INSTITUT DE FINANCEMENT DU DÉVELOPPEMENT DU MAGHREB ARABE

معهد تمويل التنمية للمغرب العربي

Mémoire de fin d'Etudes

Thème :

Méthodes de Machine Learning pour la détection et la prédiction de fraude à l'assurance automobile

« Cas de la CAAT »

Présenté et soutenu par :

Afaf ANNANE

Encadré par :

M. Ghazi BEL MUFTI

Etudiant(e) parrainé(e) par :

Ministère des Finances Algérien

Dédicace

À mes chers parents

À mon cher mari

À mes chers frères

À mon adorable sœur

À mes amis(es) et ma famille

À toutes les personnes qui ont toujours cru en moi.

AFAF

REMERCIMENTS

À Mon encadrant Mr. Ghazi BEL MUFTI, pour son suivi, ses précieux conseils et ses orientations tout au long de ce travail.

À le personnel de la CAAT,

département Automobile dans la direction générale et régionale,

À L'ensemble de mes professeurs à l'IFID ainsi que le personnel administratif pour leur disponibilité et leurs services.

À Tous ceux et celles qui ont contribué d'une quelconque manière à l'élaboration de ce travail depuis la préparation jusqu'aux ultimes moments.

Un grand merci.

RESUME

La limitation de la couverture pour certains risques de l'assurance automobile a poussé quelques assurés de mauvaise foi à la recherche des moyens et des issues leur permettant d'être remboursés en cas de sinistre, ou seulement, d'être avantagés par rapport à la souscription d'un contrat d'assurance de façon irrégulière et frauduleuse.

A cause de cet acte, les compagnies d'assurance perdent, chaque année, des montants non négligeables. Ce qui impacte négativement sur leur rentabilité, leur concurrence, leur crédibilité. La fraude à l'assurance représente un risque croissant pour les assureurs. Il s'agit essentiellement d'un problème d'asymétrie d'information entre les parties contractantes.

Notre choix de ce sujet se justifie par l'importance de la lutte contre la fraude pour les assurés, les assureurs et l'économie nationale. Notre étude consiste à mettre en œuvre un modèle prédictif en comparant cinq algorithmes de classification appliquée sur une base de données qui représente l'ensemble des déclarations de sinistres douteux jugées frauduleuses ou non par ALFA.

Dans l'évaluation numérique, nous avons constaté que la meilleure performance a été obtenue par XGBoost avec un rappel de 86%, et un score F1 de 81%. D'autre part, l'évaluation graphique de la courbe ROC et de l'indice UAC montre que la régression logistique est la plus performante avec un UAC de 75%.

Les variables les plus pertinentes qui sont répétées dans chaque modèle sont le montant des dommages, le DS/DE en jours (le temps écoulé entre la date du sinistre et la date d'effet), la prime, la durée de la police en mois.

Pour obtenir des modèles plus performants, augmenter l'échantillon en termes de nombre de dossiers et de nombre de variables explicatives qualitatives et quantitatives sera la bonne solution.

Mots clés : fraude, assurance automobile, apprentissage supervisé, modèles prédictifs, algorithme de classement.

LISTE DES ABREVIATIONS

ALFA Agence pour la Lutte contre la Fraude à l'Assurance

CAAT Compagnie Algérienne des Assurances

CAAR Compagnie Algérienne d'Assurance et de Réassurance

SAA Société Nationale d'Assurance

E-C Enquête Classique

R.V.V Recherche des Véhicules Volés

F.C.F Fichier Central Des Fraudeurs

A.P.F Audit de Prévention Contre la Fraude

DA Dinar Algérien

SVM Support Vector Machine

RMSE Root Mean Square Error

KNN K-Nearest Neighbors

ID Identification for Development

MLP Perception Multicouche

RF Random forest

ROC Receiver Operating Characteristic

AUC Area Under the Curve

AI Artificial Intelligence

ML Machine Learning

Pd Pandas

Np Numpy

2D Deux Dimensions

SMOTE Synthetic Minority Over-sampling Technique

GLM Generalized Linear Model

CART Classification And Regression Trees

OOB Out Of Bag

TP True Positive

TN Vrai Négatif

FP Faux Positif

FN Faux Négatif

DC Dommage Collision

DR Défense - Recours

RC Responsabilité Civile

EPE Entreprise Publique Economique

SPA Société Par Action

2A Algérienne des Assurances (GIG Algeria)

CASH Compagnie d'Assurance des Hydrocarbures

CIAR Compagnie Internationale d'Assurance et de Réassurance

CNMA Caisse Nationale de Mutualité Agricole

GAM Générale Assurance Méditerranéenne

SALAMA Islamic Arab Insurance Company

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ANNEXES

SOMMAIRE

INTRODUCTION GENERALE

Frappant tous les pays et concernant la plupart des catégories d'assurance, la fraude est un ancien phénomène qui revêt divers modes opératoires.

Bien que la fraude est aussi ancienne que l'assurance elle-même mais 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, avec la directive européenne qui a intégré le risque de fraude comme un risque opérationnel.

Contrairement aux autres types d'infractions, la fraude n'est pas un délit que l'on peut déceler facilement. Un dossier frauduleux a des caractéristiques très semblables d'un sinistre réel, et il n'existe pas d'élément précis qui permettrait de le confirmer rapidement avec certitude.

Parmi les différentes activités des compagnies d'assurance, la branche dommage est celle qui est la plus soumise par la fraude et plus particulièrement sur les contrats automobile, où la fraude dans cette branche peut être une fausse déclaration à la souscription, un sinistre fictif, une exagération des montants des dommages...

L'ensemble de ces actes intentionnels de l'assuré dans le but de dégager un profit illicite du contrat d'assurance représente un coût non négligeable.

Selon l'Agence de la Lutte contre la Fraude à l'Assurance ALFA les taux de fraude vont à l'encontre du principe de mutualisation car cela impacte la tarification de l'ensemble des assurés et dégradent la rentabilité de l'assureur dans un secteur d'activité à une sinistralité élevée et un faible marge de bénéfice, en plus il est difficile de déterminer avec un degré de certitude la valeur globale des pertes liées à la fraude dans l'assurance. D'autre coté les entreprises cherchent à réduire les coûts pour rester le plus concurrentiel possible.

Face à ce risque de fraude, généralement, les compagnies d'assurance établissent des procédures de contrôle des dommages, qui vont d'une simple expertise jusqu'au transfert des dossiers suspects vers des services d'enquête spécialisés comme l'Agence de la Lutte contre la Fraude à l'Assurance ALFA.

L'analyse classique des données par les référents opérationnels de la fraude permet de créer des indicateurs appelés des règles métiers afin de remonter des alertes de suspicion. Les règles métiers sont considérées comme une réplication des cas de fraude déjà connus. Cette

approche traditionnelle adoptée par les assureurs se confronte à la grande quantité de dossiers à analyser, ce qui se traduit généralement par un temps de production des alertes de suspicion non compatible avec l'objectif de traitement d'un sinistre simple en moins d'une semaine.

Des mesures de prévention doivent être mise en place et des systèmes de détection doivent être crées par les compagnies d'assurances pour pouvoir combattre ce phénomène. Avec l'apparition des nouvelles technologies notamment le Big Data et la Data Science, l'utilisation des techniques de Machine Learning peut pallier à ce problème.

Ces techniques permettent d'identifier les corrélations complexes entre un grand nombre de variables et à détecter les signaux faibles de la fraude. En effet, les méthodes supervisées qui se basent sur un référentiel des fraudes identifiées sont efficaces pour repérer les typologies des fraudeurs et confirmer les assertions des experts et des gestionnaires de sinistres.

En littérature, certains travaux montrent l'efficacité de quelques modèles testés dans le cadre de la recherche de fraude à l'assurance.

Donc dans ce mémoire nous cherchons à mettre en place un système automatisé de détection des comportements frauduleux en utilisant les données que la compagnie possède et en faisant appel à des techniques prédictives. Cette approche est considérée comme un complément du modèle déterministe traditionnel et vise à aider à la décision sur le choix des dossiers à investiguer en priorité et à identifier les cas de fraudes cachés afin d'agir dans les meilleurs délais.

L'objectif de notre étude est de construire un modèle qu'a la possibilité d'effectuer une classification des dossiers de déclarations de sinistres selon deux classes dossiers frauduleux et dossiers non frauduleux. Donc notre problématique sera la suivante :

« Comment détecter et prédire un dossier sinistre frauduleux en assurance automobile ? »

Pour répondre à notre problématique, nous avons été amenés à structurer notre travail en trois chapitres :

Dans le premier chapitre nous tenterons de faire une synthèse analytique, nous présenterons les notions liées à la fraude en assurance automobile .Après nous allons exposer l'organisme, les politiques de détection et de prévention contre la fraude en assurance automobile en Algérie. Et nous finirons par une brève étude bibliographique dans laquelle nous

présenterons trois articles qui vise à modéliser la détection de la fraude par les méthodes de machine Learning.

Le deuxième chapitre traitera de la partie méthodologique et nous y définirons l'apprentissage statistique ainsi que les techniques de sélection des variables, d'équilibrage des données et le fondement théorique des cinq méthodes utilisées. Pour finir nous présenterons quelques méthodes d'évaluation de modèles prédictifs.

La méthodologie adoptée dans le cadre de notre travail est à la fois descriptive et analytique, descriptive, dans le sens où une présentation théorique est faite sur l'ensemble des méthodes de prédiction utilisées, et analytique à travers l'application empirique des cinq méthodes sur la base de données qui sont la régression logistique, l'arbre de décision, la forêt aléatoire, AdaBoost et XGBoost.

CHAPITRE 1 : ETAT DE L'ART DE LA FRAUDE EN ASSURANCE AUTOMOBILE

1 ETAT DE L'ART DE LA FRAUDE EN ASSURANCE AUTOMOBILE

1.1 Introduction

Avant de traiter le phénomène de la fraude, dans ce chapitre, nous allons commencer par les garanties offertes par les compagnies d'assurance en assurances automobiles, puis nous allons tenter de proposer une définition de la fraude en assurance automobile, les types et les causes et son impact sur la compagnie d'assurance.

Nous présenterons ensuite l'organisme, les politiques de détection et de prévention contre la fraude en assurance automobile en Algérie et enfin les principales sanctions encourues par les fraudeurs.

Toutefois, il est important de s'inspirer des études déjà réalisées dans ce domaine qui utilisent une nouvelle méthode de détection de fraude en assurance automobile qui est basée sur la machine Learning, c'est pourquoi nous exposerons les travaux de certains auteurs et la manière dont ils ont appréhendé la modélisation de la fraude en assurance automobile et les résultats qu'ils ont obtenus pour à la fin citer les méthodes que nous avons choisi d'appliquer.

1.2 Les notions liées à la fraude en assurance automobile

1.2.1 Les garanties d'assurance automobile

Selon le Recueil des guides de Gestion de L'assurance « Automobile » en Algérie :

- **La Garantie de base**

L'obligation d'assurance automobile existe depuis 1958 et concerne la garantie de la responsabilité civile automobile pour les dommages causés à autrui.

- **Les dommages causés au véhicule : Assurance "Tous risques"**

En cas de collision avec un autre véhicule, de choc contre un corps fixe ou mobile, ou de versement sans collision préalable, du véhicule assuré, la Compagnie garantit le paiement de la réparation des dommages que cet événement aura causé au véhicule assuré, ou aux accessoires ou pièces de rechange prévues dans le catalogue du constructeur.

- **Dommages – Collision**

En cas de collision survenant hors des garages, remises ou propriétés, occupés par l'assuré, entre le véhicule assuré et, soit un piéton identifié, soit un véhicule ou un animal domestique appartenant à un tiers identifié, la Compagnie garantit à l'assuré. Le paiement jusqu'à concurrence de la somme indiquée aux conditions particulières, de la réparation des dommages que cette collision aura causé au véhicule assuré.

- **Bris de glaces**

La Compagnie garantit l'assuré contre les dommages causés au pare-brise, à la lunette arrière et aux glaces latérales du véhicule assuré, par projection de cailloux, de gravillons ou autres corps. L'assurance s'exerce indifféremment que ledit véhicule soit en mouvement ou à l'arrêt.

- **Le vol**

L'assureur garantit, en cas de vol ou de tentative de vol du véhicule assuré :

- Les dommages résultant de sa disparition ou de sa détérioration

- Les frais engagés par l'assuré, légitimement ou avec l'accord de la compagnie pour sa récupération.

L'assureur garantit, en outre, les pneumatiques ainsi que les accessoires et les pièces de rechange, dont le catalogue de construction prévoit la livraison en même temps que celle du véhicule, s'ils sont volés dans l'une ou l'autre des circonstances suivantes ;

- Soit en même temps que le véhicule assuré

- Soit dans les garages, s'il y a eu effraction, escalade, usage de fausses clés, tentatives de meurtre ou violences corporelles

- **Incendies et explosions**

L'assureur garantit les dommages subis par le véhicule assuré et par les accessoires et les pièces de rechange dont le catalogue du constructeur prévoit la livraison en même temps que celle du véhicule lorsque ces dommages résultent de l'un des évènements suivants : Incendie, combustion spontanée, chute de la foudre et explosion à l'exclusion de celles occasionnées par tout explosif transporté dans le véhicule assuré.

- **La défense – Recours**

L'assureur garantit, à concurrence de la somme indiquée aux conditions particulières, le paiement de tous les frais d'avocat, d'expertise, d'enquête, de consultation, d'assistance et généralement de tous frais de procédure devant les juridictions civiles et pénales pouvant incomber à l'assuré du fait du véhicule automobile assuré.

1.2.2 Définition de la fraude en assurance

Il n'existe pas de définition exacte de la notion de fraude, elle peut être définie de plusieurs façons selon le contexte utilisé.

Dans le secteur d'assurance, L'ALFA, Agence pour la lutte contre la fraude à l'assurance a défini la fraude comme un : «Un acte intentionnel, réalisé par une personne morale ou physique, afin d'obtenir indûment un profit du contrat d'assurance»

Aussi d'après L'article 13 de la Directive Solvabilité (II), ils ont classé la fraude comme un risque opérationnel que les assurances doivent évaluer et combattre.

Dans l'assurance automobile ce type de fraude implique qu'une personne tente de tromper une compagnie d'assurance au sujet d'une réclamation impliquant son véhicule à moteur personnel ou commercial. Il peut s'agir de donner des informations trompeuses ou de fournir de faux documents à l'appui de la réclamation avec une mauvaise foi.

Nous concluons donc que pour qu'un acte soit qualifié de frauduleux en assurance, les éléments suivants doivent être réunis :

- ✓ L'action est volontaire et de mauvaise foi où l'acte et l'intention doivent être joints (l'un sans l'autre n'est pas un crime).
- ✓ L'objectif est d'obtenir une prestation, un avantage ou un profit illégal en exécution du contrat d'assurance sans droit.
- ✓ La perte monétaire réelle n'est pas nécessaire tant que le suspect a commis un acte et avait l'intention de commettre le crime.

1.2.3 Les types de fraude

La fraude peut prendre toutes les formes et toutes les ampleurs. Il peut s'agir d'un acte simple impliquant une seule personne ou d'une opération complexe impliquant un groupe de personnes à l'intérieur ou à l'extérieur de la compagnie d'assurance. D'après nos lectures, nous voyons qu'il n'y a pas qu'une seule classification de l'acte frauduleux, nous avons choisi cette classification qui est la plus utilisée au niveau de la société d'assurance :

- **Fraude à la souscription** : Le but du fraudeur dans ce cas est de réduire le montant de la prime. Certains assurés omettent des informations importantes ou donnent des informations erronées volontairement, elle compte parmi les fraudes les moins détectables.
- **Fraude au sinistre** : L'objectif du fraudeur est d'obtenir de son assureur un enrichissement sans cause, c'est-à-dire une indemnité supérieure à son droit. Dans ce cas, le fraudeur fournit une description incorrecte.
- **Fraude à la multi-assurance** : Le but du fraudeur est d'obtenir un cumul de remboursement par la souscription des contrats d'assurance couvrant le même bien chez plusieurs compagnies d'assurances. Ce type est le moins fréquent et rarement découvert par les assureurs.

Des exemples de fraude à l'assurance automobile (IFPA, 2022)

Les tentatives de fraude à l'assurance automobile les plus fréquentes sont les suivantes :

- Accidents de voiture mis en scène et fausses déclarations de blessures
- Fausses déclarations de véhicules volés
- Fausses allégations selon lesquelles un accident s'est produit après l'achat d'une police ou d'une couverture.
- Fausses réclamations pour des dommages qui existaient déjà.

1.2.4 Les causes de la fraude

Le triangle de la fraude est un cadre utilisé pour expliquer la raison derrière la décision d'un individu de commettre une fraude. En tant que théorie, cela nous aide à comprendre la motivation et l'état d'esprit des fraudeurs à l'assurance en général, et à l'automobile en particulier.

Et si nous pouvons comprendre pourquoi et comment la fraude se produit dans l'assurance, alors nous pouvons la limiter.

Ce modèle a été développé par le sociologue Donald Cressey en 1950, sur la base d'entretiens avec des personnes condamnées pour fraude, en essayant d'identifier les points communs à chaque cas de fraude.

Ce modèle montre que la commission d'une fraude est le résultat de la combinaison de trois éléments :

- **La pression :** Qui représente les motivations et les besoins qui poussent à commettre la fraude. Cela peut également être appelé "incitation". Il y a deux cas :

– L'Incitation personnelle : La raison la plus évidente pour laquelle un individu commet une fraude est sa situation personnelle. Cela peut être quelque chose d'aussi simple que le sentiment qu'il devrait gagner plus, ou des facteurs plus compliqués tels que des dettes personnelles ou pour alimenter la dépendance.

– Le plaisir comme récompense : Certains fraudeurs apprécieront simplement le processus d'obtention d'un peu plus que ce qui leur est dû. Ils pourraient être ravis de l'idée qu'ils « trompent le système ».

- **L'opportunité :** La fraude à l'assurance ne peut pas se produire si une opportunité ne se présente pas. Cela dit, ces opportunités peuvent être difficiles à atténuer et encore plus difficiles à maîtriser pour les assureurs.

  Une étude commandée par l'Association des assureurs britanniques a révélé que la plupart des cas de fraude opportuniste surviennent parce que les gens croient qu'ils ne se feront jamais prendre, ce qu'ils font n'est pas réellement un crime.

- **La rationalisation** : L'état d'esprit par lequel le fraudeur légitime son acte,

  Le dernier point du triangle de la fraude concerne le processus mental et les causes pour lesquelles un fraudeur estime que ses actions sont justifiées (tout en restant dans sa zone de confort moral). Certaines rationalisations sont plus répandues que d'autres, telles que :

– Adopter l'attitude tous les autres le font : La fraude à petite échelle est un phénomène si courant sur le marché de l'assurance que la personne moyenne peut penser qu'il est facile de s'en tirer.
– Percevoir un droit à une indemnisation en raison des primes payées.
– Penser qu'il a été maltraité : une personne qui estime qu'on lui manque de respect ou qui n'a pas obtenu le résultat qu'elle espérait peut croire que commettre une fraude est une façon de « se venger » de l'assureur.

  Briser le triangle de la fraude est l'élément clé de la dissuasion. Il faut pour cela supprimer un des éléments.

1.2.5 L'impact de la fraude en assurance automobile

L'assurance a pour objet d'indemniser l'assuré qui a subi un dégât, ou de lui redonner la même situation financière qu'avant le sinistre. L'assurance repose sur le principe de mutualisation et est conçue pour éviter les pertes importantes et incertaines. La fraude à l'assurance contourne ce système et les fraudes épuisent les fonds versés par les clients honnêtes pour couvrir les pertes réelles. Donc elle a des effets :

Sur la tarification des produits

La fraude n'est pas une faute négligeable, elle n'affecte pas seulement les compagnies d'assurances. En raison des primes d'assurance plus élevées, la plupart des titulaires des polices honnêtes paient en fin de compte la malhonnêteté des fraudeurs. La réalité c'est que les actes frauduleux affectent le prix des produits d'assurance qui ont, tendance à augmenter les primes. En effet la détection des fraudes permet de tarifier avec une manière plus précise et plus basse en fonction du risque réel des assurés honnêtes. Cela rendra les compagnies d'assurance qui ont détectés la fraude, plus compétitives sur le marché d'assurance.

Sur le provisionnement

La fraude à l'assurance force les compagnies d'assurance à régler leurs assurés pour des dommages dont le montant est beaucoup plus élevé que le montant réel, même pour des pertes qui n'ont pas eu lieu. Cela réduit la capacité d'investissement de la compagnie et l'oblige à fournir des provisions dépassant sa sinistralité réelle. La détection efficace de la fraude

permettra aux compagnies d'assurance d'allouer des provisions adaptables à leur situation réelle.

1.3 L'organisme, les politiques de détection et de prévention contre la fraude en assurance automobile en Algérie

1.3.1 L'organisme de lutte contre la fraude en assurance automobile en Algérie

L'assureur est le premier acteur de la chaîne de détection de la fraude car c'est généralement lui qui est le premier à être en possession des pièces susceptibles d'être analysées. L'assureur seul n'est, en revanche, pas en moyen de détecter et prouver la fraude. C'est dans cette optique d'aide à la détection des fraudes qu'a été créée l'Agence pour la lutte contre la fraude à l'assurance ALFA.

Présentation de l'organisme et les services offerts (ALFA) (SPLCFA, 2011)

L'agence de lutte contre la fraude est une société de service à vocation nationale, sa création est intervenue en tenant compte des recommandations du comité de pilotage du Ministère des Finances en 2004. Cette société créée en mars 2007 sous forme de société publique par actions, dont les actionnaires sont les trois entreprises d'assurance publiques leaders du marché des assurances en Algérie : la CAAT, la CAAR et la SAA, a démarré son activité qu'en juin 2009, au capital social de 50.000 000,00 DA, elle dispose de trois délégations « Est » « Ouest » et « Centre », elle offre des prestations de service dans le domaine de la lutte, la détection et la prévention contre la fraude à l'assurance. A ce titre, les services offerts par ALFA sont :

- **L'enquête classique "E-C"**

Il s'agit d'investigations menées par des enquêteurs conventionnés (ex-chefs de brigades ou commissaires de police) dont l'expertise est avérée dans le domaine de la recherche et de la lutte contre la fraude et le blanchiment d'argent. Souvent, il s'agit de vérifier les circonstances des sinistres déclarés susceptibles d'engager la garantie de l'assureur.

- **La recherche des véhicules volés "R.V.V"**

Il s'agit d'une offre de service dont l'objet est la recherche des véhicules déclarés volés au niveau des sociétés d'assurances. Ces recherches sont effectuées par les enquêteurs d'ALFA pour localiser l'emplacement des véhicules (fourrières, épavistes, garagistes, chez l'assuré, au niveau des services de sécurité ou chez un tiers). Ainsi, ALFA assiste les compagnies d'assurances pour récupérer ces véhicules

- **Le fichier central des fraudeurs "F.C.F"**

C'est une base de données, constamment actualisée, mise à la disposition des sociétés d'assurances adhérentes. Elle regroupe les éléments d'identification de personnes impliquées dans des affaires de fraude à l'assurance

- **Le fichier central sinistre "Plateforme"**

Cette plate-forme constituera un partage d'informations entre les compagnies adhérentes, dont l'objectif est de repérer les cas de multiplicité de déclarations et de connaitre aussi l'historique du véhicule

- **L'audit de prévention contre la fraude "A.P.F"**

Il s'agit de missions d'enquête menées au niveau de sociétés d'assurances pour évaluer le risque d'exposition à la fraude et le degré de fiabilité des procédures de souscription et de gestion des sinistres mises en place par les compagnies.

1.3.2 Les principaux politiques de prévention contre la fraude

Chaque compagnie d'assurance devrait allouer les ressources suffisantes pour trouver des techniques efficaces de prévention de la fraude tel que

- **La formation**

De manière générale la connaissance de la fraude dans la majorité des compagnies d'assurance n'est pas satisfaisante. Avec le développement rapide des pratiques frauduleuses, et qui constituent souvent des échantillons de créativité, il est nécessaire de mener une formation continue aux différents collaborateurs des compagnies d'assurance pour pouvoir suivre ces évolutions.

- **Suivre une bonne politique de gestion des sinistres**

Lorsque l'assuré déclare à son assureur un sinistre, il réclame une indemnisation, le gestionnaire de sinistre reçoit ces informations puis procède à la constitution d'un dossier sinistre. Les informations contenues dans ces documents permettent au gestionnaire de détecter une éventuelle fraude. C'est pour cette raison il faut adopter une politique plus rigoureuse dans la gestion des sinistres.

- **Recours aux enquêteurs disposé d'un processus de prévention**

En effet chaque compagnie d'assurance doit procéder à la mise en place des services d'enquête pour investiguer les cas suspects et analyser les données collectées sur les fraudeurs.

- **La mise en place d'un dispositif de lutte contre la fraude**

La mise en place d'un dispositif anti-fraude permet de prévenir contre la fraude, de détecter et de traiter des cas de fraude et de veiller à ce que l'ensemble de système atteint l'auto apprentissage de manière dynamique. Ce système repose sur 04 piliers (zouari,2021)

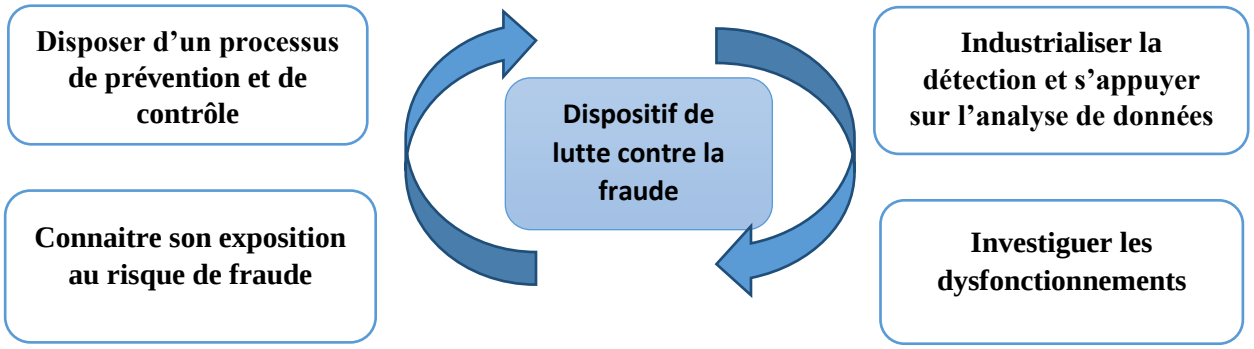


Figure 1: Dispositif de lutte contre la fraude

Source : Cours risque opérationnels

Pour détecter la fraude il faut disposer des indicateurs et des outils d'alertes (traces informatiques), réaliser des reporting réguliers pour assurer un contrôle permanant

La prévention se fait par le développement d'une culture éthique au sein de l'organisation, la formation de tous les intervenants sur le risque de fraude, la disposition d'un système de contrôle interne etc.

Pour pouvoir analyser les données, il faut commencer tout d'abord par la collecte des informations, ensuite la consolidation de ces informations et finir par l'analyse pour pouvoir comprendre le phénomène de la fraude.

L'investigation de dysfonctionnement se fait par l'analyse des défaillances et l'identification d'éventuels schémas répétitifs anormaux. Et finalement procéder à des mesures correctives.

- **La mise en place de processus de modélisation prédictive (CRISP-DM)**

Le processus de conduite d'une modélisation prédictive le plus utilisé par les entreprises et les analystes est le (CRISP-DM) pour Cross Industry Standard Process for Data Mining.

Notons que la construction d'un modèle prédictif ne se résume pas au simple choix de l'algorithme à appliquer, mais demande bien plus que ça. D'ailleurs, la réussite de n'importe quel projet sous-entend le passage par un processus précis, dans le domaine de l'apprentissage.

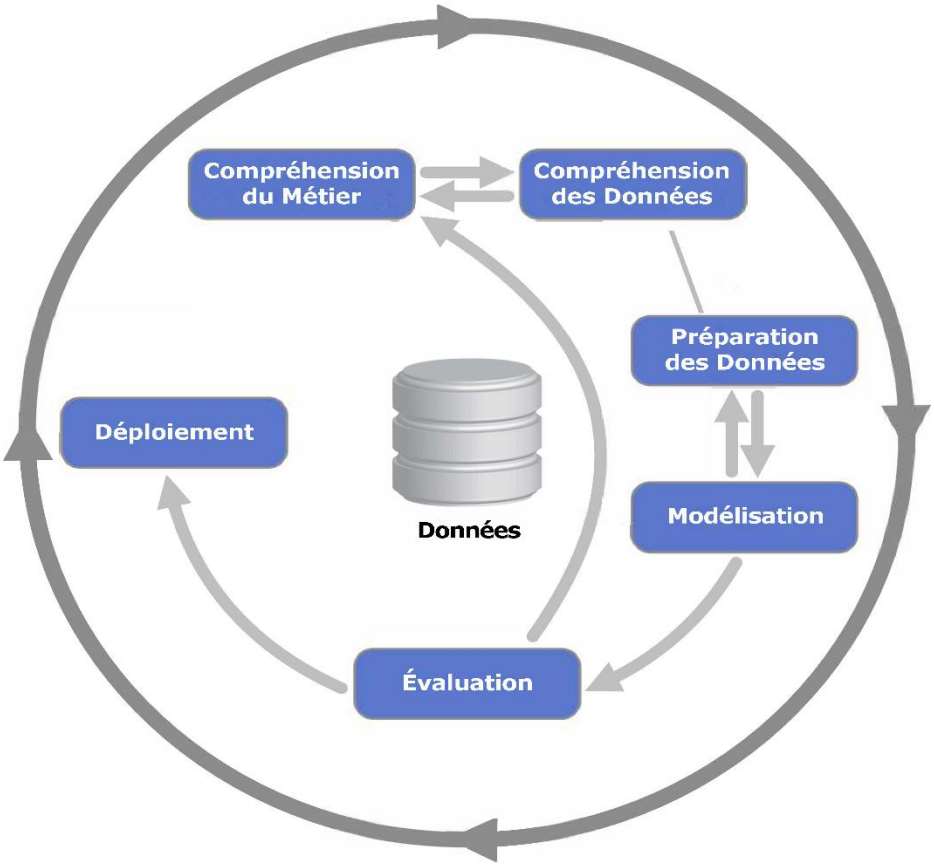


Figure 2: Processus du CRISP-DM

Source : https://www.datascience-pm.com/crisp-dm-2/

Il comporte six phases séquentielles (Hotz, 2022)

1. La compréhension du domaine: la conduite d'un projet d'analyse prédictive a pour objectif de répondre à un besoin exprimé par la firme tel que gagner de nouveaux clients, vendre plus de produits ou encore améliorer l'efficacité , ou détecté la fraude... des lors qu'au cours de cette première phase l'analyste doit être en mesure de comprendre le besoin qu'il traite afin qu'il puisse par la suite y remédier.

2. La compréhension des données : après avoir défini le but dans lequel l'analyse prédictive va servir, il est important d'avoir une compréhension fine des données issues des différentes sources.

3. La préparation des données : la création d'un modèle prédictif nécessite que les données soient organisées et structurées d'une manière spécifique. Cette phase rassemble toutes les activités de nettoyage et de prétraitement requises pour la conception d'un tableau de données bien formé à partir duquel des modèles d'apprentissages peuvent être créées.

4. La modélisation : dans cette étape plusieurs algorithmes d'apprentissage automatique sont sélectionnés donnant ainsi plusieurs modèles candidats dont le meilleur sera exploité lors du déploiement.

5. L'évaluation : avant l'adoption du modèle par l'organisation, il est judicieux de le soumettre à une évaluation minutieuse pour mesurer sa capacité de généralisation et s'assurer qu'il ne s'agisse guère d'un modèle souffrant de sur-apprentissage ou de sous-apprentissage.

6. Le déploiement : cette dernière phase correspond à l'intégration du modèle au sein de l'organisation et à sa mise à l'épreuve dans le monde réel afin d'accomplir la tâche pour laquelle il fut développé.

Dans le monde technologique actuel, la détection et la prévention de la fraude consistent à arrêter la fraude au moment où 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.

1.3.3 Les sanctions civiles et pénales contre les fraudeurs selon la loi Algérienne

Afin de lutter contre la fraude à l'assurance, la loi algérienne prévoit des sanctions contre ces fraudeurs, que ce soit sur le volet pénal ou civil

Les sanctions civiles de la fraude ou tentative de fraude à l'assurance

Conformément à l'article 21 de l'ordonnance 95-07 du 25 janvier 1995 modifiée et complétée par la loi 06-04, relative aux assurances : "Toute réticence ou fausse déclaration intentionnelle de la part de l'assuré ayant pour conséquence de fausser l'appréciation du risque de la part de l'assureur entraîne la nullité du contrat, sous réserve des dispositions prévues à l'article 75 de la présente ordonnance. On entend par réticence, l'omission volontaire de la part de l'assuré de déclarer un fait de nature à modifier l'opinion que l'assureur se fait du risque".

À titre de dommages et intérêts, les primes payées demeurent acquises à l'assureur et il peut réclamer à l'assuré le remboursement de l'indemnité déjà perçue.

La figure suivante montre les sanctions civiles de la fraude ou tentative de fraude à l'assurance selon la loi en Algérie.

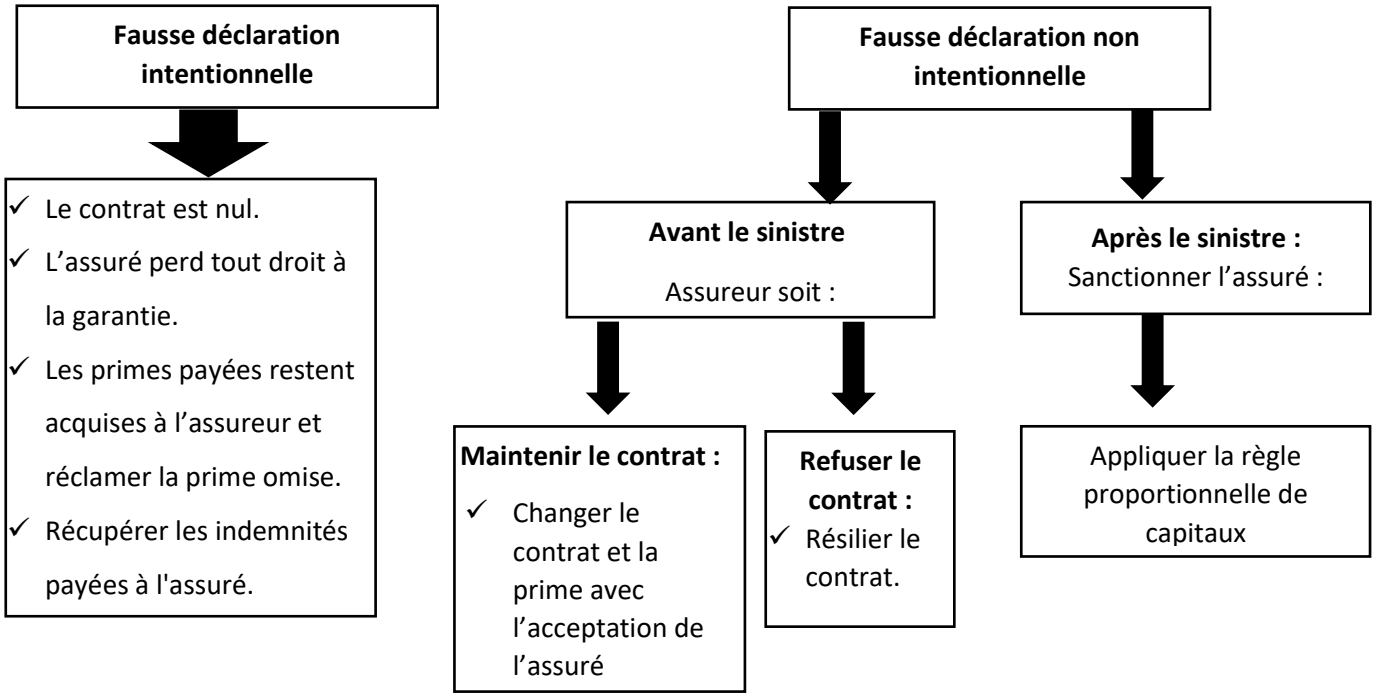


Figure 3: Sanctions civiles de la fraude ou tentative de fraude à l'assurance.

Source : réalisé par l'auteur.

Les sanctions pénales de la fraude ou tentative de fraude à l'assurance

Comme les autres formes de fraude, la fraude à l'assurance est un crime sanctionné par la loi, et selon l'article 372 du code pénal : « Quiconque, soit en faisant usage de faux noms ou de fausses qualités, soit en employant des manœuvres frauduleuses pour persuader l'existence de fausses entreprises, d'un pouvoir où d'un crédit imaginaire, ou pour faire naître l'espérance ou la crainte d'un succès, d'un accident ou de tout autre événement chimérique . Se fait remettre ou délivrer, ou tente de se faire remettre ou délivrer des fonds, des meubles ou des obligations, dispositions, billets, promesses quittances ou décharges, et, par un de ces moyens, escroque ou tente d'escroquer la totalité ou une partie de la fortune d'autrui est puni d'un emprisonnement d'un an au moins et de cinq ans au plus, et d'une amende de 500 à 20.000 DA.»

Nous pouvons donc conclure qu'il n'existe pas d'infraction spécifique à la fraude à l'assurance dans le droit algérien, par contre dans le code pénal algérien, la fraude est mentionnée sous tous ses aspects. Elle est considérée comme une escroquerie qui touche tous les domaines de l'activité d'une entreprise. Mais cette notion n'est pas spécifique à l'assurance. Il n'est pas répertorié par le législateur, il rentre dans le cadre général du délit d'escroquerie.

1.4 Etude bibliographique sur la modélisation de la fraude dans l'assurance automobile

Plusieurs chercheurs se sont orientés vers des techniques d'apprentissage statistique pour traiter cette problématique, ainsi que vers des variables clés permettant de mieux expliquer le phénomène, parmi eux :

Selon (Bouzgarne, Youssfi, Qbadou, & Bouattane, 2019) Leur article intitulé « *Performance comparative study of machine learning algorithms for automobile insurance fraud detection* »

Dans cette étude, les auteurs se sont concentrés sur la fraude en assurance automobile et la détection de comportements douteux. Les données de l'étude sont constituées par des sinistres déclarés d'une compagnie d'assurance.

Pour atteindre cet objectif, ils ont présenté une comparaison des dix algorithmes d'apprentissage automatique les plus fréquemment utilisés (Random Forest, AdaBoost M1, PART, J48, Multilayer Perceptron, Decision Table, Logistic, SGD, Naive Bayes, SVM). Et ils ont comparé leurs performances avec deux méthodes d'évaluation « F-Score » et « K-Score » ainsi que l'indicateur « Root Mean Squared Log Error » afin de juger la pertinence des résultats et pour déterminer laquelle est la plus appropriée pour la prédiction de la fraude.

Concernant l'ensemble de données, leur étude comprend 15420 demandes de remboursement de janvier 1994 à décembre 1996, avec 32 variables prédictives et une variable cible représentant "Fraude" et "Pas de fraude". Où14 497 échantillons légaux (94%) et 923 cas de fraude (6%). les variables de cette étude sont

- **Fiche de données personnelles de l'assuré** (âge, sexe, état civil, etc.)
- **Détails du contrat d'assurance** (type de police, catégorie de véhicule, franchises d'assurance, type d'agent, couvertures d'assurance, etc.)
- **Les circonstances de l'accident** (date de l'accident, lieu de l'accident, rapport de police déposé, témoin présenté, le responsable de la faute, etc.)
- **Autres donnés de l'assuré** (nombre de voitures, sinistres antérieurs, notation du conducteur, etc.)
- **Fraude constatée** (Oui ou Non) Qui est la caractéristique à prévoir.

Après la collection des données ils ont effectué un prétraitement des données pour améliorer les performances des techniques d'apprentissage automatique ensuite pour l'étape de la sélection des variables, ils ont proposé d'éliminer (7) caractéristiques qu'ils ont considéré comme non pertinentes telle que « le numéro de police, le numéro d'affiliation, les différentes

dates », après cette réduction les résultats de F-measure et de K-measure sont visiblement améliorés.

Et comme les données sont déséquilibrées où le pourcentage d'occurrences de fraude représente seulement 6% du total des sinistres déclarés, donc ils ont procédé au rééquilibrage de l'ensemble des données pour réajuster ce pourcentage et obtenir un meilleur résultat.

Les résultats de cette étude montrent que l'algorithme forêt aléatoire a la meilleure performance avec l'évaluation de la K-mesure et le meilleur score avec RMSE. Et a également le meilleur pourcentage de la Precision (de 19,7 %,) et de Recall (23,83%) c'est-à-dire la part de la fraude découverte parmi toutes les fraudes est de (23,8%).

A la fin, ils ont comparé la valeur de l'accuracy de leur étude de cas avec les résultats des recherches précédentes qui ont été effectué sur le même ensemble de données (une en 2010, et deux en 2015, et une en 2016 et l'autre en 2017). Les auteurs dans cet article ont obtenu le meilleur score.

Selon (Dhieb, Ghazzai, & Besbes, 2019) Leur article intitulé *« Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations »*

Dans cet article, ils ont proposé d'employer un nouvel algorithme d'apprentissage automatique, à savoir l'algorithme XGBoost (extreme gradient boosting), pour détecter automatiquement les réclamations frauduleuses et les classer en différents types de fraude.

Les performances de XGBoost sont comparées à celles de trois autres classifieurs utilisés dans la littérature pour la détection de la fraude dans les applications d'assurance, à savoir l'arbre de décision, le bayes naïfs et l'algorithme du plus proche voisin (KNN) .Fournir à tous les classificateurs les mêmes données pour l'entrainement et le test.

Ils ont supposé que la variable cible est catégorique (aucune fraude et autres types de fraude),

Les autres variables sont : ID de l'assuré, âge, sexe, état civil, somme assurée, prime, sinistre frauduleux, motif de la réclamation frauduleuse, ID du courtier, type de sinistre, code postal de l'assuré, dommages corporels, témoins, rapport de police disponible, montant total du sinistre.

Leur base de données contenant plus de 64 000 réclamations qui ont utilisé pour entraîner, valider et tester le classificateur. Huit classes sont retenues, où T(0) : fait référence aux réclamations non frauduleuses. Et ils ont utilisé trois types de sinistres liés à la fraude automobile : T(1) : « types de sinistres invalides », T(2) : « Pas de prime mais a un sinistre », T(3) : « Montant du sinistre frauduleux ». Les classes T (1+2), T (2+3), T (1+3), T (1+2+3) sont obtenues à partir de différentes combinaisons des trois types de fraude.

Le résultat de cette étude montre que les meilleures valeurs pour les quatre métriques définies précédemment sont obtenues par le classificateur XGBoost, mais il prend plus de temps d'apprentissage et d'évaluation par rapport à l'arbre de décision et aux bayes naïfs.
La performance de l'algorithme de l'arbre de décision est meilleure que celle des algorithmes de bayes naïfs et du plus proche voisin.
Et pour valider la performance de l'algorithme, ils ont présenté la matrice de confusion où le XGBoost a obtenu la meilleure performance avec une spécificité (accuracy) de 99.25%, une sensibilité (Recall) de 0.992%, une précision de 0.9928% et un F1-Scor de 0.9926%.
Tout ça montre que par le modèle de XGBoost même avec un ensemble de données déséquilibré, les réclamations frauduleuses et non frauduleuses sont parfaitement détectées.

Selon (Maula, Prasasti, Dhini, & Laoh , 2020) Leur article intitule « *Automobile Insurance Fraud Detection using Supervised Classifiers* »
Ce travail vise à détecter la fraude à l'assurance automobile en utilisant les méthodes de classification supervisées proposées qui sont la perception multicouche (MLP), l'arbre de décision C4.5 et la forêt aléatoire (RF).et à évaluer et comparer leurs performances par la matrice de confusion, la courbe ROC et des paramètres tels que la sensibilité…
Dans cette recherche, les données collectées sont des données du monde réel, qui représente l'ensemble des réclamations en assurance automobile d'une des compagnies d'assurance étatique en Indonésie de 2016 à 2018.
La base se compose de 1881 réclamations qui contiennent 32 fraudes et 1849 légitimes c'est-à-dire que seulement 1,7% des cas de fraude dans l'ensemble de données. A cause de ce fort déséquilibre d'ensemble des données qui est l'un des principaux problèmes de l'apprentissage qu'il provoque une grande précision pour les données de classe majeure et de faible performance pour la classe minoritaire. Pour cette raison, les auteurs ont décidé d'utiliser la technique de sur-échantillonnage des minorités synthétiques SMOTE et des méthodes de sous-échantillonnage.
L'ensemble d'entraînement après l'application de la technique de SMOTE se compose de 253 données frauduleuses et de 460 données légitimes.
Concernant les variables utilisées on a : le sexe de l'assuré, le rapport de police (Oui ou non), le type et le modèle de la voiture, prix de la voiture, âge de la voiture, la franchise, mois de l'accident, coût du sinistre, type d'accident, Classe (fraude ou non).
Après de nombreux essais avec une combinaison différente, les paramètres optimaux sont choisis pour obtenir le meilleur résultat. A la fin, les résultats montrent que : Le modèle RF

présente le plus petit nombre d'instances mal classé (7 cas contre 30 et 31 pour DT et MLP) et le plus grand nombre d'instances correctement classées avec une précision de 98,5 % (contre 93.6% et 93.4% pour DT et MLP) aussi, elle a fourni la valeur de spécificité la plus élevée qui est de 98,5%. La valeur AUC explique que tous les modèles sont bien classés.
D'après le résultat, il est prouvé que apprenants d'ensemble produisent une meilleure généralisation des performances par rapport à un apprenant unique.
A la fin de cet article, les auteurs ont proposé des recommandations pour améliorer la performance des modèles et faire face aux limites de recherche, en ajoutant plus d'instances de fraude et des caractéristiques telles que l'historique des sinistres passés, l'état civil du titulaire de la police, la cote de conduite et la police de base. Aussi, la mise en œuvre d'un plus grand nombre d'apprentissage d'ensemble tels que XGBoost, CatBoost et AdaBoost, et l'utilisation d'approches d'optimisation peuvent améliorer le modèle.

Suite à notre lecture de ces articles et d'autres, nous avons décidé d'appliquer les modèles suivants : La régression logistique est l'un des algorithmes les plus anciens et les plus fondamentaux pour résoudre un problème de classification et reste la méthode la plus utilisée dans le domaine des assurances pour prédire des variables qualitatives.
Nous avons également choisi de traiter l'arbre de décision, car il s'agit d'un élément important nécessaire au développement d'autres algorithmes d'apprentissage d'ensemble comme la forêt aléatoire. Selon les travaux de recherche précédents cette technique a eu la meilleure performance par rapport aux autres techniques classiques, ensuite nous avons traité l'algorithme d'ada boost qui est aussi basé sur l'algorithme de l'arbre de décision et à la fin nous avons appliqué un algorithme de XG Boosting qui est le plus récent par rapport à l'autre créé en 2016. Pour que nous puissions choisir l'algorithme le plus performant qui détecte mieux le dossier frauduleux.

1.5 Conclusion

Nous avons présenté la fraude à l'assurance automobile en rappelant ses types, ses causes et son impact. Aussi, nous avons donné l'ampleur des sanctions monétaires et pénales qui peuvent être appliquées en cas de détection. Au terme de ce chapitre nous concluons que 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, au niveau de

la fraude en assurance automobile l'organisme avec lequel les compagnies algériennes travaillent est ALFA.

Cet organisme a été créé dans l'optique d'aide à la détection des fraudes, ainsi il existe d'autres méthodes de détection et prévision comme la machine Learning. Ceci fera l'objet du chapitre suivant

CHAPITRE 2 : METHODES ET TRAITEMENTS

2 METHODES ET TRAITEMENTS

2.1 Introduction

Grâce aux nombreux apports théoriques dont les méthodes d'apprentissage ont pu bénéficier et développer des puissants outils de traitements informatiques la mise en œuvre de modèles prédictifs permet aux entreprises d'anticiper la survenue des évènements imprévus ce qui leur offre une marge de manouvre pour guider leur actions à venir.

Dans ce chapitre, nous présenterons la définition et les types d'apprentissage automatique ainsi que les bibliothèques clés, ensuite nous exposons la phase de sélection des variables et de rééquilibrage des données et les fondements théoriques des méthodes prédictives utilisées pour la détection de la fraude, enfin le dernier élément nous concentrons sur les métriques d'évaluation numériques et graphique.

2.2 Machine Learning

Définition

Le machine Learning est un sous-ensemble de l'AI. Ce terme a été introduit en 1959 par Arthur Samuel qui en donne la définition suivante « *l'apprentissage automatique permet à une machine d'apprendre automatiquement à partir de données, d'améliorer ses performances par un processus d'apprentissage et de fournir ensuite des résultats (prédictions) qui n'ont pas été explicitement programmés* » (Gérard, Gondran, Lacomme, & Samir, 2022)

En d'autres termes, la ML consiste à créer des algorithmes qui permettent aux systèmes informatiques (ordinateurs) d'apprendre à partir de données historiques au lieu d'être programmés. L'objectif est d'extraire des informations, et de créer des modèles qui peuvent fournir des prédictions sur des nouvelles données. Elle fait simultanément appel à plusieurs sciences à la fois, comme les mathématiques, les méthodes statistiques, l'informatique.

Types d'apprentissage automatique

Il n'existe pas un seul type d'algorithme d'apprentissage automatique. Ils sont classés essentiellement en trois grandes catégories :

1 L'apprentissage non supervisé « *est généralement utilisé pour décrire la structure des données ou pour y découvrir des modèles latents. L'objectif de l'apprentissage non supervisé est d'améliorer notre compréhension des données et d'en tirer des informations exploitables* » (Shailesh, 2019)

C'est-à-dire que l'apprentissage non supervisé est basé sur le concept de laisser la machine apprendre par elle-même les modèles cachés dans des ensembles de données non étiquetées.

On parle de données non étiquetées si nous n'avons que les entrées et pas de sortie, ce qui signifie qu'il n'y a pas de réponses correctes. Le but de ce type est de résoudre deux problèmes soit : par le regroupement (clustering) ou par les règles d'association.

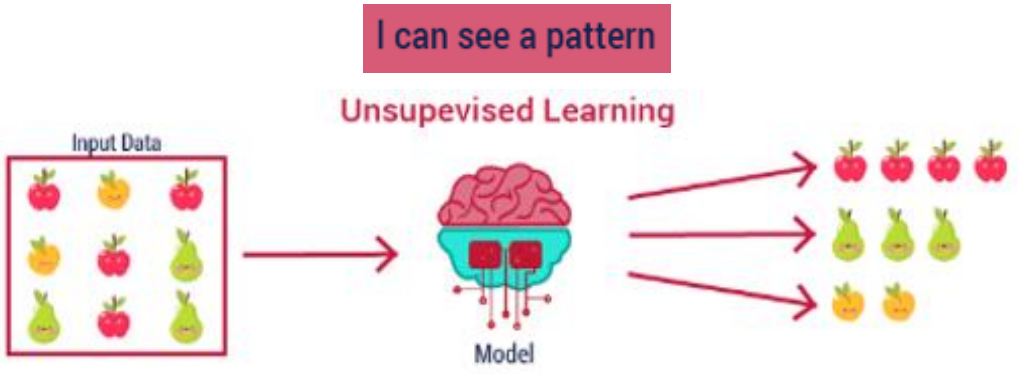


Figure 4: Apprentissage non supervisé

2 L'apprentissage supervisé *« est généralement utilisé pour apprendre une correspondance entre une observation et une prédiction menant à une décision. La plupart des systèmes de décision actuels dans divers domaines sont basés sur des modèles d'apprentissage supervisé ».* (Shailesh, 2019)

L'apprentissage supervisé repose sur le concept d'enseigner à un algorithme d'apprentissage automatique comment prédire à l'aide de données déjà étiquetées. C'est-à-dire, développer un modèle prédictif basé sur les données où nous avons la sortie correspondante pour chaque entrée.

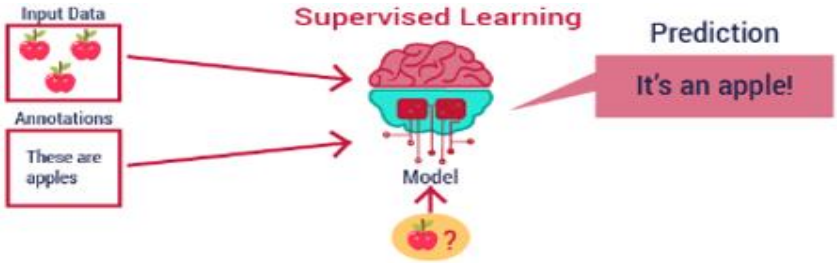


Figure 5: Apprentissage supervisé

On distingue principalement deux types de problèmes d'apprentissage supervisé : les algorithmes de classification et de régression.

- **Dans la régression** : On prédit une valeur numérique en fonction des données réelle observées précédemment. Où :

– La variable d'entrée et de sortie sont des valeurs continues.

– La fonction de mappage (Shailesh, 2019) est une fonction mathématique.

– Dans les problèmes de régression, nous pouvons utiliser la régression linéaire, multilinéaire et polynomiale. Concernant le choix d'un algorithme de régression par les data scientistes, il se fait à l'aide des représentations graphiques.

- **La classification :** C'est une technique pour prédire à quelle classe appartient notre cible en se basant sur la similarité des caractéristiques.

– La variable de sortie est une variable catégorielle. (se présentent sous forme de catégorie)

– La fonction de mappage est basée sur des probabilités (la mesure des similarités entre une paire d'objets).

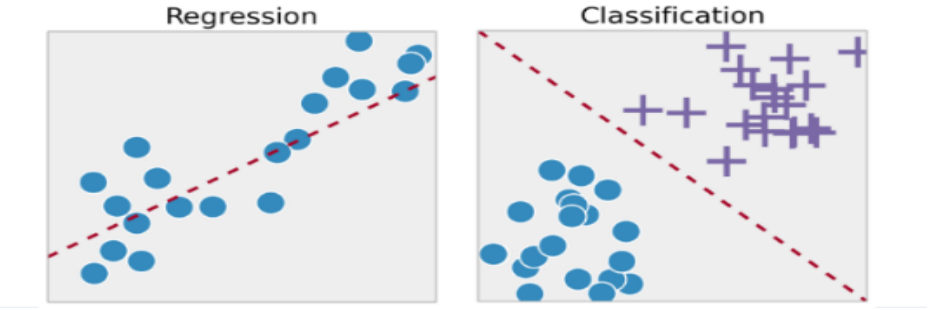


Figure 6: Deux types de problèmes d'apprentissage supervisé

3 L'apprentissage par renforcement : Plutôt que de simplement manipuler des données, les algorithmes d'apprentissage par renforcement fonctionnent en introduisant un logiciel, appelé agent d'apprentissage automatique, dans un environnement et en lui apprenant comment agir. Ainsi, l'agent apprend les actions à prendre, à partir d'expériences, de façon à optimiser une récompense quantitative au cours du temps.

2.3 Les bibliothèques clés de Machine Learning

Python a été créé par Guido van Rossum, et publié en 1991. C'est l'un des langages de programmation les plus populaires et le plus préféré chez les développeurs d'application et les data analystes en sciences de données et en intelligence artificielle.

Nous avons choisi d'utiliser Python car :

- Il est plus pratique avec une syntaxe claire, un langage de haut niveau c'est-à-dire écrits sous une forme proche de notre langage humain.
- Il prend en charge la gestion et la manipulation de données volumineuses et nous permet de créer des différents projets d'IA.
- Il nous offre une grande variété de bibliothèques et d'installations telles que :

– **Pandas (Pd)** est un outil d'analyse et de manipulation de données open source rapide, puissant, flexible et facile à utiliser.

- **Numpy (Np)** offre une large collection de fonctions mathématiques de haut niveau pour fonctionner sur de grands tableaux multidimensionnels comme les matrices.
- **Matplotlib** est construite sur des tableaux Numpy. Elle est conçue pour créer une grande variété de graphiques et de tracés en 2D pour la visualisation des données.
- **Seaborn** est construit sur Matplotlib et s'intègre avec les structures Pandas. Cette bibliothèque est aussi performante que Matplotlib. Elle introduit des types de graphiques supplémentaires et de haut niveau.
- **Scikit-learn** est une bibliothèque d'apprentissage automatique qui vise à la modélisation des données. Elle propose une large sélection d'algorithmes d'apprentissage supervisé et non supervisé, elle fournit également divers outils pour l'ajustement des modèles, le prétraitement des données, la sélection des modèles, l'évaluation des modèles et de nombreuses autres fonctionnalités.

2.4 La sélection des variables

2.4.1 Les tests de corrélation (Test de Khi-deux, Test ANOVA)

Dans une démarche d'un projet de Machine Learning, les tests d'indépendance permettent d'exclure des variables explicatives potentiellement non porteuses d'informations.

Il existe différents test d'indépendance qui permettent de définir s'il existe un lien entre deux variables. Le test appliqué dépend du type de deux variables traitées :

Test du khi 2

Il est utilisé pour tester l'hypothèse d'indépendance entre deux variables catégorielles (une variable qualitative et la variable cible). C'est-à-dire que si ces deux variables dépendent l'une de l'autre, la variation de l'une influence la variation de l'autre.

Le résultat de ce test est interprété par p-value qui indique la probabilité d'obtenir une valeur de la statistique du Khi 2. Les hypothèses du test sont donc :

H0 : Variables indépendantes si p-value > α

H1 : Variables non indépendantes si p-value < α

Test ANOVA

De la même logique que le test de Khi 2, nous pouvons utiliser le test ANOVA pour tester l'hypothèse d'indépendance entre deux variables : une quantitative et autre catégorielle (la variable cible) une analyse de variance. Les hypothèses du test sont donc :

• L'hypothèse H0 : p-value < α, il existe une relation de dépendance

• L'hypothèse H1 : p-value > α, il n'existe pas de relation de dépendance

2.4.2 Matrice de corrélation

C'est un outil pour résumer un grand ensemble de données .Il se présente sous forme de table contenant des lignes et des colonnes qui représentent les variables. Chaque cellule du tableau contient le coefficient de corrélation des variables deux à deux, plus le coefficient se rapproche des valeurs extrêmes (-1,1) plus la corrélation linéaire est forte de sens positif ou négatif. Plus les valeurs sont proches de 0 la corrélation est faible ou nulle (Zipporah, 2021). Quand les deux variables explicatives sont fortement corrélées. Seule la variable la plus discriminante doit être sélectionnée.

2.5 Techniques d'équilibrage des données

Souvent, lors du traitement des données, nous rencontrons le problème du déséquilibre, où la difficulté réside dans le fait qu'il existe une classe minoritaire ; c'est généralement elle qui est au centre des préoccupations. (Dans notre cas la déclaration frauduleuse en assurance auto).

Ce problème influence les performances de l'algorithme car le modèle tente d'apprendre uniquement la classe majoritaire. Il doit donc être résolu avant l'élaboration du modèle (Emmanuel, Vanessa, Emilie, & Bruna, 2021)

Il existe plusieurs techniques d'échantillonnage pour gérer l'équilibre des classes, notamment le sur-échantillonnage, le sous-échantillonnage.

Le premier type est les techniques de sous-échantillonnage qui consistent à réduire la taille de la classe majoritaire afin d'arriver à deux classes avec des tailles relativement proches ou bien égales.

Dans ce travail, nous avons choisi d'utiliser le deuxième type, une méthode de sur-échantillonnage qui peut être définie comme suit : « *Une approche pour gérer le déséquilibre consiste à générer (créer) des nouveaux données supplémentaires à partir de la classe minoritaire, afin de surmonter son manque de données* » (Elreedy & Atiya, 2019) l'une de ces méthodes est :

La technique SMOTE qui a été proposée par Chawla et al. Son idée principale consiste à générer de nouvelles données "synthétiques" en suivant ces étapes :

- Identifier une observation initiale x_1 de la classe minoritaire et son plus proche voisin. (ou ses k plus proches voisins).
- Tracer le segment joignant ce(s) voisin(s) et le point considéré,
- Prendre la différence entre la ligne.
- Multiplier la différence par un nombre aléatoire compris entre 0 et 1.

- Identifier un nouveau point sur le segment de ligne en ajoutant le résultat de multiplication à l'observation initiale x_1.
- Répéter le processus pour les observations identifiées.

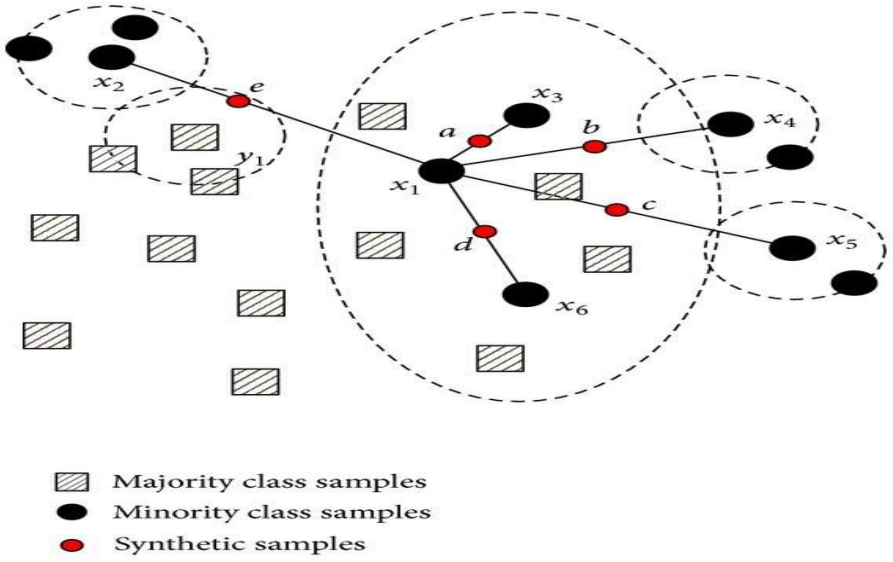


Figure 7: Technique SMOTE

Source : Data Augmentation Techniques in Deep Learning (Shamsudheen, 2020).

Cette technique ne fonctionne que pour les ensembles de données quantitatives, même avec l'encodage de données catégorisées sous forme numérique (par one-hot encoding ou d'autres encodeurs) lorsque nous appliquons le SMOTE, il générera de fausses modalités qui n'ont pas de signification ; par exemple, des chiffres avec une virgule.

Parce que l'encodage des variables catégorielles donne des classes finies, par exemple la classe 1 (dossier frauduleux) et la classe 0 (dossier non frauduleux).

C'est pourquoi nous devons utiliser SMOTE-NC lorsque nous avons des données mixtes.

Cet algorithme nous permet de traiter les variables catégorielles de manière plus adaptée sans faire du prétraitement. Il applique la même approche de la technique SMOTE pour le traitement des variables numériques (création d'un nouvel individu entre l'observation initiale et le plus proche voisin) mais pour les variables catégorielles, il attribue la modalité qui est majoritaire parmi les k plus proches voisins. Dans la suite de notre travail, nous utilisons la technique SMOTE-NC (Tremblay & Clément, 2022)

2.6 Les modèles prédictives : (principe, avantages, inconvénients)

2.6.1 La régression logistique

Nous pouvons considérer la régression logistique comme un cas particulier de régression linéaire généralisé (GLM) lorsque la variable cible est catégorique,

Cet algorithme décrit la relation entre une variable dépendante Y qualitative généralement binaire que nous cherchons à expliquer à travers p variables explicatives X = (X_1, . ., X_p) qu'elles soient qualitatives ou quantitatives avec un échantillon de n observations indépendantes

Elle est utilisée pour résoudre des problèmes de classification avec deux classes possibles (ou plus dans le cas de la régression logistique polynomiale), c'est-à-dire pour prédire à quelle classe appartient notre ensemble de données. Pour cela, nous allons utiliser l'équation linéaire comme une ligne qui sépare nos deux classes :

$$\hat{y} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Notons que la régression linéaire et la régression logistique son objectif est de déterminer les coefficients βj pour satisfaire l'équation ŷ.

β_i : Coefficient à estimer qui permet de mesurer l'influence de chaque variable et par la suite de déterminer les plus discriminantes.

ŷ : Les valeurs prédites.

Dans le cas de la régression linéaire ŷ appartient à ℝ une variable continue alors que pour la régression logistique ŷ appartient à [0,1] ce qui pose le problème de l'inégalité entre les deux parties de l'équation.

Pour cela, les statisticiens ont utilisé la fonction sigmoïde (inverse de la fonction logit) car elle a la même représentation graphique que la régression logistique: (Wiley Online Library, 2019)

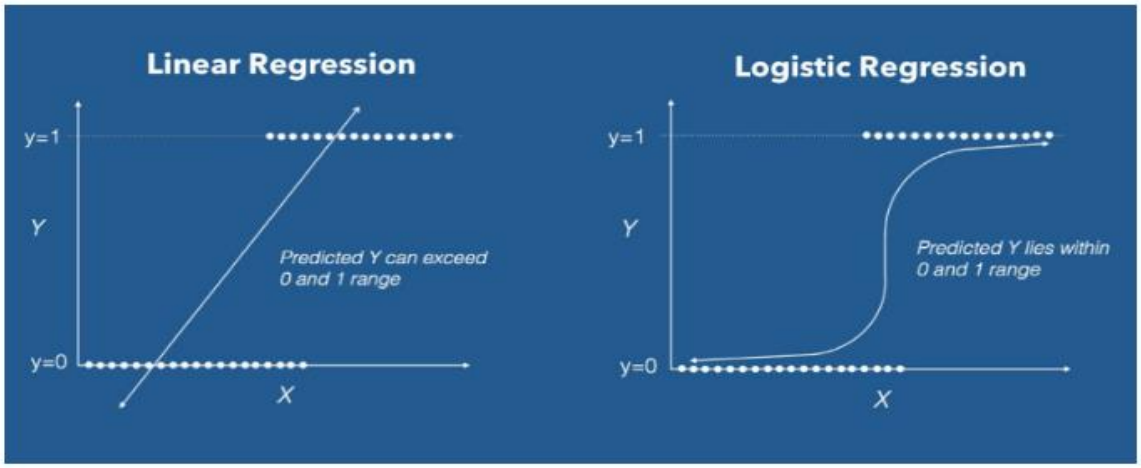


Figure 8: Représentation graphique de la régression linéaire contre la régression logistique

La fonction sigmoïde est utilisée pour transformer des valeurs sur (-∞, ∞) en des nombres sur (0, 1).

$$\hat{y} = g(\beta^T X) = \frac{1}{1 + e^{-\beta^T x}}$$

Plus ŷ est proche de 1 plus l'observation a de chance d'appartenir à la classe 1.

Plus ŷ est proche de 0 plus l'observation a de chance d'appartenir à la classe 0.

Le modèle Logit propose une modélisation de la loi de Y/X=x par la loi de Bernoulli de paramètre tel que :

$log\left(\frac{\pi(xi)}{1-\pi(xi)}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \beta_o + \sum_{i=1}^{P} X_i\, \beta_i = \beta^T xi$

Avec :

$$E(Y|X = x) = 1 \times P(Y = 1|X = x) + 0 \times P(Y = 0|X = x) = P(Y = 1|X = x) = \pi(x)$$

$$\pi(x) = P(Y = 1|X = x) = \frac{1}{1 + e^{-\beta^T x}}$$

Pour entraîner un modèle, nous devons toujours utiliser la fonction cout[1]. Dans le cas de la régression linéaire, la fonction cout est convexe (avec un seul minimum global ; si nous convergeons à partir de n'importe quel point, nous convergeons vers le même minimum) ; c'est pourquoi nous pouvons utiliser la décente du gradient[2].

Nous cherchons à minimiser cette expression :

$$J(\beta) = \frac{1}{2m}\sum_{i=1}^{m} (\hat{y}\text{i} \quad - y_i)^2$$

Mais dans le cas de régression logistique le changement de l'expression de modèle change la fonction de cout change selon $\hat{y}^i$ qui nous donne une fonction non convexe (un ou plusieurs minimums locaux) donc il faut trouver une nouvelle fonction cout convexe pour continuer de l'utilisation de la décente de gradient comme algorithme d'optimisation.

$$J(\beta) = \frac{1}{2m}\sum_{i=1}^{m} (\frac{1}{1 + e^{-\beta^T x^{(i)}}} - y_i)^2$$

La fonction de coût pour la régression logistique est proportionnelle à l'inverse de la vraisemblance des paramètres. Par conséquent, nous pouvons en obtenir une expression, en utilisant l'équation de log vraisemblance comme une fonction de coût vu que le but est de la minimiser. C'est à dire de maximiser le produit de la $P(y_i|X = xi)$

[1]La fonction cout qui est la moyenne des erreurs de notre algorithme

[2] Le but est pour chaque itération d'algorithme d'apprentissage de mettre à jour notre paramètre Bi à fin de soit plus proche de minimum globale.

$$L(\beta) = \prod_{i=1}^{n} P(y_i|X = xi) \forall i \in [1, \mathrm{n}]$$

Nous remplaçons la probabilité avec son expression binomiale (fonction de densité). On obtient :$L(\beta) = \prod_{i=1}^{n} \big(\pi(\mathrm{x_i})\big)^{y_i} \big(1 - \pi(\mathrm{x_i})\big)^{1-y_i}$

Log $(L(\beta))$= $\sum_{i=1}^{n} y_i \, Log\left(\frac{\pi(\mathrm{x_i})}{1-\pi(\mathrm{x_i})}\right) + Log\big(1 - \pi(\mathrm{x_i})\big)$

$$= \sum_{i=1}^{n} y_i log\left(\pi(\mathrm{x_i})\right) + (1 - y_i) Log(1 - \pi(\mathrm{x_i}))$$

A partir de cette équation de log vraisemblance, nous concluons la fonction de cout qui est

$$J(\beta) = -\frac{1}{2m} \left(\sum_{i=1}^{n} y_i log\left(\pi(\mathrm{x_i})\right) + (1 - y_i) Log(1 - \pi(\mathrm{x_i}))\right)$$

Pour trouver la valeur de béta, on calcule la dérivé partiel on obtient

$$\frac{\partial J(\beta)}{\partial \beta} = \sum_{i=1}^{n} (\pi(xi) - yi).xi$$

Avantage de La régression logistique

- La régression logistique est considérée comme l'une des méthodes de classification binaire les plus fiables.
- Facile à interpréter, on compare avec le seuil 0,5. Tout ce qui est supérieur à 0,5 sera considéré comme un vote pour la classe 1, et ce qui est inférieur (ou égal) à 0,5 sera considéré comme un vote pour la classe 0.

Inconvénients de la régression logistique

- Dans certains cas, elle nécessite d'effectuer un grand nombre de prétraitements et de visualisations de données avant l'application de modèle : elle ne traite pas les valeurs manquantes aussi les prédicteurs doivent être linéairement indépendants, nous réalisons une étude de corrélations lors du traitement de la base pour sélectionner les variables significatives, nécessite la normalisation des données.
- Elle a des résultats limités puisqu'elle ne peut prédire que des cibles catégorielles.
- Dans un nombre très important de données, la régression devient instable dans certains cas.

En raison de ces inconvénients, nous avons également choisi d'appliquer l'algorithme de l'arbre de décision dans notre cas pratique pour choisir le modèle le plus performant, cet algorithme ne nécessite pas le prétraitement des données.

2.6.2 Arbre de décision

Créé par Morgan et Sonquist, en 1963, l'arbre de décision est un modèle d'apprentissage non paramétrique qui représente un ensemble de décisions hiérarchisées pour obtenir un résultat finale qui peut être une explication ou/et une prédiction par la construction d'un arbre de régression ou d'un arbre de classification.

Donc l'objectif est de sélectionner parmi les variables explicatives qui peuvent être quantitatives ou qualitatives (X 1, ..., X p) celles qui sont les plus discriminantes pour la variable cible Y (Bel Mufti, 2021)

Pour construire l'arbre de décision : À chaque nœud, nous allons essayer de créer une division binaire, puis de sélectionner la caractéristique qui nous donne la division binaire optimale.

Si nous visualisons l'ensemble des décisions, nous pouvons clairement voir la forme de l'arbre. Mais il est à l'envers. Cette arbre est composé de :

Les composants d'arbre de décision

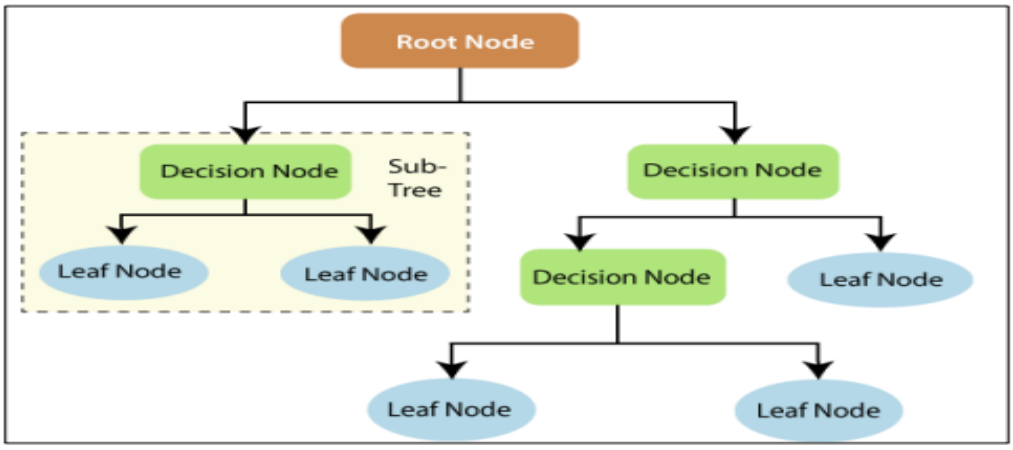


Figure 9: Composants d'arbre de décision

Source: Classification Based on Decision Tree Algorithm for Machine Learning (Bahzad & Adnan Mohsin, 2021)

- **Un nœud racine (Roodnode)** qui contient toutes les observations de notre base.
- **Un nœud non terminal (Decision note)** qui représente un test sous forme de questions sur nos caractéristiques qui les découpe en deux parties : les individus vérifiant la condition du nœud sont affectés à la branche gauche et les restants à la branche droite.
- **Un sous arbre (Sub_tree)** qui est tout arbre obtenu en élaguant arable initial obtenu en supprimant d'un nombre arbitraire de ses nœuds non terminaux.
- **Les nœuds terminaux /feuilles (Leafnode)** qui représentent une classification ou une décision finale.

Il existe plusieurs algorithmes associés aux arbres de décision, l'un de ces algorithmes créé par Breiman en 1984 est connu sous le nom de CART (classification and regressiontrees); le principal avantage de cette méthode est la possibilité de modéliser des variables qualitatives ou quantitatives.

Le processus de l'algorithme CART s'étend sur deux phases

- La première consiste à construire un arbre maximal, qui définit la famille de modèles au sein de laquelle nous chercherons à sélectionner le meilleur.

  Pour déterminer la distribution la plus optimale, nous devons calculer combien chaque distribution nous coûtera en termes de précision. Pour ce faire, nous utiliserons une fonction de coût.

  La distribution qui coûte le moins cher est choisie car nous voulons toujours maximiser notre précision. Il existe de nombreuses fonctions de coût que nous pouvons utiliser, mais la plus courante est l'indice de Gini.

  L'algorithme CART utilise l'indice de Gini, dont la formule est la suivante :

$$G(p) = 1 - \sum_{k=1}^{k} P_k^2$$

  Il quantifie la pureté du nœud/de la feuille. Un indice supérieur à zéro implique que les échantillons de ce nœud appartiennent à des classes différentes.

- La seconde étape est dite Élagage (Pruning) qui est l'étape finale de la création de notre arbre est le raccourcissement de ses branches. Ce processus s'appelle l'élagage et nous l'utilisons pour éviter le sur-ajustement.

  L'élagage peut être effectué en réduisant le nombre de nœuds feuilles ou en réduisant la profondeur de l'arbre pour construire un arbre généralisé et optimal tout en tenant compte du taux d'erreur.

  Dans cette étude, nous nous intéresserons à l'arbre de classification pour but d'expliquer une variable qualitative (la fraude) par des variables qui peuvent être qualitatives ou non.

  Les paramètres utilisés pour ce modèle sont (Scikit-learn, s.d.):

- Critère "gini", "entropie", "log loss", défaut="gini" : la fonction pour mesurer la qualité d'une division. Les critères pris en charge sont "gini" pour l'impureté Gini et "log loss" et "entropy" pour le gain d'informations de Shannon.
- Max depth entier, par défaut=Aucun : la profondeur maximale de l'arbre. Si aucun, les nœuds sont développés jusqu'à ce que toutes les feuilles soient pures ou jusqu'à ce que toutes les feuilles contiennent moins de min samples split échantillons.

- Min samples split int ou float, default=2 : le nombre minimum d'échantillons requis pour diviser un nœud interne :

1. Si int, alors on considère min samples split comme le nombre minimum.

2. Si flottant, alors min samples split est une fraction et représente le nombre minimum d'échantillons pour chaque fractionnement. Ceil (min samples split * n samples)

- Min samplesleaf entier ou flottant, par defaut=1.Le nombre minimum d'échantillons requis pour être à un nœud feuille. Un point de partage à n'importe quelle profondeur ne sera pris en compte que s'il laisse au moins min samplesleaf des échantillons d'apprentissage dans chacune des branches gauche et droite.

Cela peut avoir pour effet de lisser le modèle, notamment en régression.

1. Si int, alors on considère min samplesleaf comme le nombre minimum.

2. Si float, alors min samplesleaf est une fraction et est le nombre minimum d'échantillons pour chaque nœud. ceil(min samplesleaf * n samples)

Les avantages de l'arbre de décision

- Il ne nécessite pas beaucoup de prétraitement car il a la capacité de traiter des données brutes: il peut traiter des données numériques et catégorielles. Il n'est donc pas nécessaire de transformer les variables et la normalisation, il a la résistance aux valeurs manquantes, l'étude des corrélations est facultative, cela nous fait gagner du temps.
- Facile à comprendre et à interpréter grâce à sa forme. À chaque nœud, nous pouvons voir exactement quelle décision prend notre modèle.

Les inconvénients de l'arbre de décision

- L'arbre manque de robustesse car les modèles obtenus dépendent fortement de l'échantillon. On constate souvent des différences importantes entre les arbres obtenus par des échantillons différents. Les variables sélectionnées peuvent changer d'un échantillon à l'autre.
- Ils peuvent dans certains cas devenir extrêmement complexes et exposer au sur-apprentissage ce qui ne leur permettra pas une bonne généralisation.

2.6.3 Les techniques d'ensemble d'apprentissage

Leo Breiman était conscient du manque de robustesse des arbres de décision expliqués précédemment, qui nuit parfois à la qualité de l'apprentissage. Plutôt que de lutter contre ce défaut, il a eu l'idée d'utiliser des approches de type bagging et forêt aléatoire.

C'est-à-dire utiliser des techniques d'ensemble d'apprentissage qui consistent à entraîner plusieurs sous-modèles d'apprentissage automatique qui ont un faible niveau de performance

pour créer le modèle le plus performant possible par la combinaison entre eux afin que la force de ce modèle compense la faiblesse de l'autre.

Ce concept est largement utilisé, on le trouve non seulement dans le bagging mais aussi dans le boosting…

2.6.3.1 Bagging

Avant la création de la forêt aléatoire, Leo Breiman (1996) a inventé premièrement le concept de bagging.

- La première étape du bagging consiste à appliquer le bootstrapping, qui est une technique d'échantillonnage (Efron 1979) permettant de créer des sous-ensembles aléatoires avec N échantillons à partir de l'ensemble de données original. Les N échantillons sont choisis par des tirages avec remise.
- La deuxième étape est le Bagging Aggregating consiste à appliquer un algorithme à chaque sous-ensemble aléatoire. La moyenne de toutes les prédictions des différents modèles est prise comme résultat final pour des variables quantitatives ou le vote majoritaire pour des variables qualitatives.

Algorithm : Bagging

Entrée B le nombre de classifieurs.

- Faire pour k = 1 à B.
- Tirer un échantillon bootstrap.
- Estimer ϕ k sur l'échantillon bootstrap.

Fin pour

- Calculer l'estimation moyenne pour des variables quantitatives.

$$\phi = \frac{1}{B}\sum_{k=1}^{B} \phi_k$$

- Prendre le vote majoritaire parmi les φk pour des variables qualitatives.

La forêt aléatoire

Peut être considérée comme plusieurs arbres de décision entraînés ensemble par la "bagging", avec une étape supplémentaire.

Dans une forêt aléatoire, nous prenons une sélection aléatoire de caractéristiques plutôt que d'utiliser toutes les caractéristiques pour faire croître les arbres. Cet algorithme peut être utilisé pour des problèmes de régression ou bien de classement.

La figure suivante montre la différence entre le bagging et le foret aléatoire :

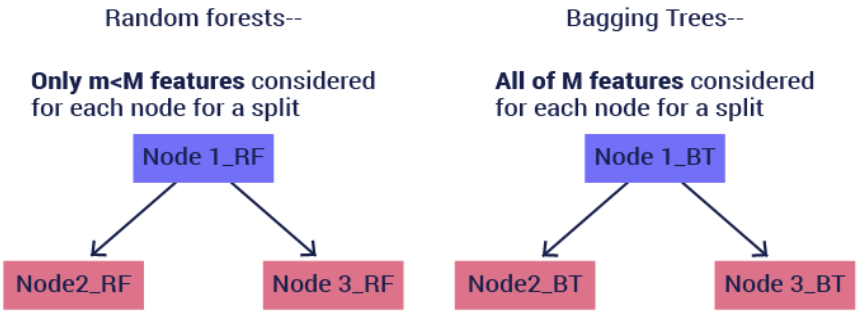


Figure 10: Différence entre le bagging et le foret aléatoire

Étapes de la mise en œuvre d'une forêt aléatoire

Supposons qu'il y ait N observations et M caractéristiques dans l'ensemble de données d'apprentissage.

- Tout d'abord, un échantillon de l'ensemble de données d'apprentissage est pris au hasard avec remise.
- Un sous-ensemble de M caractéristiques est sélectionné aléatoirement et la caractéristique qui donne la meilleure division est utilisée pour diviser le nœud de manière itérative.
- L'arbre croît jusqu'au plus grand niveau possible.

Les étapes ci-dessus sont répétées et une prédiction finale est donnée sur la base de l'agrégation des prédictions d'un nombre n d'arbres.

Algorithm : forêt aléatoire

Entrée B le nombre de classifieurs.

s le nombre de variables aléatoires.

Faire pour k = 1 à B.

Tirer un échantillon bootstrap.

Tirer un nombre s de variables aléatoires.

Construire un arbre de décision binaire ϕK sur l'échantillon bootstrap et les s variables.

Fin pour

Calculer l'estimation moyenne : $\phi = \frac{1}{B}\sum_{k=1}^{B} \phi_k$, pour des variables quantitatives.

Prendre le vote majoritaire parmi les ϕ_K pour des variables qualitatives.

Les avantages de forêt aléatoire

- Elle résout le problème de sur apprentissage de l'arbre de décision.
- Un modèle robuste face aux données aberrantes et manquantes
- La normalisation des données n'est pas requise
- Traitement des variables quantitatives et qualitatives
- Permet de calculer l'erreur OOB qui veut dire (Out Of Bag)[3]

Les inconvénients de forêt aléatoire

- L'interprétation du modèle est moins évidente.
- Il prend beaucoup de temps puisqu'il faut : construire un nombre énorme d'arbres de décision, passer en revue tous les arbres tout en prédisant une nouvelle valeur.
- Le réglage des paramètres n'est pas simple et ne se fera qu'avec l'expérience. Et pour une raison bien précise, il dispose d'une multiplicité de paramètres (Hounsinou, 2021). Son implantation dans la librairie *scikit-learn* (Scikit-Learn, s.d.) de Python fait appel à pas moins de 14 paramètres dont les plus importants sont :

- **N_estimators** : c'est les nombres d'arbres différents à entraîner.
- **Criterion** : c'est le critère statistique utilisé pour couper les feuilles de chaque arbre en cours de construction.
- **Max_depth** : c'est la profondeur maximale de chaque arbre, un critère très important qui dépend du niveau d'interaction entre les variables
- **Min_samples_split** : le nombre d'observations qu'il faut dans une feuille avant séparation, ce critère permet d'éviter le sur-apprentissage.
- **Max_features** : c'est le nombre maximum de variables qu'on tire aléatoirement pour chaque arbre.
- **Nb_jobs** : indique le nombre de cœurs de CPU que nous utiliserons pour la construction des arbres.
- **Verbose** : ce paramètre nous permet de surveiller la construction des arbres.

2.6.3.2 Boosting

Comme le principe du Bagging, le boosting combine les sorties de plusieurs modalités faibles pour obtenir un résultat plus précis. Sauf que ce dernier est un processus itératif où chaque modèle généré est directement influencé par le modèle précédemment produit. Après la

[3] Erreur Out-of-bag Pour prévenir le sur-apprentissage. Cette erreur permet de contrôler le nombre d'arbres B. On dit alors que cet individu est out-of-bag. L'erreur est calculée en évaluant la proportion d'individus mal classés.

construction d'un modèle Gm, les poids des données d'apprentissage sont réajustés de manière à attirer l'attention du modèle Gm+1 sur les observations mal classées par le modèle Gm. Le modèle final G combine les votes des T modèles pondérés par leurs précisions. C'est-à-dire on détermine l'un après l'autre plusieurs modèles relativement faibles, on demande à chaque modèle de corriger les erreurs de classification commises par celui qui le précède.

Adaboost

Freund et Schapure en 1996 on développé la version originale du boosting qui est l'algorithme d'adaboost (Adaptative Boosting)

Dans le cadre d'un problème de classification binaire, Soit x la variable à prévoir et z = {(x1, y1), …, (xn, yn)} l'échantillon d'apprentissage. L'algorithme commence par initialiser les poids de chaque observations à 1/n et puis ce poids est ajusté en fonction de la nouvelle estimation autrement dit à chaque itération l'algorithme corrige au fur et à mesure l'importance de l'observation en fonction de la qualité de son classement. Le résultat du classifieur boosté est une combinaison des classifieurs g pondérée par les qualités d'ajustement de chaque modèle. Cet algorithme repose sur le choix d'un modèle faible g comme classifieur de base (appelé parfois une règle) parmi une famille de modèles G.

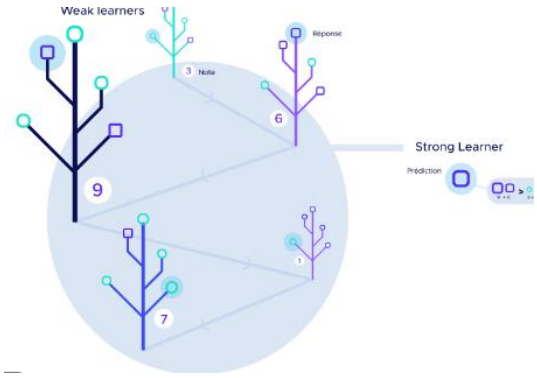


Figure 11: Adaboost (datascientest, 2020)

Source : https://datascientest.com/algorithmes-de-boosting-adaboost-gradient-boosting-xgboost

Extreme Gradient Boosting (XGBoost)

Cet algorithme a été créé par Tianqi Chen et Carlos Guestrin en 2016, cette méthode est une implémentation optimisée de l'algorithme Gradient Boosting. Ce modèle a eu un grand succès, il est conçu pour améliorer l'efficacité, la vitesse de calcul et les performances du modèle. La nouveauté dans cet algorithme est l'introduction d'un « terme de régularisation » L qui permet de rajouter différentes régularisations dans la fonction de perte, limitant un phénomène qui arrive assez souvent lors de l'utilisation d'algorithmes de Gradient Boosting : Le sur-apprentissage. En effet XGBoost propose un panel d'hyper paramètres très important ; il est

ainsi possible grâce à cette diversité de paramètres, d'avoir un contrôle total sur l'implémentation du Gradient Boosting.

Parmi ces paramètres :

- **Nrounds** est le nombre d'itérations à effectuer. Plus la valeur est grande, plus le modèle est lent.
- **Max depth** correspond à la profondeur de l'arbre maximal. Une grande valeur de max depth conduit à un modèle trop complexe et engendre un phénomène de sur-apprentissage. En contrepartie, une valeur faible de max depth augmente le risque du sous-apprentissage. La valeur par défaut est fixée à 6.
- **Colsamplebytree** est le pourcentage des variables choisies aléatoirement parmi l'ensemble de tous les attributs au moment de la construction de l'arbre. C'est l'équivalent de mtry dans l'algorithme de forêt aléatoire.
- **Eta (learning rate)** est un paramètre qui contrôle les poids des arbres conçus dans le modèle. Par d´défaut il est fixé à 0.3.
- **Subsample** détermine le pourcentage des observations à utiliser pour construire l'arbre. Par défaut, ce paramètre est égal à 1.
- **Gamma** détermine la réduction minimale des pertes (la fonction cout) requise pour effectuer une partition supplémentaire sur un nœud terminal de l'arbre. Une grande valeur conduit à un modèle plus conservateur / prudent. Sa valeur par défaut est fixée à 0.

Les Avantages d'XGBoost

- Absence du risque de sur-apprentissage (avec le paramètre de la régularisation).
- Traitement des données manquantes.
- Gestion des interactions entre les variables.
- Capacité de détection des relations non linéaires entre les données.
- XGBoost aura quasi tout le temps de meilleurs résultats que son modèle faible de base.

Les inconvénients d'XGBoost

- Complexité d'interprétation : l'algorithme est qualifié de boite noire.
- Temps de calcul élevé.
- Un nombre important des paramètres à définir.
- L'optimisation de l'algorithme nécessite la prise en compte de plusieurs paramètres d'où la complexité de ce modèle.

2.7 Métriques d'évaluation

2.7.1 Métriques d'évaluations numériques « Matrice de confusion, Accuracy, Recall, Précision, F1 score »

Une méthode scientifique serait d'utiliser la matrice de confusion. La matrice de confusion montre le nombre d'étiquettes réel et prédit et combien d'entre eux sont classées correctement.

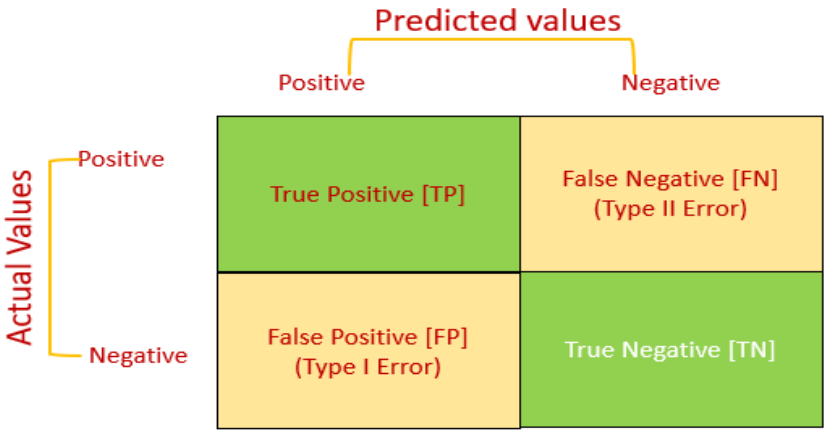


Figure 12: Matrice de confusion

Les colonnes représentent les valeurs réelles, les lignes représentent la prédiction. Chaque case individuelle représente l'une des suivantes :

- **True Positive (TP) :** Le modèle prédit correctement que le résultat est positif.
- **Vrai négatif (TN)** : Le modèle prédit correctement que le résultat est négatif.
- **Faux positif (FP)** : Le modèle a prédit à tort que le résultat était positif, mais le résultat réel est négatif.
- **Faux négatif (FN)** : Le modèle a prédit à tort que le résultat était négatif, mais le résultat réel est positif.

Informations sur la matrice de confusion : Nous allons l'utiliser pour calculer d'autres métriques, chacune de ces métriques répond à une certaine question :

- **Accuracy (Taux de succès)** est définie comme la somme de toutes les prédictions correctes divisée par le nombre total de prédictions.

Accuracy = (TP+TN)/ (TP+FP+FN+TN)

- **Recall (sensibilité)** Parmi ces événements positifs, combien ont été prédits correctement, Cette métrique permet d'éliminer le maximum le taux de faux négatifs.

Recall= TP/ (TP+FN)

- **Precision (Précision)** parmi les prédictions positives, combien ont été réellement prédites correctement, Cette métrique Permet d'éliminer le maximum le taux de faux positifs.

Précision= TP/ (TP+FP)

- **F1 score (F-measure)** qui fait le rapport entre la Précision et le Recall. C'est un bon moyen de résumer l'évaluation de l'algorithme en un seul chiffre.

2.7.2 Métrique d'évaluation graphique

La courbe ROC et l'indice de AUC

Les méthodes graphiques sont réputées pour être les plus riches en informations, mais en même temps les plus compliqué dans la pratique soulignent (Japkowicz & Shah, 2014)

L'utilisation de la courbe ROC (Receiver Operating Characteristic) dans le but de comparer entre la performance des différents classifieurs est devenue assez courante. La surface engendrée par la courbe ROC, appelé en anglais AUC pour Area Under the Curve donne un score de pertinence.

L'avantage de la courbe ROC réside dans l'exhaustivité de l'information qu'elle divulgue ; en effet, elle représente la probabilité de détecter la classe positive (sensibilité) et la classe négative (1-spécificité) pour tous les seuils de classification possibles. Or, la surface sous la courbe est la moyenne de la sensibilité à travers toutes les valeurs de spécificité plus la surface est grande plus notre modèle est précis.

Selon (Hosmer et Lemeshow, 2000), proposent dans leur livre l'échelle de mesure suivante :

- AUC=0.5→discrimination aléatoire.
- $0.5 \leq AUC < 0.7$ → discrimination n'est pas aléatoire
- $0.7 \leq AUC < 0.8$ → discrimination acceptable.
- $0.8 \leq AUC < 0.9$ → discrimination excellente.
- $AUC \geq 0.9$ →discrimination exceptionnelle.

La courbe d'apprentissage

- Elles sont utilisées pour diagnostiquer les performances du modèle d'apprentissage automatique sur les ensembles de données d'entrainement et de validation.
- Elles montrent les changements dans les performances d'apprentissage au fil du temps en termes d'expérience.
- Elles sont utilisées pour diagnostiquer un modèle sous-ajusté, surajusté ou bien ajusté aussi Elles sont utilisées pour diagnostiquer si les ensembles de données d'entraînement ou de validation ne sont pas relativement représentatifs du domaine du problème.

2.8 Conclusion

Au cours de ce chapitre nous sommes parvenus à comprendre le principe de fonctionnement des algorithmes d'apprentissage Nous avons présenté le fondement théorique des méthodes aussi nous avons eu recours à des métriques d'évaluations pour apprécier la capacité de généralisation de chaque modèle.

Ces modèles vont être utilisés pour attribuer un statut à chaque demande d'indemnisation à travers l'émission des signes d'alertes

CHAPITRE 3 : APPLICATION ET RESULTATS

CHAPITRE 3 : APPLICATION ET RESULTATS

3 APPLICATION ET RESULTATS

3.1 Introduction

Après avoir eu une idée globale et claire de l'apprentissage statistique et une explication détaillée du principe de fonctionnement de chaque algorithme utilisé dans notre analyse.

Dans ce chapitre nous allons traiter trois sections, la première sera consacrée à la plus longue phase du processus de modélisation qui est le prétraitement de la base de données qui représente l'ensemble des déclarations de sinistres douteux jugées frauduleuses ou non par ALFA. Nous avons obtenu cette base de données auprès de la compagnie d'assurance CAAT

Quant à l'application des méthodes de classification suivantes : régression logistique, arbre de décision, forêts aléatoires, Ada boost et XG boosting, elle fera l'objet de la deuxième section.

Enfin, la dernière section sera consacrée à la comparaison entre les résultats obtenus à partir des différents modèles en termes de performances.

Lors de la construction de notre modèle, nous suivrons le processus de modélisation prédictive le plus utilisé par les entreprises et le plus apprécié par les analystes (CRISP-DM)

Rappelons que l'objectif de notre étude est de prédire la fraude à l'assurance automobile, c'est-à-dire de construire un modèle qui a la possibilité de classer les dossiers de sinistres selon deux classes : une classe comprenant les dossiers douteux jugés frauduleux par ALFA ; et une seconde classe comprenant les dossiers douteux jugés non frauduleux par ALFA

3.2 Présentation de la CAAT

La Compagnie Algérienne des Assurances - CAAT est une Entreprise publique économique, Société par actions (EPE/SPA) au capital de 20.000.000.000 DA, détenue entièrement par l'Etat algérien, actionnaire unique.

Pour rappel, la CAAT a été créée en avril 1985 pour pratiquer les assurances liées aux transports du fait de la spécialisation des compagnies et de l'exercice du monopole de l'Etat sur l'activité d'assurance en Algérie.

Après la levée de la spécialisation et l'entrée en vigueur de la loi instaurant la séparation des assurances de personnes des assurances de dommages, la CAAT est devenue, depuis le 1er juillet 2011, une compagnie d'assurance pratiquant, uniquement, les branches d'assurances "dommages".

La structure du portefeuille

La structure globale du portefeuille de la CAAT qui demeure diversifiée se présente, en 2021, comme suit sur la base du chiffre d'affaires :

- Les assurances Incendie et Risques Divers totalisent 63%;

- L'assurance Automobile, occupe la deuxième position avec 25%;
- Les assurances Transports représentent 8%;
- Les assurances Engineering totalisent 4%.

Position sur le marché

La CAAT est un des acteurs majeurs sur le marché algérien des assurances et un de ses principaux leaders. Au fur et à mesure de son évolution, la CAAT est parvenue par diversifier son portefeuille et à conforter sa dimension d'assureur des risques d'entreprises.

Très vite, la CAAT a pu s'implanter sur pratiquement tout le territoire national pour présenter des offres de couverture diversifiées et adaptées aux besoins de la clientèle avec un intérêt particulier pour la prévention des risques et la qualité de la prestation.

La part de marché de la CAAT, en 2020, est de l'ordre de 18 % avec un chiffre d'affaires de 24,750 milliards de dinars. Ce résultat lui permet de se maintenir à la deuxième position au sein du marché algérien des assurances dommages. Les premières estimations, au titre de l'exercice 2021, semblent confirmer le maintien de ce taux à 18%.

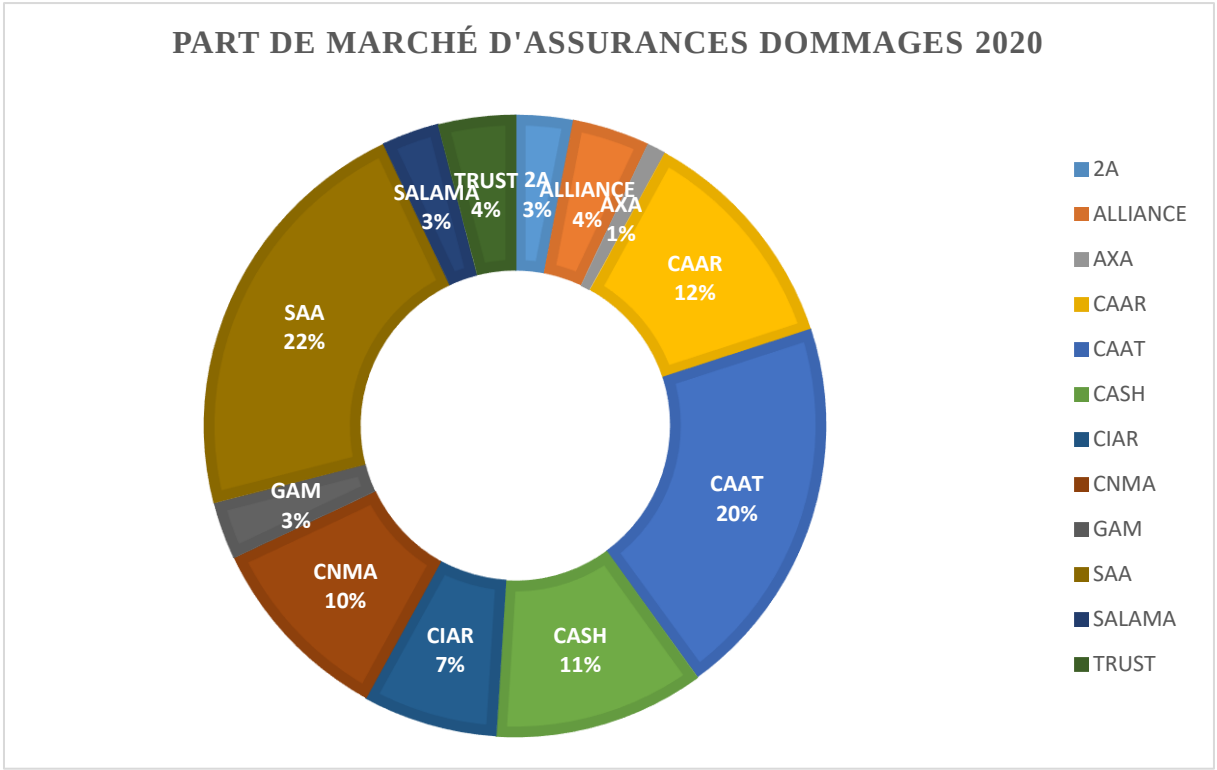


Figure 13: Part de marché exercice 2020 (Assurances dommages)

Pour 2020, la CAAT demeure leader en assurances des risques d'entreprise avec une part de marché de 28%. En assurance transports, elle détient une part de 32% du marché en assurance automobile, sa part de marché est de l'ordre de 10%.

CHAPITRE 3 : APPLICATION ET RESULTATS

3.3 Le prétraitement des données

La qualité des données affecte énormément la capacité d'apprentissage de notre modèle, c'est pourquoi l'étape de prétraitement est très importante dans la procédure de modélisation, et c'est l'étape la plus longue.

Lors de la réalisation de cette étape nous passerons par des sous-étapes, en commençant par la visualisation, afin de comprendre au mieux le phénomène étudié et les différentes variables puis le nettoyage des données qui comprend l'identification des données manquantes, des données bruyantes et des données incohérentes. Le recodage et la création de variables ont également été réalisés.

3.3.1 Présentation et analyse de la base de données

Notre étude de cas a été menée sur les données de l'assurance CAAT, où nous avons voulu élaborer un modèle de prédiction de fraude.

C'est pourquoi nous avons demandé toutes les données concernant les dossiers de sinistres réglés (c'est-à-dire que nous supposons qu'il s'agit de dossiers non frauduleux tant qu'ils sont indemnisés) qui se trouvent dans le département informatique et nous les avons fusionnés avec tous les dossiers détectés comme frauduleux qui se trouvent dans le département automobile. Mais malheureusement, nous n'avons pas eu la chance d'obtenir la base au bon moment.

Pour cette raison, nous avons adapté notre étude avec la base que nous avons pu obtenir par le département automobile qui est l'ensemble des dossiers douteux transmis à l'organisme d'enquête ALFA qui décide si le dossier est frauduleux ou non suite d'une enquête.

Cette base de données concerne la période de 2010 à 2018, pour le reste des années à partir de 2018 la mise à jour de la base de données n'a pas encore été faite.

Les données imputées dans cette base représentent des polices d'assurance qui ont fait l'objet de sinistres déclarés par les assurés avec la désignation éventuelle d'un ingénieur expert pour évaluer les dégâts subis et déterminer si les relations entre les circonstances déclarées et les dommages sont conformes ou non. Par conséquent, à l'issu de cette observation, croisée avec l'avis d'un gestionnaire sinistres.

Si, après l'expertise, le gestionnaire doute que le dossier soit frauduleux, mais qu'il ne dispose pas de toutes les preuves, il l'envoie à la cellule ALFA pour une enquête plus approfondie ; à ce moment-là, la cellule ALFA établit un rapport détaillé. Sur cette base, le gestionnaire qualifie chaque dossier de potentiellement frauduleux ou non.

Notons que dans la pratique des assurances automobiles, les rapports d'enquête d'ALFA sont considérés comme un moyen parmi d'autres moyens d'appréciation d'un dossier sinistre

quelle que soit sa nature. Autrement dit, un rapport d'enquête n'est qu'un simple document versé au dossier, dont le gestionnaire sinistre doit l'exploiter comme tout autre document. Dans ce contexte, il faut signaler que les rapports d'enquête d'ALFA n'ont pas une force juridique comme un rapport d'enquête des autorités, tel que le PV de gendarmerie, PV de police, et le PV de la protection civile. À ce titre, le juge ne les prend pas en considération comme un moyen de preuve de matérialité, mais il le prend en compte comme un moyen d'appréciation ou d'indication, et dans la plus part du temps il est rejeté catégoriquement.

Donc, au cours de cette étude, nous cherchons à construire un modèle qui a la possibilité d'effectuer une classification des dossiers de déclarations de sinistres selon deux classes :

- ✓ Une classe qui comprend les dossiers douteux jugés frauduleux par ALFA.
- ✓ Une deuxième classe qui comprend les dossiers douteux jugés non frauduleux par ALFA.

La base de données utilisée contient des dossiers jugés frauduleux par ALFA et confirmés par le gestionnaire comme étant des dossiers frauduleux.

Afin de comprendre nos données, nous allons effectuer une analyse descriptive de notre base de données.

Notre base de données brute contient des dossiers frauduleux, des dossiers réglés et des dossiers en cours de traitement. Dans cette étude, nous ne sommes intéressés que par les deux premières classes, nous avons éliminé les dossiers en cours de traitement.

Le nombre de variables explicatives brutes comprend les caractéristiques du véhicule assuré ainsi que les caractéristiques du sinistre. Ces caractéristiques sont traduites en attributs de différents types :

- 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 : Variable nominale avec seulement deux modalités.

- Les variables quantitatives

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

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. Donc nous commencerons cette étape par l'analyse de forme :

CHAPITRE 3 : APPLICATION ET RESULTATS

La distribution de la variable cible

Notre base de données est constituée de 239 observations, dont 131 cas de fraude qui représentent 55% de notre base, et 107 cas de non frauduleux c'est-à-dire une réclamation légale qui représentent 45%. Comme le montre la figure suivante :

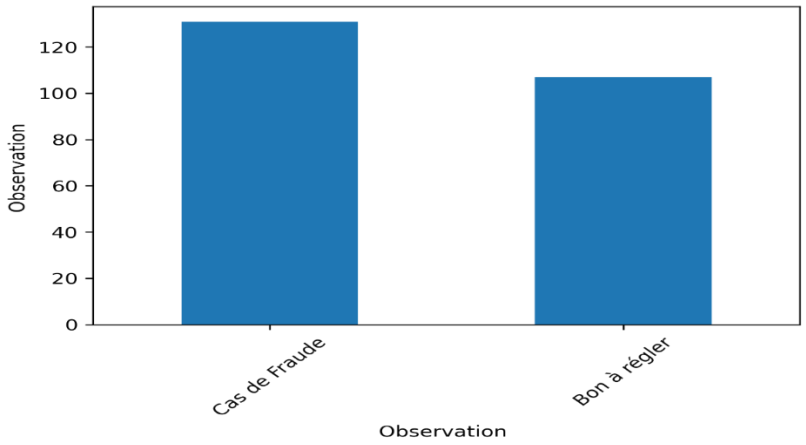


Figure 14: Répartition de la variable cible

Nous pouvons remarquer que plus de la moitié des dossiers douteux envoyés à ALFA (Sans compter les dossiers en cours de traitement), ces dossiers après l'enquête ont été jugés frauduleux. Avec une proportion qui n'est pas négligeable. Si la société n'a pas détecté l'acte frauduleux pour ces dossiers, elle paiera un montant supplémentaire d'indemnité chaque année comme le montre le graphique

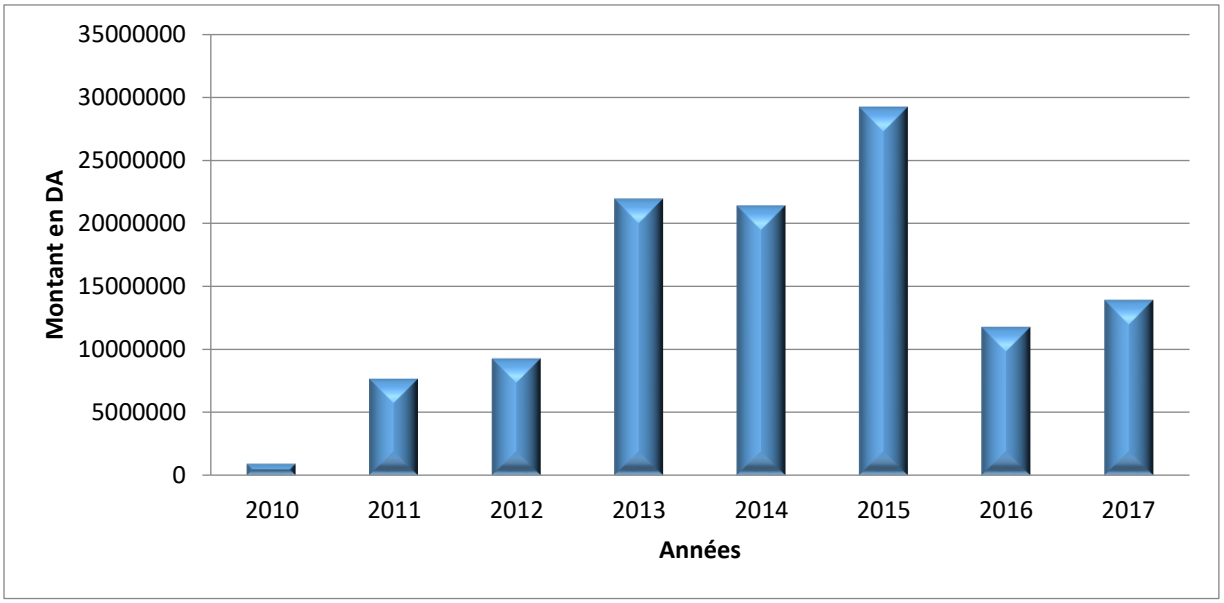


Figure 15: Totale de montant de dommages déclaré de chaque année des dossiers frauduleux

Ces chiffres ne sont pas négligeables, ils peuvent couvrir d'autres charges légales de l'entreprise, exemple : payer les fonctionnaires plusieurs mois...c'est pourquoi on peut dire que l'intérêt de la lutte contre la fraude touche plusieurs axes au niveau de l'entreprise.

La répartition de la fraude par années de survenance du sinistre

Passons maintenant à la répartition des dossiers en fonction de l'année de survenance du sinistre. Elle est présentée dans le graphique suivant :

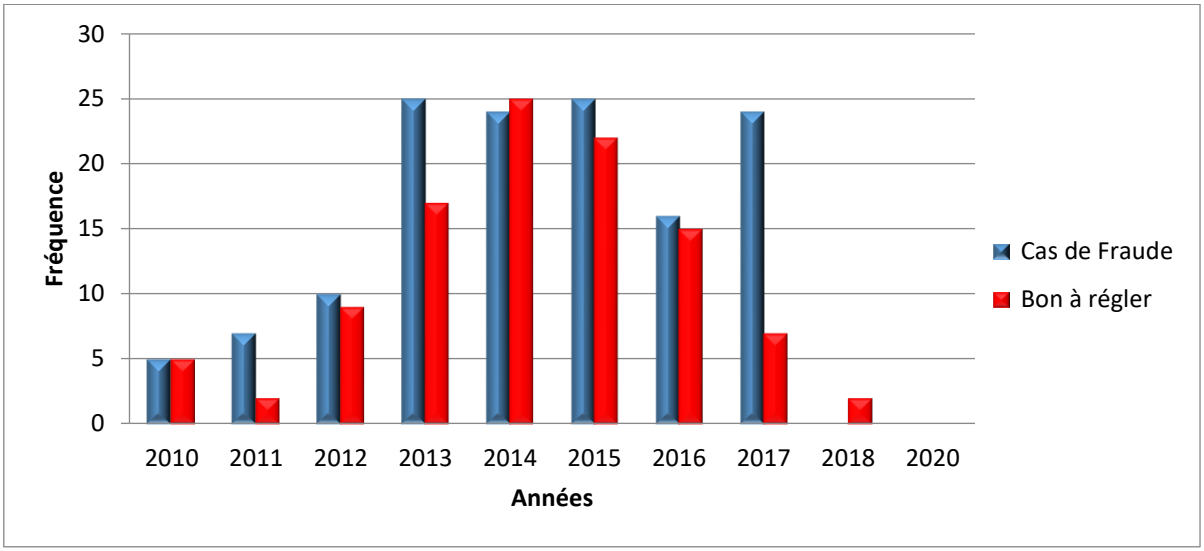


Figure 16: Fraude par années de survenance du sinistre

D'après le graphique, nous remarquons qu'à partir de l'année 2013, le nombre de dossiers envoyés à alfa a augmenté de manière très remarquable.

La répartition des dossiers frauduleux selon le type de garantie souscrite

Nous traitons la répartition des dossiers frauduleux selon le type de garantie souscrite

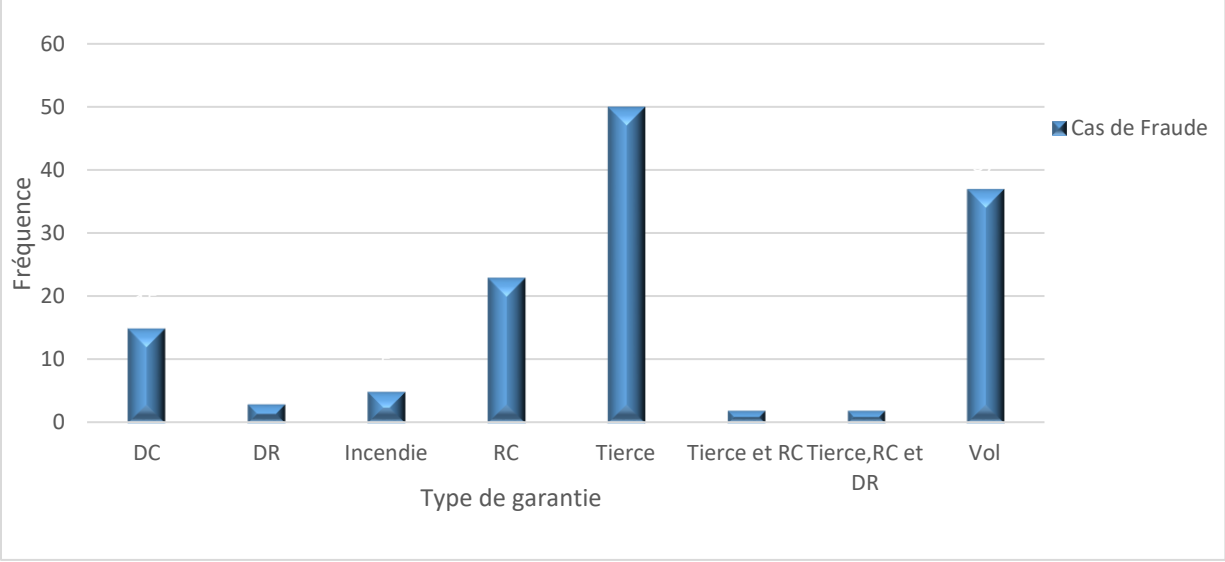


Figure 17: Fraude par type de garantie

Nous observons que la tranche des fraudeurs qui sont souscrits a une garantie tiers (tous risques) sont les plus fréquents car dans cette garantie : les sinistres fictifs sont faciles à mettre en scène, l'assuré sera indemnisé indépendamment de sa responsabilité (fautif ou non fautif)

Suivie par la garantie vol puisque l'opportunité trouvée par les fraudeurs dans cette garantie réside dans la difficulté pour les experts de prouver qu'un vol volontaire, a été commis par l'assuré surtout dans le cas de vol total notamment lorsque le véhicule n'est pas retrouvé.

Ces deux garanties sont faciles car elles ne nécessitent pas d'adversaires, généralement le tiers n'est pas identifié dans le cas des fraudes, le fraudeur pourra soit agir seul soit faire recours à un tiers (un garagiste) pour produire de fausses factures ou des factures gonflées. Par contre la RC qui prend la troisième position, les souscripteurs de cette garantie font un recours a la fraude quand ils sont fautifs, ils essayent de trouver une autre personne qui possède une couverture tout risque pour qu'elle endosse la responsabilité à leur place et à la fin ils seront tous les deux gagnants.

Aussi, dans le cas de DC (dommages collision) pour qu'un assuré soit couvert il faut qu'il y ait une collision, dans le cas des accidents sans adversaire l'assuré fait un recours à la fraude…

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

Nous voulons aussi savoir les marques de voiture les plus fréquentes dans la fraude.

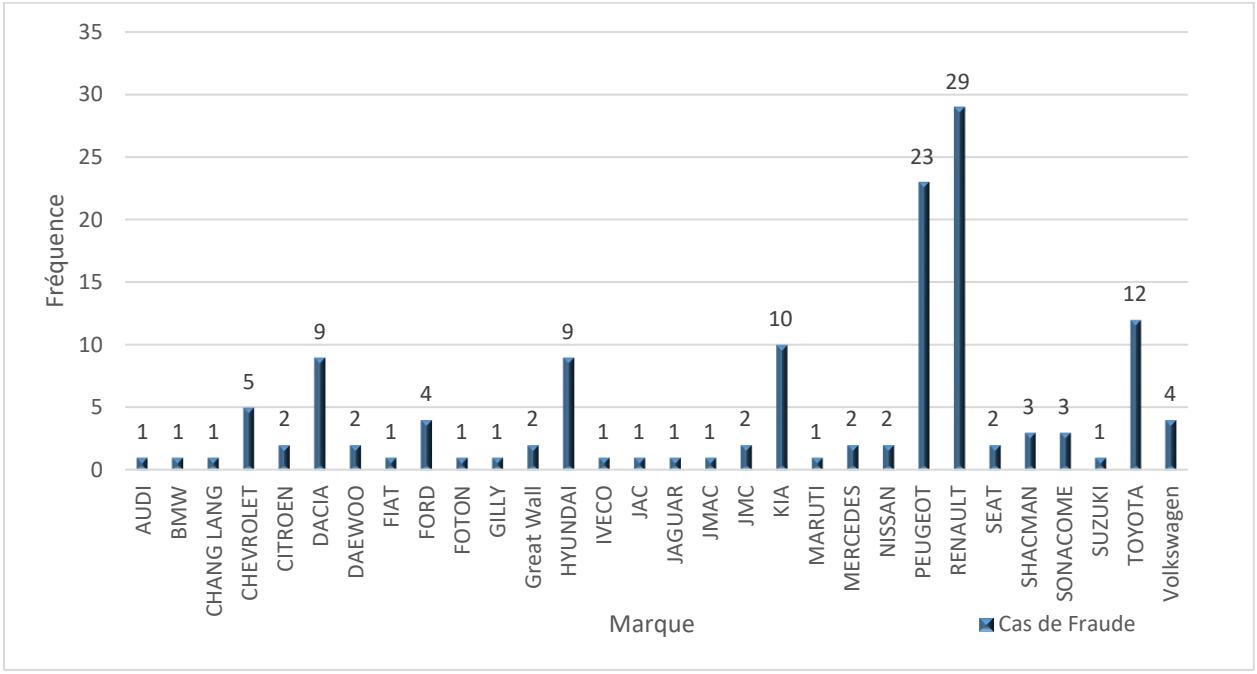


Figure 18: Répartition du nombre de fraude par marque

Une analyse de la distribution des fraudes par marque a permis d'établir une liste de marques présentant un nombre élevé de cas de fraude. Certes, cette distribution dépend du nombre de réclamations qui ont eu lieu pour chaque marque.

Le graphique ci-dessous montre que les deux marques RENAULT, PEUGEOT sont les plus présentes dans la base de données en termes de nombre de fraudes. Suivies par TOYOTA, KIA, HYUNDAI, DACIA. Et le 3ème groupe les marques : Chevrolet, Ford, Volkswagen.

En comparant cette liste avec d'autres études réalisées sur la fraude à l'assurance automobile dans d'autres pays, on constate que : l'étude réalisée en France montre que les trois marques RENAULT, PEUGEOT et Citroën sont les plus présentes dans la base de données en termes de nombre de fraudes. Aussi, l'étude au niveau de la Tunisie, montre que les marques ayant les fréquences les plus élevées sont : RENAULT, PEUGEOT, KIA, VOLSWAGEN, SEAT.

Nous constatons que ces deux marques sont les plus fréquemment utilisées dans les fraudes

La base de départ

- Notre base de données contient 239 observations, avec un certain nombre de variables indiquant: montant des dommages, date d'effet, date de déclaration, date de survenance du sinistre, compagnie/agence tierce, branche, couverture souscrite, marque du véhicule, type de véhicule, prime, immatriculation du véhicule, nom de l'expert, honoraires d'expert et frais enquîtes, date envoi de fichier à alfa et la variable observation (qui est la variable cible).

Rappelons que nous avons utilisé dans le traitement et la modélisation de nos données le langage python (justification de notre choix précédemment mentionné dans le chapitre deux).

3.3.2 Traitement des données manquantes /aberrantes (Outliers)

Le nettoyage de la base

Dans la base que nous avons, il y a quelques variables qui ne présentent aucun intérêt pour la modélisation de la fraude. Une phase de nettoyage a dû être réalisée pour restreindre le nombre de variables explicatives :

- Il s'agit de variables qui n'apportent aucune information intéressante pour expliquer l'événement "fraude". C'est le cas du nom de l'expert, des frais d'expertise et d'enquête, honoraires d'expert et frais enquîtes, date envoi de fichier à ALFA.
- Il s'agit des variables qui fournissent la même information mais dans un format différent Nous avons utilisé la matrice de corrélation pour supprimer l'une des deux variables qui ont une forte corrélation : comme dans le cas de la variable type de voiture, et marque de voiture, nous avons supprimé le type et nous avons gardé la marque.
- D'autres sont supprimées par la suite car de nouvelles variables seront créées pour les tester et en tirer des résultats.

CHAPITRE 3 : APPLICATION ET RESULTATS

Identification des valeurs manquantes

Nous recourons à cette figure présentée ci-dessous pour visualiser les valeurs manquantes qui sont des tiers blancs.

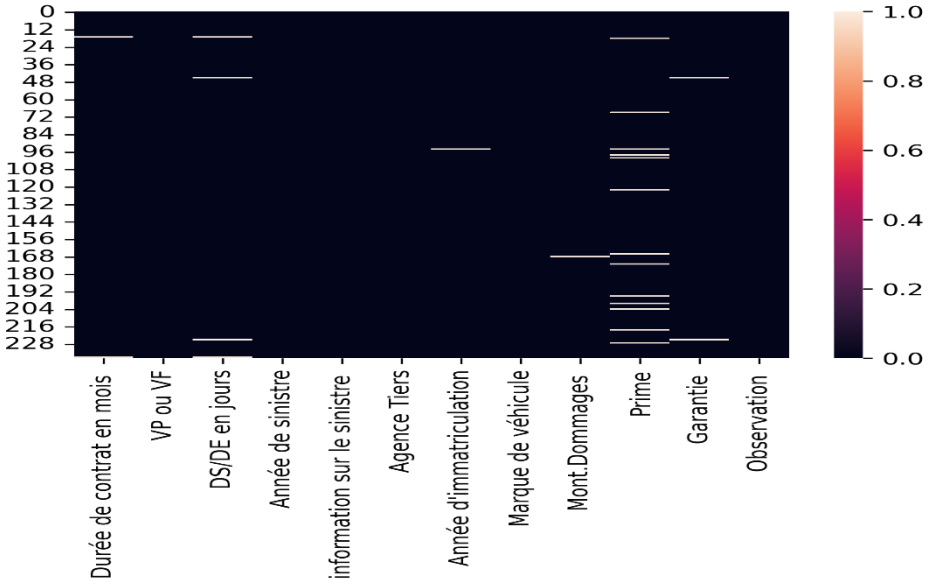


Figure 19: Valeurs manquantes

- Nous remarquons que notre base de données ne contient pas beaucoup de valeurs manquantes mais lorsque nous visualisons graphiquement les différentes modalités de chaque variable, nous trouvons des champs remplis de X ou bien par Néant ou vide.
- Par conséquent, afin d'obtenir une base de données complète et fiable, nous avons étudié chaque variable au cas par cas. En effet, pour certaines variables, une valeur manquante a une signification particulière et doit donc être considérée comme une modalité à part entière, comme dans le cas de la variable agent tiers, une valeur manquante peut être interprétée comme l'absence d'un tiers.
- Pour d'autres cas, nous avons remplacé la valeur vide par la moyenne ou le mode, ou nous l'avons lissée comme une modalité de valeur manquante, car il y a des variables avec lesquelles nous ne pouvons pas utiliser ces techniques, l'information remplie sera incohérente avec le reste des renseignements des dossiers.

Traitement des données aberrantes « Outliers »

Par définition, un outliers est une observation qui dévie nettement du comportement général par rapport au critère sur lequel l'analyse est réalisée. En d'autres termes, une valeur aberrante est une valeur extrême, anormalement différente de la distribution d'une variable (la valeur de cette observation diffère grandement des autres valeurs de la même variable) (Benzaki, 2017).

Il faut noter qu'il est important de détecter les valeurs aberrantes et de les extraire avant

l'exécution de l'algorithme d'apprentissage car leur présence dans l'échantillon d'entrainement aura des conséquences sur la performance du modèle prédictif.

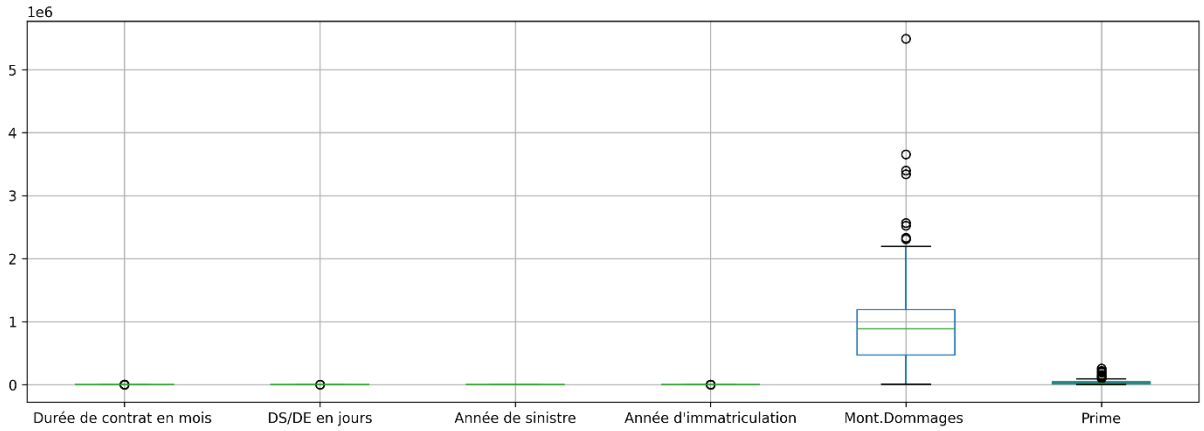


Figure 20: Valeurs aberrantes

Ces valeurs aberrantes sont détectées dans chacune des variables par la méthode de la boîte à moustaches de TUKEY dont le principe consiste à détecter toutes les valeurs en dehors des bornes inférieures, celles qui sont plus petites que

Q1-1.5x (Q3-Q1)) et supérieure (celle qui sont plus grande que Q3+1.5 x (Q3-Q1)).

3.3.3 Création et recodage des variables

L'étape de création des données

Nous avons créé de nouvelles variables à partir des variables existantes pour avoir plus de variables explicatives porteuses d'informations :

- Nous utilisons le numéro de police pour créer une variable qui mentionne le type de garantie particulier ou flotte. Et nous éliminons la variable Numéro de police.
- Nous utilisons l'immatriculation de véhicule pour créer une variable qui contient l'année de fabrication de voiture, nous avons éliminé aussi la variable immatriculation.
- Variables date : Les variables de type date ne sont pas adaptées à une utilisation directe en modélisation. Dans ces conditions, certaines d'entre elles ont été manipulées sous forme d'une variable quantitative, par exemple :
  - ✓ DS/d'en jours : qui représente la différence entre la date de déclaration du sinistre et la date d'effet.
  - ✓ Durée du contrat en mois : qui représente la différence entre la date d'effet et la date d'échéance.

    Après, nous avons éliminé les différentes dates : la date d'Effet et date d'échéance.
  - ✓ Nous supprimons le jour et le mois de la date de réclamation, nous laissons uniquement l'année de sinistre pour réduire le nombre de modalité.

- Nous avons remarqué que certaines variables qualitatives ont pris plusieurs modalités comme la marque du véhicule. Nous savons que certains algorithmes comme l'arbre de décision ont des difficultés à exploiter l'information contenue dans ces variables lorsque le nombre de modalités est élevé. Ils peuvent donc être plus discriminants si nous réduisons le nombre des modalités. Afin de s'en tirer avec ce nombre élevé de modalités, nous avons combiné les modalités à faibles fréquences en une seule modalité appelée « Autres ».

  Nous avons laissé les modalités suivantes les plus fréquentes : RENAULT, PEUGEOT, TOYOTA, DACIA, KIA, HYUNDAI, VOLKSWAGEN, CHEVROLET, SEAT, FORD.

L'encodage des variables

Certaines variables qualitatives portent des modalités textuelles. Il est nécessaire pour quelques modèles de les convertir en format numérique. Il est nécessaire de spécifier alors après l'encodage que ces variables sont des valeurs catégoriques.

Durant cette partie nous avons utilisé la fonction : Label encoding, MultiColumn Label Encoder.

Notons que la variable cible est une variable catégorielle à deux modalités, pour cette variable nous retiendrons la recodification suivante :

« Classe 1 » pour dossiers frauduleux.

« Classe 0 » pour client règlementaire (non frauduleux).

Base de données finale

Les variables regroupées dans le tableau suivant représentent notre base de données finale :

Tableau 1: Base des données finale

Variable	Type de variable
Observation	Variable catégorielle binaire
Mont Dommages	Variable quantitative
Marque de véhicule	Variable catégorielle (11 modalités)
Garantie souscrite	Variable catégorielle (8 modalités)
Durée de contrat en mois	Variable quantitative (8 modalités)
Prime	Variable quantitative
Compagnie de tiers/où agence tierce	Variable catégorielle (10 modalités)
Année d'immatriculation de voiture	Variable quantitative (27 modalités)
Année de sinistre	Variable quantitative (9 modalités)

DS/DE en jours (Période entre date d'effet et date de sinistre)	Variable quantitative
Information sur le sinistre	Variable catégorielle (4 modalités)
VP ou VF (véhicule particulier ou flotte)	Variable catégorielle binaire

3.3.4 Répartition des données et traitement des classes déséquilibrées

Construire un estimateur requiert de répartir notre base de données en deux sous-échantillons ; le premier servira d'exemple ou d'entrainement à partir duquel notre modèle tirera les règles d'attribution, et le second servira à mesurer sa performance.

Il existe plusieurs façons de découper le jeu de donnés, dans notre cas nous avons opté pour le protocole 80-20. 80% de la BDD servira à entrainer le modèle et le reste à l'évaluer.

Concernant le rééquilibrage des données, dans la réalité il est évident que les fraudes avérées sur les contrats auto sont peu fréquentes.

Dans le cas où nous avons pu avoir la totalité des sinistres réglés, cette étape sera nécessaire pour avoir un bon modèle parce que ce déséquilibre entre les observations positives et négatives représente un risque de sur-apprentissage pour la modélisation. Mais avec notre échantillon cette méthode est inutile par ce que notre échantillon est équilibré.

3.3.5 Normalisation ou standardiser les données

Le Feature Scaling qui comprend la Standardisation et la Normalisation.

Il s'agit d'une bonne pratique, pour ne pas dire d'une obligation, lors de la modélisation avec le machine learning comme le cas de la régression logistique les data set proviennent avec des ordres de grandeurs différents. Cette différence d'échelle peut conduire à des performances moindres. Pour pallier à cela, des traitements préparatoires sur les données existent.

Pour effectuer cette transformation, nous soustrayons aux données leur moyenne empirique m et les divisons par leur écart-type σ.

Notons que la mise à l'échelle n'est pas nécessaire pour les arbres de décisions aussi pour les forêts aléatoires, tous les modèles qui sont basés sur les arbres.

3.3.6 Sélection des variables

La présence de caractéristiques non pertinentes dans vos données peut réduire la précision de nombreux modèles, en particulier les algorithmes linéaires tels que la régression linéaire et logistique.

Pour la phase de sélection des variables, nous avons commencé l'étude par un simple test de corrélation entre chaque variable explicative qualitative et la variable cible par le test du chi 2,

puis nous avons appliqué le test ANOVA pour tester la relation entre les variables explicatives quantitatives et la variable cible. Mais nous n'éliminons aucune variable sur la base de ces tests, nous voulons utiliser la méthode Kbest select.

La sélection des variables par select Kbest (chi2 et ANOVA)

Au lieu de traiter chaque deux variables indépendamment de l'ensemble des données.

Nous avons choisi l'analyse de corrélation de l'ensemble des données pour extraire les meilleures variables et éliminer la partie la moins importante des données et pour réduire le temps de formation par la méthode SelectKBest qui sélectionne les variables en fonction du score k le plus élevé.

Elle a effectué les deux tests pour sélectionner les caractéristiques ayant les deux meilleures valeurs F ANOVA, et Chi2.

Nous avons sélectionné les variables explicatives présentant les meilleurs scores de chi2 et d'ANOVA et nous avons éliminé les modalités dont le score était inférieur à 1.

Nous avons obtenu la liste suivante :

Tableau 2: Listes des variables qualitatives sélectionnés

Variable de base	Variable dommies
information sur le sinistre	Information sur le sinistre _Vol Total
	Information sur le sinistre _X
Agence Tiers	Agence Tiers_ CAAR
	Agence Tiers_ CIAR
	Agence Tiers_ SAA
	Agence Tiers_ SALAMA
	Agence Tiers_ TRUST
Marque de véhicule	Marque de véhicule_ ISUZU
	Marque de véhicule_ Volkswagen
Garantie	Garantie _Tierce et RC
	Garantie _Vol
	Garantie _Tierce et RC
	Garantie _Tierce, RC et DR
	Garantie _Tierce, RC et DR

Tableau 3: Listes des variables quantitatives sélectionnés

Variable de base
Durée de contrat en mois
DS/DE en jours
Montant Dommages
Prime

Nous constatons que les variables qui ont un score supérieur à 1 sont : la prime, le montant des dommages, le DS/DE en jours, la durée du contrat en mois, et quelques modalités des variables suivantes : la garantie souscrite, la marque de la voiture, l'agence tierce, ainsi que les informations sur le sinistre. Les variables qui sont complètement éliminées sont l'année d'immatriculation, l'année de sinistre, VP ou VF.

Comme il existe plusieurs méthodes pour réduire le nombre de variables dans le modèle. Dans cette étape, nous avons essayé d'autres méthodes de sélection de variables pour déterminer la meilleure fonction à appliquer qui nous donne les meilleurs résultats. Au final, nous avons décidé d'utiliser les résultats de la méthode selectkbest pour la sélection des variables.

3.4 Construction des modèles

La deuxième grande phase après le prétraitement est le choix des algorithmes utilisés pour la prédiction ; dans notre étude nous avons opté pour la régression logistique, l'arbre aléatoire, la forêt d'arbres aléatoires, puis Adaboost et enfin XGboost.

3.4.1 Résultats de la modélisation avant et après d'optimisation des hyper paramètres

Nous avons d'abord exécuté le modèle avec les paramètres par défaut. Puis nous avons utilisé la méthode GridSearch CV pour automatiser le réglage des hyper paramètres. Afin de déterminer les valeurs optimales pour un modèle donné. Car la performance d'un modèle dépend significativement de la valeur des hyper paramètres.

Nous avons fait une étude comparative entre les deux modèles avant et après l'optimisation et nous avons choisi le plus performant entre les deux,

puis nous confirmons notre choix avec la visualisation des courbes d'apprentissage.

Nous avons utilisé le modèle le plus performant dans la dernière étape de la comparaison entre tous les types de modèles de notre étude de cas.

3.4.1.1 La régression logistique

Tout d'abord, nous avons effectué la normalisation de la base de données avant d'appliquer la régression logistique.

Les hyper paramètres après l'optimisation de la régression logistique avec le CV Gridsearch sont les suivants :

Tableau 4: Paramètres de la régression logistique.

Les paramètres	Les valeurs
C	0.001
fit_ intercept	True
penalty	l2
solver	newton-cg

Les variables qui ont un effet sur la variable cible avant et après l'optimisation des paramètres avec Grid search

Notons que l'Odds est le rapport des chances d'avoir une fraude pour ceux qui ont une caractéristique X d'une part et ceux qui n'en ont pas d'autre part.

Ce rapport est l'expo de (coefficient)

$$Odds_P = \frac{P}{1-P}$$

CHAPITRE 3 : APPLICATION ET RESULTATS

Selon le tableau suivant, nous avons chaque variable et son Odds Ratio avant et après l'optimisation :

Tableau 5: Odds Ratio avant et après optimisation de la régression logistique

Variable	Odds Ratio avant optimisation	Odds Ratio Après optimisation
Information sur le sinistre Vol Total	1	0.99
Information sur le sinistre X	1	1.0008
Agence Tiers CAAR	1	1.0009
Agence Tiers CIAR	1	0.99
Agence Tiers SAA	1	1.0014
Agence Tiers SALAMA	1	0.99
Agence Tiers TRUST	1	0.99
Marque de véhicule ISUZU	1	0.99
Marque de véhicule Volkswagen	1	0.99
Garantie Tierce et RC	1	1.0007
Garantie Tierce RC et DR	1	1.0006
Garantie Vol	1	0.99
Mont Dommages	0.99	0.99
Prime	1.000024	1.00002
Durée de contrat en mois	1	0.99
DS/DE en jours	1.00000003	0.99

L'exponentielle des coefficients des paramètres du modèle, comme nous pouvons le voir dans le tableau, indique que :

Avant l'optimisation

- Si l'odds ratio est inférieur à 1 donc il y a un effet bénéfique,
- Le montant des dommages a un effet négatif puisque l'Odds ratio < 1 (légèrement inférieur à 1) cela dit un accroissement du montant de dommage d'une unité réduit la probabilité que le dossier soit frauduleux de 0.99.
- Si l'odds ratio est supérieur à 1 pour un effet délétère.
- Pour la prime son coefficient Odds ratio >1 (légèrement supérieur à 1) cela signifie que leur contribution dans la prédiction de la fraude est positive. Si elle augmente d'une unité la probabilité que le dossier soit frauduleux augmente de 1, 0000245

- De même pour la variable DS/DE en jours qui est la durée entre la date d'effet et la date de sinistre, c'est-à-dire l'accroissement de la durée entre la date d'effet et la date de sinistre d'un jour augmente la probabilité que le dossier soit frauduleux par 1.00000003.
- Pour le reste des variables explicatives dans le tableau, l'Odds Ratio est égal à 1, c'est-à-dire l'absence d'effet entre ces variables et la fraude.

Nous pouvons écrire l'équation comme suit :

Y*= 4.10615891e-10 - 8.19961574e-07 M + 2.45317749e-05 P + 2.54214097e-08 D

Avec M : Montant de dommage , P : Prime, D : la variable DS/DE en jours

Par le même principe, nous pouvons classer les variables du côté après optimisation, nous remarquons que, également après optimisation, les valeurs d'Odds restent proches de zéro. Cela signifie que l'effet est très faible (soit du côté positif ou négatif).

Nous savons que plus l'odds est éloigné de 1, plus l'effet est important.

D'après nos entretiens avec les gestionnaires du département automobile, nous pouvons dire que parmi ces variables, certaines ont un effet sur la fraude, qui n'apparaît pas dans nos résultats car notre échantillon est très petit.

Comparaison entre les métriques d'évaluation avant et après l'optimisation

Tableau 6: Comparaison entre les métriques d'évaluation avant et après l'optimisation de RL.

	Precision	Recall	F-measure (F1 score)	Accuracy
Avant optimisation :				**0.63**
0	0.46	0.79	0.58	
1	**0.84**	0.55	**0.67**	
Après optimisation :				0.57
0	0.65	0.67	0.47	
1	0.72	0.63	0.64	

Lorsque nous avons comparé les résultats avant et après l'optimisation, nous décidons d'utiliser le modèle avec les paramètres par défaut, nous basons sur le critère du score f1 qui combine subtilement la précision et le rappel. Il est plus intéressant que la précision car le nombre de vrais négatifs (TN) n'est pas pris en compte.

Les courbes d'apprentissage :

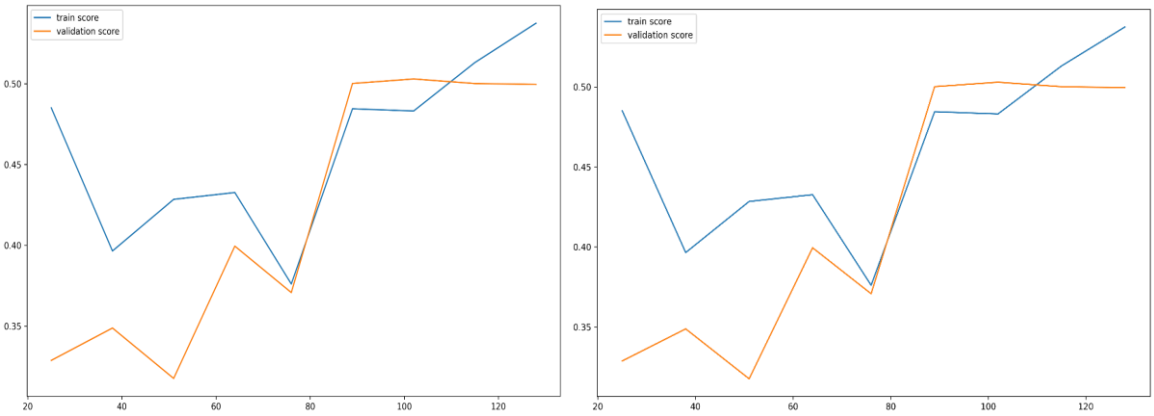


Figure 21: Courbes d'apprentissage avant et après optimisation

D'après les deux courbes, nous remarquons que l'optimisation des paramètres ne donne aucune amélioration au modèle de régression logistique.

3.4.1.2 L'arbre de décision

Nous avons exécuté le modèle avec les paramètres par défaut, nous obtenons une représentation graphique de l'arbre de décision (voir annexes 1).

La sortie de ce modèle, nous permet de visualiser l'architecture de l'arbre mais nous remarquons que cet arbre de décision est grande taille avec des nœuds qui contiennent peu d'éléments ce qui pose le problème de la complexité de l'arbre dû à un sur-ajustement.

Optimisation de l'arbre de décision avec la fonction Gridsearch CV

Après l'imputation de la fonction Gridsearch CV, nous avons obtenu les hyper paramètres suivants :

Tableau 7: Paramètres de l'arbre de décision

Les paramètres	Les valeurs
Max_ depth	9
Min_ samples_ leaf	20
Min_samples_split	20
Max_leaf_nodes	10
Max_leaf_nodes	gini

Après cette optimisation, nous avons obtenu une représentation graphique plus claire et plus lisible de l'arbre de décision (voir annexes 2), avec un nombre réduit de variables.

La comparaison entre les variables importantes avant et après l'optimisation des paramètres avec Grid search

Nous avons utilisé la fonction "feature_importances" pour avoir les variables les plus importantes pour cet arbre de décision. Plus le score attribué à chaque variable. C'est à dire identifier les variables pertinentes dans la modélisation de la fraude.

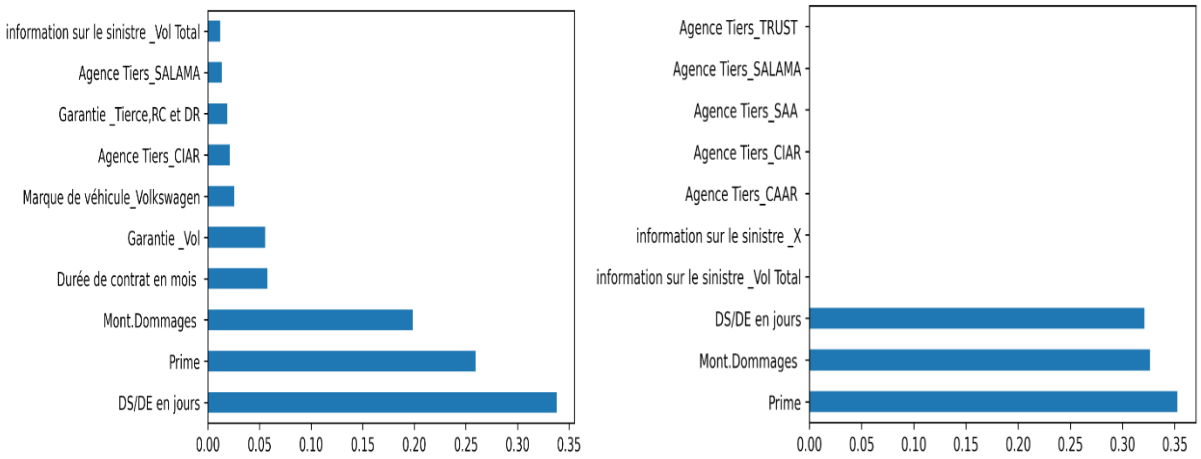


Figure 22: Variables importantes avant et après l'optimisation des paramètres avec Grid search.

Nous remarquons qu'avant l'optimisation des paramètres (comme le montre l'histogramme de gauche), les 3 variables les plus importantes pour l'arbre de décision sont la durée entre la date du sinistre et la date d'effet, la prime et le montant du dommage avec les scores suivants : 34%, 26%, 20%.

Elle est suivie par deux autres variables qui sont la durée du contrat et la garantie contre le vol avec un taux d'importance pour chacune d'elles de près de 6%. Il y a également cinq autres variables mais leur niveau d'importance est très faible, inférieur à 3%. Ce sont : la marque du véhicule Volkswagen, l'agence tierce CIAR, la garantie tierce RC et DR, l'agence tierce SALAMA, et l'information disponible sur le sinistre est un vol total.

Concernant la phase après optimisation nous remarquons que le nombre de variables importantes est réduit, il reste que 3 qui sont les mêmes 3 variables les plus importantes pour l'arbre avant l'optimisation sauf que l'ordre change et le degré d'importance augmente, ils atteignent des taux supérieurs à 30%.

Comparaison entre les métriques d'évaluation avant et après l'optimisation

Tableau 8: Comparaison entre les métriques d'évaluation avant et après l'optimisation de l'arbre de décision

	Precision	Recall	F1 score	Accuracy
Avant optimisation :				0.63
0	0.44	0.57	0.50	
1	**0.76**	0.66	0.70	
Après optimisation :				**0.65**
0	0.46	0.43	0.44	
1	0.73	**0.76**	**0.75**	

Dans le cas de l'arbre de décision, nous avons choisi le modèle après l'optimisation avec un F1 de 75%.

Les courbes d'apprentissage

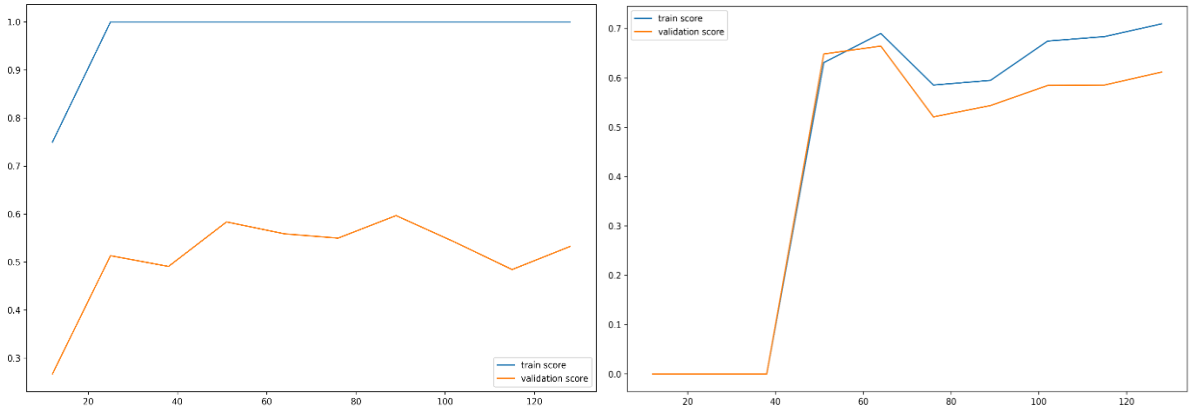


Figure 23: Courbes d'apprentissage d'arbre de décision avant et après optimisation

Nous remarquons que notre modèle avec les paramètres par défaut, il souffre d'un sur ajustement, mais après l'optimisation des paramètres le modèle sera mieux ; cette représentation graphique nous aidera à confirmer notre choix entre avant ou après optimisation.

3.4.1.3 La forêt aléatoire

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

L'optimisation des hyper paramètres du foret aléatoire avec la Gridsearch CV donne les hyer paramètre suivantes :

Tableau 9: Paramètres du foret aléatoire

Les paramètres	Les valeurs
Max_depth	7
Max_features	2
Random_state	30
N_estimators	400
Random_state	30

La comparaison entre les variables importantes avant et après l'optimisation des paramètres avec Grid search

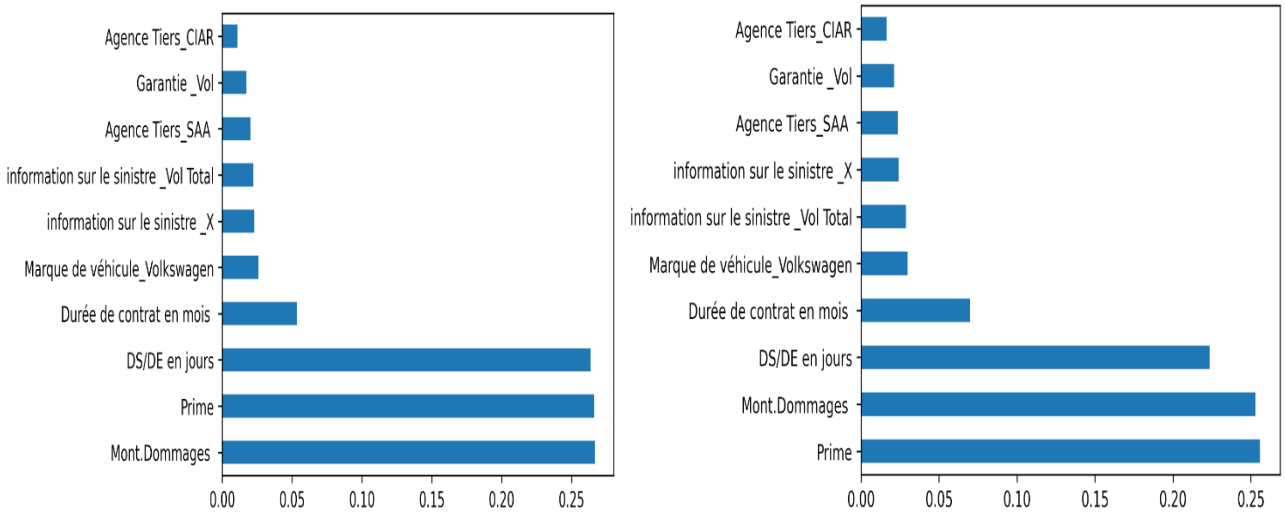


Figure 24: Variables importantes avant et après l'optimisation des paramètres avec Grid search

Les variables les plus influentes sur la prédiction sont clairement les trois mêmes variables importantes pour l'arbre de décision avec des taux différents autour de 27%, après ces trois variables nous trouvons la durée du contrat en mois, son score est inférieur à 10%.

Le reste de ces variables explicatives leur taux d'importance est inférieur à 3%.

Nous remarquons qu'il y a un certain changement dans la liste de ces variables par rapport à l'arbre de décision. Ces variables sont : la marque du véhicule Volkswagen, l'information disponible sur le sinistre est un vol total. L'information disponible sur le sinistre est contre X. Agence tierce SAA, la couverture souscrite est le vol, donc l'agence tierce est CIAR.

CHAPITRE 3 : APPLICATION ET RESULTATS

Comparaison entre les métriques d'évaluation avant et après l'optimisation

Tableau 10: Comparaison entre les métriques d'évaluation avant et après l'optimisation de la foret aléatoire

	Precision	Recall	F1 score	Accuracy
Avant optimisation				0.65
0	0.47	0.57	0.52	
1	0.77	0.69	0.73	
Après optimisation				**0.67**
0	0.50	0.64	0.56	
1	**0.80**	**0.69**	**0.74**	

Pour le foret aléatoire le choix est le modèle après optimisation qui à un taux F1 de 74%.

Les courbes d'apprentissage :

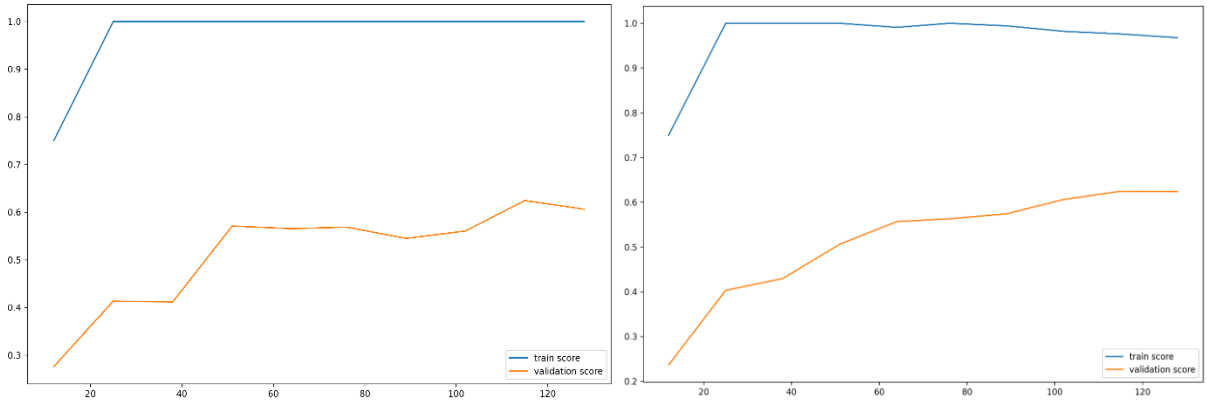


Figure 25: Courbes d'apprentissage avant et après l'optimisation

Selon la courbe d'apprentissage, nous remarquons que notre modèle à gauche, qui est avant l'optimisation, est sur ajusté car il a un score de presque 100% dans l'ensemble d'entraînement mais il est incapable de généraliser sur de nouveaux cas.

Après l'optimisation des paramètres, la courbe d'entraînement commence à converger vers la courbe de validation. Nous pouvons donc dire qu'avec l'augmentation de l'échantillon, sa performance peut augmenter.

3.4.1.4 Modilisation XGBoost

L'optimisation de XG XGBoost avec Gridsearch CV fournit les hyper paramètres suivants :

Tableau 11: Paramètres d' XGBoost

Les paramètres	Les valeurs
Learning_rate	0.1
Max_depth	6
N_estimators	5
Nthread	4
Min_child_weight	11

La comparaison entre les variables importantes avant et après l'optimisation des paramètres avec Grid search :

Pour l'alghorithme XGboost la liste des variables importantes sont :

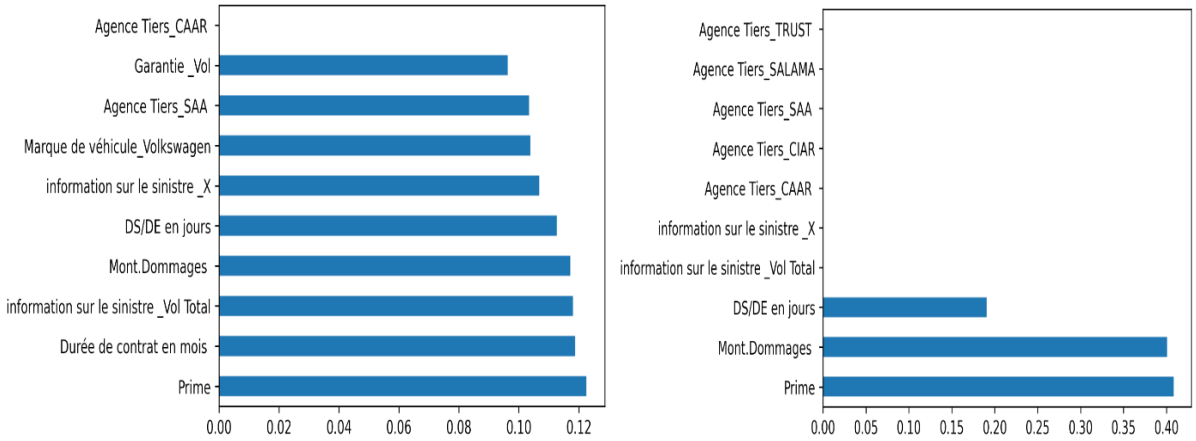


Figure 26: Variables importantes pour XGboost avant et après l'optimisation des paramètres avec Grid search

Nous remarquons qu'avant l'optimisation des paramètres, le modèle XGboost accorde un faible niveau d'importance à plusieurs variables dans une fourchette de 10% à 12%.

Ces variables sont : la prime, la durée du contrat en mois, l'information sur le sinistre est le vol total, le montant des dommages, le DS/DE en jours, l'information sur le sinistre contre X, la marque du véhicule volkswagen, l'agence tierce SAA, et la couverture du vol.

Mais après l'optimisation le nombre de variables importantes est devenu 3, c'est 3 variables sont les mêmes pour l'arbre de décision, le changement réside dans le taux d'importance ici le modèle XGboost donne moins d'importance à la variable DS/DE en jours.

CHAPITRE 3 : APPLICATION ET RESULTATS

Comparaison entre les métriques d'évaluation avant et après l'optimisation

Tableau 12: Comparaison entre les métriques d'évaluation avant et après l'optimisation de l'XGBoost

	Precision	Recall	F1 score	Accuracy
Avant optimisation				0.58
0	0.40	0.57	0.47	
1	0.74	0.59	0.65	
Après optimisation				**0.72**
0	0.60	0.43	0.50	
1	**0.76**	**0.86**	**0.81**	

Lors de la comparaison entre les résultats des deux modèles avant et après l'optimisation, On peut voir que XGBoost après optimisation ne trouve pas beaucoup de difficultés à classer les cas frauduleux car selon le rapport, le modèle peut identifier 86% des cas frauduleux. Alors qu'il manque des dossiers dans les cas non frauduleux, il ne détecte que 43%. Comme dans le cas des modèles précédents, le critère de choix est le Scor F1 qui a un taux de 81%.

Les courbes d'apprentissage

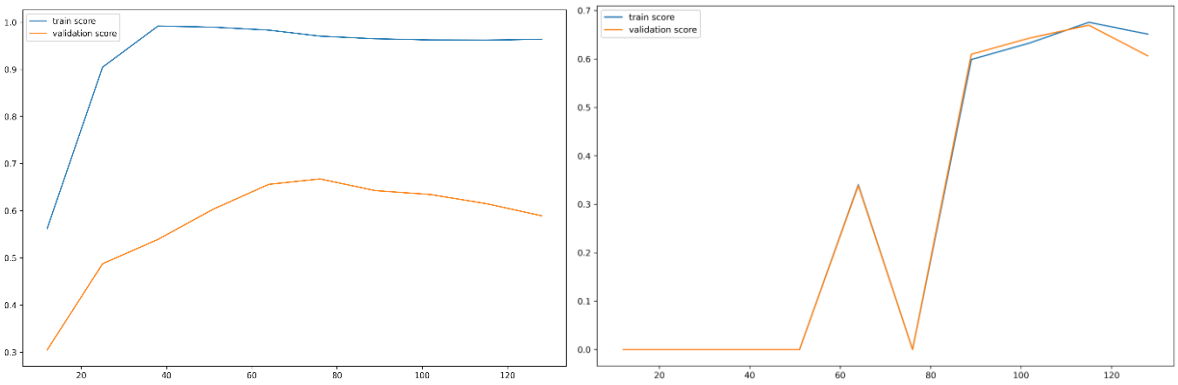


Figure 27: Courbes d'apprentissage avant et après optimisation

Nous remarquons que notre modèle avec les paramètres par défaut, il souffre d'un sur ajustement, mais après l'optimisation des paramètres le modèle sera mieux à la fin.

3.4.1.5 La modilisation par Adaboost

Les hyper paramètres optimale après l'application de Grid search sont :

Tableau 13: Paramètres d'Adaboost

paramètre	valeur
base_estimator__max_depth	2
base_estimator__min_samples_leaf	10
Learning rate	0.01
N_estimators	250

La comparaison entre les variables importantes avant et après l'optimisation des paramètres avec Grid search

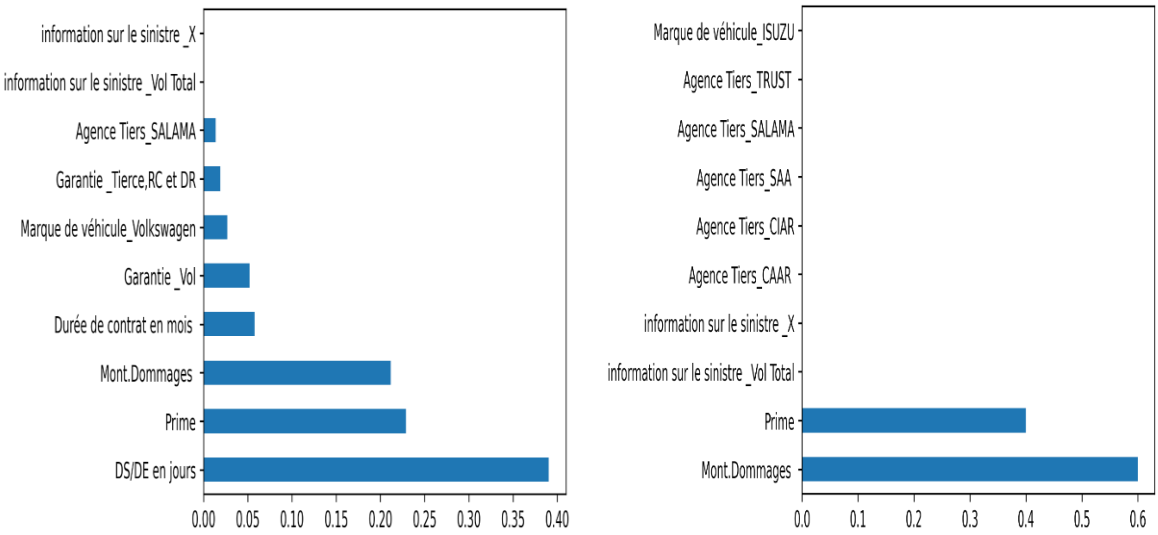


Figure 28: Variables importantes avant et après l'optimisation des paramètres avec Grid search

Nous remarquons qu'avons l'optimisation la variable la plus importante est DS/DE en jours avec un taux presque 40%, suivi par la variable prime et montant de dommages par des taux d'importances de 24% et 22%. Et nous avons également 5 autres variables explicatives qui ont un niveau d'importance inférieur à 5% qui sont la durée de contrat, la garantie vol, la marque de véhicule volkswagen, la garantie tierce, RC et DR, l'agence tiers SALAMA, et l'information disponible sur le sinistre est un vol total.

Mais après l'optimisation des paramètres nous remarquons qu'il y a que deux variables importante pour l'algorithme d'adaboost qui sont le montant de dommages avec un taux de 60% et la prime avec un taux de 40%

CHAPITRE 3 : APPLICATION ET RESULTATS

Les métriques d'évaluation numérique

Tableau 14: Courbe roc après l'optimisation des paramètres du AdaBoost

	Precision	Recall	F1 score	Accuracy
Avant optimisation				**0.63**
0	0.44	0.57	0.50	
1	**0.76**	0.66	**0.70**	
Après optimisation				0.58
0	0.36	0.36	0.36	
1	0.69	**0.69**	0.69	

Dans le cas de l'algorithme ada boost, nous remarquons que le modèle avant optimisation avec les paramètres par défaut est meilleur qu'après optimisation ; c'est pourquoi nous avons choisi d'utiliser le modèle avant optimisation dans la phase de comparaison entre les performances des modèles.

Les courbes d'apprentissage

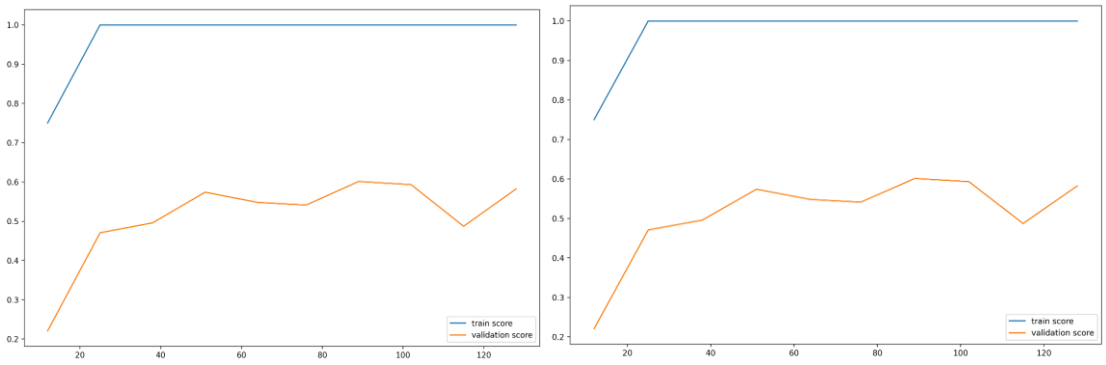


Figure 29: Courbes d'apprentissage avant et après optimisation

D'après les deux courbes, nous remarquons que l'optimisation des paramètres ne donne aucune amélioration au modèle, il souffre d'un sur ajustement dans les deux cas. Parmi les solutions est l'augmentation de la base de l'étude.

3.4.2 Évaluation et comparaison entre les modèles

Nous allons maintenant procéder à l'application de mesures d'évaluation numériques et graphiques dans le but de comparer les performances des modèles achevés.

Évaluation numérique

Tableau 15: Evaluation numérique des cinq modèles

	Precision	**Accuracy (Taux de succès)**	**Recall (Rappel) sensibilité**	**F-measure (F1 score)**
La régression logistique		**0.63**		
0	0.46		0.79	0.58
1	**0.84**		**0.55**	**0.67**
Arbre de décision		**0.65**		
0	0.46		0.43	0.44
1	**0.73**		**0.76**	**0.75**
Foret aléatoire		**0.67**		
0	0.50		0.64	0.56
1	**0.80**		**0.69**	**0.74**
XG boost		**0.72**		
0	0.60		0.43	0.50
1	**0.76**		**0.86**	**0.81**
Adaboost		**0.63**		
0	0.44		0.57	0.50
1	**0.76**		**0.66**	**0.70**

Avant de commencer la comparaison entre les différents modèles nous rappelons que selon les résultats de la partie précédente : nous avons choisi d'utiliser la régression logistique et Adaboost avec les paramètres par défaut, par contre pour la forêt aléatoire et l'arbre de dicision ainsi que XG boost après l'optimisation des paramètres.

En analysant les résultats dans le tableau qui regroupe les scores d'évaluation numériques des cinq modèles sur l'échantillon de test, nous constatons que XGBoost a conduit au taux le plus élevé qui est défini comme la somme de toutes les prédictions correctes divisée par le nombre total de prédictions. Avec un taux de :

Accuracy = (TP+TN)/ (TP+FP+FN+TN)= 72%

Suivi par la forêt Aléaoire, arbre de décision, régression logistique, Adaboost avec des taux respectifs de (67%), (65%), (63%), (63%).

Nous trouvons aussi que la sensibilité (recall) présente un taux satisfaisant pour certain modèle comme XGBoost avec un taux de (86%**),** Arbre de décision (76%) , suivi par la Foret aléatoire (69%), adaboost (66%), d'autre par le score le plus faible enregistré par la régression logistique (55%).

Recall= TP/ (TP+FN)

Ça se voit dans la matrice de corrélation où parmi les 29 dossiers frauduleux dans la base test, le modèle XGboost peut classer correctement 25 dossiers (true possitif=25).

Et pour la régression logistique, elle ne peut classer correctement que 16 des 29 dossiers.

Plus ce recall est élevé, plus le modèle de Machine Learning maximise le nombre de vrai positif. Mais cela ne veut pas dire que le modèle ne se trompe pas. Quand le recall est haut, cela veut plutôt dire qu'il ne ratera aucun positif. Néanmoins cela ne donne aucune information sur sa qualité de prédiction sur les négatifs**.**

En ce qui concerne la précision, la régression logistique a le meilleur score (84%) parce qu'elle a bien classé les dossiers non frauduleux sur les 14 dossiers, elle peut classer correctement 11 dossiers, donc son faux positive est seulement 3.

Précision= TP/ (TP+FP)

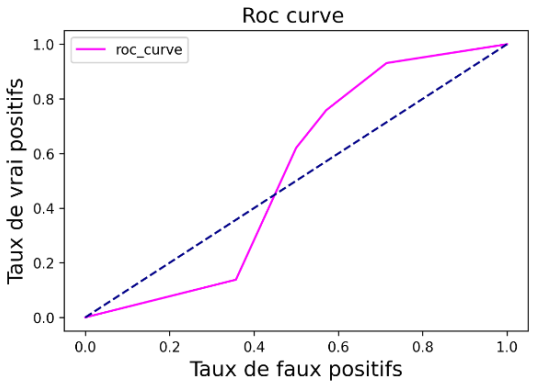
Passons maintenant au la F-mesure, elle a classé XG Boost comme le modèle qui réalise le meilleur compromis dans la prédiction des deux classes avec un taux de (81%), suivi par Arbre de décision (75%), forêt aléatoires(74%), ensuite avec le même classement des autres métrique d'évaluation, nous trouvons ada boost (70%), et la régression logistique de (67%).

Précision= TP/ (TP+FP)

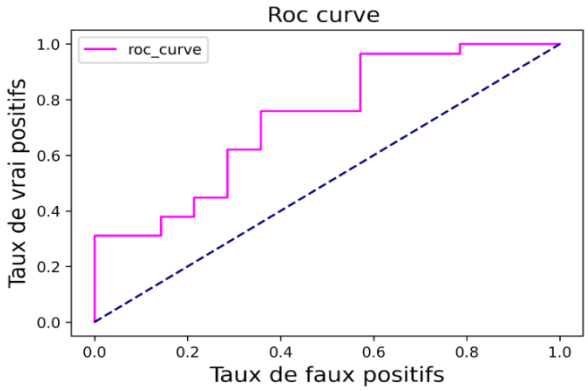
Ainsi, après cette petite analyse des métriques d'évaluation numérique, nous concluons que le meilleur modèle est le XGbosst (qui a les meilleurs scores en F1 mesure, en recalle), puis l'arbre de décision, la forêt aléatoire, puis l'ada bosst et à la fin la régression logistique.

CHAPITRE 3 : APPLICATION ET RESULTATS

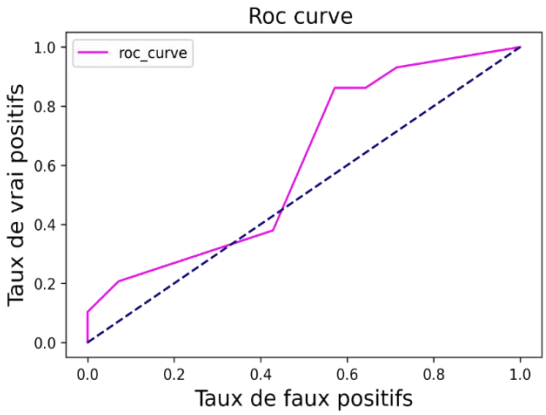
Evaluation graphique



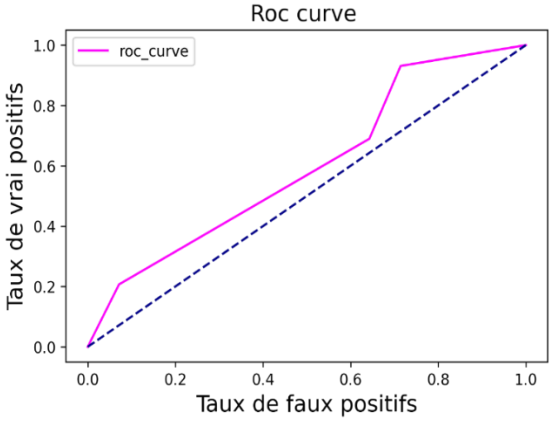
AD : UAC=0.524



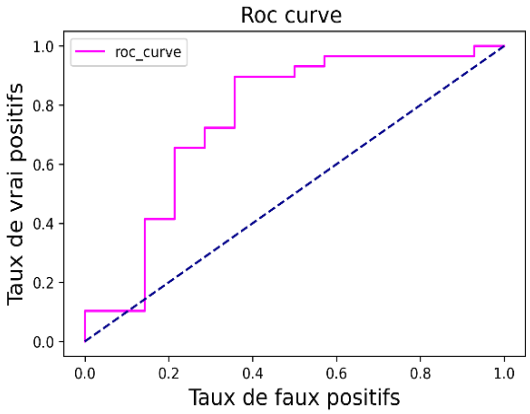
FA : AUC= 0.731



XG Boost: UAC=0.605



Ada Boost: UAC=0.630



RL : UAC=0.75

Figure 30: Courbes rocs des cinq modèles

En outre, nous remarquons que les courbes ROC ne coïncident pas avec la diagonale, ce qui montre que les modèles ne sont pas aléatoires.

Cependant, nous remarquons qu'elles sont également éloignées du coin supérieur gauche, ce qui signifie que les modèles ne sont pas très efficaces. De même pour la valeur de l'AUC.

Selon l'échelle de mesure suivante :

- AUC=0.5 → discrimination aléatoire.
- $0.5 \leq AUC < 0.7$ → discrimination n'est pas aléatoire, performance médiocre.
- $0.7 \leq AUC < 0.8$ → discrimination acceptable, performance moyenne.
- $0.8 \leq AUC < 0.9$ → discrimination excellente, bonne performance.
- $AUC \geq 0.9$ → discrimination exceptionnelle, très bonne performance.

Donc selon ce critère, nous pouvons dire que tout nous modèles sont supérieurs à 0.5 c'est-à-dire ils ne sont pas dans le cas de discrimination aléatoire. Sauf l'arbre de décision son taux est de 0.52 qu'est proche 0.5.

Concernant la régression logistique (0.75) et la forêt aléatoire (0.731), ils sont considérés comme des modèles de discrimination acceptable avec une performance moyenne. Et les autres 3 modèles ont une performance médiocre.

Comme nous avons vu au chapitre deux que parmi les inconvénients du modèle XGBoost figure la complexité d'interprétation, cet algorithme est qualifié de boîte noire. En revanche, la régression logistique est considérée comme l'une des méthodes de classification binaire les plus fiables et les plus faciles à interpréter.

3.5 Conclusion

Au terme de ce chapitre dédié à la partie empirique, nous avons conduit une analyse par les méthodes d'apprentissage supervisé dans le but de modéliser la fraude .Pour y parvenir nous avons suivi le processus CRISP-DM qui est le plus utilisé. La préparation des données a constitué la plus grande partie du travail ; en effet, nous avons inspecté les variables une à une, recodé quelques une, créé de nouvelles et même exclu les moins pertinentes, éliminé des valeurs aberrantes,…

Cette phase a comporté également la sélection des variables explicatives pour non seulement éviter qu'il y ait une redondance dans l'information apportée mais aussi diminuer le temps d'entrainement des classifieurs. Dans cette perspective nous avons utilisé la matrice de corrélation, et la fonction selecte Kbest qui est basée sur les tests de KHI2 et ANOVA. Pour détecter l'existence d'une éventuelle corrélation entre les prédicteurs et de pouvoir l'éliminer

CHAPITRE 3 : APPLICATION ET RESULTATS

La deuxième grande phase nous l'avons consacré à la modélisation dont le choix des algorithmes ; dans notre cas : la régression logistique, l'arbre de décision aussi les trois algorithmes d'ensembles : baging comme la forêt aléatoire, et le boosting comme adaboost et XG boost.

Nous avons comparé la performance de ces modèles avant et après l'optimisation des hyper paramètres par Grid search cv, et le plus performant entre les deux nous l'utilisons dans la phase de comparaison entre les cinq modèles. Nous avons tiré les variables les plus pertinentes pour chaque modèles nous avons remarqué que les variables les plus répétées pour tous les modèles sont le montant de dommages, DS/DE en jours (la durée entre la date de sinistre et la date d'effet), la prime, durée de contrat en mois.

La comparaison des cinq modèles nous a mené à choisir selon le critère de F1 score et recall le modèle de XGboost comme le modèle le plus performant dans la modélisation de fraude en assurance automobile , selon AUC la performance de ce modèle est médiocre et selon ce même critère il y a autres modèles qui sont plus performants que XGboost comme la régression logistique, et la forêt aléatoire.

Comme nous avons vu au chapitre deux que parmi les inconvénients du modèle XGBoost figure la complexité d'interprétation, cet algorithme est qualifié de boîte noire. En revanche, la régression logistique est considérée comme l'une des méthodes de classification binaire les plus fiables et les plus faciles à interpréter.

CONCLUSION GENERALE

Le but de notre étude a été de développer un modèle fiable permettant de prédire la fraude en assurance automobile.

Dans ce travail, nous avons puisé dans la littérature pour tenter de comprendre les notions de base en matière de fraude. En effet, le premier chapitre porte sur le cadre conceptuel de fraude et comment l'appréhender à des fins de modélisation. Nous avons tenté de présenter la fraude à l'assurance automobile en rappelant ses types, ses causes et son impact. Aussi, nous avons donné l'ampleur des sanctions monétaires et pénales qui peuvent être appliquées en cas de détection, les moyens de luttes contre ce phénomène, et les résultats des études déjà réalisées sur la fraude par les méthodes de machine Learning.

Le deuxième chapitre a été consacré aux fondements théoriques propres aux méthodes d'apprentissage supervisé utilisées, à savoir leur principe de fonctionnement d'une part et les métriques de leur fiabilité et performance d'autre part. Cela a contribué à enrichir nos connaissances et à accentuer les différents aspects stratégiques et opérationnels sur ce sujet.

Enfin dans le dernier chapitre nous avons parcouru les différentes phases permettant la concrétisation de nos modèles du prétraitement au post traitement passant par l'analyse d'une base de données qui représente l'ensemble des déclarations de sinistres douteux jugées frauduleuses ou non par ALFA (et confirmé par le gestionnaire d'assurance).

Pour ce faire, nous avons utilisé cinq méthodes d'apprentissage supervisé, dont la régression logistique, qui est l'un des algorithmes les plus anciens et les plus utilisés dans le domaine des assurances pour prédire les variables qualitatives.

Nous avons également choisi de traiter l'arbre de décision, car il s'agit d'un élément important nécessaire au développement d'autres algorithmes d'apprentissage d'ensemble comme la forêt aléatoire. Selon les travaux de recherche précédents cette technique avait la meilleure performance par rapport aux autres techniques classiques, ensuite nous avons traité l'algorithme d'ada boost qui est aussi basé sur l'algorithme de l'arbre de décision et à la fin nous avons appliqué un algorithme de XG Boosting qui est le plus récent par rapport à l'autre créé en 2016. Pour que nous puissions choisir l'algorithme le plus performant qui détecte mieux le dossier frauduleux.

Les résultats de la recherche :

Enfin nous avons abouti aux conclusions suivantes :

La fraude en assurance automobile implique qu'une personne tente de tromper une compagnie d'assurance au sujet d'une réclamation impliquant son véhicule à moteur personnel ou commercial. Il peut s'agir de donner des informations trompeuses ou de fournir de faux documents à l'appui de la réclamation avec une mauvaise foi.

L'une des méthodes de détection de la fraude en assurance automobile est l'apprentissage automatique qui est un champ de l'intelligence artificielle. Il utilise un nombre limité d'entrée d'un système avec les valeurs de leurs sorties pour apprendre une fonction qui décrit la relation fonctionnelle existante, mais non connue, entre les entrées et les sorties du système.

La statistique associée à de puissantes machines de traitement et de stockage a révolutionné le monde de l'apprentissage automatique donnant naissance à des modèles capables d'agir aussi bien sur les petits échantillons que sur les grands

La phase de prétraitement est indispensable dans la démarche de modélisation et a des répercussions directes sur la capacité de prédiction d'un algorithme.

Les cinq algorithmes appliqué s'accordent sur l'ordre d'importance des variables explicatives, notamment celles qui permettent une meilleure discrimination entre les assurés, nous citons le montant de dommages, DS/DE en jours (la durée entre la date de sinistre et la date d'effet), la prime, durée de contrat en mois.

En basons sur le recall et F1 score, nous sommes arrivés à la conclusion que le modèle le plus performant est le XGBoost. Mais selon l'indice d'AUC, nous trouvons d'autres algorithmes plus performants, comme la régression logistique, le foret aléatoire. Nous obtenons ces résultats par ce que notre échantillon est de petite taille.
Nous pouvons donc conserver les cinq modèles et les utiliser sur une base plus large pour obtenir de meilleurs résultats.

Limites de la recherche

Cependant, de tels apports ne doivent pas occulter les limites inhérentes, voire propres à toute contribution qui se veut scientifique. Avant toute chose la principale restriction de notre

travail de recherche relève au non obtention de la base qui répond à notre problématique de départ donc nous trouvons dans le cas de manque de nombre de variables explicatives et de nombre d'observations.

La deuxième limite porte sur les méthodes de prétraitement utilisé, il s'agit d'un domaine très vaste regorgeant de théories que nous n'avons pas pu investiguer en détail faute de temps.

Les perspectives de la recherche

A des fins d'approfondissement, nous proposons, d'une part, d'élargir le choix des variables exogènes en intégrant, des variables qualitatives et quantitatives en plus de celles utilisée, d'explorer les méthodes de détections des valeurs aberrantes multivariées, utiliser d'autres fonction de sélection des variables comme la méthode d'FRE CV.

Aussi pour améliorer les modèles utilisés faire appel au Stacking. Il s'agit de combiner plusieurs algorithmes afin de concevoir un modèle plus robuste et plus performant.

Ou utiliser une deuxième catégorie d'apprentissage automatique qui est l'apprentissage non supervisé qu'il s'agit de détecter les similarités dans les données non étiquetées et de créer ainsi des classes ou bien des groupes d'individus homogènes présentant des caractéristiques similaires et communes. Par ailleurs, la détection d'anomalies constitue une catégorie d'apprentissage non supervisé. Cette méthode permet de résoudre des problèmes liés à la fraude. En effet, en raison du faible nombre de cas de fraudes avérées, la fraude est considérée comme une anomalie ou un évènement rare.

Pour améliorer ces résultats aussi , les compagnies doivent renforcer leurs moyens de détection de fraude en améliorant leurs système d'information, en l'occurrence la qualité des données nécessaires dans la modélisation, en intégrant les démarches de recherche de fraude dès la souscription du contrat, et également en se dotant des technologies appropriées. La détection de fraude est devenue un enjeu stratégique pour l'assureur dont objectif est de disposer d'un système efficace et automatisé en matière de lutte contre la fraude.

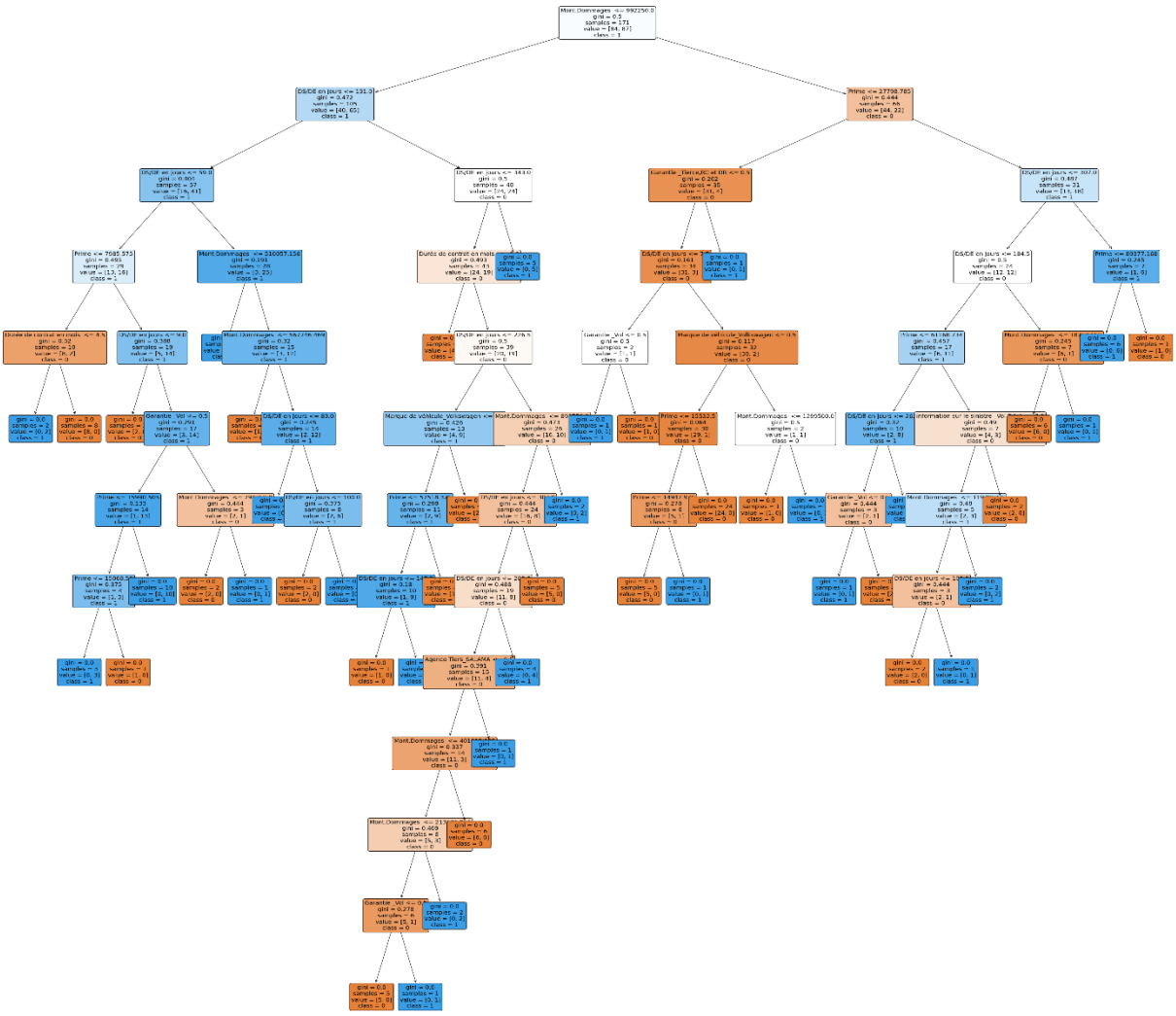
En effet, la démonstration de la fraude est le produit d'une analyse humaine basée sur la constatation d'un scénario, la fraude n'existe que si le conseiller l'a préalablement identifiée et analysée. Ceci constitue le principal biais dans cette étude et permet de mettre en évidence la nécessité d'un investissement métier préalable pour améliorer l'échantillon.

Nous ne terminerons pas notre présent travail sans souligner que cette expérience nous a été très bénéfique. Bien que ce fut compliqué de construire un modèle performant en utilisant des données appartenant au monde réel. Il n'y a aucun processus linéaire permettant d'arriver à

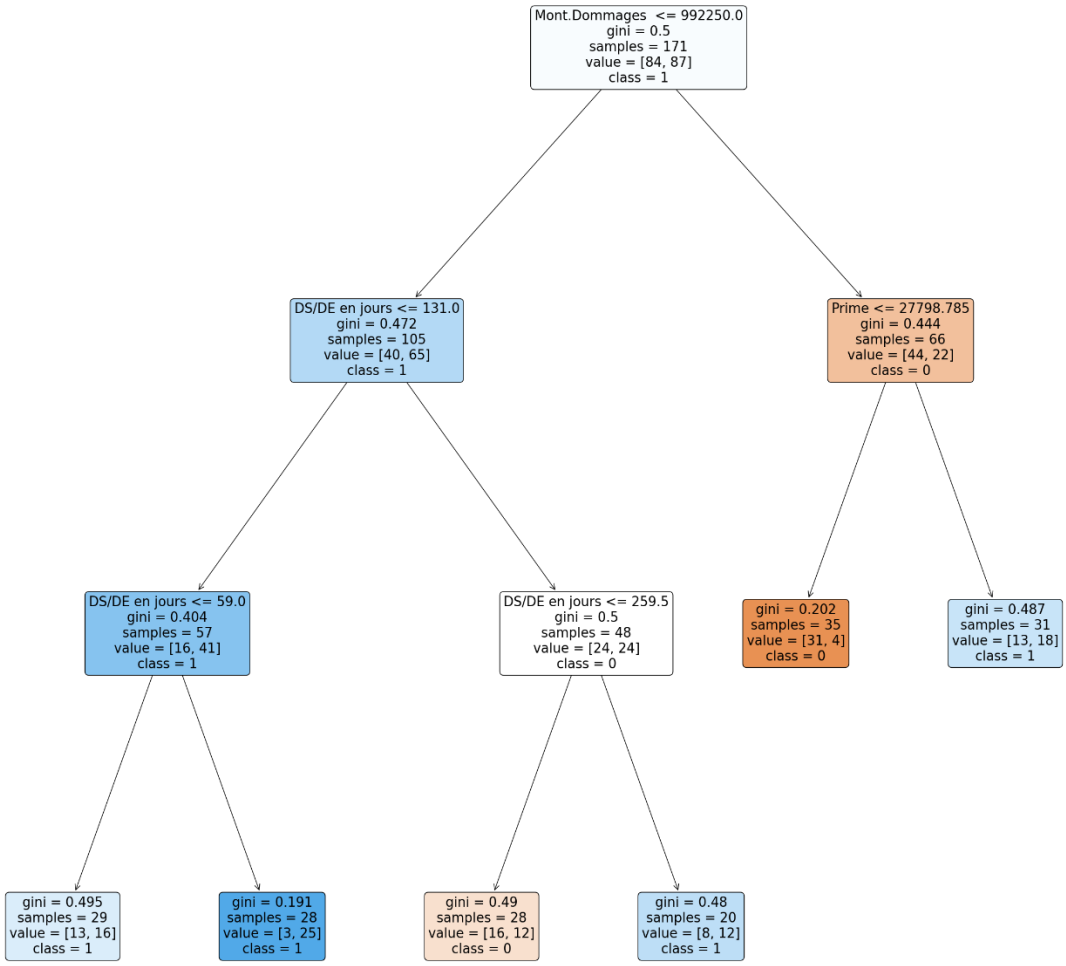
des conclusions rapides. Il est nécessaire de combiner plusieurs méthodes et faire des va-et-vient sur chaque phase pour tenter d'arriver à un résultat pertinent tant est que celui-ci existe.

LES ANNEXES

Annexe 1: Résultats de l'Arbre décision avant l'optimisation des paramètres avec Grid search



Annexe 2: Résultats de l'Arbre décision après l'optimisation des paramètres avec Grid search



BIBLIOGRAPHIE

- Bahzad, T. J., & Adnan Mohsin, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends, 02*(01), 20-28.

- Bel Mufti, G. (2021). Cours sur data minig,ESSAI

- Benzaki, Y. (2017, novembre 27). *Tout savoir sur les Valeurs Aberrantes (Outliers)*. Récupéré sur https://mrmint.fr/outliers-machine-learning

- Bouzgarne, I., Youssfi, M., Qbadou, M., & Bouattane, O. (2019). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. *IEEE*.

- *datascientest*. (2020, oct 19). Récupéré sur Algorithmes de Boosting – AdaBoost, Gradient Boosting, XGBoost: https://datascientest.com/algorithmes-de-boosting-adaboost-gradient-boosting-xgboost

- Dhieb, N., Ghazzai, H., & Besbes, H. (2019). Extreme Gradient Boosting Machine Learning Algorithm For Safe Auto Insurance Operations. *IEEE*.

- Elreedy, D., & Atiya, A. F. (2019). *Information Sciences Volume 505*. (W. Pedrycz, Éd.) Récupéré sur Pages 32-64 A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance.

- Emmanuel, R., Vanessa, V., Emilie, D., & Bruna, M.-C. (2021). *Revue des principales approches de résolution du.* Récupéré sur https://hal.archives-ouvertes.fr/hal-03480661/document

- Gérard, F., Gondran, M., Lacomme, P., & Samir, C. (2022). *ellipses.* (é. m. ellipses, Éd.) Récupéré sur Informatique - découverte du machine learning - Les outils de l'apprentissage automatique cedex 15 Page16: https://www.editions-ellipses.fr/accueil/13524-24830-informatique-decouverte-du-machine-learning-les-outils-de-l-apprentissage-automatique-9782340047334.html#/1-format_disponible-broche

- Hotz, N. (2022, Aout 8). *Qu'est-ce que CRISP DM ?* Récupéré sur Alliance des processus de science des données: https://www.datascience-pm.com/crisp-dm-2/

- Hounsinou, M. (2021). *Memoire online.* Récupéré sur Amélioration l'estimation des sinistres responsabilité civile automobile par machine learning:

https://www.memoireonline.com/02/22/12649/m_Amelioration-lestimation-des-sinistres-responsabilite-civile-automobile-par-mac1.html#_Toc78577112

- *IFPA*. (2022). (Pennsylvania) Récupéré sur Insurance Fraud Prevention Authority: https://helpstopfraud.org/insurance-fraud/types-of-insurance-fraud/
- Japkowicz, N., & Shah, M. (2014). *Evaluating Learning Algorithms: A Classification Perspective* (éd. 1 st). (E-book, Éd.) New york, USA: Cambridge University Press.
- Maula, I., Prasasti, N., Dhini, A., & Laoh , E. (2020). Automobile Insurance Fraud Detection using Supervised Classifiers. *IEEE*.
- Recueil des guides de Gestion de L'assurance « Automobile » en Algérie
- Scikit-Learn. (s.d.). Récupéré sur Randomforest classifier scikit-learn: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
- Scikit-learn. (s.d.). *Scikit-learn*. Récupéré sur Decisiontree classifier scikit-learn: https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- Shailesh, K. (2019). Essentials of Business Analytics. (S. Seshadri, & P. Bhimasankaram, Éds.) p 507–568 chapitre Machine Learning (Supervised). Récupéré sur pp 507–568 chapitre Machine Learning.
- Shamsudheen, M. (2020, fevrier 9). *DataDrivenInvestor*. Récupéré sur Techniques d'augmentation des données dans l'apprentissage en profondeur: https://medium.datadriveninvestor.com/data-augmentation-techniques-in-deep-learning-d78c59be8ea5
- SPLCFA. (2011, 03 07). Agence de lutte contre la fraude (SPA-ALFA). *Séminaire sur la Prévention et la Lutte Contre la Fraude à l'Assurance en Algérie*. Centre de formation CAAR: Benis Messous.
- Tremblay, C., & Clément, C. (2022, Janvier 20). Récupéré sur SMOTE et données mixtes, traiter les variables catégorielles avec SMOTE-NC: https://kobia.fr/imbalanced-data-smote-nc/
- *Wiley Online Library*. (2019). Récupéré sur Supervised Learning—Classification Using Logistic Regression Python® Machine Learning p151–175.

- Zipporah, L. (Éd.). (2021). *Feature Selection in Machine Learning*. Récupéré sur https://medium.com/geekculture/feature-selection-in-machine-learning-correlation-matrix-univariate-testing-rfecv-1186168fac12

- Zouari M.cours.(2021). risque opérationels, IFID

TABLE DES MATIÈRES