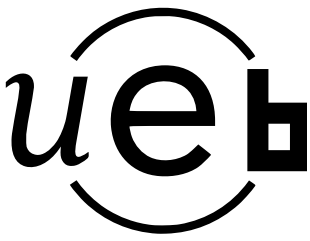
ANNÉE 2020



la citation , voir www.supercitations.com
par bob

Remerciements

Je remercie Prenom Nom, Titre, qui me fait l'honneur de présider ce jury.

Je remercie Prenom Nom, Titre, et Prenom Nom, Titre, d'avoir bien voulu accepter la charge de rapporteur.

Je remercie Prenom Nom, Titre, et Prenom Nom, Titre, d'avoir bien voulu juger ce travail.

Je remercie enfin Pasteur Poda, Maitre de Conférences, et Prénom Nom, Titre, qui ont dirigé ma thèse.

Il faut vraiment que je dise merci ? :o)

vous n'êtes pas obligés d'utiliser cette commande mais elle vous donnera une idée de la chose Je remercie Prenom Nom, Titre, qui me fait l'honneur de présider ce jury.

Je remercie Prenom Nom, Titre, et Prenom Nom, Titre, d'avoir bien voulu accepter la charge de rapporteur.

Je remercie Prenom Nom, Titre, et Prenom Nom, Titre, d'avoir bien voulu juger ce travail.

Je remercie enfin Pasteur Poda, Maitre de Conférences, et Prénom Nom, Titre, qui ont dirigé ma thèse.

TABLE DES MATIÈRES

INTRODUCTION

Traditionnellement, de nombreuses grandes banques se sont appuyées sur d'anciens systèmes experts fondés sur des règles pour détecter la fraude, mais ces systèmes se sont révélés trop faciles à battre ; Le secteur des services financiers s'appuie sur des algorithmes de détection des fraudes de plus en plus complexes. Les opérations bancaires à l'étranger comprennent de nombreuses activités et transactions quotidiennes, périodiques et apériodiques effectuées par, ou touchant à de nombreuses parties prenantes telles que les employés, les clients, les débiteurs et des entités externes. La nature complexe de ces activités et de ces activités et transactions nécessite une surveillance constante pour s'assurer que ni la banque, ni ses employés ne sont exposés à des risques.

Pour prévenir ces différents risques, la Société Générale Burkina Faso, à travers le service Innovation nous a confié le thème ci-après :

> *Application des techniques de Machine Learning à l'analyse de la conformité des dossiers de transfert à l'étranger.*

L'étude que nous allons mener devra nous permettre de montrer s'il est possible oui ou non d'utiliser les techniques de Machine Learning dans l'analyse de la conformité d'un dossier de transfert d'argent qui aura été transmis au service de la Banque.

Dans la suite de notre travail nous présenterons tout d'abord le contexte et le sujet de notre étude. Par la suite, nous présenterons la conformité dans le domaine bancaire. Après avoir compris les concepts clés, nous proposerons une méthode d'analyse des dossiers ainsi que les résultats que nous obtenons la méthode que nous allons utiliser.

CHAPITRE 1

SUJET ET CONTEXTE GÉNÉRALE DE L'ÉTUDE

Introduction

L'un des critères les plus importants de la réussite d'un projet est la satisfaction du client et des utilisateurs finaux. Nous ne pouvons satisfaire le demandeur sans avoir compris les problèmes qui ont suscité la naissance du projet. Dans ce chapitre, nous présenterons dans un premier temps la structure dans laquelle nous avons effectué notre stage. En second lieu, mous parlerons du projet qui nous a été confié, de son intéret pour l'entreprise, de ses objectifs et de la problématique y afférente.

1.1 La structure d'accueil

1.1.1 La structure d'accueil

Présentation de la Société Générale Burkina Faso

Nous avons effectué notre stage au sein de la Société Générale Burkina Faso (SGBF) une filiale du groupe français Société Générale (SG). La Société Générale Burkina Faso exerce dans la banque de détail et les services financiers, la gestion d'actifs et services aux investisseurs et dans la banque de financement et d'investissement. Elle est présente au Burkina Faso depuis mai 1998. Elle est née de la cession par l'état de 51% du capital de la Banque pour le Financement du Commerce et des investissements au Burkina (BFCI-B). De mai 1998 au 08 février 2013, la filiale du Burkina s'appelait Société Générale des Banques du Burkina (SGBB). A partir du 08 février 2013, elle change de dénomination sociale et prend désormais le nom de Société Générale Burkina Faso (SGBF).

L'organnigramme de la SOciété Générale Burkina Faso se présente comme suit : **??**.

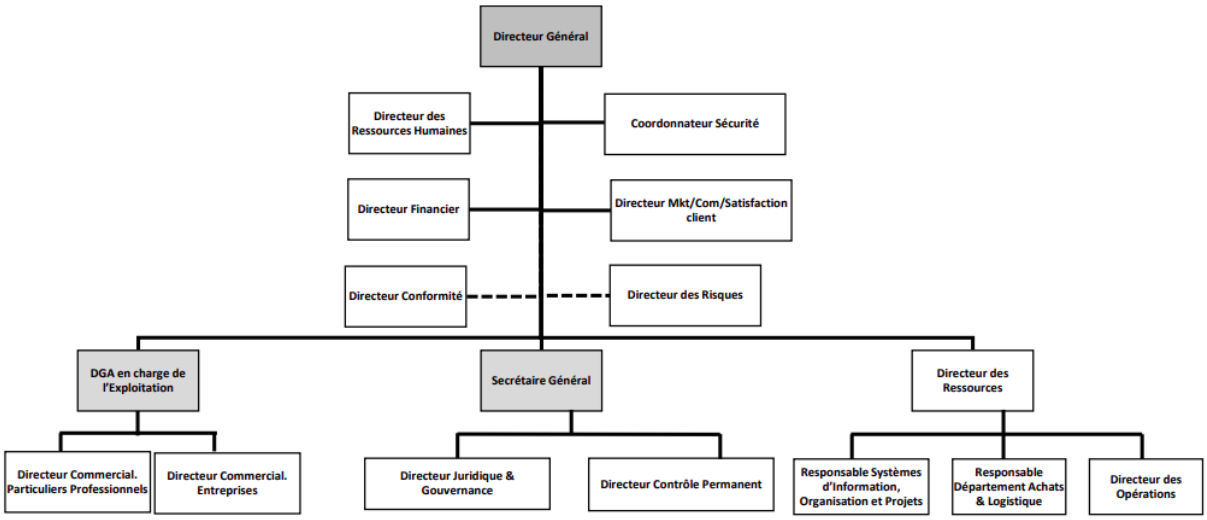


FIGURE 1.1 – Organigramme de la société générale Burkina Faso.

La direction des resources

Nous avons effectué notre stage au sein de la Direction des Ressources plus précisément dans la cellule Innovation de cette direction. La Direction des Ressources est l'entité chargée de gérer toutes les ressources matérielles et logicielles de la banque. Les missions de la cellule Innovation, service dans lequel nous avons effectué notre stage sont les suivantes :

— Continuer de diffuser la culture de l'innovation

— Identifier de nouveaux business et services pour les clients

— Développer de nouveaux process optimaux dans la réalisation des tâches quotidiennes des collaborateurs.

— Favoriser l'émergence d'innovations de rupture et tirer parti des technologies et de la gestion des données.

1.2 Présentation du sujet

1.2.1 Libellé du sujet

Application des techniques de Machine Learning à l'analyse de la conformité des dossiers de transferts à l'étranger.

1.2.2 Contexte du sujet

L'esprit d'innovation traverse toutes les activités de la banque d'où son nouveau slogan « C'est vous l'avenir ». A travers cette approche, la Société Générale cherche des moyens qui lui per-

mettront d'améliorer la satisfaction client par traitement qualitatif et rapide des opérations. La Société Générale Burkina Faso dans ses missions quotidiennes, effectue pour ses clients des opérations transferts à l'étranger(pays n'appartenant pas à la zone UEMOA). Ces opérations comportent de nombreux risques de violation des différentes règlementations en cours.

Effectuer une opération de transfert à l'étranger commence par la constitution d'un dossier appellé dossier de transfert. Lorsqu'un client veut effectuer un transfert vers un pays étranger, la règlementation exige qu'il transmette à l'intermédiaire agréé qui est la banque un certain nombre d'éléments permettant de justifier le transfert qu'il voudrait initier. L'ensemble de ces éléments constitue le dossier de transfert. L'analyse du dossier par les experts de la règlementation s'effectue après que le client soit reparti. Vingt-quatre heure peuvent s'écouler entre le moment où le dossier est déposé au guichet et celui où le transfert est réellement effectuer et ceci lorsque le dossier ne comporte aucune irrégularité.

En cas d'irrégularités. le dossier est transmis au conseiller clientèle qui se chargera de contacter le client pour lui demander de passer corriger les irrégularités si elle peuvent être corrigés.

Dans un domaine où la satisfaction du client et la célérité dans le traitement des opérations est l'une des exigences, un délai de vingt-quatre heures pour traiter une opération est long. Lorsque le dossier comporte certaines irrégulariés ce délai sera rallonger. En effet, entre le jour où l'irrégularité est levée et corrigée, une semaine peut s'écouler et pendant ce temps le processus de transfert de fond est à l'arrêt.

Lorque un dossier non conforme à la conformité passe malgré le système d'analyse mis en place par l'institution financière ce sont d'importantes sanctions financières qui seront imposées à la banque. En 2018, la SG a payé une amende de 1.2 milliards d'Euros pour des opérations en dollars vers des entités sanctionnées par les autorités américaine.

Dans cet ordre d'idée nous nous sommes posés la question comment permettre aux collaborateurs du services des opérations internationales de donner aux clients l'état de leur dossier vis-à-vis de la conformité immédiatement au dépôt de son dossier. Cela permettra de diminuer la charge de travail des collaborateurs de la banque.

Le Service des Opérations internationales est un service dépendant de la Direction des Opérations (DOPE). C'est ce service qui a pour rôle de recevoir tous les dossiers de transferts à l'étranger et de traiter toutes les transactions financières à destination ou venant des pays hors de la zone UEMOA.

1.2.3 Intérêt du sujet

Ce projet est d'un intérêt très élevé pour la SGBF car elle vise à apporter des solutions pour le traitement rapide des opérations à l'international. Les résultats attendus sont :

— Un gain en temps,

— Une facilité de prise de décision,

— Une grande disponibilité et de performance

— Un système toujours actuel et compatible

— Faciliter la gestion des réclamations

1.2.4 Problématique du sujet

Le service des opérations internationales (OPI) reçoit en moyenne soixante-dix (70) dossiers d'opérations de transferts par jour. Plusieurs types d'opérations sont effectuées au quotidien par les collaborateurs de OPI.

Une analyse rigoureuse est appliquée après réception d'un dossier au guichet. Cette analyse concerne aussi bien les différents intervenants de l'opération que la nature l'opération elle-même. Cette analyse peut être décrites suivant trois axes majeurs :

— Analyse de la complétude du dossier

— Vérification des hits

— Analyse de la conformité règlementaire du dossier

C'est après toutes ces analyses que le dossier est transmis pour saisie et validation dans le CBS (Core Bank System). Si le dossier comporte des irrégularités, il est transmis au conseiller ou à la conformité pour demande d'accord. Ces derniers devront recontacter le client pour lui demander de venir corriger les irrégularités. Sinon la saisie est validée et le transfert éffectué.

Pour pouvoir mettre en place un système qui puisse analyser des dossiers, il nous faudra trouver des réponses aux questions suivantes :

— Comment codifier un dossier de transfert à l'étranger ?

— Quels sont les caractéristiques qui entre en jeux dans l'analse d'un dossier de transfert ?

— La règlementation financière est un outil qui évolue. Comment mettre en place un système qui puissent suivre ces évolutions ?

— La Banque doit dire au client la raison pour laquelle son dossier a été rejeté. Quels méthodes d'apprentissage sied le mieux à ce problème ?

Pour répondre à ces interrogations, nous allons explorer tous les contours du domaines d'étude, afin de fixer les bases solides qui nous permettrons de mener notre projet à son terme.

1.2.5 Objectifs

L'objectif de l'étude est de développer un environnement permettant une analyse des dossiers et des intervenants des différentes opérations de transferts. Spécifiquement, il s'agit de

— analyser le processus d'analyse de dossier afin de dégager les grandes étapes

— Proposer un modèle de machine learning permettant d'analyser un dossier de transfert fourni en paramètre.

— Mettre en place une interface web permettant d'utiliser le modèle qui aura été mis en place.

Conclusion

Il a été question dans ce chapitre de présenter la structure d'accueil, la SGBF, qui a suivi et coordonner tous les travaux .Ensuite nous avons présenter le sujet qui nous a été confié, dégager sa problématique et l'intérêt qu'il suscite pour la Société Générale Burkina Faso. Dans la suite, il sera question d'exposer les différents concepts echniques qui s'articulent autour de ce sujet.

CHAPITRE 2

LA CONFORMITÉ DANS LE SECTEUR BANCAIRE

Introduction

Dans ce chapitre, nous allons faire une présentation générale des concepts liés à notre domaine d'étude. Nous ferons d'abord une étude de l'existant dans le domaine bancaire. Nous présenterons par la suite les différentes opérations que la SGBF effectue vers l'étranger. Nous ferons enfin une étude de la conformité dans le secteur bancaire beaucoup plus précisement dans le secteur des transferts à l'étranger.

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

2.1 Vocabulaire

2.1.1 Dossier de transfert

Un dossier de transfert est un ensemble de documents qui peut être composé

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. Il demande à la banque de débiter une somme de son compte pour créditer un autre compte. Celui-ci peut se trouver dans la meme banque ou dans un réseau.

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 dévise, ce document permet autorise le change vers la dévise dans laquelle le transfert sera effectué.

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.

La délaration 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 : Les Autorisations Spéciales d'Importer concernent des produits dont la liste est fixée par avis ministériel : Elles concernent le sésame, les céréales, les amande de Karité.

L'attestation d'importation : C'est un document qui est utile dans les règlements financiers liés aux opérations d'importation. L'attestation d'Importation est signé par la Douane pour attester l'effectivité de l'importation.

L'analyse de la conformité d'un dossier de transfert à l'étranger s'appuie sur les informations présentes sur l'ensemble de ces documents. On y retrouve :

— La date du transfert
— L'identité du donneur d'ordre
— L'activité du donneur d'ordre
— Le pays du donneur d'ordre
— L'identité du bénéficiaire de l'ordre
— L'activité du bénéficiaire de l'ordre
— Le pays de résidence du bénéficiaire
— L'objet de la transaction
— Le montant de la transaction
— La devise dans laquelle la transaction est faite (Euros, Dollars, CAD,...)
— etc.

2.1.2 Pays étranger

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

2.1.3 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 habituelles à l'étranger.

2.2 Etude de l'existant

2.2.1 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 règlementtion autorise l'opération qui est entrain d'être menées entre les diffrents acteurs.

2.2.2 Plateformes existantes dans le milieu bancaire

De nombreuse plateformes permettent de juger le risque AML d'un acteur d'une opération de transfert.

ComplianceBond

ComplianceBond est une plateforme web qui aide à rester à jour au niveau des exigences réglementaires et des normes en constante évolution afin de renforcer la conformité à l'échelle de l'organisation. Elle permet de réduire le temps passer à documenter et à tester la conformité. Elle automatise les tests de surveillance de conformité en temps réel. Elle est 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 toute institution traitant les transactions de paiement électronique.

ELCA

La plateforme ELCA de détection de fraudes est un système novateur d'aide à la décision. S'appuyant sur l'analyse de données, elle permet aux entreprises de transport de définir une stratégie claire pour planifier et gérer les missions de contrôle, réduisant ainsi les fraudes.

kdprevent

La plateforme kdprevent permet de lutter contre le blanchiment d'argent et le financement du terrorisme. Elle a été mise en œuvre 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 fonctionalités 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.

2.3 Les différentes opérations à l'étranger

Plusieurs types d'opérations à l'étranger sont effectuées par les services de la banque.

2.3.1 Les transferts émis

Il s'agit d'une opération émise par la banque résidente en l'occurence la SGBF à destination d'une autre banque présente dans un autre pays. Il peut s'agir d'une simple correspondance dans laquelle le client ordonne le transfert. On distingue quatres 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'occurence dans notre cas la SGBF.

— La banque du bénéficiaire de l'ordre

— Le Bénéficiaire de l'ordre

2.3.2 Les transfert reçu

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

2.3.3 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

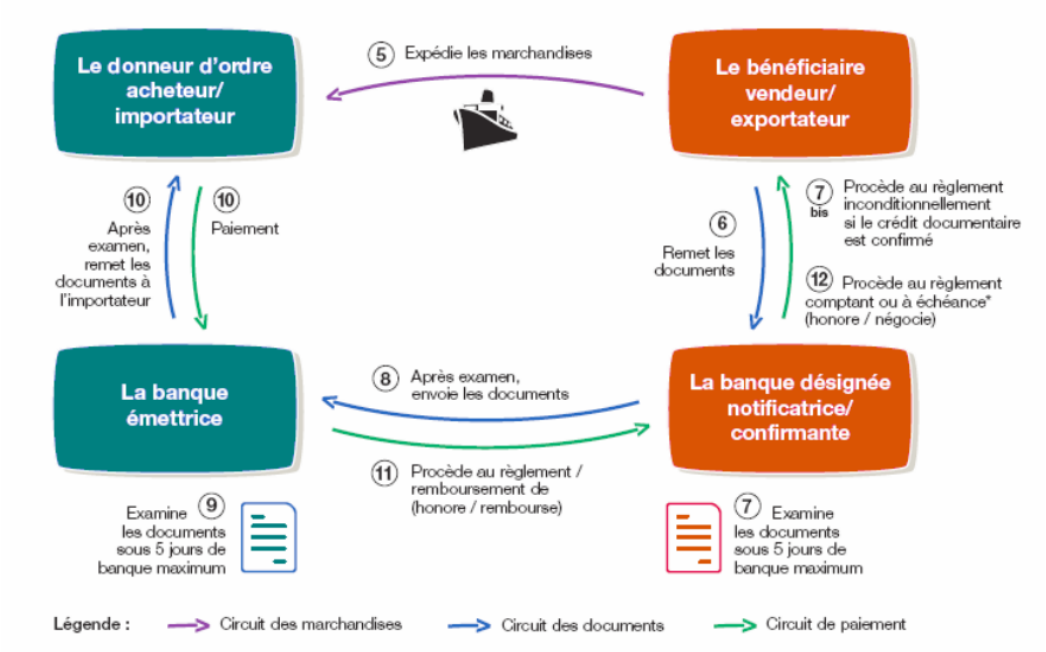


FIGURE 2.1 – Fig1.

2.3.4 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)

2.3.5 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.

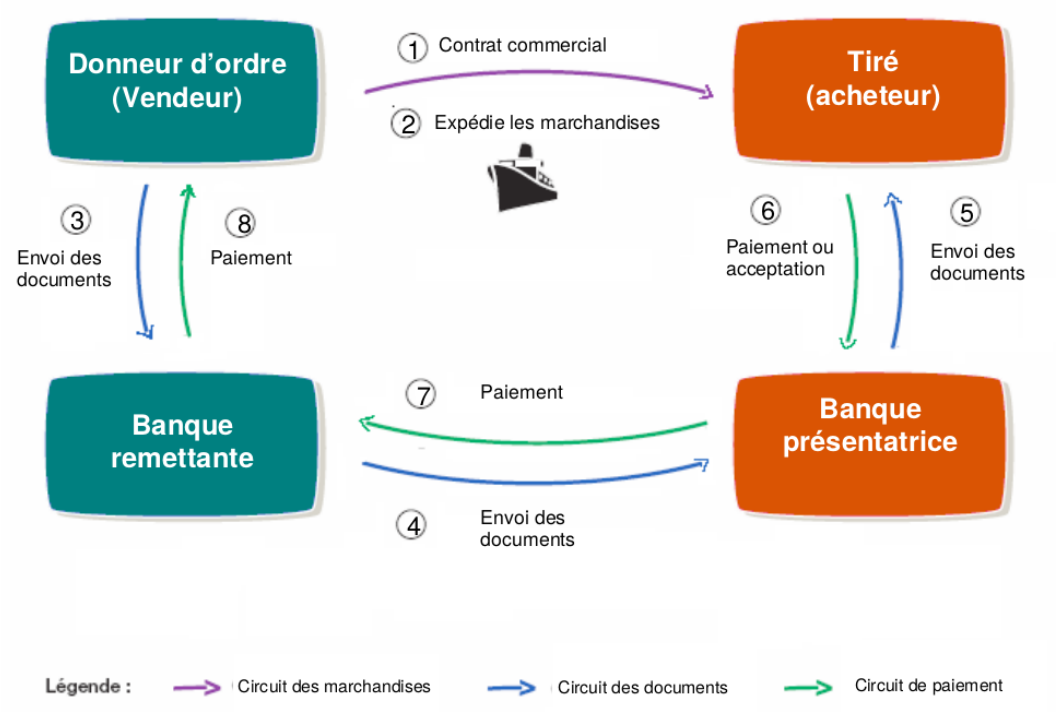


FIGURE 2.2 – Fig1.

2.4 La conformité ou compliance

Le cadre règlementaire autour des activités financières a été fortement renforcé, faisant de la Conformité (ou Compliance) un pilier indispensable de la protection des banques et de leurs clients.

2.4.1 Définition

La conformité en anglais compliance est un concept qui a fait naître de nouvelles obligations pour le banquier. Elle se définit comme l'obligation de veiller à ce que les collaborateurs des différentes banques s'assurent en permanence que soient respectées :

— Les dispositions législatives et règlementaires 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 cette entité économique pour qu'elle se mette en conformité avec les normes législatives et règlementaire.

En somme, la conformité est l'ensemble des actions visant à l'intégration, dans la structure bancaire des exigences issues des règlementations 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.

2.4.2 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.

2.5 Présentation du cadre règlementaire

La société générale est un groupe international français. Il est soumis aussi bien aux règles de la zone UEMOA que celles européennes. Les transferts à l'étranger sont soumis à un ensemble de textes.

2.5.1 Les textes généraux régissant les transfert à l'étranger

Un ensemble de textes structure, encadre l'activité bancaire en ce qui concerne les opérations à l'étranger. De ces textes, nous pouvons dires que les opérations à l'étranger sont soumis à trois règlementations

1. Celle relative à la règlementation de change
2. Celle concernant les sanctions internationales(lutte anti-blanchiment et lutte contre le financement du terrorisme, sanction embargo. . .).
3. Et celle sur les importations et les exportations des marchandises.

2.5.2 La règlementation de change

La règlementation 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 règlementation 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 règlementations 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 règlementation de change sur les opérations d'investissement et d'emprunt

La règlementation 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.

2.5.3 La règlementation sur les importations et les exportations

Il s'agit d'un outil juridique permettant de fixer les règles et procédures en matière d'importation et d'exportation des marchandises. Cet outil permet de définir également la liste des produits soumis à ASI (Autorisation Spéciale d'Importation) et à ASE (Autorisation Spéciale d'Exportation).

Une importation est une entrée dans un pays de biens ou services provenant d'un autre pays. L'exportation est l'action de vendre à l'étranger une partie de la production de biens ou de services d'un ensemble économique, pays ou région.

2.5.4 La règlementation sur la lutte contre le blanchiment de capitaux et le terrorisme

Le blanchiement 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 Blanchiement permet notamment aux criminels de masquer une augmentation trop ostensible de leur richesse afin d'éviter d'attirer l'attention des autorités.

Le financement du terrorisme consiste à fournir ou réunir des fonds, des biens ou des services susceptibles d'être utilisés dans le but de facilité ou de perpétrer des actes de terrorisme. Ces opérations à finalité criminelle impliquent parfois des fonds d'origine parfaitement légale.

Il est du devoir des établissements financiers comme les banques de lutter efficacement contre ces pratiques. La règlementation dans ce domaine oblige les banques à :

— mettre en oeuvre une organisation approprié

— Assurer une vigilance constante à l'égard des clients et de leurs opérations.

— Déclarer les opérations suspectes à la CRF(Cellule de Renseignement Financiers)

— Conserver les données des clients

L'application des règles LCB-FT oblige les collaborateurs de la SGBF à

— Appliquer impérativement la règlementation française

— se conformer à la règlementation 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 règlementation américaine pour toute transaction vers les Etats-Unis ou impliquant le dollar Américain.

Le non-respect de ces règles entraine de-facto l'application de sanction à l'encontre de la banque. Ces sanctions peuvent être pénales, discipliaires, ou financières.

Conclusion

Cette partie nous a permis de présenter la conformité dans le domaine bancaire, de présenter l'ensemble des rglements qui régissent l'activité des banques dans le domaine des opérations à l'étranger.

CHAPITRE 3

APPRENTISSAGE AUTOMATIQUE ET AIDE À LA DÉCISION

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 apprenent des exemples. Le machine Learning est un ensemble de méthodes qui permettent aux ordinateurs d'apprendre à partir des données qui leurs 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.

3.1 Vocabulaire du machine learning

3.1.1 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.

3.1.2 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 :

x1, x2, . . . , x3

3.1.3 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.

3.1.4 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".

3.2 Les étapes d'un projet de Machine learning

Pour mener à bien un projet de Machine Learning, une méthodologie particulière doit être suivi :

3.2.1 Définition et compréhension du problème

La définition du contexte et du problème est indispensable à la compréhension de la signification des données. Cette étape est un préalable à la collecte des données.

3.2.2 Collecte des données et le prétraitement

Les données constituent la matière première d'un projet de machine learning. Il s'agit au cours de cette étape d'extraire les données de leur environnement d'origine. Ces données seront par la suite épurées afin d'extraire ou de nettoyer les données incohérentes. Il s'agit là de la phase de prétraitement.

3.2.3 La modélisation

Elle consiste en la construction d'un modèle de prédiction.Modéliser en machine learning signifie représenter le comportement d'un phénomène afin de pouvoir aider à la résolution d'un problème concret.Généralement, l'implémentation se base sur plusieurs techniques(réseaux de neurones, arbre de décision, clustering), puis on choisit le bon résultat.

3.2.4 Interprétation du modèle et Evaluation de l'algorithme

Il est nécessaire d'évaluer l'algorithme d'apprentissage choisi sur son jeu de données, en évitant au mieux le biais de sur-apprentissage. Une évaluation rigoureuse des performances d'un algorithme est une étape indispensable à son déploiement.

3.3 Les objectifs et méthodes du Machine learning

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

3.3.1 Les objectifs du machine learning

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 ?

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.

3.3.2 Les méthodes d'aprentissage statistiques

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 «enseignées» pour trouver des erreurs et modifier le modèle en conséquence. L'apprentissage supervisé utilise pour son entrainer son modèle des exemples étiquettés.

L'apprentissage non-supervisé

L'apprentissage non supervisé consiste à apprendre à classer sans supervision ; les exemples fournis sont non-étiquettés. 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.

3.4 Les différents type de classifieurs

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

3.4.1 Méthode des K plus proche voisins (KNN)

La méthode des 'K plus proche voisins' ou **k-Nearest Neighbors KNN** en anglaisest une méthode non paramétrique 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 comme c'est le cas pour la régression logistique ou la régression linéaire. C'est pourquoi cet algorithme est qualifié comme paresseux (Lazy Learning) car il n'apprend rien pendant la phase d'entrainement.

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 [Ben] 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.

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 éléctronique...), la distance euclidienne est un bon candidat. Quant à la distance de Manhattan, elle est une bonne mesure quand les données (input variables) 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.

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.

3.4.2 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 premiere 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 derniere couche, quant à elle, produit les résultats du système.

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 (boite noire).

— L'échantillon nécessaire à l'apprentissage doit être suffisamment grand et représentatif des sorties attendues.

3.4.3 Support Vector Machine (SVM)

Les SVM constituent une technique d'apprentissage supervisée introduite en fin des années 90. Grâce à ses performances, cette technique a ouvert un domaine de recherche très actif et un grand éventail d'applications. Historiquement, ils ont été développés dans les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage : la théorie de Vapnik-Tchervonenkis. L'idée est de rechercher une règle de décision basée sur une séparation par hyperplan de marge optimale. Son principe est de rechercher le meilleur hyperplan qui sépare linéairement deux classes, tout en les repoussant aux 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.

Principe des SVMs

Les SVMs peuvent être utilisés pour résoudre des problèmes de classification. La résolution d'un tel problème passe par la construction d'une fonction h définit par :

$$y = h(x_i) =< w, x_i > +b = \sum_{j=1}^{p} w_j.x_i^j + b$$

Où w est le vecteur poids et b le biais. Pendant son entrainement , le svm calculera un hyperplan vectoriel d'équation $w_1.x_1 + w_2.x_2 + \cdots + w_n.x_n = 0$ ainsi qu'un scalaire b.

Une fois l'entrainement terminé, pour classer un nouvelle exemple x, le SVM regardera le signe de la fonction :

$$h(x) = w_1a_1 + w_2a_2 \cdots + w_na_n$$

Si $h(x)$ est positif ou nul, alors x est d'un côté de l'hyperplan et appartient à la première classe de notre dataset. Sinon x appartient à l'autre classe.

En résumé, trouver la classe d'un nouvel exemple revient à réaliser la classification suivante.

$$\begin{cases} < w, x_i > +b \geq 1 \text{ si } y_i = 1 \\ < w, x_i > +b \leq -1 \text{ si } y_i = -1 \end{cases}$$

Avantage

— Les SVM possèdent des fondements mathématiques solides.

— Les exemples de test sont comparés juste avec les supports vecteur et non pas avec tout les exemples d'apprentissage.

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

Inconvénient

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.

3.4.4 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 fixés. Il permet d'émettre des prédictions sur le problème par réduction niveau après niveau du domaine.

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 a 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 brruit.

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 : D es 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).

3.5 Les arbres de décisions

Un arbre de décision est un arbre (orienté) dont les nœuds représentent un choix sur un attribut, les arcs représentent les possibilités pour l'attribut testé et les feuilles représentent les décisions en fonction des compositions d'attributs. Leur fonctionnement repose sur des heuristiques qui, tout en satisfaisant l'intuition, donnent des résultats remarquables. Leur structure arborescente les rend également lisibles par un être humain, contrairement à d'autres approches où le prédicteurs construit est une boite noire.

Un arbre de décision modélise une hiérarchie de tests sur les valeurs d'un ensemble d'attributs. À l'issue de ces tests, le prédicteur produit une valeur numérique ou choisit un élément dans un ensemble discret de conclusions. On parle de régression dans le premier cas et de classification dans le second.

3.5.1 Théorie de l'information

Les théories de Shannon sont à la base de l'algorithme ID3 et *C4.5*. L'entropie de Shannon est la plus connue et la plus appliqué. Il définit d'abord la quantité d'informations fournies par un événement : plus la probabilité d'un événement est faible, plus il fournit d'informations, plus il est important.

Entropie de Shannon

L'entropie de Shannnon correspond à la quantité d'information contenue dans une source d'information. C'est également la longueur minimale nécessaire pour coder la classe d'un membre pris au hasard dans le set d'exemple S. Sa formule est la suivante :

$$Entropy(S) = \sum_{c \in classes(S)} -p_c * log_2(p_c)$$

Tel que p_c est la proportion d'exemples de S ayant pour classes résultante (étiquette) c. Voici à quoi ressemble la fonction d'entropie pour un ensemble à deux classes possibles(par exemple Oui/Non)

Gain d'information(Information Gain G(p,T))

La fonction dpermettant de sélectionner le test qui doit étiqueter le noeud courat est la fonction *Gain*. Le gain est défini par un set d'exemples et par un attribut. Sa formule est la suivante :

$$Gain(S, A) = Entropy(S) - \sum_{v \in valeur(A)} \frac{S_v}{S} * Entropy(S_v)$$

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.

3.5.2 Les principaux algorithmes de construction d'arbres de décisions

Un arbre de décision peut être construits à partir d'un ensemble d'observations. Ainsi à partir d'un ensemble d'observations $T = (x, y)$, on souhaite construire un arbre de décision prédisant l'attribut y en fonction de nouvelles instances x. Il existe essentiellement deux 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 »

ID3

ID3 (Iterative Dichotomiser 3) a été developpé par Ross Quinlan en 1986. Il se base qur 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. Le critère utilisé par Quinlan pour trouver l'attribut à tester est l'entropie de shannon.

L'algorithme crée un arbre multivoie, trouvant pour chaque nœud (i-e 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. Le pseudo-code de ID3 est le suivant

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 éléments en plus.

— 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 valeur continues

La nouvelle fonction de gain devient :

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)}$$

tel que

$$SplitInfo(p, test) = -\sum_{j=1}^{n} P'(\frac{j}{p}) * log_2(P'(\frac{j}{p}))$$

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.

CART

CART (Classification and Regression Trees) 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).

Dans 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.

L'impureté de Gini d'un nœud est la probabilité qu'un échantillon choisi au hasard dans un noeud soit mal étiqueté s'il était étiqueté par la distribution des échantillons dans le nœud. Sa formule est la suivante

$$I_G = 1 - \sum_{i=1}^{n} (p_i)^2$$

Les différences entre CART et C4.5 sont :

— Les tests CART sont toujours binaire, mais C4.5 permet plusieurs résultats.

— L'algorithme CART utilise l'indice de Gini pour classer les données alors que C4.5 utilise des critères basés sur les informations.

— CART recherche des tests alternatifs qui se rapprochent des résultats lorsque l'attribut testé a une valeur inconnue, mais C4.5 calcule la probabilité des différentes sorties et choisit la meilleure.

Des différents algorithmes présentés plus haut, l'algorithme le mieux adaptés à notre contexte est C4.5.

L'« overfitting »dans les arbres de décisions se produit lorsque nous disposons d'un modèle très flexible qui mémorise essentiellement les données de formation en les ajustant étroitement. Le problème est que le modèle apprend non seulement les relations réelles dans les données d'entraînement, mais aussi tout bruit qui est présent. Un modèle rigide est dit avoir un biais élevé parce qu'il fait des hypothèses sur les données de formation. Par exemple, un classificateur linéaire fait l'hypothèse que les données sont linéaires et n'a pas la flexibilité nécessaire pour s'adapter à des relations non linéaires. Un modèle rigide peut ne pas avoir la capacité de s'adapter même aux données de formation et dans les deux cas - variance et biais élevés - le modèle n'est pas capable de bien généraliser aux nouvelles données.

La raison pour laquelle l'arbre de décision est enclin à se surajuster lorsque nous ne limitons pas la profondeur maximale est qu'il a une flexibilité illimitée, ce qui signifie qu'il peut continuer à croître jusqu'à ce qu'il ait exactement un nœud de feuille pour chaque observation, les classant toutes parfaitement.

Comme alternative à la limitation de la profondeur de l'arbre, qui réduit la variance (bonne) et augmente le biais (mauvaise), nous pouvons combiner plusieurs arbres de décision en un seul modèle d'ensemble connu sous le nom de forêt aléatoire.

3.6 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'échantillonage aléatoire des données d'entrainement 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.

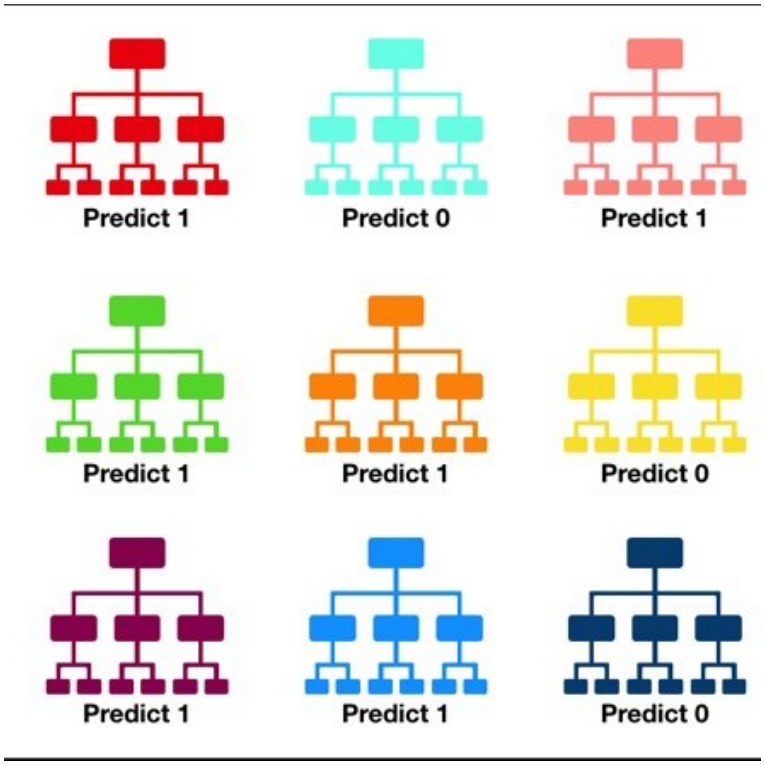


FIGURE 3.1 – Forêt aléatoire.

3.6.1 Echantillonage aléatoire des données d'entrainement

Lors de la phase d'entrainement, 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 sus le nom de *bagging* abbréviation de *bootstrap aggregating*

3.6.2 Fractionnement des noeuds

Le deuxième concept principal de la for 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 raçine carré du nombre de caractéristiques pour une classification. Ainsi si nous avons 25 caractéristiques seulement 5 seront pris aléatoirment pour diviser le noeud.

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 prouver qu'il est possible de mettre en place un système permettant de donner l'état d'un dossier à la réception de celui-ci.

Pour cela, nous considérons un dossier comme l'ensemble des informations des intervenants et de l'opération elle-même. Ces informations sont relevées directement sur le dossier physique. L'idée est de prendre le dossier et de récupérer tous les éléments d'analyse qui sont présents sur le documents. Notre dataset initiale nous a été extrait des données du CBS. Il comportait initialement 10000 lignes.

Nous allons nettoyer nos données en écartant les anomalies, tout en restant vigilant sur ce que contiennent les données mises de côté. La notion d'anomalie dans les transferts à l'étranger est très délicate. En effet, ce n'est parce que M. X a toujours efectué des transferts de 1000 Euros vers un pays A que demain il n'enverra pas 5 Euros vers le même pays pour le même motif.

La détection d'anomalie dans notre cas n'a pas fourni de résultats satisfaisants (ils éliminent une énorme partie des données d'entrée , qui ne sont pas forcément des anomalie)

Après l'analyse de la conformité bancaire et de son application sur les transferts à l'étranger, nous avons donc du changer de méthode. L'étude de notre sujet et des contraintes qui nous sont

imposées nous permet de faire un apprentissage supervisé sur nos données.

Pour une meilleure prédicion, nous posons comme hypothèse de départ que tous dossiers qui passent par notre modèle est considérés comme complet. Cette hypothèse nous permet de nous focaliser sur les autres aspects de la conformité.

4.2 Réalisation

4.2.1 Le jeu de données

Acquisition des données

Les données initiales fournies par la SGBF sont constituées d'un fichier de 10000 lignes au format tableur excel. Chaque ligne de ce fichier représente un dossier de transfert. Ces données extraites du CBS contiennent exclusivement les informations permettant d'effectuer l'opération de transfert i-e les coordonnées du donneurs, ceux du bénéficiaires, le montant de la transaction et la dévise dans laquelle l'opération est menée. Les caractéristiques de nos datas à cette étape sont :

— Le nom du donneur d'ordre

— L'activité du donneur d'ordre

— Le pays du donneur d'ordre

— Le nom du beneficiaire de l'ordre

— le pays du bénéficiaire de l'ordre

— Le pays de la banque du bénéficiaire

— L'objet de l'opération

— Le montant de l'opération

— La dévise de l'opération

Les informations qui nous ont été fournies ne permettaient malheureusement pas à elles seule de juger de la conformité d'une opération de transfert de fonds prise à part. Il y manque certaines informations telles l'activité du bénéficiaire, la banque du bénéficiaire et les résultats des multiples contrôles de sanctions. Pour pouvoir mener à bien l'analyse d'un dossier, ces éléments devront être ajoutés aux caractéristiques de notre dataset. Ce sont :

— Le résultat des contrôles sur l'identité du donneur

— Le pays du donneur d'ordre

— Le résultat des contrôle sur le pays du donneur

— Le résultat des contrôle sur l'identité du bénéficiaire

— L'activité du bénéficiaire de l'ordre

— le pays du bénéficiaire de l'ordre

— Le résultat du contrôle du pays du bénéficiaire

— La banque du bénéficiare

— Le résultat des contrôle sur la banque du bénéficiaire

Mis à part les contrôles qui doivent être effectués sur des plateformes utilisées par l'ensemble des filiales du groupe(force-online par exemple), les autres informations sont pour la plupart disponible directement sur le dossier physique.

En conséquence, les caractéristiques de notre dataset seront :

— Le nom du donneur d'ordre

— Le résultat des contrôles sur l'identité du donneur

— L'activité du donneur d'ordre

— Le pays du donneur d'ordre

— Le résultat des contrôles sur le pays du donneur

— Le nom du bénéficiaire de l'ordre

— Le résultat des contrôles sur l'identité du bénéficiaire

— L'activité du bénéficiaire de l'ordre

— le pays du bénéficiaire de l'ordre

— Le résultat des contrôles sur le pays du bénéficiaire

— La banque du bénéficiare

— Le resultat des contrôles sur la banque du bénéficaire

— Le pays de la banque du bénficiaire

— L'objet de l'opération

— Le montant de l'opération en dévise

— La dévise de l'opération

Notons que la SGBF ne souhaitant pas être mêlée à une opération non-conforme, la banque ne dispose pas de base de donnée d'opération non-conforme. Toute les données issus du CBS sont des données de dossiers conformes. Pour constitué un set de données de dossier non-conforme, les collaborateurs du service OPI et conformité nous ont aidé à constituer un dataset des dossiers non-conformes.

4.2.2 Prétraitrement des données

Les caractéristiques recencées ci-dessus nous permettent de juger de la conformité d'un dossier de transferts. Certaines caractéristiques sont très distinctives et pourraient entrainées un sur-apprentissage sur nos données. Nous choisissons de les supprimer de notre dataframe. Il s'agit de l'identité du donneur d'ordre et de l'identité du bénéficiaire, du pays du donneur et de celui du bénéficiaire.

Les caractéristiques en entrée de notre algorithme seront donc :

♢ le résultat des contrôles sur l'émetteur

- ♢ le résultat des contrôles sur le bénéficiaire
- ♢ le résultat des contrôles sur la banque du bénéficiare
- ♢ le résultat des contrôles sur le pays du beneficaire
- ♢ le résultat des contrôles sur le pays du donneur
- ♢ l'activite du donneur,
- ♢ l'activite du bénéficiaire,
- ♢ l'objet de transaction,
- ♢ le montant de l'opération,
- ♢ la devise de l'opération

Notre jeu de données final est constitué de huit cent cinquante(850) dossiers.

4.2.3 Les outils

Nous allons présenter les outils qui nous ont permis de mettreen place l'outil que nous avons proposé.

Le langage python

Le langage python a été utilisé pour les codes d'implémentation de notre modèles. De nombreuses bibliothques pythons ont été également utilisé.

Le langage Javascript

A travers le framework angular, il nous a permis de réaliser une interface conviviale qui permettrait l'utilisation de notre modèle.

Flask

Il s'agit d'un serveur web python qui nous a permir de mettre à la disposition de notre client javascript le modèle python qui a été implémenté.

4.3 Résultats

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. La validation croisée va nous permettre d'utiliser l'intégralité de notre jeu de données pour l'entraînement et pour la validation.

Pratiquement, On découpe 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 des k-1 autres parties) est utilisé pour l'entraînement.

Les mesures détaillés des test par categorie sont représentéssur les figures suivantes. **1** pour dossier conforme **0** pour dossier non conforme.

```
Accuracy score for fold 0: 91.3043%
[[55  3]
 [ 3  8]]
              precision    recall  f1-score   support

           0       0.95      0.95      0.95        58
           1       0.73      0.73      0.73        11

    accuracy                           0.91        69
   macro avg       0.84      0.84      0.84        69
weighted avg       0.91      0.91      0.91        69
```

```
Accuracy score for fold 1: 82.6087%
[[51  6]
 [ 6  6]]
              precision    recall  f1-score   support

           0       0.89      0.89      0.89        57
           1       0.50      0.50      0.50        12

    accuracy                           0.83        69
   macro avg       0.70      0.70      0.70        69
weighted avg       0.83      0.83      0.83        69
```

```
Accuracy score for fold 2: 91.3043%
[[55  4]
 [ 2  8]]
              precision    recall  f1-score   support

           0       0.96      0.93      0.95        59
           1       0.67      0.80      0.73        10

    accuracy                           0.91        69
   macro avg       0.82      0.87      0.84        69
weighted avg       0.92      0.91      0.92        69
```

```
Accuracy score for fold 3: 94.1176%
[[59  2]
 [ 2  5]]
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        61
           1       0.71      0.71      0.71         7

    accuracy                           0.94        68
   macro avg       0.84      0.84      0.84        68
weighted avg       0.94      0.94      0.94        68
```

```
Accuracy score for fold 4: 92.6471%
[[57  0]
 [ 5  6]]
              precision    recall  f1-score   support

           0       0.92      1.00      0.96        57
           1       1.00      0.55      0.71        11

    accuracy                           0.93        68
   macro avg       0.96      0.77      0.83        68
weighted avg       0.93      0.93      0.92        68
```

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 est globale est de 74%.

Les arbres de décisions étant considérés comme des classifieurs faibles, nous avons apliqué sur nos données les Random Forest Classifier. Les résultats obtenues sont présentés dans la figure ci-dessous.

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

4.4 Interprétation des résultats

4.4.1 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.

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

Le rappel : En anglais *recall*, il donne la propertion 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.

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

4.4.2 Limites et difficultés

4.4.3 Perspectives

CONCLUSION

C'était très chouette !

PUBLICATIONS PERSONNELLES

[Ben] Younes Benzaki, « Data-scientist : Du rêve à la réalité », .

[Cha20] Yannis Chaouche, « Initiez-vous au machine learning ». https ://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning, Mis à jour le 13 février 2020, consulté le 19 février 2020.

[Mel20] Melepe, « Un peu de machine learning avec les svm ». https ://zestedesavoir.com/tutoriels/1760/un-peu-de-machine-learning-avec-les-svm, Mis à jour le 15 février 2020, consulté le 19 février 2020.

[TT00a] Toto et Titi, « Efficient thesis writing », *in somebook*, p. 1–10, 2000.

[TT00b] Tutu et Tata, « Latex rox », *in Extra book*, vol. 42 *in Lectures*, p. 1–10, 2000.

BIBLIOGRAPHIE

Résumé

Résumé en français...

Abstract

Abstract in english...

Numéro d'ordre : 00000