ÉCOLE SUPERIEURE DE COMMERCE

Pôle Universitaire de KOLEA

Mémoire de fin de cycle en vue de l'obtention du diplôme de Master en Sciences financières et comptabilité

Spécialité : Finance d'entreprise

Thème :

La gestion du risque des crédits d'exploitation des PME par le Machine Learning

Étude de cas : La Banque CPA

Elaboré par :

Lyne Imene SOUADDA

Manel MENARI

Encadré par :

Dr Yasser Moussa BERGHOUT

Dr Billel BENILLES

Année Universitaire : 2021 / 2022

Remerciement

Il nous importe de remercier les personnes suivantes :

Je remercie en premier lieu mes inspirants enseignants, pour l'expertise, les conseils et la générosité qu'ils ont investis dans ma formation. En particulier, je remercie Mr Benilless et Mr Berghout. Grace à leur encadrement, j'ai appris à apprécier le sens de la recherche scientifique, leur enthousiasme et engagement a rendu ce travail plus attrayant. Je remercie également Mr Touati Mr Chouike et Mr Melzi pour leur bienveillance, leurs encouragements et leur aide.

À ma famille, à mes parents, pour leur éduction, leur amour et leur présence réconfortante. Merci à ma sœur, Souha, pour sa compagnie réjouissante.

Un merci tout spécial à toi Maman. Ton amour et ton soutien comptent pour moi bien plus que tout prix académique.

A tous mes amis, votre amitié, votre tolérance et vos beaux esprits me sont si chers que je ne peux m'en passer.

A vous tous, merci. Chacune des lignes que j'ai composé dans ce travail de recherche est en partie grâce à vous.

Cette expérience m'a apporté bien plus que je l'espérais. Or, ce n'est que le début d'un parcours, certes incertain, mais surement passionnant, qu'on ne peut pas s'arrêter en si bon chemin…

Lyne

Je remercie ma chère mère qui m'a appris à ne jamais laisser tomber, pour son soutien, ses sacrifices, ses encouragements tout au long de mon cursus et surtout pour son amour sans limite, qui m'ont amené à ce que je suis aujourd'hui.

A ma sœur Sana et mon frère Mehdi qui ont fait que cette année soit remplie de joie et de rire.

Je remercie mes amis qui m'ont épaulé tout au long de mes études avec leur présence et leurs encouragements.

Je remercie toutes les personnes qui ont contribué de près ou de loin pour l'élaboration de ce modeste travail.

Que ce travail soit un enrichissement et source de savoir.

Manel

SOMMAIRE

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

ACT :	Actif à Court Terme
ACP :	Analyse en Composantes Principales
ANN :	Artificial Neural Network
ASC :	Aire Sous la Courbe
BFR :	Besoin en Fonds de Roulement
CA :	Chiffre d'Affaires
CAF :	Capacité d'Autofinancement
CPA :	Crédit Populaire d'Algérie
DCT :	Dette à Court Terme
DLT :	Dette à Long Terme
DT :	Decision Trees
EBE :	Excédent Brut d'Exploitation
EURL :	Entreprise Unipersonnelle à Responsabilité Limité
FR :	Fonds de Roulement
IA :	Intelligence Artificielle
KNN :	K Nearest Neighbors
LR :	Log Vraisemblance
ML :	Machine Learning
NA :	Not Available
OCDE :	Organisme de Coopération et de Développemnt Economique
PME :	Petite et Moyenne Entreprise
RF :	Random Forest
RL :	Régression Logistique
ROA :	Return On Asset
ROC :	Receiver Operating Characteristic
ROE :	Return On Equity
SARL :	Société A Responsabilité Limitée
SNC :	Société a Nom Collectif
SPA :	Société Par Action
SVM :	Support Vector Machine
VA :	Valeur Ajoutée

Résumé

Le développement technologique, la disponibilité des données et la puissance de calcul amènent la plupart des banques à moderniser leurs modèles de notation de crédit. L'octroi de crédit constitue leurs principales activités, ce qui fait de la notation des créanciers une compétence clé pour la continuité des banques commerciales. Cette opération s'avère plus risquée pour les entreprises de moyenne et petite taille. En effet, de par l'instabilité de leur activité et le manque de garanties, elles deviennent une menace pour ces emprunteurs. Étant donné que même une petite amélioration de la précision, entraîne une réduction significative des pertes, l'utilisation du meilleur modèle de classification s'avère d'une grande importance. En intelligence artificielle, le scoring de crédit fut historiquement l'un des premiers champs d'application des techniques de Machine Learning. En effet, il a montré son efficacité en atteignant les résultats souhaités. L'objectif de ce travail est d'évaluer quelques modèles d'apprentissage automatique supervisé dans la classification des emprunteurs en défaillant et non défaillant. Par le biais de trois modèles : RL, KNN et ANN, 282 dossiers de crédits d'exploitation sont entrainés à l'aide de 31 variables économiques afin d'aboutir à cette classification. Notre échantillon est composé exclusivement de PME algériennes. Des méthodes de prétraitement telle que l'analyse en composantes principales (ACP) sont utilisées pour trouver le nombre optimal de caractéristiques nécessaires pour une prédiction précise. Dans la validation et la comparaison des modèles, des mesures comme la matrice de confusion et la courbe ROC sont appliquées. Les résultats de l'étude montrent que la technique "neuronale" est meilleure en termes de prévision.

Mots clés : Scoring, Risque de crédit, Machine Learning, PME, Crédit d'exploitation.

Abstract

Technological development, data availability and computing power are leading most banks to modernize their credit rating models. Credit granting constitues their core business, which makes credit scoring a key skill for the continuity of commercial banks. This operation proves to be more risky for small and medium-sized companies. In fact, the instability of their business and lack of guarantees make them a threat to these borrowers. Since even a small improvement in accuracy leads to a significant reduction in losses, using the best classification model is of great importance. In artificial intelligence, credit scoring was historically one of the first fields of application of Machine Learning techniques. Indeed, it has shown its efficiency in achieving the desired results. The objective of this work is to evaluate some supervised machine learning models in the classification of borrowers into defaulters and non-defaulters. By means of three models: LR, KNN and ANN, 282 operating credit files are trained using 31 economic variables to achieve this classification. Our sample is composed exclusively of algerian SMEs. Preprocessing methods such as principal component analysis (PCA) are used to find the optimal number of features needed for an accurate prediction. Regarding the validation and comparison of the models, measures such as the confusion matrix and the ROC curve are applied. The results of the study show that the "neural" technique is better in terms of prediction**.**

Keywords : Scoring, Credit risk, Machine Learning, SME, Operating credit.

Introduction générale

Les banques par la nature de leur activité s'occupent principalement de l'octroi de crédit, elles sont ainsi confrontées au risque de non remboursement. Ce risque, notamment lorsque le crédit est important, peut causer la faillite totale d'une banque. D'une façon générale, le risque de crédit se définit comme étant le risque qu'un emprunteur manque à ses engagements. Ainsi, sa maîtrise et son maintien constituent la principale préoccupation des organismes bancaires.

Habituellement, l'approche générique de l'évaluation du risque de crédit consiste à analyser les dossiers des anciens clients de la banque, afin de trouver les caractéristiques d'un profil défaillant soit par des ratios financiers et comptables ou par des modèles linéaires, puis d'attribuer différentes notes de crédit aux entreprises ayant différentes probabilités de défaut, c'est la notation de crédit.

Aujourd'hui, l'intelligence artificielle est une composante existentielle de la finance. Ses progrès dans ce domaine sont remarquables. Elle a rendu la finance moins chère, plus rapide, plus grande et plus rentable, en particulier une de ses branches ; le Machine Learning. Cette discipline contribue à plusieurs applications : tarification des assurances, trading, détection des fraudes et gestion de portefeuille. Elle est notamment efficace dans la gestion de crédits, elle élabore des modèles fiables capables de détecter et de prédire avec précision les profils défaillants. Des travaux empiriques récents s'activent à vérifier le pouvoir prédictif des méthodes d'intelligence artificielle dans les problèmes de gestion et notation de ce risque. En 2020, Wang et al., ont publié une étude sur le processus de décision d'octroi de crédit aux petites entreprises en utilisant des modèles classiques et des modèles supervisés d'apprentissage automatique, l'étude consiste à comparer les modèles : score-carte, RF, ANN ainsi que d'autres méthodes. Selon l'article, le modèle de forêts aléatoires boostés à présenter le plus de précision.

En Guinée, Ampountolas et al. (2021), ont procédé à une étude comparative entre différentes méthodes d'apprentissage supervisé sur un échantillon de 4450 dossiers de crédit d'exploitation sur une période de 4ans. Plusieurs mesures de performances sont utilisées. Dans l'ensemble, les résultats ont montré que l'algorithme des arbres de décision était le plus efficace. Ainsi, ces méthodes peuvent constituer des solutions de rechange prometteuses aux méthodes traditionnelles.

En Algérie, nous assistons à une croissance notable du nombre de PME, 4,8% entre 2020 et 2021.[1] Cette croissance a été possible grâce au financement du secteur bancaire. Néanmoins, seulement la moitié des crédits sollicités a été accordée[2].Cette méfiance des institutions financières est due en partie à l'absence de marché financier, d'où un manque d'informations et un manque de garanties. Il en résulte, la nécessité de développer un outil de gestion de ce risque.

1. La problématique de recherche

L'objectif de la présente étude, est l'adaptation de différents modèles d'apprentissage automatique supervisé dans la classification des demandeurs de crédit. Ainsi, notre mémoire s'inscrit dans le contexte de développement de méthodes modernes et puissantes pour la gestion du risque auquel sont confrontées les banques, plus précisément, le risque de non remboursement des PME. Dans cette perspective, la problématique de recherche est formulée de la manière suivante :

« Dans quelle mesure les algorithmes du Machine Learning sont capables de modéliser le risque de crédit d'exploitation des PME ? »

Il en résulte les questions ci-dessous :

- A quel point le risque de crédit d'exploitation impacte -t- il l'activité bancaire ?
- Parmi les méthodes de ML, laquelle apporte le plus de précision ?
- Quelles sont les méthodes de prétraitement à appliquer sur l'ensemble de données pour arriver à une meilleure précision ?

2. Les hypothèses

Afin de répondre à ces questions, nous avons émis les hypothèses suivantes :

- H_1 : le crédit d'exploitation présente un risque de défaut qui peut menacer l'activité de la banque.
- H_2 : a priori, les réseaux de neurones artificiels sont plus susceptibles de donner une performance supérieure dans une problématique de classification binaires.

[1] Bulletin PME N° 39 : novembre 2021, Bulletins d'information statistique de la PME, https://www.industrie.gov.dz/ 24/04/2022 à 14 : 00.
[2] Bulletin PME N° 33 : novembre 2018, Bulletins d'information statistique de la PME, https://www.industrie.gov.dz/ 24/04/2022 à 14 : 15.

- H_3 : la codification des variables qualitatives, le traitement des valeurs manquantes ainsi que la réduction des dimensionnalités peuvent fortement améliorer la qualité de l'ensemble de données et ainsi augmenter la précision du modèle.

3. Choix du thème

Différentes raisons nous ont incité à mener cette recherche :

- Les modèles scoring sont standardisés, les même ratios et coefficients sont généralisés sur différents environnements. Nous pensons qu'il est nécessaire de les adapter à chaque type de marché.
- En général, ces modèles sont principalement basés sur l'information financière. Pour notre part nous avons introduit d'autres variables économique (secteur d'activité, forme juridique, …).
- La littérature algérienne n'a pas accordé beaucoup d'attention à la gestion du risque de crédit par l'outil de l'apprentissage automatique.

4. Méthodologie de la recherche

Nous avons observé dans l'ensemble des études antérieures que les chercheurs s'intéressent à la comparaison de différents modèles automatiques supervisés dans le domaine de scoring des crédits. Toutefois, certains les ont complétés par l'optimisation de ces modèles par des méthodes de réduction de dimensionnalité et différentes techniques de prétraitement de données. Cette dernière démarche est celle que nous souhaitons adopter afin de vérifier nos hypothèses.

D'abord, une base de données constituée de 282 dossiers de crédit d'exploitation est recueillie auprès de la direction des petites et moyennes entreprises du CPA. Par la suite, ces données sont traitées et l'analyse en composantes principales (ACP) est introduite pour trouver le nombre optimal de caractéristiques pour une prédiction précise, cela permet d'utiliser plus efficacement les ressources limitées disponibles.13 modèles sont ainsi configurés, testés et évalués. Une fois arrivé à la phase de modélisation, trois algorithmes d'apprentissage supervisé sont mis en place sur un ensemble de 19 variables explicatives, il s'agit des algorithmes de régression logistique, les k voisins les plus proches et les réseaux de neurones artificiels. Avant leurs implémentations, ces modèles de notation de crédit doivent être évalués pour leurs cohérence et précision. Ce processus est appelé la validation,

il est généralement entrepris par une partie indépendante du processus de développement et est réalisé à partir de données qui n'ont pas fait partie de la phase d'apprentissage du modèle. Dans cette validation, nous avons opté pour les mesures ROC ASC et la matrice de confusion.

5. Objectif de la recherche

L'objectif principal de l'étude consiste à tester la performance des modèles intelligents supervisés dans la notation des crédits d'exploitation, en d'autres termes, l'étude vise à quantifier et comparer par des mesures de précision le pouvoir prédictif de différents modèles d'apprentissage automatique dans la gestion des crédits d'exploitation dans le secteur bancaire algérien des PME.

Ces outils peuvent servir d'aide à la gestion pour les gouvernements, les analystes de crédit, les investisseurs, les gestionnaires et d'autres parties prenantes dans la prise des décisions économiques.

6. Structure globale du mémoire

Afin d'avoir une réponse à notre problématique nous avons scindé notre recherche en trois chapitres :

Le premier chapitre intitulé : « La gestion de risque de crédit » présente la notion de crédit, sa typologie et la gestion du risque de crédit, le cadre règlementaire international et national sont ensuite aborder, ainsi que les différentes techniques de gestion.

Le deuxième chapitre : « Intelligence Artificielle, Machine Learning et leurs applications dans la finance » porte sur les notions et l'historique de l'intelligence artificielle et sur les méthodes les plus célèbres du Machine Learning. Enfin, quelques applications dans le secteur bancaire sont citées.

Le troisième chapitre est consacré à l'application des méthodes RL, KNN et ANN dans la classification des bénéficiaires de crédit d'exploitation par le biais de l'organisme bancaire algérien, le CPA. Il s'énonce : « Applications du Machine Learning dans le scoring des crédits d'exploitation destinés aux PME : Banque CPA » le chapitre vise à mesurer et comparer la performance des différents modèles ainsi que leur optimisation.

Chapitre 01 : La gestion du risque de crédit

Introduction

L'activité principale de la banque consiste à distribuer des crédits aux entreprises publiques et privées et aux individus afin de faire face à un besoin de financement, ces crédits sont par leurs natures, multiples et diverses.

A travers son activité d'octroi de crédit, la banque est confrontée au risque de non-remboursement, ce dernier peut générer une perte financière pour la banque. Ce risque devient de plus en plus récurant d'où la nécessité de le gérer. Afin de faire face à ce risque, les autorités du comité de Bâle ont mis en place un ensemble de normes prudentielles dans le but de réduire l'exposition à ce risque.

Le risque de crédit doit être identifié et mesuré par un système de notation en utilisant différentes méthodes notamment le crédit scoring. La notation permet d'évaluer la situation financière d'une entreprise, elle permet également d'avoir un suivi et une prévision face aux risques de non remboursement.

L'objectif de chapitre est de faire ressortir les différentes techniques de gestion du risque de crédit ainsi que l'évolution de la règlementation prudentielle liée à ce risque.

Le chapitre est présenté en trois sections. En premier lieu, nous allons voir les généralités du crédit, ensuite nous allons aborder le risque lié au crédit et présenter les aspects réglementaires relatif à ce risque. Enfin, la dernière section présente les différentes méthodes d'évaluation du risque de crédit.

Section 01 : Généralités sur le crédit

Le crédit bancaire est l'un des moyens les plus importants de financement de l'économie, il contribue au développement de l'activité économique. Dans cette partie nous allons aborder la notion de crédit, son rôle ainsi que sa typologie.

1. Définition du crédit

Les activités de la banque sont multiples et diverses. Elles reposent en grande partie sur les opérations de crédit et la collecte des dépôts. Cela est mentionné dans l'article 66 de la loi n°03-11 du 26 aout 2003 : « Les opérations de banque comprennent la réception de fonds du public, les opérations de crédit ainsi que la mise à disposition de la clientèle des moyens de paiement et la gestion de ceux-ci ».

Par ailleurs, la loi de la monnaie et du crédit, dans l'article 68 définit le crédit comme : « Tout acte à titre onéreux par lequel une personne met ou promet de mettre des fonds à la disposition d'une autre personne ou prend, dans l'intérêt de celle-ci, un engagement par signature tel qu'aval, cautionnement ou garantie ».[1]

« Faire un crédit c'est faire confiance, c'est donner librement la disposition effective et immédiate d'un bien réel, ou d'un pouvoir d'achat, contre la promesse que le même bien, ou un bien équivalent, vous sera restitué dans un certain délai, le plus souvent avec rémunération du service rendu et du danger encouru, danger de perte partielle ou totale que comporte la nature même de ce service ».[2]

De cette définition on constate que le crédit repose sur trois éléments :

La confiance : le créancier fait confiance au débiteur.

La promesse : tenir la promesse de rendre le prêt.

Le temps : la durée pendant laquelle le débiteur utilisera et profitera du prêt.

[1] Loi n°03-11 du 26 aout 2003 relative à la monnaie et au crédit, article 68.
[2] BOUYAKOUB Farouk : « *L'entreprise et le financement bancaire* », édition Casbah, Alger, 2000, p.17.

2. Rôle du crédit

Le crédit joue un rôle important dans l'économie et son financement. Il intervient dans différent aspects : [1]

2.1. L'échange : lors d'un besoin de financement, une entreprise fait recours à un crédit en passant par la banque. Cette dernière lui transfert un pouvoir d'achat. Cet échange constitue une anticipation sur ses recettes futures.

2.2. La stimulation de la production : la production est l'activité principale des entreprises manufacturières. Ces dernières ont recours à un crédit, cela leurs permet d'acquérir des équipements, de couvrir les charges directes et indirectes. D'autre part, les consommateurs aussi ont recourt à un crédit pour pouvoir consommer cette production et la stimuler.

2.3. L'amplification du développement : les effets d'un prêt pour l'achat d'un bien de production ou de consommation se manifestent chez l'agent économique bénéficiaire de l'opération, mais s'étendent également à d'autres agents économiques d'une manière indirecte.

2.4. La création monétaire : le crédit est considéré comme un instrument de création de la monnaie. L'octroi de crédit génère des flux monétaires, dont la source provient des dépôts d'autres agents. De cette façon, la monnaie n'est pas immobilisée.[2]

3. Classification du crédit

Les crédits proposés par les établissements bancaires sont diversifiés et cela est due aux différents besoins des demandeurs. Le crédit est classé selon différents critères, les principaux étant la durée et l'objet.

3.1. La durée

On distingue trois types de crédit :

- Les crédits à court terme (1 jour à 2ans).

[1] CAUDAMINE Guy et MONTIER Jean : « *Banque et marché financiers* », édition Economica, 1998, p.142.
[2] Ibid., p.143.

- Les crédits à moyen terme (2ans à7ans).
- Les crédits à long terme (plus de 7ans).

3.2. L'objet

On distingue deux types de crédit : [1]

3.2.1. Les crédits aux particuliers : ils sont accordés par les banques aux ménages pour combler les besoins personnels. Ils sont appelés crédits à la consommation. Dans notre étude on s'intéresse essentiellement sur le crédit accorder aux entreprises.

3.2.2 Les crédits aux entreprises : dont on distingue le crédit d'exploitation et le crédit d'investissement. Nous allons aborder ces deux types de crédits dans le point suivant.

4. La typologie des crédits

Il existe quatre types de financement : les crédits aux particuliers, le crédit au commerce extérieur, le crédit de l'exploitation et le crédit de l'investissement. Nous allons définir les deux derniers types de crédit en raison de leurs importances dans notre étude.

4.1. Le crédit de l'exploitation

Le crédit d'exploitation est une source de financement par endettement. Généralement ce mode de financement intervient au cours du cycle de production qui se caractérise par la transformation d'un flux monétaire en flux marchandises puis en flux monétaire.

Les crédits de l'exploitation permettent à l'entreprise de financer les activités à court terme, d'où l'actif circulant du bilan, leurs remboursements sont ainsi assurés par les recettes d'exploitation. Parmi les crédits d'exploitation, nous distinguons deux grandes catégories :[2]

4.1.1. Les crédits par caisse : c'est une forme de crédit ou la banque autorise d'aller au négatif sur un compte à vue d'une entreprise, jusqu'à un moment déterminé et à certaines conditions. Il existe 2 types de crédit par caisse :

[1] AUBIER Maud et CHERBONNIER Frédéric : *« L'accès des entreprises au crédit bancaire »*, Economie et prévision, 2007.
[2] Ibid.

A. Les crédits par caisse globale : ils servent principalement à couvrir les insuffisances momentanées du fonds de roulement. L'utilisation de ce type de crédit se fait par le débit du compte courant de l'emprunteur. On retrouve plusieurs formes de ce type de crédit :

La facilité de caisse : ce crédit assure l'élasticité nécessaire au bon fonctionnement de la trésorerie courante et permet de faire face aux décalages de très courte durée. S'agissant d'une souplesse de trésorerie, son utilisation doit être limitée et doit s'accompagner en contrepartie d'un mouvement significatif.[1]

Le découvert : est un concours bancaire destiné à financer un besoin de trésorerie né d'une insuffisance en fonds de roulement. Pour mettre en évidence les mécanismes du découvert, nous avons retenu la définition suivante : « Le découvert consiste pour le banquier, à laisser le compte de son client devenir débiteur dans la limite d'un maximum qui, le plus souvent, est fixé à titre indicatif sans qu'il y'ait engagement d'assurer le concours pendant une période déterminée ».[2]

Le crédit de compagne : le crédit de compagne est un concours bancaire destiné à financer un besoin de trésorerie né d'une activité saisonnière. Ce type de concours est généralement utile pour les entreprises qui, dans leur activité, sont soumises à une distorsion entre leur production et la consommation.[3]

B. Crédits spécifiques : ils ont des objets bien précis, qui ont des garanties dont la forme diffère selon le crédit sollicité. Ces garanties ne sont autres que le gage de certains actifs circulants (marchandise, créance). Les crédits spécifiques se présentent sous les formes suivantes :

L'escompte commercial : l'escompte est « une opération de crédit à court terme par laquelle un banquier appelé banquier escompteur, paie le montant d'un effet de commerce à son client qui en est porteur et il lui remet en contrepartie ».[4]

Le factoring (ou affacturage) : le factoring est un acte auquel une société spécialisée appelée factor, devient subrogée aux droits de son client, appelé adhérant, en payant ferme

[1] BEGUIN Jean-Marc et BERNARD Arnaud : « *L'essentiel des techniques bancaires* », édition Groupe Eyrolles, Paris, 2008, p.50.
[2] BRANGER Jacques : « *Traité d'économie bancaire : Instruments juridiques, techniques fondamentales* », Presses Universitaires de France, Paris, 1975, p.511.
[3] BENHALIMA Ammour : « *Pratique des techniques bancaires avec référence à l'Algérie* », éditions Dahleb, Alger, 1997.
[4] Ibid.

à ce dernier le montant intégral d'une facture à échéance fixe résultant d'un contrat et en prenant à sa charge moyennant une rémunération le risque de non remboursement.[1]

L'avance sur titres : c'est une opération qui consiste pour le possesseur de placement à obtenir un prêt dont les titres en portefeuille constitueront la garantie. Le propriétaire des titres les remet en gage avec un acte de nantissement signé par le propriétaire de bons.[2]

Enfin, il y a d'autres types d'avance comme : avance sur marché public, avance sur marchandises et avance sur facture administrative …etc.

4.1.2. Le crédit par signature : le crédit par signature appelé aussi engagement par signature, représente la garantie d'un banquier envers un tiers. Par conséquent, dans le cas où le client s'avère défaillant la banque paie à sa place. On distingue quatre formes de crédit par signature :

A. L'aval : il se définit par le droit des cambiaires comme étant : « un engagement fourni par un tiers qui se porte garant de payer tout, ou une partie du montant d'une créance, généralement un effet de commerce ». L'aval est soit donné sur le titre, soit sur un acte séparé.[3] Il se matérialise par la signature de la banque précéder par la mention « bon pour aval ».

B. Le cautionnement : « c'est un contrat par lequel une personne garantie l'exécution d'une obligation, en s'engageant envers le créancier à satisfaire cette obligation si le débiteur n'y satisfait pas lui- même ».[4] Il existe plusieurs types de cautions bancaires : les cautions fiscales, les cautions pour impôt, les cautionnements en douane.

C. L'acceptation : l'acceptation bancaire est l'engagement donné par une banque (tiers garant) de régler la lettre de change tirée sur elle à l'échéance convenue. Cette opération est fréquente tant dans le cadre du commerce extérieur que dans celui du commerce interne.

D. Le crédit documentaire : le crédit documentaire est un mode très utilisé dans le commerce extérieur. Il représente un engagement par le banquier de la part d'un importateur pour garantir à un exportateur le paiement des marchandises.

[1] Code de commerce algérien, article 543 bis 14.
[2] LAZARUS Jeanne : « *L'épreuve du crédit* », édition Sociétés contemporaines, Paris, 2009, p.25.
[3] Code de commerce algérien article 409.
[4] Ibid.

4.2. Le crédit d'investissement

Les crédits d'investissement servent à financer le haut du bilan qui représente les immobilisations que l'entreprise met en œuvre. Ils permettent la création, l'extension, ou la modernisation de l'unité de production. Parmi les différents types de crédits d'investissement existant, nous pouvons distinguer :

4.2.1. Les crédits à moyen terme : d'une durée de 02 à 07 ans, les crédits à moyen terme sont destinés à financer le matériel et les installations légères, ainsi qu'à l'acquisition d'équipements et d'outillages.[1]

4.2.2. Les crédits à long terme : « le crédit à long terme s'inscrit dans la fourchette huit à vingt ans. Il finance des immobilisations lourdes, notamment de constructions ».[2]

Le crédit-bail : « Le crédit-bail est une technique de financement d'une immobilisation par laquelle une banque ou une société financière acquiert un bien pour le louer à une entreprise, cette dernière ayant la possibilité de racheter le bien loué pour une valeur résiduelle généralement faible en fin de contrat ».[3] Il s'agit d'un contrat de location d'un équipement pour une période déterminée, avec option d'achat de ce dernier à la fin du contrat (chaque bien à des conditions spécifiques).

La prochaine section va porter sur les risques auxquels la banque est confrontée. Nous mettrons l'accent sur les spécificités de ce risque vis-à-vis des PME.

[1] GERMAIN-MARTIN Henry : « *Le crédit à moyen terme* », Revue d'économie politique,1958.
[2] BEGUIN Jean-Marc et BERNARD Arnaud : op.cit., p.50.
[3] DESMICHT François : « *Pratique de l'activité bancaire* », édition Dunod, France, 2004, p.68.

Section 02 : Les risques liés aux crédits

Le risque de crédit est considéré comme le plus important des risques auxquels est confrontée la banque et les établissements de crédit. Ce risque comprend tous les risques relatifs au défaut du non remboursement. Dans cette section, nous allons définir les risques liés à l'activité bancaire, principalement le risque de crédit et sa gestion par la banque, puis nous allons citer les normes prudentielles mis en place par les autorités bancaires.

1. Les risques liés à l'activité bancaire

L'activité de la banque est multiple et diverse, elle fait face à plusieurs risques. Ces risques sont classés en trois grandes catégories : le risque de marché, le risque opérationnel et le risque de contrepartie. Ce dernier est le risque le plus dangereux auquel est exposé la banque, c'est pourquoi on s'intéressera dans notre étude à ce type de risque.

1.1. Les risques de marché

Le risque de marché évoque le risque de pertes sur les positions de bilan ou de hors bilan résultant de la variation des prix sur le marché dans le cadre d'une activité de négociation.[1] Ce risque inclus : le risque relatif aux taux d'intérêt, titres de propriété et le risques de change liés aux transactions en devises.

1.2. Le risque opérationnel

Le comité de Bâle définit le risque opérationnel comme : « le risque de pertes résultant de carences ou de défauts attribuables à des procédures, aux personnels et aux systèmes internes ou à des évènements extérieurs ».[2]

1.3. Le risque de contrepartie

Le risque de contrepartie est aussi appelé le risque de crédit, il est représenté comme le premier risque auquel la banque fait face. Il est défini comme « un engagement portant une incertitude dotée d'une probabilité de gain et de préjudice, que celui-ci soit une dégradation ou une perte ».[3]

[1] RONCALLI Thierry : « *La gestion des risques financiers* », édition Economica, 2009, p.127.
[2] LAMARQUE Eric et MAURER Frantz : « *Le risque opérationnel bancaire* », Revue française de gestion, 2009.
[3] NAULLEAU Gérard et ROUACH Michel : « *Contrôle de gestion bancaire* », édition Revue banque, 2020, p.30.

2. Le risque de crédit

L'activité principale de la banque est le financement, lorsque la banque donne un prêt à une entreprise, elle prend le risque de ne pas récupérer l'intégralité du prêt accordé, c'est le risque de crédit.

2.1. Définition du risque de crédit

Comme on l'a cité en haut, le risque de crédit peut être défini comme la perte potentielle supportée par un agent économique suite à une modification de la qualité de crédit de l'une de ses contreparties, ou d'un portefeuille de contreparties, sur un horizon donné.[1] Ainsi, le risque de crédit traduit la défaillance possible d'un emprunteur, d'un émetteur d'obligation ou d'une contrepartie dans une transaction financier.[2] En effet, le risque apparait dès qu'on atteint la première échéance.

Le risque de crédit est estimé en fonction de trois paramètres : [3]

- Le montant de la créance.
- La probabilité de défaut.
- La proportion de la créance qui sera recouverte en cas de défaut.

Le terme « risque de crédit » est un terme qui contient le risque de défaut et le risque de contrepartie à savoir :

- **Risque de défaut :** ce risque se définie comme « l'occurrence d'un défaut qui se caractérise par l'incapacité du débiteur à faire face à ses obligations ».[4] Le comité de Bâle considère que ce défaut intervient lorsque l'un ou plusieurs des évènements suivants sont réalisés : [5]
    - Le débiteur ne remboursera vraisemblablement pas en totalité ses dettes (principal, charge financière).

[1] MAURER Frantz : « *L'impact du risque de marché sur le résultat de* l'entreprise », Revue française de gestion, 2005, p.77.
[2] HULL C John et al. : « *Gestion des risques et institutions financières* », 8ème édition Pearson, 2018, p.229
[3] VERNIMMEN Pierre et al. : « *Finance d'entreprise* », 20ème édition Dalloz, 2021, p.1050
[4] Ibid., p.1051
[5] FEKIR Hamza : « Présentation du nouvel accord de Bâle sur les fonds propres », Revue MIF, 2005, p.9.

- L'emprunteur est en défaut de paiement depuis quatre-vingt-dix (90) jours sur l'une de ses dettes.
- Le débiteur introduit une procédure de faillite ou une procédure similaire pour protéger ses créances.

➢ **Risque de contrepartie :** ce risque correspond à : « la défaillance de la contrepartie sur laquelle une créance ou un engagement est détenu ».[1] Ce risque est lorsque le débiteur n'arrive pas à honorer ses engagements financiers.

3. Les risque liés aux financements des PME

Les PME sont plus vulnérables que les autres entreprises, elles ont une probabilité de défaillance nettement plus importante que les grandes entreprises.[2] Le choix de financement est considéré comme un facteur déterminant de la stratégie financière de la PME.

Différentes sources de financement se présentent à l'entreprise dont le financement interne, qui consiste à ce que l'entreprise se finance par ses fonds internes. Le second type est le financement externe, cela consiste à faire appel aux parties externes à l'entreprise. Dans ce type de financement on retrouve le financement bancaire.

Actuellement, les banques proposent des emprunts bancaires qui se différencient par les durées, les modalités de remboursement, les taux d'intérêt, les garanties et les conditions de remboursement. Ainsi, on distingue deux grandes catégories de crédit bancaire : le crédit à court terme (crédit d'exploitation) et le crédit à moyen et long terme (crédit d'investissement).[3]

Cependant, le recours à l'emprunt bancaire est en forte relation avec la capacité de remboursement et d'endettement et le risque encouru par le prêteur.

Les PME ont un accès limité aux différentes sources de financement[4], plusieurs éléments de risque sont à prendre en considération en raison du manque de garantie ainsi qu'au problème informationnel existant entre la banque et l'entreprise. Il s'agit de l'asymétrie d'information.

[1] DE COUSSERGUES Sylvie et BOURDEAUX Gautier : « *Gestion de la banque : du diagnostic à la stratégie* », 6ème édition Dunod, Paris,2010, p.13.
[2] DE LA BRUSLERIE Hubert : « *Analyse financière* »,5ème édition, Dunod, Paris, 2014 p.429.
[3] Nous avons défini ce type de crédit dans la section 01, p.10-13.
[4] ANG James S : « *Small Business Uniqueness and the Theory of Financial Management* », The journal of entrepreneurial finance,1991, p.7.

3.1. L'asymétrie ex ante : il s'agit d'un problème de sélection adverse qui apparait avant la signature du contrat, le préteur ne peut pas évaluer la vraie valeur de l'entreprise et sa capacité exacte de remboursement, ceci dit l'emprunteur peut fournir de fausses informations selon ses convenances.[1]

L'asymétrie informationnelle se traduit par l'application d'un prix moyen unique pour des produits de qualité différente.[2] En appliquant ce principe sur le secteur bancaire, lorsqu'une entreprise dissimule certaines informations à la banque, cette dernière aura du mal à sélectionner le projet le plus rentable. De ce fait une sélection adverse se produit. Par conséquent les bons emprunteurs préfèrent se retirer du marché car ils considèrent que le taux d'intérêt moyen appliquer doit être moins élevé, tandis que les mauvais emprunteurs restent sur le marché avec des projets plus risqués.

3.2. L'asymétrie ex post : il s'agit d'un problème d'aléa moral, il apparait suite à l'opération de l'octroi de crédit, il s'agit d'une forme d'opportunisme ex post.[3]

L'aléa moral résulte de l'incapacité de la banque à observer les actions de l'emprunteur. Par conséquent, le risque de non remboursement dépendra du comportement de celui-ci, il sera supporté par la banque.[4] Ainsi, les banques auront plutôt tendance à financer un emprunteur qu'un projet. Toutefois, l'analyse des divers ratios financiers issus des bilans de l'entreprise, ainsi que l'interrogation du fichier de la centrale des risques, ne peuvent servir qu'à analyser sa santé financière passée (cela pose un grand problème lorsque l'entreprise est récemment créée).

Quel que soit le degré d'implication d'une banque dans la recherche d'information et la surveillance de son client, elle ne pourra jamais supprimer complètement l'asymétrie d'information et cela dépendra en partie, de la volonté du client de divulguer l'information à sa banque.[5]

[1] STIGLITZ Joseph E et WEISS Andrew : « *Credit rationing in markets with imperfect information* », American Economic Review, 1981.
[2] AKERLOF George A : « *The market for lemons : quality uncertainity and the market mechanisms* », The quarterly journal of ecnomics,1970.
[3] WILLIAMSON Oliver : « *La théorie des coûts de transaction* », Revue française de gestion, 2003, p.50.
[4] STIGLITZ Joseph E et WEISS Andrew : op. cit.
[5] GUILLE Marianne et al. : « *Structure financière et dépenses de R&D* », Economie et prévision, 2011, p.135.

4. La gestion du risque au sein de la banque

Le risque de crédit engendre des conséquences négatives sur l'activité bancaire. Précisément, les pertes liées au non remboursement des créances provoquent une diminution du résultat de la banque. Des agences de notation attribuent un rating aux banques commerciales en fonction de leurs résultats et solvabilité. Cette notation est fortement affectée par le degré de risque de crédits auquel la banque est exposée.

Il en découle que la maîtrise du risque de crédit constitue la préoccupation principale de la banque. Dans ce qui suit, les moyens les plus fréquemment employés pour se prémunir de ce risque sont présentés :

- Les supports à exiger dans l'administration du crédit.
- Les garanties pour se prémunir du risque de défaillance.

4.1. Les supports (documents)

Un ensemble de documents accompagne la demande de crédit et constitue la liasse nécessaire à l'établissement de la base de données. Cette dernière est consultée également lors d'un futur renouvellement du crédit ou d'autres formes d'assistance que les entreprises rechercheront auprès des banques. Les documents les plus importants dans le dossier de crédit sont la convention de crédit et l'assurance-crédit.[1]

4.2. Les garanties

Les banques exigent des garanties pour se protéger en cas de défaillance de l'emprunteur, on distingue trois types de garanties : [2]

4.2.1. Les garanties réelles : elles sont liées à la propriété sous forme d'actifs et de biens donnés par le débiteur à son créancier et prennent la forme d'hypothèques ou de nantissement.

[1] CHANT Elizabeth M et WALKER David A : « *Small business demand for trade credit* », Applied economics,1988, p.870.
[2] GOURIEROUX Christian et TIOMO André : « *Risque de crédit : Une approche avancée* », édition Economica, 2007, p.68.

4.2.2. Les garanties personnelles : c'est lorsque la banque demande au débiteur une caution, afin d'éviter que ce dernier s'oppose face à ses engagements. Elles prennent la forme de cautionnement ou d'aval.

4.2.3. Les garanties financières : elles permettent aux banques de récupérer facilement leurs crédits en suivant le cycle des ressources de l'entreprise. Elles comprennent les dépôts à terme et autres placements et les bons de caisse.

5. La réglementation prudentielle bancaire

La réglementation bancaire internationale et nationale en matière de gestion des risques veille sur la stabilité et la sécurité du système financier en établissant des normes prudentielles destinées aux banques.

5.1. La réglementation prudentielle internationale

La réglementation prudentielle internationale, résulte des travaux du comité de Bâle, cette dernière a connu une grande évolution ces dernières années. Elle a pour objectif de s'assurer que les banques sont suffisamment capitalisées au regard des risques pris. Même s'il est impossible d'éliminer complètement le risque de faillite bancaire, les autorités publiques mettent tout en œuvre pour réduire au minimum la probabilité de défaillance.

5.1.1. L'accord de Bâle I

En 1988, le premier accord de Bâle a mis en place des normes internationales en matière de régulation bancaire.[1] Cet accord a été conçu afin d'encadrer le risque de crédit par le respect d'un ratio unique et simple qui est le ratio de Cooke. Au nom de l'ex directeur de la banque d'Angleterre, Peter Cooke, qui fut, par la suite nommé le premier président du comité de Bâle. Ce ratio, également nommé le ratio de capital, ou le ratio de solvabilité internationale est imposé aux banques sur la base d'observations historiques du risque de crédit. Ce ratio prescrit que le capital social et les éléments assimilables à des fonds propres de toute banque doivent constituer au moins 8% du montant de leurs actifs et engagements hors bilan, ces derniers étant pondérés pas des coefficients de risque individuels.[2]

[1] HULL C John : op.cit., p.25

[2] ARTUS Patrick : « *De Bâle 1 à Bâle 2 : Effet sur le marché du crédit* », Revue économique, 2005, p.82.

$$Ratio\ Cooke = \frac{\text{Fond propre réglementaire}}{\text{La somme des risque de crédit pondérés}} > 8\%$$

5.1.2. L'accord de Bâle II

L'accord de Bâle de 1988 a été jugé insuffisant pour de multiples raisons : [1]

- Une conception des risques bancaires trop étroite, du fait qu'elle se limite au seul risque de crédit et depuis 1996 au risque de marché.
- Une mesure de risque insuffisamment affinée : pondération uniforme des entreprises à 100%, même si elles étaient dotées de toutes les garanties et bien notées, alors que certains états de l'OCDE, pourtant pondérés à 0% ont pu se révéler risqués.
- Une grille de pondération rigide qui ne prend pas en compte les techniques de réduction des risques (garanties).

Et c'est pour cela qu'un nouvel accord prudentiel, à savoir ; l'accord de Bâle II, est apparu en 2004. Ce dernier vise à mieux évaluer les risques bancaires et à imposer un dispositif de surveillance prudentiel transparent. Outre le risque de crédit, le ratio McDonough doit avoir une solvabilité minimale de 8% qui se calcule par rapport de la somme des fonds propres réglementaire et la somme des risques de crédit, opérationnels et de marché pondéré.

Les recommandations du nouvel accord s'appuient en trois piliers : [2]

- **Pilier I :** Exigence minimale de fonds propres :
  - Risque de crédit (nouvelles approches de calcul).
  - Risque de marché (inchangé).
  - Risque opérationnel(nouveau).
- **Piller II :** Surveillance par les autorités prudentielles :
  - Évaluation des risques et dotation en capital spécifique à chaque banque.
  - Communication plus soutenue et régulière avec les banques.
- **Pilier III :** Discipline de marché :
  - Renforcement de la communication financière.
  - Obligation accrue de publication (notamment de la dotation en fonds propres et des méthodes d'évaluation des risques).

[1] ARTUS Patrick : op. cit., p.90.
[2] NOUY Danièle : « *Bâle II face à la crise : quelle réformes* », Revue d'économie financière, 2008, p.369

5.1.3. L'accord de Bâle III

La crise financière de 2007- 2008 a poussé le comité de Bâle a publié une série de documents visant l'optimisation des trois piliers de l'accord de Bâle II dans le but de renforcer le système financier. Plus précisément, l'accord vise à améliorer la qualité des fonds propres bancaires et d'assurer une gestion plus stricte du risque de liquidité.[1]

La mise en place de Bâle III va : [2]

- ✓ Améliorer la qualité des fonds propres des banques.
- ✓ Réduire le risque systémique.
- ✓ Prévoir un délai suffisant pour que le passage au nouveau régime s'opère sans heurts.

5.2. La réglementation nationale

La règlementation prudentielle nationale s'inspire principalement de la réglementation internationale tout en l'adaptant au contexte Algérien, ce dernier s'avère plus rigide.

5.2.1. Le ratio de couverture des risques

Les banques et établissements financiers Algériens doivent toujours respecter une solvabilité minimale de 9,5%. Elle se calcule par le rapport de la somme des fonds propres réglementaires et la somme des risques de crédit, opérationnels et de marché pondérés (Total de l'actif net pondéré).[3]

5.2.2. Ratios de division et de concentration des risques

Les banques doivent s'assurer en permanence de la diversification de leur portefeuille d'investissement en respectant deux ratios de risque (par client et par groupe) afin d'éviter toute concentration des risques sur un même client ou un groupe de clients tel que stipulé dans le règlement n° 14-02 du 16 février 2014 relatif aux grands risques et aux participations :

- Le montant des risques encourus sur un même bénéficiaire ne doit pas excéder 25% des fonds propres.

[1] COUPPEY-SOUBEYRAN Jézabel : « *De Bâle 2 à Bâle 3 : la nouvelle réglementation bancaire internationale* », La documentation française, 2013, p.45
[2] RUGEMINTWARI Clovis et al. : « *Bâle 3 et la réhabilitation du ratio de levier des banques* », Revue économique, 2012, p.814.
[3] Règlements de la Banque d'Algérie N°14-01 du 16 février 2014 portant coefficients de solvabilité applicables aux banques et établissements financiers, article 2

- Le total des grands risques encourus par une banque ou un établissement financier ne doit pas dépasser huit 8 fois le montant de ses fonds propres réglementaires.

5.2.3. Classement et provisionnements des créances

Selon le règlement de la Banque d'Algérie n°14-03 du 16 février 2014 relatif aux classements et provisionnement des créances, ces derniers sont classés en créances courantes et en créances classées.

A. Les créances courantes : Elles représentent les créances dont le recouvrement intégral dans les délais parait assuré, les créances appartenant à cette catégorie sont : [1]

- Les créances assorties de la garantie de l'Etat.
- Les créances garanties par les dépôts constitués auprès de la banque ou de l'établissement financier prêteur.
- Les créances garanties par les titres nantis pouvant être liquidés sans que leur valeur ne soit affectée.

B. Les créances classées : Ce sont des créances qui ont un taux de provisionnement plus élève à cause de leurs caractères de risque. Nous distinguons trois types de ces créances dans le tableau suivant : [2]

Tableau N° 1 : Les types des créances classées

Créances	Retard de remboursement	Taux d'approvisionnement
Créances à problèmes potentiels	De 3 mois à 6 mois	20%
Créances très risquées	De 6 mois à 1 an	50%
Créances compromises	Plus d'un an	100%

Source : Elaboré par les auteurs selon le règlement de la banque d'Algeria n°14-03 du 16 février 2014 relatif au classements et provisionnements des créance

[1] Règlement 14-03 du 16 février 2014 relatif au classement et provisionnement des créances et des engagements par signature des banques et établissements financiers, article 4.
[2] Ibid., article 5.

Section 03 : Les méthodes d'évaluation du risque de crédit

La mesure du risque de crédit des entreprises est un enjeu important pour les banques, ces dernières se doivent de mettre en place des méthodes pour pouvoir le quantifier. Cette section va traiter les différentes approches d'évaluation du risque de non remboursement tel que l'analyse financière et les nouvelles méthodes de scoring.

1. L'approche traditionnelle : (L'analyse financière)

L'analyse financière est une technique de l'approche traditionnelle. Il s'agit probablement de la méthode à la fois la plus ancienne et la plus utilisée dans l'analyse du risque. Elle a pour objet d'étudier le passé afin de prévoir le présent et l'avenir.[1]

« L'analyse financière est un ensemble de concepts, méthodes et outils qui permettent de traiter des informations internes et externes, en vue de formuler des recommandations pertinentes concernant la situation d'un agent économique spécifique, le niveau et la qualité de ses performances, ainsi que le degré de risque dans un environnement fortement concurrentiel ».[2]

L'analyse financière s'effectue traditionnellement sur la base des états financiers des entreprises. Donc son objectif est de savoir si, à travers quelques ratios, la banque peut attribuer ou non un crédit. Cette évaluation est basée sur la situation de l'entreprise et non pas sur le risque de non remboursement.

1.1. Limites de l'approche traditionnelle

L'analyse financière est la technique de diagnostic la plus utilisée. Cette méthode analyse la rentabilité la solvabilité de l'entreprise pour des fins de décisions d'investissements ou de financement. Cependant, elle est limitée car différents ratios peuvent donner des prévisions différentes pour une même entreprise. Il existe également le problème d'asymétrie d'information entre prêteurs et emprunteurs.[3] C'est pourquoi les banques peuvent être incitées à développer d'autres modèles tel que le crédits scoring comme outil moderne dans la prévision de la défaillance des entreprises et la mesure du risque de crédit.

[1] VERNIMMEN Pierre et al., op.cit., p.950.
[2] DAYAN Armand : « *Manuel de gestion* », édition Collectif ellipses, volume 2, 2004, p.59.
[3] Ibid., p.20.

2. Les nouvelles méthodes d'évaluation du risque de crédit : le crédit scoring

Le crédit Scoring un outil de gestion et d'évaluation utilisée par les organismes bancaires pour estimer le risque de défaut et mesurer la solvabilité de chaque entreprise en la classent soit comme une entreprise saine ou une entreprise défaillante.

2.1. Définition du crédit scoring

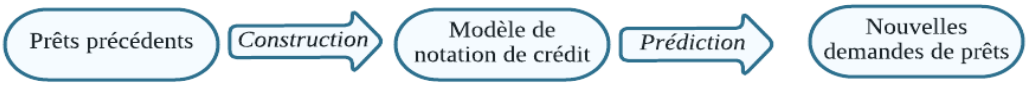
Bardos (2008) définit le scoring comme étant une analyse statistique permettant de prédire la qualité d'un emprunteur.

Quant à Flaman, (1997), le crédit scoring est le processus d'assignation d'une note (ou score) à un emprunteur potentiel pour estimer la performance future de son prêt.

Thomas et al. (2002) stipulent que le crédit scoring représente un ensemble de modèles de décision et des techniques sous-jacentes qui permettent de décider l'octroi des crédits de consommation.

À travers l'historique des données des prêts précédents, le scoring permet d'attribuer une note appelée « score », qui permet de prédire la probabilité de défaut pour les nouveaux crédits. Le processus du crédit scoring peut être résumé comme suit : [1]

Figure N° 1 : Processus du crédit Scoring



Source : LIU Yang : « New issues in credit scoring application », 2001

L'objectif des modèles de notation de crédit est d'attribuer aux clients une note (client sain ou défaillant). Par conséquent, les problèmes de notation sont liés à l'analyse de classification.[2]

[1] LOTFI Sihem et MESK Hicham : « *Prédiction du risque de crédit : étude comparative des techniques de Scoring* » International Journal of Accounting, Finance, Auditing, Management and Economics, 2020, p.514.
[2] HUSSEIN Abdou et al. : « On the applicability of credit scoring models in egyptian banks », Banks and Bank Systems,2002, p.7.

2.2. Historique de crédit scoring

En 1958, le cabinet Fair et Issac a mis en place un système de traitement de données de masse capable de calculer le risque d'un prêt grâce à un calcul d'un score appelé FICO.

Le score FICO est un système de notation de crédit américain, il est utilisé par les principales agences de crédit pour évaluer la solvabilité des entreprises, il se situe entre 300 et 850. Un score FICO de 703 est considéré comme bon.

Le pionnier du crédit scoring est attribuable à Beaver (1966). Il utilise l'analyse univariée afin de classer les entrepris en entreprises saines ou en difficultés selon le taux d'erreur le plus faible.[1] Pour développer son modèle, Beaver a utilisé cinq catégories de ratios financiers, soit : les ratios de flux monétaires, les ratios de revenu, les ratios d'endettement, les ratios de liquidité et les ratios de rotation.[2] Bien que cette méthode fournisse des résultats performants, elle a été énormément critiquée du fait qu'elle se limite à quelques ratios uniquement, ce qui ne décrit pas la situation réelle de l'entreprise. Malgré les critiques, cette méthode a été le point de départ pour le développement d'autres modèles tel que le modèle z-score publié par Altman (1968) et qui apparaît le modèle de prédiction des défauts le plus populaire de la littérature.

Ce modèle est développé en utilisant un échantillon de 66 entreprises réparties en deux classes de 33 chacune : une classe pour des entreprises considérées comme défaillantes, l'autre classe pour celles considérées comme saines. Le modèle utilise la technique statistique de l'analyse discriminante multivariée. Il détermine une fonction de score qui est une combinaison linéaire de cinq ratios financiers considérés comme les plus pertinents, pour discriminer au mieux les deux groupes d'entreprises (saines ou défaillantes). Cette fonction de score, nommée Z-score, s'exprime par la relation :

$$Z = 1.2R_1 + 1.4R_2 + 3.3R_3 + 0.6R_4 + 0.9R_5$$

$R_1 =$ Fond de net / Actif total

$R_2 =$ Bénéfice non réparti / Actif total

$R_3 =$ Bénéfice avant intérêts et impôts / Actif total

[1] LOTFI Sihem et MESK Hicham : op.cit., p.520.

[2] Ibid., p.516.

$R_4 = \text{Capitaux propres / Dettes totalesl}$

$R_5 = \text{Chiffre d'affaires H. T / Actif total}$

Le risque encouru par la banque varie dans le sens contraire de Z, avec 3 comme valeur critique.[1] Pour un score supérieur à 3, l'entreprise a peu de risque de faire défaut, entre 2,7 et 3, l'entreprise est à risque. S'il est compris entre 1,8 et 2,7, la probabilité de faire défaut est importante et l'entreprise est jugée à haut risque. Enfin pour un score inférieur à 1,8 la probabilité d'un problème financier est très élevée.

Toutefois, Le problème majeur dans l'application de ces méthodes, est que la validité des résultats trouvés par ces techniques est tributaire de leurs hypothèses restrictives qui sont rarement satisfaites dans la vie réelle, en l'occurrence l'hypothèse de la normalité de la distribution de chacune des variables retenues et l'hypothèse de l'indépendance entre celles-ci ce qui peut rendre ces méthodes théoriquement invalides.[2] Par conséquent, le caractère contraignant des hypothèses de base nécessaires pour une mise en œuvre efficace de l'analyse discriminante a conduit certains chercheurs à tester l'efficacité d'autres outils statistiques.

Dans les années 80, le calcul du crédit Scoring est revenu grâce à l'émergence des systèmes expert. Par conséquent, les banques ont commencé à utiliser ce genre de systèmes dans leurs méthodes de contrat de crédit mais aussi dans d'autres domaines d'expertise pour calculer un risque particulier (par exemple, le calcul de la rentabilité par rapport au risque ...). Un système expert consiste en un logiciel développé pour réaliser une analyse financière à l'aide d'une base de connaissances, conduit par un spécialiste /expert du domaine, enrichie d'analyses préalablement réalisées.[3]

2.3. Le choix de la technique à utiliser

Avec le développement des différents besoins des systèmes de crédit scoring, on assiste à une variété des méthodes d'évaluation des risques de crédit avec un objectif identique qui est l'augmentation de l'efficacité des prises de décision.

[1] LOTFI Sihem et MESK Hicham : op.cit., p.516.

[2] HUANG Chen-Lung et al. : « *Credit scoring with a data mining approach based on support vector machines* », Expert systems with applications, 2004, p.848.

[3] LOTFI Sihem et MESK Hicham : op.cit., p.518.

Le principe de ces méthodes est d'identifier les variables qui déterminent la probabilité de défaut afin de pondérer leurs poids en un score quantitatif.

Ces systèmes de crédit scoring sont mis en place à partir de quatre principales formes de modélisation multivariées :

- Le scoring par le modèle linéaire.
- Le scoring par le modèle d'analyse discriminante.
- Le scoring par le modèle Logit.

En plus de ces techniques, on trouve d'autres méthodes d'intelligence artificielle : [1]

- Les réseaux de neurones.
- Les arbres de décision.
- Les systèmes experts.
- Algorithmes Génétique.

Par ailleurs, le secteur bancaire présente des caractéristiques majeures (clients, historiques, informatisation poussée, volumes de données homogènes) qui en font un terrain de jeu privilégié pour le déploiement d'outils basés sur l'IA.

[1] TUFFERY Stéphane : « *Data Mining et statistique décisionnelle : L'intelligence des données* », éditions TECHNIP, 2012, p.89.

Conclusion

Par le moyen de l'intermédiation financière, le risque de crédit est le risque principal auquel est confronté la banque, il correspond à une menace de pertes dans le cas de la défaillance de l'emprunteur.

De nombreuse études ont mis le point sur la gestion du risque de crédit vue la forte exposition de la banque face à ce risque. Pour se prémunir de ce dernier différentes politiques ont été mises en place par la réglementation prudentielle. D'autre part, l'utilisation des méthodes d'évaluation tel que l'analyse financière, actuellement utilisées par les banques algériennes ainsi que le crédit scoring, comme outil plus moderne, a pour finalité de mieux prévoir les pertes potentielles pouvant mettre la banque dans une situation d'insolvabilité.

Dans le second chapitre, nous allons nous intéresser aux nouvelles méthodes de gestion du risque de crédit, les méthodes IA.

Chapitre 02 : Intelligence Artificielle, Machine Learning et leurs applications dans la finance

Introduction

Le Machine Learning est un des champs d'application de l'intelligence artificielle. Cette discipline a pour but de développer des modèles algorithmiques capables d'automatiser l'activité humaine, notamment, dans les problématiques de classification et/ou de régression.

Les systèmes basés sur l'apprentissage atteignent cet objectif en prospectant sur les motifs réguliers et récurrents d'un sous-ensemble de données de formation, ce qui permet de générer des hypothèses sur l'ensemble de données. Ainsi, la performance d'un tel système devrait s'améliorer à mesure qu'il acquiert de l'expérience ou des données.

L'objectif de cette introduction est de dresser un panorama du machine learning ainsi que les méthodes les plus utilisées, elle a aussi pour objectif de donner une vue globale sur ses applications dans le secteur bancaire.

Le chapitre se présente comme suit :

Dans une première section, un aperçu global sur l'IA est d'abord exposé. Ensuite, des concepts de base sur l'apprentissage automatique sont présentés, avec un accent particulier sur la classification binaire supervisée. Enfin, le chapitre se termine par des applications de l'IA dans les problématiques du secteur bancaire.

Section 01 : Aperçu sur l'intelligence artificielle

Dans une première section, nous allons nous positionner par rapport au terme « IA » à travers une présentation générale de cette discipline.

1. Historique de l'intelligence artificielle

Le Machine Learning (ou l'apprentissage automatique) fait son apparition dans les années 80, c'est une branche de l'intelligence artificielle dont les débuts remontent à l'après-guerre. L'intelligence artificielle est introduite par Alan Turing, un mathématicien britannique ayant inventé l'architecture mathématique des ordinateurs. En 1950, il publia l'article « Machines informatiques et intelligence », ou il posa la question suivante : « est ce que les machines peuvent penser ? » accompagné d'une mesure de l'intelligence des machines ; le test de Turing.[1]

Cet article ouvra le débat aux curieux et les regroupa dans la conférence de « Dartmouth Summer Research Project on Artificial Intelligence » en 1956. Cet événement marqua la véritable naissance du terme IA.[2] Dans lequel le principe suivant est formulé : « chaque aspect de l'apprentissage ou toute autre caractéristique de l'intelligence peut être si précisément décrit qu'une machine peut être conçue pour le simuler ».[3]

Dans les années 60, un psychologue et chercheur : Frank Rosenblatt réussit à imiter le fonctionnement du cerveau humain en créant un réseau de neurones artificiels appelé « perceptron ». La particularité de cet algorithme est que la relation et le poids de la connexion entre les neurones pourraient être ajustés suite aux erreurs de perception. Cette invention a permis une avancée importante dans l'histoire de l'IA.

La fin des années 60 marque une période de crise dans le secteur, la recherche et les investissements se gèlent, les premières critiques et questions d'étiques émergent. En 1966, le rapport US ALPAC souligne l'absence de progrès dans la recherche en traduction automatique visant la traduction immédiate du russe dans le contexte de la guerre froide. De

[1] BOUCHARD Guillaume : « *Les modèles génératifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle »,* Interface homme-machine université Joseph-Fourier - Grenoble I, 2005, p.146

[2] MILANA Carlo et ASHTA Arvind *: « Artificial intelligence techniques in finance and financial markets : A survey of the literature »*, Strategic change, 2021.

[3] BOUCHARD Guillaume : op. cit., p.149.

nombreux programmes financés par le gouvernement américain sont annulés. De même, en 1973, le gouvernement britannique publie le rapport "Lighthill", soulignant la déception de la recherche sur l'IA. Encore une fois, les compressions budgétaires réduisent les programmes de recherche. Cette période de scepticisme durera jusqu'en 1980, ce que l'on appelle aujourd'hui « le premier hiver de l'intelligence artificielle ».[1]

Cet hiver se termine avec l'arrivée des systèmes experts, apportant une idée neuve : au lieu de tester toutes les solutions possibles à un problème on va créer des heuristiques, c'est des façons de ne tester que les solutions pertinentes, qui ont une chance d'être juste.

Le Japon et les États-Unis investissent massivement dans la recherche sur l'IA. Les entreprises dépensent plus d'un milliard de dollars par an en systèmes experts, et le secteur est en pleine croissance.[2]

La fin des années 90 marque l'entrée dans l'IA moderne, celle qu'on utilise aujourd'hui. La discipline de l'IA se fractionne en plusieurs spécialités, ce qui coïncide avec la révolution numérique, et l'apparition du concept de data.

En 1997, l'histoire de l'intelligence artificielle est marquée par un évènement majeur, le programme Deep Blue de l'IMB remporte la victoire dans le monde des échecs contre un Homme, le champion Gary Kasparov.

Une décennie plus tard, les avancées technologiques ont permis de mettre à jour l'intelligence artificielle. En 2008, Google a fait d'énormes progrès dans la reconnaissance vocale et a introduit la fonctionnalité dans son application pour smartphone.

En 2012, Andrew Ng a fourni un réseau neuronal utilisant 10 millions de vidéos YouTube comme ensemble de données de formation. Grâce au Deep Learning, ce réseau de neurones a appris à reconnaître les chats sans avoir à apprendre ce que sont ces derniers. C'est le début d'une nouvelle ère d'apprentissage en profondeur.

En 2016, une autre victoire de la machine frappa et cela dans le monde des jeux, le système alpha DeepMind de Google sur Lee Sedol qui été le champion de jeu de Go.

[1] MILANA Carlo et ASHTA Arvind : op. cit.
[2] Ibid.

En 2017, 80% des grandes entreprises du monde avaient déjà investi dans l'IA, aujourd'hui ses champs d'application sont divers.[1]

2. Définition de l'intelligence artificielle

L'intelligence artificielle (IA) est un domaine d'étude qui s'appuie sur de nombreuses disciplines dont la philosophie, l'informatique, les mathématiques et la théorie de l'information.[2] Le test de Turing, inventé par Alan Turing en 1960, consiste à dire que l'IA est atteinte lorsqu'un interlocuteur n'est plus capable de distinguer la conversation d'un humain d'une machine.

L''intelligence est la capacité de penser, d'imaginer, de mémoriser, de reconnaître les schémas, de faire des choix, de s'adapter aux changements et d'apprendre de l'expérience.[3]

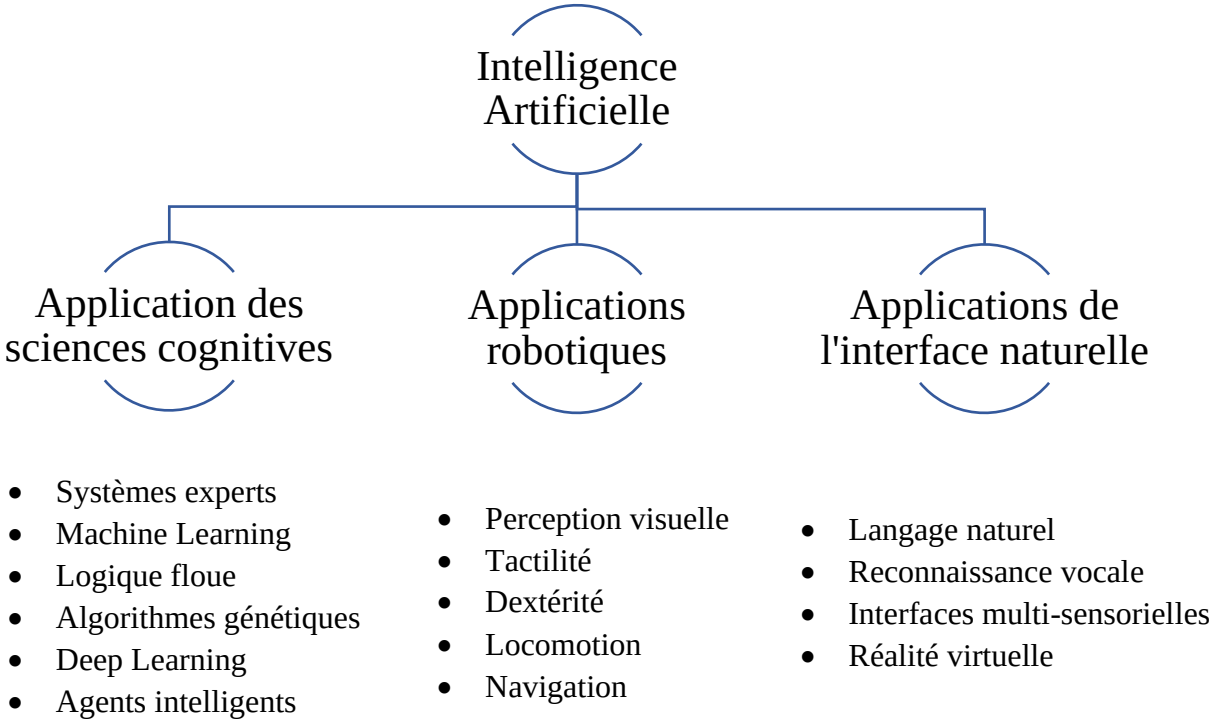
3. Les branches de l'intelligence artificielle

Pour cerner les capacités actuelles de cette technologie, il est important de différencier deux types d'IA ; l'IA faible et l'IA forte. L'IA faible, également connue sous le nom d'IA étroite, se concentre sur l'exécution d'une tâche spécifique, programmées à l'avance sans aucune forme d'improvisation. Ces modèles sont préalablement alimentés à l'ordinateur par un humain comme répondre à des questions basées sur l'entrée de l'utilisateur ou jouer aux échecs. A la différence de l'IA faible, l'IA forte correspond à un programme informatique capable de raisonner, apprenant éventuellement à résoudre de nouveaux problèmes de manière autonome. Bien que l'apport humain accélère la phase de croissance de l'IA forte, il n'est pas nécessaire et, au fil du temps, l'algorithme développe une conscience de type humain au lieu de la simuler.[4]

Les systèmes basés sur l'IA sont la manifestation d'un large éventail de technologies et de stratégies axées sur le développement : [5]

[1] BOUCHARD Guillaume : op. cit.
[2] COOK Diane J et al. : « *Graph based hierarchical conceptual clustering* », Journal of machine learning research, 2001.
[3] BORANA Jatin, « *Applications of artificial intelligence & associated technologies* », Department of electrical engineering, Jodhpur National University ,2016, p.65.
[4] Ibid., p.66.
[5] https://www.softwaretestinghelp.com/what-is-artificial-intelligence/ 25/03/2022 à 10 : 53

Figure N° 2 : Disciplines de l'intelligence artificielle



Source : BORANA Jatin, « Applications of artificial intelligence & associated technologies », Department of electrical engineering, Jodhpur National University ,2016.

Le machine learning est une application de l'IA faible, qui permet aux ordinateurs d'apprendre à partir de l'expérience et d'améliorer l'exécution de tâches spécifiques. Il permet aux ordinateurs d'analyser des données et d'employer des techniques statistiques pour apprendre à partir de ces données afin d'améliorer leur capacité d'exécution.[1] Une présentation plus détaillée sera exposer dans la section suivante.

[1] BORANA Jatin : op. cit., p.66.

Section 02 : Généralités sur le Machine Learning

La section suivante présente l'aspect générale du machine learning et ses trois branches ainsi que leurs méthodes les plus populaires.

1. Le Machine Learning (apprentissage automatique)

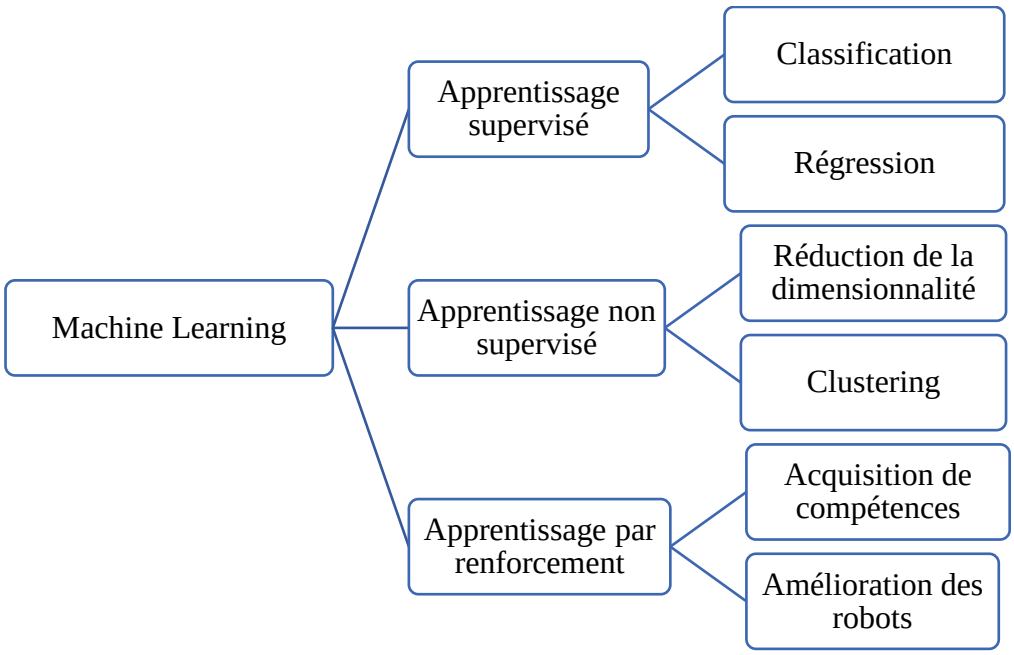
Il s'agit d'une branche d'étude de l'intelligence artificielle qui permet à une machine d'apprendre par elle-même, à partir de données, à atteindre un certain objectif sans être explicitement programmée.

- L'apprentissage automatique se produit lorsque les ordinateurs modifient leurs paramètres/algorithmes lors de l'exposition à de nouvelles données sans que les humains aient à les reprogrammer.[1]
- Arthur Samuel a décrit l'apprentissage automatique comme l'étude qui permet aux ordinateurs d'apprendre sans être explicitement programmés.[2]
- Quant à Hurlin et Pérignon, il pense qu'il s'agit d'un ensemble d'algorithmes destinés à résoudre des problèmes dont la performance s'améliore avec l'expérience et les données sans intervention humaine a posteriori.[3]

Tous ce qui peut être stocké numériquement peut servir de données pour le ML, de ce fait on le trouve dans plusieurs applications.[4]

[1] REHFELDT Ruth Anne et al. : « *Observational learning and the formation of classes of reading skills by individuals with autism and other developmental disabilities D. Latimoreb, Research in development disabilities* », Res dev disabil, 2003.
[2] CORNUEJOLS Antoine et al. : « *Apprentissage artificiel* », édition Eyrolles, 2003.
[3] HURLIN Christophe et PERIGNON Christophe : « *Machine learning et nouvelles sources de données pour le scoring de crédit* », Revue d'économie financière, 2019.
[4] MILANA Carlo et ASHTA Arvind *:* op. cit.

Figure N° 3 : Types du Machine Learning



Source : MILANA Carlo et ASHTA Arvind : « Artificial intelligence techniques in finance and financial markets : A survey of the literature », Strategic change, 2021.

Le ML est également très utilisé dans l'analyse des données et data science. Notamment, dans les problèmes de classification et régression. C'est deux disciplines feront l'objet de notre mémoire.

2. Types de Machine Learning

L'apprentissage automatique peut être divisé en trois classes : [1]

- L'apprentissage supervisé.
- L'apprentissage non supervisé.
- L'apprentissage par renforcement.

2.1. Apprentissage supervisé

Dans les algorithmes d'apprentissage supervisé, un ensemble de données d'apprentissage étiquetées est d'abord utilisé pour entraîner l'algorithme sous-jacent. Cet algorithme entraîné

[1] UlLLAH Hayat et al. : « *Comparative study for machine learning classifier recommendation to predict political affiliation based on online reviews* » CAAI Trans Intell Technol, 2021.

est ensuite alimenté par l'ensemble de données de test non étiqueté pour les classer dans des groupes similaires.[1]

Ce type d'apprentissage est subdivisé en régression et classification.[2] Dans les problèmes de classification, la variable de sortie sous-jacente est discrète. Cette variable est classée en différents groupes ou catégories, tels que : rouge ou noir, solvable ou non solvable.[3]

La variable de sortie correspondante est une valeur réelle dans les problèmes de régression, tels que le risque de développer une maladie cardiovasculaire pour un individu.[4] Le but est de construire un modèle à partir d'un ensemble d'apprentissage permettant de prévoir la sortie d'une nouvelle entrée non étiquetée.[5]

Parmi les modèles supervisés, on distingue :

2.1.1. Les modèles de classification supervisés

Dans les problèmes de classification, la variable de sortie sous-jacente est discrète. Cette variable est classée dans différents groupes ou catégories.

Comme évoqué précédemment, la finalité d'un tel modèle est de prédire si une observation X appartient à telle ou telle catégorie définie au préalable. Lorsqu'il n'y a que deux choix possibles, on parle de classification binaire (cas de notre étude).

Les algorithmes les plus fréquemment utilisés sont :

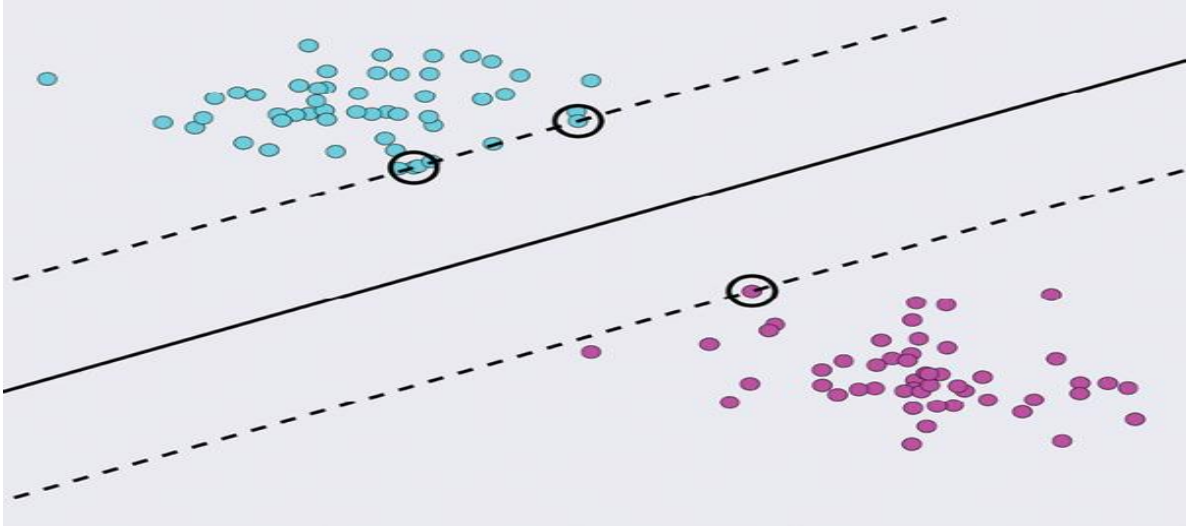
- Machines à vecteurs de support **(**SVM).
- K plus proches voisins (KNN).
- Arbres de décision (DT).
- Forêt aléatoire (RF).
- Réseaux de neurones artificiels (ANN).
- Les Classificateurs Bayésiens Naïfs.

[1] UDDIN Shahadat et al. : « *Comparing different supervised machine learning algorithms for disease prediction* », BMC Med Inform Decis Mak, 2019.
[2] MHLANGA David : « Financial inclusion in emerging economies : The application of machine learning and artificial intelligence in credit risk assessment, International journal of financial studies,2021.
[3] UlLLAH Hayat et al. : op. cit.
[4] MHLANGA David : op. cit.
[5] MILANA Carlo et ASHTA Arvind : op. cit.

A. Machines à Vecteurs de Support (SVM)

SVM est un modèle paramétrique permettant de traiter des problèmes de classification et de régression linéaires et non linéaires. L'algorithme crée la meilleure ligne ou hyperplan qui sépare les données en classes, en maximisant la distance entre les deux classes. Cette distance est appelée la marge et les points qui tombent exactement sur cette marge sont appelés les vecteurs de support.[1]

Figure N° 4 : Représentation graphique de SVM



Source : GLASSNER Andrew : « Deep learning : A visual approach », édition No strach press, 2021.

De toutes les lignes qui séparent les deux ensembles de points, la ligne présentée sur la figure n° 4 est la plus éloignée de chaque ensemble, car elle présente la plus grande distance par rapport à ses vecteurs de support. Les observations encerclées sont les observations les plus proches, ou les vecteurs de support.

La distance entre la ligne continue et les lignes pointillées qui passent par les vecteurs de support est appelée la marge. Intuitivement, plus les points de données sont éloignés de l'hyperplan, plus nous sommes sûrs qu'ils ont été correctement classés. Nous voulons donc que nos points de données soient aussi éloignés que possible de l'hyperplan, tout en restant du bon côté.[2]

[1] VAPNIK Vladimir N et al. : « A training algoritm for optimal margin classifiers », Computational learning theory, 1992.
[2] GLASSNER Andrew : « *Deep learning : A visual approach* », édition No strach press, 2021, p.445.

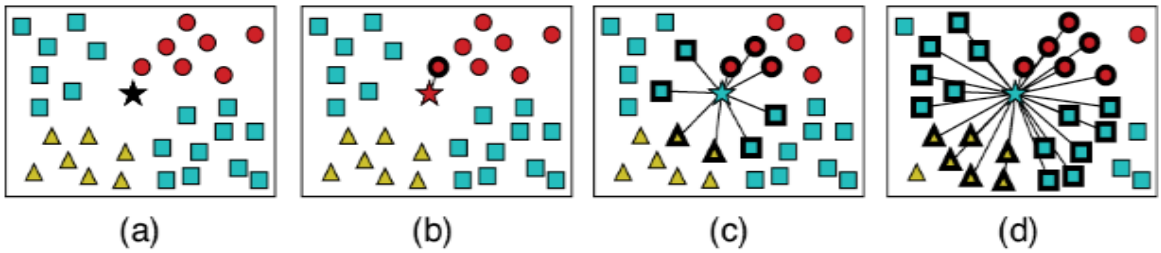
D'autre part la qualité des résultats est sensible au paramètre C (d'où le caractère paramétrique du modèle) qui spécifie le nombre d'observations autorisés près de la frontière. Plus la valeur de C est grande, plus l'algorithme exige une zone vide autour de la ligne et vice-versa. Nous pouvons utiliser la validation croisée pour essayer différentes valeurs de C et choisir la meilleure.

SVM est limité par les formes qu'il est capable de trouver. L'algorithme, par exemple, ne peut trouver que des formes linéaires, comme les lignes et les plans. L'astuce Kernel[1] ou l'astuce du noyau, peut parfois nous permettre d'utiliser une frontière linéaire là où il semble initialement que seule une frontière non linéaire peut être ajustée.

B. K plus proches voisins (KNN)

KNN est un classificateur non linéaire d'apprentissage supervisé qui prédit à quelle classe appartient un nouveau point de données X en identifiant la classe de ses k voisins les plus proches en termes de distance (euclidienne, Manhattan...).[2]

Figure N° 5 : Représentation graphique de KNN



Source : GLASSNER Andrew : « Deep learning : A visual approach », édition No strach press, 2021.

La figure n° 5 : pour trouver la classe d'une nouvelle observation, représenté par une étoile, nous trouvons le plus proche de ses k voisins. Dans la figure n°5(a), nous avons une nouvelle observation (une étoile) au milieu d'un groupe d'autres observations qui représentent trois classes (cercle, carré et triangle). Pour déterminer une classe pour l'étoile, nous regardons ses k voisins les plus proches et nous comptons leurs classes. La classe qui est la plus peuplée devient la classe de la nouvelle observation.[3]

[1] GLASSNER Andrew : op. cit., p.446.
[2] ZHANG Min-Ling et ZHOU Zhi-Hua : « *ML-KNN : A lazy learning approach to multi-label learning* », Pattern recognition ,2007.
[3] GLASSNER Andrew : op. cit., p.417.

Dans la figure n°5(b), nous avons fixé k à 1, ce qui signifie que nous voulons utiliser la classe de l'observation la plus proche. Dans ce cas, il s'agit d'un cercle rouge, donc l'étoile est classée comme un cercle. Dans la figure n°5(c), nous avons fixé k à 9, ce qui signifie que nous examinons les neuf points les plus proches. Nous trouvons ici 3 cercles, 4 carrés et 2 triangles. Comme il y a plus de carrés que toute autre classe, l'étoile est classée comme un carré. Dans la figure n°5(d), nous avons fixé k à 25. Nous avons maintenant 6 cercles, 13 carrés et 6 triangles, et l'étoile est à nouveau classée comme un carré. Plus deux points sont proches l'un de l'autre, plus ils sont similaires et vice versa, le fonctionnement peut être assimilé à l'analogie suivante « dis-moi qui sont tes voisins, je te dirais qui tu es ».[1]

Pour effectuer une prédiction, l'algorithme KNN n'a pas besoin de construire un modèle prédictif et va se baser sur le jeu de données en entier en calculant la similarité entre une observation en entrée et les différentes observations du jeu de données. Ainsi, il n'existe pas de phase d'apprentissage proprement dite, c'est pour cela qu'on le catégorise comme un algorithme paresseux. Le contre coût est qu'il doit garder en mémoire l'ensemble des observations pour pouvoir effectuer sa prédiction ce qui peut ralentir l'algorithme de manière significative, ainsi qu'augmentais le cout de l'opération.[2]

Également, le choix de la méthode de calcul de la distance ainsi que l'hyper paramètrent K (le nombre de voisins) peut ne pas être évident. Il n'y a pas de méthode officielle, Il faut essayer plusieurs combinaisons et faire des validations croisées pour avoir un résultat satisfaisant en fonction des types de données qu'on manipule.[3]

Un autre problème avec cette technique est qu'elle dépend de la présence de nombreux voisins à proximité. Cela signifie que nous avons besoin de beaucoup de données d'apprentissage.

C. Arbre de décision (DT)

Un arbre de décision est une suite d'algorithmes d'apprentissage se basant sur une représentation visuelle de classification de données suivant différents critères qu'on

[1] GLASSNER Andrew : op. cit., p.418.
[2] Ibid., p.422.
[3] Ibid., p.424.

appellera décisions (ou nœuds) placées dans les feuilles.[1] L'ensemble des nœuds se divise en trois catégories :

- Nœud racine : l'accès à l'arbre se fait par ce nœud.
- Nœuds internes : les nœuds qui ont des descendants, qui sont à leur tour des nœuds.
- Nœuds terminaux (ou feuilles) : nœuds qui contiennent la classe.

Il est également courant d'utiliser les termes associés aux arbres familiaux, Chaque nœud (à l'exception de la racine) a un nœud au-dessus de lui. Nous appelons cela le parent de ce nœud. Les nœuds situés immédiatement sous un nœud parent sont ses enfants, les nœuds qui partagent le même parent immédiat sont appelés frères et sœurs.

Un nœud racine contient des observations de différentes classes, distinguées par leurs formes et leurs couleurs. Pour construire l'arbre de décision, nous divisons les observations de chaque nœud en deux groupes à l'aide d'un test quelconque (on gardera le test qui donnera le maximum de gain d'information), ce qui aboutit à une décision, l'objectif est de continuer à diviser chaque nœud jusqu'à ce qu'il ne reste plus que des observations d'une seule classe. À ce moment-là, nous déclarons que ce nœud est une feuille et nous arrêtons de le fractionner.[2]

Lorsqu'une nouvelle observation arrive, il suffit de commencer à la racine et de descendre dans l'arbre, en suivant la branche appropriée à chaque nœud en fonction du test jusqu'à atteindre la feuille, nous arrivons ainsi à la classe de cette observation.

Les arbres de décision ont pour avantage d'être simples à interpréter, d'être non paramétrique, et de nécessiter très peu de prétraitement de données. Parfois, ils sont utilisés, même lorsque les résultats sont inférieurs à ceux d'autres classificateurs, parce que leurs décisions sont transparentes, et faciles à comprendre.[3]

D. Forêts aléatoires (RF)

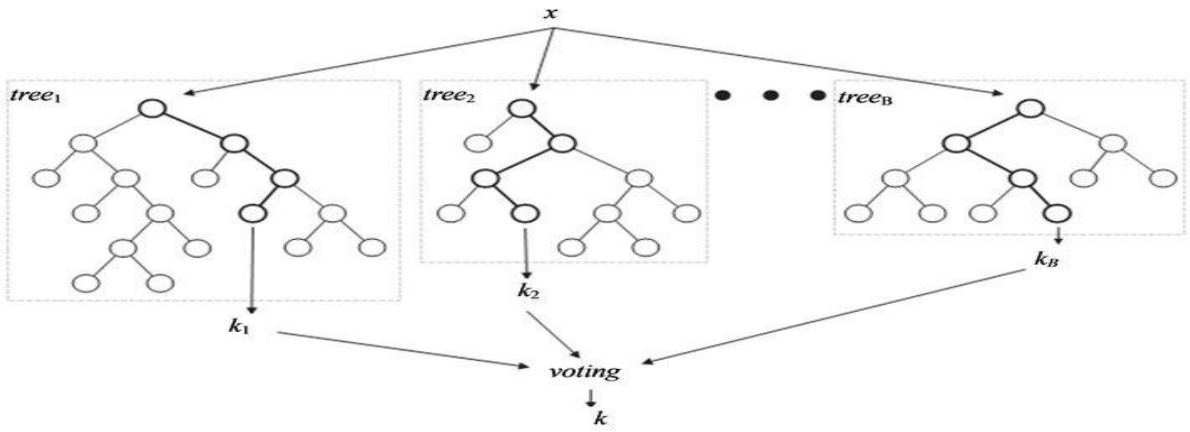
Comme évoqué dans l'arbre de décision, lorsqu'il est temps de diviser en deux un nœud, nous pouvons choisir n'importe quelle fonctionnalité (variable ou ensemble de fonctionnalités) pour créer le test qui dirige les éléments vers un enfant ou un autre. Nous

[1] GLASSNER Andrew : op. cit., p.426.
[2] Ibid., p.428.
[3] Ibid., p.428.

recherchons souvent le meilleur test en considérant chaque fonctionnalité. Mais nous pouvons également utiliser une technique appelée feature bagging. Avant de rechercher le meilleur test à un nœud, nous choisissons d'abord un sous-ensemble aléatoire de fonctionnalités à celui-ci. Nous sommes maintenant prêts à rechercher le meilleur test, basé uniquement sur ces fonctionnalités.[1]

Plus tard, lorsque nous passons au prochain nœud, nous choisissons à nouveau un tout nouveau sous-ensemble de fonctionnalités et déterminons notre nouvelle division en utilisant uniquement celles-ci. L'idée est illustrée à la figure n° 6.

Figure N° 6 : Représentation graphique d'une Foret Aléatoire



Source : https://topdata.news/how-the-random-forest-algorithm-works-in-machinelearning/ 18/04/2022

En ne choisissant au hasard que quelques-unes des caractéristiques, nous pouvons éviter de faire le même choix pour ce nœud dans chaque arbre que nous formons, et ainsi nous pouvons augmenter la diversité de nos décisions. Lorsque nous construisons des ensembles de cette manière, nous appelons le résultat une forêt aléatoire. La partie aléatoire du nom fait référence à notre choix aléatoire de caractéristiques à chaque nœud, et le mot forêt fait référence à la collection résultante d'arbres de décision.[2] Ce choix aléatoire du point de séparation nous permet de sacrifier un peu de précision pour réduire le surajustement.

[1] GLASSNER Andrew : op. cit., p.474.
[2] Ibid., p.475.

E. Réseau de neurones artificiels (ANN)

Les réseaux de neurones artificiels sont composés de couches et de nœuds. La première couche d'un réseau de neurones est appelée la couche d'entrée, suivie des couches cachées, puis enfin la couche de sortie. Chaque nœud est conçu pour se comporter de la même manière qu'un neurone dans le cerveau humain. Une fois qu'un neurone reçoit ses entrées des neurones de la couche précédente il additionne chaque signal multiplié par son poids correspondant et les transmet à une fonction d'activation.[1]

La fonction d'activation calcule la valeur de sortie pour le neurone. Cette valeur de sortie est ensuite transmise à la couche suivante du réseau. Les connexions entre les nœuds portent une valeur (poids), initialement, ces poids sont attribués de manière aléatoire, par la suite, ils sont adaptés pour s'aligner sur la valeur d'entrée pour donner la sortie souhaitée. C'est l'apprentissage.[2]

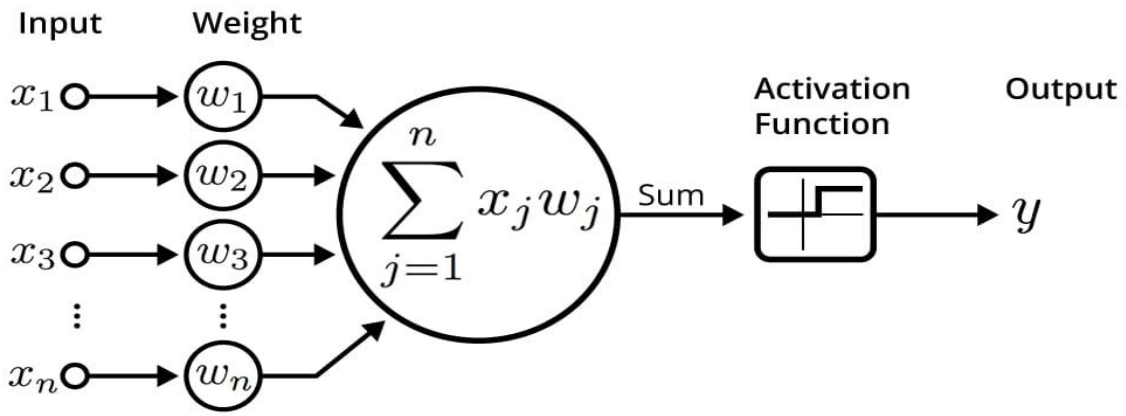
L'algorithme d'apprentissage le plus populaire en réseaux de neurones est appelé rétropropagation (Backpropagation). Fondamentalement, il vérifie et ajuste les poids de manière à obtenir le moins de perte possible.

Les poids ne sont pas ajustés de manière radicale, l'ajustement vers l'objectif souhaité se fait à travers des petites étapes. Ces dernières sont appelées le taux d'apprentissage.[3]

Voici une visualisation simplifiée pour illustrer comment cela fonctionne :

[1] KHEMAKHEM Sihem et BOUJELBENE Younes : « *Artificial intelligence for credit risk assessment : Artificial neural network and support vector machines* » ACRN Oxford, Journal of finance and risk perspectives ,2017.
[2] Ibid.
[3] GLASSNER Andrew : op. cit., p.494.

Figure N° 7 : Représentation graphique de ANN



Source : https://becominghuman.ai 05/04/2022

Un des avantages d'ANN est qu'il prend en charge l'apprentissage séquentiel, cela signifie que l'ANN peut se mettre à jour en permanence à mesure que de nouvelles données deviennent disponibles au fil du temps sans avoir à reformer l'ensemble du modèle à zéro. Par ailleurs, il fonctionne pour les 3 types d'apprentissages.

On distingue différents types de réseaux de neurones, ils sont classés selon le nombre d'épaisseurs qui séparent l'entrée de données du résultat, et en fonction du nombre de nœuds cachés du modèle : [1]

- **Réseau de neurones feed-forward :** les informations passent directement de l'entrée aux nœuds de traitement puis aux sorties sans retour en arrière de l'information.
- **Les réseaux de neurones récurrents :** ils traitent l'information en cycle et cela permet au réseau de traiter l'information plusieurs fois, donc ils sauvegardent les résultats produits par les nœuds de traitement et nourrissent le modèle à l'aide de ces résultats. Ce mode d'apprentissage est un peu plus complexe, il se compose d'une ou plusieurs couches.
- **Un réseau de neurones convolutif :** permet d'analyser des informations complexes et en grand nombre. Il est composé de plusieurs réseaux de neurones artificiels

[1] GLASSNER Andrew : op. cit., p.538.

distincts les uns des autres, qui prennent en charge chacun une partie de l'information. Il s'agit du Deep Learning.[1]

F. Classifieur Bayésien Naïf

C'est un modèle de classification paramétrique, probabiliste, suivant un algorithme d'apprentissage supervisé, il puise ses racines du célèbre théorème de Bayes. Ce dernier est un classique de la théorie des probabilités, il est fondé sur les probabilités conditionnelles :

$$P(A\backslash B) = \frac{P(B\backslash A) * P(A)}{P(B)}$$

P(A\B) : probabilité postérieure : est la probabilité conditionnelle d'un futur événement incertain A fondé sur des preuves pertinentes qui s'y rapportent historiquement a B_i. Cela suppose que A n'est pas indépendant de B, il s'est produit après l'événement original B, d'où le post dans postérieur. Ainsi, nous pouvons utiliser les résultats des données historiques pour fonder les croyances que nous utilisons pour tirer des probabilités nouvellement mises à jour.[2]

Le modèle a comme hypothèse l'indépendance entre les caractéristiques (variables), cette hypothèse naïve sur les caractéristiques est la raison pour laquelle cet algorithme est appelé naïf.[3] Il existe une deuxième hypothèse sur la distribution normale des caractéristiques du modèle. Naïf Bayes est populaire parce que cette hypothèse s'avère correcte, ou presque correcte, assez souvent.[4] Ainsi nous supposons que chaque caractéristique suit une distribution gaussienne, si nos caractéristiques suivent réellement des distributions gaussiennes, alors cette hypothèse produit un bon ajustement.

En général, les algorithmes de Bayes font souvent un bon travail sur tous les types de données. Cela s'explique probablement par le fait que de nombreuses données du monde réel proviennent de processus qui sont bien décrits par les gaussiennes.[5]

2.1.2. Les modèles de régression

[1] Le Deep-Learning fait référence à l'apprentissage automatique utilisant des algorithmes de réseaux de neurones avec un nombre important de couches cachées. Il s'agit donc d'un type particulier d'algorithme de Machine Learning.
[2] GLASSNER Andrew : op. cit., p.458.
[3] MILANA Carlo et ASHTA Arvind : op. cit.
[4] CHOW Jacky C. K : « *Analysis of financial credit risk using machine learning* », Aston university, 2018.
[5] GLASSNER Andrew : op. cit., p.464.

Un modèle de régression se distingue d'un modèle de classification par la variable de sortie qui peut prendre des valeurs numériques continues.[1]

A. Régression Linéaire

La régression linéaire est une méthode statistique qui permet de réaliser des prédictions sur la base de valeurs existantes. À partir d'un algorithme d'apprentissage supervisé, une relation linéaire est établie entre une variable expliquée et une variable explicative.

La régression linéaire prend des entrées continues et produit une variable continue. Cela peut déjà être utile pour prévoir les difficultés financières d'une entreprise.

Le choix d'un modèle linéaire n'est généralement pas contraignant. D'une part, beaucoup de problèmes commerciaux importants peuvent être exprimés à l'aide d'une relation linéaire, et d'autre part, les modèles non linéaires complexes peuvent être linéarisés à l'aide du développement en série de Taylor du premier ordre.[2]

B. La régression logistique

La régression logistique est un modèle statistique classique de classification binaire. Cette technique est utilisée pour calculer la probabilité d'occurrences mutuellement exclusives d'une variable cible. Ainsi, nous proposons de l'estimer par la probabilité $P(x)$ car une probabilité est toujours comprise entre 0 et 1.[3]

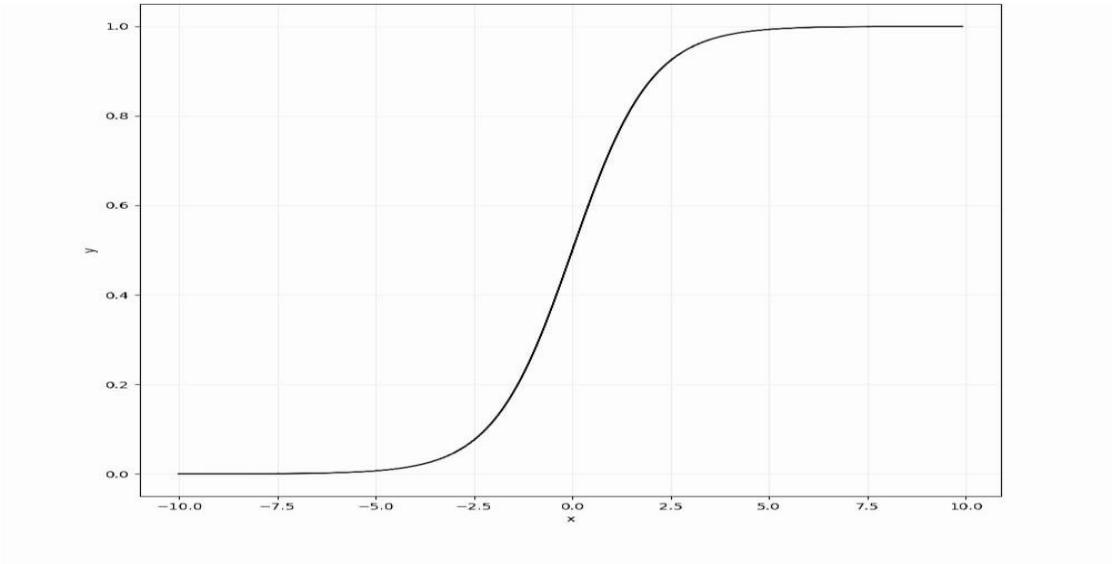
Dans ce cas, un modèle linéaire n'est pas bien adapté pour estimer la variable d'intérêt, car il peut en résulter une probabilité négative ou supérieur à 1.

Pour ce faire, la littérature propose que la probabilité $P(x)$ a une forme sigmoïdale. La fonction sigmoïdale peut être approché par différents modèles, elle peut ressembler à une fonction de distribution cumulative par exemple. Dans un modèle de RL, nous prenons la fonction de répartition d'un modèle logistique F_L.[4]

$$Y = P(x) = F_L(\beta_0 + \beta_1 x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} \text{ ==========> Modèle Logit}$$

[1] Ibid., p.465.
[2] RUSSOLILLO Giorgio : « *Régression logistique* », CNAM, 2018.
[3] Supposons le cas d'une régression logistique simple avec une seule variable explicative x
[4] MARAZZI Alfio : « *Introduction à la Clogistique* », IUMSP, 1989.

Figure N° 8 : Courbe Sigmoïde



Source : https://datascience.eu/fr/mathematiques-et-statistiques/regression-logistique/ 18/04/2021.

La transformation inverse, la fonction bijective, du modèle Logit nous permet de transformer la sigmoïde en fonction linéaire (afin d'estimer les coefficients et évaluer le modèle), elle se définie de]0; 1[→ R par :

$$F_L^{-1} = \ln(y/(1-y)) = \beta_0 + \beta_1 x$$

Rappelons-nous que l'objectif est de trouver la valeur de Y, d'où la valeur de P (Y =1/X=x_i). Pour ce faire, il suffit d'estimer les paramètres du modèle. Cela se fait à partir du modèle linéaire F_L^{-1} par le moyen de la méthode de vraisemblance.[1] Il s'agit du logarithme népérien des odds ratios : [2]

$$\begin{cases} odds\ ratio = o(1,0) = \exp(\beta_1) \\ \beta_1 = \ln\left(o(1,0) = logit\big(p(1)\big) - logit(p(0)\right) \end{cases}$$

La statistique LR est approximativement une distribution de khi-deux avec k-h+1 degrés de liberté ou l'hypothèse nulle signifie la nullité de tous les coefficients du modèle. Ainsi, la vraisemblance du modèle complet est comparée à la vraisemblance du modèle réduit.[3]

$$LR = -2\ln\left(\frac{vraisemblance\ du\ modèle\ courant}{vraisemblance\ du\ modèle\ saturé}\right)$$

[1] BOURBONNAIS Régis : « *econometrie cours et exercices corrigés* », Dunod,2014.
[2] Les chances, communément appelées odds, sont la proportion de la réalisation de Y=1 sachant que X=x_i
[3] BOURBONNAIS Régis : op. cit.

Tandis que l'ajustement du modèle se mesure par la statistique McFadden's appeler également Pseudo $-R^2$

$$\text{Pseudo} - R^2 = 1 - \frac{\text{Ln}(vraisemblance\ du\ modèle\ saturé)}{\text{Ln}(vraisemblance\ du\ modèle\ courant)}$$

- **La fonction d'activation Sigmoïde**

Il s'agit d'une fonction qui peut convertir n'importe quelle valeur donnée en probabilité comprise entre 0 et 1. cette transformation n'est pas linéaire, elle prend la forme S pour capturer tous les points de données. Sigmoïde est centrée sur la valeur 0.5. Ainsi, toute valeur supérieure à 0.5 peut être classée comme 1 et vice versa (ce seuil peut changer en fonction des données/du cas d'utilisation commerciale).[1] Cette fonction est également utilisée comme fonction d'activation dans divers modèles de ML tel que les réseaux de neurones.

Une étude comparative des points forts et faibles de ces différentes méthodes d'apprentissage automatique supervisé est présentée dans l'annexe n° 1.

2.2. Apprentissage non supervisé

Un algorithme d'apprentissage automatique non supervisé utilise des données dont les sorties sont inconnues, et sélectionne lui-même les classes qui lui semblent les plus judicieuses (clustering). Ce qui permet une découverte des tendances, des réponses et des distributions cachées.[2]

Il existe plusieurs méthodes d'apprentissage non supervisé tel que : Clustering Kmeans, classification hiérarchique, DBSCAN, …[3]

2.3. Apprentissage par renforcement

En général, un agent d'apprentissage par renforcement apprend à atteindre un objectif dans un environnement incertain et complexe. Dans ce cas, une intelligence artificielle est confrontée à une situation de jeu. L'ordinateur procède à des essais et des erreurs pour

[1] RUSSOLILLO Giorgio : op. cit.
[2] MHLANGA David : op. cit.
[3] HAMDAD Leila : « *Introduction au machine learning* » Ecole supérieure d'informatique ,2021.

trouver une solution au problème par un système de récompense comme régule de jeu son objectif est de maximiser la récompense totale.[1]

3. La différence fondamentale entre le Machine Learning et les modèles statistiques

La quasi-totalité de l'apprentissage automatique est une implémentation automatisée de méthodes statistique classiques. "L'apprentissage" est juste un ajustement de courbes fantaisistes.[2] Tandis que la statistique insiste sur une méthodologie appropriée et rigoureuse et est encadrée pas des hypothèses. De plus, elle se soucie de la manière dont les données sont collectées, et des relations entre les variables. L'apprentissage automatique concerne la prédiction. Il s'intéresse à l'utilisation des prédictions dans la prise de décision. Le ML ne trouve pas d'inconvénient a traité les algorithmes comme une boîte noire tant qu'il fonctionne. La prédiction, la performance et la prise de décision sont reines, et l'algorithme n'est qu'un moyen pour y parvenir. Il est très important en ML de s'assurer que la performance du modèle s'améliore.

Quand l'objectif des statistiques global est d'arriver à de nouvelles connaissances scientifiques basées sur les données. L'objectif du Machine Learning, est de résoudre une tâche de calcul complexe en "laissant la machine apprendre". Au lieu d'essayer de comprendre suffisamment bien le problème. Pour ce faire, la collecte d'un maximum de donnée doit être faite. Souvent, les algorithmes d'apprentissage sont de nature statistique. Mais tant que la prédiction fonctionne bien, aucune sorte d'aperçu statistique des données n'est nécessaire.[3]

De plus, les statistiques et l'apprentissage automatique sont pratiqués par deux communautés différentes, qui suivent des pratiques de publication différentes (conférences pour ML, revues pour Stat) et utilisent des lieux de publication différents (par exemple, JACM vs Ann. Stat), et sont approprié par des départements différents (CS/EE vs Stat/Math).[4]

---

[1] CUNNINGHAM Sally Jo « *Machine learning and statistics : A matter of perspective* » Hamilton, New Zealand : University of Waikato, Department of Computer Science, 1995.
[2] MILANA Carlo et ASHTA Arvind : op. cit.
[3] https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3 https://onlinestats.canr.udel.edu/machine-learning-vs-statistics/ 11/04/2022 à 15 :38
[4] www.quora.com/ 29/03/2022 à 11 : 59

Des modèles tel que les modèles RL et KNN sont des modèles statistiques applicable au ML dans un contexte classification/régression.[1] L'avantage est un degré d'automatisation plus élevé ainsi qu'une capacité de traitement des données de plus grande dimension à grande échelle avec une plus grande capacité de calcul.

[1] Il y a un débat sur l'appartenance de RL aux algorithmes du machine learning. Nous appartenons au courant qui valide cela.

Section 03 : L'intelligence artificielle dans le secteur bancaire

Après cinq décennies de transformation numérique, englobant la numérisation et la datafication, la finance est le segment le plus mondialisé, le plus numérisé et le plus documenté de l'économie mondiale.[1]

Le tableau suivant représente les mots clés les plus fréquemment utilisés dans un échantillon de 191 recherches et publications, liées à l'IA et la finance.

Tableau N° 2 : Les mots-clés les plus fréquemment utilisés dans la recherche

IA	79
Finance	21
Machine Learning	18
Prévision	11
Systèmes experts	10
Réseaux neuronaux artificiels	9
Big data	9
Systèmes d'aide à la décision	9
Réseau de neurones	9
Comptabilité	8
Prise de décision	8
Marchés financiers	6

Source : MILANA Carlo et ASHTA Arvind : « Artificial intelligence techniques in finance and financial markets : A survey of the literature », Strategic change, 2021.

1. Intelligence artificielle dans l'économie

L'IA est au cœur d'une nouvelle révolution industrielle qui touche tous les secteurs de l'économie. Sa contribution à la recherche et au développement est cruciale pour une croissance économique et sociale durable. Au sein de la modernisation numérique qu'on vit aujourd'hui, l'IA offre une innovation continue pour l'amélioration et le remplacement potentiel des tâches et activités humaines par un large éventail d'applications en finance, en soins de santé, industrie lourde, vente au détail, chaîne d'approvisionnement, logistique et services publics, système de recommandation, moteur de recherche, …[2]

[1] MILANA Carlo et ASHTA Arvind : op. cit.
[2] Ibid.

Dans le secteur de la santé, l'IA ouvre les portes d'une médecine personnalisée, adapté à chaque individu, permettant aux médecins de choisir le traitement le plus adapté. De l'élaboration de diagnostique à la prévention, ces outils viennent en premier lieu aider le corps médical à rendre l'expérience patient plus dynamique, personnalisée et collaborative. Dans un domaine qui vise également à améliorer l'expérience des ménages, l'IA fait ces preuves en marketing. Notons que le processus d'achat numérique est loin d'être linéaire. Comprendre le comportement du consommateur final, c'est donc être en mesure de mieux rejoindre sa clientèle cible et permettre son intégration dans l'activité, dans le mindset, de l'offreur. [1]

Dans un tout autre domaine, la géographie et le climat, l'IA a également fait ses preuves. Tandis que les méthodes traditionnelles utilisent des équations complexes et prévoient seulement entre six heures et deux semaines, les solutions IA font preuve de plus de précision. Ceci permet également de faire face aux catastrophes naturelles dues au changement climatique.[2] Entre l'art et la nature, dans la santé et la finance, il existe des expériences d'humeurs et de bien être complètement gérées par l'IA.[3]

2. Applications de l'IA dans le secteur bancaire

Comme indiqué précédemment, l'IA a émergé il y a plus de 60 ans et n'a cessé de croître et d'évoluer. Depuis quelques années, les entreprises ont décidé d'intégrer cette technologie dans leurs outils informatiques, entre autres le secteur bancaire, ce qui a permis de créer la rupture avec les banques traditionnelles. L'IA s'applique sur différents domaines du secteur bancaire, on peut citer :

- **Secteur des assurances**

  Si les données ont toujours été essentielles aux activités d'assurances, l'intelligence artificielle renforce encore leur valeur aux yeux des actuaires. Plusieurs solutions d'IA sont en effet susceptibles d'affiner les offres d'assurance, notamment en matière de segmentation client. L'IA serait ainsi employée pour mieux évaluer les risques des profils des clients et optimiser les systèmes de tarification.

[1] https://www.forbes.com/sites/forbesagencycouncil/2019/08/21/how-artificial-intelligence-is-transforming-digital-marketing/?sh=6163eef021e1 le 04/05/2022 à 13 : 13
[2] https://developer.nvidia.com/blog/global-ai-weather-forecaster-makes-predictions-in-seconds 30/04/2022 à 16 : 20
[3] Ibid.

- **Trading algorithmique**

  La première tâche du trading algorithmique consiste en la réduction des risques sur un marché à forte volatilité. La deuxième réside dans la réduction de l'asymétrie de l'information et objectiver les opérations d'investissement de trading.

- **Prévention de la fraude**

  En utilisant des systèmes modernes de protection contre la fraude basée sur le ML, les institutions financières réduisent considérablement les risques de manquer des transactions suspectes, des erreurs humaines et des cas de failles de sécurité. La précision accrue de ces algorithmes offre aux entreprises financières une large couverture vis-à-vis des risques divers.

- **La gestion du cyber risque**

  Les acteurs financiers, notamment bancaires, utilisent de plus en plus l'intelligence artificielle pour se prémunir contre les cyberattaques. Les données de sécurité sont très difficilement compréhensibles pour un humain, les données sont donc semi-structurées (sous format texte) mais peuvent être traitées avec efficacité par des algorithmes intelligents. Ces outils de codifications présentent l'avantage de s'adapter en temps réel sans demander une mise à jour par le fabricant de logiciel.

- **Paiements intelligent**

  Les consommateurs se sont habitués à la commodité d'acheter et de payer quand et où ils veulent. Les commerçants intelligents ont répondu à cette préférence en proposant des paiements multicanaux rationalisés. L'apprentissage automatique et l'IA amènent les paiements sur différents canaux tel que le paiement mobile, paiement par carte, …

- **Prévision des défaillances d'entreprises**

  Une application bien connue dans la prédiction de la faillite d'entreprise est celle d'Altman (1968) comme cité précédemment.[1]

- **Évaluation du risque de crédit**

  Les opérations de crédits sont des anticipations de revenus futurs comportant un risque que ces revenus ne se produisent pas ou que leurs remboursements à

[1] Voir page 25.

> l'échéance ne soient que partiels en cas de défaillance de l'emprunteur. Ainsi, la gestion de ce risque est une composante essentielle de l'activité bancaire.[1]

Des travaux empiriques récents visaient à vérifier le pouvoir prédictif des méthodes d'intelligence artificielle pour les problèmes de gestion et notation de ce risque. L'évaluation des crédits à court terme est l'une des applications les plus importantes des modèles de notation de crédit, elle a attiré l'attention des chercheurs au cours des dernières décennies.[2]

3. Évaluation des performances du ML dans la gestion des risques de crédit : Revue de la littérature

Dans le milieu des années 80, de nombreuses études académiques ont cherché à évaluer la performance prédictive des méthodes de ML par rapport aux méthodes plus classiques comme la régression logistique dans la gestion du risque de crédit.

Dans l'objectif de synthétiser les modèles modernes de ML, Thomas (2000) propose la première étude comparative des modèles de scoring par les modèles supervisés. L'auteur reporte le pourcentage de classification correcte de six méthodes dont : les arbres de décision, les réseaux de neurones, la régression logistique, la régression linéaire, ... Les résultats indiquent qu'aucune méthode ne domine les autres et que les différences entre les méthodes sont très faibles. Par ailleurs, une étude similaire réalisée par Baesens et al., (2003) à donner d'autres résultat. La publication propose une analyse de 17 algorithmes de classification à partir de huit ensembles de données sur des crédits de banques internationales. Les machines à vecteurs de support (SVM) et les réseaux de neurones ont affichés les meilleures performances prédictives avec des ASC de 66% et 91% respectivement.

Dans une autre étude sur le pouvoir prédictif des algorithmes ML dans la gestion du risque de crédit, Moscatelli et al. (2020) ont comparer la performance de classification de modèles traditionnels comme la régression logistique avec celle des modèles de ML. L'étude a révélé

[1] HUSSEIN Abdou et al. : op. cit.

[2] LOUZADA Francisco et al. : « *On the impact of disproportional samples in credit scoring models : An application to a Brazilian bank data* », Academia,1989.

que les modèles d'apprentissage automatique procurent des gains significatifs en puissance et en précision discriminatoires par rapport aux modèles plus classique.

En 2009, Yeh et Lien ont comparé le pouvoir discriminant de la probabilité de défaut de paiement entre six techniques d'exploration de données : LDA, RL, ANN, KNN, naïf bayes et DT. Les résultats ont montré que la prévision de la probabilité de défaut par l'ANN est la seule méthode qui pourrait être utilisée pour représenter la probabilité réelle de défaut. Par conséquent, cette technique devrait être utilisée pour déterminer les scores des clients plutôt que d'autres techniques d'exploration de données.

En Tunisie, Matoussi à analyser 1435 dossiers de crédit octroyés aux entreprises industrielles tunisiennes entre 2003 et 2006. Les modèles scoring, RL et ANN sont comparés. Les résultats montrent la supériorité des réseaux de neurones artificiels par rapport aux autres méthodes classiques avec un taux de bon classement global de 89.9%.[1]

Dans de plus petites dimensions, plusieurs éléments favorisent l'implémentation des techniques de scoring intelligent par les institutions bancaires lors de l'étude des crédits aux PME. Selon Berger et al. (2011), en 1998, 62 % des plus grandes banques américaines utiliser déjà le scoring pour faire l'évaluation des demandes de financement des petites entreprises.

En 2020, Wang et al., ont publié une étude sur le processus de décision d'octroi de crédit aux petites entreprises en utilisant des modèles classiques et des modèles supervisés d'apprentissage automatique, l'étude consiste à comparer les modèles : score- carte, RF, ANN ainsi que d'autres méthodes. Selon l'article, le modèle de forets aléatoires booster à présenter le plus de précision.

Dans leurs étude, Ampountolas et al., ont procédé à une étude comparative entre différentes méthodes d'apprentissage supervise sur un échantillon de 4450 dossiers de crédit d'exploitation sur une période de 4ans, en guinée. Plusieurs mesures de performances sont utilisées. Dans l'ensemble, les résultats ont montré que l'algorithme des arbres de décision était le plus efficace.

[1] MATOUSSI Hamadi et ABDELMOULA Aida Krichène : « *La prévision du risque de défaut dans les banques tunisiennes : Analyse comparative entre les méthodes linéaires classiques et les méthodes de l'intelligence artificielle : les réseaux de neurones artificiels* », Crises et nouvelles problématiques de la valeur, 2010.

De nos jours, les solutions de l'IA s'activent à changer l'expérience client et diversifier l'offre de produits bancaires. Avec cette technologie, les banques auront plus de moyens pour se concentrer sur leurs clients et leur offrir plus de proximité. De grandes sociétés bancaires comme la BNP Paribas réfléchissent actuellement à la mise en place d'assistants virtuels pour mieux répondre aux besoins des clients, les assister dans leurs démarches et les orienter vers des produits mieux adaptés à leurs besoins.[1]

[1] MILANA Carlo et ASHTA Arvind : op. cit.

Conclusion

La classification est un axe de recherche important dans le domaine de l'apprentissage automatique. Son objectif est d'attribuer des étiquettes de classe à un ensemble d'observations décrites par leurs fonctionnalités. En apprentissage supervisé, les entrées du modèle sont étiquetées. Le modèle est ensuite entrainé et évalué. Il a pour finalité l'amélioration de la performance des modèles.

Ce chapitre traite des méthodes traditionnelles telle que la régression logistique et la régression linéaire ainsi que des méthodes plus récentes comme SVM, KNN, Naïve Bayes, DT, RF et ANN.

Pour terminer, les applications les plus courantes dans le secteur bancaire sont présentées. Celles-ci abordent la gestion des risques de crédit d'exploitation par les PME.

Le chapitre suivant est une application empirique de quelques-unes de ses méthodes.

Chapitre 03 : Applications du Machine Learning dans le scoring des crédits d'exploitation destinés aux PME : Banque CPA

Introduction

L'apprentissage automatique est une boîte à outils puissante dans les problématiques de classification. À l'aide de données réelles, une comparaison des différents outils d'apprentissage supervisé : **RL, KNN et ANN** sera conduite. À travers une analyse binaire détaillée, 282 demandes de crédit d'exploitation auprès de l'organisme bancaire, le CPA, seront examinées et partagées en deux classes : défaillant et non défaillant.

Les méthodes de réduction de dimensionnalité et traitement des valeurs manquantes NA sont des méthodes de prétraitement populaire. Elles seront testées sur notre ensemble de données dans une approche expérimentale afin de sélectionner l'ensemble de variables avec lequel nous allons modéliser.

Les résultats analytiques ont révélé que les algorithmes d'apprentissage automatique sont capables d'être utilisés dans la modélisation du risque de crédit d'exploitation.

Le chapitre se présente comme suit :

La section 01 décrit la méthodologie expérimentale de la recherche, la section 02 est une analyse descriptive de l'ensemble de données. Enfin, la section 03 présente les résultats expérimentaux et les mesures utilisées pour évaluer la performance des classificateurs, ces résultats sont ensuite discutés et analysés.

Section 01 : Démarche méthodologique

Cette section présente la démarche adoptée dans cette recherche et énonce les outils de travail utilisés dans la notation de crédit. Bien qu'il n'y ait pas de meilleures techniques ou processus d'IA utilisées dans l'ensemble pour construire des modèles de notation, ce qui est le mieux dépendra du problème en question, de la structure des données, des variables utilisées et de l'objectif de la classification.[1]

1. Ensemble de données

Afin de réaliser une étude comparative en appliquant les algorithmes ML, des données secondaires ont été exploitées.

1.1. Présentation de la base de données

L'ensemble de données a été obtenu auprès de l'institution bancaire Algérienne « le CPA », il se compose de 282 demandes de crédit d'exploitation d'entreprise de taille PME entre la période 2015-2019.

La collecte de données a été réalisé comme suit : d'abord une base de données numérisée contenant des dossiers de crédit d'exploitation a été mise à notre disposition. Par la suite, un échantillonnage stratifié est conduit sur Excel. L'échantillon est composé de 80% de clients non défaillant et 20% de clients défaillant proportionnellement à la composition originale de la population.

La base de données comporte une multitude de variables financières et comptables. Nous avons synthétisé 26 ratios inspirés de la littérature dans le tableau N° 3.

1.2. Caractéristiques de l'ensemble de données

Les entreprises de l'échantillon utilisées dans cette étude sont des PME actives dans divers secteurs d'activités. Ainsi, les deux critères retenus lors de la formation de l'échantillon se présentent comme suit :

- Taille de l'entreprise : PME.
- Nature du crédit sollicité : crédit d'exploitation.

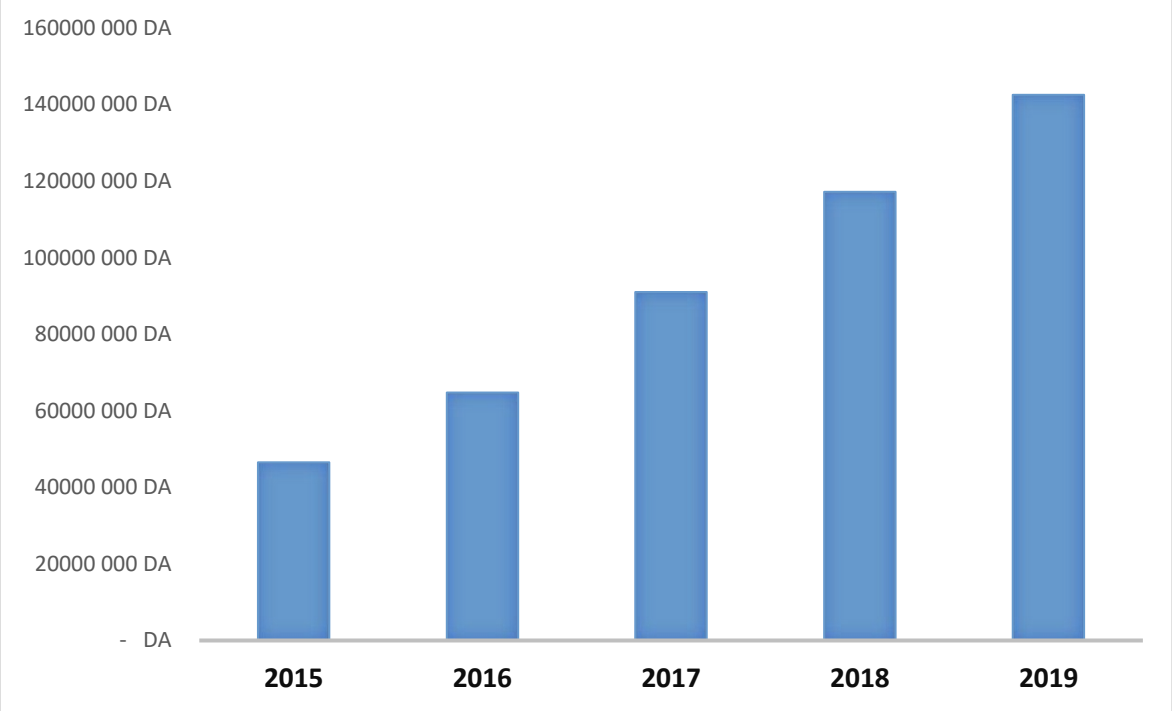
[1] MHLANGA David : op. cit.

L'échantillon se compose exclusivement de PME car elles sont généralement sous capitalisées, elles ne disposent principalement que d'une source externe de financement : la banque. Par ailleurs, elles font face à la contrainte de production de l'information, cette contrainte surgit à cause du caractère familial d'une grande partie de ces entreprises ainsi qu'au défaut d'accès aux marchés financiers producteurs d'informations publiques et le poids négligeable de ce dernier en Algérie.[1]

De plus, les PME ont une probabilité de défaillance nettement plus importante que les grandes entreprises.[2] Dans leurs articles « Estimation du risque de crédit et qualité de l'information comptable en Algérie » Gliz et Touati-Tliba (2012) annoncent : « le scoring est utilisé de façon intensive pour le crédit à la consommation et de plus en plus pour le crédit aux entreprises, en particulier celles de petite taille ».[3]

En effet, on constate une forte évolution des crédits d'exploitation octroyés par l'organisme bancaire, le CPA, entre la période 2015-2019.

Figure N° 9 : Évolution des crédits d'exploitation CPA



Source : Elaboré par les auteurs à l'aide d'une documentation interne du CPA.

[1] NAHMIAS Laurent : « *Impact économique des défaillances d'entreprise* », Bulletin de la banque de France n°137, 2005.
[2] Ibid.
[3] GLIZ Abdelkader et TOUATI-TLIBA Mohamed : « *Estimation du risque de crédit et qualité de l'information comptable en Algérie* », A. Gliz et M. Touati-Tliba, Les cahiers du cread n°98-99, 2012.

1.3. Critère de défaillance des entreprises

Les clients de la banque sont classés en catégories de risque par le système de notation comportementale de la banque.

Afin d'obtenir deux classes d'entreprise (entreprise saine et entreprise défaillante), nous avons arrêté comme critère de défaut tout retard de remboursement de 90 jours à partir de la date de survenance de l'incident de paiement, et ce conformément aux exigences du comité de Bâle et au règlement 2014-03 relatif au classement et au provisionnement des créances.[1]

2. Présentation des variables

Dans un premier lieu, 26 ratios économiques déduits d'un traitement comptable et 5 variables qualitatives sont retenus conformément à la littérature.

Au total, nous disposons de 31 variables explicatives résumer dans le tableau suivant :

[1] Voir chapitre 1, section 02, p.22.

Tableau N° 3 : Présentation des variables quantitatives

Ratio	Nom	Calcul	Code	Reference
Trésorerie	Valeur ajoutée	VA / CA	V1	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
	Fonds de roulement	(FR x 360) /CA	V2	CHOW Jacky C. K, 2017.BERRAIH Radia, 2020.
	Besoin de fonds de roulement	(BFR x 360) /CA	V3	BERRAIH Radia, 2020.
	Délai de règlement client	(Créances clients x 360) /CA	V4	ELHAMMMA Azzouz, 2011. BERRAIH Radia, 2020.
	Délai de règlement fournisseur	(Dettes fournisseurs x 360) /CA	V5	ELHAMMMA Azzouz, 2011. BERRAIH Radia, 2020.
Liquidité	Ratio de liquidité réduite	ACT-Stocks/DCT	V6	CHOW Jacky C.K, 2017. BERRAIH Radia, 2020.
	Disponibilité	Disponibilité (net) / Actif	V7	CHOW Jacky C. K, 2017. BERRAIH Radia, 2020.
	Ratio de liquidité immédiate	Disponibilité (net) / DCT	V8	ELHAMMMA Azzouz, 2011. BERRAIH Radia, 2020.
	Ratio de couverture	Charges financières / EBE	V9	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017.BERRAIH Radia, 2020.
	Dette à court terme	DCT/Total dette	V10	BERRAIH Radia, 2020.
	Rotation de stock	(Stock moyen de marchandises HT / coût d'achat des marchandises vendues HT) * 360	V11	CHOW Jacky C. K, 2017. BERRAIH Radia, 2020.
Rentabilité	ROA	Résultat net après impôt / Total actif	V12	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
	ROE	Résultat net / Fonds propres	V13	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
	Marge brute d'autofinancement	CAF / CA	V14	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
	Taux d'intégration	Valeur Ajoutée d'exploitation / CA	V15	BERRAIH Radia, 2020.
	Rentabilité opérationnelle	EBE / CA	V16	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.

Solvabilité	Marge opérationnelle	Résultat opérationnel / CA	V17	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
	Taux de marge financière	Résultat financier / CA	V18	CHOW Jacky C.K, 2017. BERRAIH Radia, 2020.
	Rentabilité commerciale avant impôt	Résultat net avant impôts / CA	V19	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
	Rentabilité commerciale après impôt	Résultat net après impôts / CA	V20	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
Structure	Solvabilité générale	Total dettes / Total Actif	V21	LOTFI Siham et MESK Hicham ,2020. BERRAIH Radia, 2020.
	Levier financier	Dette financière / Fonds propres	V22	BERRAIH Radia, 2020.
	Autonomie financière	Fonds propres / Passif Non courant	V23	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
	Ratio d'adéquation des fonds propres	Fonds propres / Total actif	V24	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.
	Capacité de remboursement	Dette financière / Résultat net	V25	BERRAIH Radia, 2020.
	Capacité de remboursement	DLT / CAF	V26	KHEMAKHEM Sihem et BOUJELBENE Younes, 2017. BERRAIH Radia, 2020.

Source : Elaboré par les auteurs.

Tableau N° 4 : Présentation des variables qualitatives

Variable	Modalité	Code	Référence
Forme juridique	SARL	1	BERRAIH Radia, 2020. KHEMAKHEM Sihem et BOUJELBENE Younes, 2017.
	EURL	2	
	SNC	3	
	SPA	4	
Centrale des risques	Néant	0	
	Existant	1	
Impayés confrères	Non	0	
	Oui	1	
Secteur d'activité	Industrie agroalimentaire	1	
	Industries manufacturières	2	
	Pharmaceutique	3	
	Hydrocarbure, Energie, Mines et Services liés	4	
	Service	5	
	Commerce	6	
	Import/Export	7	
	BTPH	8	
	Santé	9	
Mouvements confiés	Quasi intégral	1	
	Partiel	0	

Source : Elaboré par les auteurs.

2.1. Prétraitement de l'ensemble de données

Ce processus préalable à la formation permettra une qualité supérieure des données de formation. Par conséquent, un modèle plus apte à effectuer des classifications correctes.

2.1.1 Codification de variables qualitatives

Une variable qualitative réfère à une caractéristique qui n'est pas quantifiable. Cette donnée n'a pas le même type d'analyse statistique qu'une donnée numérique, elle n'a pas d'échelle d'intervalle standardisée. Il existe une variété de système de codage. Dans notre cas, nous avons fait correspondre un chiffre pour chaque variable qualitative. De cette façon, elles peuvent être considérées comme entrées dans nos modèles.

2.1.2. Valeurs manquantes (NA)

Les valeurs manquantes sont les observations manquantes de l'ensemble de données, leur total est de 12. La méthode de prétraitement choisie consiste en leur suppression.

2.1.3. Avis d'expert

La base de données a été présenté à un expert du domaine bancaire qui a donné des orientations, ainsi qu'une première sélection de variables explicatives. Ce qui nous a amené à supprimer les variables redondantes et peu significatives.

Par conséquent, 11 variables sont supprimées. Nous retenons donc 14 variables quantitatives (V2, V4, V5, V8, V9, V10, V11, V12, V13, V14, V15, V16, V21, V26) et 5 variables qualitatives (Forme juridique, centrale des risques, impayés confrère, mouvement confiés et activité).

2.1.4. Création d'un ensemble de données d'entraînement et de test

L'étape suivante consiste à répartir les observations en un ensemble de formation et un ensemble de test. Le premier ensemble (80 % de l'ensemble de données) est utilisé pour concevoir divers modèles et élaborer des règles d'affectation d'un individu en fonction de ses caractéristiques. Quant au deuxième (20 % de l'ensemble de données), il sert à vérifier si le modèle basé sur l'échantillon de formation est statistiquement fiable.[1]

3. Méthodes utilisées

Nous allons rappeler les différents aspects des trois méthodes utilisées dans notre étude ainsi que leurs principaux avantages.

3.1. Démarche méthodologique

Cette recherche va adopter une démarche expérimentale, du fait que plusieurs modèles seront testés et discutés et les meilleurs seront retenus pour la comparaison (tableau n°11).

3.1.1. Méthodes de classification

Cette étape concerne la sélection d'un algorithme d'apprentissage automatique supervisé et l'adaptation de nos données d'entraînement à celui-ci.[2] Cette analyse prédictive aura recours à différents algorithmes, et elle sera réalisée sur R (résultats dans la section 03).

[1] ADDO Peter et al. : « *Credit risk analysis using machine and deep learning models* », Maison des sciences économiques,2018.
[2] Ces modèles sont détaillés dans le chapitre 02.

Il est toujours conseillé d'avoir une variété d'algorithmes pour les différents problèmes, le même algorithme ne sera pas toujours le meilleur.[1]

A. RL

La régression logistique permet d'étudier la relation entre une variable qualitative dépendante et des variables explicatives indépendantes à travers une transformation sigmoïde des probabilités.

La régression logistique sera utilisée en raison de ses aspects intéressants tels que la performance, le faible cout computationnel, la documentation, le background littéraire, la rapidité et surtout la transparence.

B. KNN

L'algorithme KNN repose sur le calcul des distances entre des paires d'observations, il est utilisé dans les problèmes de classification où les données de formation sont disponibles avec des valeurs cibles connues. La valeur de k indique le nombre de voisins les plus proches, une façon de le trouver consiste à tracer un graphique contenant différentes valeurs de k et leurs taux d'erreur correspondant (figure n°28).

Parmi les motivations du choix de ce modèle, nous avons notamment la simplicité de son exécution et de son interprétation. De plus, il est libre de toutes hypothèses de départ. Ainsi, l'avantage majeur de cette approche est qu'il n'est pas nécessaire d'établir un modèle prédictif avant la classification, d'où sa désignation de modèle paresseux.

C. ANN

Les réseaux de neurones artificiels utilisent des équations pour développer successivement (d'une couche à une autres) des relations significatives entre les variables d'entrée et de sortie à travers un processus d'apprentissage. Nous avons appliqué des réseaux de rétropropagation pour classer les données.

Les ANN peuvent facilement gérer les effets non linéaires et interactifs entre les variables. Ils seront utilisés en raison de leur nature numérique, de l'absence de toute exigence concernant les hypothèses de distribution des données (pour les entrées) et leurs capacités de mettre à jour les données sans retraiter les anciennes.

[1] MHLANGA David : op. cit

4. Méthode de réduction de dimensionnalité

La classification avec seulement une, deux, ou même trois caractéristiques dimensionnelles est généralement intuitive, parce que les données de formation peuvent être visualisées et interprétées. Or, l'information sur la santé financière des entreprises se trouve dans un espace à dimensions plus élevées.[1]

Dans notre ensemble de données, nous avons 14 mesures comptables quantitatives. L'élimination de certaines de ces variables peut améliorer la performance des modèles de prédiction. Afin d'avoir une meilleure visibilité et interprétation du modèle de régression logistique, une réduction de dimensionnalités par la méthode ACP sera conduite.

4.1. Analyse en composantes principales

Le principe de l'ACP est de réduire la dimension des données initiales (qui est (p) si l'on considère p variables quantitatives), en remplaçant ces variables initiales par (q) vecteurs appropries (q < p), dit composantes principales. Les (q) vecteurs cherchés sont des combinaisons linéaires des variables initiales. Leur choix se fait en maximisant la dispersion des individus selon ces facteurs (variance maximum) dans le but de synthétiser l'information et créer plus de visibilité. Ainsi, chaque nouvelle caractéristique après l'ACP est une projection linéaire de nombreuses caractéristiques. Un inconvénient notable de l'ACP est qu'une partie importante de l'information pourrait être perdue.

5. Mesure de performance

C'est une étape très importante dans l'analyse prédictive. Il s'agit de la vérification de l'exactitude des valeurs prédites dans l'échantillon test pour savoir si elles correspondent aux valeurs connues dans l'échantillon d'apprentissage. Il existe plusieurs mesures pour comparer la performance des modèles, notamment ASC, Gini, fonction logLost, matrice de confusion et RMSE. Dans ce mémoire, nous mettrons d'avantage l'accent sur les critères de la matrice de confusion et la courbe ROC/ASC.

[1] ZAKI H et al. : « Méthodologie générale d'une étude ACP : Généralités, concepts et exemples », Revue interdiscplinaire, volume 1, 2016.

5.1. Matrice de confusion

Une matrice de confusion est un moyen évaluateur de la performance d'un modèle. Elle illustre le nombre total de valeurs vrai positif, vrai négatif, faux positif et faux négatif. D'autres mesures telles la courbe ROC peuvent être dérivées d'une matrice de confusion.

Figure N° 10 : Matrice de confusion

		Classe Réelle	
		+	-
Classe Prédite	+	Vrai Positif	Faux Positif
	-	Faux Négatif	Vrai Négatif

Source : Elaboré par les auteurs.

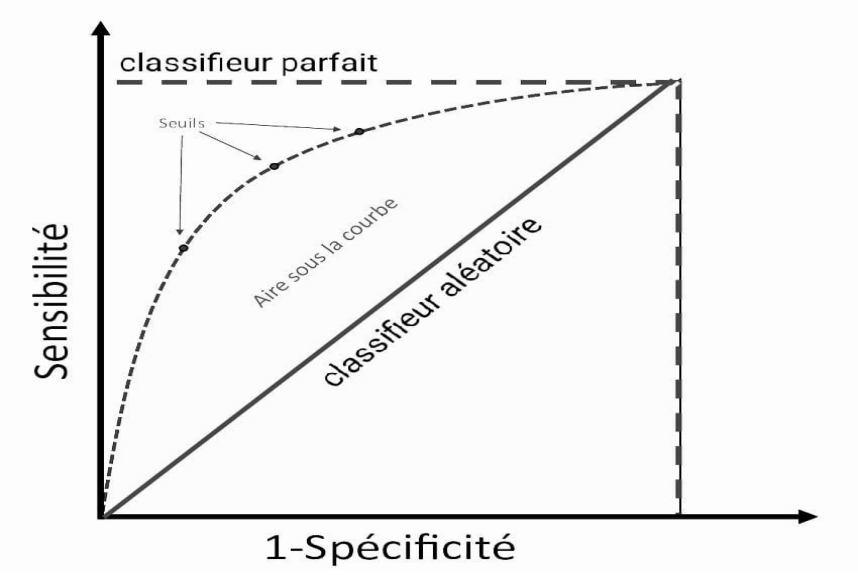
5.2. Courbe ROC

La courbe caractéristique de fonctionnement du récepteur (ROC) illustre la comparaison schématique entre le taux de vrais positifs et le taux de faux positifs à tous les seuils de classification.

$$\text{Sensibilité} = \text{le taux de vrais positifs} = \frac{\text{TruePositive}}{(\text{TruePositive} + \text{FalseNegative})}$$

$$\text{Spécificité} = \text{le taux de vrais negatif} = \frac{\text{TrueNegative}}{(\text{FalsePositive} + \text{TrueNegative})}$$

Figure N° 11 : Illustration de la courbe ROC



Source : https://www.semanticscholar.org/paper/Analyse-d%E2%80%99un-test-diagnostique-%3A-courbe-ROC%2C-ou-%C2%AB-%C2%BB-Perneger-Perrier/8c48ffc3e9868f6b143dc7d08cae104748ce02ab 11/04/2022

Pour effectuer la tâche de classement du modèle, une méthode courante consiste à calculer la zone sous la courbe ROC, abrégée en ASC.

5.3. ASC

Il s'agit d'une métrique importante qui permet d'analyser la performance d'un modèle. ASC décrit l'aire totale couverte sous la courbe ROC. Une valeur ASC plus élevée représente la supériorité d'un classificateur et vice - versa.

La performance globale de tous les classificateurs sera évaluée lors d'une analyse comparative dans la section 03.

6. Outil de travail

R est un environnement intégré de manipulation de données, de calcul et de préparation de graphiques. Toutefois, ce n'est pas seulement un « autre » environnement statistique (comme SPSS ou SAS, par exemple), mais aussi un langage de programmation complet et autonome.

Il est développé depuis les années 90 par un groupe de volontaires de différents pays et par une large communauté d'utilisateurs et utilisatrices. C'est un logiciel libre, publié sous licence GNU GPL.

Le R est un langage particulièrement puissant pour les applications mathématiques et statistiques puisqu'il est précisément développé dans ce but. Parmi ses caractéristiques particulièrement intéressantes, on note :

- R est un logiciel gratuit et a code source ouvert.
- Dispose d'une documentation détaillée disponible sur le CRAN. En effet, il existe plusieurs communautés statistiques qui répondent aux questions et publient des tutoriaux.
- R ne nécessite pas de typage ni de déclaration obligatoire des variables.

6.1. À propos de RStudio

RStudio n'est pas à proprement parlé une interface graphique pour R, il s'agit plutôt d'un environnement de développement intégré, qui propose des outils facilitant l'usage de R au quotidien.

Section 02 : Analyse descriptive de l'ensemble de données

Dans cette section, nous allons procéder à une analyse descriptive des variables quantitatives et qualitatives retenues pour notre étude.

1.Les variables quantitatives

Les variables quantitatives représentent des ratios financiers et comptables. Ces variables sont jugées pertinentes dans l'explication de la situation financière des PME, elles sont extraites des bilans, des comptes de résultats et des tableaux de flux de trésorerie des différentes entreprises de l'échantillon de l'étude.

Tableau N° 5 : Statistiques descriptives

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
V2	282	-333.68	4961.357	-81497.766	-48.912	140.75	2730
V4	282	86.58	131.322	0	8.078	98.244	881.251
V5	282	168.663	730.568	0	16.649	138.75	10546.555
V8	282	0.443	1.576	0	0	0.162	16.412
V9	282	0.409	1.468	-4.34	0.028	0.333	13.736
V10	282	0.747	0.277	0	0.58	1	1
V11	276	9.267	31.639	0	1.118	4.98	330.53
V12	282	0.039	0.078	-0.178	0.005	0.046	0.484
V13	282	0.144	1.026	-5.983	0.019	0.175	14.942
V14	282	-0.028	0.466	-2.383	-0.095	0.044	4.919
V15	282	0.304	0.435	-0.106	0.149	0.391	6.896
V16	282	0.127	0.245	-2.548	0.065	0.196	1.168
V21	281	0.61	0.274	0	0.41	0.844	1.27
V26	278	0.263	54.8	-811.91	0	0.612	271.35

Source : Elaboré par les auteurs à l'aide du logiciel R.

Nous distinguons quatre groupes de ratios financiers : les ratios de trésorerie, les ratios de rentabilité, les ratios de liquidité et les ratios de structure.

On constate que les variables V2, V4 et V5 qui représentent le fonds de roulement, le délai client et le délai fournisseur sont très dispersées, ces variables sont regroupées dans l'indicateur de trésorerie. Dans cette même catégorie, on constate que le délai fournisseur a une durée moyenne de 168 jours qui est supérieur au délai client avec une durée moyenne de 86 jours.

Pour ce qui est de l'indicateur de liquidité, on remarque que la variable V11 : rotation de stock, se renouvelle en moyenne chaque 9 jours du CA. Nous constatons également que cette variable est très dispersée comparer aux autres variables du groupe. Elle s'étale de 0 jours jusqu'à 330jours.

L'indicateur de rentabilité affiche une rentabilité moyenne assez faible : V12 et V13. Dans ce même indicateur, les entreprises génèrent une CAF moyenne de -0,028, elles disposent ainsi d'une capacité de financement interne moyenne négative. Cela justifie le recours le financement par crédit.

Finalement, pour l'indicateur de structure : la solvabilité générale V21 indique que la banque peut récupérer en moyenne 60% du montant des crédits octroyés en prélevant de l'actif de ses créanciers. Tandis que V26 présente une forte dispersion.

2. Les variables qualitatives

Pour faciliter leurs interprétations, les variables qualitatives sont transformées en variables dichotomiques.

2.1. Forme juridique

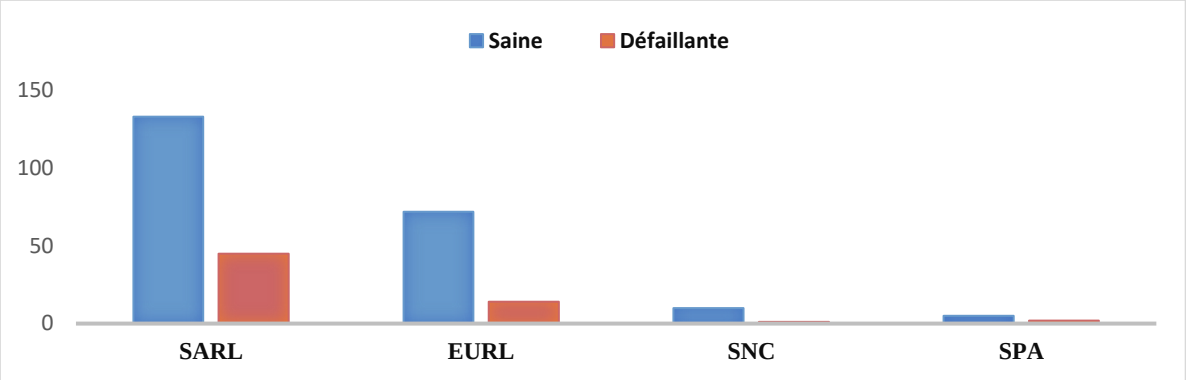
Nous remarquons d'après ce graphique que les entreprises de forme juridique : SARL occupent une place prépondérante dans notre échantillon, constitué auprès du crédit populaire d'Algérie, à hauteur de 63,12%. En deuxième position, on retrouve les entreprises de forme juridique EURL : 25,53%. Suivis en dernier, par les SNC et SPA respectivement. Dans ce qui suit, nous avons assigné une modalité à chaque forme juridique.

Tableau N° 6 : Les modalités de la variable forme juridique

La forme	Code
SARL : Société à responsabilité limitée	**1**
EURL : Entreprise unipersonnelle responsabilité limitée	**2**
SPA : Société par actions	**3**
SNC : Société par action	**4**

Source : Elaboré par les auteurs.

Figure N° 12 : Histogramme de la variable forme juridique



Source : Elaboré par les auteurs à l'aide du logiciel R.

2.2. Centrale des risques

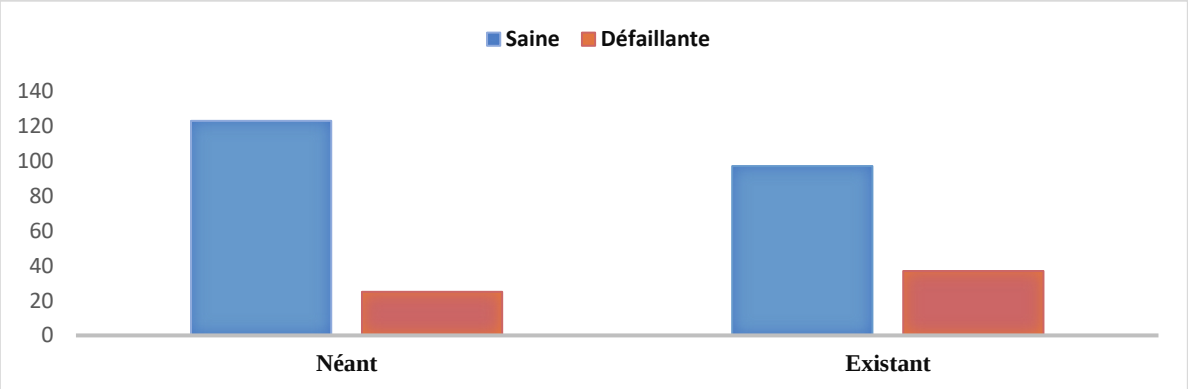
Cette variable nous permet de savoir si l'entreprise a déjà bénéficier d'un crédit auprès d'une autre institution financière ces dernières années. Elle prend les deux modalités présentées ci-dessous :

Tableau N° 7 : Les modalités de la variable centrale des risques

Centrale des risques	Code
Néant	**0**
Existant	**1**

Source : Elaboré par les auteurs.

Figure N° 13 : Histogramme de la variable centrale des risques



Source : Elaboré par les auteurs à l'aide du logiciel R.

2.3. Impayés confrères

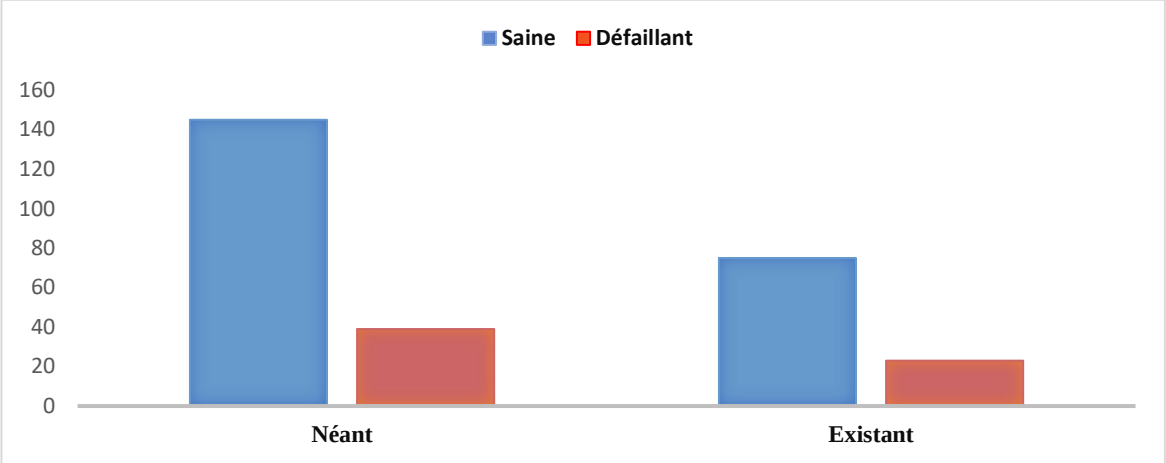
Cette variable nous permet de savoir si l'entreprise a enregistré un incident de paiement sur un crédit contracté chez une autre banque. Elle prend les modalités présentées ci-dessous :

Tableau N° 8 : Les modalités de la variable impayés confrères

Impayés confrères	Code
Néant	**0**
Existant	**1**

Source : Elaboré par les auteurs.

Figure N° 14 : Histogramme de la variable impayés confrères



Source : Elaboré par les auteurs à l'aide du logiciel R

2.4. Secteurs d'activités

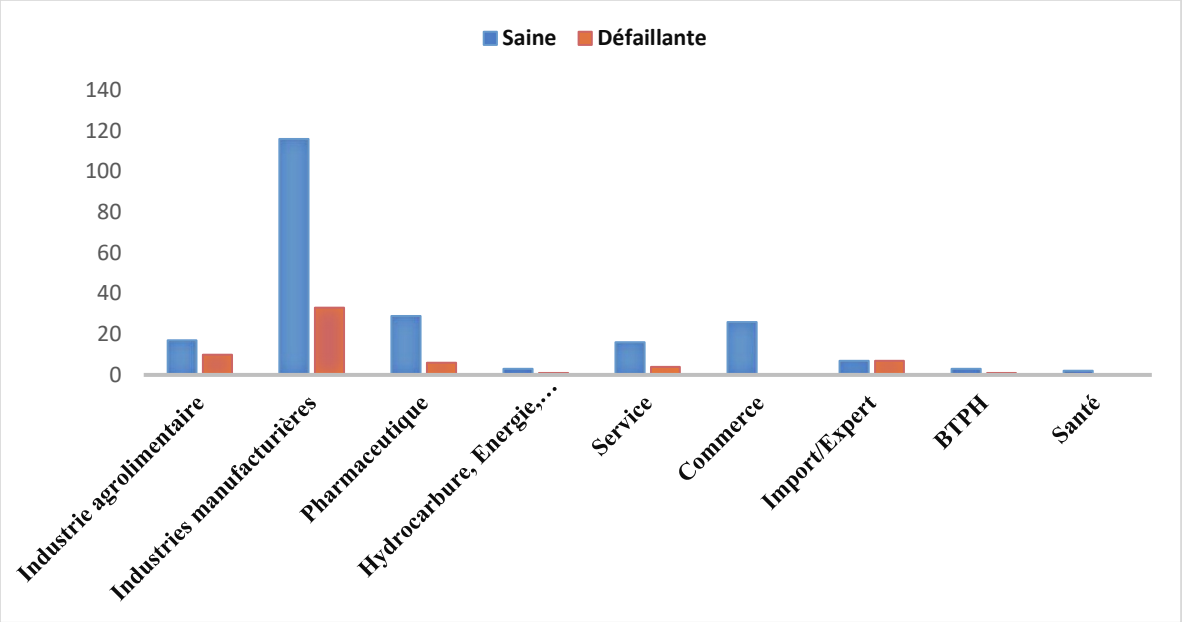
La base de données comporte des PME de différents secteurs d'activités. La variable Activité prend les modalités suivantes :

Tableau N° 9 : Les modalités de la variable activités

Les activités	Code
Industrie agroalimentaire	1
Industries manufacturières	2
Pharmaceutique	3
Hydrocarbure, Energie, Mines et Services liés	4
Service	5
Commerce	6
Import/Export	7
BTPH	8
Santé	9

Source : Elaboré par les auteurs.

Figure N° 15 : Histogramme de la variable activités



Source : Elaboré par les auteurs à l'aide du logiciel R.

2.5. Mouvements confiés

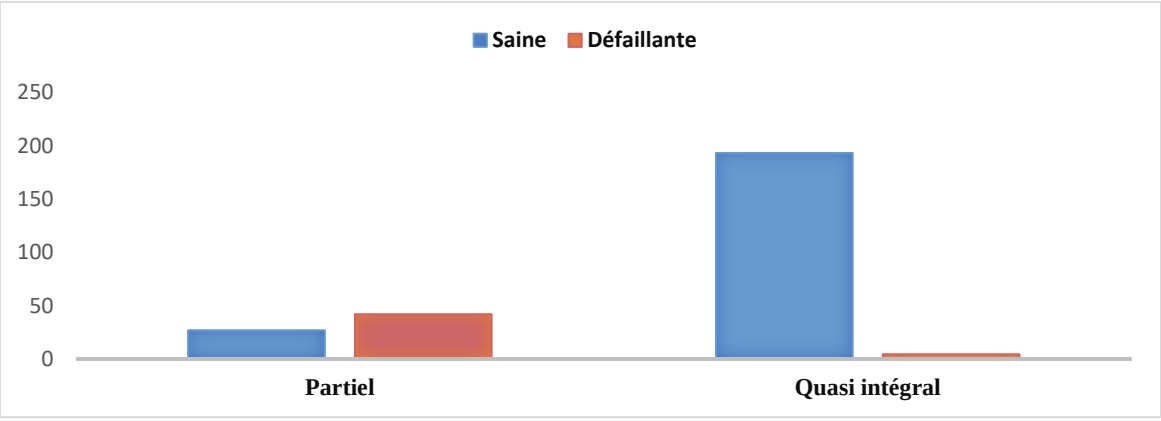
Cette variable vise à traquer les mouvements du compte d'un client par rapport à son chiffre d'affaires. Elle prend les modalités présentées ci-dessous :

Tableau N° 10 : Les modalités de la variable mouvements confiés

Mouvements confiés	Code
Partiel	**0**
Quasi intégral	**1**

Source : Elaboré par les auteurs.

Figure N° 16 : Histogramme de la variable mouvements confiés



Source : Elaboré par les auteurs à l'aide du logiciel R.

Section 03 : Résultats et Discussion

Dans cette section, nous présentons les résultats analytiques des différents modèles d'apprentissage automatique. Rappelons que l'objectif de la section est l'évaluation des diffèrent modèles ML dans la prévision des risques de crédit d'exploitation. Il s'agit des crédits octroyés aux PME par le CPA durant la période 2015-2019.

1.Préparation des données

Une étape préalable à la construction d'un modèle d'apprentissage automatique consiste à générer un ensemble de fonctionnalités adaptées à l'apprentissage du modèle. Cette tâche implique des processus de manipulation de données tels que la transformation des caractéristiques qualitatives, le traitement des valeurs manquantes, la détection des valeurs aberrantes, … un traitement correct peut améliorer fortement la qualité des données et des résultats**.**

1.1. Codification de variables qualitatives

L'encodage des variables qualitatives est l'une des principales étapes à réaliser avant d'alimenter tout modèle. Différentes techniques d'encodage peuvent être utilisées pour rendre ces données lisibles pour un algorithme d'apprentissage automatique. La façon la plus courante de gérer les catégories consiste simplement à mapper chaque catégorie avec un numéro, il s'agit du codage d'étiquette. Suite à une telle transformation, le modèle va traiter les catégories comme des entiers ordonnés. Voir tableau n°4.

Figure N° 17 : Codification des variables qualitatives

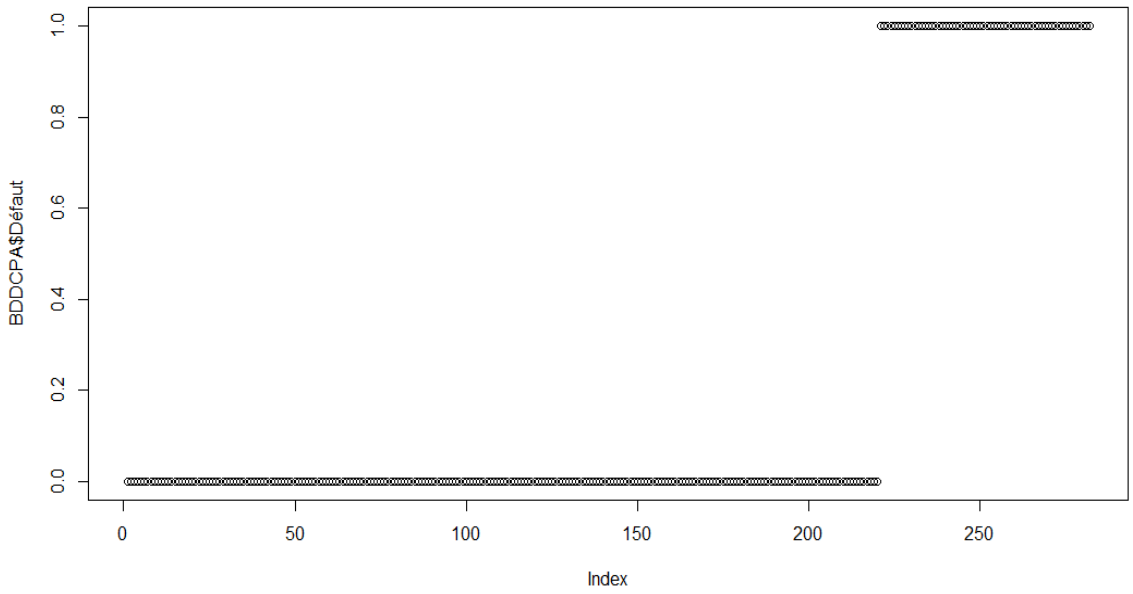
```
#codification des variables qualitatives
BDDCPA$FormeJuridique=factor(BDDCPA$FormeJuridique,level=c("SARL","EURL","SNC","SPA"),labels=c(1,2,3,4))
BDDCPA$centraleRisques=factor(BDDCPA$centraleRisques,level=c("Néant","Existant"),labels=c(0,1))
BDDCPA$ImpayésConfrères=factor(BDDCPA$ImpayésConfrères,level=c("Non","Oui"),labels=c(0,1))
BDDCPA$Activité=factor(BDDCPA$Activité,level=c("indusries Agroaliemntaires","industries manufacturières","Pharmacetique",
                                               "Hydrocarbures, Energie, Mines et services liés","Service","Commerce","Import/Export",
                                               "BTPH","Santé"),labels=c(1,2,3,4,5,6,7,8,9))
BDDCPA$MouvementsCconfiés=factor(BDDCPA$MouvementsCconfiés,level=c("Quasi intégral","Partiel"),labels=c(1,0))
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

Par ailleurs, le point de coupure sélectionné pour permettre de distinguer les deux groupes, qui est la variable cible, est intitulé défaut. Elle prend la valeur 0 si l'entreprise est saine et

1 si elle est défaillante. Soit les bons payeurs (BP) et les mauvais payeurs (MP). C'est la mesure adoptée par Altman (1968) ainsi que par divers autres auteurs.

Figure N° 18 : Distribution de la variable défaut



Source : Elaboré par les auteurs à l'aide du logiciel R.

1.2. Valeurs manquantes (NA)

Dans le même contexte, l'ensemble de données de travail est incomplet dans le sens où il manque des attributs pour certaines entreprises (12). Ces valeurs manquantes sont alors supprimées. De cette façon, le nombre d'observations diminue à 271 observations.

Figure N°19 : Les valeurs manquantes

```
> sum(is.na(BDDCPA))
[1] 12
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

1.3. Création d'un ensemble de données d'entraînement et de test

Une dernière étape avant d'appliquer le modèle consiste à partager l'ensemble de données en deux, 80 % pour l'ensemble de formation et 20 % pour l'ensemble de test. Les 20 % de données de test ne sont jamais utilisé dans aucune des étapes d'apprentissage automatique, sauf dans l'évaluation finale de la précision.

Figure N°20 : Script R - création d'un ensemble de données d'entraînement et de test

```
#Création d'un ensemble de données d'entraînement et de test :
sample <- sample(2,nrow(BDD5PC),
                 replace=T,
                 prob=c(0.8,0.2))
apprentissage <- BDD5PC[sample==1,]
test <- BDD5PC[sample==2,]
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

1.4. Sélection de l'ensemble de variables

Afin de sélectionner les caractéristiques de l'étude empirique, nous avons eu recours à l'avis d'un expert du secteur bancaire. Ainsi, le nombre de variables a été réduit de 31 à 19 variables explicatives, puis nous avons opté pour une réduction de dimensionnalité par la méthode ACP comme indiqué dans ce qui précède.

L'ensemble de variables qui sera utilisés par la suite est restreint à 19 variables, les motifs de ce choix seront discutés dans le tableau n° 11.

2. Modélisation

Dans ce qui suit, trois modèles de machine learning seront modélisés, mesurés et enfin discutés.

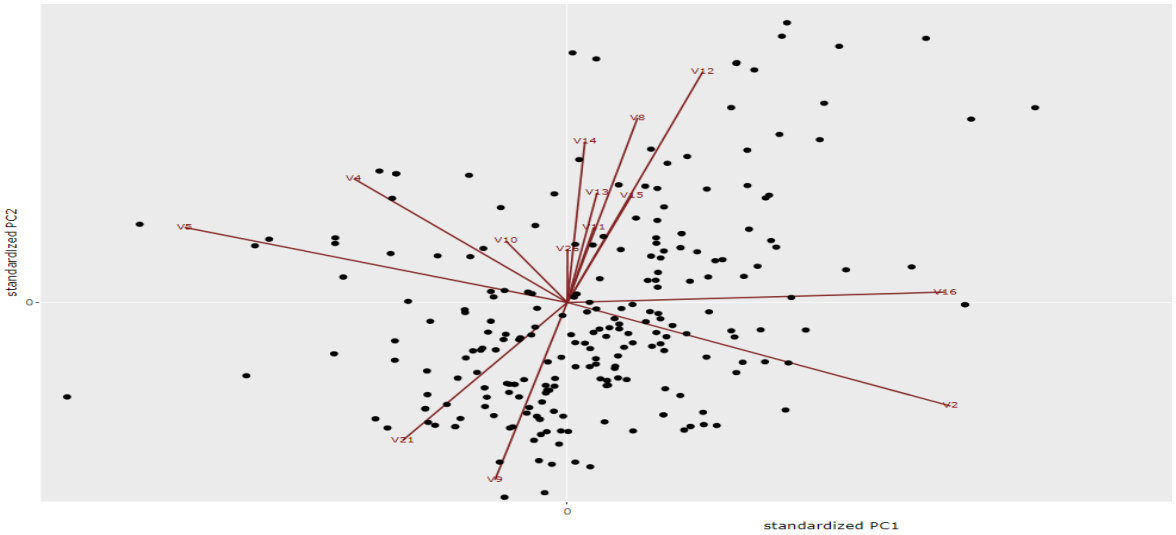
2.1. La régression logistique (RL)

Avec la création et le traitement des ensembles de données d'entraînement et de test, l'entraînement des modèles prédictif peut commencer.

2.1.1. Réduction des dimensionnalités (ACP) : afin de simplifier l'interprétation du modèle statistique RL et ses résultats, nous proposons la méthode ACP, dans laquelle les variables d'origine sont intégrées dans des vecteurs de plus faibles dimensions.

En effet, dans le cas de plusieurs dimensions, une évaluation visuelle de la classification n'est pas possible :

Figure N° 21 : Carte des vecteurs de variables

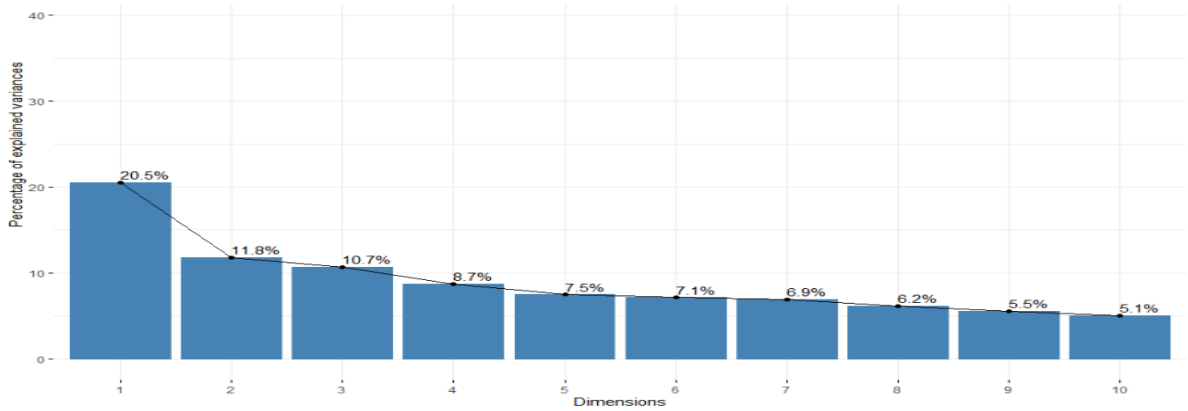


Source : Elaboré par les auteurs à l'aide du logiciel R.

Sur la figure n°20, la longueur des flèches en rouge indique la contribution de chaque variable à la première et la deuxième composante principale. Les variables corrélées positivement sont du même côté du graphique, quant aux variables corrélées négativement, elles se situent sur le côté opposé du graphique. Par ailleurs, une interprétation graphique permet de comprendre la structure des données analysées. Cette interprétation sera guidée par un certain nombre d'indicateurs numériques tel que les valeurs propres (Eigen vectors).[1]

Pour choisir un nombre approprié de vecteurs, les composantes principales sont tracées dans l'ordre décroissant selon le pourcentage de variance expliquée par chaque axe principal, comme illustré à la figure n° 22.

Figure N° 22 : Pourcentage de la variance expliquée



Source : Elaboré par les auteurs à l'aide du logiciel R.

[1] Voir l'annexe n°2

Visuellement, 32 % de la variance est résumée par le premier couplet de composantes principales, ce qui est très faible par rapport aux études antérieures. Cela peut être interprété par le fait que les variables soient peu corrélées dans l'espace d'origine. Chacune porte une information distincte.

Les composantes principales 10 à 14 ne représentent ensemble que 10 % de la variance dans l'ensemble de données. Par conséquent, il est raisonnable d'éliminer certaines dimensions moins informatives pour diminuer la complexité du problème.

Par conséquent, le modèle de régression logistique sera appliqué aux seules 5 plus grandes composantes principales, pour faciliter son interprétation et fonctionnement. Ce qui devraient nous laisser 59,20% de l'information tout en réduisant la dimensionnalité de 63%.[1] Cependant, un inconvénient surgit : 41% de l'information sera perdue, cela pourrait influencer négativement sur la performance des modèles.

En réalité, ces composantes principales n'ont pas beaucoup de sens pour un spécialiste financier. Une solution possible est d'étudier la fonction de projection et indirectement relier la structure de chaque vecteur principal aux variables originales.

En effet, chaque nouvelle caractéristique après l'ACP est une projection linéaire de nombreuses variables explicatives. Les poids des variables de chaque vecteur principal sont présentés dans l'annexe n°3. Une interprétation de ces résultats peut être résumé comme suit :

- PC1 : FR (26) + Délai fournisseur (26) + EBE (24) => **Vecteur de trésorerie**
- PC2 : ROA (22) => **vecteur/variable de rentabilité des actifs**
- PC3 : Délai client (11) + Rotation des stocks (2.28) + ROE (34) + VAExp (35) =>**Vecteur de rentabilité commerciale**
- PC4 : Disponibilité nette/ DCT (20) + total D/actif (16) =>**Vecteur de liquidité**
- PC5 :CF/EBE (12) + DCT/Total D (18) + CAF (21) + DLT/CAF (24) => **Vecteur de structure**

On peut observer que la 1ère composante est dominée par le fonds de roulement et le délai fournisseur .il s'agit donc d'un vecteur de trésorerie.

[1] Voir l'annexe n°2

La 2ème composante est facile à interpréter, du fait qu'elle n'a qu'une seule variable dominante qui est ROA, ainsi, il s'agit d'un vecteur/variables de rentabilité des actifs.

La 3ème composante exprime principalement la valeur ajoutée d'exploitation et ROE, ainsi le caractère de rentabilité commerciale du vecteur.

La 4ème composante principale est donc un vecteur de liquidité, et finalement, la 5ème est un vecteur de structure.

2.1.2. Modélisation et résultat RL

La régression logistique est peut-être l'algorithme le plus couramment utilisé dans le secteur de la notation du crédit à court terme.[1] Pour sa modélisation sur R, la fonction glm est utilisée. Le modèle linéaire généralisé (GLM) est une généralisation de la régression linéaire ordinaire sur d'autres distributions d'erreurs autres qu'une distribution normale, comme la distribution binomiale dans le cas d'une régression logistique. C'est une fonction de base dans R, ainsi elle ne nécessite pas de package.[2]

Figure N° 23 : Script R - modélisation RL

```
LR <- glm(Défaut~FormeJuridique +centraleRisques+ImpayésConfrères+MouvementsCconfiés+Activité+
              V2+V4+V5+V8+V9+V10+V11+V12+V13+V14+V15+V16+V21+V26,
          data=apprentissage,
          family='binomial')
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

Où :

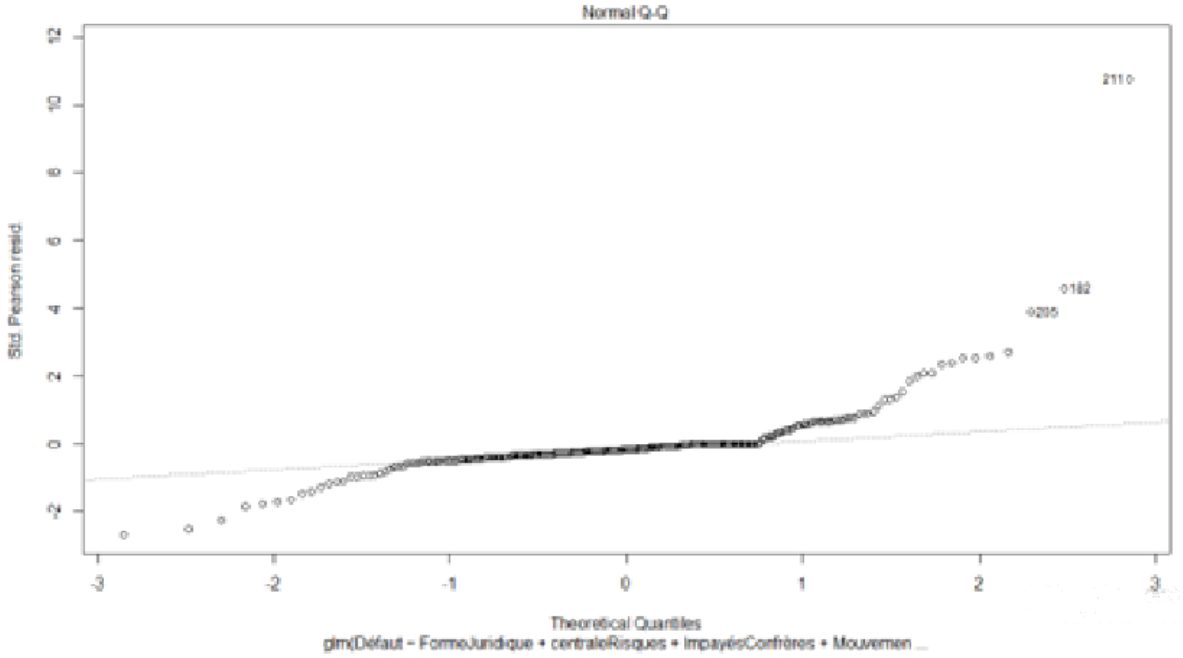
- **data** : l'ensemble d'apprentissage
- **family :** indique la distribution des erreurs du modèle. Elle peut prendre plusieurs distributions : poisson, binomial.

L'annexe n°4 est un tableau récapitulatif, il résume les résultats du modèle RL composé de 5 vecteurs principaux et 5 variables qualitatives.

[1] KENNEDY Kenneth : « *Credit scoring using machine learning* », Doctoral thesis, Technological University Dublin, 2013.
[2] Members of the R Core team, GLMR

La figure suivante représente le graphique du modèle de régression logistique. On constate qu'il se rapproche de la forme d'une fonction sigmoïde (S) :

Figure N° 24 : Modèle RL



Source : Elaboré par les auteurs à l'aide du logiciel R.

2.1.3. Interprétation du modèle

L'ensemble des variables sélectionnées étaient non significatives, à l'exception de trois variables et une constante : la variables mouvements confiés à le plus de signification statistique à hauteur de 0,1%, suivi par la constante à 1%, puis PC4 pour 5%, et finalement la variable Activité7 : import-export à 10%.

Or, en raison de leur importance potentielle (tableau n° 3) et de l'objectif du machine learning qui tente de montrer que le pouvoir prédictif du modèle est prépondérant aux statistiques,[1] nous les avons conservés dans le modèle.

La pertinence du vecteur de liquidité PC4 montre, par le signe de son coefficient, que ce vecteur influence bien la situation de l'entreprise et permet de différencier les deux catégories d'entreprises (saine et défaillante). La liquidité est inversement proportionnelle au risque de défaillance des crédits à court terme d'exploitation. Elle joue donc un rôle

[1] Voir page 49.

prépondérant dans l'appréciation de la situation financière actuelle et future d'un demandeur de crédit auprès du CPA. Ainsi, quand le coefficient de PC4 augmente, le Logit des odds (chances) de faire défaut diminue de 0,55.

La constante est très significative. Elle dispose d'un coefficient négatif, cela signifie que le score moyen d'un nouveau client est de -1,7 lorsque toutes les variables explicatives du modèle sont nulles. Cela a du sens à interpréter, car l'octroi de crédit est classiquement basé sur la confiance et le favoritisme. Un client peut avoir un crédit d'où, il peut être classé 0 : non défaillant, sans présenter un dossier de crédit à la banque (d'où la nullité des variables explicatives).

Pour ce qui est des variables qualitatives, la variable centrale des risques indique que si le demandeur de crédit figure dans la base de la banque centrale, il a déjà bénéficier d'un crédit au paravent. Tandis que la variables impayées confrère indique si le demandeur de crédit a des impayés de dette dans d'autres banques. Ces deux données sont disponibles au niveau de la banque centrale d'Algérie.

Malgré sa forte signification statistique, la variable mouvement confié ne correspond pas à nos attentes. En effet, quand celle-ci augmente elle signifie que l'entreprise dépose la quasi-totalité de son CA d'activité au CPA ou la banque d'accueil dans laquelle elle sollicite un crédit, si elle présente un coefficient positif le Logit(odds) du risque de crédit augmentera dans ce modèle. Cela peut être interprété par le fait que les fonds d'activité (CA) sont tellement faibles qu'ils ne parviennent pas à couvrir le crédit sollicité.

Les secteurs d'activités 6 et 9 sont les moins susceptibles de présenter un risque de non remboursement de crédit, à cause de leurs coefficients négatifs et importants, tandis que les secteurs 7 et 8 sont risqués.

Or, l'examen des résultats montre que les variables : forme juridique, centrale des risques et impayés confrères et les vecteurs PC1, PC2, PC3 et PC5 ne sont pas significatif au seuil de 10%, ce qui prouve qu'elles n'ont pas d'effet statistiquement significatif sur la classification des crédits d'exploitation dans cet ensemble de données.

Pour la validation statistique du modèle, la statistique Log vraisemblance (LR) est calculé :

$$LR = -2\ln\left(\frac{vraisemblance\ du\ modèle\ courant}{vraisemblance\ du\ modèle\ saturé}\right) = 254{,}20 - 161{,}59 = 92{,}61$$ [1]

Cette statistique suit une distribution khi-deux à 10 degrés de liberté, elle est donc comparé à la valeur tabulée du khi-deux :

$X^2_{10}(0{,}05) = 18{,}31 < 92{,}61$ -> Rejet de H_0 [2] → Modèle valide statistiquement.

La statistique McFadden's ou Pseudo$-R^2$ est une mesure de la capacité d'un modèle d'apprentissage automatique à se généraliser à des données similaires à celles sur lesquelles il a été formé. Un modèle bien ajusté produit des résultats plus précis. Un modèle sur-ajusté correspond trop étroitement aux données. Enfin, un modèle sous-ajusté ne correspond pas assez étroitement.[3]

$$\text{Pseudo} - R^2 = 1 - \frac{Ln(vraisemblance\ du\ modèle\ saturé)}{Ln(vraisemblance\ du\ modèle\ courant)} = 1 - \frac{161{,}59}{254{,}20} = 0{,}364$$

Une valeur Pseudo $- R^2$ entre 0.2 et 0,4 indique un excellent ajustement du modèle [4], cette statistique est de 0,364 dans notre RL, cela prouve l'ajustement du modèle.

La matrice de confusion et la précision de l'échantillon de validation se présentent comme suit :

Figure N° 25 : Matrice de confusion du modèle RL

```
> MatriceDeConfusion <- table(predicted=PPCATest,actual=testPCA$Défaut)
> MatriceDeConfusion
         actual
predicted  0  1
        0 43  6
        1  5  4
> Précision=sum(diag(MatriceDeConfusion))/sum(MatriceDeConfusion)
> Précision
[1] 0.8103448
>
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

On observe que le taux moyen de classification correcte –Accuracy- de l'échantillon test est de 81 % avec un seuil de 0,5 qui est le seuil par défaut de la matrice de confusion, ce qui

[1] Voir l'annexe n°4.

[2] H_0 : cette hypothèse suppose que le modèle composer uniquement de la constate est plus significatif que le modèle complet.

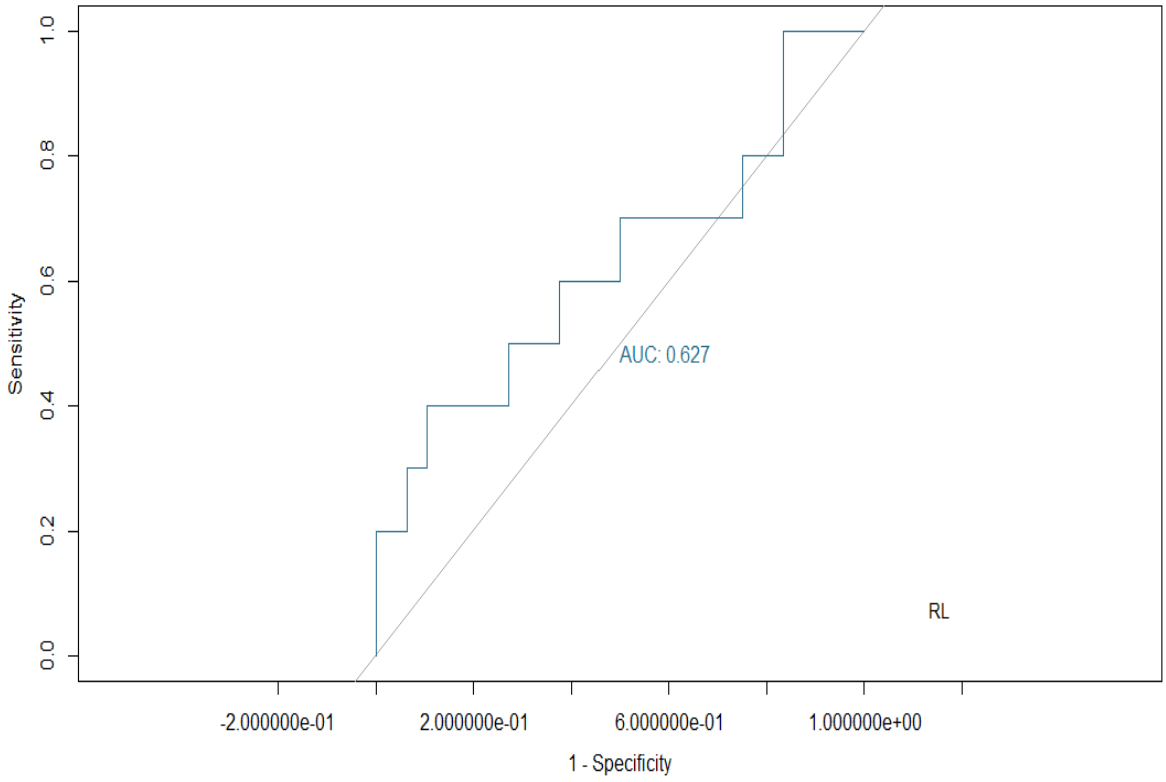
[3] https://www.datarobot.com/wiki/fitting/#:~:text=Model%20fitting%20is%20a%20measure,doesn't%20match%20closely%20enough, 13/05/2022 à 7 : 50

[4] https://stats.stackexchange.com/questions/82105/mcfaddens-pseudo-r2-interpretation, 11/05/2022 à 10 :45.

indique que si la probabilité de défaut est supérieure à 0,5, le client est classé défaillant, sinon, il est bon.

Afin de généraliser les seuils de classification, on a recours à la courbe ROC qui simule ses derniers sur une plage de seuils comprise entre 0 et 1. Ainsi, le pouvoir prédictif du modèle apparait dans la surface en dessous de la courbe : ASC, qui est dans ce cas 62,7%.

Figure N° 26 : Courbe ROC ASC du modèle RL



Source : Elaboré par les auteurs à l'aide du logiciel R.

2.2. K plus proches voisins (KNN)

Le classificateur des k plus proche voisins sert d'illustration d'une approche statistique basée sur le calcul des distances pour faire les prévisions. Pour sa réalisation, nous avons fait usage du package Caret du langage R.[1]

[1] KUHN Max : « *Building Predictive Models in R Using the caret package* », Journal of statistical software, 2008.

Figure N° 27 : Script R - modélisation KNN

```
KNN <- train(Défaut ~  EURL+SARL+SNC+SPA+CentraleRisq1+CentraleRisq2+MouvementsConfiés1+MouvementsConfiés2+impayésConfréres1+
                       impayésConfréres2+ActAgroAlim+ActManuf+ActPharma+ActEnergie+ActService+ActCommerce+ActImportExport+ActBTPH+
                       ActSanté+V2+V4+V5+V8+V9+V10+V11+V12+V13+V14+V15+V16+V21+V26,
             data = trainingKNN,
             tuneGrid   = expand.grid(k = 1:50),
             method = "knn",
             tuneLength = 20,
             metric     = "ROC",
             trControl = trControl)
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

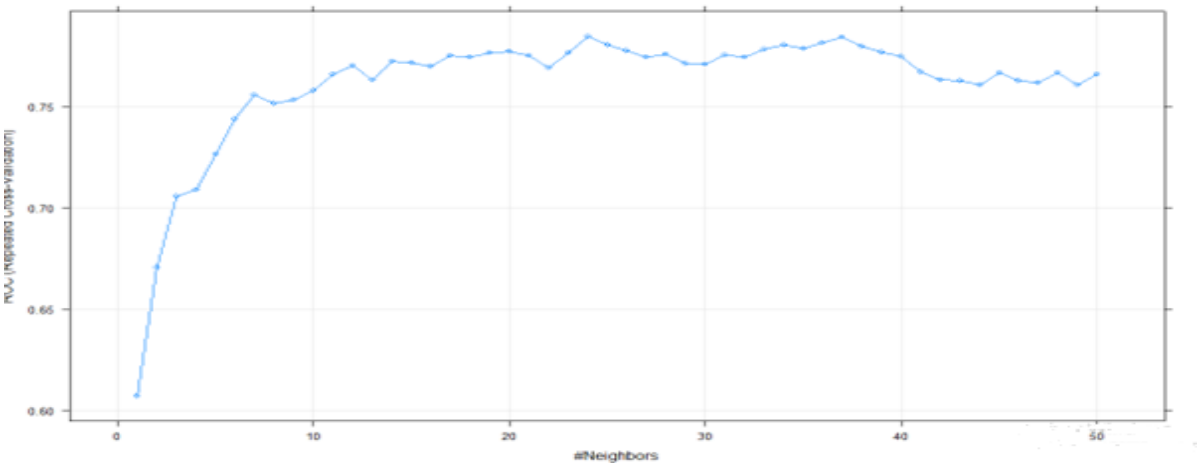
Où :

- **data :** l'ensemble d'apprentissage.
- **TrControl** : la fonction ajuste les paramètres de la phase d'apprentissage.
- **tuneGrid :** la fonction ajuste les valeurs de réglage possibles.
- **tuneLength :** Un entier indiquant la finesse de l'intervalle des valeurs du paramètre K dans la grille des paramètres de réglage.
- **method :** le modèle choisi.
- **metric :** Une chaîne qui spécifie qu'elle métrique récapitulative sera utilisée pour sélectionner le modèle optimal.

Le choix du paramètre k affecte la performance de l'algorithme. Celui-ci peut être déterminé expérimentalement. En partant de k=1, nous utilisons un test pour estimer le taux d'erreur du classificateur. Ce processus est répété en incrémentant le paramètre k. La valeur k qui donne le taux d'erreur minimum peut être sélectionnée. En général, plus la taille de l'échantillon d'apprentissage est grande, plus la valeur de k sera grande.

Le package propose une fonction qui automatise cette sélection par la commande : « metric ». Comme on peut le voir sur la figure n°28, le nombre optimal des k voisins semble être autour des 33 voisins les plus proches.

Figure N° 28 : Choix du paramètre k par optimisation ROC



Source : Elaboré par les auteurs à l'aide du logiciel R.

La matrice de confusion de l'échantillon de test atteint un taux d'erreur de 11 % ainsi qu'une précision de 74%. Elle se présente comme suit :

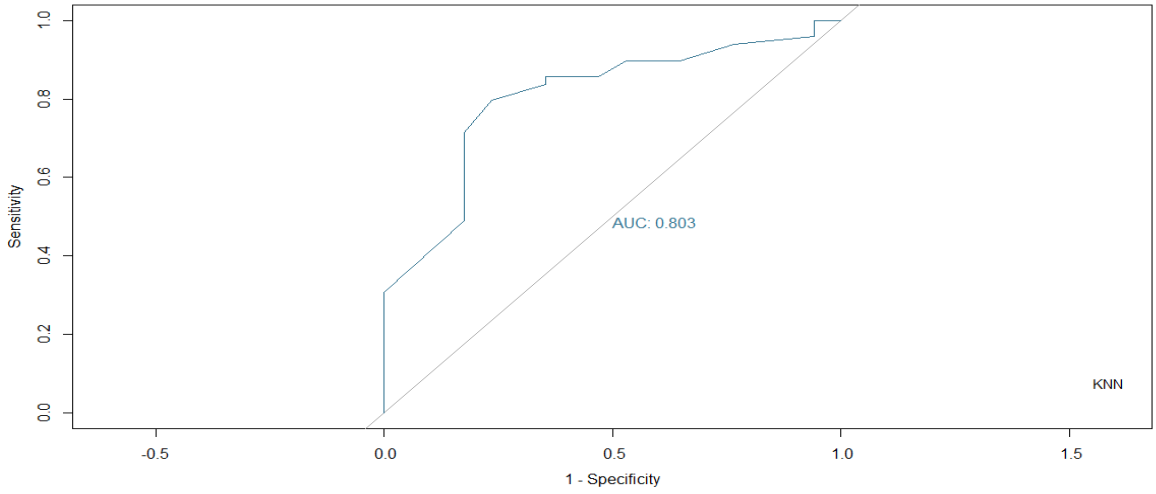
Figure N° 29 : Matrice de confusion du modèle KNN

```
Confusion Matrix and Statistics

          Reference
Prediction Non Oui
       Non   1   1
       Oui  16  48

               Accuracy : 0.7424
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

La surface ASC du modèle est de 80,3%. Comme montrée sur la figure suivante :

Figure N° 30 : Courbe ROC ASC du modèle KNN



Source : Elaboré par les auteurs à l'aide du logiciel R.

Un problème pratique de l'application du modèle KNN à notre ensemble de données est le faible nombre d'enregistrements qui peut mal impacter la performance du modèle.

2.3. Réseau de neurones artificiels (ANN)

Les réseaux de neurones sont des modèles mathématiques qui utilisent des algorithmes d'apprentissage inspirés du cerveau humain.

Le type de réseau de neurone utilisé dans cette recherche est le réseau de neurones récurrent, puisé du package « neuralnet » dans R.[1] L'algorithme est ainsi entraîné sur tout l'ensemble d'apprentissage par la méthode de rétropropagation. La fonction permet des réglages flexibles grâce aux choix personnalisés de la fonction d'erreur et d'activation. Par ailleurs, la fonction d'activation utiliser par défaut par le package est la fonction sigmoïde pour les neurones cachés dans les couches cachées et le neurone de sortie. Cette commande crée le réseau et initialise ses poids.

Figure N° 31 : Script R - modélisation ANN

```
n3 <- neuralnet(Défaut~ EURL+SARL+SNC+SPA+CentraleRisq1+CentraleRisq2+MouvementsConfiés1+MouvementsConfiés2+
                    impayésConfréres1+impayésConfréres2+ ActAgroAlim+ActManuf+ActPharma+ActEnergie+ActService+
                    ActCommerce+ActImportExport+ActBTPH+ActSanté+V2+V4+V5+V8+V9+V10+V11+V12+V13+V14+V15+V16+V21+V26,
                data = trainingANN,
                hidden = c(5,1),
                linear.output = F,
                rep=1)
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

Où :

- **data :** l'ensemble d'apprentissage.
- **hidden :** un vecteur d'entiers spécifiant le nombre de neurones cachés dans chaque couche.
- **linear.output :** logique. Si activation function ne doit pas être appliqué aux neurones de sortie, définissez la sortie linéaire sur TRUE, sinon sur FALSE.
- **rep :** le nombre de répétitions pour l'entraînement du réseau de neurones.

Lors de cette recherche, plusieurs variétés de réseaux de neurones sont testées en faisant varier le nombre de couches et le nombre de neurones cachés dans chaque couche afin de choisir la meilleure architecture qui présente un taux d'erreur minimal, puisqu'il n'existe

[1] FRITSCH Stefan et GUENTHER Frauke : « Training of neural networks », Neuralnet package, 2019.

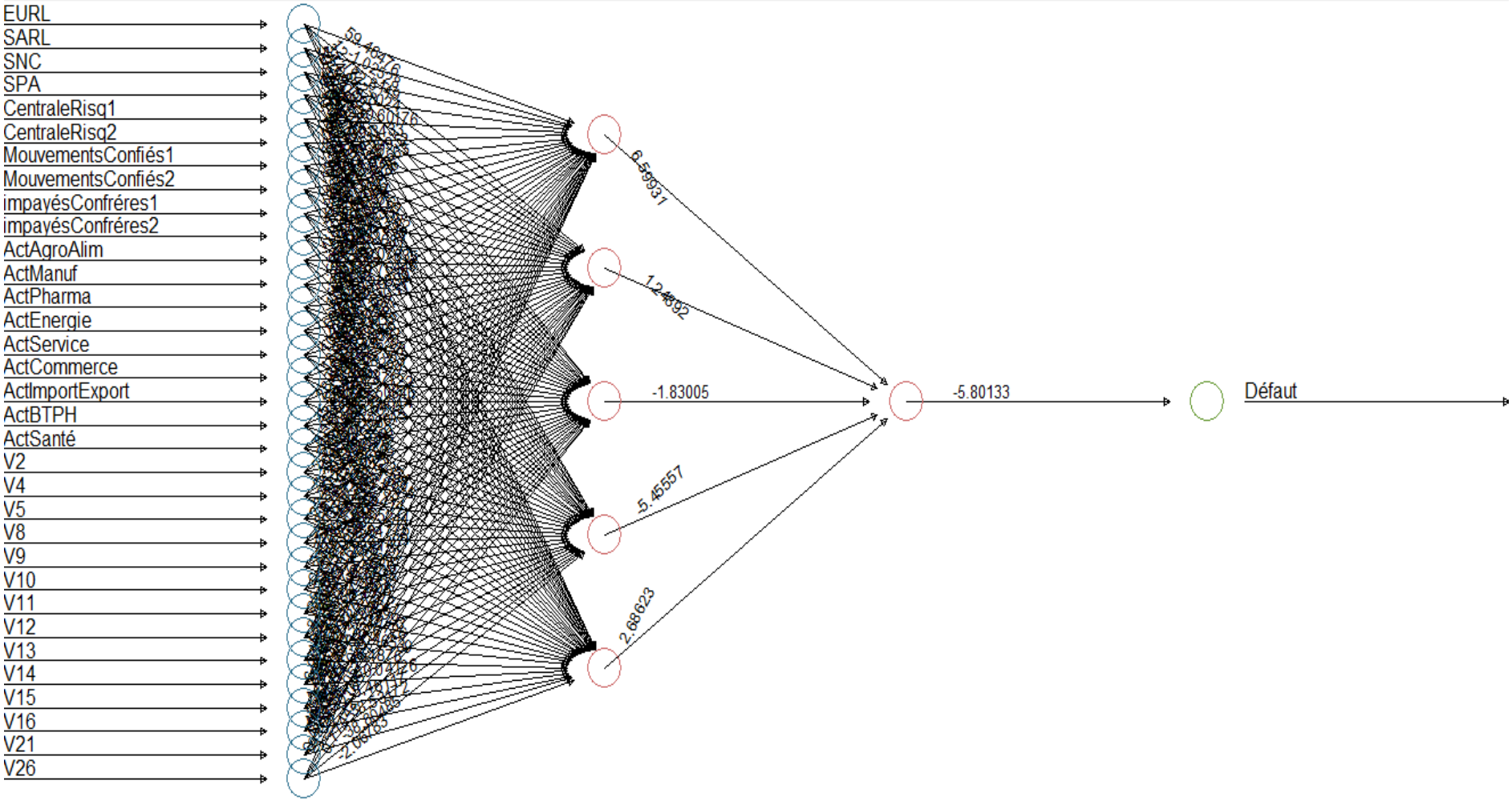
aucune loi ou théorème qui permettrait de déterminer le nombre de couches cachées et le nombre de neurones optimal à placer dans la couche cachée.[1]

En effet, dans notre programme, nous avons fixé un nombre de couches cachées égal à 2 composés de 5 et 1 neurones respectivement. Car suite à quelques simulations, ce modèle a montré le plus de précision.

Le modèle se présente comme suit :

[1] ADDO Peter et al. : op. cit.

Figure N° 32 : Présentation du modèle ANN



Source : Elaboré par les auteurs à l'aide du logiciel R.

Le modèle ANN utilisé automatise le développement de modèles analytiques avec une intervention humaine minimale, il fonctionne en recevant un ensemble de variables dans la couche d'entrée, une combinaison linéaire est utilisée pour générer de nouvelles caractéristiques et une fonction d'activation résulte de la génération d'un ensemble de neurones en sortie, ces derniers constituent les entrées de la couche suivante. Dans les couches cachées, de nouvelles caractéristiques des couches précédentes sont générés, jusqu'à l'atteinte de la couche de sortie, en obtenant une valeur prédite (propagation directe). La valeur prédite est réajustée pour minimiser la marge d'erreur des itérations précédentes jusqu'à ce que le modèle converge, il s'agit de l'optimiseur de la rétropropagation.[1] Le modèle présente un taux de classement positif de 81,39%.

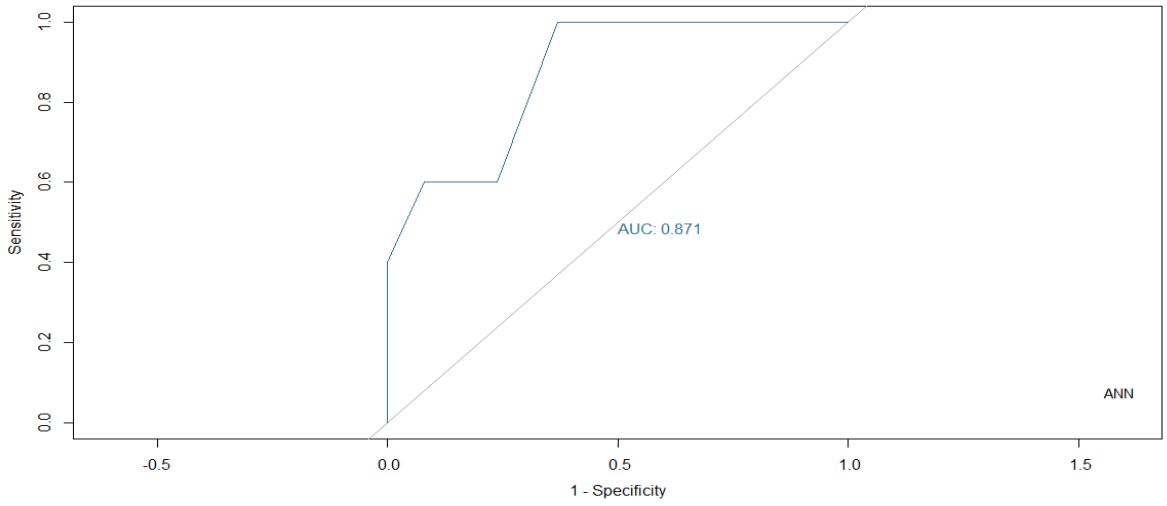
Figure N° 33 : Matrice de confusion du modèle ANN

```
> MatriceDeConfusion <- table(predicted=Pn3Test,actual=testingANN$Défaut)
> MatriceDeConfusion
         actual
predicted  0  1
        0 32  2
        1  6  3
> Précision=sum(diag(MatriceDeConfusion))/sum(MatriceDeConfusion)
> Précision
[1] 0.8139535
>
```

Source : Elaboré par les auteurs à l'aide du logiciel R.

Compte tenu des résultats de la métrique ASC ROC, la performance de l'algorithme est de 87,1%.

Figure N° 34 : Courbe ROC ASC du modèle ANN



Source : Elaboré par les auteurs à l'aide du logiciel R.

[1] ADDO Peter et al. : op. cit.

2.4. Mesures de précision

En ML la qualité des modèles se manifeste par leurs pouvoirs prédictifs.

Des études arrivent à la conclusion que les mêmes classificateurs et ensembles de données pourraient arriver à des résultats différents suite de l'utilisation de méthodes différentes de validation suivants des seuils différents.

La matrice de confusion mesure les performances du classificateur à un seuil précis (50% par défaut sur R), par extension, les méthodes d'analyse graphique (ROC) et leurs performances associées sont utilisées pour mesurer les performances du classificateur sur une plage de valeurs seuils.

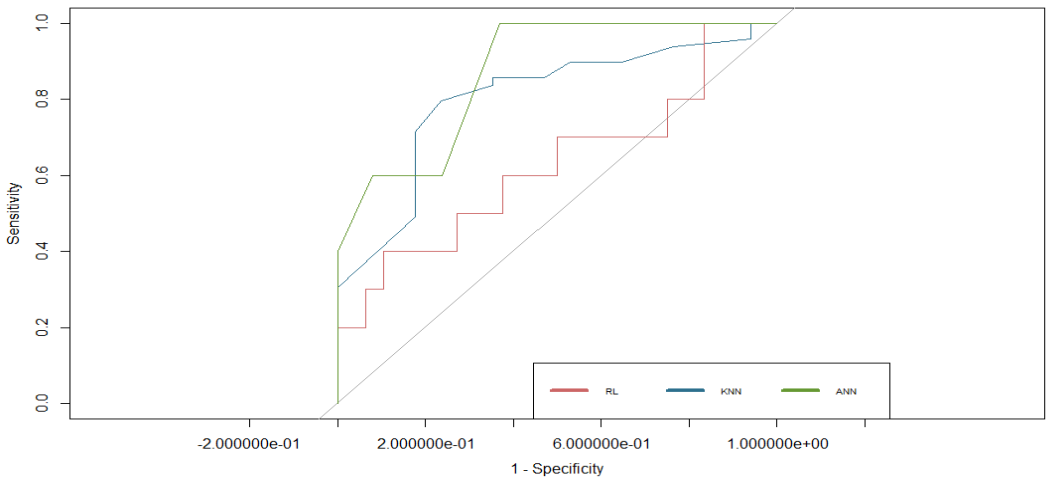
2.4.1. ROC

Nous présentons ici les courbes de caractéristique de fonctionnement du récepteur (ROC) et leurs aires sous la courbe (ASC). L'objectif étant, la détermination du modèle qui fonctionne le mieux avec nos données.

L'ASC ROC considère à la fois les résultats positifs et négatifs, ce qui permet de mesurer le niveau d'affirmation de manière plus appropriée dans l'octroi de crédits d'exploitations. Sur cette courbe, la valeur de seuil optimale se trouve généralement dans le coin supérieur gauche de la courbe où le taux de vrais positifs est beaucoup plus élevé que le taux de faux positifs. Plus l'aire est proche de 1, meilleur est le pouvoir discriminant du classificateur.

La figure n° 35 rassembles les 3 modeles précédents :

Figure N° 35 : Comparaison des trois courbes ROC ASC



Source : Elaboré par les auteurs à l'aide du logiciel R.

La comparaison du pouvoir prédictif des trois modèles, **RL, KNN, ANN** montre la performance de la technique neuronale par rapport aux deux autres modèles. En effet, le pourcentage de bon classement, issu de l'application des réseaux de neurones artificiels, est meilleur que celui présenté par RL et KNN. Leurs pourcentages respectifs sont de : 62,7%, 80,3% et 87,1%.

2.4.2. Matrice de confusion

Dans une matrice de confusion, la diagonale représente les classifications correctes, tandis que les valeurs hors diagonale sont celles ayant été mal classés. Plus nous enregistrons de valeurs sur la diagonale principale, plus nous avons de preuves d'une classification correcte. Ainsi, Il s'agit des relations entre les vrais positifs, les faux positifs, les faux négatifs et les vrais négatifs.

A partir de cette matrice, on calcule la précision du modèle dite en anglais Accuracy. La précision est la mesure de la capacité de l'algorithme à détecter les vrais positifs. Dans le cas de cette thèse, cela peut se traduire par la capacité du modèle à prédire qu'une entreprise est en faillite alors qu'elle fait faillite. Par exemple, une précision de 100 % signifie que les sociétés déclarées 1 : défaillante, connaîtront certainement ou avec une grande certitude une défaillance à l'avenir.[1]

Les dégrées de précisions obtenues à partir des matrices de confusions apparaissent dans les figures n°25, n°29 et n°33.

L'étude de ces matrices révèle que les modèles RL et ANN se ressemblent en précisons, ainsi, ils affichent une Accuracy de 81%. Alors que le modèle KNN peut trouver les positifs des positifs dans 74% des cas.

En accord à l'opinion populaire sur le modèle de régression logistique dans la littérature, le pouvoir prédictif de ce modèle est souvent classé en dernier.

3. Discussion des résultats

Dans ce qui suit, l'analyse qui a permis de faire le choix de l'ensemble des variables correspondant le mieux à nos données.

Nous allons donc passer en revue 13 configurations de modèles. Ils seront examinés en termes de méthodes utilisées, de mesures de performance ainsi que du nombre de variables

[1] UDDIN Shahadat et al. : op. cit.

et de prétraitement. Deux classes de clients sont prédites et 3 algorithmes : M2, M5, M13, sont sélectionnés. En effet, ces algorithmes ont montré les meilleures précisions.
Un tableau synthétisant 13 configurations de modèles est ainsi présenté, accompagnée des résultats auxquels nous avons abouti :

Tableau N°11 : Présentation des configurations des modèles

Codage	Méthode	Nombre de variable explicatives	Vecteur ACP	NA	ASC	Précision
M1	KNN	31	-	Supprimer	77%	90%
M2	KNN	19	-	Supprimer	80,3%	74,24%
M3	KNN	5	5	Supprimer	71%	76%
M4	ANN	31	-	Supprimer	61%	66%
M5	ANN	19	-	Supprimer	87,1%	81,39%
M6	ANN	5	5	Supprimer	55%	73%
M7	RL	31	-	Ne converge pas	Ne converge pas	Ne converge pas
M8	RL	31	-	Garder	61%	78%
M9	RL	19	-	Garder	62%	68%
M10	RL	19	-	Supprimer	55,5%	76%
M11	RL	5	5	Garder	61%	84%
M12	RL	5	5	Garder	66%	85%
M13	RL	5	5	Supprimer	62,7%	81,03%

Source : Elaboré par les auteurs à l'aide du logiciel R.

Dans le tableau n°11, nous observons la comparaison des performances et le test de différents algorithmes d'apprentissage automatique supervisé dans la classification des demandeurs de crédits d'exploitation.

Dans un modèle KNN, on remarque que la surface sous la courbe ROC atteint son optimum quand le nombre de variables est de 19. Les résultats pour chaque groupe de variables sont respectivement : 80%, 77% et 71% pour un nombre de variables et/ou vecteurs de 19, 31 et 10.

Tandis que, la matrice de confusion indique que la précision varie entre 76% et 90%, avec plus de précision pour le modèle M1. Cela signifie que, pour un seuil de 50%, le nombre de

client classé défaillant quand ils sont réellement défaillant par apport au total des observations est de 90%. Donc, le pouvoir discriminant de l'algorithme KNN est plus pertinent avec 19 variables explicatives seulement.

Les modèles M4, M5 et M6 sont des algorithmes de réseaux de neurones, les deux mesures, ASC et Accuracy, indiquent que le modèle le plus performant en prédiction est M5, pour un ensemble de 19 variables explicatives.

Remarquons que tous les modèles de KNN de ANN ont subi un prétraitement de valeurs manquantes. Ce prétraitement est une étape incontournable du fait que les packages utilisés dans leur modélisation n'acceptent les NA.

La régression logistique est ensuite appliquée sur 6 modèles d'ensembles de données différents. On remarque que le premier modèle, M7, n'a pas convergé, probablement en raison de l'hétérogénéité des intervalles des caractéristiques, car, une fois normalisé, il aboutit à des résultats dans M8.

On remarque une grande amélioration de performance entre les modèles 9 et 10, suite au prétraitement des NA. Leurs suppressions à nettement amélioré le modèle.

Malgré la performance du modèle M12, c'est M13 qui est sélectionner. Et ce, à cause de la nécessiter du traitement des NA pour permettre la comparaison des 3 méthodes.[1]

Par ailleurs, on observe que KNN a montré la précision (Accuracy) supérieure dans la majorité des cas pour les 3 ensembles de variables explicatives.
Pour la mesure ASC, ANN présente une meilleure performance dans le modèle M5, de 87,1%. Mais perd en pouvoir prédictif dans M6 une fois l'ACP utilisée. Dans le cas contraire, l'introduction de la méthode de réduction des dimensionnalités sur le modèle RL a amélioré tous ses résultats pour les deux méthodes de validation, par exemple, une ASC optimale dans M12 suivi d'une haute précision.
Bien que RL a été considéré comme le modèle le moins prédictif en termes de mesure ASC, on remarque qu'il a montré des pourcentages élevés et réguliers de précision.
La précision des modèles de machine learning peut avoir des comportements très différents, ce qui rend difficile de juger quel est le modèle le plus performant. De nombreux classificateurs dans le tableau n°11 ont une ASC élevée (c'est-à-dire supérieure à 70 %),

[1] Rappelons que les algorithme ANN et KNN n'acceptent pas les NA

mais leur précision peut varier considérablement. Par exemple : M1, M2, M3, M4. Par ailleurs, le contraire est vrai, plusieurs modèles présentent des précisions très favorable, mais leurs ASC ne le sont pas toujours : M5, M6, M8, M10, M11, M12, M13.

Cela peut être due au seuil appliqué dans ces deux méthodes. Tandis que la précision issue de la matrice de confusion applique par défaut un seuil de 50%, les probabilités issues de chaque modèle seront comparées avec ce seuil et seront ainsi classé en 0 ou 1. La courbe ROC ASC applique une variété de seuil (toutes les combinaisons comprises entre 0 et 1) pour comparer cette probabilité.

Le seuil optimal, tranchant les classes, varie d'un secteur d'activité à un autre. Il est difficile de dire quel est le meilleur seuil à appliquer au risque de crédit d'exploitation. Une manière de se faire est de procéder à une validation croisée sur différents seuils.

Ainsi, les réseaux de neurones artificiels semblent constituer un outil de prévision puissant en matière de gestion de risque de crédit d'exploitation. Ce travail confirme ainsi les études empiriques antérieures (Oden et Sharada (1990), Kerling et Podding (1994), Abdou et al, (2008)).

Les modèles des réseaux de neurones artificiels sont de plus en plus utilisés en scoring. Cependant, bien que ces nouvelles méthodes soient intéressantes et parfois plus performantes que les techniques statistiques traditionnelles, elles sont moins déchiffrables, les réseaux de neurones sont incapables d'expliquer les résultats qu'ils fournissent, ils se présentent comme des boîtes noires dont les règles de fonctionnement sont inconnues.[1]

En termes d'interprétation des pondérations, la régression logistique semble être plus performante.

En effet, dans un réseau de neurones artificiels, les liaisons internes n'ont pas de signification économique. Les pondérations des ratios figurant dans les fonctions logistiques sont par contre transparentes et faciles à interpréter du point de vue de l'analyste financier.

Ainsi, la comparaison finale s'est effectuée sur les modèles optimisés : M2, M5, M13, opérant avec un ensemble de variables de 5 composantes principales et 5 variables qualitatives dans la RL et un ensemble de 19 variables explicatives dans les deux autres algorithmes. A noter que les 19 variables sont exploitées dans les 3 modèles, même dans l'ACP, la dimension est réduite car une rotation de l'espace de données est faite. Mais

[1] KENNEDY Kenneth : op. cit.

rappelons-nous que toutes les variables rentrent dans la composition des 5 vecteurs principaux, ce sont les pourcentages de contribution qui varient.

Pour conclure, on peut dire que l'approche neuronale et la régression logistique se révèlent être deux techniques complémentaires. La RL nous permet de sélectionner les variables les plus pertinentes et le réseau de neurones peut reprendre ces variables et calculer le taux d'erreur le moins élevé.

Conclusion

Ce chapitre a permis d'abord de présenter le cadre méthodologique de la recherche, ensuite, de montrer la procédure de préparation des données et enfin, de discuter et analyser les résultats obtenus.

Cette discussion traverse trois étapes :

Premièrement, nous comparons les résultats de classification des modèles ML formés à l'aide de différentes combinaisons de variables économiques. Suite à cette comparaison, un ensemble de variables explicatives est sélectionné et transmis à l'étape d'évaluation.
Par la suite, un prétraitement de la base de données est conduit sur les valeurs manquantes, une approche de suppression de ces valeurs est adoptée, les résultats sont discutés dans la section 03.
Dans une dernière étape, une comparaison des modèles retenus nous permet de sélectionner l'algorithme disposant du meilleur pouvoir discriminant sur nos données.
Les résultats analytiques ont révélé que les algorithmes d'apprentissage automatique sont capables de modéliser le risque de crédit d'exploitation, ANN a donné les meilleures précisions avec 87,1% d'aire sous la courbe ROC et 81,39% de précision, suivi de KNN (80,3% ;74,24%) et RL (62,7% ;81,03%).

Conclusion générale

Le financement bancaire est la principale source de financement externe sollicitée par les PME. Ce financement n'est toutefois pas facilement accessible, l'évaluation des institutions bancaires du risque de crédit est souvent plus sévère pour cette taille d'entreprise en raison de leur susceptibilité à refléter une éventuelle faillite, et l'instabilité de leurs chiffres.

Ainsi, la maîtrise de ce risque demeure une préoccupation majeure et un objectif recherché par les emprunteurs de crédit. En effet, la mise en place des meilleurs mécanismes, les plus modernes doit être envisagée. Pour y parvenir, les modèle ML peuvent constituer un outil d'aide à la décision pour le conseiller bancaire. Étant donné que l'objectif de ce mémoire est de vérifier la capacité des modèles d'intelligence artificielle à prédire le risque de crédit vis-à-vis des demandeurs de prêt d'exploitation afin d'atteindre la probabilité de défaut de chaque demandeur, des modèles intelligents de l'apprentissage supervisé sont sélectionnés, entraînés, validés et évalués à travers différentes métriques d'évaluations. Par ailleurs, l'importance de ce crédit est reflétée par son rôle vital tout au long du cycle d'exploitation des entreprises. Il représente 10% de l'ensemble des crédits accordé par le CPA en 2019. [1] Cela concorde avec notre hypothèse de départ H_1.

A travers 13 configurations de modèles supervisés de classification appliquer sur les données de l'organisme bancaire algérien le CPA, 282 demandes de crédit sont examinées à travers 31 variables économiques et financières. Les techniques de prédiction utilisées dans l'étude empirique sont RL, KNN et ANN. Les résultats indiquent que le modèle le plus affirmé dans ce processus de notation est le réseau de neurones artificiels (87,1%), suivi du modèle des K voisins les plus proches (80,3%) et finalement, la régression logistique (62,7%). Pour des précisions respectives de 81,03%, 74,24% et 81,39%, ce résultat confirme ainsi notre 2ème hypothèse H_2.

La procédure de prétraitement des données à travers la méthode de réduction des dimensionnalité ACP ainsi que la codification des variables et la suppression des valeurs manquantes a eu un impact manifeste sur la performance des modèles, cela nous a permis de poursuivre l'étude comparative avec des modèles optimisés. En effet, ces résultats convergent avec l'hypothèse H_3.

De cette étude, nous avons déduit que l'approche neuronale et la régression logistique sont deux méthodes complémentaires. Tandis que la régression logistique nous permet de

[1] Figure N°9 : Evolution des crédits d'exploitation CPA.

sélectionner l'ensemble de variables, le réseau de neurones permet une modélisation optimale. Ces informations peuvent aider les gouvernements, les investisseurs, les gestionnaires et d'autres parties prenantes dans la prise des décisions économiques. D'un point de vue Marketing, l'intelligence artificielle permettra aux banques d'identifier les préférences de leurs clients. Ainsi, elle renforcera la relation Banque-Client et améliorera « l'expérience client » en proposant des produits et des solutions de manière proactive.

Par ailleurs, sur un plan théorique, notre recherche a pour but d'introduire le concept de la notation des crédits d'exploitation en général, et de cette notation pour les PME en particulier, en présentant les différents risques et spécificités liés à cette dernière, en matière de garanties, fonctionnement, typologie, …

Dans un deuxième aspect, cette recherche offre une présentation géométrique des modèles les plus célèbre de l'intelligence artificielle dans la classification binaire, on peut citer : SVM, KNN, DT, RF, ANN, RL, ... et résume leurs forces et faiblesse dans un tableau. Cette présentation simplifiée permet aux chercheurs des instituts de commerce de se familiariser et avoir une vue panoramique claire de cette discipline.

Sur le plan empirique, on présente une étude expérimentale de 13 configurations de modèles d'apprentissage supervisé ayant subis ou non un prétraitement sur les données. L'analyse en composantes principales (ACP) est utilisée afin de sélectionner le nombre de caractéristiques adéquat pour une prédiction optimale. Cette analyse se termine par une étude comparative des résultats des différents modèles appliqués ainsi qu'une discussion des résultats.

- **Problèmes ouverts aux travaux futurs**

A travers cette étude, nous avons implémenter trois modèles d'apprentissage supervisé dans un contexte de classification des clients de la banque en classes binaire : défaillant et non défaillant. Les résultats nous ont amené à la conclusion de la primauté du modèle de réseau de neurones dans la bonne classification des créanciers. Néanmoins, cette étude ne s'est pas réalisée sans embuches, ces difficultés peuvent constituer des problématiques à de nouvelles études :

Dans un premier lieu, l'ensemble de donnée peut être étendu en tenant compte d'un plus grand nombre et d'une multitude de variables explicatives. Nous pouvons, par exemple, introduire les modèles pourraient également étudier l'influences des chocs économiques, tels que la pandémie mondiale COVID-19 sur la notation des crédits d'exploitation.

Des facteurs plus subjectifs doivent être pris en compte dans les critères de notation, comme la réputation, la longévité de la relation avec la banque, l'entente avec le dirigeant de la banque, …

Par ailleurs, la méthode de réduction de dimensionnalité (ACP) est appliquée au modèle RL. Il en résulte comme limite, la perte de 41% de l'information. Ça peut être la raison pour laquelle les résultats du modèle de régression logistique sont les moins efficace.

D'autre part, la codification des variable, le traitement des valeurs manquantes et des valeurs aberrantes, le traitement du déséquilibre de classe et bien d'autres méthodes de prétraitement de données peuvent être introduite ou optimiser dans le modèle, cette étape peut complètement changer les résultats et améliorer les modèles.

Finalement, on propose l'implémentation d'autres modèles supervisés et non supervisés pour de futures comparaisons.

Bibliographie

Ouvrages

- BEGUIN Jean-Marc et BERNARD Arnaud : « *L'essentiel des techniques bancaires* », édition Groupe Eyrolles, 2008.
- BENHALIMA Ammour : « *Pratique des techniques bancaires avec référence à l'Algérie* », éditions Dahleb, Alger, 1997.
- BOURBONNAIS Régis : « *econometrie cours et exercices corrigés* », Dunod,2014.
- BOUYAKOUB Farouk : « *L'entreprise et le financement bancaire* », édition Casbah, Alger, 2000.
- CAUDAMINE Guy et MONTIER Jean : « *Banque et marché financiers* », édition Economica, 1998.
- CORNUEJOLS Antoine et al. : « *Apprentissage artificiel* », édition Eyrolles, 2003.
- DAYAN Armand : « *Manuel de gestion* », édition Collectif ellipses, volume 2, 2004.
- DESMICHT François : « *Pratique de l'activité bancaire* », édition Dunod, France, 2004.
- DE COUSSERGUES Sylvie et BOURDEAUX Gautier : « *Gestion de la banque : du diagnostic à la stratégie* », 6ème édition Dunod, 2010.
- DE LA BRUSLERIE Hubert : « *Analyse financière* »,5ème édition, Dunod, Paris, 2014.
- GLASSNER Andrew : « *Deep learning : A visual approach* », édition No strach press, 2021.
- GOURIEROUX Christian et TIOMO André : « *Risque de crédit : Une approche avancée* », édition Economica, 2007.
- HULL C John : « *Gestion des risques et institutions financières* », édition Pearson, 2018.
- LAZARUS Jeanne : « *L'épreuve du crédit* », édition Sociétés contemporaines, Paris, 2009.
- MARAZZI Alfio : « *Introduction à la régression logistique* », IUMSP, 1989.
- NAULLEAU Gérard et ROUACH Michel : « *Contrôle de gestion bancaire* », édition Revue banque, 2020.

- RONCALLI Thierry : « *La gestion des risques financiers* », édition Economica, 2009.
- RUSSOLILLO Giorgio : « *Régression logistique* », CNAM, 2018.
- TUFFERY Stéphane : « *Data Mining et statistique décisionnelle : L'intelligence des données* », éditions TECHNIP, 2012.
- VERNIMMEN Pierre et al. : « *Finance d'entreprise »*, édition Dalloz, 2020.

Articles de revue

- ADDO Peter et al. : *« Credit risk analysis using machine and deep learning models* », Maison des sciences économiques,2018.
- AKERLOF George A : « *The market for lemons : quality uncertainity and the market mechanisms* », The quarterly journal of ecnomics,1970.
- ANG James S : « *Small Business Uniqueness and the Theory of Financial Management »*, The journal of entrepreneurial finance,1991.
- AUBIER Maud et CHERBONNIER Frédéric : *« L'accès des entreprises au crédit bancaire* », Economie et prévision, 2007.
- BORANA Jatin, « *Applications of artificial intelligence & associated technologies* », Department of electrical engineering, Jodhpur National University ,2016.
- BOUCHARD Guillaume : « *Les modèles génératifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle »*, Interface homme-machine université Joseph-Fourier - Grenoble I, 2005.
- BRANGER Jacques : « *Traité d'économie bancaire : Instruments juridiques, techniques fondamentales* », Presses Universitaires de France, Paris, 1975.
- CHANT Elizabeth M et WALKER David A : « *Small business demand for trade credit* », Applied economics,1988.
- CHOW Jacky C. K : « *Analysis of financial credit risk using machine learning* », Aston university, 2018.
- COOK Diane J et al. : « *Graph based hierarchical conceptual clustering* », Journal of machine learning research, 2001.

- COUPPEY-SOUBEYRAN Jézabel : « De Bâle 2 à Bâle 3 : la nouvelle réglementation bancaire internationale », La documentation française, 2013.
- CUNNINGHAM Sally Jo « *Machine learning and statistics : A matter of perspective* » Hamilton, New Zealand : University of Waikato, Department of Computer Science, 1995.
- FEKIR Hamza : « Présentation du nouvel accord de Bâle sur les fonds propres », Revue MIF, 2005.
- FRITSCH Stefan et GUENTHER Frauke : « Training of neural networks », Neuralnet package, 2019.
- GERMAIN-MARTIN Henry : « *Le crédit à moyen terme* », Revue d'économie politique,1958.
- GUILLE Marianne et al. : « *Structure financière et dépenses de R&D* », Economie et prévision, 2011.
- HUANG Chen-Lung et al. : « *Credit scoring with a data mining approach based on support vector machines* », Expert systems with applications, 2004.
- HUSSEIN Abdou et al. : « On the applicability of credit scoring models in egyptian banks », Banks and Bank Systems,2002.
- KHEMAKHEM Sihem et BOUJELBENE Younes : « *Artificial intelligence for credit risk assessment : Artificial neural network and support vector machines* » ACRN Oxford, Journal of finance and risk perspectives ,2017.
- KUHN Max : « *Building Predictive Models in R Using the caret package* », Journal of statistical software, 2008.
- LAMARQUE Eric et MAURER Frantz : « *Le risque opérationnel bancaire* », Revue française de gestion, 2009.
- LOTFI Sihem et MESK Hicham : « Prédiction du risque de crédit : étude comparative des techniques de Scoring » International Journal of Accounting, Finance, Auditing, Management and Economics, 2020.
- LOUZADA Francisco et al. : « *On the impact of disproportional samples in credit scoring models : An application to a Brazilian bank data* », Academia,1989.
- MATOUSSI Hamadi et ABDELMOULA Aida Krichène : « *La prévision du risque de défaut dans les banques tunisiennes : Analyse comparative entre les méthodes*

linéaires classiques et les méthodes de l'intelligence artificielle : les réseaux de neurones artificiels », Crises et nouvelles problématiques de la valeur, 2010.

- MAURER Frantz : « *L'impact du risque de marché sur le résultat de* l'entreprise », Revue française de gestion, 2005.
- MHLANGA David : « *Financial inclusion in emerging economies : The application of machine learning and artificial intelligence in credit risk assessment* », International journal of financial studies,2021.
- MILANA Carlo et ASHTA Arvind *: « Artificial intelligence techniques in finance and financial markets : A survey of the literature »*, Strategic change, 2021.
- NAHMIAS Laurent : « *Impact économique des défaillances d'entreprise* », Bulletin de la banque de France n°137, 2005.
- NDIAYE Khadidiatou, « *Le scoring en microfinance : un outil de gestion du risque de crédit* », PAMIF CRES, 2012.
- NOUY Danièle : « *Bâle II face à la crise : quelle réformes* », Revue d'économie financière, 2008.
- REHFELDT Ruth Anne et al. : « *Observational learning and the formation of classes of reading skills by individuals with autism and other developmental disabilities D. Latimoreb, Research in development disabilities »*, Res dev disabil, 2003.
- RUGEMINTWARI Clovis et al. : « Bâle 3 et la réhabilitation du ratio de levier des banques », Revue économique, 2012.
- SCHREINER Mark : « Les vertus et faiblesses de l'évaluation statistique (Credit Scoring) en Microfinance », Microfinance risk management, 2003.
- STIGLITZ Joseph E et WEISS Andrew : « *Credit rationing in markets with imperfect information »*, American Economic Review, 1981.
- UDDIN Shahadat et al. : « *Comparing different supervised machine learning algorithms for disease prediction* », BMC Med Inform Decis Mak, 2019.
- UlLLAH Hayat et al. : « *Comparative study for machine learning classifier recommendation to predict political affiliation based on online reviews* », CAAI Trans Intell Technol, 2021.
- VAPNIK Vladimir N et al. : *« A training algoritm for optimal margin classifiers* », Computational learning theory, 1992.

- WILLIAMSON Oliver : « *La théorie des coûts de transaction* », Revue française de gestion, 2003.
- ZAKI H et al. : « Méthodologie générale d'une étude ACP : Généralités, concepts et exemples », Revue interdiscplinaire, volume 1, 2016.
- ZHANG Min-Ling et ZHOU Zhi-Hua : « *ML-KNN : A lazy learning approach to multi-label learning* », Pattern recognition ,2007.

Travaux universitaires

- BERRAIH Radia : « *Gestion du risque de défaillance des PME par le Scoring Application comparative entre la régression logistique et les réseaux de neurones* », IFID, 2020.
- ELHAMMA Azzouz : « *La gestion du risque crédit par la méthode du scoring : cas de la Banque Populaire de Rabat-Kénitra* », Maroc, 2009.
- HAMDAD Leila : « *Introduction au machine learning* » Ecole supérieure d'informatique ,2021.
- KENNEDY Kenneth : « *Credit scoring using machine learning* », Doctoral thesis, Technological University Dublin, 2013.

Les textes de lois

- Code de commerce algérien, article 409.
- Code de commerce algérien, titre 3, chapitre 3, article 543 bis 14.
- Loi n°03-11 du 26 aout 2003 relative à la monnaie et au crédit.
- Règlements de la Banque d'Algérie Nº14-01 du 16 février 2014 portant coefficients de solvabilité applicables aux banques et établissements financiers.
- Règlement 14-03 du 16 février 2014 relatif au classement et provisionnement des créances et des engagements par signature des banques et établissements financiers.

Webographie

- https://www.bank-of-algeria.dz/
- https://cran.r-project.org/web/packages/available_packages_by_name.html
- https://stackoverflow.com/
- https://www.industrie.gov.dz/
- https://fr.quora.com/

Annexes

Annexe N° 1 : Tableau comparative entre les différents modèles de classification

Modèle de classification	Avantages	Désavantages
RL	- Interprétation probabiliste facile des paramètres du modèle. - Absence d'hypothèses sur la distribution variable explicatives. - Traitement des données manquantes.	- Ne considère pas les relations non linéaires entre les variables explicatives (Multi-colinéarité). - Difficile à interpréter lorsque le nombre de dimensions du modèle est élevé.
KNN	- Facile à comprendre et à interpréter. - Modélisation des problèmes de différents types : régression et classification.	- Lenteur de l'exécution de l'algorithme surtout pour une grande base de données. - Attribution du même poids(importance) a toutes les variables du modèle. - Requiert un grand espace de stockage.
SVM	- Fonctionne sur tout type de données (structuré, non structuré et semi-structuré). - Couvre les modèles linéaire et non linéaire. -Présente un faible risque de surajustement.	- Ne convient pas aux échantillons de grande taille. - Nécessite un prétraitement sur les données. - Le modèle est difficile à interpréter.

Naïve Bayes	- Applicable sur les problèmes de classification binaires et multi-classes. - Fonctionne sur des problèmes non linéaires. - Fonctionne sur les petits échantillons. - Il peut faire des prédictions probabilistes.	- L'hypothèse naïve d'indépendance des variables explicatives. - Il suppose la distribution normale des attributs numériques.
DT	- Facile à comprendre et à interpréter. - Plusieurs types de données : nominaux, numériques et qualitatifs sont pris en charge par le modèle. - Traite les données manquantes. - Peut être validé par des tests statistiques.	-L'algorithme dépend de l'ordre des variables. -le surajustement peut facilement se produire. -nécessite un grand nombre d'observations.
RF	- Présente moins de chances de surajustement aux données d'entrainement. - Puissance du processus de classification. - Sélection des variables explicatives les plus importantes dans un problème donné.	- Complexe et couteux. -des hyperparamètres doivent être prédéterminés.
ANN	- Modélisation des relations non linéaires entre les variables. - Modélisation des problèmes de différents types ; régression et classification. - Mise à jour facile. - Flexible, prends plusieurs formes d'apprentissage.	- Les résultats ne sont pas explicites et sont difficile à comprendre par les utilisateurs. - Présente un risque de surajustement élevé. - Ne traite pas les NA. - Cout computationnel élevé.

Annexe N° 2 : Les valeurs propres et pourcentages de la variance expliquée

```
> ValeursPropres
       eigenvalue variance.percent cumulative.variance.percent
Dim.1   2.8712653       20.5090380                    20.50904
Dim.2   1.6512019       11.7942994                    32.30334
Dim.3   1.4975712       10.6969373                    43.00027
Dim.4   1.2214919        8.7249425                    51.72522
Dim.5   1.0468907        7.4777907                    59.20301
Dim.6   0.9972788        7.1234200                    66.32643
Dim.7   0.9698351        6.9273933                    73.25382
Dim.8   0.8615172        6.1536942                    79.40752
Dim.9   0.7764540        5.5461002                    84.95362
Dim.10  0.7070627        5.0504480                    90.00406
Dim.11  0.6051820        4.3227288                    94.32679
Dim.12  0.5472391        3.9088511                    98.23564
Dim.13  0.1297732        0.9269511                    99.16259
Dim.14  0.1172367        0.8374053                   100.00000
```

Annexe N° 3 : Contribution des variables explicatives dans les composantes principales

```
> head(var$contrib, 14)
                            Dim.1        Dim.2       Dim.3       Dim.4        Dim.5
FREnJrsCA              26.25962876  6.597722077  0.02682376  0.43677654 6.869107e-01
DélaiClient             7.29000905  6.571189517 11.61960525  0.04670262 4.111389e+00
DélaiFournisseur       26.34110256  4.028311870  0.13448921  1.40557575 3.615772e+00
Disponibilité(net)/DCT  1.44629915 15.006222464  4.50563206 20.06430798 4.574188e-01
CF/EBE                  1.05046760 10.653230590  1.37497698  7.14502603 1.270422e+01
DCT/Total D             0.28644184  4.470466719 11.74339979  9.89657907 1.871850e+01
RotationStocks          0.17195819  2.064825487  2.28429192  1.57727913 2.029191e+00
ROA                     5.10254582 22.918648225  1.79330910  0.76076718 1.082655e+00
ROE                     0.14932282  3.527222213  0.92455692 34.71921283 5.748034e+00
CAF/CA                  0.17117605  9.499053363  9.45654686  4.16708596 2.106882e+01
VAExplt/CA              0.65579621  2.881109395 35.40733411  2.34682911 5.344418e+00
EBE/CA                 24.38209941  0.001138943 11.00742554  0.01680576 4.222751e-05
TotalD/TotalActif       6.67702012 10.628334165  7.73632453 16.99970946 3.367656e-01
DLT/CAF                 0.01613241  1.152524971  1.98528398  0.41734257 2.409587e+01
>
```

Annexe N° 4 : Table récapitulative du modèle RL

```
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.74326    0.65562  -2.659  0.00784 **
FormeJuridique2      -0.75794    0.48758  -1.554  0.12007
FormeJuridique3      -0.79952    1.17137  -0.683  0.49489
FormeJuridique4       1.07130    1.27045   0.843  0.39909
centraleRisques1      0.13874    0.49187   0.282  0.77790
ImpayésConfrères1     0.02410    0.49628   0.049  0.96127
MouvementsCconfiés0   2.77929    0.43034   6.458 1.06e-10 ***
Activité2            -0.43181    0.62661  -0.689  0.49075
Activité3            -0.11211    0.82442  -0.136  0.89183
Activité4            -0.99666    1.54476  -0.645  0.51881
Activité5            -0.96212    0.94484  -1.018  0.30854
Activité6           -17.75651 1246.77668  -0.014  0.98864
Activité7            1.86782    1.01164   1.846  0.06484 .
Activité8            0.57032    1.99325   0.286  0.77478
Activité9           -16.55458 6522.63864  -0.003  0.99797
PC1                  -0.07389    0.17845  -0.414  0.67885
PC2                   0.02487    0.18549   0.134  0.89332
PC3                   0.36998    0.22761   1.626  0.10405
PC4                  -0.55106    0.27143  -2.030  0.04234 *
PC5                   0.23463    0.19994   1.173  0.24061
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 254.20  on 227  degrees of freedom
Residual deviance: 161.59  on 208  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 201.59
```

Table des matières

TABLE DES MATIERES