RÉPUBLIQUE ALGERIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

École Nationale Polytechnique

Département Génie Industriel

Entreprise Société Générale

Mémoire de projet de fin d'études

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

Etude et implémentation des modèles d'apprentissage supervisé pour la prédiction du risque ''Faillite'' des clients corporate

Présenté par

Bilel CHERKI (Management de l'Innovation)

Abdelrahmane BOUTERANE (Management de l'Innovation)

Sous la direction de :

M. Hakim FOURAR LAIDI

M. Oussama ARKI

Composition du Jury :

Président	Mme. Noual BOUKADOUM	MAA	ENP
Promoteur	M. Hakim FOURAR LAIDI	MCB	ENP
Promoteur	M. Oussama ARKI	MAB	ENP
Examinateur	Mme. Woujdene NAHILI	MAB	ENP

ENP 2021

RÉPUBLIQUE ALGERIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

École Nationale Polytechnique

Département Génie Industriel

Entreprise Société Générale

Mémoire de projet de fin d'études

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

Etude et implémentation des modèles d'apprentissage supervisé pour la prédiction du risque ''Faillite'' des clients corporate

Présenté par

Bilel CHERKI (Management de l'Innovation)

Abdelrahmane BOUTERANE (Management de l'Innovation)

Sous la direction de :

M. Hakim FOURAR LAIDI

M. Oussama ARKI

Composition du Jury :

Président	Mme. Noual BOUKADOUM	MAA	ENP
Promoteur	M. Hakim FOURAR LAIDI	MCB	ENP
Promoteur	M. Oussama ARKI	MAB	ENP
Examinateur	Mme. Woujdene NAHILI	MAB	ENP

ENP 2021

Dédicaces

A mes **parents** qui sont ma fierté, mon bonheur et ma joie.
A mon **frère et ma sœur**.
A mon Binôme l'Ingénieur Artiste '' Matkhafch ''
A ma chère Team **COLLO** la **LEGENDAIRE**.
A mes frères :
Bilel Maouche et Abdeldjalil Mahmoudi
A mon groupe **Indus 4.0** et mon club **IEC**
A toute personne qui m'a aidé tout au long de mon parcours.
A tous mes chers.

Je dédie ce travail

Abdelrahmane BOUTERANE

الملخص

انطلقت البنوك الجزائرية في طريق التحديث الذي يعتمد على نشر تقنيات المعلومات .يتم هذا العمل ضمن هذا الإطار ويهدف إلى تحسين نظام إدارة المخاطر ، حتى نكون قادرين أولاً وقبل كل شيء على تلبية الطلب المتزايد وضمان أمن محافظهم الاستثمارية ، ثم في الخطوة الثانية ، ليكونوا قادرين على تلبية الطلب المتزايد وتحقيق الامتثال للأحكام الاحترازية الدولية من خلال أداة صنع القرار

يتكون النهج المعتمد من تصميم وتقييم وتصنيف خوارزميات التعلم تحت الإشراف المختلفة ؛ ثم اختيار النموذج الأكثر ملاءمة للمشكلة التي يطرحها هذا المشروع من أجل تحقيق الأهداف التي وضعها البنك

الكلمات المفتاحية :الذكاء الاصطناعي ، التعلم الآلي ، التنقيب في البيانات ، البنك ، العميل ، المخاطر

Abstract

Algerian banks are committed to the path of modernization based on the deployment of information technologies. This work takes place within this framework and aims to improve the risk management system, in order to be able, firstly, to meet the growing demand and to guarantee the security of their investment portfolios, and secondly, to comply with the international prudential provisions through a decision support tool.

The approach adopted consists in the design, evaluation and ranking of various supervised learning algorithms; to then select the most suitable model for the problem posed by this project in order to achieve the objectives set by the bank.

Keywords: Artificial Intelligence, Machine Learning, Data Mining, Bank, Customer, Risk.

Résumé

Les banques Algériennes se sont engagées dans la voie de la modernisation qui s’appuie sur le déploiement des technologies de l’information. Ce travail prend place dans ce cadre et vise à améliorer le dispositif de gestion du risque, afin de pouvoir, dans un premier temps, répondre à la demande croissante et garantir la sécurité de leurs portefeuilles d'investissement, puis dans un second temps, se mettre en conformité avec les dispositions prudentielles internationales par le biais d'un outil d'aide à la décision.

L'approche adoptée consiste en la conception, l'évaluation et le classement de divers algorithmes d'apprentissage supervisé ; pour ensuite sélectionner le modèle le plus adapté à la problématique posée par ce projet en vue d'atteindre les objectifs fixés par la banque.

Mots clés : Intelligence artificielle, Machine Learning, Data Mining, Banque, Client, Risque.

Sommaire

Liste des figures

Liste des tableaux

Liste des Annexes

Abréviation

ANN : Artificial Neural Network
BDD : Base de données
CRISP-DM : Cross Industry Standard Process for Data Mining
CRS : Common Reporting Standard
DCCIT : Dossier Electronique De Crédit Commercial A L'international
EBE : Excédent Brut D'exploitation
EBIT: Earnings Before Interest and Taxes
FATCA: Foreign Account Tax Compliance Act
HLAD : Hors limite à divers
IBFS: International Banking and Financial Services
IBM: International Business Machines Corporation
KNN : K-Nearest Neighbors
KYC : Know Your Customer
KYS : Know Your Supplier
LAD : (Limite à divers) Montant de l'autorisation qui est déléguée par le Directeur de IBFS aux Responsables d'Implantations.
Logit : Logistic Regression
LDA : Linear Discriminant Analysis
PNB : Produit National Brut
PNC : processus de notation classique
PND : processus de notation dérivée
PNE : processus de notation à dire d'expert
RAROC: l'acronyme anglais RAROC signifie « Risk Adjusted Return on Capital » Ratio qui mesure la rentabilité des fonds propres ajustée au risque.
RESO : Réseau Société Générale France
SGA : Société Générale Algérie
SSC : (Secteur de suivi de clientèle) désigne les Directions opérationnelles de la SGA ou de toute autre filiale IBFS auxquelles sont rattachés des clients (ou groupes clients). Les SSC ont pour tâche d'établir la stratégie commerciale et risque de leur portefeuille.
SIG : Solde Intermédiaire de Gestion
SVM : Support Vector Machine
TPE : Terminal de Paiement Electronique

Introduction Générale

INTRODUCTION GENERALE :

La gestion des risques est l'essence même des activités bancaires. En fait, le rôle des banques dans le système financier est de convertir les dépôts en prêts. Ce rôle expose les banques à de multiples risques comme le risque de change, risque de taux d'intérêt et risque de contrepartie.

À travers un grand nombre de revues descriptives et prédictives, la gestion du risque de crédit a toujours été un thème traditionnel des théoriciens bancaires, et c'est un sujet riche en littérature. Aujourd'hui, ce sujet est plus important que jamais.

Il a quatre raisons principales de revenir à l'actualité économique et financière :

- ✓ L'attractivité de la bourse pour les meilleurs emprunteurs entraîne une concurrence entre la bourse et le marché bancaire
- ✓ La dimension internationale de la banque nécessite l'application des dispositions prudentes de Bâle II
- ✓ Utiliser les systèmes d'information pour développer des techniques de modélisation statistique
- ✓ Problèmes de concurrence interbancaire et excès de liquidité.

D'un point de vue opérationnel, toutes les banques se sont engagées dans la modernisation des systèmes de gestion du risque de contrepartie, d'abord pour répondre à la demande croissante et assurer la sécurité de leurs portefeuilles d'investissement, puis pour les qualifier pour se conformer aux dispositions prudentielles internationales.

Les banques algériennes, comme les banques internationales, ont pris en compte la réforme globale du système financier du pays et ont initié le niveau de système de gestion des risques conformément à l'approche graduelle et pragmatique menée par la banque centrale.

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.

Afin d'optimiser l'efficacité du processus décisionnel, toutes les parties impliquées dans ce processus doivent se soucier en permanence de la qualité des risques encourus et prendre toutes les mesures nécessaires pour les maîtriser conformément à cette politique de crédit.

Que le risque porte sur un client ou sur une transaction, le principe d'une intervention conjointe des lignes commerciales et de la filière Risques s'applique tout au long du processus d'approbation. Il est important que chacune d'elles s'organise de façon à rendre ce circuit le plus court et le plus productif possible.

L'objectif de notre mémoire est de permettre un développement sain des engagements de la banque, basé sur une meilleure convergence entre les objectifs commerciaux et l'impératif d'une bonne maîtrise des risques.

A cet effet, nous avons organisé le présent travail en trois (3) chapitres :

Au cours du premier chapitre, nous avons entrepris la présentation des acteurs impliqués dans ce projet. Nous avons ensuite entamé un diagnostic en décrivant de manière concise les procédures internes de Société Générale. Puis, nous avons analysé de manière détaillée le processus d'octroi de crédit et les critères d'approbation d'un dossier de crédit. Ce diagnostic nous a permis de choisir la problématique qui sera traitée dans ce travail.

Le deuxième chapitre est consacré à l'état de l'art et traite des outils que nous avons utilisés pour répondre à la problématique choisie. Ainsi, des notions sur l'intelligence artificielle seront introduites de même que différentes techniques issues de l'apprentissage automatique et du data mining.

Dans le troisième chapitre, nous nous consacrons au développement de la solution qui répondra à la problématique choisie. Pour cela, nous avons développé un outil d'aide à la décision reposant sur des techniques d'apprentissage automatique supervisé en vue d'automatiser l'analyse du profil risque des clients corporate ainsi que la prédiction de la probabilité de leur défaillance (banqueroute).

Le constat des résultats de la mise en œuvre des programmes développés nous amène à proposer des perspectives d'amélioration en guise de conclusion.

Chapitre 1 :
Etat des lieux et diagnostic

Chapitre 1 : Etat des lieux et diagnostic

Introduction :

Dans cette partie, nous analyserons l'environnement de Société Générale, ses diverses démarches internes ainsi que ses domaines d'activités pour aboutir ainsi à un diagnostic du processus d'analyse d'un dossier de crédit et aux pistes d'amélioration et ce, dans le but de définir le périmètre de notre étude ainsi que son cadre.

1. Etude de l'existant :

1.1. Définition d'une Banque :

« 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. » [1]

« 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. » [2]

1.2. Présentation du Groupe Société Générale :

Société Générale est l'un des tous premiers groupes européens de service financier 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.

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 Fixed Income, 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

1.3. 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. (L'organigramme de l'entreprise est en **Annexe D)**

1.3.1. 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.

Son organisation est illustrée dans l'organigramme suivant :

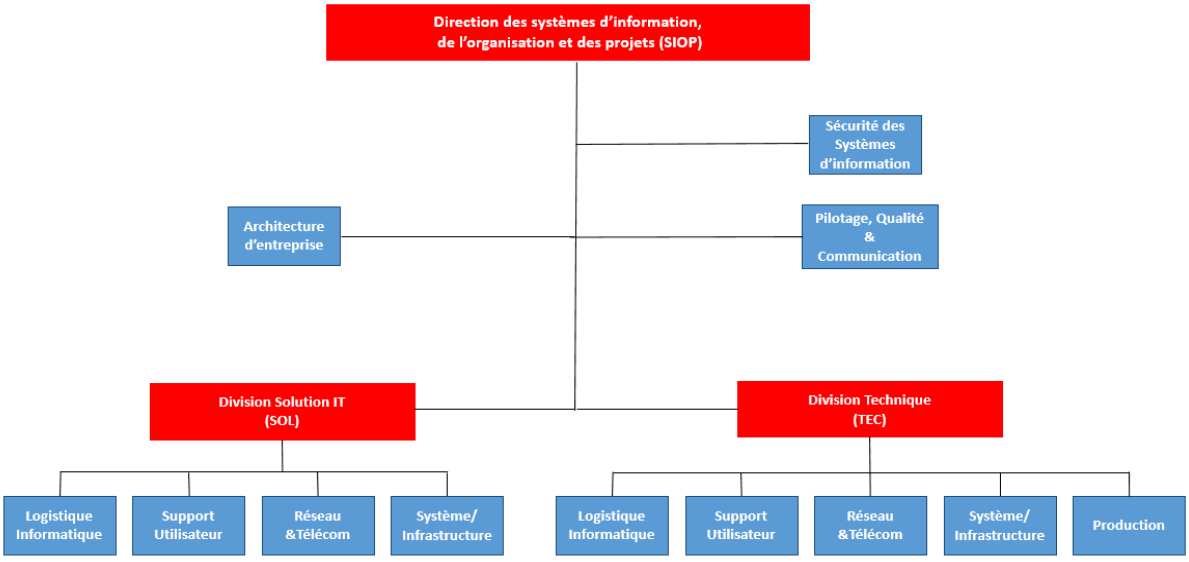


Figure 1 : Organigramme de la Direction SIOP

1.3.2. Missions et activités de la SIOP :

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

- 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.

1.3.3. 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.

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.

- **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 :

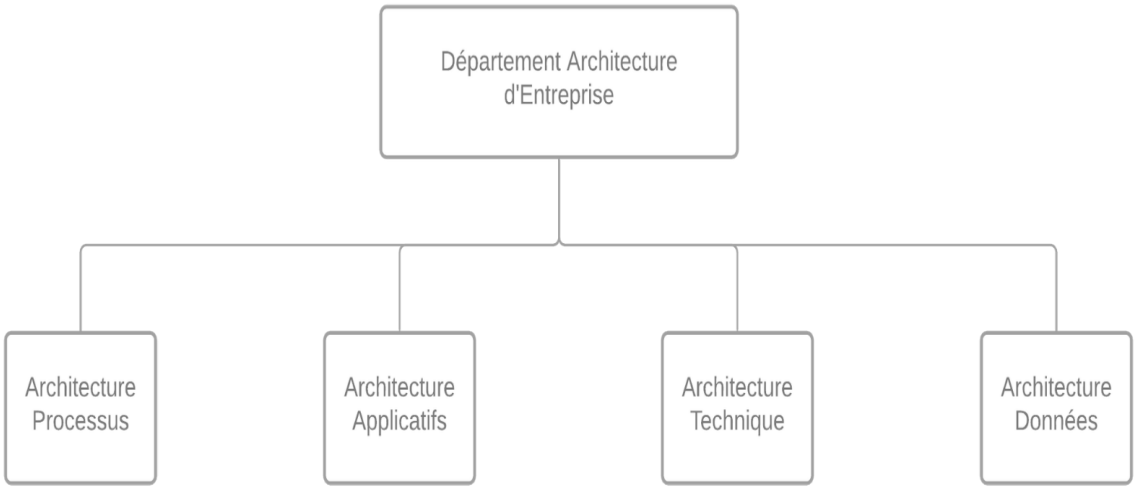


Figure 2: Activités du département Architecture de l'entreprise

- **Architecture Données :**

  Cette activité a pour rôle de définir et piloter le déploiement de la stratégie « Données » au sein de la Société Générale Algérie, dans le respect de la stratégie « Données » du Groupe.

- **Architecture Processus :**

  En étroite collaboration avec les équipes projets, cette activité participe à la transformation de la banque en accompagnant des projets d'amélioration de processus ou des projets de déploiement liées à la démarche processus.

- **Architecture Applicative :**

  L'architecture applicative accompagne et oriente les projets de transformation dans le respect du cadre d'urbanisme global afin de répondre aux exigences de performances, de pérennité et d'évolutivité. Elle a pour rôle de garantir l'évolution cohérente du SI tant que sur les aspects fonctionnels et applicatifs que les autres contraintes de couts, délai, risque, …

- **Architecture Technique :**

  L'architecture Technique veille sur le respect du cadre Groupe pour le choix des solutions techniques. Elle participe aussi à l'évaluation du patrimoine technique pour proposer des pistes d'amélioration.

2. Concepts de base et fonctionnement interne :

2.1. Notions Générales :

Le client :
Est considéré client SGA, au regard de la présente politique de crédit, tout emprunteur remplissant l'un des critères suivants :

- Ancienneté bancaire : compte à vue ouvert à SGA depuis au moins 3 mois et dont le KYC est validé.
- Société appartenant à un client SGA : entreprise nouvellement créée, dont plus de 25 % des parts sociales sont détenues par un client SGA dont le CA est > à 100 Millions de DZD, et ce quelle que soit la durée de l'ancienneté du compte.
- Société dont le mouvement confié sur le compte de SGA serait ≥ à 1/5 de son dernier CA fiscal, et ce quelle que soit la durée de l'ancienneté du compte.
- Entreprise entrant dans le cadre d'un partenariat signé entre SGA et un partenaire (exemple : leasing), définissant les modalités spécifiques de financement d'une cible bien identifiée (prospect recommandé).

Sont considérées comme « prospects » :

- Les entreprises ne remplissant pas l'une des conditions précitées ;
- Les entreprises domiciliées à la SGA, disposant d'un compte inactif depuis plus d'une année.

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.

Emprunteur & Garant :

L'Emprunteur : Est qualifié d'emprunteur toute entreprise bénéficiaire de financements et / ou d'engagements de la part de la SGA.

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 pouvoir assumer seul la charge du crédit en cas de défaillance de l'emprunteur.

2.2. Segmentation Clients :

Définition :

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.

Segmentation du groupe SG :

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 :

- 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 ».

Tableau 1: Segmentation des clients

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

Le processus de segmentation :

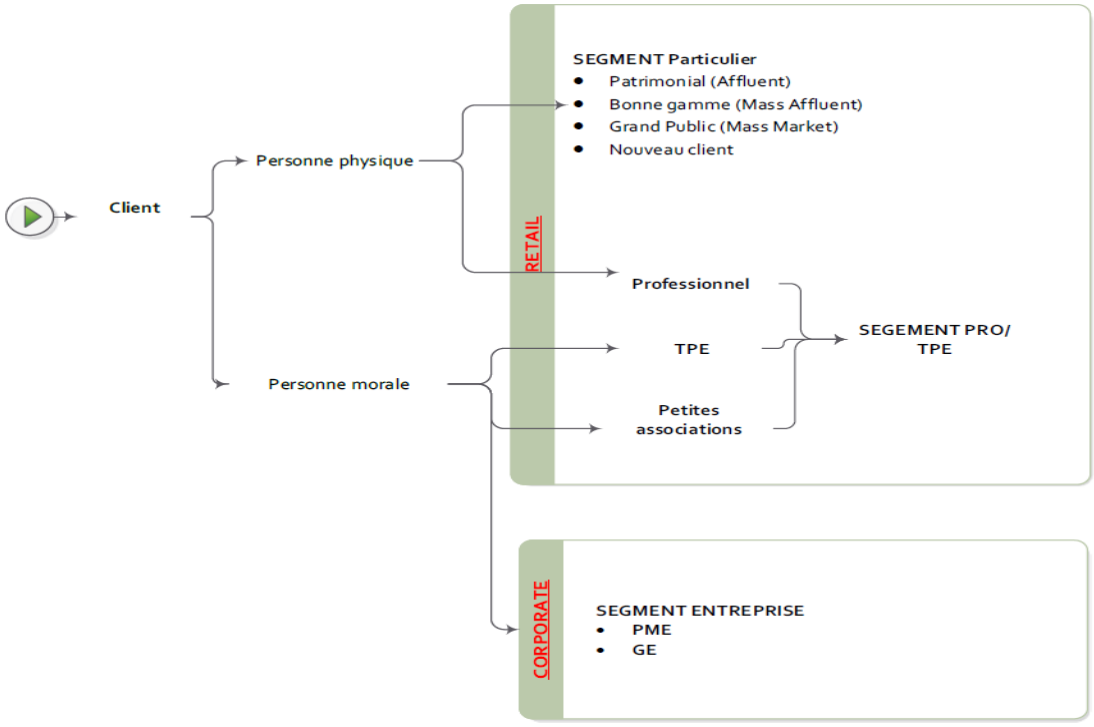


Figure 3: Schématisation du processus segmentation des clients

L'explication détaillée des différents segments sont en **Annexe B**

2.3. Crédit Bancaire :

Définition :

« Un crédit est une mise à disposition d'argent sous forme de prêt, consentie par un créancier (prêteur) à un débiteur (emprunteur). Pour le créancier, l'opération donne naissance à une créance sur l'emprunteur, en vertu de laquelle il pourra obtenir remboursement des fonds et paiement d'une rémunération (intérêt) selon un échéancier prévu. Pour l'emprunteur, qu'il s'agisse d'une entreprise ou d'un particulier, le crédit consacre l'existence d'une dette et ouvre la mise à disposition d'une ressource financière à caractère temporaire. » [3]

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]

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 nous pouvons constater 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, dont nous allons détailler ultérieurement.

Selon la durée du crédit, nous distinguons 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 nous trouvons différents types de crédit selon leurs caractéristiques, par exemple nous avons les crédits selon la monnaie c'est-à-dire soit en monnaie nationale ou en devise, nous trouvons aussi des crédits selon le type de taux (fixe ou variable) …

Cycle de vie d'un crédit bancaire :

Le cycle de vie de chaque crédit diffère selon sa nature, mais nous pouvons distinguer des phases communes à tous les types qui sont les suivantes :

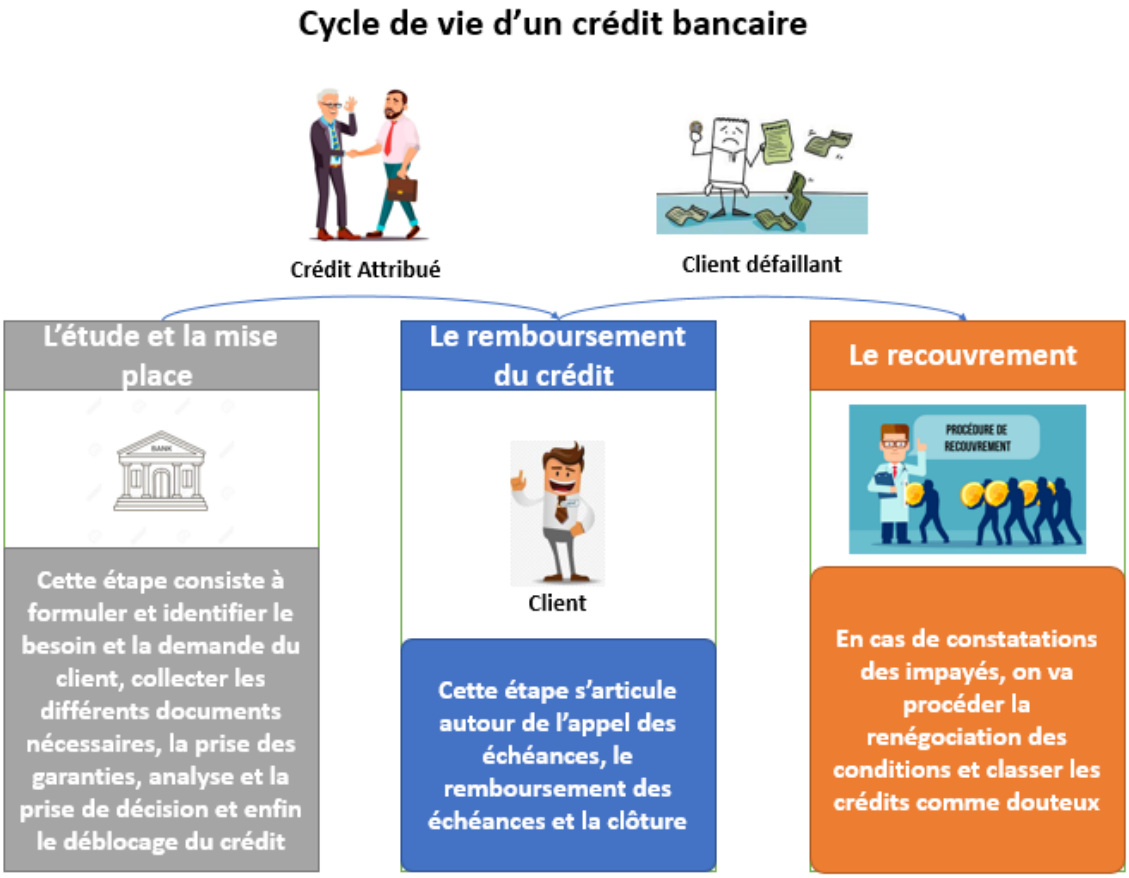


Figure 4 : Phases communes du cycle de vie d'un crédit

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.

Le processus doit également prévoir le pilotage ainsi que l'archivage des éléments. Comme présenté ci-dessus :

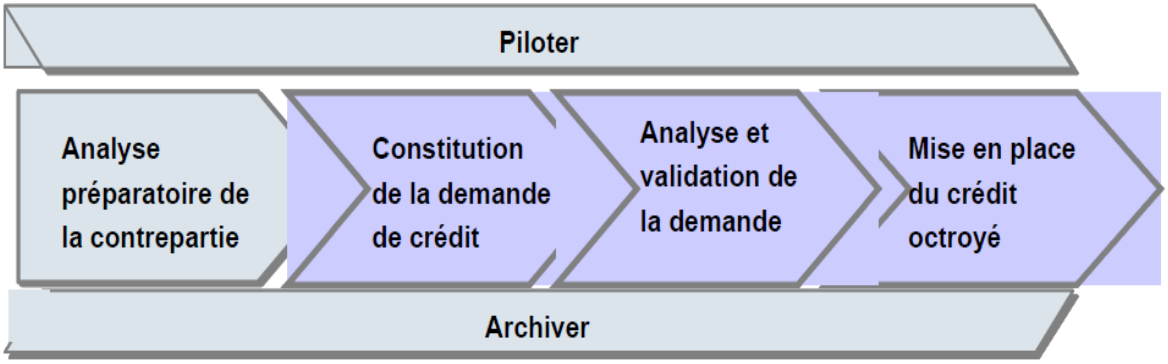


Figure 5 : Processus d'octroi de crédit

2.4. Types de crédit proposé par la banque :

Les financements proposés aux clients peuvent être scindés en trois catégories :

- Les financements liés au cycle d'exploitation ou crédits de fonctionnement.
- Les financements d'investissement (biens mobiliers et immobiliers).
- La ligne de couverture de risque de change (CVaR).

- **Les financements d'exploitation :**

Les crédits de fonctionnement sont accordés aux entreprises pour faire face à leurs besoins d'exploitation et assurer un déroulement normal de leur activité.
Ils ont pour objet de financer le cycle d'exploitation de l'entreprise (phase d'approvisionnement, de transformation, de stockage et de commercialisation) en complément de son fonds de roulement.
Ainsi, la banque propose des solutions de financement adaptées à chacune de ces phases en appréciant la nature du besoin, l'importance des montants en jeu et la maturité des utilisations.

Ces crédits peuvent être classé selon leurs nature en deux grandes sous – catégories :

a) Les Crédits par caisse :

Ils se matérialisent par un décaissement effectif et immédiat des fonds de la part de la banque, sous différentes formes, dont nous retrouvons essentiellement :

- Les crédits de trésorerie : découvert, crédit de campagne, crédit spot.
- Les crédits de mobilisation de créances clients : l'escompte de papier commercial, l'avance sur facture.
- Les crédits de financement des stocks : le spot de refinancement d'achat locaux et d'importation, le crédit de campagne.
- Les crédits de financement des marchés publics : l'avance sur situation de marché (créance née constatée C.N.C).

Parmi cette sous-catégorie, Les principales formes utilisées au niveau de la banque sont :

- Découverts et facilités de caisse : pour gérer et combler les décalages de trésorerie sur des durées courtes adaptées selon les besoins.

- Crédit spot : est un crédit à court terme (de 30 à 90 jours) répondant à des besoins de financement du cycle d'exploitation, il est matérialisé par la signature de billet à ordre.
- Escompte commercial : pour mobiliser les créances commerciales réglées par effets de commerce.
- Avances sur factures : avances sur situations pour mobiliser des créances sur présentation de factures ou situations de travaux.

b) Les Crédits par signature :

Il s'agit de crédits qui engagent la signature de la banque vis-à-vis des tiers (Fournisseurs, Administrations). Ces concours peuvent participer à alléger la trésorerie des clients puisqu'ils leurs permettent soit de différer des règlements (aval, Crédoc), soit d'éviter des décaissements (cautions en douanes…), soit d'activer des rentrées de fonds (CBE, CRA…).
Il s'agit d'engagements irrévocables de la part de la banque qui peuvent donner lieu à des décaissements certains ou probables de fonds sur les comptes des clients d'où l'importance de les gérer avec beaucoup de précaution. Le dimensionnement des crédits par signature doit donc être cohérant avec la taille des lignes de crédit par caisse que nous sommes disposés à octroyer.
Les principales formes utilisées au niveau de la banque sont :

- Les cautions : garantie donnée sur l'ordre du client, lui permettant d'exécuter ses obligations contractuelles, il existe plusieurs types de cautions par exemple : cautions diverses de soumission, de bonne exécution et de restitution d'avances ; les cautions administratives et les cautions en douanes.
- Les avals locaux et étrangers : est une sûreté propre aux effets de commerce, le paiement à échéance de ces derniers sera garanti aux fournisseurs.
- Les CREDOC/SBLC : est un engagement donné dans le cadre du paiement des opérations d'importation.

Le plus important à retenir pour ces crédits de fonctionnement, très prisés par les clients, est leur caractère extrêmement risqué, qui impose un dimensionnement rigoureux des lignes en fonction :

- Des besoins réels de l'entreprise
- De l'évolution de son besoin en fonds de roulement
- Du niveau de contribution des autres partenaires bancaires
- De l'historique de l'affaire
- De l'expérience et de la moralité du promoteur

- **Les financements d'investissement :**

Il s'agit des crédits mis à la disposition des entreprises pour financer leurs projets de création, d'extension, de modernisation ou de reconstitution du fonds de roulement.
Le développement de financements à moyen voire long terme constitue un ticket souvent indispensable pour fidéliser les clients existants et conquérir de nouveaux prospects.
En fait, le point principal pour une bonne appréhension des risques liés à ce type de financement réside dans notre capacité à vérifier le réalisme et la cohérence des études prévisionnelles proposées par les clients et à s'assurer ainsi de l'adéquation entre leur capacité de remboursement et la charge de la dette future. Il est également important de vérifier les garanties.
Parmi cette catégorie, Les principales formes utilisées au niveau de la banque sont :

- Le crédit à moyen terme (CMT) est destiné au financement des projets de création, d'extension et de modernisation des entreprises.

- Le crédit-bail ou Leasing est un moyen de financement des investissements de biens, d'équipements, de matériels et d'outillage (matériel roulant, engins de travaux publics, équipements industriels, matériel médical…). L'investissement en question est acquis par Société Générale Algérie qui reste propriétaire et le met à la disposition du client moyennant un loyer mensuel. En fin de contrat, il aura la possibilité d'acquérir cet investissement contre une valeur résiduelle.
- L'enveloppe cadre pour les prévisions d'investissements : une enveloppe de financements prédéterminée peut être mise en place et utilisable sur une période de 12 mois.

- **La couverture du risque de change CVaR :**

Il concerne principalement les opérations de change à terme, les swaps de change et de taux. Il commence le jour de la transaction et se termine à l'échéance de la transaction.

L'objectif étant de privilégier les crédits adossés et les engagements par signature sans pour autant sous-estimer les autres lignes (découvert, SPOT…) très rentables pour la banque tout en évitant un recours excessif à ce type de crédits. De même, le bon dosage des lignes doit permettre d'éviter un surdimensionnement des crédits par rapport aux besoins réels des clients, pouvant entrainer la banque dans le risque d'un soutien abusif. Pour autant, il convient de ne pas sous-estimer le besoin réel du client et, par conséquent, d'éviter les demandes ponctuelles ou une perte de courant d'affaires avec le client.

2.5. Le risque bancaire :

Définition :

Le risque est un événement aléatoire à date incertaine que les assureurs et banquiers essaient de minimiser par le biais de calculs mathématiques (calcul de probabilité/statistique) et de scoring.

« Un risque bancaire est un risque auquel s'expose un établissement bancaire lors d'une activité bancaire. L'activité bancaire, par son rôle d'intermédiation financière et ses services connexes, expose les établissements bancaires à de nombreux risques. » [5]

Typologies des risques bancaires :

Dans le domaine bancaire les principaux risques qu'on peut distinguer sont :

- Le risque de liquidité :

« Ce type de risque désigne l'insuffisance de liquidité bancaire pour faire face à ces besoins inattendus. En effet, ce risque peut conduire à la faillite de la banque suite à un mouvement de panique des déposants, qui peuvent demander leurs dépôts au même temps. Le recoure aux retraits massifs des fonds par les épargnants, ainsi que leurs inquiétudes sur la solvabilité de l'établissement bancaire, peut aggraver la situation de cette dernière et entraîne ce que nous appellons ‘'une crise de liquidité brutale'' » [6]

- Le risque du marché :

« Il correspond à la baisse de la valeur du portefeuille d'actifs (obligation, action, ...) détenu par la banque à la suite d'une évolution défavorable de la valeur des cours sur le marché, en d'autres termes ce risque provient de l'incertitude de gains résultant de changement dans les conditions

du marché. Ce type de risque découle principalement de l'instabilité des paramètres du marché (taux d'intérêt, indices boursiers et taux de change), d'où l'effet des marchés volatils, de la libéralisation, et des nouvelles technologies sont accompagnés par un accroissement remarquable de risque de marché. » [6]

- Le risque pays :

« La défaillance éventuelle du pays de l'acheteur. Il englobe l'ensemble des aléas pouvant affecter un investissement financier dans le déroulement de ses opérations en relation avec un pays dit ''à risque'', indépendamment de la qualité du débiteur ou du projet. » [4]

- Le risque de crédit :

Appelé aussi risque de contrepartie ou risque de défaut, c'est le principal risque qui menace le bien être des établissements de crédit, d'où il désigne le risque de défaut des clients ainsi que la dégradation de la situation financière d'un emprunteur face à ces obligations.

Selon (Godlewiski C. J.) « le risque de crédit peut être défini comme une non performance de la contrepartie engendrant une perte probable au niveau de la banque ». [6]

De plus ce risque dépend de la probabilité de défaillance de contrepartie que ce soit un pays, un particulier, une entreprise ou un établissement de crédit avec laquelle la banque est engagée.

- Le risque de solvabilité :

« Désigne l'insuffisance des fonds propres afin d'absorber les pertes éventuelles par la banque, en effet, ce risque ne découle pas uniquement d'un manque de fonds propres mais aussi des divers risques encourus par la banque tel que, le risque de crédit, du marché, du taux et de change. L'exposition des banques à ce type de risque peut mettre en danger son activité, d'où l'objectif recherché par les institutions financières c'est d'essayer d'ajuster les fonds propres aux risques afin de faire face à ce genre de risque d'insolvabilité. » [6]

Gestion des risques bancaires :

La gestion des risques bancaires regroupe l'ensemble des techniques permettant à une banque donnée de contenir ou de réduire le risque de perte financière dans ses activités, parmi les moyens de réduire ce risque c'est l'exigence des sûretés et assurances afin de couvrir la perte encourue par la banque en cas d'insolvabilité.

2.6. Les critères d'approbation d'un dossier de crédit :

L'approbation d'un dossier de crédit passe par 2 étapes essentielles qui sont :
Premièrement, les informations à recueillir où nous allons collecter tous les documents nécessaires pour l'instruction d'un dossier de crédit, retenir les critères qui permettent d'apprécier la qualité d'une contrepartie (âge de l'entreprise, situation financière, forme juridique …) ensuite recueillir et analyser les informations reprises sur les bases internes comme le fonctionnement du compte courant, la prise des garanties, la note attribuée sur STARWEB… et enfin le recueille des informations qualitatives via la procédure KYC.

Deuxièmement l'analyse de l'information financière elle consiste à collecter et traiter l'information financière afin de permettre au chargé de dossier d'apprécier la santé financière de l'entreprise emprunteuse ; pour ce faire il convient d'observer son évolution sur au moins 3 exercices en analysant les ratios présents sur les notices financières (compte de résultat et le bilan) et les comparer aux normes arrêtés par SGA.

Concernant les détails de cette procédure (Information à recueillir et analyse de l'information financières) sont en **Annexe C**

3. 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'améliorations au sein de l'entreprise.

3.1. 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 des deux départements (Risque et Commerciale) ; étudié les différentes procédures internes de l'entreprise « Business Understanding » pour comprendre le fonctionnement détaillé de la banque, pour ensuite étudié les différentes sources de données qui nous ont été fournies (BDD transactionnel, les outils : DCCIT, Work Flow et Starweb).

3.2. Entretiens :

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, vue que la problématique proposée initialement était très vaste (Data mining et profilage client corporate), ce qui nous a amené à procéder par le biais de brainstorming afin de mieux déceler un axe d'amélioration pertinent pour l'état actuel de la banque.

L'idée de départ était de travailler sur les crédits Corporate, car ce type de crédit est le plus répandue au sein de la banque comme nous pouvons le constater dans l'histogramme suivant qui représente le nombre de crédits octroyés selon le segment des clients depuis l'an 2000, et il représente selon les statistiques internes de la banque la majeure partie de son profit.

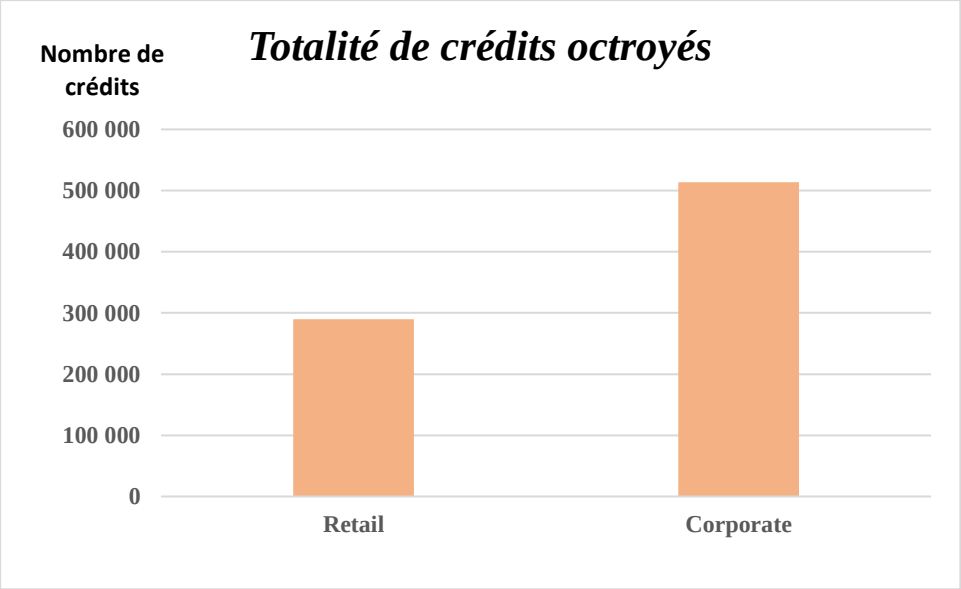


Figure 6 : Histogramme du nombre de crédits octroyés de 2006 à 2020

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) pour cela nous n'avions pas pu faire des extractions directement de la BDD de la banque pour des raisons de confidentialité, mais ils nous ont fourni les dataset (fichiers Excel format ***''csv''***) nécessaires pour faire notre analyse ; après cela nous comptions généraliser ce modèle et l'ajuster sur le reste des types de crédit. Mais après un brainstorming effectué avec l'analyste risque Mr. Bakir qui nous a expliqué que cette démarche n'est pas faisable au sein de SGA car nous pouvons 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.

Nous avons aussi effectué d'autres entretiens avec des employés de la banque, parmi eux y'avait l'adjoint directeur du crédit leasing, qui nous avait présenté le crédit, son activité, l'état actuel de ce dernier en Algérie qui n'arrête pas de se dégrader et auquel la banque devrait faire face pour remonter la pente et les différentes mesures qu'elle a mis en place mais qui n'ont pas encore porter leurs résultats à cause de quelques obstacles au sein de la banque que nous allons aborder ultérieurement.

3.3. Etude des sources de données :

Cette étape du diagnostic était un peu délicate dans les premiers temps à cause des règlementations de confidentialités internes de la banque, ainsi l'obtention des données a été faites en passant par notre encadreur (auditrice interne) qui faisait le requêtage de la BDD de la banque et nous envoyait les tables demandés (fichier csv), ces fichiers étaient volumineux, complexes à manier, difficiles à comprendre et à chaque fois nous revenions au dictionnaire des données de la banque.

Les fichiers les plus importants lors de notre travail sont les suivants :

a) La table de tous les clients de la banque (Corporate et Retail)

b) La tables des impayés (les clients qu'ont été considéré comme déclassés)

c) La table des garanties recueillies

Après l'étude des fichiers précédents nous avons constaté que la quantité d'information n'était pas suffisante pour mettre en place un modèle précis, car les informations et les ratios financiers des clients étaient sauvegardés comme pièce jointe dans l'outil DCCIT, et nous devions extraire les ratios et les données nécessaires pour notre étude, et ainsi constitué une nouvelle BDD contenant les informations présentes dans les notices financières pour pouvoir mener notre étude d'une manière optimale.

3.4. Identification des utilisateurs du modèle :

L'étude du risque d'un dossier de crédit se fait principalement au niveau des deux départements Risk et marketing, mais dans tous les cas c'est l'analyste qui s'en charge d'une manière très concrète de cette tâche, donc il va être le principal utilisateur de ce modèle qui va l'accoter durant sa prise de décision (un outil d'aide à la décision).

3.5. Difficultés rencontrées :

Lors de cette phase, nous avons rencontré les difficultés suivantes :

a) La confidentialité très élevée au sein de la banque (données, outils demandés, procédures…)

b) Non disponibilité des responsables et des employeurs

c) Données complexes et multitude des outils informatiques de gestion.

d) Manque d'informations décrivant le savoir utilisé lors de l'étude du risque d'un dossier de crédit comme c'est un savoir propre à l'analyste risque lui-même.

3.6. Critique de l'environnement existant :

Après plus d'un mois de diagnostique au sein de SGA (procédures internes, entretien, observations du travail des analystes risques…) nous avons pu constater les axes d'améliorations suivants :

a) Automatisation de la collecte des garanties :

La collecte des garanties ne se fait pas d'une manière automatique, c'est-à-dire à chaque fois qu'un client arrive sur place et selon le crédit vers lequel il serait orienté, les garanties se collectent toujours d'une manière manuelle (une check-list) à remplir par l'agent commercial selon les possessions du client, nous avons pensé à automatiser cette tâche par un modèle qui en se basant sur les garanties déposées par les anciens clients qu'ont le même profil et qu'ont bénéficié du même type de crédit afin de proposer les garanties nécessaires d'une manière automatique, ce qui va permettre de :

- Automatiser la tâche répétitive ;
- Faciliter et assurer la remontée instantanée de l'information vers le système d'information ;
- Proposé d'une manière automatique les garanties demandées sans devoir passé par un questionnaire et une check-list.

La figure suivante montre une schématisation du nouveau process :

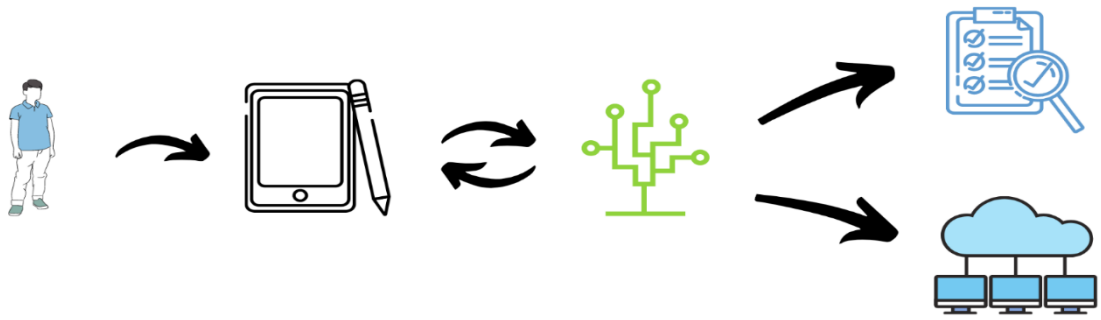


Figure 7: Schématisation du nouveau processus de traitement des garanties

En premier lieu le client va introduire les informations nécessaires sur une tablette qu'il va trouver au niveau des agences commerciales de la banque où il fait sa demande de crédit (ça pourrait être une fonctionnalité à ajouter sur l'application de la banque APPLI SGA), ces informations sont : nom, prénom, âge, secteur d'activité, type de crédit… Après cela ces informations vont être stocker et traiter par le modèle afin de proposer les garanties nécessaires à ce type de crédit et selon le profil du client d'une manière automatique, tout en stockant ces informations (données du client et garanties proposées).

b) Analyse sectorielle annuelle :

Le marché bancaire en Algérie est sans doute dominé par les banques publics 77% contre seulement 23% de banques privées (en 2015), ce qui restreint les secteurs d'activités de ces

derniers, donc ils essayent d'en tirer profit et de s'accentuer sur tous les secteurs possibles sans pour autant désigner un secteur d'activité mère « Océan Bleu » sur lequel se concentrer chaque année et fournissent plus d'efforts dans celui-ci pour assurer une croissance de profit régulière et continue, ce qui nous a amené à penser qu'une étude sectorielle (secteur d'activité des clients SGA) pour mettre en lumière quel secteur leur rapporte le plus et celui qui leur rapporte le moins chaque année, tout en déterminant les éléments économiques déclencheurs de cette croissance/baisse à fin de pouvoir déterminer les patterns qu'ont précédé à ces éléments et ainsi prévoir les prochaines fluctuations dans les secteurs dans laquelle elle exerce.

Malheureusement, cet axe d'amélioration ne pourrait pas voir le jour pour les raisons suivantes: le marché algérien et son état économique connait un changement continue de jour en jour et en plus de ça il n'existe pas des sources d'informations à son propos pour pouvoir faire le suivi et détecter les fluctuations prochaines ; la deuxième raison est que SGA étant une filiale du groupe Société Générale elle n'a pas la main libre sur sa stratégie et ses objectifs annuelles (stratégie Business Unit).

c) Le Leasing :

Le crédit-bail (leasing) est un des modes de financement récemment introduit en Algérie (1996), sa taille du marché a été estimée à 162 milliards de dinars à la fin de 2020 contre 135 milliards de dinars afin de 2019. Nous observons donc une croissance de 20 % en 2020, et ce, malgré des conditions économiques et sanitaires difficiles. Société générale Algérie a voulu encore mettre la barre plus haut dans ce segment de marché en sortant de l'ordinaire, c'est-à-dire les secteurs types concernés par le leasing (BTP, l'industrie, les véhicules et équipements…), elle a donc signé plusieurs conventions avec divers fournisseurs pour lancer le crédit-bail dans les secteurs médicales et électroniques, des secteurs abondants et avec peu de concurrence.

Le problème auquel SGA fait face, est que le crédit-bail est un type particulier de crédit moins risqué que les autres types de crédit pour les corporate pour plusieurs raisons qui sont :

- L'équipement procuré aux clients restera toujours la propriété de la banque durant tout le contrat avec des options à la fin (achat pour un prix symbolique, renouvellement du contrat ou renoncement) la plupart des clients optent pour la première option car l'achat va se faire à 1 DA symbolique, ce qui veut dire que dans tous les cas la banque est propriétaire du bien et peut le récupérer en cas de problème, et pour faire face au problème que le bien perd sa valeur initiale lorsqu'il sera déjà utilisé et récupéré en cas de problème, la banque impose un premier loyer majoré aux clients (30% du montant de l'équipement).
- Le contrat entre SGA et ses fournisseurs qui leurs imposent de racheter l'équipement en cas de désistement ou de problème de la part du client, si le bien en question est un objet technique et non pas un produit, en d'autres termes si le bien ne pourra pas se revendre dans le marché algérien en raison de sa complexité par exemple : matériel médical sophistiqué, la machine Evcon de production d'eau de Cevital…

Ce qui implique que ce crédit ne nécessite pas une étude approfondie d'un point de vu risque comme les autres types de financement proposés par la banque, car en signant de nouvelles conventions avec de nouveaux fournisseurs, son objectif est de vendre le maximum possible des contrats crédit-bail et ce n'est pas le cas, car les analystes risques abordent le Leasing de la même façon que les autres crédit et étudient minutieusement l'informations financières de la même sorte qu'ils le font avec le CMT, SPOT, ASF… ce qui freine l'atteinte des objectifs fixés par le directeur du Leasing.

d) L'analyse risque d'un dossier crédit :

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 SGA (code vert, orange, rouge), ces indicateurs se situent sur les notices financières (compte de résultat + bilan) que nous trouvons sur le DCCIT sous forme de fichiers Excel en pièce jointe, 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.

Dans cette démarche nous avons pu constater les goulots suivants :

- L' étude de ces ratios se fait d'une manière standard et répétitif pour chaque clients, en se référant à l'échelle d'acceptations internes de SGA, mais nous avons constaté que des clients ont été déclassé malgré le fait que leurs indicateurs financiers étaient en bonne santé, c'est-à-dire dans les normes et acceptables (code vert), ce qui signifie l'existence de nouvelles patterns qui n'ont pas étaient pris en considération par la banque, des ratios auxquels les analystes ne donnaient pas beaucoup d'importance par rapport aux autres métriques présentent sur la notice financière ; et pour pouvoir détecter si vraiment ces patterns entre les profils des clients existent, la participation de l'apprentissage automatique est nécessaire.
- Un travail manuel sur une quantité d'informations volumineuse (la notice financière est composé de 9 feuilles Excel et chaque feuille comporte plusieurs ratio) 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.

3.7. Enoncé de la problématique :

Parmi les critiques de l'environnement existant vu dans le point précédent, la problématique choisie 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 et non pas un seul, ce qui explique pourquoi le problème du leasing a été laissé en attente pour le moment ; la deuxième raison est que pour pouvoir mener ce projet d'une manière efficace ils nous est obligé de procéder à l'enrichissement de la BDD existante de la banque en extrayant les données présentes sur les notices financières les regrouper sur des fichiers format **''csv''** à fin de constituer une nouvelle BDD complémentaire à l'ancienne.

Ainsi, le travail que nous réaliserons dans le cadre de ce projet visera à développer un outil d'aide à la décision permettant d'orienter de manière optimale les analystes risques vers les ratios et indicateurs optimaux selon le profil du client pour une prise de décision plus rapide, adéquate et justifiée et ainsi une amélioration saillante dans la détection des red flags.

3.8. 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.

L'étude de l'environnement existant et des besoins exprimés nous a amenés à révéler plusieurs axes d'améliorations, les encadrer afin de proposer une solution adaptée à la problématique exposée.

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 02 :
Etat de l’art

Chapitre 2 : Etat de l'art

Introduction :

Le présent chapitre portera sur les aspects théoriques des différents concepts et terminologies utilisés dans ce travail. En effet, il existe une multitude de définitions et de typologies qui peuvent être considérées en vue de la définition d'un concept, d'un processus ou même en amont de la détermination d'une problématique liée à un thème particulier.

Par ailleurs, la revue bibliographique nous a permis de prendre connaissance des réalisations déjà effectuées dans le cadre du contexte envisagé, et de mieux appréhender les défis auxquels sont confrontés les acteurs du même domaine, afin de faire preuve d'originalité dans la contribution qui résultera de ce travail.

A. ÉTUDE BIBLIOGRAPHIQUE :

1. 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 Du Jardin (2010) [7] 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 (Virág, 2004) [8] La prévision des faillites est considérée comme une discipline limitative entre la finance d'entreprise et les statistiques, qui tente de prédire la solvabilité future des entreprises en utilisant des ratios financiers comme variables explicatives en appliquant des méthodes multivariées (Nyitrai, 2015) [9]

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 (Fitzpatrick) [10]. La première percée méthodologique s'est produite lorsque Durand [11] 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 Beaver (1966) [12].

En réalisant que la classification des observations à l'aide d'une variable ne fournit pas un résultat fiable, Myers et Forgy (1963) [13] ont appliqué l'analyse de régression multivariée et l'DA pour élaborer un système de notation de crédit 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 d'Altman (1968) [14], qui a été en mesure de classer les entreprises dans l'échantillon avec 95 % de précision de classification. Depuis sa première publication, le odèle 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 (Nyitrai 2015a) [9].

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 Chesser (1974) [15] sur une base de données de risque de crédit. Dans la distribution mondiale de logit, la publication d'Ohlson (1980) [16] 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 (Zmijewski 1984) [17].

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 Frydman et al. (1985) [18].

Les années 1990 ont posé de nouveaux défis aux spécialistes et aux praticiens de la prévision des faillites. 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 . Les NN ont été appliqués pour la première fois à la solvabilité des clients par Odom et Sharda (1990) [19]. Les auteurs ont prouvé que les performances des réseaux de backpropagation à trois couches surpassaient les résultats des méthodes antérieures. Depuis lors, les NNs ont été largement distribués, ont connu des développements importants 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 (Vlachos et Tolias, 2003) [20]. 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 (Fan et J. Risk Financial Manag.) [21] En outre, les méthodes de rough set theory (RST) (Dimitras et al. 1999), k Nearest Neighbors (KNN) (Ardakhani et al. 2016) [22], les réseaux de Bayes , les algorithmes génétiques (GA) (Lensberg et al. 2006) [23], la quantification des vecteurs d'apprentissage (LVQ) (Neves et Vieira 2016) et le raisonnement fondé sur des cas (CBR) (Bryant, 1997) [24] 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 (Marqués et al. 2012) [25]. 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, random subspace, random forest, Gauss-processes et autoencoder appartenant à la famille des procédures d'apprentissage automatique (Nyitrai 2015a, Wang 2017) [26] [9]. Les recherches actuelles sur les prévisions de faillite sont sans ambiguïté dominées par la machine learning, data mining, l'intelligence artificielle et le hybrid modelling par la combinaison créative de différentes nouvelles méthodes (Barboza et al. 2017) [27]. 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.

2. Une vue générale sur les articles étudiés :

A travers ce chapitre nous avons veillé à faire une lecture approfondie et pertinente des articles scientifiques de valeurs, nous avons essayé une intervalle de temps qui représente l'année de l'édition, allant de 1997 jusqu'à l'année 2020. Et à travers ce tableau ci-dessous vous allez trouver les différentes clés qui caractérisent chacun des articles revus.

Tableau 2 : Les différents mots-clés dans les articles scientifiques étudiés

	Année	Auteurs	Mots clé
Utilizing machine learning for improved bankruptcy predictions in the Norwegian market with an emphasis on financial, management and sector statements	2019	Eystein Nordby Meese, Torbjørn Viken	Bankruptcy Prediction, Norwegian Markets, SVM, Random Forest, Neural Network, Confusion Matrix, Multiyear Model, Financial-Management- and Sector Statements
A Comprehensive Review of Corporate Bankruptcy Prediction in Hungary	2020	Tamás Kristóf, Miklós Virág	bankruptcy prediction; classification; credit risk modelling; corporate failure; rating systems
MACHINE LEARNING MODELS FOR PREDICTING FINANCIAL DISTRESS	2018	Joseph BONELLO, Xavier BRÉDART, Vanessa VELLA,	Financial Distress, Financial Ratios, Decision Trees, Naïve Bayes, Neural Networks, Previous-Year Comparison, Accuracy.
Machine learning models and bankruptcy prediction	2017	Flavio Barboza , Herbert Kimura , Edward Altman	Bankruptcy prediction, Machine learning, Support vector machines, Boosting, Bagging ,Random forest

3. Les modèles et les données utilisés :

Le tableau ci-dessous présente les différentes données utilisées dans les articles revus, ainsi que différent algorithme d'apprentissage adopté.

Tableau 3 : Données et algorithmes d'apprentissage utilisés dans les articles étudiés

Titre	Données	Algorithmes Utilisés
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
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, l'ensemble de données EDGAR classe les entreprises en échec ou en activité.	Decision Tree, Naive Bayes, ANN
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, DecisionTree, ANN.
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

4. Résultat :

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

Tableau 4 *:* Tableau comparatif des résultats obtenus dans chaque article

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%

5. Discussion :

Tout au long de cette étude nous avons pu extraire en sus des nouvelles connaissances des nouvelles approches et méthodes pour réussir sa stratégie numérique. Dans l'ensemble des articles que nous avons choisi, la base de données utilisée appartenait à différentes banques à travers le monde ce qui a rendu ce travail assez fiable en termes de multiplicités de données et de méthodes utilisées. Par ailleurs les résultats obtenus ont révélé que certains algorithmes ou techniques du Machine Learning peuvent servir efficacement à prédire la solvabilité des clients. 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 Decision Tree a été préféré selon l'étude qu'a été menée 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 nous pouvons conclure que les techniques du Machine Learning peuvent renforcer et accélérer le processus d'analyse de solvabilité des clients en adoptant les méthodes adéquates et en explorant les données nécessaires pour réussir sa stratégie numérique.

B. NOTIONS FONDAMENTALES :

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 donnes présent au sein des organisations, ceci a rendu la prise de décision de plus en plus complexe, d'où la nécessité du machine learning 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

1. 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.

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. 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.

2. Machine Learning :

2.1. Introduction :

Arthur Samuel un informaticien d'IBM a écrit un programme informatique pour jouer aux dames en 1959. Même si les règles qui régissent le jeu sont assez peu complexes, il faut savoir qu'il existe des milliards de situations possibles et il est par conséquent quasiment impossible de programmer explicitement l'ordinateur avec des instructions sur ce qu'il faut faire dans chaque situation. Au début, les scores étaient basés sur une formule utilisant des facteurs tels que le nombre de pièces de chaque côté et le nombre de rois. Cette approche fonctionnait, cependant Samuel avait déjà une vision de comment améliorer ses performances. Il a fait jouer le programme des milliers de parties contre lui-même et a utilisé les résultats (victoire ou défaite) pour évaluer la probabilité de gagner en effectuant un mouvement donné.

Arthur avait donc écrit un programme informatique capable d'améliorer ses propres performances grâce à l'expérience et c'est ainsi que l'apprentissage automatique a vu le jour.

2.2. Définition et typologie :

L'apprentissage automatique (machine learning) est une branche de l'intelligence artificielle (I.A) 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.

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.

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.

2.3. 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 le surajustement.
- L'ensemble de test : il est constitué de candidats sur lesquels sera appliqué le modèle d'apprentissage pour tester et corriger l'algorithme.

Méthodologie :

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

2.4. Algorithmes de Machine Learning :

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.

2.4.1. Decision Tree (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".

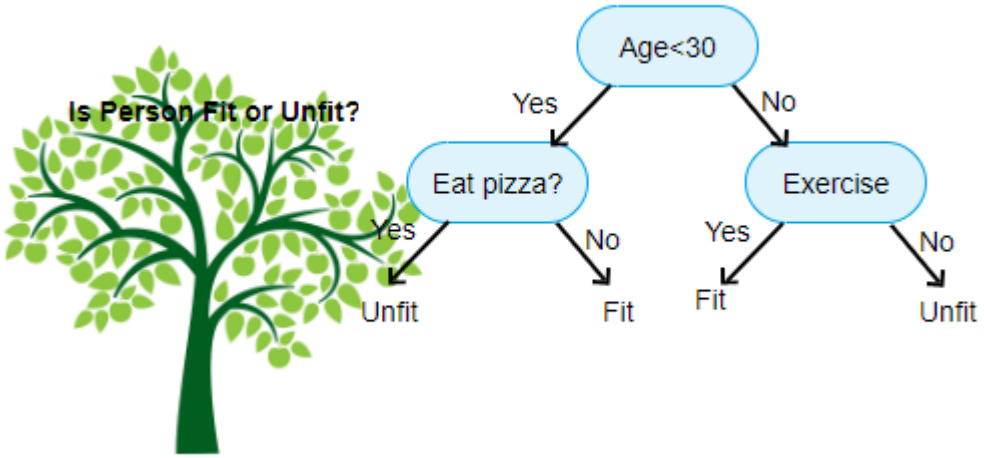


Figure 8 : Arbre de décision pour classification fit/unfit

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. Nous disons alors que les arbres produisent des estimateurs de variance élevée.

2.4.2. Random Forest :

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.

- **Principe de l'algorithme :**

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

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

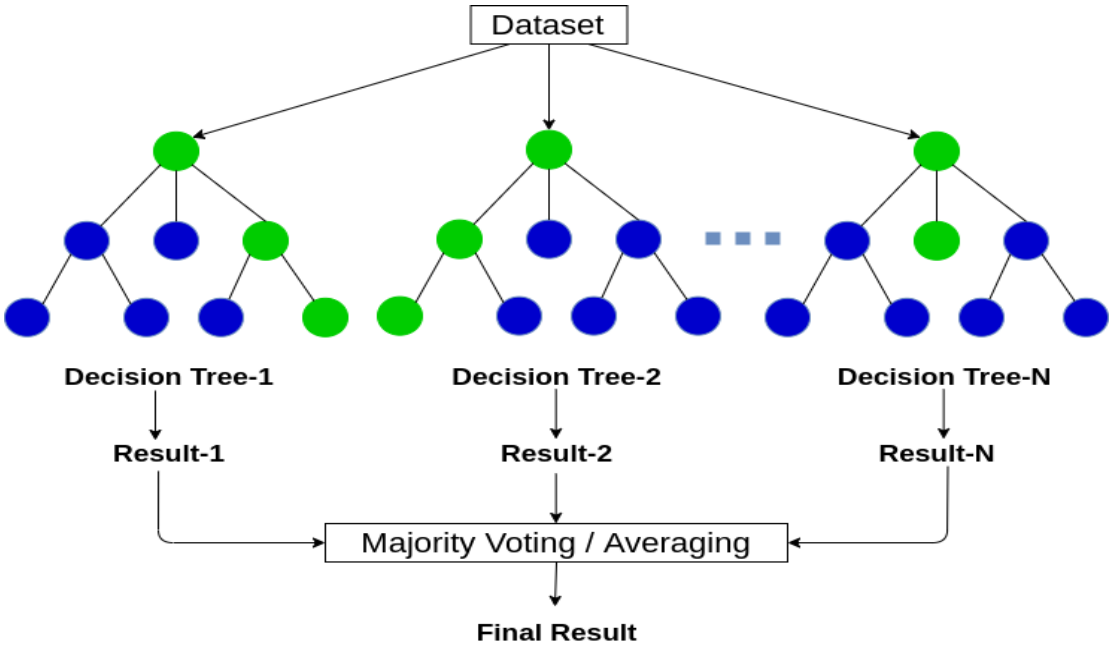


Figure 9: Exemple illustrant le déroulement de Random Forest

2.4.3. 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.

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

$$\mathbf{P} = \frac{\mathbf{1}}{\mathbf{1} + \mathbf{e}^{-\mathbf{ywx}}}$$

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.

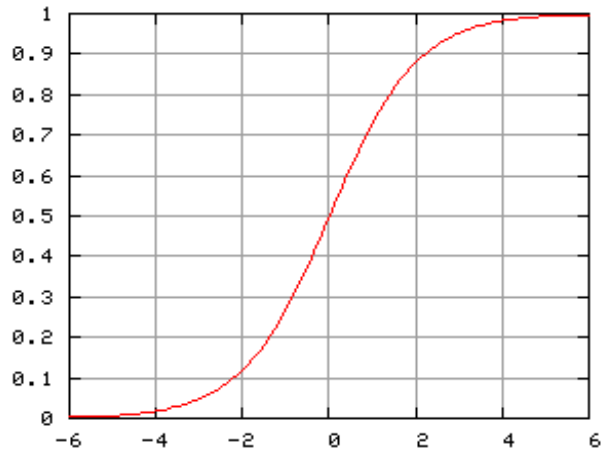


Figure 10 : Variations de la fonction logistique

2.4.4. Support Vector Machine (SVM) :

La machine à vecteurs de support (SVM) a été introduite par Vapnik (1995). 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 surajustement 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.

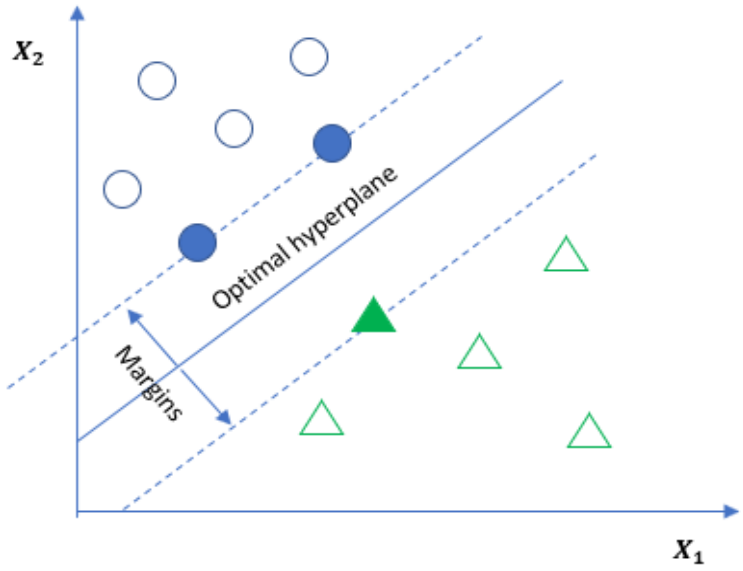


Figure 11 : Fonctionnement du modèle Support Vector Machine

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'entraînement très long, ce qui fait qu'elles ne conviennent pas à des jeux de données volumineux.

2.4.5. 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.

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.

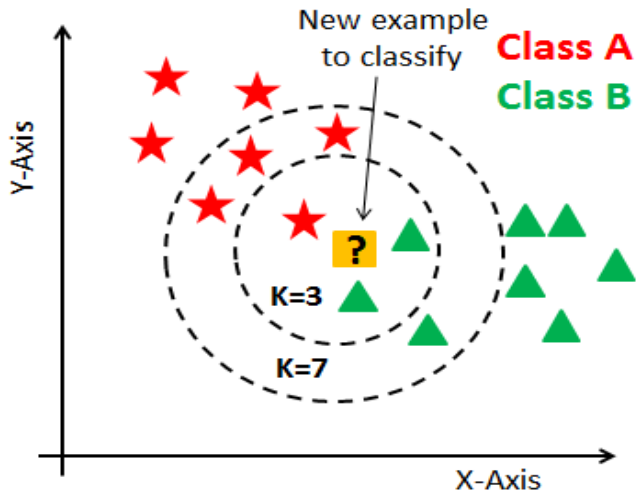


Figure 12 : Exemple de classification KNN (K=3 et K=5)

2.4. 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.

2.5. Mesure de performance :

2.5.1. 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).

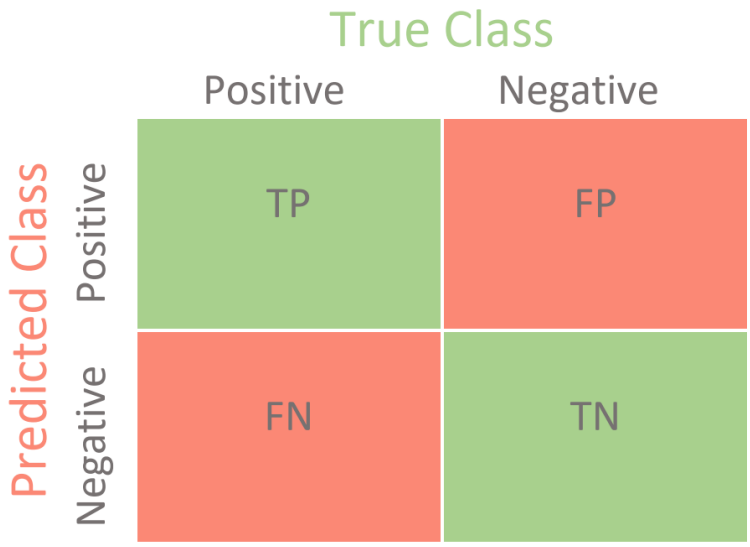


Figure 13 : Confusion Matrix

- TP (True Positive) : Vrai positifs, c'est-à-dire les cas où la prédiction est positive, et la valeur réelle est aussi positive.
- TN (True Negative) : 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.

2.5.2. Métriques d'évaluation :

Maintenant que nous avons les 4 valeurs TP, TN, FP, FN qui sont nécessaire mais pas suffisante pour juger un modèle, c'est pour ça qu'on devrait calculer d'autres métriques qui nous permettront de mesurer la performance du modèle comme il se doit.

- **TPR (True Positive Rate) - La Sensibilité** :

C'est le taux des vrais positifs, c'est-à-dire la proportion de cas positifs correctement identifiés, calculée par la formule suivante :

$$\boldsymbol{TPR = \frac{TP}{TP + FN}}$$

- **TNR (True Negative Rate) - La Spécificité** :

C'est le taux des vrais négatifs, c'est-à-dire la proportion de cas négatifs correctement identifiés, calculée par la formule suivante :

$$\boldsymbol{TNR = \frac{TN}{TN + FP}}$$

- **FPR (False Positif Rate) – Erreur du Type I** :

C'est le taux des faux positifs, c'est-à-dire la proportion de cas négatifs non identifiés ou en d'autres termes c'est les cas négatifs qui ont été classé incorrectement comme cas positifs, calculée par la formule suivante :

$$\boldsymbol{FPR = \frac{FP}{TN + FP}}$$

- **FNR (False Negative Rate) – Erreur du Type II**:

C'est le taux des faux négatifs, c'est-à-dire la proportion de cas positifs non identifiés ou en d'autres termes c'est les cas positifs qui ont été classé incorrectement comme cas négatifs, calculée par la formule suivante :

$$\boldsymbol{FNR = \frac{FN}{TP + FN}}$$

- **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 :

$$\textbf{Précision} = \boldsymbol{\frac{TP}{TP + FP}}$$

- **F1-Score (Le F-mesure) :**

Il est utilisé pour mesurer la qualité du test. Il s'agit d'une moyenne pondérée de la précision et de la sensibilité. Lorsque le score F1 est de 1, c'est le meilleur score, tandis que lorsqu'il est égal à 0, c'est le pire.

La précision et la sensibilité se font généralement au détriment l'une de l'autre. Autrement dit, une grande précision se fait au détriment de la sensibilité, et une grande sensibilité au détriment de la précision. Le modèle idéal doit présenter une sensibilité et une précision élevées, et c'est

là que le F1-Score intervient pour trouver l'harmonie idéal entre ces deux métriques pour avoir un modèle robuste.

Le F1-Score est calculé par la formule suivante :

$$\mathbf{F1-Score} = \frac{\mathbf{2} \times (\boldsymbol{pr}\acute{\mathbf{e}}\boldsymbol{cision} \times \boldsymbol{sensibilit}\acute{\mathbf{e}})}{(\boldsymbol{pr}\acute{\mathbf{e}}\boldsymbol{cision} + \boldsymbol{sensibilit}\acute{\mathbf{e}})}$$

2.5.3. Courbes 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. » [29]

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 $\alpha(S) = prob(score(x) \geq S \setminus x = évènement)$, c'est-à-dire bien détecter un événement au seuil S.
- La spécificité = $prob(score(x) \leq S \setminus x = non\text{-}évènement)$, 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 :
$$\alpha(S) - 1 - \beta(S) = 1$$

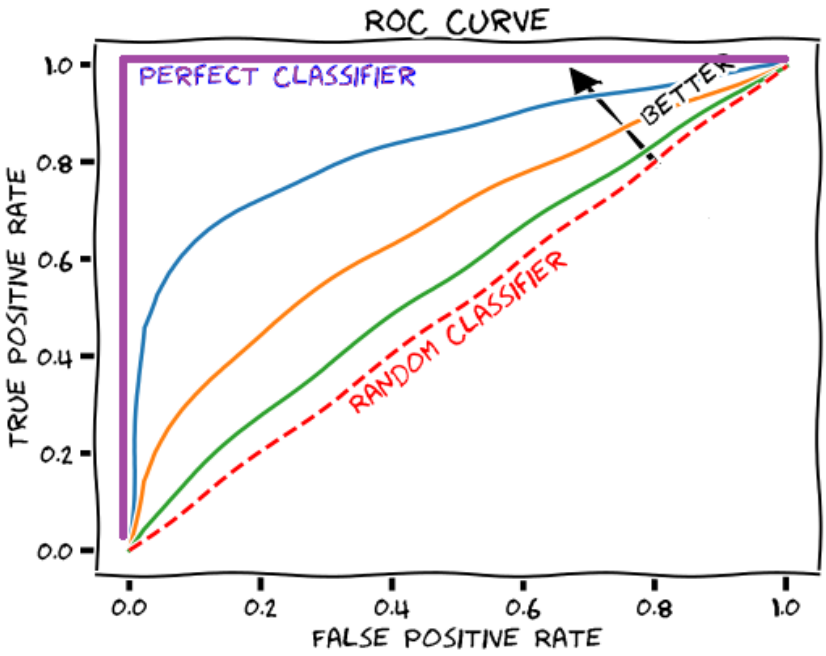


Figure 14 : Courbe ROC

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.

2.5.4. 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.

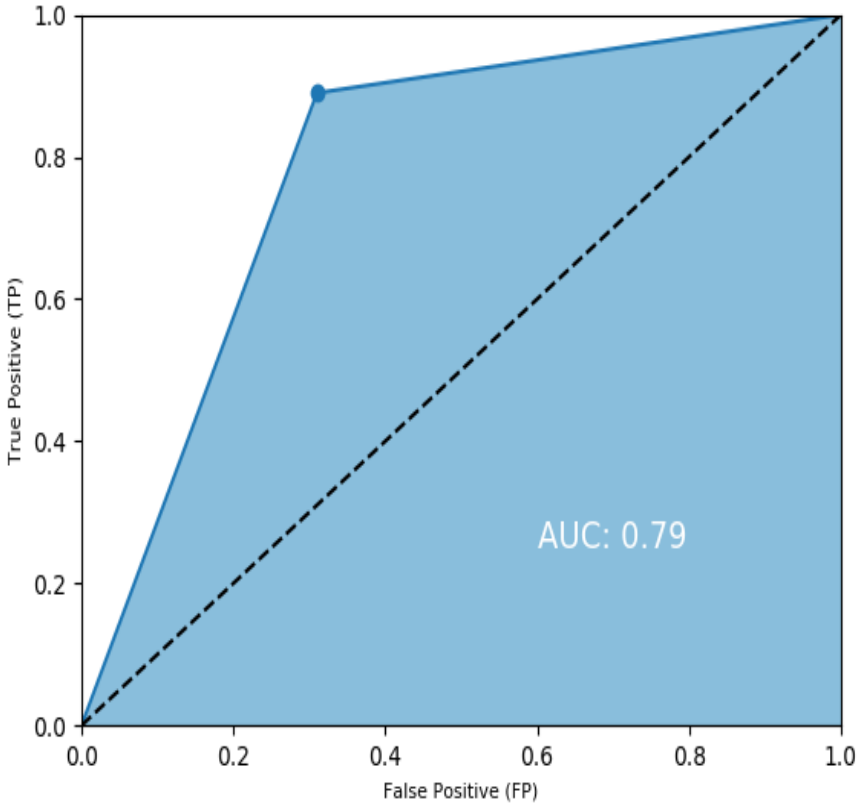


Figure 15 : *AUC*

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.

2.5.5. Cross Validation (CV) :

La validation croisée (CV) est importante pour garantir la bonne validité des modèles créés. Cependant, elle n'a pas toujours été aussi importante ou réalisable. Les premières adoptions consistaient généralement à tester et à entraîner le modèle sur les mêmes données, car les informations et la puissance de calcul étaient rares. Il s'agit là d'un cas classique de surajustement du modèle ou, en d'autres termes, l'adaptation du modèle au point que les nouvelles observations introduites seraient très probablement faussement classées. La première tentative de validation croisée primitive a été introduite avec la division stricte entre les données de test et de formation. Cela réduit le problème de l'overfitting et crée donc des modèles plus robustes qui conservent un pouvoir prédictif en dehors de l'environnement d'apprentissage. Cette méthode est encore utilisée aujourd'hui, mais des méthodes plus avancées de CV sont souvent préférées.

La méthode utiliser pour la validation croisée de notre modèle est k-fold cross-validation

La validation croisée à *k* blocs, « *k*-fold cross-validation » : elle consiste à diviser l'échantillon original en *k* échantillons (ou « blocs »), puis sélectionner un des *k* échantillons comme ensemble de validation pendant que les *k-1* autres échantillons constituent l'ensemble d'apprentissage. Après apprentissage, on peut calculer la performance de validation déjà vu auparavant. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les blocs prédéfinis. À l'issue de la procédure nous obtenons ainsi *k* scores de performances, un par bloc. La moyenne et l'écart type des *k* scores de performances peuvent être calculés pour estimer le biais et la variance (expliqués en annexe) de la performance de validation. [30]

Le tableau suivant illustre le fonctionnement d'une validation croisée à 3 blocs :

Tableau 5 : Le fonctionnement d'une validation croisée à 3 blocs

K	Bloc 1	Bloc 2	Bloc 3
1	Validation	Apprentissage	Apprentissage
2	Apprentissage	Validation	Apprentissage
3	Apprentissage	Apprentissage	Validation

3. Logiciels utilisés :

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

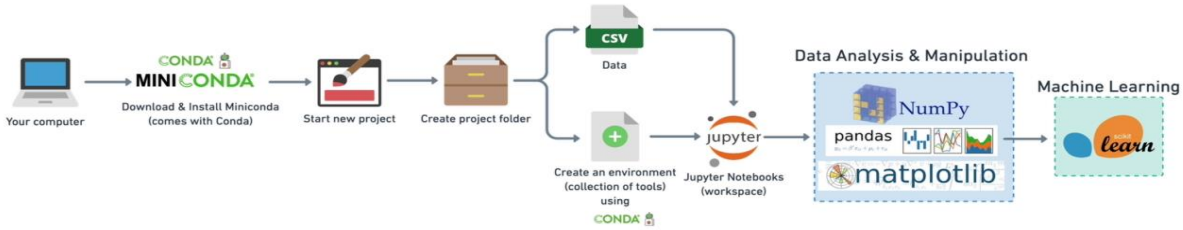


Figure 16: Logiciels utilisés lors du Datamining

3.3. 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 "conda install" est utilisée pour installer plus de 720 packages conda supplémentaires à partir du référentiel Anaconda. [33]

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

```
conda create -n monenv jupyter pandas numpy matplotlib scikit-learn
```

3.4. 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". [34]

3.5. 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". [35]

3.6. 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.

3.7. 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

3.8. 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 Random forests, l'amplification de gradient et k-means. Elle est conçue pour interagir avec les bibliothèques numériques et scientifiques Python NumPy et SciPy. Certains problèmes sont résolus rapidement, tandis que d'autres sont plus complexes et doivent être traités séparément comme de vrais projets. Cependant, il existe de nombreuses façons de résoudre ces problèmes complexes, mais elles suivent généralement le même processus (c'est-à-dire définir le problème, trouver la cause, trouver et mettre en œuvre la solution, analyser et assurer le suivi).

Pour notre travail le choix d'une méthodologie ou d'une approche de traitement des problèmes de Datamining nous aidera pour mieux orienter et organiser les phases d'exploration des données. Donc notre choix c'est orienter vers la méthodologie CRISP-DM.

4. Methodology CRISP – DM ''Cross Industry Standard Process for Data Mining'':

CRISP-DM, qui signifie Cross-Industry Standard Process for Data Mining est une méthode qui permet d'orienter les travaux d'exploration de données, en tant que méthodologie elle comprend des descriptions des phases typiques d'un projet et les taches dans chaque phase. En tant qu'un modèle de processus elle nous offres un aperçu du cycle de vie de l'exploration de données.

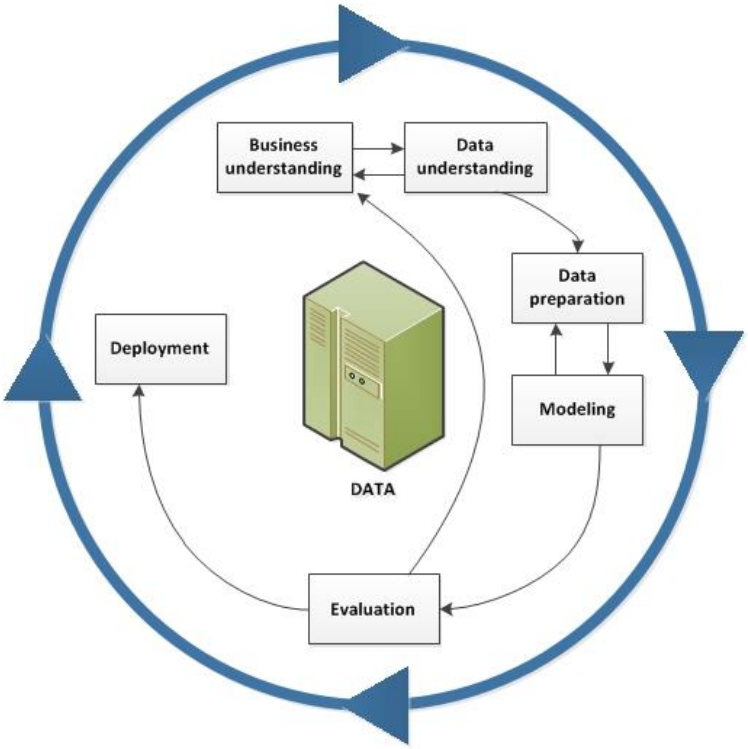


Figure 17: Le cycle de vie de l'exploration des données

4.1. Histoire de la méthode :

« La méthode CRISP-DM est conçue en 1996. En 1997, elle se développe en tant que projet de l'Union européenne financé par le programme ESPRIT. Le projet est conduit par quatre compagnies : ISL, NCR Corporation, Daimler-Benz et OHRA. Ce cœur du consortium apporte différentes expériences au projet : ISL, plus tard acquis et intégré dans SPSS Inc. produit ses progiciels d'analyse prédictive du même nom, intégré de nos jours au groupe IBM. Le géant informatique NCR Corporation créa la division Teradata spécialisée dans les entrepôts de données et son propre progiciel de data mining. Daimler-Benz avait une importante équipe de data miners. OHRA, une compagnie d'assurance, venait juste de commencer à explorer le potentiel d'utilisation du data mining. La première version de la méthode fut publiée sous le numéro de version CRISP-DM 1.09 en 1999. » [36]

4.2. Les 6 étapes de la méthode CRISP – DM :

4.2.1. Business Understanding :

La phase de compréhension du métier se concentre sur la compréhension des objectifs et des exigences du projet. Les tâches de cette phase sont des activités fondamentales de gestion de projet qui sont universelles à la plupart des projets :

- Déterminer les objectifs commerciaux : Il faut d'abord « comprendre en profondeur, du point de vue de l'entreprise, ce que le client veut vraiment accomplir ». Puis définir les critères de réussite de l'entreprise.
- Évaluer la situation : Déterminez la disponibilité des ressources, les exigences du projet, évaluez les risques et les imprévus, et effectuez une analyse coûts-avantages.
- Déterminer les objectifs de l'exploration des données : En plus de définir les objectifs de l'entreprise, il faut définir ce à quoi ressemble le succès d'un point de vue technique de l'exploration de données.
- Produire un plan de projet : Sélectionnez les technologies et les outils et définissez des plans détaillés pour chaque phase du projet.

Bien que de nombreuses équipes de projet se dépêchent de passer par cette phase mais établir une solide compréhension du métier est comme construire les fondations d'une maison, c'est absolument indispensable.

4.2.2. Data Understanding :

La phase de compréhension des données. S'ajoutant à la base de la compréhension du métier, elle permet d'identifier, de collecter et d'analyser les ensembles de données qui peuvent vous aider à atteindre les objectifs du projet. Cette phase comporte également quatre tâches :

- Collecter les données initiales : Acquérir les données nécessaires et (si nécessaire) les charger dans votre outil d'analyse.
- Décrire les données : Examinez les données et documentez leurs propriétés de surface, comme le format des données, le nombre d'enregistrements ou l'identité des champs.
- Explorer les données : Creusez plus profondément dans les données. Interrogez-les, visualisez-les et identifiez les relations entre les données.
- Vérifiez la qualité des données : les données sont-elles propres ou sales ? Documentez tout problème de qualité.

4.2.3. Data Preparation :

Cette phase, souvent appelée « broyage des données », prépare le/les ensembles de données finaux pour la modélisation. Elle comporte cinq tâches :

- Sélectionner les données : Déterminer les ensembles de données qui seront utilisés et documenter les raisons de leur inclusion/exclusion.
- Nettoyer les données : C'est souvent la tâche la plus longue. Sans elle, vous risquez d'être victime du phénomène de « Garbage-in, Garbage-out ». Une pratique courante au cours de cette tâche consiste à corriger, imputer ou supprimer les valeurs erronées.

- Construire des données : Déterminez de nouveaux attributs qui seront utiles. Par exemple, déduisez l'indice de masse corporelle d'une personne à partir des champs taille et poids.
- Intégrer des données : Créez de nouveaux ensembles de données en combinant des données provenant de plusieurs sources.
- Formatez les données : Reformatez les données si nécessaires. Par exemple, on peut convertir des chaînes de caractères contenant des chiffres en valeurs numériques afin de pouvoir effectuer des opérations mathématiques.

4.2.4. Modeling :

Dans cette phase on va probablement construire et évaluer divers modèles basés sur plusieurs techniques de modélisation différentes. Cette phase comporte quatre tâches :

- Sélectionner les techniques de modélisation : Déterminez les algorithmes à essayer (par exemple, la régression, le réseau neuronal).
- Générer la conception du test : En fonction de l'approche de modélisation, on peut diviser les données en ensembles de formation, de test et de validation.
- Construire le modèle : Aussi prestigieux que cela puisse paraître, il peut s'agir simplement d'exécuter quelques lignes de code comme *« reg = LinearRegression().fit(X, y) »*.
- Évaluer le modèle : Généralement, plusieurs modèles sont en concurrence les uns avec les autres, et le DataScientist doit interpréter les résultats du modèle en se basant sur la connaissance du domaine, les critères de réussite prédéfinis et la conception du test.

Bien que le guide CRISP-DM suggère d'itérer la construction et l'évaluation du modèle jusqu'à ce que on soit convaincu d'avoir trouvé le(s) meilleur(s) modèle(s), dans la pratique, les équipes doivent continuer à itérer jusqu'à ce qu'elles trouvent un modèle « suffisamment bon », passer par le cycle de vie CRISP-DM, puis améliorer encore le modèle dans les itérations futures.

4.2.5. Evaluation :

Alors que la tâche d'évaluation du modèle de la phase de modélisation se concentre sur l'évaluation du modèle technique, la phase d'évaluation examine plus largement quel modèle répond le mieux à l'activité et ce qu'il faut faire ensuite. Cette phase comporte trois tâches :

- Évaluer les résultats : Les modèles répondent-ils aux critères de réussite de l'entreprise ? Lequel ou lesquels devons-nous approuver pour l'entreprise ?
- Réviser le processus : Examinez le travail accompli. Y a-t-il eu des oublis ? Toutes les étapes ont-elles été correctement exécutées ? Résumez les résultats et corrigez si nécessaire.
- Déterminez les prochaines étapes : Sur la base des trois tâches précédentes, déterminez s'il faut passer au déploiement, poursuivre l'itération ou lancer de nouveaux projets.

4.2.6. Deployment:

Un modèle n'est pas particulièrement utile si le client ne peut pas accéder à ses résultats. La complexité de cette phase est très variable. Cette phase finale comporte quatre tâches :

- Planifier le déploiement : Développer et documenter un plan de déploiement du modèle.
- Planifier le suivi et la maintenance : Développer un plan de suivi et de maintenance approfondi pour éviter les problèmes pendant la phase opérationnelle (ou phase post-projet) d'un modèle.
- Produire un rapport final : L'équipe de projet documente un résumé du projet qui peut inclure une présentation finale des résultats de l'exploration des données.
- Revoir le projet : Effectuez une rétrospective du projet sur ce qui s'est bien passé, ce qui aurait pu être mieux, et comment s'améliorer à l'avenir.

Le travail de notre organisation ne s'arrête peut-être pas là. En tant que cadre de projet, CRISP-DM ne décrit pas ce qu'il faut faire après le projet (également appelé -opérations-). Mais si le modèle est mis en production, assurer de le maintenir en production. Une surveillance constante et un ajustement occasionnel du modèle sont souvent nécessaires.

Conclusion :

En guise de conclusion, dans ce chapitre, il nous a été donné de présenter les concepts liés au Machine Learning qui constitue une des techniques de l'intelligence artificielle, puis de spécifier le type d'apprentissage utilisé dans le cadre de ce projet qui est l'apprentissage supervisé et ensuite d'apporter quelques notions théoriques sur les algorithmes de classification et les méthodes utilisées en vue d'évaluer les modèles qui en découlent.

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 03 : Solution proposée

Chapitre 03 : Solution proposée

Introduction :

Dans cette partie, nous allons répondre à la problématique formulée auparavant. Pour cela, nous allons suivre les étapes de la méthodologie CRISP-DM, afin de structurer le travail d'extraction de l'information et expliquer en détail chaque étape du travail que nous avons effectué.

1. Business Understanding :

Cette phase est constituée des points suivants :

1.1. Détermination des objectifs :

L'objectif de notre travail ci-dessous, selon une procédure de data mining, est d'améliorer le processus de traitement et d'attribution d'un crédit par un outil d'aide à la décision qui prend en compte l'historique et l'importance de la globalité des critères décisionnels afin de consolider les décisions prises par les analystes et leur apporter un gain de temps significatif en vue de se conformer aux objectifs fixés par le groupe.

1.2. Détermination des objectifs de l'exploration des données :

L'objectif technique est de mettre au point un outil reposant sur l'intelligence artificielle pour assister les analystes risque à travers l'analyse d'un dossier de crédit, afin d'évaluer la santé financière du client et en prévoir la viabilité.

Pour y parvenir, des algorithmes de Machine Learning seront élaborés en utilisant les données fournies par la banque dans lesquelles figurent les clients avec toutes les informations qui les caractérisent ainsi que leurs lignes de crédit.

1.3. Plan de projet :

Le schéma que nous allons adopter reflète les étapes de la méthodologie CRISP-DM.

Pour ce faire, nous allons :

- Importer la base de données et recueillir les différentes données sur le profil des clients, et étant donné que la base de données actuelle de la banque ne nous permet pas de mettre en place un modèle pertinent pour la prise de décision, il sera donc nécessaire de l'enrichir avec de nouvelles données sur la santé financière de chaque entreprise cliente, en procédant à l'extraction des informations financières se trouvant sur les bilans et les comptes de résultats de chaque entreprise, qui sont stockés sous forme de pièce jointe sur le DCCIT ; on va donc faire du DataScrapping.

- Ensuite, nous procéderons au nettoyage des données, la sélection de variables pertinentes et à la génération de nouvelles variables destinées à renforcer les résultats des modèles proposés.

- Nous procéderons ensuite à la sélection des modèles dont les performances se révèlent être les plus fiables, que nous adapterons en fonction des besoins et exigences du problème.

Pour ce faire, nous avons donc choisi de faire appel au Langage de programmation Python qui comporte diverses librairies permettant de simplifier la manipulation des données et la création des modèles.
Enfin, ce processus aura comme entrée une base de données et comme sortie des modèles permettant de mettre en évidence l'importance de différents ratios et critères de validation ainsi que la probabilité que le client soit défaillant.

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

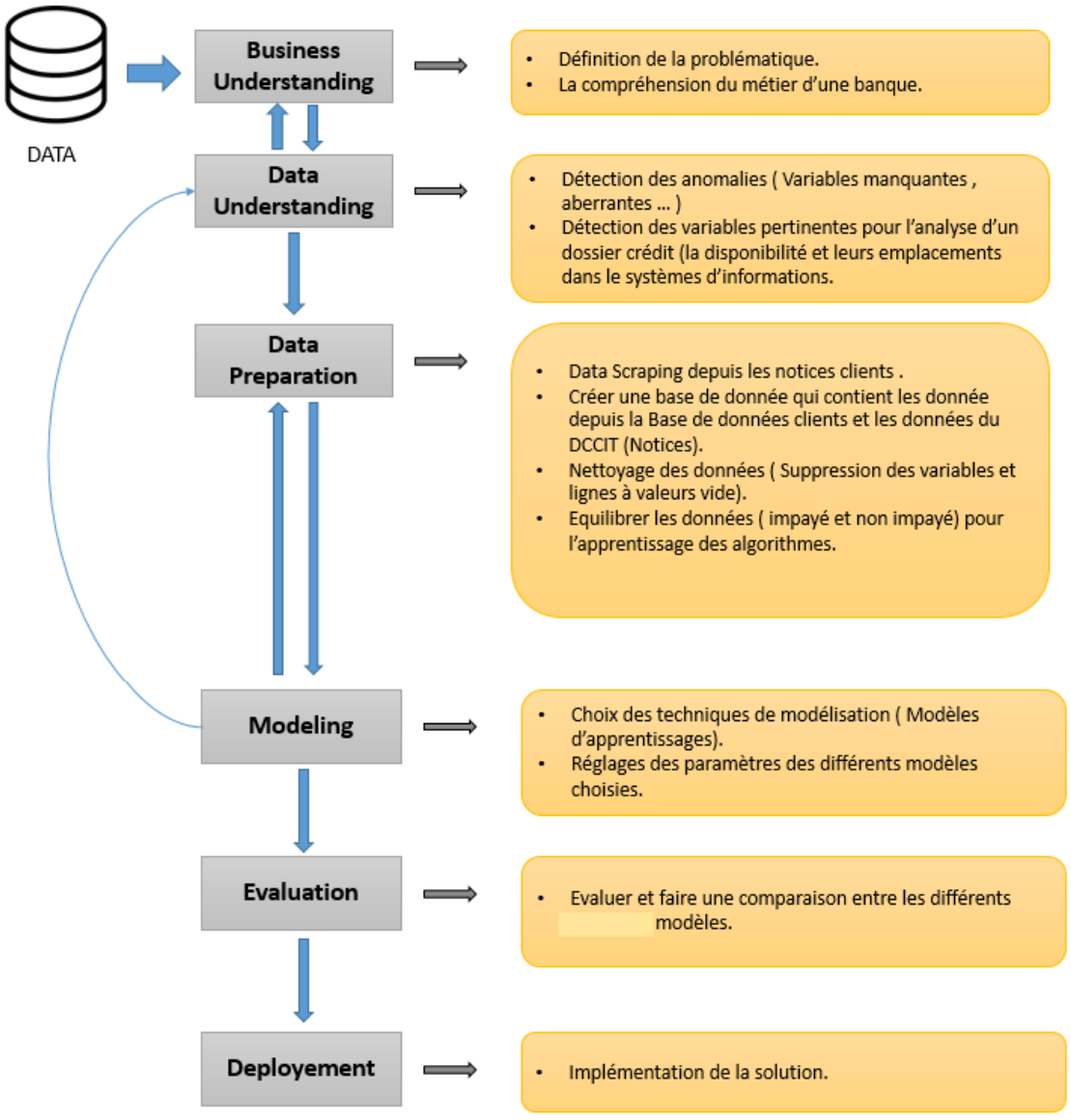


Figure 18 : Les phases de traitement de la problématique

2. Data Understanding :

A partir de la base de données, nous avons extrait les données sur tous les clients actuels, leurs lignes de crédit autorisées et nous l'avons complété par une jointure afin d'extraire les garanties proposées. Cette extraction était sous forme d'un fichier Excel format ***''csv''***.

Tandis que les données relatives à la situation financière des clients ne sont pas intégrées dans la base de données mais se trouvent dans l'outil DCCIT de la banque sous la forme d'un fichier Excel joint qui comporte les bilans financiers, les comptes de résultat du client pour les trois dernières années d'activité et le tableau des ratios financiers, afin d'extraire les données, nous nous sommes servis des programmes de DataScraping en utilisant la librairie Pandas du langage Python afin d'obtenir une nouvelle table de données plus riche et adéquate que celle déjà proposé par SGA. Chaque ligne de la nouvelle table correspond à un client qui s'est vu accorder un crédit tandis que chaque colonne illustre les différents taux financiers, ainsi qu'une colonne précisant si le client a été déclassé vers statue d'impayé ou s'il a bien honoré ses crédits.

2.1. DataScraping et constitution du DataSet :

La première étape de notre étude, une fois que la phase Business Understanding est terminée, va consister à l'extraction de différentes données du système d'information :

- Tout d'abord, nous avons demandé à disposer le dictionnaire de données afin de bien comprendre les attributs et les différentes tables existant dans la base de données de la banque.

- La 1ère extraction était une extraction directe de la table des autorisations de toutes lignes de crédit ***« bkautc ».*** Cette table contient tous les clients courants de la banque (Corporate et Retail) avec les dates d'autorisation de la ligne crédit, le montant, numéro de compte, date d'échéance, date de fin garantie, l'identité et autres attributs associé à la gestion administrative du département commercial.

- La 2ème extraction visait à déterminer les clients impayés qui n'ont pas honoré leurs engagements. (Les données sont des clients de l'historique 2019-2020).

- La 3ème extraction était différente des extractions précédentes où les données se trouvaient déjà dans la BDD et leur extraction était faite uniquement par des requêtes SQL, celle-ci a été faite à partir des notices financières (bilan et compte de résultat du client) ; ces notices étaient enregistrées dans des fichiers Excel (PDF dans certains cas) comme pièce-jointe sur le DCCIT. L'objectif de cette extraction est de constituer une BDD complémentaire à la première pour l'enrichir ; elle permet également de réaliser une étude approfondie du profil et de la santé financière de l'entreprise cliente à des fins d'apprentissage.

  Pour cette extraction, nous avons utilisé le langage Python par le biais de la bibliothèque Pandas et le module Os.

  Donc algorithmiquement nous avons :
- Importer les deux bibliothèques adéquates pour l'extraction des données et de manipulation des fichiers dans leurs emplacements.
- Nous avons stocké le fichier client dans une variable qu'on va la manipuler.

- A l'aide d'une boucle ***''for''*** nous avons extrait les données financières (Bilan financier et compte résultat Template unis.) en citant les cellules cibles.
- Les données extraites vont être stocké dans un DataSet où les ligne représente les clients et les colonnes représentent les inputs financiers, pour calculer les taux financiers pour notre étude.
- Apres calcul des taux financiers nous avons eu un DataSet final qui contient chaque client et ces taux financiers.

- La nouvelle colonne qui exprime si l'entreprise sera en faillite ou pas, a été obtenu grâce à une jointure entre le DataSet des impayés et le nouveau DataSet des données financières à l'aide de l'identifiant client. Cette colonne est de type booléen où la valeur 1 réfère que le client n'a pas honoré ces engagement (faillite '' Bankruptcy '') tandis que la valeur 0 signifie que le client a bien honoré ces engagements.

- Avec la collaboration d'un analyste de la direction de la gestion des risques, nous avons sélectionné les données les plus pertinents pour notre étude ainsi que calculer de nouveaux ratios pour constituer tous les taux et indicateurs financiers nécessaires à l'apprentissage.

  Ces taux financiers se présentent comme suit (Suite des taux dans **Annexe F**) :

- 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

Les données brutes sont souvent sujettes à des valeurs manquantes et incomplètes. Il est donc impératif de les analyser afin de déceler ces irrégularités et cela avant même de s'engager davantage dans un processus Data Mining.

2.2. Les valeurs manquantes :

En employant la fonction seaborn.heatmap() de la bibliothèque Seaborn, nous avons repéré des cellules vides dans notre DataSet, elle nous permet également de visualiser ces valeurs manquantes comme indiqué ci-dessous.

Code :

```
Entrée [1]: sb.heatmap(dataset.isnull(), cbar=False)
```

Le résultat :

Figure 19: Les valeurs manquantes – Heatmap fonction

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

3. Data Preparation :

Afin d'obtenir des résultats pertinents, interprétables et performants, il est essentiel de disposer d'une base de données de qualité, et pour cela, il est indispensable de passer par cette phase de préparation.

3.1. La correction des anomalies :

Une fois les anomalies détectées, nous avons entrepris leur correction comme suit :

Afin de traiter les valeurs manquantes on peut procéder de plusieurs manières :

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

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

L'application était faite à l'aide de la bibliothèque Pandas dans le langage Python, en utilisant la fonction « Drop » ; cette fonction va supprimer toutes les lignes qui contiennent des valeurs manquantes.

Script :

```
Entrée [4]: Dataset = dataset.dropna()
```

Résultat :

Figure 20: Les valeurs manquantes après traitement – Heatmap fonction

3.2. Test de Pearson :

Puisque nos variables sont toutes des variables contenues, nous pouvons donc tester la corrélation entre celles-ci en calculant la corrélation de Pearson, entre chacune des variables prises deux à deux. Ce test nous permet de déterminer si les deux variables évoluent dans le même sens, c'est-à-dire si les fortes valeurs de l'une sont associées aux fortes valeurs de l'autre (corrélation positive), ou si les fortes valeurs de l'une sont associées aux faibles valeurs de l'autre (corrélation négative), ou si les deux valeurs sont indépendantes. Dans le cas où deux variables sont dépendantes, il serait judicieux de ne garder qu'une seule d'entre elles, car les deux apportent presque la même information.

Les résultats du test de Pearson (voir **Annexe I**) sont présentés dans une matrice de dimension 57 x 57 présentée en **Annexe I**, 57 étant le nombre de variables dont nous disposons.

Pour chaque couple de variable, nous avons comparé leur corrélation deux à deux et avons éliminé l'une des variables du couple présentant une forte corrélation. Après ceci nous en sommes sorties avec 52 variables.

3.3. L'échantillonnage :

En vue de la réalisation de l'apprentissage automatique, il a été nécessaire dans un premier temps de constituer un échantillon de notre base de données car celle-ci est déséquilibrée. En effet, les clients défaillants ne représentent que 4% de l'ensemble des données ce qui équivaut à 2337 clients impayés, Quant aux clients avec un statue de non-impayé, ils sont au nombre de 56088.

Ce déséquilibre va conduire les modèles d'apprentissage à déclarer tous les clients comme fidèles afin de converger rapidement, et donc à ne pas prendre en compte ces cas minoritaires afin de maximiser l'indicateur utilisé pour juger la fiabilité du modèle.

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

- Le sous échantillonnage : il équilibre l'ensemble de données en réduisant la taille de la classe présente en majorité et en n'y collectant qu'un échantillon aléatoire de données tout en conservant toutes les occurrences de la classe rare.
- Le sur-échantillonnage : dans ce cas, il s'agira de maintenir les occurrences de la classe présente en majorité tout en augmentant le nombre d'occurrences du cas rare en répétant aléatoirement les cas déjà inclus dans l'échantillon.

Dans le cas d'un sur-échantillonnage, il faudrait prendre les 56088 clients non défaillants et prendre autant d'impayées en les reproduisant aléatoirement pour atteindre ce nombre ; Nous obtiendrons alors un échantillon total de taille 56088 x 2=112176 clients. Cet échantillon est alors trop important pour pouvoir s'en servir dans le développement de nos algorithmes (nécessité de machines robustes).

En ce qui nous concerne, nous avons opter pour un sous-échantillonnage aléatoire où nous avons procédé comme suit :

1er cas : Nous avons pris un échantillon de taille égale des deux classes (2337 clients) et nous avons visualisé les résultats des matrices de confusion que nous obtiendrons avec les classifieurs que nous avons développé. Le tableau suivant représente la matrice de confusion pour le modèle régression logistique :

Tableau 6 : Confusion matrice‘’Régression logistique’’ (1er cas d’échantillonnage)

Régression Logistique	+	-
+	286	173
-	212	264

2ème cas : Nous avons dans ce cas doublé la taille des clients non défaillants avec un échantillon totale de 7011 clients (2337 clients défaillants et 4674 clients non défaillants). Les résultats de la matrice de confusion obtenus sont les suivants :

Tableau 7 : Confusion matrice ‘’Régression logistique’’ (2ème cas d’échantillonnage)

Régression Logistique	+	-
+	653	87
-	108	554

3ème cas : Un échantillon de 9348 clients où la proportion est de 75% clients non défaillants et 25% sont des clients défaillants. La matrice de confusion est comme suit :

Tableau 8 : Confusion matrice ‘’Régression logistique’’ (3ème cas d’échantillonnage)

Régression Logistique	+	-
+	1250	60
-	90	470

4ème cas : Nous avons pris une proportion de 20% clients défaillants et 80% non défaillants, pour un échantillon total de taille de 11685.

Tableau 9 : Confusion matrice ‘’Régression logistique’’ (4ème cas d’échantillonnage)

Régression Logistique	+	-
+	1553	221
-	93	370

5ème cas : Nous avons pris une proportion de 20% défaillants et 80% non défaillants, taille de l’échantillon est de 14022.

Tableau 10 : Confusion matrice ‘’Régression logistique’’ (5ème cas d’échantillonnage)

Régression Logistique	+	-
+	1853	591
-	103	520

D’après les résultats des 5 tableaux précédents, nous avons pu constater dans notre cas de sous échantillonnage aléatoire que le 3ème cas où nous avons pris la proportion des clients non défaillants 3 fois plus grande que celle des clients défaillants a donné les meilleurs résultats.

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

Pour l'équilibre des données nous avons utilisé la bibliothèque Pandas, algorithmiquement nous avons :

- Diviser notre DataSet en deux listes la 1ère contient les clients défaillant '' Faillite = 1 '' et le 2ème pour les non défaillants '' Faillite = 0 ''.
- Nous avons détecté le déséquilibre des données des deux classes que nous pouvons voir dans l'histogramme suivant.

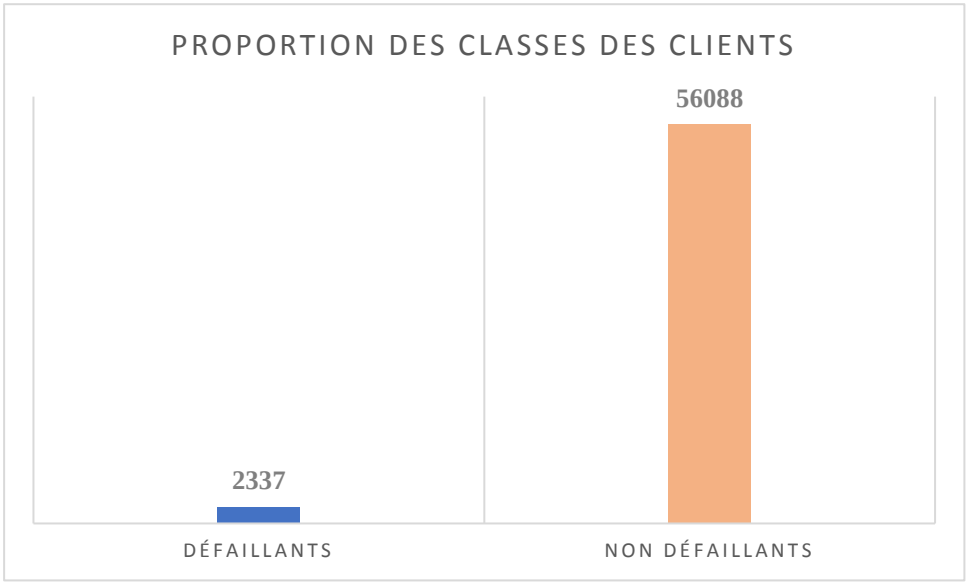


Figure 21 : La proportion des clients

- Nous avons procédé à l'échantillonnage des données, par le biais du sous-échantillonnage aléatoire, en prenant 7011 clients non défaillants aléatoirement (3 fois la taille des clients défaillants) et la totalité des clients défaillants.
- La concaténation des deux classes (défaillants et non-défaillants).
- La division du Dataset en deux ensembles, un pour l'apprentissage qui représente 80% de l'échantillon et un autre pour le test des modèles qui représente les 20% restantes.

4. Modeling :

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 Datamining et du 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, de choisir cinq modèles à appliquer sur nos données, étant donné étant donné le fait que notre problématique est une classification, nous en sommes partis pour les algorithmes suivants :

- Decision Tree
- Random Forest
- Support Vector Machine *''SVM''*
- K-Nearest Neighbours *''KNN''*
- La Régression logistique

Le script de tous l'apprentissage des 5 modèles et de la préparation des données est en **Annexe G** en langage Python.

Les données utilisées pour l'apprentissage ainsi que pour le test ont été les mêmes pour l'ensemble des algorithmes.

Grace à la fonction « *train_test_split* » nous avons divisé nos données en deux ensembles, 80% de données pour faire l'apprentissage sous forme de tableaux numpy, 2 tableaux pour l'apprentissage *« X_train et y_train »* et 20% de données pour le test sous forme de 2 autres tableaux *« X_test et y_test »* ; où

Y représente notre variable de sortie qu'on veut prédire *« Faillite »*

X représentes nos 52 variables d'entrées *« Taux et indicateurs financiers »*

Voici le script concernant le split des données :

```
from sklearn.model_selection import train_test_split
```

```
X = Datasetfinal.drop(["Faillite"],axis =1)
y = Datasetfinal["Faillite"]
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=101)
```

Pour la validation et le bon choix des paramètres, nous avons opté pour la Cross Validation prenant en compte la valeur du *R-Carré.*

- Le R-carré est une mesure statistique qui représente la qualité de l'ajustement d'un modèle de régression. La valeur idéale du carré R est de 1. Plus la valeur du carré R est proche de 1, plus le modèle est bien ajusté.
  Le R-carré est une comparaison entre la somme résiduelle des carrés et la somme totale des carrés. La somme totale des carrés est calculée par la somme des carrés de la distance perpendiculaire entre les points de données et la ligne moyenne
  « $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ ».

- La validation croisée est une méthode statistique utilisée pour estimer la compétence des modèles d'apprentissage automatique. Elle est couramment utilisée dans l'apprentissage automatique appliqué pour comparer et sélectionner un modèle pour un problème de modélisation prédictive donné, car elle est facile à comprendre, facile à mettre en œuvre et donne des estimations de compétence qui ont généralement un biais plus faible que les autres méthodes.

Dans notre cas nous allons utiliser Validation croisée k-fold.

Validation croisée k-fold :

La validation croisée est une procédure de ré échantillonnage utilisée pour évaluer les modèles d'apprentissage automatique sur un échantillon de données limité.

La procédure a un seul paramètre appelé k qui fait référence au nombre de groupes dans lesquels un échantillon de données donné doit être divisé. En tant que telle, la procédure est souvent appelée validation croisée k-fold. Lorsqu'une valeur spécifique pour k est choisie, elle peut être utilisée à la place de k dans la référence au modèle, par exemple k=10 pour une validation croisée 10 fois.

La validation croisée est principalement utilisée en apprentissage automatique appliqué pour estimer la compétence d'un modèle d'apprentissage automatique sur des données non vues. En d'autres termes, il s'agit d'utiliser un échantillon limité afin d'estimer comment le modèle devrait se comporter en général lorsqu'il est utilisé pour faire des prédictions sur des données qui n'ont pas été utilisées pendant la formation du modèle.

La procédure générale est la suivante :

- Mélangez l'ensemble de données de façons aléatoire.
- Divisez l'ensemble de données en k groupes

Pour chaque groupe unique :

- Prenez le groupe comme un ensemble de données d'attente ou de test.
- Prenez les groupes restants comme un ensemble de données d'entraînement
- Ajustez un modèle sur l'ensemble d'apprentissage et évaluez-le sur l'ensemble de test.
- Conservez le score d'évaluation et éliminez le modèle.
- Résumez la compétence du modèle en utilisant l'échantillon des scores d'évaluation du modèle.

Il est important de noter que chaque observation de l'échantillon de données est affectée à un groupe individuel et reste dans ce groupe pendant toute la durée de la procédure. Cela signifie que chaque échantillon peut être utilisé 1 fois dans l'ensemble de maintien et utilisé pour entraîner le modèle k-1 fois.

Le Script utilisé pour procéder à une validation croisée dans notre projet afin de déterminer les best_parameters (paramètres optimaux) est le suivant :

```
from sklearn.model_selection import cross_val_score
```

```
Logg=LogisticRegression(C = 70, max_iter = 100,solver ='liblinear')
cross_val_score(Logg, X_train, y_train, cv=5, scoring='accuracy')
```

4.1. 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 :

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{52} X_{52}$$

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 Outcome « *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.

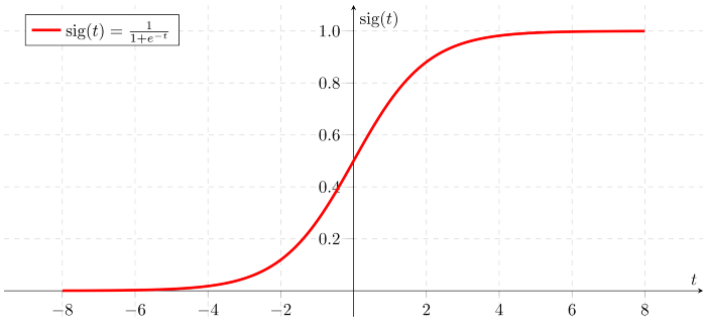


Figure 22: Graphe et expression de la fonction sigmoïde.fonction

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.

4.1.1. Modèle d'apprentissage sur python :

Comme 1er étape de l'apprentissage il faut d'abord sélectionner le modèle et ces paramètres, autrement dit la validation.

Le modèle est importé du packages *Scikit-learn '' bibliothèque libre Python destinée à l'apprentissage automatique ''* sous cette forme :

```
from sklearn.linear_model import LogisticRegression
```

```
model = LogisticRegression(C = 70, max_iter = 100,solver ='liblinear')
```

Comme on peut le voir, notre fonction est composée de 3 paramètres :

- 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.

4.1.2. Cross Validation et la validation curve :

Afin d'aboutir à des résultats adéquates, il est impératif de procéder par le cross validation pour pouvoir trouver le choix optimal des paramètres et estimateurs, pour cela :

Nous avons importé la fonction cross_val_score du package sklearn.model_selection

Nous avons choisi un k-fold de 5 (on découpe sur 5 échantillons, cv = 5)

```
from sklearn.model_selection import cross_val_score
```

```
Logg=LogisticRegression(C = 70, max_iter = 100,solver ='liblinear')
cross_val_score(Logg, X_train, y_train, cv=5, scoring='accuracy')
```

```
array([0.86607143, 0.84821429, 0.91071429, 0.85714286, 0.86607143])
```

A partir des résultats obtenus, nous pouvons constater que le dans l'ensemble des 5 échantillons, le 3ème s'est démarqué avec un score de R^2= 0.91071429, tandis que la moyenne de l'ensemble est égale à 0.8696

Afin d'exécuter la curve validation, nous avons utilisé le script suivant qui contient une boucle for pour varier les valeurs des deux hyperparametres de la fontion du modele et calcule à chaque valeur, la valeur du R-carré, afin de déterminer le meilleur R-carré.

```
val_score = []
for k in range(100, 1000):
    score = cross_val_score(LogisticRegression(C = 60, max_iter = k,solver ='liblinear'), X_train, y_train, cv=5).mean()
    val_score.append(score)

plt.plot(val_score)
```

```
val_score = []
for k in range(1, 100):
    score = cross_val_score(LogisticRegression(C = k, max_iter = 100,solver ='liblinear'), X_train, y_train, cv=5).mean()
    val_score.append(score)

plt.plot(val_score)
```

Pour déterminer les paramètres les plus optimaux pour notre algorithme, nous avons utilisé une boucle (for) qui varie les valeurs des 2 paramètres de la fonction logistique le (C et max_iter), ensuite en faisant appel à la bibliothèque matplotlib.pyplot, nous avons représenté les résultats obtenues que nous pouvons voir dans ce qui suit.

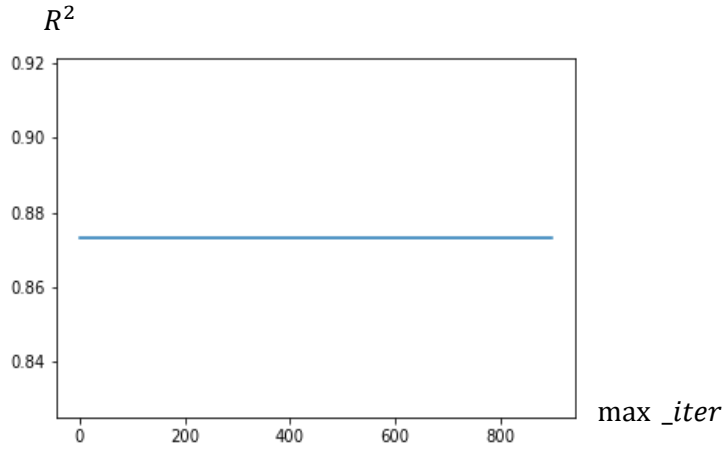


Figure 23: Variation du score R^2 en fonction du paramètre max_iter

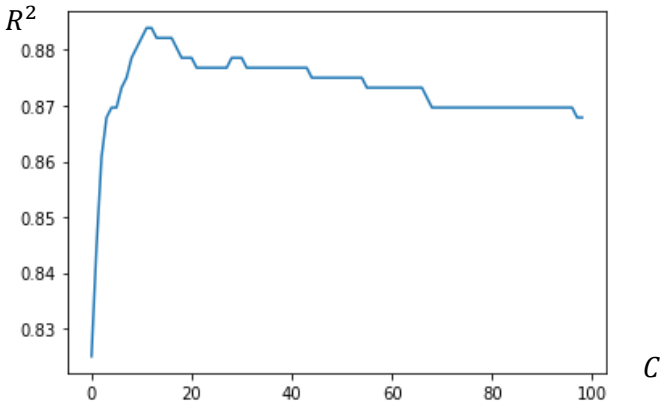


Figure 24: Variation du score R^2 en fonction du paramètre C

Comme nous pouvons le constater sur les 2 graphes précédents, la variation du paramètre « max_iter » n'affecte par la valeur du score R^2, par défaut nous lui avons affecté une valeur de 100 ; tandis que la valeur du score R^2 varie avec la variation du paramètre C, et nous obtenons une valeur maximale du $R^2 = 0.919786$ pour un C égale à 11.

Pour déterminer la valeur exacte du paramètre C, nous utiliserons la fonction GridSearchCV de la même bibliothèque, comme montré ci-dessous.

```
from sklearn.model_selection import GridSearchCV
```

```
param_grid = {'C': np.arange(1, 100)}

grid = GridSearchCV(LogisticRegression(C = 1, max_iter = 100,solver ='liblinear'), param_grid, cv=5)

grid.fit(X_train, y_train)
```

```
GridSearchCV(cv=5, estimator=LogisticRegression(C=1, solver='liblinear'),
             param_grid={'C': array([ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16, 17,
       18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
       35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
       52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68,
       69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85,
       86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99])})
```

```
grid.best_params_
```

```
{'C': 11}
```

Une fois les paramètres les plus adaptés à l'algorithme déterminés, nous entamons l'apprentissage en vue d'obtenir les outputs souhaités.

4.1.3. Apprentissage et résultats :

L'entraînement des données est devenu simple grâce à la bibliothèque scikit-learn dans laquelle nous trouvons toutes les fonctions dont nous avons besoin, qui sont prédéfinies, faciles à utiliser et à paramétrer training figure dans le script ci-dessus :

```
# Logistic Regression

from sklearn.linear_model import LogisticRegression

model = LogisticRegression(C = 11, max_iter = 100,solver ='liblinear')

model_training=model.fit(X_train,y_train)

prediction = model.predict(X_test)
model.score(X_test,y_test)
```

```
0.9197860962566845
```

```
model = LogisticRegression(C = 70, max_iter = 100,solver ='liblinear')
model.fit(X_train,y_train)
y_pred_logistic = model.decision_function(X_test)
```

Figure 25: Algorithme et training du modèle logistique

Après l'ajustement des hyperparamètres dans la fonction de notre modèle, nous utiliserons la fonction ***model.fit(X_train,y_train)*** afin de faire notre apprentissage.

LogisticRegression(C =11,max_iter =100,solver = 'liblinear').

Ensuite, nous avons utilisé la fonction score(X_test,Y_test) pour calculer le R^2 du modèle, le résultat obtenu est de 0.9192 donc notre modèle est d'une précision de 91.92%.

Vu que le modèle de régression logistique contient des coefficients pour constituer l'hyper-droite, ces coefficients représentent l'importance donnée à X_i dans la classification.

On peut voir dans la figure suivante une partie des coefficients obtenus par le script ci-dessous ; (le reste des coefficients se trouve en **Annexe J**)

```
a = pd.DataFrame()
a["Columns"] = X.columns
a
coef = model.coef_
a["Coef"] = coef[0]
a
```

10	Taux de croissance du résultat net après impôts	3.067341
11	Taux de croissance du résultat net	3.391394
12	Taux du total des actifs (Année 1)	-0.018568
13	Taux de croissance du rendement total des acti...	1.308456
14	Réinvestissement en espèces %	4.736146
15	Le ratio courant	-0.404615
16	Taux d'intérêt débiteur	6.715971
17	Total du passif / Valeur nette	-2.325268
18	Ratio d'endettement %	13.145383
19	Valeur nette / actifs	-8.673518
20	Ratio d'adéquation des fonds à long terme	5.912253
21	Résultat d'exploitation	-2.872214
22	Bénéfice net avant impôt par capital	-9.842852

Figure 26: Coefficient du modèle logistique utilisé

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, comme c'est le cas pour le **« *Bénéfice net avant impôt par capital* »** avec un coefficient négatif égale à (-9. 842852). Tandis qu'un coefficient positif signifie que cet indicateur fera converger le résultat vers (1) donc vers une situation de banqueroute de l'entreprise, comme c'est le cas du ***« Ratio d'endettement % »*** qui est égal à (13,145383) compte tenu de l'importance de ce ratio dans la détermination de la santé financière de l'entreprise.

Grace à la fonction ***predict_proba(X_test)*** qui a été appliqué sur l'échantillon du test, le résultat obtenu sera la probabilité en % que notre outcome « faillite » serait égale à 1 c'est-à-dire une situation de banqueroute ou pas (donc égale à 0) ; la figure suivante représente les résultats obtenus pour les 10 premiers clients avec la probabilité (non_faillite , faillite) :

```
proba = model.predict_proba(X_test)
proba[:10]
```

```
array([[0.98954895, 0.01045105],
       [0.12308417, 0.87691583],
       [0.89370127, 0.10629873],
       [0.95147245, 0.04852755],
       [0.37584596, 0.62415404],
       [0.96822361, 0.03177639],
       [0.68156467, 0.31843533],
       [0.89016574, 0.10983426],
       [0.97808079, 0.02191921],
       [0.43889627, 0.56110373]])
```

Figure 27: Probabilité de défaillance à l'aide du modèle régression logistique

Pour terminer, nous évaluerons la robustesse du modèle à l'aide de la matrice de confusion, que nous appelons par la fonction *confusion_matrix(y_test,prediction)* de la bibliothèque *metrics*, les résultats de cette matrice sont d'une nécessité impérative pour évaluer chaque modèle.

Tel que nous pouvons le voir ci-dessous, le script utilisé.

```
from sklearn.metrics import confusion_matrix
logistic_cm = confusion_matrix(y_test,prediction)
logistic_cm
```

Figure 28: Script matrice de confusion du modèle régression logistique

4.2. Decision Tree :

Les arbres de décision sont une méthode d'apprentissage supervisé non paramétrique utilisée pour la classification et la régression. Son objectif est de créée 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).

4.2.1. Modèle et apprentissage sur Python :

La façon de procéder afin de déterminer les paramètres les plus optimaux va être la même suivait dans le modèle précédent.

Le modèle est représenté comme suit :

```
from sklearn.tree import DecisionTreeClassifier
```

```
dtree = DecisionTreeClassifier(max_depth=10,min_samples_leaf=17)
```

Pour parvenir à des résultats optimaux, nous utiliserons la validation croisée afin de déterminer les meilleurs paramètres' permettant d'obtenir un score R^2 optimal.

Les hyperparamètres utilisés 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.

4.2.2. Cross validation et la validation Curve :

Nous allons décomposer notre taille sur 5 échantillons, comme nous l'avons déjà effectué pour le modèle précédent. Les résultats obtenus sont les suivants :

```
dtree=DecisionTreeClassifier()
cross_val_score(dtree, X_train, y_train, cv=5, scoring='accuracy')
```

```
array([0.79464286, 0.83035714, 0.85714286, 0.85714286, 0.83035714])
```

D'après les résultats obtenus, nous remarquons que le score R^2 moyen est égale à : $R^2 = 0.8339$.

De la même façon dont nous avons procédé dans le modèle précédent, nous obtenons les graphes suivants :

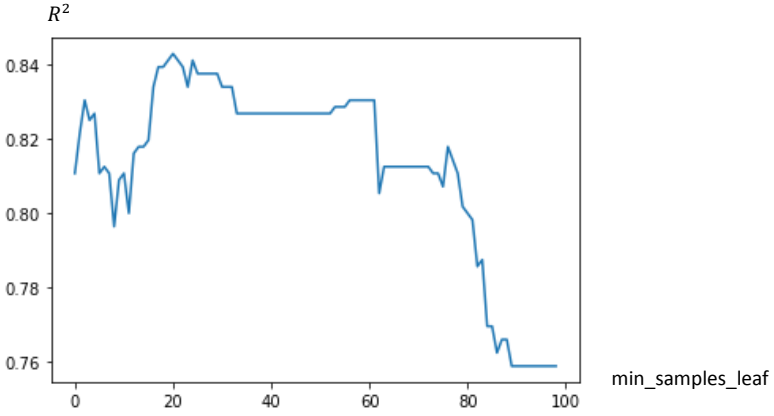


Figure 29: Variation du score R^2 en fonction du paramètre min_samples_leaf

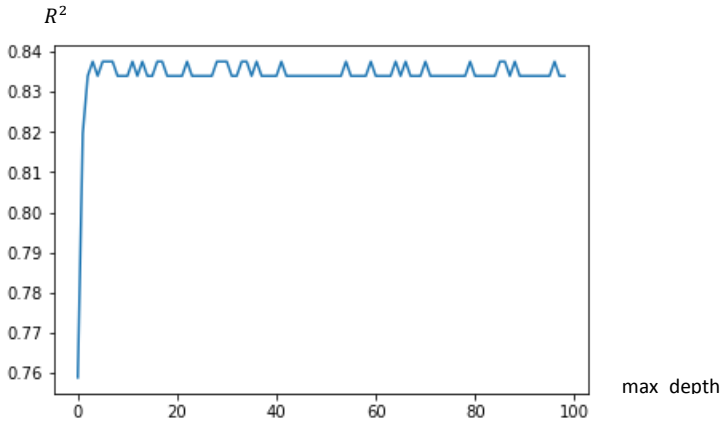


Figure 30: Variation du score R^2 en fonction du paramètre max_depth

A partir des deux graphes précédents nous pouvons constater la variation du score R^2 en fonction de l'évolution des deux paramètres, et afin de déterminer les valeurs exactes des paramètres pour le score maximal qui est égale à $R^2 = 0.85$; nous utiliserons la fonction *GridSearchCV*, dont les résultats sont les suivants :

```
grid.best_params_
{'max_depth': 3}
```

```
grid.best_params_
{'min_samples_leaf': 21}
```

4.2.3. Apprentissage et résultats :

Suite au choix des hyperparamètres optimaux ***DecisionTreeClassifier(max_depth=3, min_samples_leaf=21),*** nous débutons le Training de notre modèle via la fonction ***model.fit(X_train,y_train)*** ; puis, nous calculons le R^2 en utilisant la fonction ***score(X_test,y_test)*** ; celui-ci est égale à **0.8717** qui indique une precision de **87,17%** pour ce modele

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_importances_*** et la bibliothèque ***matplotlib.pyplot*** afin de faciliter leur interprétation sur des visuels. (Visuels disponibles en **Annexe K)**.

A travers les visualisations obtenues, nous pouvons remarquer que le modèle n'a pas considéré toutes les variables (Juste 7 variables) et a donc jugé certaines d'entre elles comme étant sans importance pour la décision finale.

Grace à la fonction ***predict_proba(X_test)*** qui a été appliqué sur l'échantillon du test, nous avons obtenu comme avec le dernier modèle la probabilité en % que notre outcome « faillite » serait égale à 1 c'est-à-dire une situation de banqueroute ou pas (donc égale à 0) ; la figure suivante représente les résultats obtenus pour les 10 premiers clients.

```
                [Non Faillite,     Faillite]
array([[0.81481481, 0.18518519],
       [0.09803922, 0.90196078],
       [0.99586777, 0.00413223],
       [0.88095238, 0.11904762],
       [0.09803922, 0.90196078],
       [0.99586777, 0.00413223],
       [0.99586777, 0.00413223],
       [0.35714286, 0.64285714],
       [0.99586777, 0.00413223],
       [0.88095238, 0.11904762]])
```

Figure 31 : Probabilité de défaillance à l'aide du modèle Decision Tree

4.3. 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.

4.3.1. Modèle et apprentissage sur Python :

La façon de procéder afin de déterminer les paramètres les plus optimaux va être la même suivait dans le modèle précédent.

Le modèle est représenté comme suit :

```
from sklearn.ensemble import RandomForestClassifier
```

```
forest = RandomForestClassifier(n_estimators = 550,min_samples_leaf=3,max_depth=18)
```

Pour parvenir à des résultats optimaux, nous utiliserons la validation croisée afin de déterminer les meilleurs paramètres' permettant d'obtenir un score R^2 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.

4.3.2. Cross validation et la validation Curve :

Nous allons décomposer notre taille sur 5 échantillons, comme nous l'avons déjà effectué pour le modèle précédent. Les résultats obtenus sont les suivants :

```
rforest=RandomForestClassifier(n_estimators = 3500)
cross_val_score(rforest, X_train, y_train, cv=5, scoring='accuracy')
```

```
array([0.85714286, 0.85714286, 0.89285714, 0.86607143, 0.89285714])
```

D'après les résultats obtenus, nous remarquons que le score R^2 moyen est égale à :$R^2 = 0.8732$.

De la même façon dont nous avons procédé dans le modèle précédent, nous obtenons les graphes suivants :

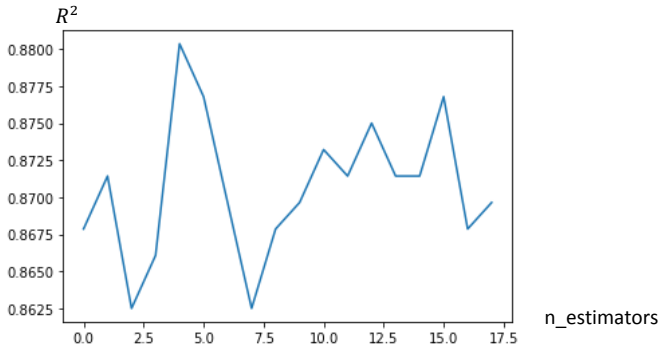


Figure 32: Variation du score R^2 en fonction du paramètre n_estimators

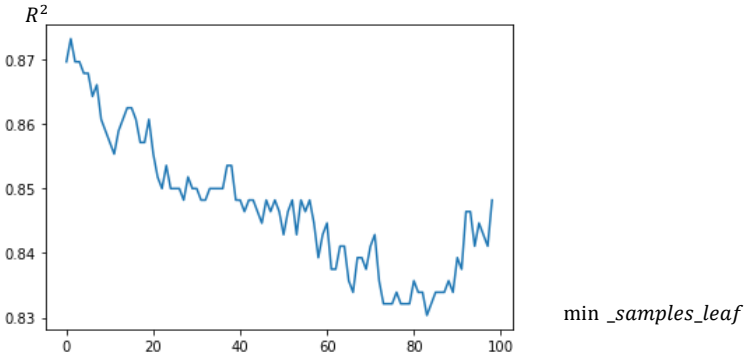


Figure 33: Variation du score R^2 en fonction du paramètre min_samples_leaf

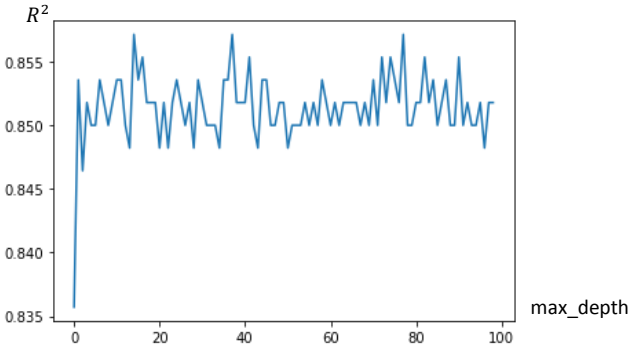


Figure 34: Variation du score R^2 en fonction du paramètre max_depth

A partir des deux graphes précédents nous pouvons constater la variation du score R^2 en fonction de l'évolution des deux paramètres, et afin de déterminer les valeurs exactes des paramètres pour le score maximal qui est égale à $R^2 = 0.85$; nous utiliserons la fonction ***GridSearchCV***, dont les résultats sont les suivants :

```
{'n_estimators': 550}        {'min_samples_leaf': 3}        {'max_depth': 18}
```

On peut visualiser la courbe ROC (AUC variation en variant les hyperparametres en **Annexe Q**) et voir l'influence du choix optimal sur le AUC.

4.3.3. Apprentissage et résultats :

Suite au choix des hyperparamètres optimaux, *RandomForestClassifier(n_estimators = 550,min_samples_leaf=3,max_depth=18)*. Nous débutons le Training de notre modèle via la fonction *model.fit(X_train,y_train) ;* puis, nous calculons le R^2 en utilisant la fonction *score(X_test,y_test) ;* celui-ci est égale à une précision de 90,37% .

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_importances_*** et la bibliothèque ***matplotlib.pyplot*** afin de faciliter leur interprétation sur des visuels. (Visuels disponibles en **Annexe L**).

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

Grace à la fonction ***predict_proba(X_test)*** qui a été appliqué sur l'échantillon du test, nous avons obtenu comme avec le dernier modèle la probabilité en % que notre outcome « faillite » serait égale à 1 c'est-à-dire une situation de banqueroute ou pas (donc égale à 0) ; la figure suivante représente les résultats obtenus pour les 10 premiers clients.

```
foresta=forest.predict_proba(X_test)
foresta[:10]
```

```
array([[0.99234848, 0.00765152],
       [0.06827371, 0.93172629],
       [0.99836364, 0.00163636],
       [0.91097573, 0.08902427],
       [0.03472872, 0.96527128],
       [0.99739394, 0.00260606],
       [0.97778788, 0.02221212],
       [0.89238344, 0.10761656],
       [0.9852619 , 0.0147381 ],
       [0.98693074, 0.01306926]])
```

Figure 35: Probabilité de défaillance à l'aide du modèle Random Forest

Pour terminer, nous évaluerons la robustesse du modèle à l'aide de la matrice de confusion, que nous appelons par la fonction ***confusion_matrix(y_test, forest.predict(X_test))*** de la bibliothèque ***metrics***, les résultats de cette matrice sont d'une nécessité impérative pour évaluer chaque modèle.

4.4. Support Vector Machines :

Le modèle SVM va être utiliser avec les différents noyaux suivants : Linear, Polynomial, Radial.

4.4.1. Modèle et apprentissage sur Python :

La façon de procéder afin de déterminer les paramètres les plus optimaux va être la même suivait dans le modèle précédent.

Le modèle est représenté comme suit **: *SVC (C, gamma, kernel)***

Pour parvenir à des résultats optimaux, nous utiliserons la validation croisée afin de déterminer les meilleurs paramètres' permettant d'obtenir un score R^2 optimal.

Les hyperparamètres utilisés sont :

- ***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.
- ***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.

4.4.2. Cross validation et la validation Curve :

Nous allons décomposer notre taille sur 5 échantillons, comme nous l'avons déjà effectué pour le modèle précédent. Les résultats obtenus sont les suivants :

```
svmm=SVC(C=3,gamma=3)
cross_val_score(svmm, X_train, y_train, cv=5, scoring='accuracy')
```

```
array([0.82142857, 0.82142857, 0.86607143, 0.83928571, 0.86607143])
```

Dans ce cas nous avons utilisé le *Kernel* par défaut ***'' rbf - Radial ''***. D'après les résultats obtenus, nous remarquons que le score R^2 moyen est égale à :$R^2 = 0.8429$.

De la même façon dont nous avons procédé dans le modèle précédent, nous obtenons les graphes suivants :

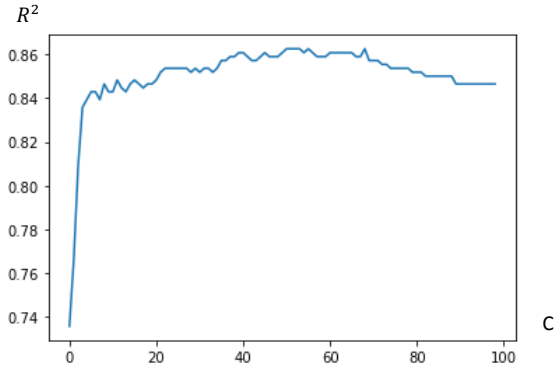


Figure 36: Variation du score R^2 en fonction du paramètre C

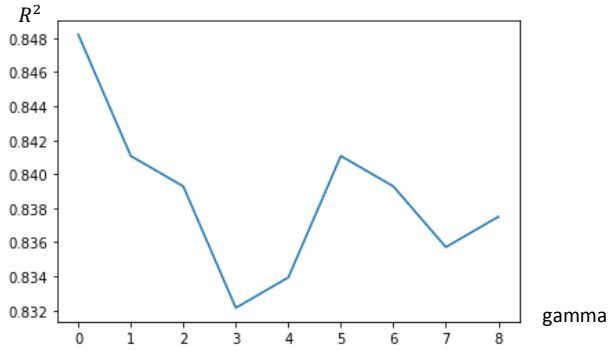


Figure 37: Variation du score R^2 en fonction du paramètre gamma

A partir des deux graphes précédents nous pouvons constater la variation du score R^2 en fonction de l'évolution des deux paramètres, et afin de déterminer les valeurs exactes des paramètres pour le score maximal qui est égale à $R^2 = 0.8429$; nous utiliserons la fonction *GridSearchCV*, dont les résultats sont les suivants :

```
grid.best_params_
```

```
{'C': 4}
```

```
grid.best_params_
```

```
{'gamma': 1}
```

4.4.3. Apprentissage et résultats :

Suite au choix optimaux des deux des hyperparamètres *C et gamma*, ***SVC (C=4, gamma=1, kernel)*.** Nous débutons le Training de notre modèle tout en faisant varier la valeur de l'hyperparamètre *kernel* et ainsi calculer à chaque fois le score et la matrice de confusion. Les résultats obtenus sont les suivants :

- *Kernel = ' rbf ' :*

```
svm = SVC(C=4,gamma=1)
```

```
svm.fit(X_train,y_train)
```

```
SVC(C=4, gamma=1)
```

```
svm.score(X_test,y_test)
```

```
0.9251336898395722
```

- *Kernel = ' poly ' :*

```
svm = SVC(C=5,gamma=1,kernel='poly')
```

```
svm.fit(X_train,y_train)
```

```
SVC(C=5, gamma=1, kernel='poly')
```

```
svm.score(X_test,y_test)
```

```
0.839572192513369
```

- *Kernel= ' linear ' :*

```
svm = SVC(C=5,gamma=1,kernel='linear')
```

```
svm.fit(X_train,y_train)
```

```
SVC(C=5, gamma=1, kernel='linear')
```

```
svm.score(X_test,y_test)
```

```
0.893048128342246
```

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_importances_*.**

Après le test de nos 3 modèles, nous avons constaté à partir des matrices de confusions illustrées et les scores obtenus que le modèle à noyaux *Radial '' rbf ''* a donné les meilleurs résultats en termes de matrice de confusion et de paramètre R-carré. C'est pourquoi, seul ce modèle sera gardé pour le comparer au reste des algorithmes dans la phase suivante.

Le score R^2 du modèle de 0.9251 ce qui veut dire que la précision de notre modèle est de 92,51%.

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 ***mutual_info_classif (X_train, y_train)*** et la bibliothèque ***matplotlib.pyplot*** afin de faciliter leur interprétation sur des visuels. (Visuel de diagramme à barre disponible en **Annexe M**)

A travers les visualisations obtenues, nous pouvons remarquer que le modèle n'a pas pris en considération les 3 variables (Passif courant/Passif, Taux de frais de recherche et développement et le Taux de dépense d'exploitation) dans sa prise de décision.

Grace à la fonction ***predict_proba(X_test)*** qui a été appliqué sur l'échantillon du test, nous avons obtenu comme avec le dernier modèle la probabilité en % que notre outcome « faillite » serait égale à 1 c'est-à-dire une situation de banqueroute ou pas (donc égale à 0) ; la figure suivante représente les résultats obtenus pour les 10 premiers clients.

```
probability[:10]
```

```
array([[0.99718715, 0.00281285],
       [0.0612141 , 0.9387859 ],
       [0.89046545, 0.10953455],
       [0.87371505, 0.12628495],
       [0.21734621, 0.78265379],
       [0.95940467, 0.04059533],
       [0.81549415, 0.18450585],
       [0.95391312, 0.04608688],
       [0.9726356 , 0.0273644 ],
       [0.95330359, 0.04669641]])
```

Figure 38: Probabilité de défaillance à l'aide du modèle SVM

4.5. Les k plus proches voisins « K-Nearest Neighbours » :

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.

4.5.1. Modèle et apprentissage sur Python :

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(n_neighbors=5).***

Pour parvenir à des résultats optimaux, nous utiliserons la validation croisée afin de déterminer le paramètre le plus optimal permettant d'obtenir un score R^2 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.

4.5.2. Cross validation et la validation Curve :

Nous allons décomposer notre taille sur 5 échantillons, comme nous l'avons déjà effectué pour le modèle précédent. Les résultats obtenus sont les suivants :

```
knnn=KNeighborsClassifier(n_neighbors=9)
cross_val_score(knnn, X_train, y_train, cv=5, scoring='accuracy')
```

```
array([0.84821429, 0.86607143, 0.83928571, 0.85714286, 0.8125    ])
```

D'après les résultats obtenus, nous remarquons que le score R^2 moyen est égale à :$R^2 = 0.8732$.

De la même façon dont nous avons procédé dans le modèle précédent, nous obtenons le graphe suivant :

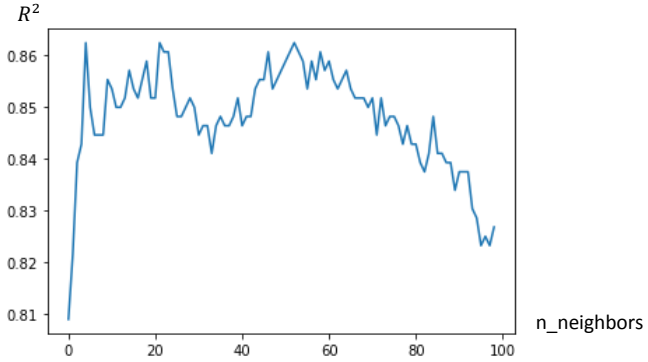


Figure 39: Variation du score R^2 en fonction du paramètre n_neighbors

A partir du graphe précédent nous pouvons constater la variation du score R^2 en fonction de l'évolution du paramètre n_neighbors, et afin de déterminer la valeur exacte du paramètre pour le score maximal qui est égale à $R^2 = 0.8503$; nous utiliserons la fonction ***GridSearchCV***, dont le résultat est le suivant : `'n_neighbors': 16`

On peut visualiser la courbe ROC (AUC variation en variant les hyperparametres en **Annexe P**) et voir l'influence du choix optimal sur le AUC.

4.5.3. Apprentissage et résultats :

Suite au choix du hyperparamètre optimal, ***KNeighborsClassifier(n_neighbors=16).*** Nous débutons le training de notre modèle via la fonction *model.fit(X_train,y_train) ;* puis, nous calculons le R^2 en utilisant la fonction ***score(X_test,y_test)*** *;* celui-ci est égale à 0.8503.

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_importances_*** et la bibliothèque ***matplotlib.pyplot*** afin de faciliter leur interprétation sur des visuels. (Visuel de diagramme à barre disponible en **Annexe N**).

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

Grace à la fonction ***predict_proba(X_test)*** qui a été appliqué sur l'échantillon du test, nous avons obtenu comme avec le dernier modèle la probabilité en % que notre outcome « faillite » serait égale à 1 c'est-à-dire une situation de banqueroute ou pas (donc égale à 0) ; la figure suivante représente les résultats obtenus pour les 10 premiers clients.

```
knn= KNN.predict_proba(X_test)
knn[:10]
```

```
array([[1.    , 0.    ],
       [0.125 , 0.875 ],
       [0.75  , 0.25  ],
       [0.9375, 0.0625],
       [0.0625, 0.9375],
       [1.    , 0.    ],
       [0.6875, 0.3125],
       [0.9375, 0.0625],
       [1.    , 0.    ],
       [1.    , 0.    ]])
```

Figure 40: Probabilité de défaillance à l'aide du modèle KNN

Pour terminer, nous évaluerons la robustesse du modèle à l'aide de la matrice de confusion, que nous appelons par la fonction *confusion_matrix(y_test, forest.predict(X_test))* de la bibliothèque *metrics*, les résultats de cette matrice sont d'une nécessité impérative pour évaluer chaque modèle.

5. Evaluation :

Dans ce qui suit, nous souhaitons tester les 5 modèles que nous avons élaborés afin de sélectionner celui qui présente les meilleures performances par rapport aux métriques d'évaluation.

Dans un premier temps, nous visualisons les courbes ROC des différents modèles avec l'aire sous chaque courbe :

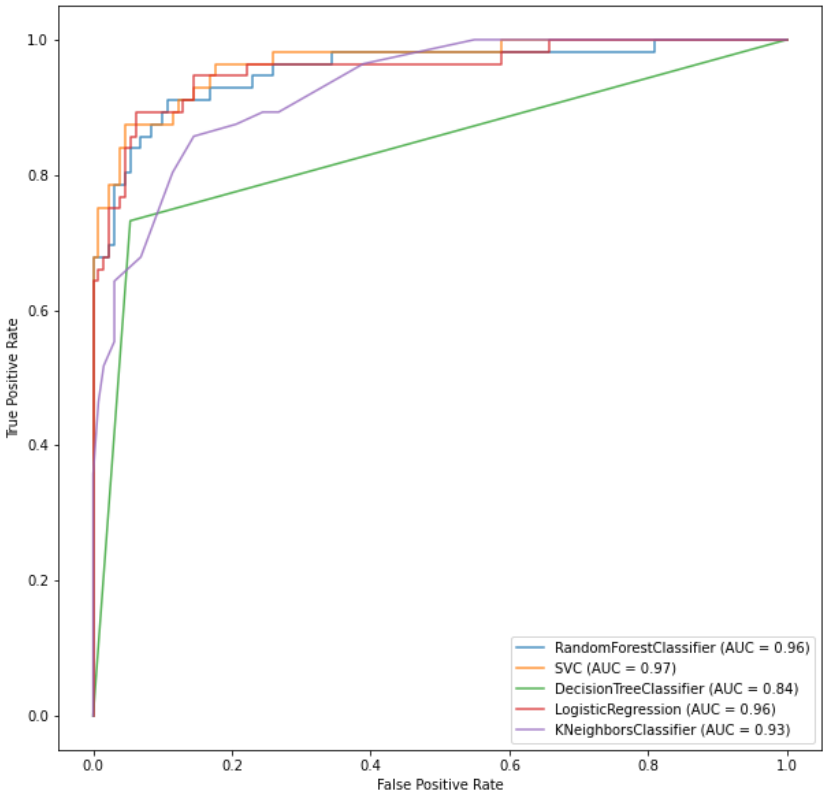


Figure 41 : ROC courbes des 5 modèles d'apprentissage utilisés

A première vue, nous pouvons constater que les modèles utilisés aboutissent tous à des résultats bons, avec des AUC qui se rapproche de 1.

Nous calculons ensuite les indices de performances présentés dans la partie état de l'art (voir Chapitre 2.1.7) de chaque modèle en nous basant sur les matrices de confusion obtenues et présentées ci-après :

(+ non faillite, - faillite)

Tableau 11: Matrice de confusion « Régression logistique »

Régression Logistique	+	-
+	1250	60
-	90	470

Tableau 12: Matrice de confusion « Random Forest »

Random Forest	+	-
+	1230	80
-	100	460

Tableau 13 : Matrice de confusion « SVM »

SVM	+	-
+	1250	60
-	80	480

Tableau 14: Matrice de confusion « KNN »

KNN	+	-
+	1220	90
-	180	380

Tableau 15 : Matrice de confusion « Décision Tree »

Decision Tree	+	-
+	1220	90
-	160	400

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

Tableau 16 : Indices de performance des cinq modèles utilisés

	Accuracy	Précision	Spécificité	Sensibilité	AUC
Logistic Regression	0.9197	0.9291	0,83928571	0,9541985	0.96
Decision Tree	0.8663	0.8717	0,71428571	0,9312977	0.84
Random Forest	0.9037	0.9037	0,82142857	0,9389313	0.96
SVM	0.9251	0.9251	0,85714286	0,9541985	0.97
KNN	0.8556	0.8503	0,67857143	0,9312977	0.93

En plus des indices de performances calculés précédemment, l'importance donnée aux variables joue un rôle important dans le choix du modèle. Après discussion avec les analystes risque sur la contribution de chaque variable dans la décision finale, et de la façon dont chaque modèle les a considérés, pour choisir le modèle d'apprentissage dont la manière d'interpréter et d'analyser les ratios se rapproche de celle de l'analyste et donc de la réalité.

5.1. Comparaison de l'importance des variables entre les différents modèles :

Dans ce qui suit, nous allons comparer les résultats de chaque modèle **(Voir les Annexe J, Annexe K , Annexe L, Annexe M, Annexe N)** vis-à-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 six variables (Capitaux propres au passif, Rentabilité des capitaux propres, Flux de trésorerie par rapport au passif, Fonds de roulement par rapport aux actifs, Résultat d'exploitation, Taux d'intérêt débiteur, Constant taux d'intérêt (après impôts)) ont contribué 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 pas le plus optimaux.

Contrairement au reste des modèles où nous avons constaté des importances avec des taux différents :

Pour les modèles, Régression logistique et KNN, 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 SVM, Nous avons obtenu des importances à valeurs positive ; sachant que le SVM a donné des taux de contribution plus signifiant que du modèle Random Forest.

5.2. Comparaison des modèles à la base des scores attribués :

Afin de trancher parmi les modèles élaborés, nous avons décidé de leur attribuer des scores allant de 1 à 4 selon leurs performances et calculé la moyenne générale de ces scores, afin de pouvoir les classer et choisir le modèle le plus performant pour notre problématique.

Les résultats obtenus sont présentés dans le tableau ci-après :

Tableau 17: Comparaison entre les modèles après l'attribution des scores

	Accuracy	Precision	Specificité	Sensibilité	AUC	Importance des variables
Logistic Regression	4	4	3	4	4	4
Decision Tree	2	2	2	3	1	1
Random Forest	3	3	3	3	4	3
SVM	4	4	4	4	4	3
KNN	2	1	1	3	3	3

Tableau 18: Score moyen des modèles utilisés

	Score moyen
Logistic Regression	3,833333333
Decision Tree	1,833333333
Random Forest	3,166666667
SVM	3,833333333
KNN	2,166666667

5.3. Comparaison entre la partie étude bibliographique et solution :

Dans ce qui va suivre nous allons comparer la précision obtenue par les modèles utilisés dans la partie ‘‘étude bibliographique’’ et dans notre solution.

Tableau 19 : Comparaison de la précision

Modèle	Partie Bibliographique	Notre Projet
Logistic Regression	76.29%	92.91%
Decision Tree	78.46%	87.17%
Random Forest	87.06%	90.37%
SVM	79.77%	92.51%
KNN	97.2%	85.03%

Comme montré dans le tableau ci-dessus, nous remarquons que les résultats obtenus lors de notre intervention au sein de SGA se rapprochent d'une manière aux résultats constatés dans les articles étudiés, nous pouvons citer à titre exemple le modèle Random Forest où la précision obtenue par ce modèle est presque identique dans les 2 cas (87.06% pour la partie étude bibliographique et 90.37% pour notre cas) ; bien sûr nous constatons aussi une différence majoritairement importante dans quelques modèles comme le Logit, SVM. Cette différence revient beaucoup plus à la nature et la composition des données utilisées pour les modèles, en y'ajoutant aussi les ''features'' utilisés.

Pour conclure nous pouvons affirmer que la partie ''étude bibliographique'' nous a été d'une grande aide lors du choix de nos modèles, il existe aussi d'autres modèles assez performant dans ce contexte comme les réseaux de neurones dont nous n'avons pas pu choisir lors de notre intervention pour des raisons de restrictions rencontrées au sein de SGA.

Conclusion :

A travers les résultats des deux tableaux précédents, nous constatons que le modèle de Logistic Regression et SVM obtiennent les meilleurs résultats avec une moyenne égale à 3.833. Toutefois, pour la suite du travail, nous privilégierons le modèle de régression logistique, puisqu'il comporte moins de paramétrages, d'ajustements et temps d'exécution.

6. Deployement :

Dans cette partie, nous allons expliquer comment, notre modèle prédictif obtenue avec Logistic regression va être assigner au sein de SGA afin de répondre à la problématique que nous avons définie précédemment.

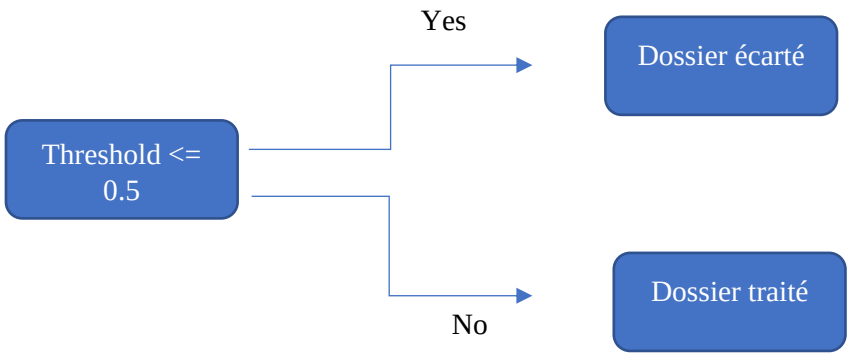
Pour des raisons de réglementations strictes imposés par le groupe société générale, le modèle mis en place ne pourra pas être utilisé directement sur les nouvelles demandes de crédit avant qu'il soit validé par le top management, et pour cela une demande a été remontée au groupe pour qu'il puisse constater par eux même les résultats obtenus par le modèle, sa précision et le gain du temps que peut engendrer un tel modèle d'aide à la décision.

Une fois validé, le modèle va prendre place sur la 2ème ligne défensive interne de SGA à côté de la direction des risques (comme indiqué sur la figure suivante).

Figure 42 : Contrôle interne de Société Générale

En effet, étant donnée l'important volume des dossiers de crédit corporate passant par le département risque, et la nécessité d'effectuer une inspection minutieuse de chaque dossier spécialement l'information financière des demandeurs ce qui engendre une perte de temps majeur si la santé financière du demandeur ne serait pas adéquate pour une demande de crédit et donc un freine pour atteindre les objectifs fixés par le groupe, donc l'objectif majeur de ce modèle étant de permettre un gain de temps pour les analystes risque vu le volume énormes des demandes de crédits auxquels ils font face durant l'année, et la détection de patterns présents parmi les demandeurs en se basant sur tout l'historique de la banque, une chose qui ne pourrait pas être faites par la compétence-humaine.

Pour commencer nous allons fixer un threshold de 0.5, c'est-à-dire toute entreprise ayant une probabilité de défaillance supérieur ou égale à 50% selon le modèle son dossier va être écarté et mis de côté par les analystes risques, pour ensuite le rejeter ou demander un complément de dossier ou des informations supplémentaires pour renforcer et expliquer l'état actuel de leur entreprise.



La valeur du seuil va varier selon les résultats obtenus par le threshold précédent, pour soi l'augmenter si l'apport du dernier seuil n'a pas été très important ou le baisser si les demandes de crédits dans une période ne vont pas être très importante.

6.1. Validation, utilisation et suivi :

Après la proposition du projet réalisé pour le compte de notre responsable, nous avons proposé des points à prendre en considération pour concrétiser la fiabilité du modèle, l'un de ces points, c'est de mener des extractions des clients archivés (Année 2018 et 2017) et de tester le modèle sur ces derniers vu que la population sur laquelle nous avons travaillé ci-dessus était des clients qui n'ont pas été encore archivés (Année 2019 et 2020). Un autre point est de tester les résultats sur les nouveaux clients et évaluer la décision à la fin de l'échéance.

Concernant l'utilisation et l'exploitation de notre travail, c'est une chose très simple à réaliser. Il suffit juste d'intégrer les valeurs des variables du nouveau client dont on veut tester la santé financière et prédire sa probabilité de tomber en situation ‘'impayés'' dans une liste de données ***''Newclient[]''***, ensuite il suffit seulement d'exécuter la fonction ***model.predict_proba(Newclient)*** afin d'obtenir la probabilité de ce dernier. Nous avons aussi proposé une option pour le choix du modèle, vu que nous avons obtenus de bons résultats avec la Logistique Régression et la Support Vecteur Machine, un volet option qui permet de choisir entre les deux modèles et de calculer l'écart entre les deux modèles.

Quant à la surveillance et à la maintenance du modèle, plus le DataSet sera alimenté en données, et plus il sera robuste et apte à détecter de nouveaux patterns, à condition que les données utilisées soient exemptes d'anomalies. Cependant, l'impact sur le comportement du modèle doit être surveillé en fonction de la quantité de données alimentées : pour une petite quantité de données, l'impact est négligeable, mais une quantité considérable de 100 000 données ou plus affectera le modèle. Une révision est nécessaire (importance des variables, R-carré) et une évaluation des performances et de la précision.

6.2. Projet et Axes d'améliorations :

Notre projet s'est bien déroulé, notamment dans la phase de modélisation où nous avons obtenu de bons résultats en termes de score, de matrice de confusion, de prédiction et de recherche des hyper paramètres optimaux.

En ce qui concerne le traitement des données, il aurait été préférable de disposer d'une base de données toute prête contenant toutes les informations financières des clients afin d'obtenir un ensemble de données d'apprentissage plus fiable et riche en termes de cas de faillite et de quantité de données.

Comme tout projet, des axes d'amélioration ont été proposés afin d'assurer une amélioration continue du service et de la qualité de l'outil, pour ce faire nous avons proposé quelques

améliorations concernant l'outil et l'intégration d'autres fonctionnalités qui prennent en considération d'autres lignes de crédit.

- La 1ère amélioration réside dans la mise en place d'une interface pour faciliter son utilisation par les non-informaticiens
- Le 2ème axe d'amélioration est l'intégration de la ligne de crédit " leasing " dans le modèle, afin de raccourcir le temps de traitement du dossier en le considérant d'une manière différente des autres crédits, en lui ajoutant de nouveaux critères.
- Le 3ème axe d'amélioration concerne l'exploitation des données du STARWEB afin d'enrichir le modèle par de nouvelles informations sur les profils des clients, ces informations sont stockées sous forme PDF sur les fiches STARWEB pour chaque client, cette quantité d'information va permettre à l'outil de prendre en compte le profil du client vis-à-vis la position dans le marché, le management, le domaine d'activité …

6.3. Conclusion

Dans ce chapitre, nous avons d'abord suivi les étapes de la méthodologie d'exploration de données CRISP-DM décrite dans le chapitre précédent. La préparation de notre base de données s'est faite selon les étapes dictées par la méthodologie et son utilisation a permis de construire des modèles prédictifs. Ces modèles ont été testés et leurs résultats étaient globalement très satisfaisants. Deux algorithmes se sont démarqués des autres en présentant les meilleures performances en termes de métriques que nous avons utilisées pour les évaluer. Néanmoins, nous avons choisi la régression logistique afin d'apporter les modifications nécessaires pour l'adapter aux besoins de notre problème. Une fois fait, nous avons proposé une méthode pour son déploiement au sein de la SGA une fois validé par le groupe, tout en ajoutant des axes d'amélioration continue à mettre en place dans un futur proche.

Conclusion Générale

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 Société Générale Algérie.

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 SGA.

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.

On est ensuite passé à la mise au point de notre solution en choisissant le cheminement CRISP-DM comme guide dans notre démarche de solution.

Après avoir récupéré la base de données existante et l'avoir traitée, nous nous sommes appliqués à l'enrichir en extrayant les informations financières présentes sur les notices financières (attachées sur le DCCIT) pour les consolider sur à la base de données actuelle, ainsi qu'en calculant de nouveaux ratios pour mettre en évidence la santé financière des clients avec la collaboration des analystes risque. Une fois terminée, nous avons exploité cette nouvelle base de données pour construire 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 pour le déployer au sein du département risque où il sera utilisé par les analystes risque lors des prochaines analyses.

Dans le même esprit, des perspectives ont été proposées afin de s'inscrire dans une démarche d'amélioration continue, ainsi nous recommanderons d'améliorer continuellement les modèles

que nous avons proposés en intégrant de nouvelles fonctionnalités et de nouveaux jeux de données pour améliorer l'apprentissage et éventuellement un jour automatiser cette tâche de manière définitive.

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.

Références

Références :

[1] «Définitions : banque - Dictionnaire de français Larousse,» [En ligne]. Available: https://www.larousse.fr/dictionnaires/francais/banque/7863.

[2] «Banque : définition, traduction et synonymes - JDN,» [En ligne]. Available: https://www.journaldunet.fr/business/dictionnaire-economique-et-financier/1198859-banque-definition-traduction-et-synonymes/#:~:text=D%C3%A9finition%20du%20mot%20Banque&text=Elle%20constitue%2C%20juridiquement%2C%20une%20institution,g%C3%A9rer%20les%20moy.

[3] «Crédit — Wikipédia,» [En ligne]. Available: https://fr.wikipedia.org/wiki/Cr%C3%A9dit.

[4] «Le crédit bancaire : Définition du crédit bancaire - WikiMemoires,» [En ligne]. Available: https://wikimemoires.net/2011/05/credit-bancaire-definition-de-credit-bancaire/#:~:text=Le%20mot%20%C2%AB%20Cr%C3%A9dit%20%C2%BB%20%C3%A0%20la,attend%20le%20remboursement%20du%20pr%C3%AAt..

[5] «Risque bancaire — Wikipédia,» [En ligne]. Available: https://fr.wikipedia.org/wiki/Risque_bancaire.

[6] «Le risque de crédit: évaluation à partir des ... - Memoire Online,» [En ligne]. Available: https://www.memoireonline.com/07/09/2318/m_Le-risque-de-credit-evaluation-a-partir-des-engagements-des-banques-aupreacut5.html.

[7] P. Du Jardin, «Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy,» *Neurocomputing,* 2011.

[8] M. Virág, «A cs″odmodellek jellegzetességei és története,» *Vezetéstudomány.*

[9] T. Nyitrai, «Dinamikus pénzügyi mutatószámok alkalmazása a cs″odel″orejelzésben. Ph.D. thesis,» *Budapesti Corvinus Egyetem,* 2015.

[10] P. J. Fitzpatrick, «A Comparison of the Ratios of Successful Industrial Enterprises with Those of Failed Companies,» *Washington: The Accountants' Publishing Company.*

[11] D. Durand, «Risk Elements in Consumer Instalment Financing,» *National Bureau of Economic Research.*

[12] W. H. Beaver, «Financial ratios as predictors of failure. Empirical research in accounting: selected studies,» *Journal of Accounting Research* .

[13] J. H. a. E. W. F. Myers, «The development of numerical credit evaluation systems,» *Journal of the American Statistical Association* .

[14] E. I. Altman, «Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,» *The Journal of Finance.*

[15] D. L. Chesser, «Predicting loan noncompliance,» *Journal of Commercial Bank Lending* .

[16] J. A. Ohlson, «Financial ratios and the probabilistic prediction of bankruptcy,» *Journal of Accounting Research* .

[17] M. E. Zmijewski, «Methodological issues related to the estimation of financial distress prediction models,» *Journal of Accounting Research* .

[18] H. E. I. A. a. D.-L. K. Frydman, «Introducing recursive partitioning for financial classification: The case of financial distress.,» *The Journal of Finance* .

[19] M. D. a. R. S. Odom, «A neural network model for bankruptcy prediction. Paper present at the International Joint Conference on Neural Networks,» *Ann Arbor: IEEE Neural Networks Council.*

[20] D. a. Y. A. T. Vlachos, «Neuro-fuzzy modeling in bankruptcy prediction.,» *Yugoslav Journal of Operational Research* .

[21] A. a. M. P. Fan, «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.*

[22] M. N. V. Z. M. M. S. a. E. S. Ardakhani, «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.*

[23] T. A. E. a. T. E. M. Lensberg, «Bankruptcy theory development and classification via genetic programming,» *European Journal of Operational Research.*

[24] S. M. Bryant, «A case-based reasoning approach to bankruptcy prediction modeling, Intelligent Systems in Accounting,» *Finance and Management* .

[25] A. I. V. G. a. J. S. S. Marqués, «Exploring the behaviour of base classifiers in credit scoring ensembles,» *Expert Systems with Applications ,* 2012.

[26] N. Wang, «Bankruptcy Prediction Using Machine Learning,» *Journal of Mathematical Finance* .

[27] F. H. K. a. E. I. A. Barboza, «Machine learning models and bankruptcy prediction,» *Expert Systems with Applications.*

[28] «Bootstrap aggregating — Wikipédia,» [En ligne]. Available: https://fr.wikipedia.org/wiki/Bootstrap_aggregating#:~:text=Le%20bootstrap%20aggregating%2C%20%C3%A9galement%20appel%C3%A9,permet%20d'%C3%A9viter%20le%20surapprentissage..

[29] «Courbe ROC,» [En ligne]. Available: https://fr.wikipedia.org/wiki/Courbe_ROC.

[30] «Validation_croisée,» [En ligne]. Available: https://fr.wikipedia.org/wiki/Validation_croisée.

[31] «Qu'est-ce que le processus ETL | Oracle France,» [En ligne]. Available: https://www.oracle.com/fr/database/processus-etl-

definition.html#:~:text=Extraction%2C%20transformation%2C%20chargement%20(ETL,charge%20dans%20un%20Data%20Warehouse..

[32] «ETL (Extract, Transform, Load) - Wikipédia,» [En ligne]. Available: https://fr.wikipedia.org/wiki/Extract-transform-load.

[33] «miniconda [Wiki ubuntu-fr] - Documentation Ubuntu,» [En ligne]. Available: https://doc.ubuntu-fr.org/miniconda.

[34] [En ligne]. Available: www.jupyter.org. Retrieved 2020-11-13..

[35] «Pandas,» [En ligne]. Available: https://pandas.pydata.org/pandas-docs/pandas documentation. 28 January 2020..

[36] «Cross Industry Standard Process for Data Mining — Wikipédia,» [En ligne]. Available: https://fr.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining.

[37] C. R. Harris, K. J. Millman, S. J. v. d. Walt et e. al, Array programming with NumPy, 16 September 2020.

[38] W. McKinney, pandas: a Foundational Python Library for Data Analysis and Statistics, 2 August 2018.

[39] a. Pedregosa et G. Varoquaux., «Scikit-learn: Machine Learning in,» *Journal of Machine Learning Research.*

[40] «Le risque de crédit: évaluation à partir des ... - Memoire Onlin,» [En ligne]. Available: https://www.memoireonline.com/07/09/2318/m_Le-risque-de-credit-evaluation-a-partir-des-engagements-des-banques-aupreacut5.html.

[41] E. I. Altman, «Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,» *The Journal of Finance* .

[42] J. A. Ohlson, «Financial ratios and the probabilistic prediction of bankruptcy,» *Journal of Accounting Research.*

[43] L. a. P. P. S. Sun, «Using Bayesian networks for bankruptcy prediction: Some methodological issues,» *European Journal of Operational Research.*

[44] A. I. R. S. R. S. a. C. Z. Dimitras, «Business failure prediction using rough sets,» *European Journal of Operational Research* .

[45] E. a. C. M. M. Neophytou, «Predicting Corporate Failure in the UK: A Multidimensional Scaling Approach.,» *Journal of Business Finance and Accounting.*

[46] K. Kiviluoto, «Predicting bankruptcies with the self-organizing map.,» *Neurocomputing* .

Annexes

Annexe A

Le biais :

C'est l'erreur provenant d'hypothèses erronées dans l'algorithme d'apprentissage. Un biais élevé peut être lié à un algorithme qui manque de relations pertinentes entre les données en entrée et les sorties prévues (sous-apprentissage).

La variance :

C'est l'erreur due à la sensibilité aux petites fluctuations de l'échantillon d'apprentissage. Une variance élevée peut entraîner un surapprentissage, c'est-à-dire modéliser le bruit aléatoire des données d'apprentissage plutôt que les sorties prévues.

Coefficient de Gini :

C'est une mesure statistique permettant de rendre compte de la répartition d'une variable au sein d'une population. Autrement dit, il mesure le niveau d'inégalité de la répartition d'une variable dans la population. Sa formule est la suivante :

$$\text{Gini index} = 1 - \sum_{i=1}^{n} (p_i)^2$$

Entropie :

La quantité d'information qui est nécessaire afin de décrire un échantillon. Si cet échantillon est homogène, et donc ne contient que des éléments similaires, l'entropie est nulle. Dans le cas contraire, si l'échantillon est uniformément réparti entre ses éléments, alors l'entropie atteint son maximum, qui est de 1. Sa formule est la suivante :

$$\text{Entropy} = -\sum_{i=1}^{n} p_i * \log(p_i)$$

Classification and Regression Trees or CART:

Arbres de classification et de régression ou CART en abrégé est un terme introduit par Leo Breiman pour désigner les algorithmes d'arbres de décision qui peuvent être utilisés pour les problèmes de modélisation prédictive de classification ou de régression. Classiquement, cet algorithme est appelé "arbres de décision", mais sur certaines plateformes comme R, il est désigné par le terme de CART.

Annexe B

Segment Marché des particuliers :

Le marché des particuliers comprend des personnes physiques dont les besoins bancaires sont liés à leur vie privée. Il se décompose en 03 principaux segments :

1. Patrimonial 2. Bonne Gamme 3. Grand Public

Cette segmentation du marché « Particuliers » est calculée en se basant sur les critères ci-dessous :

- Total des avoirs détenus par le client - Flux régulier

- Zone d'habitation du client - Age du client - Profession du client

La frontière entre les trois segments sera déterminée à l'issue d'une analyse de données réalisée par le département marketing stratégique études & datamining de la banque, en effet. La segmentation est calculée en automatique tous les 6 mois avec 3 mois de données complètes.

Segment marché des professionnels :

Les Professionnels et les TPE (PRO/TPE) constituent un marché qui comprend des entités exerçant sous la forme de personne physique ou sous la forme de personne morale.

Le critère de segmentation ici est le chiffre d'affaires, ainsi : ce marché comprend des entités dont le chiffre d'affaires est inférieur à 150MEURO.

Professionnels personnes morales :

Ce sont des personnes morales structurées sous forme de société commerciale ou industrielle (TPE).

Professionnels personnes physiques :

Il s'agit des professionnels qui exercent leur activité sous forme d'entreprise individuelle.

La personne physique exerce alors un rôle d'exploitant direct dans une entreprise individuelle. Ce dernier et cette même entreprise individuelle ne forment alors juridiquement qu'une seule et même personne.

Les professionnels personnes physiques sont rattachées au marché des Professionnels dès lors qu'ils sont en relation avec la Banque pour tout ou une partie de leur activité professionnelle.

Petites associations :

Il s'agit des Associations dont le montant des ressources annuelles est inférieur ou égal à 1000 EURO.

Segment Marché Entreprises :

Il comprend les personnes morales pour lesquelles les approches commerciales et marketing relèvent du périmètre Non Retail. Par principe, ce sont des personnes morales structurées sous forme de sociétés commerciales et industrielles à capitaux privés ou publics y compris celles du groupe SG

Les personnes morales appartenant au marché « Entreprises » font l'objet d'une approche commerciale sélective. Il convient en effet de privilégier des sociétés de bonne qualité ayant

une santé financière solide et qui représentent un potentiel de PNB non seulement en termes de marge d'intérêt mais également en termes de commissions.

Le marché Entreprises se décompose en :

Petite Entreprises (PE) :

Personnes morales ou groupes de personnes morales dont le chiffre d'affaires ne dépasse pas 5 MEUR.

Moyennes Entreprises (ME) :

Personnes morales ou groupes de personnes morales dont le chiffre d'affaires se situe entre 5 MEUR et 50 MEUR.

Grandes Entreprises (GE) :

Personnes morales ou groupes de personnes morales dont le chiffre d'affaires est supérieur à 50 MEUR.

Segment collectivités locales, États et Institutions publiques :

Sont définis comme souverains les Etats et Administrations centrales incluant les banques centrales et éventuellement :

- Etats et administrations centrales - Gouvernements nationaux - Ministères

Sont définis comme supranational tout Organisme qui se situe au-dessus des autorités nationales ; qu'il dépasse ainsi, la souveraineté de l'état, ex. : l'Union européenne.

Segment institutions financières :

Ce segment regroupe des institutions financières soumises à la supervision d'autorités de tutelle et de contrôle de leur activité tel que les Banques et les Assurances.

Principes relatifs à la gestion et aux relations avec les tiers :

Les tiers sont des personnes morales ou physiques extérieures au groupe Société Générale et incluent les clients, les fournisseurs de services financiers, les fournisseurs (de biens et de produits ou de services non-financiers) ainsi que toutes les personnes impliquées dans une opération effectuée par une Entité du Groupe qui ne sont ni des clients, ni des fournisseurs de services financiers, ni des fournisseurs (appelées parties connexes). Ne sont pas compris dans la notion de tiers, les employés du Groupe et joint-ventures du Groupe ainsi que les personnes qui font l'objet d'opérations de fusion/acquisition avec ou par le Groupe.

Les tiers sont soumis soit au processus KYC (know your customer), soit au processus KYS (know your supplier). Ces deux processus ont notamment vocation à répondre aux obligations réglementaires en matière de Lutte Contre le Blanchiment et Financement du Terrorisme (LCB-FT), Sanctions et Embargos, Lutte contre la Corruption, et Trafic d'Influence (ABC), Responsabilité Sociale et d'Entreprise (RSE), FATCA, CRS, classification EMIR, etc.

▪ Le processus KYC vise notamment à répondre aux obligations réglementaires en matière de sécurité financière. Il s'applique aux clients et aux fournisseurs de services financiers. Un processus KYC spécifique s'applique aux parties connexes.

▪ Le processus KYS vise notamment à répondre aux obligations relatives à la lutte contre la Corruption et le Trafic d'influence et participe à la gestion du risque d'image du Groupe. Le processus KYS s'applique aux fournisseurs (de biens, produits ou services non-financiers).

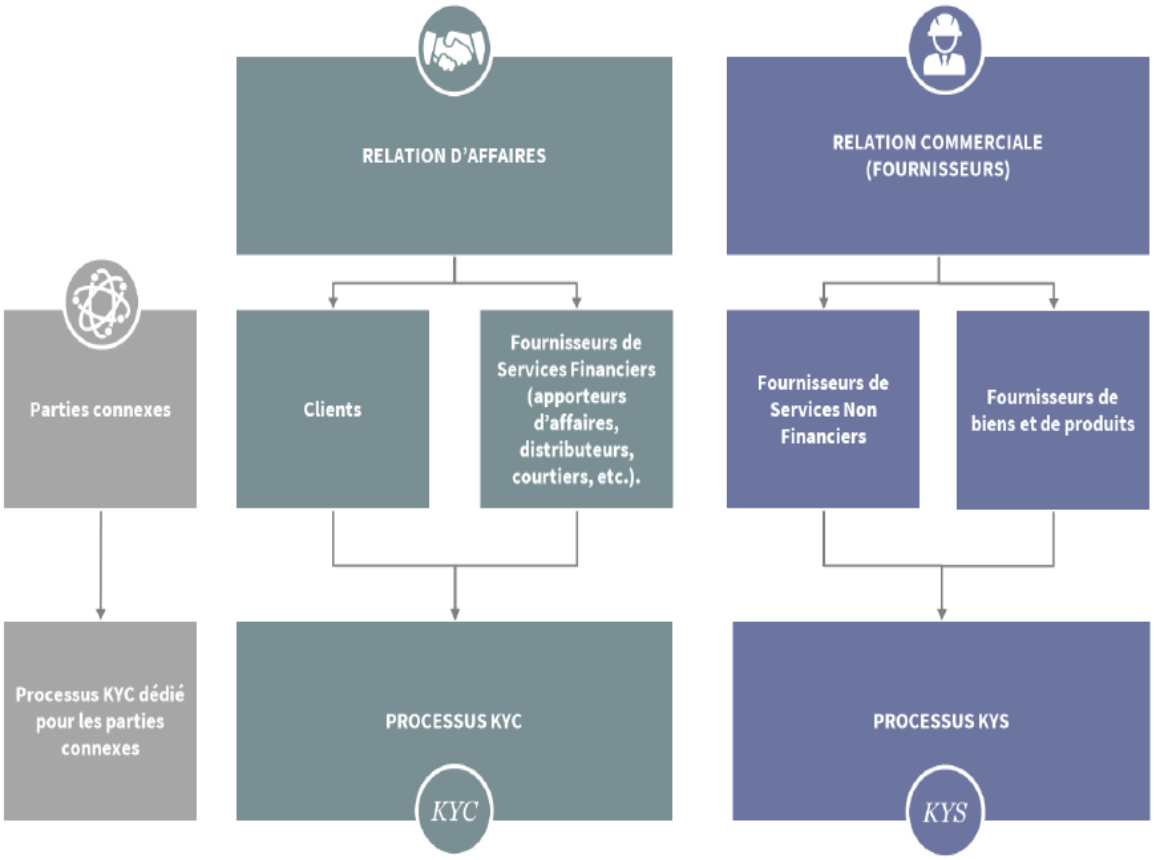


Figure 43 : Processus de gestion des relations avec les tiers

Principes KYC :

La connaissance client ou Know Your Customer (KYC) regroupe un ensemble de règles, d'origine réglementaire ou non, visant à recueillir des informations sur la Relation d'Affaires, évaluer son profil de risque et de prendre une décision quant à l'entrée en relation ou au maintien d'une relation d'affaires existante.

Le processus KYC permet également d'offrir aux clients un produit ou un service qui leur est adapté tout en assurant la maîtrise du risque de non-conformité, d'une part, et de réputation, d'autre part.

▪ Le risque de non-conformité découle notamment de la non-application des règles relatives :

- à la protection de la clientèle ;

- à la sécurité financière ;

- à la lutte contre la corruption et le trafic d'influence ;

- aux risques environnementaux et sociaux ;

- à la transparence fiscale (en particulier Foreign Account Tax Compliance Act (FATCA) et Common Reporting Standard (CRS)) ;

- à toute autre réglementation contenant des obligations relatives à la connaissance du client.

▪ le risque de réputation découle non seulement des sanctions liées au non-respect de ces obligations, mais aussi de tout autre élément relatif au client comme par exemple des soupçons ou des poursuites judiciaires pour des crimes ou délits, des atteintes à l'environnement, des comportements non éthiques.

Le processus KYC ne doit pas être entendu comme la simple collecte de documents ou d'informations mais comme l'analyse circonstanciée d'un ensemble d'éléments de nature à évaluer le profil de risque du client et les mesures de vigilance à mettre en œuvre.

Principes De Connaissance Des Tiers Non Clients (KYS) :

Le dispositif KYS regroupe un ensemble de règles qui doit permettre de connaître le fournisseur en vue d'évaluer son profil de risque et d'être en mesure de prendre une décision quant à l'entrée en relation commerciale ou au maintien d'une relation commerciale existante (contrat en cours).

Ces règles doivent notamment permettre au Groupe SG d'identifier les fournisseurs exposés principalement au risque de corruption et de gérer son risque.

L'analyse circonstanciée des informations collectées doit permettre d'établir une classification des risques présentés par un fournisseur, reposant sur une méthodologie de notation fondée sur des critères objectifs spécifiques au dispositif KYS. Identifiés par le Groupe comme des facteurs de risque au regard de la corruption.

Annexe C

Tous les dossiers de crédit concernant la clientèle corporate doivent être établis et transmis via DCCIT.
Les dossiers de crédit doivent contenir :

- Les lignes de crédit avec leur date de validité et les modalités d'utilisation ;
- Les documents de base qui sont la notice financière et le dossier de demande d'autorisation ;
- Tous les documents, pièces et informations utiles à l'étude et à la compréhension du dossier de crédit.

Pour les contreparties composées de plusieurs entités juridiques, une notion de « groupe-client » est indispensable pour une meilleure appréciation. Elle se compose de :

- Une présentation du groupe ;
- Les autorisations mises en place par la banque sur le groupe ;
- Les relations entretenues par la banque avec le groupe ;
- Une synthèse présentant notamment la stratégie du groupe et la politique de la banque à son égard.

Tout dossier doit comporter une analyse financière et un diagnostic économique qui portera sur l'analyse des éléments suivants :

- La présentation de l'activité de la contrepartie ;
- Le schéma du cycle d'exploitation ;
- La stratégie de la contrepartie justifiant la mise en place des lignes ;
- Les risques et facteurs d'atténuation ;
- Le management.

1-Informations à recueillir :

a) Documents à collecter :

La documentation exhaustive à recueillir pour l'instruction d'un dossier de crédit a été regroupée dans une check liste mise à la disposition des commerciaux. Cette check liste doit obligatoirement accompagnée tous les dossiers de crédit et faire l'objet d'un contrôle minutieux de la part des instances décisionnelles.
La complétude du dossier de crédit est un préalable pour sa réception, afin d'enclencher son étude. Le fait d'accepter un dossier de crédit engage la banque en matière de date de réception et de délais de décision. A cet effet, tout dossier incomplet ne doit pas être accepté.
Aussi, la liste des documents à réclamer par nature de crédit est consignée sur la check-list, ceci n'exclut pas de réclamer au besoin tout autre document probant afin de consolider le fonds documentaire du dossier.
Le montage du dossier de crédit doit se faire dans le respect des canevas arrêtés par la Banque.

b) Les critères à retenir pour apprécier la qualité d'une contrepartie :

- Âge de la société : les sociétés de création récente (< à 3 ans d'existence) doivent faire l'objet d'une attention particulière et d'une analyse approfondie en raison d'un taux de défaillance élevé sur cette catégorie ;
- Forme juridique : favoriser les sociétés de capitaux aux sociétés de personnes ;

- Situation économique de la société : position de la contrepartie sur son marché ;
- Situation financière : obtention des trois derniers bilans sauf si investissement récent (de création ou d'extension), et dans ce cas il est important de se concentrer sur la capacité à générer des cash flows ;
- Secteur d'activité compatible avec la politique commerciale de SGA ;
- Qualité des relations bancaires : absence d'incidents de paiement (chez SGA ou chez les autres banques) et respect des délais de paiement.

Pour pouvoir se prononcer sur la faisabilité d'un montage de dossier de crédit, le recueil d'informations disponibles en internes ou en externe est un exercice obligatoire. La contrepartie doit respecter un certain nombre de critères d'éligibilité, qui permettent d'encadrer la prise de risque pour la banque.
Ces critères ont pour objectifs d'exclure du périmètre du financement, les emprunteurs ou opérations jugés trop risqués de par leur localisation géographique, leur appartenance sectorielle, ou leur faible ancienneté professionnelle.
Le chargé du dossier doit s'assurer du respect des critères d'éligibilité et de l'exhaustivité des données. La collecte de ces éléments et les interrogations requises visent à donner une appréciation globale du risque et de la situation du client. De ce fait, le non-respect de l'un de ces critères ou de l'absence d'un document, devra être argumenté ainsi que la décision de poursuivre le processus de la demande de financement.
Une approche dynamique est aussi à privilégier dans l'appréciation des différents équilibres économiques et financiers, ce qui exige une analyse dans le temps, qui ne peut se faire sur un seul exercice (analyse obligatoire sur les trois derniers exercices clos).

c) Les informations internes à la banque :

Lors du traitement d'un dossier de crédit, quel que soit le type de financement, il est nécessaire de recueillir et analyser les informations reprises sur les bases internes suivantes :

- La dernière notification pour vérifier le respect des covenants et conditions et recueil des garanties exigées ;
- Le fonctionnement du compte courant de la relation et des comptes annexes pour vérifier le respect des limites autorisées, l'utilisation des lignes, la fréquence et justification des dépassements et impayés, et le respect des prévisions de régularisation. Il est nécessaire d'élargir l'analyse sur les deux derniers exercices ;
- Le calcul des Mouvements Créditeurs Confiés : Il s'agit de la somme des flux créditeurs générés par l'activité commerciale du client ;
- Le calcul du PNB
- La consultation des lignes métiers et filiales SG dans le cas d'un groupe ayant des implantations à l'étranger ;
- La consultation de la note STARWEB en cours de validité et du projet de notation précédent pour l'appréciation du développement de l'affaire ;
- En cas de notion « Groupe – client », l'analyse du fonctionnement des comptes ouverts au nom des filiales du même groupe.

Cependant, tout dossier sur lequel s'applique un des critères ci-dessous, ne sera pas recevable, sauf dérogation du responsable du marché sur la base d'un mémo explicatif dûment justifié :

- Existence d'un impayé en cours ;
- Existence d'un prêt en procédure de recouvrement ;

En plus de ces éléments, une attention particulière devra être portée en cas de survenance d'un ou de plusieurs signaux d'alerte, et qui doivent être argumentés :

- Une notation STARWEB ≥ 6 + ;
- La présente d'un découvert structurel et / ou de dépassements récurrents des autorisations, avec multiplication des impayés et difficultés à les absorber ;
- Une baisse substantielle et non saisonnière des mouvements confiés, et l'absence de flux durant une durée > à 3 mois ;
- L'existence de conflits entre les associés et / ou dirigeants avec risque de compromettre l'activité de la relation ;
- Une baisse significative des opérations confiées (baisse des opérations Trade, des marchés domiciliés, etc…) ;
- Une mise en jeu des cautions par les bénéficiaires ;

Dans le cas où, sur les 6 derniers mois, le client a effectué une ou plusieurs demandes de crédits sans finalisation, le commercial doit en prendre connaissance pour juger de sa pertinence et évaluer la fiabilité des données déclarées : si les demandes effectuées et / ou saisies présentent des données déclaratives différentes de la demande actuelle, le commercial doit être vigilant et s'assurer de la bonne foi du client.

d) Les informations qualitatives :

La procédure KYC est désormais obligatoire pour toute nouvelle mise en place ou renouvellement de crédit.
La fiche KYC à rattacher au dossier doit permettre de s'assurer de la présence de l'ensemble des justificatifs y énumérés. La validation du KYC est un préalable à la saisie des lignes.

2-Analyse de l'information financière :

a) Informations financières :

Le chargé du dossier doit être en mesure d'apprécier la santé financière de l'entreprise emprunteuse. Il convient donc de pouvoir observer son évolution sur au moins trois exercices, et éventuellement pouvoir les approcher des moyennes observées dans les mêmes secteurs d'activité.

Trois bilans définitifs devront être recueillis, ou bien à minima deux bilans définitifs et une situation comptable intermédiaire de l'exercice en cours lorsque le client ne dispose pas de trois bilans.

b) Traitement de l'information financière :

Les bilans fiscaux fournis doivent être certifiés par l'administration fiscale. La situation intermédiaire comptable doit également être dument estampillée par le comptable de l'entreprise. Elle doit être recueillie périodiquement comme suit :

- Au 30 juin de l'exercice en cours (n) pour chaque demande introduite sur le S2 de chaque exercice ;
- Au 31 décembre de l'exercice précédent (n-1), pour chaque dossier introduit sur le S1 de l'exercice en cours.

Les situations intermédiaires devront être datées au plus de 06 mois. Dans cet intervalle, certains clients disposent des informations financières à fréquence moins large (par exemple mensuelle, ou trimestrielle). Dans ce cas, et sans obligation, le chargé du dossier peut les réclamer si elles sont disponibles au moment du montage du dossier.
Dans le cas où le client ne peut pas formaliser une situation comptable sous forme de bilan (actif, passif, TCR), le chargé du dossier peut compléter le dossier, par tout document comptable fourni par le client et dûment estampillé par le management de l'entreprise, ou à minima son comptable.

A cet effet, les documents comptables acceptables par la SGA sont :

- Les déclarations mensuelles du chiffre d'affaires G50 ;
- Les tableaux d'activité ;
- Les tableaux de bord.

c) Analyse du compte de résultat :

Le compte de résultat est l'outil qui enregistre les flux à destination (produits) et en provenance (charges) de l'entreprise. Ces charges et produits sont structurés par nature : ceux liés à l'exploitation, à l'activité financière, et les charges et produits de nature exceptionnelle. Cette présentation fait ressortir pour chaque nature de charge et de produit un résultat ou un solde intermédiaire de gestion SIG.

L'analyse du compte de résultat revient donc à évaluer :

- L'évolution des principaux soldes intermédiaires de gestion ;
- La composition du chiffre d'affaires ;
- L'existence d'un carnet de commandes, d'un plan de charge ;
- La confirmation d'un positionnement commercial, ou d'un changement dans la politique commerciale

- Les éventuels éléments exceptionnels, … etc.

En fonction du type de crédit ou du montant du financement, un compte de résultat prévisionnel sur l'exercice en cours (projections de clôture) peut être transmis par le client et doit faire l'objet d'une analyse. Les principaux agrégats / ratios analysés, et évolutions sont les suivants :

- CA
- marge brute
- EBE
- EBE/CA
- EBE/Frais financiers
- Résultat net
- Marge brute d'autofinancement (MBA)

d) Analyse des Bilans :

Le bilan donne une image sur l'actif (emploi) et le passif (ressource) de l'entreprise selon leur degré d'immobilisation et d'exigibilité. Il permet de visualiser la surface de l'actif qui permettra en cas de liquidation de faire face à l'exigible de la relation, le principe étant que les ressources à moyens et longs termes financent les actifs de même durée, et que ceux à court terme financent les actifs dit cycliques ou d'exploitation.

Les principaux agrégats/ratios analysés, et évolutions sont les suivants :

- Le total bilan : le total actif doit être égal au total passif ;
- Les immobilisations nettes ;
- Les fonds propres ;
- Les fonds propres / total bilan ;
- Les dettes financières ;
- Les dettes financières / total bilan ;
- Le gearing net (ou dettes nettes/ fonds propres) ;
- Le leverage net (ou total des dettes financières + CCA/EBE) ;
- Le fonds de roulement ;
- Le besoin en fonds de roulement

Selon le cas, d'autres ratios s'imposent, et sont repris sur la notice financière.

e) Règles relatives à l'information financière :

Les règles d'analyse doivent être approfondies lorsque les principaux agrégats financiers présentent des évolutions ou des niveaux dégradés par rapport aux normes arrêtées par la SGA.

Les niveaux de dégradations retenus par les codes couleurs « orange » et « rouge » nécessitent une appréciation approfondie des différents ratios.

- ✓ <u>Code « vert » :</u>

Correspond aux zones de confort, qui représentent les normes admises dans l'appréciation du risque financier d'une contrepartie à SGA, lesquels n'appellent pas de commentaires par les différents intervenants.

- ✓ <u>Code « orange » :</u>

Ce sont des indicateurs d'alerte sur la dégradation de la situation financière et qui nécessite des explications du client. Le chargé du dossier, une fois qu'il aura constaté ces éléments, devrait

réfléchir à limiter les lignes des crédits à blanc, imposer des covenants financiers et / ou renforcer les garanties. Si nouveau client : exiger systématiquement une garantie réelle ou financière pour la couverture totale des lignes.

- ✓ Code « rouge » :

Ce sont des signaux qui indiquent une situation critique, qui imposera systématiquement plus de vigilance donc potentiellement la revue des lignes à la baisse, une visite sur site avec vérification physique de l'état des stocks, un état comptable des stocks et prévisions d'écoulement, l'obtention des balances âgées clients / fournisseurs et des prévisions de régularisation. Il appartient au chargé du dossier d'argumenter ces éléments.

f) Analyse des données prévisionnelles :

Toute entreprise souhaitant démarrer son activité, ou procéder à un investissement recherche des solutions de financement appropriées, et passe souvent par une demande de financement d'investissement. Pour cela, le chargé du dossier doit réclamer un business plan étayé et appuyé par des bilans prévisionnels de préférence sur une période adossée à celle du remboursement du prêt sollicité.

Après avoir apprécié le contexte économique de l'entreprise et sa situation financière d'ensemble, le banquier devra critiquer les projections financières du projet et s'assurer du potentiel de ce dernier à dégager suffisamment de flux nets de trésorerie.

Il convient d'apporter une étude critique sur les éléments suivants :

- L'objet du financement est-il clairement défini ?
- L'affaire dispose-t-elle des compétences et des moyens nécessaires pour mener à bien le projet ?
- La situation financière et positionnement de l'affaire lui permettent- ils cet investissement ?
- Quel sera l'impact du projet ?
- Est-ce que le prévisionnel apparait réaliste : l'évolution du CA et de la rentabilité sont-ils cohérents avec l'investissement projeté ?
- L'évolution des paramètres d'exploitation sont-ils cohérents avec la réalité économique du projet ?
- Les hypothèses retenues dans la construction des chiffres prévisionnels sont-elles détaillées et suffisamment crédibles ?

En tout état de cause, il est fortement recommandé de dégrader le prévisionnel fourni par le client et ainsi d'élaborer des projections avec des hypothèses pessimistes.

g) Analyse du tableau des flux de trésorerie :

Le tableau des flux de trésorerie apporte un éclairage assez puissant sur la situation de la trésorerie de l'entreprise, et permet de mieux cerner le risque de faillite de celle-ci.

Ce tableau regroupe les encaissements et les décaissements liés aux activités courantes de l'entreprise, à ses investissements et ses financements. Il constitue de ce fait un état de synthèse à part entière dont la finalité première est d'expliquer la variation de la trésorerie.

Le tableau de flux de trésorerie participe à évaluer l'aptitude de l'entreprise à générer des liquidités et complète l'analyse des états financiers traditionnels que sont le bilan et le compte

de résultat. Il traduit la capacité de l'entreprise à générer de la trésorerie et à payer dans le futur ses obligations.

Il s'agit d'un outil d'aide à la décision pour plusieurs raisons :

- Le découpage par fonction fait ressortir les flux d'investissement et de financement, que l'on retrouve dans le plan de financement, utiles pour les prises de décisions stratégiques
- En écartant l'influence des charges et produits non décaissés (amortissement, transfert de charges, etc.), le tableau gomme en partie les choix comptables, eux-mêmes souvent liés aux choix fiscaux, qui brouillent la lecture du résultat comptable ;
- Le tableau permet aux investisseurs d'apprécier plus facilement la capacité de l'entreprise à générer des liquidités, à faire face aux remboursements des emprunts contractés, à honorer les actionnaires par le versement des dividendes mais également à investir sans forcément en appeler aux financements externes ;
- Il procure des renseignements indispensables aux prévisions de flux de trésorerie et contribue à une démarche budgétaire, enfin, il permet d'apprécier plus objectivement la part des mouvements créditeurs en neutralisant l'effet de décalage d'encaissement.

h) Analyse du plan de financement :

Un plan de financement doit être fourni pour tout concours moyen / long terme, et doit apporter une réponse aux points suivants :

- La quotité d'autofinancement exigée ;
- Le « loan to value » ou le ratio autofinancement / coût global du projet : la cible étant d'atteindre une quotité de 20 % sauf cas exceptionnel à déroger ;
- La période et modalité de remboursement et éventuellement celle du différé ;
- La cohérence entre le projet envisagé et l'activité du client.

Le plan de financement est généralement établi par le client, et stressé par la banque, sur la base des projections à l'occasion d'une demande de crédit d'investissement. Ce dernier traduit la capacité de l'entreprise à générer de la trésorerie et à honorer ses obligations à moyen / long terme.

En plus d'enregistrer les opérations d'encaissements et de décaissement, et les opérations courantes (MBA et variation du BFR), le plan de financement reprend toutes les entrées et sorties d'argent, qui résultent d'opérations financières et d'investissements, et qui affectent la trésorerie de l'entreprise : achats / cessions d'investissements, nouveaux apports en numéraire et crédit, …etc.

Le plan de financement est censé présenter des soldes nuls ou positifs durant toute la période de projection, autrement, il faut adopter d'autres scénarios que celui convenu initialement : exiger des apports supplémentaires, accorder des différés de paiements, exiger des covenants financiers, …etc.

3-La notation des contreparties :

La notation interne des contreparties constitue un enjeu stratégique de première importance pour Société Générale. Elle répond à la fois à un besoin interne et à une contrainte externe visant ainsi à :

- Doter notre Etablissement d'un instrument d'évaluation et de pilotage du risque fondé sur une nouvelle approche de type RAROC, largement opérationnelle sur l'ensemble des activités de crédit du Groupe SOCIETE GENERALE et sur ses filiales.
- Se conformer aux exigences réglementaires dans le cadre de la réforme Bâle II ; incitant fortement les banques à s'appuyer sur des outils de cotation interne pour évaluer les risques.

L'outil « STARWEB » a été développé pour permettre de noter les contreparties et mesurer la rentabilité ajustée au risque selon la démarche « RAROC ».

Les objectifs de la notation des contreparties :

La mise en place de la notation interne répond aux objectifs suivants :

Figure 44 : Objectifs de la notation des contreparties

Annexe D

Figure 45 : Organigramme de l'entreprise SGA

Annexe E

Groupe clients : l'ensemble des sociétés ayant des liens en capital. Lorsque la société –mère contrôle la majorité du capital ou des droits de vote des sociétés filiales, ou détient dans ces filiales la minorité de blocage et les responsabilités de gestion, aucun actionnaire n'y détenant une part supérieure à la sienne.

Notation d'une contrepartie : appréciation objective et subjective du risque d'insolvabilité de la contrepartie.

Risque de crédit appelé aussi risque de contrepartie : cela fait référence au montant de la perte probable que supporterait la SGA dans le cadre des transactions qui sont effectuées avec un client/contrepartie. Ce risque se traduit par une probabilité de défaut (PD). Cette dernière sert de base à la notation.

EBIT (Earnings Before Interest And Taxes) : Il s'agit du résultat duquel sont déduits les intérêts des débiteurs et les impôts. Il correspond au chiffre d'affaires net duquel sont déduites les charges d'exploitation. Il se distingue du résultat net par le fait que les charges et produits financiers ainsi que les impôts sur le bénéfice ne sont pas pris en compte.

Bâle II :

Les normes Bâle II (le second accord de Bâle) constituent un dispositif prudentiel destiné à mieux appréhender les risques bancaires et principalement le risque de crédit ou de contrepartie et les exigences, pour garantir un niveau minimum de capitaux propres, afin d'assurer la solidarité financière. Ces directives ont été préparées depuis 1988 par le Comité de Bâle, sous l'égide de la Banque des règlements internationaux et ont abouti à la publication de la Directive CRD (Capital Requirements Directive).

Annexe F

- Taux de frais de recherche et 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 ‘
- Réinvestissement en espèces
- Taux d'intérêt débiteur
- Total du passif / Valeur nette
- Valeur nette / actifs
- Ratio d'adéquation des fonds à long terme (A)
- Liquidités / total des actifs
- Actif à court terme/passif à 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
- 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
- Rentabilité des capitaux propres
- Des capitaux propres au passif

Annexe G

Code complet de l'implémentation et d'apprentissage

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sb

%matplotlib inline

dataset = pd.read_excel('SGADATAFR.xlsx')

Dataset = dataset.dropna()

del Dataset['ROA(A) % avant intérêts après impôts']

del Dataset['ROA (B) avant dépréciation avant impôt']

Dataset.head(5).drop("Faillite", axis = 1)

sb.heatmap(Dataset.isnull(), cbar=False)

corr=Dataset.corr()
corr

p_value=np.corrcoef(Dataset)
p_value

Correlation=pd.DataFrame(corr)
Corrrr = 'correlationPFE.xlsx'
Correlation.to_excel(Corrrr, index=False)

Dataset_faillite = Dataset[Dataset["Faillite"] == 1]
Dataset_non_faillite = Dataset[Dataset["Faillite"] == 0]
Dataset_frame = Dataset_non_faillite.sample(frac = 0.08)
Datasetfinal = pd.concat([Dataset_faillite,Dataset_frame])

DATASET = Dataset[Dataset["Faillite"] == 0].sample(frac = 0.009)
DATASET;

wa = sb.countplot(x = "Faillite", data = Dataset)
for p in wa.patches:
    wa.annotate(f'\n{p.get_height()}', (p.get_x()+0.2, p.get_height()), color='black', size=15, ha="center")

from sklearn.model_selection import train_test_split

X = Datasetfinal.drop(["Faillite"],axis =1)
y = Datasetfinal["Faillite"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=101)

# Logistic Regression

from sklearn.linear_model import LogisticRegression

model = LogisticRegression(C = 10, max_iter = 100,solver ='liblinear')

model_training=model.fit(X_train,y_train)

prediction = model.predict(X_test)
model.score(X_test,y_test)

model = LogisticRegression(C = 10, max_iter = 100,solver ='liblinear')
model.fit(X_train,y_train)
y_pred_logistic = model.decision_function(X_test)

from sklearn.metrics import confusion_matrix

from sklearn.metrics import confusion_matrix
logistic_cm = confusion_matrix(y_test,prediction)
logistic_cm

a = pd.DataFrame()
a["Columns"] = X.columns
a
```

```
coef = model.coef_
a["Coef"] = coef[0]
a
```

```
proba = model.predict_proba(X_test)
proba[:10]
```

```
from sklearn.model_selection import cross_val_score
```

```
Logg=LogisticRegression(C = 10, max_iter = 100,solver ='liblinear')
cross_val_score(Logg, X_train, y_train, cv=5, scoring='accuracy')
```

```
val_score = []
for k in range(100, 1000):
    score = cross_val_score(LogisticRegression(C = 60, max_iter = k,solver ='liblinear'), X_train, y_train,
cv=5).mean()
    val_score.append(score)

plt.plot(val_score)
```

```
val_score = []
for k in range(1, 100):
    score = cross_val_score(LogisticRegression(C = k, max_iter = 100,solver ='liblinear'), X_train, y_train,
cv=5).mean()
    val_score.append(score)

plt.plot(val_score)
```

```
from sklearn.model_selection import GridSearchCV
```

```
param_grid = {'C': np.arange(1, 100)}

grid = GridSearchCV(LogisticRegression(C = 11, max_iter = 100,solver ='liblinear'), param_grid, cv=5)

grid.fit(X_train, y_train)
```

```
grid.best_params_
```

```
param_grid = {'max_iter': np.arange(100, 1000)}

grid = GridSearchCV(LogisticRegression(C = 11, max_iter = 100,solver ='liblinear'), param_grid, cv=5)

grid.fit(X_train, y_train)
```

```
grid.best_params_
```

```
# Decision Tree
```

```
from sklearn.tree import DecisionTreeClassifier
```

```
dtree = DecisionTreeClassifier(max_depth=3,min_samples_leaf=26)
```

```
Tree_Training=dtree.fit(X_train,y_train)
```

```
dtree.score(X_test,y_test)
```

```
from sklearn.metrics import confusion_matrix
dtr = confusion_matrix(y_test,prediction)
dtr
```

```
fig = plt.figure(figsize = (35,25))
ax = fig.add_subplot(121)
dt = dtree.feature_importances_
impo_dff = pd.Series( dt,X_train.columns)
impo_dff.plot(color="teal", kind="barh",ax = ax)
```

```
dtree.get_n_leaves()
```

```
dtree.get_depth()
```

```
prob=dtree.predict_proba(X_test)
prob[:10]
```

```
dtree=DecisionTreeClassifier()
cross_val_score(dtree, X_train, y_train, cv=5, scoring='accuracy')

val_score = []
for k in range(1, 100):
   score = cross_val_score(DecisionTreeClassifier(max_depth=10,min_samples_leaf=k), X_train, y_train,
cv=5).mean()
   val_score.append(score)

plt.plot(val_score)

val_score = []
for k in range(1, 100):
   score = cross_val_score(DecisionTreeClassifier(max_depth=k,min_samples_leaf=17), X_train, y_train,
cv=5).mean()
   val_score.append(score)

plt.plot(val_score)

param_grid = {'max_depth': np.arange(1, 100)}


grid = GridSearchCV(DecisionTreeClassifier(), param_grid, cv=5)

grid.fit(X_train, y_train)

grid.best_params_

param_grid = {'min_samples_leaf': np.arange(1, 100)}


grid = GridSearchCV(DecisionTreeClassifier(), param_grid, cv=5)

grid.fit(X_train, y_train)

grid.best_params_

#Random Forest

from sklearn.ensemble import RandomForestClassifier

forest = RandomForestClassifier(n_estimators = 550,min_samples_leaf=3,max_depth=18)

forest1 = RandomForestClassifier(n_estimators = 150,min_samples_leaf=3,max_depth=18)

forest2= RandomForestClassifier(n_estimators = 350,min_samples_leaf=3,max_depth=18)

forest3 = RandomForestClassifier(n_estimators = 450,min_samples_leaf=3,max_depth=18)

forest4 = RandomForestClassifier(n_estimators = 650,min_samples_leaf=3,max_depth=18)

forest_training=forest.fit(X_train,y_train)

forest.score(X_test,y_test)

fig = plt.figure(figsize = (35,25))
importances = forest.feature_importances_
ax = fig.add_subplot(121)
impo_df = pd.Series(importances,X_train.columns)
impo_df.plot(color="teal", kind="barh")

pp = forest.predict(X_test)
forest_cm = confusion_matrix(y_test,pp)
forest_cm

foresta=forest.predict_proba(X_test)
foresta[:10]

rforest=RandomForestClassifier(n_estimators = 3500)
cross_val_score(rforest, X_train, y_train, cv=5, scoring='accuracy')

val_score = []
for k in range(100, 1000,50):
```

```
    score = cross_val_score(RandomForestClassifier(n_estimators = k), X_train, y_train, cv=5).mean()
    val_score.append(score)

plt.plot(val_score)

val_score = []
for k in range(1, 100):
    score = cross_val_score(RandomForestClassifier(n_estimators = 550,min_samples_leaf=k), X_train, y_train,
cv=5).mean()
    val_score.append(score)

plt.plot(val_score)

param_grid = {'min_samples_leaf': np.arange(1, 100)}


grid = GridSearchCV(RandomForestClassifier(n_estimators = 100,min_samples_leaf=26,max_depth=3),
param_grid, cv=5)

grid.fit(X_train, y_train)

grid.best_params_

val_score = []
for k in range(1, 100):
    score = cross_val_score(RandomForestClassifier(n_estimators = 550,min_samples_leaf=26,max_depth=k),
X_train, y_train, cv=5).mean()
    val_score.append(score)

plt.plot(val_score)

param_grid = {'max_depth': np.arange(1, 100)}


grid = GridSearchCV(RandomForestClassifier(n_estimators = 100,min_samples_leaf=26,max_depth=3),
param_grid, cv=5)

grid.fit(X_train, y_train)

grid.best_params_

param_grid = {'n_estimators': np.arange(100, 1000 , 50)}


grid = GridSearchCV(RandomForestClassifier(n_estimators = 100), param_grid, cv=5)

grid.fit(X_train, y_train)

grid.best_params_

#SVM

from sklearn.svm import SVC

svm = SVC(C=4,gamma=1,probability = True)

SVC(probability = True)

svm.fit(X_train,y_train)

svm.score(X_test,y_test)

svm = SVC(C=5,gamma=1)
svm.fit(X_train,y_train)
y_pred_svm = svm.decision_function(X_test)

from sklearn.feature_selection import mutual_info_classif

imp = mutual_info_classif(X_train,y_train)

fig = plt.figure(figsize = (35,25))
ax = fig.add_subplot(121)
```

```
pl = pd.Series(imp,X_train.columns)
pl.plot(kind = "barh", color = "teal")
```

```
from sklearn.calibration import CalibratedClassifierCV
model = CalibratedClassifierCV(svm)
model.fit(X_train,y_train)
```

```
import pickle
filename = 'linearSVC.sav'
pickle.dump(model, open(filename, 'wb'))
```

```
model = pickle.load(open(filename, 'rb'))
```

```
pred_class = model.predict(X_test)
probability = model.predict_proba(X_test)
```

```
probability[:10]
```

```
eee = svm.predict(X_test)
svm_cm = confusion_matrix(y_test,eee)
svm_cm
```

```
svmm=SVC(C=3,gamma=3)
cross_val_score(svmm, X_train, y_train, cv=5, scoring='accuracy')
```

```
val_score = []
for k in range(1, 100):
    score = cross_val_score(SVC(C=k,gamma=0.1), X_train, y_train, cv=5).mean()
    val_score.append(score)

plt.plot(val_score)
```

```
val_score = []
for k in range(1, 10):
    score = cross_val_score(SVC(C=17,gamma=k), X_train, y_train, cv=5).mean()
    val_score.append(score)

plt.plot(val_score)
```

```
param_grid = {'C': np.arange(1, 100)}


grid = GridSearchCV(SVC(C=17,gamma=1), param_grid, cv=5)
```

```
grid.fit(X_train, y_train)
```

```
grid.best_params_
```

```
param_grid = {'gamma': np.arange(1, 10)}


grid = GridSearchCV(SVC(C=17,gamma=1), param_grid, cv=5)
```

```
grid.fit(X_train, y_train)
```

```
grid.best_params_
```

```
#KNN
```

```
from sklearn.neighbors import KNeighborsClassifier
```

```
KNN = KNeighborsClassifier(n_neighbors=16)
```

```
KNN1 = KNeighborsClassifier(n_neighbors=10)
```

```
KNN2 = KNeighborsClassifier(n_neighbors=1)
```

```
KNN3 = KNeighborsClassifier(n_neighbors=5)
```

```
KNN4 = KNeighborsClassifier(n_neighbors=28)
```

```
KNN.fit(X_train,y_train)
```

```
KNN.score(X_test,y_test)
```

```
knn= KNN.predict_proba(X_test)
knn[:10]
```

```
cc = KNN.predict(X_test)
KNN_cm = confusion_matrix(y_test,cc)
KNN_cm
```

```
KNN.predict(X_test)
```

```
from sklearn.inspection import permutation_importance
results = permutation_importance(KNN, X, y,scoring='accuracy')
```

```
a = pd.DataFrame()
a["Columns"] = X.columns
a
importance = results.importances_mean
coef = importance
a["Coef"] = coef[[x for x in range(len(importance))]]

a.plot(kind ="bar",figsize = (7,7))
```

```
knnn=KNeighborsClassifier(n_neighbors=9)
cross_val_score(knnn, X_train, y_train, cv=5, scoring='accuracy')
```

```
val_score = []
for k in range(1, 100):
   score = cross_val_score(KNeighborsClassifier(n_neighbors=k), X_train, y_train, cv=5).mean()
   val_score.append(score)

plt.plot(val_score)
```

```
from sklearn.model_selection import GridSearchCV
```

```
param_grid = {'n_neighbors': np.arange(1, 20),
          'metric': ['euclidean', 'manhattan']}

grid = GridSearchCV(KNeighborsClassifier(), param_grid, cv=5)

grid.fit(X_train, y_train)
```

```
grid.best_params_
```

```
from matplotlib.pyplot import figure
from sklearn.metrics import plot_roc_curve
forest.fit(X_train, y_train)
svm.fit(X_train,y_train)
dtree.fit(X_train,y_train)
model.fit(X_train,y_train)
KNN.fit(X_train,y_train)
plt.figure(figsize=(10, 10))
ax = plt.gca()
forest_disp = plot_roc_curve(forest, X_test, y_test, ax=ax, alpha=0.9)
svm_disp = plot_roc_curve(svm,X_test, y_test, ax=ax, alpha=0.9)
dtree_disp = plot_roc_curve(dtree,X_test, y_test, ax=ax, alpha=0.9)
logistic_disp = plot_roc_curve(model,X_test, y_test, ax=ax, alpha=0.9)
KNN_disp = plot_roc_curve(KNN,X_test, y_test, ax=ax, alpha=0.9)
plt.show()
```

```
#Logisticc regression
Log_Accuracy = 125
Log_Score = 0.9291
Log_AUC = 0.93
Log_sensitivity= 125/(125+6)
Log_specificity= 47/(9+47)
print("Accuracy : " , Log_Accuracy)
print("Score : " , Log_Score)
print("AUC : " , Log_AUC)
print("Sensibilité : " , Log_sensitivity)
```

```
print("Specificité : " , Log_specificity)

#Random Forest
forest_Accuracy = forest_cm[0,0]
forest_Score = forest.score(X_test,y_test)
forest_AUC = 0.94
forest_sensitivity= forest_cm[0,0]/(forest_cm[0,0]+forest_cm[0,1])
forest_specificity= forest_cm[1,1]/(forest_cm[1,0]+forest_cm[1,1])
print("Accuracy : " , forest_Accuracy)
print("Score : " , forest_Score)
print("AUC : " , forest_AUC)
print("Sensibilité : " , forest_sensitivity)
print("Specificité : " , forest_specificity)

#KNN
KNN_Accuracy = KNN_cm[0,0]
KNN_Score = KNN.score(X_test,y_test)
KNN_AUC = 0.92
KNN_sensitivity= KNN_cm[0,0]/(KNN_cm[0,0]+KNN_cm[0,1])
KNN_specificity= KNN_cm[1,1]/(KNN_cm[1,0]+KNN_cm[1,1])
print("Accuracy : " , KNN_Accuracy)
print("Score : " , KNN_Score)
print("AUC : " , KNN_AUC)
print("Sensibilité : " , KNN_sensitivity)
print("Specificité : " , KNN_specificity)

#SVM
svm_Accuracy = svm_cm[0,0]
svm_Score = svm.score(X_test,y_test)
svm_AUC = 0.95
svm_sensitivity= svm_cm[0,0]/(svm_cm[0,0]+svm_cm[0,1])
svm_specificity= svm_cm[1,1]/(svm_cm[1,0]+svm_cm[1,1])
print("Accuracy : " , svm_Accuracy)
print("Score : " , svm_Score)
print("AUC : " , svm_AUC)
print("Sensibilité : " , svm_sensitivity)
print("Specificité : " , svm_specificity)

from matplotlib.pyplot import figure
from sklearn.metrics import plot_roc_curve
forest.fit(X_train, y_train)
forest1.fit(X_train, y_train)
forest2.fit(X_train, y_train)
forest3.fit(X_train, y_train)
forest4.fit(X_train, y_train)
plt.figure(figsize=(10, 10))
ax = plt.gca()
forest_disp = plot_roc_curve(forest, X_test, y_test, ax=ax, alpha=0.8)
forest_disp1 = plot_roc_curve(forest1, X_test, y_test, ax=ax, alpha=0.8)
forest_disp2 = plot_roc_curve(forest2, X_test, y_test, ax=ax, alpha=0.8)
forest_disp3 = plot_roc_curve(forest3, X_test, y_test, ax=ax, alpha=0.8)
forest_disp4 = plot_roc_curve(forest4, X_test, y_test, ax=ax, alpha=0.8)
plt.show()

from matplotlib.pyplot import figure
from sklearn.metrics import plot_roc_curve
KNN.fit(X_train,y_train)
KNN1.fit(X_train,y_train)
KNN2.fit(X_train,y_train)
KNN3.fit(X_train,y_train)
KNN4.fit(X_train,y_train)
plt.figure(figsize=(10, 10))
ax = plt.gca()
KNN_disp = plot_roc_curve(KNN,X_test, y_test, ax=ax, alpha=0.8)
```

```
KNN_disp1 = plot_roc_curve(KNN1,X_test, y_test, ax=ax, alpha=0.8)
KNN_disp2 = plot_roc_curve(KNN2,X_test, y_test, ax=ax, alpha=0.8)
KNN_disp3 = plot_roc_curve(KNN3,X_test, y_test, ax=ax, alpha=0.8)
KNN_disp4 = plot_roc_curve(KNN4,X_test, y_test, ax=ax, alpha=0.8)
plt.show()
```

Annexe H

	ROA(C) avant dépréciation avant impôt	Marge brute d'exploitation	Marge brute (Gross Margin)	Constant interest rate (after tax)	Taux de dépenses d'exploitation	Taux de frais de recherche et développement	Ratio de flux de trésorerie	Taux de la dette portant intérêt	Tax rate (A)	Taux Croissance marge brute (Année 2)	...	Des capitaux propres au passif à long terme	Flux de trésorerie par rapport à l'actif total	Flux de trésorerie par rapport aux passif
0	0.464291	0.610235	0.610235	0.781506	0.000290	0.000000	0.461867	0.000647	0.0	0.022080	...	0.120916	0.641100	0.459001
1	0.426071	0.601450	0.601364	0.780284	0.000236	0.000003	0.458521	0.000790	0.0	0.022760	...	0.117922	0.642765	0.459254
2	0.399844	0.583541	0.583541	0.781241	0.000108	0.000000	0.465705	0.000449	0.0	0.022046	...	0.120760	0.579039	0.448518
3	0.465022	0.598783	0.598783	0.781550	0.000079	0.000000	0.462746	0.000686	0.0	0.022096	...	0.110933	0.622374	0.454411
4	0.388680	0.590171	0.590251	0.781069	0.000157	0.000000	0.465861	0.000716	0.0	0.021565	...	0.112917	0.637470	0.458499

Figure 46 : Visualisation des 5 premiers clients dans le DataSet

Annexe I

Figure 47 : Matrice des corrélations '' Test de Pearson ''

Annexe J

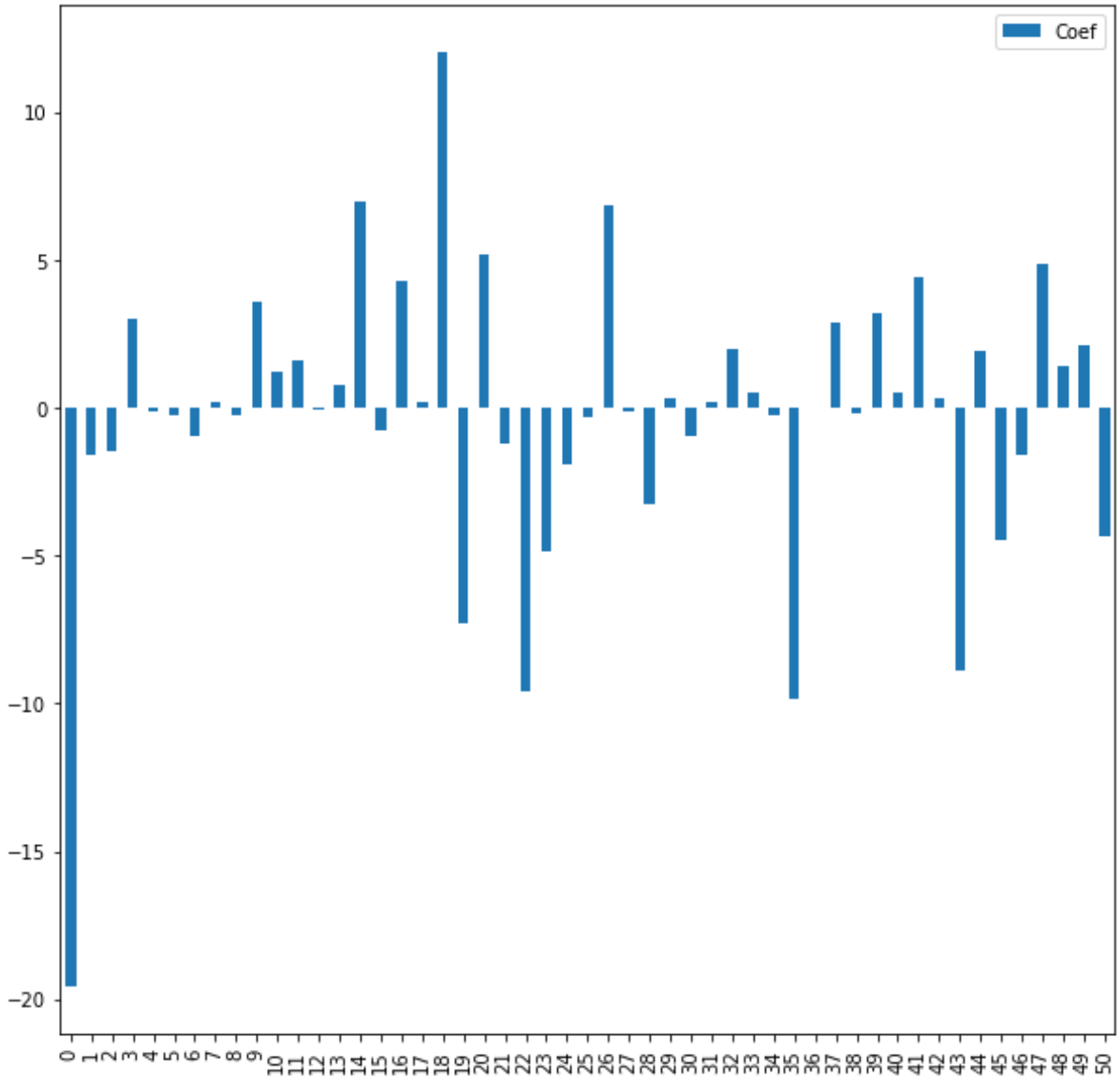


Figure 48 : Diagramme des coefficients d'importance de la régression logistique

Annexe K

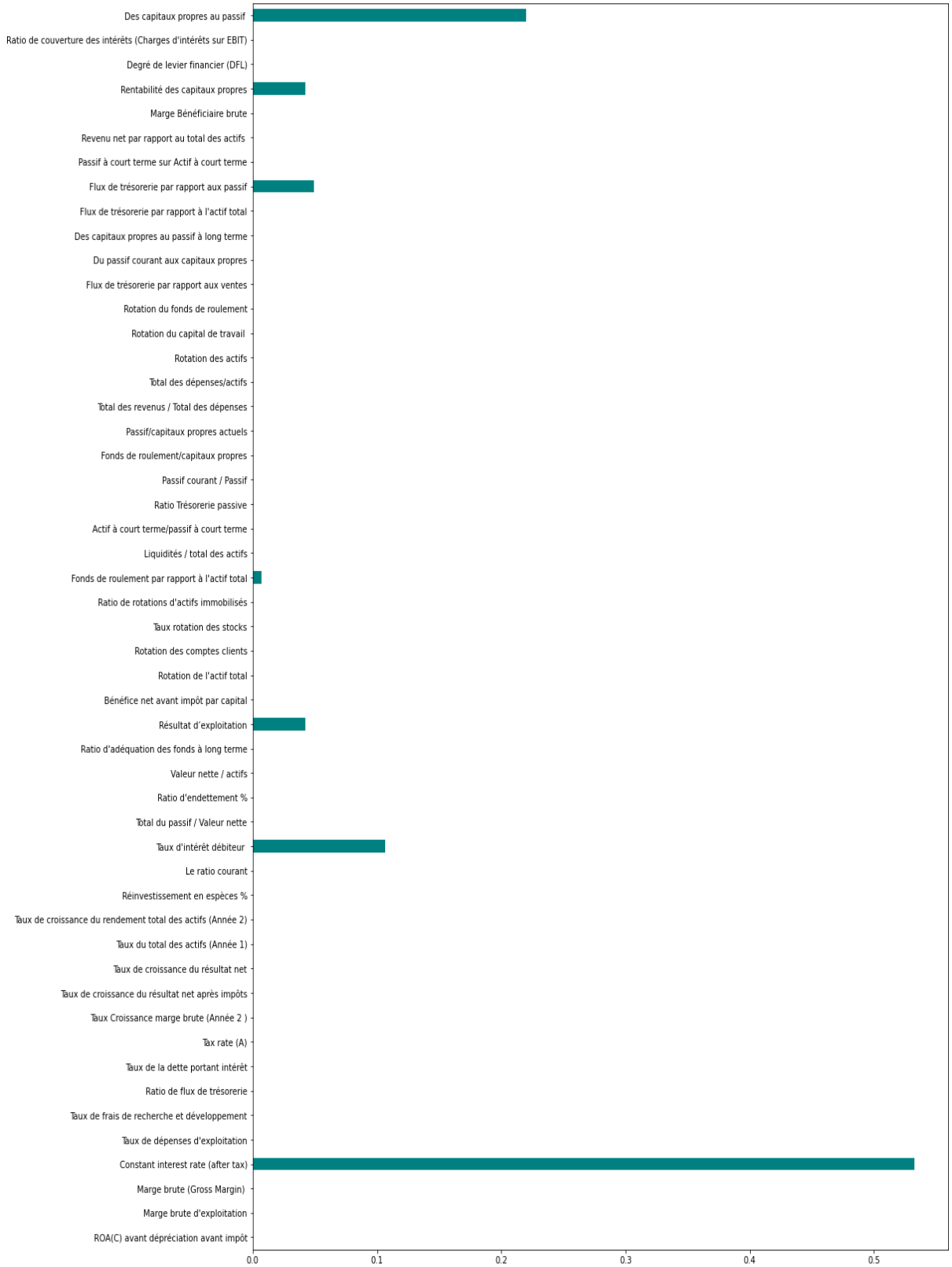


Figure 49 : Diagramme d'importances des variables du modèle Decision Tree

Annexe L

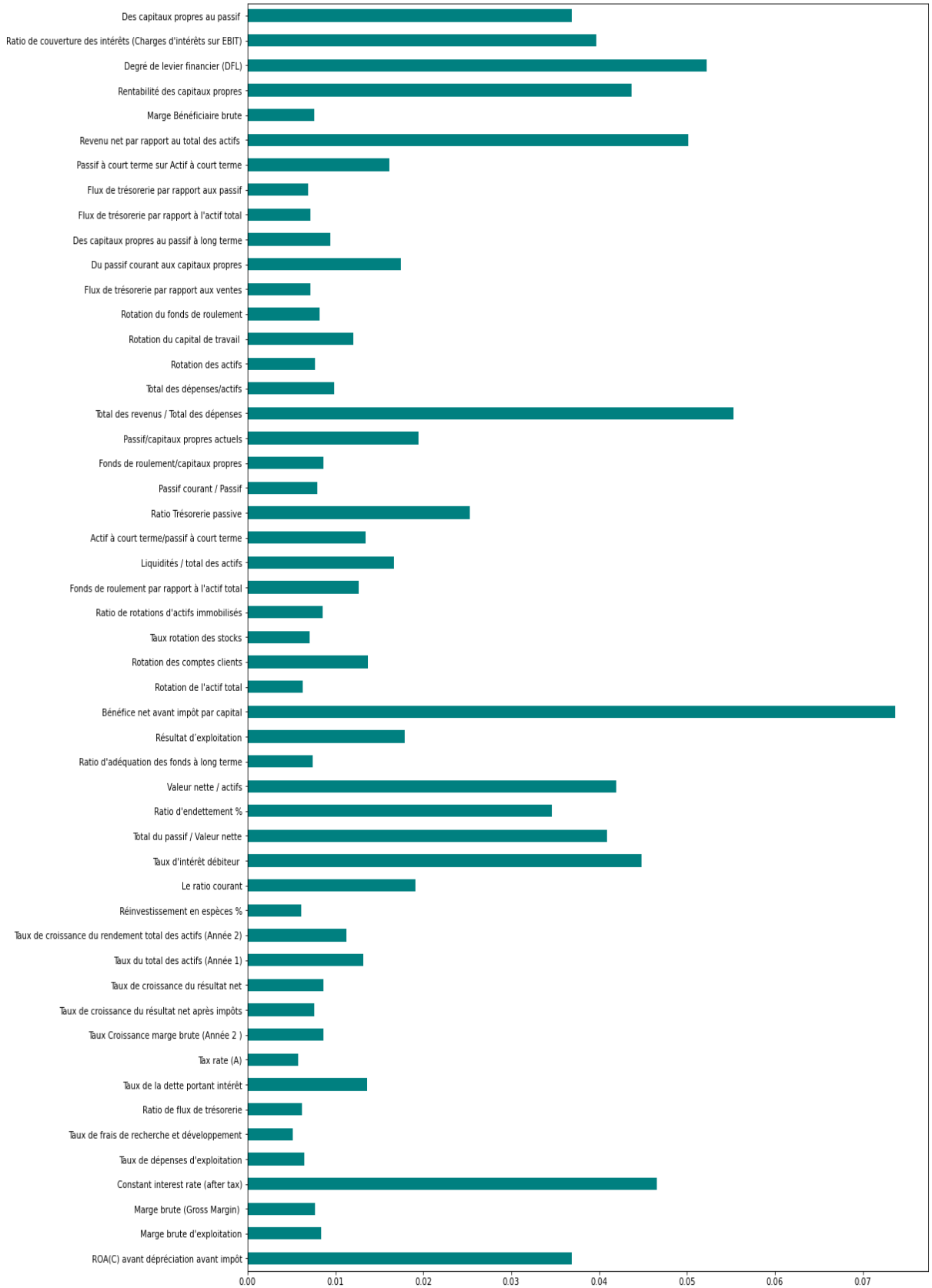


Figure 50 : Diagramme d'importance des variables du modèle Random Forest

Annexe M

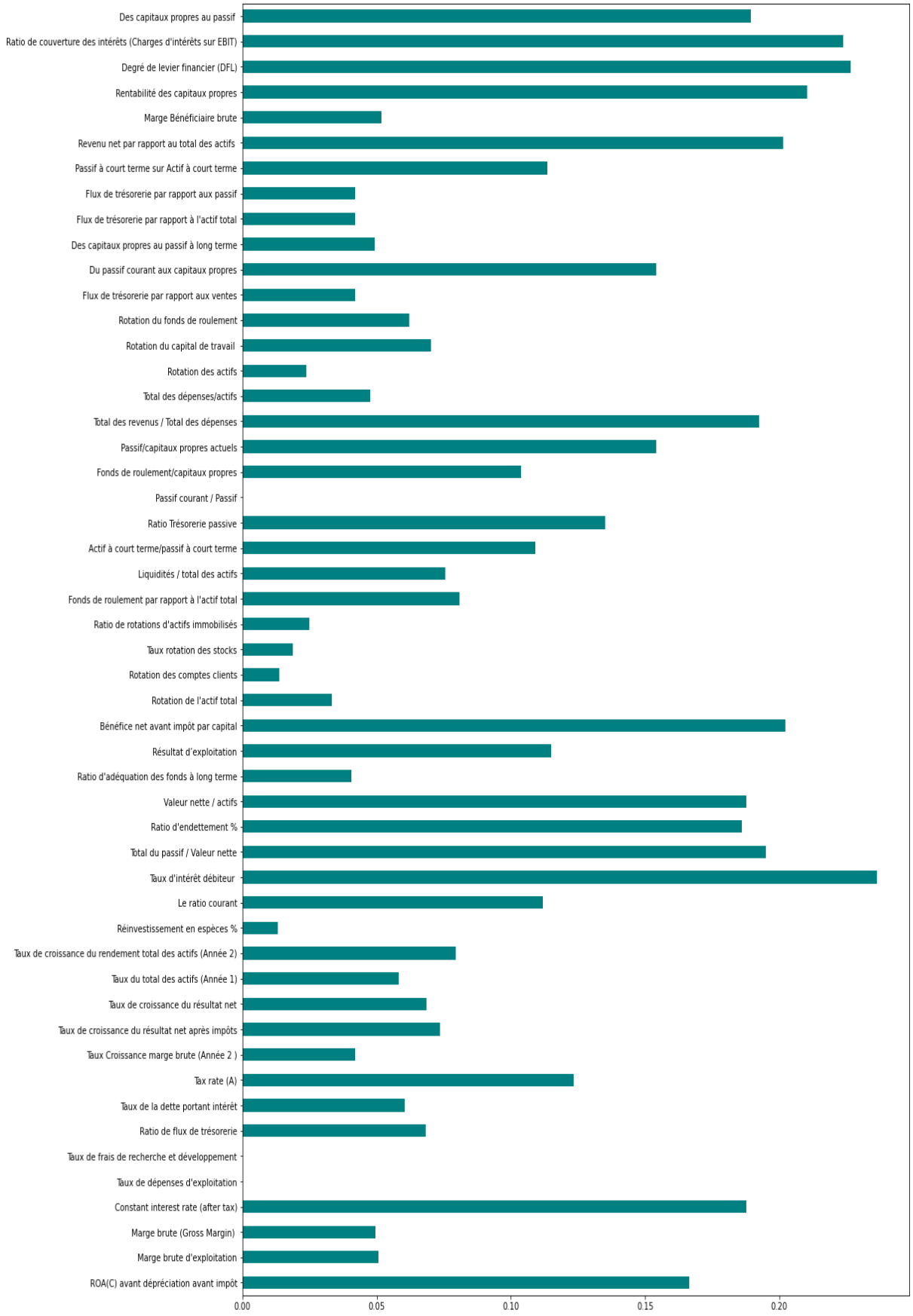


Figure 51 : Diagramme d'importance des variables du modèle SVM

Annexe N

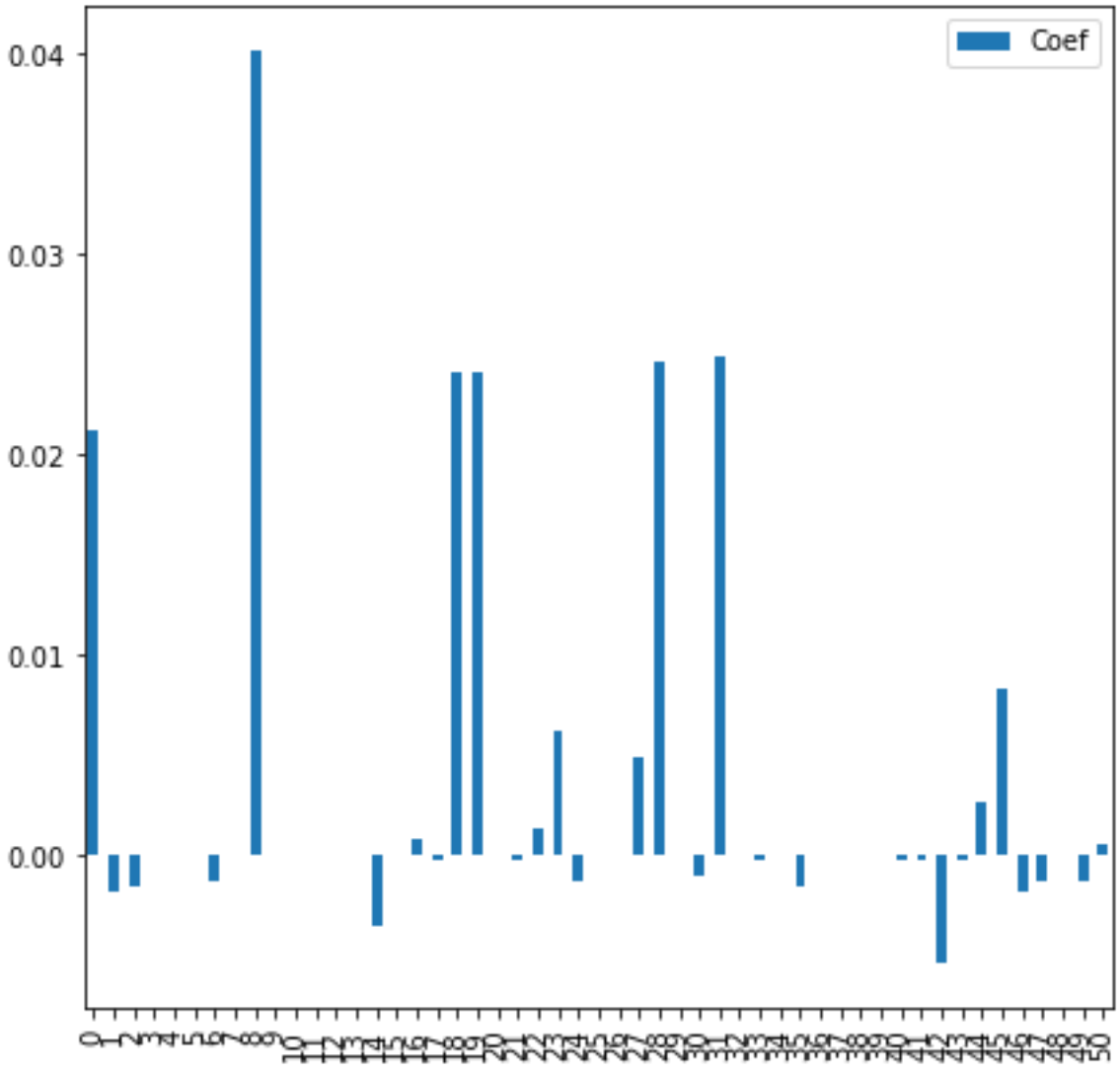


Figure 52 : Diagramme d'importance des variables du modèle KNN

Annexe O

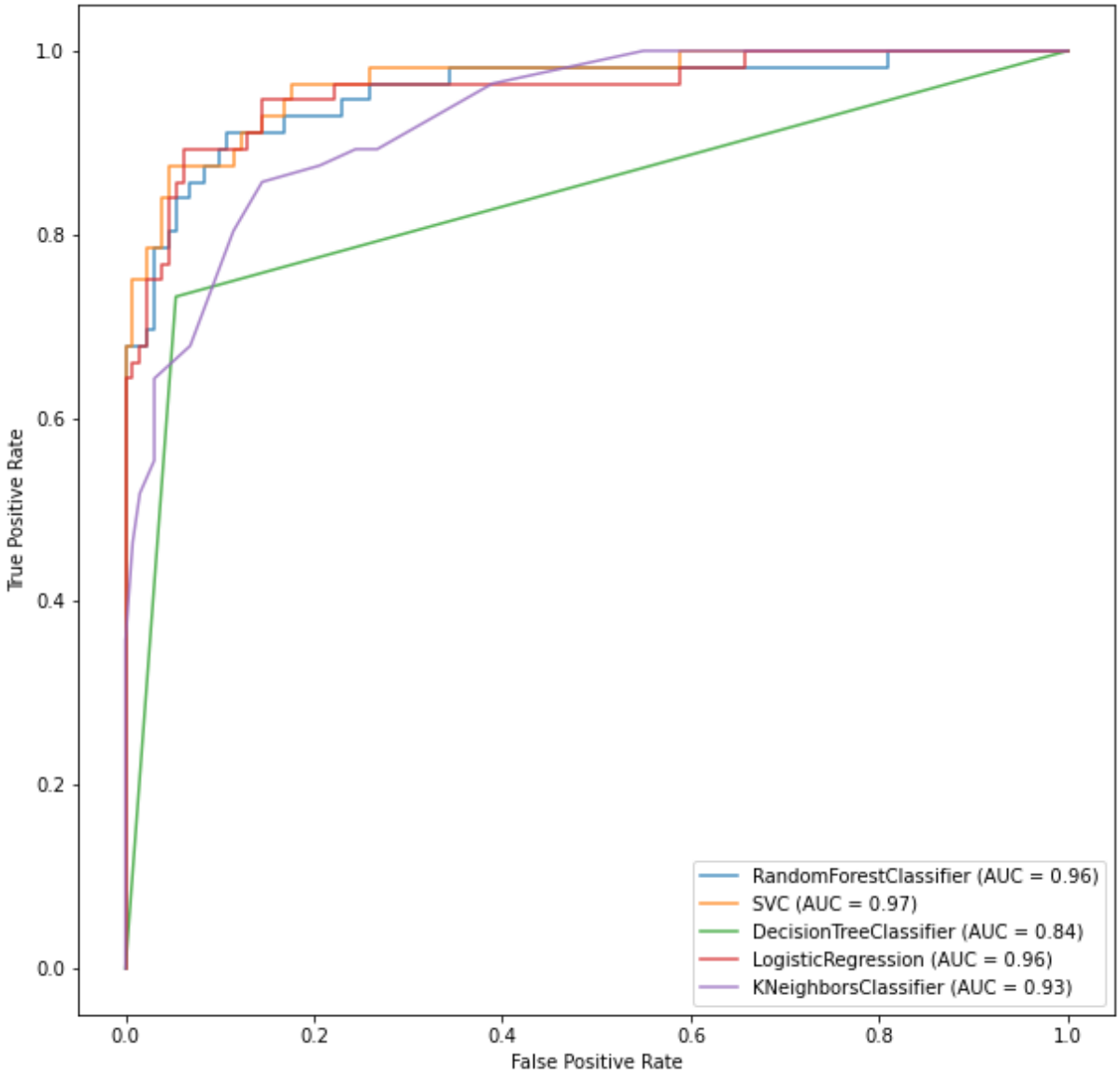


Figure *41* : ROC courbes des 5 modèles d'apprentissage utilisés

Annexe P

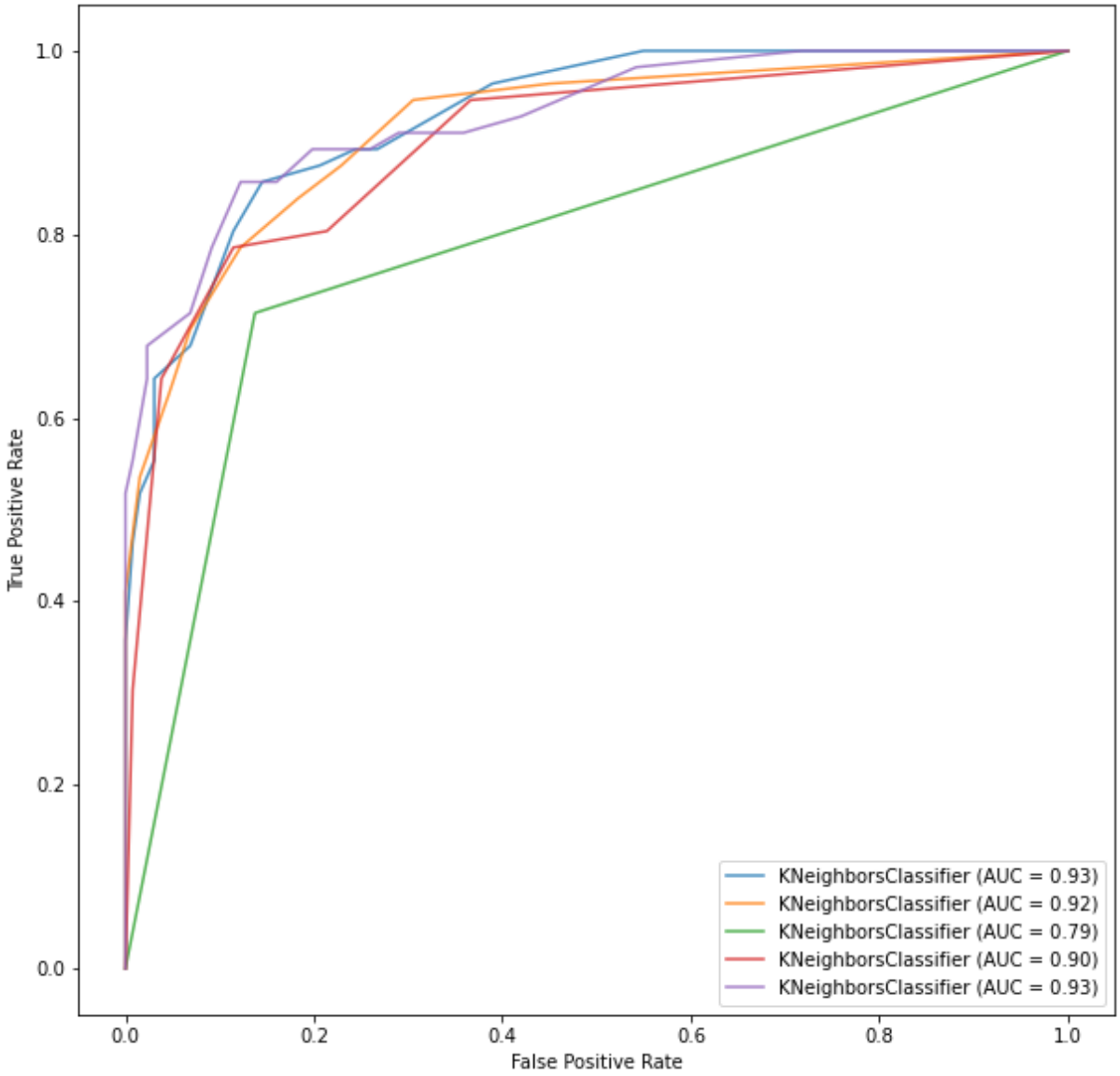


Figure 53 : Comparaison ROC (AUC) du modèle KNN en variant les hyperparametres

Annexe Q

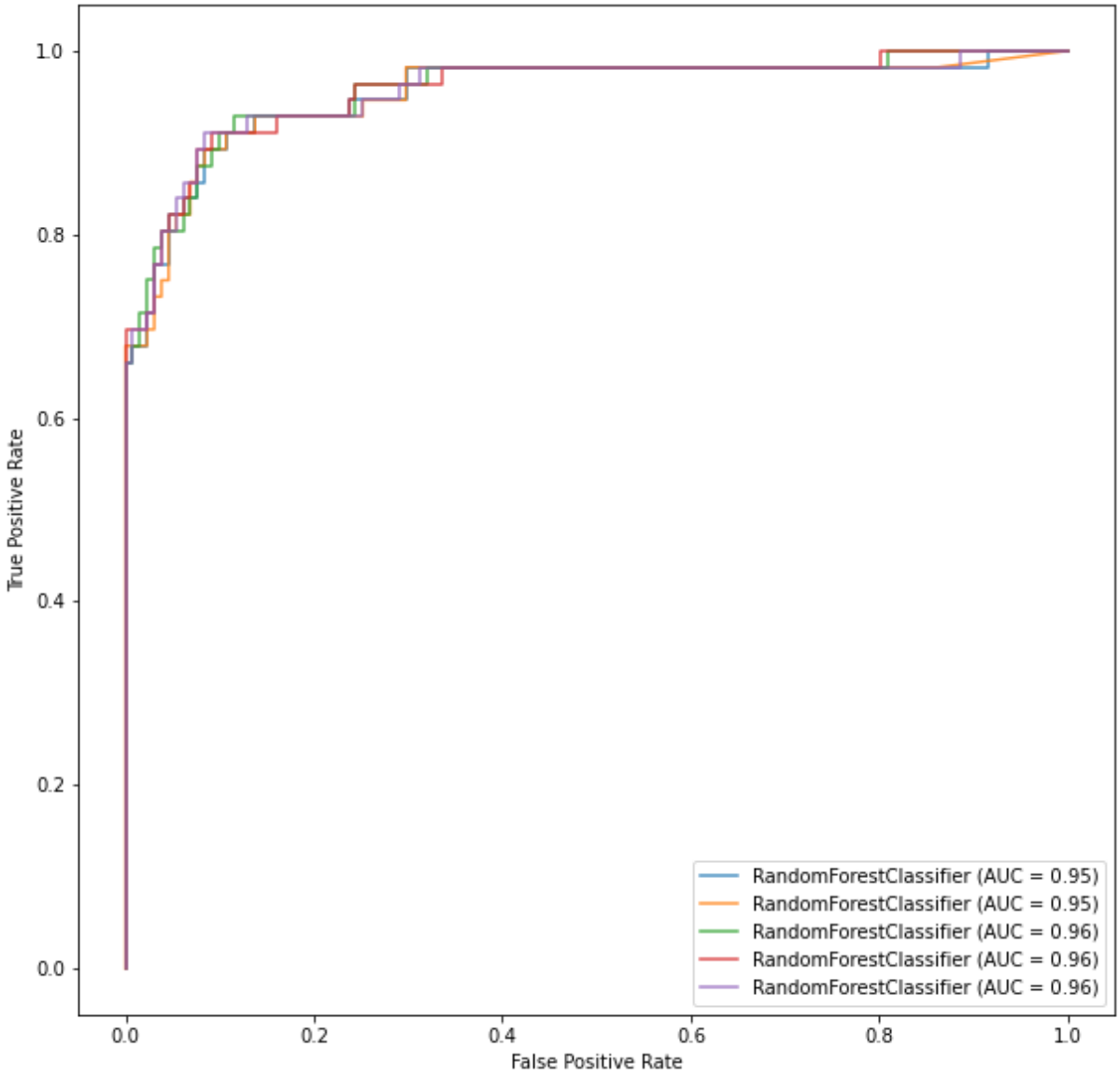


Figure 54 : Comparaison AUC du Random Forest en variant les hyperparamètres