REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique

Département Génie Industriel

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'ingénieur d'Etat en Génie Industriel

Amélioration du processus de gestion du risque douanier et de lutte contre la fraude par les outils de l'Intelligence Artificielle

Présenté par :

Fatima Zohra AOUALI (Management Industriel)

Sandra Lydia METSAHA (Management Industriel)

Présenté et soutenue publiquement le (01/07/2019)

Composition du jury :

Président	Mme Nacéra ABOUN	MAA	ENP
Promoteur	M. Mabrouk AIB	DOCTEUR	ENP
Examinateur	M. Wassim BENHASSINE	MCA	ENP

ENP 2019

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique

Département Génie Industriel

Mémoire de projet de fin d'études

Pour l'obtention du diplôme d'ingénieur d'Etat en Génie Industriel

Amélioration du processus de gestion du risque douanier et de lutte contre la fraude par les outils de l'Intelligence Artificielle

Présenté par :

Fatima Zohra AOUALI (Management Industriel)

Sandra Lydia METSAHA (Management Industriel)

Présenté et soutenue publiquement le (01/07/2019)

Composition du jury :

Président	Mme Nacéra ABOUN	MAA	ENP
Promoteur	M. Mabrouk AIB	DOCTEUR	ENP
Examinateur	M. Wassim BENHASSINE	MCA	ENP

ENP 2019

DEDICACES

À la mémoire de ma grand-mère TITIS, que Dieu l'accueille dans son vaste paradis.

À mon grand-père DJADIS, que Dieu t'accorde une longue vie parmi nous.

À mon très cher père, qui m'a accompagnée par sa patience sans fin, son amour, ses longues années de sacrifices pour me voir réussir, ses prières, sa bénédiction, que dieu te protège et t'accorde une longue vie.

À ma source de bonheur, la femme qui m'a ramenée au monde et a fait de moi ce que je suis aujourd'hui qui m'a comblée par son amour, qui s'est sacrifiée pour mon bonheur et ma réussite : ma mère.

À mon frère BILLEL, toi qui m'a donné tant de choses et tu continues à le faire, sans jamais te plaindre. Je te dois ce que je suis aujourd'hui et ce que je serai demain et je ferai toujours de mon mieux pour rester ta fierté et ne jamais te décevoir.

A ma tante SOUSOU Je suis reconnaissante, pour ton soutien moral et tes encouragements m'ont toujours aidée à aller de l'avant,

À ma meilleure amie, ma sœur et éternelle confidente MANOU qui n'a jamais cessé de partager avec moi les bons et pires moments tout au long de ce cursus, et d'être constamment là pour moi.

À mon binôme SANDRA, avec qui j'ai partagé des moments très tchatcheurs

Et toute la famille AOUALI, KHELLOUL et ARZANI.

Ainsi à mes très chers amis YASMINE, IMEN, HOCINE, DRISS.

Je dédie ce mémoire

FATIMA ZOHRA

DEDICACES

À mes très chers parents, pour tous leurs sacrifices, leur dévouement, et leur soutien tout au long de mes études.
Que dieu les protège.
À mes chers sœurs Yasmine et Maya ainsi qu'à mon unique frère Yanis
À la mémoire de mes grands-parents que Dieu les accueille dans son vaste paradis.
A mon binôme qui a eu la gentillesse d'endurer mes sautes d'humeur
À tous mes amis, et à tous ceux que j'aime

Je dédie ce mémoire

Sandra

REMERCIEMENTS

Nous remercions en premier lieu Dieu le tout puissant qui nous a doté de la merveilleuse faculté du raisonnement et nous a donné le courage et la volonté de mener à terme le présent travail

Nos plus vifs remerciements sont adressés à notre promoteur ; Dr Mabrouk AIB, pour nous avoir encadré. Les recommandations qu'il nous a prodigué nous ont été d'un grand apport et nous ont permis d'améliorer considérablement la qualité de ce mémoire.

Nous tenons à témoigner notre profonde gratitude envers Monsieur Ammar MELLIANI et Mme Assia IBOU qui nous ont guidés attentivement tout au long de notre projet. Leur rigueur, leurs conseils éclairés, leurs suggestions pertinentes et leur expertise nous ont permis de mener à bien ce travail.

Nous exprimons toutes notre reconnaissance aux experts de la Banque Mondiale ; Pr. Dalila BENAHACHNOU ainsi que Mr. Jean François ARVIS pour nous avoir fait bénéficier de leurs compétences scientifiques, et pour nous avoir accordé leurs constantes disponibilités, et leur suivi permanent dans l'élaboration de ce travail.

Nous tenons aussi à remercier l'ensemble des enseignants du département du Génie Industriel qui ont contribué à notre formation ainsi qu'aux membres du jury qui ont accepté d'évaluer notre travail.

الملخص

شرعت الجمارك الجزائرية في عملية تحديث تعتمد على نشر تكنولوجيا المعلومات. هذا العمل جزء من هذا البرنامج ويهدف إلى تحسين عملية إدارة المخاطر الجمركية ومحاربة الاحتيال؛ وذلك باستخدام أداة دعم القرار. يهدف هذا العمل إلى توجيه وترشيد الرقابة الجمركية بصفة تلقائية وسيستند إلى أدوات الذكاء الاصطناعية.

يتكون النهج المستخدم من تصميم وتقييم وتصنيف خوارزميات التعلم الخاضعة للإشراف المختلفة؛ ثم اختيار النموذج الذي سيتم تكييفه لمشكلة هذا المشروع والتي ستحقق الأهداف التي وضعتها المديرية العامة للجمارك الجزائرية.

ستتم مناقشة حدود الأداة التي تم تطويرها وسيتم اقتراح حل لمواجهتها. والذي يعتمد على طرق تعلم غير خاضعة للإشراف وسيتم استخدامه للكشف عن عمليات احتيال غير مسبوقة.

لكلمات المفتاحية: الذكاء الاصطناعى، التعلم الآلى، التنقيب فى البيانات، المخاطر الجمركية، الاحتيال.

Abstract

The Algerian Customs is engaged in a modernization process based on the deployment of information technologies. This work is part of this program and aims to improve the process of customs risk management and the fight against fraud; using a decision-making tool. It helps to guide and rationalize customs control automatically and is based on Artificial Intelligence techniques.

The approach used consists in the design, evaluation and classification of different supervised learning algorithms; then the selection of the model that is adapted to the goals and constraints of this project and that allow to achieve the objectives set by the Algerian Customs.

The limitations of the tool developed will be discussed and a solution proposed to address them. It will be based on unsupervised learning methods and will be used to detect undetected types of fraud.

Keywords: Artificial Intelligence, Machine Learning, Data Mining, Customs risk, Fraud.

Résumé

La Douane Algérienne est engagée dans un processus de modernisation qui s'appuie sur le déploiement des technologies de l'information. Ce travail s'inscrit dans le cadre de ce programme et tend à améliorer le processus de gestion du risque douanier et de lutte contre la fraude à l'aide d'un outil d'aide à la décision. Il a pour but d'orienter et de rationaliser le contrôle douanier de manière automatique et se base sur les outils de l'Intelligence Artificielle.

L'approche employée consiste en la conception, l'évaluation et le classement de différents algorithmes d'apprentissage supervisé ; puis la sélection du modèle qui sera adapté à la problématique de ce projet et qui permettra d'atteindre les objectifs fixés par la Direction Générale des Douanes Algériennes.

Les limites de l'outil développé seront abordées et une solution sera proposée pour les contrer. Elle se basera sur les méthodes d'apprentissage non supervisé et servira à détecter des fraudes inédites.

Mots clés : Intelligence artificielle, Machine Learning, Data Mining, Risque douanier, Fraude.

Table des matières

Liste des figures

Liste des tableaux

Listes des abréviations

INTRODUCTION GENERALE 12

Chapitre 1: Etat des lieux et diagnostic 15

I Présentation des entreprises et des parties concernées 15
I.1 Le groupe de la Banque 15
I.2 La direction générale des douanes Algériennes 15
I.2.1 Organisation de la direction générale des douanes 16
I.2.2 Les missions de l'administration des Douanes 18

II Diagnostic 20
II.1 Les procédures de dédouanement 20
II.1.1 Les formalités préalables au dédouanement 20
II.1.2 Les formalités du dédouanement 20
II.2 Les facilitations douanières 23
II.2.1 La gestion des risques en douane 24
II.2.2 L'Opérateur Economique Agréé OEA 29
II.3 La fraude 29
II.3.1 Définition 29
II.3.2 Les causes de la fraude commerciale 30
II.3.3 Les types de fraude commerciale 31
II.4 Résultat du diagnostic 33

III Enoncé de la problématique 35

Chapitre 2: Etat de l'art 37

I Machine Learning 37
I.1 L'intelligence artificielle 37
I.2 Définition du Machine Learning 37
I.3 Les données d'apprentissage 38
I.4 Les types d'apprentissage automatiques 38
I.5 Les algorithmes d'apprentissage supervisé 39
I.5.1 Les arbres de décision 40
I.5.2 Les forêts aléatoires 42
I.5.3 AdaBoost 44
I.5.4 Le Gradient Boosting 45
I.5.5 Extrem Gradient Boosting (XGBoost) 48
I.5.6 Les k plus proches voisins 51
I.5.7 Les machines à vecteur support SVM 52
I.6 Algorithmes d'apprentissage non supervisé 55
I.6.1 DBSCAN 55

I.6.2 Les k-moyennes (k-means) 57
I.6.3 Classification ascendante hiérarchique (CAH) 57
I.6.4 Méthodes d'apprentissage non supervisé issues des réseaux de neurones 59
I.7 Evaluations des performances d'un classifieur 63
I.7.1 La matrice de confusion 63
I.7.2 Les mesures de base dérivées de la matrice de confusion 64
I.7.3 La courbe ROC (Receiver Operating Characteristic) 65
I.7.4 ROC AUC (The area under the ROC curve) 66
II Le Data Mining 66
II.1 Définition du Data Mining 67
II.2 Les objectifs du Data Mining 67
II.3 Le processus de Data Mining 67
II.4 Les types de variables 69
III Le logiciel R 70
Chapitre 3: Solution proposée et son application 72
I La compréhension du métier 72
II La compréhension des données 74
III La préparation des données 77
IV La modélisation 83
IV.1 L'auto-encodeur 83
IV.2 Les modèles supervisés 87
IV.2.1 Random Forest 87
IV.2.2 Les machines à vecteurs supports 88
IV.2.3 Les k plus proches voisins 90
V L'évaluation 92
VI Le déploiement 94
Chapitre 4: Traitement des limites du modèle supervisé 102
I Implémentation des modèles avec toutes les variables 102
I.1 Conception des modèles 102
I.1.1 K-means 103
I.1.2 DBSCAN 104
I.1.3 Carte de Kohonen 105
I.2 Discussion des résultats 106
II Implémentation des cartes de Kohonen avec un nombre restreint de variables 106
III Implémentation des cartes de Kohonen avec des variables de dimensions réduites 108
Conclusion générale 111
Bibliographie 113
ANNEXES 116

Liste des figures

Figure 1-1 : Organigramme de l'administration des douanes Algériennes 17
Figure 1-2 : Le processus de la gestion du risque douanier 24
Figure 1-3 : Le fonctionnement du système d'orientation des déclarations vers les circuits d'inspection 26
Figure 1-4 : Hiérarchisation des risques 27
Figure 1-5 : La répartition des déclarations sur les circuits de dédouanement 33
Figure 1-6 : La distribution des déclarations sur les trois circuits de dédouanement 34
Figure 2-1: Les algorithmes d'apprentissage supervisés 39
Figure 2-2 : Exemple d'un arbre de décision 40
Figure 2-3 : Exemple illustrant le déroulement de Random Forest 43
Figure 2-4 : Exemple illustrant le déroulement d'AdaBoost 45
Figure 2-5 : Exemple illustrant le déroulement du Gradient Boosting 48
Figure 2-6 : Classification avec l'algorithme des k plus proches voisins 51
Figure 2-7 : Hyperplan optimal avec une marge maximale 52
Figure 2-8 : Cas linéairement séparable 53
Figure 2-9 : Cas non linéairement non séparable 53
Figure 2-10 : les algorithmes d'apprentissage non supervisés 55
Figure 2-11 : Point central 55
Figure 2-12 : Point bordure 56
Figure 2-13 : Exemple des différents points 56
Figure 2-14 : Dendrogramme de la Classification Ascendante hiérarchique 58
Figure 2-15 : Perceptron monocouche 59
Figure 2-16 : Réseau de Kohonen 60
Figure 2-17 : Carte de Kohonen 62
Figure 2-18 : Auto-encodeur 63
Figure 2-19 : Courbe ROC 66
Figure 2-20 : Le modèle CRISP-DM 69
Figure 3-1 : Etapes et outils de la méthodologie CRISP-DM 73
Figure 3-2 : Les valeurs manquantes 75
Figure 3-3 : Boite à moustaches pour la comparaison des taux 76
Figure 3-4 : Anomalies dans les codes postaux 76
Figure 3-5 : Courbe de la variance intra-classe 78
Figure 3-6 : Dendrogramme résultant de la CAH 79
Figure 3-7 : La relation entre la variable section majoritaire et la variable cible fraude 80
Figure 3-8 : Transformation One-Hot-Encoding 82
Figure 3-9: Constitution des bases de données de l'auto-encodeur 84
Figure 3-10 : L'algorithme de l'auto-encodeur 85
Figure 3-11 : Distribution de l'erreur de reconstruction (MSE) des données frauduleuses et non frauduleuses 86
Figure 3-12 : Algorithme Random Forest 87
Figure 3-13 : Algorithme SVM 89
Figure 3-14 : Algorithme Knn 90

Figure 3-15 : L'algorithme XGBoost 91
Figure 3-16 : Coubre ROC des modèles d'apprentissage supervisé 92
Figure 3-17 : Le cutt-off 94
Figure 3-18 : Cut-off optimum selon Random Forest 95
Figure 3-19 : Algorithme de redéfinition du cutt-off 95
Figure 3-20 : Affectation des marchandises aux circuits de vérification 96
Figure 3-21 : Mode opératoire de l'outil d'aide à la décision proposé 97
Figure 3-22 : Diagramme en camembert représentant la proportion des déclarations en chaque circuit 98
Figure 3-23 : Schéma récapitulatif de l'outil proposé ..... Erreur ! Signet non défini.
Figure 4-1 : Algorithme k-means 103
Figure 4-2 : Résultats obtenus par k-means 103
Figure 4-3 : Algorithme DBSCAN 104
Figure 4-4 : Résultats obtenus par DBSCAN 104
Figure 4-5 : Algorithme Kohonen 105
Figure 4-6 : Distribution des déclarations sur la carte de Kohonen selon l'ensemble des variables 105
Figure 4-7 : Classement des variables selon le gain de Gini par ordre d'importance ... 107
Figure 4-8 : Distribution des déclarations sur la carte de Kohonen selon les variables sélectionnées grâce à l'indice de Gini 107
Figure 4-9 : La visualisation en 3D des déclarations avec la méthode de réduction T-SNE 108
Figure 4-10 : Distribution des déclarations sur la carte de Kohonen selon les trois dimensions obtenues par T-SNE 109
Figure 4-11 : Cluster obtenus par la carte de Kohonen 109

Liste des tableaux

Tableau 2-1 : Matrice de confusion 63
Tableau 2-2 : Types de variable 69
Tableau 3-1: Transformation subies par la variable pays 79
Tableau 3-2 : Découpage de l'Algérie en cinq zones géographies 81
Tableau 3-3 : La matrice de confusion du modèle à noyaux linéaires 90
Tableau 3-4 : La matrice de confusion du modèle à noyaux polynomiale 90
Tableau 3-5 : Matrices de confusion des modèles d'apprentissage supervisé 93
Tableau 3-6 : Performance des modèles d'apprentissage supervisé 93
Tableau 3-7 : Classement des modèles d'apprentissage supervisé 93
Tableau 3-8 : Affectation des déclarations aux circuits de vérification par l'outil d'aide à la décision. 97

Listes des abréviations

- **AdaBoost** : Adaptive Boosting
- **AID** : Association internationale de développement
- **AUC** : Area Under the ROC curve
- **BIRD** : Banque internationale pour la reconstruction et le développement
- **CAH** : Classification Ascendante Hiérarchique
- **CIRDI** : Centre international pour le règlement des différends relatifs aux investissements
- **CRISP-DM** : Cross Industry Standard Process for Data Mining
- **DBSCAN** : Density-Based Spatial Clustering of Applications with Noise
- **DGD** : Direction Générale des Douanes
- **kNN** : k Nearest Neighbors
- **MIGA** : Agence multilatérale de garantie des investissements
- **OEA** : Opérateur Economique Agréé
- **PIB** : Produit Intérieur Brut
- **ROC** : Receiver Operating Characteristic
- **SIGAD** : Système d'Information et de Gestion Automatisée des Douanes
- **SVM** : Support-Vector Machine
- **XGBoost** : Extrem Gradient Boosting

INTRODUCTION GENERALE

La mondialisation des échanges et le poids de plus en plus déterminant du commerce extérieur dans toutes les économies nationales imposent un assouplissement de la circulation transfrontalière des marchandises et des facteurs de production. Dans ce contexte de libéralisation des échanges entre pays et d'ouverture des marchés, les pays en développement dont l'Algérie fait partie, sont contraints d'accroitre leur compétitivité en matière de commerce international afin de faire face à la concurrence globale qui en résulte.

Dans ce contexte, l'Algérie a dû transiter d'un système économique dirigiste et centralisé, marqué par le monopole de l'Etat pour tout ce qui concernait le commerce extérieur, vers une économie de marché en ouvrant ses frontières à la concurrence internationale.

Cependant, l'ouverture du marché national par la levée des restrictions à l'importation et à l'exportation, pose de multiples problèmes liés à l'inexpérience des intervenants nationaux dans la chaîne logistique. De plus, la complexité des procédures commerciales dans le pays implique une augmentation des coûts et des délais de dédouanement compte tenu de la densité des flux commerciaux générée par cette ouverture.

L'administration des douanes, qui est une institution stratégique dans la mise œuvre d'une politique commerciale, a alors été dotée d'une nouvelle mission d'ordre économique pour accompagner cette évolution structurelle. Dans cette optique, une stratégie de modernisation a été mise au point depuis 1994 à son niveau. Elle a notamment intégré dans sa démarche les recommandations de la convention internationale de Kyoto révisée pour la simplification et l'harmonisation des régimes douaniers, en apportant de nombreuses simplifications et facilitations à ses procédures. Néanmoins, ces facilitations ne doivent pas se faire au détriment des impératifs de lutte contre la fraude dont l'administration des douanes a la responsabilité.

En effet, l'un des rôles majeurs de la douane est d'assurer le respect des dispositions légales et règlementaires que les pouvoirs publics ont édicté. Elle doit alors identifier et réprimer tout fraudeur essayant de réaliser des profits illégaux au dépend de l'intérêt économique de l'Etat en lui causant des pertes de recette fiscale d'une part, et empêcher l'introduction de produits prohibés sur le territoire national d'autre part.

D'après la loi de finance 2018, les recettes douanières représentent 3 milliards US$. Etant donné l'importance de ces recettes pour l'économie nationale et de la sensibilité sécuritaire des opérations, la douane avait choisi d'opérer un contrôle exhaustif des containers de marchandises. Bien que l'on puisse comprendre l'intérêt de cette approche, ceci engendre une augmentation significative des coûts et des délais de dédouanement et constitue une entrave au commerce extérieur ainsi qu'une contradiction avec les objectifs de facilitation que la douane s'est fixée.

Afin d'arriver à concilier ses deux missions en contrôlant efficacement tous les flux sans pour autant perturber la fluidité des échanges, un système moderne de gestion du risque douanier a été mis en place selon les recommandations de la convention de Kyoto révisée. Il repose sur la rationalisation du contrôle douanier dans la sélection des opérations à soumettre au contrôle.

En 2018, l'Algérie n'occupait que la 157ème place au classement *Doing Business* établi par la banque mondiale (Groupe de la Banque mondiale, 2019), Cet indice classe les économies de 1 à 190, la première place étant la meilleure. Un classement élevé signifie que l'environnement réglementaire du pays est favorable aux activités commerciales. Ce qui n'est pas le cas de l'Algérie, malgré les efforts de modernisation mis en place au niveau de l'administration des douanes.

C'est donc dans ce contexte que nous nous sommes intéressées au processus de gestion du risque douanier. Nous l'avons analysé et y avons identifié certaines pistes d'amélioration. Nous avons tâché de proposer une solution qui améliore ce processus et optimise, dans la mesure du possible, les procédures de dédouanement.

Pour cela, nous avons organisé le présent travail en quatre (04) chapitres :

- Dans le premier chapitre nous avons entrepris de présenter les parties prenantes concernées par ce projet. Nous avons, par la suite, entamé un diagnostic en décrivant de manière concise les procédures de dédouanement, ainsi que le processus de gestion du risque qui oriente le contrôle douanier. Puis, nous avons analysé ce processus en indiquant notamment les résultats obtenus de son application sur l'orientation des marchandises vers les différents circuits d'inspection. Cette analyse nous a permis de choisir la problématique qui sera traitée dans ce présent travail.

- Le chapitre 2 est consacré à l'état de l'art et porte sur les outils que nous avons utilisés pour répondre à la problématique que nous avons choisi de traiter. Des notions sur l'intelligence artificielle seront introduites ainsi que diverses techniques issues de l'apprentissage automatique et du Data Mining.

- Le chapitre 3 est dédié au développement de la solution qui répondra à la problématique retenue. Nous avons élaboré un outil d'aide à la décision basé sur les techniques de Machine Learning supervisé afin d'automatiser le processus d'orientation des marchandises vers les différents circuits de dédouanement et en y supprimant la subjectivité humaine.

- Le chapitre 4 est, quant à lui, consacré à pallier aux limitations inhérentes à l'outil développé dans le chapitre qui le précède. Nous y avons mis en œuvre des méthodes issus du Machine Learning non supervisé pour détecter des fraudes inédites en se basant sur l'hypothèse que les fraudeurs auraient tendance à avoir des comportements et des méthodes similaires qui les différencient des non fraudeurs. Cette hypothèse, si elle est toutefois confirmée ultérieurement, servira à mieux identifier les fraudeurs et à fiabiliser la base de données.

Enfin nous conclurons par un résumé du travail que nous avons effectué ainsi que les perspectives à explorer et les recommandations à suivre.

Chapitre 1 : Etat des lieux et diagnostic

Chapitre 1: Etat des lieux et diagnostic

Introduction :

Nous consacrons ce chapitre dans sa première partie, à la présentation des entreprises et organisations concernées par ce projet. Puis nous passerons à la présentation du processus de dédouanement ainsi que les démarches déployées par l'administration des douanes pour le rendre plus efficace et moins contraignant. Nous aborderons ensuite le concept de la fraude en douane ainsi que ses causes et ses types. Enfin, nous pointerons les faiblesses du processus de dédouanement et définirons la problématique qui fera l'objet de notre travail.

I Présentation des entreprises et des parties concernées

I.1 Le groupe de la Banque (Larders et Boeck, 2005)

Le Groupe de la Banque mondiale (World Bank Group) est un regroupement de cinq organisations internationales réalisant des prêts à effet de levier pour les pays en développement. Le groupe fondé le 4 juillet 1944, est basé à Washington D.C. Il a fourni environ 61 milliards de dollars en prêts pour les pays en développement pour l'année 2014 et il est lié à l'Organisation des Nations Unies (ONU). Il travaille dans plus de 100 économies en développement, y apportant idées et programmes de financement destinés à améliorer le niveau de vie et à éradiquer la pauvreté.

Ces institutions qui forment le Groupe de la Banque mondiale sont spécialisées, chacune a sa manière, dans différents aspects du développement mais œuvrent toutes dans le même sens : lutter contre la pauvreté. Les termes «Banque mondiale » et« la Banque» ne désignent que la BIRD et l'IDA, tandis que les expressions « Groupe de la Banque mondiale » et « Groupe de la Banque » englobent les cinq institutions suivantes :

- La Banque internationale pour la reconstruction et le développement (BIRD) ;
- L'Association internationale de développement (AID) ;
- L'Agence multilatérale de garantie des investissements (MIGA) ;
- Le Centre international pour le règlement des différends relatifs aux investissements (CIRDI).
- La Société financière internationale (SFI).

Les missions de la Banque mondiale selon (Larders et Boeck, 2005)

- Lutter contre la pauvreté avec passion et professionnalisme pour obtenir des résultats durables ;
- Aider les populations à se prendre en charge et à maitriser leur environnement par la fourniture de ressources, la transmission de connaissances, le renforcement des capacités et la mise en place de partenariats dans les secteurs public et privé ;
- Exceller en tant qu'institution capable d'attirer, de motiver et de développer un personnel dévoué, aux compétences exceptionnelles, qui soit à l'écoute et capable d'apprendre.

I.2 La direction générale des douanes Algériennes

La douane est une institution fiscale chargée de la perception des droits et taxes dus à l'entrée de marchandises sur un territoire. Son activité est réglementée par le droit national, mais aussi par des accords internationaux.

L'administration des douanes est sous la tutelle du ministère des finances. Elle est dirigée par la Direction Générale des Douanes (DGD) qui est chargée de la mise en œuvre des mesures légales et règlementaires permettant d'assurer l'application de la loi douanière et de la loi tarifaire et celles mises à sa charge en vertu des textes législatifs et règlementaires en vigueur. (Décret exéctif n° 17-90, 2017)

I.2.1 Organisation de la direction générale des douanes

L'administration centrale de la direction générale des douanes est composée de :

- Une inspection générale ;
- 02 directeurs d'études ;
- 06 Chefs d'études ;
- 10 directions centrales.

(Décret exéctif n° 17-90, 2017)

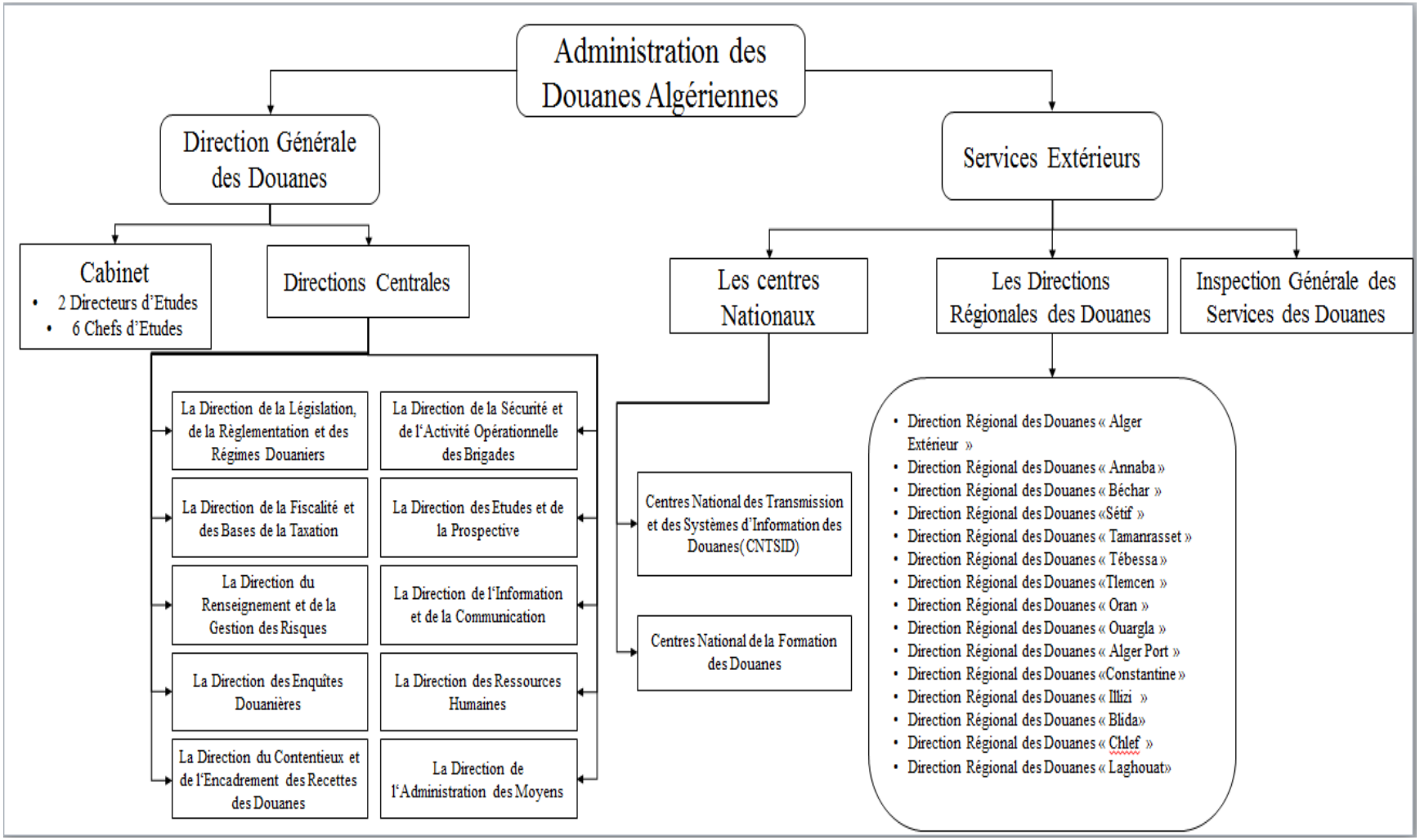

Figure 1-1 : Organigramme de l'administration des douanes Algériennes

Dans le cadre de notre mission au sein de la Direction Générale des Douanes, nous avons eu à travailler exclusivement avec la direction du renseignement et de la gestion des risques.

La direction du renseignement et de la gestion des risques

Cette direction est chargée : (Décret exéctif n° 17-90, 2017)

- De participer à l'élaboration des textes législatifs et d'initier les textes réglementaires en matière du renseignement et de gestion des risques et d'en soumettre les projets à la direction de la législation, de la réglementation et des régimes douaniers pour garantir leur cohérence ;
- De veiller à la recherche, à la collecte et à l'exploitation du renseignement douanier et de l'information en matière de fraude (commerciale, de contrefaçon, de contrebande, de trafic illicite de stupéfiants…) et de veiller à la constitution de bases de données en la matière ;
- De concevoir et d'actualiser le système de gestion et d'analyse des risques ;
- De mettre en œuvre les conventions d'assistance mutuelle en vue de la recherche et de la répression de la fraude douanière et commerciale et d'en assurer le suivi ;
- De veiller à la collaboration avec les services et les institutions de l'état chargés de la recherche et de la répression des fraudes et ceux chargés de la lutte contre la contrebande et le trafic illicite des stupéfiants ;
- D'élaborer des normes d'élaboration des procédures, et de les soumettre, à la direction de la législation pour en garantir la cohérence.

I.2.2 Les missions de l'administration des Douanes

Traditionnellement chargée de la perception des droits et taxes douanières, du recouvrement des impositions fiscales et parafiscales, de la lutte contre les trafics illicites et du contrôle des marchandises et des personnes aux frontières, la Douane s'est vue confier de nouvelles missions à forts enjeux économiques et sécuritaires.

a) Les missions économiques

- Appliquer en collaboration avec les institutions concernées, la législation et la réglementation régissant la circulation transfrontalière des marchandises;
- Rechercher et réprimander les pratiques déloyales et frauduleuses;
- Encourager les investissements, nationaux et étrangers, à travers les facilitations douanières et les régimes douaniers économiques institués à cet effet;
- Contrôler l'authenticité de l'origine des marchandises lorsque des conventions prévoyant l'octroi de préférences commerciales et tarifaires sont conclues avec un pays;
- Appliquer les mesures de prohibitions édictées tant à l'importation qu'à l'exportation.

b) Les missions fiscales

- Recouvrer les droits et taxes auxquels sont soumises les marchandises à leur importation ;
- Recouvrer les redevances douanières spécifiques (redevance pour prestation de services et redevance d'utilisation du système d'information et de gestion automatisée des douanes SIGAD);
- Recouvrer les pénalités (amendes et confiscations) dues à la violation des lois et règlements que l'administration est chargée d'appliquer ;

- Assurer l'application de la loi douanière régissant la circulation des marchandises à l'entrée ou à la sortie du territoire douanier et réprimer tous les actes des personnes morales ou physiques qui enfreignent cette loi ;
- Lutter contre la fraude douanière par la justification de l'origine des marchandises, leur espèce et leur valeur en douane, pour le contrôle de l'assiette des droits et taxes;
- Appliquer les mesures de rétorsion édictées à l'encontre des pays qui pourraient soumettre les produits nationaux à des mesures discriminatoires et moins favorables que celles appliquées à d'autres pays (surtaxes) ;
- Suivre et contrôler les avantages fiscaux :

  - ✓ Institués par les lois de finances et les lois spécifiques (secteur pétrolier, secteur minier, ANDI, ANSEJ…) afin d'éviter le détournement des biens importés de leur destination privilégiée
  - ✓ Prévus par les accords tarifaires préférentiels pour s'assurer des conditions de leur bénéfice légal.

c) **Les missions de protection**

- Participer à la préservation de l'ordre et de la sécurité publics (armes, explosifs, substances chimiques et produits dangereux);
- Participer à la protection du consommateur en veillant à ce que les produits de consommation non alimentaires et les produits domestiques soient soumis au contrôle de conformité aux normes de fabrication et de sécurité ;
- Lutter contre le trafic illicite des stupéfiants, la contrebande, le blanchiment d'argent et de manière générale le crime organisé transfrontalier ; trafic de cigarettes, trafic de drogues etc… .

II Diagnostic

Avant-propos

Nous allons dans ce qui suit décrire succinctement la procédure de dédouanement puis nous déroulerons le processus de gestion du risque que la douane a adopté après la signature de la convention de Kyoto. Nous citerons quelques outils que la douane a mis en place pour faciliter les opérations de dédouanement puis nous relèverons les points pouvant être améliorés.

II.1 Les procédures de dédouanement

Toutes les marchandises importées ou destinées à l'exportation doivent être soumises à des dispositions législatives et réglementaires mises en place en vue d'assurer une correcte perception des droits et taxes. Elles doivent pour cela, passer par les procédures de dédouanement après être passées par certaines formalités au préalable. (La loi n° 17-04, 2017)

II.1.1 Les formalités préalables au dédouanement

Il s'agit de l'ensemble des procédures qui précèdent la déclaration en détail des marchandises. Ce sont, des formalités qui doivent être respectées depuis l'introduction des marchandises sur le territoire douanier jusqu'à leur placement sous un régime douanier et elles se présentent comme suit :

a) La conduite en douane des marchandises

Toute marchandise importée ou exportée doit être conduite auprès d'un bureau de douane pour y être soumise au contrôle douanier. Elle doit donc être acheminée au bureau de douane le plus proche de la frontière du territoire douanier par son transporteur dans le cas de d'une importation ou par le déclarant dans le cas d'une exportation.

b) La mise en douane des marchandises

Il s'agit de la prise en charge des marchandises par l'administration des douanes lui permettant ainsi d'identifier et de garder sous sa surveillance les marchandises jusqu'au dédouanement ou l'enlèvement.

II.1.2 Les formalités du dédouanement

Après avoir été soumises aux formalités préliminaires au dédouanement, les marchandises doivent à présent passer par les formalités de dédouanement permettant de les placer sous un régime douanier autorisé et de garantir l'application de la législation douanière. Ses formalités se présentent comme suit :

a) La déclaration en détail de la marchandise

La déclaration en détail est l'acte juridique par lequel le déclarant marque sa volonté de placer les marchandises sous un régime douanier et s'engage à accomplir les obligations qui en découlent. Il s'agit aussi d'indiquer toutes les énonciations nécessaires à l'identification de la marchandise et à l'application des droits et taxes. Cette déclaration doit être écrite, signée et déposée par le déclarant qui peut être le propriétaire de la marchandise, son transporteur ou un commissionnaire mandaté.

b) Contrôle de la recevabilité

Avant d'être enregistrée, la déclaration en détail doit être contrôlée au niveau de la forme et des documents justificatifs du régime douanier assigné aux marchandises. Le contrôle préalable de la déclaration constitue ce qu'on appelle l'opération de recevabilité. Il consiste à vérifier si toutes les énonciations sont mentionnées et que tous les documents nécessaires sont annexés à la déclaration. Si la déclaration est jugée recevable, elle est immédiatement enregistrée par les agents des douanes dans le système d'information et de gestion automatisée de la douane (SIGAD) qui se chargera de l'affecter au circuit de vérification qui lui convient (rouge, orange ou vert).

c) Contrôle consécutif à l'enregistrement de la déclaration

Les déclarations enregistrées sont réparties entre les inspecteurs vérificateurs pour procéder à un contrôle se déclinant en deux phases : d'une part, un contrôle de fond des éléments constitutifs de la déclaration et des documents qui lui sont annexés, d'autre part, un éventuel contrôle physique de la marchandise pour vérifier entre autre l'adéquation avec ce qui a été déclaré. Ces contrôles peuvent se faire au moment du dédouanement ou après, avec ce que l'on appelle un contrôle à postériori exécuté de manière générale dans les locaux de l'opérateur économique.

Le contrôle au moment du dédouanement se divise en deux types :

- **Le contrôle documentaire :**

Ce contrôle permet de vérifier la concordance des informations fournies par la déclaration avec celles figurant dans les documents y annexés. Les principaux éléments de la déclaration qui sont inspectés par les agents de vérification sont les suivants :

- ✓ **La valeur en douane :**

La valeur en douane est déterminée par la valeur transactionnelle qui est définie comme étant le prix effectivement payé ou à payer pour les marchandises importées. Le prix étant celui mentionné sur la facture avec des ajustements éventuels. Cette valeur est le montant de base à partir duquel les droits et taxes imposables sont calculés. Elle constitue un élément important lors de la taxation des marchandises.

- ✓ **L'espèce tarifaire :**

L'espèce tarifaire est la dénomination attribuée à chaque marchandise par le tarif douanier en fonction de ses caractéristiques et en fonction de la nomenclature tarifaire. Le tarif douanier étant un constitué d'un tableau contenant les diverses marchandises avec l'indication des droits applicables à chacune d'elles. Les produits y sont classés en sections, chaque section est divisée en chapitres et chaque chapitre comprend des positions tarifaires qui sont-elles mêmes divisées en sous positions (**voir *Annexe A*)**.
L'identification de l'espèce tarifaire détermine le montant des droits et taxes à acquitter ainsi que la nature des formalités douanières à accomplir lors du dédouanement.

- ✓ **L'origine en douane :**

L'origine est définie comme la «nationalité» de la marchandise. En d'autres termes, elle désigne le pays ou la fabrication du produit concerné a été effectuée. Elle permet de déterminer les taux des droits de douane à appliquer sur la marchandise, selon la nature de son origine, à savoir préférentielle ou non.

On dit qu'une marchandise est d'origine préférentielle lorsqu'elle est originaire d'un pays avec lequel l'Algérie a conclu des accords d'associations, leurs conférant des avantages allant d'un abattement jusqu'à une franchise des droits de douanes et des taxes d'effet équivalent.

- **Le contrôle physique :**

Il s'agit de l'ensemble des opérations matérielles effectuées, afin de s'assurer de la conformité des marchandises déclarées par rapport aux énonciations de la déclaration en détail et aux documents commerciaux y annexés.

La visite des marchandises permet de déceler d'éventuelles manœuvres frauduleuses dans :

- L'origine des marchandises ;
- L'espèce tarifaire des marchandises ;
- La valeur (majoration ou minoration) des marchandises ;
- La quantité et le poids réel des marchandises non déclarées ou prohibées à l'importation.
- Déclaration ou non du fret (frais de transport).

La vérification peut porter sur la totalité des marchandises (vérification intégrale) ou sur une partie d'entre elles (vérification partielle ou par épreuve).

Le mode opératoire de la vérification doit être choisi en fonction des résultats du contrôle documentaire et des facteurs de risques de fraude liés à la nature de la marchandise déclarée (système de gestion des risques).

En effet, si les marchandises sont affectées au circuit rouge, elles doivent être soumises à une visite intégrale tandis qu'à l'orange elles ne subiront qu'une visite partielle. En revanche, le circuit vert n'est pas concerné par le contrôle physique.

Une fois la vérification terminée, l'agent ayant effectué la visite établit un certificat de visite au verso de la déclaration. Ce certificat contient une description précise de l'ensemble des opérations et constatations effectuées lors de la vérification. Dans le cas où les résultats de la vérification sont conformes aux énonciations de la déclaration en détail, l'agent vérificateur le mentionne dans le certificat de visite. Les droits et taxes peuvent alors être acquittés et les marchandises enlevées.

Dans le cas contraire, le service des douanes constate des différences entre les marchandises présentées et les informations fournies dans la déclaration en détail, et cela donne lieu à la naissance d'un litige.

Ainsi, on invite le déclarant à approuver les constats des agents vérificateurs sur la marchandise et à accepter les éventuelles suites contentieuses. Si le déclarant accepte, il devra le faire par écrit et le litige prendra fin par un arrangement transactionnel. En d'autres termes, il aura une amande à payer. S'il refuse, deux éventualités sont envisageables :

- Si la différence constatée entre la déclaration et la marchandise se situe au niveau des éléments matériels facilement vérifiables tels que le poids ou le volume, un procès-verbal de saisie est rédigé et l'affaire sera portée en justice.
- Si la différence constatée entre la déclaration et la marchandise se situe au niveau de l'espèce, l'origine ou la valeur des marchandises, le litige est soumis à la commission nationale de recours.

d) Liquidation et acquittement des droits et taxes :

Les droits et taxes sont liquidés - autrement dits calculés- en fonction de l'espèce tarifaire, de l'origine et de la valeur des marchandises. Le déclarant s'acquitte des droits et taxes auprès de la caisse au niveau de la recette principale qui lui remet une quittance.

e) Enlèvement des marchandises :

Le déclarant présente la quittance de paiement et la déclaration afin que l'Inspection Principale au Contrôle des Opérations Commerciales lui délivre un bon à enlever, qui constitue l'autorisation d'enlèvement des marchandises. Ensuite, muni de la déclaration, de la quittance et du bon à enlever, il se présente à la brigade commerciale qui lui délivre un bon de sortie et le déclarant peut alors disposer librement de sa marchandise.

II.2 Les facilitations douanières

L'ensemble de règles et procédures formelles destinées à encadrer l'acheminement de la marchandise entre le moment du passage physique de la frontière et celui où elle sera libérée de toute sujétion douanière présente un caractère contraignant. En effet, dans la partie précédente, nous n'avons fait que résumer et simplifier les procédures douanières, qui sont en réalité beaucoup plus complexes.

Néanmoins, du fait de l'évolution des échanges via l'ouverture du commerce extérieur, la diversification des échanges et la révolution des technologies de l'information, ces règles ont été adaptées, de nouvelles procédures ont été mises en place tandis que d'autres sont encore des exigences récurrentes des entreprises.

Ces facilitations visent à réduire la complexité et le coût des procédures commerciales et à améliorer l'environnement du commerce en rationalisant systématiquement les procédures et en normalisant les documents et les informations.

Afin de réaliser cela, l'Algérie qui est membre de l'Organisation Mondiale de la Douane depuis 1966 a adopté la plupart des instruments élaborés sous son égide et a notamment signé la convention de Kyoto.

La convention de Kyoto ()

La Convention internationale pour la simplification et l'harmonisation des régimes douaniers (Convention de Kyoto révisée) a été adoptée en 1974, puis révisée en 1999 ; et entrée en vigueur en 2006. Elle promeut la facilitation des échanges et l'efficacité des contrôles grâce à ses dispositions juridiques qui détaillent l'application de procédures simples mais efficaces. Elle prône entre autre une utilisation maximale des technologies de l'information et le recours à la gestion du risque pour orienter le contrôle douanier.

La ratification de la convention de Kyoto revisitée a amené l'administration des douanes à promouvoir l'utilisation de la gestion des risques pour le ciblage des marchandises à vérifier. Car le contrôle physique immédiat des marchandises a toujours été le moyen le plus efficace d'assurer le respect de la réglementation douanière.

Néanmoins, les opérations d'importation et d'exportation sont en constante augmentation et les moyens physiques et humains limités des douanes ne leur permettent pas de contrôler immédiatement toutes ces opérations. Le recours à la gestion des risques est donc primordial pour concilier les objectifs antagonistes de l'administration des douanes à savoir : le contrôle et les facilitations.

En effet, les douanes sont responsables de la lutte contre la fraude en la détectant et en la réprimant par le biais de contrôle douanier aussi divers qu'efficaces, car la fraude en plus d'encourager la concurrence déloyale, a une incidence sur l'économie ainsi que le trésor public. D'autre part les douanes sont censées encourager le trafic légal en simplifiant les procédures et en allégeant les contrôles. Ces deux tâches sont, à première vue, contradictoires, néanmoins les facilitations peuvent améliorer l'efficacité des procédures de contrôle en suivant le principe qui stipule que contrôler moins c'est contrôler mieux et cela passe par l'adoption d'une gestion des risques moderne qui prône le recours aux technologies de l'information dans la détection de la fraude.

II.2.1 La gestion des risques en douane

- **Définition de la gestion des risques :**

Il s'agit de l'ensemble des politiques, des stratégies, des dispositifs de maîtrise, de contrôle et de suivi, et des moyens humains, financiers et matériels mis en œuvre par une entité organisationnelle, visant à identifier, détecter, limiter et maîtriser les risques liés, directement ou indirectement, à ses activités.

- **Définition du risque douanier :**

La notion de risque en douane fait référence à la probabilité que se produise un fait conduisant à enfreindre la loi que la douane est chargée d'appliquer.

- **Définition de La gestion des risques en douane** :

L'OMD définit la gestion des risques comme l'application systématique des pratiques et procédures en matière de gestion permettant à la douane de recueillir les renseignements nécessaires au traitement des mouvements ou des envois de marchandises qui présentent un risque.

II.2.1.1 Le processus de gestion des risques en douane (Organisation Mondiale des Douanes, 2013)

La figure ci-après décrit le processus standard de la gestion du risque douanier selon l'Organisation Mondiale des Douanes.

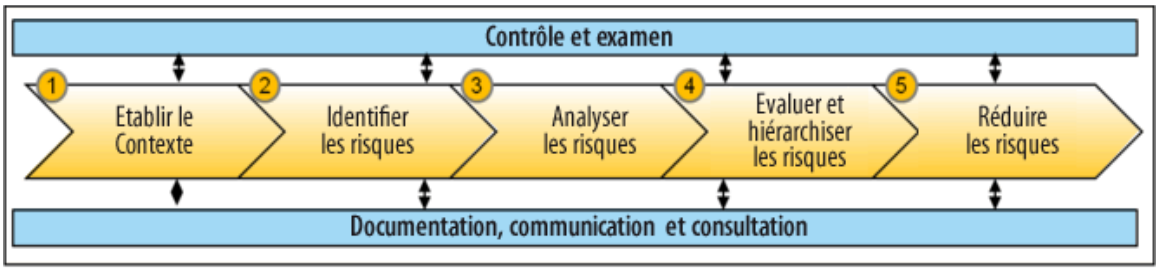

Figure 1-2 : Le processus de la gestion du risque douanier

Etape 1 - Etablir le contexte :

Cette étape consiste à définir le cadre stratégique et organisationnel à travers lequel le processus de gestion du risque se déroulera. Pour y parvenir, il suffit de se poser les questions suivantes :

- Quelles sont les capacités et les moyens disponibles pour gérer les risques ?
- Quels sont les critères utilisés pour évaluer les risques ?

- Quelles sont la portée et les limites de la gestion des risques ?
- Quelles sont les attentes des parties prenantes comme le gouvernement, les opérateurs et autres groupes concernés ?

Etape 2 - Identifier les risques :

Il s'agit de réaliser une liste exhaustive de tous les risques de fraudes, et d'en déterminer les causes pour ensuite établir les profils de risques. Pour cela, les questions suivantes doivent être étudiées :

- Quelles sont les sources du risque ?
- Quels risques pourraient se présenter, pourquoi et comment ? de quelle manière ?
- Quels contrôles pourraient déceler ou prévenir ces risques ?
- Quels sont les contrôles et les mécanismes mis en place pour détecter ces risques ?

Le résultat de cette étape est un inventaire qui documente les risques et garantit que l'ensemble des risques sont pris en considération.

Etape 3- Analyser les risques :

L'analyse des risques consiste essentiellement à quantifier les risques en considérant la probabilité qu'un événement survienne, et les conséquences potentielles d'une telle survenance. La combinaison de ces éléments donne une estimation du degré de risque.

Pour estimer cette probabilité, un système automatisé d'analyse et de gestion des risques a été introduit au niveau des douanes. Il est chargé d'orienter les marchandises vers les différents circuits de vérification selon un mode opératoire que nous expliciterons ultérieurement.
Initialement deux architectures du système ont été envisagées :

- Utilisation d'un logiciel d'analyse économétrique externe au système d'information et de gestion automatisée de la douane (SIGAD),
- Intégration dans le SIGAD d'un sous module d'analyse économétrique.

La deuxième option a été retenue pour les raisons suivantes :

- Disponibilité de logiciels d'analyse économétrique puissants,
- Nécessité d'une étude conceptuelle approfondie pour l'élaboration d'un sous module d'analyse économétrique,
- Difficultés techniques d'intégration d'un tel module dans le SIGAD.

Avant de décrire plus en détail ce système, il convient d'introduire le système d'information et de gestion automatisée de la douane (SIGAD).

a) Définition du SIGAD

Le SIGAD est un système informatique de gestion mis en place depuis Octobre 1995 et venant en remplacement de l'ancien système jugé peu performant car limité dans ses applications et dans son implantation géographique (port et aéroport d'Alger seulement). Il comporte un ordinateur central, où sont enregistrés tous les éléments de la réglementation douanière, et des postes installés dans les bureaux des douanes et les locaux de certains transitaires ayant signés une convention avec la douane. (Moussaoui, et al., 2017)

Le SIGAD est constitué de 4 sous-systèmes :

- Le système tarif intégré : qui rassemble l'ensemble de la réglementation, la fiscalité ainsi que la classification des produits ;
- Le système dédouanement des marchandises : qui prend en charge le dédouanement de la marchandise ;
- Le système contentieux : qui concerne la gestion et le suivi des litiges éventuels.
- Le système statistique : qui prend en charge l'ensemble des outils de manipulation et d'interprétation des informations recueillies sur le système de dédouanement.

b) Les avantages du SIGAD :

- Il permet d'accélérer le processus de dédouanement avec le traitement informatisé des déclarations ;
- Il permet de maitriser le flux d'informations sur les sorties et les entrées des marchandises ;
- Il permet d'alléger les manipulations des documents, d'éviter leurs transport et leurs stockage tout en améliorant la qualité du support d'informations.
- Il permet la réduction des relations directes entre les opérateurs et les fonctionnaires des douanes ce qui tend à supprimer leur subjectivité.

c) Le fonctionnement du SIGAD :

Le SIGAD assure le traitement de la déclaration en temps réel. Il contrôle sa recevabilité ainsi que les droits et taxes exigibles par référence au tarif enregistré dans le système central et informe le déclarant quant aux documents qui doivent être annexés à la déclaration. Ensuite, il affecte les marchandises à l'un des circuits de contrôles (rouge ou orange) ou en circuit d'admission pour conforme (circuit vert). Nous présenterons ultérieurement ces trois circuits.

d) Fonctionnement du système automatisé d'analyse et de gestion des risques :

Le schéma ci-après illustre le fonctionnement du système :

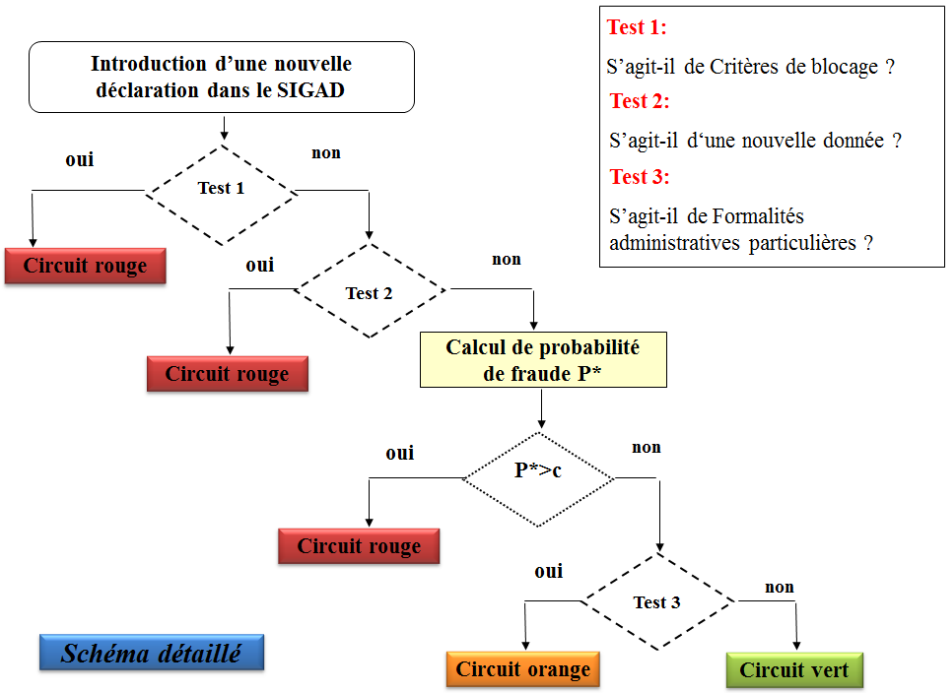

Figure 1-3 : Le fonctionnement du système d'orientation des déclarations vers les circuits d'inspection

Lorsqu'une nouvelle déclaration est introduite dans le système, il vérifie que les produits importés ne sont pas concernés par les critères de blocage. Ces derniers représentent des caractéristiques propres aux produits ou aux importateurs qui poussent les douaniers selon leur connaissance du métier à se méfier de ce genre de transactions. Les critères de blocage en vigueur sont les suivants :

- Les produits destinés à la revente en l'état.
- Les produits concernés par une fourchette de valeur et qui sont déclarés en dehors de cette fourchette.
- Toute déclaration engagée par un opérateur figurant dans la liste « fraudeurs ».

S'il s'avère que la déclaration est concernée par un critère de blocage, elle est orientée vers le circuit rouge. Sinon elle passe au second test qui consiste à vérifier si la déclaration présente au moins une nouveauté par référence à l'historique des transactions (nouvel opérateur, nouvelle marchandise,...), si c'est le cas, elle est orientée d'office vers le circuit rouge. Dans le cas contraire, le système compare la probabilité de fraude qui n'est rien d'autre que la projection des résultats de l'analyse économétrique sur la nouvelle déclaration et la compare à un seuil fixé au préalable. Si elle est supérieure au seuil, la déclaration est orientée au circuit rouge, sinon elle passe au dernier test.

A ce stade, la déclaration en question ne présente pas de risques, mais elle peut être concernée soit par une ou plusieurs formalités administratives particulières (Autorisations, certificats, documents spécifiques…) ou un avantage fiscal. Dans ce cas, les agents de vérifications doivent s'assurer de la présence des documents justifiants le droit aux franchises et cela induit le passage vers le circuit orange. Si la déclaration n'est pas concernée par cette formalité, elle sera dirigée vers le circuit vert.

Etape 4 - Evaluer et hiérarchiser les risques :

Cette étape consiste à comparer les risques dont on a estimé les degrés avec des critères d'importance préalablement déterminés, cela permettra d'évaluer et de prioriser les principaux risques à analyser de façon plus pointue et ainsi prévoir un déploiement adéquat des ressources pour se préparer aux risques, les prévenir et y faire face.
A la fin de cette étape les risques sont en général classés en trois catégories : faible, moyen, élevé selon la matrice suivante :

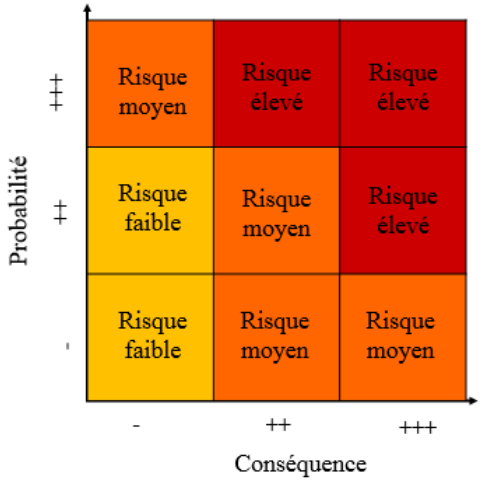

Figure 1-4 : Hiérarchisation des risques

Etape 5 - Réduire les risques :

Il s'agit de la phase opérationnelle de la gestion du risque où l'on décide de la manière dont on va traiter les risques qui ont été identifiés et évalués. Cette phase consiste donc à orienter la marchandise compte tenu du risque qu'elle présente à l'un des trois circuits suivants :

- **Circuit vert** : qui implique que la marchandise est admise pour conforme au moment du dédouanement et seul un contrôle documentaire est effectué à postériori. Ce type de circuit est envisagé pour les exercices ayant peu de chance d'être frauduleux.
- **Circuit orange** : qui implique une simple vérification documentaire pouvant être suivi d'un contrôle physique sur la marchandise si une anomalie apparait lors de ce contrôle. Ce circuit est octroyé aux opérations à risque moyen.
- **Circuit rouge** : dans ce circuit la vérification est aussi bien physique que documentaire. Il est octroyé aux opérations présentant un risque élève de fraude.

Contrôle et examen :

Pour garder son efficacité, tout système de gestion des risques doit être contrôlé et examiné sous tous ses aspects ; y compris son rendement, les changements qui pourraient l'affecter et l'émergence de nouveaux risques. Pour cela, la douane doit se tenir à jour en consultant les diverses sources de renseignement dont elle dispose.

Les questions traitées à ce stade sont énumérées ci-dessous :

- Les hypothèses de risque sont-elles toujours fondées ?
- Existe-t-il des risques inédits ou émergents ?
- Les opérations de traitement destinées à réduire les risques sont-elles performantes ?
- Les traitements sont-ils rentables ?
- Les opérations de traitement sont-elles conformes aux exigences légales et aux politiques gouvernementales et organisationnelles ?
- Comment le système peut-il être amélioré ?

Documentation, communication et consultation :

La communication entre les parties prenantes et leurs consultations sont primordiales à chaque étape du processus pour garantir son efficacité.
Il convient aussi de documenter les activités de gestion des risques à tous les stades du processus afin de constituer un historique qui pourra être consulté en cas de besoin. Cet historique devra contenir les hypothèses, les méthodes utilisées, les sources des données, les résultats obtenus, et les décisions prises ainsi que les raisonnements qui les étayent.

II.2.1.2 Utilité de la gestion du risque en douane

L'application de la gestion du risque en matière douanière permet de bénéficier des avantages suivants :

- ✓ Le ciblage sélectif permet de détecter rapidement les infractions ;
- ✓ L'accélération de la circulation des marchandises et des personnes ;
- ✓ La réduction du coût de l'intervention des douanes ;
- ✓ La rationalisation des contrôles en fonction du risque ;
- ✓ La révision permanente des procédures et introduction de mesures correctives ;
- ✓ La décongestion des aires de transit (Ports et aéroports).

Outre l'avènement de la gestion du risque dans le domaine douanier, l'administration des douanes s'est efforcée de mettre en place une panoplie de mesures tendant à encourager le commerce extérieur. Nous citerons par exemple le statut d'opérateur économique agréé qu'elle octroie à certains opérateurs.

II.2.2 L'Opérateur Economique Agréé OEA

Dans le cadre de son programme de modernisation, l'administration douanière s'est engagée dans une politique de partenariat avec les entreprises économiques et les autres acteurs de la chaîne du commerce extérieur afin de libéraliser le commerce international et promouvoir les investissements.

Le statut de l'Opérateur Economique Agréé est une démarche volontaire et partenariale avec la douane. Il permet à toute entreprise exerçant une activité liée au commerce international d'acquérir un label de qualité sur les processus douaniers qu'elle met en œuvre. Il permet de distinguer les entreprises les plus fiables, de faciliter les échanges et de mieux sécuriser les flux de marchandises entrant ou sortant du territoire national. (Décret exéctif n°12-93, 2012)

Ce statut est délivré par la douane sur demande de l'opérateur économique, à partir de l'examen d'un questionnaire détaillé d'auto-évaluation rempli par ce dernier et d'un audit approfondi de ses processus internes.

Pour ces opérateurs, les opérations de dédouanement sont ainsi rendues plus rapides, plus simples et plus sûres ce qui contribue à leur compétitivité. En effet, l'un des principaux avantages de ces opérateurs réside dans le fait que leurs déclarations sont affectées au circuit vert ; ce qui implique que leur marchandise ne subira aucun contrôle immédiat.

Cette différenciation entre opérateurs est également utile à la douane car elle lui permet de concentrer ses efforts de contrôle sur les opérateurs et flux les plus porteurs de fraude.

II.3 La fraude

Introduction

Le processus de dédouanement ainsi que les outils visant à le simplifier présentés, nous passerons dans ce qui suit à la définition de la fraude commerciale, ses causes ainsi que ses différents types.

II.3.1 Définition

En matière douanière, la fraude correspond à toute infraction aux dispositions législatives ou règlementaires que l'administration des douanes est chargée de faire appliquer.

Elle se manifeste sous plusieurs formes :

- La contrebande qui consiste à faire entrer à travers les frontières douanières des marchandises en les soustrayant au contrôle douanier.

- La fraude commerciale qui est la plus répandue et qui est commise à l'occasion des opérations commerciales, en présentant par exemple de faux documents ou en utilisant des manœuvres frauduleuses afin :
  - D'éluder le paiement des droits et taxes applicable aux marchandises.
  - De tenter de percevoir des remboursements ou des subventions.
  - De tenter d'obtenir des avantages commerciaux.

II.3.2 Les causes de la fraude commerciale

Les causes de la fraude commerciale sont diverses et peuvent avoir plusieurs facteurs. En effet, il se pourrait qu'elles soient liées aux actions entreprises par la douane pour la contrer ou au contraire être de nature totalement extra-douanière (MILIANI, 2001) :

i. Facteurs douaniers :

L'administration des douanes souffre de beaucoup d'insuffisances qui entravent le bon déroulement du contrôle douanier et incitent les opérateurs à tenter des actes frauduleux. Ces insuffisances sont causées par :

- Un personnel insuffisant et mal formé ;
- Un moyen matériel dépassé ;
- Une organisation rigide.

ii. Facteurs économiques :

L'Homme pouvant être cupide par essence, l'une des raisons principales qui le poussent à frauder est le gain d'argent. En effet s'il n'a pas d'avantage économique à la clé, il ne prendra pas un tel risque.
Ses avantages peuvent se présenter dans les cas suivants :

- La rareté des produits :

Si certains produits devenaient rares pour cause de prohibition par exemple, cela engendrerait une forte demande de la part du marché, ce qui inciterait les fraudeurs à satisfaire cette demande en usant de moyens illégaux tout en réalisant des gains considérables en jouant sur les prix de ces marchandises.

- L'intensification de la concurrence :

Pour pouvoir s'aligner à la concurrence, un opérateur devra jouer sur la qualité ou sur le prix de son produit. Et préférant les pratiques frauduleuses à une concurrence loyale, il procédera à de fausses déclarations pour diminuer les frais qu'il doit à la douane et appliquera des prix imbattables.

iii. Facteurs psychologiques :

L'administration des douanes fait objet d'autorité publique au niveau des frontières vis-à-vis des opérateurs du commerce international, et les « tromper » en commettant une fraude serait une satisfaction psychique en soit. De plus, nombre d'entre eux trouvent injustifiées les charges fiscales imposées à la marchandise, et commettre une fraude serait pour eux un moyen d'exprimer le rejet de cet impôt.

iv. Facteurs sociologiques :

Des observations tentent à prouver que la fraude diffère d'un pays à un autre voire d'une région à une autre. En effet, pour les gens vivant aux frontières, la contrebande est un métier qui se transmet de père en fils et est considéré dès lors comme un travail traditionnel.

De plus, les bénéfices tirés des opérations frauduleuses sont plus importants et moins pénibles à réaliser qu'en travaillant légalement, dans une société où justement l'offre d'emploi se fait rare. La fraude ne provoque donc pas une réprobation de l'opinion publique comme l'aurait fait le vol ou l'escroquerie.

II.3.3 Les types de fraude commerciale

Il existe trois types d'infractions majeurs dont découlent plusieurs types de fraudes (MILIANI, 2001) :

- Les fraudes portant sur les éléments principaux de la taxation (valeur, espèce tarifaire, origine).
- Les fraudes portant sur les autres éléments de la taxation (provenance, quantité...etc.)
- Les fraudes en matière de régime douanier économique.

a. Les fraudes portant sur les éléments principaux de la taxation :

La fraude sur la valeur, l'espèce tarifaire, et l'origine de la marchandise consiste pour l'opérateur à effectuer de fausses déclarations pour avoir certains bénéfices pécuniaires.

a.1 La fraude portant sur la valeur :

- **La sous-évaluation :**

La minoration de la valeur consiste à faire apparaitre sur la facture un prix de vente inférieur au prix réel que l'opérateur a payé pour sa marchandise afin de diminuer les droits et taxes qu'il aura à payer, et de ce fait s'imposer sur le marché en appliquant des prix de ventes qui défient la concurrence. Pour cela, le fraudeur usera de diverses méthodes allant de la minoration ou la suppression de certains éléments qui constituent la transaction (frais de transport, assurance…) jusqu'à la production de faux documents commerciaux.

- **La surévaluation :**

La majoration de la valeur permet au fraudeur de masquer la fuite de capitaux à destination de pays à faible cours d'impôt et à haut taux d'intérêt.

a.2 La fraude portant sur l'origine :

- **Fausses déclaration sur l'origine :**

Les fraudeurs peuvent falsifier le certificat d'origine de leur marchandise afin d'échapper aux mesures de restriction du commerce extérieur en introduisant des marchandises dont les importations sont restreintes ou prohibées lorsqu'elles proviennent de certains pays en particuliers. Comme ils pourraient bénéficier d'un taux réduit ou nul de droit en douane en déclarant que leur marchandise provient d'un pays avec lequel l'Algérie a signé des conventions.

Pour cela, les fraudeurs useront de certaines pratiques dont le transbordement, le ré-emballage, le ré-étiquetage des produits mais aussi l'élaboration de documents visant à dissimuler leur véritable origine.

a.3 La fraude portant sur l'espèce tarifaire :

- **Fausses déclarations sur l'espèce :**

Il s'agit de classer la marchandise dans une position tarifaire plus avantageuse que la position réelle de sa marchandise afin de bénéficier d'une réduction des taxes à payer ou d'éviter certaines mesures de restriction.

Pour cela, le fraudeur tâchera de choisir une position tarifaire voisine de celle de sa marchandise pour ne pas éveiller les soupçons tout en omettant de manière délibérée certains détails concernant sa marchandise lors de la rédaction de sa déclaration. Ou au contraire utiliser des mots techniques concernant sa marchandise que les agents de contrôle ne pourraient déchiffrer.

b. Fraude portant sur les autres éléments de la taxation :

Les autres éléments présents dans la déclaration peuvent aussi constituer matière à frauder pour les importateurs à travers :

b.1 La fraude portant sur la provenance :

Les importateurs font de fausses déclarations sur le pays de provenance de leur marchandise qui n'est autre que le dernier pays duquel elle est expédiée afin d'éviter la encore des mesures restrictives et profiter de régimes préférentiels.

Pour cela, ils useront de transbordement et de faux documents pour justifier l'itinéraire déclaré.

b.2 Fraude portant sur la quantité et la qualité de la marchandise :

Les importateurs présentent aux autorités douanières de fausses déclarations quant aux caractéristiques physiques de la marchandise qu'ils importent telles que la nature, le volume, la quantité en vue d'éviter des restrictions ou d'obtenir des avantages financiers.

Cela peut aussi impliquer l'introduction de produits dont les caractéristiques ne répondent pas aux normes et exigences techniques exercées sur les marchandises dans le but de protéger le consommateur et échapper à la concurrence déloyale.

c. Fraude portant sur les régimes douaniers économiques :

Le régime douanier désigne selon le glossaire des termes douaniers internationaux, le traitement applicable par la douane aux marchandises assujetties à son contrôle.
Les régimes douaniers économiques sont quant à eux un moyen utilisé par les douanes pour encourager les producteurs nationaux et renforcer leur capacité concurrentielle en leur offrant des avantages variés sous forme de suspension de droits et taxes, avantages fiscaux….

Les opérateurs peuvent profiter de cette faveur et s'adonner à des pratiques frauduleuses, en faisant de fausses déclarations ou en manquants aux engagements qu'ils ont dû prendre pour bénéficier de tels régimes. En effet, la complexité des règles régissant de tel régimes, et la multiplicité des privilèges dont jouissent ses opérateurs facilitent les manœuvres frauduleuses.

Dans le cas du régime économique *transit* par exemple, ou les opérateurs peuvent faire circuler leurs marchandises d'un bureau de douanes à un autre. Ils peuvent en profiter pour substituer leurs marchandises par une autre de moindre valeur et écouler leur marchandise au marché local sans payer de taxes au lieu de la faire parvenir à la destination prévue.

II.4 Résultat du diagnostic

La douane algérienne, dans un effort de modernisation et d'une tentative de conciliation entre le contrôle douanier et les facilitations au commerce extérieur a mis en place un dispositif de gestion du risque pour la guider dans ses missions.

Nous avons décrit dans ce qui précède les étapes constituant ce processus et avons développé plus particulièrement la partie concernant l'analyse du risque.

Nous avons ensuite décidé de mettre en œuvre la technique de data visualisation qui nous permet de représenter nos données et d'effectuer des premières analyses afin d'apprécier les résultats auxquels a abouti ce processus.

Nous avons commencé par représenter le graphique illustrant la distribution des déclarations selon les trois circuits de vérifications sur les trois années 2016, 2017 et 2018. Les résultats sont les suivants :

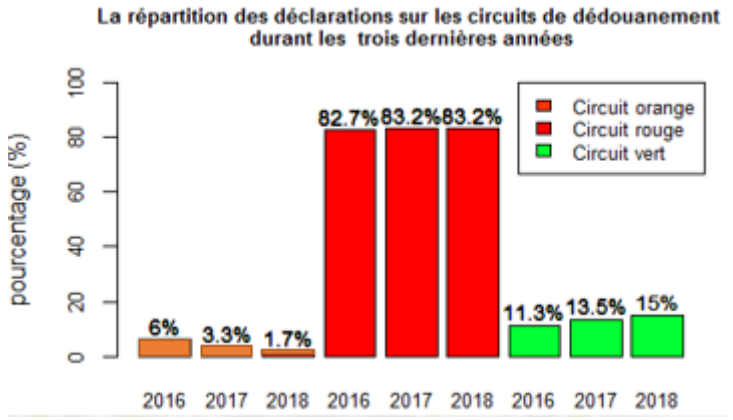

Figure 1-5 : La répartition des déclarations sur les circuits de dédouanement

A partir de ce graphique, nous avons pu tirer les conclusions suivantes :

- On remarque clairement que la majorité écrasante des déclarations, soit 83 % en moyenne, passent par le circuit rouge. La proportion est quasiment figée sur les trois années : il n'y a donc aucune amélioration.

- En moyenne, 3,7% des déclarations passent par le circuit orange. Ce dernier évolue néanmoins avec une proportion de déclarations qui baissent d'environ 50 % chaque année au profit du circuit vert.

- Le circuit vert quant à lui se voit affecter en moyenne 13,4 % des déclarations, cette proportion augmente chaque année. En effet, de 2016 à 2017 la proportion a augmenté de 20 % et seulement de 11% de 2017 à 2018. Néanmoins, ces augmentations ne sont pas gages de réussite de la démarche qui consiste à faciliter les opérations de dédouanement. En effet, la proportion des déclarations au niveau du circuit vert augmente seulement parce que celle du circuit orange diminue. Or, ce que la douane souhaite, c'est réduire la proportion des déclarations du circuit rouge au profit des autres circuits, et plus particulièrement du circuit vert.

Nous avons ensuite visualisé la proportion de fraudes présente dans chaque circuit en agrégeant les données des années 2016 et 2017 et 2018. Nous avons obtenu le graphique suivant :

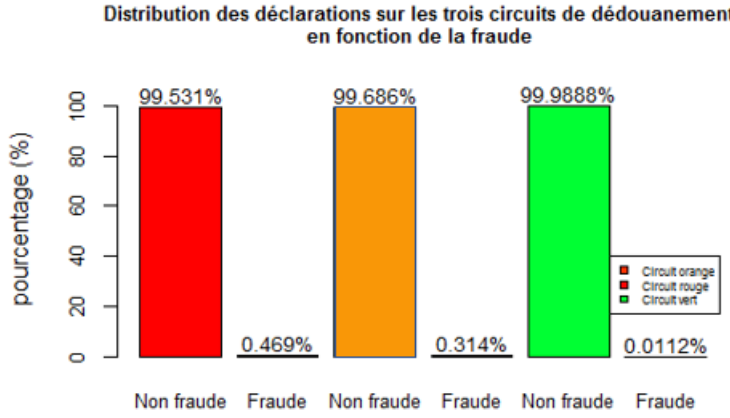

Figure 1-6 : La distribution des déclarations sur les trois circuits de dédouanement

A partir de la ***Figure 1-6,*** nous avons pu remarquer que :

- Seulement 0,47% des déclarations passants par le circuit rouge sont frauduleuses.
- Le circuit orange cumule sur les 3 années 97 fraudes sur 3532 soit un pourcentage de 0.31%.
- Le circuit vert ne comporte quasiment aucune fraude, en ne cumulant sur les 3 années seulement 13 fraudes parmi les 3532 transactions frauduleuses soit un pourcentage de 0.0112%.

Ces analyses nous ont permis de constater que les déclarations sont orientées à tort vers le circuit rouge, car seulement 0,47% d'entre elles sont réellement frauduleuses, ou du moins ont pu être détectées et labélisées comme telles ; car rappelons-le, les fraudes ne sont pas toute détectées et il se pourrait que des déclarations aient été considérées à tort comme non frauduleuses. La vérification systématique qui se fait au niveau de ce circuit engendre bon nombre de préjudices à travers :

- Le remboursement des entreprises ayant encouru des pertes financières à cause de la rétention de leurs marchandises au niveau des ports afin d'effectuer les vérifications nécessaires au circuit rouge, alors que les transactions opérées n'étaient pas frauduleuses.

- La congestion au niveau des ports, causées par l'important volume des déclarations passant par le circuit rouge, ne permet pas aux agents de vérification d'effectuer leur travail de vérification avec autant d'application que nécessaire, ce qui implique que les vérifications peuvent être bâclées et que des transactions frauduleuses passent par les mailles du filet.

- Le temps que doivent attendre les importateurs pour disposer de leur marchandise engendre des coûts qui s'ajoutent aux couts d'emmagasinage au niveau des ports. Ils augmentent donc les prix de leurs produits pour absorber ces coûts et partager cet inconvénient avec les consommateurs qui au final pâtissent de la situation au même titre que les importateurs.

- Les temps d'attente au niveau des ports découragent les investisseurs et entravent donc le développement du commerce extérieur et de l'économie en général.

- L'importance des temps d'attente pousse les opérateurs à verser des pots de vins pour disposer de leurs marchandises, ce qui a pour effet de répandre la corruption au niveau des ports.

III Enoncé de la problématique

Il est évident que l'affectation qui se fait des dossiers d'importation au niveau des circuits n'est pas pertinente car d'une part, le modèle économétrique utilisé dans la prévision de la fraude n'a pas été actualisé depuis qu'il a été mis en place au milieu des années 90 (on peut raisonnablement considérer que des changements structurels ont eu lieu depuis cette époque). D'autre part, les critères drastiques de blocage poussent les douaniers à orienter dès le départ une proportion importante des déclarations vers le circuit rouge.

Le travail que nous effectuerons dans le cadre de ce projet aura donc pour but de développer un outil d'aide à la décision qui orientera de manière optimale les marchandises aux différents circuits de vérifications tout en essayant d'atteindre les objectifs fixées par la DGD et qui se présentent comme suit :

- La modification de la distribution des déclarations au niveau des circuits pour atteindre une proportion de 50 % au niveau du circuit rouge et de 20% au niveau du vert.
- Amélioration de la détection des transactions frauduleuses.

Conclusion :

Ce chapitre nous a permis, dans un premier temps, d'introduire les parties prenantes de ce projet, puis d'exposer les étapes du processus de dédouanement en soulignant les contraintes qui l'entouraient.

Nous avons ensuite introduit le système de gestion du risque adopté par la douane algérienne à la suite de sa ratification de la convention de Kyoto dans une démarche de modernisation. Nous avons plus particulièrement développé la partie du processus qui porte sur l'analyse du risque, et avons conclu grâce à la visualisation de la distribution des déclarations sur les différents circuits, compte tenu de la fraude, que cette distribution n'était pas optimale.

Nous avons souligné les inconvénients qu'impliquaient une telle distribution et avons pu définir la problématique à laquelle nous devrons répondre dans ce présent travail.

Pour ce faire, et étant donnée le volume et la complexité de nos données, nous avons décidé de faire appel aux méthodes de l'intelligence artificielle. En effet, pour résoudre cette problématique, il est impératif d'intégrer le facteur de l'expertise métier. Et l'intelligence artificielle a justement pour but de reproduire le comportement de l'humain dans ses activités de raisonnement et à résoudre des problèmes à forte complexité logique et algorithmique.

Nous ferons donc, le chapitre qui suit une revue de littérature des différentes approches de l'intelligence artificielle.

Chapitre 2 : Etat de l’art

Chapitre 2: Etat de l'art

Introduction :

Ce chapitre est dédié à l'état de l'art et va nous permettre d'introduire les concepts liés au Data Mining et au Machine Learning. Nous présenterons et expliciterons plusieurs algorithmes d'apprentissage supervisés et non supervisés ainsi que leurs avantages et leurs inconvénients, puis nous développerons les méthodes utilisées pour les évaluer.

I Machine Learning (Brink, et al., 2017)

En 1959, un informaticien d'IBM, Arthur Samuel, a écrit un programme informatique pour jouer aux dames. Bien que les règles du jeu soient relativement simples, il existe des milliards de positions possibles .Il est donc impossible de programmer explicitement l'ordinateur en lui indiquant quoi faire dans chaque situation. Au début, les scores étaient basés sur une formule utilisant des facteurs tels que le nombre de pièces de chaque côté et le nombre de rois. Cela a fonctionné, mais Samuel avait une idée sur la façon d'améliorer sa performance. Il a fait jouer le programme des milliers de parties contre lui-même et a utilisé les résultats (victoire ou défaite) pour évaluer la probabilité de gagner en jouant un coup donné.

En 1962, le programme battait le champion du Connecticut. C'était une nouvelle d'une importance majeure : pour la première fois, une simple machine électronique, était capable de challenger la supériorité intellectuelle de l'être humain.

Samuel avait alors écrit un programme informatique capable d'améliorer ses propres performances grâce à l'expérience et c'est ainsi que le Machine Learning naquit.

I.1 L'intelligence artificielle (Cornuéjols et Miclet, 2003)

Né dans les années 1950 avec les travaux d'Alan Turing, puis de John McCarthy et Marvin Lee Minsky, le concept d'intelligence artificielle se trouve à l'interface de la logique mathématique, de la science des réseaux neuronaux et de l'informatique. Ses divers modes opératoires visent tous à l'élaboration d'algorithmes permettant de résoudre des problèmes souvent très complexes.

Marvin Lee Minsky l'a défini comme étant: « la construction de programmes informatiques qui s'adonnent à des tâches qui sont pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique ».

Le terme « artificiel » faisant référence à l'usage des ordinateurs ou autres processus électroniques et le terme « intelligence » à la capacité d'imiter le comportement humain que ce soit dans le raisonnement ou dans la compréhension.

I.2 Définition du Machine Learning

Définition 1 : (Cornuéjols et Miclet, 2003)

Dans son livre sur l'apprentissage automatique, Cornuéjols le définit comme étant « La notion qui englobe toute méthode permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle partiel ou moins général, soit en créant complétement le modèle ».

Définition 2 : (Gacôgne, 2015)

L'apprentissage automatique est un sous-domaine de l'intelligence artificielle qui fait référence au développement de méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de résoudre les problèmes pour lesquels les algorithmes classiques n'offrent pas de solution.

En d'autres termes, lorsque nous ne connaissons pas de solution exacte, nous ne pouvons pas écrire de programme informatique. L'apprentissage automatique consiste alors à programmer des algorithmes permettant d'apprendre automatiquement de données et d'expériences passées pour proposer des solutions à ces problèmes.

I.3 Les données d'apprentissage

Les données d'apprentissage sont, souvent, réparties en 3 catégories :

- L'ensemble d'apprentissage ou population d'entrainement qui constitue l'ensemble des candidats ou exemples utilisés pour générer le modèle d'apprentissage.
- L'ensemble de validation qui est un sous ensemble de l'ensemble d'entrainement utilisé lors de la phase d'apprentissage pour corriger l'algorithme et éviter le sur-ajustement.
- L'ensemble de test qui est constitué de candidats sur lesquels sera appliqué le modèle d'apprentissage pour tester et corriger l'algorithme.

I.4 Les types d'apprentissage automatiques

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient :

a) L'apprentissage supervisé (Géron, 2017)

Les techniques d'apprentissage supervisé permettent de construire des modèles à partir d'exemples d'apprentissage ou d'entrainement dont on connait la réponse. En d'autres termes, les données d'entrées de l'algorithme sont étiquetées. Ce dernier s'en servira pour pouvoir étiqueter d'autres données non labélisées.

Formulation :

Une expression simplifiée du problème d'apprentissage peut être énoncée de la manière suivante (Cornuéjols, et al., 2003):

Soit un ensemble d'apprentissage $S = \{x_i, y_i\}_{1,n}$ dont les éléments obéissent à la loi jointe : $P(x, y) = P(\ )(y|x)$.

On cherche à approcher une loi sous-jacente $f(x)$ telle que $y_i = f(x_i)$ par une hypothèse $h_a(x)$ aussi proche que possible, où les $\boldsymbol{a}$ sont les paramètres du système d'apprentissage, tel que :

- Si $\boldsymbol{f(x)}$ est discrète, c'est-à-dire, lorsque l'ensemble des valeurs de sortie est fini, on parle d'un problème de classification. Le modèle prévisionnel construit dans ce cas est appelé classifieur.

Exemple : la classification d'individus en deux catégories : atteint d'une maladie, ou non.

- Si $\boldsymbol{f(x)}$ est une fonction continue, c'est-à-dire lorsque la sortie qu'on cherche à estimer est une valeur dans un ensemble continu de réel, on parle alors de régression.

Exemple : L'estimation de la taille d'un individu.

b) L'apprentissage non supervisé (Géron, 2017)

On parle d'apprentissage non supervisé lorsque l'on dispose d'un ensemble de données sans aucune valeur cible associée (étiquette). Il s'agira donc de concevoir un modèle non seulement capable de découvrir par lui-même la structure des données et d'extraire les régularités qui y sont présentes, mais aussi d'apprendre des caractéristiques, qui permettent à la machine intelligente de découvrir automatiquement les représentations nécessaires pour classer les nouvelles données brutes.

Exemple : Essayer de déterminer des segments d'individus ayant le même comportement d'achat selon certains paramètres (salaire, origine…etc.)

c) L'apprentissage semi supervisé (Géron, 2017)

Dans ce type d'apprentissage, les données d'entrée sont constituées d'exemples étiquetés et non étiquetés, il se situe donc entre l'apprentissage supervisé et non supervisé. Ce qui peut être très utile quand on a deux types de données (étiquetées et non étiquetées), car cela permet de ne pas en laisser de côté et d'utiliser toute l'information.

d) L'apprentissage par renforcement (Géron, 2017)

L'apprentissage par renforcement correspond au cas où l'algorithme apprend un comportement étant donné une observation afin de trouver la solution optimale. Il va donc interagir avec « l'environnement », et essayer plusieurs solutions (on parlera « d'exploration »), puis observer la réaction de l'environnement et adapter son comportement (les variables) pour trouver la meilleure stratégie (on parlera « d'exploitation »).

Exemple : Un robot apprend à marcher en ayant pour stimulant la distance parcourue en avant.

I.5 Les algorithmes d'apprentissage supervisé

Dans cette partie, nous présentons une liste des méthodes d'apprentissage supervisé appartenant à diverses familles d'algorithmes (***Figure 2-1)*** afin de donner un aperçu sur les différents principes appliqués.

Nous décrirons le fonctionnement de ces algorithmes ainsi que leurs avantages et leurs inconvénients.

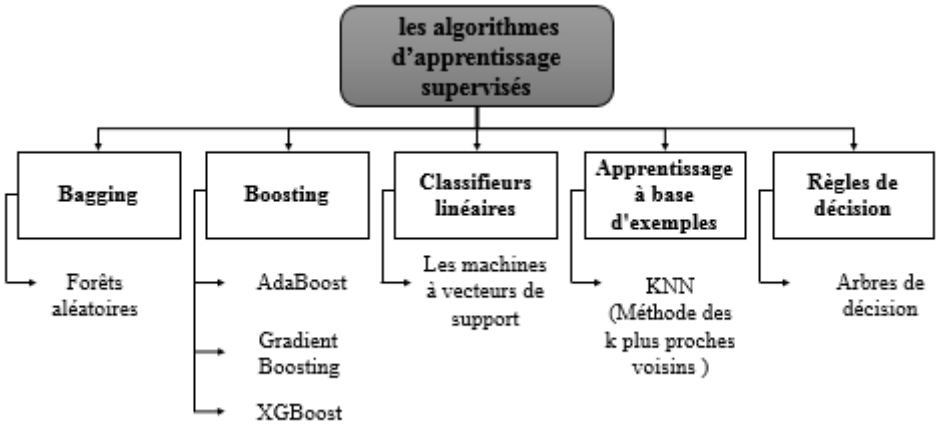

Figure 2-1: Les algorithmes d'apprentissage supervisés

I.5.1 Les arbres de décision (Quinlan, 1986)

Les arbres de décision sont des modèles d'apprentissage supervisé utilisés à la fois pour la classification et la régression. Ils arrivent à prédire la valeur d'une nouvelle instance en appliquant des règles de décision simples déduites à partir des données d'entrainement.
Un arbre de décision est constitué des éléments suivants :

- **Nœud racine :** il s'agit du nœud initial de l'arbre qui contient l'ensemble des données.
- **Nœud interne :** aussi appelé nœud test, car il applique un test sur un des attributs de la variable cible afin de l'orienter vers une des branches de l'arbre.
- **Branche :** elle correspond à une des réponses possibles à la question posée au niveau du nœud duquel part la branche.
- **Nœud terminal (feuille) :** il représente la valeur que prend la variable cible à la fin de son parcours de l'arbre. Cela peut être une classe dans le cas d'une classification, ou un nombre dans le cas d'une régression.

Au départ de l'arbre, tous les individus sont contenus dans le nœud racine, puis ils sont séparés à chaque itération au niveau des nœuds en se basant sur la valeur de l'attribut qui est testé. Ce dernier est choisi de manière à mieux discriminer les individus et à homogénéiser les sous-ensembles qui découlent du partitionnement effectué.

Il existe des métriques qui nous permettent de tester l'homogénéité d'un jeu de données, et ainsi choisir le critère (attribut) qui nous permettra d'avoir le meilleur découpage. Nous citerons l'indice de Gini et l'entropie de Shannon.

- L'entropie de Shannon : $(y, K) = -\sum_{l \in Y}((P(y = l) \times \log_2(P(y = l))$
- L'indice de diversité Gini : $gini(y, K) = 1 - \sum_{l \in Y} P(y = l)^2$

Avec $\boldsymbol{K}$ représentant l'ensemble d'apprentissage, $\boldsymbol{Y}$ représentant l'ensemble des valeurs que $\boldsymbol{y}$ peut prendre et $\boldsymbol{P(y = l)}$, est la probabilité qu'une instance sélectionnée au hasard de $\boldsymbol{K}$ fasse partie de la classe $\boldsymbol{l}$. Ses deux métriques sont à minimiser.

La construction de l'arbre se fait donc en ajoutant au fur et à mesure de nouveaux nœuds jusqu'à ce que : soit tous les éléments des sous-ensembles ont la même valeur de la variable cible, ou lorsque la séparation n'améliore plus la prédiction.
Le processus que nous venons de décrire est appelé induction descendante d'arbres de décision.

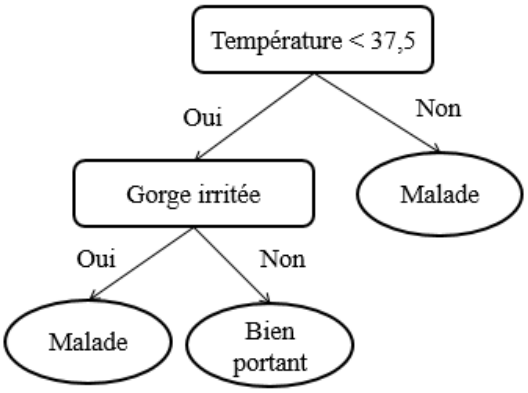

Figure 2-2: Exemple d'un arbre de décision

Les types des arbres de décision :

- **CART (Classification And Regression Tree) :**

Cet algorithme a été introduit par Breiman en 1984. Il produit des arbres binaires, c'est-à-dire qu'un nœud ne peut fournir que deux branches (deux réponses au test) et le critère utilisé pour le choix de la variable discriminante est l'indice de diversité de Gini.
Il traite aussi bien les variables catégorielles que continues.

- **ID3 (Iterative Dichotomiser 3) :**

Cet algorithme a été introduit par Quinlan en 1986, il produit des arbres de décision dont le critère de découpage est l'entropie de Shannon. Il ne prend en compte que les variables catégorielles.

- **C4.5 (Iterative Dichotomiser 4.5) :**

Cet algorithme a été introduit par Quinlan en 1993 comme une amélioration du modèle ID3, il produit des arbres de décision dont le critère de découpage est l'entropie de Shannon. Il prend en compte les variables catégorielles et numériques, néanmoins ces dernières doivent être discrétisées pour garantir l'efficacité du modèle.

Les avantages des arbres de décision :

- Ils prennent en compte aussi bien les variables qualitatives que quantitatives.
- Les données fournies aux modèles ne nécessitent aucun prétraitement (traitement des valeurs manquantes, normalisation et sélection des variables significatives)
- Le nombre de classes que peuvent fournir les arbres n'est pas limité à deux.
- Ils nous permettent de distinguer quelles variables influencent le plus la segmentation des données.
- La méthode est peu gourmande en termes de ressource de calcul.

Les inconvénients des arbres de décision :

- Il se pourrait que l'arbre généralise mal l'ensemble d'apprentissage, et l'on se retrouverait donc avec un problème de sur-apprentissage[1], un élagage[2] approprié de l'arbre est donc primordial.
- Lorsqu'une variable catégorielle possède plusieurs niveaux, elle risque de biaiser le modèle en sa faveur.
- La première segmentation qui se fait au niveau du nœud initial influence grandement la suite du modèle, et si les données prises pour l'apprentissage du modèle sont peu représentatives, il se pourrait que le modèle obtenu présente des résultats erronés.

[1] **Sur-apprentissage** : On parle de sur-apprentissage lorsque le modèle apprend par cœur les données d'apprentissage, perdant ainsi son pouvoir de prédiction sur les nouvelles entrées.
[2] **Elagage** : cela consiste à supprimer les branches de l'arbre qui ne sont pas représentatives, c'est-à-dire qui ne serviront pas à prédire de nouveaux cas, afin de garder de bonnes performances prédictives.

I.5.2 Les forêts aléatoires (Random Forest)

Avant d'entamer une description de l'algorithme Random Forest, nous avons jugé nécessaire de développer certaines notions essentielles à la compréhension de ce type de modèle.

- **Le Bootstrapping** : il s'agit d'une méthode d'inférence statistique qui consiste à créer de nouveaux échantillons avec des tirages avec remise à partir de l'échantillon initial, afin de simuler la distribution d'un estimateur lorsque l'on ne connait pas la loi qu'il suit.

  Supposons que l'on dispose d'un ensemble $Z = (x_1, x_2, \dots, x_n)$ de $\boldsymbol{N}$ données observées de notre population, et que l'on veut calculer une statistique $\boldsymbol{S(T)}$ (une moyenne, une variance...).

  Le Bootstrap consistera donc à former $\boldsymbol{L}$ échantillons $Z_k^* = (x_1^*, x_2^*, \dots, x_{N'}^*)$ pour $k = 1, \dots, L$, tel que chaque $\boldsymbol{Z_k^*}$ est constitué à partir d'un tirage aléatoire avec remise de $\boldsymbol{N'}$ données issues de l'ensemble initial $\boldsymbol{Z}$**.** On pourra alors calculer $\boldsymbol{S(T_k^*)}$ pour chaque échantillon bootstrap et obtenir ainsi $\boldsymbol{L}$ estimations de la statistique que l'on cherche à calculer au lieu d'une seule, on fera la moyenne empirique de ces $\boldsymbol{L}$ valeurs et pourrons estimer avec plus de précision $\boldsymbol{S(T)}$.

- **L'agrégation** : cette méthode est utilisée pour améliorer les performances d'un modèle prédictif en agrégeant plusieurs algorithmes d'apprentissage.

  Soit $\boldsymbol{Y}$ une variable à prédire où $x_1, x_2, \dots, x_p$ sont ses attributs et soit $\boldsymbol{f(X)}$ un modèle fonction de $X_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in R^p$, $i = 1, n$ **.**

  On note $\boldsymbol{n}$ le nombre d'observations, tel que $Z = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ est un échantillon de loi $\boldsymbol{F}$.

  En considérant $\boldsymbol{B}$ échantillons indépendants notés $\{Z_b\}_{b=1,B}$, une prévision par agrégation de modèles est définie comme suit :

  - Dans le cas où la variable à prédire est quantitative, on calculera :

$$\widehat{f_B}(.) = \frac{1}{B} \sum_{b=1}^{B} \widehat{f_{Z_b}}(.)$$

    Il s'agit de la moyenne calculée sur les résultats du modèle pour chaque échantillon.
  - Dans le cas où la variable à prédire est qualitative, on calculera :

$$\widehat{f_B}(.) = \arg\max_j card\{b \mid \widehat{f_{Z_b}}(.) = j\}$$

    Il s'agit d'une sorte de vote majoritaire des modèles, en d'autres termes, nous choisirons la prédiction qui est apparue le plus de fois.

- **Le Bagging** : cette méthode a été introduite par Breiman en 1996. Elle consiste à appliquer le principe de Bootstrap à l'agrégation de classifieurs ; d'où son nom Bagging pour Bootstrap Aggregating. Dans ce cas, la statistique que nous chercherons à approximer est un algorithme d'apprentissage noté $\boldsymbol{V(x)}$**.** Pour cela, plusieurs prédicteurs élémentaires notés $\boldsymbol{V_k(x)}$ seront entrainés sur des échantillons bootstrap de données, pour ensuite en agréger les résultats comme expliqué ci-dessus.

Les forêts aléatoires ont été proposées par Breiman en 2001, elles sont composées comme leurs noms l'indiquent d'arbres de décisions binaires. Il s'agit d'une amélioration du Bagging, car en plus d'utiliser des échantillons Bootstrap pour entrainer les arbres de décision qui le constituent, elles ajoutent un élément aléatoire dans le choix des variables explicatives intervenant dans chaque arbre, le but étant de les rendre indépendants. (Breiman, 1996)

Description de l'algorithme :

Soit $S = \{X_1, Y_1; \ldots . ; X_m Y_m\}$ l'ensemble d'entrainement, $\boldsymbol{a}$ étant le nombre d'attributs de chaque individu.

Considérons $\boldsymbol{S_t}$ un bootstrap contenant m instances obtenues par rééchantillonnage avec remplacement de $\boldsymbol{S}$ et soit $\{h_1, \ldots, h_T\}$ un ensemble de $\boldsymbol{T}$ arbres de décision tels que chaque arbre $\boldsymbol{h_t}$ est construit à partir de $\boldsymbol{St}$.

Pour chaque nœud de l'arbre, l'attribut de discrimination est choisi en considérant un nombre k $(k < a)$ d'attributs choisis aléatoirement (parmi les $\boldsymbol{a}$ attributs), et parmi ces derniers, est choisi celui qui optimise le critère d'homogénéité considéré par les arbres utilisés (entropie de Shannon ou indice de Gini).

Pour prédire un nouvel individu, le modèle obtenu utilise la majorité des votes des arbres de décisions dans le cas de la classification, ou calcule la moyenne des résultats obtenus par chaque arbre.

Les avantages des forêts aléatoires :

- Elles évitent le sur-apprentissage.
- Elles améliorent les performances des arbres de décisions.

Les inconvénients des forêts aléatoires :

- Perte de lisibilité des arbres de décisions (effet boite noire)
- Importants temps de calcul.

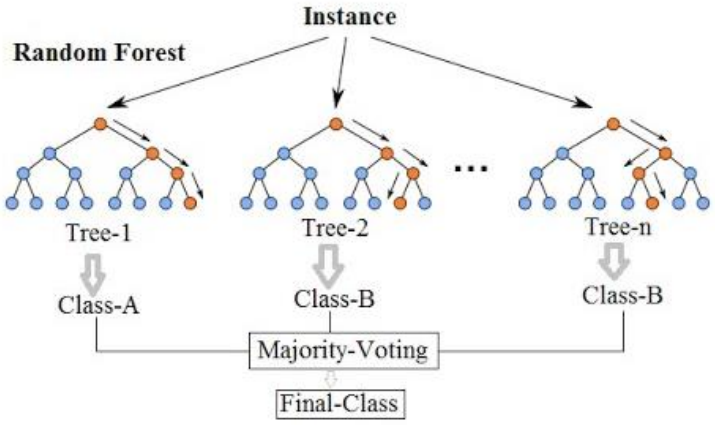

Figure 2-3 : Exemple illustrant le déroulement de Random Forest

I.5.3 AdaBoost (Freund et Schapire, 1996)

Cet algorithme repose sur la méthode du Boosting qui a été initié en 1996 par Freund et Schapire qui voulaient améliorer les performances d'un prédicteur jugé faible (weak learner), qui pourrait certes aboutir à une classification meilleure qu'une affectation aléatoire, mais dont les résultats seraient néanmoins peu robustes. AdaBoost utilise les principes du Bagging et l'améliore dans le sens où à chaque itération, un nouveau modèle prédictif est créé afin d'améliorer les prédictions du modèle qui le précède ; et ce en accordant plus d'importance aux individus mal prédits. Ces derniers auront donc une probabilité plus importante d'être échantillonnés dans l'itération qui suit.

Cette méthode permet d'aboutir à de meilleurs résultats en concentrant les efforts du modèle sur les instances les plus difficiles à estimer, tout en évitant le sur-apprentissage grâce à l'agrégation de modèles prédictifs.

Nous allons décrire dans ce qui suit la première implémentation de cette méthode dans l'algorithme AdaBoost élaboré par Freund et Schapire pour la prévision d'une variable binaire.

Soit $z = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ un ensemble d'apprentissage, où $x_i \in X$ qui est l'ensemble des observations et $y_i \in \{-1, +1\}$ qui est la classe de chaque observation.

Les individus de l'ensemble d'apprentissage sont pondérés uniformément de la manière suivante :

$$\omega_t(i) = \frac{1}{n} \ , \forall\, i = 1, n.$$

Pour $\boldsymbol{t}$ allant de 1 à $\boldsymbol{T}$, où $\boldsymbol{T}$ est le nombre de modèles :

- Trouver le prédicteur $h_t(x): X \to \{-1, +1\}$ qui minimise l'erreur de prédiction $\boldsymbol{\varepsilon_t}$ en fonction du poids des exemples $\omega_t(i)$ avec :

$$\varepsilon_t = \sum_{i=1}^{m} \omega_t(i) * ind\,[y_i \neq h(x_i)]$$

$$h_t = arg \min_{h \in H} \sum_{i=1}^{m} \omega_t(i) * ind\,[y_i \neq h(x_1)]$$

- Si $\varepsilon_t < 0.5$ **,** le prédicteur est sélectionné, sinon l'algorithme s'arrête.
- On choisit alors le poids du prédicteur $\alpha_t \in R$ avec $\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$.
- La pondération des exemples est alors mise à jour comme suit :

$$\omega_{t+1}(i) = \frac{\omega_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t}$$

Où $Z_t = 2 * \sqrt{\varepsilon_t(1-\varepsilon_t)}$ est un facteur de normalisation.

- Le prédicteur qui résulte du processus de sélection est le suivant :

$$H(x) = sign \sum_{t=1}^{T} \alpha_t h_t(x)$$

Ce qui correspond à un vote pondéré des prédicteurs sélectionnés.

Les avantages d'AdaBoost :

- Il est simple à implémenter et donne des résultats très rapidement.
- Il ne nécessite le réglage que d'un seul paramètre (***T***).
- Il garantit théoriquement de meilleurs résultats que Random Forest qui est un algorithme très performant.
- Il améliore les performances de n'importe quel prédicteur faible.

Les inconvénients d'AdaBoost :

- Le choix d'un faible prédicteur adapté au problème n'est pas toujours évident.
- Il est parfois sensible au bruit présent dans les données, dans le sens ou si la qualité des données est altérée, l'algorithme fournira des résultats qui le sont tout autant.

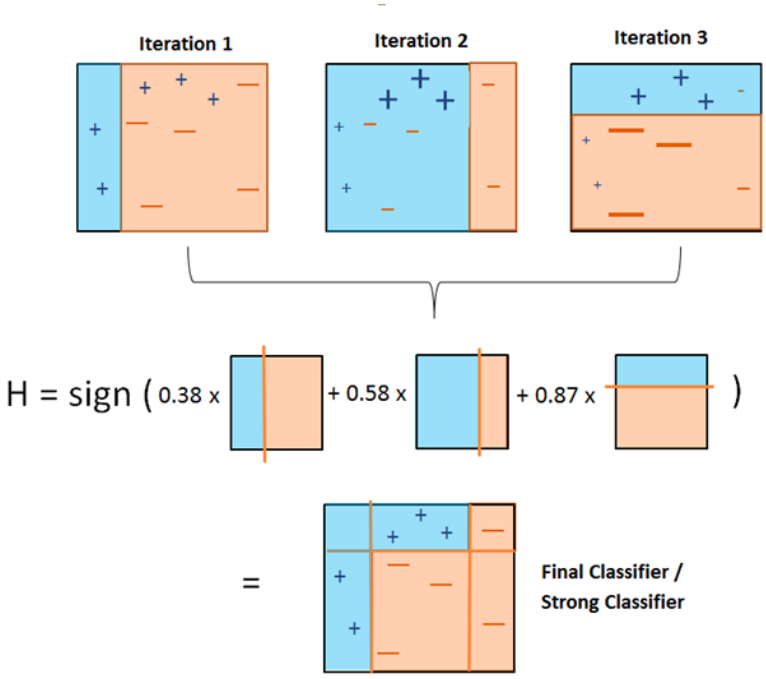

Figure 2-4 : Exemple illustrant le déroulement d'AdaBoost

I.5.4 Le Gradient Boosting (Friedman, 1999)

Il s'agit d'une méthode d'apprentissage supervisée utilisée dans la régression et la classification proposée par Friedman en 1999. Tout comme AdaBoost, il se base sur la construction en séquence de faibles prédicteurs. Cependant, cet algorithme ne réalise pas une somme pondérée ou un vote majoritaire sur les résultats des prédicteurs. Dans ce cas-ci les prédicteurs régressent à chaque itération l'erreur de prédiction commise par leurs prédécesseurs respectifs. Le modèle final sera la somme de la valeur initiale prise par la valeur cible et les résultats de tous les faibles prédicteurs.

Cet algorithme combine donc les deux méthodes de Boosting et de descente de gradient[3]. Car ; premièrement, les modèles sont réalisées en séquence et améliorées à chaque itération. En second lieu l'algorithme régresse le gradient d'une fonction coût, tel que :

Soit $\{(\boldsymbol{x_i}, \boldsymbol{y_i})\}_{\boldsymbol{i=1,n}}$ l'ensemble d'apprentissage,

Soit M le nombre de prédicteurs.

Soit $\boldsymbol{L(y, F(x))}$ une fonction de coût, qui en général est de la forme :

$$\sum_i \frac{1}{2}\left(y_i - F(x_i)\right)^2$$

Soit initialement $\boldsymbol{F_0(x) = f_1(x)}$ un faible prédicteur tel que :

$$y_i = F_0\left(x_i\right) + \varepsilon_1$$

$\boldsymbol{\varepsilon_1}$ représente le résidu (erreur de prédiction du prédicteur$F_0\left(x_i\right)$).

L'idée de l'algorithme est de construire un nouveau modèle $f_2\left(x_i\right)$ pour modéliser ce résidu et l'associer au modèle précédent en compensant additivement l'erreur de prédiction. On aura donc :

$$y_i = F_1(x_i) + f_2(x_i)$$

L'algorithme du gradient cherchera donc à estimer $f_2(x_i) = y_i - F_1$ ce qui revient à estimer :

$-\frac{\delta L(y,F(x_i))}{\delta F_1(x_i)} = -(F_1(x_i) - y_i) = y_i - F_1(x_i)$.

Le prédicteur amélioré sera donc :

$$F_2(x_i) = F_1(x_i) - \frac{\delta L\left(y, F(x_i)\right)}{\delta F(x_i)}$$

Cependant, pour éviter le sur-apprentissage, une constante d'apprentissage $0 < \gamma < 1$ est ajoutée, on obtient alors :

$$F_2(x_i) = F_1(x_i) - \gamma\ \frac{\delta L\left(y,F(x_i)\right)}{\delta F(x_i)}.$$

Nous avons montré comment se déroule une certaine partie de la première itération et nous allons décrire dans ce qui suit les étapes de l'algorithme :

[3] Il s'agit d'un algorithme itératif qui chercher à estimer le minimum d'une fonction différentiable. Pour cela, on part de la fonction d'un point choisi aléatoirement puis on lui ajoute le négatif de la dérivée de cette fonction en ce point multiplié par un facteur appelé pas d'apprentissage. Cette procédure est réalisée jusqu'à ce que la solution converge vers le minimum de la fonction.

Initialiser $f_0(x)$ à $argmin_\gamma\ (L(y,\gamma))$. La résolution de cette équation dans le cas de la régression aboutit à la moyenne des y_i et dans celui de la classification le $\log(\frac{p}{1-p})$ de la population majoritaire.

1. Pour m allant de 1 à M faire :

i. Calculer $r_{im} = -\left[\frac{\delta L(y,F(x_i))}{\delta F(x_i)}\right]_{F(x_i)=F_{m-1}}$ pour $i = 1, n$.

ii. Ajuster un arbre de régression f_m au couple (x_i, r_{im}).

iii. Calculer le multiplicateur γ_m en résolvant l'équation :

$$\gamma_m = argmin_\gamma[L(y_i, F_{m-1}\ (x_i) + \gamma f_m(x_i)].$$

iv. Mettre à jour le modèle :

$$F_m(x_i) = F_{m-1}(x_i) + \gamma_m f_m(x_i).$$

Lorsqu'il s'agit d'une classification ; binaire par exemple, les modèles régressent la valeur $t = \log(\frac{p}{1-p})$ tel que $\boldsymbol{p}$ représente la probabilité d'occurrence de l'évènement.

Pour obtenir la classe à laquelle appartiendra un individu, il suffira d'insérer le résultat du modèle dans la fonction $\frac{e^t}{1+e^t}$.

Nous avons décrit précédemment l'algorithme générique du Gradient Boosting. Cependant il utilise généralement des arbres de décision (de type CART en particulier) comme apprenant faible. Friedman a donc modifié l'algorithme générique pour l'adapter aux arbres de décision de la manière suivante :

Soit $\boldsymbol{J_m}$ le nombre de feuilles de l'arbre $\boldsymbol{m}$ qui partitionne l'espace des données en $\boldsymbol{j_m} = \boldsymbol{1, J_m}$ régions $\boldsymbol{R_{j\,m}}$ et soit $\boldsymbol{b_{jm}}$ la valeur prédite dans cette région.

L'arbre de régression peut alors être modélisé de la forme :

$$f_m(x) = \sum_{j=1}^{J_m} b_{jm} 1_{R_{jm}}(\mathrm{x}).$$

Friedman a modifié l'étape 2.ii en assignant à chaque région $\boldsymbol{R_{jm}}$ au lieu de l'arbre au complet un coefficient d'ajustement γ_{jm} qui est obtenu en résolvant l'équation :

$$\gamma_{jm} = argmin_\gamma[\ \sum_{x_i \in R_{jm}} l(y_i, F_{m-1}\ (x_i) + \gamma)].$$

Le modèle est alors mis à jour de la manière suivante :

$$F_m(x_i) = F_{m-1}(x_i) + \alpha \sum_{j=1}^{J_m} \gamma_{jm} 1_{R_{jm}}(x_i)$$

$\boldsymbol{\alpha}$ étant une constante d'apprentissage fixée en général à 0.1.

Les avantages du Gradient Boosting :

- Il permet de choisir la fonction coût la mieux adaptée aux données.
- De manière générale, il offre de meilleurs résultats que la plupart des autres modèles.

Les inconvénients du Gradient Boosting :

- Il est sensible au bruit et aux valeurs extrêmes.
- Il faut bien choisir le nombre de faibles prédicteurs pour éviter le sur-apprentissage.

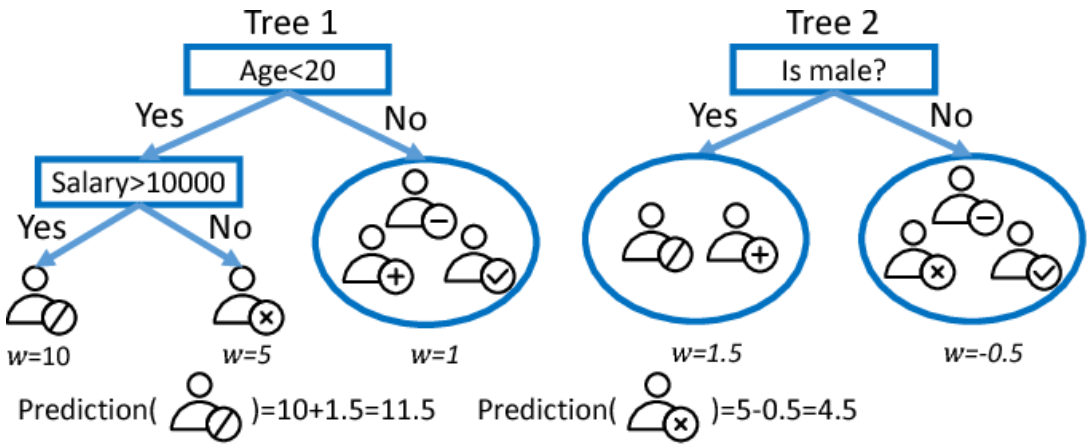

Figure 2-5 : Exemple illustrant le déroulement du Gradient Boosting

I.5.5 Extrem Gradient Boosting (XGBoost) (Chen et Carlos , 2016)

XGBoost a été introduit en 2014 par Tianqi Chen et est devenu depuis l'un des algorithmes les plus utilisés pour la prédiction en remportant plusieurs compétitions d'apprentissage automatique ; car il gère efficacement une grande variété de types de données et possède un grand nombre d'hyper-paramètres qui peuvent être modifiés et réglés à des fins d'amélioration.

Il s'agit d'une implémentation rapide et évolutive de l'algorithme du Gradient Boosting dont le fonctionnement est le suivant :

Soit $\{(x_i, y_i)\}, i = 1, n.$ l'ensemble d'apprentissage et soit $L(y_i, \widehat{y_i})$ une fonction de perte.

Cette fonction dans le cas de la régression est l'erreur quadratique moyenne égale à :

$$L(y_i, \widehat{y_i}) = \frac{1}{P}\sum_{i=1}^{P} (y_i - \widehat{y_i})$$

Avec $\boldsymbol{P}$ le nombre de paramètres des variables.

Dans le cas de la classification la fonction cross entropie définie comme suit est utilisée :

$$L(y_i, \widehat{y_i}) = \frac{-1}{N}\sum_{i=1}^{P} y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

Où $\boldsymbol{p}$ représente la probabilité d'occurrence de l'évènement.

Comme pour dans le cas du Gradient Boosting, nous avons :

$$\widehat{y_i} = \sum_{k=1}^{M} f_k(x_i)$$

Avec $\boldsymbol{f_k}$ étant un arbre de décision tel que $f = \omega_{q(x)}$ où $\boldsymbol{q(x)}$ représente la structure de chaque arbre (une fonction) qui rattache une observation à la feuille qui lui correspond, $\boldsymbol{\omega}$ la valeur que prend cette feuille et **M** est le nombre d'arbres du modèle.

Soit la fonction objectif qui est une fonction représentant une perte à optimiser suivante :

$$obj^t(\theta) = \sum_{i=1}^{n} L(y_i, \widehat{y_i}) + K(f_k)$$

Où $\boldsymbol{\theta}$ représente les paramètres du modèle à optimiser et t l'itération.
$\boldsymbol{K}$ est un terme de régulation permettant de contrôler la complexité du modèle et d'éviter le sur-apprentissage, il est défini par :

$$K(f_k) = \gamma T + \frac{1}{2}\alpha\|\omega\|^2$$

Avec $\boldsymbol{T}$ représentant le nombre de feuilles de l'arbre et $\boldsymbol{\omega_j}$ la prédiction de la ***j*éme** feuille.
$\boldsymbol{\alpha}$ et $\boldsymbol{\gamma}$ sont des paramètres de pénalisation du modèle.

La fonction objectif peut se réécrire de la manière suivante :

$$obj^t(\theta) = \sum_{i=1}^{n} L(y_i^t, y_i^{t-1} + f_t(x_i)) + K(f_t) + constante.$$

Afin de pouvoir utiliser les techniques d'optimisation traditionnelles, cette fonction doit être transformée en utilisant le développement de Taylor qui se présente comme suit :

$$f(x+a) = f(x) + f'(x) \times a + \frac{1}{2}f''(x) \times a^2$$

La fonction objective devient :

$$obj^t(\theta) = \sum_{i=1}^{n} L(y_i^t, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t(x_i)^2 + K(f_t) + constante.$$

Où $g_i = \partial_{\hat{y}_i^{t-1}} L(y_i^t, \hat{y}_i^{t-1})$ et $h_i = \partial^2_{\hat{y}_i^{t-1}} L(y_i^t, \hat{y}_i^{t-1})$ sont les dérivées partielles respectivement d'ordre 1 et 2 de la fonction de perte.

En enlevant les termes constants et en remplaçant K par sa valeur on obtient :

$$obj^t(\theta) = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2}h_i f_t(x_i)^2] + \gamma T + \frac{1}{2}\alpha \sum_{j=1}^{T} \omega_j^2$$

Qui devient :

$$obj^t(\theta) = \sum_{j=1}^{T} [(\sum_{i \in S_i} g_i \omega_j + \frac{1}{2}\left(\sum_{i \in S_i} h_i + \alpha\right)\omega_j^2] + \gamma T$$

Avec $S_j = \{i|q(x_i) = j\}$ est l'ensemble des individus affectés à la même feuille $\boldsymbol{j}$ pour un arbre de structure $\boldsymbol{q}$.

Ce qui précède est devenu une somme de fonctions d'une variable quadratique et simple pouvant être minimisée en utilisant des techniques connues. Nous allons donc chercher la valeur optimale de la prédiction $\boldsymbol{\omega_j}$ pour la feuille $\boldsymbol{j}$ qui minimise la fonction objectif. En dérivant et annulant cette dernière par rapport à $\boldsymbol{\omega}$ nous obtenons :

$$\boldsymbol{\omega_j^* = \frac{\sum_{i\in j} g_i}{\sum_{i\in S_j} h_i + \alpha}}.$$

Ce qui correspond à une valeur de la fonction objectif égal à :

$$obj^{*t}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in j} g_i)^2}{\sum_{i\in S_j} h_i + \alpha} + \gamma T$$

C'est cette dernière équation qui mesure la qualité de l'arbre ajouté ; tel que plus cette valeur sera petite et plus l'ajout de l'arbre associé sera pertinent.

Maintenant que l'on possède un moyen de mesurer la qualité d'un arbre, il reste à déterminer quelle structure de l'arbre choisir. Cette question peut se décomposer en deux autres : comment choisir les variables à chaque division au niveau d'un nœud ? Et quand arrêter la division ?
Pour répondre à ces deux questions il suffit d'évaluer pour chaque variable à chaque nœud la valeur du gain défini par :

$$gain = perte\ avant\ le\ split - perte\ aprés\ le\ split$$
$$gain = perte\ au\ niveau\ du\ noeud\ pére - (perte\ au\ niveau\ du\ \ noeud\ fils\ droit + perte\ au\ niveau\ du\ \ noeud\ fils\ gauche)$$

Nous obtenons :

$$gain = \frac{1}{2}\left(\frac{G_R^2}{H_R + \alpha} + \frac{G_L^2}{H_L + \alpha} - \frac{(G_R + G_L)^2}{H_R + H_L + \alpha}\right) - \gamma$$

Avec : $G_R = \sum_{i\in S_R} g_i$, $G_L = \sum_{i\in S_L} g_i$, $H_R = \sum_{i\in S_R} h_i$ et $H_L = \sum_{i\in S_L} h_i$.

Il s'agira donc de choisir la variable qui maximise ce gain et l'on peut remarquer que si le gain est plus faible que le terme de régularisation, alors il n'y a pas d'intérêt à effectuer cette subdivision. On répond alors à la deuxième question évoquée ci-dessus.

Les avantages de l'Extrem Gradient Boosting :

- Il est globalement plus performant que tous les algorithmes d'apprentissage supervisé et remporte souvent les compétitions de Machine Learning.
- Il converge rapidement vers la solution optimale et est donc peu gourmand en temps de calcul.
- Il traite les valeurs manquantes et les données n'ont pas besoin de prétraitement.

Les inconvénients de l'Extrem Gradient Boosting :

- Il a tendance à faire du sur-apprentissage.

I.5.6 Les k plus proches voisins (Cover et Hart, 1967)

Avant d'entamer la description de cette méthode, il est essentiel de définir les deux notions suivantes :

Espace de caractéristiques : Chaque observation de l'ensemble d'apprentissage est représentée par un point dans un espace qui comporte autant de dimensions que l'individu a d'attributs.

Mesure de similarité : Une notion de distance est utilisée dans cet espace pour trouver les ***k*** plus proches voisins d'une nouvelle entrée. La métrique utilisée par défaut est la distance euclidienne.

L'algorithme des k plus proches voisins (k Nearest Neighbours (kNN)) est une méthode d'apprentissage supervisée qui consiste simplement à prédire de nouveaux individus sur la base de leur similarité avec les exemples d'apprentissage. De ce fait, cet algorithme ne construit aucun modèle. A chaque fois qu'il devra estimer une nouvelle instance il utilisera l'ensemble d'apprentissage.

Dans le cas de la classification, la classe d'un nouvel individu sera la classe majoritaire de ses k-plus proches voisins.

Dans le cas d'une régression, la classe d'un nouvel individu sera la moyenne des valeurs de ces k plus proches voisins.

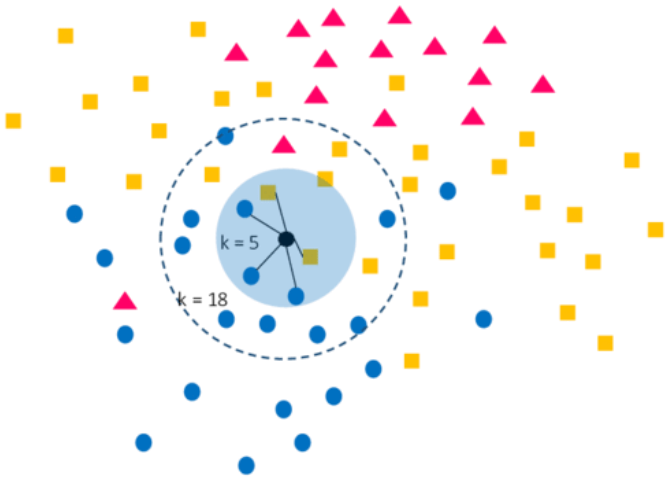

Figure 2-6 : Classification avec l'algorithme des k plus proches voisins

Dans l'exemple illustré dans la ***Figure 2-6***, un nouvel individu sera affecté à la classe « cercle bleu » dans les deux cas : k=5 ou k=18, car la majorité de ses k voisins appartiennent à cette classe.

Les avantages de l'algorithme des k plus proches voisins :

- Il ne nécessite aucune phase d'apprentissage.
- Son principe de classification est trivial et donc facilement compréhensible.

Les inconvénients de l'algorithme des k plus proches voisins :

- La prédiction est lente et gourmande en mémoire, car elle implique de passer en revue tous les exemples d'apprentissage.
- Cette méthode est sensible aux variables non pertinentes et corrélées.
- Lorsque la dimension de l'espace de caractéristique est grande, le calcul de la distance peut s'avérer très couteux.

I.5.7 Les machines à vecteur support SVM (Guyon et al., 1992)

Les machines à vecteurs supports ou les séparateurs à vastes marges sont des méthodes de classification binaire supervisée introduite par Boser, Guyon et Vapnik dans les années 1990.

Leur principe de base consiste à trouver un classifieur sous la forme d'un hyperplan pouvant séparer deux groupes d'observations de telles sorte que la marge entre les plus proches individus des deux ensembles et l'hyperplan soit maximale. Les points les plus proches, qui sont seuls utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support, ils se trouvent à une distance égale à la marge d'un côté ou de l'autre de l'hyperplan de séparation comme le montre la figure qui suit :

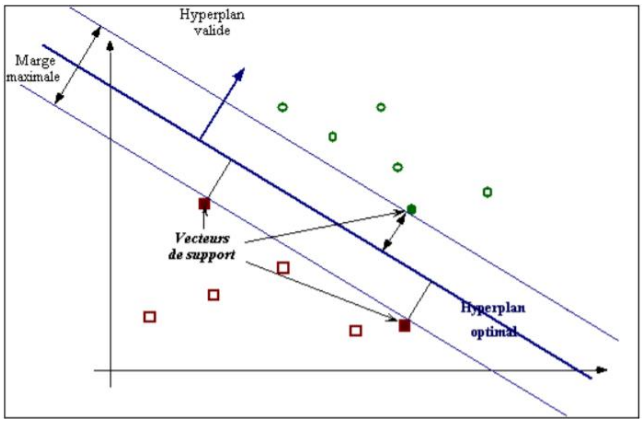

Figure 2-7 : Hyperplan optimal avec une marge maximale

Si on arrive à trouver un séparateur linéaire ; c'est-à-dire qu'il existe un hyperplan séparateur alors, le problème est dit linéairement séparable sinon, il n'est pas linéairement séparable et il n'existe pas un hyperplan séparateur.

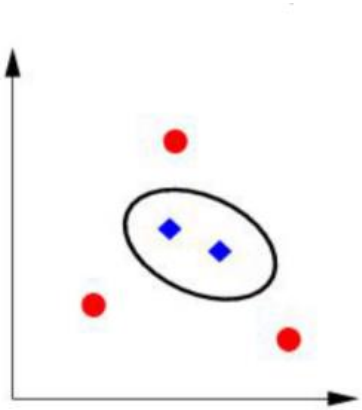

Figure 2-8 : Cas non linéairement séparable

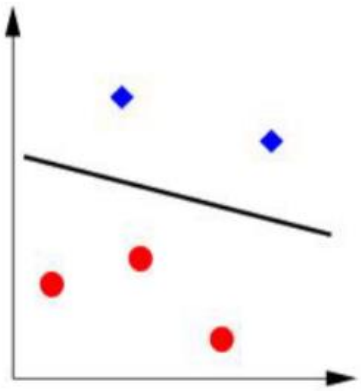

Figure 2-9 : Cas linéairement séparable

a) Le cas linéairement séparable :

Soit $\{(x_i, y_i), i = 1, n\}$ l'ensemble d'apprentissage, avec $x_i \in R^P$ et $y_i = \{-1, +1\}$, P étant le nombre d'attributs d'un individu.
On dit qu'un problème est linéairement séparable lorsqu'il existe une fonction de décision linéaire pouvant classer correctement toutes les observations de l'ensemble d'apprentissage. Elle aura pour forme :

$$f(x) = (\omega^T x + a\ )$$

$avec\ \omega \in R^P\ et\ a\ \in R$, $\boldsymbol{\omega}$ étant le vecteur poids et $\boldsymbol{a}$ le déplacement par rapport à l'origine. Lorsque $f(x) > o$, l'individu sera de classe 1 et lorsque $f(x) < o$, l'individu sera de classe -1.

L'apprentissage de l'algorithme consiste à estimer l'hyperplan optimal à partir des données d'entrainement. C'est-à-dire celui qui maximisera la marge (qui est la distance entre le plus proche exemple d'apprentissage et l'hyperplan de séparation), il aura pour équation :

$$\omega^T x + a = 0.$$

Formulation du problème :

- Soit $\frac{\omega^T x + a}{\|\omega\|}$, la distance de tout point par rapport à de l'espace de l'hyperplan.
- Un point est bien classé si et seulement si $y_i(\omega^T x_i + a\ ) > 0$.
- Maximiser la marge *M* est équivalent à maximiser la somme des distances des deux classes par rapport à l'hyperplan, et maximiser cette distance reviendrait à maximiser la somme des distances entre les vecteurs supports et l'hyperplan. On obtient la formule suivante :

$$M = \min_{x_i | y_i = 1} \frac{\omega^T x_i + a}{\|\omega\|} - \max_{x_i | y_i = -1} \frac{\omega^T x_i + a}{\|\omega\|}$$

Cependant les paramètres $\boldsymbol{\omega}$ et $\boldsymbol{a}$ ne sont pas uniques, $\boldsymbol{k\omega}$ et $\boldsymbol{ka}$donnent la même surface de séparation :

$$k\omega^T x_i + ka = w^T x_i + a = 0\ .$$

On impose alors : $|\omega^T x_s + a| = 1$, $\boldsymbol{x_s}$ étant un vecteur support.

On obtient :

$$M = \frac{1}{\|\omega\|} - \frac{-1}{\|\omega\|} = \frac{2}{\|\omega\|}$$

- Trouver l'hyperplan optimal revient donc à maximiser $\frac{2}{\|\omega\|}$. Ce qui est équivalent à minimiser $\frac{1}{2}\ \|\omega\|^2$ sous la contrainte $y_i(\omega^T x_i + a\ ) > 1$.

b) Le cas non linéairement séparable :

Lorsque les observations ne peuvent pas être séparées linéairement par un hyperplan, l'idée des SVM sera de projeter les données dans un espace de dimension supérieure où il serait très probablement possible de les séparer linéairement.
Pour cela on appliquera à nos données une transformation non linéaire $\emptyset$, tel que l'espace d'arrivée de $\emptyset(\boldsymbol{x})$, sera appelé espace de description.

Dans cet espace, on cherche l'hyperplan :

$$h(x_i) = \omega^T \emptyset(x_i) + a \text{ qui vérifie} \quad h(x_i)y_i > 0.$$

La solution à ce problème dépend du produit scalaire entre vecteurs $\langle \emptyset(\boldsymbol{x}), (\emptyset\ \boldsymbol{x}') \rangle$ dans l'espace de redescription qui est de grande dimension, ce qui conduit à des calculs volumineux.

Néanmoins il est possible d'estimer ce produit scalaire sans pour autant expliciter $\emptyset$, et ce à l'aide de l'astuce du noyau qui implique l'emploi d'une fonction noyau facile à calculer notée $k(x_1, x_2) = \langle \emptyset(x_1), (\emptyset\ x_2) \rangle$ devant satisfaire plusieurs conditions et correspondant à un produit scalaire dans un espace de grande dimension .

Voici quelques exemples de noyaux utilisés :

- **Noyau linéaire** : $k(x_1, x_2) = x_1 * x_2$.
- **Noyau polynomial** : $k(x_1, x_2) = c + (x_1 * x_2)^n, \ \ c \in R$.
- **Noyau gaussien (radial)** : $k(x_1, x_2) = e^{-\frac{(x_1 - x_2)^2}{\sigma}}, \ \sigma \in R$.
- **Noyau laplacien** : $k(x_1, x_2) = e^{-\frac{(x_1 - x_2)}{\sigma}}$.

Les avantages des machines à vecteur support :

- Elles jouissent d'une solide base théorique.
- Elles permettent de résoudre des problèmes non linéaires complexes grâce aux choix des noyaux.
- Elles sont très efficaces pour les problèmes à grandes dimensions.
- Leur apprentissage n'est pas coûteux en termes de temps.

Les inconvénients des machines à vecteur support :

- Il n'existe pas de méthode bien définie pour choisir la fonction noyau.
- Elles ne traitent que des problèmes de classification binaires.

I.6 Algorithmes d'apprentissage non supervisé

Nous allons dans cette partie, décrire certains algorithmes d'apprentissage non supervisé ***Figure 2-10*** et comme nous l'avons fait pour les modèles supervisés, nous citerons leurs avantages et leurs inconvénients.

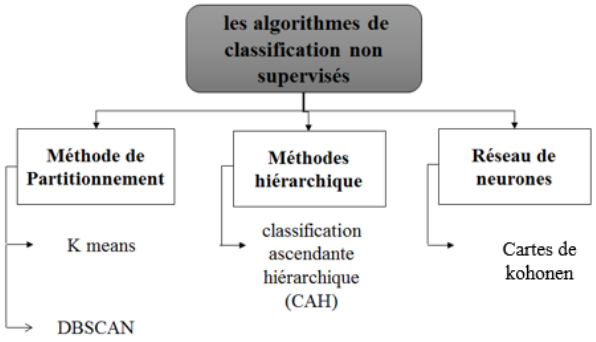

Figure 2-10 les algorithmes d'apprentissage non supervisé

I.6.1 DBSCAN (Ester 1996)

DBSCAN (density-based spatial clustering of applications with noise) est un algorithme de clustering proposé en 1996 par Martin Ester, Hans-Peter Kriegel, Jörg Sander et Xiaowei Xu. Il se base sur l'estimation de la densité des clusters pour établir un partitionnement des données. Cette densité est représentée par les deux paramètres suivants :

L'$\boldsymbol{\varepsilon}$-voisinage d'un point : il s'agit de l'ensemble des points appartenant à la boule de rayon $\boldsymbol{\varepsilon}$ centrée sur ce point.

Le $\boldsymbol{Minpts}$: il s'agit du nombre minimum d'individus devant se trouver dans la boule de rayon $\boldsymbol{\varepsilon}$ pour créer un cluster.

L'ensemble des individus représentés par des points se classent en 3 catégories :

- **Point central (core) :** un point p est dit central, s'il contient au moins $Minpts$ individus dans son voisinage formé par un cercle de rayon $\boldsymbol{\varepsilon}$. Un point du ε-voisinage de p est dit directement densité accessible depuis p.

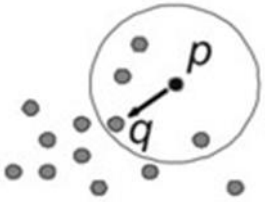

Figure 2-11 : Point central

- **Point bordure (border) :** un point q est dit bordure, s'il contient moins de $Minpts$ individus dans son ε –voisinage mais qui est dans le voisinage d'un point central. Il est alors dit que q est densité-accessible depuis p s'il existe un chemin $p = p_1, p_2, \ldots, p_n = q$ de sorte que chaque p_{i+1} est directement densité accessible depuis p_i, et tous les points du chemin sont des points centraux.

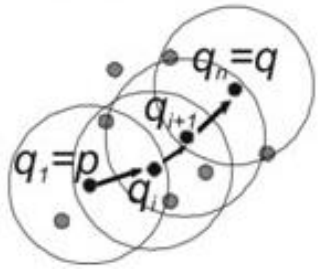

Figure 2-12 : Point bordure

- **Point bruit (noise) :** un point est dit bruit s'il n'est ni central ni voisinage (c'est à dire, il n'est atteignable par aucun type de point).

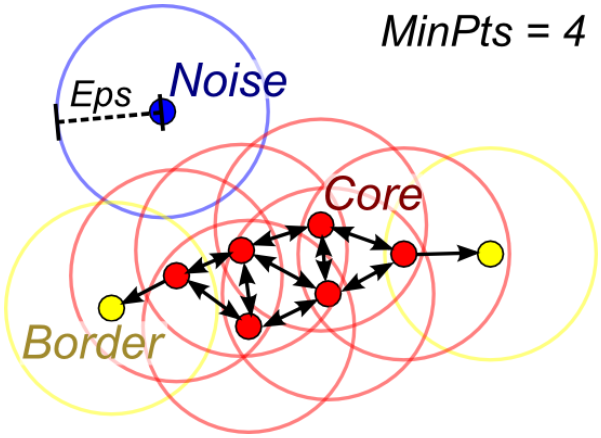

Figure 2-13 : Exemple des différents points

L'algorithme DBSCAN fonctionne comme suit :

- Un point est choisi aléatoirement. Si son ε-voisinage contient $Minpts$ individus, un cluster commence à se former. Sinon, ce point est considéré temporairement comme bruit jusqu'à ce qu'il apparaisse dans l'ε-voisinage d'une autre classe.

- Les points faisant partie du cluster créé sont ensuite examinés. S'ils sont centraux leur ε-voisinage viendra rejoindre le cluster auquel ils appartiennent. Le processus continue jusqu'à la complétion de la classe avec les points densités-accessible.

- Un autre point n'ayant pas encore été visité est alors choisi et le processus se renouvèle jusqu'à ce que tous les points soient étiquetés.

Les avantages de l'algorithme DBSCAN :

- Il ne nécessite pas qu'on spécifie le nombre de classes.
- Il détecte les données aberrantes (outliers) et les élimine.

Les inconvénients de l'algorithme DBSCAN :

- le choix des valeurs de $Minpts$ et ε est délicat et nécessite une parfaite connaissance des données.
- Il fonctionne mal avec les dimensions élevées.
- Il ne peut pas gérer des clusters ayant des densités différentes.

I.6.2 Les k-moyennes (k-means) (Forgy, 1965)

La classification k-means a été introduite par E. W. Forgy en 1965. Elle permet de répartir les données en k classes homogènes en minimisant la fonction :

$$f = \sum_{i=1}^{k} \sum_{x_j \in c_i} \|x_j - u_i\|$$

Qu'on appelle variance intra classe, avec u_i représentant la moyenne.

Soit l'ensemble ***D*** composé de ***N*** individus $D = \{x_1, x_2, \dots, x_n\}, \forall x_{i=1,n}: x_i \in R^d$, **d** étant le nombre d'attributs de chaque individu.

L'algorithme classiquement utilisé pour résoudre cette optimisation est le suivant :

- Initier aléatoirement k centres de clusters c_k.
- Répéter jusqu'à convergence de l'algorithme vers une partition stable :
  - Générer un nouveau cluster en assignant chaque individu au centre auquel il est le plus proche : $x_i \in c_k$ si $\forall j \ |x_i - \mu_k| = \min |x_i - \mu_j|$, avec μ_k le centre de la classe k.
  - Mettre le barycentre du cluster à jour en le remplaçant par la moyenne des individus qui font partie du cluster.

Les avantages de la méthode des k-means :

- Il est très facile à comprendre et à mettre en œuvre.
- Il peut gérer une grande quantité de donnée.

Les inconvénients de la méthode des k-means :

- Le nombre de classes doit être préalablement fixé.
- Le résultat dépend de l'initialisation du centre des classes.

I.6.3 Classification ascendante hiérarchique (CAH) (Lance et Williams, 1967)

Il s'agit d'une méthode d'apprentissage automatique non supervisé introduite par Lance et William en 1967. Elle est utilisée en analyse de données pour assigner les individus à des classes en prenant en compte un critère de ressemblance bien défini qui s'exprime sous la forme d'une matrice de distance représentant les dissemblances entre les individus pris deux à deux.

On parle de classification ascendante car au départ, chaque individu forme une classe à lui seul. L'algorithme cherchera à chaque itération à fusionner les classes dont la dissemblance est la plus faible jusqu'à n'obtenir à la fin qu'une seule classe.

Le terme hiérarchie fait référence au dendrogramme auquel aboutit l'algorithme et qui consiste en une imbrication de classes.

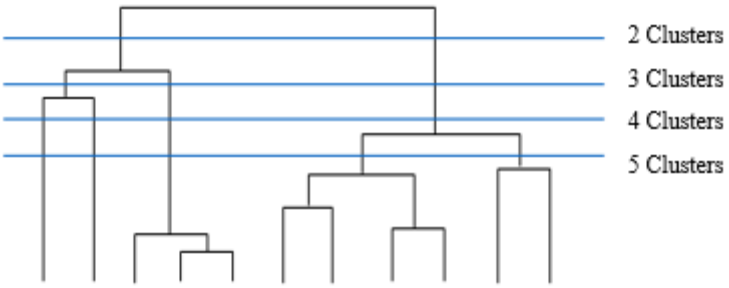

Figure 2-14 : Dendrogramme de la Classification Ascendante hiérarchique

De nombreuses métriques sont utilisées pour mesurer la distance entre individus (indice de dissimilarité), nous citerons les plus utilisées :

- La distance Euclidienne : $d(I_i, I_k) = \sum_k \sqrt{(x_{ik} - x_{jk})^2}$
- La distance du City-block (Manhattan) : $d(I_i, I_k) = |x_{ik} - x_{jk}|$
- La distance de Tchebychev : $d(I_i, I_k) = \max(x_{ik} - x_{jk})$
- Percent disagreement : $d(I_i, I_k) = \frac{nombre\ de\ (x_{ik} \neq x_{jk})}{K}$, cette distance est utilisée avec les données catégorielles.

Quant à la mesure de dissimilarité entre classes (indice d'agrégation), elle se mesure à l'aide des métriques suivantes :

- Le plus proche voisin : $d(c_1, c_2) = \min((d\ (i,j)\ ;\ i\ \in\ c_1;\ ;\ j\ \in\ c_2)$.
- le diamètre maximum : $d(c_1, c_2) = \max((d\ (i,j)\ ;\ i\ \in\ c_1;\ ;\ j\ \in\ c_2)$.
- La distance des centres de gravité : $d(c_1, c_2) = \mathrm{d}(\mu_{c_1}, \mu_{c_2})$.
- La distance de Ward = : $d(c_1, c_2) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} * \mathrm{d}(\mu_{c_1}, \mu_{c_2})$.

Principe général de l'algorithme :

- Créer une classe pour chaque individu.
- Construire la matrice des distances en utilisant le critère de dis similarité choisi.
- Répéter jusqu'à n'avoir qu'une seule classe :

  - Regrouper les deux classes qui sont les plus proches en termes de distance choisie.
  - Remplacez les deux classes regroupées par la nouvelle classe créée et calculer les distances la séparant des autres classes.

Les avantages de la classification ascendante hiérarchique :

- Le choix de la dissimilarité peut être fait selon le type de données manipulées.
- Le nombre adéquat de classes devant être sélectionné est facilement repérable grâce au dendrogramme.

Les inconvénients de la classification ascendante hiérarchique :

- Le temps de calcul augmente avec le nombre d'individu.

I.6.4 Méthodes d'apprentissage non supervisé issues des réseaux de neurones

Les réseaux de neurones artificiels dans leur forme la plus basique ont été introduits dans le but de reproduire le fonctionnement du cerveau humain. Ces réseaux sont composés de processus élémentaires de calcul appelés neurones connectés les uns aux autres et échangeant des informations à l'aide de connexions qui les lient. Ces connexions sont munies de poids représentant la mesure de ces liens. Les poids sont déterminés en utilisant un ensemble d'apprentissage qui va spécialiser le réseau de neurones. Cet apprentissage peut être supervisé ou non.

Le réseau de neurone dans sa forme la plus basique est appelé Perceptron monocouche. Il s'agit d'un classifieur linéaire qui ne possède qu'une seule sortie booléenne. Sa formulation mathématique est illustrée dans la figure ci-dessous :

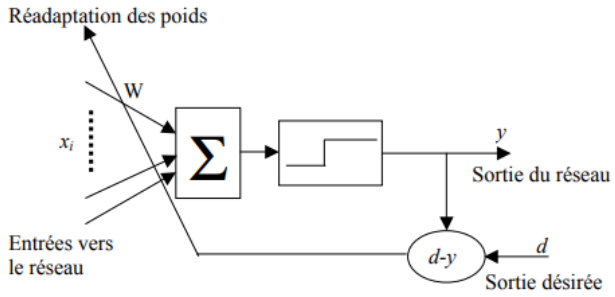

Figure 2-15 : Perceptron monocouche

Soit $x_i^t = (P_1, P_1, \dots, P_R), i = 1, \boldsymbol{n}$ des entrées ayant **R** attributs.

Soit $w_{1,i}^t = (w_{1,1}, w_{1.2}, \dots, w_{1,R})$ le vecteur poids du neurone.

Le neurone est constitué d'un intégrateur qui effectue la somme pondérée des $\boldsymbol{R}$ entrées définie par : $y = \sum_{j=1}^{R} w_{1,j} P_j - b$, ou $\boldsymbol{b}$ est un facteur correctif trouvé par tâtonnement appelé biais du neurone et $\boldsymbol{n}$ le niveau d'activation du neurone. Le résultat obtenu passe ensuite par une fonction d'activation f qui produit la sortie du neurone $\boldsymbol{y}$.

Les fonctions d'activations les plus utilisées sont les suivantes :

- La fonction seuil : $y = \begin{cases} 0, & n < 0 \\ 1, & n \geq 0 \end{cases}$
- La fonction sigmoïde : $y = \frac{1}{1-e^{-n}}$

L'adaptation des poids se fait comme suit :

$$w(t+1) = w(t) + \Delta w(t)$$

Avec $\Delta = \delta\big(d(t) - y(t)\big) * x(t)$, où δ est le pas d'apprentissage et $d(t)$ est la sortie désirée.

Dans cette section nous allons aborder les modèles d'apprentissage non supervisé basés sur les réseaux de neurones.

I.6.4.1 Les cartes auto-adaptatives de Kohonen (Kohonen, 1982)

Les cartes auto adaptatives, auto-organisatrices, topologiques ou tout simplement cartes de Kohonen sont des méthodes d'apprentissage non supervisé basées sur les réseaux de neurones artificiels. Elles ont été introduites par le statisticien Teuvo Kalevi Kohonen en 1982 qui cherchait à représenter des données multidimensionnelles et de grande taille. Pour cela, il a utilisé l'apprentissage pour partitionner ses données en groupements similaires dont la structure de voisinage peut être matérialisée et visualisée en les projetant dans un espace discret de faible dimension (1, 2 ou 3D) appelé «carte topologique». Cette carte respecte la notion de voisinage entre classes dans le sens où des observations voisines dans l'espace des variables (de grande dimension) appartiennent après classement à la même classe ou à des classes voisines.

Les cartes de Kohonen sont représentées sous forme d'une grille unidimensionnelle ou plus souvent bidimensionnelle notée A. Chaque nœud de cette grille correspond à un neurone, et chaque neurone possède une position fixe sur la carte donnée par son numéro de ligne et son numéro de colonne (si elle est bidimensionnelle). Cet emplacement est donné par le vecteur $r = (i, j)^T$, $\boldsymbol{i}$ représentant la ligne, $\boldsymbol{j}$ la colonne.

Le neurone est aussi associé à un vecteur adaptable appelé vecteur référent dans l'espace d'entrée. Il est noté par $w_r, r \in A$ avec $w_r \in R$, l'ensemble des réels .

Le réseau de Kohonen est constitué d'une couche d'entrée et d'une couche de sortie appelée couche de compétition. Tout individu devant être classé sera présenté à la couche d'entrée sous forme d'un vecteur à plusieurs dimensions.

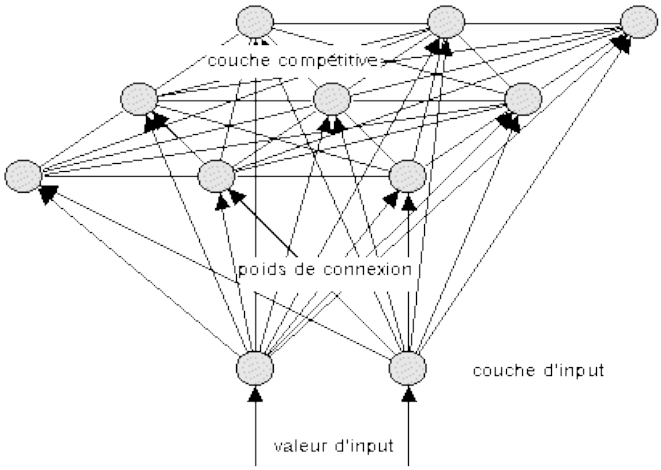

Figure 2-16 : Réseau de Kohonen

L'objectif de la carte sera de mettre à jour les vecteurs référents, de façon à approximer au mieux la distribution des vecteurs d'entrée tout en reproduisant la topologie des données sur la carte. L'algorithme présenté ci-après décrit ce processus d'apprentissage :

- **Initialisation** : chaque individu représenté par un vecteur ayant $\boldsymbol{d}$ attributs doit être centré et réduit (sa moyenne étant nul et son écart type égal à1). Les vecteurs référents seront initialisés aléatoirement et seront également centrés et réduits.
- **Présentation des individus :** les individus sont présentés aléatoirement à la couche d'entrée.

- **Recherche du neurone vainqueur :** chaque vecteur en entrée (stimulus) v sera comparé à chaque neurone w_r de la couche de compétition. Le neurone gagnant sera le neurone le plus similaire au vecteur d'entrée :

$$w_{gagnant}^{t} = argmin_{r \in A} \| v^t - w_r^t \|.$$

- **Mise à jour des prototypes des vecteurs référents** : le vecteur référent du neurone gagnant est modifié ainsi que tous les autres vecteurs référents des autres neurones mais de manière moindre pour se rapprocher du vecteur stimulus comme suit :

$$w_r^{t+1} = w_r^t + \alpha(t) h(c, r, t)(v_t - w_r^t)$$

Tel que :

$$\alpha = \alpha_0 \frac{K}{C + t}$$

Il s'agit du pas d'apprentissage, il règle la vitesse d'apprentissage et doit décroitre avec le temps comme l'indique l'équation. $\boldsymbol{t}$ est l'itération actuelle, $\boldsymbol{\alpha_0}$ le pas d'apprentissage initial et $\boldsymbol{K}$ une constante arbitraire.

$$h(c, r, t) = e^{-\frac{d(r_c - r_r)^2}{2\sigma(t)^2}}$$

Il s'agit de la fonction de voisinage, elle correspond à la pondération du neurone $\boldsymbol{r}$ lorsque le neurone gagnant est le neurone $\boldsymbol{c}$. Elle force les neurones qui se trouvent dans le voisinage de $\boldsymbol{c}$ à rapprocher leurs vecteurs référents du vecteur d'entrée $\boldsymbol{v}$. Plus un neurone est loin du vainqueur dans la grille, moins son déplacement est important.

$d(r_c - r_r)^2$ représente la distance entre les neurones $\boldsymbol{c}$ et $\boldsymbol{r}$ sur la carte, $\boldsymbol{\sigma(t)}$ représente le rayon de voisinage ; son rôle est de déterminer un rayon de voisinage autour du neurone vainqueur, il décroit avec le temps et est défini par l'équation suivante :

$\sigma(t) = \sigma_0 \frac{T-t}{T}$ avec T étant le nombre d'itération totale et $\boldsymbol{\sigma_0}$ le rayon de voisinage initial.

- **Utilisation d'un critère d'arrêt :** Il se fait usuellement en définissant au préalable le nombre d'itérations car l'algorithme continue à converger même en absence de changement d'affectation.

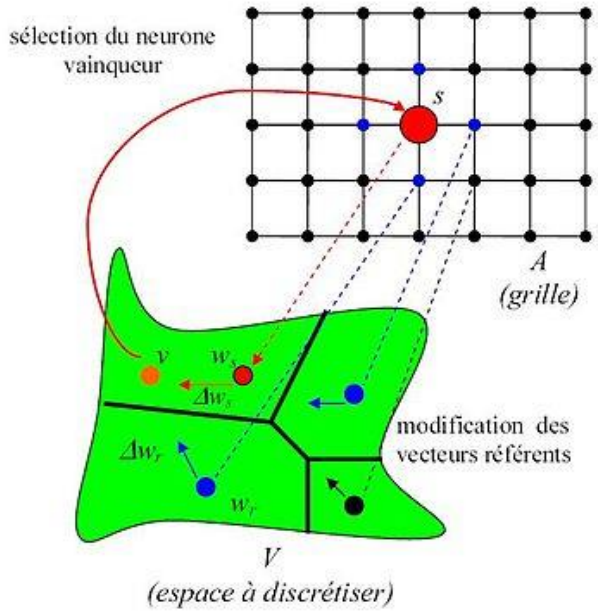

Figure 2-17 : Carte de Kohonen

Les avantages des cartes Kohonen :

- Elles permettent une réduction non linéaire des dimensions.
- Elles utilisent des opérations simples et sont donc économiques en temps de calcul.

Les inconvénients des cartes Kohonen :

- Elles sont sensibles aux valeurs extrêmes.
- La topologie de la carte doit être spécifiée dès le début.
- La convergence vers une solution globale n'est pas assurée.

I.6.4.2 Les auto-encodeurs (Dufour et al ., 2016)

Un auto-encodeur est un réseau de neurones artificiels utilisé pour l'apprentissage non supervisé de caractéristiques discriminantes. En d'autres termes, il permet de construire une nouvelle représentation d'une donnée en la transformant en une version plus compacte (ayant moins de dimensions) au lieu de l'utiliser pour prédire une variable cible.

Dans sa forme la plus basique, un auto-encodeur est composé de 3 couches de neurones : la couche d'entrée, la couche cachée et la couche de sortie.

La première couche ainsi que la 2éme forment l'encodeur. Ce dernier est chargé de traiter les données pour en construire une nouvelle représentation dite encodée. La dernière couche et la couche cachée forment le décodeur, il est chargé de traiter la donnée encodée pour en restituer la version de départ.

L'encodeur construit le vecteur $h = \sigma(Wx + b)$ de taille $\boldsymbol{m}$ ($\boldsymbol{m}$ correspond au nombre de neurones de la couche cachée).

Où $x \in R^n$ avec $\boldsymbol{n}$ le nombre d'attributs de l'individu, $W \in R^m \times R^n$ est une matrice de poids, $b \in R^m$ est un vecteur biais et $\boldsymbol{\sigma}$ est une fonction d'activation de type sigmoïde.

Le décodeur utilise le vecteur $\boldsymbol{h}$ pour construire $x' = \sigma'(W'h + b')$, ou les paramètres peuvent ou non être les mêmes que ceux de l'encodeur.

L'auto-encodeur construit les nouvelles versions des données en minimisant l'erreur de reconstruction donnée par :

$$e = \|x - x'\|^2.$$

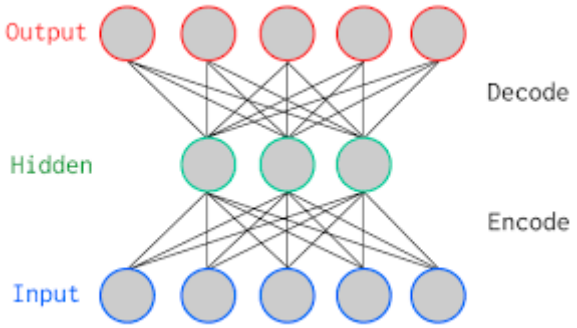

Figure 2-18 : Auto-encodeur

I.7 Evaluations des performances d'un classifieur

Une fois le modèle de classification créé, il est obligatoire de le valider avant de le mettre en application réelle. Pour cela, plusieurs critères sont utilisés, et le choix qui se fait parmi eux est fonction du problème traité.
Dans cette partie, nous allons évoquer les critères de performance en classification.

I.7.1 La matrice de confusion

La matrice de confusion est un tableau utilisé pour décrire les performances d'un classifieur sur un ensemble de données test. La matrice indique le nombre de prédictions correctes et incorrectes établies par le modèle de classification par rapport aux résultats réels (valeur cible) dans les données. Le tableau suivant affiche une matrice de confusion 2x2 pour deux classes (positive et négative) :

	Classe réelle (+)	**Classe réelle (-)**
Classe prédite (+)	TP	FP
Classe prédite (-)	FN	TN

Tableau 2-1 : Matrice de confusion

- **TP** (True Positive) : les cas où la prédiction est positive, et où la valeur réelle est positive.
- **TN** (True Negative) : les cas où la prédiction est négative, et où la valeur réelle est négative.
- **FP** (False Positive) : les cas où la prédiction est positive, mais où la valeur réelle est négative.
- **FN** (False Negative) : les cas où la prédiction est négative, mais où la valeur réelle est positive.

I.7.2 Les mesures de base dérivées de la matrice de confusion

À partir de la matrice de confusion, nous aboutissons aux mesures suivantes pour comparer les modèles :

- **L'Accuracy (Exactitude):** il s'agit de la proportion du nombre total de prédictions correctes qui est cependant peu descriptif lorsqu'elle est utilisée pour mesurer la performance d'un jeu de données très déséquilibré. Un modèle peut avoir des niveaux élevés d'exactitude, mais peut ne pas obtenir de hauts niveaux d'identification de la classe dont la prédiction nous intéresse. L'Accuracy est déterminé en utilisant l'équation :

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Dans le cas des multi-classes, la table de confusion est généralisée sur plusieurs classes. Si **M** est une matrice de confusion, l'Accuracy est alors :

$$Accuracy = \frac{\sum_i M_{ii}}{\sum_i \sum_j M_{ij}}$$

-
- **L'erreur de classification :** il s'agit du complément de *l'Accuracy*, qui évalue un classificateur par son pourcentage de prédictions incorrectes. L'erreur de classification est déterminée en utilisant l'équation :

$$\text{Erreur de classification } = \frac{FP + FN}{TP + TN + FN + FP} = 1 - Accuracy$$

Dans le cas des multi-classes, la table de confusion est généralisée sur plusieurs classes. Si **M** est une matrice de confusion, l'erreur de classification est alors :

$$\text{Erreur de classification} = \frac{\sum_{i \neq j} M_{ij}}{\sum_i \sum_j M_{ij}} = 1 - Accuracy$$

- **La sensibilité (Recall ou True positive rate):** il s'agit de la proportion de cas positifs correctement identifiés, calculée à l'aide de l'équation :

$$\text{Sensibilité} = \frac{TP}{TP + FN}$$

- **La spécificité (True negative rate) :** il s'agit de la proportion de cas négatifs correctement identifiés, calculée à l'aide de l'équation :

$$\text{Spécificité} = \frac{TN}{TN + FP}$$

- **Taux de faux positif (False positive rate)** : il s'agit de la proportion de cas négatifs classés incorrectement comme positifs, calculée à l'aide de l'équation:

$$\text{fpr} = \frac{FP}{TN + FP}$$

- **Taux de faux négatifs** (**False negative rate**) : il s'agit de la proportion de cas positifs classés incorrectement comme négatifs, calculée à l'aide de l'équation :

$$\text{fnr} = \frac{FN}{FN + TP}$$

- **Précision** : il s'agit de la proportion des cas positifs prédits qui étaient corrects, calculée à l'aide de l'équation:

$$\text{Précision} = \frac{TP}{TP + FP}$$

La précision et la sensibilité se font souvent au détriment l'une de l'autre. En d'autres termes, une précision élevée est obtenue au détriment de la sensibilité et une sensibilité élevée est obtenue au détriment de la précision. Un modèle idéal aurait à la fois une sensibilité élevée et une grande précision.

- **Le F-mesure** : c'est la mesure qui combine la précision et la sensibilité. Elle mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres. Elle est calculée à l'aide de l'équation:

$$\text{F} - \text{mesure} = \frac{2 \times \text{Précision} \times \text{Sensibilité}}{\text{Précision} + \text{Sensibilité}}$$

I.7.3 La courbe ROC (Receiver Operating Characteristic) (Tufféry, 2012)

La courbe ROC nous permet de visualiser le compromis entre spécificité et sensibilité en traçant l'évolution de cette dernière (taux de vrais positifs) en fonction de $\mathbf{1} - \boldsymbol{sp\acute{e}cificit\acute{e}}$ (taux de faux positifs) selon les valeurs d'un certain seuil ***s***.

La courbe va donc représenter sur l'axe des ordonnées la proportion d'événements déclarés positifs (par exemple frauduleux) car leur score est supérieur au seuil (cut-off) ***s***, en fonction de la proportion de non-événements déclarés positifs parce que leur score est inférieur au seuil ***s***.

Soit $\boldsymbol{x}$ un individu et soient les fonctions suivantes :

- La sensibilité $\alpha(s) = prob(score(x) \geq s \backslash x = \acute{e}v\acute{e}nement)$, qui implique de bien détecter un événement au seuil **s**.
- La spécificité $(s) = prob(score(x) \leq s \backslash x = non - \acute{e}v\acute{e}nement$, qui implique de savoir détecter un non-événement au seuil s. On dira alors que la proportion des non-événement déclarés comme évènement est $1 - \beta(s)$.

La courbe ROC représente donc $\alpha(s)$ en fonction de $1 - \beta(s)$ pour des valeurs de s allant du maximum où l'on considère tous les individus comme non-événement et donc :

$$\alpha(s) = 1 - \beta(s) = 0$$

Au minimum où l'on considère tous les individus comme événement et donc :

$$\alpha(s) = 1 - \beta(s) = 1$$

- Au point (0, 0) le classificateur déclare toujours non-événement : il n'y a aucun faux positif, mais également aucun vrai positif.
- Au point (1, 1) le classificateur déclare toujours évènement : il n'y a aucun vrai négatif, mais également aucun faux négatif.

Un classificateur aléatoire tracera une droite allant de (0, 0) à (1, 1).

- Au point (0, 1) le classificateur n'a aucun faux positif ni aucun faux négatif, et est par conséquent parfaitement exact, ne se trompant jamais.
- Au point (1, 0) le classificateur n'a aucun vrai négatif ni aucun vrai positif, et est par conséquent parfaitement inexact, se trompant toujours.

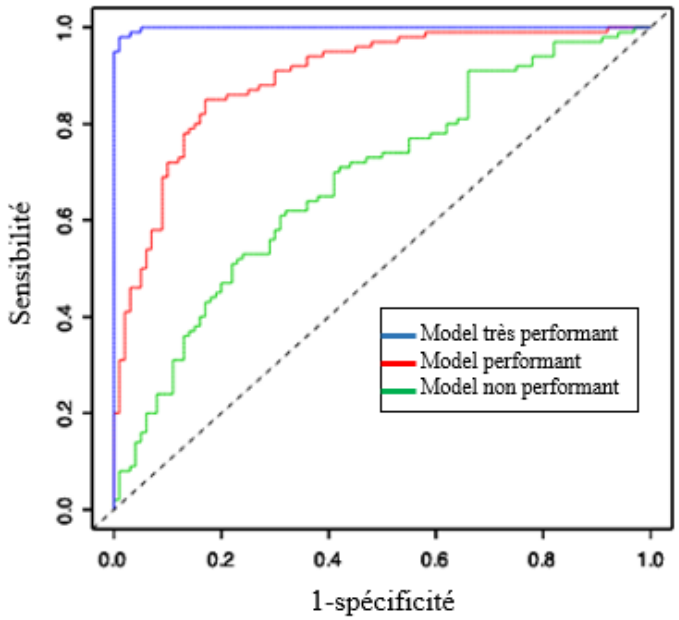

Figure 2-19 : Courbe ROC

I.7.4 ROC AUC (The area under the ROC curve)

Comme son nom l'indique, il s'agit de l'aire sous la courbe de ROC et un classifieur est d'autant plus performant que cette aire est proche de 1.

Plus précisément, cette aire est la probabilité que le score d'un individu **x** tiré aléatoirement de l'ensemble des individus libellés comme événement soit supérieur au score d'un individu **y** tiré aléatoirement de l'ensemble des individus libellés comme non-événement. Si l'air est égale à 1, cela veut dire que tous les scores des individus **x** sont supérieurs aux scores des individus **y**.

II Le Data Mining

Aujourd'hui, des quantités astronomiques de données sont collectées chaque jour, et grâce aux capacités de stockage offertes par l'informatique moderne, les entreprises peuvent en disposer dans de gigantesques bases de données. Ces données une fois stockées doivent être exploitées pour en extraire les informations stratégiques pour les entreprises et c'est là qu'intervient le Data Mining.

Il permet de générer de la connaissance en découvrant des modèles implicites dans ces données en se basant entre autre sur les méthodes de Machine Learning que nous avons décrites précédemment.

II.1 Définition du Data Mining

Définition 1 :

Le "Data Mining" que l'on peut traduire par "fouille de données" est une discipline apparue dans les années 90 aux Etats Unis. Elle se positionne à l'interface des sciences statistiques et des technologies de l'information telles que : les bases de données, l'intelligence artificielle et l'apprentissage automatique.

Définition 2 :

« Le Data Mining désigne l'ensemble des méthodes destinées à l'exploration et l'analyse de grandes bases de données informatiques, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles d'associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données permettant les prises de décision » (2017).

II.2 Les objectifs du Data Mining

Les objectifs du Data Mining sont regroupés en six catégories :

A. **La classification :** elle implique la construction d'une fonction qui, selon ses caractéristiques, affecte un individu dans une classe prédéterminée.

B. **La description :** elle consiste à découvrir les tendances cachées dans les données à travers des analyses exploratoires.

C. **La segmentation (clustering) :** elle sert à segmenter l'ensemble des données qui est hétérogène en sous-groupes relativement homogènes à l'aide de mesures de distances. Contrairement à la classification, ces sous-populations ne sont pas préétablies.

D. **L'estimation :** elle est similaire à la classification, sauf que la variable cible est numérique plutôt que catégorique. Autrement dit, le modèle élaboré va affecter à une nouvelle entrée une valeur numérique continue.

E. **La prévision :** elle est semblable à la classification et l'estimation, sauf que pour la prévision, les résultats se situent dans le futur.

F. **L'association :** elle consiste à rechercher les relations ou des dépendances existent entre plusieurs caractéristiques (variables) d'un individu.

II.3 Le processus de Data Mining (Kelleher, 2015)

L'élaboration de solutions pour la création de connaissance implique bien plus que le choix du bon algorithme d'apprentissage automatique. Comme tout autre projet important, les chances de succès d'un projet de Data Mining augmentent considérablement si un processus standard est utilisé pour le gérer tout au long de son cycle de vie. Et l'un des processus les plus couramment utilisés est le processus interprofessionnel standard d'exploration de données (CRISP-DM). Les principales caractéristiques qui le rendent si attractif pour les praticiens de l'analyse de données sont le fait qu'il soit non exclusif. En d'autres termes, il est adapté à tous types d'applications, et d'industries.

De plus, il offre une analyse explicite du processus de Data Mining d'un point de vue technique et applicatif, c'est pour cela que nous l'avons choisi pour en donner une description détaillée.

CRISP-DM (Cross Industry Standard Process for Data Mining) (Pacheco, 2015)

CRISP-DM est un processus de découverte de connaissance élaboré dans les années 1996. Il est applicable à tous les secteurs et découpe le processus de fouille de données en six étapes afin de le structurer.

Les phases principales de ce processus sont les suivantes :

2. **La compréhension du métier :** cette phase s'intéresse à la compréhension des objectifs et besoins pour les convertir en un problème de Data Mining. Elle se découpe elle-même en quatre étapes comme suit :

- La détermination des objectifs ;
- L'évaluation de la situation en termes de ressources (contraintes et hypothèses) ;
- La conversion des objectifs généraux en buts techniques à réaliser par les outils de Data Mining ;
- La production d'un plan de projet pour réaliser les objectifs qui comprend la description des étapes avec leurs durées ainsi que les ressources requises, les entrées et les sorties.

3. **La compréhension des données :** cette étape consiste à collecter les données, les visualiser pour vérifier leur qualité et déterminer un ordre de grandeur pour chaque variable, et éventuellement tester quelques hypothèses.

4. **La préparation des données :** cette étape consiste à sélectionner les variables pertinentes, les nettoyer et les transformer si nécessaire mais aussi créer d'autres variables en combinant certaines d'entre elles.

5. **La modélisation :** cette phase consiste à sélectionner les techniques de Data Mining qui seront utilisées pour réaliser l'objectif, les adapter selon les variables dont on dispose et les paramétrer pour aboutir à des résultats optimaux, mais aussi définir un protocole de test sur les modèles selon le problème étudié.

6. **L'évaluation :** cette étape consiste à vérifier si les résultats obtenus par le modèle sélectionné satisfont les objectifs déterminés au préalable, auquel cas l'étape suivante pourra être lancée. Si les résultats s'avèrent non satisfaisant une revue des étapes précédentes du processus est effectuée pour les améliorer.

7. **Le déploiement :** cette étape a pour but d'intégrer la connaissance obtenue au processus de prise de décision. La forme que peut prendre ce déploiement dépend des objectifs du projet. Cela peut aller de la rédaction d'un rapport décrivant la connaissance obtenue jusqu'à l'élaboration d'une application informatique qui permet d'utiliser le modèle élaboré pour réaliser des prévisions sur des données futures.

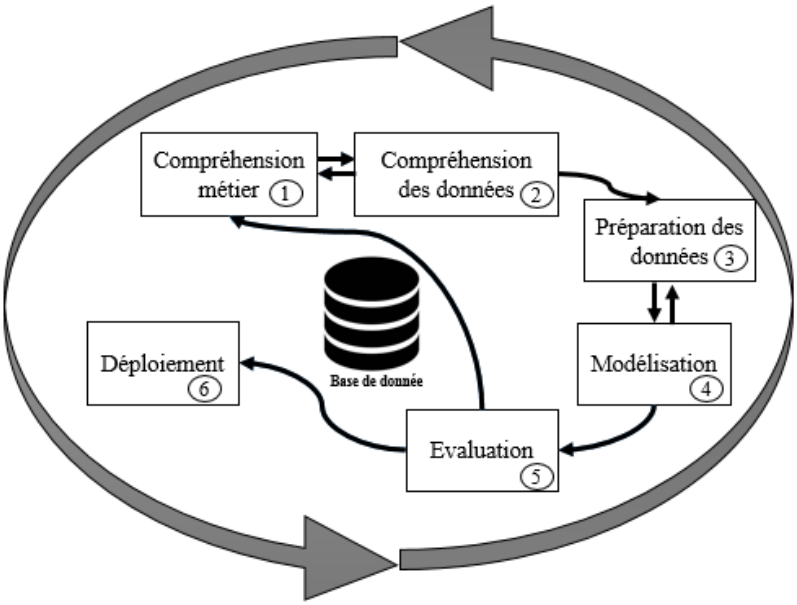

Figure 2-20 : Le modèle CRISP-DM

La ***Figure 2-20*** en plus d'illustrer les phases du processus CRISP-DM, met en avant le flux entre chacune des phases et souligne le fait que les données y sont au cœur. Certaines de ces phases sont plus étroitement liées que d'autres, telles que la compréhension du métier avec la compréhension des données, et la préparation des données avec la modélisation et beaucoup de temps est consacré à ces phases lors du projet.

II.4 Les types de variables

Une variable désigne toute caractéristique d'une entité qui peut être exprimée par une valeur numérique (mesure) ou codée (attribut). Elles peuvent prendre plusieurs types qui conditionnent les techniques utilisées dans le processus de Data Mining. En effet, les algorithmes de Machine Learning, comme nous avons pu le constater précédemment, ne peuvent pas tous intégrer n'importe quel type de variables.

Nous trouverons dans le tableau suivant une description des différents types de variables :

Types de variable	**Caractéristiques**
Disjonctives (binaire ou booléenne)	Elles peuvent prendre deux états (modalités), par exemple : vrai ou faux.
Catégoriques non ordonnées	Les différentes catégories ne contiennent pas de notion d'ordre, par exemple: la couleur des yeux.
Catégoriques ordonnées	Les différentes catégories peuvent être classées, par exemple: les tranches d'âges)
Continues	Elles prennent des valeurs numériques sur lesquelles des calculs, tels que la moyenne, peuvent être effectués.

Tableau 2-2 : Types de variable

III Le logiciel R

R est un langage de programmation et un logiciel gratuit (open source), multiplateforme (fonctionnant sous Windows et Linux) destiné aux statistiques et à la data science. Il est à la fois un langage informatique et un environnement de travail tel que les commandes sont exécutées grâce à des instructions codées dans un langage relativement simple. Les résultats sont affichés sous forme de texte et les graphiques sont visualisés directement dans une fenêtre qui leur est propre.

Nous avons choisi de travailler avec Rstudio qui est une application proposant un environnement de développement et des outils adaptés au langage et à l'environnement de programmation R. Ce qui nous facilite le développement de nos algorithmes. Il intègre :

- Une fenêtre d'affichage des fichiers sources où l'on peut notamment écrire nos scripts, et sélectionner les parties à exécuter avec Ctrl-Enter.
- Une console où l'on écrit nos lignes de commandes directement, puis nous appuyons sur la touche Enter pour exécuter chaque ligne écrite. C'est aussi à cet endroit que s'affichent les éventuels messages d'erreur.
- Une fenêtre environnement où s'affiche la liste des objets que nous avons créés. Ces objets peuvent être de simples variables, des tableaux ou des fonctions.
- Une fenêtre multi onglets qui nous sert à naviguer dans le système de fichiers de notre ordinateur, afficher les graphiques que l'on a construit et les paquets disponibles et chargés (lorsque cochés).

Les avantages du logiciel R

- Le langage est basé sur la notion de vecteur, ce qui simplifie les calculs mathématiques et réduit considérablement le recours aux structures itératives (boucles for, while, etc.) ;
- Il ne nécessite pas que l'on définisse le type de variable ;
- Les programmes sont courts, en général quelques lignes de code seulement suffisent;
- Il permet l'utilisation des méthodes statistiques classiques à l'aide de fonctions prédéfinies ;
- Il permet de créer ses propres programmes dans un langage de programmation intuitif et simple d'utilisation (proche de Matlab),
- Il permet d'utiliser des techniques statistiques innovantes et récentes à l'aide de package en développement permanent par la communaut des utilisateurs de R et disponibles sur le site du CRAN (http://cran.r-project.org/).

Conclusion :

Ce chapitre, en plus d'introduire le domaine de l'intelligence artificielle, nous a permis d'expliciter les étapes de la démarche de Data Mining « CRISP-DM », car nous avons décidé de nous en servir comme base de notre méthodologie dans le déploiement de notre solution.

Il nous a aussi permis de faire le tour des algorithmes de Machines Learning, de les détailler mathématiquement tout en vulgarisant leurs principes de fonctionnement afin de faciliter leurs compréhensions.

Nous y avons décrit les méthodes utilisées pour l'évaluation des algorithmes de classifications supervisées, dont nous nous servirons dans la suite de notre étude pour faire un arbitrage entre les modèles que nous aurons construits et testés.

Nous avons clôturé ce chapitre par la présentation du logiciel que nous utiliserons pour la réalisation de notre solution.

Chapitre 03 : Solution proposée et son application

Chapitre 3: Solution proposée et son application

Introduction :

Dans cette partie, nous allons répondre à la problématique formulée au premier chapitre. Pour cela, nous allons suivre les étapes de la méthodologie CRISP-DM, car elle nous permet de structurer le travail d'extraction de l'information et de bien expliquer chaque étape du travail que nous avons effectué.

I La compréhension du métier

Cette phase du processus CRISP-DM comprend les éléments suivants :

- **La détermination de l'objectif :**

L'objectif que l'on poursuit en suivant une procédure de Data Mining est d'aider la douane à concilier entre les deux missions antagonistes dont elle est chargée et qui sont d'une part le contrôle douanier pour protéger le consommateur et l'intérêt économique du pays et, d'autre part, les facilitations pour favoriser le développement du commerce extérieur et la réduction des coûts du contrôle douanier.

- **Contraintes et hypothèses :**

- Nous pouvons considérer que nous disposons d'une base de données déséquilibrée, car elle ne contient que 0,4% de déclarations frauduleuses seulement.

- Nous savons, notamment, d'après les informations recueillis par les experts, que de nombreuses déclarations frauduleuses ne sont pas détectées par les agents du contrôle douanier en raison notamment de la surcharge au niveau du circuit rouge ce qui signifie que notre base est entachée d'erreurs.

- Nous savons aussi que cette base n'est pas mise à jour lorsque des contrôles à postériori sont opérés sur les déclarations du circuit vert notamment ; il se pourrait donc que des déclarations soient faussement déclarées comme non frauduleuses (fraudes non déclarées dans le système).

- Les opérateurs économiques agréés sont affectés d'office aux circuits verts.

- **Détermination des objectifs techniques :**

L'objectif technique consiste à développer un outil basé sur l'intelligence artificielle afin d'aider l'administration des douanes à mieux orienter les dossiers d'importation vers les différents circuits de vérification.

Pour cela, des algorithmes de Machine Learning seront développés et testés en utilisant la base de données fournies par la douane contenant les déclarations avec l'ensemble des informations qui les caractérisent.

-

- **Plan de projet :**

Le plan que nous allons suivre correspond à la mise en œuvre des étapes énoncées dans la méthodologie CRISP-DM.

Nous allons donc commencer par importer notre base de données. Nous la nettoierons puis tacherons de sélectionner des variables pertinentes tout en en créant d'autres afin d'alimenter les modèles que nous construirons et testerons.

Nous procéderons ensuite à la sélection du modèle dont les performances sont globalement les plus satisfaisantes, puis nous l'adapterons selon les besoins de la problématique.

Afin de réaliser cela, nous avons décidé d'utiliser le logiciel R car il contient plusieurs bibliothèques qui nous faciliteront la manipulation de nos données et la création de nos modèles.

Enfin, ce processus aura pour entrée une base de données et pour sortie un modèle qui affecte les déclarations en douane aux circuits de vérifications qui leur correspondent pour satisfaire les objectifs de la Direction Générale des Douanes.

Le déroulement de cette démarche, ainsi que les différents outils statistiques et algorithmiques pouvant être déployés à chaque étape est illustrée ci-dessous :

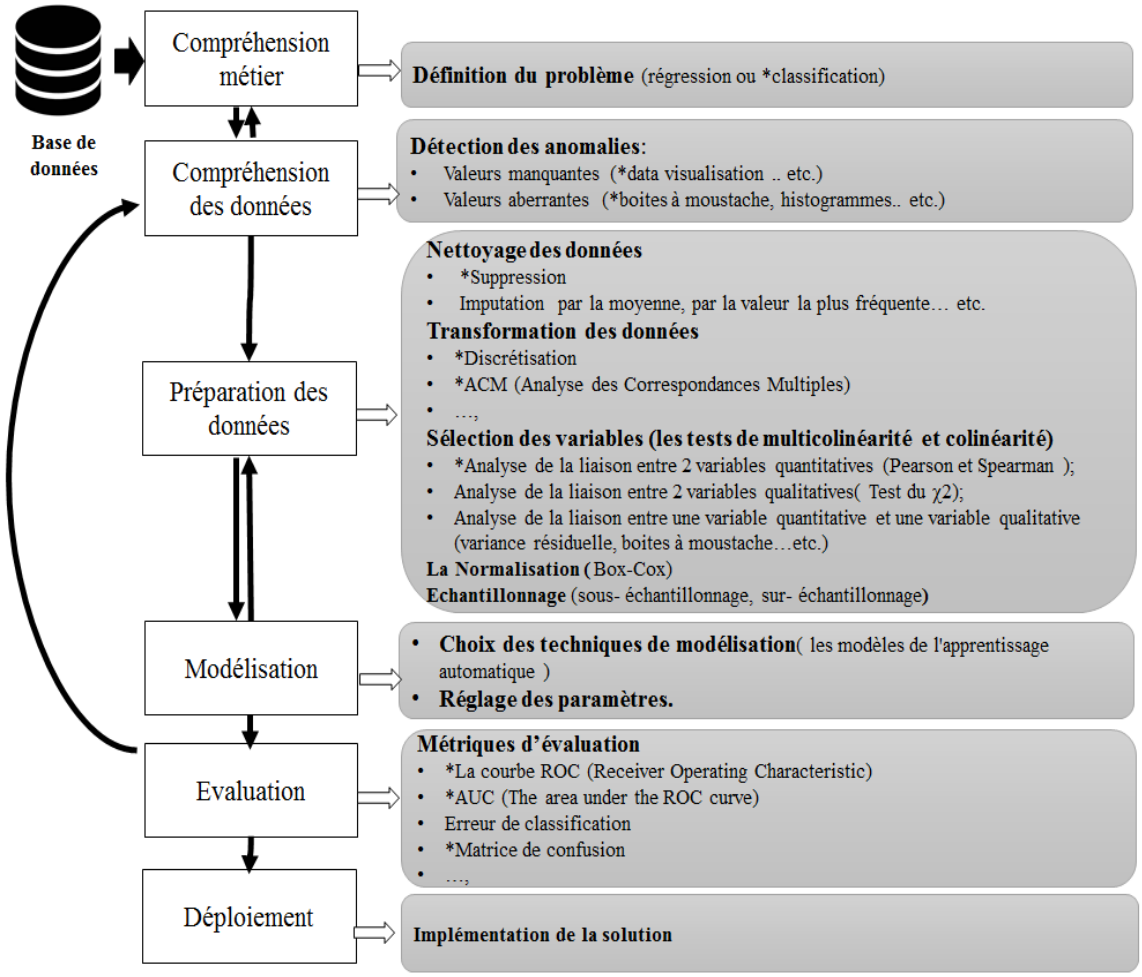

Figure 3-1 : Etapes et outils de la méthodologie CRISP-DM

II La compréhension des données

Nous avons récupéré les données concernant les déclarations des 2 années 2016 et 2017 (*voir **Annexe B*** au niveau du Centre National des Transmissions et Systèmes Informatiques des Douanes (CNTSID) sous format informatique DBF. Nous les avons importées puis concaténées avec le logiciel R et nous avons obtenu une base de données contenant plus de 5 millions d'enregistrements. Chaque ligne représente l'importation d'un produit et chaque colonne contient une information particulière relative à cette importation, notamment si elle est frauduleuse ou non (telle qu'identifiée par l'administration des douanes).

Avec la collaboration de l'équipe de la direction de la gestion des risques et des renseignements, nous avons sélectionné certaines de ces données en colonne pour constituer les variables explicatives brutes que nous allons traiter, exploiter puis étoffer afin de développer nos modèles prédictifs. Ces variables se présentent comme suit :

- **Code bureau :** il correspond au bureau de douane qui est le service de l'administration des douanes où sont accomplies les formalités douanières, notamment les actes de constatation, de liquidation, de contrôle documentaire et de recouvrement des droits et taxes et pénalités de toute nature conformément à la législation en vigueur.
- **Date de la déclaration :** elle correspond à la date de validation de la déclaration de l'importateur.
- **Code régime :** il correspond au régime douanier économique qui est défini par l'OMD comme étant le traitement applicable par la douane aux marchandises assujetties au contrôle douanier.
- **Type d'opération :** elle indique l'utilisation qui se fera de la marchandise (revente en l'état, transformation de matière première…)
- **La banque domiciliataire :** c'est la banque choisie par l'importateur pour assurer le règlement des opérations financières.
- **Type d'incoterm :** il s'agit du type d'incoterm adopté par l'importateur.
- **La position tarifaire :** elle correspond à la dénomination attribuée à une marchandise dans le tarif des douanes, à partir de laquelle dépend le taux de base des droits qui sont prélevés.
- **Le pays de provenance :** il s'agit du dernier pays par lequel la marchandise a transité avant d'arriver en Algérie.
- **Le pays fournisseur :** il s'agit du pays avec lequel l'importateur effectue la transaction (le pays où va l'argent).
- **Le pays d'origine :** il s'agit du pays où a été fabriquée la marchandise.
- **Le taux de droit commun :** il s'agit de l'impôt prélevé sur une marchandise importée lors de son passage à la frontière.
- **Le taux appliqué :** il s'agit du taux effectivement appliqué sur la marchandise importée. Ce taux peut être inférieur au taux de droit commun si la marchandise a des origines préférentielles, sinon les deux taux sont égaux.
- **Le code fiscal :** il s'agit du code fiscal de l'importateur.
- **Code postal :** le code postal de l'opérateur.
- **Fraude :** il s'agit d'une variable binaire, telle qu'une valeur de 1 signifie que la déclaration est frauduleuse.

Les données brutes étant très souvent sujettes à des valeurs manquantes, aberrantes et incomplètes. Il est primordial de les analyser afin de détecter ces anomalies, avant de s'engager plus loin dans un processus de Data Mining, car une qualité médiocre de données aboutirait à la construction de modèles prédictifs peu robustes. Et afin de détecter ces particularités, nous avons procédé comme suit :

- **Pour les valeurs manquantes :**

Nous avons tout d'abord remplacé les cellules vides de notre base de données par des valeurs nulles pour qu'elles puissent être détectées par la fonction « vis_miss » du logiciel R. Cette fonction nous permet de visualiser la proportion de valeurs manquantes pour chaque variable. Nous avons obtenu le graphique suivant :

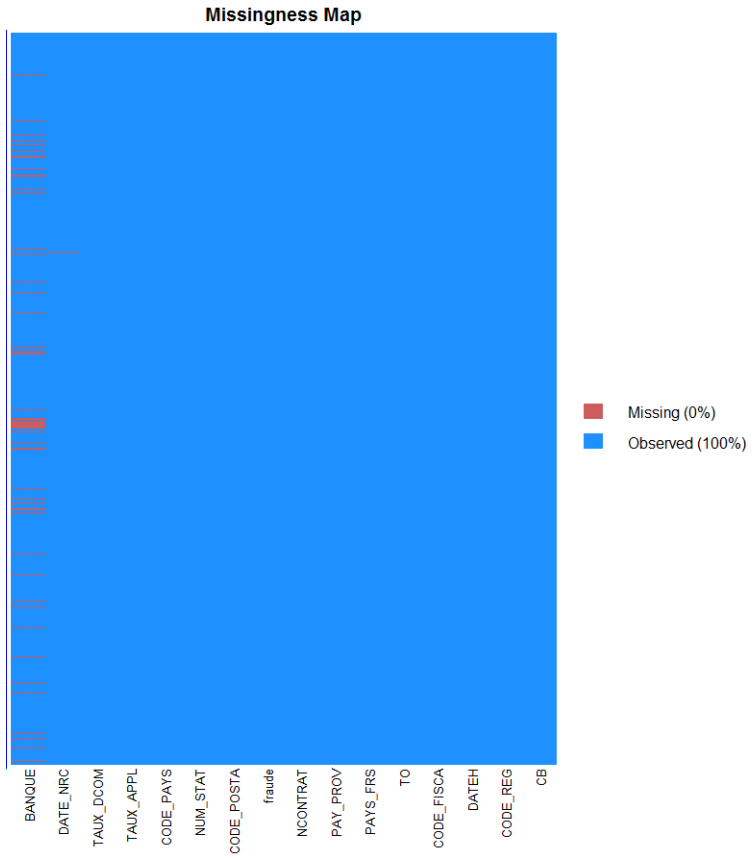

Figure 3-2 : Les valeurs manquantes

Ce graphique classe les variables selon le nombre de valeurs manquantes dans l'ordre croissant. Nous pouvons constater qu'il n'y que la variable « Banque » qui présente des valeurs manquantes (en rouge) et que le pourcentage total de ces valeurs avoisine les 0%.

- **Pour les variables aberrantes :**

Une valeur aberrante est une valeur erronée correspondant à une mauvaise mesure, une erreur de saisie ou une fausse déclaration. Et si les valeurs extrêmes ne sont pas toujours aberrantes, ces dernières ne sont pas extrêmes non plus : ceci rend leur détection plus délicate et peut nécessiter une bonne connaissance des données.

Afin de détecter les aberrances dans les valeurs numériques, nous avons utilisé la boite à moustache (voir ***Annexe C***). Nous avons pu remarquer par exemple des enregistrements où le taux appliqué était supérieur au taux de droit commun, ce qui constitue une anomalie étant donné que le taux appliqué doit être soit inférieur au taux de droit commun ou lui être égal. Afin de déceler cela, nous avons visualisé grâce à la boite à moustache la différence entre ces deux taux. Les valeurs négatives et celles se situant au-delà d'un seuil sont considérées comme aberrantes. Le résultat obtenu est présenté ci-après :

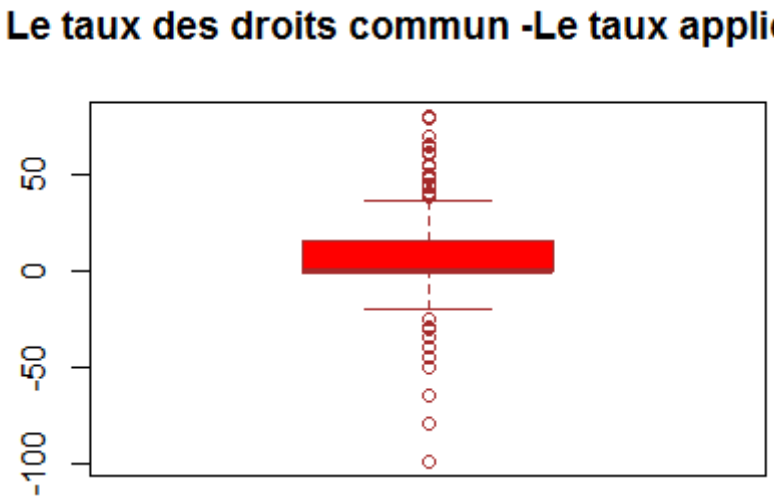

Figure 3-3 : Boite à moustaches pour la comparaison des taux

Pour les variables catégorielles, l'examen de la distribution des données ne nous a pas fourni de résultats concluants. Nous avons dû examiner tous les types de valeurs que chaque variable catégorielle prend et nous avons remarqué par exemple les anomalies suivantes dans les codes postaux :

DATEH	DATEI	DATE_QUIT	CODE_POSTA	TO2	AGENCE	BANQUE	CODE_BNQ
2018-12-27 14:11		08/01/2019	5000	3	803	161	161803
2018-05-24 15:36	2018-05-31 13:19	05/06/2018	g0900	2	504	160	160504
2018-09-18 13:51	2018-09-24 14:19	26/09/2018	A1000	3	201	161	161201
2018-10-28 13:43	2018-10-31 08:49	04/11/2018	KHRA0	1	201	162	162201
2018-10-28 13:43	2018-10-31 08:49	04/11/2018	KHRA0	1	201	162	162201
2018-10-28 13:43	2018-10-31 08:49	04/11/2018	KHRA0	1	201	162	162201
2018-10-29 11:42	2018-11-05 13:11	06/11/2018	KHRA0	1	216	160	160216
2018-10-29 11:42	2018-11-05 13:11	06/11/2018	KHRA0	1	216	160	160216

Figure 3-4 : Anomalies dans les codes postaux

Nous savons que le code postal est composé de cinq chiffres, et nous avons identifié dans les données, des codes postaux constitués de lettres, ce qui représente une anomalie.

III La préparation des données

Cette étape du processus est la plus longue et la plus délicate, car pour que les modèles que nous développerons aboutissent à de bonnes performances, il est primordial de disposer d'une base de données de bonne qualité. Et pour y parvenir, il s'agira de suivre les étapes suivantes :

a) La correction des anomalies

Une fois les anomalies détectées, nous avons entrepris de les corriger comme suit :

En ce qui concerne le traitement des valeurs manquantes, plusieurs options s'offrent à nous :

- Ne pas utiliser les variables concernées dans la construction de nos modèles.
- Imputer les valeurs manquantes par des valeurs déterminées statistiquement, grâce à la connaissance des données ou par une source externe de données.
- Supprimer les enregistrements contenants les valeurs manquantes.

Etant donné que les données manquantes représentent quasiment 0% du volume de données dont nous disposons, nous avons jugé préférable d'appliquer la 3ème option.

En ce qui concerne le traitement des valeurs aberrantes, nous avons décidé de les traiter au cas par cas, car elles n'étaient pas nombreuses. Nous avons par exemple, éliminé les enregistrements contenant des valeurs de taux de droit commun et de taux appliqué négatif, et nous avons corrigé l'erreur dans les codes postaux, en extrayant ses derniers du code fiscal de l'importateur au lieu d'utiliser la colonne qui lui est dédiée.

b) La redéfinition de l'enregistrement

Nous avons ensuite groupé les importations par déclaration, et nous avons donc réduit le nombre d'enregistrements à 500.000 occurrences (une déclaration contenant en moyenne 10 produits importées). Cependant, en agrégeant ainsi les données, nous perdons des informations spécifiques aux produits suivants : le pays d'origine, la position tarifaire, le taux de droit commun et le taux appliqué.

Pour capter une partie des informations perdues, nous avons créé certaines variables qui expriment dans la mesure du possible ces informations. Nous n'en citerons que quelques-unes pour des raisons de confidentialité :

- ***DIF_CODE_PAYS :*** elle indique le nombre de pays d'origine différents dans une déclaration- car rappelons-le, une déclaration contient plusieurs marchandises importées et ces dernières peuvent être fabriquées dans différents pays.
- ***ONE_CODE_PAYS :*** si toute la marchandise a la même origine, cette variable prend comme valeur le code du pays d'origine, sinon elle prend une valeur de ''0000''.
- ***DIF_NUM_STAT :*** elle indique le nombre de types d'articles différents dans une déclaration.
- ***TOT_ART :*** elle indique le nombre d'articles importés.
- ***MIN_TAUX_APPLIQUEE :*** elle indique le minimum de taux appliqués sur la marchandise.

c) Le regroupement des modalités des variables catégorielles

Avant de pouvoir tester le pouvoir discriminant des variables, nous avons dû réduire la dimension des variables catégorielles qui présentaient un nombre trop élevé de modalités et qui de ce fait ne peuvent être intégrées dans les modèles prédictifs. En effet, ces modèles ne peuvent inclure les variables dont le nombre de modalité dépasse un certain seuil (53 modalités).

Pour les pays fournisseurs, d'origine et de provenance, nous avons choisi de calculer la fréquence d'apparition de chaque pays dans les importations. Chaque pays s'est donc vu attribuer trois (3) valeurs de fréquences avec :

$$fréquence = \frac{\sum_i importation\ par\ pays_i}{\sum des\ importations}.$$

De cette manière, les variables pays fournisseur, pays de provenance et pays d'origine sont transformées de variables catégoriques en variables continues.

Nous avons procédé ainsi pour pouvoir utiliser un algorithme de clustering (car nous n'avons pas de variable cible et donc pas d'étiquettes), afin de réduire le nombre de modalités d'origine de ces variables. Ces algorithmes ne peuvent intégrer que des valeurs continues, nous avons donc été amenées à remplacer les codes des pays par une grandeur représentative, et avec l'accord des experts de la DGD nous avons choisi la fréquence définie ci-dessus.

Les trois (3) valeurs de fréquences obtenues ont été introduites dans un algorithme de classification ascendante hiérarchique (CAH) afin d'obtenir des clusters homogènes. La visualisation de la variance intra-classe présentée dans la ***Figure 3-5*** nous a permis de choisir le nombre de clusters. En effet, la cassure au niveau de la courbe qui représente cette variance nous indique quel nombre optimal choisir pour le nombre de cluster. Dans notre cas, nous retiendrons huit (8) clusters. Il s'agit du nombre au delà duquel une partition supplémentaire n'améliorait que très peu la variance intra-classe.

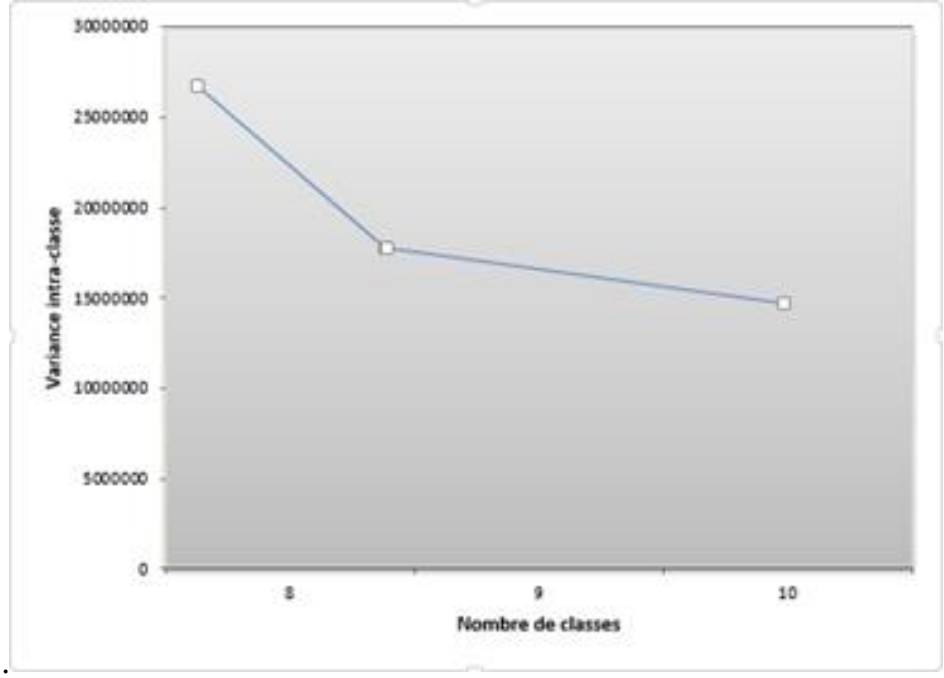

Figure 3-5 : Courbe de la variance intra-classe

Le dendrogramme présenté dans la ***Figure 3-6*** suivante illustre les 8 clusters retenus. Il s'agit de la représentation graphique d'une classification ascendante hiérarchique sous forme d'un arbre binaire. L'affectation des pays à chaque classe est présentée en ***Annexe D .***

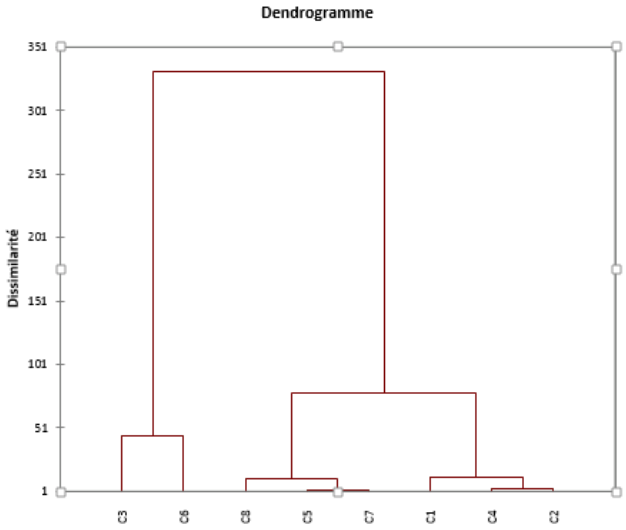

Figure 3-6 : Dendrogramme résultant de la CAH

Le tableau suivant récapitule les transformations que l'on a effectuées sur les pays :

Type de départ	Type en transition	Type final
Variable catégorielle possédant 300 modalités.	Variables continues allant de 0 à 1.	Variable catégorielle possédant 8 modalités.

Tableau 3-1: Transformation subies par la variable pays

D'autres modifications ont été apportées sur les variables pour réduire le nombre de modalités mais qui n'ont pas nécessité une technique statistique particulière. Nous avons juste modifié le niveau d'agrégation de nos données.

Ainsi, nous avons choisi de travailler avec les sections des produits présentées dans ***l'Annexe A*** au lieu des positions tarifaires. En sachant que chaque section regroupe plusieurs chapitres incluant à leurs tours plusieurs positions tarifaires. Nous travaillerons donc au final avec une variable ayant seulement 28 modalités.

Nous avons procédé de la même manière avec la variable « codes banque ». Nous disposions au départ d'un code comprenant six (6) caractères. Exemple : 162702.

Le 16 représente la Wilaya où se situe l'agence bancaire, le 27 représente la banque et le 02 représente l'agence de la banque en question.

Nous avons choisi de travailler avec la partie du code représentant le nom de la banque au lieu du code entier. Dans l'exemple qui précède il s'agit du numéro 27.

d) La création de nouvelles variables

Après ces modifications, nous avons visualisé la relation entre la fraude constatée et chaque variable à l'aide de diagrammes en barres. Pour des raisons de concision, nous ne présenterons qu'un seul diagramme, le reste étant présenté dans ***l'Annexe E*** .

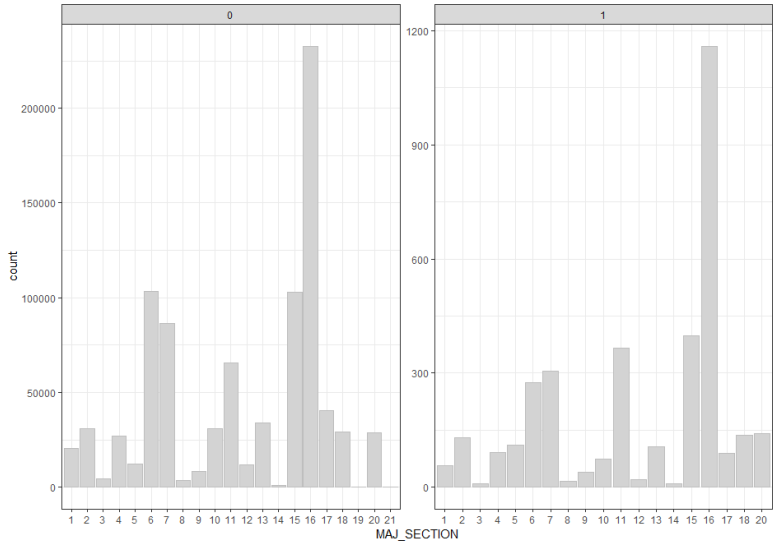

Figure 3-7 : La relation entre la variable section majoritaire et la variable cible fraude

A partir du diagramme présenté dans la ***Figure 3-7*** nous avons constaté que la variable représentant la section majoritaire dans la déclaration n'explique pas vraiment la fraude. En d'autres termes, on ne remarque pas un comportement particulier dans le cas de la fraude qui n'est pas présent dans le cas de la non fraude. C'est aussi le cas pour le reste des variables mais à des degrés différents.

Pour pallier à cela, nous avons décidé de créer de nouvelles variables en combinant celles dont nous disposons. Pour cela, nous avons mené plusieurs entretiens avec les membres du département des renseignements de la DGD et nous avons pu, grâce à leur expertise du domaine, proposer des variables qui expriment certains comportements du fraudeur. Cependant, pour des raisons de confidentialité, nous ne présenterons que quelques-unes de ces variables :

- Nous avons comparé le taux appliqué avec le taux de droit commun, et nous avons créé une variable binaire qui prend la valeur 1 si le taux appliqué est inférieur au taux de droit commun ou 0 si ils sont égaux. Car deux des fraudes définies dans le ***Chapitre 1:II.3.3***, à savoir la majoration et la minoration peuvent être détectées en comparant ces deux taux. En effet, si l'importateur à un taux avantageux, il aura tendance à majorer le taux en faisant de fausses déclarations quant à la marchandise qu'il aura importée pour pouvoir s'adonner à la fuite de devises. Dans le cas contraire, il voudra au contraire minorer le taux pour avoir un montant en droits et taxes moins important à payer.

- Nous avons modélisé le comportement de l'importateur qui préfère, pour des raisons inconnues, déclarer sa marchandise dans un bureau de douane qui se situe dans une région éloignée de celle où il réside. Pour cela, nous avons divisé l'Algérie en cinq régions : Nord, Est, Ouest, Sud-est et Sud-ouest. L'affectation de chaque Wilaya aux régions est présentée dans ***l'Annexe F .*** Nous avons ensuite créé la variable qui modélise ce comportement selon les cas présentés dans le tableau suivant :

Région de l'importateur	Région du bureau de douane	Valeur de la variable
Nord	Sud-est	1
Nord	Sud-ouest	1
Nord	Est	1
Nord	Ouest	1
Est	Ouest	1
Ouest	Est	1
Est	Sud-ouest	1
Ouest	Sud-est	1

Tableau 3-2 : Découpage de l'Algérie en cinq zones géographies

Pour le reste des cas qui ne sont pas cités dans le ***Tableau 3-2,*** la variable prend la valeur de 0.

- Nous avons créé une variable qui indique si le continent fournisseur ainsi que le continent d'origine sont les mêmes, mais que la marchandise provient d'un autre continent. Cette information est particulièrement pertinente lorsque la marchandise provient d'Asie, mais que l'argent va en Europe. Cela indique une possibilité de tentative de fuite de devises est opérée en faisant de fausses déclarations concernant l'origine de la marchandise.

- Nous avons modélisé le changement de comportement de l'importateur en introduisant un ensemble de variables qui compare l'état actuel des différentes caractéristiques des importations qu'il effectue à l'état ***n*** avec celles qu'il effectuait à l'état ***n-1***. Par exemple si l'importateur change de type d'opération entre deux déclarations la variable prendra la valeur 1 ; si ce n'est pas le cas la valeur prendra la valeur 0.

e) L'encodage des variables catégorielles

Certains modèles prédictifs sont basés sur des notions de distance entre variables. De ce fait ils ne peuvent admettre les variables catégorielles. Nous les avons donc transformés en utilisant la méthode One Hot Encoding.

Elle consiste à éclater la variable catégorielle en autant de variables binaires qu'elle possède de modalités. La valeur qu'elle prend à chaque enregistrement est traduite par une valeur de 1 prise par la variable binaire représentant la modalité qui lui correspond ; le reste des variables représentants les autres modalités quant à elles prennent la valeur de 0.

Exemple :

La variable qui représente les régions de l'Algérie comme nous l'avons décrit ci-dessus comporte cinq (5) modalités. Ces modalités sont éclatées alors en cinq (5) variables représentant chacune une modalité. La figure ci-après illustre cette transformation.

Numéro déclaration	Région
1223	3

Numéro déclaration	Région 1	Région 2	Région 3	Région 4	Région 5
1223	0	0	1	0	0

Figure 3-8 : Transformation One-Hot-Encoding

f) La corrélation entre les variables

Maintenant que nous disposons de variables ayant toute le même format (valeurs continues), nous pouvons tester la corrélation entre elles en calculant la corrélation de Pearson entre chacune des variables prises deux à deux. Ce test nous permet de déterminer si deux variables évoluent dans le même sens, c'est-à-dire si à des valeurs fortes de l'une sont associées des valeurs fortes de l'autre (corrélation positives), ou bien si a des valeurs fortes de l'une sont associées des valeurs faibles de l'autres (corrélation négative), ou encore si les deux valeurs sont indépendantes. Dans le cas où deux variables sont dépendantes, il serait judicieux de n'en garder qu'une seule, car toutes deux apportent presque la même information.

Les résultats du test de Pearson (***voir Annexe G***) sont présentés dans une matrice de dimension 146×146 présentée en ***Annexe H*** , 146 étant le nombre de variables dont nous disposons.

Pour chaque couple de variable, nous avons comparé leur corrélation deux à deux et avons éliminé l'une des variables du couple présentant une forte corrélation.

Après avoir effectué ce tri, il nous reste 120 variables.

g) L'échantillonnage

Pour pouvoir construire nos modèles prédictifs, il a d'abord été nécessaire de constituer un échantillon à partir de notre base de données car celle-ci est déséquilibrée En effet, la fraude ne représentant que 0,4% de l'ensemble des données ce qui est équivalent à 2621 fraudes. Les déclarations non frauduleuses quant à elles sont au nombre de 569258.

Cette particularité implique que les modèles prédictifs, afin de converger rapidement ne prennent pas en compte ces cas minoritaires et déclarent toute déclaration non frauduleuse afin de maximiser l'indicateur utilisé pour juger de la fiabilité du modèle (le modèle réussira alors à classer correctement 99.6% des déclarations).

Pour éviter ce problème, plusieurs méthodes d'échantillonnage sont envisagées, parmi lesquelles :

- **Le sous échantillonnage :** il équilibre l'ensemble de données en réduisant la taille de la classe présente en majorité et en n'y collectant qu'un échantillon aléatoire de données tout en conservant toutes les occurrences de la classe rare.

- **Le sur-échantillonnage :** dans ce cas, il s'agira de maintenir les occurrences de la classe présente en majorité tout en augmentant le nombre d'occurrences du cas rare en répétant aléatoirement les cas déjà inclus dans l'échantillon.

Si nous décidions d'appliquer la première méthode, nous obtiendrons un échantillon de taille $2621 \times 2 = 5242$. Car il faudrait prendre un nombre de déclarations non frauduleuses égal au nombre de déclarations frauduleuses dont nous disposons. Cette échantillon est alors trop petit étant donnée la taille de notre base de données.

Si nous décidions d'appliquer la deuxième méthode, il faudrait alors prendre les 569258 déclarations non frauduleuses et prendre autant de déclarations frauduleuses en les reproduisant aléatoirement pour atteindre ce nombre. Nous nous retrouverions avec un échantillon de $569258 \times 2 = 1138516$ déclarations. Cet échantillon est alors trop important pour pouvoir s'en servir dans le développement de nos algorithmes (nécessité de machines robustes).

En ce qui nous concerne, nous avons choisi une solution intermédiaire appelée méthode hybride. Nous avons fait varier la proportion de la fraude dans l'échantillon entre 5 et 25%, et nous avons visualisé le résultat (matrice de confusion) qu'on obtiendrait avec les classifieurs que nous avons développé pour chaque valeur prise de cette proportion. Et nous avons constaté que nous obtenons de meilleurs résultats avec une valeur de 10%.

Une fois notre échantillon prêt, nous l'avons divisé en deux parties. 80% des données seront consacrées à l'entrainement et le reste au test.

IV La modélisation

Nous allons décrire dans ce qui suit les différents algorithmes que nous avons utilisé lors de la construction des modèles prédictifs.

IV.1 L'auto-encodeur

Pour pouvoir répondre à la problématique de la douane**, il** faudrait tout d'abord concevoir un modèle capable de discerner les déclarations frauduleuses des autres déclarations.

Nous avons effectué plusieurs recherches sur les méthodes utilisées dans la détection de la fraude et avons pu constater une utilisation fréquente des auto-encodeurs, notamment en ce qui concernait le domaine de la fraude bancaire et la fraude à l'assurance. Nous avons alors décidé de l'appliquer sur nos données.

Lors de son apprentissage, cet algorithme cherche à reconstruire l'entrée d'origine avec le minimum de pertes d'informations. Une fois le modèle formé, les données peuvent être compressées en utilisant uniquement le composant encodeur de l'auto-codeur qui réduit la dimension des données et c'est pour cela que cet algorithme est principalement utilisé. Néanmoins, une autre utilisation est envisagée pour cet algorithme dans la détection de la fraude. Pour cela, la procédure suivante est appliquée :

a) Diviser les données en 3 groupes :

- L'ensemble d'apprentissage qui contient deux tiers des données sur les déclarations non frauduleuses.
- L'ensemble de test qui représente un dixième de l'ensemble d'apprentissage.
- L'ensemble de validation contenant le tiers restant des données concernant les déclarations non frauduleuses ainsi que toutes les déclarations frauduleuses.
-

b) Entrainer le modèle avec l'ensemble d'apprentissage, et le tester à chaque itération avec l'ensemble test jusqu'à ce que l'erreur de reconstruction converge.

c) Utiliser l'ensemble de validation pour déterminer la valeur du seuil K de l'erreur de reconstruction au-delà de laquelle une déclaration sera considérée frauduleuse.

Une fois le modèle prêt, il pourra être utilisé sur des nouvelles déclarations pour trouver les déclarations suspicieuses de la manière suivante :

- Entrer la nouvelle déclaration $\boldsymbol{x_k}$ dans l'auto-encodeur. Une reproduction de la donnée est récupérée au niveau de sa couche de sortie.
- L'erreur de reconstruction $\boldsymbol{\varepsilon_k}$ est calculée entre l'entrée d'origine et celle construite par l'auto encodeur.
- Pour décider si la déclaration est frauduleuse ou non, il faut utiliser la réglé suivante :

$$\begin{cases} x_k \text{ est frauduleuse} & \text{si } \varepsilon_k > K. \\ x_k \text{ est non frauduleuse} & \text{sinon.} \end{cases}$$

La transaction frauduleuse est alors considérée comme étant une anomalie étant donnée son erreur de reconstruction.

La figure suivante illustre comment les différentes bases de données du modèle sont constituées :

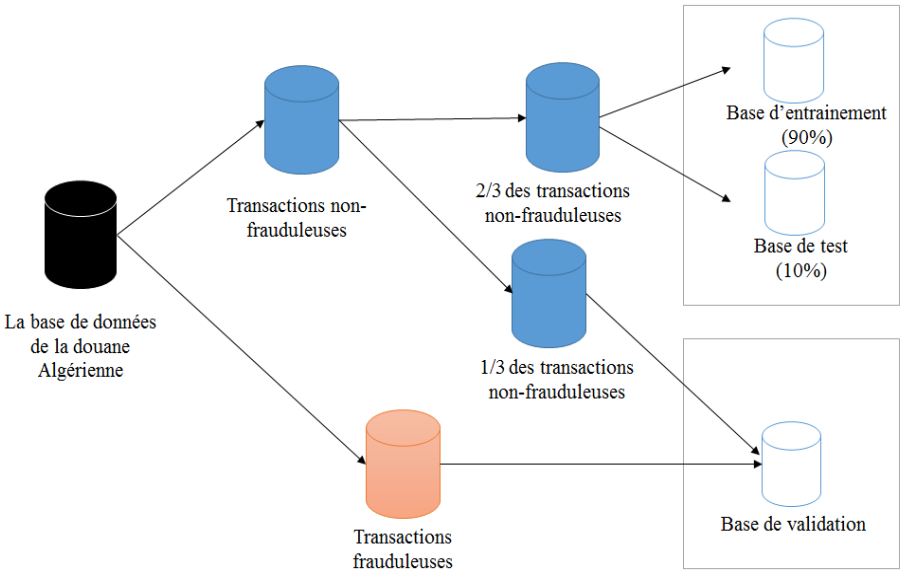

Figure 3-9: Constitution des bases de données de l'auto-encodeur

Afin de pouvoir réaliser les étapes décrites ci-dessous, nous avons construit l'algorithme suivant :

```r
library(keras)
library(caret)
library(dplyr)
library(pROC)
library(ROCR)
DataDouane_ae=read.csv(file.choose(),header = T,sep = ",")
DataDouane_ae=DataDouane_ae[,-1]
#normalization_minmax
DataDouane_ae$DIF_CODE_PAYS <- scales::rescale(DataDouane_ae$DIF_CODE_PAYS,to=c(0,1))
DataDouane_ae$SUM_DIF_PAYS <- scales::rescale(DataDouane_ae$SUM_DIF_PAYS,to=c(0,1))
DataDouane_ae$SUM_DIF_CONTI <- scales::rescale(DataDouane_ae$SUM_DIF_CONTI,to=c(0,1))
DataDouane_ae$TOT_ART <- scales::rescale(DataDouane_ae$TOT_ART,to=c(0,1))
DataDouane_ae$LAG_AVG_TAUX_APPL <- scales::rescale(DataDouane_ae$LAG_AVG_TAUX_APPL,to=c(0,1))
DataDouane_ae$PROP_DIF_CONTI <- scales::rescale(DataDouane_ae$PROP_DIF_CONTI,to=c(0,1))
DataDouane_ae$DATEH_NRC <- scales::rescale(DataDouane_ae$DATEH_NRC,to=c(0,1))
DataDouane_ae$LAG_DATEH <- scales::rescale(DataDouane_ae$LAG_DATEH,to=c(0,1))
DataDouane_ae$LAG_TOT_ART <- scales::rescale(DataDouane_ae$LAG_TOT_ART,to=c(0,1))
DataDouane_ae$AVG_TAUX_APPL <- scales::rescale(DataDouane_ae$AVG_TAUX_APPL,to=c(0,1))
DataDouane_ae$PROP_DIF_PAYS <- scales::rescale(DataDouane_ae$LAG_TOT_ART,to=c(0,1))
DataDouane_ae$FRAUDE_DEC=as.factor(DataDouane_ae$FRAUDE_DEC)

DataDouane_ae_fraude=subset(DataDouane_ae,FRAUDE_DEC==1)
DataDouane_ae_pas_fraude=subset(DataDouane_ae,FRAUDE_DEC==0)

set.seed(12)
x=sample(1:2,nrow(DataDouane_ae_pas_fraude),replace = TRUE,prob =c(2/3,1-(2/3)))
train=DataDouane_ae[x==1,]
DataDouane_ae_pas_fraude_val=DataDouane_ae[x==2,]
validation= bind_rows(DataDouane_ae_pas_fraude_val,DataDouane_ae_fraude)
x_train=as.matrix(train[,-27])
y_train=as.matrix(train[,27])
x_val=as.matrix(validation[,-27])
y_val=(validation[,27])

autoencoder <- keras_model_sequential()
autoencoder %>%
  layer_dense(units = 70, activation = "tanh", input_shape = ncol(x_train)) %>%
  layer_dense(units = 50, activation = "tanh") %>%
  layer_dense(units = 20, activation = "tanh") %>%
  layer_dense(units = 50, activation = "tanh") %>%
  layer_dense(units = 70, activation = "tanh") %>%
  layer_dense(units = ncol(x_train))
autoencoder %>% compile(loss = "mean_squared_error",optimizer = "adam")
checkpoint <- callback_model_checkpoint(filepath = "resultat.hdf5", save_best_only = TRUE,period = 1)
early_stopping <- callback_early_stopping(patience = 8)

autoencoder %>% fit(x_train, y_train, epochs = 50, batch_size = 500,
                    validation_split = 0.1,callbacks = list(checkpoint,early_stopping))
pred_train <- predict(autoencoder, x_train)
mse_train <- apply((x_train - pred_train)^2, 1, sum)

pred_val <- predict(autoencoder, x_val)
mse_val <- apply((x_val - pred_val)^2, 1, sum)
pred_nfraude <- predict(autoencoder, as.matrix(DataDouane_ae_fraude[,-27]))
auc(y_train,mse_train)
mse_nfraude <- apply((as.matrix(DataDouane_ae_fraude[,-27]) - pred_nfraude)^2, 1, sum)
auc(DataDouane_ae_fraude$FRAUDE_DEC,mse_nfraude)
plot(mse_train,col=2,main ="Distibution du MSE des données frauduleuses et non frauduleuses"
     ,xlab = "MSE")
plot(mse_nfraude,col=3)
legend(0.3,0.4,legend = c("non fraude","fraude"),col = c(2,3))

possible_k <- seq(0, 0.5, length.out = 100)
precision <- sapply(possible_k, function(k) {
  predicted_class <- as.numeric(mse_val > k)
  sum(predicted_class == 1 & y_val == 1)/sum(predicted_class)
})

qplot(possible_k, precision, geom = "line")
+ labs(x = "Threshold", y = "Precision")
y_val=as.numeric(y_val)
recall <- sapply(possible_k, function(k) {
  predicted_class <- as.numeric(mse_val > k)
  sum(predicted_class == 1 & y_val == 1)/sum(y_val)
})
qplot(possible_k, recall, geom = "line")
+ labs(x = "Threshold", y = "Recall")

```

Figure 3-10: L'algorithme de l'auto-encodeur

En ce qui concerne cet algorithme, nous n'avons pas effectué d'échantillonnage. Nous avons travaillé avec toute la base de données.

Nous avons donc importé les données puis nous avons transformé les variables numériques de telle sorte à ce qu'elles varient entre 0 et 1, pour que le réseau de neurone puisse les prendre en entrée.

Nous avons ensuite partitionné les données en ensemble d'entrainement et de validation. L'ensemble test sera quant à lui créé au niveau de l'apprentissage (voire la ligne 48 du code) avec un pourcentage que nous lui avons indiqué (10% de l'ensemble d'entrainement).

Nous avons construit un auto-encodeur symétrique contenant 7 couches avec pour fonction d'activation la fonction tangente hyperbolique. Le choix qui s'est fait quant au nombre de couches et de la fonction d'activation s'est fait de manière empirique (*trial and error*), car il n'existe pas de méthode directe pour déterminer le choix optimal.

Nous avons ensuite construit deux fonctions qui nous permettent respectivement de garder le meilleur modèle obtenu après la phase d'entrainement et de stopper l'entrainement si aucune amélioration n'est observée après 8 itérations consécutives.

Nous avons entrainé le modèle avec comme paramètres :

epoch : Il représente le nombre de fois où toutes les données sont passées à travers le réseau de neurone.

batch_size : Il représente le nombre d'individus dans un seul lot. En effet, l'ensemble d'entrainement est divisé en lots pour faciliter au réseau de neurones la phase d'apprentissage.

Nous avons ensuite décidé, avant de poursuivre la démarche, d'afficher les graphes représentant les erreurs de reconstruction des nouvelles versions des données d'entrainement puis des données frauduleuses. Nous avons obtenus les deux graphes suivants :

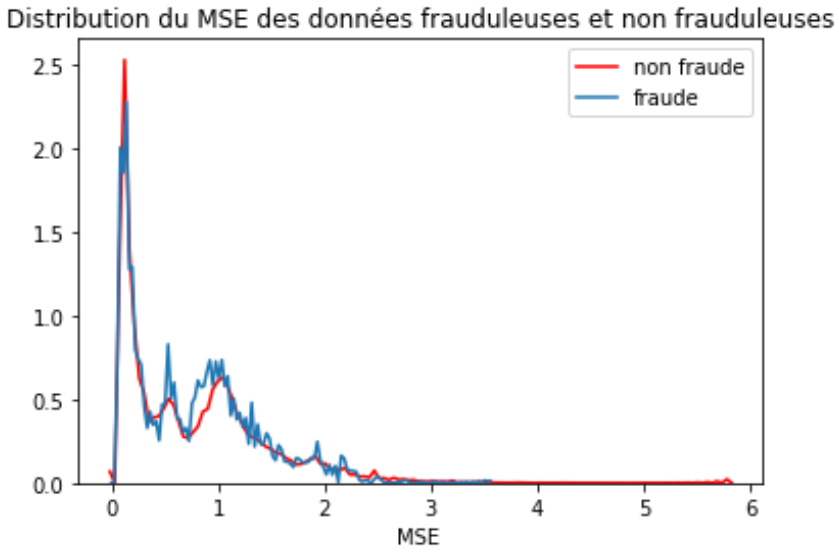

Figure 3-11 : Distribution de l'erreur de reconstruction (MSE) des données frauduleuses et non frauduleuses

Comme nous pouvons le remarquer sur la figure ci-dessus, la distribution de l'erreur dans le cas des données non frauduleuses utilisées dans la phase d'entrainement est quasiment la même que celles des données frauduleuses, alors que l'auto encodeur n'a pas appris à reconstruire ces dernières. Il devrait logiquement avoir plus de mal à les reproduire.

Il existe aussi certaines données dans la base d'entrainement dont l'erreur de reconstruction dépasse celles des données contenant les fraudes. Ce sont des valeurs extrêmes qui n'indiquent cependant pas une transaction frauduleuse mais qui perturbent néanmoins l'apprentissage. Les données non frauduleuses n'étant pas abondantes, il est moins probable de tomber sur de telles valeurs dans une si petite base. Ceci implique donc que l'on ne peut choisir un seuil d'erreur pour délimiter les fraudes des non fraudes.

Nous avons néanmoins essayé de faire varier le seuil *K* en fonction des deux métriques complémentaires *precision* et *recall* pour avoir la valeur de K qui maximise ces deux dernières. Et nous avons confirmé ce que nous avons conclu grâce au graphe de la ***Figure 3-11,*** car les différentes valeurs de K n'amélioraient en rien les valeurs de la *precision* et du *recall*, ces deux dernières étant constantes.

Etant donné que nous n'avons pas pu déterminer le seuil qui sépare les données frauduleuses du reste des données, nous savons que la piste que nous avons explorée n'aboutira pas. Nous passons donc aux modèles supervisés où nous utiliserons des échantillons représentatifs dans lesquels les données frauduleuses seront mieux exploitées.

IV.2 Les modèles supervisés

Nous avons décidé de développer des modèles d'apprentissage supervisé appartenant à des familles de méthodes différentes (voir ***Figure 2-1 Chapitre 2:I.5***) car nous ne savons pas quel type de modèle est adapté à notre problématique.

IV.2.1 Random Forest

L'algorithme que nous avons construit est reproduit dans la figure qui suit :

```
library(ROSE)
library(caTools)
library(randomForest)
library(caret)
library(pROC)
library(ROCR)

DataDouane=read.table(file.choose(),header = TRUE,sep = ",")
sample_data_both=ovun.sample(FRAUDE_DEC~.,data=DataDouane_ae,method="both",p=0.1,N=26210,seed = 1)$data
set.seed(12)
x=sample(1:2,nrow(sample_data_both),replace = TRUE,prob =c(0.8,0.2)  )
train=sample_data_both[x==1,]
test_rf=sample_data_both[x==2,]
train$FRAUDE_DEC= make.names(train$FRAUDE_DEC)
tunegrid <- expand.grid(mtry=c(5,6,7,8,9,10), ntree=c(200,300,400,500))
control_ROC <- trainControl(method="repeatedcv", number=10, repeats=3,
                            summaryFunction = twoClassSummary)
set.seed(12)
best_RF_model_ROC <- train(FRAUDE_DEC~., data=train
                           , method= "RF"
                           , metric="ROC"
                           , tuneGrid=tunegrid
                           ,trControl=control_ROC)


set.seed(1)
train$FRAUDE_DEC=as.factor(train$FRAUDE_DEC)
best_rf_model<-randomForest(FRAUDE_DEC~.,data = train,
                            na.action=na.omit,
                            mytry= 7,
                            ntree= 400)
```

Figure 3-12: Algorithme Random Forest

Nous avons tout d'abord importé nos données, créé un échantillon puis nous l'avons séparé en deux parties pour l'entrainement et le test.

Nous avons tout d'abord déclaré la fraude comme étant une variable catégorielle pour que l'algorithme de Random Forest fasse une classification et non une régression.

Nous avons utilisé la fonction *expand.grid* pour qu'elle nous fournisse les différentes combinaisons possibles des paramètres du modèle. Ces derniers sont définis comme suit :

$\boldsymbol{mtry}$: qui représente le nombre de variables choisies pour tester la pertinence de la division à chaque split.

$\boldsymbol{ntree}$ **:** qui représente le nombre d'arbres de décisions utilisés dans Random Forest.

Le résultat de la fonction est présenté en ***Annexe I*** .

Nous avons choisi de maximiser la surface sous la courbe ROC, indiquée par l'abréviation AUC, car elle nous propose la combinaison qui nous garantirait un résultat Pareto optimal. Nous optons donc pour la combinaison $mtry = 7\ et\ ntree = 400$.

Nous avons aussi fait appel à la fonction *trainControl* pour choisir la méthode avec laquelle seront entrainées les données. Nous avons choisi une méthode de validation croisée (cross validation) à 10 plis répétées 3 fois expliquée dans ***Annexe J*** , tel qu'à chaque itération, le découpage des données est effectué différemment.

Suite à cela, nous lançons le modèle et laissons la machine faire son apprentissage.

IV.2.2 Les machines à vecteurs supports

Nous avons construit 3 algorithmes de machine à vecteur support, tels que pour chaque algorithme, nous avons fait appel à une fonction noyaux différente à savoir : linéaire, radiale et polynomiale. L'algorithme est décrit dans la figure suivante :

```
library(magrittr)
library(dplyr)
library(Matrix)
library(ROSE)
library(SparseM)
library(kernlab)

svmvar=read.csv(file.choose(),header=TRUE,sep=",")
set.seed(123)
echantillon=ovun.sample(FRAUDE_DEC~.,data=svmvar,method="both",p=0.1,N=26210,seed = 1)$data
echantillon$FRAUDE_DEC = factor(echantillon$FRAUDE_DEC, levels = c(0, 1))
inde=sample(1:2,nrow(echantillon),replace = TRUE,prob =c(0.8,0.2)  )
entrainer=echantillon[inde==1,]
tester=echantillon[inde==2,]
entrainer=entrainer[,apply(entrainer,2,function(x) !all(x==0))]
tester=tester[,names(entrainer)]
tuned <- tune.svm(FRAUDE_DEC ~., data = entrainer, gamma = 10^(-6:-1), cost = c(1:200), degree=c(1:8), metric="ROC")
print(tuned)
svm_1= svm(formula = FRAUDE_DEC ~ .,
           data = entrainer,
           kernel = 'linear',
           preProcess = c("center", "scale"),
           cost=100
           )
test_pred_1<-predict(svm_1,newdata=tester)
table(test_pred_1, tester$FRAUDE_DEC )
svm_2 = svm(formula = FRAUDE_DEC ~.,
           data = entrainer,
           kernel = 'radial',
           preProcess = c("center", "scale"),
           cost=100,
           gamma=0.1 )
test_pred_3<-predict(svm_2,newdata=tester)
table(test_pred_2, tester$FRAUDE_DEC )
svm_3 = svm(formula = FRAUDE_DEC ~ .,
           data = entrainer,
           kernel = 'polynomial',
           preProcess = c("center", "scale"),
           cost=100,
           gamma=0.1,
           degree=1)
test_pred_3<-predict(svm_3,newdata=tester)
table(test_pred_3, tester$FRAUDE_DEC )

```

Figure 3-13 : Algorithme SVM

Après avoir importé nos données et préparé nos bases d'apprentissage et de test, nous avons fait appel à la fonction *tune.svm* pour nous donner les paramètres qui vont faire en sorte d'aboutir à la meilleure courbe ROC possible, et ils se définissent comme suit :

C : Il contrôle le compromis entre la maximisation de la marge et la minimisation de l'erreur de reconstruction. Une grande valeur de C implique une marge faible mais une erreur de classification moins importante tandis que dans le cas contraire l'inverse se produit.

Si C'est trop grand, le modèle va sur-apprendre et donc mal classer les données de test, une valeur trop petite va quant à elle entrainer un mauvais apprentissage de l'algorithme.

gamma : Il est utilisé pour adapter l'hyperplan aux données et est responsable de son degré de linéarité, c'est pour cela qu'il n'est pas utilisé dans le cas d'une fonction noyau à base linéaire.

Plus gamma est petit, plus l'hyperplan aura l'air d'une ligne droite ; si au contraire, il est trop grand, l'hyperplan sera plus courbé et pourrait trop bien délimiter les données et conduire à du sur-apprentissage.

degree **:** Il représente le degré de la fonction noyau de base polynomiale.

Une fois les paramètres optimaux sélectionnés, nous faisons appel à la fonction *svm* et indiquons les paramètres choisis ainsi que le type de la fonction noyau. Nous indiquons aussi que les données doivent être centrées et réduites, car cet algorithme utilise une notion de distance et donc les données doivent avoir la même grandeur (entre -1 et 1).

Nous avons testé nos 3 modèles et avons rapidement remarqué à partir des matrices de confusions illustrées dans les ***Tableau 3-3*** et ***Tableau 3-4 :*** que les modèles à noyaux linéaires et polynomiales ne sont pas adaptés à la nature de nos données car ils n'arrivent pas à séparer les transactions frauduleuses de celles qui ne le sont pas. Nous ne garderons que le modèle à noyau radial. Il sera comparé au reste des algorithmes dans la phase suivante.

	0	1
0	4732	500
1	0	0

Tableau 3-3 : La matrice de confusion du modèle à noyau linéaire

	0	1
0	4732	500
1	0	0

Tableau 3-4 : La matrice de confusion du modèle à noyau polynomial

IV.2.3 Les k plus proches voisins

L'algorithme que nous avons construit est présenté dans la figure suivante :

```
library(caret)
library(e1071)
library(ROSE)
library(pROC)
library(ROCR)
DataDouane=read.csv(file.choose(),header = TRUE,sep = ",")
sample_data_both<-ovun.sample(FRAUDE_DEC~.,data=DataDouane,method="both",p=0.1,N=26210,seed = 1)$data

set.seed(12)
ind<-sample(2,nrow(sample_data_both),replace=T,prob=c(0.8,0.2))
train<-sample_data_both[ind==1,]
test<-sample_data_both[ind==2,]
set.seed(12)
KNN_trcontrol<-trainControl(method="repeatedcv", number=10, repeats=3,
                            summaryFunction = twoClassSummary)

train$FRAUDE_DEC=as.factor(train$FRAUDE_DEC)
train$FRAUDE_DEC= make.names(train$FRAUDE_DEC)
KNN_model<- train(FRAUDE_DEC~., data = train, method ="knn",
                        preProcess = c("center","scale"),
                        trControl =KNN_trcontrol ,
                        metric ="ROC",
                        tuneGrid = expand.grid(k = 2:20))
```

Figure 3-14: Algorithme Knn

Nous avons tout d'abord importé, échantillonné puis séparé nos données en ensemble d'entrainement et de test.

Nous avons suivi les mêmes étapes qu'avec Random Forest pour le choix de la méthode d'entrainement, puis nous avons utilisé la fonction *train* en précisant que nous souhaitons utiliser l'algorithme des plus proches voisins. Nous précisons également que les données doivent être centrées et réduites sachant que nous souhaitons utiliser la métrique ROC. Le choix du *K* quant à lui est laissé à la fonction *expand.grid* qui le fera varier de 2 à 20 pour enfin choisir la valeur qui maximisera la courbe ROC (voir **Annexe K).**

En ce qui concerne les algorithmes basés sur le Boosting, nous avons choisi de ne présenter que l'algorithme de l'Extrem Gradient Boosting, car il présentait de meilleurs résultats que les autres, à savoir AdaBoost et Gradient Boosting. L'algorithme que nous avons construit se présente comme suit :

```
library(tidyverse)
library(xgboost)
library(caret)
library(readxl)
library(Matrix)
library(ROSE)
library(pROC)
library(ROCR)
DataDouane<-read.csv(file.choose(),header = T,sep = ',')

DataDouane<-ovun.sample(FRAUDE_DEC~.,data=DataDouane,method="both",p=0.1,N=26210,seed = 1)$data
set.seed(12)
ind<-sample(2,nrow(DataDouane),replace=T,prob=c(0.8,0.2))
train<-DataDouane[ind==1,]
test<-DataDouane[ind==2,]
train=train[,apply(train,2,function(x) !all(x==0))]
test=test[,names(train)]

xgbGrid <- expand.grid(nrounds = c(300,400,500),
                       max_depth = c(8,9,10,11,12),
                       colsample_bytree = seq(0.5, 0.9, length.out = 5),
                       eta = 0.1,
                       gamma=0,
                       min_child_weight = c(1,2,3,6,10))

train$FRAUDE_DEC=as.factor(train$FRAUDE_DEC)
train$FRAUDE_DEC= make.names(train$FRAUDE_DEC)
xgb_trcontrol_ROC = trainControl(method="repeatedcv", number=10, repeats=3,
                                 summaryFunction = twoClassSummary)
set.seed(12)
xgb_model_roc_plus = train(FRAUDE_DEC~., data=train
                           , method="xgbTree"
                           , metric="ROC"
                           , tuneGrid=xgbGrid
                           ,trControl=xgb_trcontrol_ROC)
xgb_model_roc$bestTune
```

Figure 3-15: L'algorithme XGBoost

Après avoir importé et partitionné nos données, nous avons fait appel à la fonction *expand.grid* pour nous donner les valeurs optimales des paramètres du modèle, à savoir :

nrounds : Il s'agit du nombre maximum d'itérations.

max _*depth* : Il correspond à la profondeur maximale d'un arbre. Elle est par défaut égal à 6 et son augmentation rendra le modèle plus complexe et plus susceptible d'être sujet au sur ajustement.

colsample_bytree : Il correspond à la proportion de variables échantillonnées aléatoirement qu'un arbre peut utiliser pour se construire. S'il est égal à 1 cela veut dire que toutes les variables sont utilisées.

eta : Il correspond au pas d'apprentissage. Il contrôle la contribution de chaque arbre en multipliant ce dernier par une valeur variant entre 0 et 1 avant de l'ajouter au modèle précédent. Sa valeur par défaut est de 0,3 et une valeur plus faible signifierait que le modèle est plus robuste au sur ajustement mais plus lent à converger.

gamma : Il s'agit du gain minimal requis pour effectuer une partition supplémentaire sur un nœud de l'arbre. Plus ce nombre est grand, plus l'algorithme est conservateur.

min ***_child_weight*** : Il s'agit du nombre minimum d'individus devant appartenir à une feuille pour que celle-ci puisse être sauvegardée. Plus ce nombre est grand, plus l'algorithme est conservateur.

Nous utilisons ensuite la fonction *trainControl* pour définir la méthode d'apprentissage du modèle. Là encore, nous choisissons une validation croisée à 10 split répétée 3 fois.

Nous finissons par inclure les deux fonctions précédentes dans le modèle et nous lui précisons que nous voulons optimiser la courbe de ROC (voir ***Annexe L)***.

V L'évaluation

Nous allons, dans ce qui suit, tester les 4 modèles que nous avons construits en nous basant sur plusieurs paramètres afin de sélectionner celui dont les performances sont les meilleures au regard de ces métriques.

Nous commençons par visualiser les courbes de ROC des différents modèles avec la surface sous chaque courbe :

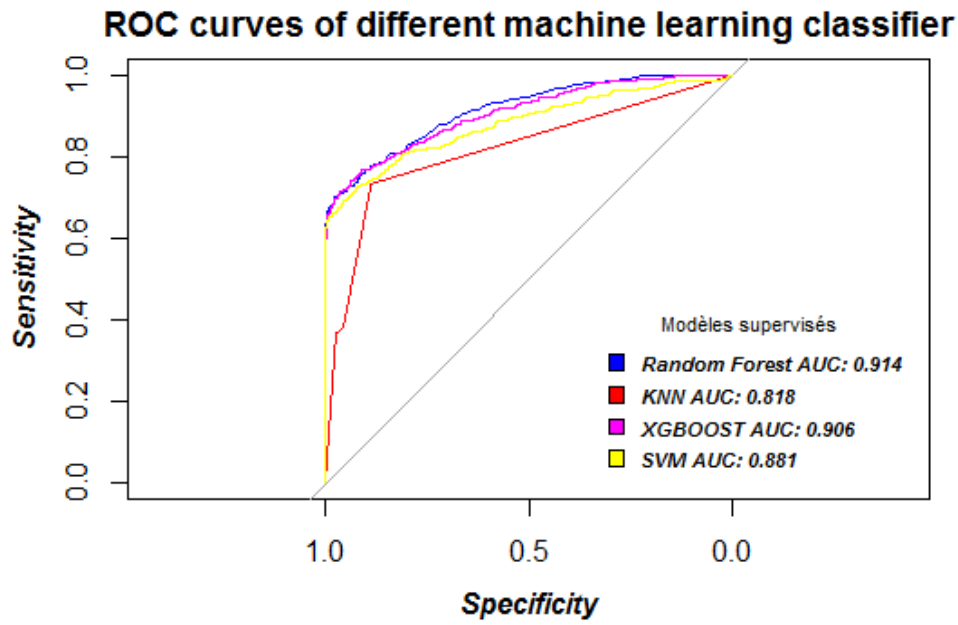

Figure 3-16 : Coubre ROC des modèles d'apprentissage supervisé

Nous remarquons que, de manière générale, les modèles aboutissent à de bons résultats avec des valeurs de surfaces sous la courbe proches de 1.

Nous calculons ensuite des indices de performance (voir ***Chapitre 2:I.7***) de chaque modèle en nous basant sur les matrices de confusion obtenues et présentées ci-après :

Matrices de confusion	classe réelle(-)	classe réelle(+)
	XGBoost	
classe prédite (-)	4711	176
classe prédite (+)	21	324
	Random Forest	
classe prédite (-)	4728	182
classe prédite (+)	5	316
	KNN	
classe prédite (-)	4361	234
classe prédite (+)	371	266
	SVM	
classe prédite (-)	4706	179
classe prédite (+)	26	321

Tableau 3-5 Matrices de confusion des modèles d'apprentissage supervisé

Le tableau suivant contient les indices de performances calculés pour chaque modèle à partir des matrices ci-dessus :

	Accuracy	**Precision**	**Spécificité**	**Sensibilité**	**AUC**
XGBoost	0,962347095	0,939130435	0,99556213	0,648	0,906
Random Forest	0,964251577	0,984423676	0,998943588	0,635	0,914
KNN	0,884365443	0,417582418	0,921597633	0,532	0,818
SVM	0,960818043	0,925072046	0,994505495	0,642	0,881

Tableau 3-6 Performance des modèles d'apprentissage supervisé

Pour trancher parmi ces modèles, nous avons décidé de leur attribuer des scores allant de 1 à 4 selon leurs performances puis nous avons calculé une moyenne sur ces scores pour ne garder qu'une seule valeur. Les résultats obtenus sont présentés dans le tableau ci-après :

	Accuracy	**Précision**	**Spécificité**	**Sensibilité**	**AUC**	**Score**
XGBoost	3	3	3	4	4	3,4
Random Forest	4	4	4	2	3	3,4
KNN	1	1	1	1	1	1
SVM	2	3	2	3	2	2,4

Tableau 3-7 Classement des modèles d'apprentissage supervisé

Il apparait que XGBoost et Random Forest obtiennent les meilleurs résultats avec une moyenne de 3.4. Néanmoins, nous choisirons pour la suite du travail le modèle Random Forest, puisqu'il comporte moins de paramètres à régler.

VI Le déploiement

Dans cette partie, nous allons expliquer comment, à partir des prédictions obtenues avec Random Forest, nous assignerons les déclarations aux différents circuits de vérification afin de répondre à la problématique que nous avons définie précédemment.

Nous avons tout d'abord commencé par redéfinir les cas de fraude car la base dont nous disposons n'est pas fiable dans le sens où une déclaration étiquetée comme non frauduleuse ne l'est pas forcément (fraude non détectée).

En effet, étant donnée l'important volume des importations passant par le circuit rouge, il n'est pas toujours possible d'effectuer une inspection minutieuse des conteneurs et donc il est probable qu'une fraude réelle ne soit pas détectée par les agents du contrôle douanier.

Par ailleurs, il existe des collusions entre certains opérateurs économiques et certains agents des douanes qui se détournent de leurs fonctions originelles, permettant ainsi à ses opérateurs, moyennant paiement, de commettre des fraudes sans être signalés.

De plus, la base n'est pas actualisée dans le cas où un contrôle s'opère à postériori. Une déclaration étiquetée comme étant non frauduleuse au moment de son enregistrement dans la base ne sera pas modifiée même si le contrôle à postériori révèle une fraude.

Néanmoins, des fraudes similaires (ayant les mêmes caractéristiques) que celles que nous venons d'évoquer (n'ayant pas été décelées) auraient pu être détectées dans une période antérieure, et celles-ci ont été apprises par le modèle qui s'est entrainé sur toutes les fraudes durant les 2 dernières années. Et donc, afin de capter ce genre de fraudes, nous allons jouer sur le *cut-off* illustré dans la figure suivante :

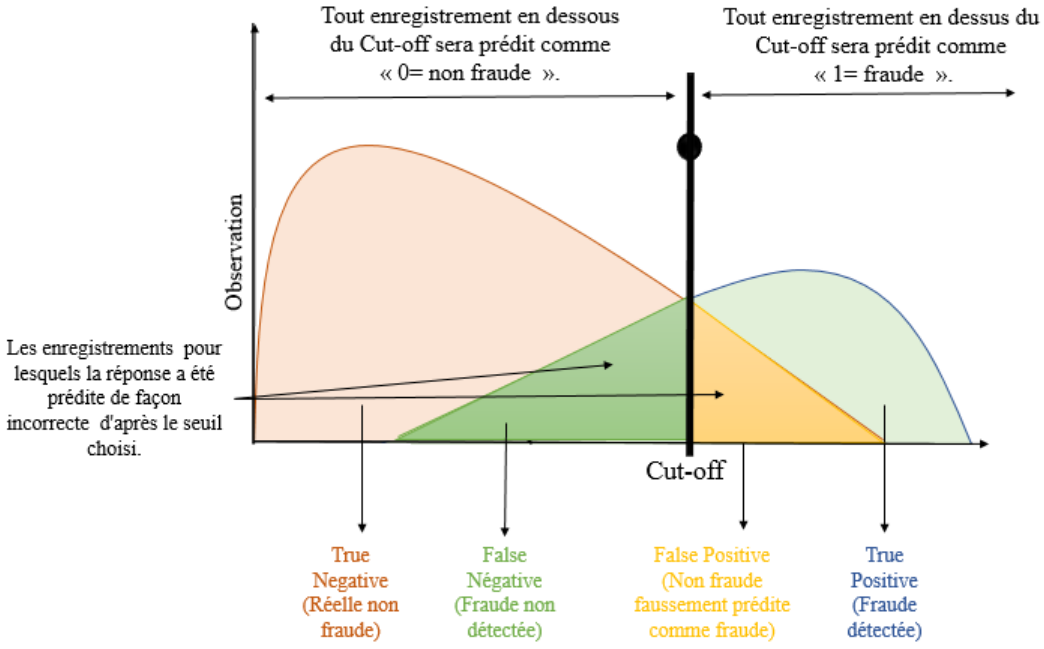

Figure 3-17: Le cut-off

Lorsque nous demandons à la machine de fournir le modèle optimal, l'algorithme cherchera le cut-off qui maximise la sensibilité et la spécificité en même temps. Il cherche donc l'optimum au sens de Pareto.

Le cut-off estimé par Random Forest est de 0.276 (***Figure 3-18)*** pour une sensibilité de 0.977 et une spécificité de 0.3 et on ne peut améliorer l'une des valeurs sans détériorer l'autre.

Néanmoins, nous cherchons à minimiser la proportion de faux négatif (fraude non détectée) quitte à augmenter la proportion de faux positif (déclaration faussement désignée comme frauduleuse) ; ce qui revient à minimiser 1 − spécificité (axe des abscisse) qui est égal au fnr et donc augmenter la spécifié au détriment de la sensibilité (axe des ordonnées). Car notre objectif est de capter le plus de fraudes même si cela implique de pousser le modèle à déclarer faussement certaines déclarations comme étant frauduleuses. Cela se traduira alors par une diminution de la sensibilité du modèle.

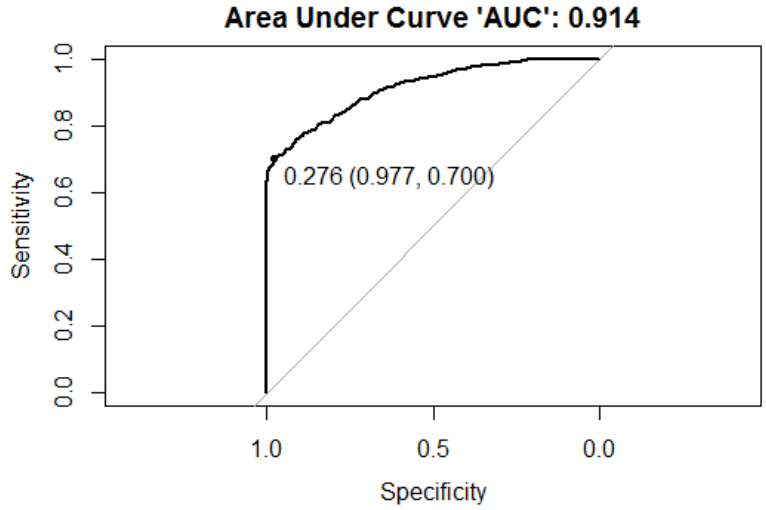

Figure 3-18: Cut-off optimum selon Random Forest

Nous allons donc modifier la valeur du cut-off indiqué sur la ***Figure 3-18*** afin de capter le maximum de fraude. Pour cela, nous avons utilisé l'algorithme suivant :

```
set.seed(1)
model1<-randomForest(FRAUDE_DEC~.,data = train, ntree =400,mtry =7,cutoff=c(1-0.035,0.035) )
train$fraude_rede=predict(model1,train)
train$new_fraude=apply(train[, c("fraude_rede","FRAUDE_DEC")], 1, max)
```

Figure 3-19 : Algorithme de redéfinition du cutt-off

Après plusieurs essais, nous avons choisi une valeur de cut-off égal à 0.035, puis nous avons créé de nouvelles fraudes en labélisant une déclaration comme frauduleuse si sa probabilité de fraude est supérieure ou égal à 0.035. Ensuite nous avons comparé l'ancienne version de nos données avec la nouvelle, telle que :

- Si la marchandise est frauduleuse et que sa probabilité de fraude soit supérieure à 0.035. Dans ce cas elle sera considérée comme frauduleuse (pas de changement sur la base).
- Si la marchandise est non frauduleuse et que sa probabilité de fraude soit supérieure à 0.035. Dans ce cas elle sera considérée comme frauduleuse (Mise à jour de la base).
- Si la marchandise est frauduleuse et que sa probabilité de fraude soit inférieure à 0.035. Dans ce cas elle sera considérée comme frauduleuse (pas de changement sur la base).

- Si la marchandise est non frauduleuse et que sa probabilité de fraude soit inférieure à 0.035. Dans ce cas elle sera considérée comme non frauduleuse (pas de changement sur la base).

Après avoir créé notre nouvelle base, nous lui avons intégré une variable qui contient les affectations des déclarations aux 3 circuits. Puis nous l'avons modifiée en y ajoutant un 4ème circuit qui impliquera une vérification de la marchandise à l'aide d'un scanner[4], ce qui sera moins contraignant qu'une fouille physique mais plus efficace qu'une simple revue documentaire. Nous avons utilisé les règles suivantes pour les nouvelles affectations :

- Une déclaration est affectée au circuit vert si elle était auparavant dans ce même circuit ou dans le circuit orange et qu'elle soit déclarée comme étant non frauduleuse par l'outil.
- Une déclaration est affectée au circuit orange si cette dernière est non frauduleuse et que son affectation originale était le circuit rouge.
- Une déclaration est affectée au circuit jaune si elle est frauduleuse et que son ancien circuit soit vert ou orange.
- Une déclaration est affectée au circuit rouge si elle est frauduleuse et que son ancien circuit soit rouge.

L'affectation des déclarations aux différents circuits est illustrée dans la ***Figure 3-20***.

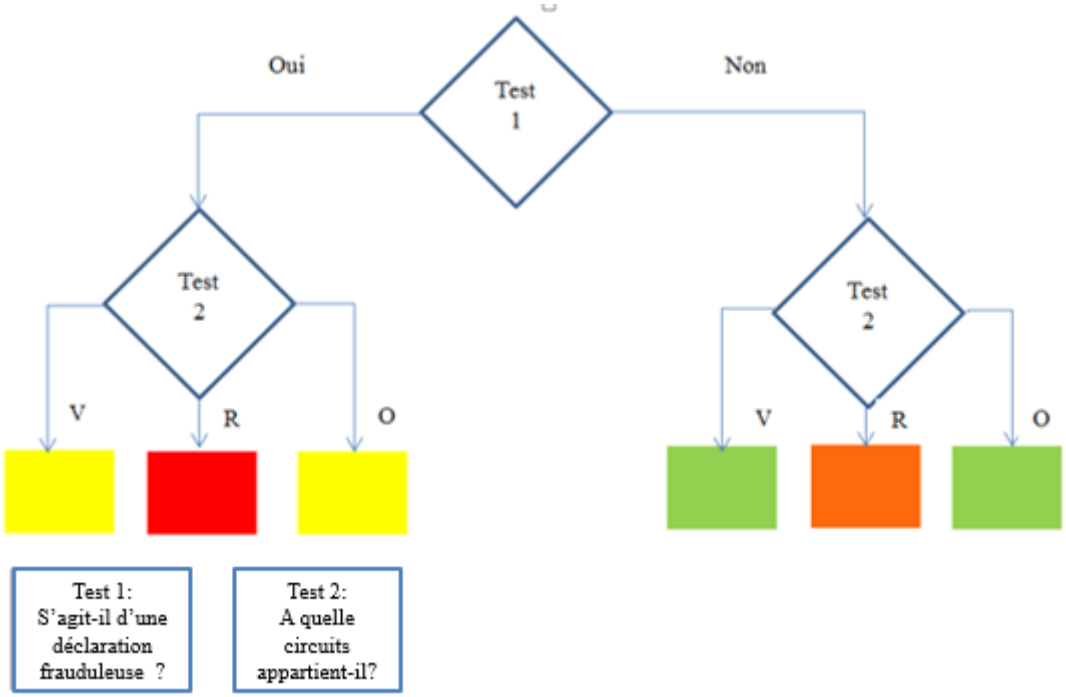

Figure 3-20: Affectation des marchandises aux circuits de vérification

Nous nous retrouvons donc avec une variable contenant quatre modalités représentant les quatre types de circuits de vérification : rouge, orange, jaune et vert.

[4] Les scanners sont des outils qui permettent l'accélération de la procédure de contrôle et de vérification et facilitent les opérations de dédouanement. Il s'agit d'équipements sophistiqués qui permettront de visualiser, d'analyser les cargaisons et d'en connaitre le contenu sans avoir recours au déchargement.

Nous avons ensuite entrainé notre modèle Random Forest avec la nouvelle base avec pour variable cible cette fois-ci les types de circuits et non pas l'éventualité de la fraude. Random Forest va ainsi nous fournir l'affectation des déclarations aux différents circuits selon les règles que nous lui avons inculquées en l'entrainant et non pas nous indiquer si la déclaration est frauduleuse ou non.

Le schéma ci-après résume le mode opératoire de l'outil que nous proposons :

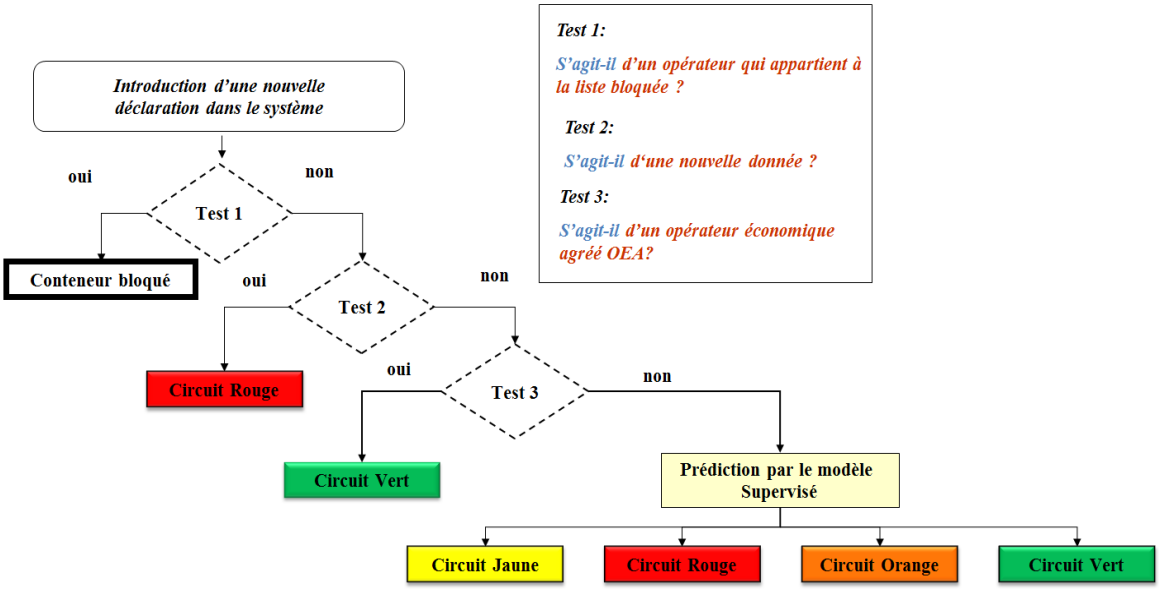

Figure 3-21 : Mode opératoire de l'outil d'aide à la décision proposé

La liste bloquée étant une liste contenant les opérateurs ayant commis une fraude et n'ayant toujours pas réglé l'amende qui leur a été adressée.

Nous l'avons ensuite testé sur toutes les déclarations de l'année 2018, après avoir récupéré et traité la base de données contenant les déclarations de cette année-là.

Nous avons visualisé la fraude présente dans chaque circuit et les résultats obtenus sont présentés dans le tableau suivant :

	Fraude	
	0	**1**
J	83450	52
O	23707	5
R	137425	845
V	59575	9

Tableau 3-8 : Affectation des déclarations aux circuits de vérification par l'outil d'aide à la décision.

Nous avons ensuite visualisé la proportion des déclarations affectées à chaque circuit pour vérifier si nous avons atteint l'objectif fixé par la douane, et nous avons obtenu le digramme en camembert suivant :

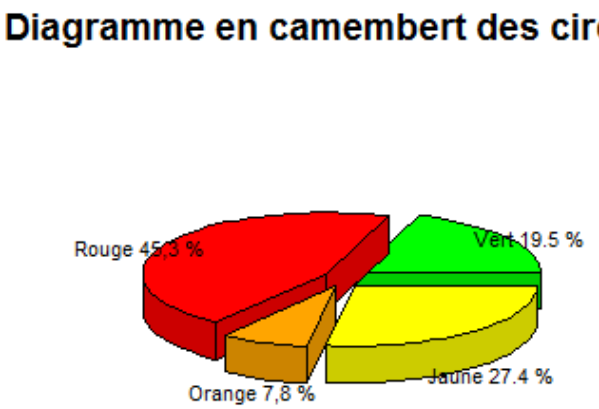

Figure 3-22 : Diagramme en camembert représentant la proportion des déclarations en chaque circuit

Nous remarquons que 93% des fraudes passent par le circuit rouge, elles seront donc théoriquement toutes interceptées par les agents du contrôle douanier.

De plus, le circuit rouge comporte 45,3% de l'ensemble des déclarations ; ce qui représente une diminution de 38% par rapport à la proportion qui lui était attribué au départ.

Ceci implique donc que le circuit rouge sera moins encombré et que les vérifications sur la marchandise se feront de manière plus minutieuse et ainsi plus de fraudes pourront alors être détectées.

En effet, parmi les 137.425 déclarations non frauduleuses qui lui ont été affectés, un comportement suspect a été détecté par le modèle et il se pourrait que des fraudes se cachent dans cette population. Car rappelons le, le modèle a été entrainé avec la fraude redéfinie. Il est donc « méfiant » envers la fraude et si une déclaration a été affectée au circuit rouge c'est qu'il y'a quelque part risque de contentieux.

Quant au circuit vert, le modèle lui a affecté environ 20% des déclarations. Et, parmi ces déclarations se trouvent 9 fraudes (1% du total des fraudes) : ce que les spécialistes au niveau de la douane ont jugé acceptable au regard des améliorations générales des affectations aux circuits.

Le circuit jaune que nous avons ajouté allégera la charge sur le circuit rouge en traitant 27,4% des déclarations et les 52 fraudes qu'il contient pourront normalement être détectées grâce au scanner. En effet, ce dernier permettra de visualiser les anomalies présentes dans la marchandise et une inspection physique s'en suivra pour une vérification plus poussée si nécessaire.

Nous avons élaboré le schéma illustré dans la **Figure 3-23** pour expliciter de manière succincte le travail que nous avons effectué dans ce chapitre.

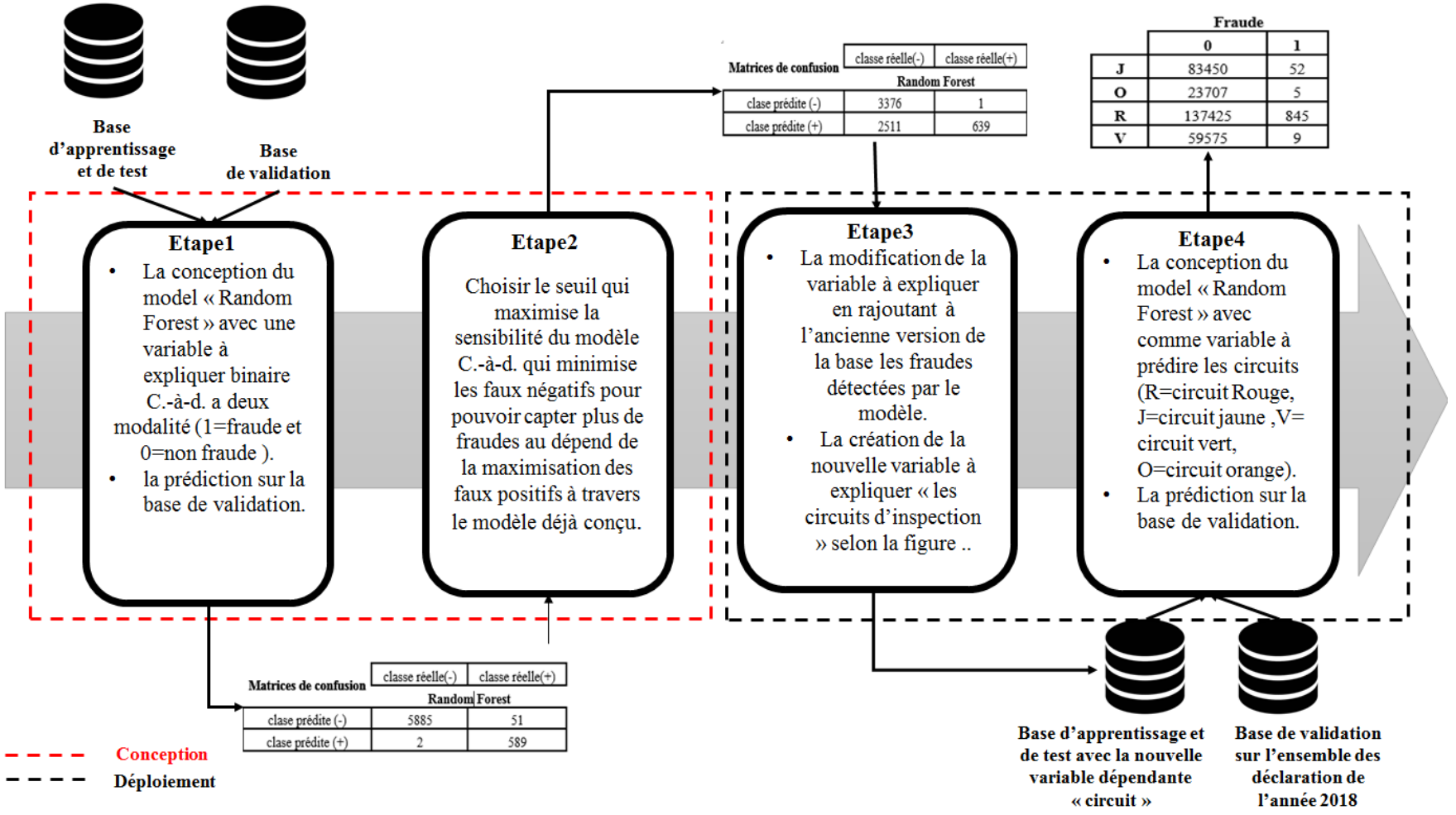

Figure 3-23 : Schéma récapitulatif de l'outil proposé

Conclusion

Nous avons dans cette première partie du chapitre 3 suivi les étapes de la méthodologie de Data Mining CRISP-DM décrite dans le précédent chapitre.

Nous avons préparé notre base de données selon les étapes dictées par la méthodologie et nous nous en sommes servies pour construire des modèles prédictifs.

Nous avons testé nos modèles, et les résultats se sont avérés globalement plutôt satisfaisants. Deux algorithmes se sont démarqués des autres en présentant les meilleures performances aux regards des différentes métriques que nous avons utilisées pour les évaluer.

Néanmoins, nous avons sélectionné Random Forest pour opérer les modifications que nécessite son adaptation aux besoins de notre problématique.

Une fois effectuées, nous avons utilisé les résultats de l'algorithme pour redéfinir la fraude et avons entrainé une nouvelle fois Random Forest. Mais cette fois si avec pour variable cible les différents circuits de vérification (rouge, jaune, orange et vert).

L'entrainement terminé, nous avons testé notre modèle avec l'ensemble des données de l'année 2018 et avons constaté que nous avons atteint l'objectif fixé par la douane.

Chapitre 4 : Traitement des limites du modèle supervisé

Chapitre 4: Traitement des limites du modèle supervisé

Introduction :

L'étiquetage préalable des données repose largement sur l'expérience humaine et la connaissance de la fraude ce qui réduit la possibilité d'en détecter de nouvelles.

De plus, nous savons que les fraudes détectées a postériori ne sont pas mises à jour et que beaucoup d'entre elles échappent à la vigilance des agents du contrôle douanier à cause du volume excessif de déclarations passant par le circuit rouge. Ou encore, celles-ci sont dissimulées par des agents de douanes corrompus.

L'apprentissage non supervisé est utilisé lorsqu'on ne dispose pas de données labélisées. Il nous permet de modéliser la structure ou la distribution sous-jacente dans les données afin d'en apprendre davantage sur les populations.

Le clustering est l'une des méthodes d'apprentissage non supervisées qui nous permet de regrouper des instances ayant des traits communs sans connaitre au préalable le nombre de groupe que l'on peut trouver ni leurs structures caractéristiques.

Nous allons utiliser dans cette partie quelque unes de ces méthodes malgré le fait que nous disposons de données étiquetées. Car, nous chercherons dans ce qui suit à détecter des fraudes inédites en observant le comportement des données et non pas en les classant selon des catégories définies auparavant.

En identifiant les regroupements anormaux, le clustering est capable de distinguer le modèle de fraude même s'il est totalement nouveau. Il nous permettra donc de palier à ce problème.

Afin d'y parvenir, nous avons utilisé des modèles d'apprentissage non supervisé de différentes manières.

Nous allons présenter dans ce qui suit leur implémentation. Puis nous commenterons les résultats obtenus par chaque modèle, afin de désigner celui que l'on devra améliorer.

I Implémentation des modèles avec toutes les variables

Nous avons commencé par introduire les déclarations dont nous disposons dans notre base de données, avec l'ensemble des variables utilisées dans la classification, dans trois différents modèles de clustering.

Ces modèles sont : les moyennes mobiles (K-means), DBSCAN et les cartes auto-organisatrices de Kohonen.

I.1 Conception des modèles

Ci-dessous sont présentés les modèles d'apprentissage non supervisé que nous avons développés.

I.1.1 K-means :

L'algorithme se présente ainsi :

```
library(cluster)
library(kselection)
library(e1071)
library(dplyr)
library(caTools)

kvar=read.csv(file.choose(),header=TRUE,sep=",")
set.seed(1)
kechantillon=sample(kvar,20000,replace = FALSE)
kechantillon$FRAUDE_DEC = factor(kechantillon$FRAUDE_DEC, levels = c(0, 1))
kechantillon_scal=scale(kechantillon[,-"FRAUDE_DEC"])
k<-kselection(kechantillon_scal,parallel=FALSE,k_threshold=0.9,max_centers=20)

clusters <- kmeans(kechantillon_scal, k)
clusplot(kechantillon_scal,clusters$cluster,main = "Clusters K-means")
```

Figure 4-1 : Algorithme k-means

Nous avons utilisé un échantillon de 20.000 déclarations pour ne pas encombrer le graphique des clusters puis nous avons fait appel à la fonction ***kselection*** pour avoir une valeur optimale du nombre de clusters **k**. Nous avons choisi la valeur 20 comme nombre maximal de clusters et un seuil de 0,9 comme résultat à obtenir d'une partition pour qu'une valeur de **k** soit sélectionnée. Ce seuil est une mesure qui permet de tester la qualité du partitionnement pour un **k** donné.

Pour cet algorithme, nous avons trouvé le résultat suivant :

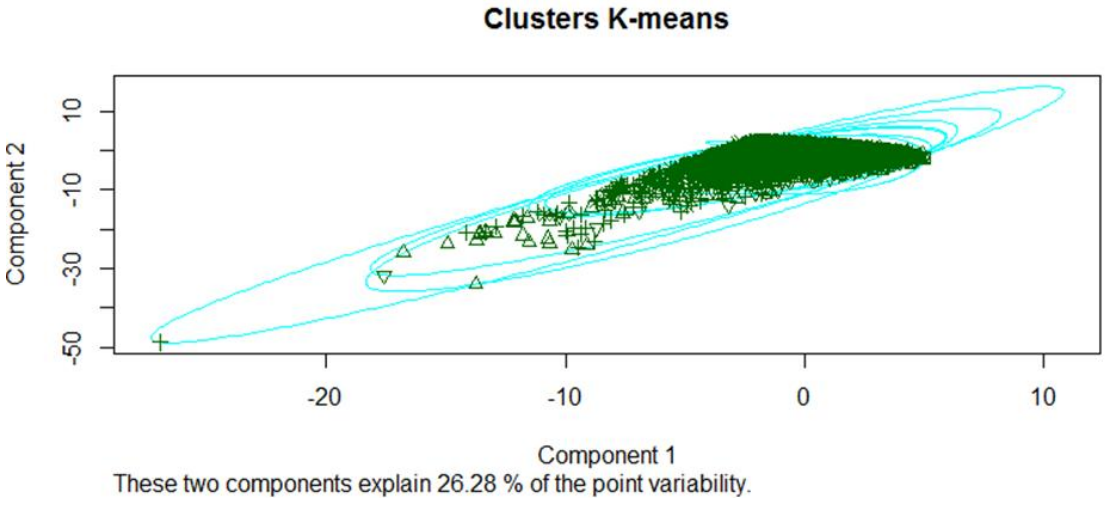

Figure 4-2 : Résultats obtenus par k-means

I.1.2 DBSCAN :

L'algorithme se présente ainsi :

```
library(fpc)
library(dbscan)
library(factoextra)
library(dplyr)
library(tidyverse)
library(caret)

dbscanvar=read.csv(file.choose(),header=TRUE,sep =",")

dbscanechantillon=sample(dbscanvar,20000,replace=FALSE)
dbscanechantillon$FRAUDE_DEC = factor(dbscanechantillon$FRAUDE_DEC, levels = c(0, 1))

dbscanechantillon=scale(dbscanechantillon[,-27],)
dbscanechantillon=as.matrix(dbscanechantillon[,-27])

dbscan::kNNdistplot(dbscanechantillon,k =8)
abline(h =5)

res.db <- dbscan::dbscan(dbscanechantillon, 5, 8)
clusplot(dbscanechantillon,res.db$cluster,main="cluster DBSCAN")
```

Figure 4-3: Algorithme DBSCAN

Nous avons inséré dans l'algorithme un échantillon de 20.000 observations et nous avons utilisé la fonction *kNNdistplot* pour choisir la valeur du rayon de voisinage optimale. Cette fonction nous a retournés une valeur de 5 (***Annexe M***) pour un nombre maximal de clusters égal à 8. Nous avons obtenu les résultats suivants :

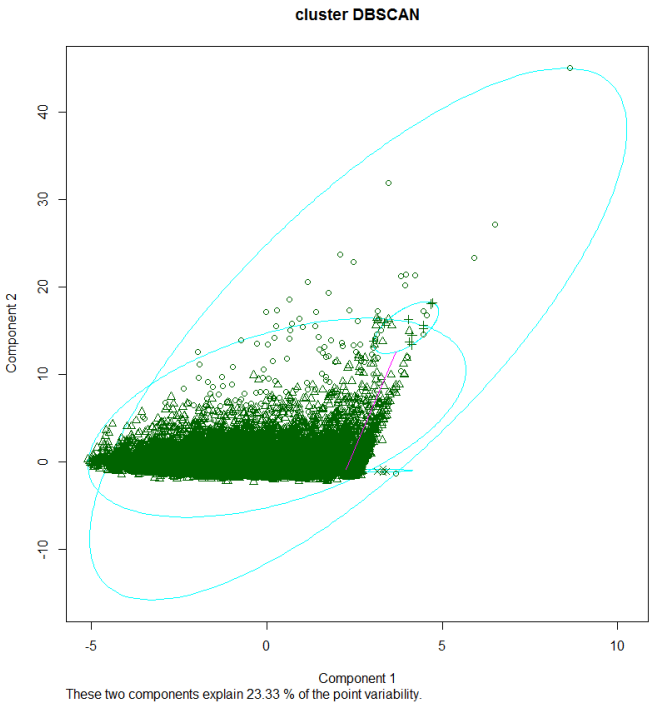

Figure 4-4 : Résultats obtenus par DBSCAN

I.1.3 Carte de Kohonen

L'algorithme utilisé est le suivant :

```
# Unsupervised Self-Organizing Maps
library(kohonen)
library(dplyr)
library(ggplot2)
library(ROSE)
library(dimRed)
library(dplyr)
# Data
data_kohonan <- read.csv(file.choose(), header = T,sep = ",")
data_kohonan_y <-data_kohonan[,1:145]
data_kohonan_x<- as.factor(data_kohonan$FRAUDE_DEC)
set.seed(123)
SOM_grid <- somgrid(xdim = 6, ydim =6, topo = "rectangular" )
SOM_model <- som(as.matrix(data_kohonan_y),
                 grid = SOM_grid,
                 alpha = c(0.05, 0.01,0.005),
                 rlen=400)
plot(SOM_model, type="mapping",
     col = c("blue", "red")[data_kohonan_x],
     pch=16,cex=0.01,
     main = "Carte de KOHONEN en fonction \n de la variable cible 'Fraude'")
legend("bottomright", legend=c("Fraude", "non Fraude"),
       col =   c("blue", "red"), pch = 16)
plot(SOM_model, type="counts" ,
     main = "Carte de KOHONEN en fonction \n de la cardinalité des neurones")
```

Figure 4-5 : Algorithme Kohonen

Pour les cartes de Kohonen, nous n'avons pas eu besoin d'utiliser un échantillon car la carte topologique offre un bon aperçu de la dispersion des données grâce à la flexibilité de la taille de la grille utilisée. Dans notre cas, nous avons choisi une grille carrée de 6*6 neurones. Nous avons laissé l'algorithme choisir le pas d'apprentissage entre les valeurs de 0.05, 0.01 et 0.005 et avons fixé le nombre d'itérations maximal à 400 car l'algorithme ne possède pas de condition de convergence. Nous avons obtenu la carte suivante :

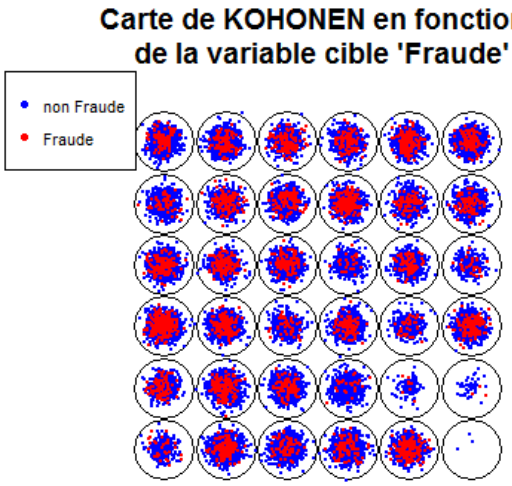

Figure 4-6 : Distribution des déclarations sur la carte de Kohonen selon l'ensemble des variables

I.2 Discussion des résultats

Les résultats obtenus par DBSCAN et K-means sont très similaires et tout aussi décevants. Les données sont présentées sous forme d'un nuage de points indissociables ou aucun cluster ne semble apparaitre de manière distincte. De plus, les deux axes qui ont été gardés pour représenter les données ne présentent que 26% et 23% de l'information pour K-means et DBSCAN respectivement.

Quant aux cartes de Kohonen, nous pouvons remarquer que la fraude est distribuée de manière quasi uniforme sur la carte ce qui en fait un résultat tout aussi insatisfaisant car inexploitable. Néanmoins, avec ces nombreuses sorties graphiques, les résultats sont de manière générale faciles à comprendre et à expliquer. En effet, elles nous permettent de visualiser l'effectif dans chaque nœud, ainsi que d'établir le rôle des variables dans la définition des différentes zones qui composent la carte topologique.

De plus, malgré le fait que la fraude est présente un peu partout sur la carte et qu'il serait difficile de trouver des clusters, nous savons que des déclarations regroupées au sein du même nœud ont des caractéristiques communes ce qui pourrait nous fournir des informations utiles. C'est pour cela que nous allons tâcher d'améliorer ce dernier modèle dans la suite de l'étude.

II Implémentation des cartes de Kohonen avec un nombre restreint de variables

En intégrant toutes les variables, nous avons réduit le pouvoir discriminatoire des modèles. En effet, plus le nombre de variables augmente, plus un algorithme a du mal à trouver les similarités entre les données.

Pour y remédier, nous avons décidé d'utiliser les variables que Random Forest a utilisé dans la classification. En effet, lors de sa construction ce dernier se base sur le gain de Gini pour sélectionner les variables qui découperont le mieux les données.
Le gain de Gini est défini comme suit :

$$Gain_{gini}(P,T) = gini(S_P) + \sum_{j=1}^{2} p_j gini(S_{Pj})$$

Où $gini$ représente l'indice de Gini présenté précédemment ; $\boldsymbol{T}$ représente le test considéré ; et $\boldsymbol{P}$ la position sur l'arbre de décision. $\boldsymbol{S_P}$ est l'échantillon associé à la position $\boldsymbol{P}$ et $\boldsymbol{p_j}$ est la proportion de l'échantillon $\boldsymbol{S_P}$ qui satisfont la $j$ème branche du test $\boldsymbol{T}$.

Le gain est maximal lorsque le choix d'un attribut permet de classer correctement toutes les données et il est nul lorsque les données sont aussi mal classées après le test qu'avant.

Nous avons donc sélectionné les meilleures variables au sens de Gini présenté dans la ***Figure 4-7*** après avoir flouté leurs noms pour des raisons de confidentialité.

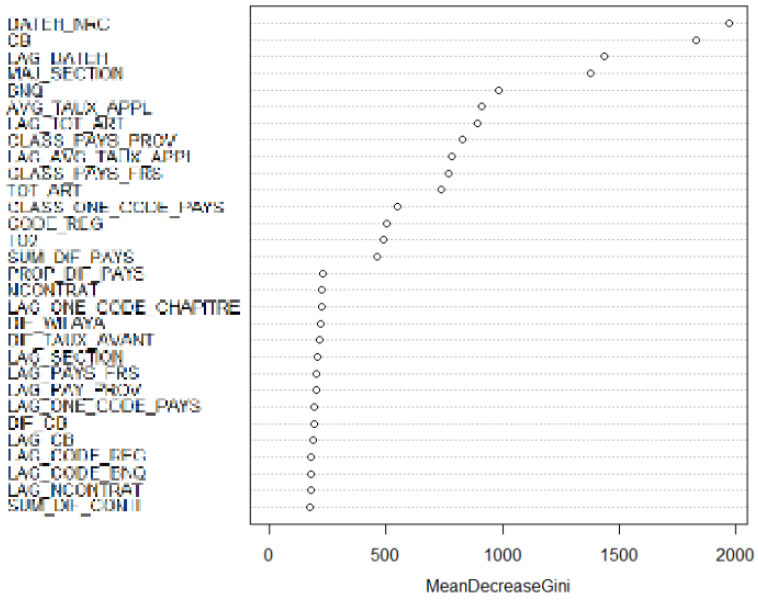

Figure 4-7 : Classement des variables selon le gain de Gini par ordre d'importance

Pour ne pas encombrer le modèle, nous n'avons sélectionné que les quinze (15) premières variables ayant obtenues le gain de Gini le plus important. Car certaines de ces variables sont catégoriques et comportent donc plusieurs modalités. Avec l'application de la méthode One-Hot-Encoding, ces modalités seront transformées en autant de variables, ce qui encombrerait le modèle.

Ensuite, nous avons introduit ces variables sélectionnées dans le modèle des cartes de Kohonen qui nous a fourni la carte topologique suivante :

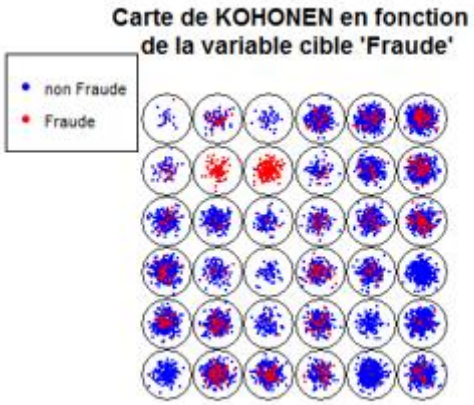

Figure 4-8 : Distribution des déclarations sur la carte de Kohonen selon les variables sélectionnées grâce à l'indice de Gini

Nous pouvons remarquer que bien que la fraude soit là encore distribuée sur quasiment l'ensemble de la carte topologique, il semble y avoir des nœuds où la concentration de la fraude est plus prononcée par rapports aux autres nœuds ; ce qui est une amélioration du modèle précédent.

Néanmoins, là encore, nous ne pouvons pas générer des clusters où les déclarations frauduleuses et non frauduleuses soient séparées de manière distincte. Nous allons donc essayer d'améliorer encore ce résultat en ayant recours aux techniques de réduction dimensionnelle sur l'ensemble des variables que nous avons sélectionné grâce au gain de Gini.

III Implémentation des cartes de Kohonen avec des variables de dimensions réduites

Nous avons tout d'abord essayé d'appliquer une ACP, qui est une méthode linéaire de réduction de dimension sur nos données. Mais nous avons remarqué en visualisant le gain de l'information en fonction de l'ajout des axes, que celle-ci évoluait de manière proportionnelle avec ces ajouts. En d'autres termes, il n'y a pas d'axes dominants qui résumeraient l'information, et nous ne pouvons donc en sacrifier à notre guise sans que l'information en pâtisse.

Nous avons donc décidé d'utiliser une technique non linéaire de réduction de dimension et nous avons choisi la t-SNE (***Annexe N***, car la fonction disponible sur R nous permet de visualiser les données avec leurs dimensions réduites sur des espaces de 2 à trois dimensions.

L'algorithme que nous avons utilisé est disponible en ***Annexe O*** et les résultats obtenus grâce à cet algorithme sont illustrés sur la figure suivante :

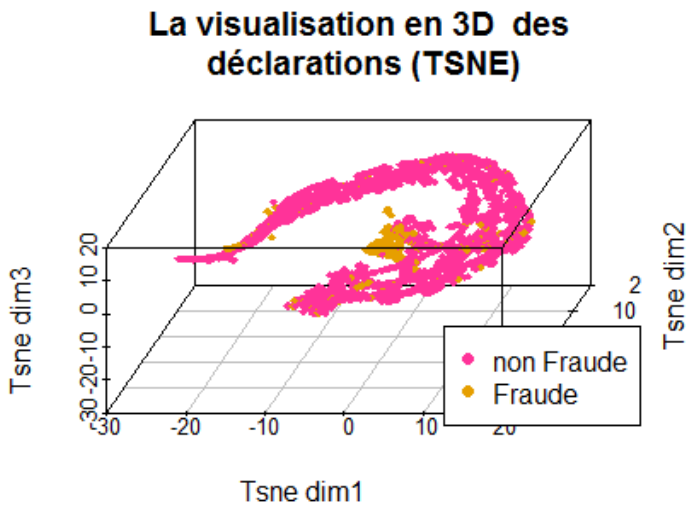

Figure 4-9 : La visualisation en 3D des déclarations avec la méthode de réduction T-SNE

Il apparait clairement qu'une grande partie des déclarations frauduleuses (en jaune) est séparée du reste des données en (en rose). Ce qui veut dire que l'information quant à la nature de la déclaration peut être obtenue en utilisant seulement trois dimensions.

Nous avons donc récupéré ces données avec cette fois-ci seulement 3 dimensions et les avons injectées dans le modèle des cartes de Kohonen. Nous avons obtenu les résultats suivants:

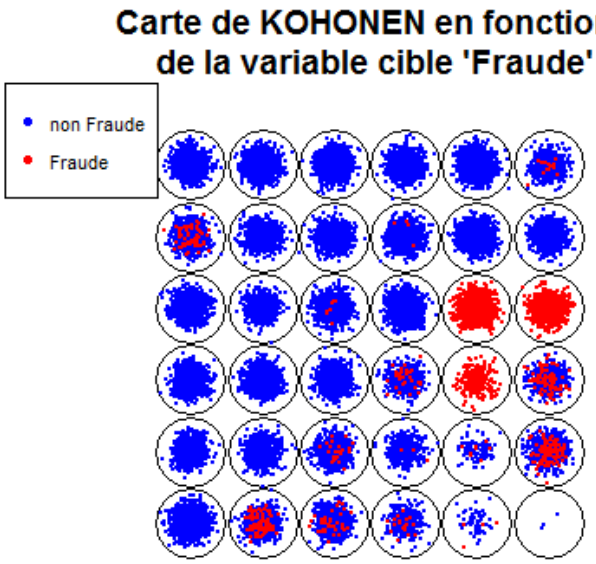

Figure 4-10 : Distribution des déclarations sur la carte de Kohonen selon les trois dimensions obtenues par T-SNE

Nous pouvons remarquer une nette amélioration de la carte topologique telle qu'il existe trois neurones où il n'y a que la fraude, ainsi que d'autres où il n'y en a aucune. Les données sont donc mieux séparées en comparaison avec les résultats du modèle obtenu précédemment.

Nous pouvons aussi remarquer que les neurones se trouvant au Sud-Est de la carte ont tendance à avoir plus de fraude que ceux du Nord-ouest. Ils sont aussi proches des neurones où il n'y a que de la fraude, et étant donné que la carte respecte la topologie des données, leur proximité indique une certaine ressemblance des déclarants au niveau de certains paramètres.

On peut donc supposer que certaines de ces déclarations étiquetées comme non frauduleuses peuvent être en fait des fraudes qui n'ont pas été détectées lors du contrôle douanier.

Une fois nos données positionnées sur la carte topologique, nous les partitionnons en 3 classes selon la cassure de la courbe représentant l'inertie intra-classe (***Annexe P*** . Nous avons choisi nos clusters selon la ressemblance qui excise entre les neurones. Nous obtenons la figure suivante :

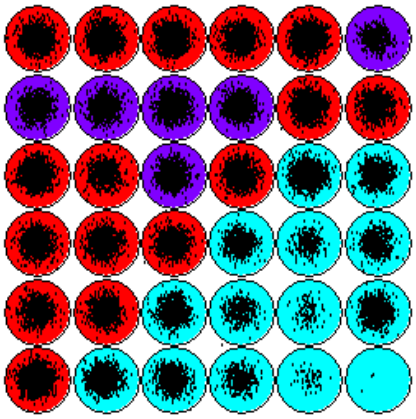

Figure 4-11 : Clusters obtenus par la carte de Kohonen

Les clusters obtenus vont servir à détecter de nouvelles fraudes éventuelles. En effet, toute déclaration qui sera affectée aux clusters représentés en bleu sur la carte devra être inspectée de manière minutieuse, car ils correspondent aux neurones où il n'y a que de la fraude, ainsi que les neurones contenant des données ayant un comportement similaire. Ce qui correspond à une forte probabilité de fraude.

Les données groupées dans le cluster représenté en rouge sont quant à elles non frauduleuses et une déclaration qui sera affectée à ce cluster ne sera pas suspecte.

Quant au cluster représenté en violet, il contient des données où la concentration de fraude est moins importante que dans le cluster représenté en bleu. Cette particularité implique qu'une déclaration qui lui sera affecté sera aussi inspectée minutieusement même si la probabilité qu'elle soit frauduleuse n'est pas très importante- car le but est d'en capter le maximum.

Conclusion :

Nous avons consacré ce chapitre à sélectionner et développer un outil qui permet de détecter le plus de fraudes. L'idée principale étant d'identifier des comportements frauduleux que l'expérience humaine n'a pas su reconnaitre.

Nous avons utilisé pour cela différents algorithmes de Machine Learning non supervisés et avons sélectionné celui qui nous fournissait une meilleure représentation de nos données.

Nous avons amélioré ce modèle en modifiant la base de données utilisées et avons pu obtenir des clusters où la fraude est très présente.

Ces clusters vont servir à détecter des fraudes inédites, telles que si une déclaration future est assignée à ces clusters et que le modèle supervisé indique qu'il ne s'agit pas d'une fraude, il se pourrait qu'en fait il s'agisse d'une fraude que le modèle supervisé ne connait pas.

Conclusion générale

En ratifiant la convention de Kyoto, les douanes algériennes se sont engagées dans un projet de modernisation dont le principal objectif est la rationalisation du contrôle douanier grâce à l'intégration de la gestion du risque. En théorie, ce processus permet d'orienter les efforts et les ressources de l'administration des douanes vers les transactions à risque évitant ainsi un contrôle exhaustif des déclarations et permettant un ciblage efficace de la fraude.

Néanmoins, lors du diagnostic que nous avons effectué dans le premier chapitre, nous avons constaté que les résultats auxquels a permis d'aboutir ce processus ne correspondaient pas aux attentes des douanes algériennes. En effet, l'implémentation du modèle économétrique lors de la phase de l'analyse du risque a abouti à une distribution non cohérente des déclarations sur les circuits de vérification (rouge, orange, vert). Cette faible performance est due aux restrictions (liste de blocage) imposées sur les marchandises avant d'être évaluée par le modèle et aux variables non mises à jour qui le composent.

Ces constatations nous ont amené à orienter nos travaux vers la proposition d'un outil d'aide à la décision basé sur l'intelligence artificielle en remplacement du modèle économétrique actuellement appliqué afin de répondre aux objectifs antagonistes de la douane, à savoir les facilitations au commerce extérieur et le contrôle douanier.

Pour cela, nous avons effectué une revue de littérature sur les différents concepts et techniques de l'intelligence artificielle, notamment le Machine Learning dont nous avons détaillé certains algorithmes ; ce qui nous, a par ailleurs, permis de bien assimiler leur fonctionnement avant leur implémentation.

Nous sommes ensuite passées à la conception de notre solution et avons choisi le processus CRISP-DM pour nous guider dans notre démarche de résolution.

Nous avons alors récupéré la base de données des déclarations et nous l'avons traité, puis nous avons consacré une grande partie de notre travail à l'enrichir en concevant des variables qui expliquent le comportement des fraudeurs avec la collaboration des experts. Une fois prête, nous avons utilisé cette base pour la construction de différents algorithmes d'apprentissage supervisé censés discerner entre les déclarations frauduleuses et non frauduleuses. Nous les avons évalués puis comparés et avons fini par sélectionner l'algorithme de Random Forest pour ensuite le modifier afin qu'il détecte plus de fraudes quitte à détériorer ses performances globales.

Random Forest étant désormais plus méfiant à l'égard de la fraude, nous nous en sommes servis pour redéfinir cette dernière variable en y ajoutant les fraudes détectées par l'algorithme - même si ces transactions étaient au préalable définies comme non frauduleuses.

Nous avons ensuite entrainé Random Forest avec cette nouvelle base mais avec pour variable cible les différents circuits de vérification initiaux (rouge, orange et vert) auxquels nous avons ajouté sous les orientations du directeur des renseignements de la DGD un quatrième circuit (jaune), qui consiste à vérifier les documents accompagnant la transaction ainsi que soumettre la marchandise à l'inspection au scanner afin de détecter d'éventuelles anomalies.

Nous avons testé ce modèle avec l'ensemble des déclarations de l'année 2018 et avons pu constater que nous avons atteint de manière satisfaisante les objectifs fixés par la douane.

En effet, 45% des déclarations seulement sont affectées au circuit rouge tandis que 20% le sont au circuit vert, ce qui correspond aux proportions que l'on nous avait chargées d'atteindre. De plus seulement 1% des fraudes sont présentes dans le circuit vert tandis que le circuit rouge en cumule 93%. Ceci indique qu'elles devraient pour la majorité écrasante être interceptée au moment de la vérification. De plus, le circuit rouge étant désormais moins engorgé, les vérifications seront menées plus efficacement et plus de fraudes pourront être interceptées.

Lors des différents entretiens que nous avons menés avec l'équipe de la gestion du risque, on nous a confié que la base de données disponible n'était pas fiable. En effet, plusieurs raisons (congestion du circuit rouge, corruption,…) portent à croire qu'une certaine partie des déclarations serait faussement déclarée comme étant non frauduleuse.

Ce genre de transaction douteuse ne peut être toujours détecté par un algorithme d'apprentissage supervisé, car la base d'apprentissage est ainsi faussée, ce qui constitue d'ailleurs une limite au modèle que nous avons proposé. Pour cette raison, nous avons eu recours aux techniques de l'apprentissage non supervisé afin essayer de cerner un comportement commun entre les fraudeurs et avons pu générer des clusters assez représentatifs des importateurs. Cette segmentation permettra détecter des fraudes inédites et d'appuyer le modèle supervisé.

Le travail que nous avons effectué a pour objectif de défendre les intérêts économiques de l'Algérie en aidant les douanes à mieux intercepter la fraude et à augmenter les recettes douanières de l'état tout en améliorant la qualité de services des douanes auprès des opérateurs économiques, et d'encourager le commerce extérieur.

Néanmoins des perspectives restent à explorer, et nous recommanderons donc d'améliorer continuellement les modèles que nous avons proposés en intégrant de nouvelles variables pertinentes et en rafraichissant l'apprentissage qui a été effectué en ajoutant à la base d'entrainement les nouvelles fraudes ayant été détectées.

Un autre aspect du travail, qui devra être mené dans le futur, consistera à confirmer que les déclarations identifiées comme potentiellement frauduleuses par le modèle non supervisé, le sont effectivement. L'hypothèse centrale selon laquelle les fraudeurs auraient des comportements similaires devra par conséquent également être validée. Suite à cela, des profils à risques devront être identifiés avec l'ensemble des informations qui les caractérisent afin de comprendre le comportement de la fraude. Ce modèle viendra alors en appui à l'outil d'aide à la décision que nous avons proposé en alimentant la base de données des fraudes.

Bibliographie

Afifi, Abdelmonem. 2011. *Practical Multivariate Analysis (Chapman & Hall/CRC Texts in Statistical Science).* s.l. : Chapman and Hall/CRC, 2011. p. 537. ISBN :978-1439816806.

Arthur, David and Vassilvitskii, Sergei. 2009. Worst-Case and Smoothed Analysis of the ICP Algorithm, with an Application to the k-Means Method. *en ligne.* 2009, Vol. 13, pp. 766-782 .

Breiman, Leo and Friedman, Jerome. 1984. *Classification and Regression Trees (Wadsworth Statistics/Probability).* s.l. : Chapman and Hall/CRC; 1 edition, 1984. p. 368. ISBN : 978-0412048418.

Breiman, LEO. 1996. Bagging Predictors. *en ligne.* 1996, p. 18.

Brink, Henrik and Richards, Joseph. 2017. *Real World Machine Learning.* s.l. : Manning Publications, 2017. ISBN : 9781617291920.

Chen, Tianqi and Guestrin, Carlos. 2016. XGBoost: A Scalable Tree Boosting System. *en ligne.* 2016, pp. 785-794.

Chiboubi, Meziane and Harrouche, Dahman. 2013. *Impact des facilitations douaniéres sur la promotion du commerce extérieur UNIVERSITE ABDERRAHMANE MIRA DE BEJAIA.* 2013. p. 88.

Cornuéjols, Antoine and Miclet, Laurent. 2003. *Apprentissage artificiel.* s.l. : Eyrolles, 2010, 2003. p. 803. ISBN :782212124712.

Cover, T. M. and Hart, P. E. 1967. Nearest Neighbor Pattern Classification. *en ligne.* 1967, pp. 21-27.

Décret exéctif n° 17-90. **2017.** 2017, JOURNAL OFFICIEL DE LA REPUBLIQUE ALGERIENNE N° 13, p. 9.

Décret exéctif n°12-93. **2012.** 2012, JOURNAL OFFICIEL DE LA REPUBLIQUE ALGERIENNE.

Devijver, Pierre A. and Kittler, Josef. 1982. Pattern Recognition : A Statistical Approach. *en ligne.* 1982.

Dietterich, Thomas G. 2000. *An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization.* s.l. : Kluwer Academic Publishers, 2000. ISBN : 1007607513941.

Dobson, Annette J. 2008. *An Introduction to Generalized Linear Models, Third Edition (Chapman & Hall/CRC Texts in Statistical Science).* s.l. : Chapman and Hall/CRC, 2008. p. 320. ISBN : 978-1584889502.

Dufour &al . 2016. Auto-encodeurs pour la compréhension de documents parlés. *en ligne* . 2016, pp. 73-80.

Efron, Bradley. 1993. *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics and Applied Probability).* s.l. : Chapman and Hall/CRC, 1993. p. 456. ISBN :978-0412042317.

Ester, Martin and Kriegel, Hans-Peter. 1996. A Density-Based Algorithm for Discovering Clusters. 1996, pp. 226–231.

Forgy, E.W. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. 1965, pp. 768–769.

Freund, Yoav and Schapire, Robert. 1996. Experiments with a New Boosting Algorithm. *en ligne.* January 22, 1996, p. 16.

Friedman, Jerome. 1999. Stochastic Gradient Boosting. *en ligne.* March 1999, p. 10.

Gacôgne, Louis. 2015. *Intelligence artificielle Cours, exercices et projets.* s.l. : Ellipses, 2015.

Géron, Aurélien. 2017. *Machine Learning avec Scikit-Learn.* s.l. : Dunod, 2017. p. 256 .

Groupe de la Banque mondiale. 2019. *DOING BUSINESS 2019 Training for Reform.* washington : A World Bank Group Flagship Report, 2019.

Guyon et al. 1992. A Training Algorithm for Optimal Margin Classifiers. *en ligne.* 1992, p. 9.

Jambu, Michel. 1999. *Introduction au Data Mining.* s.l. : Editions Eyrolles et France Télécom-CENT, 1999. ISBN :2-212-05255-3.

Johnson, Kjell and Kuhn, Max. 2016. *Applied Predictive Modeling.* 2016. ISBN :978-1-4614-6848-6.

Kassambara, Alboukadel. 2017. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning.* 2017. p. 188. ISBN : 978-1542462709.

Kelleher, John D. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.* s.l. : The MIT Press, 2015. p. 624. ISBN : 978-0-262-02944-5.

Kohonen, Teuvo. 1982. Self-Organized Formation of Topologically Correct Feature Maps. *en ligne.* 1982, pp. 59-69.

La loi n° 17-04. **2017.** 2017, JOURNAL OFFICIEL DE LA REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE N°11, Vol. CHAPITRE 6, p. 16.

Lafaye, Pierre. 2017. *Le logiciel R Maitriser le langage Effectuer des analyse (bio)statistiques.* s.l. : Springer, 2017. p. 674. ISBN :978-2746248182.

Lance, G. N and Williams, W T. 1967. A General Theory of Classificatory Sorting. *en ligne.* February 1967, Vol. 9, pp. 373–380.

Lefébure, René and Venturi, Gilles . 2001. *Le Data Mining.* s.l. : Eyrolles, 2001. ISBN : 978-2212091762.

Maimon, Oded Z. *Data Mining with Decision Trees: Theroy and Applications (Machine Perception and Artificial Intelligence).* s.l. : World Scientific Pub Co Inc. p. 244. ISBN :978-9812771711.

MILIANI, Amar. 2001. *gestion des risques dans les contrôles douaniers, mémoire de fin d'étude, Institut d'Economie Douanière et Fiscale.* Tipaza : s.n., 2001. p. 128.

Moussaoui, Hanane and Oumakhlouf, Kenza. 2017. *Etude de l'impact des facilitations douanières à l'importation sur la performance de l'entreprise :cas de CEVITAL.* Université Abderrahmane Mira de Béjaïa. 2017.

Organisation Mondiale des Douanes. 2013. Recueil de l'OMD sur la gestion des risques en matière douanière. *World Customs Organization.* [Online] 2013. http://www.wcoomd.org/fr.aspx.

Pacheco, Erik Rodrigues. 2015. *Unsupervised Learning with R.* s.l. : Packt Publishing - ebooks Account, 2015. p. 192. ISBN :978-1785887093.

Quinlan, J. Ross. 1992. *Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning) 1st Edition.* s.l. : Morgan Kaufmann, 1992. p. 302. ISBN : 978-1558602380.

Tufféry, Stéphane. 2012. *Data mining et statistique décisionnelle.* Paris : Editions technip, 2012. p. 826. ISBN : 978-2710810179.

Van der Maaten, Geoffrey and Hinton, Geoffrey. 2008. Visualizing Data using t-SNE. [ed.] Yoshua Bengio. *en ligne.* 2008, pp. 2580-2605.

World Customs Organization. *Wikipedia.* [Online] [en ligne]. [Cited: Avril 12, 2019.]

ANNEXES

Annexe A : La nomenclature tarifaire

SECTION I — **ANIMAUX VIVANTS ET PRODUITS DU REGNE ANIMAL**

Notes de section

Chapitre 1	Animaux vivants.
Chapitre 2	Viandes et abats comestibles.
Chapitre 3	Poissons et crustacés, mollusques et autres invertébrés aquatiques.
Chapitre 4	Laits et produit de la laiterie; oeufs d'oiseaux; miel naturel; produits comestibles d'origine animale, non dénommés ni compris ailleurs.
Chapitre 5	Autres produits d'origine animale, non dénommés ni compris ailleurs.

Chapitre 1

Animaux vivants

Note.

1.- Le présent Chapitre comprend tous les animaux vivants, à l'exclusion :

a) des poissons et des crustacés, des mollusques et des autres invertébrés aquatiques, des n°s 03.01, 03.06 ou 03.07 ;

b) des cultures de micro-organismes et des autres produits du n° 30.02 ;

c) des animaux du n° 95.08.

Notes complémentaires :

1- Sont admis dans les sous positions intitulées « destinés aux parcs zoologiques », les animaux vivants destinés aux parcs zoologiques nationaux ou des collectivités locales, importés ou exportés à leur ordre ou pour leur compte et conduits directement à ces parcs.

2- Tout cheval de moins d' 1,51 m au garrot (ou 1,52 m ferré) est classé « poney ».

3- Les antilopes de la sous-famille Bovinae du n° 0102.90.91.00, s'entendent l'antilope tétracère (Tetracerus quadricornis) et l'antilope à cornes spiralées des genres Taurotragus et Tragelaphus. Les antilopes autres que de la sous-famille Bovinae relèvent du n° 01.06.

Position & Sous Position	U.Q.N	Désignation des Produits	D.D
01.01		**Chevaux, ânes, mulets et bardots, vivants.**	
		- Chevaux :	
		- - Reproducteurs de race pure :	
		- - - De course :	
0101.21.11.00	U	- - - - De pur-sang arabe	5
0101.21.19.00	U	- - - - Autres que de pur-sang arabe	5
		- - - Autres que de course :	
0101.21.91.00	U	- - - - De pur-sang arabe	5
0101.21.99.00	U	- - - - Autres	5
		- - Autres :	
0101.29.10.00	U	- - - De course	30
0101.29.20.00	U	- - - Pour abattage	30
		- - - Pour parcs zoologiques :	
0101.29.31.00	U	- - - - Etalons et hongres	30
0101.29.32.00	U	- - - - juments	30
0101.29.33.00	U	- - - - Poulains et pouliches	30
0101.29.34.00	U	- - - - Poneys et ponettes	30
		- - - Autres :	
0101.29.91.00	U	- - - - Etalons et hongres	30
0101.29.92.00	U	- - - - juments	30
0101.29.93.00	U	- - - - Poulains et pouliches	30

Figure A.1 : Nomenclature tarifaire

Annexe B : Extrait de la base de données de la DGD

NUM_FISCA	DATEH	CB	AN_DECL	CODE_BNQ	CODE_REG	NUM_DECL	TO2	PAYS_FRS	PAY_PROV	NCONTRAT	DATE_NRC	FRAUDE_DEC
1001[illegible]	2019-01-15 08:46	10	2019	162702	1000	940	1	321	321	FOB	11/02/2010	0
1001[illegible]	2019-01-23 11:02	10	2019	162702	1000	836	1	321	321	FOB	11/02/2010	0
1001[illegible]	2019-01-23 11:04	10	2019	162702	1000	840	1	321	321	FOB	11/02/2010	0
1001[illegible]	2019-01-23 11:20	10	2019	162702	1000	841	1	321	321	FOB	11/02/2010	0
1002[illegible]	2019-02-04 14:22	10	2019	20601	1000	2910	1	358	358	CFR	14/04/2010	0
1002[illegible]	2019-03-03 13:27	10	2019	21901	1000	3019	1	532	532	CFR	14/04/2010	0
1002[illegible]	2019-01-21 13:35	12	2019	162007	1000	4279	2	321	336	FOB	23/12/2010	0
1002[illegible]	2019-03-18 15:09	12	2019	162007	1000	17717	2	336	321	FOB	23/12/2010	0
1002[illegible]	2019-01-06 15:39	10	2019	160301	1000	188	4	525	525	FOB	08/08/1988	0
1002[illegible]	2019-01-06 15:58	79	2019	160301	1000	351	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-01-24 14:25	79	2019	160301	1000	1464	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-02-03 15:44	79	2019	160301	1000	1974	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-02-20 14:51	79	2019	160301	1000	7018	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-03-13 14:27	12	2019	160301	1000	16718	4	525	525	FOB	08/08/1988	0
1002[illegible]	2019-03-18 15:51	79	2019	160301	1000	4484	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-03-19 14:13	79	2019	160301	1000	4607	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-03-20 16:14	79	2019	160301	1000	4730	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-03-28 10:07	79	2019	160301	1025	5271	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-03-28 10:12	79	2019	160301	1000	5272	2	525	525	FOB	08/08/1988	0
1002[illegible]	2019-03-28 10:18	79	2019	160301	1000	5273	2	525	525	FOB	08/08/1988	0

Figure B.1 : Base de données des déclarations

Annexe C : La boite à moustache

La boite à moustache est une façon normalisée d'afficher la distribution des données en fonction de quelques indicateurs de position du caractère étudié à savoir : le minimum, le premier quartile (Q1), la médiane, le troisième quartile (Q3), et le maximum.

La boite à moustache comprend un rectangle central s'étendant du premier quartile au troisième quartile. Un segment à l'intérieur du rectangle indique la médiane et les "moustaches" au-dessus et au-dessous de la case indiquent l'emplacement du minimum et du maximum. La figure suivante illustre les différents composants d'une boite à moustache :

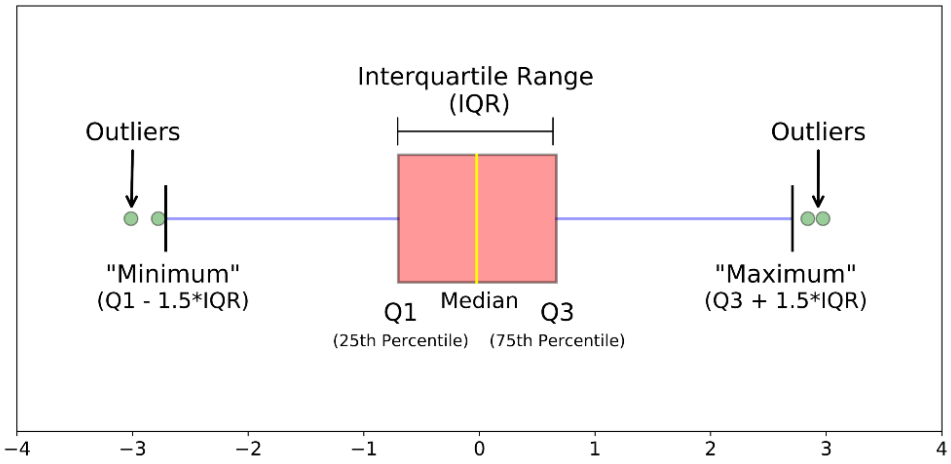

Figure C.1 : Boite à moustache

- La médiane Q2 est la valeur moyenne de l'ensemble de données.
- Le premier quartile Q1 est le nombre moyen entre le plus petit nombre et la médiane de l'ensemble de données.
- Le troisième quartile Q3 est la valeur moyenne entre la valeur médiane et la valeur la plus élevée de l'ensemble de données.
- L'écart interquartile (IQR) s'étend de Q1 à Q3.
- Les moustaches (en bleu) représentent le maximum et le minimum.
- Les valeurs aberrantes sont représentées par des cercles verts.

Annexe D : L'affectation des pays aux clusters

Résultats par objet :									
Code pays	Classe	Code pays	Classe	Code pays	Classe	Code pays	Classe	Code pays	Classe
558	8	74	1	219	1	343	1	505	1
550	7	80	1	221	1	349	1	510	1
580	7	92	1	224	1	355	1	514	1
597	7	94	1	227	1	358	1	518	1
532	6	96	1	228	1	359	1	520	1
525	5	107	1	231	1	361	1	522	1
331	4	108	1	233	1	362	1	527	1
508	4	110	1	240	1	366	1	529	1
321	3	113	1	243	1	369	1	536	1
186	2	116	1	250	1	371	1	538	1
236	2	121	1	253	1	377	1	540	1
327	2	126	1	255	1	378	1	544	1
336	2	131	1	261	1	380	1	547	1
337	2	135	1	269	1	381	1	552	1
586	2	139	1	272	1	384	1	553	1
592	2	140	1	275	1	387	1	554	1
25	1	145	1	277	1	388	1	563	1
28	1	165	1	279	1	390	1	567	1
31	1	168	1	280	1	394	1	570	1
36	1	171	1	287	1	400	1	573	1
38	1	174	1	291	1	401	1	576	1
39	1	177	1	302	1	405	1	578	1
43	1	180	1	306	1	409	1	583	1
46	1	183	1	307	1	409	1	587	1
49	1	195	1	308	1	411	1	589	1
54	1	200	1	310	1	415	1	590	1
56	1	202	1	313	1	416	1	594	1
58	1	206	1	317	1	419	1	601	1
61	1	209	1	319	1	501	1	603	1
65	1	210	1	324	1	502	1	607	1
68	1	213	1	335	1	503	1	710	1
71	1	216	1	340	1	504	1	711	1

Figure D.1 : Classe pays

Annexe E : La visualisation des relations être les variables explicatives et la variable cible « fraude » représentée par le diagramme en barres

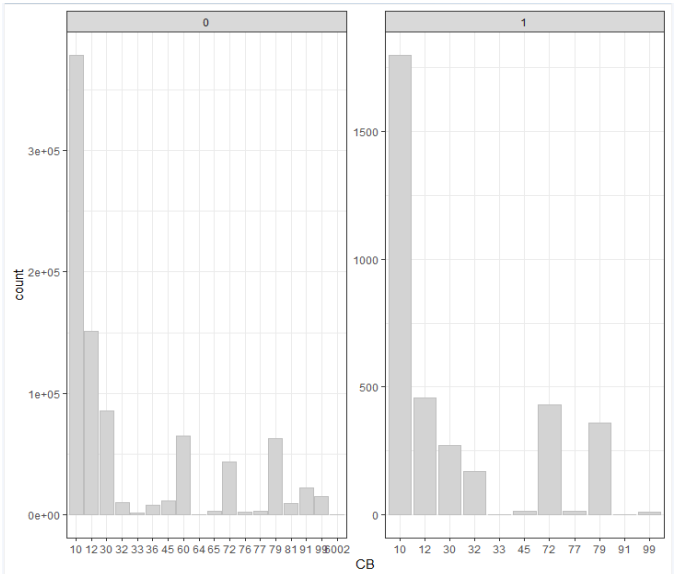

Figure E-1 : Relation entre la variable code bureau et la variable cible fraude

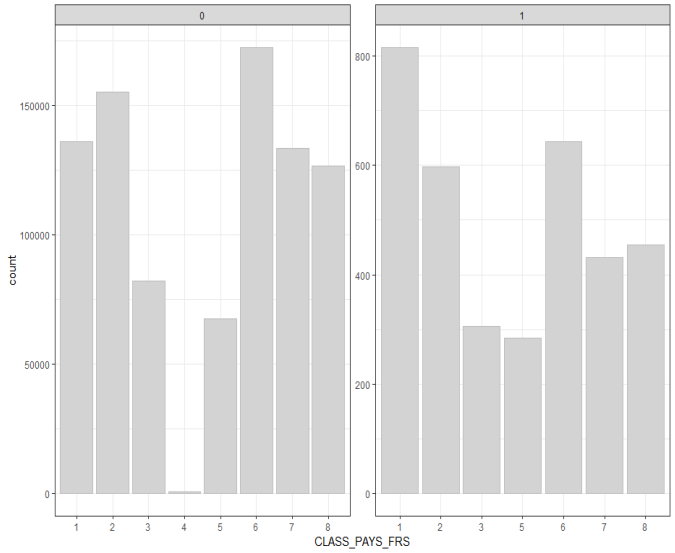

Figure E-2 : Relation entre la variable pays fournisseur et la variable cible fraude

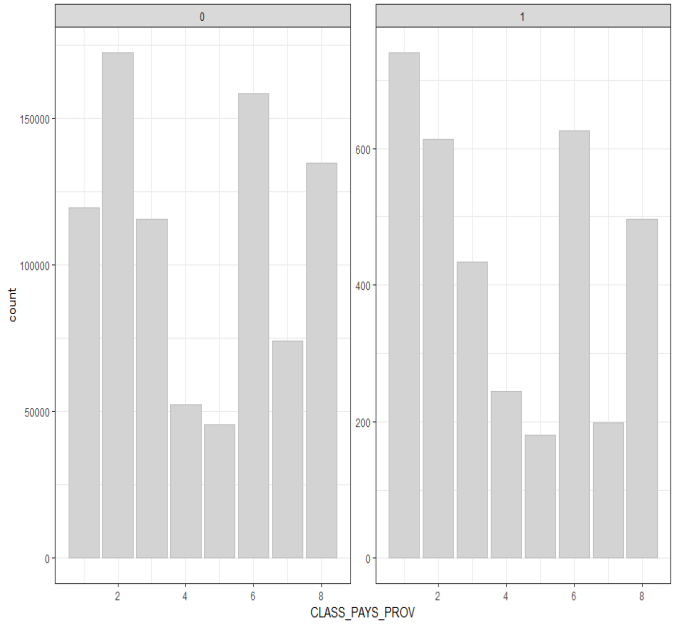

Figure E-3 : Relation entre la variable pays provenance et la variable cible fraude

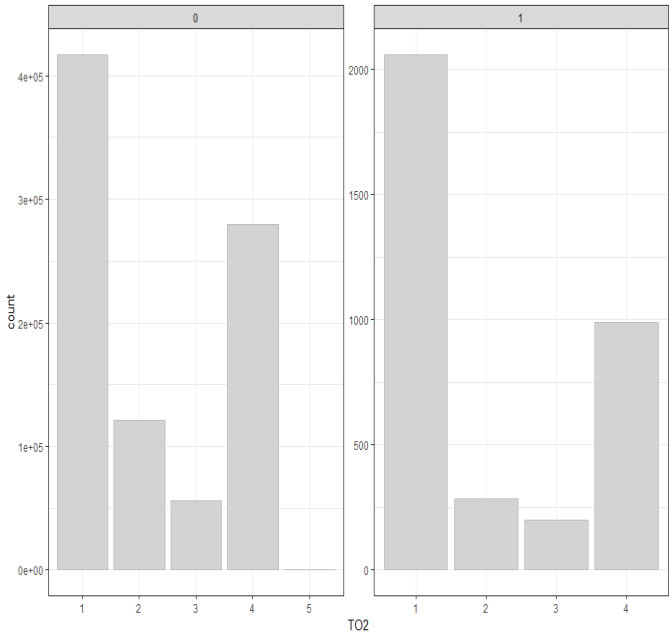

Figure E-4 : Relation entre la variable type d'opération et la variable cible fraude

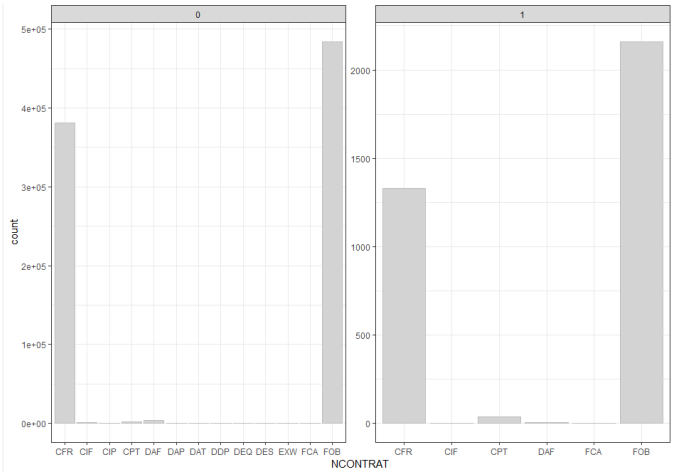

Figure E-3 : Relation entre la variable incoterm et la variable cible fraude

Annexe F : L'affectation de chaque Wilaya aux cinq régions

Region	Wilaya	CODE WILAYA	Region	Wilaya	CODE WILAYA
NORD	CHELF	2	OUEST	MOSTAGANEM	27
NORD	BLIDA	9	OUEST	MASCARA	29
NORD	BOUIRA	10	OUEST	ORAN	31
NORD	TIZI OUZOU	15	OUEST	EL BAYDH	32
NORD	ALGER	16	OUEST	NAAMA	45
NORD	DJELFA	17	OUEST	AIN TEMOUCHENT	46
NORD	MEDEA	26	OUEST	RELIZANE	48
NORD	BOUMERDES	35	EST	OUM EL BOUAGHI	4
NORD	TISSEMSILT	38	EST	BATNA	5
NORD	TIPAZA	42	EST	BEJAIA	6
NORD	AIN DEFLA	44	EST	BISKRA	7
SUD OUEST	ADRAR	1	EST	TEBESSA	12
SUD EST	LAGHOUAT	3	EST	JIJEL	18
SUD OUEST	BECHAR	8	EST	SETIF	19
SUD EST	TAMANRASSET	11	EST	SKIKDA	21
SUD EST	OUARGLA	30	EST	ANNABA	23
SUD EST	ILLIZI	33	EST	GUELMA	24
SUD OUEST	TINDOUF	37	EST	CONSTANTINE	25
SUD EST	EL OUED	39	EST	M'SILA	28
SUD EST	GHARDAIA	47	EST	BOURDJ BOU ARRE	34
OUEST	TLEMCEN	13	EST	EL TARF	36
OUEST	TIARET	14	EST	KHENCHLA	40
OUEST	SAIDA	20	EST	SOUK AHRAS	41
OUEST	SIDI BEL ABBES	22	EST	MILA	43

Figure F.1 : Régions des Wilaya

Annexe G : La corrélation de Pearson

Le coefficient de corrélation de Pearson est une mesure de la corrélation linéaire entre deux variables X et Y. Il varie entre +1 et -1 tel que s'il est égal à :

- 1 : il existe une forte relation linéaire positive entre les deux variables.
- 0 les deux variables ne sont pas corrélées linéairement.
- -1 il existe une forte relation linéaire négative entre les deux variables.

Il est calculé à l'aide de la formule : $r_{x,y} = \frac{cov(x,y)}{\sigma_x \times \sigma_y}$.

Avec $cov(x, y)$ étant la covariance entre x et y et σ_x, σ_y sont respectivement les écarts types de x et y.

Annexe H : Matrice de corrélation obtenue par le test de Pearson

Matrice de corrélation (Pearson (n)) :

Variables	DIF_CODE_PAYS	DIF_NUM_STAT	TOT_ART	DIF_TAUX_AVANT	AVG_TAUX_APPL	SUM_DIF_PAYS	PROP_DIF_PAYS	DATEH_NRC	LAG_AVG_TAUX_	LAG_TOT_ART	LAG_DATEH	LAG_DIF_TAUX_A	LAG_CB	LAG_CODE_REG	LAG_TO2	LAG_CODE_BNQ	LAG_NCONTRAT	LAG_SECTION	LAG_PAY_PROV	LAG_PAYS_FRS	LAG_ONE_CODE_	LAG_ONE_CODE_	CLASS_PAYS_PR	CLASS_PAYS_PR	CLASS_PAYS_PR	CLASS_PAYS_PR	CLASS_PAYS_PR	CLASS_PAYS_PR	CLASS_PAYS_FR	CLASS_PAYS_FR	CLASS_PAYS_FR	CLASS_PAYS_FR	CLASS_PAYS_FR	CLASS_PAYS_FR	CLASS_PAYS_FR	CLASS_ONE_COD	CLASS_ONE_COD
DIF_CODE_PAYS	1	0,510	0,566	-0,040	-0,052	0,173	-0,131	0,068	-0,004	0,244	-0,006	0,059	0,007	0,046	0,009	0,018	0,002	0,007	-0,001	-0,020	-0,014	0,011	-0,099	0,177	-0,
DIF_NUM_STAT	0,510	1	0,740	-0,107	-0,012	0,464	-0,035	-0,125	-0,036	0,342	0,065	0,014	0,014	-0,017	-0,042	-0,027	-0,037	0,028	-0,046	-0,061	-0,080	-0,072	0,054	0,027	-0,
TOT_ART	0,566	0,740	1	-0,072	-0,008	0,633	-0,023	-0,056	-0,023	0,382	0,086	0,034	0,030	0,004	-0,026	0,001	0,010	0,020	-0,008	-0,028	-0,045	-0,032	0,024	0,045	-0
DIF_TAUX_AVANT	-0,040	-0,107	-0,072	1	-0,785	-0,042	0,045	0,144	-0,366	-0,018	-0,070	0,195	0,024	0,169	0,064	0,058	0,009	0,038	0,016	0,055	0,029	0,094	-0,262	0,075	[
AVG_TAUX_APPL	-0,052	-0,012	-0,008	-0,785	1	0,000	0,268	-0,201	0,408	-0,038	0,061	-0,145	-0,066	-0,135	-0,112	-0,087	-0,037	-0,059	-0,052	-0,081	-0,038	-0,148	0,256	-0,090	-0,
SUM_DIF_PAYS	0,173	0,464	0,633	-0,042	0,000	1	0,268	-0,077	-0,017	0,251	0,115	0,025	0,013	0,013	-0,013	-0,010	0,001	0,046	-0,004	0,004	-0,053	-0,022	0,146	-0,031	-0
PROP_DIF_PAYS	-0,131	-0,035	-0,023	0,045	0,027	0,268	1	-0,066	-0,010	-0,001	0,080	-0,005	0,009	0,016	0,007	-0,015	0,015	0,008	-0,045	0,028	-0,029	-0,068	0,257	-0,125	0,
DATEH_NRC	0,068	-0,125	-0,056	0,144	-0,201	-0,077	-0,066	1	-0,019	-0,012	-0,124	0,108	0,061	0,100	0,102	0,104	0,072	0,098	0,132	0,176	0,120	0,211	-0,197	0,173	-0
LAG_AVG_TAUX_	-0,004	-0,036	-0,023	-0,366	0,408	-0,017	-0,010	-0,019	1	-0,047	0,001	-0,045	0,004	-0,001	0,002	-0,007	0,031	0,001	0,002	-0,019	-0,009	0,001	0,003	-0,006	-0,
LAG_TOT_ART	0,244	0,342	0,382	-0,018	-0,038	0,251	-0,001	-0,012	-0,047	1	0,008	0,018	0,031	0,004	-0,001	0,010	0,000	0,014	0,021	0,019	0,010	0,005	0,001	0,012	-0
LAG_DATEH	-0,006	0,065	0,086	-0,070	0,061	0,115	0,080	-0,124	0,001	0,008	1	-0,025	0,064	-0,006	0,001	-0,033	-0,024	0,089	-0,025	0,017	-0,024	-0,029	0,081	-0,039	-0
LAG_DIF_TAUX_A	0,059	0,014	0,034	0,195	-0,145	0,025	-0,005	0,108	-0,045	0,018	-0,025	1	0,138	0,638	0,231	0,117	0,089	0,199	0,149	0,178	0,217	0,249	0,081	0,070	0,
LAG_CB	0,017	0,014	0,030	0,024	-0,066	0,013	0,009	0,061	0,004	0,031	0,064	0,138	1	0,228	0,277	0,190	0,261	0,227	0,343	0,305	0,230	0,270	-0,080	-0,014	0
LAG_CODE_REG	0,046	-0,017	0,004	0,169	-0,135	0,013	0,016	0,100	-0,001	0,004	-0,006	0,698	0,228	1	0,306	0,169	0,133	0,198	0,214	0,238	0,217	0,316	-0,142	0,076	0,
LAG_TO2	0,009	-0,042	-0,026	0,064	-0,112	-0,013	0,007	0,102	0,002	-0,001	0,001	0,231	0,277	0,306	1	0,190	0,180	0,204	0,200	0,232	0,242	0,223	-0,065	0,001	0,
LAG_CODE_BNQ	0,018	-0,027	0,001	0,058	-0,087	-0,010	-0,015	0,104	-0,017	0,010	-0,033	0,117	0,190	0,169	0,190	1	0,134	0,144	0,225	0,245	0,160	0,213	-0,057	0,041	0
LAG_NCONTRAT	0,002	-0,037	0,010	0,009	-0,037	0,001	0,015	0,072	0,031	0,000	-0,024	0,089	0,261	0,133	0,180	0,134	1	0,170	0,265	0,273	0,196	0,241	-0,059	0,014	-0,
LAG_SECTION	0,007	0,028	0,020	0,038	-0,059	0,046	0,008	0,098	0,001	0,014	-0,025	0,199	0,227	0,198	0,204	0,144	0,170	1	0,253	0,270	0,611	0,270	-0,032	0,056	-0,
LAG_PAY_PROV	-0,001	-0,046	-0,008	0,016	-0,052	-0,004	-0,045	0,132	0,002	0,021	0,015	0,149	0,349	0,214	0,210	0,225	0,285	0,253	1	0,684	0,287	0,591	-0,127	-0,054	-0,
LAG_PAYS_FRS	-0,020	-0,061	-0,028	0,055	-0,081	0,004	0,028	0,176	-0,019	0,019	-0,017	0,178	0,305	0,238	0,232	0,245	0,273	0,270	0,684	1	0,293	0,604	-0,097	0,007	-(
LAG_ONE_CODE_	-0,014	-0,080	-0,045	0,029	-0,038	-0,053	-0,029	0,120	-0,009	0,010	-0,024	0,217	0,230	0,217	0,242	0,160	0,196	0,611	0,287	0,293	1	0,331	-0,040	0,039	-0,
LAG_ONE_CODE_	0,011	-0,072	-0,032	0,094	-0,148	-0,022	-0,056	0,211	0,001	0,005	-0,029	0,249	0,270	0,316	0,223	0,213	0,241	0,270	0,591	0,604	0,331	1	-0,236	0,121	0,
CLASS_PAYS_PR	-0,099	0,054	0,024	-0,262	0,256	0,146	0,257	-0,197	0,003	0,001	0,081	-0,142	-0,080	-0,142	-0,065	-0,057	-0,059	-0,032	-0,127	-0,097	-0,040	-0,236	1	-0,205	-0
CLASS_PAYS_PR	0,177	0,027	0,045	0,075	-0,090	-0,031	-0,125	0,173	-0,006	0,012	-0,039	0,070	-0,014	0,076	0,001	0,041	0,014	0,056	-0,054	0,007	0,039	0,121	-0,205	1	-0
CLASS_PAYS_PR	-0,046	-0,028	-0,017	0,101	-0,050	-0,015	0,027	-0,001	-0,006	-0,014	-0,013	0,044	0,031	0,039	0,006	0,014	-0,008	-0,025	-0,006	-0,011	-0,023	0,006	-0,143	-0,160	
CLASS_PAYS_PR	-0,055	0,123	0,067	-0,141	0,153	-0,078	-0,269	-0,167	0,000	0,045	0,016	-0,097	0,007	-0,095	-0,071	-0,047	-0,042	-0,021	-0,031	-0,097	-0,023	-0,046	-0,119	-0,133	-0,
CLASS_PAYS_PR	0,044	-0,003	-0,019	0,095	-0,044	-0,059	-0,078	0,002	0,009	-0,033	0,001	0,007	0,037	0,017	0,029	-0,029	0,005	-0,003	0,060	0,011	0,020	0,058	-0,111	-0,124	-0,
CLASS_PAYS_PR	0,039	0,007	0,019	0,033	-0,059	0,082	0,177	0,042	0,023	0,033	0,014	0,071	0,083	0,100	0,071	0,027	0,062	0,057	0,078	0,086	0,070	0,120	-0,181	-0,214	-0
CLASS_PAYS_PR	-0,038	-0,074	-0,041	0,068	-0,093	-0,023	0,106	0,040	-0,026	-0,009	-0,044	0,012	-0,024	-0,006	0,000	0,007	0,004	-0,015	0,020	0,022	-0,002	0,012	-0,139	-0,156	-0
CLASS_PAYS_PR	-0,052	-0,091	-0,078	0,097	-0,060	-0,069	-0,100	0,052	-0,001	-0,029	-0,018	0,010	-0,026	-0,013	0,016	0,022	0,009	-0,037	0,074	0,047	-0,043	0,032	-0,188	-0,211	-0
CLASS_PAYS_FR	-0,113	0,135	0,069	-0,283	0,270	0,107	0,145	-0,259	-0,003	0,013	0,075	-0,169	-0,084	-0,172	-0,105	-0,091	-0,084	-0,031	-0,130	-0,191	-0,068	-0,307	0,631	-0,222	-0,
CLASS_PAYS_FR	0,161	0,023	0,047	0,062	-0,060	-0,025	-0,104	0,150	-0,007	0,021	-0,026	0,074	-0,006	0,083	0,017	0,032	0,035	0,043	-0,036	-0,020	0,035	0,120	-0,185	0,745	-0,
CLASS_PAYS_FR	-0,034	-0,036	-0,022	0,112	-0,059	0,016	0,149	0,017	-0,005	-0,013	-0,010	0,046	0,040	0,039	0,021	0,016	0,016	-0,013	0,000	0,021	-0,015	0,030	-0,106	-0,107	0,
CLASS_PAYS_FR	0,038	0,030	0,062	-0,013	0,027	-0,009	-0,030	0,012	0,008	0,034	-0,010	-0,005	0,007	0,006	0,004	0,005	-0,015	-0,005	-0,009	0,015	0,000	-0,010	-0,013	0,012	-0
CLASS_PAYS_FR	0,026	-0,001	-0,016	-0,039	0,021	-0,072	-0,227	-0,050	-0,003	-0,017	-0,003	-0,044	-0,014	-0,035	-0,012	-0,024	-0,010	-0,038	0,037	0,017	0,004	0,003	0,018	-0,111	-0,
CLASS_PAYS_FR	0,021	-0,010	0,009	0,068	-0,094	0,065	0,122	0,020	0,025	0,006	-0,006	0,065	0,067	0,089	0,062	0,014	0,037	0,028	0,059	0,062	0,028	0,094	-0,199	-0,139	-0
CLASS_PAYS_FR	-0,043	-0,071	-0,051	0,028	-0,047	-0,070	-0,096	0,078	0,002	-0,005	-0,052	0,020	-0,010	-0,015	0,014	0,019	-0,003	0,030	0,020	0,030	0,055	0,005	-0,079	-0,119	-0,
CLASS_PAYS_FR	-0,021	-0,071	-0,055	0,100	-0,064	-0,047	-0,009	0,051	-0,014	-0,020	0,018	0,019	0,018	0,020	0,010	0,038	0,018	-0,036	0,074	0,113	-0,045	0,084	-0,146	-0,127	-0,
CLASS_ONE_COD	0,632	0,274	0,287	-0,030	-0,087	0,071	-0,187	0,108	-0,014	0,251	0,024	0,054	0,038	0,048	0,024	0,016	0,019	0,037	0,029	0,026	0,022	0,087	-0,151	0,231	-0,
CLASS_ONE_COD	-0,139	0,129	0,055	-0,365	0,335	0,049	-0,026	-0,320	0,021	0,008	0,066	-0,188	-0,091	-0,189	-0,106	-0,094	-0,084	-0,063	-0,142	-0,175	-0,068	-0,336	0,708	-0,251	-0,

Annexe I : La courbe de ROC en fonction des différents paramètres de Random Forest

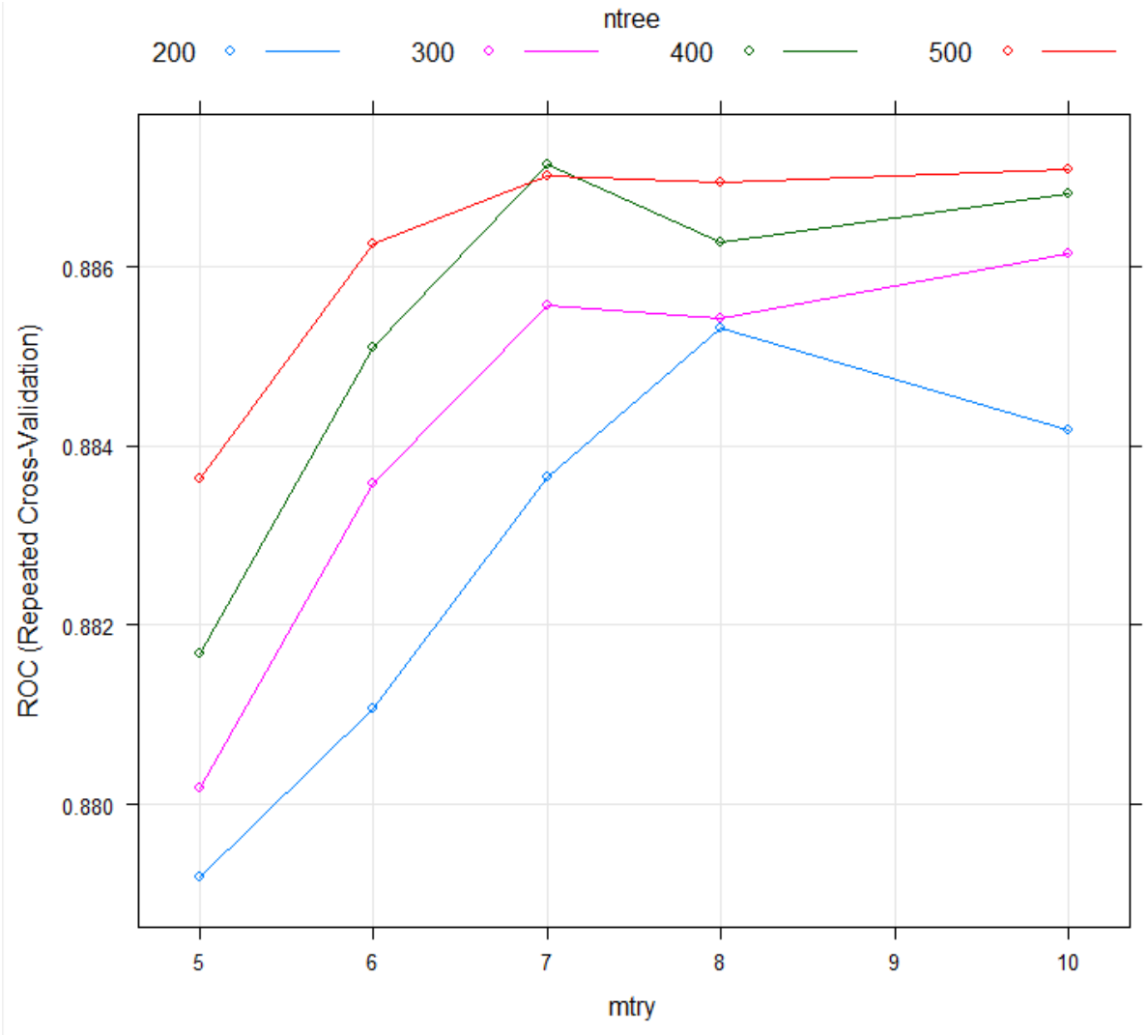

Figure I.1 : courbe Roc de Random Forest

Annexe J : Validation croisée à k plis (Cross Validation) (Devijver, et al., 1982)

Il s'agit d'une méthode d'échantillonnage qui consiste à partitionner les données d'apprentissage en k partitions. Une de ces partitions est choisie pour la validation alors que les $k-1$ autres sont utilisées pour l'entrainement. Cette procédure est répétée k fois jusqu'à ce que chaque partition soit utilisée comme ensemble de validation.

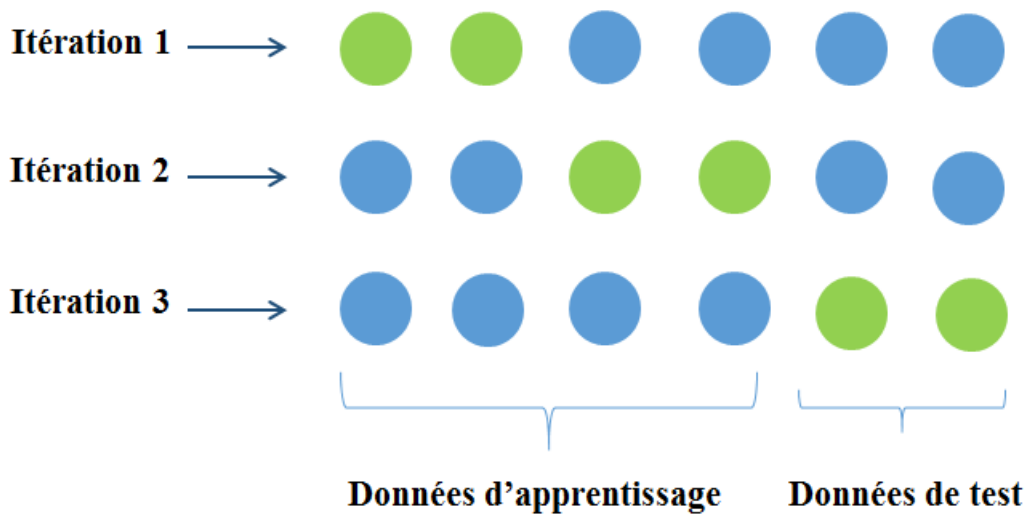

Figure J-1 : Exemple d'application de la validation croisée à 3 plis sur un ensemble de données de taille 6.

Annexe K : La courbe de ROC en fonction des différents paramètres de kNN

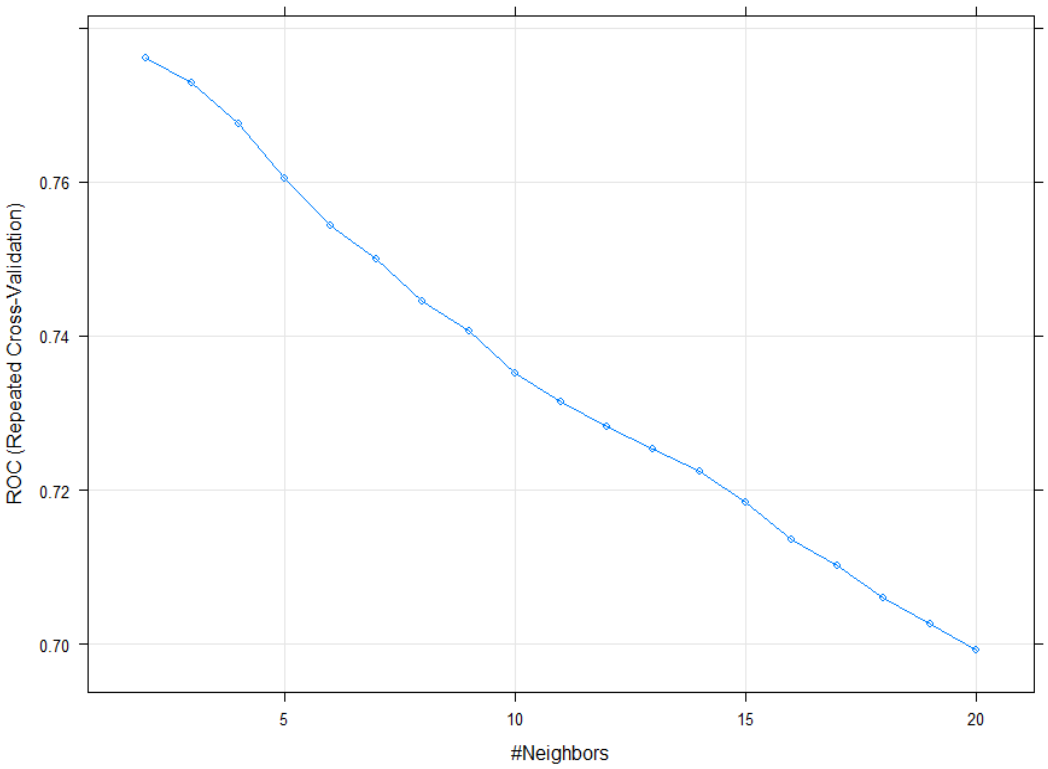

Figure K.1 : Courbe de Roc kNN

Annexe L : La courbe de ROC en fonction des différents paramètres de XGBoost

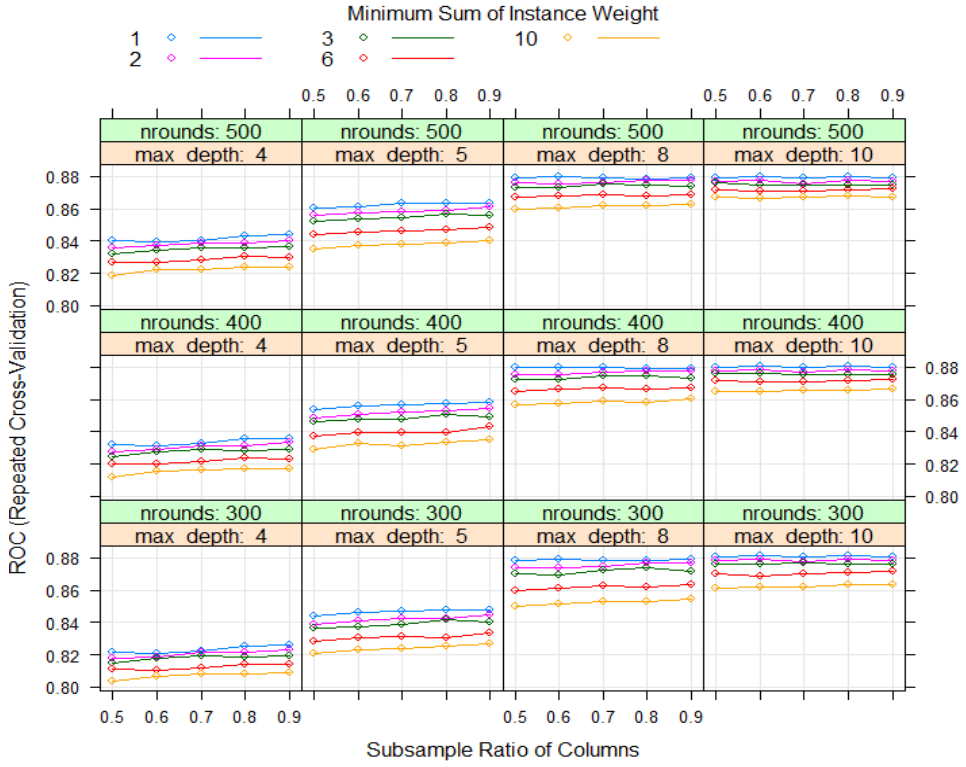

Figure L.1 : Courbe Roc XGBoost

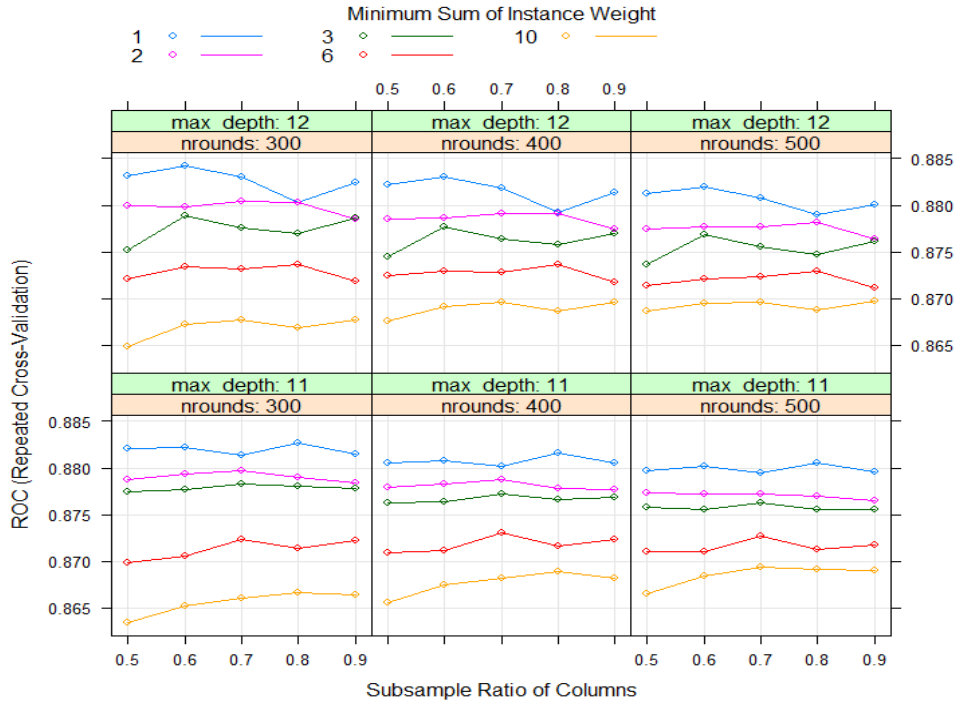

Figure L.2 : Courbe Roc XGBoost

Annexe M : Résultat de knndistplot

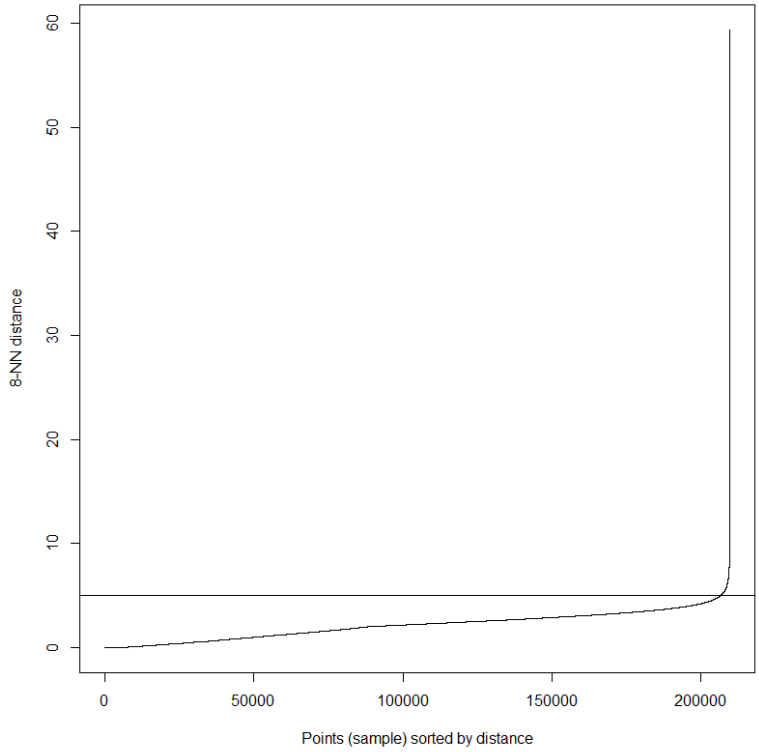

Figure M.1 : Cassure obtenue par kNN displot

Annexe N : L'algorithme T-SNE (Van der Maaten, et al., 2008)

L'algorithme t-SNE (t-distributed stochastic neighbor embedding) est une technique de réduction de dimension pour la visualisation de données développée par Geoffrey Hinton et Laurens van der Maaten en 2008.

Son objectif est de faire correspondre les distributions des distances entre les points en haute dimension avec leur distribution dans un espace à faible dimension grâce à des probabilités conditionnelles. Son principe de fonctionnement est le suivant :

Etant donné un ensemble de données N de grande dimension, le t-SNE calcule d'abord les probabilités p_{ij} qui sont positivement proportionnelles à la similarité entre deux objets x_i et x_j comme suit :

$$p_{j/i} = \frac{e^{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}}}{\sum_{k \neq i} e^{-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}}} \text{ et } p_{ij} = \frac{p_{j/i} + p_{i/j}}{2N}.$$

σ_i est associé à un paramètre appelé perplexité qui peut être interprété approximativement comme le nombre de voisins proches que chaque individu.

Il procède de la même manière avec les données y_i, cette fois ci en faible dimension et avec pour probabilité de similarité :

$$q_{ij} = \frac{\left(1 + \|x_i - x_j\|^2\right)^{-1}}{\sum_{k \neq l} (1 + \|x_i - x_l\|^2)^{-1}}$$

En fin de compte, l'algorithme minimise la divergence de Kullback-Leibler (KL) par rapport aux points en utilisant la descente du gradient avec :

$$KL(P\backslash\backslash Q) = \sum_{i \neq i} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right)$$

Annexe O : Le code de l'algorithme T-SNE

```
library(readr)
library(Rtsne)
library(ROSE)
library(plot3D)
library("scatterplot3d")

training_set <- read.csv(file.choose(), header = T)

sample(replace = )
training_set=training_set[,c("LAG_DATEH", "DATEH_NRC", "LAG_MAT_AGT", "AVG_TAUX_APPL", "TO2",
                             "LAG_AVG_TAUX_APPL", "DIF_CHAPITRE", "TOT_ART", "LAG_TOT_ART",
                             "LAG_CB", "LAG_PAYS_FRS", "LAG_PAY_PROV","FRAUDE")]
ech=sample(training_set,N=100000,replace =FALSE)
set.seed(1) # for reproducibility
tsne <- Rtsne(ech[,1:12], dims = 3, perplexity=70,
              verbose=TRUE, max_iter = 500,
              check_duplicates = FALSE,
              eta = 200)
z=as.data.frame(tsne$Y)
ech$FRAUDE <- as.factor(ech$FRAUDE)
v= as.data.frame( bind_cols(z,fraude=ech$FRAUDE))

colors <- c("#FF3399", "#E69F00")
colors <- colors[as.numeric(v$fraude)]
scatterplot3d(v[,1:3], pch = 18, color=colors ,angle = 200,grid=TRUE, cex.symbols =0.4,
              main="La visualisation en 3D  des \ndéclarations (TSNE)",
              xlab = "Tsne dim1",
              ylab = "Tsne dim2",
              zlab = "Tsne dim3")
legend("bottomright", legend=c("Fraude", "non Fraude"),
       col =  c("#FF3399", "#E69F00"), pch = 16)
```

Figure O.1 : Algorithme T-SNE

Annexe P : la courbe de l'inertie intra-classe

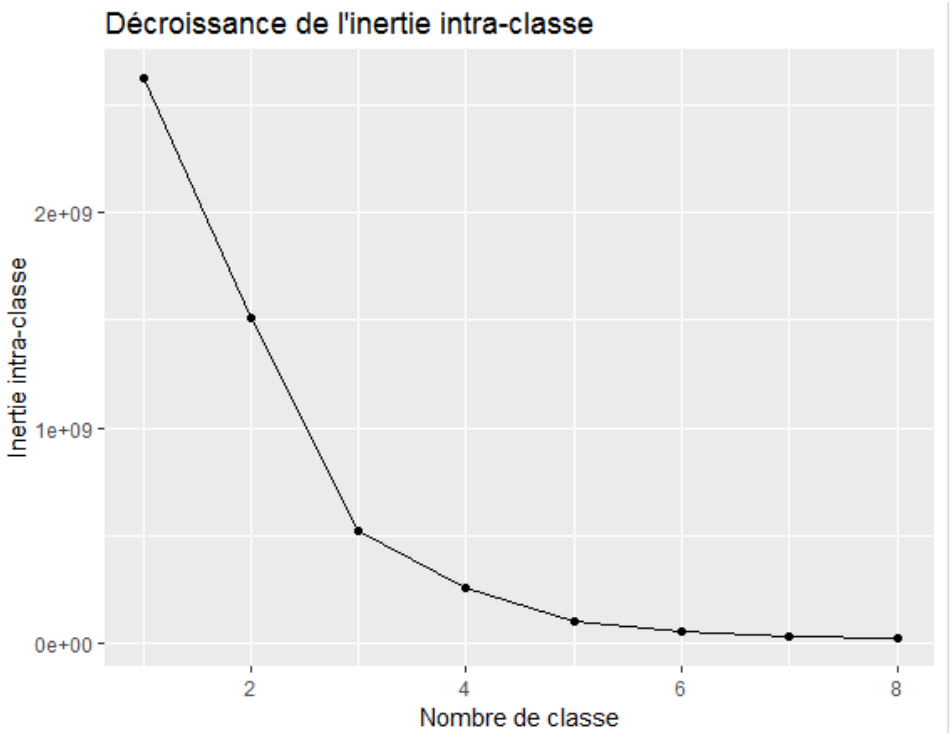

Figure P.1 : Variance intra-classe