République Algérienne Démocratique et Populaire

Ministère de l'enseignement Supérieur et de la

Recherche Scientifique

Université Saad Dahleb Blida -1-

Faculté des Sciences

Département d'Informatique

MEMOIRE DE FIN D'ETUDES EN VUE DE

L'OBTENTION DU DIPLÔME DE MASTER EN

INFORMATIQUE

Options : Ingénierie des logiciels +
Sécurité des systèmes d'information

Technique d'apprentissage et de prédiction intégrant Blockchain dans la gestion du risque crédit

Réalisé par :

- Ouacif Oualid Ali
- Temzi Riyadh

Promotrice :

- Mme. Zahra Fatma Zohra

Encadreur :

- Mr. Aitoufroukh Mohand

Résumé

Les banques étant tenues de respecter strictement la réglementation financière de leur pays, la gestion du risque de crédit est confrontée à de nombreux défis. Il s'agit notamment d'une gestion inadéquate de données, d'un manque de cadre de modélisation des risques à l'échelle du groupe, d'un travail constant sur le ratio d'efficacité, d'outils de risque insuffisants et de rapports inexacts.

Les institutions financières cherchent à trouver une solution sécurisée pour automatiser les processus dans l'ensemble de l'entreprise afin d'aider leurs analystes à prendre des décisions correctes sur la base des prédictions fiables. Pour cela, les Blockchains et les techniques de Machine Learning peuvent offrir un large éventail d'opportunités et de solutions dans différents aspects de la gestion du risque de crédit**.**

Sur la base des résultats empiriques, l'étude tire plusieurs propositions normatives sur la manière d'assembler une base de données de prédiction de faillite ainsi que sur la sélection de la/les méthodes de classification appropriées pour réaliser une prédiction efficace de la faillite d'entreprise.

Mots-clés : **A**pprentissage automatique, Classification, gestion du risque crédit, prédiction des faillites, Blockchain.

Abstract

As banks are required to strictly adhere to their country's financial regulations, managing credit risk faces many challenges. These include inadequate data management, a lack of a group-wide risk modeling framework, constant work on the effectiveness ratio, insufficient risk tools and reporting inaccurate.

Financial institutions are looking to find a secure solution to automate processes across the business to help their analysts make correct decisions based on reliable predictions. For this, Blockchains and Machine Learning techniques can offer a wide range of opportunities and solutions in different aspects of credit risk management.

Based on the empirical results, the study draws several normative proposals on how to assemble a bankruptcy prediction database as well as on the selection of the appropriate classification method (s) to achieve effective bankruptcy prediction.

Keywords : Machine Learning, credit risk, bankruptcy prediction, classification, Blockchain

ملخص

نظرًا لأن البنوك مطالبة بالالتزام الصارم باللوائح المالية لبلدها ، فإن إدارة مخاطر الائتمان تواجه العديد من التحديات. وتشمل هذه عدم كفاية إدارة البيانات ، وعدم وجود إطار عمل لنمذجة المخاطر على مستوى المجموعة ، والعمل المستمر على نسبة الكفاءة ، وأدوات المخاطر غير الكافية والتقارير غير الدقيقة.

تحتاج المؤسسات المالية إلى إيجاد حل آمن لأتمتة العمليات عبر المؤسسة. لهذا ، يمكن أن يوفر التعلم الآلي و تقنية سلسلة البلوك مجموعة واسعة من الفرص والتحديات في جوانب مختلفة من مخاطر الائتمان.

بناءً على النتائج التجريبية ، ترسم الدراسة العديد من المقترحات المعيارية حول كيفية تجميع قاعدة بيانات التنبؤ بالإفلاس واختيار طريقة (طرق) التصنيف المناسبة لتحقيق التنبؤ الفعال بفشل الأعمال.

الكلمات المفتاحية :التعلم الآلي ,تقنية سلسلة البلوك,إدارة مخاطر الائتمان,التنبؤ بالإفلاس,التصنيف

Remerciement

Nous tenons à remercier.

En premier lieu le bon dieu de nous avoir donné la force et

le courage pour réaliser à terme ce travail.

Nous remercierons notre promotrice Mme Zahra Fatma Zohra

pour nous avoir aidé dans cette étude, nous lui sont très

reconnaissants pour ses remarques et conseils

Table de Matière

Liste des Figures

Liste des Tableaux

Introduction générale

La banque exerce son activité dans un domaine en constante évolution. Ce qui lui permet de disposer d'une base de données riche et étendue. Afin de mieux exploiter la donnée, un prétraitement est nécessaire. Il permettra d'extraire les connaissances pour en tirer profit, Bien que les banques ont employées des outils d'analyse statistiques avec un peu de succès pendant plusieurs années, les modèles précédemment invisibles des comportements des clients deviennent maintenant plus clair à l'aide des nouveaux outils d'apprentissage.

Société Générale Algérie, représentant typique des grandes banques nationales, travaille actuellement à la mise à niveau de son système de gestion des risques pour réussir dans une nouvelle direction stratégique.

En effet, l'ouverture de la banque au financement de tous les secteurs rend nécessaire l'adaptation des méthodes et des outils de gestion du risque de crédit dans le cadre d'une politique de risque globale claire et en accord avec la stratégie globale de la banque.

La procédure actuel de l'étude d'un dossier client demande un travail manuel sur une quantité d'informations volumineuse (grand nombre de ratios et d'indicateurs financiers) est nécessairement de longue durée, ce qui peut engendrer un frein dans l'atteinte des objectifs annuelles, et pour aborder cette situation un modèle d'apprentissage automatique pourra intervenir en tant qu'outil d'aide à la décision, en mettons en lumière les ratios à considérer selon le profil du client demandeur pour attirer l'attention de l'analyste risque sur les métriques les plus importants du cas actuel pour lui faire gagner du temps et lui faciliter sa prise de décision.

Dans ce travail, on s'est focalisé sur les grands clients Corporate. Compte tenu des problèmes liés à la politique de confidentialité de la banque, on a proposé de travailler avec des Datasets gratuits ayant la même structure de données que le dataset prévu (ratios et indicateurs financiersannuels des clients), afin d'évaluer les résultats sur des données externes et les négocier avec l'analyste risque pour après refaire les mêmes étapes directement sur les données de la banque.

Concernant les normes de sécurité, la BlockChain sera aussi appliquée en se basant sur la cryptographie qui permet d'enregistrer des échanges transactionnels, Le principe est que toutes les informations sont stockées dans des fichiers qui plutôt que d'être centralisés dans la

même base de données, ils sont distribués auprès des utilisateurs du réseau d'échange. Ainsi, il existe autant de copies que d'utilisateur sur le réseau (pear to pear network), ce qui rend l'information non seulement transparente mais ultra sécurisée. Pour mettre fin à toute difficulté de synchronisation, les ordinateurs doivent résoudre un problème mathématique qui se complexifie en fonction de la charge de la Blockchain (plus elle est active plus les problèmes sont complexes), le premier ordinateur qui résout son problème peut valider son fichier, et la boucle recommence. Ce principe est appelé la preuve de travail ou Proof of Work.

Organisation du mémoire

Hormis l'introduction le mémoire se reparti en quatre chapitres :

Chapitre 1 : « Machine Learning »

Ce chapitre est consacré à la présentation de l'apprentissage automatique (Machine Learning) avec ses différents types et modèles.

Chapitre 2 : « Blockchain »

Le deuxième chapitre présente le Blockchain, ses types ainsi que la sécurité des transactions en se basant sur la cryptographie

Chapitre 3 : « Prédiction de la faillite d'entreprise (Solvabilité des clients)»

Dans le troisième chapitre, nous avons présenté l'entreprise et défini les besoins afin de procéder à la solution

Chapitre 4 : « Expérimentation »

Ce chapitre, sera réservé pour exposer nos solutions proposées.

Conclusion générale :

Un bilan récapitulatif du travail fait durant ce mémoire, ainsi que des axes d'amélioration

Chapitre 1 : Machine Learning

1.1 Introduction

Le contexte mondial a connu plusieurs bouleversements durant les deux dernières décennies dues aux révolutions technologiques, ce qui a déclenché une croissance sans précédent du volume des données présent au sein des organisations, ceci a rendu la prise de décision de plus en plus complexe, d'où la nécessité du Machine Learning (ML) et de la business intelligence, ce qui va permettre aux décideurs de prendre les décisions adéquates 'Data Driven Décision' pour conquérir de nouveaux marchés et assurer une performance durable.Dans ce chapitre, nous définirons l'intelligence artificielle et l'apprentissage automatique, puis nous présenterons les réseaux de neurones, ainsi que les mesures de performances.

1.2 Intelligence artificielle

Intelligence artificielle (IA), c'est la faculté d'un ordinateur numérique ou d'un robot contrôlé par ordinateur d'effectuer des tâches généralement associées à des êtres intelligents. Le terme est couramment appliqué au projet de développement de systèmes dotés des processus intellectuels propres aux êtres humains, tels que la capacité de raisonner, de découvrir un sens, de généraliser ou d'apprendre à partir d'expériences passées.[1]

Depuis le lancement de l'ordinateur numérique dans les années 1940, il a été démontré que les ordinateurs peuvent être programmés pour effectuer des tâches très complexes comme, par exemple, jouer aux échecs avec une grande compétence. Pourtant, malgré les progrès constants de la vitesse de traitement et de la capacité de mémoire des ordinateurs, il n'existe pas encore de programmes capables d'égaler la flexibilité de l'homme dans des domaines plus vastes ou dans des tâches exigeant de grandes connaissances quotidiennes. [2]

D'autre part, certains programmes ont atteint et ont même dépassé les niveaux de performance des experts et des professionnels humains dans l'exécution de certaines tâches spécifiques, de sorte que l'intelligence artificielle dans ce sens limité se retrouve dans des applications aussi diverses que le diagnostic médical, les moteurs de recherche informatiques et la reconnaissance de la voix ou de l'écriture.

1.3 Machine Learning

L'apprentissage automatique (Machine Learning en langue anglaise) est une branche de l'intelligence artificielle et un domaine d'étude qui vise à donner aux ordinateurs la capacité d'apprendre et d'améliorer leurs performances à partir de l'expérience (entraînement) sans être explicitement programmés. À plus grande échelle, l'apprentissage automatique est le processus qui consiste à enseigner aux systèmes informatiques comment faire des prédictions précises lors de la réception de données grâce à une analyse statistique et sans intervention ou assistance humaine.[3]

L'objectif de l'apprentissage automatique est généralement de comprendre la structure des données et d'adapter ces données à des modèles qui peuvent être compris et utilisés par des personnes, en d'autres termes il permet d'apprendre ce qui se fait naturellement chez les humains et se révèle utile lorsque nous avons des tâches ou des problèmes complexes impliquant une grande quantité de données.

L'apprentissage automatique couvre de multiples applications, telles que la reconnaissance d'images, la reconnaissance vocale, le traitement du langage naturel, etc.

Tout comme l'écriture de code ordinaire, nous écrivons un algorithme, la machine exécute l'algorithme sur des données spécifiques, puis elle peut effectuer la même tâche avec de nouvelles données qu'elle n'a jamais vues auparavant. Cependant, au lieu d'écrire manuellement du code à l'aide d'un ensemble d'instructions spécifique, grâce à l'apprentissage automatique, les machines sont entraînées à l'aide de grandes quantités de données et apprennent à effectuer des tâches sans leur dire explicitement comment le faire.[4]

Les ordinateurs doivent être formés pour atteindre leurs objectifs et, au cours du processus d'apprentissage, ils essaieront d'accéder à plus de données sur une période de temps pour créer une logique et l'améliorer.

L'apprentissage automatique peut se faire de plusieurs manières : il peut s'agir d'un apprentissage supervisé, d'un apprentissage non supervisé, d'un apprentissage semi-supervisé ou d'un apprentissage par renforcement.

1.4 Types d'apprentissage

Dans ce qui suit, nous présentrons les types d'apprentissage, mais vue que nous utiliserons que l'apprentissage supervisé, nous parlerons en détails que de ce dernier,avec une bréve présentation des autre types.

1.4.1 Apprentissage supervisé

Dans l'apprentissage supervisé, les données sont étiquetées pour guider la machine vers les tendances exactes qu'elle doit rechercher, par exemple, des images de figures manuscrites marquées pour indiquer à quel numéro elles correspondent, un système d'apprentissage supervisé apprendrait les caractéristiques de chaque chiffre, pour finalement reconnaître les chiffres écrits à la main et faire la distinction entre eux. Nous alimentons le système avec des exemples d'entrées et leurs sorties souhaitées, et l'objectif est d'apprendre une règle générale qui relie les entrées aux sorties.[5]

Le processus d'apprentissage se poursuit jusqu'à ce que le modèle atteigne le niveau de correction souhaité sur les données d'apprentissage. En outre, l'algorithme d'apprentissage peut comparer sa sortie avec la sortie correcte, trouver les erreurs, et les corriger. Nous pouvons représenter les données sous la forme suivante : di = (xi ; yi) avec x l'entrée, y la cible associée, i l'indice de l'observation et (xi ; yi) représente un exemple d'apprentissage. Autrement dit, l'algorithme produit une fonction appelée fonction d'hypothèse, et cette fonction prend l'entrée x, essaie de sortir la valeur estimée de y[6].

En fonction de la nature de la sortie, il existe deux types d'apprentissage supervisé : la régression et la classification.

- Lorsque la valeur de sortie est discrète, on parle de tâche de classification.La fonction d'hypothèse prédit la classe ou la catégorie pour une observation donnée. En général, les modèles de classification prédisent une valeur continue comme la probabilité qu'une entrée donnée appartienne à chaque classe de sortie.

- Inversement, lorsque la variable de sortie est réelle, il s'agit d'une tâche de régression. L'objectif de la fonction d'hypothèse est de faire correspondre des variables d'entrée à des variables de sortie de valeur réelle.

1.4.1.1 Classification

La modélisation prédictive de classification est la tâche d'approximation d'une fonction de mappage (f) des variables d'entrée (X) aux variables de sortie discrètes (y).Les variables de

sortie sont souvent appelées étiquettes ou catégories. La fonction de mappage prédit la classe ou la catégorie pour une observation donnée.Par exemple, un e-mail de texte peut être classé comme appartenant à l'une des deux classes : « spam » et « pas de spam ».

1.4.1.2 Régression

La modélisation prédictive de régression est la tâche d'approximation d'une fonction de mappage (f) des variables d'entrée (X) à une variable de sortie continue (y).Une variable de sortie continue est une valeur réelle, telle qu'une valeur entière ou à virgule flottante. Ce sont souvent des quantités, telles que des quantités et des tailles.Par exemple, on peut prévoir qu'une maison se vendra pour une valeur en dollars spécifique, peut-être entre 100 000 $ et 200 000 $.

1.4.2 Apprentissage non-supervisé

Parfois, les cibles ne sont pas disponibles, et nous avons seulement di = xi.Par ailleurs, il est plus facile d'obtenir des données non étiquetées que des données étiquetées et plus pratique d'éviter l'intervention manuelle ; c'est là qu'intervient l'apprentissage non supervisé.

L'apprentissage non supervisé consiste à former une machine à l'aide de données non étiquetées, de sorte que la machine doit apprendre en utilisant ces données sans aucune orientation. L'apprentissage non supervisé tente d'explorer les données et de trouver une certaine structure à l'intérieur, cela peut être un objectif en soi de découvrir des modèles cachés inconnus.

Le clustering est une forme d'apprentissage non supervisé. Il regroupe les données qui présentent des modèles ou des structures similaires en groupes appelés clusters. Dans un problème de clustering, nous essayons de trouver un algorithme pour regrouper des données non étiquetées en clusters cohérents. L'algorithme K-means est l'algorithme de clustering le plus populaire et le plus utilisé.

Le principe du fonctionement de l'algorithme K means est comme suit :

La première étape consiste à initialiser aléatoirement k points appelés centroïde du cluster. K-means est un algorithme itératif qui répète deux étapes : la première est l'affectation des clusters, puis une étape de déplacement du centroïde. L'étape d'affectation de cluster boucle sur chaque exemple et calcule la distance entre les centroïdes et l'ensemble de données, puis affecte chaque ensemble de données au centroïde le plus proche. L'étape de déplacement du centroïde prend les centroïdes et les déplace vers la moyenne de l'ensemble de données. Ces

deux étapes sont répétées jusqu'à ce que les clusters cessent de changer. L'algorithme K-means attend deux entrées : la valeur de K et le jeu de données. Un bon résultat de clustering dépend principalement de la valeur K[7].

1.4.3 Apprentissage par renforcement

L'apprentissage par renforcement consiste à prendre des mesures appropriées pour maximiser une récompense dans une situation spécifique. L'agent apprend à effectuer une action dans un environnement incertain. Le modèle reçoit des retours d'information sous forme de récompenses ou de punitions lorsqu'il opère dans son espace de problème.

Il existe deux types de renforcements : le renforcement positif est lorsque l'agent est récompensé pour l'encourager à suivre un comportement particulier, et le renforcement est négatif lorsqu'il est entraîné à éviter un comportement négatif pour augmenter la probabilité d'un comportement positif.[5]

L'entrée est l'état initial à partir duquel le modèle démarre et la sortie dépend de l'état de l'entrée actuelle, à chaque pas de temps, l'entrée suivante utilise la sortie de l'étape précédente. La meilleure solution est celle qui fournit la récompense maximale.

1.5 Les données d'apprentissage

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

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

1.6 Méthodologie

Dans cette section nous allons voir deux points qui sont très important dans le processus du machine learning, dans le premier nous allons présenter les différents algorithmes abordés durant notre implémentation qui sont des algorithmes supervisés de classification. En deuxième partie nous allons voir la manière dont les modèles sont évalués, validés, et les caractéristiques particulières à prendre en compte.

1.6.1 Algorithmes d'apprentissage supervisé

Tout au long de ce travail, nous serons amenés à utiliser des algorithmes issus de l'apprentissage supervisé. Il y a diverses méthodes pour des fins différentes.

Dans cette section, nous aborderons en détail les algorithmes de classification (apprentissage supervisé) et leurs formulations mathématiques.

1.6.1.1 DecisionTree (arbre de décision)

Un arbre de décision est un diagramme ou un graphique qui aide à déterminer un plan d'action ou à montrer une probabilité statistique. Le diagramme est appelé arbre de décision en raison de sa ressemblance avec la plante en question. Il se présente généralement sous la forme d'un diagramme vertical ou horizontal qui se ramifie. À partir de la décision elle-même (appelée "nœud"), chaque "branche" de l'arbre de décision représente une décision, un résultat ou une réaction possible. Les branches les plus éloignées de l'arbre représentent les résultats finaux d'une certaine voie de décision et sont appelées les "feuilles".[8]

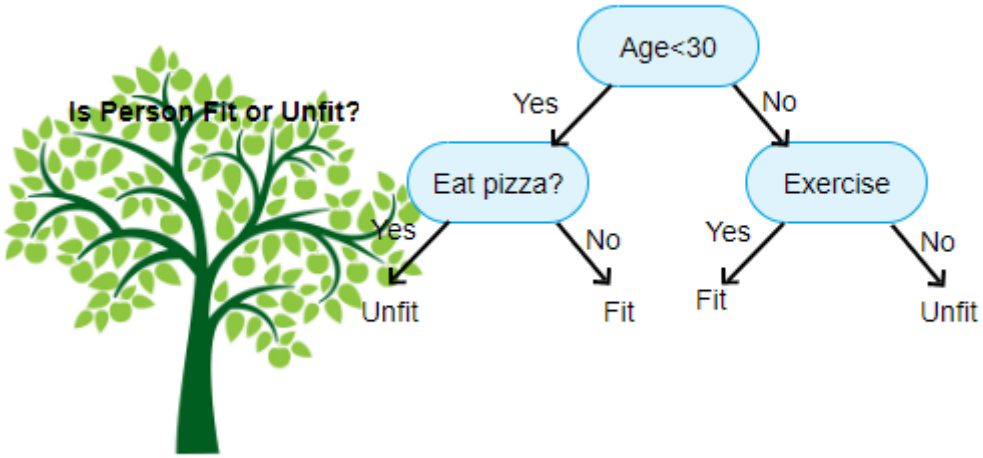

Figure 1: Arbre de décision pour classification fit/unfit.[8]

L'arbre de décision est la mise en œuvre de la stratégie "diviser pour mieux régner" sur un ensemble d'instances indépendantes pour apprendre le problème. L'arbre de décision décisionnel est composé d'une racine, de nœuds de décision internes et de feuilles terminales. Chaque nœud d'un nœud de décision représente un test d'un attribut particulier ou une fonction d'un ou plusieurs attributs dans l'ensemble d'instances à classer. Le résultat du test représente des branches, ainsi chaque branche représente donc la valeur de test que le nœud peut prendre. Ce processus commence à la racine et est répété de manière récursive jusqu'à ce qu'un nœud feuille soit atteint, puis l'instance est classée selon la classe attribuée à la feuille.

En théorie des graphes, un arbre est un graphe non orienté, acyclique et connexe. L'ensemble des nœuds se divise en 3 catégories :

- Nœud racine : l'accès à l'arbre s'effectue par ce nœud
- Nœud interne : les nœuds qui ont des descendants qui sont à leur tour des nœuds
- Nœuds terminaux (feuilles) : nœuds qui n'ont pas de descendant
- Branche : définit le résultat d'un test effectué sur les nœuds internes

Les gens utilisent les arbres de décision pour clarifier, cartographier et trouver une réponse à un problème complexe. Les arbres de décision sont fréquemment utilisés pour déterminer un plan d'action dans les domaines de la finance, de l'investissement ou des affaires. En mathématiques, les arbres de décision sont également appelés diagrammes en arbre.

Le problème à élucider avec les arbres de décision est de déterminer la répartition d'une population d'individus en groupes homogènes en fonction d'un ensemble de variables discriminantes et conformément à un objectif fixe qui est la variable cible.

Comme tout algorithme, les arbres de décision ont leurs points forts et faiblesses.

Les points forts des arbres de décision sont les suivants :

- Les arbres de décision sont capables de générer des règles compréhensibles.
- Les arbres de décision effectuent la classification sans nécessiter beaucoup de calculs.
- Les arbres de décision sont capables de traiter des variables continues et catégorielles.
- Les arbres de décision fournissent une indication claire des champs les plus importants pour la prédiction ou la classification.

Les faiblesses de la méthode sont :

- Les arbres de décision sont moins appropriés pour les tâches d'estimation où l'objectif est de prédire la valeur d'un attribut continu.
- Les arbres de décision sont sujets à des erreurs dans les problèmes de classification avec de nombreuses classes et un nombre relativement faible d'exemples d'apprentissage.
- La formation d'un arbre de décision peut être coûteuse en termes de calcul. Le processus de croissance d'un arbre de décision est coûteux en termes de calcul. À chaque nœud, chaque champ candidat à la division doit être trié avant que la meilleure division puisse être trouvée. Dans certains algorithmes, des combinaisons de champs sont utilisées et une recherche doit être effectuée pour trouver les poids de combinaison optimaux.

- Ils sont instables, c'est-à-dire que de petits changements dans les données peuvent produire des arbres très différents. Les modifications apportées aux nœuds proches de la racine peuvent grandement affecter l'arborescence résultante. On dit que les arbres produisent desestimateurs de variance élevée. [9]

1.6.1.2 Random Forest

Avant de se lancer dans la description de l'algorithme Random Forest, il nous a semblé nécessaire de mettre au point certaines notions indispensables à la compréhension de ce type de modèle.

- Le Bootstrapping :

Si on considère un ensemble N, le bootstrap de cet ensemble, est l'ensemble des données obtenues à la suite d'un tirage aléatoire n fois des éléments d'avec remise.[10]

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

Supposons que l'on dispose d'un ensemble N $\{x_1, x_2, x_3, \ldots, x_{n-1}, x_n\}$de $\boldsymbol{N}$ données observées de notre population, et que l'on veut calculer une statistique $\boldsymbol{S}(\boldsymbol{T})$ (une moyenne, une variance...)[11].

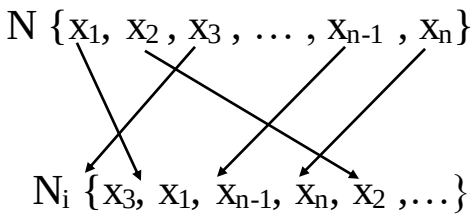

Figure 2: Exemple du Bootstrap.[11]

Le Bootstrap consistera donc à former $\boldsymbol{L}$ échantillons pour (i= 1, …, L) tel que chaque $\mathbf{N_i}$est constitué à partir d'un tirage aléatoire avec remise de $\boldsymbol{N}$données issues de notre ensemble initial. On pourra alors calculer $\boldsymbol{S}(\boldsymbol{T}_i)$ pour chaque échantillon bootstrap et obtenir ainsi $\boldsymbol{L}$ estimations de la statistique que l'on cherche àcalculer au lieu d'une seule, on fera la moyenne empirique de ces $\boldsymbol{L}$ valeurs et pour ronsestimer avec plus de précision $\boldsymbol{S}(\boldsymbol{T})$.

- Le Bagging :

Le bootstrapaggregating, également appelé bagging (de bootstrap aggregating), est un méta-algorithme d'apprentissage ensembliste conçu pour améliorer la stabilité et la précision des algorithmes d'apprentissage automatique. Il a été introduit Breiman en 1996,il permet de réduire la variance et d'éviter le surapprentissage. Bien qu'il soit généralement appliqué aux méthodes d'arbres de décision, il peut être utilisé avec n'importe quel type de méthode. Le bootstrapaggregating est un cas particulier de l'approche d'apprentissage ensembliste[12].

Le Random Forest est un Algorithme de classification composé de nombreux arbres de décisions. Formellement proposé en 2001 par Leo Breiman et Adèle Cutler, il fait partie des techniques d'apprentissage automatique. Cet algorithme combine les concepts de sous-espaces aléatoires et de rééchantillonnage avec remise ensembliste (bagging). L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.

La forêt aléatoire consiste en un ensemble d'arbres de décision binaires qui introduisent le caractère aléatoire. Ces arbres se dressent distinguer les uns des autres par les sous-échantillons de données sur lesquels ils sont entraînés. Ces sous-échantillons sont tirés au hasard (d'où le terme « aléatoire ») dans un jeu de données.

La technique des forêts aléatoires modifie la méthode du Bagging appliquée aux arbres en ajoutant un critère de décorrélation entre ces arbres. L'idée de cette méthode est de réduire la corrélation sans augmenter trop la variance. Le principe consiste à choisir de façon aléatoire un sous-ensemble de variables qui sera considéré à chaque niveau de choix du meilleur nœud de l'arbre[13].

Principe de l'algorithme :

Considérons un ensemble d'entrainement S= $\{(x_1, y_1), \ldots, (x_m, ym)\}$, et **a** le nombre d'attributs des exemples de **X** (individus).

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

Pour chaque nœud de l'arbre, l'attribut de partitionnement est choisi en considérant un nombre **f (f<a)** d'attributs choisis aléatoirement (parmi les attributs). Celui choisi c'est celui qui va

optimiser le critère d'homogénéité considéré par les arbres utilisés (entropie de Shannon ou indice de Gini)

Pour classifier une nouvelle instance, le classificateur des forêts aléatoires effectue un vote de majorité uniformément pondéré des classificateurs de cet ensemble.

Comme tout algorithme, les Random Forest ont leurs points forts et faiblesses.

Les points forts des arbres de décision sont les suivants :

- Elles permettent d'éviter le sur-apprentissage.
- Elles permettent d'améliorer les performances des arbres de décision.

Les faiblesses de la méthode sont :

- La perte de lisibilité des arbres de décisions.
- Importants temps de calcul.

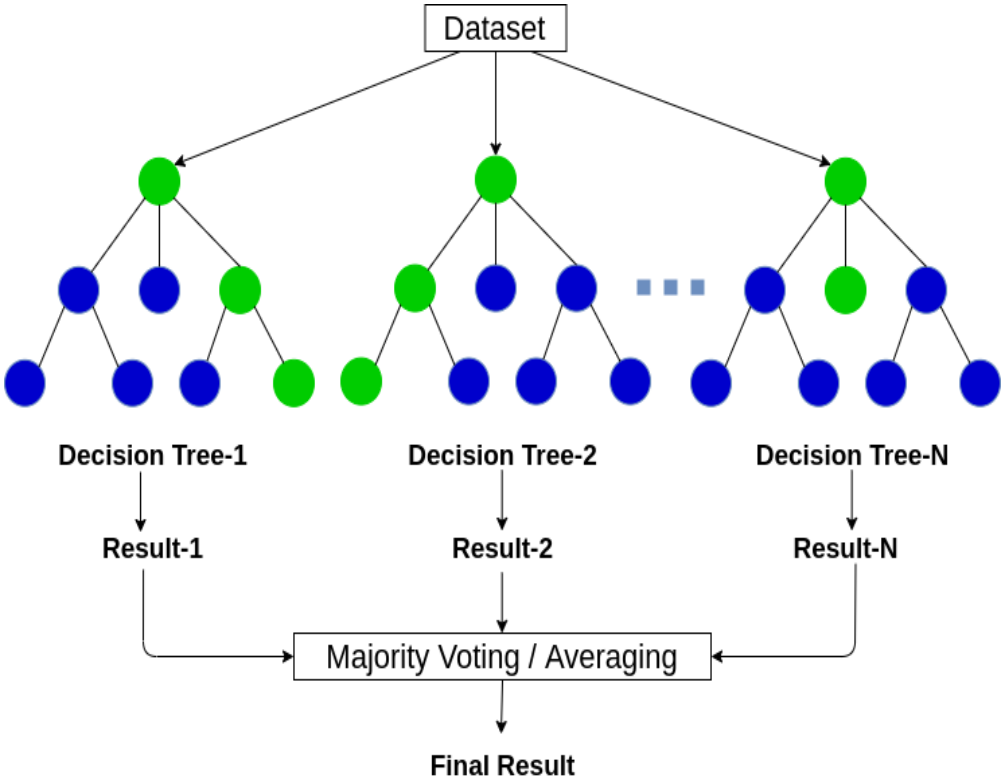

Figure 3: Exemple illustrant le déroulement de Random Forest. [14]

1.6.1.3 Logistic regression

La régression logistique ou modèle logit est une méthode statistique puissante qui peut modéliser des résultats binomiaux avec une ou plusieurs variables explicatives. Il mesure la relation entre une variable dépendante catégorielle et une ou plusieurs variables indépendantes en utilisant une fonction logistique (c'est-à-dire une distribution logistique cumulative) pour estimer la probabilité.

La régression logistique a été utilisée dans les sciences biologiques au début du vingtième siècle. Elle a ensuite été utilisée dans de nombreuses applications des sciences sociales. La régression logistique est utilisée lorsque la variable dépendante (cible) est catégorique[14].

Par exemple : Pour prédire si un email est un spam (1) ou non(0)

Si la tumeur est maligne (1) ou non (0)

La régression logistique indique généralement où se trouve la frontière entre les différentes classes, en plus de ça elle indique que les probabilités des classes dépendent de la distance de la frontière.

La régression logistique est une méthode prédictive. Cependant, par régression logistique, cette prédiction conduira à une dichotomie. La régression logistique est l'un des outils les plus couramment utilisés en statistiques appliquées et analyse de données discrètes.

La fonction qui régit le modèle de régression logistique est le suivant :

$$P = \frac{1}{1 + e^{-ywx}} \quad (1)$$

Tel que : x est le vecteur de la donnée où $x_i \in R^n$, y est le vecteur de l'étiquette de la classe où $y_i \in \{1,-1\}$ et $w \in R^n$ est le vecteur des poids.

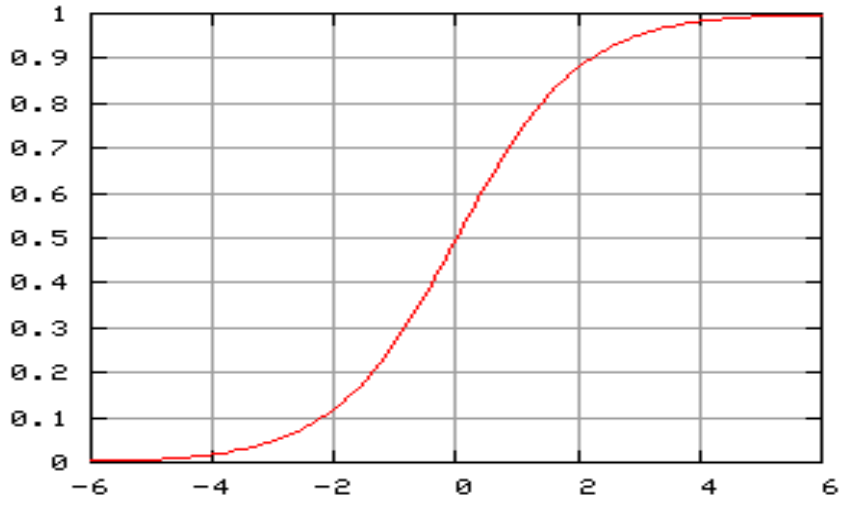

Figure 4 : Variations de la fonction logistique. [14]

1.6.1.4 Support Vector Machine (SVM)

La machine à vecteurs de support (SVM) a été introduite par [15]. Il s'agit d'un mariage entre la modélisation linéaire et l'apprentissage basé sur les instances. Il sélectionne un petit nombre d'instances limites critiques, appelées vecteurs de support, dans chaque classe et construit une fonction discriminante linéaire qui sépare chaque classe aussi largement que possible. Le système transcende les limites des frontières linéaires en rendant pratique l'inclusion de termes non linéaires dans la fonction, ce qui permet de former des frontières de décision quadratiques, cubiques et d'ordre supérieur.

Soit un ensemble de points de 2 types dans **N** lieu dimensionnel, SVM génère un hyperplan dimensionnel (**N - 1**) pour séparer ces points en 2 groupes. Supposons que certains points de 2 types peuvent être séparés linéairement. SVM trouvera la ligne droite qui sépare ces points en 2 types et qui se situe le plus loin possible de tous ces points et ce problème est dit linéairement séparable sinon il n'est pas linéairement séparable et il n'existe pas un hyperplan séparable.

L'idée de base du SVM est d'utiliser un modèle linéaire pour mettre en œuvre des limites de classe non linéaires par le biais d'une mise en correspondance non linéaire du vecteur d'entrée dans un espace de caractéristiques de haute dimension. Un modèle linéaire construit dans le nouvel espace peut représenter une limite de décision non linéaire dans l'espace original. Dans le nouvel espace, un hyperplan de séparation optimal est construit. Ainsi, le SVM est connu comme l'algorithme qui trouve un type spécial de modèle linéaire, l'hyperplan de la marge maximale. L'hyperplan de marge maximale donne la séparation maximale entre les classes de décision. Les exemples d'apprentissage qui sont les plus proches de l'hyperplan de la marge maximale sont appelés vecteurs de support. Tous les autres exemples de formation ne sont pas pertinents pour définir les limites des classes binaires.

Le SVM repose donc sur deux notions principales qui sont : la notion de marge maximale et la notion de fonction noyau.

Les machines à vecteurs de support, comme les réseaux neuronaux, ne subissent pas les contraintes des distributions statistiques. Avec les machines à vecteurs de support, il est peu probable qu'il y ait sur ajustement et elles produisent souvent des classificateurs très précis. En revanche, leur calcul est très complexe et elles sont lentes par rapport à d'autres algorithmes d'apprentissage automatique lorsqu'elles sont appliquées dans un cadre non linéaire.[16]

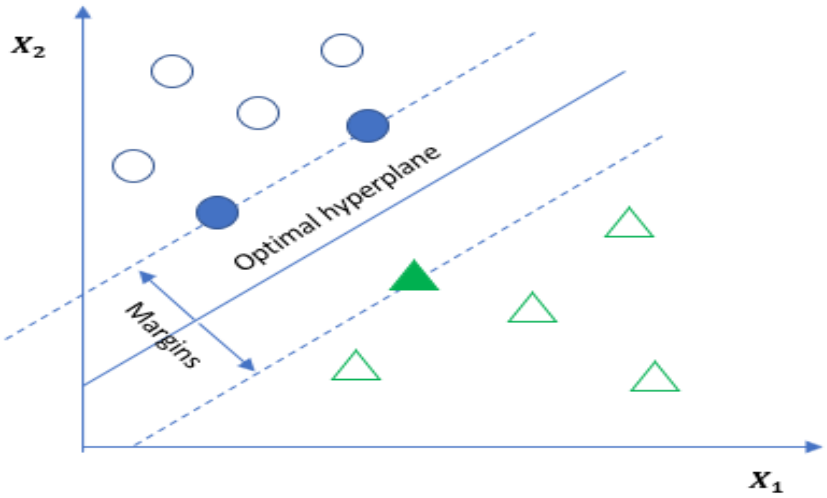

Figure 5: Support Vector Machine. [18]

Comme tout algorithme, les SVM ont leurs points forts et faiblesses.

Ses points forts sont les suivants :

- Elles sont très efficaces du fait qu'elles utilisent un sous-ensemble de points d'entraînement.
- Elles reposent sur une solide base théorique.

Les faiblesses de la méthode sont :

- La complexité des formules mathématiques utilisées lors de la classification.
- Un temps d'entraînementtrès long, ce qui fait qu'elles ne conviennent pas à des jeux de données volumineux.

1.6.1.5 KNN (K-Nearest-Neighbors)

L'algorithme K-Nearest Neighbors (KNN) est l'une des méthodes de classification les plus fondamentales et les plus simples, basée sur les exemples d'apprentissage les plus proches dans l'espace des caractéristiques. K-NN est un type d'algorithme algorithme basé sur l'instance dans la catégorie des algorithmes d'apprentissage paresseux (Aha, 1997). K-NN classifie un objet en fonction de sa similarité avec d'autres objets. La logique suppose que les objets similaires sont proches les uns des autres et que les objets dissemblables sont éloignés les uns des autres. Un objet est donc étiqueté en fonction de l'étiquette de la majorité de ses

voisins. La similarité des objets est évaluée à l'aide d'une métrique de distance appropriée, généralement la distance euclidienne. La distance euclidienne est utilisée comme métrique de distance pour les variables continues. Cependant, il n'existe pas de concept commun pour définir le nombre de voisins les plus proches, il est défini afin d'obtenir une bonne précision de classification ; mais il est intuitif d'utiliser plus d'un voisin le plus proche si la taille de l'ensemble d'apprentissage est grande.[17]

Cette méthode simple présente quelques problèmes pratiques : elle a tendance à être lente pour les grands ensembles d'apprentissage, elle est peu performante avec les données bruyantes et elle est peu performante avec les attributs non pertinents car chaque attribut a la même influence sur la décision, tout comme dans la méthode. D'autre part, l'avantage de cette méthode simple par rapport à la plupart des autres méthodes d'apprentissage automatique est qu'elle permet d'ajouter de nouveaux exemples à l'ensemble d'apprentissage à tout moment.

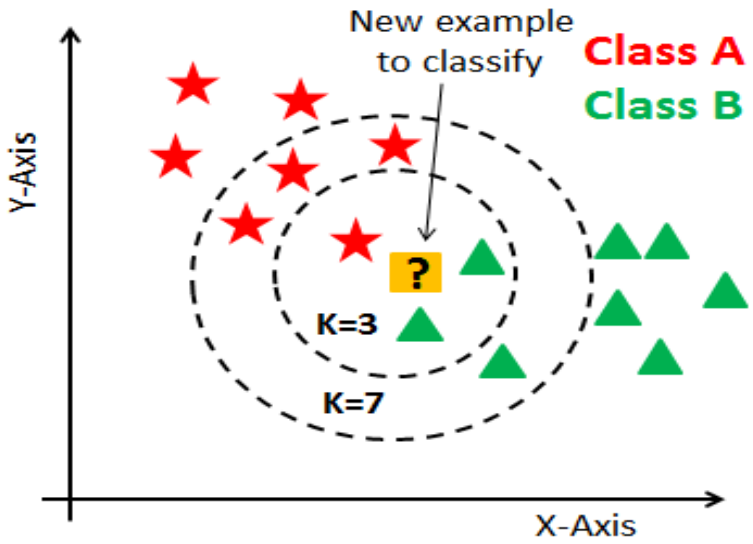

Figure 6: Exemple de classification KNN (K=3 et K=7).

1.6.1.6 Réseaux de neurones artificiels et apprentissage profond

J. M. Keller & al [18] a décrit Les Réseaux de neurones (NN) comme : "A network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes".

Ou comme décrit plus récemment par R. Shah [19]: "Artificial neural network is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing".

Les réseaux de neurones artificiels ont été inspirés par l'architecture neuronale d'un cerveau humain et, comme dans un cerveau humain, l'unité de base s'appelle un neurone artificiel. C'est une fonction mathématique qui prend une ou plusieurs entrées qui sont multipliées par des valeurs appelées poids et additionnées. Cette valeur est ensuite transmise à une fonction, appelée fonction d'activation, pour devenir la sortie du neurone.

Ce modèle est basé sur une approche de calcul appelée connectionism [20]. Dans un sens plus pratique, les NN sont des modèles de données statistiques non linéaires, ou des outils d'aide à la décision, qui sont utilisés pour modéliser des relations complexes entre les entrées et les sorties. La figure suivante représente une architecture d'un réseau de neurones artificiel avec multiple couches cachées (hiddenlayers).

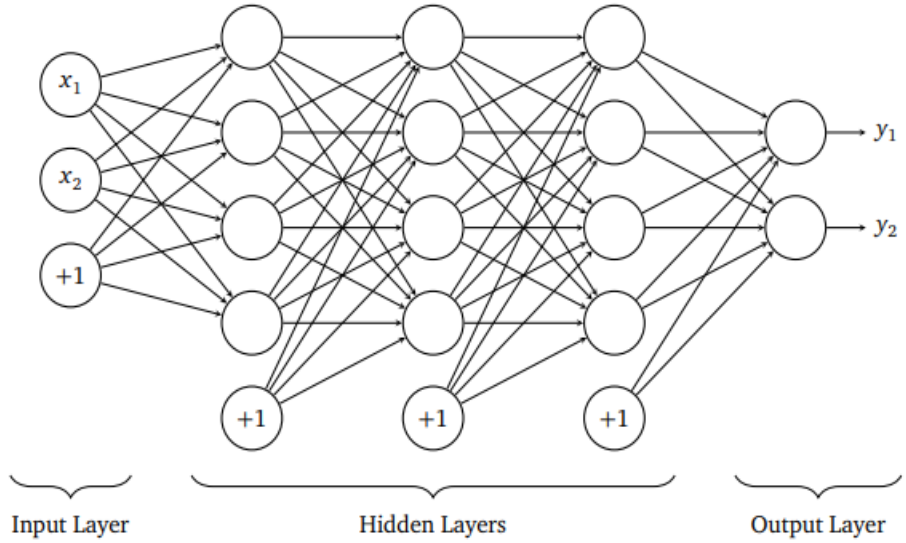

Figure 7: Architecture d'un réseau de neurones artificiel[20].

Comme nous pouvons le voir, chaque couche contient des perceptrons ou des nœuds, il n'y a pas de connexions entre les nœuds d'une même couche, cependant, chaque nœud d'une couche donnée est connecté à tous les nœuds de la couche suivante. Les couches cachées sont des couches intermédiaires entre la couche d'entrée et la couche de sortie. Chaque nœud de la couche cachée est le résultat de l'application d'une fonction sur les activations obtenues en multipliant les entrées par les poids. Les choix courants pour la fonction d'activation incluent la fonction sigmoïde et la fonction tanh.

1.6.1.6.1 Fonctions d'activation

Les fonctions d'activation peuvent faire toute la différence entre un réseau de neurones actif et un réseau de neurones qui ne fonctionne pas. Ils sont des équations mathématiques qui transforment leur entrée en une sortie utile pour le réseau de neurones. Les fonctions d'activation sont attachées à chaque neurone et décident si l'entrée de chaque neurone est pertinente pour la prédiction du modèle.

La figure ci-dessous présente les fonctions d'activation les plus utilisée :

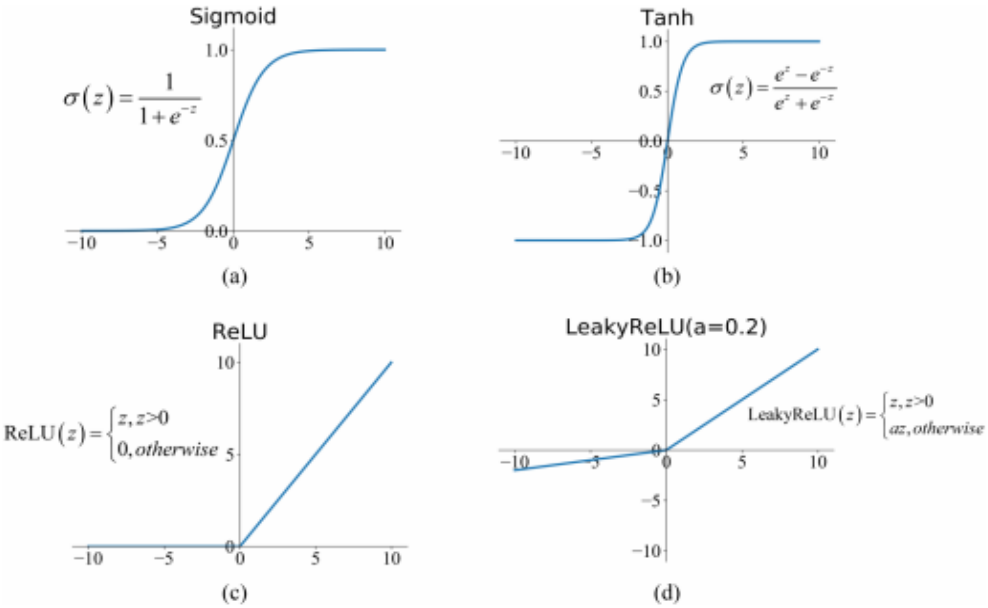

Figure 8: Les fonctions d'activations fréquemment utilisées. [25]

1.6.1.6.2 Loss function

Les machines apprennent au moyen d'une fonction de perte. elle est une méthode d'évaluation des modèles de la façon dont l'algorithme bien spécifiques, les données fournies. Si les prédictions s'écartent trop des résultats réels, la fonction de perte cracherait un très grand nombre. Progressivement, à l'aide d'une fonction d'optimisation, la fonction de perte apprend à réduire l'erreur de prédiction[21].

Dans notre cas la fonction de perte (lossfunction) la plus compatible avec notre problème c'est Binary Cross-entropycostfunction (BCE).

BCE également connue sous le nom de classification binaire, généralement utilisée pour les problèmes de prédiction où la valeur attendue pour chaque entrée est l'une des deux valeurs. L'équation suivante explique comment la calculer.

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i) \quad (2)$$

1.6.1.6.3 Gradient descent (La Descente de gradient)

Le gradient descent est un algorithme d'optimisation qui permet de calculer le minimum local d'une fonction en changeant au fur et à mesure (itérations) les paramètres de cette fonction. En d'autres termes, le gradient descent est un algorithme permettant de trouver le minimum local d'une fonction différentiable. La descente de gradient est simplement utilisée pour trouver des valeurs aux paramètres d'une fonction permettant d'atteindre ce minimum local. [22]

1.6.1.6.4 Convolutional Neural Networks(CNN)

Nous avons présenté dans cette partie le modèle CNN d'apprentissage profond qui s'adapt bien a notre problématique

Un Convolutional neural network (CNN) est un réseau de neurones artificiels le plus couramment utilisé pour analyser des images. Bien que l'application la plus courante des CNN ait été l'analyse d'images, elle peut également être utilisée pour d'autres problèmes d'analyse de données ou de classification. Plus généralement, nous pouvons considérer CNN comme un réseau de neurones artificiels doté d'un certain type de spécialisation pour pouvoir détecter des modèles et leur donner un sens. CNN a des couches cachées appelées couches convolutives, et ces couches sont précisément ce qui en fait un CNN.

1.6.2 Evaluation et validation

Vous devez toujours évaluer le modèle pour déterminer s'il vous aidera à prédire correctement la cible dans les nouvelles données à venir. Étant donné que la future instance a une valeur cible inconnue, vous devez vérifier l'indice de précision du modèle ML sur les données qui connaissent déjà la réponse cible, puis utiliser ce niveau comme indice de la précision prédictive des données futures.

La mesure de précision choisie influe sur l'applicabilité des modèles et sur leurs performances hors échantillon.

1.6.2.1 Mesure de performance

1.6.2.1.1 Confusion Matrix (matrice de confusion)

En classification, la matrice d'évaluation habituelle est la matrice de confusion. C'est un tableau qui va décrire les performances d'un modèle et mesurer sa qualité sur un ensemble de données d'entraînement, Dans la matrice, les prédictions absolues sont divisées en prédictions correctes et fausses ; c'est-à-dire le nombre de prédictions correctes et fausses établies par le modèle de classification par rapport aux résultats réels (valeur cible) dans les données. La figure suivante affiche une matrice de confusion pour deux classes (2x2).

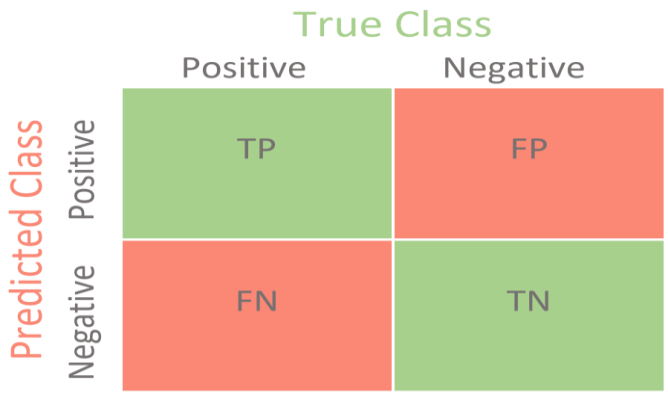

Figure 9: Confusion Matrix. [23]

- **TP (True Positive) :** Vrai positifs, c'est-à-dire les cas où la prédiction est positive, et la valeur réelle est aussi positive.
- **TN (TrueNegative) :** Vrai négatif, c'est-à-dire les cas où la prédiction est négative, et la valeur réelle est aussi négative.
- **FP (False Positive) :** Faux positif, c'est-à-dire les cas où la prédiction est positive, mais la valeur réelle est négative.
- **FN (False Negative) :** Faux négatif, c'est-à-dire les cas où la prédiction est négative, mais la valeur réelle est positive.

- **La précision :**

La précision représente sur tout l'ensemble des points qui sont déclarés positifs, le pourcentage de ces points qui sont réellement positifs, calculée par la formule suivante :

$$\text{Précision} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

1.6.2.1.2 Courbe d'évaluation-Receiver Operating Characteristics Curve (ROC Curve)

La fonction d'efficacité du récepteur, plus fréquemment désignée sous le terme -courbe ROC- dite aussi caractéristique de performance (d'un test) ou courbe sensibilité/spécificité, est une mesure de la performance d'un classificateur binaire, c'est-à-dire d'un système qui a pour objectif de catégoriser des éléments en deux groupes distincts sur la base d'une ou plusieurs des caractéristiques de chacun de ces éléments. Graphiquement, on représente souvent la mesure ROC sous la forme d'une courbe qui donne le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR). Les courbes ROC furent inventées pendant la Seconde Guerre mondiale pour montrer la séparation entre les signaux radar et le bruit de fond.[23]

La courbe ROC nous permettra de visualiser le compromis entre spécificité et sensibilité en représentant graphiquement l'évolution de la (sensibilité) en fonction de (1-spécificité) selon les valeurs d'un certain seuil S (Threshold).

Diminuer la valeur du seuil de classification S va permettre de classer plus d'éléments comme positifs, ce qui va augmenter le nombre de faux positifs et de vrais positifs.

Soit x un individu et soient les fonctions suivantes :

- La sensibilité $\alpha(S)$ = prob(score(x) $\geq$ S \ x = évènement), c'est-à-dire bien détecter un événement au seuil S.
- La spécificité = prob(score(x) $\leq$ S \ x = non-évènement), c'est-à-dire qui implique de bien détecter un non-événement au seuil S.

On dira alors que la proportion des non- événement déclarés comme évènement est $1 - \beta(S)$.

La courbe ROC va donc représenter $\alpha(S)$ en fonction de $1 - \beta(S)$ pour des valeurs de S allant du :

- Maximum où l'on considère tous les individus comme non-événement ce qui implique que :

$$\alpha(S) - 1 - \beta(S) = 0.$$

- Minimum où l'on considère tous les individus comme événement ce qui implique que :

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

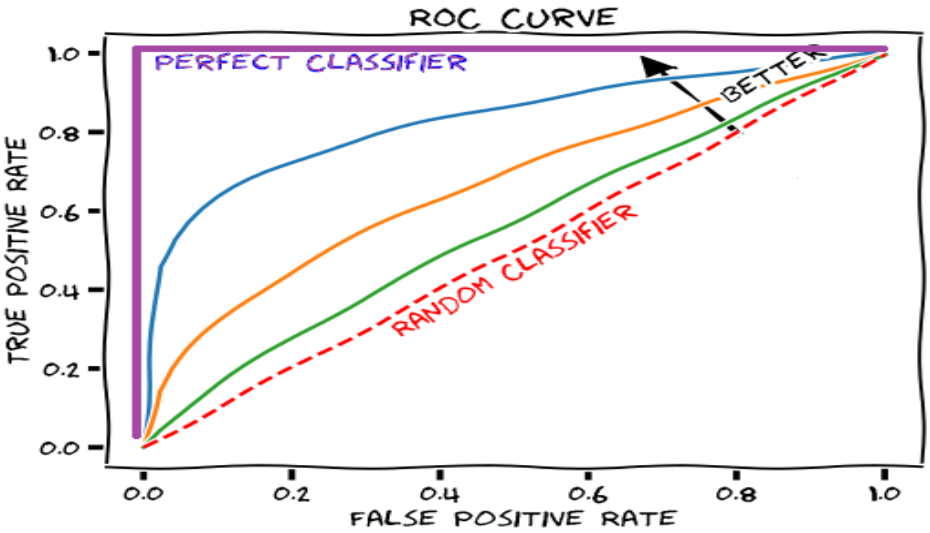

Figure 10 : Courbe ROC. [23]

Interprétation des points critiques dans le graphe précédent :

- Au point (0, 0) le classificateur déclare tous les individus comme non-événement : c'est-à-dire il n'y a aucun faux positif, mais également aucun vrai positif.
- Au point (1, 1) le classificateur déclare tous les individus comme évènement : c'est-à-dire il n'y a aucun vrai négatif, mais également aucun faux négatif.
- Au point (0, 1) le classificateur n'a aucun faux positif ni aucun faux négatif, et est par conséquent parfaitement exact, c'est-à-dire ne se trompant jamais.
- Au point (1, 0) le classificateur n'a aucun vrai négatif ni aucun vrai positif, et est par conséquent parfaitement inexact, c'est-à-dire se trompant toujours.

1.6.2.1.3 Aire sous la courbe ROC (Area Under Curve - AUC)

AUC - La courbe ROC est une mesure de performance pour les problèmes de classification à différents seuils. La courbe ROC étant une courbe de probabilité, l'AUC représente le degré ou la mesure de la séparabilité. Elle indique dans quelle mesure le modèle est capable de faire la distinction entre les classes. Plus l'AUC est élevée, plus le modèle est capable de prédire les 0 comme 0 et les 1 comme 1. Par exemple, plus l'AUC est élevée, plus le modèle est capable de distinguer les patients atteints de la maladie de ceux qui ne le sont pas.

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

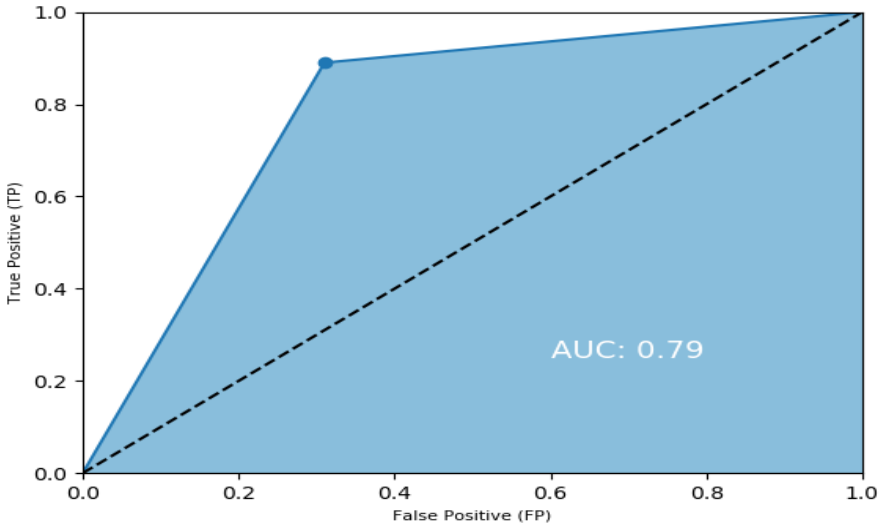

Figure 11: AUC. [24]

Lors de la création et de l'entraînement d'un modèle de machine learning, l'objectif est de choisir le modèle qui fait les meilleures prédictions, c'est-à-dire de choisir le modèle avec les meilleurs paramètres (paramètres ou hyperparamètres du modèle de machine learning). Cependant, si vous choisissez des paramètres de modèle qui produisent les « meilleures » performances prédictives sur les données de l'évaluation, vous risquez de surajuster le modèle. Lorsque le modèle mémorise les tendances qui apparaissent dans les sources de données de formation et d'évaluation, mais ne parvient pas à généraliser ces tendances dans les données, un surapprentissage se produit. Cela se produit généralement lorsque les données d'entraînement incluent toutes les données utilisées dans l'évaluation. Les modèles de surapprentissage ont bien fonctionné pendant la période d'évaluation, mais n'ont pas été en mesure de faire des prédictions précises sur des données inconnues.

Pour éviter de sélectionner un modèle surajusté comme meilleur modèle, vous pouvez conserver d'autres données pour vérifier les performances du modèle ML. Par exemple, vous pouvez séparer les données en utilisant 60 % pour la formation, 20 % pour l'évaluation et 20 % pour la validation.

Cependant, l'utilisation des données du processus de formation pour l'évaluation et la vérification réduira la quantité de données disponibles pour la formation. Ceci est particulièrement problématique pour les petits ensembles de données, car il est préférable d'utiliser autant de données que possible pour la formation. Pour résoudre ce problème, vous pouvez effectuer une **validation croisée**.

1.7 Conclusion

Dans ce chapitre nous avons parlé de l'apprentissage automatique, ses principaux types ainsi que les algorithmes de classification les plus utilisés. Le chapitre suivant est consacré pour présenter une technologie récente, intéressante de le domaine de banking,qui est le blockchain.

Chapitre 2 : Blockchain

2.1 Introduction

De temps en temps des percées technologiques se produisent qui ouvrent un tout nouveau monde de possibilités. Par exemple, l'invention d'Internet a été une percée comme celle-ci qui a changé le monde de presque tous les points de vue. La technologie Blockchain est encore une fois l'une des percées technologiques émergentes qui devrait révolutionner la façon dont les transactions sont effectuées, affectant ainsi une grande variété de domaines d'application potentiels. [24]

Ce chapitre donne un aperçu général sur la technologie Blockchain. En outre, il présente le concept, la définition et les types de Blockchain.

2.2 Definition de blockchain

Drescher [29] a définit Blockchain comme suit : «The blockchain is a purely distributed peer-to-peer system of ledgers that utilizes a software unit that consist of an algorithm, which negotiates that informational content of ordered and connected blocks of data together with cryptographic and security technologies in order to achieve and maintain integrity». [29]

Tandis que Ølnes[25] définit Blockchain comme une base de données ouverte, distribuée et sans confiance sur Internet. D'autres chercheurs et praticiens tels que [26] conviennent que la blockchain est une structure de données distribuées, une base de données ou un système. Wang [27] également d'accord avec la définition précédente, mais il a ajouté plus d'explications sur la définition montrant certaines fonctionnalités Blockchain, il a défini Blockchain comme une base de données distribuée comprenant des enregistrements de transactions ou d'événements numériques qui ont été exécutés et partagés entre les parties participantes.

De plus, la partie intéressante de la Blockchain est sa nature d'ouverture. Signifie Blockchain par elle-même que la technologie est une composante nécessaire, mais pas suffisant pour créer la véritable magie comme une plate-forme de confiance qui n'est pas

contrôlée par quiconque et ne peut pas être effacé ou modifié après que les transactions ont été enregistrées.

2.2.1 Signification de l'appellation « Blockchain »

Chaque nouveau bloc qui recense les dernières transactions effectuées fait référence au bloc précédent. Une fois qu'un bloc de transaction est validé, on lui applique la fonction de hachage, et ainsi, une simple ligne de caractères fait référence au bloc précédent, qui lui-même faisait référence au bloc précédent, et ainsi de suite. Cette série de block permetde créer ce que on a appelle une blockchain.

2.3 Types de Blockchain

Il existe essentiellement trois types de Blockchain : blockchain privé, blockchain consortium et blockchain public. Alors que ces types parfois regroupés à nouveau en fonction de la perspective d'autorisation d'appartenir à trois catégories ainsi : blockchain ouvert pour l'accès ouvert, blockchain fermé pour l'accès restreint et blockchain hybride pour l'accès personnalisé tombe entre les anciens types. [28]

Le premier type classique de Blockchain existe ouvertement est le public Blockchain (Permissionless) tels que le blockchain Bitcoin [29]. Cela est appelé public dans le sens où les réseaux sont ouverts au grand public pour se joindre en tant qu'utilisateurs ou servir de nœuds, également dans le sens que les données blockchain est publiquement transparent [28].

Cependant, cette ouverture permet plus de possibilités de pirater la blockchain. Ça veut dire que tous ceux qui souhaitent rejoindre le réseau peuvent voir les données. Ce problème de confidentialité des données ne peut pas être complètement protégé, car il n'est pas possible d'anonymiser toutes les données. Par conséquent, Les blockchains privées (souvent retrouvées sous le terme anglo-saxon "permissioned blockchains") ne disposent que d'une seule entité capable d'ajouter de nouvelles données. Comme pour les blockchains consortium, les droits de lecture peuvent être publics ou restreints à des utilisateurs définis.

Les blockchains peuvent également être créées plus ou moins flexibles, ou spécifiques, dans quelles actions Sont autorisés. C'est exactement où le troisième type de Blockchain se situe entre les blockchains privé et public, qui est appelé blockchain hybride [30] où il y a une certaine flexibilité pour les solutions hybrides à mettre en œuvre.

Bien qu'il n'existe pas encore de taxonomie officielle définie pour de nombreux aspects des modèles / types de conception Blockchain. Comme l'étude illustrée à la figure 23 qui

donne un aperçu de la matrice des blockchains permissionless/permissioned et généralisées/spécialisées.

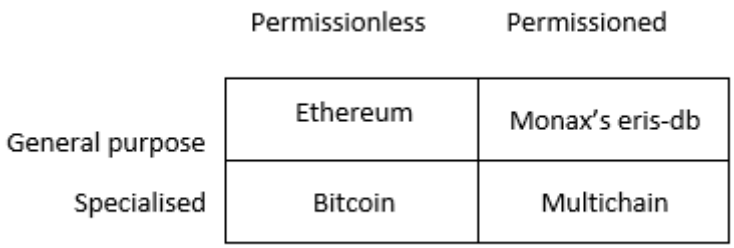

Figure 12: Types de chaînes de blocs autorisés et sans autorisation. [31]

La figure 12 donne également des exemples d'une application ou d'une plateforme blockchain pour chaque catégorie, qui sont décrits brièvement ci-dessous :

• **Ethereum:** est une plate-forme logicielle ouverte à usage général et sans autorisation basée sur la technologie blockchain qui permet aux développeurs de construire et de déployer des applications décentralisées qui exécutent des contrats intelligents. [31]

• **Bitcoin:** est une application à but spécialisé (crypto-monnaie), une forme de trésorerie électronique. C'est une monnaie numérique décentralisée sans une banque centrale ou un administrateur unique (ouvert au public pour se joindre et utiliser) qui peut être envoyé par un utilisateur à un autre utilisateur sur le réseau peer-to-peer bitcoin sans le besoin d'intermédiaires [29].

• **Multichain:** est une plateforme de création et de déploiement de blockchains privées. Il résout les problèmes liés à l'exploitation minière, à la protection de la vie privée et à l'ouverture grâce à la gestion intégrée des autorisations des utilisateurs [28], [32].

2.3.1 Le rapport Blockchain/Bitcoin

Bitcoin a été la principale mise en œuvre de la technologie de la blockchain [33]. Bitcoin est une monnaie numérique, utilisé principalement comme une nouvelle méthode de paiement qui utilise la technologie Blockchain comme l'un des concepts fondamentaux. En d'autres termes, cela signifie simplement Blockchain est une technologie tandis que Bitcoin est une application. Cependant, la Blockchain n'est pas seulement une technologie qui prend en charge Bitcoin, il est bien au-delà de cela car il peut être utilisé pour enregistrer et enregistrer en toute sécurité toute transaction de valeur et pas seulement les transactions financières [34], [35]. Il semble y avoir des possibilités infinies pour les applications basées

sur la Blockchain que les entreprises et les gouvernements sont en train de développer. Puisque l'objectif de cette recherche est de comprendre, expliquer la technologie Blockchain et l'analyse de ses futures orientations de recherche, cette étude ne se concentrera pas sur Bitcoin.

2.4 La Cryptographie dans la sécurité des blockchains

2.4.1 Terminologie de base

Un chiffrement de données (ou chiffrement) est une paire d'algorithmes qui créent le chiffrement et le déchiffrement inversé. Le fonctionnement détaillé d'un chiffrement est contrôlé à la fois par l'algorithme et dans chaque cas par une "clé". La clé est un secret (idéalement connu seulement pour les communiants), habituellement une courte chaîne de caractères, qui est nécessaire pour déchiffrer le chiffrement. [36]

Le chiffrement des données peut être classé en trois branches : clé non saisie, clé symétrique et clé asymétrique [37], comme le montre la figure 15. Les primitives non saisies sont des fonctions qui n'utilisent pas de clé pour chiffrer un message, par ex. hachage de longueur arbitraire et permutations. Les primitives à clé symétrique utilisent la même clé pour le chiffrement et le décryptage alors que la cryptographie à clé asymétrique utilise le système d'une clé publique et d'une clé privée. La figure suivante résume les différents types de chiffrement.

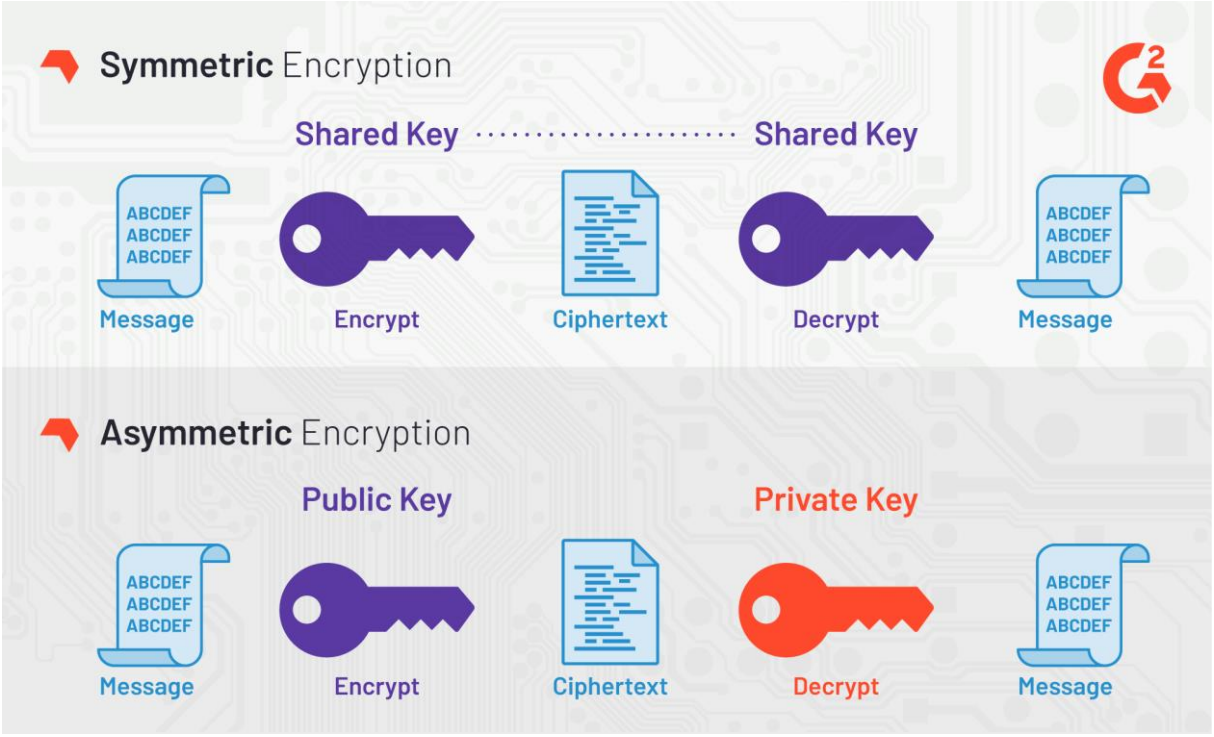

Figure 13:Les trois principaux types de chiffrement : hachage, symétrique, asymétrique. [36]

2.4.2 Fonctions de hachage

Une fonction de hachage est simplement une fonction qui prend la valeur d'entrée, et à partir de cette entrée crée une valeur de sortie déterministe de la valeur d'entrée. Pour toute valeur d'entrée x, la sortie recevra toujours la même valeur de sortie y chaque fois que la fonction de hachage sera exécutée. De cette façon, chaque entrée a une sortie déterminée. Comme le montre la figure 13 Une fonction de hachage est donc quelque chose qui prend une entrée (qui peut être n'importe quelle donnée comme des nombres, des fichiers, etc) et produit un hachage. Un hash est généralement affiché sous forme de nombre hexadécimal.

Les différents algorithmes de hash tels que (MD, SHA) sont les fonctions de hash les plus populaires [38]. Les fonctions de hachage sont généralement irréversibles (à sens unique), ce qui signifie qu'on ne peut pas comprendre l'entrée si on ne connaît que la sortie – à moins d'essayer toutes les entrées possibles (ce qui est appelé une attaque par force brute [37]). Il y'a de nombreuses applications pour les fonctions de hachage, mais le contrôle de l'intégrité des données est une application la plus Common. Elle est utilisée pour générer les sommes de contrôle sur les fichiers de données. Cette application fournit une assurance à l'utilisateur sur l'exactitude des données. [38]

2.4.3 Infrastructure à clés publiques (ICP)

L'infrastructure à clés publiques (ICP) [37], [38] est un ensemble d'exigences qui permettent (entre autres) la création de signatures numériques. Grâce à l'ICP, chaque transaction de signature numérique comprend une paire de clés : une clé privée et une clé publique. La clé privée, comme son nom l'indique, n'est pas partagée et n'est utilisée que par le signataire pour signer électroniquement les documents. La clé publique est librement accessible et utilisée par ceux qui doivent valider la signature électronique du signataire. L'ICP applique des exigences supplémentaires, comme l'autorité de certification (AC), un certificat numérique, un logiciel d'inscription de l'utilisateur final et des outils de gestion, de renouvellement et de révocation des clés et des certificats. [39]

2.4.4 Signatures numériques

La signature numérique est un processus qui garantit que le contenu d'un message n'a pas été modifié en transit [38]. Lorsque l'expéditeur, le serveur, signe numériquement un document, il ajoute un hachage unidirectionnel (cryptage) du contenu du message à l'aide de la paire de clés publiques et privées de l'expéditeur. Le client peut toujours le lire, mais le processus crée une "signature" que seule la clé publique du serveur peut décrypter. Le client, en utilisant la clé publique du serveur, peut alors valider l'expéditeur ainsi que l'intégrité du contenu du message. Si la transmission arrive mais que la signature numérique ne correspond pas à la clé publique du certificat numérique, alors le client sait que le message a été modifié.

La signature numérique est donc un schéma mathématique de présentation de l'authenticité des messages ou documents numériques. Une signature numérique valide donne au destinataire des raisons de croire que le message a été créé par un expéditeur connu (authentification), que l'expéditeur ne peut pas nier avoir envoyé le message (non-répudiation) et que le message n'a pas été modifié en transit (intégrité). [40]

Lorsqu'un signataire signe électroniquement un document, la signature est créée à l'aide de la clé privée du signataire, qui est toujours conservée en toute sécurité par le signataire. L'algorithme mathématique agit comme un chiffrement, créant des données correspondant au document signé, appelé hachage, et cryptant ces données. Les données chiffrées qui en résultent sont la signature numérique. La signature porte également l'heure à laquelle le document a été signé. Si le document change après la signature, la signature numérique est invalidée. [38]

2.4.5 Architecture de réseau

Caractéristique clé de la technologie Blockchain réside dans sa nature distribuée [41]. Contrairement aux réseaux centralisés et décentralisés, la figure 14 montre la différence entre ces trois architectures de réseau. Un système de réseau informatique réparti est un système où les données et les ressources sont réparties sur divers nœuds matériels. Il est structuré comme une architecture de réseau peer-to-peer au sommet d'Internet. Le terme peer-to-peer, ou P2P, signifie que les ordinateurs qui participent au réseau sont des pairs entre eux, qu'ils sont tous égaux, qu'il n'y a pas de nœuds spéciaux. [42]

Tous les nœuds de réseau partagent l'effort de fournir des services de réseau. Où, les nœuds fournissent et consomment des services en même temps. Distribution de contrôle dans la technologie Blockchain et l'application Bitcoin est un principe de conception de base qui ne peut être réalisé et maintenu par l'architecture réseau P2P.

En outre, dans le scénario Blockchain chaque nœud maintient une base de données (grand livre) de toutes les transactions valides, qui sont envoyées entre les nœuds dans le réseau. Bien que chaque nœud contienne une copie du grand livre, seuls les utilisateurs qui détiennent la signature peuvent accéder à l'information [43]. Signifie que le grand livre partagé peut être considéré comme des conteneurs (blocs) où les données sont stockées. Cependant, ces conteneurs sont scellés et leur contenu ne peut être vu que par ceux qui détiennent la permission. [44]

Les nœuds s'identifient par leur adresse IP, tandis que les utilisateurs s'identifient par leur clé publique [45]. Par conséquent, chaque nœud peut envoyer une transaction à tous les autres nœuds du réseau s'il connaît la clé publique du récepteur, sans qu'aucune autorité centrale ne soit impliquée dans la transaction.

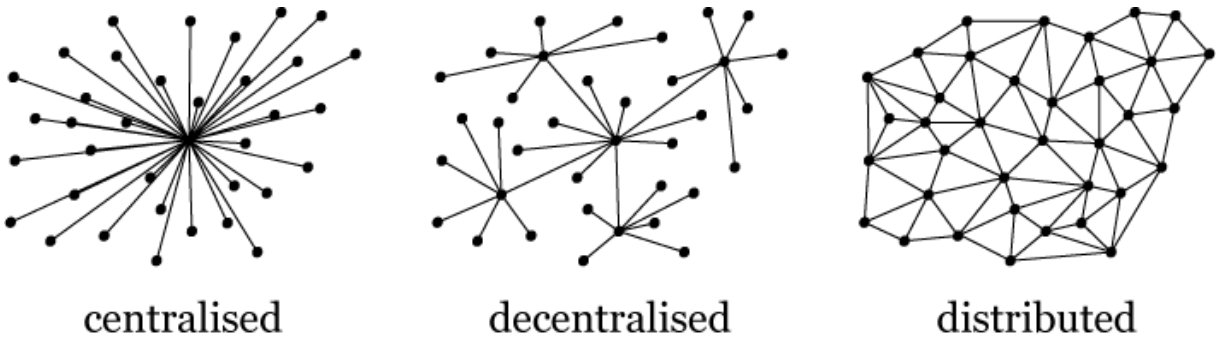

Figure 14 : Architectures réseau centralisées (a), décentralisées (b) et distribuées (c). [43]

2.4.6 Transactions

Les transactions sont la partie principale du grand livre Blockchain [42]. Toutes les technologies de la Blockchain sont conçues pour garantir que les transactions peuvent être créées, propagées sur le réseau, validées et finalement ajoutées au block transactions (la Blockchain). D'un point de vue technique, la définition la plus fondamentale d'une transaction est un événement atomique autorisé par le protocole sous-jacent [46]. Dans le cas de Bitcoin, les transactions sont généralement des paiements individuels, par exemple: - Alice envoie Bob 0.1 BTC (montré sur la figure suivante).

Figure 15: Transaction d'Alice au café de Bob. [43]

Donc, simplement, une transaction est juste un tas de données qui décrit le mouvement de Bitcoin. Ou autre Cependant, Bitcoin est seulement l'un des nombreux Blockchains. En d'autres termes, toutes les chaînes de blocs ne limitent pas leur utilité aux transactions de paiement. L'Ethereum blockchain [31] est similaire à Bitcoin (en ce qu'il peut gérer les paiements), mais il stocke également d'autres types d'informations.Par exemple, ceux-ci pourraient stocker un programme sur la blockchain Ethereum qui garde une trace de qui possède quels titres à quelles propriétés de logement. Le programme pourrait également être responsable du retour à la propriété si un prêt hypothécaire n'est pas payé à temps, ou il pourrait également transférer la propriété.

Le bloc de construction fondamental d'une transaction blockchain est une sortie de transaction Les sorties de transactions sont des morceaux indivisibles de devises utilisées, enregistrées sur la Blockchain et reconnues comme valides par l'ensemble du réseau. [42]

2.4.7 Consensus distribué

Une transaction est valide pour qu'elle ait lieu et ajouté à la blockchain, cette approbation majoritaire appelée un consensus distribué. Ainsi, au lieu d'une entité approuvant toutes les transactions et maintenant la base de données exacte, dans l'environnement Blockchain cette responsabilité est partagée entre tous les participants du réseau. C'est ainsi que la nature distribuée de Blockchain fonctionne pour l'approbation des transactions, toutes les personnes connectées au réseau sont en mesure d'avoir un mot à dire dans si une transaction doit être acceptée à la blockchain ou non.

En général, il ne serait pas viable pour tout le monde sur le réseau d'accepter d'approuver la transaction de faux comme valide quand quelqu'un essaie de tricher le système. Comme pour de nombreuses chaînes de blocs, le seuil de consensus est supérieur à 50 %, si plus de 50 % des participants au réseau conviennent qu'une transaction est valide, alors elle est acceptée comme valide. Cependant, le risque et les dangers potentiels si plus de 50% du réseau accepte une transaction invalide (qui est discuté comme l'un des défis Blockchain en vertu de la section 2.5.1)

2.4.8 Preuve de travail (Mining and Proof of Work)

Lorsque les demandes de transactions sont envoyées à chaque ordinateur sur le réseau pour valider (approuver) et inclure ensuite dans Blockchain. Afin de valider une transaction et d'ajouter à la Blockchain, les ordinateurs du réseau doivent rivaliser pour résoudre un « puzzle » connecté au bloc suivant, avant d'être ajoutés à la Blockchain. Ce processus de résolution du puzzle connu sous le nom de Preuve de travail. L'ordinateur qui résout ce puzzle en premier, ils reçoivent une récompense, généralement payée dans la crypto-monnaie ou jeton utilisé sur ce réseau (Car c'est comme extraire de petites quantités de valeur d'un bloc).

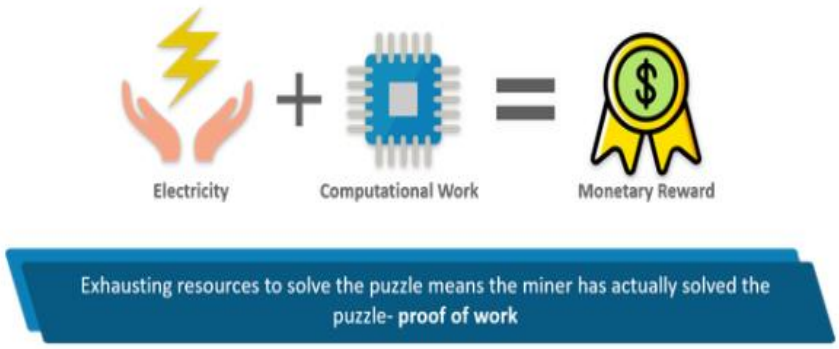

Figure 16: Les mineurs sont récompensés pour avoir économisé les ressources du réseau. [47]

Les mineurs qui résolvent ce problème et qui ont ajouté un bloc valide au réseau sont récompensés pour avoir fourni de l'énergie informatique, de l'électricité et des ressources au réseau, ce qui aide à maintenir le réseau en marche [voir figure 16].

Cet ordinateur qui résout le puzzle en premier peut alors ajouter des transactions dans le bloc, puis ajouter le bloc à la blockchain. Ainsi, Preuve de travail est le processus par lequel les transactions sont vérifiées et ajoutées au grand livre public qui est coûteuse à exécuter, mais facile à vérifier pour d'autres et qui satisfait à certaines exigences. [42]

2.4.9 Portefeuille numérique (Digital Wallet)

Un portefeuille numérique est simplement l'interface utilisateur du système Blockchain, voir la figure 17. Tout comme un navigateur Web qui est l'interface utilisateur la plus courante pour le protocole HTTP pour utiliser Internet. Techniquement, il s'agit d'un logiciel qui stocke les clés privées et publiques et interagit avec diverses blockchain pour permettre aux utilisateurs d'envoyer et de recevoir des valeurs numériques (monnaie en cas de Bitcoin, euro, dinar ect…) et de surveiller leur solde [11]. Si quelqu'un veut utiliser Bitcoin ou tout autre crypto-monnaie, il / elle devra avoir un portefeuille numérique. [48]

Il existe de nombreuses implémentations et marques de portefeuilles Bitcoin. Tout comme il existe de nombreuses marques de navigateurs Web (p. ex., Chrome, Safari, Firefox et Explorer).

Figure 17: Un exemple d'interface de portefeuille numérique Bitcoin.

2.5 Avantages et défis de la blockchain

La technologie Blockchain est considérée comme une technologie émergente, où il Ya toujours une place pour l'amélioration Cependant, ses caractéristiques peuvent être facilement identifiées et discutées sur la base de ses concepts techniques. Ces caractéristiques sont résumées dans la figure 18.

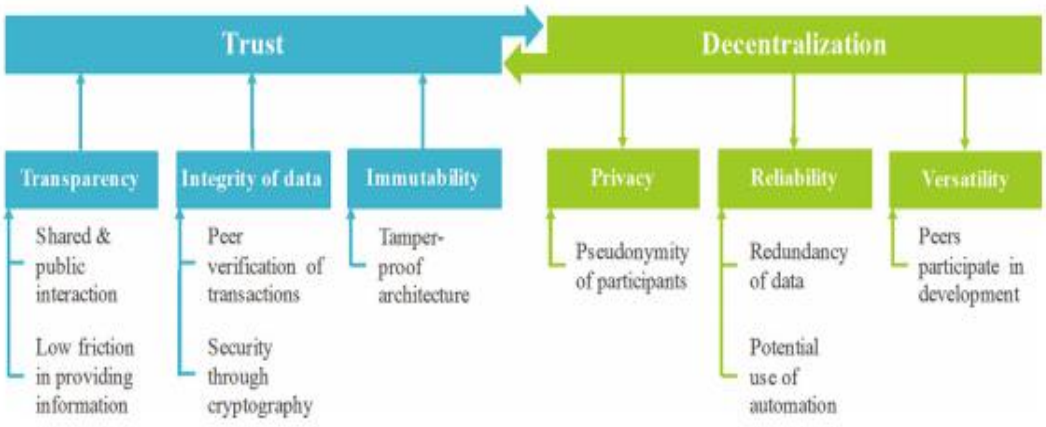

Figure 18: Caractéristiques de la technologie Blockchain.

2.5.1 Avantages

2.5.1.1 Protection des renseignements personnels

Dans la technologie Blockchain la confidentialité est comme une pièce de monnaie avec deux visages opposés. C'est un avantage et un défi en même temps. Du côté positif, les identités des participants restent anonymes [24], ce qui a permis un haut degré de confidentialité pour ses participants.

2.5.1.2 Fiabilité

Les mécanismes blockchain assurent la fiabilité par le biais de cela, il n'y a pas de transactions ajoutées au grand livre à moins qu'il ne soit exploité, validé et confirmé, puis il est répliqué à travers le réseau suite au mécanisme de stockage de la décentralisation (ou des grands livres distribués) de la Blockchain.

2.5.1.3 Polyvalence

Le concept de la technologie Blockchain utilisé bien au-delà de la Bitcoin car il est applicable aux travaux comme un grand livre public pour tout type de transactions non seulement la monnaie numérique.

2.5.1.4 Transparence

La technologie Blockchain offre la transparence à toutes les personnes sur le réseau, comme les transactions visibles à tous les ordinateurs connectés sans être contrôlé par un tiers. Où, la majorité de ces ordinateurs doivent approuver les transactions de tout changement à la Blockchain, cependant les transactions ne peuvent pas être modifiés ou supprimés une fois enregistrées. [24]

2.5.1.5 Intégrité des données

L'intégrité des données est l'une des principales caractéristiques qui facilitent et renforce de la confiance dans la technologie Blockchain. Cette intégrité des données est obtenue par une astucieuse idée de cryptographie qui constitue le mécanisme de consensus.

2.5.1.6 Immutabilité

Blockchain sont conçus pour être immuables, moyens d'avoir des données écrites une fois qu'il s'est produit et être disponible pour tout le monde. Techniquement, une fois qu'une transaction est ajoutée à un bloc, qui à son tour est ajoutée à la Blockchain, cette transaction ne peut pas être modifiée. [42]

2.5.2 Défis de blockchain

Comme toute autre technologie révolutionnaire, Blockchain a ses avantages et inconvénients ainsi. Tout comme outre technologie blockchain a ces propres problèmes tel que la sécurité, attaque 51% etc.… On peut Considérer ces problèmes comme des défis qui doivent être résolus.

2.5.2.1 Défis techniques

A. **Sécurité**

La sécurité est l'un des principaux sujets de recherche dans la technologie Blockchain.

Mais tout aspect a des biens faits et des points faibles parmi les problèmes qui ont été cernés, les attaques à 51 %, les problèmes de malléabilité des données, les problèmes d'authentification et de cryptographie. Où chacun de ces problèmes de sécurité sera discuté plus en détail dans les paragraphes suivants :

- **Attaque à 51 %** :Les mécanismes de Blockchain conçu avec l'hypothèse que les nœuds honnêtes contrôlent le réseau [50]. Si les nœuds attaquants contrôlent collectivement plus de puissance de calcul que les bons, le réseau est vulnérable à la soi-disant attaque à 51%. **Problèmes de malléabilité des données** : L'intégrité des données est un problème essentiel dans l'environnement Blockchain car les données doivent être envoyées à toutes les parties du réseau pour vérification, il est donc important de ne pas être altéré

- **Problèmes d'authentification et de cryptographie** : Dans la blockchain, la clé privée est le principal élément d'authentification. Cependant, il y a eu quelques incidents avec l'authentification , comme le cas bien connu dans Mt.Gox, où une société de portefeuille Bitcoin a été attaqué et les clés privées de leur client ont été volé [50]. Pour résoudre ce problème, de nombreux chercheurs ont suggéré diverses solutions pour renforcer l'authentification dans Blockchain

B. Ressources gaspillées : exploitation des Blockchain nécessite une grande quantité d'énergie gaspillées pour calculer et vérifier les transactions en toute sécurité. Il existe diverses solutions proposées pour le problème des ressources gaspillées dans la Blockchain et la littérature Bitcoin tels que [47], [51].

C. Confidentialité Comme Blockchain est basé sur un réseau de consensus distribué sans une partie de confiance centralisée où, toutes les transactions sont transparentes et annoncées au public. Par conséquent, la confidentialité dans Blockchain est maintenue en brisant le flux d'informations.

Conslusion

Dans ce chapitre nous avons défini le blockchain, ses principaux types, le Domaine de la cryptographie dans la sécurité des transactions ainsi que les avantages et les défit des blockchains.

Le chapitre suivant est dédié à la problématique et à la présentation de l'environement bancaire ainsi que l'organisme d'accueil

Chapitre 3 : Prédiction de la faillite d'entreprise (Solvabilité des clients)

Introduction

L'étude de l'existant est une étape très importante lors de la réalisation d'un projet. L'étude du système actuel et des processus, permettra de mieux cerner la problématique afin de proposer la solution idéale pour l'entreprise.

Cette section est le résultat de plusieurs entretiens des employeur et analyste risque. Ils nous ont expliqué le fonctionnement du système et les différents échanges d'informations. En se basant sur cette étude, nous allons diagnostiquer la procedure actuelle et identifier les anomalies afin d'arriver à une solution adéquate.

1. Définition d'une Banque

Larousse a décrit la banque comme suit :

> « Établissement financier qui, recevant des fonds du public, les emploie pour effectuer des opérations de crédit et des opérations financières, et est chargé de l'offre et de la gestion des moyens de paiement ». [52]

Ou comme cité dans le Journal du net :

> Une banque est une entreprise qui a une activité financière. Elle constitue, juridiquement, une institution financière régie par le code monétaire et financier. Sa fonction principale consiste à proposer des services financiers tels que collecter l'épargne, recevoir des dépôts d'argent, accorder des prêts, gérer les moyens de paiement.Chaque banque est spécialisée selon son activité principale et sa clientèle. Il peut s'agir d'une banque de dépôt, qui est le secteur bancaire le plus connu. Ce type de banque reçoit l'épargne de ses clients et accorde des prêts. L'établissement peut également être une banque d'investissement, qui a une activité de conseil et de financement des entreprises. Elle opère aussi des opérations sur les marchés financiers. Enfin, il peut s'agir d'une banque privée, qui est spécialisée dans la gestion de gros portefeuilles. Cette dernière propose des services haut de gamme pour la gestion de

patrimoines dont la valeur est importante. Une banque peut également proposer des services annexes tels que l'assurance, la mutuelle ou encore le cautionnement. [53]

2 Client Bancaire

Le client bancaire est au centre des préoccupations des banques, ce concept était il y a encore quelques années l'apanage d'un nombre réduit de banque. Actuellement tout le secteur bancaire place le client au centre de l'univers économiques, c'est le client qui fournit les ressources nécessaires à la pérennité de la banque et c'est toujours le client qui utilise les services de la banque moyennant le règlement des frais y afférent.

2.1 Le résident/non-résident

Le statut de résident/ non résident est définie par la règlementation algérienne comme suit :

- Est considéré résident toute personne physique ou morale dont le centre principal d'activité est situé en Algérie ;
- À contrario, le non-résident est toute personne dont le centre principal d'activité est situé en dehors de l'Algérie.

2.2 Emprunteur & Garant

2.2.1 L'Emprunteur

Est qualifié d'emprunteur toute entreprise bénéficiaire de financements et / ou d'engagements de la part de la banque.

2.2.2 Le Garant

Est une personne physique (associé ou tierce) ou personne morale, qui s'engage à honorer les obligations de l'emprunteur en cas de défaillance de celui-ci. L'étude de la capacité juridique et financière du garant est préalable à toute acceptation de celui-ci. Les revenus et charges du garant ne sont pas cumulés avec ceux de l'emprunteur. Le garant doit pourvoir assumer seul la charge du crédit en cas de défaillance de l'emprunteur. Le garant peut faire partie d'une des sociétés du même groupe.

3 Segmentation Clients

3.1Définition:

La segmentation est une méthode de découpage des domaines d'activités stratégiques d'une entreprise en segments mais également de ses clients en sous-ensembles (segment

clientèle). Les segments sont dits stratégiques quand ils ont pour vocation le découpage des activités de l'entreprise en marché, ils sont dits marketing quand ils divisent les clients de l'entreprise.

Le but de la segmentation stratégique étant d'engager, au mieux, les ressources à moyens long terme vers la création et/ou la conservation d'avantages concurrentiels de chaque segment stratégique.

Dans ce qui suit, nous avons utilisé comme exemple la segmentation du groupe Société Générale Algérie (Section 3.5).

3.2 Segmentation du groupe SGA

La Banque au même titre des autres filiales des groupes Société Générale doit respecter des critères de segmentation. Ainsi sont définis 5 marchés de référence(Figure 19) :

- Particuliers
- Professionnels/TPE
- Entreprises
- Collectivités locales, États et Institutions Publiques
- Institutions Financières

Ces marchés de référence sont eux-mêmes divisés en segments de marché. La segmentation par marché, comme définit ci-dessus, permet :

- d'organiser la force de vente en adéquation avec cette répartition, en termes de technicité,
- d'élaborer des gammes d'offres adéquates,
- de communiquer de manière adéquate.

Les marchés de référence et les segments de marché affectés aux clients sont déterminés à partir des informations collectées au cours de la relation client. Cette segmentation distingue également les marchés selon un mode d'exploitation « Retail » ou « non Retail »

- Les marchés « Particuliers » et « Professionnels/TPE » font partie du périmètre « Retail ».

- Les marchés « Entreprises », « Collectivités locales, États et Institutions Publiques » et « Institutions Financières » font partie du périmètre « Non Retail ». [54]

le tableau et la figure suivante résume la segmentation des clients de SGA.

Tableau 1 : Principaux marchés SGA.

Mode d'exploitation	Banque de détail (Retail Banking)		Banque Commerciale (Non Retail Banking)		
Principaux Marchés	Particuliers	Professionnels/TPE	Entreprises	Collectivités Locales, États, Institutions Publiques	Institutions financières

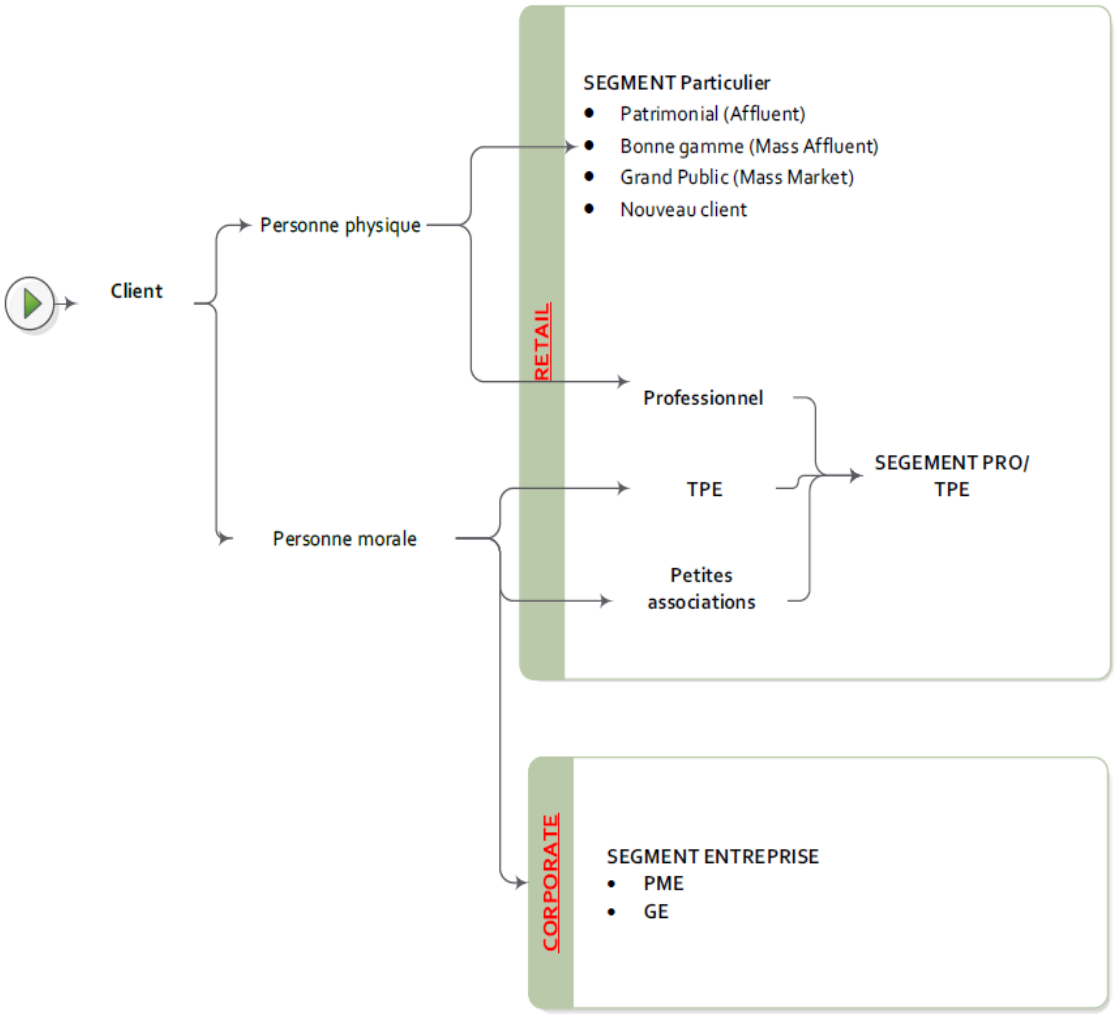

Figure 19: Exemple de segmentation client de Société Générale Algérie. [54]

4 Crédit bancaire

4.1 Définition

D'après Larousse [55]

« Acte par lequel une banque ou un organisme financier effectue une avance de fonds ; délai accordé pour un remboursement ; montant de l'avance. ».

Ou Comme cité dans Wikimemoire [56]

« Le mot **Crédit** à la même étymologie que le mot 'Croire' (en latin, crédo = je crois, j'ai confiance).C'est donc une activité qui repose la confiance, celle que le prêteur accorde à l'emprunteur, de qui, il attend le remboursement du prêt.

En finance, le crédit englobe les diverses activités de prêt d'argent, que se croit sous la forme de contrats de prêts bancaire ou de délais de paiement d'un fournisseur à un client.

Le crédit est généralement porteur d'un intérêt que doit payer le débiteur. (Le bénéficiaire du crédit, appelé aussi emprunteur) au créancier (celui qui accorde le crédit, appelé aussi prêteur).

Dans le domaine bancaire, un crédit bancaire est une mise (ou une promesse) à disposition de fonds à une date ou une période donnée contre obligation de remboursement moyennant une rémunération.

Un crédit se conclu par l'intermédiaire d'un contrat entre un emprunteur et un préteur. Les banques sont les principaux fournisseurs de crédit, tant aux particuliers qu'aux entreprises. »

4.2 Typologie des crédits bancaires

Il existe plusieurs types de crédit qui sont distingués selon les critères suivants : l'objet, la durée et les caractéristiques.

Selon l'objet du crédit on trouve deux types de crédits qui sont :

- Les crédits pour particuliers (ex : crédit-bail, crédit à la consommation, crédit immobilier…)
- Les crédits pour les entreprises et les professionnels (crédit d'exploitation, crédit d'investissement), et ce sont ces derniers qui nous intéressent dans notre travail et on va les détailler ultérieurement.

Selon la durée du crédit, on peut citer 3 principales types qui sont :

- Crédit à court terme (moins de 2 ans)
- Crédit à moyen terme (entre 2 et 7 ans)
- Crédit à long terme (de 10 à 20 ans)

Enfin on trouve différents types de crédit selon leurs caractéristiques, par exemple on a les crédits selon la monnaie c'est-à-dire soit en monnaie nationale ou en devise, on trouve aussi des crédits selon le type de taux (fixe ou variable). [57]

4.3 Cycle de vie d'un crédit bancaire

Le cycle de vie de chaque diffère selon sa nature, mais on peut distinguer des phases communes à tous les types qui sont les suivantes [57]:

- L'étude et la mise place : Cette étape consiste à formuler et identifier le besoin et la demande du client, collecter les différents documents nécessaires, la prise des garanties, analyse et la prise de décision et enfin le déblocage du crédit.
- Le remboursement du crédit : cette étape s'articule autour de l'appel des échéances, le remboursement des échéances et la clôture.
- Le recouvrement :en cas de constatations des impayés, on va procéder la renégociation des conditions et classer les crédits comme douteux.

4.4 Processus d'octroi de crédit

Le processus d'octroi de crédit se décompose en quatre étapes :

- Analyse préparatoire de la contrepartie.
- Constitution de la demande de crédit.
- Analyse et validation de la demande de crédit par la filière risques.
- Mise en place du crédit octroyé.

4.5 Définition du risque Crédit

Pour un agent économique donné, le risque clients mesure ou évalue l'exposition à une créance impayée. Autrement dit, c'est un calcul ou une évaluation d'une probabilité d'un défaut de paiement sur un client potentiel. Cela sous-entend que le créancier a prédéfini des critères ou paramètres pour calculer ou estimer l'éventualité d'un défaut de paiement.

En simplifié, il attribue une note ou un score à un client potentiel, déterminé à partir de diverses données combinées – revenus stables ; historique des remboursements ou de gestion du compte bancaire ; montant de la dette ; type et statut d'emploi ; etc. Pour une banque, le risque client concerne le risque d'impayé d'une ou de plusieurs mensualités par un ou plusieurs emprunteurs potentiels sur une période donnée. Plus généralement, les mesures algorithmiques du risque constituent le cœur de l' activité de tout établissement de crédit, voire de tout agent économique. [58]

5 Présentation de l'organisme d'accueil

5.1 Présentation du Groupe Société Générale

Société Générale c'est l'un des tout premiers groupes européens de services financiers et acteur important de l'économie depuis plus de 150 ans, accompagne au quotidien 30 millions de clients grâce à ses 133 000 collaborateurs présents dans 61 pays. Le Groupe allie

solidité financière, dynamique d'innovation et stratégie de croissance durable avec pour objectif la création de valeur pour l'ensemble de ses parties prenantes. [59]

Le groupe comporte 3 piliers essentiels de l'activité de développement :

- Les réseaux de détail en France (Société Générale, Crédit du Nord et Boursorama).
- Les réseaux de détail à l'international (IBFS : International Banking and Financial Services).
- La banque de financement et d'investissement (SG CIB, GBIS, SGSS) qui gère d'un côté la Banque de financement et les Revenu fixe, le financement structuré, la dette, le forex, et de l'autre côté les Equity et les activités de conseil.

En soutien au développement de ses trois piliers, les deux autres lignes métiers du Groupe sont :

- Services financiers spécialisés & assurances
- Banque privée, Gestion d'actifs et Services aux investisseurs

5.2 Présentation de la Filiale Société Générale Algérie

Société Générale Algérie, détenue à 100% par le Groupe Société Générale, est l'une des toutes premières banques privées à s'installer en Algérie, soit depuis 2000. Son réseau, en constante extension, compte actuellement 91 agences réparties sur 31 wilayas dont 13 Centres d'Affaires dédiés à l'activité de la clientèle des Entreprises. Société Générale Algérie propose des services dans les 2 segments Retail et Corporate, elle offre une gamme diversifiée et innovante de services bancaires à plus de 230 000 clients Particuliers, Professionnels et Entreprises.L'effectif de la banque est de plus de 1 500 collaborateurs au 31 décembre 2019. [60]

5.3 Présentation de la direction SIOP

La direction des Systèmes d'informations, Organisations et Projets est rattachée directement au Pôle Support et Opération de SOCIETE GENERALE ALGERIE. Elle a pour but de définir et de contrôler l'application de la politique informatique et la gestion des processus métiers, des normes standards en matière de technologies de l'information et de systèmes d'informations. Elle assure aussi le pilotage et le suivi des projets internes et externes.[61]

Son organisation est illustrée dans l'organigramme suivant :

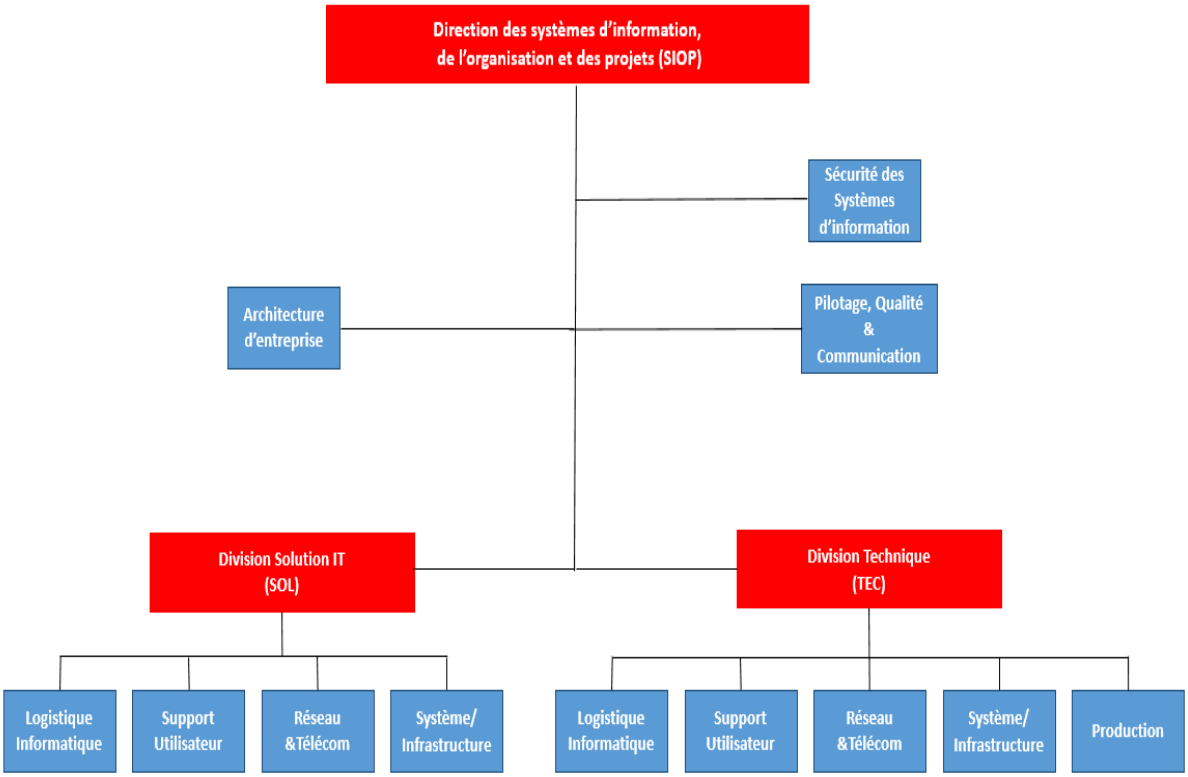

Figure 20 : Organigramme de département SIOP. [61]

5.3.1 Missions et activités de la SIOP

Parmi les missions essentielles de la Direction des Systèmes d'informations, Organisations et Projets figurent [61]:

- Maintenir en bon état de fonctionnement le patrimoine applicatif et technique de la banque.
- Piloter et délivrer les projets des métiers en adéquation avec la stratégie de la banque.
- Assurer une veille technologique afin d'identifier les nouvelles opportunités d'évolution qui répondront aux besoins futurs des métiers.
- Garantir la protection des actifs informationnels de la banque en termes de confidentialité, Intégrité, Disponibilité et traçabilité.
- Maintenir la cohérence de l'infrastructure technologique en accord avec les besoins et la stratégie de la banque.

5.3.2 Le Département Architecture de l'entreprise

La Société Générale s'inscrit dans un environnement réglementaire, commercial et technologique en constante évolution. Cette évolution impacte significativement les organisations, les processus métier et les systèmes d'information de la banque et connait une accélération forte due à la transformation numérique. Pour faire face à ce contexte général, la Société générale a défini et mis en place une démarche d'Architecture d'Entreprise afin de garantir l'alignement des projets de transformation des modèles opérationnels aves les ambitions des métiers.

Dans ce cadre la Société Générale Algérie a mis en place le Département de l'Architecture d'Entreprise, qui comprend l'architecture applicative et fonctionnelle, l'architecture technique, la gestion de la donnée et la gestion des référentiels et des processus.[61]

Ses principales missions sont les suivantes :

- Assurer la cohérence d'ensemble en termes d'urbanisation.
- Veiller au déploiement des paternes en conformité avec les pratiques du groupe SG.
- Identifier les données sensibles.
- Veiller à la disponibilité, l'intégrité, la sécurité et la qualité des données.
- Maintenir à jour des référentiels données, Applications et Processus, Organisations, Acteur
- Déployer la Gouvernance des données.

5.3.2.1 L'organisation du Département Architecture de l'entreprise

Rattachée à la direction des Systèmes d'information, Organisation et Projets, l'Architecture d'Entreprise se compose de 4 activités, organisées comme suit :

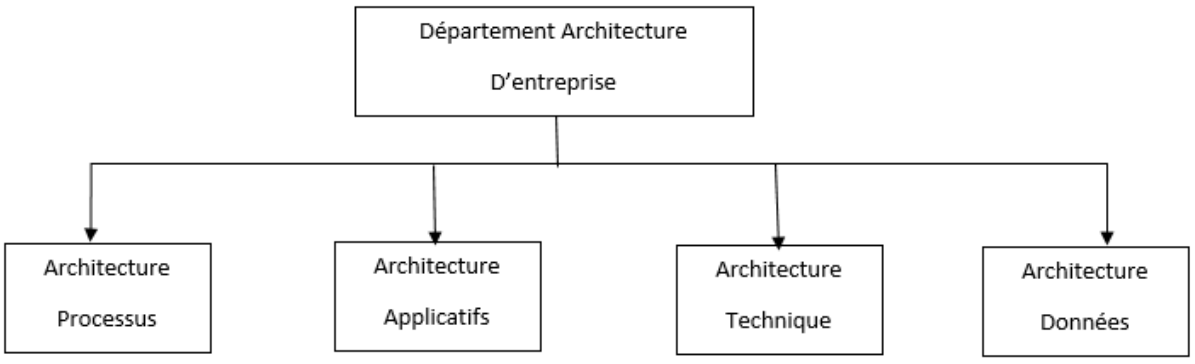

Figure 21 : Organigramme du département Architecture Entreprise.

5.4 Diagnostic et analyse des besoins

Indispensable à tout projet, le diagnostic comprend une évaluation de l'existant, et sur cette base, la mise en place des moyens nécessaires pour optimiser et proposer des solutions au sein de l'organisation.

Après avoir étudié les processus existants et les procédures décisionnel lors de l'octroi d'un crédit, nous avons collecté et analysé les besoins exprimés pendant des entretiens avec les analystes risques principalement et avec les autres parties concernées (département risque et département commerciale). L'analyse des documents internes de SGA a permet également de recueillir les besoins et de bien comprendre le fonctionnement détaillé des activités de la banque pour pouvoir détecter des axes d'amélioration au sein de l'entreprise.

5.4.1 Collecte des besoins

Il existe de nombreuses approches de collecte des besoins, qui varient d'une entreprise à l'autre en fonction de son organisation et de son mode de fonctionnement. Dans le cadre de notre étude, nous avons mené des entretiens avec des employés du départements Risk, étudié les différentes procédures internes de l'entreprise pendant le premier mois afin de comprendre le fonctionnement détaillé de la banque et étudié les différentes sources de données qui nous ont été fournies.

Dans un premier temps, tout a commencé par une dizaine d'entretiens et de meeting avec nos deux encadreurs au sein de SGA qui sont :

- Responsable de l'architecture de l'entreprise.
- Auditrice interne.

Pour pouvoir proposer et se mettre d'accord sur un axe d'amélioration, L'idée de départ sur laquelle on s'était mise d'accord était de travailler sur les crédits Corporate et c'est une chose qui n'a pas changé, car ce type de crédit est le plus répandue au sein de la banque.

L'objectif initiale était de désigner les crédits corporate les plus utilisés (SPOT, ASF et CMT), ensuite mettre en place un modèle d'apprentissage spécialement dédié à un seul type de ces crédits qu'est le plus important vis-à-vis la banque (le plus profitable ou dans lequel la banque a connu un nombre important de déclassement). Après cela on comptait généraliser ce modèle et l'ajuster sur le reste des types de crédit.

Un analyste risque nous a expliqué que cette démarche n'est pas faisable, car on peut pas prendre en considération un seul type de crédit et laisser les autres de côté sinon les informations vont être biaisé, car la plupart des clients corporate bénéficient de plusieurs lignes de crédit en même temps, chaque ligne pour répondre à un besoin précis de leur activité et dès que le client aille des retard de paiement sur un crédit parmi ceux dont il bénéficie, tout le client sera considéré comme déclassé et non pas le crédit en question, exemple : supposant qu'un client (A) a bénéficié de 3 lignes de crédit (SPOT,ASF,CMT) et qu'il a honoré l'échéance de la première ligne de crédit à temps mais au moment de l'échéance de la 2ème ligne de crédit pour des raison (X) il n'a pas pu les honorer, pour la banque ce client et toutes les lignes de crédit dont il en a bénéficié vont être considérer comme déclassé, ce qui va biaiser l'apprentissage du modèle si on considère chaque crédit séparément.

5.4.1.1 L'analyse risque d'un dossier crédit

Ensuite, on a décidé d'étudier plus profondément comment se déroule l'analyse risque d'un dossier crédit en générale, qui suit la procédure suivante :

Afin d'octroyer un crédit à un client, l'analyste risque étudie minutieusement le dossier fourni par le client pour pouvoir déterminer si ce client aurait la capacité de rembourser ses échéances à temps ou pas ; pour les clients corporate l'élément important à prendre en compte lors de cette analyse c'est la santé financière de l'entreprise emprunteuse.

Pour analyser la santé financière d'une entreprise, l'analyste risque étudie et compare chaque ratio/indicateur avec l'échelle d'acceptation interne de la banque, ces indicateurs se situent sur les notices financières sous forme de fichiers Excel, cette démarche se fait manuellement par l'analyste risque qui selon son expertise va juger si la santé financière du client est adéquate pour rembourser ses échéances ou permettra en cas de liquidation de faire face à l'exigible de la relation

5.4.2 Enoncé de la problématique

La problématique sur laquelle notre intervention s'est basée est celle de l'analyse risque d'un dossier de crédit et ce pour les raisons suivantes :

- La banque dans son état actuel a exprimé le besoin d'apporter une amélioration au processus d'octroi des crédits corporate d'une manière globale, afin d'engendrer un progrès important sur la plupart des crédits octroyés.

- Un travail manuel sur une quantité d'informations volumineuse (grand nombre de ratios et d'indicateurs financiers) est nécessairement de longue durée, ce qui peut engendrer un frein dans l'atteinte des objectifs annuelles, et pour aborder cette situation un modèle d'apprentissage automatique pourra intervenir en tant qu'outil d'aide à la décision, en mettons en lumière les ratios à considérer selon le profil du client demandeur pour attirer l'attention de l'analyste risque sur les métriques les plus importants du cas actuel pour lui faire gagner du temps et lui faciliter sa prise de décision.

Néanmoins, on a rencontré quelques problèmes liés à la politique de confidentialité de la banque, ainsi que l'extraction et la normalisation des données coutent énormément de temps. En constatent l'ampleur de ces problèmes, on a proposé de travailler avec des Datasets gratuits ayant la même structure de données que le dataset prévu (ratios et indicateurs financiers annuels des clients), afin d'évaluer les résultats sur des données externes et les négocier avec l'analyste risque pour après refaire les mêmes étapes directement sur les données de la banque.

6 - Développement méthodologique dans la littérature internationale

La prédiction de faillite d'entreprise a attiré une grande attention dans la science pendant de nombreuses décennies. Selon les recherches de [62] tout au long du développement historique de la prédiction de faillite, des modèles ont été publiés dans le monde entier en appliquant plus de 50 méthodes différentes et 500 variables. L'article englobe les méthodes les plus distribuées ayant le plus d'impact sur la recherche scientifique et application pratique.

D'un point de vue méthodologique, la prévision des faillites est un problème de classification binaire visant à différencier le mieux possible les groupes de sociétés solvables et insolvables [63]. La prévision des faillites est considérée comme une discipline limitative entre la finance d'entreprise, les statistiques et Data Mining, qui tente de prédire la solvabilité future des entreprises en utilisant des ratios financiers comme variables explicatives en appliquant des méthodes multivariées. [64]

Tout au long de la première moitié du XXe siècle, il n'existait ni méthodes statistiques ni d'ordinateurs disponibles pour prédire la faillite. Les ratios financiers des entreprises

défaillantes et non défaillantes ont été comparés, et il a été conclu qu'en cas de faillite des sociétés, les ratios les plus fréquemment utilisés avaient le plus mauvais comportement [65].

La première percée méthodologique s'est produite lorsque [66] a publié un modèle de notation de crédit fondé sur une analyse discriminante univariée (DA). Cette méthode s'est propagée dans le monde entier plus tard avec le modèle univarié DA de [67]. En réalisant que la classification des observations à l'aide d'une variable ne fournit pas un résultat fiable, [68] ont appliqué l'analyse de régression multivariée et l'DA pour élaborer un crédit système de notation pour les clients bancaires.

Dans le cas de clients plus risqués DA multivariée a montré de meilleurs résultats, en particulier par rapport au système d'évaluation des experts appliqué précédemment, donc de plus en plus d'attention a été accordée à la méthode. Le succès a été atteint par le modèle DA multivariée de renommée mondiale [69], qui a été en mesure de classer les entreprises dans l'échantillon avec 95 % de précision de classification. Depuis sa première publication, le modèle a subi plusieurs révisions. Cependant, malgré le grand nombre d'applications réussies, les limites du modèle se sont concrétisées, qui peut d'abord être ramené à l'hypothèse statistique rigoureuse de l'DA, ensuite à l'application d'une définition par défaut comme variable cible, et troisièmement la facilité d'utilisation du modèle a été réduite par le fait qu'il avait été développé dans un éventail relativement restreint de sociétés (sociétés boursières américaines), limitant ainsi son applicabilité à des populations différentes de la base de données de modélisation.

Depuis les années 1970, le développement du domaine a été dominé par la modernisation des méthodes de classification mathématique-statistique et des solutions informatiques qui les soutiennent [64].

En passant par la distribution et les hypothèses de variance de l'DA, la régression logistique (logit) est devenue une méthode de prédiction de faillite de plus en plus populaire, qui a d'abord été appliquée par [70] sur une base de données de risque de crédit. Dans la distribution mondiale de logit, la publication [71]a représenté une étape importante, qui a développé un modèle logit sur un échantillon de 105 entreprises insolvables et de 2058 entreprises solvables, exprimant ainsi que les sociétés insolvables représentent une part plus faible de la population que les sociétés solvables. L'application de la régression probit a commencé dans les années 1980 pour des raisons méthodologiques similaires [72].

Les méthodes non paramétriques n'ayant pas de postulat statistique sont apparues dans la prévision de la faillite depuis les années 1980. Les arbres décisionnels, qui sont encore aujourd'hui des outils répandus pour résoudre les problèmes de classification et pour effectuer un datamining efficace, ont d'abord été utilisés pour la prévision de la faillite par[73].

Les années 1990 ont posé de nouveaux défis aux spécialistes et aux praticiens de la prévision des faillites [74].Plusieurs critiques concernaient des modèles linéaires (ou linéarisables), des modèles robustes et les méthodes appliquées précédemment. En conséquence, les réseaux neuronaux (NN) appartenant à la famille des méthodes d'intelligence artificielle ont été stimulés pour améliorer la fiabilité des modèles [75]. Les NN ont été appliqués pour la première fois à la solvabilité des clients par [76]. Les auteurs ont prouvé que les performances des réseaux de backpropagation à trois couches surpassaient les résultats des méthodes antérieures. Depuis lors, les NNsont été largement distribués, ont connu des développements important et représentent l'une des méthodes les plus populaires d'aujourd'hui.

Depuis le début des années 2000, l'application des systèmes neuro-fuzzy à la prévision de la faillite est devenue un objet de recherche intensive, offrant de meilleurs résultats que les NNs traditionnels [77]. En parallèle, la procédure de Support Vector Machine (SVM) a également démontré une plus grande précision de classification que les méthodes appliquées précédemment, qui a d'abord été publié sur la base d'un échantillon d'entreprises australiennes utilisant vingt fois la validation croisée[78]. En outre, les méthodes de rough set theory (RST) [79],k Nearest Neighbors (KNN) [80],les réseaux de Bayes[81], les algorithmes génétiques (GA) [82], la quantification des vecteurs d'apprentissage (LVQ) [83]et le raisonnement fondé sur des cas (CBR) [83]a également commencé à se répandre dans les années 2000.

Dans les années 2010, les méthodes d'ensemble en tant que cas particulier de combinaisons de méthodes ont gagné en importance au lieu d'appliquer individuellement certaines méthodes de classification [84]. L'essence d'entre eux est le bootstrapping multiple et l'application des procédures de classification sur plusieurs sous-échantillons.

La puissance de classification du modèle final est la moyenne de celle des modèles individuels, généralement supérieure à la puissance de classification sans utiliser de méthodes d'ensemble.

Les méthodes d'ensemble les plus fréquemment appliquées sont le boosting, bagging, randomsubspace, randomforest, Gauss-processes et autoencoder appartenant à la famille des procédures d'apprentissage automatique [64]. Les recherches actuelles sur les prévisions de faillite sont sans ambiguïté dominées par le machine learning, data mining, l'intelligence artificielle et le hybridmodelling par la combinaison créative de différentes nouvelles méthodes [85]. La prédiction de faillite en tant que problème de classification multivarié est un sujet très populaire dans les concours de Data Mining visant à trouver des algorithmes de plus en plus fiables et contemporains, c'est ainsi qu'un éventail toujours plus large de solutions innovantes devient public de jour en jour.

6.1 Une vue générale sur les travaux étudiés

A travers ce chapitre on a veillé à faire une lecture approfondie et pertinente a des articles scientifiques de valeur Et à travers ce tableau ci-dessous nous avons résumé les différentes clés qui caractérisent chacun des articles revus.

Le tableau ci-dessous présente les différentes données utilisées dans les articles revus, ainsi que les différents algorithmes d'apprentissage adoptés.

Tableau 2: Résumé des travaux étudiés sur L'application du Machine Learning sur la Solvabilité Clients

Etude	Titre	Données	Algorithmes Utilisés
Eystein Nordby Meese, TorbjørnViken[86]	Utilizing machine learning for improved bankruptcy predictions in the Norwegian market with an emphasis on financial, management and sector statements	Les données ont été fournies par le Centre de recherche appliquée (SNF) de la Norwegian School of Economics (NHH). La base de données comprend tous les comptes d'entreprise norvégiens de 1991 à 2016, quelle que soit leur taille.	KNN, ANN, Random forest, SVM
TamásKristóf, MiklósVirág[87]	MACHINE LEARNING MODELS FOR PREDICTING FINANCIAL DISTRESS.	Le dataset utilisé dans cette étude est le SEC EDGAR (2017) dataset. La base de données offre un accès public gratuit aux informations d'entreprise relatives aux États-Unis. Elle contient des données financières trimestrielles s'étalant sur plusieurs années. Pour chaque entreprise,	Decision Tree, Naive Bayes, ANN

		l'ensemble de données EDGAR classe les entreprises en échec ou en activité.	
Joseph BONELLO, Xavier BRÉDART, Vanessa VELLA[88]	Using Machine Learning, Neural Networks, and Statistics to Predict Corporate Bankruptcy.	Les expériences ont été réalisées avec un grand nombre de rapport annuel Belges. Depuis 1987, la Banque nationale de Belgique a mis les rapports annuels sur CD-ROM. Ces CDROM contiennent des informations sur environ 175 000 entreprises. A l'aide de ces CD-ROM, une collection qui contient des informations sur 576 entreprises de construction a été réformé. Les informations sur chaque société sont constituées de 10 ratios financiers qui ont été calculés à partir d'un rapport annuel.	LDA, Decision Tree, ANN
Flavio Barboza , Herbert Kimura , Edward Altman[85]	Machine learning models and bankruptcy prediction	Les auteurs ont collecté des données financières sur des entreprises américaines et canadiennes de 1985 à 2013 à l'aide de Compustat. Les informations sur l'insolvabilité des entreprises ont été collectées à partir de la base de données Salomon Center de NYU.	Logit, SVM, ANN, Random forest

6.2 Discussion

Le tableau suivant montre les résultats des meilleures valeurs de précision de chaque algorithme après apprentissage et réglage des hyperparametres.

Tableau 3: Résultats des différents modèles d'apprentissage utilisés.

Titre	Algorithmes	Précision
Utilizing machine learning for improved bankruptcy predictions in the Norwegian market with an emphasis on financial, management and sector statements	KNN	97.2%
	ANN	76.5%
	Random forest	77.7%
	SVM	69.6%
MACHINE LEARNING MODELS FOR PREDICTING FINANCIAL DISTRESS.	Decision Tree	78.46%
	ANN	75.88%
	Naive Bayes	75.13%

Using Machine Learning, Neural Networks, and Statistics to Predict Corporate Bankruptcy.	LDA,	71%
	ANN	73%
	Decision Tree	71%
Machine learning models and bankruptcy prediction	Logit	76.29%
	SVM	79.77%
	ANN	72.98%
	Random forest	87.06%

Tout au long de cette étude on a pu extraire en sus des nouvelles connaissances des nouvelles approches et méthodes pour réussir sa stratégie numérique. Dans l'ensemble des articles qu'on a choisit la base de données qui était utiliser appartient différente banques à travers le monde ce qui a rendu ce travail assez faible en termes d'unicité des données et de multiplicités de méthodes utilisées.

Par ailleurs les résultats obtenus on révélé que certains algorithmes ou techniques du Machine Learning peuvent servir à monter un client assez solvable avec succès. En effet, le KNN a donné des résultats impressionnants avec une précision de 97% contrairement au SVM qui a eu un mauvais comportement dans la première étude, en outre le DecisionTree a été préféré selon l'étude qu'a été mener dans le deuxième article et ce pour sa simplicité et sa maintenabilité, alors pour la quatrième étude le Random Forest a démontré sa meilleure capacité de prédiction parmi les quatre classificateurs utilisés. En ce qui concerne les ANNs ils ont montrés des résultats assez satisfaisants dans la deuxième et troisième étude avec des précisions respectivement 75.88% et 73%.

Dans l'ensemble on peut conclure que les techniques du Machine Learning peuvent renforcer et accélérer le processus de solvabilité des clients en adoptant les méthodes adéquates et en explorant les données nécessaires pour réussir sa stratégie numérique.

7 Conclusion

En guise de conclusion, cette partie nous a d'abord permis de définir le champ de l'étude avec la présentation de l'organisme d'accueil, ses principales activités, en mettant l'accent sur l'octroi de crédit et tous les critères d'approbation d'un dossier de crédit.

Au cours du chapitre suivant, nous détaillerons notre approche de la résolution de la problématique, et donc de la conception de modèles d'apprentissage automatique.

Chapitre 4 : Expérimentation

Introduction

Le présent chapitre présente une analyse ainsi qu'une discussion des résultats des différents modèles développés.La première section expose un récapitulatif des statistiques descriptives des informations contenues clans notre base de données. La seconde, présentelesrésultats de l'analyse comparative entre Les performances des cinq différentesméthodes. La troisième, révèleles résultatsdétailles de chaquemodèle, ainsi que les variables lesplus prédictives et les caractéristiques distinguant les meilleurs modèles générés pour chacune de nos cinq méthodesutilisées.

4.1 Présentation descriptive des données

Le DataSet utilisé dans cette étude est TEJ Normalized 2013 TaiwanDataSet[1]. La base de données offre un accès public gratuit aux informations d'entreprise concernant Taïwan. Elle contient 52 variables « Taux et indicateurs financiers » trimestrielles s'étalant sur plusieurs années. Pour chaque entreprise, l'ensemble de données TEJ les classe en échec ou en activité,

Les données concernant les ratios financiers, le secteur d'activité et la taille de l'entreprise ont été obtenues ou calculées dans la base de données TEJ et sont utilisées dans les modèles ML.Ces taux financiers se présentent comme décrit dans l'Annexe A.

4.2 Architecture générale de la solution

[1] https://www.kaggle.com/chihfongtsai/taiwanese-bankruptcy-prediction

L'objectif de l'étude est d'identifier une technique d'aide à la décision efficace pour la prédiction des faillites des entreprises (Solvabilité clients), tout en essayant de déterminer les variables sur lesquelles les modèles finaux semblent s'appuyer le plus pour prédire les probabilités. La figure ci-dessous présente le workflow de Machine Learning utilisé :

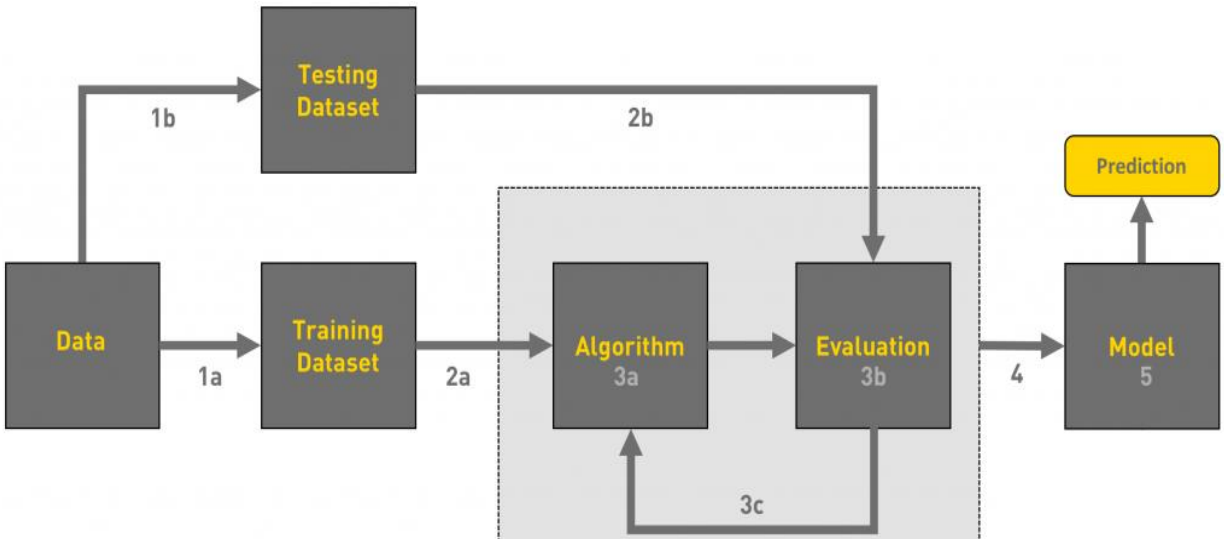

Figure 22 : Workflow de Machine Learning utilisé

Figure 23: Workflow de Machine Learning utilisé

La collection d'entreprises est répartie au hasard en deux sous-ensembles, un ensemble d'apprentissage et un ensemble d'essai (1a, 1b). Pour obtenir des résultats de test plus fiables, 20% des données est utilisée pour les tests, en supposant que 80% est une quantité raisonnable de données d'entraînement pour chacune des méthodes d'apprentissage. Nous utilisons une 5-fold cross-validation sur l'ensemble d'apprentissage pour choisir les « bonnes » valeurs de paramètres pour les différents algorithmes. Dans la 5-fold cross-validation, l'ensemble d'apprentissage D est divisé de manière aléatoire en 5 sous-ensembles D1, D2, D3, D4, D5 de taille (approximativement) égale.

Le modèle est entraîné et testé 5 fois ; à chaque fois t = 1, 2, 3, 4,5 il est entraîné sur D/Dt et testé sur Dt. Les 5 mêmes sous-ensembles en cross-validation sont utilisés chaque fois qu'une valeur ou, en cas de plus d'un paramètre, une combinaison de valeurs est essayée. Ceci est fait afin d'exclure les variations dues à la sélection aléatoire de sous-ensembles (2a, 3a).

L'évaluation du modèle fait partie intégrante du processus de développement du modèle. Cela aide à trouver le meilleur modèle qui représente nos données et à quel point le modèle choisi fonctionnera à l'avenir, Pour améliorer le modèle, nous ajustons les hyper-paramètres du modèle et essayons d'améliorer la précision et examinons également la matrice de confusion pour essayer d'augmenter le nombre de vrais positifs et de vrais négatifs(3b,3c).

4.2.1 Logiciels utilisés

Afin de réaliser notre travail, on s'est basé sur les outils présentés par la figure ci-dessous :

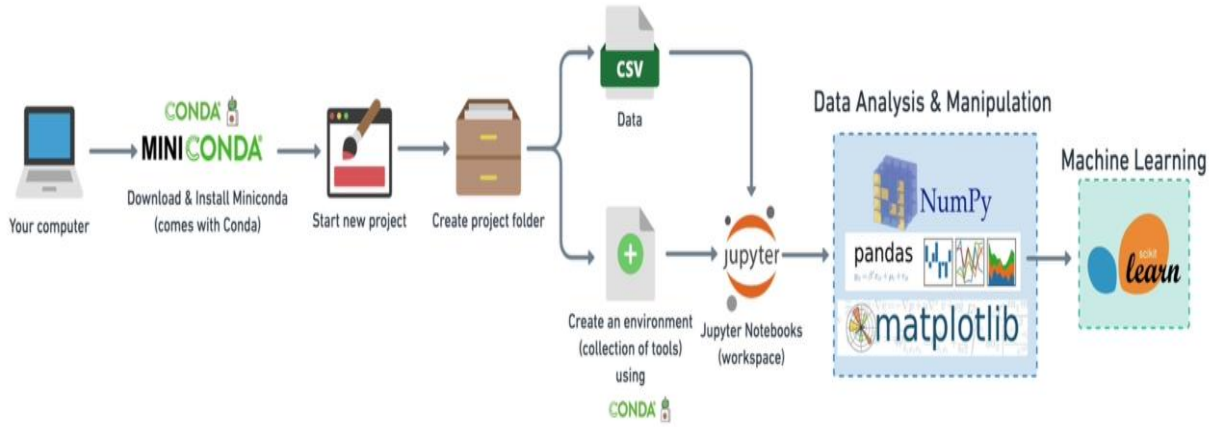

Figure 24 : Architecture technique de la solution.

4.2.1.1 MiniCONDA

Miniconda est un installateur minimal gratuit pour conda. Il s'agit d'une petite version bootstrap d'Anaconda qui inclut uniquement conda, Python, les packages dont ils dépendent et un petit nombre d'autres packages utiles, notamment pip, zlib et quelques autres. La commande "condainstall" est utilisée pour installer plus de 720 packages conda supplémentaires à partir du référentiel Anaconda. [89]

Pour créer un environnement avec des packages spécifiques on utilise cette commande :

```
conda create -n monenvjupyter pandas numpymatplotlibscikit-learn
```

4.2.1.2 Jupyter Notebook

Jupyter Notebook (anciennement IPython Notebooks) est un environnement de programmation interactif basé sur le Web permettant de créer des documents Jupyter Notebook. Le terme "notebook" peut faire référence à de nombreuses entités différentes, adaptées au contexte, telles que l'application web Jupyter, le serveur web Jupyter Python ou le format de document Jupyter.

Un document Jupyter Notebook est un document JSON. Il suit un schéma contenant une liste ordonnée de cellules d'entrée/sortie. Celles-ci peuvent contenir du code, du texte (à l'aide de Markdown), des formules mathématiques, des graphiques et des médias interactifs. Ce document se termine généralement par l'extension ".ipynb"."[90]

4.2.1.3 Pandas

Pandas est une bibliothèque écrite pour Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles.

Pandas est un logiciel libre sous licence BSD. Son nom est dérivé du terme "données de panel", un terme d'économétrie pour les jeux de données qui comprennent des observations sur plusieurs périodes de temps pour les mêmes individus. Son nom est également un jeu de mots sur l'expression "analyse de données Python".[91]

4.2.1.4 NumPy

NumPy est une bibliothèque pour le langage de programmation Python, ajoutant la prise en charge de grands tableaux et matrices multidimensionnels, ainsi qu'une vaste collection de fonctions mathématiques de haut niveau pour opérer sur ces tableaux. NumPy est un logiciel open source et compte de nombreux contributeurs.[92]

4.2.1.5 Matplotlib

Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy. Matplotlib est distribuée librement et gratuitement sous une licence de style BSD. [93]

4.2.1.6 Scikit-Learn

Scikit-learn (anciennement scikits.learn et également connu sous le nom de sklearn) est une bibliothèque logicielle gratuite d'apprentissage automatique pour le langage de programmation Python Elle comporte divers algorithmes de classifications, de régressions et de clustering, notamment les Support Vector Machines, les Randomforests, l'amplification de gradient et k-means. Elle est conçue pour interagir avec les bibliothèques numériques et scientifiques Python NumPy et SciPy. [94]

4.2.1.7 Keras

Keras est une API d'apprentissage en profondeur écrite en Python, s'exécutant sur la plate-forme d'apprentissage automatique TensorFlow . Il a été développé dans le but de permettre une expérimentation rapide. Pouvoir passer de l'idée au résultat le plus rapidement possible est essentiel pour faire de bonnes recherches. [95]

4.3 Modèles de Machine Learning Utilisés

Dans ce qui suit, nous décrirons en détail les différents algorithmes employés lors de notre apprentissage ainsi que les résultats obtenus. A la suite des différents articles consultés dont les thématiques concernent l'application du Machine Learning aux profilage des clients dans le secteur bancaire, nous avons décidé, après les avoir comparés et désigné ceux qui sont les plus adaptés à notre problématique, et de choisir six modèles à appliquer sur nos données :

- LogisticRegression
- DecisionTree.
- Random Forest
- Support Vector Machine ‘’SVM‘’
- K-NearestNeighbours ‘’KNN’’
- Neural Network ‘’NN’’

4.3.1 Logistic Regression

Le modèle logistique est un modèle utilisé pour modéliser la probabilité d'une certaine classe ou d'un événement existant tel que réussite/échec, victoire/perte pour résumé une variable binaire 0 ou 1 donc on peut appliquer le modèle sur notre cas vu que l’évènement qu’on veut prédire (0 non faillite, 1 faillite) avec toutes les variables nécessaires. Lors de

l'apprentissage le modèle va chercher à calculer les coefficients de l'hyper droite pour arriver à l'output qu'on fixe.

Dans notre étude nous avons 52 variables d'entrée on décrit donc cette « hyper-droite » comme suit :

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta 52 X52 \quad (4)$$

X_i : i^{eme} variable explicative, dans notre exemple, c'est une colonne qui contient un taux financier « Marge brute d'exploitation, Rotation du capital de travail, Le ratio courant … »

β_i : i^{eme} coefficient directeur de l'hyper-droite associé à la i^{eme} variable explicative et β_0 l'ordonnée à l'origine. Ici, on peut interpréter β_i comme une mesure de l'importance donnée à X_i dans la classification : plus ce coefficient est élevé, plus X_i joue un rôle important dans l'output du modèle.

y : la variable expliquée, ici il s'agit de l'Output « Faillite » Pour transformer le nombre que l'hyper droite fournie en une classification, on utilise une fonction que l'on nomme fonction sigmoïde $\frac{1}{1+e^{-t}}$ et qui a la propriété intéressante de transformer les nombres passés à l'intérieur en nombres entre 0 et 1.

Rappelons qu'une probabilité est un nombre entre 0 et 1, un point important en classification est qu'on cherche à estimer la probabilité d'appartenance à chaque classe pour ainsi être en mesure de classifier chaque nouvelle observation dans la classe associée à la probabilité la plus forte.

L'idée principale de la régression logistique est de se servir de la fonction sigmoïde pour transformer le nombre que donne l'hyper-droite en une probabilité de se retrouver en faillite. Ainsi si cette probabilité est de l'ordre supérieure à 0.7 donc l'entreprise a 70% va être en faillite, notre objective est donc de donner une estimation « un pourcentage » que l'entreprise cliente se retrouve en faillite afin que l'analyste prenne en considération ce taux et décide quelle décision prendre selon la situation.

Les trois principaux paramètres de ce modèle sont :

- C : float, default=1.0 ; Inverse de la force de régularisation ; doit être un flottant positif. Comme dans les machines à vecteurs de support, des valeurs plus petites indiquent une régularisation plus forte. (Régularisation : est le processus qui

régularise ou réduit les coefficients vers zéro, elle décourage l'apprentissage d'un modèle plus complexe ou plus flexible, afin d'éviter l'overfitting.)

- max_iter : int, default=100 ; Nombre maximal d'itérations nécessaires pour que les solveurs convergent.
- Solver : {'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'}, default='lbfgs' ; Algorithme à utiliser pour le problème d'optimisation.

[96]

Après l'ajustement des hyperparamètres a l'aide de la classe GridSearchCV de Scikit-learn et l'entrainement du modèle avec l'ensemble d'apprentissage, Le résultat obtenu sur les données d'essai est de 0.9304 donc notre modèle est d'une précision de 93.04%.

Les figure ci-dessous présente la matrice de confusion de la régression logistique

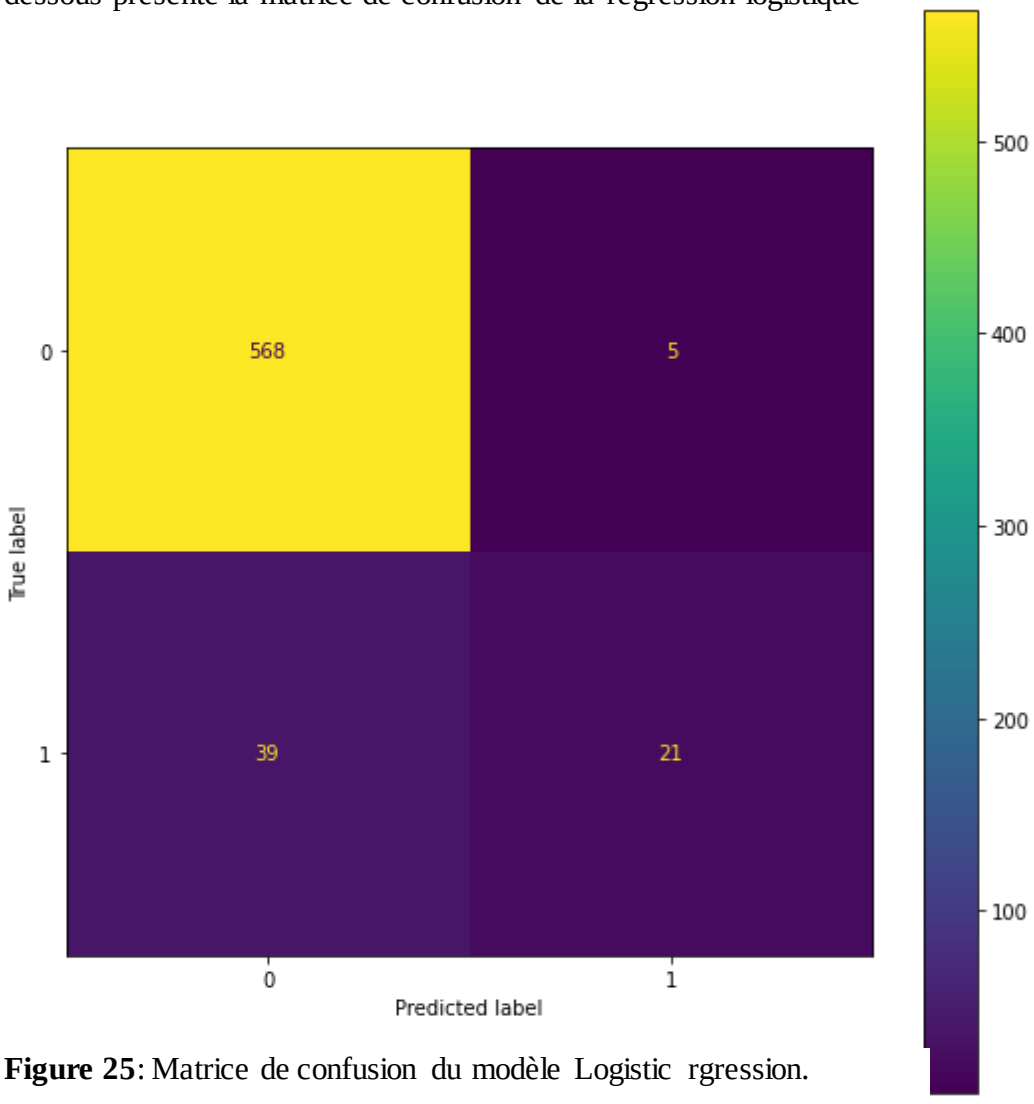

Figure 25: Matrice de confusion du modèle Logistic rgression.

Vu que le modèle de régression logistique contient des coefficients pour constituer l'hyperdroite, ces coefficients représentent l'importance donnée à X_i dans la classification. On peut voir dans Figure 26 les coefficients qu'ont été obtenus après apprentissage du modèle.

A travers ces coefficients, nous arrivons à mieux appréhender la manière dont le modèle doit procéder par rapport à l'analyse de chaque indicateur financier. En effet, chaque variable est dotée d'un coefficient positif ou négatif. Un coefficient négatif implique que cet indicateur fera converger la sortie vers (0)et donc la non-banqueroute de l'entreprise,Tandis qu'un coefficient positif signifie que cet indicateur fera converger le résultat vers (1) donc vers une situation de banqueroute de l'entreprise.

La figure suivante montre le Diagramme des coefficients d'importance de la régression logistique.

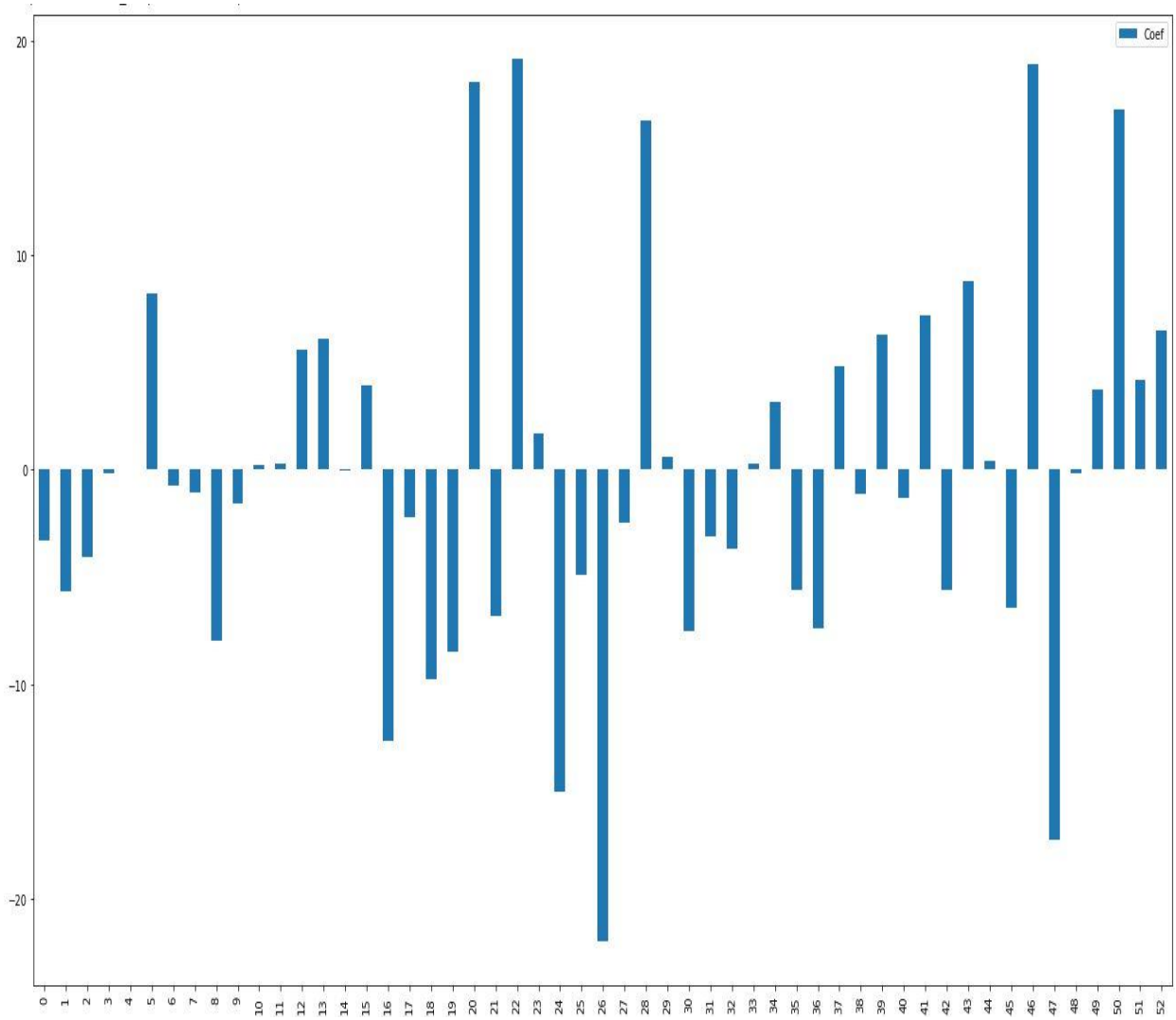

Figure 26: Diagramme des coefficients d'importance de la régression logistique.

4.3.2 DecisionTree

Les arbres de décision sont une méthode d'apprentissage supervisé non paramétrique utilisée pour la classification et la régression. Leur objectif est de crée un modèle qui prédit la valeur d'une variable cible en apprenant des règles de décision simples déduites des caractéristiques des données, d'où vient l'idée de tester cette méthode sur notre DataSet. Cet algorithme va prendre nos variables d'entrée (les taux et indicateurs financiers) pour en constituer des combinaisons de variables qui correspondent aux embranchements qui vont mener aux feuilles qui contient les valeurs de la variable cible (banqueroute).

Les principaux paramètres du DescisionTree sont :

- max_depth : int, default=None ; La profondeur maximale de l'arbre. Si None, alors les nœuds sont développés jusqu'à ce que toutes les feuilles soient pures ou jusqu'à ce que toutes les feuilles contiennent moins de min_samples_split échantillons.
- min_samples_leaf : int or float, default=1 ; Le nombre minimum d'échantillons requis pour être à un nœud feuille. Un point de séparation à n'importe quelle profondeur ne sera considéré que s'il laisse au moins min_samples_leaf échantillons d'entraînement dans chacune des branches gauches et droite. Cela peut avoir pour effet de lisser le modèle, en particulier dans la régression.

[97]

Suite au choix des hyperparamètres optimaux, et au trainig du modèle avec les données d'apprentissage, nous obtenons un résultat 0.9225 qui indique une précision de 92,25% pour ce modèle

La figure ci-dessous présente la matrice de confusion de l'arbre de décision

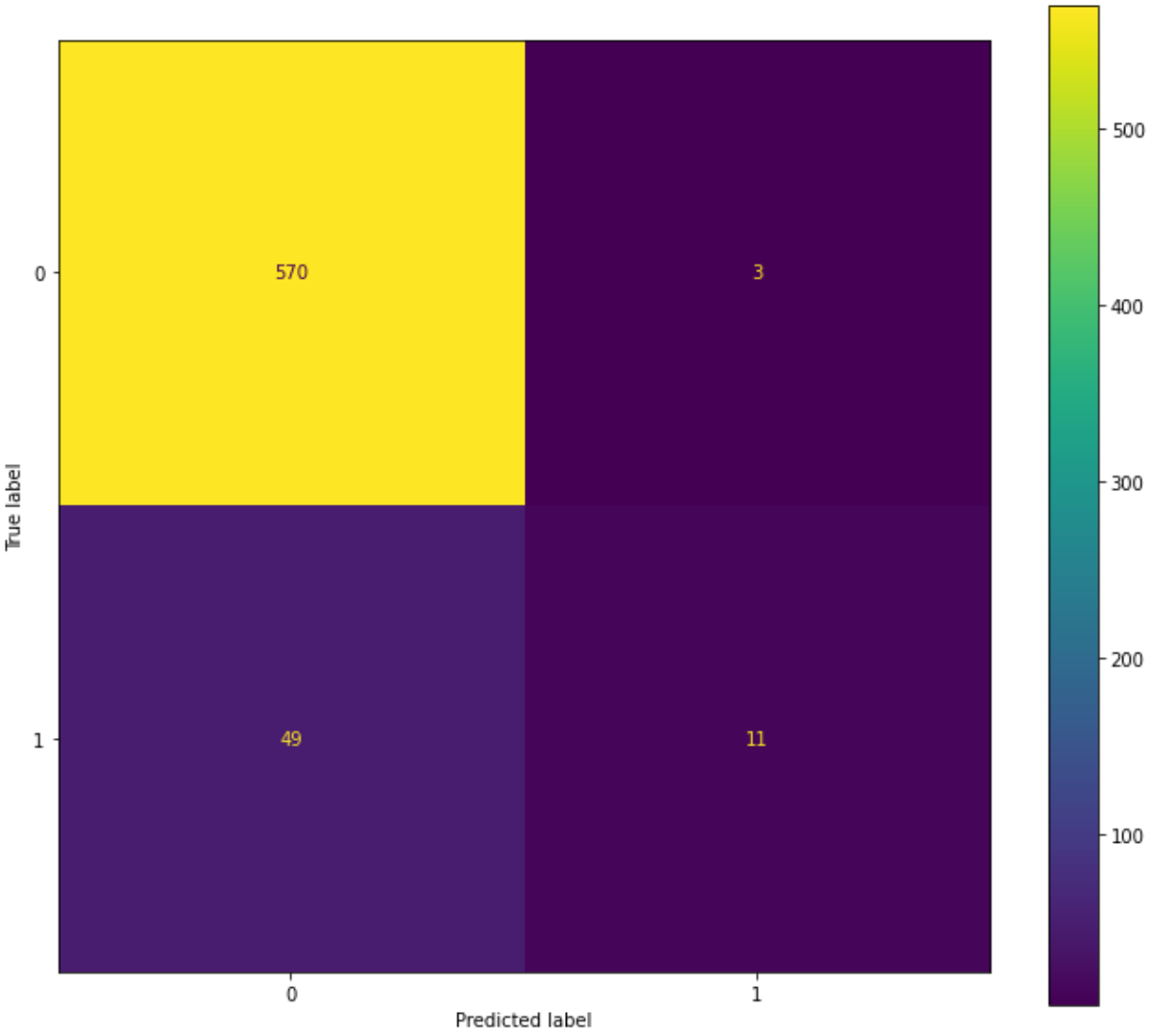

Figure 27 : Matrice de confusion du modèle DecisionTree

En vue de mieux appréhender le comportement du modèle dans l'analyse de chaque indicateur financier, et l'importance accordée à chaque variable afin de pouvoir la juger avec le soutien d'un analyste risque, nous avons utilisé l'attribut « DecisionTree.feature_importances_ » afin de faciliter leur interprétation sur des visuels(Figure 28)

A travers les visualisations obtenues, nous pouvons remarquer que le modèle n'a pas considéré toutes les variables (Juste 16 variables) et a donc jugé certaines d'entre elles comme étant sans importance pour la décision finale.

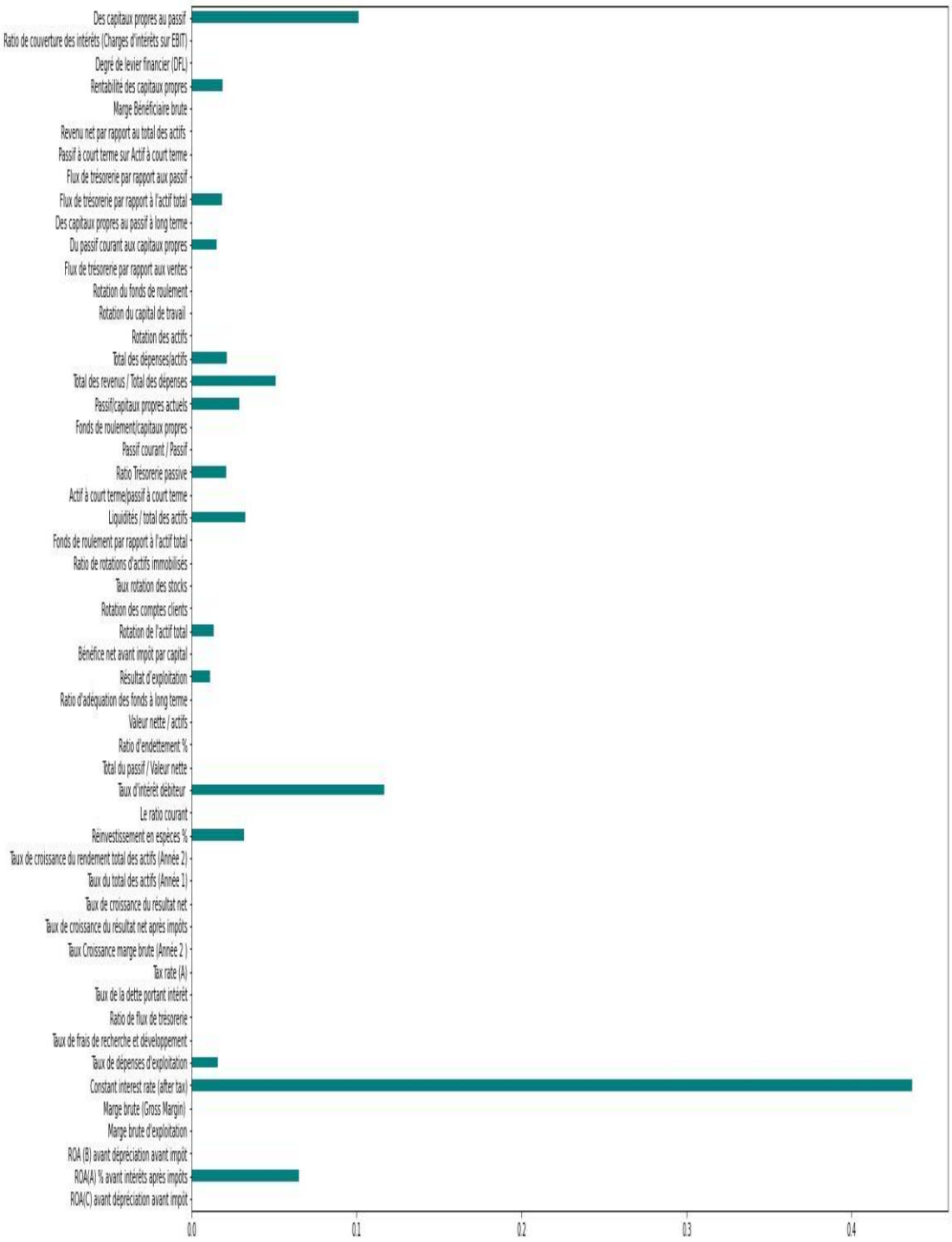

Figure 28: Diagramme d'importances des variables du modèle DecisionTree

4.3.3 Random Forest

Si nous choisissons de recourir au modèle Random Forest, tout simplement parce que dans le but d'avoir une prédiction optimale, il est sûrement nécessaire de faire plusieurs exécutions sur les données et pas qu'une seule, ce qui est le cas du Random Forest qui lui, fera en sorte de faire exécuter à plusieurs reprises l'algorithme de l'arbre de décision en utilisant à chaque fois un sous-ensemble différent de données.

Pour parvenir à des résultats optimaux, nous utiliserons la validation croisée afin de déterminer les meilleurs paramètres' permettant d'obtenir un score optimal.

Les hyperparamètres utilisés sont :

- n_estimators : int, default=100. Le nombres des arbres dans la forêt. • max_depth: int, default=None. La profondeur maximale de l'arbre. Si None, alors les nœuds sont développés jusqu'à ce que toutes les feuilles soient pures ou jusqu'à ce que toutes les feuilles contiennent moins de min_samples_split échantillons.
- min_samples_leaf : int ou float, default=1. Le nombre minimum d'échantillons requis pour être à un nœud feuille. Un point de séparation à n'importe quelle profondeur ne sera pris en compte que s'il laisse au moins min_samples_leaf échantillons de formation dans chacune des branches gauches et droite. Cela peut avoir pour effet de lisser le modèle, en particulier dans la régression.[98]

Après détermination des valeurs exactes des paramètres pour le score optimal en utilisant la Classe GridSearchCV,Nous débutons le Training de notre modèle,puis, nous calculons le score en utilisant la fonction score() celui-ci est égale à une précision de 94,47% .

La figure ci-dessous présente la matrice de confusion du Random Forest

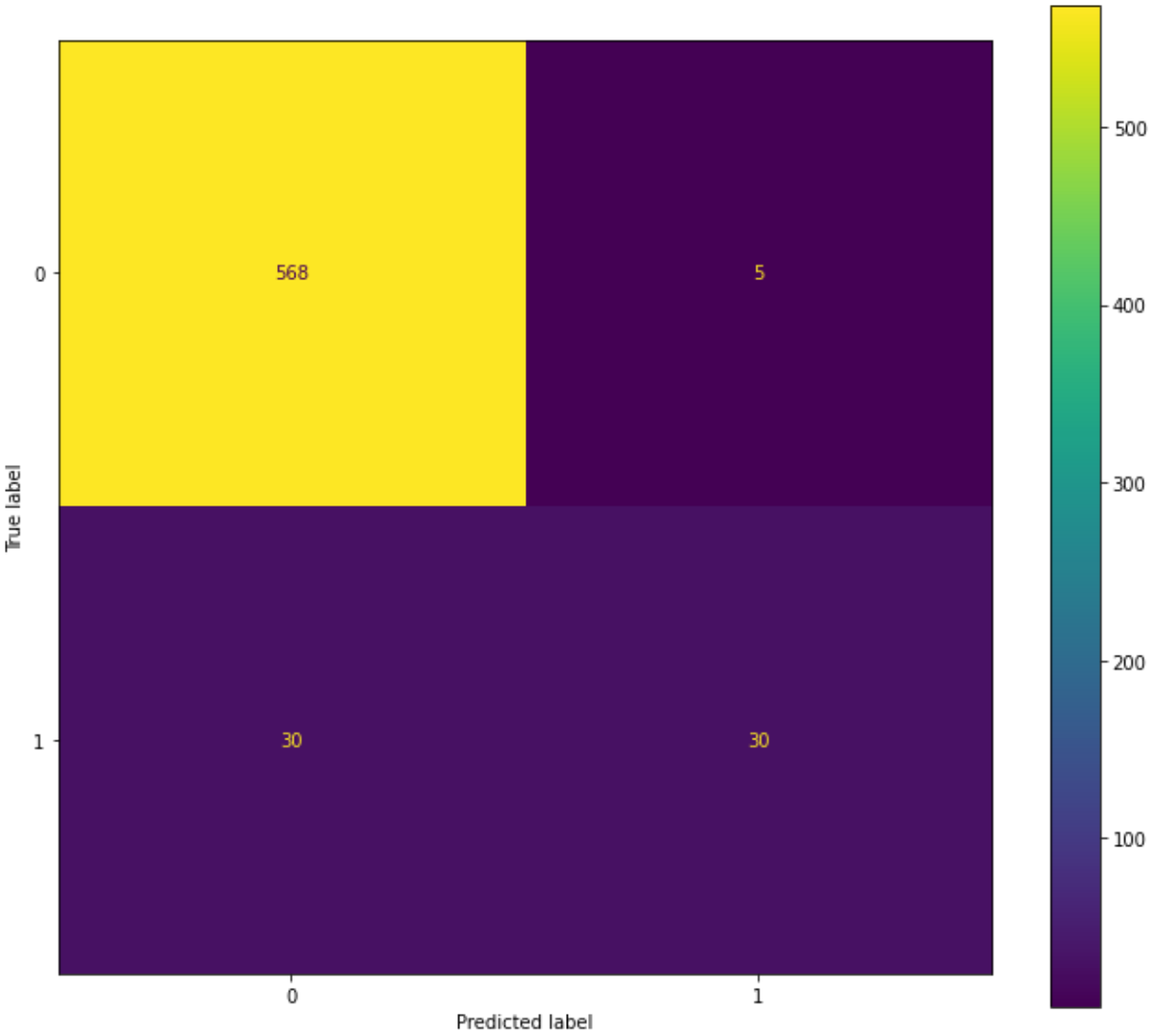

Figure 29: Matrice de confusion du modèle Random Forest.

Similairement au modèle précèdent le Random Forest utilise l'attribut feauture_importance pour appréhender le comportement du modèle dans l'analyse de chaque indicateur financier,

Nous avons utilisé la bibliothèque **matplotplib** afin de faciliter les interprétations sur des visuels.

A travers les visualisations obtenues(Figure 30) nous pouvons remarquer que le modèle a pris en considération toutes les variables mais chacun avec une importance différente des autres.

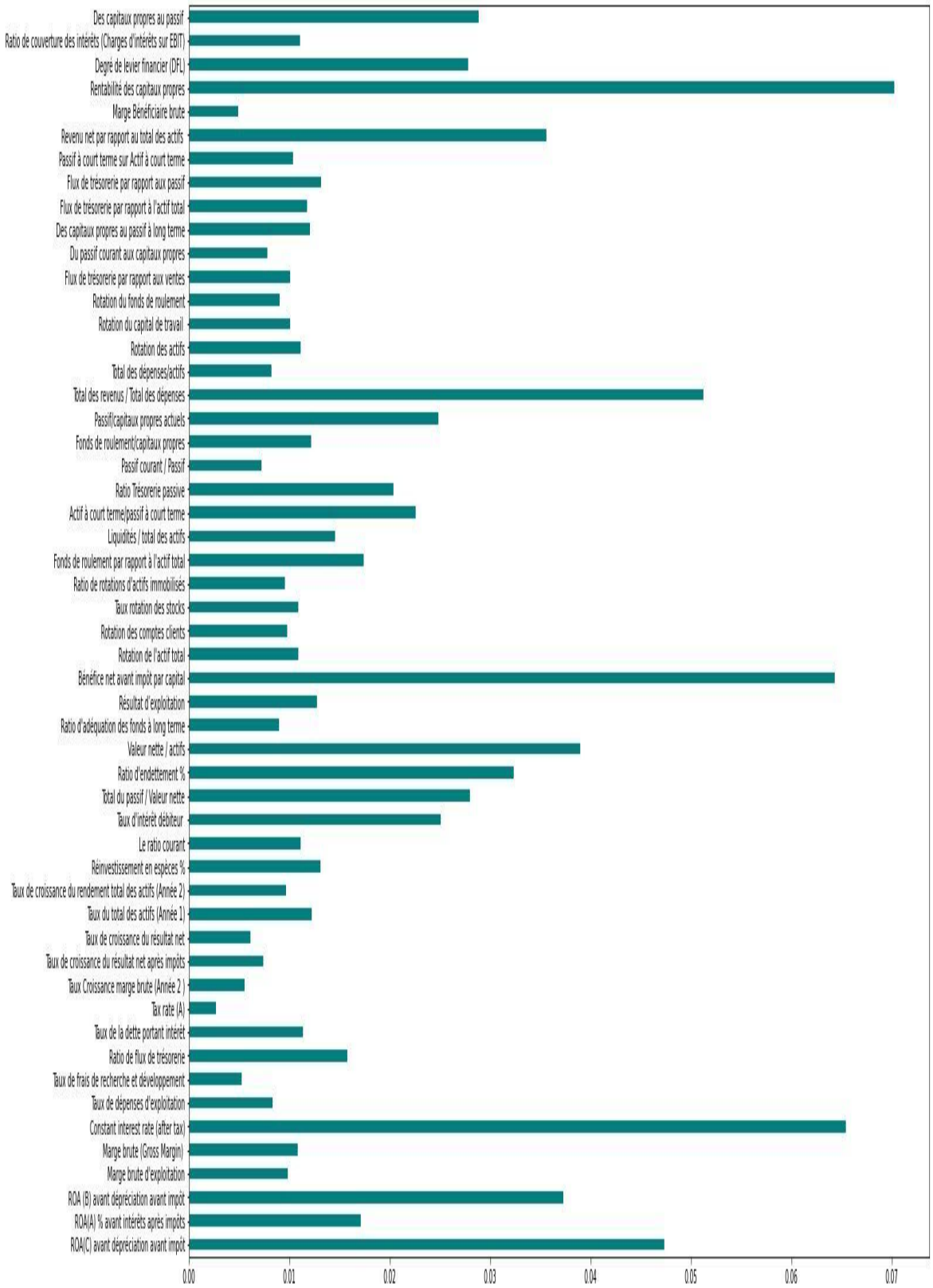

Figure 30 : Diagramme d'importance des variables du modèle Random Forest.

4.3.4 K-NearestNeighbours (KNN)

Le KNN nous permet d'estimer la classe d'une nouvelle donnée à partir de la classe majoritaire des k données les plus proche dans son voisinage. Pour ce modèle il existe un seul paramètre à fixer qui est k, le nombre de voisins à considérer.

La façon de procéder afin de déterminer les paramètres les plus optimaux va être la même suivait dans le modèle précédent. La fonction du modèle dans la bibliothèque Scikit-learn est : KNeighborsClassifier ().

Pour parvenir à des résultats optimaux, nous utiliserons la validation croisée afin de déterminer le paramètre le plus optimal permettant d'obtenir un score optimal. Les hyperparamètres utilisés sont :

• n_neighbors : int, default=5 ;Nombre de voisins à utiliser par défaut pour les requêtes kneighbors, comme les modèles précédents, nous avons passé par la cross validation pour déterminer la meilleure valeur du n_neighbors.[99]

Suite au choix du hyperparamètre optimal, Nous débutons le training de notre modèle via la fonction model.fit() ; puis, nous calculons le score en utilisant la fonction score(X_test,y_test) ; celui-ci est égale à 0.9289.

La figure ci-dessous présente la matrice de confusion du KNN

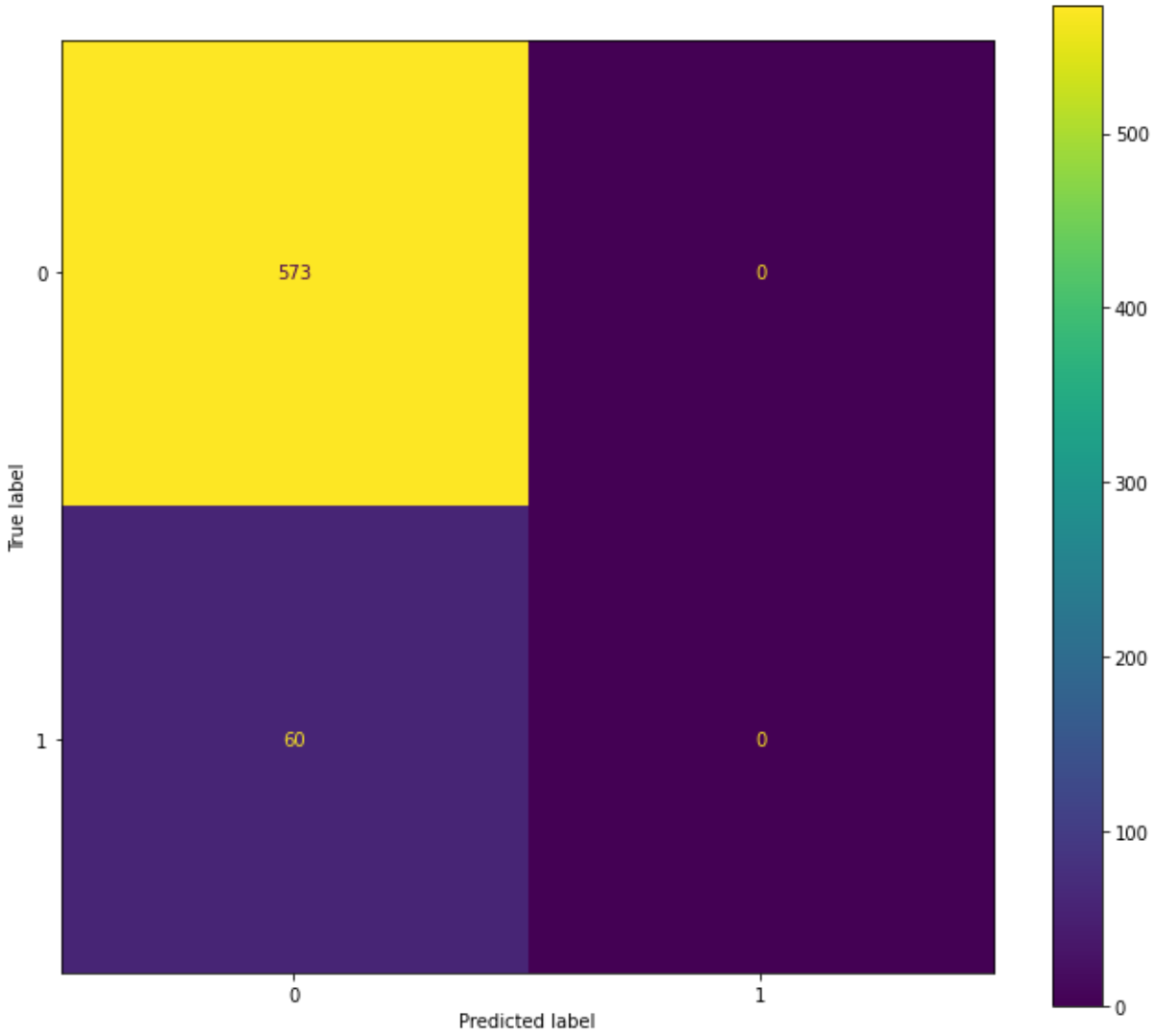

Figure 31 :Matrice de confusion du modèle KNN

En vue de mieux appréhender le comportement du modèle dans l'analyse de chaque indicateur financier, et l'importance accordée à chaque variable afin de pouvoir la juger avec le soutien d'un analyste risque,nous avons utilisé la fonction feature_mean_ et la bibliothèque **matplotlib** afin de faciliter leur interprétation sur des visuels.

La figure ci-dessous présente le Diagramme d'importance des variables du modèle KNN

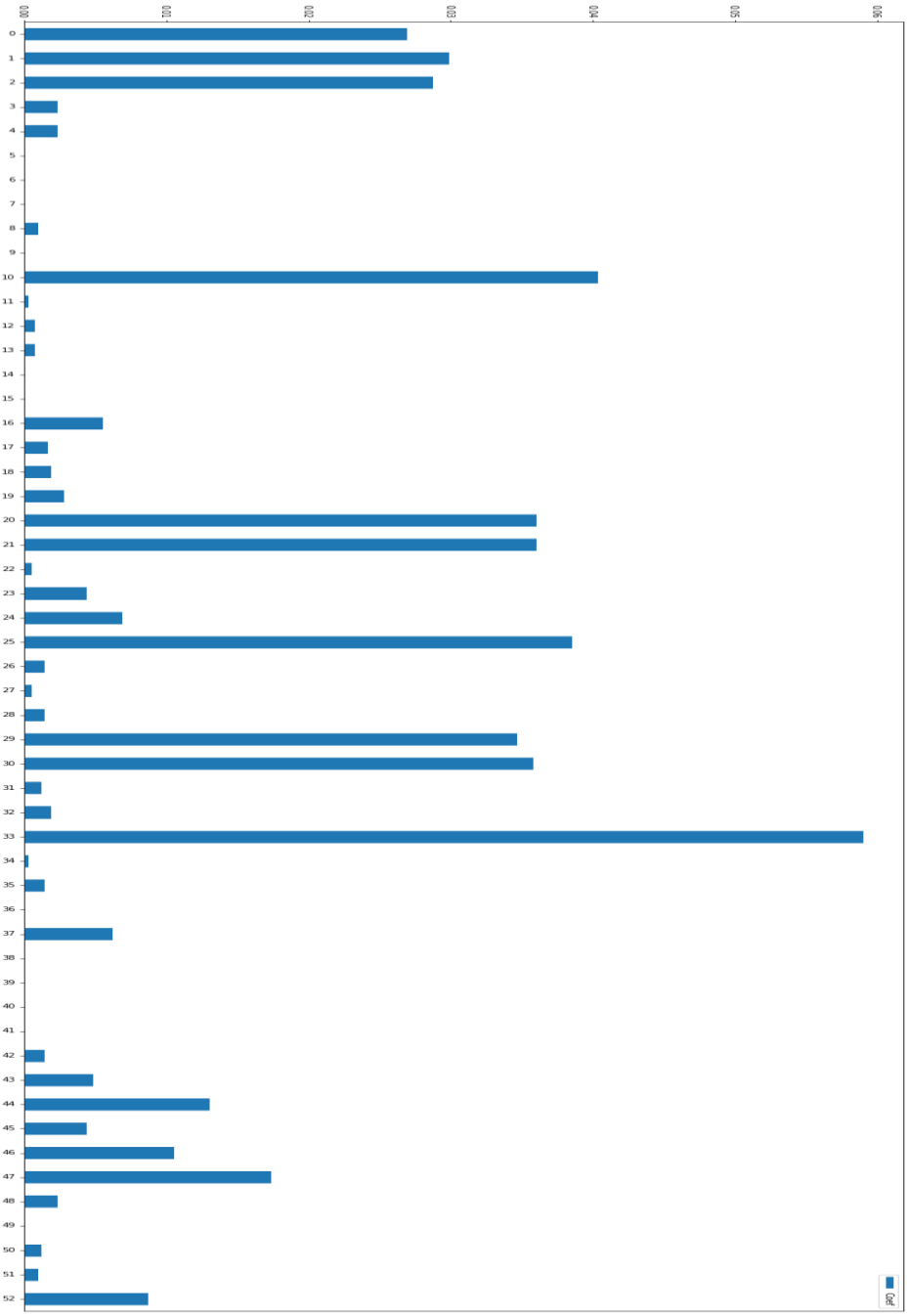

Figure 32 : Diagramme d'importance des variables du modèle KNN

4.3.5 Support Vector Machine

Suivant [16], le modèle d'optimisation SVM est basé sur la transformation d'une fonction mathématique par une autre fonction, appelée "Kernel", par laquelle on identifie la plus grande distance entre les observations les plus proches et opposées classifié.

Un critère commun est de savoir si les groupes sont complètement séparables, car cela permettrait au SVM de construire un modèle avec une précision de 100 %. En finance, cela

est pratiquement impossible car les variables économiques sont influencées par le bruit des données empiriques et sont souvent biaisées. Pour les problèmes de classification impliquant des groupes partiellement séparables, la méthode SVM permet d'inclure une marge d'Erreur.

La façon de procéder afin de déterminer les paramètres les plus optimaux va être la même suivait dans les modèles précédents.

Les hyperparamètres utilisés sont :

- $\boldsymbol{C}$: float, default=1.0 ; Paramètre de régularisation. La force de la régularisation est inversement proportionnelle à C. Doit être strictement positif. Il contrôle le compromis entre la maximisation de la marge et la minimisation de l'erreur de reconstruction. Une grande valeur de C implique une marge faible mais une erreur de classification moins importante tandis que dans le cas contraire l'inverse se produit. Si C'est trop grand, le modèle va sur-apprendre et donc mal classer les données de test, une valeur trop petite va quant à elle entrainer un mauvais apprentissage de l'algorithme.
- Kernel : {'linear', 'poly', 'rbf', 'sigmoïde', 'precomputed'}, default='rbf' ; Spécifie le type de noyau à utiliser dans l'algorithme.
- $\boldsymbol{gamma}$: Il est utilisé pour adapter l'hyperplan aux données et est responsable de son degré de linéarité, c'est pour cela qu'il n'est pas utilisé dans le cas d'une fonction noyau à base linéaire. Plus gamma est petit, plus l'hyperplan aura l'air d'une ligne droite ; si au contraire, il est trop grand, l'hyperplan sera plus courbé et pourrait trop bien délimiter les données et conduire à du sur-apprentissage[100].

Après l'ajustement des hyperparamètres a l'aide de la classe GridSearchCV de Scikit-learn et l'entrainement du modèle avec l'ensemble d'apprentissage, Le résultat obtenu sur les données d'essai est de 0.9192 donc notre modèle est d'une précision de 91.46%.

La figure ci-dessous présente la matrice de confusion du SVM

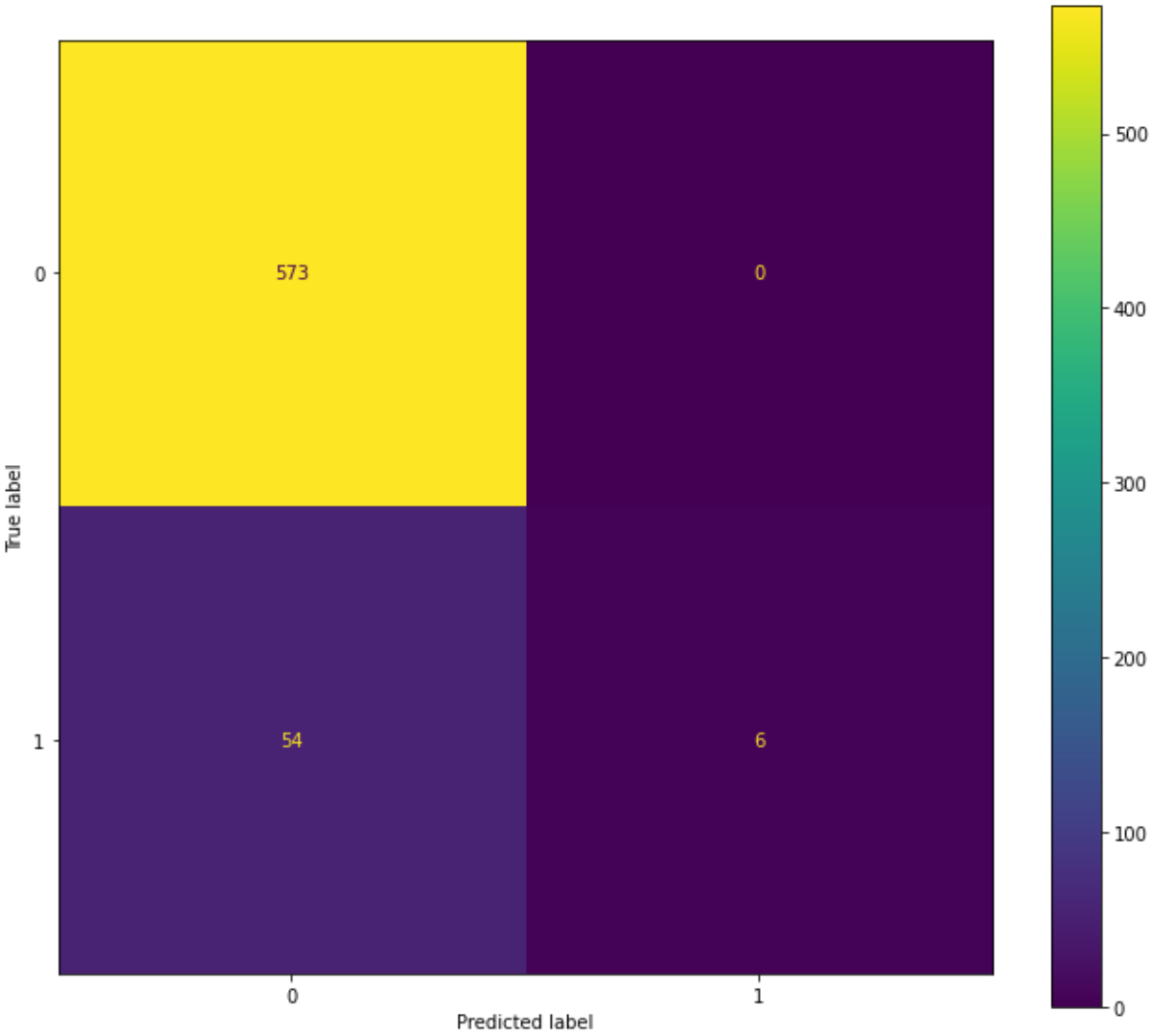

Figure 33 : Matrice de confusion du modèle SVM

En vue de mieux appréhender le comportement du modèle dans l'analyse de chaque indicateur financier, et l'importance accordée à chaque variable afin de pouvoir la juger avec le soutien d'un analyste risque, nous avons utilisé l'attribut **coef_**

La figure Suivante présente le Diagramme des coefficients d'importance du modèle SVM.

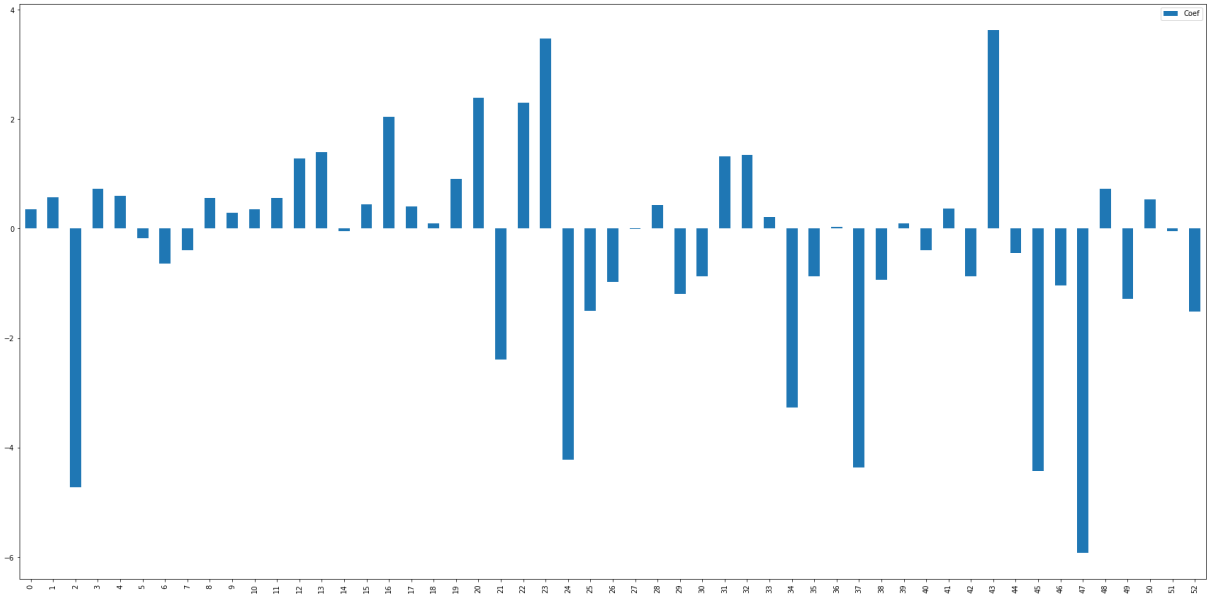

Figure 34: Diagramme des coefficients d'importance du modèle SVM

4.3.6 Réseaux de neurones

La méthode de NN n'est pas aussi simple que certaines des autres méthodes, comme décrit dans la section 3.2.6. Par conséquent, lors de l'ajustement du modèle, nous utilisons la logique proposée par [101]

Les hyperparamètres utilisés sont :

- Hyperparamètres du modèle qui influencent la sélection du modèle, tels que le nombre et la largeur des couches cachées
- Hyperparamètres d'algorithme qui influencent la vitesse et la qualité de l'algorithme d'apprentissage, tels que le taux d'apprentissage pour la descente de gradient stochastique (SGD)

La figure ci-dessous présente la matrice de confusion du CNN

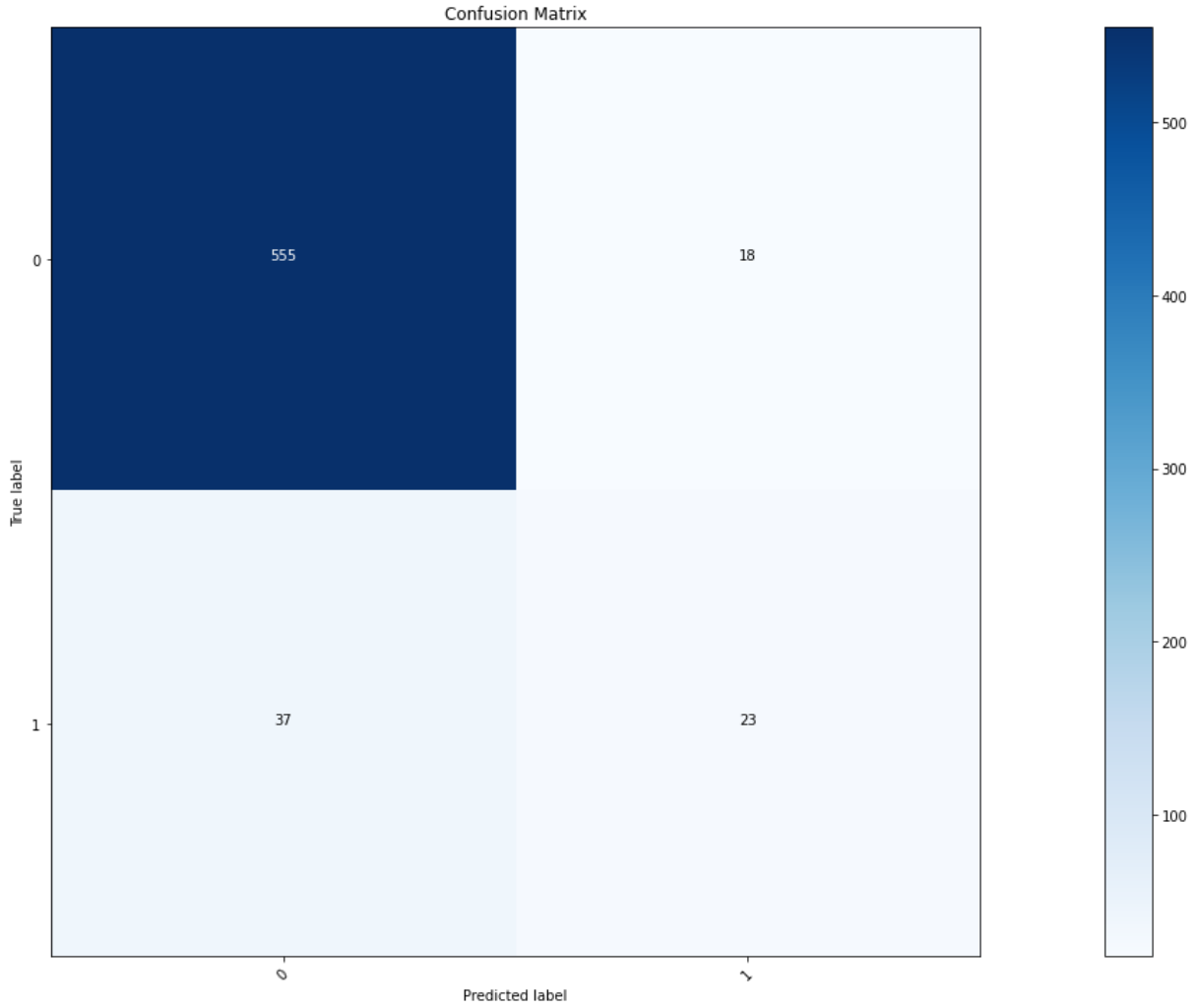

Figure 35 : Matrice de confusion du modèle CNN

4.4 Evaluation

A première vue, sur les matrices de confusion nous pouvons constater que les modèles utilisés aboutissent tous à des résultats bons, avec des valeurs de True positive et True négatives assez importante par rapport aux valuers du False positives et False négatives

ROC Curve

La figure suivante présente une visualisons des courbes ROC des différents modèles avec l'aire sous chaque courbe :

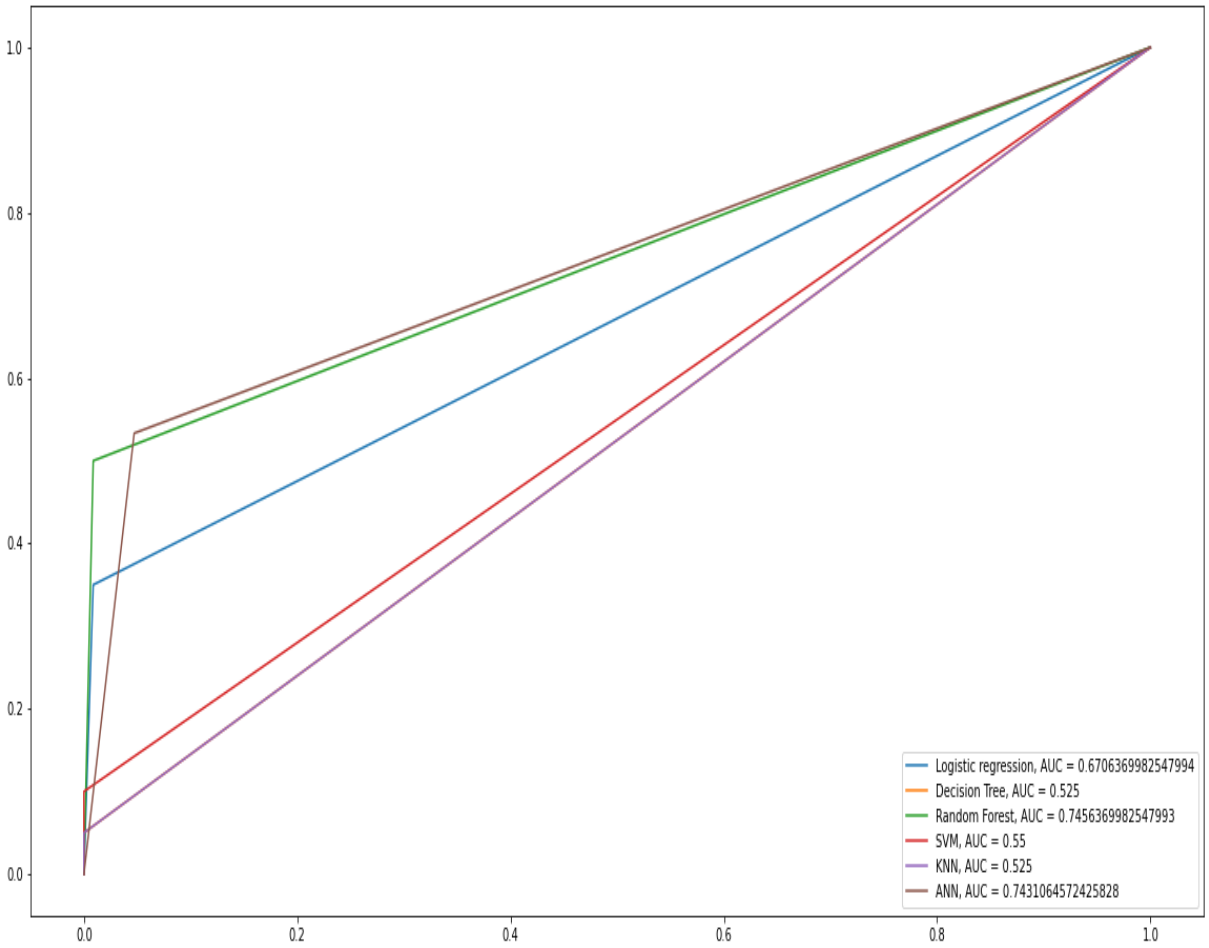

Figure 36 : ROC courbes des 6 modèles d'apprentissage utilisés.

Le tableau ci-dessous présente les indices de performance obtenus pour chaque modèle à partir des matrices de confusion précédentes :

Tableau 4 : Indices de performances obtenues pour chaque modèle.

Modèle	Précision	AUC
LogiscticRegression	93.04%	0.67
DecisionTree	92.25%	0.525
Random Forest	94.47%	0.746
SVM	91.09%	0.55
KNN	92%	0.525
CNN	91.92%	0.743

En observant la table 4, nous concluons que tous nos modèles fonctionnent bien en termes de précision, où la prédiction est stable bien au-dessus de 90 %. Cependant, il est plus logique d'optimiser l'AUROC au prix d'une précision moindre. Les valeurs AUROC présentées dans le tableau 5.10 montrent que les meilleurs modèles sont le réseau neuronal et la Random Forest, avec un AUROC de 0,743 et légèrement pire de 0,746, respectivement.

4.4.1. L'importance des variables entre les différents modèles

Dans ce qui suit, nous allons comparer les résultats de chaque modèle visà-vis l'importance des variables qui va contribuer dans le choix du modèle le plus réaliste : En ce qui concerne le modèle de Decision Tree où 16 variables ont contribuées dans l'output de ce dernier. Donc, ce résultat implique que le modèle n'est pas adaptable à la problématique. Par contre, si on modifie les paramètres, le comportement du modèle va changer, mais, la précision va être diminuée vu que les paramètres ne sont plus optimaux. Contrairement au reste des modèles où nous avons constaté des importances avec des taux différents : Pour les modèles : Régression logistique,SVM nous avons obtenu des importances avec des valeurs négatives et positives. Une importance négative signifie que cet indicateur contribue pour que la valeur de '' Faillite '' converge vers 0, c'est-à-dire une situation de non défaillance ; tandis qu'une importance positive signifie la contribution de la variable pour que l'output converge vers 1, c'est-à-dire une situation de banqueroute. Pour les modèles, Random Forest et KNN, Nous avons obtenu des importances à valeurs positive.

Pour le CNN nous ne sommes pas en mesure d'extraire ces informations du modèle.

De première vue on constate que les valeurs d'importance donnée a chaque variable du dateset utilisé diffère d'un modèle a un autre, ce qui montre que le choix du modèle optimal sera en fonction de la précision, AUC et une éventuelle intervention de l'analyste rique afin de determiner les valeurs d'importance les plus logiques en cas ou ce problème se produirait

Pour cette étude nous optons pour le Random Forest pour etre le modèle optimal ayant la meuilleur précision et meuilleur valeur de AUC, et des valeurs d'importances de variables assez logique.

4.5 Intégration de Blockchain

Nous avons fusionné la Blockchain avec notre base de données. En effet, nous avons crée une base de données qui est utilisée pour stocker les données sensibles de notre

blockchain (tels que le timestmp, data, hash, et previous_hash et nonce) Afin d'empêcher l'indisponibilité et assurer l'intégrité en dupliquant les données obtenues.

La BD est constituée de deux tables blockchain et users. La première table est utile pour stocker des données qui interagissent avec l'application. Tandis que la seconde se résume à introduire un utilisateur blockchain au système.

La Figure suivante montre Le MCD de cette BD

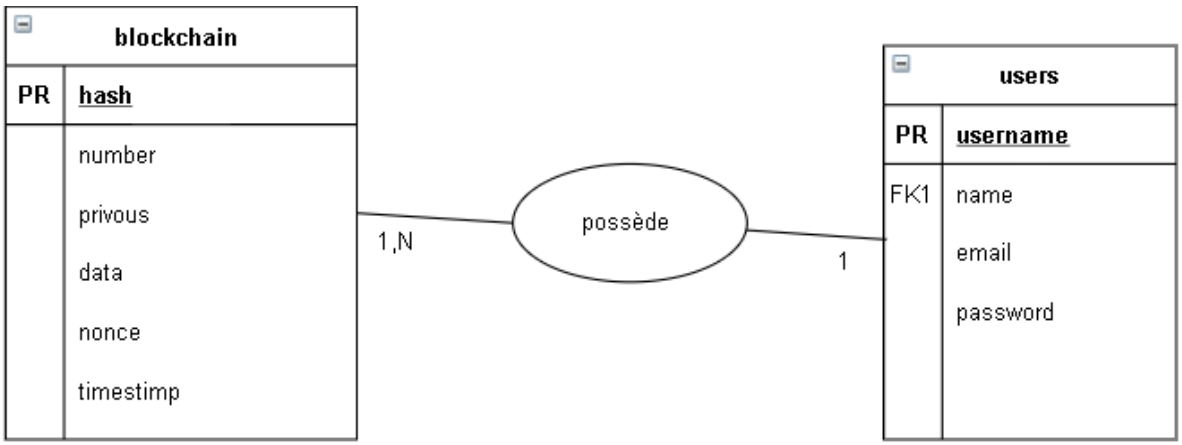

Figure 37: MCD de la BD crée

```
1 •  select * from  blockchain
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

number	hash	previous	data	nonce
1	f0201b56df5624fac321135acd03c7644a3d204...	0000000000000000000000000000000000000000...	BANK-->riyadh96-->100.0	101212
2	d5285325c79ee601f1596b6beef95d3e37594f5...	00002556183a1373d7c4b337f0094cfee17290c...	BANK-->riyadh96-->2000.0	10463
3	33fcaad1d51612e56d40af2ed8138532faefed2...	00009e589c586afd92e72e870f2cc2ccb40b25b...	BANK-->walid99-->200.0	74660
4	8a32241bc9c507478e828568a6a65a4b0d087e...	00006cd72825368b584447f9a9a97cc6d95a306...	walid99-->riyadh96-->50.0	8548
5	0000cbe20ccd5858eae57701ec5f73cbfeba648...	8a32241bc9c507478e828568a6a65a4b0d087e...	BANK-->riyadh96-->2000.0	59144
6	0000ffbf05a8c8c37b90fe21cb47bac506f21720...	0000cbe20ccd5858eae57701ec5f73cbfeba648...	riyadh96-->walid99-->1200.0	65551
7	000037e448ad7aa0569f102f98455ea7ceab0f3...	0000ffbf05a8c8c37b90fe21cb47bac506f21720...	riyadh96-->walid99-->2.0	52389
8	000050e51e2865e502dad8d19c2a060a601496...	000037e448ad7aa0569f102f98455ea7ceab0f3...	BANK-->riyadh96-->1000.0	46715
9	00003cd5457d88f3a93539f4548874ee99edf11...	000050e51e2865e502dad8d19c2a060a601496...	BANK-->CEVITAL-->1200000000.0	18239
10	000024f165e7a39437aff204b6d6fae5596a6a4...	00003cd5457d88f3a93539f4548874ee99edf11...	BANK-->KIA20-->20000000000000.0	34417
11	00000ad6012c054dd30c5cbaba6ca82c2cc7f7cf...	000024f165e7a39437aff204b6d6fae5596a6a4...	CEVITAL-->KIA20-->50000000.0	9223

Figure 38 : représentation d'une table qui contient 6 attributs d'un block.

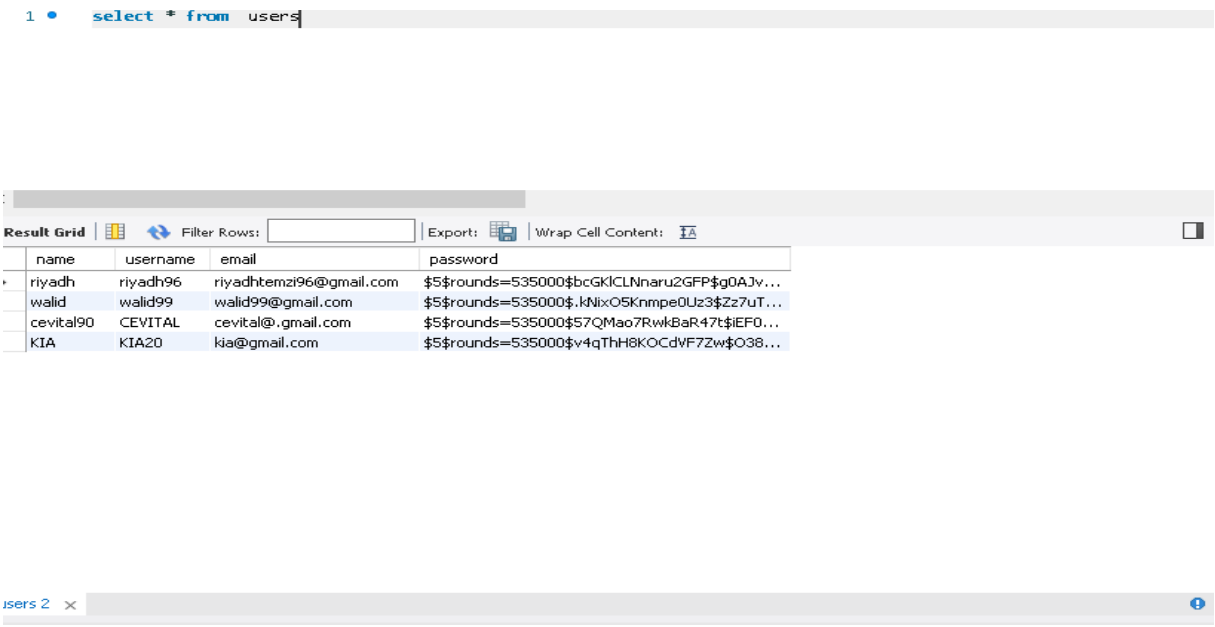

Figure 39 : Représentation des données d'un utilisateur BlockChain.

4.5.1 Les constitutions d'un block

Hash : Un hachage, comme un nonce ou une solution, est l'épine dorsale du réseau blockchain. Un hachage est développé en fonction des informations présentes dans l'en-tête de bloc. Fonctionnement d'un hachage Les fonctions de hachage typiques prennent des entrées de longueurs variables pour retourner des sorties d'une longueur fixe.

Previous hash: Dans une blockchain, le hachage d'un bloc précédent dans une séquence est le hachage précédent

Data: La structure de données blockchain est principalement basée sur un pointeur de hachage et implique le bloc comme structure de données principale. Les structures de données aident à l'organisation et au stockage des données de manière à ce qu'elles soient facilement accessibles et modifiées. D'une manière générale, la structure de données blockchain peut être décrite comme une liste de transactions liée, organisée en blocs.

Nonce: Un nonce est une abréviation pour « nombre utilisé une seule fois », qui, dans le contexte de l'extraction de crypto-monnaie, est un nombre ajouté à un bloc haché - ou crypté - dans une blockchain qui, lorsqu'elle est rehashed, répond aux restrictions de niveau de difficulté. Le nonce est le nombre que les mineurs blockchain résolvent.

Block number: Le numéro de bloc est le numéro de version du protocole réseau blockchain. Il est utilisé pour l'identification de la version du logiciel qui sera utile pour identifier les fonctionnalités ou les protocoles pris en charge

Timestamp: Timestamp dans blockchain peut être utilisé comme une preuve d'existence et il conserve la preuve de la notarisation. Le processus prouvera qu'un certain document existe, depuis une période de temps. Toute modification non authentifiée peut être détectée facilement sous ceci et à partir du moment où le document existe.

4.5.2 Analyse du résultat de cette implémentation

1- Analyse

Après une étudie générale des besoins des institutions financières de la blockchain en particulier dans les transactions quand il s'agit des différents types de crédit. Cependant le fonctionnement de la blockchain classe implémentée consiste à miner des blocks en se concentrantdes différentes données hash previous hash nonce ect…. (Qui sont créées au début de la classe) mais tout d'abord la fonction update hash est nécessaire pour la création des différents hash 256 les figures en dessous résument tout

```
#to use sha256 hash for the blockchain
from hashlib import sha256
import datetime

#it take  any number of arguments and produces a  hash of sha256  as a result
def updatehash(*args):
    hashing_text = ""; h = sha256()

    for arg in args:
        hashing_text += str(arg)

    h.update(hashing_text.encode('utf-8'))
    return h.hexdigest()

class Block():


    def __init__(self,number=0, previous_hash="0"*64, data=None, nonce=0,timestamp=0):
        self.data = data
        self.number = number
        self.previous_hash = previous_hash
        self.nonce = nonce
        self.timestamp = timestamp


    def hash(self):
        return updatehash(
            self.number,
```

Figure 40:la fonction update hash.

```
def mine(self, block):

    try: block.previous_hash = self.chain[-1].hash()
    except IndexError: pass

    while True:
        if block.hash()[:self.difficulty] == "0" * self.difficulty:
            self.add(block); break
        else:
            block.nonce += 1
```

Figure 41 : la fonction mine Block.

En revanche, la classe sqlhelpers se résume en transformant les données blockchain en données SQL. Pour cela, les fonctions get_blockchain et sync_blockchain sont de bons exemples pour ce type de transfert.

Par contre, la fonction send_money est obligatoire afin d'ajouter des transactions au blocks, plus précisément les figures en dessous sont conçues dans un but précis.

```
#get the blockchain from mysql and convert to Blockchain object
def get_blockchain():
    blockchain = Blockchain()
    blockchain_sql = Table("blockchain", "number", "hash", "previous", "data", "nonce","timestamp")
    for b in blockchain_sql.getall():
        blockchain.add(Block(int(b.get('number')), b.get('previous'), b.get('data'), int(b.get('nonce')),b.

    return blockchain

#update blockchain in mysql table
def sync_blockchain(blockchain):
    blockchain_sql = Table("blockchain", "number", "hash", "previous", "data", "nonce","timestamp")
    blockchain_sql.deleteall()

    for block in blockchain.chain:
        blockchain_sql.insert(str(block.number), block.hash(), block.previous_hash, block.data, block.nonce,
```

Figure 42:les fonction sysnc_blockchain et getblockchain

```
#send money from one user to another
def send_money(sender, recipient, amount):
    #verify that the amount is an integer or floating value
    try: amount = float(amount)
    except ValueError:
        raise InvalidTransactionException("la Transaction est invalid .")

    #verify that the user has enough money to send (exception if it is the BANK)
    if amount > get_balance(sender) and sender != "BANK":
        raise InsufficientFundsException("Fonds insuffisants.")

    #verify that the user is not sending money to themselves or amount is less than or 0
    elif sender == recipient or amount <= 0.00:
        raise InvalidTransactionException("la Transaction est invalid.")

    #verify that the recipient exists
    elif isnewuser(recipient):
        raise InvalidTransactionException("L'utilisateur n'existe pas.")

    #update the blockchain and sync to mysql
    blockchain = get_blockchain()
```

Figure 43: la fonction send money

Tandis que diffuser des transactions, lancer le server flask, entrer dans une session qui est déjà créé et connecter à notre bd sont les piliers de la classe app

```
#import flask dependencie
from flask import Flask, render_template, flash, redirect, url_for, session, request, logging
from passlib.hash import sha256_crypt
from flask_mysqldb import MySQL
from functools import wraps

#import other functions and classes
from sqlhelpers import *
from forms import *

#others dependencie
import time

#initialize the app
app = Flask(__name__)

#configure mysql
app.config['MYSQL_HOST'] = 'localhost'
app.config['MYSQL_USER'] = 'root'
app.config['MYSQL_PASSWORD'] = '1234'
app.config['MYSQL_DB'] = 'crypto'
app.config['MYSQL_CURSORCLASS'] = 'DictCursor'

#initialize mysql
mysql = MySQL(app)

#wrap to define if the user is currently logged in from session
```

Figure 44:connexion vers une base de données déjà crée.

2- Résultat

Dans cette section on a pu intégrer 2 exemples de clients CEVITALet KIA, Le premier envoie une somme de 5 milliards de dinars au second, cette immense opération est une transaction qui sera insérer à notre blockchain

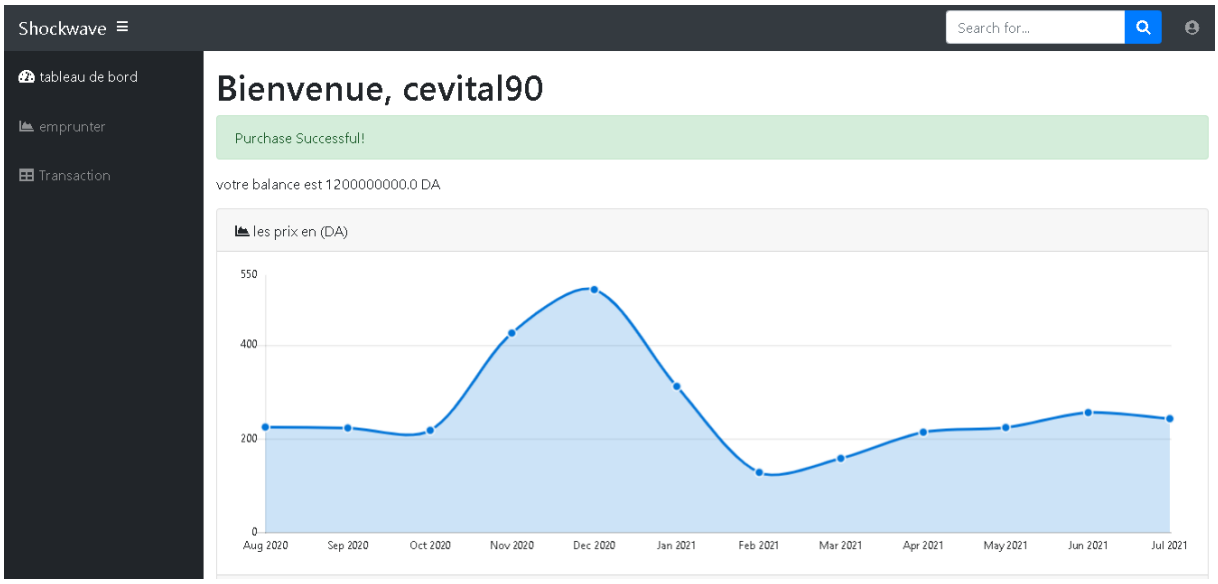

Figure 45: Le client cevital90 est introduit dans le système.

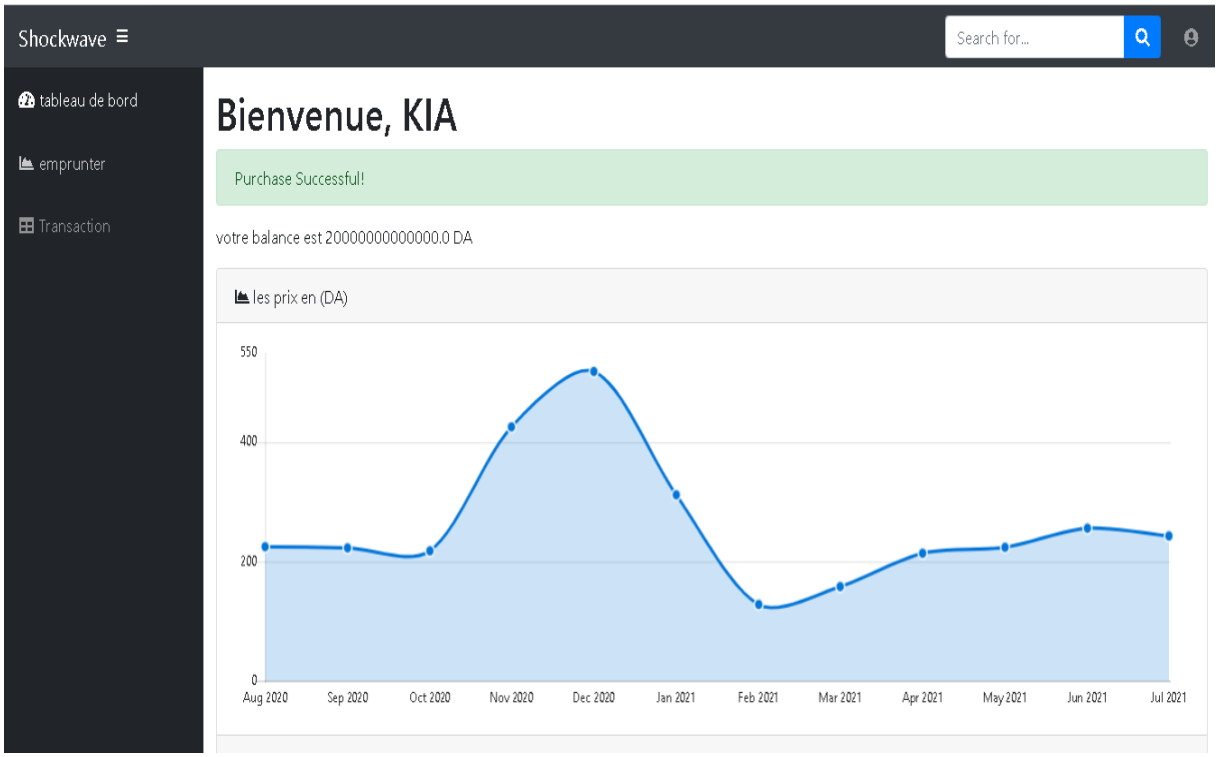

Figure 46: le client KIA est introduit dans le système.

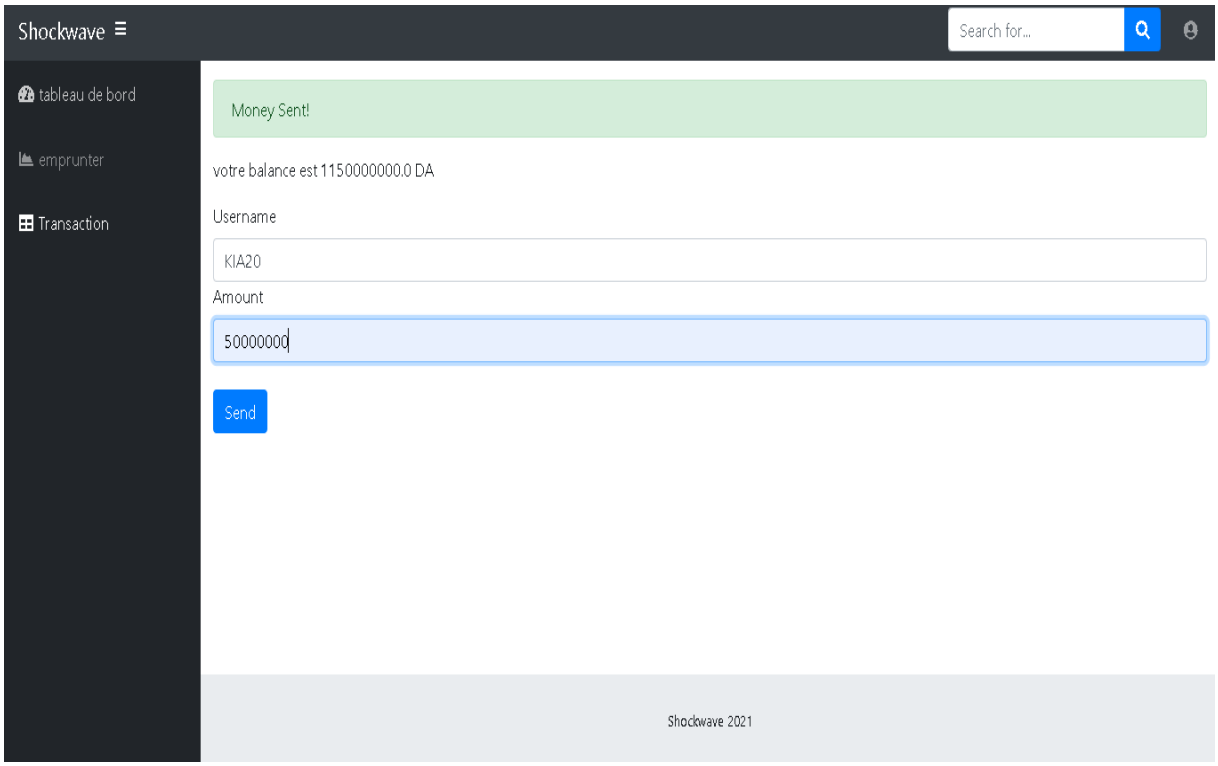

Figure 47 : le client civital envoie une somme depuis un utilisateur vers un autre.

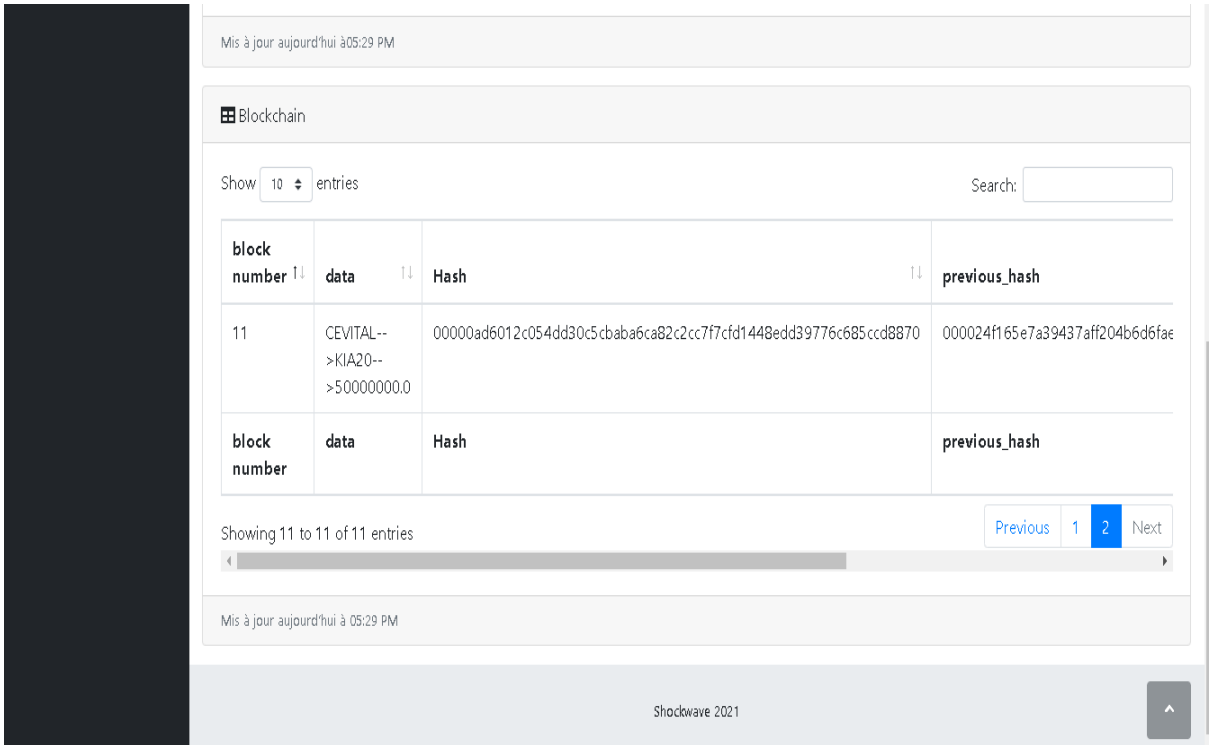

Figure 48: la transaction entre les 2 client dans un block.

Conclusion

Dans ce chapitre, notre projet atteint sa fin ? Au cours de ce dernier, nous avons décrit le processus de réalisation de nos solutions en présentant la méthode d'implémentation suivi par les différents outils de développement nécessaires à la mise en place de notre solution, par la suite nous avons présenté les résultats des modèles avec leur comparaison.

Conclusion Générale

Société Générale est particulièrement attachée à la mise en place d'une organisation rigoureuse et efficace de la gestion des risques dans l'ensemble de ses activités, marchés et régions d'intervention, et à l'équilibre entre une forte conscience des risques et la promotion de l'innovation. Cette gestion des risques, pilotée au plus haut niveau, s'effectue dans le respect des normes applicables.

Les difficultés de liquidité des entreprises et la détresse financière qui en découle constituent généralement un événement extrêmement coûteux et perturbateur pour les banques si ces derniers n'en seront pas capables de rembourser leurs crédits. Pour cette raison, cette étude tente de fournir un ensemble de caractéristiques qui peuvent aider à prédire la durabilité d'une entreprise. Cette étude implique la construction d'un système de prédiction financière qui, après avoir été entraîné sur un ensemble de comptes finaux historiques d'entreprises, les modèles construits sont ensuite utilisés pour évaluer la durabilité de l'entreprise et en seront capables d'évaluer la nature des données financières d'une autre entreprise.

Cependant, lors du processus de diagnostic que nous avons effectué, nous avons noté plusieurs points d'amélioration, dont le plus important, auprès de la banque, consiste dans le fait que l'analyse du risque d'un dossier de crédit, notamment l'analyse des informations financières, requiert beaucoup de temps, et cela ne correspond ni aux attentes ni aux objectifs de la banque.

En effet, cette contre-performance résulte du fait que le travail effectué par les analystes de risque est entièrement manuel lors de l'étude d'un dossier de crédit. Ces constats nous ont amené à aiguiller notre travail vers la mise en place d'un outil d'aide à la décision basé sur l'intelligence artificielle pour accompagner l'analyste risque dans sa tâche en vue d'apporter un souffle novateur à la démarche de gestion des risques au sein de SGA pour notamment faire en sorte d'améliorer les décisions prises par les analystes tant sur le plan de la précision que sur celui du gain de temps et ainsi répondre aux objectifs attendus par la banque.

Dans ce cadre, nous avons tout d'abord fait un recensement par une étude bibliographique de nombreux articles traitant la prédiction de faillite par l'apprentissage automatique, en vue de se faire une idée sur les solutions existantes ainsi que sur leur

efficacité. Puis nous avons réalisé une revue de littérature sur les différents concepts et techniques de l'intelligence artificielle, notamment le Machine Learning, en détaillant certains algorithmes, afin d'assimiler leur fonctionnement avant leur implémentation.

Après avoir récupéré la base de données existante et l'avoir traitée, nous nous sommes appliqués à exploiter cette dernière pour appliquer différents algorithmes d'apprentissage automatique capables de distinguer les clients qui seraient en faillite de ceux qui ne le seraient pas. Nous les avons évalués et comparés pour finalement sélectionner l'algorithme le plus performant.

Après nous avons créer une base de données afin d'intégrer des clients pour leur attribuer des trasactions et les enregistrer dans des blocs.

En conclusion, les algorithmes d'apprentissage automatique peuvent être utilisés comme un outil complémentaire pour la prédiction de la détresse financière. Toutefois, l'évaluation de la santé financière d'une entreprise en se basant uniquement sur les résultats des algorithmes d'apprentissage pourrait être trompeuse ; il convient donc de souligner que l'évaluation doit être réalisée en faisant appel à la collaboration du jugement humain et des méthodes de prédiction.

Axes d'améliorations

En ce qui concerne le traitement des données, il aurait été préférable de disposer d'une base de données prête contenant toutes les informations financières des clients réels de la banque afin d'obtenir un ensemble de données d'apprentissage plus fiable et riche en termes de cas de faillite et de quantité de données, mais on envisage de faire le meme traitement et modélisation une fois que les données de la banque seront prêtes.

Pour l'application des défirrent modèle, les seuils doivent être considérés comme des guides. Par conséquent, le seuil devrait varier en fonction du but et de l'intention de la prédiction. Vu du point de vue d'un intervenant externe, par exemple un directeur de banque, il est dans l'intérêt de prédire correctement la plupart des entreprises qui feront faillite, car les entreprises en faillite pourraient être très coûteuses. Cela implique qu'il faudrait peut-être utiliser un seuil encore plus bas. Pour les fournisseurs, lorsque le coût de la faillite n'est pas si sévère, un seuil plus élevé que celui optimisé peut convenir.

Références

[1] "https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257." https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257.

[2] groupes de travail, *France intelligence artificielle*. .

[3] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Dev.*, pp. 210–229, [Online]. Available: https://doi.org/10.1147/rd.33.0210.

[4] M. R. K. A. M. Giancarlo Zaccone, "Deep Learning with TensorFlow: Explore neural networks with Python," 2017.

[5] and A. C. Ian Goodfellow, Yoshua Bengio, "Deep Learning. MIT Press," 2016. http://www.deeplearningbook.org.

[6] "O.Mehdi et K.salim, Classification d'objets avec le Deep Learning, Univ de bouira, 2018."

[7] "Tout ce que vous voulez savoir sur l'algorithme K-Means." https://mrmint.fr/algorithme-k-means.

[8] A. Chavan, "A Comprehensive Guide to Decision Tree Learning," *aitimejournal*.

[9] "S. R. Safavian and D. Landgrebe, 'A survey of decision tree classifier methodology,' in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458."

[10] "Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324."

[11] "D. A. Freedman 'Bootstrapping Regression Models,' The Annals of Statistics, Ann. Statist. 9(6), 1218-1228, (November, 1981)."

[12] "Breiman, L. Bagging predictors. Mach Learn 24, 123–140 (1996). https://doi.org/10.1007/BF00058655."

[13] "Biau, G., Scornet, E. A random forest guided tour. TEST 25, 197–227 (2016)."

[14] *Kleinbaum D.G., Klein M. (2010) Introduction to Logistic Regression. In: Logistic Regression. Statistics for Biology and Health. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1742-3_1*. .

[15] V. N. Vapnik, *An overview of statistical learning theory*. 1995.

[16] "Noble, W. What is a support vector machine?. Nat Biotechnol 24, 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565."

[17] "J. M. Keller, M. R. Gray and J. A. Givens, 'A fuzzy K-nearest neighbor algorithm,' in IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-15, no. 4, pp. 580-585, July-Aug. 1985, doi: 10.1109/TSMC.1985.6313426."

[18] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities."

[19] "RAHBARI, Dadmehr. High Performance Data mining by Genetic Neural Network. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, [S.l.], v. 4, n. 1-4,

p. pp. 60-70, Oct. 2013. ISSN 2067-3957."

[20] G. Garson, *Neural networks: An introductory guide for social scientists*. 1998.

[21] Ravindra Parmar, "Common Loss functions in machine learning," [Online]. Available: https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23.

[22] H. Michel, "Gradient Descent – L'algorithme Du Gradient," 2020. https://ledatascientist.com/gradient-descent-lalgorithme-du-gradient/.

[23] "Courbe ROC." https://fr.wikipedia.org/wiki/Courbe_ROC.

[24] "S. Seebacher and R. Schüritz. 'Blockchain technology as an enabler of service systems: A structured literature review'. In: International Conference on Exploring Services Science. Springer. (2017)."

[25] "S. Ølnes. 'Beyond bitcoin enabling smart government using blockchain technology'. In: International Conference on Electronic Government and the Information Systems Perspective. Springer. (2016)."

[26] "R. Böhme et al. 'Bitcoin: Economics, technology, and governance'. Journal of Economic Perspectives 29.2 (2015)."

[27] "H. Wang, K. Chen, and D. Xu. 'A maturity model for blockchain adoption'. Financial Innovation 2.1 (2016)."

[28] "S. Kikitamara, M. van Eekelen, and D. I. J.-P. Doomernik. 'Digital Identity Management on Blockchain for Open Model Energy System'. MA thesis. 2017."

[29] "S. Nakamoto. 'Bitcoin: A Peer-to-Peer Electronic Cash System' (2008)."

[30] "J. Bergquist. 'Blockchain Technology and Smart Contracts: Privacy-preserving Tools'. MA thesis. 2017."

[31] "EthereumTeam. Ethereum Blockchain App Platform. (2018). url: https : / / www . ethereum.org/. (visited on 08/31/2018)."

[32] "MultiChainTeam. MultiChain Offical Website. (2017). url: https://www.multichain. com/. (visited on 08/29/2018)."

[33] "S. Bilonia. How does Bitcoin Blockchain work and what are the rules behind it? (2017). url: https://www.quora.com/How-does-Bitcoin-Blockchain-work-and-whatare-the-rules-behind-it. (visited on 06/25/2018)."

[34] "D. Drescher. Blockchain basics. Springer, (2017)."

[35] "M. Gates. Blockchain: Ultimate guide to understanding blockchain, bitcoin, cryptocurrencies, smart contracts and the future of money. CreateSpace Independent Publishing Platform, 2017."

[36] "R. L. Rivest. 'Cryptography'. In: Algorithms and Complexity. Elsevier, (1990)."

[37] "J. Katz et al. Handbook of applied cryptography. CRC press, (1996)."

[38] "TutorialsPointTeam. Cryptography Hash functions. (2016). url: https://www.tutorialspoint. com/cryptography/cryptography_hash_functions.htm (visited on 08/31/2018)."

[39] "DocuSignTeam. What are digital signatures? (2018). url: https://www.docusign. co. uk / how - it - works / electronic - signature / digital - signature / digital - signature-faq. (visited on 08/31/2018)."

[40] "E. Paul. What is Digital Signature- How it works, Benefits, Objectives, Concept. (2017). url: http://www.emptrust.com/blog/benefits-of-using-digital-signatures (visited on 08/31/2018)."

[41] "M. Swan. Blockchain: Blueprint for a new economy. O'Reilly Media, Inc., (2015)."

[42] "A. M. Antonopoulos. Mastering Bitcoin: Programming the Open Blockchain. O'Reilly Media, Inc., (2017)."

[43] "V. Morabito. 'Business Innovation Through Blockchain'. Cham: Springer International Publishing (2017)."

[44] "M. Francisconi. 'An explorative study on blockchain technology in application to port logistics'. MA thesis. 2017."

[45] "F. Glaser. 'Pervasive decentralisation of digital infrastructures: a framework for blockchain enabled system and use case analysis'. In: The Association for Information System, 2017."

[46] "J. Pitzali. What is a Blockchain transaction? (2017). url: https://www.quora.com/ What-is-a-Blockchain-transaction. (visited on 06/20/2018)."

[47] "J. Barkatullah and T. Hanke. 'Goldstrike 1: Cointerra's first-generation cryptocurrency mining processor for bitcoin'. IEEE micro 35.2 (2015)."

[48] "BlockgeeksTeam. Cryptocurrency Wallet Guide: A Step-By-Step Tutorial. (2016). url: https://blockgeeks.com/guides/cryptocurrency- wallet- guide/. (visited on 06/19/2018)."

[49] "S. Mathieson. Blockchain starts to prove its value outside of finance. (2017). url: https://www.computerweekly.com/feature/Blockchain-starts-to-prove-itsvalue-outside-of-finance. (visited on 08/31/2018)."

[50] "J. Yli-Huumo et al. 'Where is current research on blockchain technology?—a systematic review'. PloS one 11.10 (2016)."

[51] "L. Wang and Y. Liu. 'Exploring miner evolution in bitcoin network'. In: International Conference on Passive and Active Network Measurement. Springer. (2015)."

[52] "No Title." https://www.larousse.fr/dictionnaires/francais/banque/7863.

[53] "Banque : définition, traduction et synonymes." https://www.journaldunet.fr/business/dictionnaire-economique-et-financier/1198859-banque-definition-traduction-et-synonymes/.

[54] R. MARLO-MARETTE, *Connaissance Segmentation des clients (Politique interne SGA)*. 2020.

[55] "Larousse en ligne." . https://www.larousse.fr/dictionnaires/francais/cr%C3%A9dit/20314

[56] https://wikimemoires.net/2011/05/gestion-risque-secteur-bancaire/.

[57] P. M. MARLO-MARETTE, Raphael, *Politique d'octroi de credit corporate(Politique*

interne SGA). 2020.

[58] "Quelle la définition du risque client pour une banque ?" https://www.rachatducredit.com/definition-risque-client-pour-une-banque-869.html.

[59] "PRÉSENTATION SOCIETE GENERALE." https://www.societegenerale.com/fr/le-groupe-societe-generale/identite/presentation.

[60] "PRÉSENTATION DE SOCIÉTÉ GÉNÉRALE ALGÉRIE." https://particuliers.societegenerale.dz/fr/nous-connaitre/presentation-societe-generale-algerie/.

[61] SGA, *Politique Interne direction SIOP*. 2019.

[62] "Du Jardin, Philippe. 2010. Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy. Neurocomputing 70: 2047–60."

[63] "Virág, Miklós. 2004. A cs″odmodellek jellegzetességei és története. Vezetéstudomány 35: 24–32."

[64] "Nyitrai, Tamás. 2015a. Dinamikus pénzügyi mutatószámok alkalmazása a cs″odel″orejelzésben. Ph.D. thesis, Budapesti Corvinus Egyetem, Budapest, Hungary."

[65] "Fitzpatrick, Paul J. 1932. A Comparison of the Ratios of Successful Industrial Enterprises with Those of Failed Companies. Washington: The Accountants' Publishing Company."

[66] "Durand, David. 1941. Risk Elements in Consumer Instalment Financing. New York: National Bureau of Economic Research."

[67] "Beaver, William H. 1966. Financial ratios as predictors of failure. Empirical research in accounting: selected studies. Journal of Accounting Research 4: 1–111."

[68] "Myers, James H., and Edward W. Forgy. 1963. The development of numerical credit evaluation systems. Journal of the American Statistical Association 58: 799–806."

[69] "Altman, Edward I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23: 589–609."

[70] "Chesser, Delton L. 1974. Predicting loan noncompliance. Journal of Commercial Bank Lending 56: 28–38."

[71] "Ohlson, James A. 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18: 109–31."

[72] "Zmijewski, Mark E. 1984. Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research 22: 59–82."

[73] "Frydman, Halona, Edward I. Altman, and Duen-Li Kao. 1985. Introducing recursive partitioning for financial classification: The case of financial distress. The Journal of Finance 40: 269–91."

[74] "Prusak, Bła ˙zej. 2005. Modern Methods of Predicting Financial Risk in Companies. Warsaw: Difin."

[75] "Kristóf, Tamás. 2005a. A cs″odel″orejelzés sokváltozós statisztikai módszerei és empirikus vizsgálata. Statisztikai Szemle 83: 841–63."

[76] "Odom, Marcus D., and Ramesh Sharda. 1990. A neural network model for bankruptcy prediction. Paper present at the International Joint Conference on Neural Networks, San Diego, CA, USA, June 17–21; Ann Arbor: IEEE Neural Networks Council, vol. II, pp. 163–7."

[77] "Vlachos, Dimitros, and Yannis A. Tolias. 2003. Neuro-fuzzy modeling in bankruptcy prediction. Yugoslav Journal of Operational Research 13: 165–74."

[78] "Fan, Alan, and Marimuthu Palaniswami. 2000. Selecting Bankruptcy Predictors Using a Support Vector Machine Approach. In Proceedings of the International Joint Conference on Neural Networks. Neural Computing: New Challenges and Perspectives for the New Mil."

[79] "Dimitras, Augustinos I., Roman Slowinski, Robert Susmaga, and Constantin Zopounidis. 1999. Business failure prediction using rough sets. European Journal of Operational Research 114: 263–80."

[80] "Ardakhani, Mehdi N., Vahid Zare Mehrjerdi, Mohsen Sarvi, and Elias Sarvi. 2016. A survey of the capability of k nearest neighbors in prediction of bankruptcy of companies based on selected industries. Scinzer Journal of Accounting and Management 2: 27–37."

[81] "Sun, Lili, and Prakash P. Shenoy. 2007. Using Bayesian networks for bankruptcy prediction: Some methodological issues. European Journal of Operational Research 180: 738–53."

[82] "Lensberg, Terje, Aasmund Eilifsen, and Thomas E. McKee. 2006. Bankruptcy theory development and classification via genetic programming. European Journal of Operational Research 169: 677–97."

[83] "Bryant, Stephanie M. 1997. A case-based reasoning approach to bankruptcy prediction modeling, Intelligent Systems in Accounting. Finance and Management 6: 195–214."

[84] "Marqués, Ana I., Vicente García, and Javier Salvador Sánchez. 2012. Exploring the behaviour of base classifiers in credit scoring ensembles. Expert Systems with Applications 39: 10244–50."

[85] "Barboza, Flavio, Herbert Kimura, and Edward I. Altman. 2017. Machine learning models and bankruptcy prediction. Expert Systems with Applications 83: 405–17."

[86] "Meese, Eystein & Viken, Torbjørn. (2019). Machine Learning in Bankruptcy Prediction: Utilizing machine learning for improved bankruptcy predictions in the Norwegian market with an emphasis on financial, management and sector statements. 10.13140/RG.2.2.21."

[87] T. Kristóf and M. Virág, "A Comprehensive Review of Corporate Bankruptcy Prediction in Hungary," *J. Risk Financ. Manag.*, vol. 13, no. 2, p. 35, 2020, doi: 10.3390/jrfm13020035.

[88] X. BREDART, V. VELLA, and J. BONELLO, "Machine Learning Models for Predicting Financial Distress," *J. Res. Econ.*, vol. 2, no. 2, pp. 174–185, 2018, doi: 10.24954/jore.2018.22.

[89] "https://docs.conda.io/." .

[90] “www.jupyter.org. Retrieved 2020-11-13.” .

[91] Documentation Pandas en ligne “https://pandas.pydata.org/pandas-docs/pandas documentation. 28 January 2020.”

[92] Documentation Numpy en ligne “http://www.numpy.org/.”

[93] Documentation matplotlib en ligne “https://matplotlib.org.”

[94] Documentation Scikit-learn en ligne “Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.”

[95] Documentation Keras en ligne “https://keras.io/about/.”

[96] Documentation Scikit-learn en ligne “https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.” .

[97] Documentation Scikit-learn en ligne “https://scikit-learn.org/stable/modules/tree.html.” .

[98] Documentation Scikit-learn en ligne “https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.” .

[99] Documentation Scikit-learn en ligne “https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html.” .

[100] Documentation Scikit-learn en ligne “https://scikit-learn.org/stable/modules/svm.html.” .

[101] “Yuehui Chen, Bo Yang, Jiwen Dong, Time-series prediction using a local linear wavelet neural network, Neurocomputing, Volume 69, Issues 4–6, 2006, Pages 449-465, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2005.02.006. (https://www.sciencedirect.com/.”

Annexes

Annexe A

• Le Return On Asssets (ROA) : mesure le rapport entre le résultat net (outil permettant de savoir si l'entreprise est bénéficiaire ou déficitaire) et le total des actifs (ensemble des éléments générant des ressources). Il exprime la capacité d'une entreprise à générer un revenu à partir de ses ressources.

• Marge brute d'exploitation : correspond au rapport entre le résultat d'exploitation et le chiffre d'affaires. Ce ratio indique la performance économique avant prise en compte du résultat financier, des impôts, et des événements exceptionnels.

• Ratio de flux de trésorerie : Ce ratio représente la capacité d'autofinancement d'une société en fonction de la taille de cette dernière. Ce ratio témoigne de l'aptitude d'une société à générer des liquidités relativement à sa taille.

• Ratio de couverture des intérêts (Charges d'intérêts / EBIT) : Ce ratio indique dans quelle mesure les intérêts débiteurs sont couverts par les flux de trésorerie de la société. Un ratio inférieur à 1 signifie que la société a du mal à générer des flux de trésorerie suffisants pour régler ses intérêts débiteurs.

• Marge Bénéficiaire brute : ce ratio s'exprime en pourcentage, il signifie la différence entre le chiffre des ventes et le coût des marchandises vendues.

• Rotation du fonds de roulement : Le ratio du fonds de roulement correspond au quotient obtenu en divisant le chiffre d'affaires de la période par la moyenne du fonds de roulement de cette période.

• Rotation du capital de travail : On peut définir la période de rotation du capital comme l'intervalle de temps entre le moment où le capitaliste avance un capital-argent, et le moment où il récupère le capital-argent investi.

• Ratio Trésorerie passive : La trésorerie passive est égale aux soldes créditeurs de banque et aux concours bancaires. Elle correspond au passif de l'entreprise inscrit sur le bilan comptable. En d'autres termes, ce sont les dettes professionnelles à court terme.

• Fonds de roulement par rapport à l'actif total : Le fonds de roulement correspond à la différence entre les ressources stables de l'entreprise (capitaux propres et endettement à

moyen ou long terme) et les actifs immobilisés. Il constitue un élément clé de l'équilibre financier d'une entreprise.

• Ratio de rotations d'actifs immobilisés : Le ratio de rotation de l'actif immobilisé indique combien de revenus vous tirez de chaque dollar investi dans vos immobilisations corporelles.

• Taux rotation des stocks : Définition de la rotation des stocks et ses impacts sur votre entrepôt. La rotation des stocks correspond au nombre de fois que le stock de l'entrepôt est remplacé au cours d'une période donnée.

• Rotation des comptes clients : Le ratio de rotation des comptes clients est égal au rapport entre les ventes nettes réalisées par l'entreprise sur une période donnée et la moyenne des comptes clients affichés durant ladite période. Les créances clients ont un impact direct sur la santé de l'entreprise.

• Rotation de l'actif total : Le taux de rotation total de l'actif est un ratio financier qui mesure l'efficacité de l'utilisation de l'actif d'une société pour générer des revenus pour la société. Il est calculé en divisant le chiffre d'affaires net par le total de l'actif.

• Bénéfice net avant impôt par capital : Le bénéfice avant impôts (BAI) mesure la rentabilité d'une entreprise avant que les impôts soient pris en compte. Il s'agit du montant d'argent qui reste après avoir soustrait toutes les dépenses des revenus.

• Résultat d'exploitation : Le résultat d'exploitation est un solde intermédiaire de gestion qui détaille les produits et les charges de l'entreprise sur un exercice comptable écoulé. Il montre ainsi comment l'entreprise s'organise et créé de la richesse.

• Ratio d'endettement % : La ratio d'endettement est un indicateur financier qui permet de mesurer le niveau d'endettement d'une entreprise, et donc sa solvabilité. Ce ratio s'obtient en effectuant le rapport entre les dettes d'une entreprise et le montant de ses capitaux propres.

• Le ratio courant : Le current ratio, ou ratio de liquidité générale, permet d'évaluer la situation de liquidité de l'entreprise, sa capacité à faire face à ses engagements à court terme. Il se calcule en divisant l'actif courant par le passif courant.

• Taux de la dette portant intérêt : Ce ratio indique le taux d'intérêt moyen appliqué aux emprunts de la société. La comparaison du ratio actuel et de ceux des exercices antérieurs donne une idée du taux accepté par la société pour contracter de nouvelles dettes.

• Marge brute « Gross Margin » : Le taux de marge se calcule en pourcentage en divisant la marge commerciale par le prix de revient. Il permet d'extrapoler la marge dégagée sur les ventes futures.

• Degré de levier financier (DFL) : L'effet de levier se calcule en mettant en rapport le taux de rentabilité de l'actif économique après impôt et le coût de la dette

• Taux de frais de recherche et développement : Les « frais de recherche et de développement » (R&D) regroupent les dépenses correspondant à l'effort financier réalisé par un organisme en matière de recherche scientifique ou technique et de développement

• Taux d'imposition (A)

• Taux Croissance marge brute ‘’Année 2 ‘’

• Taux de croissance du résultat net après impôts

• Taux de croissance du résultat net

• Taux du total des actifs ‘’ Année 1 ‘’

• Taux de croissance du rendement total des actifs ‘’ Année 2 ‘’ : Le ratio du rendement de l’actif total compare l’actif total d’une entreprise avec le montant qu’elle remet à ses actionnaires

• Réinvestissement en espèces % : Le réinvestissement consiste pour un investisseur à replacer les revenus de son placement ou de son investissement dans le même support

• Taux d'intérêt débiteur : Le taux débiteur est un taux d’intérêt exprimé en pourcentage fixe ou variable, afin de calculer les intérêts d’un capital emprunté. C’est le taux qui permet de calculer les mensualités dues par un client

• Total du passif / Valeur nette

• Valeur nette / actifs

• Ratio d'adéquation des fonds à long terme (A) : Le ratio d'adéquation des fonds propres (CAR) permet de s'assurer que les banques disposent d'un capital suffisant pour protéger l'argent des déposants. Les exigences de fonds propres fixées par la BRI sont devenues plus strictes ces dernières années

• Liquidités / total des actifs

• Actif à court terme/passif à court terme : Le ratio de liquidité de l'entreprise permet de comparer l'actif court terme, inscrit au bilan, au passif court terme

• Passif courant / passif

• Fonds de roulement/capitaux propres

• Passif/capitaux propres actuels

• Total des revenus / total des dépenses

• Total des dépenses/actifs

• Rotation des actifs : Le ratio indique combien de revenus est tiré de chaque dollar investi dans l'actif total

• Flux de trésorerie par rapport aux ventes

• Du passif courant aux capitaux propres

• Des capitaux propres au passif à long terme

• Flux de trésorerie par rapport à l'actif total

• Flux de trésorerie par rapport aux passifs

• Passif à court terme sur Actif à court terme

• Revenu net par rapport au total des actifs : Le ratio du rendement de l'actif total compare l'actif total d'une entreprise avec le montant qu'elle remet à ses actionnaires. On le calcule en divisant le bénéfice net (revenu après impôt) de l'entreprise par actif total, et en multipliant le résultat par 100 %

• Rentabilité des capitaux propres : une notion économique à la fois très ancienne et très contestée, qui mesure en pourcentage le rapport entre le résultat net et les capitaux propres investis par les associés ou actionnaires de sociétés

• Des capitaux propres au passif : Les capitaux propres sont les ressources financières que possède l'entreprise (hors dette). Une entreprise investit et génère ses propres capitaux pour son fonctionnement mais aussi pour rémunérer ses actionnaires. C'est au passif du bilan comptable que l'on retrouve les capitaux propres