

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA
RECHERCHE SCIENTIFIQUE ET DE L'INNOVATION

SÉCRÉTARIAT GÉNÉRAL

UNIVERSITÉ NAZI BONI



Ecole Supérieure d'Informatique

MEMOIRE DE MASTER EN INFORMATIQUE

OPTION: SYSTÈME D'INFORMATION - SYSTÈME D'AIDE À LA DÉCISION
(SI-SAD)

Année académique 2018-2019

Segmentation Géométrique et Photométrie d'images Acquises par Drones

Auteur:

Hoda DABONNE

Superviseur:

Pr. Serge MIGUET
Pr. Mihaela SCUTURICI
Pr. Tiguiane YELENOU



AVANT PROPOS

L'Université Nazi BONI a été créée le 23 mai 1997. Elle est située à quinze (15) kilomètres à l'Ouest de Bobo et est composée de six (06) établissements à savoir :

- L'Ecole Supérieure d'Informatique(ESI)
- L'Institut de Développement Rural(IDR)
- l'Institut Nationale des Sciences de la Santé(INSSA)
- l'Institut Universitaire de Technologie(IUT)
- l'UFR Sciences Juridique Politique Economique et Gestion(UFR SJPEG)
- l'UFR Sciences et Technologies(UFR ST)
- l'UFR Sciences Humaines, Lettres, Arts, et Communication(UFR SH-LAM)

L'ESI a pour mission d'accompagner le Burkina Faso dans sa transformation digitale. Pour cela, elle forme depuis sa création des cadres supérieurs en informatique qui font sa fierté dans toutes les administrations du pays.

Afin de préparer les futurs diplômés à l'insertion dans la vie professionnelle, l'ESI prévoit que chaque futur diplômé effectue un stage de fin de cycle dans une entreprise ou dans un centre de recherche ou de développement.

C'est dans ce contexte, que nous avons été reçu au LIRIS pour notre stage de fin de cycle. Le présent document fait office de mémoire des travaux que nous y avons menés.

DÉDICACE

A

Mes parents

Monsieur DABONNE Hanido

Et

Madame DABONNE Née BALIMA Habibou

Je ne vous remercerais jamais assez pour tout ce que vous avez fait pour moi au
quotidien. Soyez en remercié

Madame DABONNE née COMOE Maouwa

Que ton âme repose en paix.

A ma boussole

Monsieur DABONNE Ousseni

Ton acharnement et ta détermination au travail font de toi ma source d'inspiration et
mon modèle. Merci de toujours me montrer le chemin de la droiture.

REMERCIEMENTS

La redaction de ce memoire a été possible grace a plusieurs personne et institution à qui je voudrais temoigner ma gratitude et mes sincères reconnaissances :

Je tiens à remercier le programme Erasmus+ pour m'avoir donner cette chance de sejourner en France et de faire mon stage dans l'un des plus grands laboratioire francais.

Je remercier **Professeur Serge MIGUET** et **Professeur Mihaela SCUTURICI** mes tuteur de stage, pour les conseils et suggestions, combien précieux dans la réalisation de notre travail ;

Je desire remercier **Professeur Tigiane YELEMOU** mon tuteur accademique pour ces conseils et suggestions.

Un grand merci à **Professeur Pasteur PODA**, coordonateur de notre master pour son don de soit dans le bon deroulement de notre formation.

En fin, je tien à remercier l'ensemble du corps professoral et administratif de l'ESI pour toute la disponibilité et l'encadrement dont nous avons bénéficié.

A tous ceux qui ont contribué d'une manière ou d'une autre à la réalisation de ce travail, MERCI.

10mm10mm10mm10mm11pt11mm0pt11mm

Résumé

Résumé en français...

AbstractAbstract in english...

Numéro d'ordre :

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

NOMENCLATURE

AML	Anti Money Laundering
CAI	Caisse Autonome d'Investissement
CBS	Core Banking System
CFT	Contre le Financement du Terrorisme
CNDI	Caisse Autonome des Dépôts et des Investissements
GAFI	Groupe d'Action FInancière
IA	Intelligence Artificielle
KNN	K-Nearest Neighbors
KYC	Know Your Customer
LAB	Lutte Anti Blanchiment
OPI	OPérations Internationales
PEP	Personnes Exposées Politiquement
SG	Société Générale
SGBF	Société Générale Burkina Faso
SVM	Support Vector Machine
UREBA	Union Révolutionnaire de Banques

INTRODUCTION GÉNÉRALE

CHAPITRE 1

CONTEXTE GÉNÉRAL DE L'ÉTUDE

Introduction

Ce stage resulte de plusieurs collaborations. D'une part elle resulte d'une collaboration entre l'Université Nazi Bonin et L'université Lumière Lyon 2 dans le cadre du programme erasmus+ mobilité etudiantes. C'est dans ce cadre que nous avons été accueilli au dans les locaux du LIRIS au sein de l'Université Lumière Lyon 2. Ce sujet de stage intitulé «segmentation géométrique et photométrique d'images acquises par drones», resulte aussi d'une collaboration entre la société TECNI DRONE basée à Baix (Ardèche) et l'équipe Imagine du LIRIS (site de l'Université Lumière Lyon 2 à Bron). TECNI DRONE se spécialise dans la formation de pilot de drones et dans l'acquisition de données géométriques issues de mines et de carrières. Les campagnes d'acquisition d'images par drones, en milieu naturel, permettent de produire avec un très haut niveau de qualité, des modèles numériques de terrains texturés, porteurs à la fois d'informations géométriques et d'informations photométriques extrêmement riches. De nombreuses recherches ont été menées au cours des dernières années, pour affiner la qualité du traitement de ces données visuelles, et produire des maillages texturés de plus en plus fiables. En revanche, l'exploitation de ces données reste pour l'instant limitée soit à des traitements géométriques effectués sur les maillages 3D, soit à des traitements d'images 2D, effectués par exemple sur les ortho-images obtenues par assemblage des multiples vues partielles.

1.1 La structure d'accueil

1.1.1 Présentation

Présentation du LIRIS

Nous avons effectué notre stage au sein du Laboratoire d'InfoRmatique en Image et Systeme d'information . Le est une unité miste de recherche (5205) porté par

- le CNRS
- l'INSA de Lyon
- l'Université Claude Bernard Lyon 1
- l'Université Lumière Lyon 2
- l'Ecole Centrale de Lyon

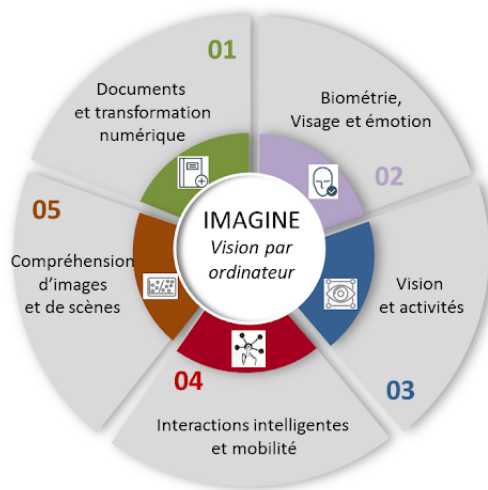
Il compte 330 membres, et a pour principal champ scientifique l'Informatique et plus généralement les Sciences et Technologies de l'Information. Une partie importante de la recherche effectuée au LIRIS s'étend à la frontière de notre discipline, au service de problématiques sociétales importantes. Certaines des ses activités de recherche se situent aux interfaces de l'ingénierie, des sciences humaines et sociales, des sciences de la vie et des sciences de l'environnement. L'ensemble des 6 pôles de compétences du LIRIS participe de façon équilibrée à la valorisation des travaux de recherche. Par ailleurs, le LIRIS entretient de nombreuses relations avec son environnement social, économique et culturel, aussi bien aux niveaux local et régional qu'au niveau national. Les interactions avec les entreprises s'établissent au travers de projets collaboratifs. Le LIRIS couvre des thématiques scientifiques structurées en 6 pôles de compétences regroupant 14 équipes.

- Simulation, virtualité & sciences computationnelles (Equipes Beagle, R3AM, SAARA)
- Géométrie & modélisation (Equipes GeoMod, M2DisCo)
- Science des données (Equipes BD, DM2L, GOAL)
- Vision intelligente & reconnaissance visuelle (Equipe Imagine)
- Interactions & cognition (Equipes SICAL, SMA, TWEAK)
- Services, systèmes distribués, sécurité (Equipes DRIM, SOC)

Les travaux des équipes de recherche trouvent aussi des applications dans les secteurs : Biologie et santé (modélisation du vivant, ingénierie pour la santé), Intelligence ambiante (systèmes pervasifs et distribués, monitoring intelligent, systèmes autonomes), Apprentissage humain (personnalisation, assistance cognitive, assistance à l'apprentissage collaboratif, jeux sérieux, loisirs numériques), Calcul scientifique (traitement de grandes masses de données – big data).

Presentation de l'équipe Imagine

Nous avons effectués notre stage au sein de l'équipe Imagine sur le site de l' université Lumière Lyon 2 à Bron. L'équipe Imagine du LIRIS est spécialisée dans la vision par ordinateur, l'apprentissage et la reconnaissance de formes. Elle réunit 21 membres permanents (8 PR, 3 MCF-HDR et 10 MCF), enseignants-chercheurs de l'Université Lyon 1, de l'Université Lyon 2, de l'École Centrale de Lyon et de l'INSA Lyon. En 2019, elle compte également parmi ses membres 29 doctorants et 17 post-doctorants. Les différentes activités de recherche menées dans l'équipe Imagine partagent les mêmes objectifs généraux visant la compréhension d'images multi-sources et multi-capteurs, intégrant ainsi une très large variété de contenus autour des images de personnes, d'objets et de scènes en 2D et 3D (scènes naturelles ou urbaines, images aériennes et satellites, visages. . .), des séquences d'images et des flux vidéos, ainsi que des documents numérisés (cartes, textes écrits et imprimés, partitions et symboles. . .). La notion d'objet visuel, au sens large, constitue ainsi un dénominateur commun de nos recherches. Les activités de l'équipe Imagine se déclinent en différentes thématiques liées à la mise en œuvre de méthodes d'indexation, de modélisation, de classification et de reconnaissance du contenu (objets, actions, concepts), avec une attention particulière portée au développement de méthodes d'apprentissage automatique pour la vision par ordinateur. Les recherches menées dans l'équipe Imagine visent à construire des passerelles pour franchir le fossé sémantique entre, d'une part, les informations de bas niveau présentes dans les données brutes (données échantillonnées issues du signal), les éventuelles données multi-sources issues d'autres modalités ou capteurs (cas notamment des applications embarquées) et les informations de plus haut niveau sémantique qui reposent sur la modélisation, la classification et l'identification des contenus. L'activité de recherche de l'équipe Imagine relève de 5 sous-thèmes majeurs qui constituent le cœur de ses applications.



1.1.2 Organisation

Organigramme

L'organigramme du LIRIS se présente comme suit : (voir Figure ??).

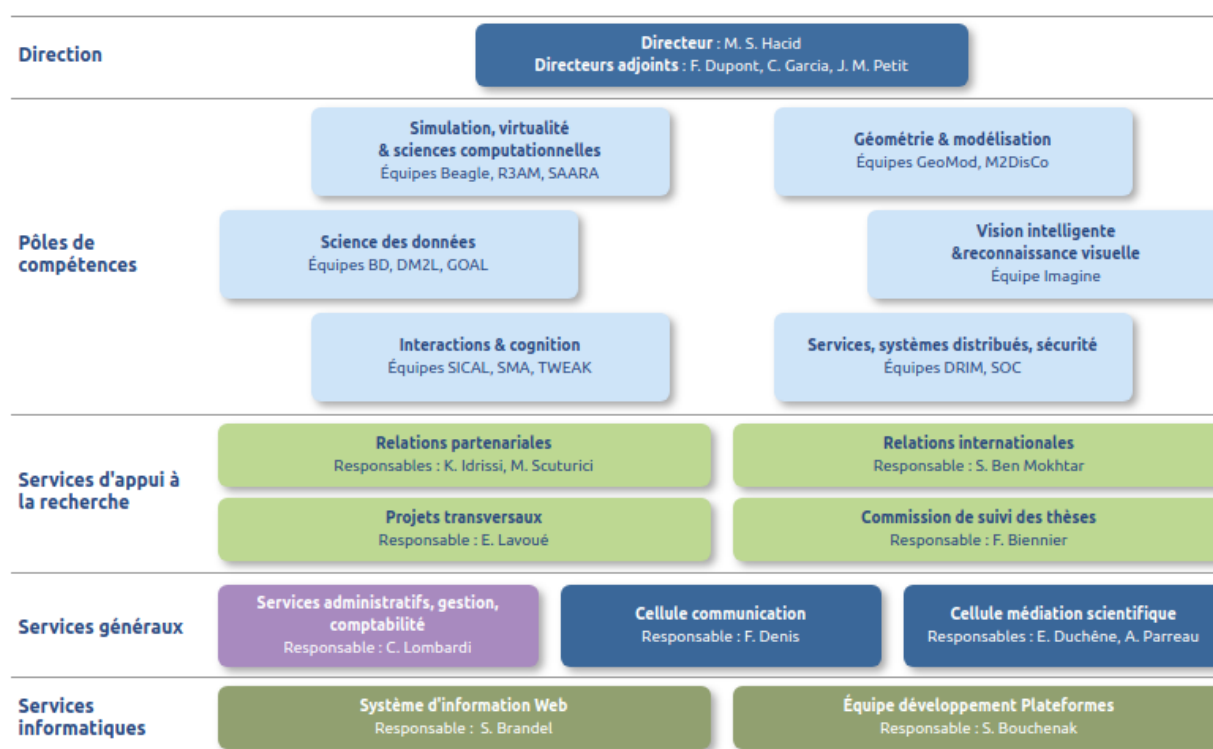


FIGURE 1.1 – Organigramme du LIRIS.

1.2 Présentation du sujet

1.2.1 Intitulé de sujet

Segmentation Géométrique et Photométrie d'images Acquisées par Drones

1.2.2 Contexte du sujet

1.2.3 Intérêt du sujet

Ce projet est très important pour Technidrone dans la mesure où elle vise à assister les techniciens de Technidrone dans leur tâche d'annotation par la même occasion réduire de façon significative le temps de travail de ces derniers. Les résultats attendus sont :

- Un gain en temps ;
- Identification automatique des formes géométriques dans les images ;
-

1.2.4 Problématique du sujet

L'exploitation de données recoltées lors des campagnes d'acquisition reste pour l'instant limitée soit à des traitements géométriques effectués sur les maillages 3D, soit à des traitements d'images 2D, effectués par exemple sur les ortho-images obtenues par assemblage des multiples vues partielles. En revanche, la qualification de ces données, indispensable aux exploitants, reste pour l'instant un travail essentiellement manuel. Il s'agit par exemple, d'identifier les ruptures de pentes qui délimitent les voies de roulement des engins, de calculer la largeur de ces voies, de délimiter les hauts et les bas de talus, de calculer le volume des tas correspondant aux matériaux extraits des carrières, etc. Ce processus qui est très important pour créer la carte d'une mine ou d'une carrière prend une dizaine d'heures pour un personnel entraîné. Une première étude effectuée au sein du laboratoire, axée essentiellement sur la géométrie contenue dans les maillages 3D a permis d'obtenir 62% de rappel et 10% de précision.

1.2.5 Objectifs

L'objectif de stage est d'utiliser d'une part les informations issues de la texture pour améliorer les résultats obtenus en se basant uniquement sur la géométrie, et d'autre part, d'utiliser de manière conjointe des informations issues des maillages, décrivant la géométrie de la scène, avec des données issues de la texture, portant des informations sur les discontinuités photométriques du terrain, pour assister les opérateurs dans leur tâche d'annotation des modèles numériques, construits à partir des images acquises par les drones. En se basant sur des terrains déjà annotés, et en entraînant des classifieurs à reconnaître les structures géométriques ou les motifs d'intérêt, nous voulons évaluer la capacité d'un système automatisé à effectuer cette identification avec un taux de succès le plus élevé possible. La tâche de l'opérateur se limiterait alors à la correction des inévitables erreurs de classification.

1.3 Concepts clés du sujet

1.3.1 Vocabulaire

La compréhension de certains concepts est indispensable à la compréhension du sujet.

Dossier de Transfert

Un dossier de transfert est l'ensemble des documents fournis par un client dans le but de l'exécution d'une opération à l'étranger. Ces documents sont de plusieurs types et sont principalement composés de :

Un ordre de virement : Il est donné par le propriétaire d'un compte bancaire qui doit payer une prestation ou un créancier ou faire un transfert. L'ordre de virement demande à la banque de débiter une somme de son compte pour créditer un autre compte. Le compte à créditer peut se trouver dans la même banque ou dans une autre banque. Ce document est obligatoire dans la réalisation d'une opération

Une autorisation de change : Il s'agit d'un document obligatoire dans la constitution d'un dossier de transfert à l'étranger. Ces opérations s'effectuant en devise, l'autorisation de change autorise le change vers la devise dans laquelle le transfert sera effectué.

La déclaration préalable d'importation : La Déclaration Préalable d'Importation (DPI) est une formalité accomplie au sein du ministère en charge du commerce préalablement à toute opération d'importation de marchandises dont la valeur FOB est supérieure ou égale à 500 000 FCFA.

L'autorisation spéciale d'importation ou d'exportation : Ces documents concernent des produits dont la liste est fixée par avis ministériel. De ces produits, nous pouvons citer le sésame, les céréales, les amande de Karité, le sucre...

Les documents justifiant l'opération : Il s'agit pour des achats de marchandise des factures par exemple, pour une inscription dans une école de l'attestation d'inscription et du passeport du concerné, pour le règlement d'un salaire du contrat de travail et du bulletin de paie

Les documents entrants en compte dans la constitution d'un dossier sont nombreux et les éléments cités ci-dessus sont loin d'être exhaustifs.

L'analyse d'une opération à l'étranger revient à analyser l'ensemble des informations contenues sur chacun de ces documents.

Pays étranger

Le terme étranger désigne tous les pays en dehors de l'UEMOA. Les transferts dans ces pays s'effectue en devise.

Selon la terminologie du règlement 09/2010/CM/UEMOA relatif aux relations financières extérieures des états membres de l'UEMOA,

Le terme étranger désigne tous les pays en dehors de l'UEMOA pour le contrôle de la position des établissements de crédit vis-à-vis de l'étranger ainsi que pour le traitement des opérations suivantes : domiciliation des exportations sur l'étranger et rapatriement du produit de leur recettes, émission et mise en vente de valeurs mobilières étrangères, importation et exportation d'or, opération d'investissement et d'emprunt avec l'étranger, exportation matérielle des moyens de paiement et de valeurs mobilières par colis postaux ou envois par la poste.

Résidents et Non-Résidents dans un Etat

Sont considérés comme résidents les personnes physiques ayant leur résidence habituelle dans l'Etat considéré. Sont considérés comme non-résidents ayant leur résidence habituelle à l'étranger.

Les opérations à l'étranger sont nombreuses et pour chaque type d'opération, les intervenants ou acteurs de la transaction sont différents.

1.3.2 Les différentes opérations à l'étranger

Plusieurs types d'opérations sont effectuées par le service des opérations internationales de la SGBF.

Les transferts émis

Il s'agit d'opérations émises par la banque résidente en l'occurrence la SGBF à destination d'une autre banque présente dans un autre pays. On distingue quatre intervenants dans une opération de transfert émis :

- L'émetteur de l'ordre qui est le donneur d'ordre
- La banque domiciliatrice de l'émetteur en l'occurrence dans notre cas la SGBF.
- La banque du bénéficiaire de l'ordre
- Le Bénéficiaire de l'ordre

Dans le cas de la SGBF et de toutes les filiales SG, il existe des hubs en l'occurrence SG New York pour les opérations en dollars et SG Paris pour toutes les autres devises. Les différents ordres sont envoyés vers ces hubs qui sont chargés de les acheminer vers les différentes banques bénéficiaires.

Les transferts reçus

Par transfert reçu, on entend tout virement en provenance de l'étranger à destination d'une banque résidente. Les transferts reçus s'effectue par transmission à la banque réceptrice d'un message SWIFT. Comme dans le cas des transferts émis nous distinguons quatre intervenants dans cette opération.

Les opérations de crédit documentaire

Le Crédit Documentaire est l'opération par laquelle une banque s'engage, à la demande et pour le compte de son client importateur, à régler à un tiers exportateur, dans un délai déterminé, un certain montant contre remise des documents strictement conformes et cohérents entre eux, justifiant de la valeur et de l'expédition des marchandises ou des prestations de services. On distingue quatre intervenants pour assurer la sécurité de l'opération :

- L'Acheteur/Importateur = Donneur d'ordre
- La Banque de l'Acheteur = Banque Emettrice
- La Banque du vendeur = Banque notificatrice et/ou Banque confirmatrice
- Le vendeur/L'Exportateur = Bénéficiaire

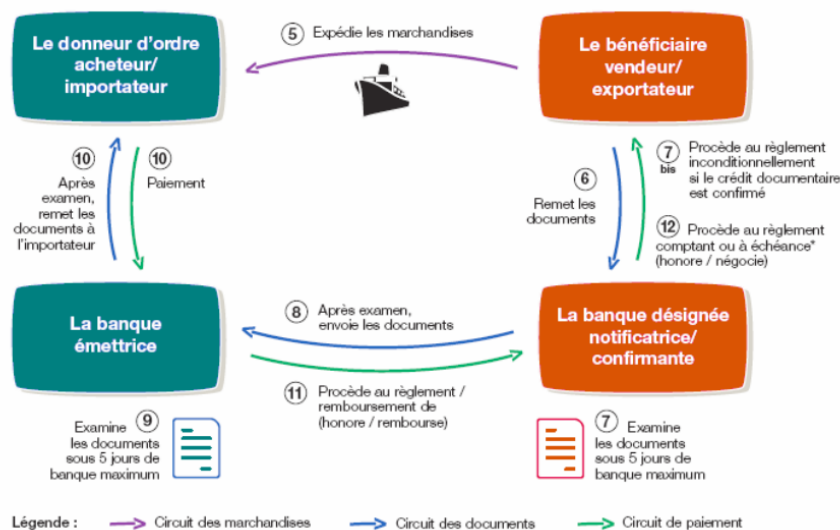


FIGURE 1.2 – Circuit d'une opération de crédit documentaire.

Les opérations de remise documentaire

La remise documentaire consiste pour le vendeur à faire encaisser par une banque le montant dû par un acheteur contre remise de documents. Les documents sont remis à l'acheteur uniquement contre paiement ou acceptation d'une lettre de change. Les intervenants dans l'opération d'encaissement sont :

- Le Donneur d'ordre (le client)
- La Banque remettante (la banque du client)
- La banque chargée de l'encaissement (autre banque que la banque remettante)
- La Banque présentatrice (banque chargée de l'encaissement)

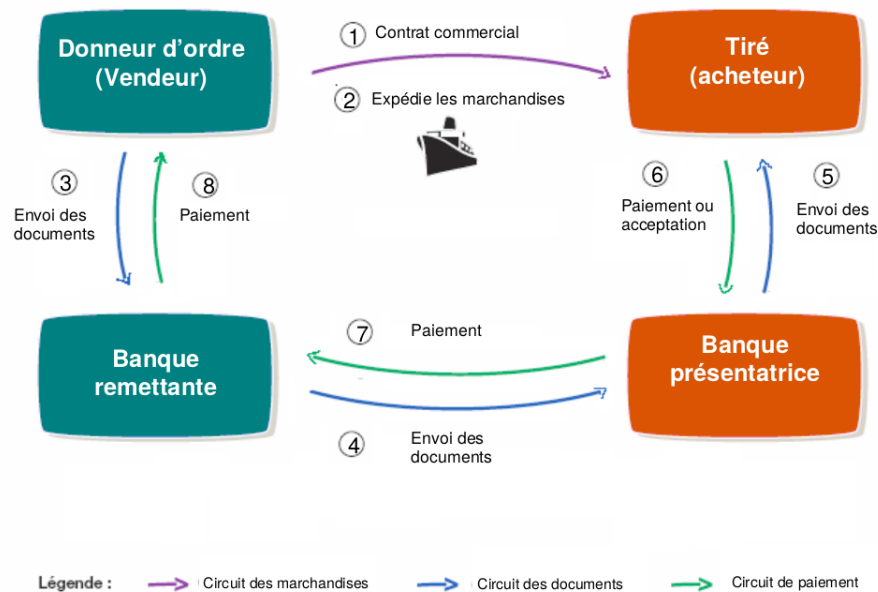


FIGURE 1.3 – Circuit d'une opération de remise documentaire.

Les remises de chèques hors UEMOA

La remise de chèques correspond au dépôt d'un ou de plusieurs chèques par un client auprès de sa banque afin que celle-ci en assure le recouvrement. Chaque chèque remis doit être signé au dos par le client bénéficiaire à qui, la banque demande, le plus souvent, d'indiquer le numéro de compte à créditer au dos du chèque.

Conclusion

Dans ce chapitre il a été question dans ce chapitre de présenter la structure d'accueil, au sein de laquelle nous avons menés nos travaux de recherche. Ensuite nous avons présenté le sujet qui fait objet de ce mémoire, dégager sa problématique et l'intérêt qu'il suscite pour la Société Technidrone. Dans le chapitre suivant, nous parlerons des techniques machine learning, de la segmentation et de la classification.

CHAPITRE 2

APPRENTISSAGE AUTOMATIQUE ET AIDE À LA DÉCISION

Introduction

Depuis plusieurs années, l'apprentissage automatique est de plus en plus exploré en vue de résoudre des problèmes complexes pour lesquels les statistiques étaient impuissante. L'objectif de l'apprentissage automatique (machine learning) est de réaliser des modèles qui apprennent des exemples. Le machine Learning est un ensemble de méthodes qui permettent aux ordinateurs d'apprendre à partir des données qui leur sont soumises. Historiquement, cette théorie a pris son essor avec les travaux des mathématiciens Vapnik et Chervonenkis dans les années 60. Avec le Machine Learning, le point de vue est différent de celui de la statistique traditionnelle. Les algorithmes d'apprentissage automatique permettent aux ordinateurs de s'entraîner sur les entrées de données et utilisent l'analyse statistique pour produire des valeurs qui se situent dans une plage spécifique.

2.1 Vocabulaire du machine learning

2.1.1 Etiquettes

Une étiquette est le résultat de la prédiction ; la variable y dans une régression linéaire simple. Il peut s'agir du cours à venir du blé, de l'espèce animale représentée sur une photo ou de toute autre chose. Dans l'analyse d'un dossier, les étiquettes sont le résultat de l'analyse d'un dossier.

2.1.2 Caractéristiques

Une caractéristique est une variable d'entrée ; la variable x dans une régression linéaire simple. Un projet de Machine Learning simple peut utiliser une seule caractéristique, tandis qu'un projet plus sophistiqué en utilisera plusieurs, spécifiées sous la forme :

$$x_1, x_2, \dots, x_3$$

2.1.3 Exemples

Un exemple est une instance de donnée particulière, x . Les exemples se répartissent dans deux catégories : les exemples étiquetés et les exemples non-étiquetés.

2.1.4 Modèles

Un modèle définit la relation entre les caractéristiques X et l'étiquette. Par exemple, un modèle de détection de spam peut associer étroitement certaines caractéristiques à du « spam ».

Les principales étapes de la durée de vie d'un modèle sont les suivants :

L'apprentissage

L'apprentissage consiste à entraîner le modèle. En d'autres termes, il s'agit de présenter au modèle des exemples étiquetés et de lui permettre d'apprendre progressivement les relations entre les caractéristiques et l'étiquette.

L'inférence

L'inférence consiste à appliquer le modèle entraîné à des exemples sans étiquette. Il s'agit d'utiliser le modèle entraîné pour faire des prédictions efficace.

2.2 Les objectifs et méthodes du Machine learning

Le choix de la méthode d'apprentissage dépend en grande partie de l'objectif poursuivi.

2.2.1 Les objectifs du machine learning

Le machine learning poursuit plusieurs objectifs qui selon le cas peut être

Une classification

Les modèles de classification prédisent des valeurs discrètes. Ils formulent, par exemple, des prédictions qui répondent à des questions telles que les suivantes :

- Un e-mail donné est-il considéré comme du spam ou non ?
- Cette image représente-t-elle un chien, un chat ou un hamster ?
- Un dossier donné est-il conforme ou pas ?

La classification est un processus en deux étapes, une étape d'apprentissage et une étape de prédiction, dans l'apprentissage machine. Dans l'étape d'apprentissage, un modèle est développé à partir d'un ensemble de données préalablement étiquetés. Dans la phase de prédiction, le modèle développé dans la phase précédente est utilisé pour prédire les étiquettes de nouvelles données.

Une regression

Les modèles de régression prédisent des valeurs continues. Ils formulent, par exemple, des prédictions qui répondent à des questions telles que :

- Quel est la valeur d'un logement au Burkina Faso ?
- Quel est la probabilité qu'un utilisateur clique sur cette annonce ?

Le clustering

Le clustering est le regroupement d'exemples en classes d'objets similaires. La différence entre clustering et classification est que les exemples sont étiquetés dans une classification alors que dans le clustering, il ne le sont pas. (Voir Figure ??)

Le but des algorithmes de clustering est de donner un sens aux données et d'extraire de la valeur à partir de grandes quantités de données structurées ou non-structurées. Ces algorithmes vont permettre de séparer les données en fonction de leurs propriétés ou fonctionnalités et de les regrouper dans différents clusters en fonction de leurs similitudes.

2.2.2 Les méthodes d'apprentissage

Les méthodes d'apprentissage automatique les plus largement adoptées sont l'apprentissage supervisé et l'apprentissage non-supervisé. Explorons donc ces méthodes plus en détail.

L'apprentissage supervisé

Le but de cette méthode est de permettre à l'algorithme de découvrir l'étiquette réelle d'un exemple à partir des étiquettes apprises pendant la phase d'entraînement, pour

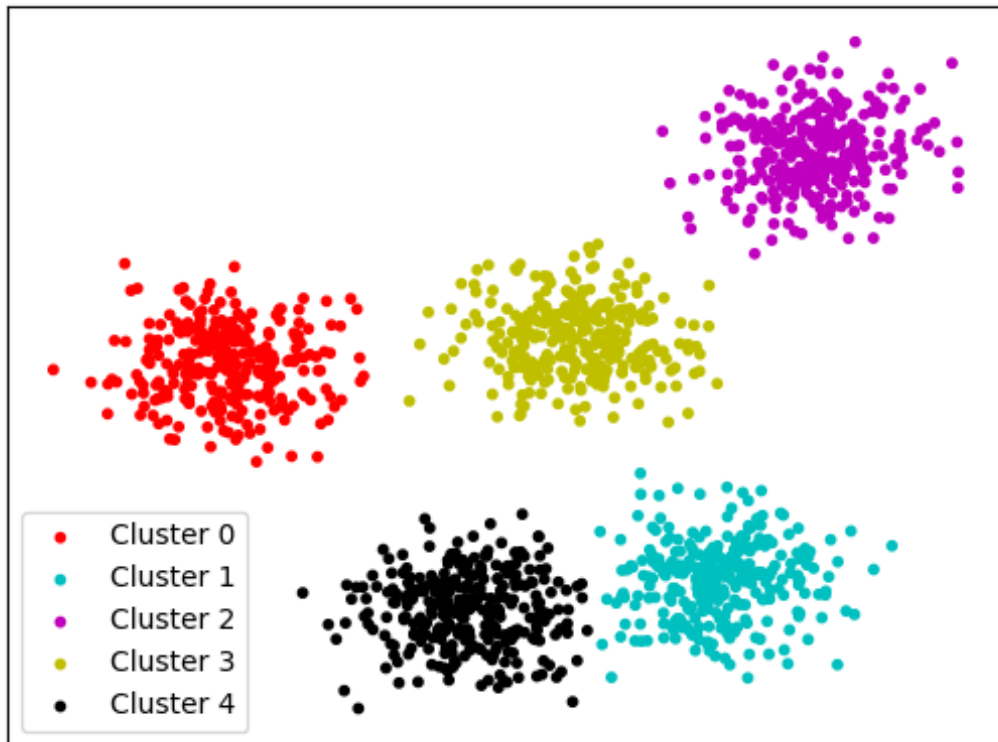


FIGURE 2.1 – Clustering des données.

trouver des erreurs et modifier le modèle en conséquence. L'apprentissage supervisé utilise pour l'entraînement de son modèle, des exemples étiquetés.

L'apprentissage non-supervisé

L'apprentissage non supervisé consiste à apprendre à classer sans supervision ; les exemples fournis sont non-étiquetés. L'objectif ici est de réunir les exemples selon des critères prédéfinis par les équipes en charge du projet. En effet, l'apprentissage non supervisé permet de regrouper des éléments non-classés dans différents groupes selon leurs caractéristiques.

2.3 Les différents type de classifieurs

Il existe plusieurs types de classifieurs. Nous présentons ici quelques classifieurs avec leurs avantages et inconvénients.

2.3.1 Méthode des K plus proche voisins (KNN)

La méthode des 'K plus proche voisins' ou **k-Nearest Neighbors KNN** en anglais est une méthode de classification dans laquelle le modèle mémorise les observations de l'ensemble d'apprentissage pour la classification des données de l'ensemble de test.[?]

Son fonctionnement peut être assimilé à l'analogie suivante : *dis moi qui sont tes voisins, je te dirais qui tu es*. Pour effectuer une prédiction, l'algorithme **K-NN** ne va pas calculer un modèle prédictif à partir d'un training set(ensemble d'apprentissage) comme c'est le cas pour la régression logistique ou la régression linéaire. C'est pourquoi cet algorithme est qualifié de paresseux (Lazy Learning) car il n'apprend rien pendant la phase d'entraînement.

Prédiction avec K-NN

K-NN se base sur le jeu de donnée entier pour effectuer une prédiction. Pour un exemple qu'on souhaite prédire qui ne fait pas parti du jeu de données [?] initiale, l'algorithme va chercher les K instances du jeu de données les plus proches de notre exemple. Ensuite pour ces K voisins, l'algorithme se basera sur leurs étiquettes pour calculer l'étiquette de l'exemple que l'on souhaite prédire.(figure ??)

Similarité dans l'algorithme K-NN

K-NN a besoin d'une fonction de calcul de distance entre deux exemples. Plus deux points sont proches l'un de l'autre, plus ils sont similaires et vice versa[?].

Il existe plusieurs fonctions de calcul de distance, notamment, la distance euclidienne, la distance de Manhattan, la distance de Minkowski, celle de Jaccard, la distance de Hamming La fonction de distance se choisit en fonction des types de données qu'on manipule. Ainsi pour des données quantitatives (poids, salaires, taille, montant de panier électronique. . .), la distance euclidienne est un bon candidat. Quant à la distance de Manhattan, elle est une bonne mesure quand les données ne sont pas de même type (age, sexe, longueur, poids. . .).

Choix de la valeur K

Le choix de la valeur K varie en fonction du jeu de données. En règle générale, si K est petit, on sera sujet au sous apprentissage (underfitting). Par ailleurs, plus on utilise de voisins (K grand) la prédiction sera plus fiable. Toutefois, si on utilise K nombre de voisins avec $K=N$ et N étant le nombre d'exemples, on risque d'avoir du overfitting et par conséquent un modèle qui se généralise mal sur des observations qu'il n'a pas encore vu.

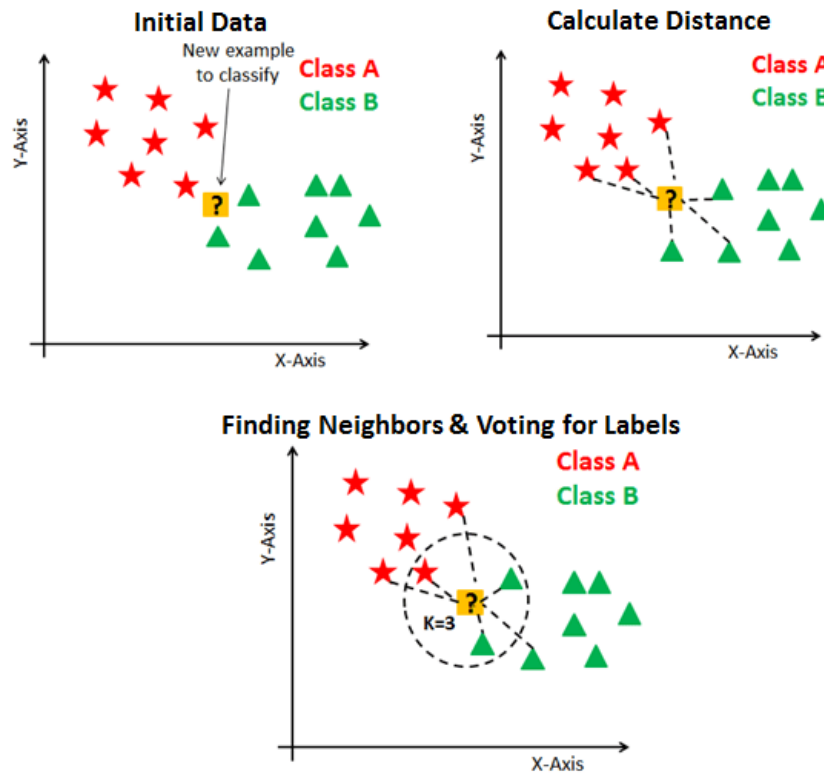


FIGURE 2.2 – Fonctionnement de l'algorithme K-NN.

Avantages

Absence d'apprentissage : Ce sont les échantillons pris en considération, qui constituent le modèle.

Clarté des résultats : bien que la méthode ne produise pas de règle explicite, la classe attribuée à un exemple peut être expliquée en exposant les plus proches voisins qui ont imposé cette attribution.

Grand nombre d'attributs : la méthode permet de traiter des problèmes avec un grand nombre d'attributs. Cependant, plus le nombre d'attributs est important, plus le nombre d'exemples doit être grand.

Inconvénients

Sélection des attributs pertinents : Pour que la notion de proximité soit pertinente, il faut que les exemples couvrent bien l'espace et soient suffisamment proches les uns des autres. Si le nombre d'attributs pertinents est faible relativement au nombre total d'attributs, la méthode donnera de mauvais résultats.

Le temps de classification : Si la méthode ne nécessite pas d'apprentissage, tous les calculs doivent être effectués lors de la classification d'un nouvel exemple.

Définir les distances et nombres de voisins : Les performances de la méthode dépendent du choix de la distance, du nombre de voisins et du mode de combinaison des réponses des voisins.

2.3.2 Les réseaux de neurones

Les réseaux de neurones sont inspirés de la structure neurophysiologique des neurones. En règle générale, un réseau de neurones repose sur un grand nombre de processeurs opérant en parallèle et organisés en tiers(couches). La première couche reçoit les entrées d'informations brutes, un peu comme les nerfs optiques de l'être humain lorsqu'il traite des signaux visuels. Par la suite, chaque couche reçoit les résultats de la couche précédente. On retrouve le même processus chez l'Homme, lorsque les neurones reçoivent des signaux en provenance des neurones proches du nerf optique. La dernière couche, quant à elle, produit les résultats du système.

Les différents cas d'usage

Les réseaux de neurones sont beaucoup utilisés dans la reconnaissance d'écriture manuscrites, la transcription « speech-to-text » ou encore dans la prévision des marchés financiers ou trading algorithmique.

Ils peuvent aussi être utilisé pour la reconnaissance faciale, la prédiction météo, la détection de cancer sur les imageries médicales. De manière générale, les réseaux de neurones excellent pour la reconnaissance de patterns.

Avantages

Classification efficace : le calcul d'une sortie à partir d'un vecteur d'entrée est un calcul très rapide.

Les données réelles : les réseaux traitent facilement les données réelles "préalablement normalisées" et les algorithmes sont robustes au bruit.

Inconvénients

- Déterminer l'architecture du réseau est complexe et les paramètres sont difficiles à interpréter (boîte noire).
- L'échantillon nécessaire à l'apprentissage doit être suffisamment grand et représentatif des sorties attendues.

2.3.3 Support Vector Machine (SVM)

Les Support Vector Machine ou Machine à Vecteur de Support constituent une technique d'apprentissage supervisée. Elles ont été inventées par Boser, Guyon et Vapnik [?] et présentées pour la première fois à la conférence Computational Learning Theory (COLT) de 1992. Grâce à ses performances [?], cette technique a ouvert un domaine de recherche très actif et un grand éventail d'applications. Les SVM utilisent une approche géométrique pour classer les données en deux catégories.

En considérant les données comme des vecteurs, les SVM construisent un plan (une frontière) qui sépare les données dans chacune des catégories. Une fois la frontière de décision construite (Hyperplan) la SVM sera capable de classer de nouvelles données en observant de quel côté de la frontière elles tombent, et en leur assignant la catégorie correspondante.

L'idée est donc de rechercher le meilleur hyperplan qui sépare linéairement deux classes, tout en les repoussant au maximum. Lors de la phase d'apprentissage, le SVM cherche à maximiser la marge entre les deux classes d'apprentissage. Ce qui lui procure une grande capacité de généralisation pendant la phase de test.

Les machines à vecteurs de support ont été appliquées dans des domaines comme la reconnaissance automatique des visages et des gestes [?], la prédiction des mouvements de la bourse... [?].

Les domaines dans lesquels, elles sont les plus efficaces sont : la reconnaissance d'objet et de d'image [?] et la catégorisation de texte [?]

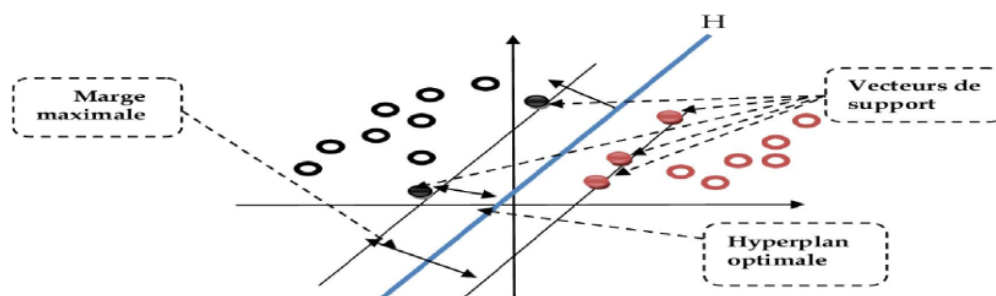


FIGURE 2.3 – Machine à vecteurs de support .

Avantage

- Grâce à leurs fondements mathématiques solides, les SVM possèdent donc une grande précision de prédiction
- les SVM fonctionnent bien sur de petits jeux de données

- Décision rapide. La classification d'un nouvel exemple consiste à voir le signe de la fonction de décision $f(x)$.

Inconvénient

Les SVM ne conviennent pas à des jeux de données très volumineux car le temps d'entraînement est très long.

Les SVM effectuent une classification binaire d'où la nécessité d'utiliser l'approche un-contre-un pour construire un classifieur multiclasse. Une grande quantité d'exemples en entrées implique un calcul matriciel important. Le temps de calcul est élevé lors d'une régularisation des paramètres de la fonction noyau.

Les SVM sont moins efficaces sur les jeux de données contenant du bruit et beaucoup d'outliers.

2.3.4 Les arbres de décisions

Un arbre de décision est un outil d'aide à la décision qui permet de répartir une population d'individus en groupes homogènes selon des attributs discriminants en fonction d'un objectif fixé. Il permet d'émettre des prédictions sur le problème par réduction de niveau après niveau du domaine.

Les arbres de décision sont facilement interprétables, toutefois, leur capacité de prédiction est presque toujours dépassée par les autres modèles de classification. Cette caractéristique a limité son utilisation. Au début des années 2000, ils ont été repris comme élément de base d'une nouvelle méthode de classification, appelée la forêt aléatoire de décision.

Cette nouvelle technique utilise de manière combinée les arbres de décision et la théorie statistique pour réduire la variance du classifieur en calculant la moyenne d'un ensemble d'arbres de décision en générant des classifieurs avec une très bonne capacité de prévision. Nous les présenterons plus largement dans les prochaines sections.

Avantages

Adaptabilité aux attributs de valeurs manquantes : les algorithmes peuvent traiter les valeurs manquantes (exemples contenant des champs non renseignés) pour l'apprentissage, mais aussi pour la classification.

Modèle white-box : D'un arbre de décision, il est possible de générer des règles permettant d'expliquer ou de comprendre le résultat d'une classification. Le résultat est facile à conceptualiser, à visualiser et à interpréter.

Classification très rapide : Le coût d'utilisation des arbres est logarithmique.

Traitement de tous type de données : Les arbres de décisions prennent en compte aussi bien les échantillons ayant des caractéristiques continues que discrètes. Il est robuste au bruit.

Donne une classification efficace L'attribution d'une classe à l'aide d'un arbre de décision est obtenu grâce au parcours d'un chemin de l'arbre.

Ils ont un bon comportement par rapport aux valeurs extrêmes (outliers).

Inconvénient

Manque d'évolutivité dans le temps : Même si les données évoluent avec le temps, il est nécessaire de relancer une phase d'apprentissage sur l'échantillon complet (anciens nouveaux exemples)

Méthode sensible au nombre de classes : les performances tendent à se dégrader lorsque le nombre de classes devient trop important.

Ils sont instables : Des changements légers dans les données produisent des arbres très différents. Les changements des nœuds proches de la racine affectent beaucoup l'arbre résultant.

Sûr-apprentissage : Les arbres générés sont trop complexes et généralisent mal (solution : élagage, contrôle de la profondeur de l'arbre et de la taille des feuilles).

2.4 Comparaison des algorithmes de classification

Le choix de l'algorithme optimal pour un problème donnée dépend de sa vitesse d'entraînement et de prédiction, de la précision de ces prévisions, de la quantité de données nécessaires à l'entraînement, de la facilité à la mettre en oeuvre, et de la capacité à expliquer le résultat de la prédiction.

Le tableau ci-dessous présente une comparaison des différents algorithmes de classification.

Le tableau ?? révèle que les algorithmes de Réseaux de neurones et ceux de forêt aléatoire ont un taux très élevé de bonne prédiction. Malheureusement, ces algorithmes fonctionnent bien sur des jeu de données énormes. De plus la vitesse d'apprentissage et de prédiction reste relativement lente par rapport aux autres algorithmes.

2.5 Les arbres de décisions

Deux techniques de classification par les arbres ont été développées au début des années 1980 par deux groupes de chercheurs. Le premier groupe, dirigé par J Ross Quinlan, a développé un algorithme d'arbres de décision en 1986 appelé ID3. Plus tard en améliorant

Algo	Interprétabilité	Précision	VE et VP	Données
Knn	Oui	Faible	Dépend de K	Beaucoup
Régression	Un peu	Faible	Rapide	Peu
Naïves bayes	Un peu	Faible	Rapide	Peu
Réseaux de neurones	non	Très élevé	Lent	Beaucoup
Arbre de décision	Oui	Moyen	Rapide	Assez
Random Forest	Non	Très élevé	Lent	Assez

TABLE 2.1 – Tableau de comparaison des algorithmes de classifications

plusieurs caractéristiques de ID3, il a développé et présenté C4.5. L. Breiman, J. Friedman, R. Olshen, et C. Stone, un groupe de statisticiens, ont développé un algorithme pour produire des arbres de décision binaires appelé CART (Classification and Regression Trees) de leur côté. Ces algorithmes ont été le début de la recherche sur la classification par les arbres de décisions. Les deux approches suivent le paradigme « Diviser pour régner »

Plusieurs objectifs concourent à la construction d'un arbre de décision. Ce sont :

- Une meilleure généralisation des exemples de la base d'apprentissage.
- Une meilleure classification de nouveaux exemples
- Une structure aussi simple que possible

La construction d'un arbre de décision consiste à partitionner un ensemble de données en des groupes les plus homogènes possible du point de vue de la variable à prédire. On prend en entrée un ensemble de données classées, et on fournit en sortie un arbre. Nous obtenons un arbre qui représente une série de noeuds en plaçant dans la partie supérieure le noeud dont la capacité de classification est la plus grande. [?] Le résultat final est un arbre renversé comme celui représenté dans la figure ??

2.5.1 Structure d'un arbre de décision

Le fonctionnement des arbres de décisions repose sur les heuristiques construites sur des techniques d'apprentissage supervisées.

Les arbres de décisions sont composés de noeuds et de feuilles reliés par des branches. Dans leur représentation graphique la racine est placée tout en haut et les feuilles en bas. Les noeuds internes sont appelés des noeuds de décision. Ils peuvent contenir une ou plusieurs règles. Les noeuds terminaux contiennent la classe aussi appelée classe à

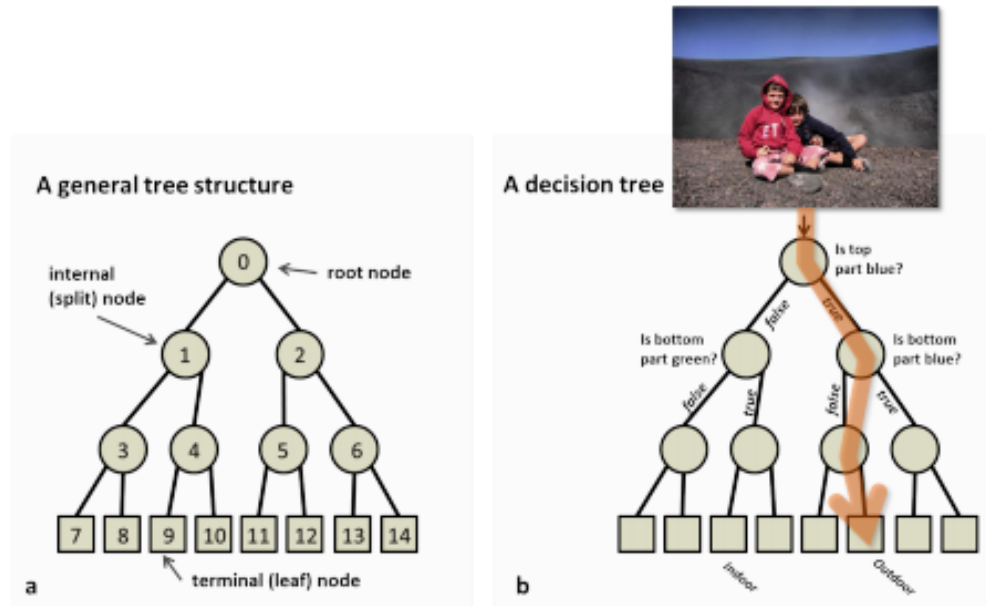


FIGURE 2.4 – Arbre de décision.

Le a représente la structure générale d'un arbre de décision. Le b montre un arbre de décision illustratif utilisé pour déterminer si une photo représente une scène d'intérieur ou d'extérieur.

prédire ou étiquette. Après sa construction, un arbre de décision peut être traduit par un ensemble de règle.

L'algorithme générique de construction d'un arbre permet de générer itérativement l'arbre en prenant à chaque itération une variable et en lui créant ses noeuds et ses feuilles. L'idée centrale est la suivante :

Diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis à l'aide des attributs jusqu'à ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples ayant tous la même étiquette.

2.5.2 Optimisation des noeuds

En général, on décide qu'un noeud est terminal lorsque tous les exemples associés à ce noeud, ou du moins la plupart d'entre eux ont la même étiquette ou s'il n'y a plus d'autres caractéristiques non utilisées dans la branche correspondante.

La sélection d'un test à associer à un noeud pour obtenir un arbre optimal est un choix crucial. En effet, construire un arbre de décision optimal consiste à construire un arbre de décision le plus petit possible rendant compte au mieux des données. Il s'agit donc

de rechercher le test qui permet de faire évoluer la tâche de classification. Pour mesurer cette évolution, *CART* utilise l'*indice de Gini*. Les algorithmes de Quinlan eux, utilisent la notion d'*entropie*.

Entropie de Shannon

L'entropie de Shannon correspond à la quantité d'information fournies par un événement : plus la probabilité d'un événement est faible (il est rare), plus la quantité d'information qu'il apporte est grande. Sa formule est la suivante [?] :

$$Entropie = - \sum_{i=1}^n p_i * \log_2(p_i)$$

Tel que p_i est la proportion d'exemples de S ayant pour classe résultante (étiquette) i .

Pour un ensemble de données T caractérisé par n classes (C_1, C_2, \dots, C_n) selon la variable cible, la quantité d'information nécessaire pour identifier la classe d'un individu correspond à l'entropie $E(P)$ où P est la distribution de probabilité de la partition (C_1, C_2, \dots, C_n) .

$$P = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \dots, \frac{|C_n|}{|T|} \right)$$

$|C_i|$ représente le cardinal de la classe i c'est-à-dire le nombre d'éléments de la classe i .

L'entropie de T est alors :

$$Entropie(T) = - \sum_{i=1}^n \frac{|C_i|}{|T|} \log_2 \frac{|C_i|}{|T|}$$

La fonction permettant de sélectionner le test qui doit étiqueter le noeud courant est la fonction *Gain*. Pour un ensemble de données T , le gain d'information de T par rapport à une partition T_j donnée est la variation d'entropie causée par la partition de T selon T_j

$$Gain(X, T) = Entropie(T) - Entropie(X, T) = Entropie(T) - \sum_{j=1}^m \frac{T_j}{T} * Entropie(T_j)$$

Le gain permet de calculer ce que l'attribut spécifié apporte au désordre du set. Plus un attribut contribue au désordre, plus il est important de le tester pour séparer le set en plus petits sets ayant une entropie moins élevée.

Indice de Gini

L'indice de Gini est une mesure statistique permettant de rendre compte de la répartition d'une variable au sein d'une population. Il mesure l'impureté qui est un concept très

utile dans la construction des arbres de décision : La qualité d'un noeud et son pouvoir discriminant peuvent être évalués par son impureté. Sa formule est la suivante :

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

2.5.3 Algorithmes d'induction d'arbres de décision

Il existe essentiellement deux grandes familles d'algorithmes permettant de construire des arbres de décisions à partir d'un set de données : les algorithmes de Quinlan (**ID3**, **C4.5**, **C5.0**) et l'algorithme **CART**. Les deux approches suivent le paradigme « diviser pour régner ». Nous présentons ici le principe des trois algorithmes de construction des arbres de décision que sont l'algorithme ID3, l'algorithme C4.5 et l'algorithme CART.

CART

L'algorithme Classification and Regression Trees(*CART*) est très similaire à C4.5, mais il en diffère par le fait qu'il prend en charge la régression en ne calculant pas des ensembles de règles. Il s'agit d'un algorithme développé par Breiman, Friedman, Olshen et Stone (1984).

Selon l'algorithme CART, un arbre de décision est construit en déterminant les questions (appelées fractionnements de noeuds) qui, lorsqu'on y répond, conduisent à la plus grande réduction de l'impureté de Gini. Cela signifie que l'arbre de décision tente de former des noeuds contenant une forte proportion d'échantillons (points de données) provenant d'une seule classe en trouvant des valeurs dans les caractéristiques qui divisent proprement les données en classes(étiquettes).

ID3

Iterative Dichotomiser 3(*ID3*) a été développé par Ross Quinlan en 1986. Il se base sur le concept d'attribut et de classe. L'algorithme recherche l'attribut le plus pertinent à tester pour que l'arbre soit le plus court et le plus optimisé possible en déterminant l'attribut qui maximise le gain d'information.[?]

L'algorithme crée un arbre multivoie, trouvant pour chaque nœud (c'est-à-dire de manière gourmande) la caractéristique catégorielle qui produira le plus grand gain d'informations pour les cibles catégorielles. Les arbres sont cultivés jusqu'à leur taille maximale, puis une étape d'élagage est généralement appliquée pour améliorer la capacité de l'arbre à généraliser les données invisibles.

Méthode	CART	C4.5
Mesure utilisé pour la sélection	index Gini	Entropie et Gain d'info
Type des variables(attributs)	discrètes et continues	discrètes et continues
Division à chaque noeud	binaire	multiple

TABLE 2.2 – Tableau comparatif des algorithmes C4.5 et CART

C4.5

L'algorithme *C4.5* est une évolution de l'algorithme ID3. Il a également été inventé par Ross Quinlan. Basé sur ID3, *C4.5* possède quelques améliorations[?]

- Une adaptation de la fonction gain qui n'a plus tendance à aller vers l'attribut avec le plus de valeur possible.
- La possibilité de gérer les valeurs manquantes.
- La possibilité de post-élaguer son arbre pour éviter l'overfitting ;
- La possibilité de manipuler des valeurs continues

C5.0 est la dernière version de Quinlan publiée sous une licence propriétaire. Elle utilise moins de mémoire et construit des jeux de règles plus petits que *C4.5* tout en étant plus précise.

Dans le cadre de notre projet, l'algorithme qui sera utilisé pour la génération de notre arbre de décision est *C4.5*. Il permet la manipulation de valeurs continues et ne génère pas un arbre de décision binaire comme *CART* (tableau ??).

Un modèle flexible mémorise essentiellement les données d'entraînement en les ajustant étroitement. Le problème d'un tel modèle est qu'il apprend non seulement les relations réelles dans les données d'entraînement, mais aussi tout bruit présent dans ces données. Un modèle rigide est dit avoir un biais élevé parce qu'il fait des hypothèses sur les données de formation. Par exemple, un classifieur linéaire fait l'hypothèse que les données sont linéaires et n'a pas de flexibilité pour s'adapter à des données non linéaires.

Dans les deux cas (modèle flexible et modèle rigide), le modèle n'est pas capable de réaliser de bonnes prédictions sur de nouvelles données. Les arbres de décisions sont des modèles d'apprentissage flexible donc sensible au bruit. Ils peuvent devenir très profonds c'est à dire croître jusqu'à ce qu'il ait exactement une feuille pour chaque observation, les classant toutes parfaitement.

Comme alternative, la forêt aléatoire empêche ce phénomène en créant des sous-ensembles aléatoires des caractéristiques et en construisant des arbres plus petits à l'aide de ces sous-ensembles. Dans la suite nous présenterons les forêts aléatoires une méthode supervisée d'apprentissage machine.

2.6 Les forêts aléatoires

La forêt aléatoire est un modèle composé de nombreux arbres de décision. Plutôt que de se contenter de faire la moyenne des prédictions des arbres (que nous pourrions appeler une « forêt »), ce modèle utilise deux concepts clés qui lui donnent le nom d'aléatoire :

- L'échantillonnage aléatoire des données d'entraînement lors de la construction de l'arbre.
- Des sous-ensembles aléatoires de caractéristiques pour le fractionnement des noeuds

L'algorithme effectue un apprentissage en parallèle sur de multiples arbres de décision construits aléatoirement et entraînés sur des sous-ensembles de données différents. Le nombre idéal d'arbres, qui peut aller jusqu'à plusieurs centaines voire plus, est un paramètre important : il est très variable et dépend du problème.

2.6.1 Fonctionnement des forêts aléatoires

La forêt aléatoire (Random Forest) fonctionne en deux phases. La première consiste à créer la forêt aléatoire en combinant N arbres de décisions. La seconde consiste à faire des prédictions pour chaque arbre créé dans la première phase.

Le processus peut être expliqué dans les étapes ci-dessous :

Etape 1 : Sélectionnez k instances dans l'ensemble d'apprentissage.

Etape 2 : Construire les arbres de décisions associés aux points de données sélectionnés.

Etape 3 : Répétez les étapes 1 et 2, N fois. (N étant le nombre d'arbres de la forêt)

Etape 4 : Pour une nouvelle instance de données, trouvez la prédiction de chaque arbre de décision de la forêt et attribuez l'étiquette qui remporte la majorité des votes.

Supposons qu'il existe un ensemble de données contenant plusieurs images de fruits. Cet ensemble de données est attribué à un modèle de forêt aléatoire. L'ensemble des données est alors divisé en sous-ensemble et donné à chaque arbre de décision. Pendant la phase d'apprentissage, chaque arbre de décision produit un résultat de prédiction. Lorsqu'une nouvelle instance apparaît, le modèle prédit la décision finale.

Echantillonnage aléatoire des données d'entraînement

Lors de la phase d'entraînement, chaque arbre d'une forêt aléatoire apprend à partir d'un échantillon aléatoire de données. Les échantillons sont tirés avec remplacement, connu sous le nom de « bootstrapping », ce qui signifie que certains échantillons seront utilisés plusieurs fois dans un seul arbre.

Les prédictions sont faites en faisant la moyenne des prédictions de chaque arbre de décision. Cette procédure est connue sous le nom de *bagging* abréviation de *bootstrap aggregating*

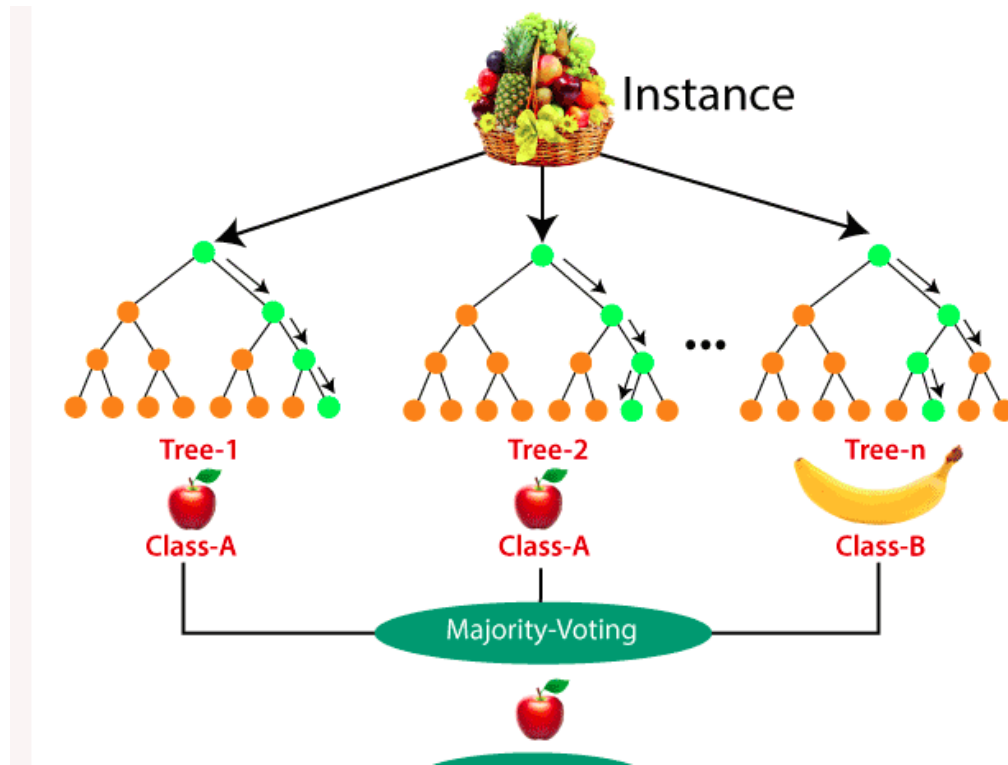


FIGURE 2.5 – Forêt aléatoire.

Fractionnement des noeuds

Le deuxième concept principal de la forêt aléatoire est que seulement un sous-ensemble de toutes les caractéristiques est pris en compte pour diviser chaque noeud d'un arbre de décision. Cette valeur est habituellement la racine carré du nombre de caractéristiques pour une classification. Ainsi si nous avons 25 caractéristiques seulement 5 seront pris aléatoirement pour diviser le noeud.

2.6.2 Les hyperparamètres dans les forêts aléatoires

Les hyperparamètres de la forêt aléatoires sont utilisés pour augmenter le pouvoir prédictif du modèle, soit pour rendre le modèle plus rapide.

Augmenter le pouvoir prédictif du modèle

Plusieurs hyperparamètres permettent d'augmenter le pouvoir de prédiction du modèle ;

n_estimators : Il s'agit du nombre d'arbres que l'algorithme construit avant de prendre le vote maximum ou la moyenne des prédictions. En général, un nombre d'arbres

élevé augmente la performance et rend les prédictions plus stables mais ralentit également le calcul.

max_features : Il s'agit du nombre maximum de caractéristiques qu'une forêt aléatoire considère pour diviser un noeud.

min_sample_leaf : Il s'agit du nombre minimum de feuilles nécessaires pour diviser un noeud interne.

Augmenter la vitesse d'exécution du modèle

Les hyperparamètres permettant d'accélérer le modèle sont les suivants :

n_jobs : Il indique au moteur le nombre de processeurs qu'il est autorisé à utiliser. Une valeur de -1 signifie qu'il n'y a aucune limite sur le nombre.

random_state : Il rend la sortie du modèle reproductible. Le modèle produira toujours les mêmes résultats si on lui donne les mêmes hyperparamètres et les mêmes données d'entraînement.

oob_score : Il s'agit d'une méthode de validation croisée des forêts aléatoires. Dans cette échantillonnage, environ un tiers des données n'est pas utilisé pour entraîner le modèle mais plutôt pour évaluer ses performances sans aucune charge de calcul supplémentaire.

2.6.3 Les avantages et les inconvénients du modèle des forêts aléatoire

Avantages

- Les forêts aléatoires permettent de surmonter le problème de sur-ajustement en faisant la moyenne ou en combinant les résultats de différents arbres de décisions.
- Les forêts aléatoires fonctionnent mieux sur un large éventail de données qu'un seul arbre de décision.
- Les forêts aléatoires présentent moins de variance qu'un arbre de décision unique
- Les forêts aléatoires possèdent une très grande précision.
- Les algorithmes de forêt aléatoire maintiennent une bonne prédiction même si certaines informations sont absentes.

Inconvénients

- La complexité est le principal inconvénient des algorithmes de Random Forest.
- La construction de forêts aléatoires est beaucoup plus difficile et longue que celle des arbres de décision.

- Il faut davantage de ressources de calcul pour mettre en oeuvre un algorithme de forêt aléatoire.
- Il est moins intuitif dans le cas où nous disposons d'une grande collection d'arbres de décision.
- Le processus de prédiction utilisant les forêts aléatoires est très long par rapport aux autres algorithmes.

Conclusion

Ce chapitre nous a permis de présenter le machine learning et quelques algorithmes de classifications. En outre, nous nous sommes attardés sur les arbres de décisions qui sont un type de classifieurs qui nous pensons sont adaptés au contexte de notre étude. Maintenant, nous poursuivrons en présentant la conformité dans le domaine bancaire.

CHAPITRE 3

LA CONFORMITÉ DANS LE SECTEUR BANCAIRE

Introduction

Dans ce chapitre, nous élaborerons d'abord une étude de l'existant. Pour cela, nous présenterons quelques plateformes permettant d'analyser des opérations de transferts de fonds. Après cela, Nous étudierons la fonction conformité ou compliance en anglais dans le secteur bancaire.

3.1 Etude de l'existant

3.1.1 Analyse des dossiers de transfert à la SGBF

Le problème posé au niveau du service OPI, c'est l'analyse en temps réel des dossiers de transferts reçus au niveau du guichet. Cette analyse suit un processus. La première phase de ce processus est la réception du dossier. A la réception du dossier, le collaborateur analyse la cohérence du dossier. Cette analyse consiste en la vérification de la cohérence et l'exactitude des informations contenues sur les éléments constitutifs du dossier.

Après cette phase, le collaborateur analyse la complétude du dossier. Une fiche, disponible au niveau du guichet permet aux collaborateurs de OPI de savoir à vue d'oeil quels justificatifs devraient être présent dans le dossier en fonction du motif de l'opération.

L'étape suivante est la vérification de la fiabilité des différents acteurs de l'opération à travers des outils comme Force-online.

La dernière étape concerne la vérification du circuit de transfert. Il s'agit à cette étape

de s'assurer que la réglementation autorise l'opération qui est entrain d'être menée entre les différents acteurs.

3.1.2 Plateformes existantes dans le milieu bancaire

De nombreuses plateformes permettent de juger le risque de non-conformité d'un acteur d'une opération de transfert. Ces plateformes sont toutes propriétaires.

ComplianceBond

ComplianceBond est un des produits de la plateforme HighBond. HighBond est une plateforme logicielle de gouvernance d'entreprise qui renforce la sécurité, la gestion des risques, la conformité et l'assurance. ComplianceBond est une solution de gestion de la conformité qui permet aux organisations de mettre en oeuvre, d'automatiser et de démontrer une assurance par rapport à un programme de conformité.

Les fonctionnalités principales de ComplianceBond sont les suivantes :

- Centraliser la documentation des besoins et des contrôles mappés. Cela permet de réduire le temps passé à documenter et à tester la conformité.
- Evaluer et surveiller la conformité en automatisant les tests de surveillance de conformité en temps réel.
- Rapport sur le statut de conformité

Il s'agit d'une plateforme propriétaire.

TraProtect

TraProtect de TraInvestment est une plate-forme multicanal, multi-activité et multi-niveau de prévention temps réel et détection de la fraude des transactions spécialement conçue pour le monitoring des transactions de paiement électronique. Elle est destinée à toute institution traitant les transactions de paiement électronique.

kdprevent

La plateforme kdprevent permet de lutter contre le blanchiment d'argent et le financement du terrorisme. Elle a été mise en oeuvre dans plusieurs pays du monde, dans plus de 50 institutions. Elle est conçue pour détecter les activités inhabituelles, inattendues et suspectes. Une fois détectée, elle envoie automatiquement des avertissements aux responsables, généralement les responsables conformité. Ses principales fonctionnalités sont :

- Analyse d'une transaction unique et d'un ensemble de transactions liées qui ont eu lieu dans une période de temps donnée.

- Détection automatique et interruption des transactions suspectes (i.e SWIFT, SEPA, SIC, etc.) et notification en temps réel.
- Génération d'alertes pour les situations suspectes détectées
- Un analyseur de relations qui vous permet d'explorer les relations potentiellement suspectes ou inconnues qui existent entre les clients, les emprunteurs ou les comptes.

3.2 La conformité ou compliance

Le cadre réglementaire autour des activités financières a été fortement renforcé, faisant de la conformité (ou Compliance) un pilier indispensable de la protection des institutions financières en particulier les banques et de leurs clients. [?]

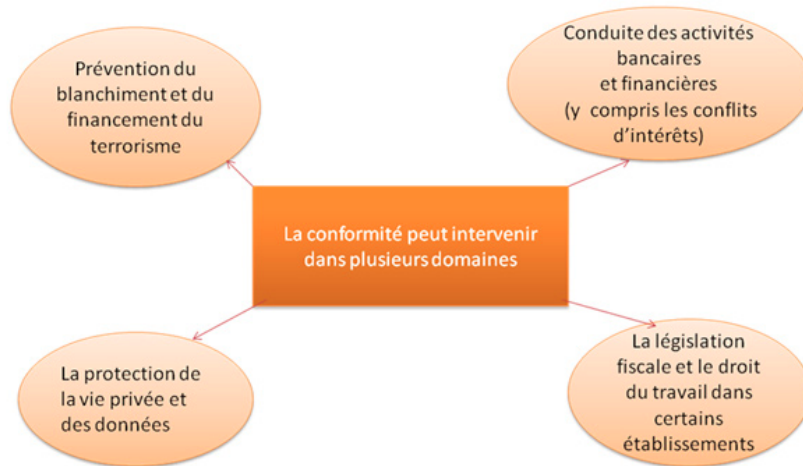


FIGURE 3.1 – Domaine d'intervention de la conformité .

3.2.1 Définition

La conformité en anglais compliance est un concept qui a fait naître de nouvelles obligations pour le banquier. En effet, face à la complexité des environnements et à « l'inflation réglementaire », la fonction conformité a pour but de prévenir tout risque de non-conformité des opérations bancaires et financières. La conformité se définit donc comme l'obligation de veiller à ce que les collaborateurs des différentes banques s'assurent en permanence que soient respectées :

- Les dispositions législatives et réglementaires propres aux activités bancaires ;
- Les normes et usages professionnels et déontologiques ;

- Les codes de conduites notamment le code éthique et les procédures internes

Dans ses grandes lignes, la conformité consiste à :

- Identifier et à jauger le degré de non-conformité d'une entité économique par rapport à l'ensemble des règles de conduite qui lui sont applicables
- Mesurer son taux d'exposition aux risques de sanction judiciaire et administrative
- Evaluer les pertes financières qu'elle pourrait subir
- Conseiller une entité économique pour qu'elle se mette en conformité avec les normes législatives et réglementaires.

En somme, la conformité est l'ensemble des actions visant à l'intégration, dans la structure bancaire des exigences issues des réglementations financières. La fonction conformité dans une banque recouvre quatre grandes activités :

Sécurité financière

Elle est attentive à la sécurité financière de la banque et lutte en ce sens contre la fraude, le blanchiment de capitaux et financement du terrorisme, les abus de marché et les embargos.

La protection Clientèle

Elle assure en parallèle, une protection continue de la clientèle en préservant aussi bien leurs intérêts propres, que ceux des marchés ou de la banque elle-même.

Le contrôle permanent

Elle appartient au dispositif global de contrôle permanent et assure la gestion des risques de non-conformité.

La déontologie

La déontologie est également une partie intégrante de la conformité. Elle permet de s'assurer du respect du recueil des règles de déontologie de l'établissement bancaire ainsi que de traiter les signalements pouvant provenir de tous les collaborateurs de la banque.

3.2.2 Rôle de la conformité dans le domaine bancaire

Le rôle de la conformité est d'abord de donner aux dirigeants de la Banque ainsi qu'au Conseil d'administration l'assurance raisonnable que les risques de non-conformité réglementaires et de réputation sont dûment surveillés, contrôlés et atténués au niveau du Groupe. C'est également s'assurer en permanence que les lois et réglementations ainsi

que les règles et normes internes définies par les pays sont respectées. In fine, il s'agit d'offrir aux clients l'assurance d'un environnement sécurisé pour réaliser leurs opérations financières, en vérifiant que celles-ci sont conformes aux règles déontologiques et aux législations.

3.3 Présentation du cadre réglementaire

La SGBF est membre d'un groupe international français. Elle est donc soumise aussi bien aux règlements de la zone UEMOA qu'à celles européennes et internationales.

Au niveau international, l'organisme de référence de contrôle est le GAFI. Le GAFI (Groupement d'Action Financière) est un organisme international, sous l'égide des nations unies, créé à Paris en 1989. Il a pour rôle d'émettre des recommandations dans le domaine de la lutte contre le blanchiment des capitaux et dans la lutte contre le financement du terrorisme. Les pays membres du GAFI acceptent de fait les recommandations et s'engagent à mettre en oeuvre les lois permettant l'application de ces recommandations. Les pays membres du GAFI s'engagent également à s'auto-évaluer à intervalles de temps réguliers dans le but d'améliorer leurs dispositifs respectifs. Il s'agit aujourd'hui la référence principale dans le domaine de l'AML (Anti Money Laundering).

Les contrôles de la fonction conformité regroupent aussi bien ceux sur la réglementation de change que ceux sur les composantes d'un programme AML .

3.3.1 La réglementation de change

La réglementation de change est un outil juridique important non seulement dans le monde des affaires, mais aussi dans la vie d'un pays compte tenu de la diversité des phénomènes économiques et de la criminalité qui pourrait se développer dans ce domaine. Elle relève de la tutelle du Ministre chargé des Finances. Elle prescrit que les règlements financiers et mouvements de capitaux entre l'UEMOA et l'Etranger, ainsi que les opérations de change manuel dans l'UEMOA, ne peuvent s'effectuer que par l'entremise de la BCEAO ou d'une banque intermédiaire agréé[?].

Le Change se définit comme l'échange d'une monnaie contre une autre, C'est le bénéfice réalisé sur la différence des cours entre deux monnaies. C'est aussi le taux de conversion entre deux monnaies.

Au Burkina Faso et dans les pays membres de l'UEMOA, les transferts à l'étranger sont régis par un ensemble de texte. Ces textes fixent les procédures à suivre par les intermédiaires agréés en matière d'exécution des opérations avec l'étranger et déterminent la procédure de domiciliation et de règlement des importations par la banque.

La réglementation de change sur les opérations d'exportation

Les opérations d'exportations d'un montant supérieur à 500000 sont soumises à domiciliation auprès d'une banque. Pour chaque opération d'exportation, les résidents sont tenus d'encaisser les recettes en devises et de les céder à la banque domiciliataire dans un délai d'un mois à compter de la date d'exigibilité du paiement.

La réglementations de change sur les opérations d'importation

Les opérations d'importation de marchandises étrangères, c'est-à-dire originaires d'un pays extérieur à la zone franc, doivent être domiciliées auprès d'une banque intermédiaire agréé, lorsque leur valeur dépasse un certain seuil variable selon les pays.[?] Pour une opération d'importation, le dossier complet de domiciliation doit contenir une copie de la facture établi par le fournisseur, une attestation d'importation, et un formulaire d'autorisation de change.

La réglementation de change sur les opérations d'investissement et d'emprunt

La réglementation de change exige que pour tout investissement, prêt, ou opération en capital par un résident, une autorisation préalable du ministère chargé des finances est obligatoire.

3.3.2 Les composantes d'un programme AML

Les programmes AML permettent de garantir la sécurité financière d'un établissement financier. Ce sont :

1. Lutte contre le blanchiment des capitaux (AML/LAB)
2. Lutte contre le financement du terrorisme (CFT)
3. Respect des embargos commerciaux et financiers
4. Surveillance des opérations de marché

La lutte contre le blanchiment de capitaux

Le blanchiment de capitaux consiste à dissimuler la provenance d'argent acquis de manière illégale, appelé communément « argent sale », en lui donnant l'apparence de fonds d'origine licite (« argent propre ») pour le réinvestir dans des activités légales. Le Blanchiment permet notamment aux criminels de masquer une augmentation trop ostensible de leur richesse afin d'éviter d'attirer l'attention des autorités. On distingue trois phase dans le processus global de blanchiment :

La phase de placement qui consiste à injecter dans le système financier les sommes d'argent issues des crimes et des délits ;

La phase d'empilement qui consiste à brouiller les pistes. Le but est d'effectuer un ensemble de transactions qui ont pour objectif d'empêcher toute traçabilité des mouvements de fonds pour remonter à l'opération d'origine.

La phase d'intégration qui consiste à réinvestir les fonds dans des placements honorables : biens immobiliers, titres, participations financières dans les entreprises.

Lutter contre le blanchiment de capitaux reviendrait donc à mettre en place des mesures d'vigilance au niveau des acteurs sociaux et économiques pour que les étapes à franchir pour blanchir les capitaux soient difficiles voire impossible. [?]

La lutte contre le financement du terrorisme

Le financement du terrorisme consiste à fournir ou réunir des fonds, des biens ou des services susceptibles d'être utilisés dans le but de faciliter ou de perpétrer des actes de terrorisme. Ces opérations à finalité criminelle impliquent parfois des fonds d'origine parfaitement légale. Alors que le blanchiment des capitaux est une opération financière qui vise à cacher l'origine des fonds, le financement du terrorisme, au contraire, utilise des techniques pour tenter de cacher la destination des fonds.

La lutte contre le financement du terrorisme s'effectue par identification et contrôle du donneur d'ordre, du destinataire effectif de la transaction, et ceci par filtrage par rapport à des listes de sanctions officielles.

Le respect des embargos commerciaux et financiers

La communauté internationale, au travers de l'Organisation des Nations Unies (ONU), s'est dotée d'un arsenal juridique pour permettre le contrôle des flux monétaires. Parmi certaines mesures figure l'embargo commercial. **Un embargo** (généralement partiel), vise à restreindre les relations des pays membres avec le pays concerné et à encadrer strictement ce qu'il est permis de faire ou non en matière de commerce et d'échange. Les embargos se traduisent généralement par des mesures d'interdiction de certains types d'opérations, comme par exemple l'interdiction de commercer sur du matériel d'origine nucléaire ou militaire, ou encore l'interdiction d'exporter les ressources pétrolières d'un pays sous embargo.

La surveillance des opérations de marché

Il s'agit d'une obligation qui vise à s'assurer que la banque ou l'établissement financier n'utilise pas son accès privilégié aux marchés financiers pour en tirer profit au détriment de ses clients. La surveillance des marchés regroupe les fonctions suivantes :

Les délits d'initiés : Pratique consistant à profiter indument d'une information privilégiée avant que celle-ci ne soit rendue publique. Une information privilégiée est une information précise sur un émetteur qui, si elle était rendu publique, serait susceptible d'influencer le cours de certains instruments financiers.

Les manipulation de marché : Cette problématique vise à s'assurer que la banque ou l'établissement financier n'utilise pas son poids financier et son effet de levier sur certains titres pour faire évoluer le marché dans un sens qui lui est favorable.

La résolution des conflits d'intérêts : Cette dernière problématique vise à identifier les éventuels conflits résultant de la multiplicité des activités bancaires au sein d'un grand groupe financier.

3.4 Machine Learning et mise en oeuvre d'un programme de conformité

Après avoir présenté les composantes d'un programme AML, nous allons analyser les moyens à mettre en oeuvre au sein des établissements de crédit pour appliquer de manière opérationnelle les recommandations du GAFI et surtout comment ces moyens pourraient être automatisés grâce au ML.

3.4.1 Le Machine Learning pour la simplification des procédures KYC

KYC est l'acronyme de Know Your Customer. Il désigne le processus permettant de vérifier l'identité des intervenants à une opération bancaire afin de s'assurer de la conformité des clients face aux législations anti-corruption, de leur probité et de leur intégrité.

Initialement mise en oeuvre par une intervention humaine, les tâches répétitives des procédures KYC pourraient être automatisées grâce au Machine learning. Les principales étapes de la procédure KYC qui peuvent être automatisées par des modèles d'apprentissage automatiques sont :

L'identification et le contrôle des informations d'identification des clients

Il s'agit au cours de cette étape de demander et d'enregistrer les informations personnelles du clients et de contrôler son identité par rapport à une pièce d'identité officielle. A ce niveau, les informations personnelles nom, prénoms, date de naissance, situation maritale, adresse... dites « bio data » sont demandées.

Le contrôle des informations d'adresse nécessitera la fourniture par le client d'une pièce probante (facture d'eau, de téléphonie fixe, etc.). La banque pourra également adresser un courrier de bienvenue ou de remerciement pour la fidélité au client et vérifier que le courrier ne revient pas.

Le contrôle des clients par rapport aux listes de sanctions

Lors de toute opération, la banque contrôle la présence éventuelle d'un des intervenants de l'opération sur une ou plusieurs listes de sanction, selon la réglementation en vigueur dans le pays. Ces listes sont établies par les autorités officielles (nationales ou supranationales comme L'ONU, L'Union Européenne). Elles regroupent des individus ou de groupes qui compte tenu de leur activités ont été frappé de mesure d'embargo nominative.

Qualification du risque de blanchiment

Il s'agit là de vérifier si le client n'existe pas sur des listes qui ne sont pas d'ordre public. Ces listes peuvent être celles des PEP ou une liste d'indésirables car en opposition avec la déontologie et les valeurs du groupe financier.

Consignation des pièces d'identification des clients

Après la phase d'identification du client et de son contrôle, l'établissement financier doit enregistrer les preuves d'identification du client et les archiver.

3.4.2 Le Machine Learning, un outil essentiel à la détection des transactions suspectes

La lutte contre le blanchiment d'argent et le financement du terrorisme est principalement basée sur l'élaboration, par des algorithmes, de scénarios d'anticipation dits « déterministes ». Les algorithmes utilisés analysent en temps réel les transactions et sont capables, en quelques instants, de déceler une transaction suspecte. Ces scénarios se basent sur des règles arrêtées, constantes et ne sont que très peu modifiés une fois mis en place. La procédure est basée sur des mots clés et il est difficile de calibrer ces logiciels à un niveau permettant une protection optimale face aux transactions frauduleuses sans générer pour autant un nombre élevé de fausses alertes qui de ce fait viendrait perturber les activités de conformité.

Pour résoudre ces problèmes particulièrement chronophages et coûteux pour les banques, des applications basées sur le Machine Learning pourraient apprendre à identifier les transactions frauduleuses en établissant des procédés standardisés et automatisés. Cela permettra de réduire la charge de travail des équipes, tout en affinant la précision de l'analyse.

Il s'agirait là pour les banques de réduire leurs coûts et leur sanctions, tout en assignant un travail à plus forte valeur ajoutée aux équipes chargées de la conformité.

Conclusion

Le blanchiment d'argent, la lutte contre le terrorisme sont des fléaux dangereux et il est du devoir des banques de lutter efficacement contre ces pratiques. Aux collaborateurs de la SGBF, il est demandé de :

- Appliquer impérativement la réglementation française(exigence du groupe)
- se conformer à la réglementation du Burkina Faso applicable à leur égard. Si Celle-ci est plus restrictive, elle s'applique en priorité tout en restant conforme avec les autres exigences du groupe.
- Appliquer la réglementation américaine pour toute transaction vers les Etats-Unis ou impliquant le dollar Américain.

Ce chapitre nous a permis de présenter la conformité dans le domaine bancaire, et les facilités qu'apporterait l'apprentissage automatique dans sa mise en oeuvre dans un établissement financier. La suite sera consacrée à l'implémentation et à la présentation des résultats.

CHAPITRE 4

APPROCHE ET IMPLÉMENTATION

Introduction

Dans ce chapitre nous présenterons premièrement les données qui ont été utilisés pour entrainer notre modèle. Ensuite, nous présenterons les résultats obtenus à la suite de l'utilisation des algorithmes sur nos données. Enfin nous présenterons les perspectives envisagées. Quelle approche est la mieux adaptée à nos besoins ? Choisir un algorithme d'apprentissage supervisé ou non supervisé dépend habituellement de facteurs liés à la structure et au volume de nos données, et le cas d'utilisation auquel nous voulons l'appliquer.

4.1 Approche

L'objectif de notre travail est de mettre en place un système permettant d'analyser ou de réaliser la mise en conformité d'une opération à l'étranger et de tous les acteurs intervenants dans cette opération. Cette analyse se base sur le dossier que le client a fourni au guichet des opérations internationales.

L'analyse conformité regroupe plusieurs activités que nous ne considérerons pas toutes dans un premier temps. En effet, dans le but d'avoir un modèle beaucoup plus efficace, nous nous focaliserons tout d'abord sur :

- La détection des transactions suspectes
- l'allègement des procédures KYC

Pour atteindre les objectifs que nous nous sommes fixés, nous considérons un dossier d'opération comme l'ensemble des informations contenues sur chaque élément du dossier.

Un dossier d'opération est donc l'ensemble des caractéristiques entrant en compte dans la détection d'une transaction suspecte et des informations d'identification de tous les intervenants à l'opération. Ces informations sont relevées directement sur le dossier physique transmis par le client.

La détection d'anomalies dans notre cas n'a pas fourni de résultats satisfaisants. En effet, elles éliminent une énorme partie des données d'entrées, qui ne représentent pas forcément des anomalies. Nous avons donc cherché d'autres méthodes pour la résolution de notre problème. Pour cela, la méthode adoptée pour la mise en place du système devra respecter les exigences bancaires à savoir que la mise à jour du système doit être possible et la décision prise explicable et le système sécurisé. En somme, le modèle ne devra pas fonctionner comme une boîte noire.

Parmi les principaux algorithmes de classification connus, ceux ne fonctionnant pas comme une boîte noire sont l'algorithme des plus proches voisins(K-NN) et celui des arbres de décisions(Décision Tree). Les arbres de décisions offrent l'avantage de pouvoir générer des règles de décisions pour toutes les différentes classifications qui sont réalisées. C'est la méthode qui sera utilisée pour réaliser notre modèle d'apprentissage.

Nous posons comme hypothèse de départ que tous les dossiers sont complets et conformes à la réglementation de change avant de passer par le modèle d'apprentissage. Cette hypothèse nous permet de nous focaliser sur les autres aspects de la conformité qui ont été cités précédemment.

4.2 Réalisation

4.2.1 Le jeu de données

Acquisition des données

Les moyens à mettre en oeuvre au sein des banques pour appliquer de manière opérationnelle les recommandations du GAFI et se conformer ainsi aux réglementations en vigueur (ordonnance 2009-104, code monétaire et financier) implique un processus approfondi de connaissance du client et le contrôle et la surveillance des transactions.

Les éléments présents sur un dossier et permettant de connaître un client intervenant dans une opération(client émetteur ou destinataire de l'opération) sont :

- le type de personne (personne physique ou morale)
- l'identité de la personne
- le pays de résidence de la personne
- la banque de la personne
- le pays dans lequel cette banque se trouve

Tous les éléments cités ci-dessus doivent être contrôlés sur les listes officielles de vérification de sanctions et d'embargo. Le résultat de chacun de ces contrôles est indispensable pour effectuer l'analyse d'une opération.

Les informations présentes sur un dossier et permettant de contrôler ou surveiller une transaction en cours sont :

- L'activité de l'émetteur de l'opération
- l'activité du bénéficiaire de l'opération
- l'objet de l'opération(salaire, achat d'une voiture, de frais médicaux...)
- le type de l'opération(transfert émis, transfert reçu, credoc, remdoc...)
- le montant de l'opération
- la devise de l'opération

En somme, Pour inférer correctement sur de nouvelles données, les jeu de données qui seront utilisés pendant la phase d'apprentissage auront les caractéristiques suivantes :

- le type de personne du donneur (personne physique ou morale)
- l'identité de l'émetteur et le résultat de son contrôle sur les listes de sanctions
- le pays de résidence de l'émetteur et sa notation
- la banque de l'émetteur et le résultat de son contrôle sur les listes de sanctions
- le pays dans lequel cette banque se trouve et sa notation
- le type de personne du bénéficiaire (personne physique ou morale)
- l'identité du bénéficiaire et le résultat de son contrôle sur les listes de sanctions
- le pays de résidence du bénéficiaire et sa notation
- la banque du bénéficiaire et le résultat de son contrôle sur les listes de sanctions
- le pays dans lequel cette banque se trouve et sa notation
- l'activité de l'émetteur de l'opération
- l'activité du bénéficiaire de l'opération
- l'objet de l'opération(salaire, achat d'une voiture, de frais médicaux ...)
- le type de l'opération(transfert émis, transfert reçu, credoc, remdoc...)
- le montant de l'opération
- la devise de l'opération

Comme nous ne disposons pas de toutes ces informations sur des fichiers, nous avons, en collaboration avec les collaborateurs du service des opérations internationales et ceux de la Direction Conformité, constitué un jeu de données afin de réaliser notre apprentissage.

4.2.2 Prétraitement des données

Les caractéristiques recensées ci-dessus nous permettent de juger de la conformité d'un dossier de transferts. Certaines caractéristiques sont très distinctives et pourraient entraîner un sur-apprentissage de notre modèle. Il s'agit par exemple de

- l'identité de l'émetteur de l'opération
- le pays de résidence de l'émetteur
- la banque de la l'émetteur
- le pays dans lequel cette banque se trouve
- l'identité du bénéficiaire de l'opération
- le pays de résidence du bénéficiaire de l'ordre
- la banque du bénéficiaire
- le pays dans lequel cette banque se trouve
- la devise de l'opération

Dans le tableau ??, nous présentons l'ensemble des caractéristiques de notre jeu de données. Celles qui sont en italiques représentent les caractéristiques d'entrée de notre algorithmes de machine learning.

Le jeu de données final qui a servi pour l'apprentissage et les tests est constitué de six cent (600) dossiers d'opérations.

Pour des questions pratiques, nous avons constitué un dictionnaire des différents secteurs d'activités ainsi que des objets de transactions. Un échantillon du dictionnaire est présenté au tableau ??.

4.2.3 Validation croisée et stratification des données

Validation croisée

Pour nous assurer que notre modèle ne souffre pas de sur-apprentissage, et qu'il saura faire des prédictions sur de nouvelles données, nous avons implémenté la validation croisée sur notre modèle de Decisions Tree. La validation croisée va nous permettre d'utiliser l'intégralité de notre jeu de données pour l'entraînement et pour la validation.

Pratiquement, il s'agit de découper le jeu de données en k parties (folds en anglais) à peu près égales. Tour à tour, chacune des k parties est utilisée comme jeu de test. Le reste (autrement dit l'union de $k - 1$ autres parties) est utilisé pour l'entraînement. La validation croisée permet d'éviter un biais potentiel lié au fait de faire une évaluation unique.

Section	Caractéristiques	Exemples
Emetteur de l'ordre	<i>Type de personne</i>	Personne physique
	Identité	XXXXXX XXXXXX
	<i>Résultat du contrôle de l'émetteur</i>	Aucune sanction
	Pays de résidence	Burkina Faso
	<i>Notation du pays de résidence</i>	LOW
	Banque de l'émetteur	SGBF
	Pays de la banque	Burkina Faso
	<i>Notation Pays de la banque</i>	LOW
	<i>Résultat du contrôle sur la banque</i>	Aucune sanction
	<i>Activité de l'émetteur</i>	Activités extractives
Bénéficiaire de l'ordre	<i>Type de personne</i>	Personne morale
	Identité	ZZZZZZZZZZ
	<i>Résultat du contrôle sur la personne</i>	Aucune sanction
	Pays de résidence	France
	<i>Notation du pays de résidence</i>	LOW
	Banque de l'émetteur	BNP Paribas
	Pays de la banque	France
	<i>Notation Pays de la banque</i>	LOW
	<i>Résultat du contrôle de la banque</i>	Aucune sanction
	<i>Activité du bénéficiaire</i>	Hébergement et hôtellerie
Opération	Type	Règlement de facture
	Objet	Frais d'hébergement
	Montant	25000
	Devise	Euros

TABLE 4.1 – Exemple de dossier d'opération conforme

Libellé secteur d'activité	Code	Libelle du secteur d'activité	Code
Activités extractives	0	Activités financières	15
Agriculture et chasse	1	Hôtels et restauration	3
Industrie	4	Activites de ménages	5
Activités des organisations extraterri- toriales	6	Activités financières	7
Commerce gros	8	Santé et action sociale	8
Administration publique	9	Commerce détail	
Transport	10	Education	11
Développements logiciels	12	Maintenance de materiels informa- tiques	13
Fabrication Produits pharmaceutiques	16	Construction	17
Fabrication de meubles	18	Activités associatives	19
Commerce détail	20	Fabrication chaussures	21
Télécommunications	22	Activités Juridiques	23
Fabrications produits alimentaires	24	Service immobilier	25
Fabrication de boissons	26	Pêche et pisciculture	27

TABLE 4.2 – Codage de quelques secteurs d'activités

Stratification

Le jeu de données dont nous disposons n'est pas équilibré i-e le nombre de dossiers non conformes est plus élevé que celui des dossiers conformes. La stratification permet d'éviter que les données d'entrainements ne contiennent que des exemples positifs et les données de test que des exemples négatifs, ce qui affecte négativement les performances du modèle.

4.2.4 Les outils

Nous allons présenter quelques outils qui nous ont permis de mettre en place le modèle et la plateforme que nous avons proposée. Chaque étape dans la mise en place d'un modèle de machinelearning possède des outils spécifiques associés.

Le langage python



Python a été utilisé pour les codes d'implémentation de notre modèles. Il s'agit d'un langage de programmation, dont la première version est sortie en 1991. Ce langage a été baptisé ainsi en hommage à la troupe de comiques les « Monty Python ». Python est un langage puissant riche en possibilités et dont les fonctionnalités peuvent être étendues grâce à de nombreuses bibliothèques. Ainsi, nous avons utilisé de nombreuses bibliothèques python afin de mettre en oeuvre notre modèle.

Pandas et numpy : Pour le nettoyage et l'exploration de nos données, les librairies **Pandas** et **Numpy** ont été utilisées. Pandas permet de créer des tableaux ou dataframes à partir des données brutes.

Scikit-learn : Tensorflow et Scikit-learn sont les librairies les plus utilisées pour la modélisation. Pour notre modèle, le choix a été fait d'utiliser **Scikit-learn**. Ce choix se justifie par le fait qu'il implémentedirectement et de manière didactique les différents algorithmes d'apprentissage automatique.

Le langage Javascript

JavaScript est un langage de programmation de scripts principalement employé dans les pages web interactives mais aussi pour les serveurs avec l'utilisation (par exemple) de Node.js. A travers le framework angular qui est un framework Javascript, il nous a permis de réaliser une interface conviviale qui permettrait l'utilisation de notre modèle.



Flask



Flask est un framework open-source de développement web en Python. Son but principal est d'être léger, afin de garder la souplesse de la programmation Python, associé à un système de templates. Ils nous a permis de mettre à disposition de notre client javascript le modèle python qui a été implémenté.

Jupyter Notebook

Le Jupyter Notebook est une application web open-source qui vous permet de créer et de partager des documents contenant du code interprété en direct, des équations, des visualisations et du texte narratif. Les utilisations comprennent : le nettoyage et la transformation de données, la simulation numérique, la modélisation statistique, la visualisation de données, l'apprentissage machine, et bien plus encore. Il est l'un des outils du projet Jupyter.



Visual studio code

Visual studio code est un éditeur de code édité par microsoft. Il est utilisé par les développeurs pour la programmer dans de nombreux langages de programmation.

4.3 Résultats

Rappelons que l'objectif de notre étude est mettre en place un système permettant de classer une opération(dossier de transfert) à l'étranger selon la conformité. Un tel système se compose de deux parties. La première partie est un modèle de machine learning réalisé grâce aux algorithmes de décision Tree. La seconde est une application permettant d'envoyer à partir d'un formulaire les éléments du dossier au modèles de Machine Learning.

Nous présenterons tout d'abord les résultats du modèle. Par la suite, nous montrerons l'application qui permettra une utilisation du modèle.

4.3.1 Résultats du modèle

Le modèle classe les dossiers opérations en deux groupes : un groupe représentant l'étiquette dossier conforme, l'autre représentant l'étiquette dossier non-conforme. Nous étiquetons un dossier conforme 1 et un dossier non-conforme 0 Les mesures détaillés des test par catégorie sont représentés sur les figures suivantes.

	precision	recall	f1-score	support
0	0.94	0.86	0.90	72
1	0.77	0.89	0.82	37

	precision	recall	f1-score	support
0	0.95	0.96	0.95	77
1	0.90	0.88	0.89	32

	precision	recall	f1-score	support
0	0.94	0.96	0.95	81
1	0.88	0.82	0.85	28

	precision	recall	f1-score	support
0	0.93	0.96	0.94	70
1	0.92	0.87	0.89	39

	precision	recall	f1-score	support
0	0.95	0.91	0.93	76
1	0.80	0.88	0.84	32

FIGURE 4.1 – Résultat du test.

Notre test nous révèle un f1-score toujours élevé pour les dossier étiquetés 0 c'est-à-dire pour les dossiers non-conformes. La moyenne de prédiction juste globale est de 61.31%.

Les arbres de décisions étant considérés comme des classifieurs faibles, nous avons appliqué sur nos données un modèle de forêts aléatoire afin de comparer ces résultats avec ceux issus d'un arbre de décision simple.

Les résultats obtenues pour un modèle de forêt aléatoire sont présentés dans la figure ci-dessous. ??

Ce second test révèle toujours un f1-score toujours élevé pour les dossiers non-conformes. La moyenne de prédiction juste est cette fois-ci de 83%.

```

=== Classification Report ===

```

	precision	recall	f1-score	support
0	0.93	0.96	0.95	72
1	0.91	0.86	0.89	37

FIGURE 4.2 – Résultat de l’entrainement avec Random Forest

4.3.2 Implémentation de la plateforme web

Pour pouvoir être utilisé par les collaborateurs de la SGBF, le modèle d’analyse des dossiers qui a été implémenté devra être utilisable à travers une interface utilisateur conviviale. Cette interface en plus d’envoyer des données au modèle, devrait permettre de contrôler l’intégrité et la fiabilité des acteurs de l’opération. Les fonctionnalités attendus sont :

- Permettre l’enregistrement de toutes les informations concernant une nouvelle opération dans une base de données.
- Faciliter la vérification sur les différentes listes officielles de sanctions et d’embargo.
- Permettre une traçabilité des opérations de l’entrée en relation jusqu’à l’exécution de l’opération

Ainsi, la plateforme qui a été mise en oeuvre permet aux collaborateurs du services OPI de renseigner les informations présent dans le dossier et permettant de juger de la conformité d’une opération.

A la réception d’un dossier, le collaborateur renseigne les informations sur la transaction sur l’interface présentée sur la figure ???. Les informations sur l’émetteur de l’ordre sont saisies sur l’écran de la figure ??, celle sur le bénéficiaire sur l’écran présenté à la figure ??.

A l’enregistrement, les informations sont transmises au modèle pour analyse. Le résultat de l’analyse est affiché sur l’écran de la figure ???. On retrouve sur cet écran, le sommaire des informations sur l’émetteur, sur le bénéficiaire et sur l’opération elle-même.

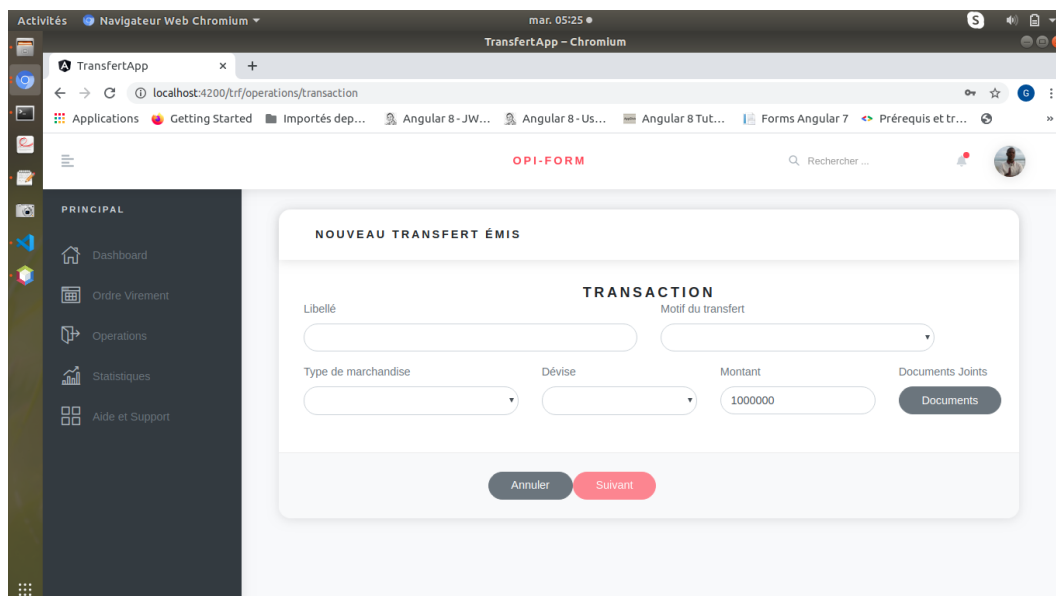


FIGURE 4.3 – Ecran de renseignement des informations sur la transaction.

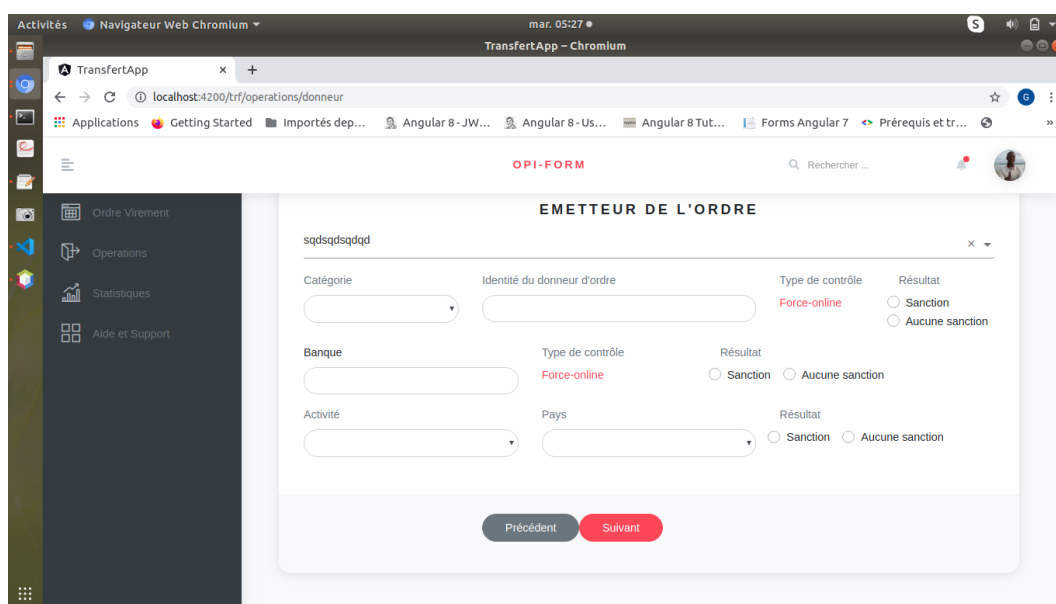


FIGURE 4.4 – Ecran de renseignement des informations sur l'émetteur de l'ordre.

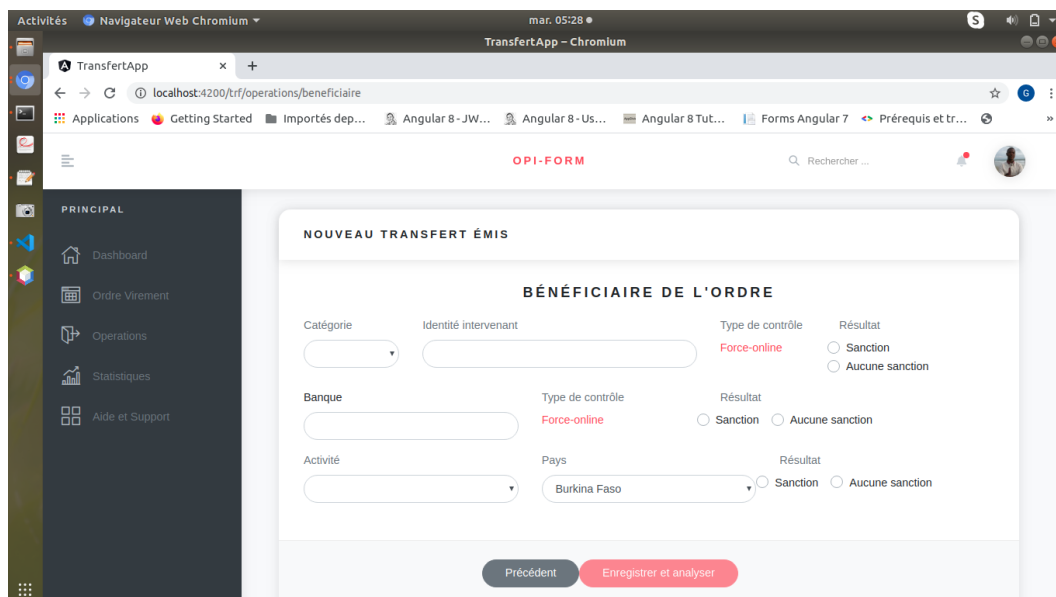


FIGURE 4.5 – Ecran de renseignement des informations sur le bénéficiaire de l'ordre.

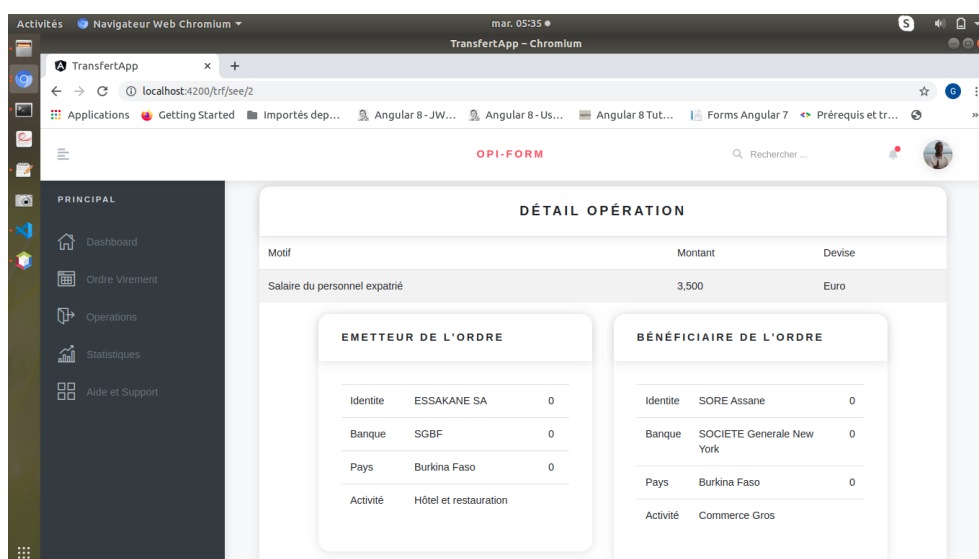


FIGURE 4.6 – Ecran de présentation des détails et du résultat de l'analyse .

4.4 Interprétation des résultats du modèle de machine learning

4.4.1 Limites et difficultés

Sans passer par mille chemins, notre stage était miné de difficultés d'ordre organisationnelles au sein de la banque et des difficultés techniques rencontrées tous les jours.

Les difficultés organisationnelles sont dues à l'acquisition et à la manipulation des données dont nous avons besoin pour l'implémentation de notre modèle. En IA, il est une barrière qui demeure impossible à franchir. « Pas de datas, pas d'IA ». La principale difficulté a été l'obtention des données pour notre modèle. Les modèles de machine learning se construisent à partir d'exemples d'apprentissage basés sur l'expérience passée. Disposer d'exemples est donc indispensable. Dans le même temps, ces exemples doivent être présents et suffisamment en grand nombre pour parvenir à une IA généralisable et applicable sur le terrain.

Cette étape a été la plus difficile car elle a mis à contribution de nombreux collaborateurs de plusieurs services différents(DCO, OPE). Cela nous a permis d'obtenir un jeu de données pour l'apprentissage.

Les difficultés techniques sont liées au choix des différents outils d'implémentations du modèle, les pare-feux de la SGBF n'autorisant pas l'installation de certaines applications sur les machines du réseau. Pour les outils dont l'installation était impossible, il fallait donc trouver d'autres permettant de faire la même tâche beaucoup plus difficilement.

4.4.2 Analyse des résultats

La matrice de confusion permet de résumer et visualiser les résultats d'un problème de classification. Des mesures permettent d'analyser la matrice de confusion. Ce sont

La précision : Elle permet de calculer le taux de classification juste. C'est la proportion des prédictions correctes parmi les points que l'on a prédit.

$$Precision = \frac{VP}{VP + FN}$$

Le rappel ou sensibilité : En anglais *recall*, il donne la proportion des exemples bien étiquetés.

$$Rappel = \frac{VP}{VP + FP}$$

L'exactitude : En anglais *accuracy*, il évalue le taux de bonnes réponses.

$$Accuracy = \frac{VP + VN}{VP + FP + VN + FN}$$

F1-score Il s'agit de la moyenne harmonique de la précision et du recall. Il reflète les différents aspects du modèle.

$$F1 = 2 * \frac{Precision * Rappel}{Precision + Rappel}$$

Les résultats obtenus après évaluation de notre modèle de classification font ressortir

quelques éléments. Globalement, la mesure du F1-score du modèle basé sur les arbres de décisions est d'environ 60%. Les scores propres pour chacune de nos étiquettes montre une certaine différence. Le modèle prédit beaucoup plus facilement les dossiers non-conforme que les dossiers conformes. En effet le F1-score pour les dossiers non-conforme est approximativement de 93% alors que celui des dossiers conformes est de moins de 86%. cela pourrait être du fait que notre jeu de donnée contient plus de données d'opérations non conformes que d'opérations conformes.

Concernant l'exactitude de notre modèle, Peter Pan [?] obtenait un score 97% sur le dataset Iris. Le dataset iris est un jeu de données ouvert plusieurs fois cité dans la littérature. L'ensemble des données contient 3 classes de 50 instances chacune. Chaque classe se réfère à un type de classe iris.

Jean philippe Vandamme et al. [?] ont mené une étude sur le taux d'échec en première année d'université. Ils ont essayé de prédire à partir de certains attributs qu'un étudiants puisse réussir son année(low-risk), réussir moyennant des actions menées par l'université(medium-risk) ou échouer(high-risk). Sur un ensemble de 533 étudiants questionnés sur un ensemble de 20 questions, ils ont obtenu un taux globale de bonne prédiction de 40,63%. Ce taux est inférieur à celui que nous avons obtenu.

Ainsi, un volume plus important de données avec une répartition égale pour chaque type de dossier permettrait d'atteindre des résultats plus beaucoup plus satisfaisant. Néanmoins, les résultats obtenus sont prometteurs et nous permettent d'affirmer qu'il est possible d'utiliser du machine learning pour analyser la conformité d'une opération à l'étranger.

4.4.3 Perspectives

L'analyse des dossiers d'opérations à l'étranger est une tâche fastidieuse. Le modèle qui a été mis en oeuvre comporte de nombreuses imperfections.

Concernant la diversité de nos données

Pour notre modèle les données que nous avons utilisées relèvent de seulement vingt cinq secteurs d'activités et 40 objets de transaction. Pour être efficace, le modèle a besoin d'apprendre du plus grands nombres de secteurs d'activités et également et de tous les objets de transaction de ces activités.

Concernant le score de prédiction juste obtenu

La conformité dans le domaine bancaire est très sensible. En effet, il s'agit d'un domaine dans lequel l'erreur n'est pas autorisée. Une transaction suspecte qui passe les mailles établies par la DCO entraine une cascade de sanction sur l'institution financière

en cause. Les structures bancaires ont donc besoin d'un modèle qui puissent leur fournir un résultat très fiable. Ainsi notre score de 63% de prédiction juste devrait être amélioré.

Conclusion

Ce chapitre nous a permis de présenter l'implémentation du modèle d'apprentissage automatique que nous proposons pour l'analyse des dossiers de transferts. Les résultats que nous obtenons sont satisfaisants mais pourraient être améliorés.

CONCLUSION GÉNÉRALE

Le problème qui nous a été posé était l'application du machine learning dans l'analyse des opérations à l'étranger. De nombreuses opérations sont menées quotidiennement par la Société Générale Burkina Faso. Ces opérations sont délicates car engageant de nombreuses personnes : la personne qui émet l'opération, son correspondant bénéficiaire de l'opération, la banque émettrice et celle bénéficiaire et les nombreux intermédiaires.

L'expérience que nous avons menée durant nos 6 mois de stage à la Société Générale, nous ont permis de comprendre le processus d'analyse et de mise en conformité d'une opération à l'étranger. Ce stage nous a également permis de détecter les processus d'analyse qui peuvent être automatisés grâce à l'apprentissage automatique.

L'objectif de notre stage était de mettre en place une plateforme intelligente qui permettrait d'analyser la conformité des dossiers de transferts déposés au guichet de la société Générale Burkina Faso. La bonne réalisation de notre projet nous contraignait à une exploration dans l'univers des algorithmes de classification afin de trouver celui qui répondait le mieux à la spécification de notre problème. Notre choix s'est porté sur les arbres de décisions et c'est grâce à eux que nous avons réalisé notre projet. Le jeu de données que nous avons utilisé pour entraîner notre modèle a été obtenu à partir des dossiers physiques et grâce à la collaboration avec les membres des services impliqués dans le processus d'analyse du dossier d'une opération à l'étranger.

Le test d'évaluation de notre modèle a donné un score de 63%. Les résultats obtenus mettent en évidence l'inégale répartition de nos données dans les différentes classes. En effet, le modèle prédit mieux les opérations non conformes que les dossiers conformes. Une seconde approche qui visera un approfondissement de notre modèle ne produira-t-elle pas de meilleurs résultats ?

ANNEXE A

ANNEXE : DOSSIERS DE TRANSFERTS

De nombreux documents entrent en compte dans la constitution d'un dossier d'opération à l'étranger. Dans cette annexe, nous présentons les principaux motifs d'opération à l'étranger et les documents composant le dossier.

A.1 Pour un règlement de facture d'achat

Les documents constituant un dossier pour une opération de règlement de facture sont :

- Un ordre de transfert précisant les coordonnées bancaires du bénéficiaire
- Une autorisation de change
- La facture réelle
- Une déclaration préalable d'importation(DPI)
- La facture proforma ayant servi à lever la DPI
- La facture définitive liée à la proforma ou à la DPI
- Une copie de l'attestation d'importation(AI SYLVIE) ou original si visée par la douane
- L'autorisation Spéciale d'Importation si produit spécifique

A.2 Pour le remboursement d'un emprunt

Le remboursement d'un emprunt est une opération permettant de régler un prêt qui avait été consenti auprès d'une banque ou d'un organisme. Les principaux documents entrant dans cette opération sont :

- Un ordre de transfert
- La convention de prêt
- Le tableau d'amortissement
- L'autorisation de change
- La preuve de rapatriement des devises reçues
- La preuve d'encaissement des fonds au Burkina Faso

A.3 Pour une prestation de service

De nombreuses entreprises extérieures effectuent des assistances techniques au Burkina Faso. Les documents permettant un transfert pour une prestation réalisée sont :

- Un ordre de virement précisant les coordonnées bancaires du bénéficiaire
- La facture de prestation
- L'autorisation de change
- Le contrat de prestation de service
- La quittance de retenue à la source si la prestation est fournie ou utilisée au Burkina Faso

A.4 Pour des frais de scolarité ou un soutien familial

- Un ordre de transfert précisant les coordonnées bancaires du bénéficiaire
- L'autorisation de change
- Un document attestant de l'inscription de l'étudiant pour l'année en cours ou de la présence du bénéficiaire à l'étranger.

A.5 Pour le virement des salaires des personnes expatriés

Les salaires des personnes de nationalité étrangères sont virés en devises. Les documents entrant en comptes dans une telle opération sont :

- L'ordre de transfert précisant les coordonnées bancaires du bénéficiaire
- L'autorisation de change
- Le bulletin de salaire fait avec l'IUTS
- Le contrat de travail
- La copie du passeport

ANNEXE B

ANALYSE ET CONCEPTION DE LA PLATEFORME WEB

Nous présentons dans cette partie les phases d'analyse et de conception de l'application web. Il s'agira d'analyser le problème posé afin de concevoir une application répondant aux besoins qui ont été exprimés.

B.1 Analyse des besoins

Les besoins exprimés par la SGBF se résume en la mise en place d'un système permettant l'analyse vis à vis de la conformité d'un dossier de transfert déposé au guichet des opérations internationales. Partant de là, nous avons pu identifier les cas d'utilisations présentés dans le tableau ??.

Les principaux acteurs qui interagiront avec le système sont :

Le guichetier : Il est chargé de la réception du dossier et de l'analyse de la cohérence et de la complétude du dossier. A la suite de cette analyse, il peut faire des observations sur le dossier.

L'administrateur : Il a pour rôle la cration de nouveaux utilisateurs sur la plateforme. Il peut également consulter les statistiques.

Le diagramme des cas d'utilisation résultant est le suivant. Il est présenté à la figure ??.

Cas d'utilisation	Description
Gérer un dossier	Enregistrer ou modifier un dossier. L'analyse conformité d'un dossier intervient juste après cette étape.
Faire des observations	Faire des observations sur la cohérence et la complétude du dossier
Gérer les accès au système	Création, modification et authentification des utilisateurs
Consulter des statistiques	Visualiser et exporter des données
Administrer le système	créer ou modifier les informations des utilisateurs. Attribuer des privilèges à un utilisateurs.

TABLE B.1 – Les principaux cas d'utilisation du système

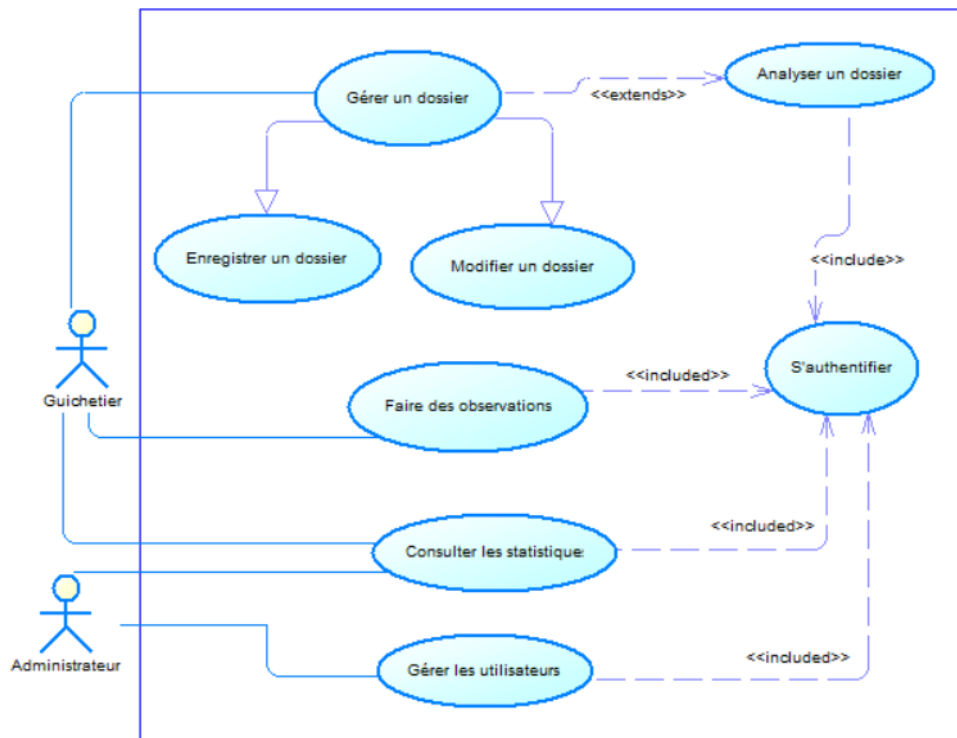


FIGURE B.1 – Diagramme des cas d'utilisation.

B.2 Conception

Après l'identification des spécifications fonctionnelles du système, nous devons procéder à sa conception. Pour y parvenir, nous avons suivi les étapes suivantes :

- Identification des entités ou concepts du domaine d'étude
- Identification et ajout des associations et des attributs
- Organisation et simplification du modèle en éliminant les classes redondantes

Le diagramme de classe obtenu est présenté à la figure ??.

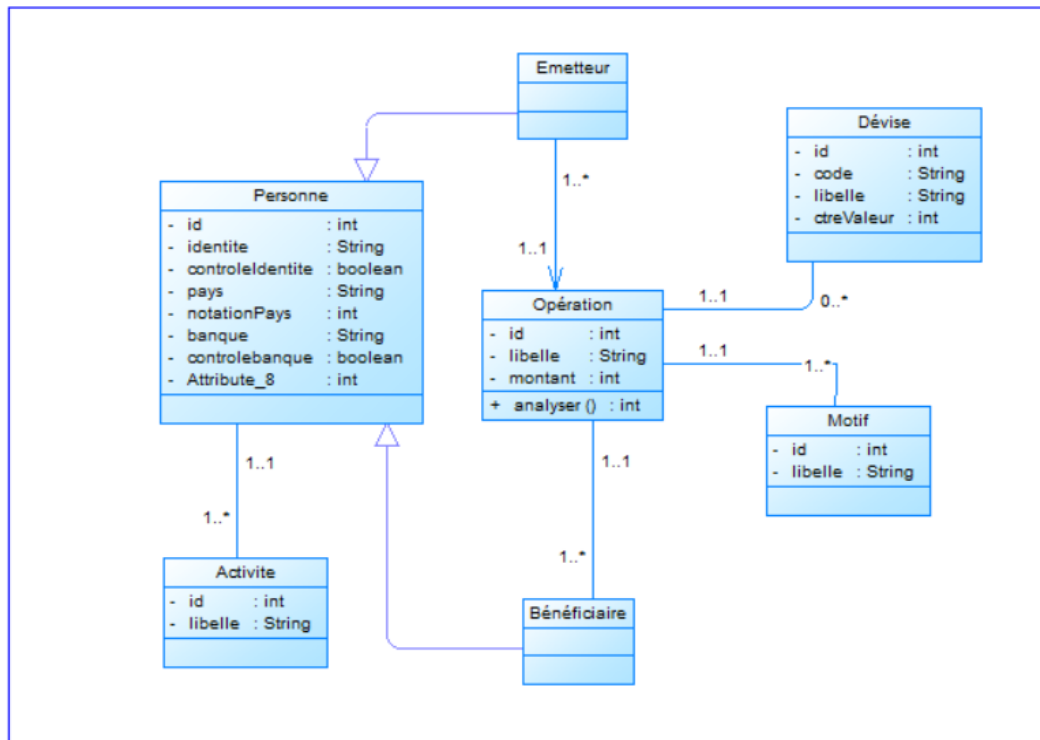


FIGURE B.2 – Diagramme de classes.