nullité du contrat/charge de la preuve de suicide sur l'assureur, celle de l'inconscience à la charge du bénéficiaire.	Article 37 du code des assurances

Source : Élaboré par moi-même

Il n'existe pas une définition juridique exacte de la notion de fraude tant en droit pénal qu'en droit civil, dans la mesure où elle recouvre une variété de manœuvres, ce qui suppose l'absence de sanctions spécifiques pour la fraude à l'assurance. Il convient alors de se référer aux articles du code pénal qui ont un effet punitif sur le fraudeur qui commet l'extorsion, le faux et l'usage de faux, l'abus de confiance et l'escroquerie. L'escroquerie est l'infraction la plus récurrente dans les cas de fraude à l'assurance, elle se définit par la volonté d'obtenir une indemnité indue ou supérieure à celle qui est due.

Le tableau ci-après contient les principales infractions pénales et les peines encourues par le fraudeur :

Exemples d'infractions	**définitions**	**Les sanctions encourues**	**Les textes applicables**

Escroquerie	C'est le fait par l'usage d'un faux nom ou d'une fausse qualité soit par l'abus d'une qualité vraie, soit par l'emploi d'une manœuvre frauduleuse	Sanctions de 5 ans d'emprisonnement	Article 291 et suivants du code pénal
Vol	La soustraction frauduleuse de la chose d'autrui	Sanctions de 5 à 10 ans d'emprisonnement	Article 258 et suivants du code pénal
Extorsion	Le fait d'obtenir par violence, menace de violences ou contrainte, soit une signature, un engagement ou une renonciation, la révélation d'un secret, soit le remise de fonds ou des valeurs	20 ans d'emprisonnement	Article 283 et suivants du code pénal
Abus de confiance	Le fait par une personne de détourner au préjudice d'autrui des fonds, des valeurs ou un bien quelconque qui lui ont été remis à charge de les rendre ou d'en faire un usage déterminé.	3 ans d'emprisonnement	Article 297 et suivants du code pénal

Source : Élaboré par moi-même

1.6. La place de la fraude dans l'assurance automobile

La branche d'assurance automobile, qui a servi d'appui à ce mémoire, est la branche la plus touchée par la fraude, surtout en cette période de crise où la plupart des assurés cherchent à obtenir le meilleur gain financier possible sur les polices d'assurance automobile. Ce point est d'ailleurs confirmé par le rapport de l'institut français de l'audit et de contrôle interne (IFACI) qui souligne que : « la crise économique, environnement déontologique fragile, failles des systèmes d'information sont autant de facteurs encourageants le passage à l'acte avec des impacts non négligeables pour les organisations ». Ainsi, les fraudeurs ont profité du fait que les assureurs ont du mal à détecter les actes frauduleux, et bien qu'ils soient capables de les identifier, le caractère frauduleux n'est pas toujours prouvé en raison du manque de preuves suffisantes, ce qui entraine la rareté des dépôts de plaintes.

Les experts supposent que la fraude aux sinistres coûte beaucoup plus chère au secteur que la fraude à la souscription. Alors, nous allons focaliser principalement sur les fraudes à la déclaration de sinistre. En outre, bien que certains assureurs et intermédiaires aient commis des fraudes, nous ne traiterons que les fraudes commises par les assurés ou par ceux qui agissent en leur nom.

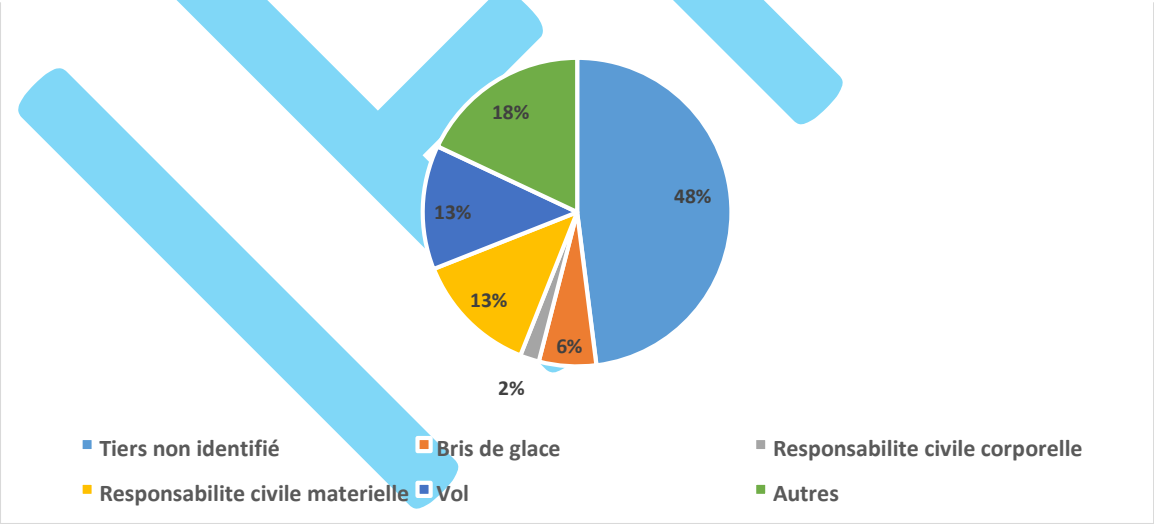
En général, la fraude à la déclaration de sinistre est une forme d'aléa moral ex-post qui implique la manipulation des informations afin de faire jouer indûment sa garantie d'assurance : elle est susceptible alors de prendre la forme d'une déclaration de sinistre mensongère ; songeons par

exemple aux fraudes commises au titre de la garantie vol, qui sont les plus fréquentes, elles prennent plusieurs formes : le preneur d'assurance prétend que sa voiture a été volée après un accident afin d'obtenir une indemnisation pour les dommages matériels, il fait disparaître son véhicule en mauvais état pour être remboursé, ou il prétend que sa voiture a été volée à l'étranger après l'avoir vendue. L'assuré peut également déplacer son véhicule pour simuler son vol et être remboursé. La déclaration mensongère porte également sur un sinistre qui n'a jamais eu lieu comme la destruction du véhicule ou un incendie…

Signalons encore l'exagération des dommages subis lors d'un accident réel. En effet, les réclamations gonflées sont probablement beaucoup plus fréquentes que les réclamations fictives ; les assureurs peuvent payer des sinistres exagérés simplement parce que le temps et l'argent consacrés au refus d'une demande d'indemnisation gonflée dépassent le simple paiement de la demande d'indemnisation gonflée.

Dans le même contexte, une étude commandée par ALFA à travers une base de données collectée, en 2015, auprès de ses adhérents, afin de connaitre la répartition des dossiers frauduleux selon le type de la garantie, se présente comme suit (figure) :

Figure 4: Répartition des dossiers frauduleux en assurance auto en 2015



Source : ALFA,2015

Parmi les fraudes les plus répandues, arrivent en tête les accidents fictifs causés par un tiers non identifié, notamment les faux accidents de stationnement avec un pourcentage de 48%.

Juste derrière, on trouve le vol qui est également une forme de fraude très fréquente où l'assuré déplace son véhicule pour faire croire à un vol. Ainsi, on remarque un pourcentage élevé de réclamations frauduleuses au titre de la garantie responsabilité civile matérielle lorsque le fraudeur fait porter la responsabilité d'un sinistre matériel à un tiers (13%). Enfin, viennent les déclarations découlant de la garantie bris de glace et la responsabilité civile corporelle avec des pourcentages successifs de 6% et 2%.

Section 2 : Détecter la fraude en assurance

3.1. Les organismes de lutte contre la fraude

Le phénomène de la fraude a pris des proportions alarmantes pour l'assureur ; il est en effet primordial de déployer des moyens de détection, de prévention et de répression en s'appuyant sur des mécanismes efficaces conçus pour fournir une orientation sur les outils opérationnels qui traquent la fraude. La coopération entre les acteurs constitue une des forces de lutte contre la fraude. Sur ce point-là, une multiplicité d'organismes œuvre conjointement pour combattre ce risque tels que :

- L'Agence pour la lutte contre la fraude à l'assurance, l'ALFA : un organisme à but non lucratif qui vise à agir contre toutes les formes de fraude au moyen d'un système de centralisation des échanges et de partage d'expériences inter compagnies aux moyens des services d'enquête ou des actions d'étude et de réflexion, de formation et de sensibilisation.
- L'Association of Certified Fraud Examiners, ou ACFE, est la plus grande organisation antifraude au monde et le premier fournisseur de formation et d'éducation en matière de lutte contre la fraude. Sa mission est de réduire l'incidence de la fraude et d'aider les membres à détecter et à dissuader la fraude.

Les organismes de l'audit interne ont devenu quant à eux largement impliqués dans la gestion du risque de fraude, ils accordent une priorité plus importante à ce risque en raison de son gravité :

- Le Committee of Sponsoring Organizations of the Treadway Commission, COSO, prône la bonne application des pratiques de gestion des risques de fraude et la construction des dispositifs antifraudes performants dans sa nouvelle version du référentiel de contrôle interne

COSO 2013 qui comporte des réflexions sur la façon d'établir les politiques de contrôle et de prévention de fraude.

- L'Institut Français des Auditeurs et Contrôleurs Internes, l'IFACI ,qui est le chapitre français de l'*Institute of Internal Auditors* (IIA), recommande l'adoption des bonnes pratiques pour l'évaluation des risques à travers la contribution de professionnels de l'audit et du contrôle interne. Il intervient notamment pour diffuser des connaissances en matière d'audit interne en proposant des feuilles de route indiquant la meilleure orientation à suivre dans la gestion des différents risques. Il propose également des « business cases » sous forme des fiches de risques contenant tous les scenarios possibles qui décrivent les schémas frauduleux comme le démontre le tableau au-dessous :

Tableau : Fiche de risque

Description de la fraude :	Descrire le schéma de la fraude	
Type de la fraude :		
	Interne	
	Externe	
	Collusion	
Catégorie de risque :		
	Revenir au référentiel de la compagnie	
Macro processus impacté :		
	Commercialisation / Souscription / Adhésion	
	Gestion de la vie du contrat	
	Gestion des prestations / sinistres	
	Gestion des apporteurs / délégataires	
Moyens de maîtrise du risque :		
Prévention	Recommander les actions correctrices à mettre en œuvre pour améliorer le dispositif de contrôle interne	
Détection	Recommander les actions de détection à mettre en œuvre pour systématiser la detection de ce risque	

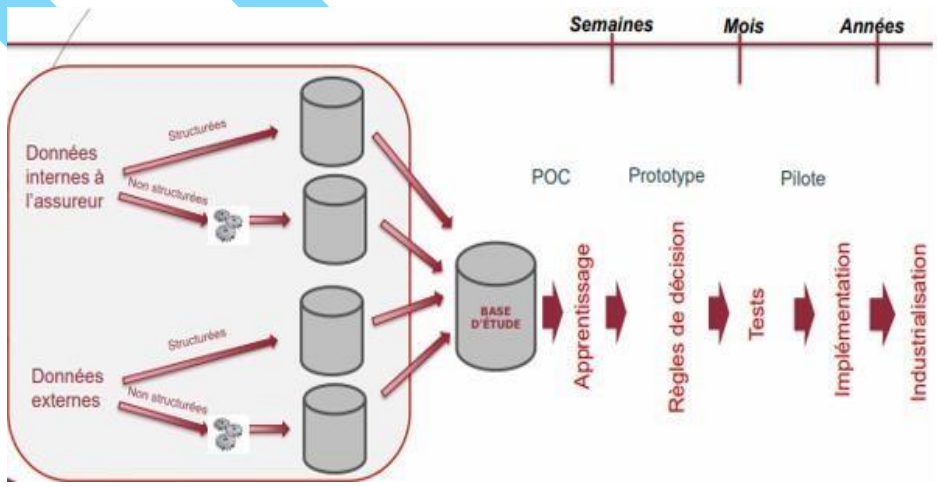
Source : Cahier de la recherche de l'IFACI

3.2. L’automatisation de la détection

La détection et la prévention de la fraude consistaient auparavant à examiner et exploiter les données historiques - en passant par les dossiers physiques pour repérer les comportements potentiellement frauduleux dans le passé. Dans le monde technologique actuel, la détection et la prévention de la fraude consistent à arrêter la fraude au moment ou même avant qu'elle ne se produise. Les solutions actuelles de gestion automatisée de la fraude visent à identifier des comportements inhabituels compatibles avec une activité frauduleuse.

Nous avons vu précédemment que les cas de fraudes sont mêlés à plusieurs formes qui évoluent et s’intensifient de plus en plus. Pour cette raison, le suivi d’une démarche d’industrialisation des dispositifs anti-fraude devient un des enjeux majeurs pour les compagnies d’assurance. Cette démarche permet de passer d’un dispositif de détection basé sur des règles métier classiques à un autre en partie automatisé de façon à systématiser le contrôle et améliorer les processus traditionnels. À ce titre, la démarche suivie pour industrialiser la détection se résume comme suit :

Figure 5: Démarche générique d’industrialisation



Source : Nicolas MARESCAUX et Marc RAYMOND (institut des actuaires)

En effet, faire appel à un système d'apprentissage automatique à base d'outils avancés d'analyse de données de type machine-learning, deep-learning, cognitive computing, web scraping, classification supervisée et non supervisée, etc. en sus des systèmes de détection artisanaux et manuels est de redoutable efficacité. Cela permet de 'faire parler' les données pour repérer au plus vite les profils frauduleux avant leur indemnisation. Bien évidemment, l'industrialisation ne peut pas être réduite uniquement à la mise en place des outils de détection informatisés...on peut s'appuyer également sur des outils de travail performants tel que la cartographie des risques utilisée pour identifier et hiérarchiser les différents risques en fonction des critères de gravite et d'intensité.

3.3. La méthodologie adoptée

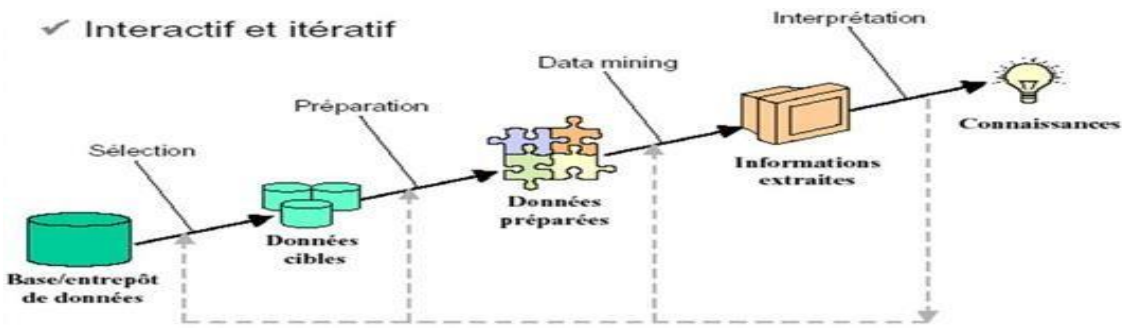
En raison de l'importance du problème de la détection des fraudes, on peut trouver plusieurs ouvrages qui traitent ce sujet. En effet, il existe des études qui cherchent à identifier les classes de fraude et à créer des méthodologies qui permettent leur classification. La fraude présente plusieurs spécificités et qu'en conséquence ces méthodologies différèrent en fonction des particularités de chaque type de fraude. Dans ce contexte, l'objectif principal de ce travail consiste à adopter une méthodologie appropriée pour mener une série de recherches ordonnée.

3.2.1. Le processus de modélisation de données

La littérature liée à la détection de fraude regroupe un certain nombre de travaux qui se penchent particulièrement sur une méthodologie basée sur un processus traditionnel de construction des modèles standards de classification qui comporte des étapes utilisées pour guider la démarche de détection des fraudes. D'ailleurs, plusieurs travaux ont été menés en ce sens à l'instar de la méthodologie connue par l'acronyme KDD : « Knowledge Discovery in Databases » développée par Oussema Fayyad en 1996 dans le but d'aider les analystes à appliquer les techniques d'apprentissage statistique et de visualisation des données, de sélection et de transformation des

variables prédictives les plus importantes, de modélisation des variables pour prédire les résultats et de confirmation de la précision du model, comme l'indique la figure au-dessous :

Figure 6: La méthode de KDD illustrée



La conduite de notre étude sera penchée sur les travaux de (Fayyad et al., 1996) relatifs à un processus d'extraction de connaissances à partir de la base de données. Dans cette lignée, nous allons suivre cette démarche méthodologique (KDD) pour réaliser notre modèle de détection de fraude. Cette démarche s'organise en 5 étapes :

Une première étape, **la sélection des données**, consiste à interroger les données pour relever des anomalies ou des informations atypiques sur des dossiers a priori considérés comme suspicieux, cette étape nécessite une analyse profonde de la base de données et une sélection des dossiers pouvant contenir des informations pertinentes pour la détection des fraudes, effectuée en collaboration avec des experts dans le domaine.

Dans le **prétraitement**, on applique des méthodes de traitement et de nettoyage des données afin d'éliminer les incohérences (corriger les doublons et les erreurs de saisie) et ce, à travers des graphiques et d'histogrammes pour identifier des tendances en matière de fraude.
Si une tendance est trouvée, l'on peut estimer que les attributs sélectionnés peuvent nous aider à identifier les dossiers potentiellement frauduleux.

La **transformation des données** est une étape liée directement au choix du modèle et inclut la transformation et le codage des données collectées à des informations utiles dans des formats appropriés à ce modèle.

Dans l'étape de **Data Mining**, nous analysons les données collectées pour trouver les causes, les comportements et les relations qui permettent d'expliquer un phénomène. D'ailleurs, le Data

Mining est utilisé fréquemment pour résoudre des problèmes décisionnels à travers des modèles polytomiques à plusieurs modalités ou des modèles dichotomiques à deux modalités. Dans notre étude, les modèles de classification sont plus intéressants, car ils permettent une meilleure compréhension du phénomène de fraude.

Enfin, viendra l'étape **d'évaluation et d'interprétation** des modèles déterminés pendant la phase précédente, elle inclut également la possibilité de retourner aux étapes précédentes pour corriger les éventuelles anomalies. Pour évaluer un modèle de classification, nous procédons à la répartition des inputs en deux groupes : un groupe d'apprentissage, un deuxième pour les tests.

Chapitre II : Théorie sur les techniques d'apprentissage automatique

Le data mining est né d'un besoin de tracer des liens pertinents dans des bases de données contenant un très grand nombre d'observations ce qui va permettre d'analyser et modéliser les données pour trouver les relations qui aboutissent à l'identification des cas de fraude potentiels. Dans ce but, l'on trouve plusieurs approches possibles de data mining qui vont servir comme un allié précieux d'aide à la détection des fraudes. Le présent chapitre est développé à ce propos, nous allons nous restreindre à quelques méthodes de Machine Learning dont le but ultime est de réaliser un classement qui qualifie le dossier de frauduleux ou non.

Ce chapitre est structuré en deux sections présentant respectivement la technique d'apprentissage automatique par arbre de décision et les méthodes d'apprentissage ensemblistes.

Section 1 : Prédiction par arbre de décision

1.1. Le choix du modèle

Le phénomène à expliquer est fortement lié au modèle choisi qui dépend quant à lui de l'objectif de l'étude ainsi que du type et de la quantité de données utilisée.

1.1.1. L'apprentissage supervisé ou non supervisé

L'apprentissage supervisé consiste à faire apprendre à un algorithme une fonction de prédiction pour laquelle la valeur cible est définie à l'avance. Cet algorithme consiste à partitionner la base de données en deux sous-ensembles distincts : une base d'apprentissage utilisée en amont pour l'entrainement du modèle et l'autre base test qui va permettre d'évaluer la qualité des prédictions de ce modèle.

Dans le cas inverse, où les observations ne sont pas classées au départ, on choisira des méthodes non supervisées qui visent à regrouper les observations en mesurant la similarité entre eux (Clustering), elles seront donc classées aux catégories les plus homogènes possibles.
En effet, les données recueillies auprès de la MAE englobent les dossiers labellisés de présence ou non d'un cas de fraude. De ce fait, une modélisation par des méthodes de classification supervisées est plus appropriée pour développer un modèle interne partiel de détection de fraude en assurance automobile.
Les plus connus des apprentissages supervisés sont les arbres de décisions et les forêts aléatoires qui seront présentés par la suite.

1.1.2. L'apprentissage paramétrique ou non paramétrique

Le modèle paramétrique suppose des hypothèses sur la forme de distribution d'entrainement. D'ailleurs, la variable expliquée Y s'écrit en fonction des variables explicatives tel est le cas des modèles linéaires généralisés qui s'écrivent sous forme de $Y = aX_1 + bX_2 + c$ et dont le principe est d'estimer les coefficients de pondération a, b et c afin d'en prédire la valeur de la variable cible.

Quant aux modèles non paramétriques on ne fait pas des hypothèses en amont sur la fonction de prédiction (loi de probabilité inconnue). Ils se basent notamment sur l'expérience ce qui explique la nécessité l'exploitation d'une base de données importante. Le modèle sera choisi donc selon des données et de leur profondeur.

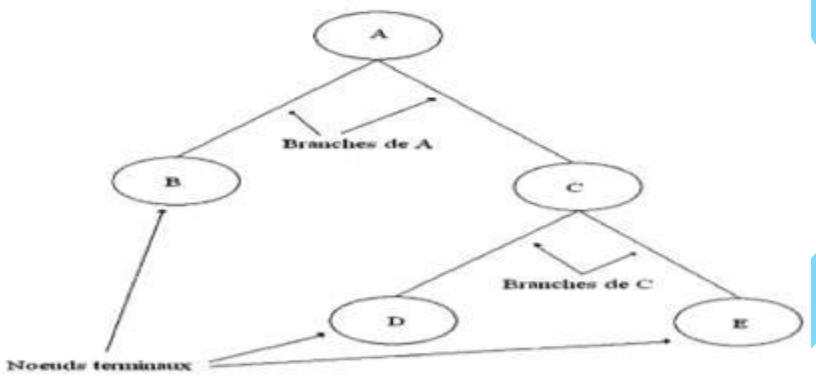
1.1.3. Définition

L'arbre de décision est un modèle d'apprentissage non paramétrique déjà ancien qui représente une suite de règles qui permettent généralement d'assister le gestionnaire dans la résolution des problèmes de classification ou de régression, il a une structure arborescente qui ressemble à un organigramme. Son fonctionnement repose sur des heuristiques construites selon des techniques d'apprentissage supervisé (Mezghani, 2015).

En effet, à partir d'un arbre de décision, on cherche à construire un modèle partant d'un ensemble d'observations relatives à des classes connues afin de trouver une description pour chacune de ces classes en se basant sur les similarités entre les observations. Après avoir construit ce classifieur,

il est possible alors d'extraire un ensemble de règles de classification qui seront utilisées pour classer de nouvelles observations dont la classe est inconnue. Le classement fait référence à un processus de partitionnement récursif de haut en bas selon le principe « diviser pour régner » en constituant des nœuds en arborescence ; chaque nœud de l'arbre représente le test d'un attribut, chaque branche qui est étiquetée avec les valeurs que peut prendre l'attribut choisi pour le nœud, représente le résultat de ce test, et chaque feuille désigne une classe ou une répartition de classes. Le nœud supérieur est le nœud racine.

Figure 7: Schéma d'un arbre de décision



1.2. Construction de l'arbre de décision par l'algorithme CART

Pour construire un arbre de décision, plusieurs algorithmes de Machine Learning dites de partitionnement récursif sont formalisés par plusieurs statisticiens tels que l'algorithme CART (Classification and Regression Tree) développé par Breiman et al. (1984) qui a trouvé son utilité dans un grand nombre d'activités d'aide à la décision automatisées et il s'est révélé prometteur particulièrement pour la prédiction des profils frauduleux à l'assurance et d'autres domaines de détection des fraudes.

Le type d'arbre inféré par CART est binaire et se base sur des partitions dichotomiques uni-variées des variables tant catégorielles que continues. Le principe général de construction d'un arbre de décision par CART consiste à séparer, d'une façon dyadique et récursive, un espace d'entrée dans lequel est défini un ensemble de données d'entraînement étiquetées (Xi, Yi). Cette séparation est faite à l'aide des tests binaires imbriqués sur les variables explicatives.

La construction d'un arbre de décision par l'algorithme CART commence par l'étape d'expansion qui se repartie quant à elle en trois étapes : il s'agit de décider en premier lieu si le nœud est terminal pour représenter les individus du jeu d'apprentissage. Pour ce faire, on doit respecter une des conditions suivantes: Tous les individus du jeu d'apprentissage appartiennent à la même classe ou tous les attributs ont été utilisés pour les nœuds précédents. Cette étape permet de déterminer si le nœud doit être étiqueté comme une feuille ou porter un test. La deuxième étape se produit lorsqu'on ne respecte pas les conditions de la première étape, elle consiste à trouver l'attribut qui représente le nœud de l'arbre ainsi que les valeurs de l'attribut associé au nœud. Toutefois, la question clé pour cette construction est de déterminer quel attribut on doit choisir pour représenter chaque nœud.

1.2.1. L'optimisation des nœuds par l'algorithme CART

Pour répondre à la question proposée à la section précédente, une méthode d'optimisation des nœuds a été conçue par l'algorithme CART pour sélectionner les tests appropriés sur les attributs appartenant à notre base d'apprentissage. Pour chacun de ces attributs, l'algorithme CART explore tous les tests possibles et choisit ensuite la coupure optimale qui maximise la réduction en impureté puis il compare les meilleures séparations pour repérer le meilleur d'entre eux. Le choix du meilleur split se fait par un critère de partition implémenté par l'algorithme CART connu sous le nom de l'indice de Gini

1.2.1.1. L'indice de Gini

L'algorithme commence par le nœud racine qui correspond à l'ensemble de la base d'apprentissage en produisant une série de tests binaires qui partitionnent cet ensemble en deux sous-ensembles disjoints, puis il compare les différents tests possibles selon un certain critère pour choisir celui qui minimise l'hétérogénéité dans chaque partition. Cette étape est itérée sur chacun des sousensembles jusqu'à atteindre le nœud terminal. Plus précisément, le critère de sélection des tests efficaces utilisé par CART est l'indice de Gini qui mesure le degré de désordre dans les différentes partitions. Les arbres CART ne génèrent que deux branches par nœud. Si un attribut a une variable discrète alors toutes les combinaisons possibles de groupes sont égales à (2n - 2) /2

(pour une variable discrète de valeur n), et en cas d'une variable continue[3], tous les points de coupure possibles sont testés (pour n enregistrements, au maximum n - 1 points de coupure sont testés.

L'indice de Gini est déterminé en déduisant la somme des carrés des probabilités de chaque classe de un , selon la formule suivante :

Pour une base d'apprentissage T qui contient des observations de m classes, Gini (T) sera défini comme suit :

$$\text{Gini}(T) = 1 - \sum_{j=1}^{m} \left(\frac{|T_j|}{|T|}\right)^2 \qquad = 1 - \sum_{1=1}^{m} p_i^2 ,$$

Où p_j : est la proportion des individus appartenant à la j[ème] classe pour un nœud particulier

L'indice de Gini varie entre les valeurs 0 et 1, où 0 exprime la pureté de la classification, c'est-à-dire que tous les éléments appartiennent à une classe déterminée (impureté nulle). Et 1 indique la répartition aléatoire des éléments entre les différentes classes.

$$\text{Gain de Gini}(X, T) = \sum_{j=1}^{m} \frac{|T_j|}{|T|} \text{Gini}(T_i)$$

Si l'ensemble de données est partitionné en deux partitions (T1, T2), alors le gain de Gini sera exprimé comme suit :

$$\text{Gain de Gini}(X, T) = \frac{|T_1|}{|T|} \text{Gini}(T_1) + \frac{|T_2|}{|T|} \text{Gini}(T_2)$$

Par ailleurs, le choix de la coupure qui maximise l'homogénéité au seins de chaque séparation se fait par le choix de l'attribut avec le gain de Gini minimum.

1.2.2. L'optimisation de la taille de l'arbre CART

[3] Enfin, dans le cas d'attributs continus, il y a une infinité de tests envisageables. Dans ce cas, on découpe l'ensemble des valeurs possibles en segments, ce découpage peut être fait par un expert ou fait de façon automatique.

Les critères de séparation qui viennent d'être décrits, ne permettent pas d'arrêter la partition prématurément ce qui cause la création d'un arbre de grande taille avec plusieurs nœuds excédentaires qui contiennent peu d'éléments voire un seul élément. Ainsi, les premières partitions sont les moins dépendantes de l'échantillon, alors que les suivantes sont trop collées à l'échantillon. Il s'agit d'une situation de sur-apprentissage ou « over-fitting » (erreur de généralisation). Ce problème nécessite des heuristiques pour limiter la complexité de l'arbre de décision et réduire les effets de sur apprentissage, on supprime les parties les moins significatives de l'arbre maximal afin de minimiser leur erreur de généralisation pour généraliser les règles extraites de l'arbre.

1.2.2.1. Les critères d'arrêt de construction d'un arbre de décision

Pour réduire la complexité de l'arbre et augmenter sa robustesse on pourrait utiliser plusieurs conditions d'arrêt lors de la construction de l'arbre. Parmi ces conditions l'on peut citer :

- Un nombre minimum d'observations par nœud terminal atteint une limite fixée.
- Un nombre minimum d'observations associé à chaque nœud pour continuer à diviser. - Un nombre de feuilles atteint un maximum fixé.
- La profondeur de l'arbre atteint une limite fixée.

Ces conditions permettent d'arrêter la construction dès qu'un seuil est atteint, ils s'inscrivent dans le cadre de la technique d'élagage et plus particulièrement dans son premier type appelé préélagage. Les méthodes d'élagage sont décrites dans la partie suivant.

1.2.2.2. La phase de pré-élagage

Le pré-élagage se produit au moment de la phase d'expansion, il propose des critères d'arrêt (détaillés au-dessus) durant cette phase en fixant une condition d'arrêt pour arrêter la construction. Cependant, en pratique, il est difficile de prédire quand il faut faire le pré-élagage parce qu'on ne sait pas comment fixer le bon seuil. C'est pourquoi il plus pertinent d'avoir recours à la technique de post-élagage.

1.2.2.3. La phase de post-élagage

Le post-élagage est une seconde approche apparue avec la méthode CART pour sélectionner l'arbre de taille optimale. Dans la mesure où l'arbre maximal est caractérisé par une complexité élevée expliquée par le sur-ajustement aux données de l'échantillon d'apprentissage (une très grande variance), il convient de trouver l'arbre optimal parmi tous les sous arbres binaires élagués de l'arbre maximal.

La méthode de post-élagage utilisée par l'algorithme CART (Breiman et al. en 1984) est basée sur le coût de complexité minimal (Cost Complexity Metric) qui combine le coût de classification erronée et le coût de pénalité pour la complexité de l'arbre. Elle est basée sur deux étapes ; la première consiste à utiliser un échantillon d'apprentissage pour construire un arbre saturé c'est-àdire avec les feuilles les plus pures possible puis générer une suite de tous les sous arbres emboités (élagués les uns des autres), Tmax = T0 ⊃ ⊃ T1 ... Tk-1⊃ Tk = racine, tous élagués de Tmax en partant des feuilles et en remontant vers la racine et en transformant un nœud en feuille à chaque étape.

La construction de cette séquence de sous-arbres repose sur une pénalisation de la complexité de l'arbre :

On compare le coût du nouvel arbre à celui du précèdent à travers le critère du coût en complexité défini comme suit :

Pour tout sous arbre T < T max et α[4] le coût de pénalité par nœud variant de 0 à l'infini on a :

$$\mathbf{R_\alpha(T) = R(T) + \alpha|L(T)|}$$

Dans cette formule, T désigne un sous arbre élagué de T_{max}, L(T) représente le nombre de nœuds terminaux de l'arbre T, R(T) représente la fraction des observations de validation mal classées par T.

Ensuite, on doit trouver la série de sous-arbres élagués T(α) de Tmax qui minimisent $R_\alpha(T)$ pour chaque valeur de α (Breiman et al., 1984) ;

[4] Paramètre de complexité et mesure la diminution du cout de mauvais classement obtenu en élaguant Ti+1 pour obtenir Ti

$$R_\alpha(T(\alpha)) = \min R_\alpha(T) \quad \text{pour tout } T \leq T_{max}$$

$$T0 \supset T1 \supset \ldots Tk\text{-}1 \supset Tk$$

Les arbres de cette série minimisent $R_\alpha(T)$ sur les plages de valeurs de α

La deuxième étape de l'élagage à complexité de coût minimale consiste à choisir, seulement parmi cette suite d'arbres, l'arbre optimal qui minimise l'erreur de généralisation. Pour estimer cette erreur pour chacun de ces arbres, on utilise un échantillon test indépendant de la base d'apprentissage.

Cependant, nous le verrons dans les sections subséquentes, même l'arbre de décision optimal peut ne pas être un classifieur efficace sur la base test. Il existe des méthodes dites d'agrégation ou ensemblistes permettant de gagner en pouvoir prédictif.

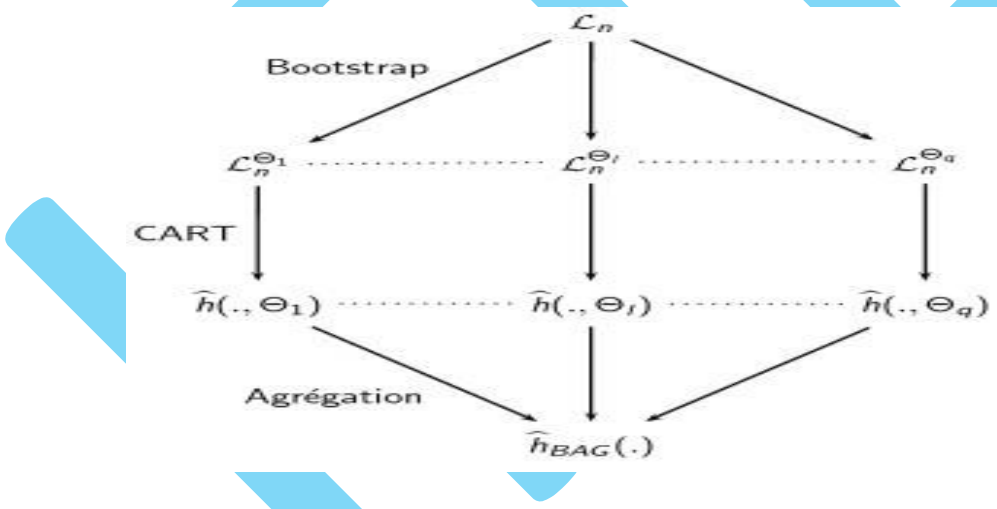
Section 2 : Prédiction par des méthodes d'agrégation des classifieurs CART

Comparées à d'autres modèles, les méthodes d'apprentissage par arbres de décision qui viennent d'être décrites nécessitent moins d'efforts pour la préparation des données lors de leur prétraitement. Ainsi, ils ont l'avantage d'être simple à interpréter et à visualiser (Modèle white box). Toutefois, ils présentent quelques inconvénients à l'instar de leur sensibilité à des petites variations dans les données, autrement dit, les petits changements, surtout ceux qui sont près de la racine, dans les données peuvent entraîner un grand changement dans la structure de l'arbre de décision, ce qui provoque une instabilité qui induit à une variance élevée, qui nuit à la qualité de l'apprentissage. Par conséquent, pour corriger ce manque de robustesse des arbres de décision, des approches de type « bagging » ou « Forêts aléatoires » ont été inventés par Breiman dans sa revue internationale Machine Learning.

2.1. Le Bagging

Avant d'introduire la méthode de random forest, Leo Breiman (1996) a inventé premièrement le concept de bagging ou Bootstrap AGGregation qui fait partie de la catégorie des algorithmes d'induction des classifieurs capables de générer automatiquement des ensembles de classifieurs.
Ce concept consiste à appliquer le principe de Bootstrap à l'aggrégation de classifieurs : Le Bootstrap est un principe de rééchantillonnage statistique (Efron 1979) utilisé pour l'estimation des paramètres statistiques dont on ne connait pas la loi. Il permet d'effectuer un rééchantillonnage de la base d'apprentissage $\mathcal{L}$n. L'objectif consiste à entraîner chaque arbre de la forêt à partir des jeux de données légèrement différents, composés des échantillons à travers des tirages avec remise de la base d'apprentissage ($\mathcal{L}_n^{\Theta 1}$, . . ., $\mathcal{L}_n^{\Theta q}$). Ensuite, on applique une fonction de prédiction (un arbre CART) sur chaque échantillon bootstrap pour obtenir une collection de prédicteurs ($\hat{h}$ (., $\mathcal{L}_n^{\Theta 1}$), . . ., $\hat{h}$ (., $\mathcal{L}_n^{\Theta q}$)), et enfin agréger ces prédicteurs par un vote majoritaire pour la classification de base tel qu'illustré dans la figure ci-dessous :

Figure 8: Exemple illustré de la méthode de Bagging



θ_l, i = 1, ..., q désignent les variables indépendantes pour l'aléa de bootstrap, $\hat{h}$(., θi) est le prédicteur construit à partir de l'échantillon $\mathcal{L}_n^{\Theta i}$

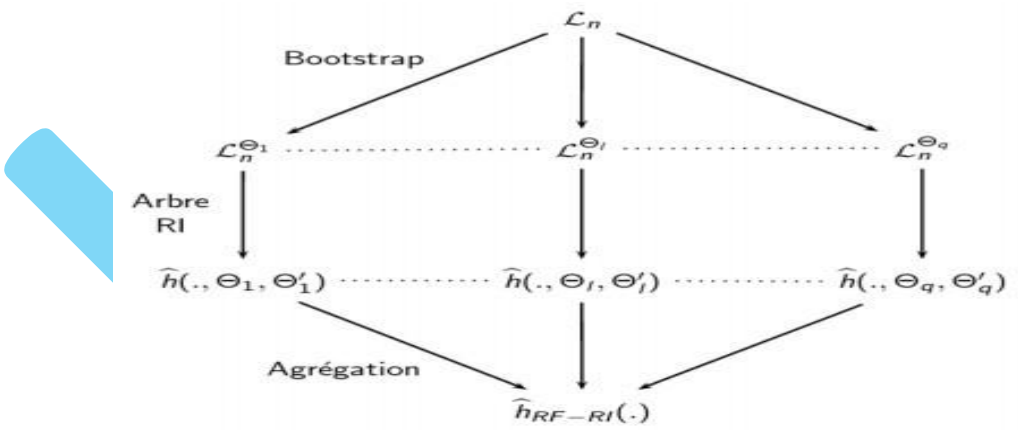
2.2. L'entrainement des forêts aléatoires

Un cas particulier de Bagging est celui appliqué à des arbres de décision avec introduction d'un tirage aléatoire parmi les variables explicatives. Il permet d'éviter de voir apparaitre toujours les mêmes variables. On a dans ce cas une double randomisation et on parle de forets aléatoires

(breiman,2001) qui sont une amélioration de Bagging à travers la dé-corrélation de la collection d'arbres provenant du Bagging en introduisant de l'aléa dans la construction de l'arbre individuel. L'objectif est donc de rendre plus indépendants les arbres de l'agrégation en ajoutant du hasard dans le choix des variables qui interviennent dans les modèles. Cette indépendance va permettre de rendre le vote plus efficace.

Le principe de construction des Random Forests consiste tout d'abord à générer plusieurs souséchantillons bootstrap ($\mathcal{L}_n^{\Theta 1}$, . . ., $\mathcal{L}_n^{\Theta q}$) d'une manière similaire à la méthode de Bagging. Plus précisément, un arbre est, ici, construit de la façon suivante ; Pour découper un nœud, on tire aléatoirement un nombre sans remise de m parmi p variables, et on cherche une coupure optimale uniquement suivant l'ensemble aléatoire de m ≤ p variables sélectionnées au début de la construction de la forêt. De plus, l'arbre construit est complètement développé et n'est pas élagué. La collection d'arbres obtenus est enfin agrégée par un vote majoritaire pour donner le prédicteur Random Forests.

Figure 9: Exemple illustré de la méthode de Random Forest



2.3. L'optimisation des forêts aléatoires

Pour comprendre plus en détail la façon dont l'estimation de l'erreur de généralisation est calculée, on fixe d'abord une observation (Xi, Yi) de l'échantillon d'apprentissage $\mathcal{L}_n$ et puis on considère l'ensemble des arbres construits sur les échantillons bootstrap ne contenant pas cette observation, c'est-à-dire pour lesquels cette observation est "Out-Of-Bag". On agrège ensuite uniquement les

prédicteurs $\hat{h}(X_i, \theta')$ de ces arbres en fonction de Xi pour fabriquer notre prédiction $\hat{Y}_i$ de Y_i. Après avoir fait cette opération pour toutes les données de $\mathcal{L}_n$, on calcule alors l'erreur OOB_i de chaque arbre i à partir de la proportion d'observations mal classées :

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\hat{Y}_i \neq Y_i}$$

.

Chapitre III : Application

Section 1 : Étude de la base des données

Cette section sera consacrée à visualiser, transformer la base des données pour optimiser la modélisation. Nous détaillerons les modèles que nous allons utiliser pour le prétraitement des données et les étapes relatives à leur préparation.

1.1. Présentation et description des données

La fraude à l'assurance est un sujet très large, c'est pourquoi nous nous sommes limités aux polices d'assurance automobile. Ces polices se démarquent par plusieurs garanties : Responsabilité civile,

bris de glace, dommages collision, incendie, vol.… Nous avons choisi de traiter une garantie contractuelle dans laquelle le risque de fraude est fréquemment répandu par les fraudeurs ; il s'agit de la garantie Tous Risques. Cette dernière est impliquée particulièrement dans la fraude pour plusieurs raisons :

- Les sinistres fictifs sont faciles à mettre en scène, l'assuré sera indemnisé indépendamment de sa responsabilité (fautif ou non fautif).
- Ces deux garanties n'impliquent pas beaucoup de complices, le fraudeur pourra soit agir seul soit faire recours à un tiers (un garagiste) pour produire des fausses factures ou des factures gonflées.
- La déclaration n'implique pas des adversaires; généralement le tiers n'est pas identifié.
- Éviter le malus affèrent à la garantie « responsabilité civile » pour ne pas toucher la prime. En effet, les accidents de stationnement avec tiers non identifiée ou le dérapage du véhicule ne font pas l'objet d'un malus

Notre base des données récence 1183 adhérents qui ont souscrit la garantie tierce complète (tous risques) à partir de l'année 2015 jusqu'au 2019. Les données imputées dans cette base représentent des polices d'assurance qui ont fait l'objet des sinistres déclarés par les adhérents avec la désignation éventuelle d'un expert ingénieur pour évaluer les dégâts subis et déterminer si les relations entre les circonstances déclarées et les dégâts sont conformes ou pas. Par conséquent, à l'issu de cette observation, croisée avec l'avis d'un gestionnaire sinistre, nous avons qualifié chaque dossier de potentiellement frauduleux ou non.

Cependant, dans la mesure où l'acte frauduleux est difficile à détecter, nous avons constaté quelques dossiers frauduleux qui ont échappé aux yeux de l'expert et qui ont fait l'objet d'un règlement. Dans ce cas l'évaluation de ces sinistres est à dire du gestionnaire.

Dans cette lignée, la base de données extraite regroupe des variables nécessaires pour modéliser et prédire le comportement frauduleux. Elle englobe les caractéristiques du véhicule assuré, les caractéristiques de l'adhèrent ainsi que les caractéristiques du sinistre. Ces caractéristiques sont traduites par des attributs avec des types différents :

- Les variables qualitatives:

- Variable nominale : Les modalités de cette variable sont sans ordre implicite.
- Variable ordinale : Variable catégorielle et ses modalités sont ordonnées.
- Variable binaire (dichotomique) : variable nominale avec seulement deux modalités.

 Les variables quantitatives:

- Variable continue : La valeur qu'elle peut prendre est réel. Il s'agit donc d'un ensemble infini non dénombrable.
- Variable discrète : il s'agit d'un ensemble numérique fini.

Tableau : Description des variables de la base des données

Attributs	Type	
Profession	Variables qualitatives	Variable catégorielle :nominale
Code postal		Variable catégorielle :nominale
Sexe		Variable catégorielle : nominale
Agent de souscription		Variable catégorielle : nominale
Nature des dégâts		Variable catégorielle : binaire
Nombre de places		Variable catégorielle : nominale
Engin		Variable catégorielle : binaire
Marque		Variable catégorielle : nominale
Dégâts		Variable catégorielle : nominale
Classe		Variable catégorielle : binaire
Délai de déclaration	Variable quantitative	Variable discrète
Age		Variable continue

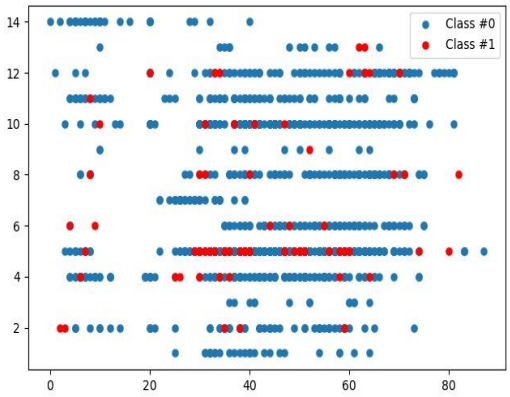
1.2. Analyses statistiques

Le prétraitement des données fait partie intégrante de la modélisation statistique vu que la qualité des données affecte énormément la capacité d'apprentissage de notre modèle ; il est donc extrêmement important que nous prétraitions préalablement nos données pour élaborer des modèles de bonne qualité.

Le jeu de données brut que nous avons recueilli revête une structure complexe. En effet, outre les problèmes classiques de traitement de données (des données bruitées, des valeurs manquantes ou aberrantes), nous constatons une distribution fortement déséquilibrée de la variable cible (figure

…) ce qui va affecter la qualité de prédiction du modèle. La négligence de toutes ces problématiques pourra engendrer de forts biais et une perte de précision du modèle donc il sera utile de suivre une démarche particulière pour rendre cette base utilisable.

Figure 10: Répartition des classes



Classe des fraudeurs
Classe des non fraudeurs

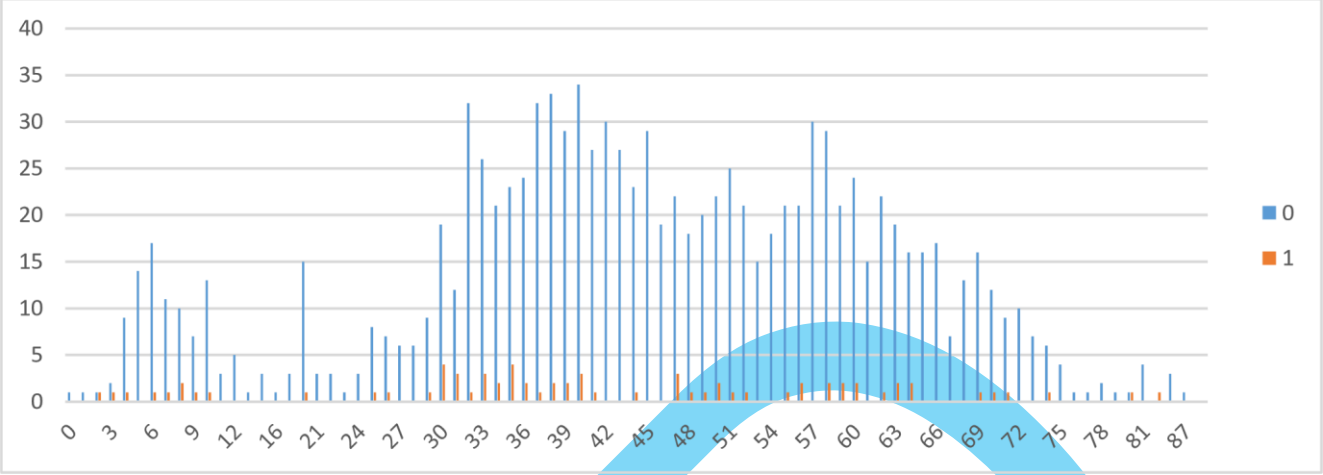
Source : Élaborée sous Python

1.3. Exploration de la distribution des variables

Pour toute variable quantitative, il est important d'étudier sa distribution en s'appuyant sur plusieurs indicateurs pour repérer des valeurs aberrantes qui seront ensuite étudiées pour évaluer leur impact sur le modèle. En ce qui concerne les variables qualitatives, l'étude du nombre des modalités ainsi que de leurs fréquences indique s'il y aura d'éventuels ajustements à faire ultérieurement. En outre, pour chaque variable, les distributions doivent être comparées de manière conditionnelle à la valeur cible Y, ce qui donne une indication du potentiel explicatif de chaque variable. Pour effectuer cette comparaison, il est possible de faire appel aux histogrammes ou des études de corrélations.

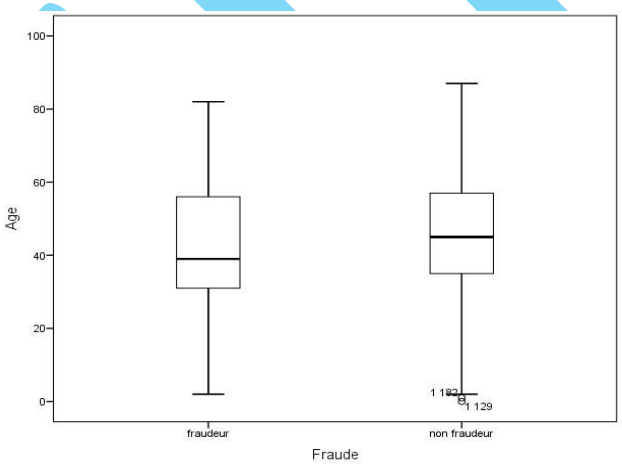
- **La variable âge**

Figure 11: Répartition des adhérents en fonction de leurs âges



Source : élaboré par nos soins à l'aide d'EXCEL

Figure 12: Boite à moustaches de la distribution de la variable « âge »



Source : Élaborée par nos soins à l'aide de SPSS

Tableau : Les caractéristiques statistiques de la variable « âge » dans les deux classes

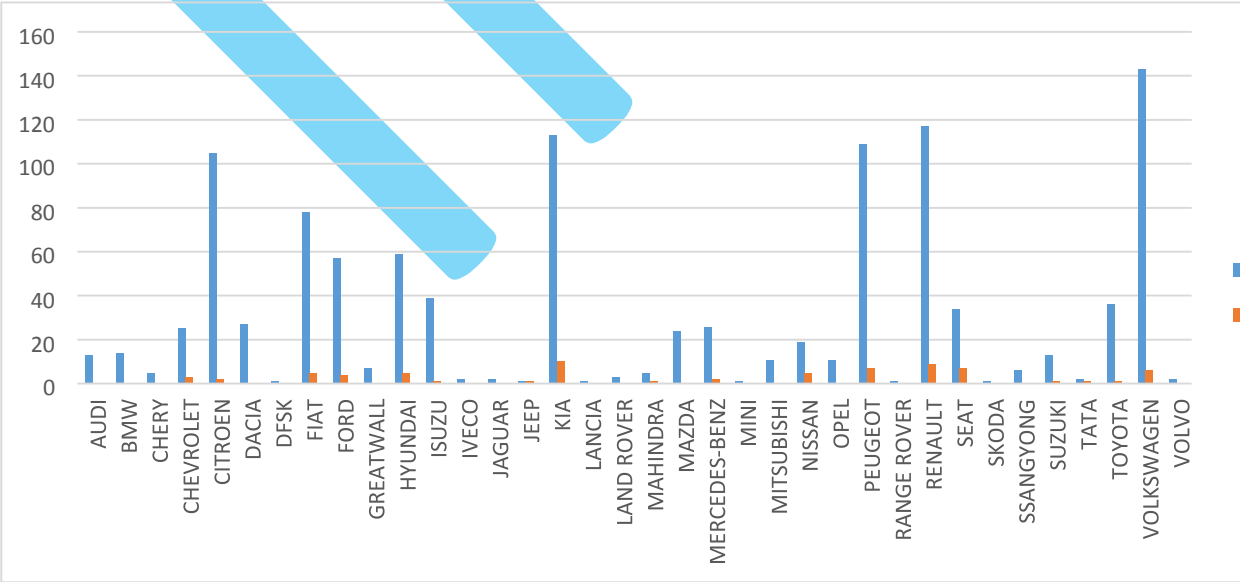
Fraude	Légitime

médiane	39	médiane	45
Q1 (quartile inférieur)	31	Q1 (quartile inférieur)	35
Q3 (quartile supérieur)	56	Q3 (quartile supérieur)	57
IQR(l'écart interquartile)	25	IQR(l'écart interquartile)	22
upper bound	93,5	upper bound	90 2
lower bound	-6,5	lower bound	

Nous avons repéré également deux âges atypiques dans la classe non fraudeurs, au niveau des adhérents A1129 et A1182. En guise de solution, nous allons modifier les âges de ces deux adhérents par la valeur de la médiane, soit 45 ans.

La variable marque

Figure 13: Répartition des marques des véhicules



Source : Élaboré par nos soins à l'aide d'EXCEL

Nous avons constaté que certaines variables qualitatives ont pris plusieurs modalités tels que la marque du véhicule et la profession. Toutefois, elles peuvent être plus discriminantes si nous réduisons le nombre des modalités vu que certains algorithmes tel que l'arbre de décision trouvent des difficultés à exploiter au mieux les informations contenues dans ces variables lorsque ce nombre de modalités est élevé.

Afin de s'en tirer avec ce nombre élevé de modalités nous avons combiné les modalités à faibles effectifs en fusionnant celles dont la fréquence est faible.

Cet histogramme illustre la répartition des marques des véhicules assurés dans toute la base des données, reparties en fonction de la valeur cible (1 : fraudeur et 0 : non fraudeur). Nous constations que les marques ayant les fréquences les plus élevées sont : KIA, VOLSWAGEN, SEAT, RENAULT, PEUGEOT.

Les fréquences des autres marques sont faibles, donc nous allons les regrouper dans une seule modalité sous le nom « autres ».

Ainsi, après avoir modifié les modalités, nous avons constaté une corrélation entre la variable cible et la variable marque de véhicule.

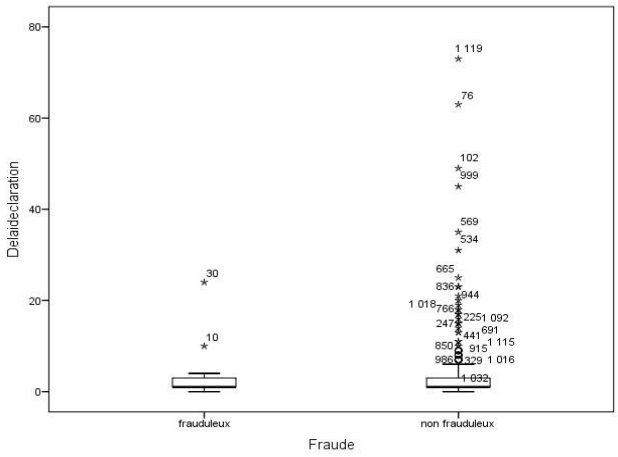
Tableau : Test de Chi-deux entre la variable cible et la marque du véhicule

	Value	df	Asymp. Sig. (2sided)
Pearson Chi-Square	12,159[a]	5	,033
Likelihood Ratio	9,422	5	,093
Linear-by-Linear Association	1,008	1	,315
N of Valid Cases	1184		

Source: élaboré à l'aide de SPSS

- **Le délai de déclaration**

Figure 14: Boîte à moustaches des délais de déclaration



Source : élaborée sous SPSS

L'analyse exploratoire de la variable « délai de déclaration » par la boite à moustaches, nous a révélé plusieurs valeurs atypiques et des Outliers légères et extrêmes représentées par des astérisques (*) sur le diagramme. Celles-ci doivent être traitées pour ne pas fausser les résultats de la modélisation. En effet, pour rendre cette variable quantitative plus discriminante nous devons la transformer par une autre catégorielle binaire qui se présente comme suit : la première modalité « délai court » et la deuxième modalité « délai long ».

2.2. Étude de la corrélation entre les variables

L'analyse de corrélation est une technique largement utilisée permettant d'identifier des relations intéressantes dans les données. Ces relations nous aident à évaluer la pertinence des attributs par rapport à la variable cible à prévoir. Toutefois, l'application des techniques de corrélation dépend du type des données. Le tableau au-dessous montre la répartition des méthodes utilisées pour l'analyse des corrélations en fonction des types de données.

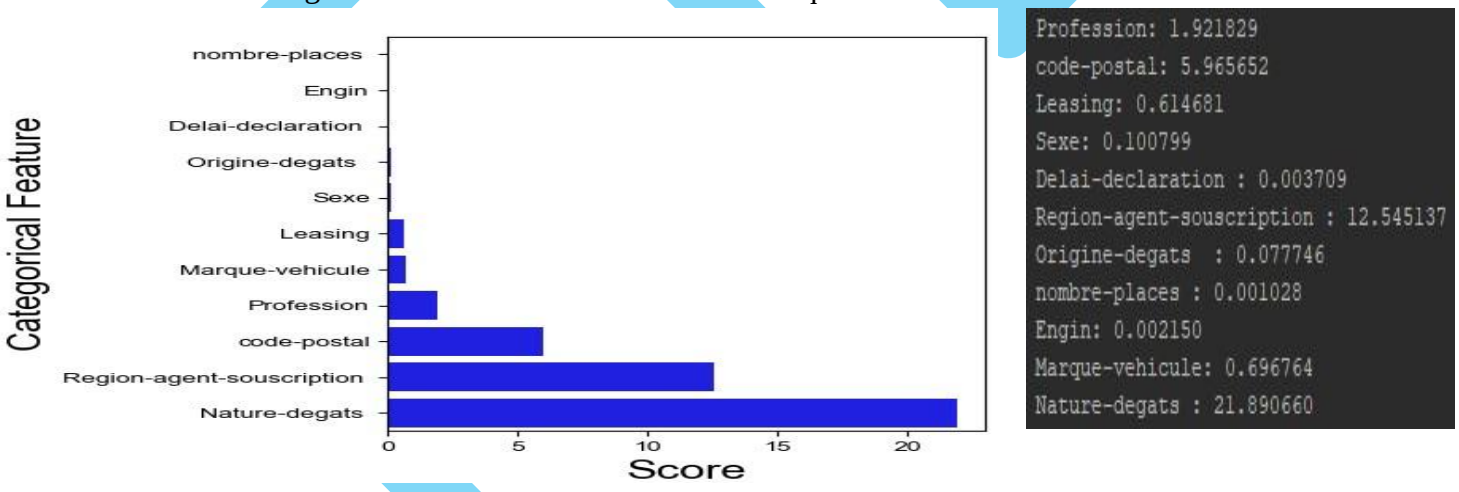
Type des données			Variable dépendante		
			catégorielle		Quantitative
			Nominale	Ordinale	
		Nominale	Test d'indépendance de KHI deux		

<table>
<tr><td rowspan="2">Variable indépendante</td><td>Catégorielle</td><td>Ordinale</td><td>Test d’indépendance de KHI deux</td><td>Corrélation de Spearman</td><td>Analyse des variances ANOVA</td></tr>
<tr><td colspan="2">Quantitative</td><td colspan="2">Test d’indépendance de KHI deux</td><td>Corrélation de Pearson</td></tr>
</table>

Afin de chercher les variables explicatives (catégorielles) les plus pertinentes de l'ensemble des données nous avons calculé le score de chi-deux entre chaque variable explicative catégorielle et la variable cible. À l’issus de ces calculs nous avons sélectionné les variables explicatives ayant les meilleurs résultats de chi-deux.

Les scores de chaque variable explicative sont détaillés dans l’histogramme qui suit :

Figure 15: Les scores de chi-deux de chaque variable



Source : Élaboré par nos soins à l’aide de Python

Les scores de chi deux les plus élevés indiquent une plus grande pertinence et importance dans la prédiction et peuvent être incluses dans le développement du modèle prédictif. Toutefois les valeurs très faibles seront exclues de la modélisation. Dans ce sens, nous avons observé des scores quasi-nulles au niveau des variables « nombre de places », « Engin », « Origine des dégâts » et « sexe » et « délai de déclaration ». Par conséquent, ces variables seront écartées de la modélisation.

Pour renforcer notre précédente interprétation, nous avons effectué une Analyse des Correspondances Multiples (ACM). Celle-ci permettra de nous aider à résumer les informations contenues dans ces variables et faciliter l'interprétation des corrélations existantes entre elles. Elle nous permet également de produire des graphiques sur lesquelles nous pouvons observer visuellement les proximités entre les modalités des variables qualitatives et les observations.

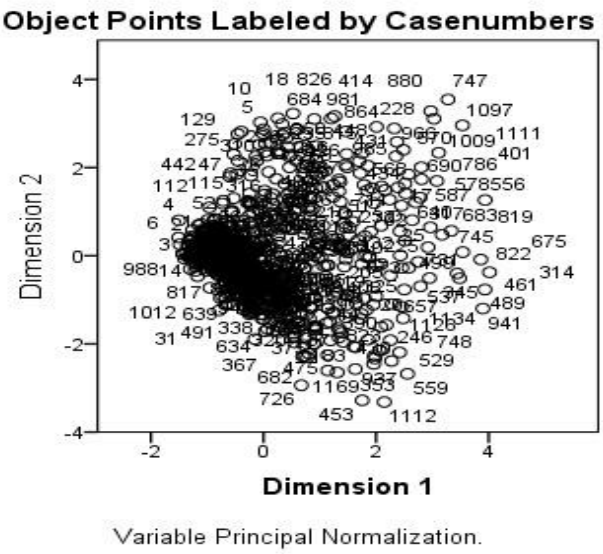
Tableau : Récapitulation des modèles

Dimension	Cronbach's Alpha	Variance Accounted For		
		Total (Eigenvalue)	Inertie	% of Variance
1	,765	3,284	,299	29,851
2	,602	2,208	,201	20,071
Total		5,491	,499	
Mean	**,7**	2,746	,250	24,961

Source: Élaboré par nos soins à l'aide de SPSS

Au regard du tableau de récapitulation des modèles, nous observons que l'ensemble de variables qu'on a entré ont dégagé deux dimensions d'un total qui est égale à 50%, ce qui est jugé satisfaisant. Donc ses deux dimensions résument 50 % de l'ensemble des informations données par l'ensemble des variables que nous avons introduit. Toujours sur le même tableau, nous remarquons que la valeur moyenne du coefficient alpha de Cronbach est de 0.7, ce qui est acceptable, puisqu'il est égal au seuil minimum requis de 0,70 (Nunnaly, 1978). Par conséquent, nous pouvons dire que nous obtenons une cohérence interne satisfaisante.

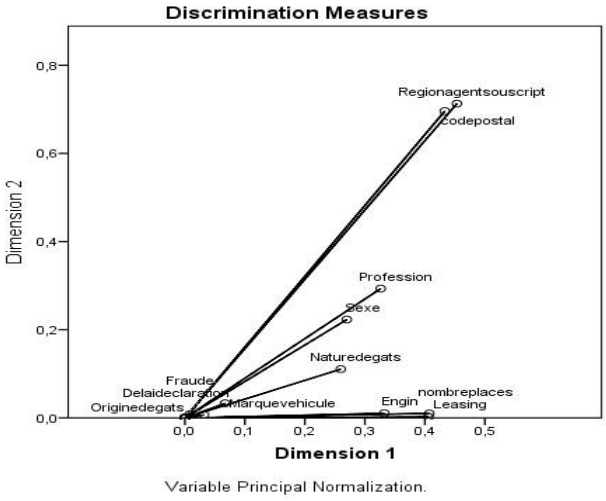
Figure 16: Tableau des objets étiquetés par nombre d'observations



Source: Élaboré par nos soins à l'aide de SPSS

En examinant l'allure générale du nuage des individus, nous observons que l'ensemble des observations sont regroupées dans la même zone. En effet, les individus proches du graphique donnent des interprétations plus précises contrairement à ceux éloignés du centre de graphique.

Figure 17: Représentation des variables sur les deux dimensions



Source: Élaboré par nos soins à l'aide de SPSS

En ce qui concerne le diagramme représentant les variables qualitatives, la visualisation des projections des variables sur les deux dimensions nous a permis d'identifier de fortes corrélations entre les variables telles que les variables « code postal » et « région de l'agent de souscription ». Ceci est confirmé par le test bi-varié de Chi-deux.

Tableau : Test de chi deux entre le code postal de l'adhérent et la région de l'agent de souscription

Chi-Square Tests

	Value	df	Asymp. Sig. (2sided)
Pearson Chi-Square	3618,279	36	,000
Likelihood Ratio	1854,960	36	,000
Linear-by-Linear Association	724,817	1	,000
N of Valid Cases	1184		

Symmetric Measures

	Value	Approx. Sig.
Contingency Coefficient	**,868**	,000
N of Valid Cases	1184	

Source: Élaboré par nos soins à l'aide de SPSS

À l'examen de la relation entre les variables « code postal de l'adhèrent » et « région de l'agent de souscription » nous remarquons que la signification asymptotique est égale à 0.000 (<0.05) donc les variables présentent une association significative (rejeter H0 de l'indépendance des deux variables).

Dans le tableau de contingence nous remarquons que la marge d'erreur est nulle et la valeur du coefficient de contingence est égale à 86.8% ce qui permet de dire qu'il y'a une forte relation entre ces deux variables.

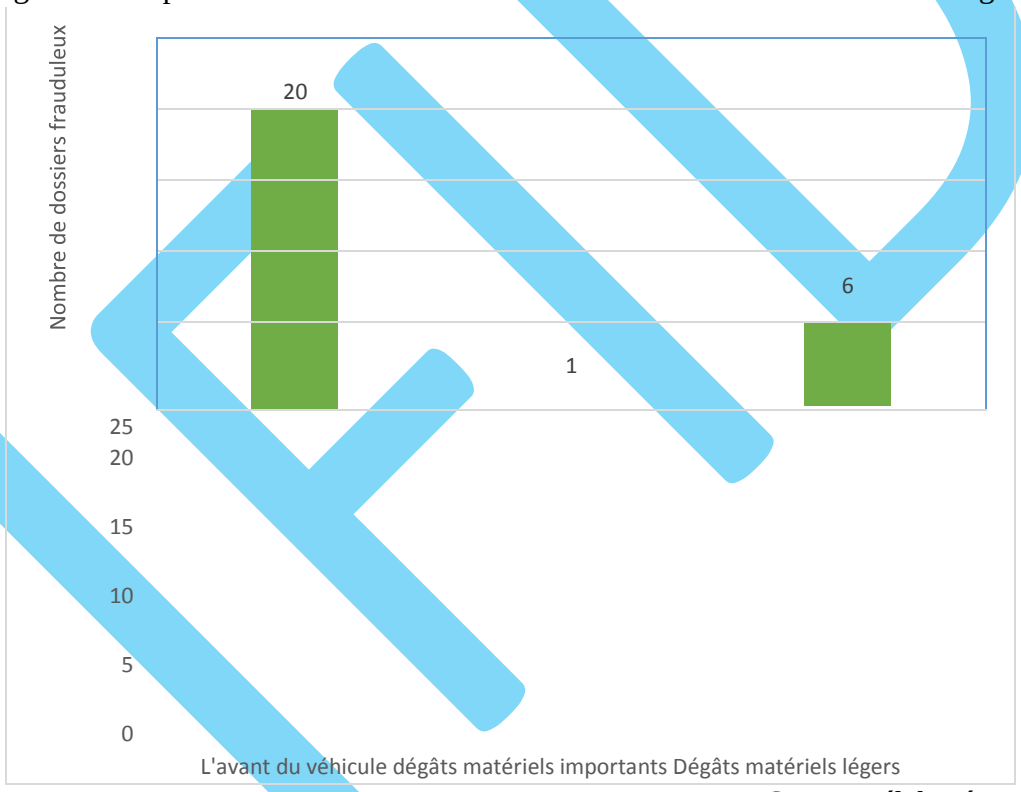
À l'issu de ce test, nous allons éliminer les variables très corrélées pour ne pas biaiser le modèle.

Section 2 : Étude d'un échantillon de dossiers frauduleux

2.1. Les dossiers avec des cas de fraude avérée

À partir de ces dossiers nous avons passé en revue tous les éléments qui caractérisent le fraudeur, détaillés comme suit :

Figure 18: Répartition des dossiers frauduleux en fonction de la nature des dégâts

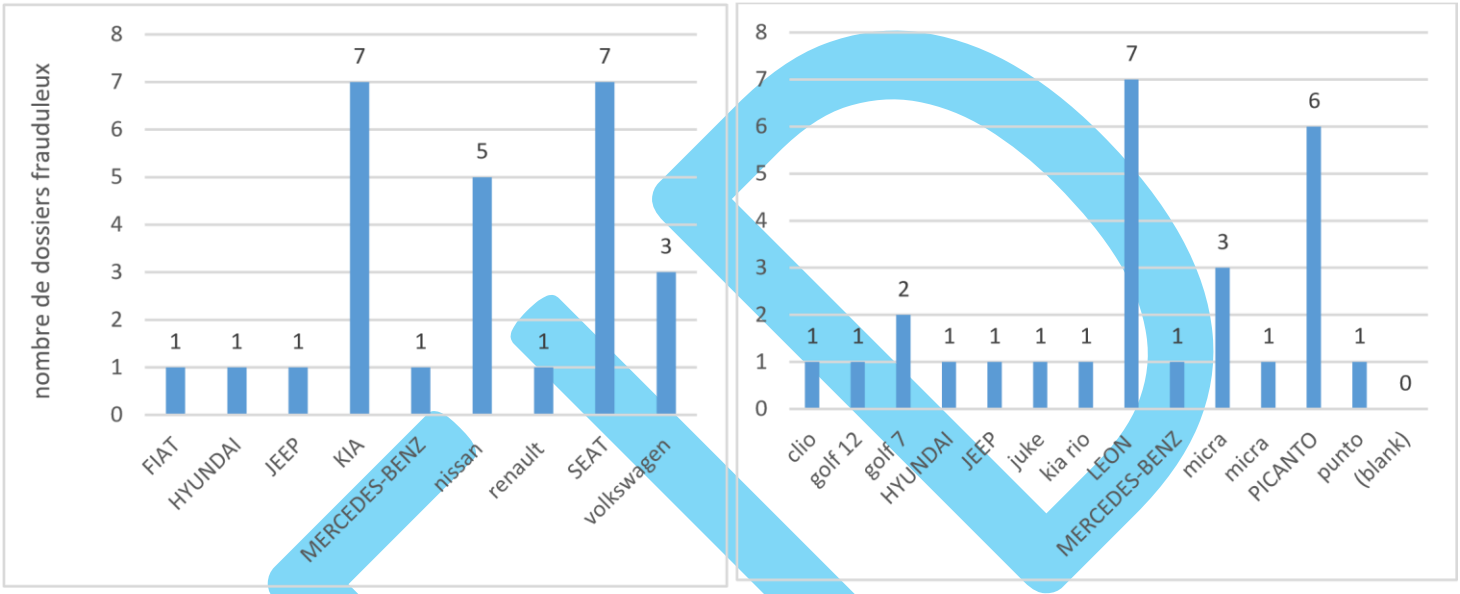


Source : élaborée sous Excel

L'évaluation a révélé que la plupart des dossiers frauduleux présentent les mêmes dégâts matériels au niveau de l'avant du véhicule accidenté, autrement dit, tous les points de choc convergent vers un choc frontal au niveau du parechoc, Roue avant, Amortisseur avant, Optique de phare, Capot moteur, Airbag, Aile avant gauche, Aile avant droite, Pare-brise… Ce résultat est attendu vu que les éléments précités sont des pièces de rechange démontables, ce qui coïncide avec

le scenario de changement des pièces intactes par des pièces accidentées avec la complicité d'un garagiste afin de bénéficier d'une indemnité pour un accident qui n'a jamais eu lieu.

Figure 19: Répartition des dossiers frauduleux selon la marque et le modèle du véhicule



Source : Élaboré sous Excel

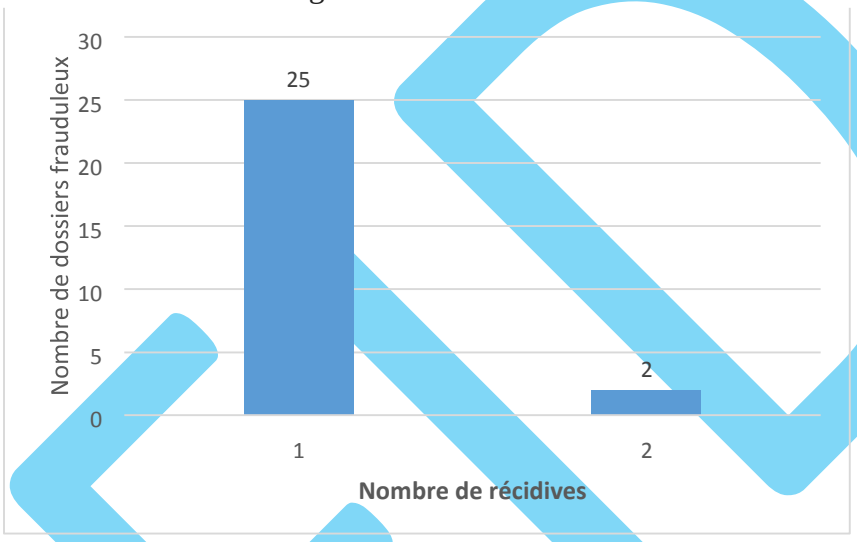
D'après la figure ci-dessus, les marques et les modèles les plus impliqués dans les actes frauduleux sont : KIA PICANTO, SEAT LEON et NISSAN MICRA. La présence fréquente de ces types de véhicules nous a laissé présumer qu'il existe plusieurs pièces démontables de ces marques au niveau des garagistes qui ont participé à la fraude.

En outre, après avoir examiné les factures de réparation, nous avons constaté que toutes les réparations des véhicules endommagés ont été effectuées dans des garages indépendants non agréés par l'assureur, ce qui laisse supposer une complicité entre les garages et les assurés frauduleux.

En sus de cela, les dossiers ont en commun le fait qu'il s'agit des nouvelles souscriptions c'est-àdire qu'il y'a un court laps de temps entre la souscription et la déclaration du sinistre suivi par une résiliation du contrat juste après l'indemnisation.

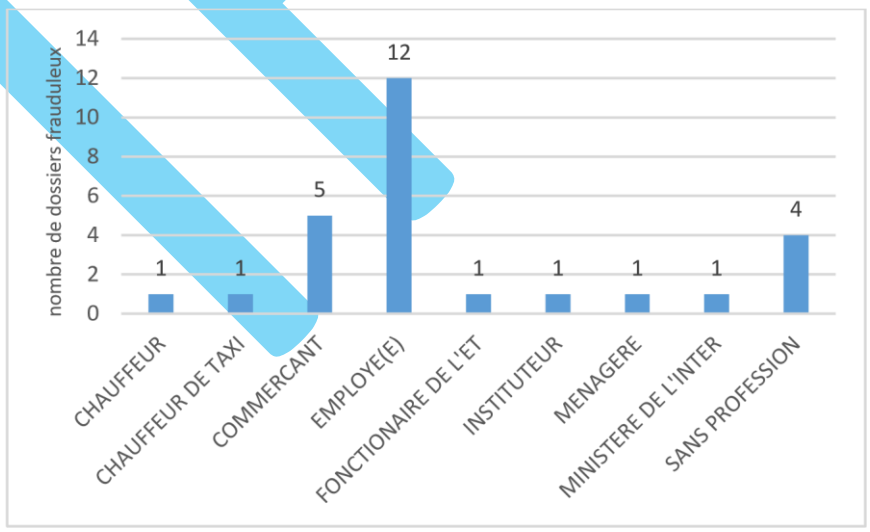
En ce qui concerne les montants des règlements, nous avons constaté que la majorité des indemnités est très élevée, allant de 10 000 000 DT à 40 000 000 DT, suivie des résiliations. Cela pourrait être expliqué par le fait que les fraudeurs réclamant des montants élevés ne peuvent pas recourir à des manœuvres frauduleuses avec la même compagnie d'assurance pour éviter de susciter l'attention des gestionnaires. Ceci s'explique également par la quasi-absence des récidives au niveau de ces dossiers.

Figure 20: Les récidives



Source : Élaboré sous Excel

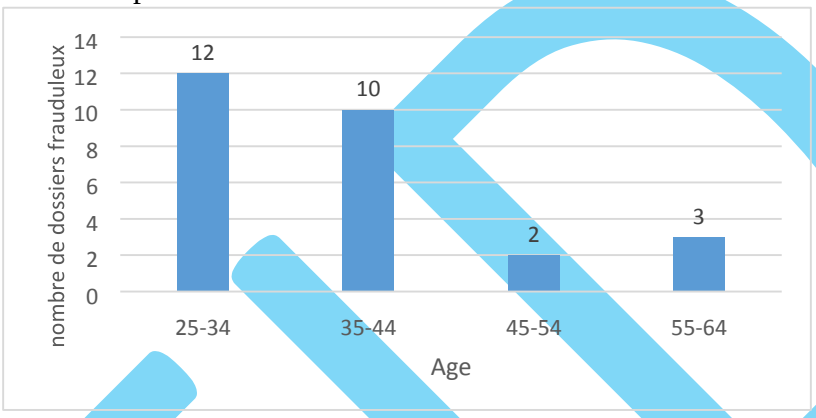
Figure 21: Répartition des dossiers frauduleux selon la catégorie socioprofessionnelle



Source : Élaboré sous Excel

Quant à la catégorie socioprofessionnelle des assurés frauduleux, nous avons constaté que la plupart des fraudeurs sont des employés, suivis par les commerçants et les chômeurs.

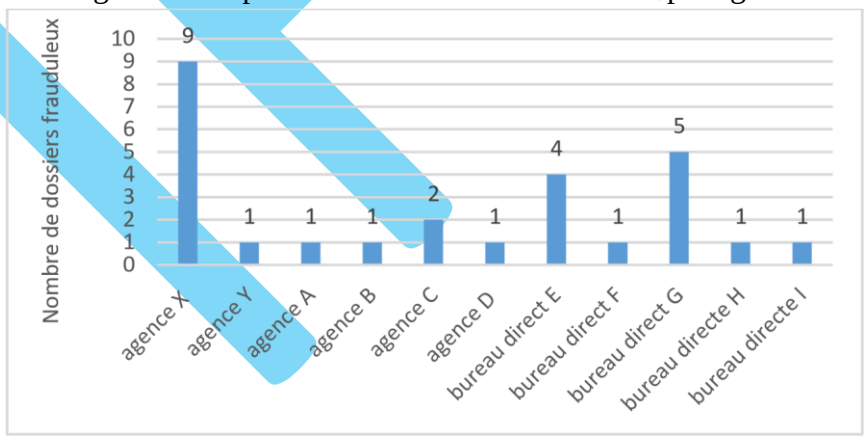
Figure 22: Répartition des dossiers frauduleux selon les tranches d'âge



Source : Élaboré sous Excel

Comme nous pouvons le voir dans l'histogramme suivant, nous constatons une forte concentration des fraudeurs étant âgés entre 25 et 44 ans.

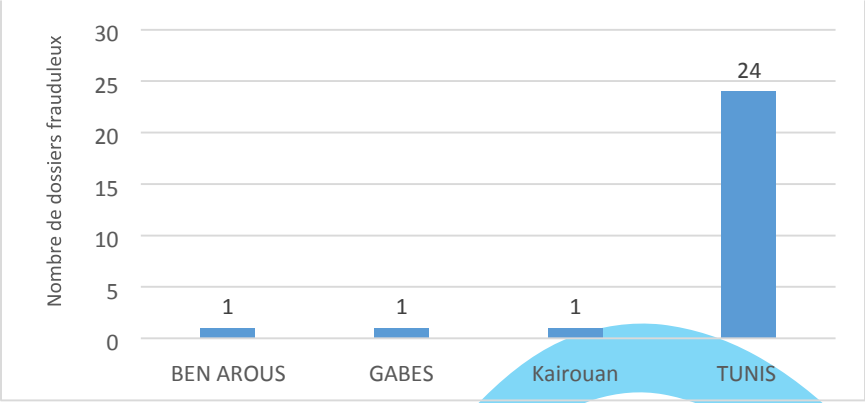
Figure 23: Répartition des dossiers frauduleux par agence



Source : Élaboré sous Excel

Nous observons une fréquence de fraude élevée au niveau de l'agence X par rapport aux autres agences suivie par le bureau direct G et le bureau direct E.

Figure 24: Répartition des dossiers frauduleux par région de résidence



Source : Élaboré sous Excel

2.2. Fraude détectée par les comportements atypiques

Comme explicité auparavant, les réclamations qui cachent des manœuvres frauduleuses présentent généralement des caractéristiques relativement similaires à celles des réclamations légitimes. Ce qui fait que le profil d'un fraudeur est difficilement repérable.

Pour ce faire, nous avons collecté une base de données qui englobe les dossiers sinistres ouverts sous la garantie Tous Risques entre 2015 et 2019 et qui ont fait l'objet d'un règlement total. Cette base nous a servi de détecter des comportements atypiques à travers le nombre de sinistres antérieurs déclarés par chaque adhérent.

Pour affiner nos résultats, nous avons ciblé les indemnisations de faibles montants (inferieurs à 1000 DT) vu que les adhérents ont toujours le pressentiment de l'existence d'un audit et qu'ils en tiennent compte lorsqu'ils se livrent à des activités frauduleuses. Les sinistres de grande ampleur font généralement l'objet d'une enquête sur l'origine du sinistre, nous nous attendons alors à ce que les fraudeurs ne tombent pas dans ce piège. Ils pensent à commettre des sinistres de moindre valeur.

En guise de solution, nous avons essayé de taguer des observations atypiques et inconsistantes par rapport à la majorité des observations, autrement dit nous avons cherché des signaux faibles qui peuvent exister dans le segment des adhérents choisi.

Comme prévu, nous avons repéré plusieurs récidives sur la même garantie pour des petits montants Nous avons isolé ces dossiers pour chercher s'il existe des caractéristiques communes entre ces récidivistes

Section 3 : Implémentation des modèles

Cette section sera consacrée à mettre en application les outils d'apprentissage, expliqués en détails dans le chapitre précédent, qui vont servir à prévoir la variable endogène Y par les variables explicatives Xi. Nous appliquerons donc l'arbre de décision et la forêt aléatoire.

3.1.1. Partition en échantillon d'apprentissage et de test

Dans un premier temps, nous avons partitionné aléatoirement le jeu de données en deux sous échantillons : la base d'apprentissage et la base de test.

- L'échantillon d'apprentissage (d'entrainement) l'échantillon principal où sont appliquées les méthodes, sur lesquelles les algorithmes apprennent. Il sert à ajuster le modèle. Il représente classiquement 80% des données.

- Une fois qu'un modèle est entrainé sur l'échantillon d'apprentissage, il sera évalué sur un sous échantillon indépendant tiré de la base totale, appelé échantillon de test (20% de la base), et ce, pour s'assurer que le modèle soit optimal.

Avant de construire le modèle, nous avons scindé les données en deux sous-ensembles ; un échantillon d'apprentissage et l'autre de test, selon les proportions suivantes :

- 80% : l'échantillon d'apprentissage soit 949 observations
- 20% : l'échantillon de test soit 237 observations

3.1.2. Construction de l'arbre de décision

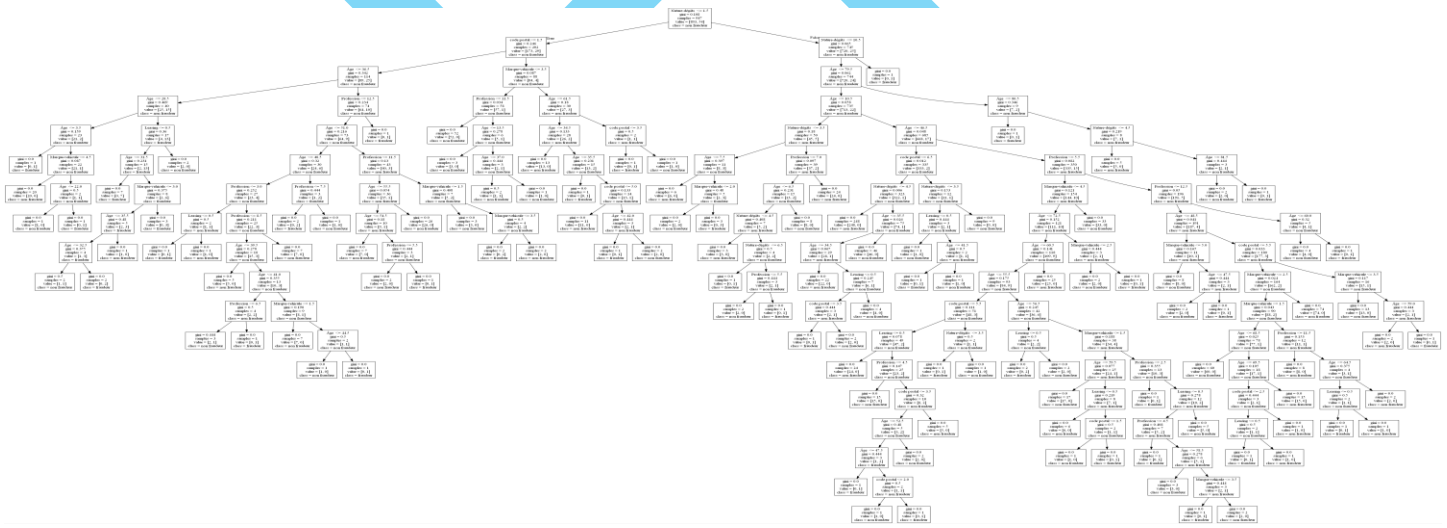
Dans le but de relier la variable à expliquer Y (binaire) à l'ensemble de variables explicatives choisies dans la section précédente nous avons mis en application l'arbre binaire de décision avec l'algorithme CART développé dans le chapitre précèdent.

Nous avons instancié l'arbre de décision avec l'algorithme CART sous le langage Python au moyen de la commande ''DecisionTreeClassifier'' de la librairie Scikit-Learn. Cette commande nécessite l'indication de quelques hyper paramètres qui revêtent une importance capitale pour atteindre le résultat attendu :

- Criterion = 'GINI' : Pour mesurer le degré de désordre dans les différentes partitions.
- max_depth = None : Nous cherchons à obtenir un arbre avec une profondeur maximale
- random_state = 0 : Pour que l'expérimentation soit reproductible
- min_samples_leaf = None : La partition est validée quelque soit le nombre des observations

Une fois le paramétrage est terminé, la sortie de cette librairie permet d'afficher l'architecture de l'arbre maximal ainsi que conditions des partitions selon le critère de GINI. Nous obtenons en effet une représentation graphique de l'arbre de décision ci-dessous:

Figure 25: Arbre de décision maximale



Source : Élaboré par nos soins sous Python

Nous remarquons que l'arbre de décision est de grande taille avec des nœuds qui contiennent peu d'éléments ce qui pose la problématique de complexité de l'arbre due au sur-ajustement.

Pour remédier à ce problème, nous avons effectué un élagage de l'arbre.

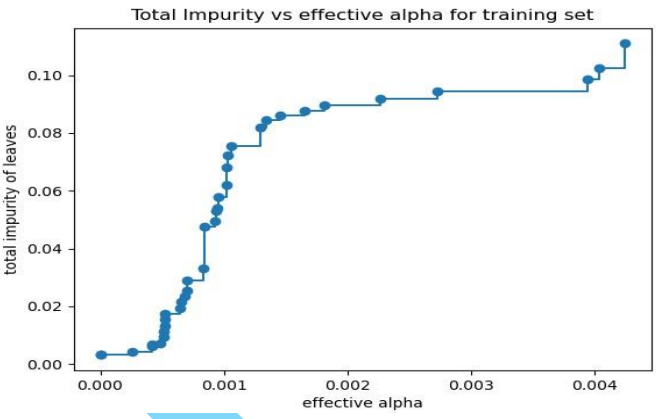
1.1 : élagage de l'arbre

Nous avons mentionné dans le chapitre précèdent que l'élagage s'inscrit dans le cadre de l'optimisation de la taille de l'arbre CART. Donc pour obtenir un arbre optimal, nous allons appliquer la technique d'élagage pour retirer les nœuds qui résultent du sur apprentissage.

Dans ce but, nous allons appliquer le paramètre Cost Complexity Metric, ccp_alpha de la librairie sktlearn..

Plus la valeur de ccp_alpha est élevée, plus le nombre de nœuds élagués est important.

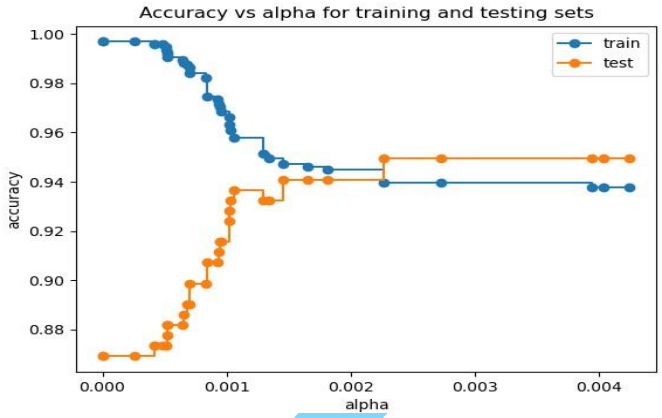
Figure 26: Le nombre des feuilles de l'arbre en fonction d'alpha



Source : Élaborée sous python

Plus l'arbre est élagué, plus l'alpha augmente, ceci indique qu'avec la diminution de la taille de l'arbre l'impureté totale de ses feuilles augmente.

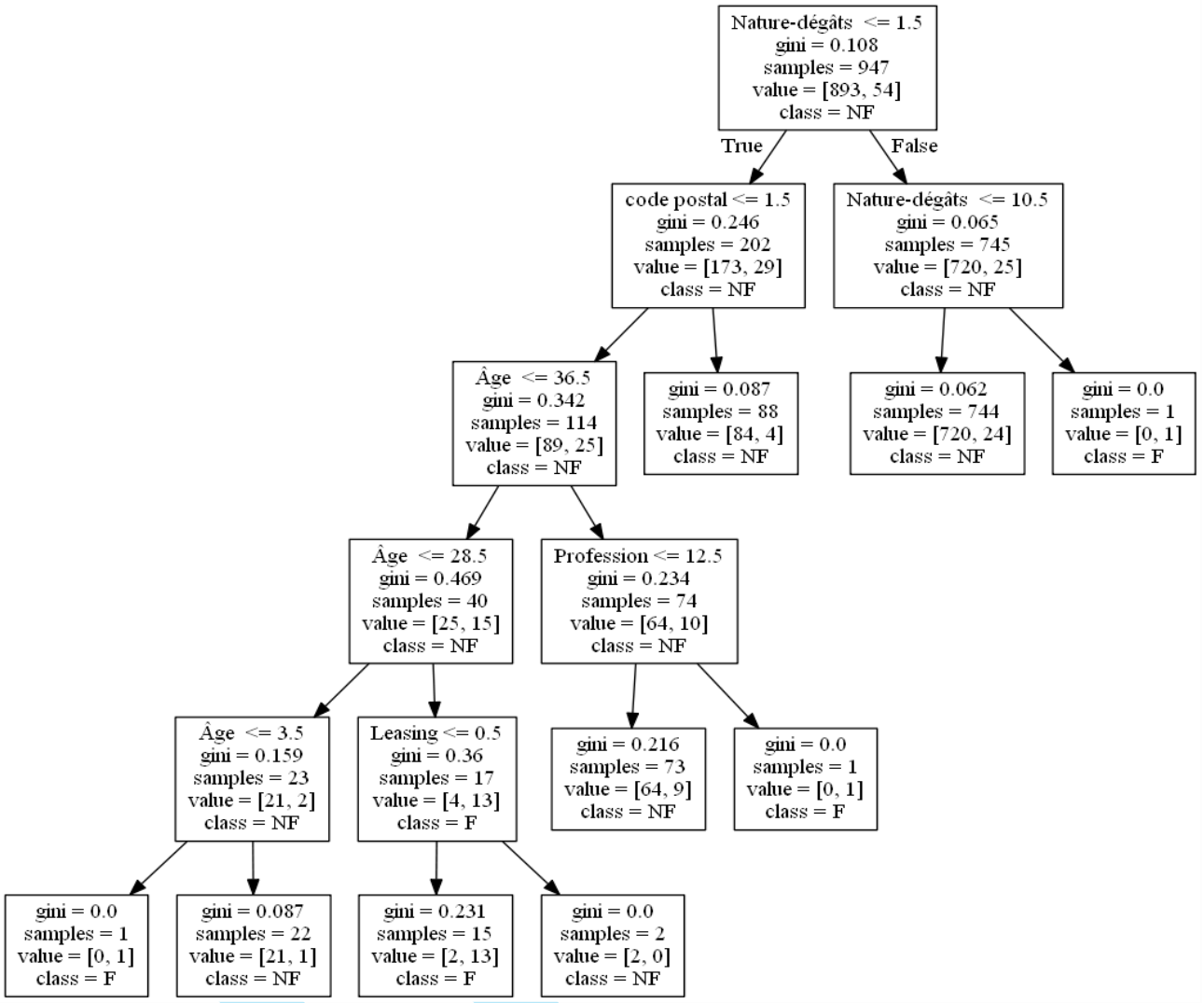
Figure 27: La précision du modèle en fonction d'alpha



Source : Élaborée sous python

La courbe au-dessus montre que lorsqu'alpha est établi à zéro, l'arbre est sur-ajustée. Il en résulte une précision de la base d'apprentissage de 100% et une précision de test de 86%. Au fur et à mesure que alpha augmente, une plus grande partie de l'arbre est élaguée, créant ainsi un arbre de décision qui se généralise mieux. Dans la figure au-dessus, à vue d'œil, alpha qui maximise la précision de la base test est égale à 0,022. Donc, nous allons construire un arbre de décision élagué avec alpha égale à 0.022

Figure 28: Arbre de décision élagué



Source : Élaboré par nos soins sous Python

3.1.3 Construction de la forêt aléatoire

Dans le chapitre précèdent, nous avons bien détaillé les étapes de construction d'une forêt aléatoire. Cet algorithme apporte la robustesse qui manque à l'arbre de décision. En effet, nous allons appliquer cet algorithme

La forêt aléatoire est implémentée avec la librairie scikit-learn sous Python « RandomForestClassifier » et doit être paramétrée par les hyper paramètres suivants :

- n_estimators : Le nombre d'arbres de décision construit
- oob_score : Estimer ou non l'erreur de généralisation OOB (Out of Bag).
- Bootstrap = True
- Criterion = "gini"
- random_state = 100

Contrairement aux arbres de décision, les forêts aléatoires ressemblent à une boite noire vu que les résultats sont difficilement interprétables

3.1.4. Évaluation en test

Après avoir implémenté les algorithmes d'induction de l'arbre de décision et de la forêt aléatoire, il est important de juger de leur qualité. Cette section présente alors les différents critères d'évaluation qui permettent de mesurer la pertinence de la classification.

3.1.4.1. Rappel sur les mesures de performance

- **La matrice de confusion**

La matrice de confusion est une mesure d'évaluation couramment utilisée en analyse prédictive, principalement parce qu'elle est facile à comprendre et peut être utilisée pour calculer d'autres paramètres essentiels tels que l'exactitude, le rappel, la précision, etc. Il s'agit d'une matrice qui décrit la performance globale d'un modèle lorsqu'il est utilisé sur un ensemble de données. Pour les données binaires nous avons une matrice de confusion 2x2, comme le montre la figure 29
Figure 29: Matrice de confusion

		La classe prédite	
		0	1
La classe réelle	0	Vrai négatif (TN)	Faux positif (FP)
	1	Faux négatif (FN)	Vrai positif (TP)

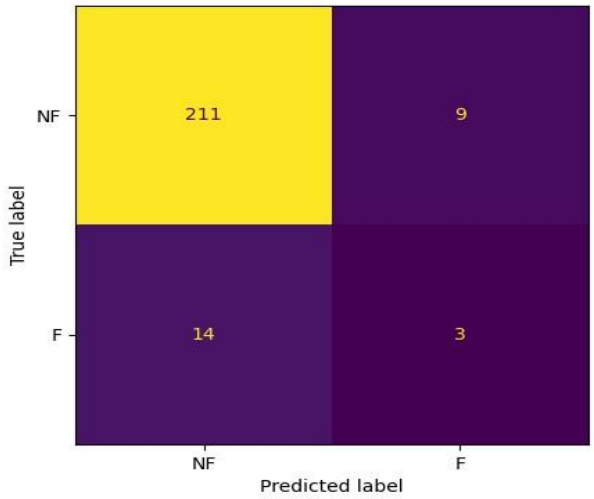
- **ROC (Receiver Operating Characteristic) et AUC (aire sous la courbe ROC)** La courbe ROC représente un graphique qui résume la performance du modèle prédictif et son pouvoir discriminant. Elle se construit en représentant sur l'axe des abscisses le taux des faux positifs (spécificité) , et en ordonnées le taux des vrais positifs (rappel), avec un seuil de classification qui varie entre 0 et 1. En effet, si cette courbe coïncide avec la diagonale, c'est que le modèle n'est pas plus performant qu'un modèle aléatoire, par contre, plus cette courbe est proche du coin du carré, plus le modèle sera meilleur, du fait qu'il permet de capturer davantage de vrais événements avec le moins possible de faux événements.

L'aire sous la courbe ROC donne un indicateur de la qualité de la prédiction (1 pour une prédiction idéale, 0.5 pour une prédiction « random »).

3.1.4.2. Résultats

Arbre de décision maximal

Figure 30: Matrice de confusion de l'arbre de décision avant élagage



Source : Élaborée par nos soins sous Python

- Vrai positif (TP) : c'est le cas où la classe réelle était positive (fraudeur) et la classe prédite est également positive.
- Faux positif (FP) : c'est le cas où la classe réelle était négative (non fraudeur) mais la classe prédite est positive
- Le vrai négatif (TN) : c'est le cas où la classe réelle était négative (non fraudeur) et la classe prédite est également négative
- Faux négatif (FN) c'est le cas où la valeur réelle était positive (fraudeur) mais la classe prédite est négative.

Tableau : Les mesures de performances de l'arbre de décision avant l'élagage

	Precision $\frac{TP}{TP+FP}$	Rappel $\frac{TP}{TP+FN}$	f1-score $\frac{2*(Precison*Rappel)}{Precision+Rappel}$	support	Taux de success (accuracy) $\frac{TP+TN}{TP+FP+FN+TN}$

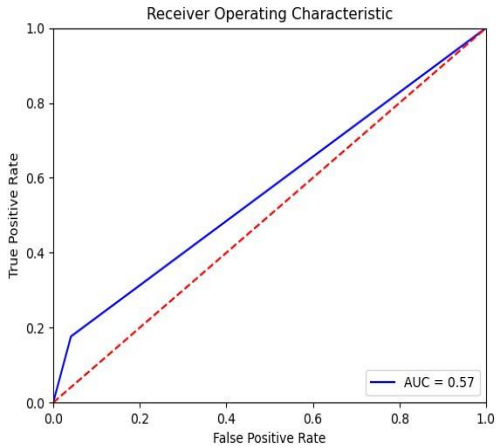
0 (non fraudeur)	0.94	0.96	0.95	220	**0.90**
1 (fraudeur)	0.25	0.18	0.21	17	

Source : élaboré par nos soins sous Python

La performance de ce modèle était satisfaisante lorsqu'il s'agissait de la classe des adhérents non frauduleux, où la précision et le rappel étaient de 0.94 et 0.96. Cependant, l'arbre de décision n'a pas bien réussi à classer les échantillons des adhérents frauduleux avec une précision, un rappel et des scores f1 de 0.25, 0.18 et 0,21 respectivement, comme le montre le tableau ….. Cela était attendu puisque l'échantillon est très réduit est déséquilibré.

En outre, nous remarquons que la courbe ROC ne coïncide pas avec la diagonale ce qui montre que le modèle n'est pas aléatoire. Cependant, nous constatons qu'il est loin du coin supérieur gauche donc le modèle n'est pas très performant. De même pour la valeur de AUC (area under curve), indiquée à la figure …,. Elle est de 0,57, cela signifie que le modèle n'a pas une bonne capacité de séparation des classes

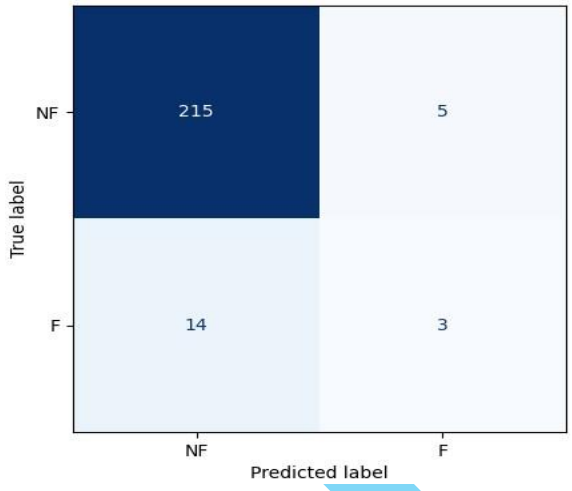
Figure 31: Représentation de la courbe ROC de l'arbre avant l'élagage



Source : Élaborée sous Python

- **Arbre de décision élagué**

Figure 32: Matrice de confusion de l'arbre élagué



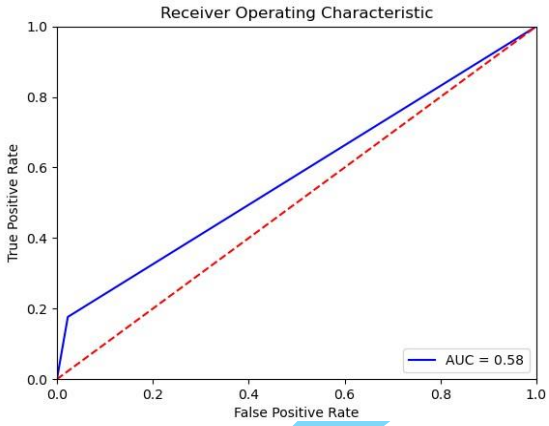
Source : Élaborée sur Python

Tableau : Les mesures de performances de l'arbre de décision après l'élagage

	Precision $\frac{TP}{TP+FP}$	Rappel $\frac{TP}{TP+FN}$	f1-score $\frac{2*(Precison*Rappel)}{Precision+Rappel}$	support	Taux de succès (accuracy) $\frac{TP+TN}{TP+FP+FN+TN}$
0 (non fraudeur)	0.94	0.98	0.96	220	0.92
1 (fraudeur)	0.38	0.18	0.24	17	

Source : élaboré sous Python

Figure 33: Représentation de la courbe ROC de l'arbre après l'élagage



Source : élaborée sous Python

Comme prévu, l'arbre de décision s'est amélioré légèrement dans la classification de la classe des frauduleux, lorsque l'élagage a été utilisé, ce qui est montré dans le tableau ….. la précision et le f-score ont progressé faiblement de 0.13 et de 0.03 respectivement.

Cependant, dans le cas de la classe négative, la précision est tombée légèrement de 2%.

De plus, la figure 33 montre une courbe ROC dans laquelle la AUC est de 0,58, ce qui n'est pas satisfaisant même après l'élagage. Ceci est dû au faible nombre d'observations en test, et de l'écart important avec l'apprentissage. En effet, non seulement le pouvoir prédictif est faible mais aussi le modèle se généralise mal. Les mauvais résultats s'expliquent par le fait que le nombre des adhérents fraudeurs est faible, nous pouvons améliorer ces résultats en appliquant une méthode d'agrégation comme le random forest

2.3. Limites et voies de recherches

Voulant construire des modèles prédictifs pouvant bien détecter les profils ayant une forte probabilité d'être frauduleux, plusieurs contraintes sont imposées :

- L'ensemble des données est fortement déséquilibré. Comme nous l'avons mentionné, sur les 1184 adhérents, seuls 71 sont des cas de fraude : à peine 6 %. Ce déséquilibre est dû en partie à la faible fréquence des fraudes détectées en général. De ce fait, la qualité de prediction des modèles appliqués est réduite, et sa prédiction n'est plus crédible à cause de l'absence de pertinence.
- Un autre point important est que les modèles peuvent perdre en performance lorsque les variables explicatives candidates pour modéliser le risque de fraude ne sont pas corrélées avec la variable cible (y), rappelons que les corrélations entre les variables explicatives et la variable à expliquer obtenues nous paraissaient faibles, d'où l'importance de sélectionner des variables plus discriminantes.

- L'autre facteur important qui peut affecter la performance des classificateurs était la limitation du jeu de données. Les résultats obtenus ont révélé des performances non satisfaisantes d'où la nécessité d'élargir le nombre d'observations.

- L'existence des valeurs aberrantes expliquées par des fautes de saisie des informations est le point le plus gênant de l'approche d'apprentissage. En effet, pour améliorer la robustesse du modèle une attention particulière doit être porter sur les variables qui comportent des données manquantes et aberrantes.

Section 4 : Recommandations

Tout au long de ce mémoire, nous avons essayé de cerner les caractéristiques et les comportements révélateurs d'un adhèrent frauduleux au moyen des analyses exploratoires des données. Eu égard

les résultats issus de ces analyses, nous proposons des recommandations qui vont déboucher sur des lignes de défense à caractère préventif et défectif. Celles-ci sont détaillées comme suit :

- L'absence ou la faiblesse du contrôle interne de l'organisation est parmi les facteurs favorisant l'acte de fraude (ACFE, 2018) donc nous recommandons la mise en place d'une architecture du dispositif de contrôle de fraude basée sur un enchaînement précis dont la première étape consiste à déployer des moyens de prévention qui permettent de s'intervenir en amont afin d'empêcher la fraude de se produire. Le point de départ de cette étape consiste à identifier et évaluer le risque en réalisant une cartographie des risques de fraude qui va aider le gestionnaire à fixer les actions nécessaires….
- L'efficacité d'un dispositif anti-fraude dépend de plusieurs choses, dont principalement l'adhésion et la responsabilisation des collaborateurs. Pour ce faire, ce dispositif doit être communiqué le plus largement possible via les différents réseaux de communication (intranet, E-mail…). En complément de ça, les collaborateurs doivent être bien formés pour identifier les comportements frauduleux au plus vite et à être capables d'agir en conséquence. Ceci nécessite la création d'un environnement propice à la prise de conscience des collaborateurs pour favoriser la gestion du risque de fraude, et ce, à travers les codes d'éthique ou de déontologie qui vont guider les collaborateurs dans l'exercice de leurs fonctions d'une façon honnête et intègre.
- La prévention peut également se traduire par la création d'une plate-forme de dénonciation ou d'un numéro vert qui offre au grand public la possibilité de signaler les soupçons de fraude sous couvert d'anonymat. Toute personne qui soupçonne certains actes frauduleux peut visiter un site web spécialisé et déposer une plainte. Par exemple, elle peut dénoncer un garagiste qui a proposé de gonfler la facture de réparation ou quelqu'un qui veut créer une complicité avec lui pour frauder l'assureur. Ce site va fournir une aide précieuse aux assureurs pour dépister les actes frauduleux.
- Comme vu précédemment, nous avons remarqué un nombre élevé de récidives dans des sinistres dont l'indemnité est inférieure à 1000 DT. Il s'agit d'un comportement anormal et inhabituel qui nous laisse penser que ces récidivistes sont frauduleux. En effet, pour réduire un tant soit peu ce type de réclamations frauduleuses, l'assureur est conseillé d'augmenter le niveau des franchises ou d'augmenter sensiblement les primes pour les assurés qui demandent une franchise étant donné qu'une franchise plus élevée implique une probabilité moindre de déclarer de petits sinistres et d'exagérer les devis. Faut-t-il noter que cette adaptation doit être retenue en considération de la

concurrence. L'assureur pourra ainsi développer des fichiers de fraudeurs qui enregistrent la fréquence des sinistres imputables à une même garantie ainsi que les personnes concernées afin de traquer les récidivistes.

- Nous avons constaté dans les dossiers frauduleux qu'il y'a quelques marques et modèles des véhicules assurés qui l'emportent sur les autres marques et qu'ils présentent également les mêmes points de chocs (avant du véhicule) au niveau des pièces de rechange facilement démontables. Dans ce cas, l'acte frauduleux consiste à changer les pièces intactes du véhicule assuré par des pièces accidentées avec la complicité du garagiste puis l'assuré déclare un faux sinistre (dérapage ou un accident contre un tiers non identifié) avec des fausses factures pour toucher une indemnisation indue sous la garantie tous risques. A la lumière de ce qui précède, l'assureur est conseillé d'insérer une clause qui prévoit la remise de toutes les pièces endommagées à la compagnie d'assurance pour éviter les faux accidents.
- Nous avons également relevé quelques dossiers étayés par des fausses factures de réparations établies par les assurés, d'une façon très soigneuse, sous le nom des garages fictifs. Sur ce point, nous suggérons d'opter pour des partenariats avec des garages agréés qui offrent des prestations en nature au lieu des versements en espèces. Cette solution permettra aux gestionnaires de sinistres d'orienter l'assuré vers un garagiste agréé par eux. L'objectif est d'éviter « les prestations en aveugle » et d'éloigner l'assuré de choisir un garagiste de son propre gré.
- Les experts désignés interviennent lors de la phase de gestion des sinistres pour constater les dégâts subis et enquêter sur les causes et les circonstances de déroulement de l'accident. Cependant, outre leur mission principale, l'assureur pourra attirer leur attention sur leur contribution notable à dépister les manœuvres frauduleuses et taguer les sinistres auxquels ils ont des doutes sur ses origines. Donc, afin de renforcer la détection de la fraude au niveau des experts l'assureur pourra prévoir un système de notation des experts
- À l'issus de notre requête sur les dossiers frauduleux, nous avons constaté que la majorité des contrats ont été résiliés immédiatement après le versement des indemnités, ce qui nous amène à penser que les fraudeurs vont absolument souscrire des polices d'assurance automobile auprès d'autres compagnies d'assurance et finiront par commettre des fraudes. Ces fraudeurs profitent de la quasi-absence de communication entre les compagnies d'assurances au sujet des dossiers frauduleux. Pour résoudre ce problème, il est donc important de lancer un consortium entre toutes

les compagnies d'assurances pour avoir une vision globale sur le risque de fraude. Tel est le cas de l'Association pour la Gestion des Informations sur le Risque en Assurance, AGIRA, qui a pour vocation la gestion des résiliations faites par les assurés ou les assureurs ainsi que leurs historiques.

- Comme nous venons de le voir, l'automatisation de la détection de fraude à travers des techniques sophistiquées de Machine Learning pourra aider le gestionnaire à contrer les tentatives de fraude à temps, elle permet ainsi de découvrir des « patterns » dont les gestionnaires ignorent l'existence. Cette solution permet d'optimiser la détection des dossiers suspicieux au moyen d'une plateforme rapide qui aide le gestionnaire à la prise de décision.

  L'analyse des données peut être appliquée pour détecter les dossiers potentiellement frauduleux. En analysant les fraudes passées, les assureurs peuvent utiliser la modélisation prédictive pour produire ce que l'on appelle un "score de suspicion", une valeur pour la propension à la fraude. Le processus fonctionne de la manière suivante : Les gestionnaires saisissent simplement les données et ils reçoivent automatiquement un score de suspicion indiquant la probabilité qu'une fraude ait eu lieu. La technologie utilisée pour ce faire implique l'utilisation d'outils d'exploration de données et l'application d'une analyse quantitative.

Conclusion

Dans cette optique, notre mémoire s'est penché sur la combinaison de l'approche classique et l'approche apprentissage automatique pour lutter contre la fraude, et ce, en adoptant deux pistes de recherche. La première s'est articulé autour du développement d'une approche classique basée

sur l'analyse approfondie d'un échantillon de dossiers frauduleux afin de cerner les caractéristiques d'un comportement frauduleux, cette analyse nous a permis d'identifier des « patterns » de manœuvres frauduleuses. Quant à la deuxième approche, elle a traité le risque de fraude par une méthode d'apprentissage automatique supervisé suivant une méthodologie d'extraction de connaissance à partir de données.

Dans le premier chapitre, certaines définitions ont été introduites. Le contexte général de la fraude, a été présenté. À ce titre, les éléments constitutifs de la fraude ont été définies, en spécifiant en particulier les principaux enjeux de la lutte contre la fraude d'une part, les sanctions encourues par les fraudeurs d'autre part. Nous nous sommes intéressés ensuite à l'étape de détection de la fraude en mettant en lumière les organismes qui participent à cette détection. Ensuite, nous avons abordé la démarche d'industrialisation de la lutte contre la fraude avant d'introduire la méthodologie adoptée afin d'y parvenir.

Dans le deuxième chapitre, nous avons fait un survol théorique sur les algorithmes les plus répandus pour effectuer l'apprentissage automatique tels que les arbres de décisions et les forets aléatoires qui vont être utilisés pour attribuer un statut à chaque demande d'indemnisation à travers l'émission des signes d'alerte.

Finalement, avant la mise en application des méthodes abordées dans le chapitre précèdent, nous avons effectué dans un premier temps, une analyse exploratoire des données recueillies auprès du département sinistre de la mutuelle assurance de l'enseignement (MAE), cette analyse nous a permis de capturer les anomalies, les valeurs extrêmes et les corrélations et les traiter pour les adapter à notre problématique.

La deuxième phase s'est focalisée sur l'approche classique d'exploitation de données, à l'issu de laquelle nous avons pu rassembler des critères de suspicion et des scenarios à travers des échantillons de dossiers potentiellement frauduleux.

Ensuite vient l'étape d'apprentissage, au cours de laquelle nous avons implémenté deux algorithmes de classifications (arbre de décision et forêt aléatoire) qui permettent de distinguer les assurés enclins à la fraude des assurés honnêtes et, in fine, de prédire l'existence d'une fraude dans une demande d'indemnisation déposée. Finalement la performance des deux modèles a été expérimentée sur la base de test pour évaluer leur contribution à ce mémoire. Cependant, les

résultats donnés par ces méthodes de classification n'étaient pas satisfaisants en raison de plusieurs limites explicitées dans la section précédente. Ces résultats confirment l'importance de la qualité des données par rapport à l'algorithme.

Finalement, bien que les algorithmes utilisés n'étaient pas performants, nous devons pas négliger le rôle grandissant du Machine Learning dans la détection de la fraude. Cela constitue une riposte automatique aux actes frauduleux. À ce titre, il est primordial pour les assureurs de comprendre les principes et les applications et du Machine Learning.