EPIGRAPHE

« Je n'ai pas besoin d'avancer à grand pas, ce qui compte c'est d'avancer dans la bonne direction »

Anonyme

DEDICACE

À celle qui m'a indiqué la bonne voie en me rappelant que le courage et la volonté font toujours de grands hommes…

Merci ma Mère.

À celui qui a attendu avec patience les fruits de sa bonne éducation…

Merci mon Père.

À tous mes frères et sœurs « Delphin» «Esperance» « Bienfait» « Arthur» « Guilaine» «Yvette» « François»…

À tous mes amis qui m'ont toujours soutenu et encouragé au cours de la réalisation de ce mémoire ainsi que tous mes camarades.

Aristide CHIZA MATUNGURU

REMERCIEMENTS

En préambule à ce mémoire nous remercions tout d'abord Dieu le père tout puissant, qui ne cesse de nous combler de sa grâce de vivre encore sur cette terre, qui exauce nos supplications et accueille nos prières.

Nos remerciements s'étendent au CT. Vianney KAMBALE WITESYAVWIRWA directeur de ce travail, pour les orientations qui ont constitué un apport considérable sans lequel ce travail n'aurait pu être réalisé de la sorte.

Nos profondes reconnaissances s'en valent aux autorités académiques de l'ULPGL, particulièrement à celles de la Faculté des Sciences et des Technologies Appliquées pour la formation scientifique de qualité qu'elles mettent à notre disposition, plus particulièrement au Prof. Dr. Ir. Olivier BARAKA MUSHAGE et Dr Ir Alain AKWIR pour leurs orientations.

Nous remercions également les autorités de la Coopéc Bonne moisson, particulièrement au Gérant KAKULE KIMBESA Jeannot et au chef de service de crédit Mr Narkison M. KAUSA pour la facilitation à l'accession des données qui ont permis la réalisation de ce travail.

Nos plus profonds remerciements s'adressent à nos parents KAGANDA WA KABWENDE Arnold et TEMBEYA SHAMAVU Marie-Joséphine qui n'ont cessé de nous soutenir à tous les niveaux durant nos études et partout ailleurs.

Nous remercions également tous les membres de l'association des Jeunes Unis pour le Développement et le Changement des Mentalités JUDCM asbl pour leur soutien et encouragements.

Nos plus profonds remerciements s'adressent également aux amis du TAO Technologies dont Ir Gloire KABALA, Ir Abednego WA MUHINDO et Joaquim MOLO, pour leurs soutiens à tous les niveaux durant notre parcours académique.

À nos frères, sœurs, amis et camarades : ESPERANCE, MERVEILLE, GRACE, ABDIAS, JOELLE, JULIE, JOHNSON, JUSTIN, LYLYANNE, JULIEN, ALAIN, ALLIANCE… et que tous ceux dont leurs noms ne sont pas cités ne se sentent pas oublier mais qu'ils trouvent également dans cette œuvre notre gratitude pour leurs contributions.

Aristide CHIZA MATUNGURU

SIGLES ET ABREVIATIONS

ANOVA : Analysis Of variance

AVEC : Association Villageoise d'Épargne et de Crédit

CART : Classification And Regression Trees

CECI : Caisse d'Épargne et de Crédit Interne

CHAID : Chi-square Automatic Interaction Detector

COOPEC : Coopérative d'Épargne et des Crédits

CRISP-DM : Cross Industry Standard Process for Data Mining

CRM : Customer Relationship Management

GUI : Graphical User Interface

IA : Intelligence Artificielle

IDE : Integrated Development Environment

JUPYTER : Julia Python Et R

MAR : Missing At Random

MCAR : Missing Completely At Random

NMAR : Non Missing At Random

OAD : Outil d'aide à la décision

PM : Personne Morale

RDC : République Démocratique du Congo

SACCOS : Saving and Credit Cooperative Society

TKINTER : Tool Kit Interface

VICOBA : Village Community Bank

LISTE DES FIGURES

LISTE DES TABLEAUX

RESUME

De nos jours, une grande quantité des données est générée par divers systèmes d'informations, et cela ne cesse de croitre du jour le jour. Cependant ces données renferment beaucoup d'informations et des connaissances qui peuvent être utilisées à plusieurs fins, y compris des fins de développement. Le Data Mining est, effectivement, un domaine d'intelligence artificielle qui s'intéresse à analyser et à exploiter les informations et tendances contenues dans des vastes quantités des données. Ce présent travail utilise des techniques de Data Mining pour palier au problème de non-remboursement des prêts sous le thème « *Modélisation de la crédibilité d'un client dans une coopérative d'épargne et de crédit avec le Data Mining*». La méthodologie Cross Industry Standard Process-Data Mining (CRISP-DM) a été utilisée pour atteindre les objectifs de cetravail. Elle est une approche structurée, utilisée pour les projets. Quatre Modèles de prédiction ont été entrainés sur notre dataset dont les arbres de décision, la régression logistique, les réseaux bayésiens et les forêts aléatoires. Pour évaluer la performance des modèles énoncées, les métriques d'évaluations suivantes : Accuracy, Precision, Recall et Taux d'erreur ont été utilisé. Ces métriques démontrent que les arbres de décision et les forêts aléatoires sont les meilleurs modèles pour cette tâche. Notons que pour réaliser ce travail, plusieurs librairies de python ont été utilisées, dont Pandas, Numpy, Scikit-learn, Matplotlip et une interface graphique a été implémentée pour faciliter l'utilisation aux décideurs de la coopérative d'épargne et des crédits avec Tkinter. Ce travail présente une opportunité d'automatisation de service et le traitement à temps-réel du dossier de sollicitation de crédit. Son application, surtout, réduira le risque d'octroyer les prêts aux clients non crédibles

Keywords: Data Mining, crédit, data-science, prédiction, remboursement, Fouille, Analyse, Métrique d'évaluation, Data set.

ABSTRACT

Nowadays, a large amount of data is generated by various information systems, and it keeps on growing every day. However, this data contains a lot of information and knowledge that can be used for many purposes, including development purposes. Data mining is, exactly, one field of artificial intelligence that is interested in analysing and exploiting information and patterns found in vast amount of data. The present work makes use of data mining techniques to respond to the challenge of non-repayment of loans under the title *"Modelling of the credibility of a customer in a savings and credit cooperative using Data Mining"*. To reach the objective of this work, the Cross Industry Standard Process-Data Mining (CRISP-DM) methodology has been used. It is a structured approach to plan data mining projects. Four models were used for the prediction on our dataset, namely the decision trees, logistic regression, Naïve Bayes and random forests. To evaluate the performance of the above-mentioned models, the following evaluation metrics were used: Accuracy, Precision, Recall and Error rate. These metrics demonstrate that the decision trees and random forests are the best for the task at hand. Let it be mentioned that this work was done by using various python libraries such as Pandas, Numpy, Sckit-learn and Matplotlib. We also created a graphical user interface using TKINTER to make the task easy for the decision makers of the savings and credit cooperative to use our model. This work presents the opportunity to automate the service and allow a real-time processing of credit application. The application will, mostly, reduce the risk of granting credit to non-credible customers.

Keywords: Data Mining, credit, data science, prediction, repayment, Search, Analysis, Evaluation metrics, Data set.

SOMMAIRE

INTRODUCTION GENERALE

Les microcrédits sont des innovations dans le domaine financier permettant aux populations démunies ou aux faibles ressources de pouvoir accéder aux crédits pour entreprendre une activité génératrice de revenus. Ceci étant d'actualité à Goma, dans d'autres villes de la RDC voire dans le monde entier, les entreprises octroyant les crédits courent certains risques dont le plus énorme est le non remboursement de la part de ses clients. Ainsi, beaucoup d'institutions imprudentes, n'ayant pas de dispositif de contrôle interne efficace avant d'octroyer un crédit, sont entrain de fermer les portes, c'est-à-dire tomber en faillite.

Certaines coopératives octroient de crédit aux clients sans pour autant analyser les paramètres pouvant influencer le remboursement de ce crédit. C'est le cas par exemple de l'activité du client, son salaire, l'affectation de l'argent emprunté, etc. C'est ainsi que certains clients ne parviennent pas à rembourser leur crédit et d'autres y arrivent après avoir dépassé l'échéance donné. Cela place la coopérative dans de mauvaises conditions de gestion et nuit même à son évolution. Aussi d'autres coopératives ne s'inspirent pas de leur historique d'octroi de crédit ; pourtant celle-ci constituerait une base des faits pouvant leur guider dans le processus d'octroi de crédit.

Ainsi donc, les questions suivantes ont guidé notre recherche :

1. Y-a-t-il moyen d'utiliser une technique permettant de minimiser le risque associé à l'octroi de crédits dans une coopérative pour ainsi se pérenniser et rester viable ?
2. Est-il possible d'utiliser ces techniques pour doter la coopérative d'un outil de prédiction de crédibilité d'un client ?
3. Comment mettre en place une solution efficace contre les risques que court une coopérative lors de l'octroi de crédit ?

L'usage des techniques de Data Mining permettrait de minimiser le risque lié à l'octroi des crédits dans une coopérative cela en supposant que la coopérative possède une grande quantité des données. Cette technique permettrait de fouiller ces données disponibles, les analyser et tirer des connaissances pour le développement de la coopérative. Ainsi, cette dernière se pérenniserait et resterait en vie si elle prend en compte ces éléments.

En outre, cette technique de Data Mining possède une méthodologie appelée Cross Industry Standard Process for Data Mining qu'on appliquerait sur les données disponibles et cela permettrait d'implémenter un modèle intelligent capable de prédire la crédibilité d'un client.

C'est dans le sens de proposer une solution efficace à ce problème que s'inscrit le présent travail. Il consiste à mettre en place un système que les décideurs des entreprises auront à utiliser pour savoir si le client est crédible, c'est-à-dire s'il peut accéder à un crédit. Cette solution sera implémentée avec les techniques de Data Mining.

Dans ce travail, nos recherches s'orientent dans le domaine du Data Mining, car de nos jours, une grande quantité des données est générée par les systèmes d'informations, et cela ne cesse de croitre du jour le jour, d'où il s'avère important de les analyser et utiliser le résultat pour des fins de développement. De même, notre travail trouve son intérêt dans la pérennisation des activités de la coopérative pour permettre seulement aux clients crédibles de continuer à accéder aux crédits et faire tourner le capital, ce qui reste aussi un avantage pour les clients.
L'objectif principal de notre travail est donc de mettre à la disposition des décideurs de la coopérative de crédit un système de prédiction permettant de réduire les risques que court l'entreprise lors de l'octroi des crédits.
Nous avons fait recours aux techniques de recherche suivante pour la réalisation de ce travail :

- ***Technique d'observation :*** C'est une technique qui consiste à mettre le chercheur en communication sur des matières déjà existantes en vérifiant bien évidemment si cela peut contribuer à notre étude.
- ***Techniques documentaires :*** par consultation à la bibliothèque et sur internet des ouvrages et articles ayant trait à notre sujet de recherche.

Comme dans toute recherche menée, les difficultés ne manquent pas, pour notre cas, les plus grandes ont été de trouver une documentation adéquate pour notre recherche, donc nous avons assisté à une rareté de la documentation.
Hormis l'introduction et la conclusion, notre travail comporte quatre grands chapitres :

1. Généralités sur le Data Mining
2. Procédure du service de prêt et remboursement dans une coopérative d'épargne et de crédit
3. Présentation et exploitation des données
4. Élaboration du modèle de prédiction et Analyse de résultats

CHAPITRE I. GENERALITES SUR LE DATA MINING

Cette partie du travail est consacrée à un aperçu sur les concepts de datamining en rapport avec l'orientation du sujet, cela dans l'optique de faire comprendre facilement la suite.

I.1. DEFINITIONS

I.1.1. Data Mining

Le Data Mining dit fouille de données en français, est un sujet qui dépasse aujourd'hui le cercle restreint de la communauté scientifique pour susciter un vif intérêt dans le monde des affaires. Plusieurs chercheurs ont pris le relais de cet intérêt et proposent pléthore de définitions générales du Data Mining dont nous pouvons cité par exemple Mr. M. Bater qui définit le Data Mining comme une discipline scientifique qui a pour but l'analyse exploratoire des grandes quantités des données, la découverte des modèles utiles, valides, inattendus ainsi que la connaissance compréhensible dans celles-ci. Outre la découverte de l'information il englobe aussi la collecte, le nettoyage et le traitement des données [7].

Le Data Mining aussi appelé « fouille de données » est une technique consistant à rechercher et extraire une information utile dans un gros volume de données stockées dans des bases de données en utilisant les techniques du Machine Learning [5].

I.1.2. Intelligence Artificiel (IA)

L'intelligence artificielle est difficilement définissable de la même façon que l'intelligence humaine. Le concept d'intelligence est complexe et relatif, et l'expression fortement controversée d' « intelligence artificielle » ne le précise pas d'avantage aussi, il est également difficile de définir la discipline scientifique qu'est l'intelligence artificielle (IA) [2].

L'intelligence artificielle (IA) est un domaine de l'informatique dédiée à la création de matériel et de logiciels capables d'imiter la pensée humaine. Le but principal de l'intelligence artificielle est de rendre les ordinateurs plus intelligents en produisant des logiciels permettant à un ordinateur d'émuler des fonctions du cerveau humain dans des applications définies. L'idée n'est pas de remplacer l'être humain mais de lui donner un outil plus puissant afin de l'aider à accomplir ses tâches [2].

L'Intelligence artificielle est un domaine de l'Informatique qui a pour but de développer des machines (ordinateurs) "intelligentes", c'est-à-dire capables de résoudre des problèmes pour lesquels les méthodes conventionnelles sont inefficaces et inapplicables [2].

I.1.3. Machine Learning

La machine Learning ou apprentissage automatique en français, est un sous domaine de l'intelligence artificielle qui constitue une manière de modéliser des phénomènes, dans le but de prendre des décisions stratégiques mais aussi représenter le comportement d'un phénomène afin de pouvoir directement aider à la résolution d'un problème concret [1].

En Machine Learning, l'idée est que l'algorithme construise tout seul une représentation interne afin de pouvoir effectuer la tâche qui lui est demandée, cette tache peut être la prédiction, l'identification, etc. Pour cela, il va d'abord falloir lui entrer un jeu de données d'exemples afin qu'il puisse s'entraîner et s'améliorer, d'où le mot apprentissage. Ce jeu de données s'appelle le ''Training set''. On peut appeler une entrée dans le jeu de données une instance ou une observation [1].

La figure I.1 représente les différentes étapes qui interviennent dans l'utilisation d'un algorithme de Machine Learning.

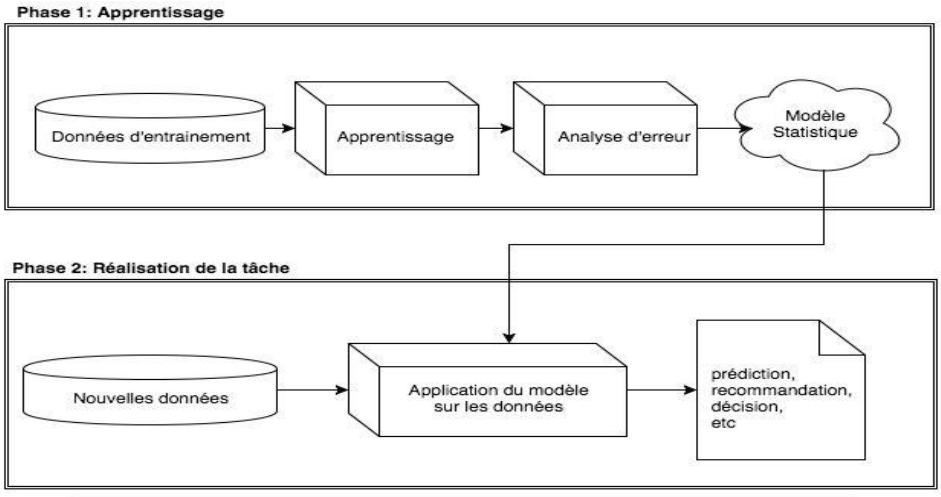

Figure 1-1: Étapes de l'utilisation d'un algorithme de Machine Learning

I.2. TECHNIQUES DU DATA MINING

Les techniques du Data Mining se divisent en 2 groupes dont les techniques supervisées dont un chercheur est là pour guider l'algorithme sur la voie de l'apprentissage en lui fournissant des exemples qu'il estime probants après les avoir étiquetés des résultats attendus ; l'intelligence artificielle apprend alors de chaque exemple, avec pour but, d'être capable de généraliser son apprentissage à de nouveaux cas et les techniques non supervisées dont l'apprentissage par la machine se fait de façon totalement autonome. Des données sont alors communiquées à la machine sans lui fournir les exemples de résultats attendus en sortie [11].

Dans cette partie, il sied important de détailler ces deux groupes ; ainsi, pour les supervisées, le recours à l'arbre de décision, aux règles de décision, à la régression, aux réseaux de neurone, aux Machines à vecteurs supports et aux réseaux bayésiens s'avère important ; quant aux non supervisées, le Clustring hiérarchique, K-means et Carte auto-organisatrice de Kohonen constutueront les détails de ce deuxième groupe.

I.2.1. TECHNIQUES SUPERVISEES

a. Arbre de décision

Un arbre de décision est un ensemble des règles de classification et de régression basant leurs décisions sur des tests associés aux attributs, organisées de manière arborescente. C'est une représentation d'une procédure d'apprentissage. Les notions suivantes sont liées aux arbres de décision : Noeud Principal ou Root Node (c'est l'attribut qui se situe au premier niveau et c'est sur base de lui que s'effectue la prédiction), spliting ou segmentation (c'est un processus consistant à diviser un nœud en 2 ou plusieurs sous nœuds sur base d'un test sur un attribut), nœud interne ou nœud de décision (c'est un nœud étiqueté par un test qui peut être appliqué à toute description d'un individu de la population), une feuille (c'est le nœud où on ne peut plus diviser ou segmenter l'arbre, il est étiqueté par une classe) [4].

Les arbres de décision fonctionnent en séparant de façon récursive la population initiale. Pour chaque groupe, ils sélectionnent automatiquement l'indicateur le plus significatif, le prédicteur qui donne la meilleure séparation par rapport au champ cible. Ils sont peut-être la technique la plus populaire de classification. Une partie de leur popularité, c'est parce qu'ils produisent des résultats transparents qui sont facilement interprétables, offrant un aperçu de l'événement à l'étude. Les résultats obtenus peuvent avoir deux formats

équivalents. Dans un format de règle, les résultats sont représentés dans un langage simple que les règles ordinaires :

SI (VALEURS PREDICTIVES)

ALORS (RESULTAT CIBLE ET SCORE DE CONFIANCE).

Dans une forme d'arborescence, les règles sont représentés graphiquement sous forme d'arbre dans laquelle la population initiale (nœud racine) est successivement divisée en des nœuds terminaux ou feuilles de sous-segments ayant un comportement similaire en ce qui concerne le champ cible [3].

Pour Construire les nœuds de l'arbre, les choix des questions les plus discriminantes peuvent se faire selon plusieurs critères : l'algorithme CART (Classification And Regression Trees) utilise l'indice de Gini, l'algorithme C4.5 utilise l'entropie, l'algorithme Chi-square Automatic Interaction Detector (CHAID) utilise le test de khi carré, et plusieurs autres algorithmes [4].

Avantages et inconvénients des arbres de décisions

Avantage :

- ***Interprétable :*** chaque élément du modèle est facile à comprendre et à analyser pour un humain, et peut donner de l'information sur les données. Ceci est surtout vrai pour les petits arbres.
- ***Flexibilité :*** les arbres de décisions peuvent être utilisés sur des données de n'importe quel type, dont les variables continues et discrètes.

Inconvénient :

Un des inconvénients principaux des méthodes d'apprentissage par arbres de décision est leur instabilité. Sur des données réelles, il s'en faut souvent de peu qu'un attribut soit choisi plutôt qu'un autre et le choix d'un attribut-test, surtout s'il est près de la racine, influence grandement le reste de la construction. La conséquence de cette instabilité est que les algorithmes d'apprentissage par arbres de décision ont une variance importante, qui nuit à la qualité de l'apprentissage. Des méthodes comme le Bagging (pour Bootstrap Aggregating) ou les Random Forests (qui consistent à utiliser plusieurs arbres et utiliser le vote classe faite par chaque arbre pour classer une instance) permettent dans une certaine mesure de remédier à ce problème [4].

b. Règles de décision

Les règles de décision sont un peu semblables aux arbres de décision qui ont été définis précédemment, la seule différence est que les règles de décision peuvent produire plusieurs règles pour chaque enregistrement et pour chaque enregistrement, une seule règle s'applique [3].

Notons que les règles de décision peuvent générer un ensemble de règles de chevauchement. Dans le cas où pour plus d'une règle où on applique des prédictions différentes, sont vrai pour chaque enregistrement, les règles sont évaluées, cela à travers une procédure intégrée, afin de déterminer l'une pour l'évaluation. Ce qui conduit à appliquer une procédure de vote qui combine les règles et les moyennes de leurs confidences individuelles pour chaque catégorie de sortie. La catégorie ayant la confiance moyenne la plus élevée sera sélectionnée à la fin comme la prédiction [3].

c. Régression

La régression est la méthode utilisée pour estimer les valeurs continues, il a comme objectif principal de trouver le meilleur modèle qui décrit la relation entre une variable continue de sortie et une ou plusieurs variables d'entrée. Il s'agit donc de trouver une fonction F qui se rapproche le plus possible d'un scénario donné d'entrées et de sorties [3].

On distingue deux types de régression, la régression linéaire et la régression logistique.

- ***Régression linéaire***

La régression est une technique consistant à prédire la sortie d'une observation en partant d'un certain nombre des variables en entrée. On part d'un certain nombre des données d'apprentissage ou d'un Training Set avec n le nombre d'éléments d'apprentissage. x_1, x_2, x_3... x_n étant les variables d'entrées qui peuvent varier de 1(pour la régression linéaire à une seule variable) à n (pour la régression linéaire à plusieurs variables) [5].

Si « y » est la variable de sortie, on notera alors l'exemple d'apprentissage : (x_1, x_2, x_3... x_n, y) et (x^i_1, x^i_2, x^i_3... x^i_n, y^i) sera la notation du $i^{ème}$ exemple d'apprentissage avec i comme index sur les données.

Le but de la régression c'est de trouver la fonction qui permet de prédire la sortie en fonction de l'entrée, cette fonction doit minimiser l'erreur entre les valeurs de la fonction aux points d'apprentissage et la valeur de sortie dans les données d'apprentissage. Cette fonction est appelée hypothèse qui sera une droite ou un polynôme selon qu'on utilise la régression linéaire ou polynomiale [5].

- ***1ère cas :*** Régression à une seule variable.

  La régression à une seule variable peut être représentée par une fonction de la forme :

  $h_\theta(x) = \theta_0 + \theta_1 x.$

  C'est donc une fonction linéaire de x avec θ_i des paramètres à déterminer.

  La figure 1-2 représente la régression linéaire mono variable.

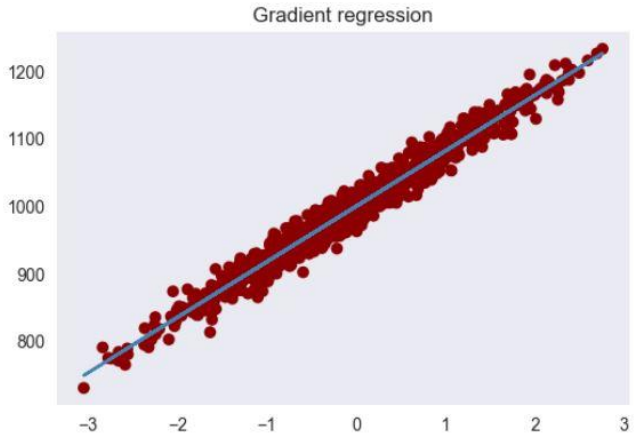

Figure 1-2: Régression linéaire mono variable

- ***2ème cas :*** Régression avec plusieurs variables

  La régression avec plusieurs variables peut être représentée par une fonction de la forme :

  $h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \ldots + \theta_n x_n$

- ***3ème cas :*** Régression polynomiale

  Dans ce cas on utilise un polynôme à n degrés qui est donné par :

  $h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x^2_2 + \theta_3 x^3_3 + \theta_4 x^4_4 + \ldots + \theta_n x^n_n$

- ***Régression logistique***

La régression logistique est une technique de classification qui consiste à appliquer une fonction h sur des éléments x_1, x_2, x_3... x_n en entrée et trouver une sortie discrète « y ». Cette sortie nous permet de déterminer la classe du tuple. Généralement « y » prend 2 valeurs 0 ou 1 et quelque fois -1 ou 1 [5].

On utilise ce type de classification souvent dans la classification des emails qui sont des spams ou non spam, des transactions en ligne de type frauduleux ou pas, dans le crédit bancaire risquant ou non, etc [5].

- ***Erreur de prédiction***

On appelle l'erreur de prédiction, la valeur définie par :

$$j(\theta_1, \theta_2, \dots \theta_n) = \frac{1}{2m} \sum_{=1}^{m} \left[h_\theta(x^i) - (y^i) \right]^2$$

Avec m le nombre des données d'apprentissage et les θ_i des valeurs qui minimisent cette erreur sur toutes les données de l'apprentissage.

Cette fonction représente l'erreur commise lors de la prédiction avec notre hypothèse par rapport à la valeur exacte.

Pour minimiser l'erreur de prédiction, Il existe différentes techniques qu'on peut appliquer à la fonction ci-haut :

- ***L'annulation de la dérivée première :*** cette technique consiste à trouver les valeurs θ_i qui annulent notre dérivée première.
  Notons que cette méthode ne convient pas pour les données d'apprentissage avec plusieurs attribues et plusieurs données [5].
- ***La descente du gradient*** *:* elle consiste à effectuer plusieurs itérations sur les valeurs de θ_i jusqu'à trouver celle qui minimise l'erreur, c'est-à-dire jusqu'à ce qu'il converge vers zéro [5].

d. Les Réseaux de neurone

Les réseaux de neurones sont des algorithmes d'apprentissage automatique qui utilisent des fonctions de cartographie complexe, non linéaire pour l'estimation et classification. Ils

sont constitués de neurones organisés en couches, où la couche d'entrée contient les prédicteurs ou neurones d'entrée et la couche de sortie est le champ cible.

Dans cette technique, la procédure de formation est un processus itératif, les enregistrements en entrée, avec des résultats connus, sont présentés sur le réseau et la prédiction du modèle est évaluée par rapport aux résultats observés. Les erreurs observées seront utilisées pour ajuster et optimiser les estimations du poids initial, ils sont considérés comme des solutions opaques ou "Boîte Noire", car ils ne fournissent pas une explication de leurs prédictions mais plutôt une analyse de sensibilité, qui résume l'importance prédictive des champs d'entrée. Cette méthode nécessite une connaissance statistique selon le problème à résoudre et un temps de traitement long pour la formation [3]. La figure 1-3 présente un réseau des neurones organisés en couche.

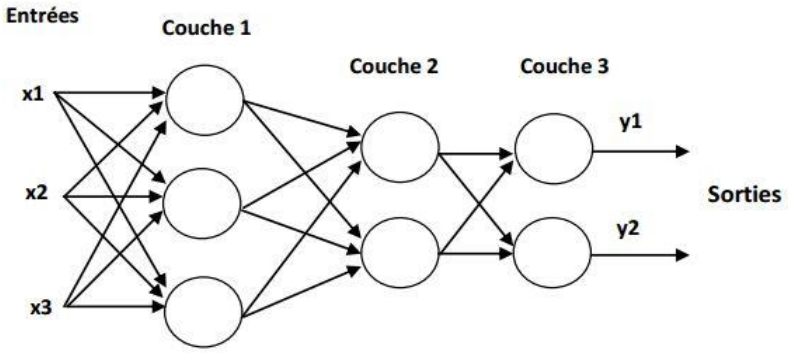

Figure 1-3:Réseaux des neurones organisés en couche

e. Machines à vecteurs supports (SVM)

Les Machines à vecteurs supports sont des algorithmes de classification qui peuvent modéliser les profils de données non linéaires hautement complexes, et d'éviter les sur-apprentissages. Les données d'entraînement d'entrée sont transformées de manière appropriée par les fonctions du noyau non linéaires et cette transformation est suivie d'une recherche de fonctions plus simples, c'est-à-dire des fonctions linéaires, qui enregistre de façon optimale distincte. Les analystes expérimentent généralement avec différentes fonctions de transformation et de comparer les résultats.

Cette technique est efficace et exigeant, en termes de ressources de mémoire et de temps de traitement mais il manque de transparence puisque les prévisions ne sont pas expliquées et seulement l'importance des prédicteurs est résumée [3].

Le SVM est un classeur discriminant qui est défini par un hyperplan, étant donné un ensemble d'apprentissage l'algorithme de SVM permet de trouver un hyperplan qui va catégoriser les éléments de l'ensemble et d'autres nouveaux éléments en maximisant la marge entre les éléments de l'ensemble d'apprentissage à la manière de la frontière de décision pour la régression logistique. On peut dire que SVM est une amélioration de la régression logistique [6]. Deux types des données peuvent s'observer ici : les données séparables linéairement et les données non séparables linéairement où intervient la notion de Kernel.

f. Réseaux bayésiens

Les modèles bayésiens sont des modèles de probabilité qui sont souvent dans des problèmes de classification. Ils ont pour but d'estimer la probabilité d'occurrences. Ce sont des modèles graphiques qui permettent de fournir une représentation visuelle des relations d'attributs tout en assurant la transparence, et une explication de la justification du modèle [3].

I.2.2 TECHNIQUES NON SUPERVISEES

a) Clustring hiérarchique

C'est une méthode qui commence avec une solution où chaque enregistrement comprend un groupe et peu à peu, les groupes se forment jusqu'au point où tous tombent dans un super-cluster. À chaque étape, il calcule les distances entre toutes les paires d'enregistrements et les groupes les plus similaires. Les étapes de regroupement et les distances respectives sont résumées dans une table appelée horaire d'agglomération ou dans un graphique dit dendrogramme [3]. La figure 1-4 présente la technique clustring hierarchique.

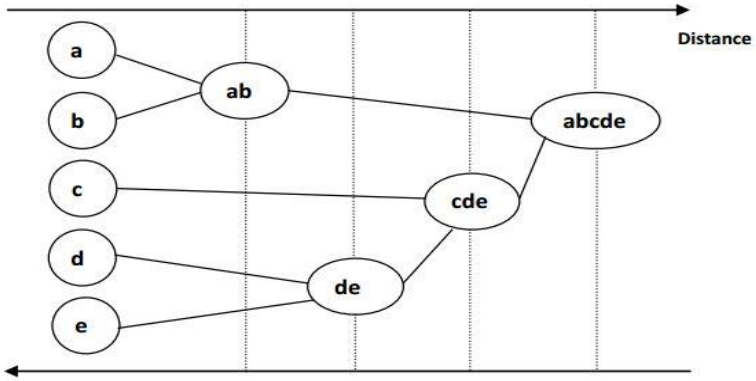

Figure 1-4: Clustring hiérarchique

b) K-means

C'est une méthode efficace, c'est l'algorithme de segmentation le plus rapide qui peut gérer plusieurs enregistrements et des ensembles de données larges donc de nombreuses dimensions de données et des champs d'entrée.

Il s'agit d'une technique de segmentation basé sur la distance et qui n'a pas besoin de calculer les distances entre toutes les paires d'enregistrements différemment de l'algorithme hiérarchique. Notons que dans cette méthode, le nombre de grappes doit être formés et est prédéterminé par l'utilisateur à l'avance. Habituellement, un certain nombre de solutions différentes doit être jugé et évalué avant d'approuver le plus approprié [3].

c) Carte auto-organisatrice de Kohonen

Les réseaux de Kohonen ou carte auto-organisatrice de Kohonen sont basés sur des réseaux de neuronaux. Ils produisent typiquement une grille à deux dimensions ou une carte des grappes ou une carte d'auto-organisation. Par rapport à la méthode K-means, celle-ci prend généralement plus de temps à la réalisation [3].

I.3. METHODOLOGIES DE MIS EN PLACE DU PROCESSUS DE DATA MINING

Le processus d'exploration consiste à découvrir, au moyen de vastes ensembles de données, des modèles, des relations et des informations qui guident les entreprises dans la mesure et la gestion de leur situation actuelle et dans la prévision de leur avenir.

Une grande quantité de données et de bases de données peuvent provenir de diverses sources de données et peuvent être stockées dans différents entreposeurs de données, et des techniques d'exploration de données telles que l'apprentissage automatique, l'intelligence artificielle et la modélisation prédictive peuvent être impliquées [9].

Le processus d'exploration de données nécessite un engagement. Mais les experts s'accordent pour dire que, dans tous les secteurs, le processus d'exploration de données est identique et devrait suivre un chemin prescrit. Voici les 6 étapes essentielles du processus d'exploration de données qui sont présentées à la figure 1-5 [9].

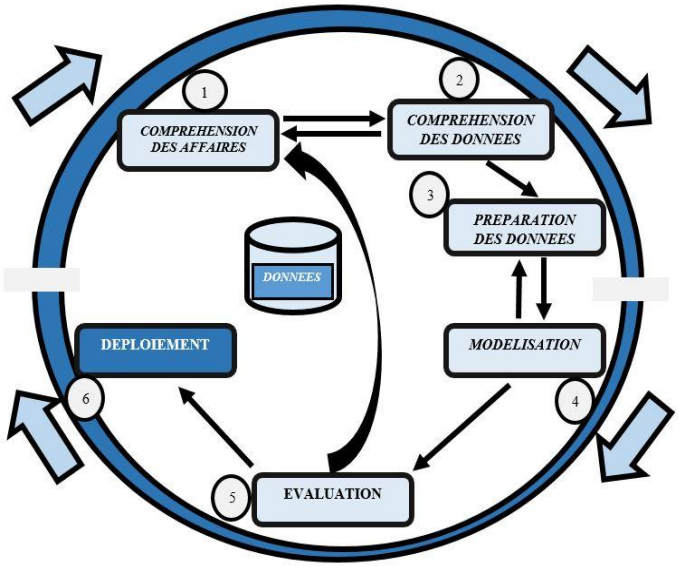

Figure 1-5: Modèle du processus de Data Mining

1. **Compréhension des affaires (Business Understanding)**

Dans la plus part de cas, il est indispensable de comprendre la signification des données et le domaine à explorer. Sans cette compréhension, aucun algorithme ne va donner un résultat fiable. C'est avec la compréhension du problème qu' on peut alors préparer les données nécessaires à l'exploration et interpréter correctement les résultats obtenus [3].

Dans la phase de compréhension des affaires :

- Tout d'abord, il est nécessaire de bien comprendre les objectifs de l'entreprise et déterminer ses besoins.
- Ensuite, évaluez la situation actuelle en recherchant les ressources, les hypothèses, les contraintes et autres facteurs importants à prendre en compte.
- Ensuite, à partir des objectifs commerciaux et des situations actuelles, créez des objectifs d'exploration de données afin d'atteindre les objectifs commerciaux dans la situation actuelle.
- Enfin, un bon plan d'exploration de données doit être mis en place pour atteindre à la fois les objectifs commerciaux et l'exploration de données. Le plan devrait être aussi détaillé que possible [9].

2. **Compréhension des données (Data Understanding)**

Sachant la définition du problème et des objectifs du data Mining, on peut avoir une idée sur les données qui doivent être utilisées. Ces données n'ont pas toujours le même format et la

même structure. On peut donc avoir des textes, des bases de données, des pages web, ...etc. Parfois, on est amené à prendre une copie d'un système d'information en cours d'exécution, puis ramasser les données de sources éventuellement hétérogènes (fichiers, bases de données relationnelles, temporelles, ...) [3].

- La phase de compréhension des données commence par la collecte initiale des données, cette opération est exécutée à partir des sources de données disponibles, afin de se familiariser avec les données. Certaines activités importantes doivent être exécutées, notamment le chargement et l'intégration des données, afin de réussir la collecte des données.
- Les propriétés des données acquises doivent être examinées avec soin et consignées.
- Les données doivent être explorées en abordant les questions d'exploration de données, qui peuvent être abordées à l'aide de requêtes, de rapports et de visualisation.
- La qualité des données doit être examinée en répondant à des questions importantes telles que «Les données acquises sont-elles complètes?», «Y a-t-il des valeurs manquantes dans les données acquises [9] ?

3. Préparation des données (Data preparation)

Les données peuvent contenir plusieurs types d'anomalies ; celles-ci peuvent être omises à cause des erreurs de frappe ou à cause des erreurs dues au système lui-même, dans ce cas il faut remplacer ces données ou éliminer complètement leurs enregistrements [3].

Des données peuvent être incohérentes c'est-à-dire qui sortent des intervalles permis, on doit les écarter ou les normaliser. Parfois on est obligé à faire des transformations sur les données pour unifier leur poids. Le prétraitement comporte aussi la réduction des données qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration. Dans la plus part de cas, le prétraitement doit préparer des informations globales sur les données pour les étapes qui suivent telles que la tendance centrale des données (moyenne, médiane, mode), le maximum et le minimum, le rang, les quartiles, la variance, ... etc. Plusieurs techniques de visualisation des données telles que les courbes, les diagrammes, les graphes,... etc, peuvent aider à la sélection et le nettoyage des données. Une fois les données collectées, nettoyées et prétraitées on les appelle entrepôt de données ou data warehouse [3].

Le résultat de la phase de préparation des données est l'ensemble de données finales. Une fois les sources de données disponibles identifiées, elles doivent être sélectionnées, nettoyées, construites et formatées sous la forme souhaitée. La tâche d'exploration de données plus en profondeur peut être effectuée au cours de cette phase pour remarquer les modèles basés sur la compréhension de l'entreprise [9].

4. **Modélisation (Modeling)**

Dans cette étape, on doit choisir la bonne technique pour extraire les connaissances (exploration) des données. Des techniques telles que les réseaux de neurones, les arbres de décision, les réseaux bayésiens, le Clustering, ... sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat [3].

- Premièrement, les techniques de modélisation doivent être sélectionnées pour être utilisées pour le jeu de données préparé.
- Deuxièmement, le scénario de test doit être généré pour valider la qualité et la validité du modèle.
- Troisièmement, un ou plusieurs modèles sont créés sur le jeu de données préparé.
- Enfin, les modèles doivent être soigneusement évalués en impliquant les parties prenantes afin de s'assurer que les modèles créés correspondent aux initiatives commerciales [9].

5. **Évaluation**

Lors de la phase d'évaluation, les résultats du modèle doivent être évalués dans le contexte des objectifs commerciaux de la première phase. Au cours de cette phase, de nouvelles exigences opérationnelles peuvent être soulevées en raison des nouveaux modèles découverts dans les résultats du modèle ou d'autres facteurs. Obtenir une compréhension des affaires est un processus itératif dans l'exploration de données. La décision d'aller ou de ne pas aller doit être prise à cette étape pour passer à la phase de déploiement [9].

6. **Déploiement (Deployment)**

Les connaissances ou informations acquises au cours du processus d'exploration de données doivent être présentées de manière à ce que les parties prenantes puissent les utiliser quand elles le souhaitent. Selon les besoins de l'entreprise, la phase de déploiement peut être aussi simple

que la création d'un rapport ou aussi complexe qu'un processus d'exploration de données pouvant être répété dans l'ensemble de l'entreprise. Au cours de la phase de déploiement, les plans de déploiement, de maintenance et de surveillance doivent être créés pour la mise en œuvre, ainsi que pour les futurs supports. Du point de vue du projet, le rapport final du projet doit résumer les expériences du projet et examiner le projet afin de déterminer la nécessité d'améliorer les enseignements tirés [9].

Ces 6 étapes décrivent le processus standard intersectoriel pour l'exploration de données, appelé CRISP-DM (Cross Industry Standard Process for Data Mining). Il s'agit d'un modèle de processus standard ouvert qui décrit les approches courantes utilisées par les experts en exploration de données. C'est le modèle d'analyse le plus largement utilisé.

I.4. DATA MINING DANS L'OCTROI ET REMBOURSEMENT DES CRÉDITS

I.4.1. L'Intelligence artificielle dans l'octroi des crédits

L'intelligence artificielle pouvant être défini comme un ensemble des technologies destinées à imiter le cerveau humain, intervient pour assister l'être humain dans le domaine de la banque. Les banques ont été même des promoteurs dans l'utilisation de l'intelligence artificielle, cela dès les années 80 avec les systèmes experts. Ces deniers sont des systèmes capables de simuler les comportements d'un expert humain et d'analyser un risque d'octroi de crédit aux particuliers, aux professionnels ou aux entreprises, C'est donc un outil d'aide à la décision [12].

Ces outils sont donc précieux afin d'avoir une meilleure maîtrise du risque. En effet, lorsqu'une banque octroie un prêt avec un fort risque de crédit (risque qu'un emprunteur ne rembourse pas la totalité ou une partie de son crédit à la date prévue), elle doit accroitre le provisionnement de ses fonds propres, ce qui est très couteux. Le recours aux systèmes experts est alors totalement justifié [12].

Pour arriver à produire un outil de prédiction, une expertise est réalisée avec des analystes financiers ou des experts du risque pour formaliser les règles et ensuite une série de critères (comme la situation professionnelle de l'emprunteur, son ancienneté, son revenu, son type de contrat de travail, etc.) sont déterminés pour former un recueil de la connaissance qui sera implanté dans le système d'analyse. Ce même système pourra alors analyser instantanément des milliers de règles métiers pour donner un niveau de risque. C'est après cela que le conseil

de l'établissement pourra décider d'octroyer le crédit ou non, en suivant les recommandations de l'outil expert [12].

L'IA permet donc d'obtenir une décision optimale en très peu de temps (entre 5 à 20 fois plus vite qu'avec un processus classique). Ainsi, les banques sont à même de réduire les coûts, améliorer la compétitivité et fidéliser les clients en apportant des réponses rapides et adaptées au dossier des emprunteurs [12].

I.4.2. Outil d'aide à la décision dans l'octroi des crédits : Crédit Scoring

a. Définitions :

- ✓ Un outil d'aide à la décision (OAD) est un outil qui est conçu pour simuler le raisonnement d'un expert risque et donne un niveau de risque pour chaque situation en suivant le raisonnement de l'expert.

- ✓ Credit Scoring ou un pointage de crédit est une expression numérique basée sur une analyse de niveau des dossiers de crédit d'une personne, afin de représenter la solvabilité d'un individu. Un pointage de crédit est principalement basé sur une information de rapport de crédit généralement fournie par les agences d' évaluation du crédit . Les prêteurs, tels que les banques ou les coopératives d'épargne et des crédits utilisent les scores de crédit pour évaluer le risque potentiel que représente le fait de prêter de l'argent aux consommateurs et pour atténuer les pertes dues aux créances irrécouvrables [12].

  En outre, le terme Crédit Scoring désigne un ensemble d'outils d'aide à la décision utilisés par les organismes financiers pour évaluer le risque de non remboursement des prêts. Un score est une note de risque, ou une probabilité de défaut [13].

b. Crédit Scoring

La décision de l'octroi du crédit repose sur l'évaluation préalable du décideur. En effet, ce dernier, en fonction de son expérience, de l'évaluation de la solvabilité du client et de son historique de remboursement, parvient à distinguer intuitivement le bon au mauvais client avant de présenter le dossier au comité de crédit. Il doit donc prendre son temps pour tirer de ses expériences (un sixième sens) lui permettant de distinguer les bons clients des

mauvais. Par contre, son humeur du jour, ses sentiments, son degré d'aversion pour le risque, etc. peuvent influer sur sa décision pouvant le conduire à sélectionner le mauvais client et laisser le bon client. D'où on fait recourt au pointage de crédit [13].

La technique de Credit Scoring, définit aussi par Mark Scheiner comme étant l'évaluation statistique se basant sur l'utilisation de connaissances quantitatives des résultats de remboursement et des caractéristiques des prêts remboursés dans le passé et enregistrés dans une base de données électronique afin de pronostiquer les résultats de remboursement des futurs prêts, consiste à attribuer une note à chaque information collectée que l'on compare à une fiche de référence préalablement établie. Ensuite le total des notes de chaque information collectée donne une appréciation sur la nature « bon » ou « mauvais », « risqué » ou « non risqué » du client considéré [13].

On distingue le Scoring avant décaissement, le Scoring après décaissement, le Scoring du recouvrement, le Scoring du risque de départ du client, etc. Dans le cas de ce travail, nous serons plus intéressés par le Scoring avant décaissement qui semble plus pertinent, car il permet de distinguer en amont les bons clients des mauvais [13].

Principe de fonctionnement du Scoring :

La théorie du Scoring part du principe selon lequel le passé est le meilleur estimateur du futur. Son fonctionnement efficace nécessite une démarche basée sur plusieurs étapes essentielles comme:

- La collecte et l'exploitation des données ;
- La définition de la fiche d'évaluation statistique ou fiche de notation;
- La classification du risque ou cotation
- Le test avec données historiques ;
- La validation de la fiche de notation

Néanmoins, les implications techniques de sa mise en place sont nombreuses. Mais elle ne peut réussir que lorsque les procédures d'octroi de prêt sont formalisées et que l'institution dispose d'une base de données fournie sur l'historique des prêts surtout individuels. Elle est donc beaucoup plus adaptée aux prêts de type individuel [13].

Pour construire le modèle Scoring, il faut :

- Disposer d'une importante base de données fiable sur l'historique des clients pour l'élaboration de la fiche d'évaluation et beaucoup de données pour chaque prêt;
- Que les données utilisées dans l'évaluation statistique soient d'une quantité suffisante ;
- L'intervention d'un expert dans le domaine des statistiques et implique donc des coûts de gestion supplémentaires ;
- Que le système d'évaluation statistique soit intégré dans le Système d'Information et de Gestion d'où la nécessité d'avoir un système automatisé ;
- Que l'évaluation statistique soit basée sur le passé, c'est-à-dire l'historique de remboursement des clients pour prédire le futur.

Comment faire la collecte de données et faire le choix du modèle :

La définition de la fiche d'évaluation se base sur la collecte des données qui peuvent influencer le comportement du remboursement de l'emprunteur. Ces données collectées sur le passé doivent renseigner les caractéristiques de l'emprunteur, du prêt et du prêteur. Ainsi, les données de l'emprunteur concernent ses caractéristiques démographiques, ses coordonnées, son activité professionnelle, le ménage auquel il appartient et les avoirs de la famille, les flux financiers du ménage et de son entreprise, ses antécédents de remboursement (arriérés de payement, expérience de l'emprunteur), etc. Pour les caractéristiques du prêt, il s'agit du montant demandé, le montant décaissé, le taux d'intérêt, du délai de remboursement, du montant des remboursements, du différé et, le cas échéant, des types de garantie et leur valeur, etc. [13].

Toutes ces informations doivent être disponibles dans la base de données de l'institution afin que le Scoring puisse en tenir compte lors de l'élaboration de la fiche d'évaluation.

I.5. LES PRINCIPALES APPLICATIONS ET LES PRINCIPAUX ALGORITHMES DE DATA MINING

Data Mining est principalement utilisé de nos jours par les entreprises fortement axées sur le consommateur, organisations de vente au détail financier, de communication et de marketing, cela pour approfondir leurs données transactionnelles et déterminer les prix, les préférences des clients et le positionnement du produit, leur impact sur les ventes, leur satisfaction, et les bénéfices des entreprises. Avec l'exploration de données, un détaillant peut utiliser les

enregistrements des achats effectués par les clients au point de vente pour développer des produits et des promotions susceptibles de séduire des segments de clientèle spécifiques [10].

Voici la liste des quelques autres domaines importants dans lesquels l'exploration de données est largement utilisée:

1. **Future Healthcare**

L'exploration de données offre un potentiel considérable pour améliorer les systèmes de santé. Il utilise des données et des analyses pour identifier les meilleures pratiques permettant d'améliorer les soins et de réduire les coûts. Les chercheurs utilisent des méthodes d'exploration de données telles que des bases de données multidimensionnelles, l'apprentissage automatique, l'informatique douce, la visualisation de données et les statistiques. L'exploitation minière peut être utilisée pour prédire le volume de patients dans chaque catégorie. Des processus sont développés pour s'assurer que les patients reçoivent les soins appropriés au bon endroit et au bon moment. L'exploration de données peut également aider les assureurs de soins de santé à détecter les fraudes et les abus [10].

2. **Analyse du panier de marché**

L'analyse du panier de marché est une technique de modélisation basée sur une théorie selon laquelle, si vous achetez un certain groupe d'articles, vous êtes plus susceptible d'acheter un autre groupe d'articles. Cette technique peut permettre au détaillant de comprendre le comportement d'achat d'un acheteur. Ces informations peuvent aider le détaillant à connaître les besoins de l'acheteur et à modifier la présentation du magasin en conséquence. En utilisant une analyse différentielle, la comparaison des résultats entre différents magasins, entre clients de différents groupes démographiques peut être effectuée [10].

3. **Éducation**

Un nouveau domaine émergent, appelé Educational Data Mining, concerne le développement de méthodes permettant de découvrir des connaissances à partir de données provenant d'environnements éducatifs. Les objectifs de l'EDM sont les suivants: prédire le comportement futur des élèves en matière d'apprentissage, étudier les effets du soutien pédagogique et faire progresser les connaissances scientifiques sur l'apprentissage. L'exploration de données peut être utilisée par une institution pour prendre des décisions précises et également pour prédire

les résultats de l'étudiant. Avec les résultats, l'institution peut se concentrer sur ce qu'il faut enseigner et comment l'enseigner. Les habitudes d'apprentissage des élèves peuvent être capturées et utilisées pour développer des techniques pour les enseigner [10].

4. Ingénierie de fabrication

La connaissance est le meilleur atout qu'une entreprise manufacturière posséderait. Les outils d'exploration de données peuvent être très utiles pour découvrir des modèles dans des processus de fabrication complexes. L'exploration de données peut être utilisée dans la conception au niveau système pour extraire les relations entre l'architecture de produit, le portefeuille de produits et les données relatives aux besoins du client. Il peut également être utilisé pour prédire la durée, le coût et les dépendances de développement du produit, entre autres tâches [10].

5. CRM (Customer Relationship Management ou Gestion de la relation de client)

La gestion de la relation client consiste essentiellement à acquérir et à fidéliser des clients, à améliorer leur fidélité et à mettre en œuvre des stratégies axées sur le client. Pour entretenir de bonnes relations avec un client, une entreprise a besoin de collecter des données et de les analyser. C'est ici que l'exploration de données joue son rôle. Avec les technologies de Data Mining, les données collectées peuvent être utilisées pour l'analyse. Au lieu de se demander où se concentrer pour fidéliser le client, les demandeurs de la solution obtiennent des résultats filtrés [10].

6. Détection de fraude

Des milliards de dollars ont été perdus à cause de la fraude. Les méthodes traditionnelles de détection des fraudes prennent du temps et sont complexes. L'exploration de données aide à fournir des modèles significatifs et à transformer les données en informations. Toute information valable et utile est un savoir. Un système de détection de fraude parfait devrait protéger les informations de tous les utilisateurs. Une méthode supervisée comprend la collecte de spécimens d'enregistrements. Ces enregistrements sont classés frauduleux ou non frauduleux. Un modèle est construit en utilisant ces données et l'algorithme est créé pour identifier si l'enregistrement est frauduleux ou non [10].

7. Détection d'intrusion

Toute action susceptible de compromettre l'intégrité et la confidentialité d'une ressource constitue une intrusion. Les mesures défensives pour éviter une intrusion incluent l'authentification de l'utilisateur, les erreurs de programmation et la protection des informations. L'exploration de données peut aider à améliorer la détection d'intrusion en ajoutant un niveau de concentration à la détection d'anomalie. Il aide un analyste à distinguer une activité de l'activité réseau courante au quotidien. L'exploration de données permet également d'extraire des données plus pertinentes pour le problème [10].

8. Détection de mensonge

Il est facile d'appréhender un criminel, alors que faire ressortir la vérité est difficile. Les forces de l'ordre peuvent utiliser des techniques minières pour enquêter sur les crimes et surveiller la communication des terroristes présumés. Ce fichier inclut également l'exploration de texte. Ce processus cherche à trouver des modèles significatifs dans les données qui sont généralement du texte non structuré. Les échantillons de données collectés lors d'enquêtes précédentes sont comparés et un modèle de détection de mensonges est créé. Avec ce modèle, des processus peuvent être créés en fonction des besoins [10].

9. Segmentation de la clientèle

Les études de marché traditionnelles peuvent nous aider à segmenter les clients, mais l'exploration de données va plus loin et augmente l'efficacité du marché. L'exploration de données aide à aligner les clients dans un segment distinct et peut adapter les besoins en fonction des clients. Le marché consiste toujours à fidéliser les clients. L'exploration de données permet de trouver un segment de clients sur la base de vulnérabilités et l'entreprise peut leur proposer des offres spéciales et améliorer leur satisfaction [10].

10. Banque financière

Avec les opérations bancaires informatisées partout, une quantité énorme de données doit être générée avec de nouvelles transactions. L'exploration de données peut contribuer à résoudre les problèmes commerciaux des secteurs bancaire et financier en détectant des schémas, des liens de causalité et des corrélations dans les informations commerciales et les prix du marché qui n'apparaissent pas immédiatement aux gestionnaires car les données de volume sont trop

volumineuses ou générées trop rapidement pour être examinées par des experts. Les responsables peuvent trouver ces informations pour mieux segmenter, cibler, acquérir, fidéliser et fidéliser un client rentable [10].

11. Surveillance d'entreprise

La surveillance d'une entreprise est la surveillance du comportement d'une personne ou d'un groupe par une entreprise. Les données collectées sont le plus souvent utilisées à des fins de marketing ou vendues à d'autres sociétés, mais sont également régulièrement partagées avec des agences gouvernementales. Il peut être utilisé par l'entreprise pour adapter ses produits à ses clients. Les données peuvent être utilisées à des fins de marketing direct, telles que les publicités ciblées sur Google et Yahoo, les publicités étant destinées à l'utilisateur du moteur de recherche en analysant l'historique de leurs recherches et leurs courriels [10].

12. Analyse de la recherche

L'histoire montre que nous avons assisté à des changements révolutionnaires dans la recherche. L'exploration de données est utile pour le nettoyage des données, le prétraitement des données et l'intégration des bases de données. Les chercheurs peuvent trouver dans la base de données des données similaires susceptibles d'apporter des changements dans la recherche. L'identification de toute séquence concomitante et la corrélation entre toutes les activités peuvent être connues. La visualisation des données et l'exploration visuelle des données nous fournissent une vue claire des données [10].

13. Enquête criminelle

La criminologie est un processus qui vise à identifier les caractéristiques de la criminalité. En réalité, l'analyse du crime comprend l'exploration et la détection des crimes et de leurs relations avec les criminels. Le volume élevé d'ensembles de données sur la criminalité et la complexité des relations entre ces types de données ont fait de la criminologie un domaine approprié pour l'application des techniques d'exploration de données. Les rapports de criminalité basés sur du texte peuvent être convertis en fichiers de traitement de texte. Ces informations peuvent être utilisées pour effectuer un processus d'appariement du crime [10].

14. Bio-Informatique

Les approches de Data Mining semblent idéales pour la bio-informatique, car elles sont riches en données. L'exploitation de données biologiques permet d'extraire des connaissances utiles d'énormes séries de données rassemblées en biologie et dans d'autres domaines connexes des sciences de la vie, tels que la médecine et les neurosciences. Les applications de l'exploration de données à la bio-informatique comprennent la découverte de gènes, l'inférence de fonction protéique, le diagnostic de la maladie, le pronostic de la maladie, l'optimisation du traitement de la maladie, la reconstruction de réseaux de protéines et d'interactions géniques, le nettoyage de données et la prédiction d'emplacement sous-cellulaire de protéines [10].

I.6. CONCLUSION PARTIELLE

Cette partie, a été consacrée aux techniques de Data Mining et il a été constaté qu'ils sont décomposés en deux catégories, la première se compose des techniques supervisées et la deuxième celle des techniques non supervisées. Dans ces deux catégories, il existe de nombreuses techniques avec des caractéristiques différentes, chacune avec des points forts et des points faibles. Ces informations seront utilisées pour choisir les techniques nécessaires afin de résoudre les problèmes précis cités dans ce travail. La présentation du processus de Data Mining suivie de quelques applications de ce dernier a été faite.

CHAPITRE II. PROCÉDURE DU SERVICE DE PRÊT ET REMBOURSEMENT DANS UNE COOPÉRATIVE D'ÉPARGNE ET DE CRÉDIT

Le crédit constitue la source de production par excellence de la coopérative. Sans crédit, une coopérative d'épargne et de crédit ne peut fonctionner, elle sera incapable de supporter ses charges. Il est donc une nécessité pour une coopérative d'octroyer des crédits en vue de sa survie. Les fonds affectés aux crédits proviennent des fonds propres de la COOPEC et d'une partie des épargnes de ses membres.

C'est pourquoi la procédure de réaliser une garantie d'un membre qui ne paye pas conformément au contrat de prêt est texte à part pour la protection de ces fonds des membres de la COOPEC.

II.1. CONDITIONS D'ACCES AU CREDIT

Pour bénéficier d'un crédit à la COOPEC BONNE MOISSON, le membre devra réunir les conditions ci-après :

a) Être majeur et juridiquement capable ;
b) Être membre depuis 3 mois au moins ;
c) Exercer régulièrement des mouvements de dépôts et de retraits durant les trois derniers mois à son compte dont la moyenne mensuelle des versements est supérieure à la tranche mensuelle de remboursements ;
d) Déposer une demande de crédit et remplir un formulaire de demande de crédit accompagné du document de garantie valable (certificat d'enregistrement, contrat de location, contrat de garantie salaire,…) ; pour les personnes morales, ce formulaire doit être précédé par une lettre de demande de crédit des répondants et/ou un procès-verbal contenant la décision pour la personne morale de solliciter le crédit ;
e) Avoir une garantie morale et matérielle suffisante ou une caution solidaire ;
f) Avoir participé à une séance d'éducation coopérative et financière organisée par la COOPEC ;
g) Avoir une activité génératrice des revenus ou une source sûre des revenus ;
h) Avoir pris connaissance profondément des conditions d'octroi de crédit ;
i) Être prêt à descendre sur terrain en cas d'appel par le service approprié ;

j) Annexer à la garantie, l'acte de cautionnement signé par le (la) conjoint(e) si le membre est marié.

k) Avoir remboursé le dernier crédit et les intérêts y relatifs.

II.2. DIFFERENTS TYPES DE CREDITS ET MODE D'OCTROI

A. Le crédit express :

C'est une facilité de caisse ouverte aux personnes physiques et morales qui, dans l'urgence, se trouvent dans l'obligation de dédouaner leurs marchandises, de payer des salaires ou de faire face à des besoins pressants à caractère économique ou autres.

Durée : 60 jours maximum

Taux par nombre des jours :

- 1 à 30 jours : 4%

- 1 à 60 jours : 4% le 1er mois et 4% le 2ème mois.

- Pour un crédit express de 60 jours remboursable en deux tranches, les intérêts seront calculés comme suit :

- 1ère tranche : 4% du montant total;
- 2ème tranche : 4% du montant restant dû

- *Dépassé l'échéance, le montant est frappé des pénalités ou intérêts de retard. Ces intérêts de retard sont les intérêts ordinaires augmentés de 4% c'est-à-dire 8% en tout.*

Mode d'étude du dossier :

- Déposer la demande de crédit et remplir le formulaire de demande d'un crédit express

- Après analyse du service technique, le gérant sollicite l'aval du Président de la Commission de crédit et si possible faire participer les autres membres de cette Commission de crédit et du Président du Conseil d'Administration pour prise de connaissance ;

- Octroi par passation des écritures.

B. Les crédits ordinaires

Ils sont appelés ordinaires parce qu'ils suivent le cours normal de traitement par la COOPEC quant à la procédure, la durée, le but et le plafond préétablis.

Le demandeur introduit sa demande de crédit à la COOPEC. Le remboursement du capital est constant et les intérêts sont payés sur le capital restant dû. Ils sont contractés plus pour les activités commerciales, l'amélioration de l'habitat, l'équipement, l'agriculture, l'agro-pastoral, les cas sociaux et l'entrepreneuriat/artisanat.

- ***Commerce***,
- ***Le crédit Habitat :*** Ce crédit est accordé aux membres qui désirent se procurer des terrains, améliorer leurs habitations ou construire des immeubles et autres infrastructures.
- ***Le crédit agro-pastoral :*** Ce crédit est accordé aux membres agriculteurs professionnels et ceux qui font l'élevage des animaux de la basse-cour, des ovins et des bovins.
- ***Le crédit scolaire :*** La COOPEC Bonne Moisson finance les besoins de scolarité de ses membres et/ou de leurs enfants. Ces besoins de scolarité comprennent les frais scolaires/académiques, les fournitures scolaires, les livres, les exigences scolaires en habillement et équipement tels les uniformes, salopettes, outillages,…
- ***Les crédits sociaux :*** Ces crédits concernent les besoins relatifs aux faits sociaux des membres comme le mariage/fête, les soins médicaux, les frais funéraires/deuil, les besoins alimentaires, le voyage,…

Durée du crédit : 18 mois maximum c'est-à-dire 18 mensualités. En effet, le crédit peut être demandé pour une durée d'1 mois, de 2 mois, de 3 mois, …, de 18 mois.

Mode de remboursement :

Le crédit est remboursé dans le respect des échéances convenues dans le contrat de prêt :

- Autant de fois par rapport à la durée du crédit ;
- Par tranches mensuelles.

Actuellement, le remboursement par tranches mensuelles est le mode le plus privilégié par la COOPEC.

Chaque mois compté, date à laquelle le compte du membre a été crédité du prêt sert de référence, le membre est débité de la tranche mensuelle du prêt et des intérêts calculés sur le montant restant dû. Il va sans dire qu'à l'octroi du crédit, le membre verse au préalable les frais préliminaires c'est-à-dire ceux relatifs au montage et à l'étude du dossier de crédit.

Au premier mois qui suit, le membre rembourse la première tranche et les intérêts du montant du crédit, les intérêts du deuxième remboursement sont calculés sur le montant total dû et sont retenus au même moment que la deuxième tranche, et ainsi de suite. La mensualité remboursée avec retard engendre des intérêts de retard calculés au prorata des jours en retard à partir du lendemain de l'échéance.

Taux des crédits : Ils varient selon le but comme l'indique le tableau 2-1 :

Tableau 2-1: Types de crédits ordinaires [25]

N°	But (Affectation)	Taux d'intérêt normal	Taux d'Intérêts de retard	Total
1	Commerce	3%	3%	6%
2	Artisanat ou entreprenariat	2,5%	2,5%	5%
3	Equipement	2,5%	2,5%	5%
4	Amélioration de l'habitat	2,5%	2,5%	5%
5	Agriculture et Elevage	2%	2%	4%
6	Crédit scolaire	2%	2%	4%
7	Mariage et fêtes	2,5%	2,5%	5%
8	Soins médicaux	2%	2%	4%
9	Consommation alimentaire	2%	2%	4%
10	Frais funéraires	2%	2%	4%
11	Voyage	3%	3%	6%
12	Autres cas sociaux (vol, incendie, emprisonnement,…)	2%	2%	4%

N.B. : La durée du crédit ordinaire peut être revue à la demande des membres sur décision du Conseil d'Administration si la santé financière de la COOPEC le permet.

C. Le crédit immobilier du personnel

Le crédit immobilier est accordé aux agents ayant accompli au moins 3 ans de service au sein de la COOPEC. La demande est adressée à la Commission de crédit. Le crédit est accordé au taux bonifié de 1% sur une période maximale de 3 ans. Les pièces relatives à l'achat du terrain ou à la construction sont confiées à la COOPEC BONNE MOISSON jusqu'au remboursement intégral du crédit. La demande de crédit est étudiée par la Commission de crédit et validée par le Conseil d'administration. En cas de démission ou de révocation de l'agent, le décompte final est retenu en premier lieu et la réalisation de la garantie pour solde de crédit.

D. Les crédits à caution solidaire

Ils sont accordés aux personnes n'ayant pas encore disposé de garantie hypothécaire. L'emprunteur doit présenter au moins trois membres garants actifs et n'ayant pas de crédits impayés. Un membre ne peut parrainer qu'un seul crédit, il en est libéré qu'au remboursement du crédit qu'il a avalisé.

E. Le crédit solidaire

Il est accordé au groupe de personnes n'ayant pas encore disposé de garantie hypothécaire. Idem que ci-dessus, les avaliseurs sont les membres du groupe et sont solidairement responsables du crédit. Le tableau 2-2 présente les différentes tailles des groupes solidaires.

Tableau 2-2: Tailles des groupes solidaires [25]

Catégorie	Nombre des membres	Taux	Intérêts de retard	Total
I	3 à 5	2%	2%	4%
II	6 à 15	2%	2%	4%
III	16 à 25	2%	2%	4%
IV	CECI/ AVEC	2%	2%	4%
V	SACCOS/VICOBA/ Mutuel	2%	2%	4%

CECI= Caisse d'épargne et de crédit interne

SACCOS= Saving and Credit Cooperative Society

AVEC= Association Villageoise d'Épargne et de Crédit

VICOBA= Village Community Bank

Durée : 12 mois maximum

Pénalités : 3%

Un groupe solidaire doit préalablement présenter son statut notarié et une hypothèque.

N.B. : - En cas de remboursement avec retard, il ne doit pas être question d'apurer d'abord le principal pour ensuite régulariser les intérêts ordinaires et les intérêts de retard. Tout avis de débit relatif à la récupération d'un impayé doit nécessairement contenir un montant retenu du principal dû, des intérêts réguliers non encore perçus et des intérêts de retard.

- La valeur et la véracité du motif de retard n'excluent pas le paiement des intérêts ordinaires et/ou des intérêts de retard à courir, sauf pour les cas de force majeur et après délibération du Conseil d'Administration.

II.3. MONTAGE DU DOSSIER DE CREDIT

a. Étude Préliminaire du dossier de crédit

La procédure de montage du dossier de crédit commence par la vérification systématique sur terrain des éléments fournis par le demandeur. Cette vérification s'effectue par les agents de crédit.

De manière spécifique, ces agents vérifient :

- Le but du crédit sollicité et la faisabilité du projet à financer ;
- La situation socio-économique et professionnelle du demandeur ;
- La résidence du demandeur ;
- L'adresse de l'activité du membre ;
- L'originalité du document de garantie ;
- La vérification d'autres éléments liés à la moralité, à la personnalité, à la sincérité et à l'honnêteté du demandeur.

L'étude préliminaire intervient juste après le dépôt de la demande et la signature du formulaire de demande par le membre afin d'éviter une longue durée pour décider de la recevabilité ou non du dossier. Cette étude ne doit pas dépasser 3 jours.

b. De la descente sur terrain

Lors de la décente sur terrain, les agents se munissent de la fiche de renseignement sur le membre et les activités de celui-ci.

L'agent doit éviter à faire un rapport de complaisance. Sur ce, la fiche de renseignement doit être sincère et exacte. Le moyen de transport lors de la descente est assuré par la COOPEC.

c. De l'examen de la Garantie

Tout prêt doit être entièrement garanti et faire l'objet d'un contrat de prêt irrévocable. Ce contrat est établi après l'analyse et la décision de la commission de crédit.

II.4. TYPES DE GARANTIES

- ***L'Épargne à terme*** *:* si le montant du crédit et les intérêts sont inférieurs à l'épargne et si l'échéance de l'épargne dépasse la durée du crédit ;
- ***Le Fonds de garantie*** *:* si ce fond est disponible dans la COOPEC et couvre totalement le montant du crédit et les intérêts y afférents jusqu'à l'échéance;
- ***Le Salaire*** *:* si l'employeur de l'emprunteur accepte la retenue mensuelle de la tranche de remboursement du crédit (principal plus intérêts) équivalant au tiers du salaire au moins pendant toute la durée du crédit;
- ***La Caution solidaire*** *:* si les avaliseurs acceptent d'être solidairement responsables du crédit sollicité par le membre ou le groupe ; si les avaliseurs sont membres de la COOPEC, ceux-ci doivent être actifs et sans crédit impayé ;
- ***L'Hypothèque*** *:* si sa valeur estimée est supérieure d'au moins 25% du montant du crédit.

a. **Documents acceptés par la COOPEC BONNE MOISSON comme garantie :**

- ***L'épargne à terme:*** le contrat du dépôt à terme
- ***Le fonds de garantie:*** le contrat de partenariat ou l'engagement du partenaire sur les fonds logés en compte ;
- ***Le salaire:*** l'attestation de retenue sur salaire, par lequel l'employeur est tenu responsable du remboursement de la totalité du montant du crédit et des intérêts y afférents pendant ou non l'exécution du contrat de l'employé envers son employeur. Pour un employé de la COOPEC Bonne Moisson, il doit changer l'hypothèque avant la résiliation du contrat

- ***La caution solidaire:*** l'attestation tenant lieu de caution solidaire dûment signée par les Avaliseurs et/ou les mandataires ;
- ***L'hypothèque:*** le certificat d'enregistrement, le contrat de location dont la validité est de plus d'une année, la fiche d'occupation parcellaire.

b. **Plafonds des crédits couverts par les titres :**

- *La fiche d'occupation parcellaire:* un crédit d'au plus 1000$;

- *Le contrat de location:* un crédit d'au plus 15 000$;

- *Le certificat d'enregistrement*

c. **Conditions d'acceptation d'un titre :**

À la réception du document, l'agent de crédit vérifie:

- l'authenticité du document matérialisant le titre de propriété foncière notamment les signatures réglementaires autorisées ainsi que la légalisation des signatures, assisté par le conseiller foncier;
- la conformité des identités ;
- si le titre porte le nom du membre ; dans le cas contraire, il devra y joindre un acte de cession du véritable propriétaire légalisé par le Notaire ;
- l'existence réelle de la propriété et procède à son évaluation ;
- la signature de l'avaliseur et le sceau de l'entreprise (employeur du demandeur) pour des crédits à garantie salariale ;
- le mouvement d'épargne de 3 mois pour tout crédit sollicité, la moyenne mensuelle des versements et la fréquence des épargnes du membre pendant l'échéance du crédit précédent.
- Si la valeur estimée de l'hypothèque dépasse d'au moins 25% le montant à octroyer.

d. Mécanisme de gestion des hypothèques

Ce sont les agents de crédit qui jugent du dépôt ou non de la garantie hypothécaire : les dossiers paraissant éligibles à l'octroi sont ceux pour lesquels la garantie hypothécaire est exigée.

Les dispositions ci-après sont observées par la COOPEC en matière de gestion des hypothèques :

- À l'issu de l'investigation, le membre dépose la garantie hypothécaire à la COOPEC ;

- L'hypothèque doit être présentée en original;

- L'hypothèque est consignée dans un Registre des hypothèques ;

- Un accusé de réception de l'hypothèque est remis au membre ;

- Le membre prendra soin de conserver une copie de sa garantie avant son dépôt à la COOPEC car toute manipulation de la garantie pendant toute la durée du crédit est strictement interdit ;

- La garantie est remise au membre après qu'il est apuré totalement le crédit ainsi que tous les intérêts et frais éventuels y afférents ;

- En cas de non remboursement du crédit par le membre, sa garantie est réalisée en suivant la procédure établie par la COOPEC.

Remarque :

- Aucun prêt n'est accordé sans l'étude préalable de la Commission de crédit.
- Pour la garantie hypothécaire, l'acte de cautionnement signé par le (la) conjoint(e) est exigé si le membre est marié au cas où le régime matrimonial est de « la communauté universelle des biens ».

e. Calcul de la rentabilité des activités du demandeur

La rentabilité représente le rapport entre les revenus du membre et les sommes qu'il a mobilisées pour les obtenir. Elle mesure la capacité des capitaux investis par le membre à dégager un certain niveau de profit. Elle est calculée en considérant le rapport entre le résultat net et les capitaux investis.

L'agent de crédit doit s'imprégner des charges engagées par les demandeurs et des produits pour dégager le résultat net.

f. De l'étude proprement dite

L'étude proprement dite est faite par la Commission de crédit. Avant d'orienter les dossiers, l'Agent de crédit procède à des investigations pour :

-Vérifier à nouveau les informations fournies par le membre et la garantie et rechercher des informations complémentaires ;

- Confirmer, par un acte, la conformité des documents et la sincérité des éléments fournis par l'investigation.

La Commission de crédit statue sur le dossier et sa décision est inscrite dans le procès-verbal de délibération. Après délibération, les membres de la Commission signent un procès- verbal des assises.

Certains dossiers peuvent nécessiter une visite au membre pour confirmer ou compléter les informations. Dans ce cas, la Commission décide du moment de la visite, désigne la personne chargée et des éléments à examiner. La décision finale ne pourra être prise qu'après le compte rendu de la visite.

À chaque réunion de la Commission de crédit, le secrétaire dresse un procès-verbal de délibération signé par les membres présents et reprenant tous les dossiers étudiés et tous les aspects traités. Les dossiers sont remis au Gérant pour exécution après prise de connaissance du Président du Conseil d'Administration.

Note importante : C'est la situation de la trésorerie qui dicte le volume de crédit à octroyer.

g. Élaboration du Contrat de prêt

Le contrat de prêt est un document reprenant les modalités d'octroi et de remboursement des prêts. C'est un engagement signé entre le prêteur (COOPEC) d'une part, et l'emprunteur (membre) d'autre part pour le crédit accordé ou reçu pendant une durée bien définie.

Le contrat de prêt est signé par le gérant et l'emprunteur.

h. Des frais

- ***Les frais d'étude***

Les frais d'étude du dossier sont à la charge du membre emprunteur et se retiennent avant l'octroi du crédit. Ils sont fixés comme suit:

- Inférieur ou égal à 1000$: 1% du montant octroyé
- Plus de 1 000$: 0,75%du montant octroyé.

- ***Les fonds d'appui au crédit :*** Les frais d'appui au crédit sont de 0,05 % du montant octroyé au membre. Ces frais sont aussi retenus pour renforcer les fonds propres de la COOPEC et augmenter sa capacité financière. La variation de ce taux est décidée par le Conseil d'Administration.

i. Éducation coopérative après étude favorable du crédit

Avant décaissement de fonds, les membres éligibles au crédit sont invités pour une séance d'éducation coopérative orientée vers la gestion de crédit.

j. De l'octroi de crédit

L'octroi de prêt est la dernière étape pour le processus de demande de prêt et il est la première pour la procédure de remboursement.

Avant la signature du contrat par les parties, la COOPEC :

- Organise la formation d'éducation coopérative au membre
- Calcule les intérêts à payer avec les tranches de remboursement;
- Établit un tableau d'amortissement du prêt (plan de remboursement) ;

Après la signature du contrat par les parties, la COOPEC :

- Vérifie si les frais d'étude et d'appui au crédit ont été retenus ;
- Établit un avis de crédit en 4 exemplaires dont les copies seront remises respectivement au membre, dans son dossier, à la comptabilité et au service de crédit pour le classement;
- Passe les écritures de prêt en créditant le compte du membre ;
- enregistre les écritures sur la fiche du membre et dans le livret d'épargne ;
- Remet les copies du plan de remboursement et du contrat au membre ;
- Classe les dossiers du membre avec tous les éléments.

N.B. : La signature du contrat de prêt par le membre donne à la COOPEC l'autorisation de récupérer sur son compte d'épargne le montant dû à l'échéance même à l'absence du membre.

II.5. REMBOURSEMENT ET RECOUVREMENT DU CREDIT

Pour le crédit en cours, l'emprunteur est rappelé deux jours avant chaque échéance par téléphone ou par message téléphonique ou en dur.

Au cas où il ne s'acquitte pas à l'échéance, un autre rappel lui sera envoyé par la même voie pour le recouvrement deux jours après l'échéance de la tranche échue.

Deux jours après, l'agent de crédit doit visiter l'emprunteur à son lieu de travail ou à domicile.

Si le membre ne rembourse pas, l'agent de crédit se fait appuyer par le Gérant et la Commission de crédit. Au cas où le dossier persiste, il est soumis au Conseil d'Administration pour dispositions.

Si le membre ne paie pas toujours, son dossier est remis entre les mains du conseiller juridique. Entretemps, les agents de crédit et la Commission de crédit renforcent les recouvrements.

Pour le crédit échu ou la tranche échue, l'emprunteur doit recevoir une note de rappel à la fin de chaque mois tout en lui indiquant les intérêts normaux et de retard ou pénalités et cela à chaque échéance. Cette note de rappel est déposée à mains propres et à domicile moyennant l'accusé de réception de l'intéressé.

Le conseiller juridique doit l'inviter pour lui signifier le risque qu'il court en cas de non remboursement. Il pourra déposer son dossier au tribunal au cas où l'emprunteur ne s'acquitte toujours pas de sa dette. Les frais de justice sont à la charge de l'emprunteur.

II.6. PROCESSUS DE GESTION DES IMPAYES

- Les relances : elles se traduisent par des appels téléphoniques, des visites et de l'envoi des lettres ;
- La gestion à l'amiable : Il s'agit des entretiens téléphoniques et physiques, de la prise de connaissance des mobiles de la défaillance et des stratégies proposées pour aider les membres à trouver une solution durable au dénouement de la défaillance ;
- La gestion par contrainte : elle se traduit par la saisie de la justice, c'est-à-dire, le débiteur est mis en demeure de payer ses dettes ;
- La vente de la garantie offerte : Cette procédure est à épuiser avant la décision d'abandon du crédit par le Conseil d'Administration sur proposition de la Commission de crédit.

II.7. LE PROVISIONNEMENT DES CREDITS

a. Causes de provisionnement des crédits

Malgré leur faible pourcentage, les emprunteurs de mauvaise foi ne peuvent pas manquer dans la coopérative. Aussi, une calamité naturelle, des conséquences malheureuses des faits politiques, la pauvreté subite due à des maladies chroniques, la disparition (le décès dans le sens juridique) sont autant des causes qui justifient le provisionnement des crédits.

b. Type des Crédits à provisionner

À l'instar de l'Instruction n°003 de la BCC aux COOPECs ainsi qu'aux microfinances relative à la classification et au provisionnement des crédits, la COOPEC BONNE MOISSON provisionne les crédits litigieux dont :

- ***Les crédits prorogés* :** leurs échéances ont été modifiées à la demande du membre avant l'échéance, la prorogation n'étant autorisée qu'une seule fois ;
- ***Les crédits impayés* :** une échéance est en retard de remboursement pendant plus d'un jour ;
- ***Les crédits douteux* :** le retard de remboursement vient de dépasser 31 jours bien que le crédit soit couvert par des garanties ;
- ***Les crédits irrécupérables* :** en retard de plus de 12 mois, comptabilisés en perte à la clôture de l'année, mais suivis-en hors bilan.

c. Du montant, de la procédure et de la période de provisionnement

➢ ***Montant***

Le montant de la provision portera sur les crédits jugés non recouvrables et sera déterminé comme suit :

- De 1 à 30 jours de retard : 5% du montant restant dû
- De 31 à 60 jours de retard : 25% du montant restant dû
- De 61 à 90 jours de retard : 50% du montant restant dû
- De 91 à 180 jours de retard : 75% du montant restant dû
- Plus de 180 jours de retard : 100% du montant restant dû

➢ ***Procédure de provisionnement***

- Le service de crédit transmettra au gérant le rapport des crédits accompagné d'une liste complémentaire des crédits litigieux;
- Après vérification, le gérant en fera rapport à la commission de crédit en vue de confirmer que le reste des crédits sont normaux et un procès-verbal sur le provisionnement sera dressé ;
- La commission de crédit soumettra le rapport assorti d'un plan de recouvrement au Conseil d'Administration à sa première convocation qui suit sa délibération.

➢ ***Période de provisionnement***

Le provisionnement des crédits sera fait chaque fin du mois.

d. De la radiation des crédits

Seront radiés :

- Le crédit qui dépasse plus de 365 jours de retard ;
- Le crédit d'un membre décédé pendant une année sans avoir un membre se portant garant de ce crédit ;
- Le crédit d'un membre frappé d'interdit.

La procédure de la radiation des crédits sera faite chaque fin d'année.

e. De la gestion de la défaillance du membre

- Le membre écrit à la COOPEC et explicite les raisons de sa défaillance au Gérant ;
- Le Gérant transmet le dossier à la Commission de crédit pour étude ;
- Le dossier est soumis au Conseil d'Administration pour délibération.
- Ne peuvent constituer des raisons d'atténuation de la charge au membre due au retard de remboursement :
    - L'incendie des biens ;
    - Le vol ;
    - La maladie prolongée, …

II.8. OUTILS DE GESTION DES CREDITS

Pour la gestion efficace des crédits, la COOPEC BONNE MOISSON retient les documents ci-après pour tout le processus du crédit c'est-à-dire de l'expression du besoin du crédit par le membre jusqu'à son remboursement. Il s'agit de :

- ***1ère étape: Montage du dossier***
    - Lettre de demande de crédit pour les personnes morales ;
    - Formulaire de demande de crédit ;
    - Registre des demandes de crédit ;
    - Procès-verbal d'investigation ;
    - Garantie offerte;
- ***2ème étape: Étude du dossier et Octroi ou non du crédit***
    - Procès-verbal de délibération de la Commission de crédit sur la demande de crédit ;
    - Procès-verbal de la réunion de la Commission de crédit ;
    - Prise de connaissance du Président du Conseil d'Administration ;
    - Contrat de prêt ou classement du dossier en cas de refus;

- Registre des crédits octroyés ;
- Note de crédit ;
- Fiche de crédit ;
- Fiche d'épargne ;
- Carnet d'épargne ;

- ***3ème étape : Remboursement du crédit***
  - Note de débit ;
  - Fiche de crédit ;
  - Fiche d'épargne ;
  - Carnet d'épargne ;
  - Registre des crédits remboursés

II.9. CONCLUSION PARTIELLE

Cette partie, a été consacrée à la présentation de la procédure du service de prêt et remboursement dans une coopérative d'épargne et de crédit (cas de la Coopéc Bonne Moisson).

Les Conditions dont la Coopéc met en place pour qu'un client accède à un crédit en les différenciant les uns aux autres ont été détaillées. Également la présentation de la manière dont la Coopéc procède à un montage du dossier de crédit, les types des garanties, les remboursements et le recouvrement, enfin les processus de gestion des impayés tout en soulignant sur les outils des gestions des crédits ont été mis en exergue.

CHAPITRE III. PRESENTATION ET EXPLOITATION DES DONNÉES

III.1.INTRODUCTION

Dans cette partie de notre travail, il est question de présenter et exploiter les données qui nous ont été fournies par le département de crédit de la Coopérative d'Épargne et des Crédits Bonne Moisson et cela dans le cadre de contribuer à la réalisation de cette étude. Le domaine de micro-crédit étant sensible à la confidentialité, nous n'avons donc pas eu accès à leur système des gestions des données, néanmoins, l'agent responsable du service des crédits nous a donné un fichier sous format Excel (.xlsx) contenant quelques données brutes sans identités des clients.

III.2.PRESENTATION DES DONNEES

Les données à notre disposition sont structurées dans un fichier Excel comme souligné dans la partie introductive de ce chapitre. Cette structure peut être comparée à une matrice dont des colonnes contiennent différent types des données liées au compte du client. C'est dont une structure appelée Data frame.

Notre matrice a une dimension de 1303 lignes et 12 colonnes au départ et après la suppression des données dupliquées nous avons 825 lignes et 12 Colonnes.

Le tableau 3-1 nous aide à présenter les données colonne par colonne :

Tableau 3-1: Présentation des données

N°	Attribut	Description
01	DATE	Contient la date à laquelle le client a pris le crédit
02	SEXE	Contient le sexe du client. Pour les hommes nous avons la notation H, pour les femmes F et pour la personne morale nous avons PM.
03	CATÉGORIE	Cette colonne représente la catégorie du membre qui peut être soit un Administrateur dans la Coopéc (noté ADM), soit un simple agent de la Coopéc (noté AG) ou un autre membre simple (noté AM).
04	AGE	L'âge du membre qui est noté en intervalle (par exemple : 19-25 ans) et pour le personne morale l'entrée contient la notation PM.
05	PRÊTS	Cette colonne contient le montant dont le client a emprunter.
06	REMBOURSEMENT	C'est la colonne contenant le montant dont le client a déjà remboursé à la Coopéc.
07	RESTE	C'est le montant restant après la tranche remboursée
08	ÉCHÉANCE	Échéance qui est en mois, c'est la durée pendant laquelle le client doit épuiser ou finir de payer le prêt.
09	GARANTIE	La garantie que laisse le client lors de la demande de crédit, sous forme de caution.
10	AFFECTATION	L'affectation de l'argent demandé, c'est le travail que va faire le prêt pris par le client.
11	ACTIVITÉ	Représente l'activité que fait le client dans la vie (ex : Commerçant, salarié, chauffeur, etc.)
12	**JRS DE RETARD**	Cette colonne contient le nombre de jours de retard qu'a le client pour payer ou le nombre de jours qui manquent pour qu'il soit en retard. Par exemple si c'est écrit 10 Jours c'est-à-dire que le client a dépassé de 10 jours pour payer mais si c'est écris -10 Jours c'est-à-dire que le client est en ordre pour la tranche en cours mais il doit payer la tranche suivante dans 10 Jours

III.2.PREPARATION DES DONNEES

Cette partie de notre travail sera consacrée au nettoyage des données. Nous allons faire les traitements des valeurs aberrantes et les valeurs manquantes.

III.2.1. Définition et technique à utiliser

a. Données Manquantes

Les données manquantes ou valeurs manquantes, apparaissent lorsqu'une observation d'une variable n'a pas de valeur [14].

Selon Little, R.J.A., and Rubin, D.B. (2002) [14], il y a trois mécanismes distincts de valeurs manquantes :

- Valeur manquante entièrement au hasard (*MCAR, Missing Completely At Random*) : le fait de ne pas avoir la valeur pour une variable, est indépendant des autres variables.
- Valeur manquante au hasard (*MAR, Missing At Random*) : le fait de ne pas avoir la valeur pour une variable, est dépendant seulement des valeurs observées.
- Valeur ne manquant pas au hasard (*NMAR, Non Missing At Random*) : le fait de ne pas avoir la valeur pour une variable ne dépendant que des valeurs manquantes.

Néanmoins, il existe trois stratégies possibles pour traiter des données manquantes :

- Utilisation des procédures de suppression : qui peut se faire par l'étude de cas complets et par Étude de cas disponible.
- Utilisation des procédures de remplacement, (substitution) des données manquantes par les valeurs présentes : qui peut se faire par l'imputation par la moyenne, l'imputation par régression, l'imputation par hot-deck, l'imputation par k-plus proche voisins, l'imputation multiple.
- Utilisation des procédures de modélisation de la distribution des données manquantes et les estimer par certains paramètres : qui peut se faire soit par le maximum de vraisemblance, ou soit par la maximisation d'espérance.

b. Données aberrantes

Les données aberrantes sont définies comme étant des données qui ne sont pas en accord avec la majorité des données. Elles peuvent être causées soit par une raison physique connue

(par exemple l'erreur d'écriture des données, le mauvais étalonnage de l`appareil de mesure...) ou par une raison non connue [14].

Pour repérer les données aberrantes, un contrôle de cohérence sur les données est nécessaire, l'usage du bon sens et de l'expérience est le plus sûr [14].

c. Les données mal orthographiées :

Les données mal orthographiées sont des données causées souvent par une erreur de frappe lors de la saisie ou incohérence des lettres majuscule et minuscule. Cela arrive souvent dans le cas où les données sont issues d'un système manuel.

III.2.2. Données en entrées

A ce stade nous traitons les problèmes avec nos données en entrées qui sont repris dans les colonnes suivantes : Date, Sexe, Catégorie, Âges, Prêts, Échéances, Garantie, Affectation et Activité. Nous parcourons donc ces différentes colonnes pour traiter les données manquantes, aberrantes et les données mal orthographiées.

En faisant le tour l'ensemble de notre Dataset, nous remarquons :

- *L'attribut Date :* présence de plusieurs données manquantes et nous avons décidé de l'effacer.
- *L'attribut Sexe :* nous avons remarqué que certaines observations avaient pour sexe « P » au lieu de « PM », tous deux voulant dire « Personne Morale » et « M » pour Masculin à la place de « H » pour homme. Ces données aberrantes ont été remplacées automatiquement.
- *L'attribut Age :* certaines observations présentaient « PM » au lieu d'un intervalle que nous avons pris en compte lors de l'encodage. Notons aussi que nous avons également traité certains intervalles qui étaient incluses dans d'autres.
- *L'attribut Échéance :* Pour nous faciliter les calculs dans la suite, nous avons décidé de rendre *float* cette attribut en commençant par supprimer le suffixe ''Mois''.
- D'autres attributs n'avaient que des données mal orthographiées que nous avons corrigées automatiquement.

III.2.3. Données en sorties

Dans cette partie nous allons concentrons sur le traitement des données manquantes, aberrantes et mal orthographiées dans la colonne de Remboursement, Reste et Jour de retard, qui représente nos données en sortie.

- Les attributs Remboursement et Reste n'avaient pas des données manquantes et aberrantes.
- L'attribut Jour de retard : Pour nous faciliter les calculs dans la suite, nous décidons de rendre *float* cette attribut en commençant par supprimer le suffixe ''Jours''.

Nous construisons une étiquette de nos données en fonction du Reste que nous avons mis en % de par le montant prêter, le montant Remboursé, de l'échéance et du nombre de jours de retard. L'Étiquette est donc notre donnée principale de sortie et comprend -1 si le client est crédible et 1 si le client n'est pas crédible. Le tableau 3-2 montre les données avec le nouvel entête.

Tableau 3-2: Présentation des données avec l'étiquette

Sexe	Categorie	Age	Prêts	Rembourssement	Reste	Garantie	Affectation	Activité	NbrJrDeRetard	EcheanceNew	RestePercentage	Etiquette
H	AM	31-40	500	356.0	144.0	Caution	Commerce	Commerçant	321.0	12.0	28.80	1
H	AM	61-70	2500	2407.0	93.0	Certificat d'enregistrement	Commerce	Commerçant	132.0	12.0	3.72	1
H	AM	41-50	8000	5436.0	2564.0	Contrat de location	Commerce	Commerçant	140.0	12.0	32.05	1
F	AM	31-40	1500	1238.0	262.0	Salaire	Commerce	Entrepreneur	130.0	12.0	17.47	1
H	AM	31-40	600	488.0	112.0	Fiche d'occupation	Commerce	Chauffeur	40.0	12.0	18.67	1

III.3. ANALYSE DES DONNEES

Cette partie sera consacrée à une analyse des données en utilisant la statistique descriptive. Nous allons donc faire une analyse statistique uni variée et une analyse statistique bi variée et nous allons visualiser les résultats par des graphiques à l'aide des certaines librairies de python.

III.3.1. Analyse Uni variée

Dans cette partie, nous analysons chaque colonne de son côté. Pour les données qualitatives nous utilisons les diagrammes circulaires et/ou les diagrammes de bar et pour les données

quantitatives nous avons visualisé les données par des histogrammes. Notons que les librairies Pandas et Matplotlib de python nous facilite la tâche à ce point.

a. ***Sexe :*** En analysant les données de cette colonne à l'aide des librairies citées ci-hauts, nous remarquons que les Hommes représentent 63.7% de notre échantillon, les femmes 24.5% et les personnes morales 11.8%, comme nous pouvons le visualiser sur la figure 3-1 :

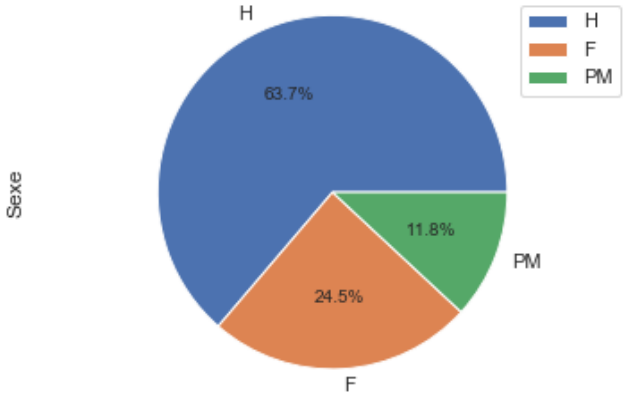

Figure 3-1: Diagramme de cercle de la distribution du sexe

b. ***Catégorie :*** Nous remarquons que nous n'avons que 3 catégories dont AM à 92.5%, AG à 5.0% et ADM à 2.5%. Cette colonne n'a pas des données mal orthographié, ni des données manquantes ou aberrantes.

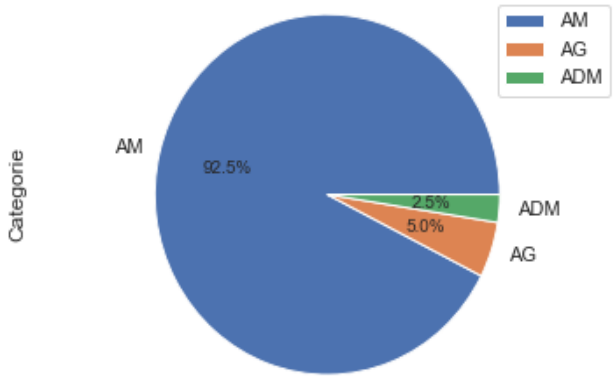

Figure 3-2: Diagramme de cercle de la distribution des catégories par le diagramme de cercle

c. ***Age :*** Le diagramme de bâton suivant montre la distribution de l'âge dans notre échantillon tout en sachant que l'encodage d'âges sera numérique et prendra en compte les PM :

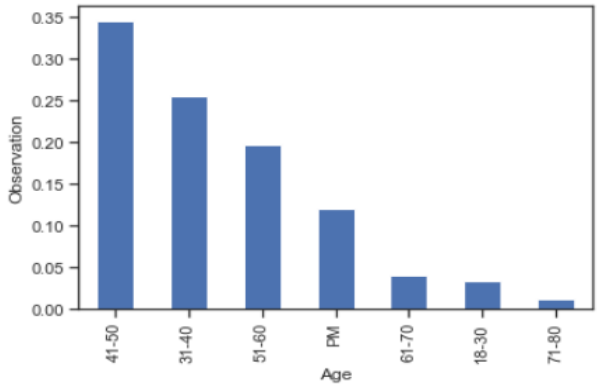

Figure 3-3: Diagramme de barre de la distribution de l'Age

d. ***Prêt, Remboursement et Reste :*** Les fugues (Histogrammes) suivantes visualisent les distributions de notre échantillon :

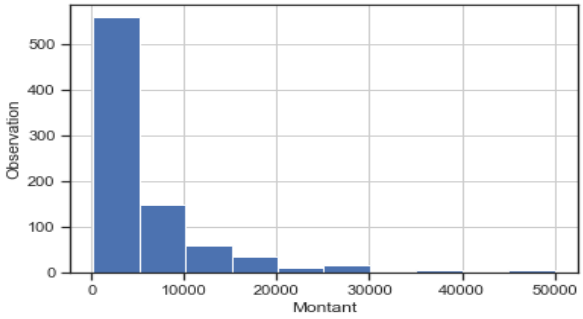

Figure 3-4: Histogramme de Prêt

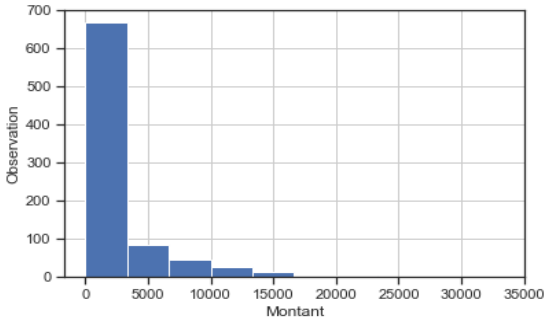

Figure 3-5: Histogramme de Remboursement

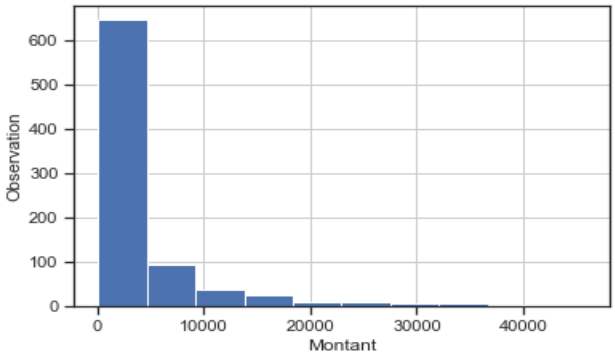

Figure 3-6: Histogramme de Reste

e. ***Échéance :*** le diagramme de barre ci-après nous aide à visualiser la distribution des échéances sur notre population:

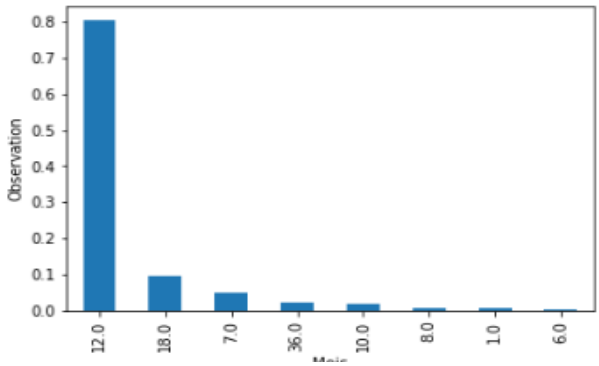

Figure 3-7: Diagramme de barre de l'échéance

f. ***Garantie :*** Nous constatons que 34.7% de notre population utilise le salaire comme garantie, 34.9% le certificat d'enregistrement, 14.8% le contrat de location, 9.9% la fiche d'occupation parcellaire, et 5.7% la caution. Nous pouvons visualiser cette distribution sur la figure suivante :

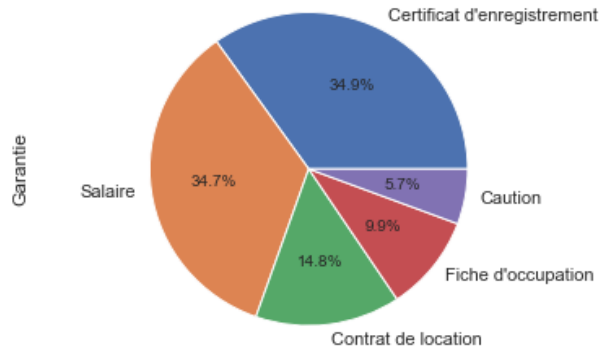

Figure 3-8: Diagramme de cercle de l'attribut Garantie

g. ***Affectation :*** 43.6% de notre échantillon affecte leur argent de prêt dans l'amélioration de l'habitation, 39.2% dans le commerce, 5.9% dans la scolarité, 5.6% dans la consommation, 4.6% achat des équipements, et 1.2% Autres affectation. Nous pouvons le voir sur la figure 3-9 la distribution des affectations :

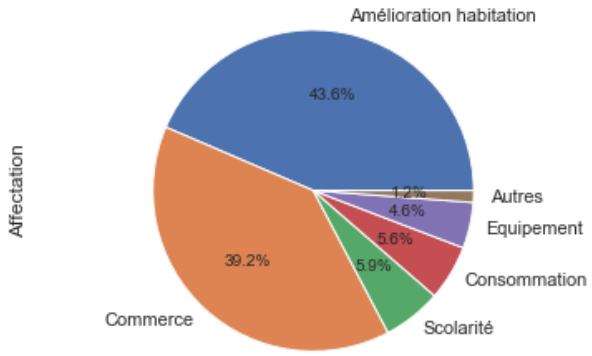

Figure 3-9: Diagramme de cercle de l'attribut Affectation

h. **Activité :** Nous avons fait aussi le traitement des données mal orthographiées comme pour les attributs précédents et le résultat peut être visualisé sur le diagramme de barre à la figure 3-10 :

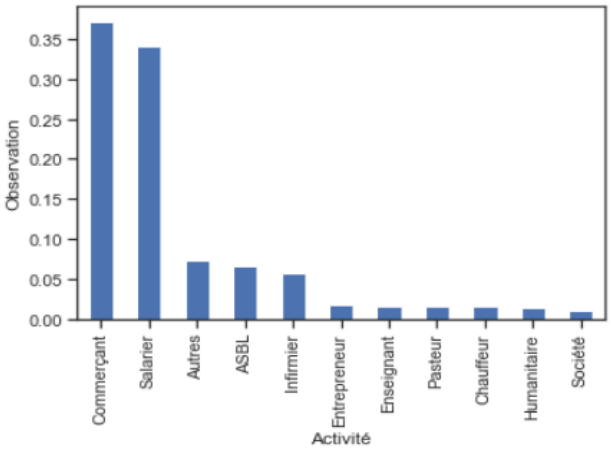

Figure 3-10: Diagramme de barre de l'attribut Activité

III.3.2. Analyse Bi-Variée

Dans cette partie nous allons faire une technique d'analyse statistique des données qui consiste à trouver la relation pouvant exister entre deux variables, dans le but de tester l'hypothèse d'association et de causalité entre ces derniers. Nous allons identifier et définir la nature de la relation qui existe entre deux variables et déterminer si la relation est importante du point de vue statistique en calculant l'intervalle de confiance [15].

Notre Dataset contient des données catégorielles ou qualitatives et des données quantitatives ou continues, nous allons donc étudier les relations qu'ont les attributs avec notre Étiquette qui est la donnée principale de sortie.

a. ***Pour étudier la relation entre notre étiquette qui est numérique et les variables catégorielles, nous utilisons le test ANOVA :*** Le but de l'analyse de la variance (ANOVA) est de tester la présence de différences significatives ou non entre des moyennes. Le cas classique le plus simple est celui de la comparaison de deux moyennes provenant d'échantillons indépendants [16].
b. ***Pour étudier la relation entre notre étiquette qui est numérique et d'autres variables numériques ou quantitatives, nous avons utilisé le coefficient de corrélation de Pearson:*** La corrélation est une quantification de la relation linéaire entre des variables

continues. Le calcul du coefficient de corrélation de Pearson repose sur le calcul de la covariance entre deux variables continues [17].

Dans le cadre de ce travail, pour implémenter ANOVA et calculer le coefficient de Pearson, nous avons utilisé la librairie *scipy* de python.

Le tableau 3-3 présente les relations entre les attributs et les Étiquettes qui représentent notre donnée de sortie principale.

Tableau 3-3: Présentation des relations entre les attributs et les étiquettes

N°	Variable	P-value	Conclusion
		Test ANOVA	
01	Sexe	Statistic=0.4078581492566402, P-value=0.6652077288586493	Pas de relation linéaire entre ces deux variables car le P est supérieur à 0.05.
02	Catégorie	Statistic=0.040029797625365925, P-value=0.9607626830965975	Pas de relation linéaire entre ces deux variables car le P est supérieur à 0.05.
03	Age	Statistic=1.4199385383699559, P-value=0.1937448356617634	Pas de relation linéaire entre ces deux variables car le P est supérieur à 0.05.
04	Garantie	Statistic=1.5577417395456905, P-value=0.18365886287947947	Pas de relation linéaire entre ces deux variables car le P est supérieur à 0.05.
05	Affectation	Statistic=0.07515425893123448, P-value=0.9959554630884898	Pas de relation linéaire entre ces deux variables car le P est supérieur à 0.05.

06	Activité	Statistic= 3.170216095026461, P-value= 0.0003189763873830969,	Il existe une relation linéaire entre ces deux variables car P est inférieur à 0.05.
		Coefficient de PEARSON	
07	Prêts	Statistic= 0.02966520883169488, P-value= 0.39478956006420207,	Il existe une relation linéaire positive entre ces variables car 0<P-value<1
08	Remboursement	Statistic= -0.10595801100009303, P-value= 0.0023084562348071095,	Il existe une relation linéaire positive entre ces variables 0<P-value<1
09	Reste	Statistic=0.09204188163610852, P-value=0.008161600592590515,	Il existe une relation linéaire positive entre ces variables car 0<P-value<1
10	Échéance	Statistic= 0.08768024494165128, P-value= 0.011753548553856562,	Il existe une relation linéaire positive entre ces variables car 0<P-value<1

III.4. CONCLUSION PARTIELLE

Dans cette partie, nous avons nettoyé notre ensemble des données et ensuite analyser les relations qui existent entre les variables de notre ensemble des données.

L'analyse uni-variée nous a permis de voir les distributions pour chaque attribut et l'analyse bi variée a permis de voir la relation qui existe entre deux attributs et là nous nous sommes focaliser sur la relation entre les attributs et la donnée principale de sortie.

Enfin, nous avons remarqué que certaines variables n'ont pas de relation avec l'Étiquette, qui représente la donnée principale de sortie prouvant prédire si le client est crédible ou pas mais elles peuvent aider à bien entrainer les modèles et améliorer la précision.

CHAPITRE IV. ÉLABORATION DU MODÈLE DE PRÉDICTION ET ANALYSE DE RÉSULTATS

IV.1. INTRODUCTION

Tout au long de ce chapitre nous allons entrainer nos algorithmes de prédiction avec des données d'entrainement et les évaluer avec les données de test puis essayer de faire une prédiction individuelle des clients. Nous allons entrainer quelques algorithmes parmi ceux-là que nous avons présentés dans notre chapitre premier.

Notons que la librairie Scikit-learn de python nous a énormément servi dans le choix des algorithmes à utiliser, passant par sa documentation officielle et par la suite nous décidons d'entrainer les modèles suivant :

- ***Arbre de Décision***
- ***Régression Logistique***
- ***Foret aléatoire***
- ***Réseaux Bayésiens***

IV.2. PREPARATION DES DONNEES

Dans cette partie, nous préparons nos données pour qu'elles soient dans le format compatible avec l'entrée du modèle soit compréhensible par nos modèles de prédiction en faisant l'encodage, puis la normalisation. Nous subdivisons ainsi notre data set en données de test (Test set) et données d'entrainement (Training set).

IV.2.1. ENCODAGE

Les modèles de Machine Learning ne travaillent qu'avec des données numériques, nous allons expliquer comment procéder à une conversion de données catégorielles en données numériques par les techniques de Label Encoding et One Hot Encoding.

a. Label Encoding

Le codage d'étiquettes aussi appelé Label Encoding, peut être réalisé à l'aide de la librairie *Skitlearn* de python. Cette librairie fournit un outil très efficace pour encoder les niveaux des caractéristiques catégorielles en valeurs numériques. Cet outil est appelé *LabelEncoder*, qui encode les étiquettes avec une valeur comprise entre 0 et n_classes-1 où n est le nombre

d'étiquettes distinctes. Si une étiquette se répète, elle affecte la même valeur que celle attribuée précédemment. [19]

Notons qu'en fonction des données, l'encodage des étiquettes pose un nouveau problème. Il y aura différents nombres dans la même colonne et le modèle comprendra mal que les données sont dans un ordre quelconque, par exemple : 0 <1 <2. C'est qui peut causer à ce que le modèle établisse une corrélation qui n'existe peut-être pas. D'où la naissance de l'encodage One Hot Encoding pour résoudre ce problème [19].

b. One Hot Encoding

Le codage à chaud ou One Hot Encoding consiste à coder une colonne contenant des données catégorielles en la divisant en plusieurs colonnes. Les attributs sont remplacés par des 1 et des 0, en fonction de la valeur de la colonne.
L'avantage principal de cet encodage est que pour passer d'un état à un autre, seules deux transitions sont nécessaires : un chiffre passe de 1 à 0, un autre de 0 à 1 et son inconvénient est qu'il faut au minimum **n bits** pour représenter **n états**, ce qui conduit à une augmentation linéaire du nombre de chiffres par rapport au nombre d'états [19].

Nous remarquons que dans le cas de ce travail, en utilisant l'encodage One Hot ou Label Encoding, les résultats restent les mêmes et à la fin nous avons utilisé Label Encoding au choix.

IV.2.2. NORMALISATION

De nombreux algorithmes de Machine Learning fonctionnent mieux ou convergent plus rapidement lorsque les fonctionnalités sont à une échelle relativement similaire et / ou presque distribuées normalement. [20]

Quelques méthodes de normalisation :

- ***MinMaxScaler :*** Pour chaque valeur d'une entité, MinMaxScaler soustrait la valeur minimale de l'entité, puis le divise par la plage. La plage est par défaut comprise entre 0 et 1. [20]
- ***RobustScaler :*** Transforme le vecteur de caractéristiques en soustrayant la médiane, puis en le divisant par la plage interquartile (valeur de 75% - valeur de 25%). Cette normalisation ne redimensionne pas les données dans un intervalle prédéterminé comme MinMaxScaler et la plage de chaque fonctionnalité après l'application de RobustScaler est plus large que pour MinMaxScaler. [20]

- ***StandardScaler:*** Permet de normaliser une caractéristique en soustrayant la moyenne puis en la mettant à l'échelle de la variance. La variance unitaire signifie la division de toutes les valeurs par l'écart type. Le StandardScaler résulte en une distribution avec un écart-type égal à 1. La variance est égale à 1 également, car variance = écart-type au carré. Et $1^2 = 1$. Pour cette normalisation, la moyenne de la distribution est 0 et environ 68% des valeurs se situeront entre -1 et 1 [20].

Dans le cadre de ce travail, nous avons essayé toutes les normalisations précitées et nous avons remarqué que le résultat reste inchangé. Voilà pourquoi nous avons utilisé StandardScaler au choix.

IV.2.3. DONNEES D'ENTRAINEMENT ET DONNEES DES TEST

Nous avons subdivisé notre ensemble des données en deux, une partie pour le training set ou données entrainement comprenant 70% de notre Data set, et une autre partie pour le test set ou données de test comprenant 30% de notre Data set. Les données d'entrainement permettrons à nos modèles de s'entrainer et les données de test seront utiliser pour évaluer les résultats de nos modèles.

IV.3. ELABORATION, EVALUATION DES MODELES ET PRESENTATION DES RESULTATS

Dans cette partie, nous allons présenter les résultats issus des algorithmes après les avoir entrainées avec les données. Nous avons utilisé la matrice de confusion pour mesurer les performances de nos modèles de prédictions.

IV.3.1. EVALUATION DES MODELES

- ***Matrice de confusion*** *(Confusion Matrix)*

Aussi appeler tableau de contingence, la matrice de confusion est un outil qui mesure les performances d'un modèle de Machine Learning en vérifiant notamment à quelle fréquence ses prédictions sont exactes par rapport à la réalité dans un problème de classification. Bref, Elle présente un résumé des résultats de prédictions sur un problème de classification [21].

La matrice se présente de la manière suivante :

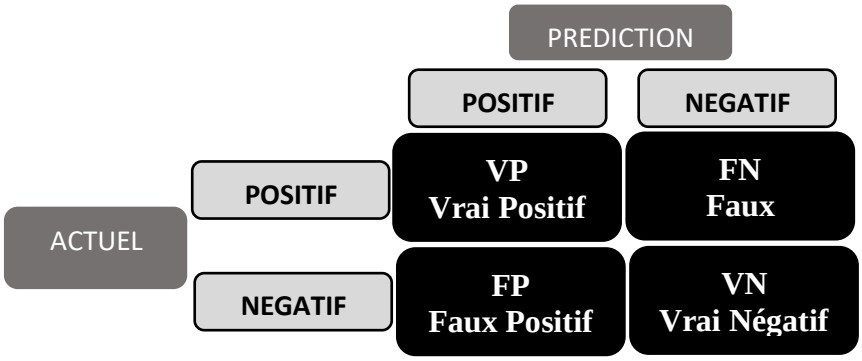

Figure 4-1: Matrice de Confusion

Commentaires : Expliquons les terminologies VP, VN, FP et FN permettant de bien comprendre le fonctionnement d'une matrice de confusion : [21]

a. **VP (Vrai Positif)** : c'est le cas où la prédiction est positive et la valeur réelle est effectivement positive. *Exemple : L'expert dit que vous êtes crédible et vous êtes bel et bien crédible*
b. **VN (Vrai Négatif)** : c'est le cas où la prédiction est négative et la valeur réelle est effectivement négative. *Exemple : L'expert dit que vous n'êtes pas crédible et vous ne l'êtes pas.*
c. **FP (Faux positif)** : c'est le cas où la prédiction est positive, mais la valeur réelle est négative. *Exemple : L'expert dit que vous êtes crédible et vous ne l'êtes pas.*
d. **FN (Faux Négatif)** : c'est le cas où la prédiction est négative, mais la valeur réelle est positive. *Exemple : L'expert dit que vous n'êtes pas crédible, mais vous êtes crédible.*

L'avantage de la matrice de confusion est qu'elle est souvent simple à lire et à comprendre et permet de visualiser très rapidement les données et les statistiques afin d'analyser les performances d'un modèle. Un bon modèle est celui qui a la valeur 0 sur FP et FN.

- ***Métriques d'évaluation***
  - ***La métrique Accuracy (Exactitude)*** : désigne simplement la proportion d'instances qui ont été classées correctement. Il s'agit généralement du premier métrique que nous avons examiné pour évaluer l'exactitude de chacun de nos modèles de prédiction.

    Notons que lorsque les données de test sont déséquilibrées (dans les cas où la plupart des instances appartiennent à l'une des classes) l'exactitude ne permet pas de déterminer véritablement l'efficacité d'un classifier. Voilà donc pour quoi Il est utile de calculer d'autres métriques capturant des aspects plus spécifiques de l'évaluation issus de la matrice de confusion. [22]

- ***La métrique Precision (Précision) :*** Elle détermine le taux de positifs qui ont été classés correctement.

$$\boldsymbol{Precision} = \frac{Vrai\ Positif}{Vrai\ Positif + Faux\ Positif} = \frac{Vrai\ Positif}{Total\ Prediction\ Positif} \quad \textbf{(Form.4.1)}$$

Immédiatement, nous avons remarquez que la Précision indique à quel point le modèle est précis par rapport aux résultats positifs prédits et combien d'entre eux sont réellement positifs. C'est un bon moyen de déterminer si le coût du faux positif est élevé [22].

- ***La métrique Recall (Rappel) :*** Elle calcule le nombre des points positifs capturés par notre modèle en le qualifiant de positif (de vrai positif).

$$\boldsymbol{Recall} = \frac{Vrai\ Positif}{Vrai\ positif + Faux\ Negatif} = \frac{Vrai\ Positif}{Total\ Actuel\ Positif} \quad \textbf{(Form.4.2)}$$

En appliquant la même interprétation, nous savons que Rappel sera la métrique du modèle que nous utiliserons pour sélectionner notre meilleur modèle lorsque les coûts associés à Faux Négatif sont élevés [22].

- ***La métrique du Taux d'erreur :*** Elle est calculée en fonction de l'exactitude.

$$\boldsymbol{Taux\ d'erreur} = 1 - Accurancy \quad \textbf{(Form.4.3)}$$

Notons que l'importance d'une métrique d'évaluation sur l'autre est fondamentalement fonction de type des données et le type d'application.

IV.3.2. PRÉSENTATION DES RÉSULTATS

a. Arbre de décision (Decision Tree)

- ***Matrice de confusion***

Tableau 4-1: Matrice de confusion du classifieur Desision Tree

PREDICTION VRAI	CREDIBLE	NON-CREDIBLE	TOTAL
CREDIBLE	82	5	87
NON-CREDIBLE	15	146	161
Total	**97**	**151**	**248**

Commentaire : Sur un échantillon de 248 observations représentant les 30% prise sur notre ensemble des données comme données de test, pour les arbres de décision, 82 observations

étaient crédibles et ont été prédits crédibles, 5 étaient crédibles et ont été prédits non crédibles, 15 étaient non crédibles et ont été prédits crédibles et 146 étaient non crédibles et ont été prédits non crédibles.

- ***Métriques***

Tableau 4-2: Les Métriques d'évaluation du classifieur Desision Tree

Métrique	*Accuracy*	*Precision*	*Recall*	*Taux d'erreur*
Valeur	91.9355	96.6887	90.6832	8.06452

b. Foret aléatoire (Random Forest)

- ***Matrice de confusion***

Tableau 4-3: Matrice de confusion du classifieur Random Forest

PREDICTION VRAI	**CREDIBLE**	**NON-CREDIBLE**	**TOTAL**
CREDIBLE	72	15	87
NON-CREDIBLE	15	146	161
Total	**87**	**161**	**248**

Commentaire : Sur un échantillon de 248 observations représentant les 30% prise sur notre ensemble des données comme données de test, pour le Random Forest, 72 observations étaient crédibles et ont été prédits crédibles, 15 étaient crédibles et ont été prédits non crédibles, 15 étaient non crédibles et ont été prédits crédibles et 146 étaient non crédibles et ont été prédits non crédibles.

- ***Métriques***

Tableau 4-4: Les métriques du classifieur Random Forest

Métrique	*Accuracy*	*Precision*	*Recall*	*Taux d'erreur*
Valeur	88.3065	92.3077	89.441	11.6935

c. Régression Logistique (Logistic Regression)

- ***Matrice de confusion***

Tableau 4-5: Matrice de confusion du classifieur Logistic Regression

PREDICTION VRAI	**CREDIBLE**	**NON-CREDIBLE**	**TOTAL**
CREDIBLE	34	53	87
NON-CREDIBLE	38	123	161
Total	**72**	**176**	**248**

Commentaire : Sur un échantillon de 248 observations représentant les 30% prise sur notre ensemble des données comme données de test, pour la régression logistique, 34 observations étaient crédibles et ont été prédits crédibles, 53 étaient crédibles et ont été prédits non crédibles, 38 étaient non crédibles et ont été prédits crédibles et 123 étaient non crédibles et ont été prédit non crédibles.

- ***Métriques***

Tableau 4-6: Les métriques du classifieur Logistic regression

Métrique	*Accuracy*	*Precision*	*Recall*	*Taux d'erreur*
Valeur	63.3065	69.8864	76.3975	36.6935

d. Réseaux bayésiens (Naive Bayes Classifier)

- ***Matrice de confusion***

Tableau 4-7: Matrice de confusion du classifieur Naive Bayes

PREDICTION VRAI	**CREDIBLE**	**NON-CREDIBLE**	**TOTAL**
CREDIBLE	46	41	87
NON-CREDIBLE	61	100	161
Total	**107**	**141**	**248**

Commentaire : Sur un échantillon de 248 observations représentant les 30% prise sur notre ensemble des données comme données de test, pour le Naive bayes classifier, 46 observations étaient crédibles et ont été prédits crédibles, 41 étaient crédibles et ont été prédits non crédibles,

61 étaient non crédibles et ont été prédit crédibles et 100 étaient non crédibles et ont été prédits non crédibles.

- ***Métriques***

Tableau 4-8: Les métriques du Naive Bayes Classifier

Métrique	*Accuracy*	*Precision*	*Recall*	*Taux d'erreur*
Valeur	58.871	70.922	62.1118	41.129

Le tableau 4-9 permet de comparer les résultats partir des métriques exprimer en %.

Tableau 4-9: Présentation des résultats avec les métriques d'évaluation

	Naive Bayes	***Logistic Regression***	***Random Forest***	***Desision Tree***
Accuracy(Exactitude)	**58.871**	**63.3065**	**87.9032**	**91.9355**
Precision	**70.922**	**69.8864**	**90.6832**	**96.6887**
Recall(Rappel)	**62.1118**	**76.3975**	**89.441**	**90.6832**
Taux d'erreur	**41.129**	**36.6935**	**12.0968**	**8.06452**

Nous pouvons visualiser ces résultats sur la figure suivante :

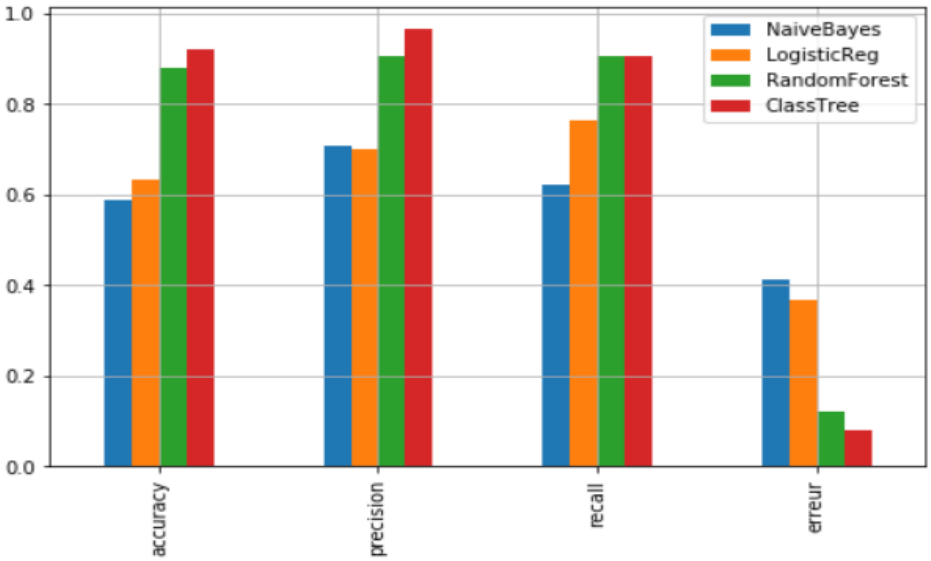

Figure 4-2: Diagramme de barre de la Présentation des résultats avec les métriques d'évaluation

En observant les résultats du tableur Tab.4.9 et le diagramme de barre à la Figure4-2, nous avons constaté que le classifier Arbres de décision est meilleur dans notre cas. Cela car il présente un degré d'exactitude élevé, ce qui implique un faible taux d'erreur. Il présente également une précision élevée par rapport aux autres modèles de prédiction. C'est donc avec lui que nous allons faire la prédiction individuelle des clients.

IV.4. REALISATION DU GUI (Graphical User Interface)

Pout tout projet de Data Mining, la dernière étape est la mise en production du modèle d'apprentissage comme souligné dans la méthodologie CRISP-DM que nous avons présentée précédemment. Dans le cas de notre travail, nous avons réalisé une interface graphique avec TKINTER (Tool Kit Interface) de python.

IV.4.1. PRÉSENTATION DE L'OUTILS UTILISÉS

- ***L'outil Tkinter***

C'est l'outil qui nous a permis réaliser des interfaces graphiques de python pour l'utilisateur, cela à l'aide d'un ensemble de composants graphiques ou widgets. Nous avons donc créé des composants graphiques sur la fenêtre facilitant l'interfaçage homme-machine de notre application [23].

Dans le cas de ce travail, nous utilisons les composants suivants : Une fenêtre (sur laquelle nous allons poser d'autres composants), un champ de saisie Entry *(*permettant à l'utilisateur d'entrer le montant de prêt*)*, six listes déroulantes (permettant à l'utilisateur de choisir le Sexe, l'Age, l'activité, l'affectation, l'échéance et la garantie), deux boutons (l'un permet de soumettre les données saisies au modèle pour la prédiction et l'autre permet d'effacer le formulaire), et quelques labels qui sont des textes permettant de communiquer avec l'utilisateur.

La figure 4-3 donne l'image de notre interface graphique :

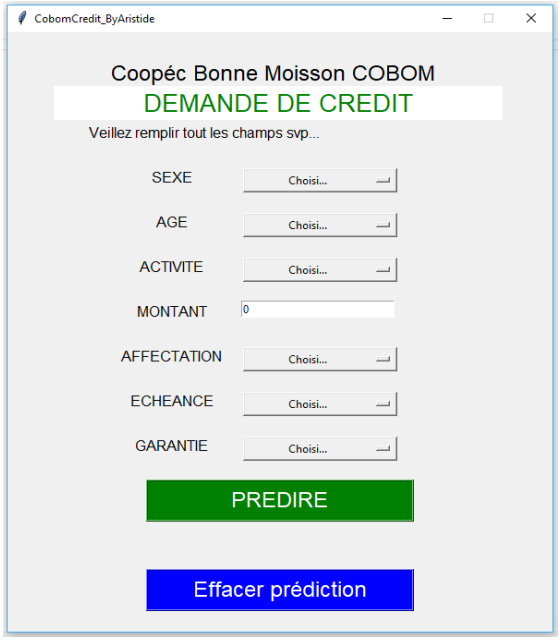

Figure 4-3: Présentation de l'Interface graphique

- ***Langage de programmation et IDE***

Dans le cas de ce travail, nous avons programmée en python avec comme IDE Jupyter Notebook.

Langage de programmation : Python est un langage de programmation interprété, multiplateformes et orientée objet. Il est doté d'un typage dynamique fort, d'une gestion automatique de la mémoire et d'un système de gestion d'exceptions excellent [24].

IDE (Integrated Development Environment) : Jupyter Notebook est issue des 3 langages de programmation Julian, Python et R. Jupyter est une application web utilisée pour programmer dans plusieurs langages de programmation, il est plus une évolution du projet IPython. Il permet de réaliser des calepins ou notebooks, c'est-à-dire des programmes contenant à la fois du texte en markdown et du code en Julian, Python, R... Ces notebooks sont utilisés en science des données pour explorer et analyser des données [24].

IV.4.2. PRÉSENTATION DES INTERFACES DU GUI

Comme dit précédemment, notre API permet au décideur de la Coopéc d'entrer les informations du client, d'avoir sous les yeux la décision proposer par notre modèle de prédiction à l'appui du bouton PREDIRE et effacer le formulaire à la fin de la prédiction en appuyant sur le bouton EFFACER.

Les figures 4-4 et 4-5 montrent les interfaces présentant la soumission des données et la réponse du modèle de prédiction.

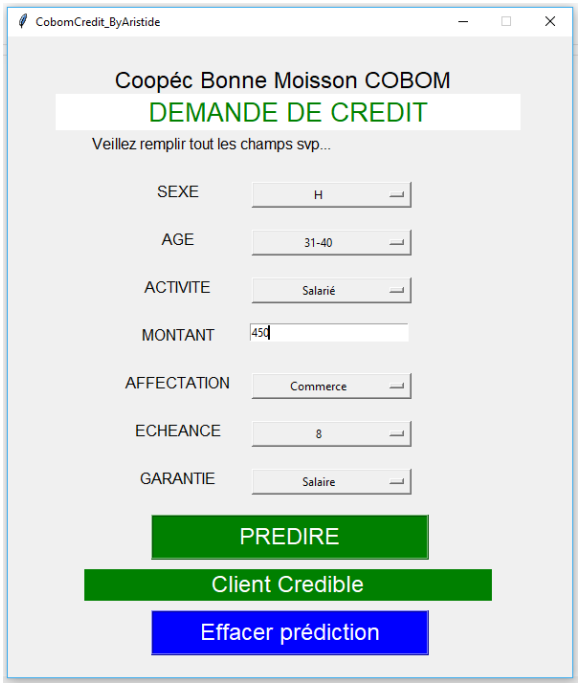

Figure 4-4: Interface: Client crédible

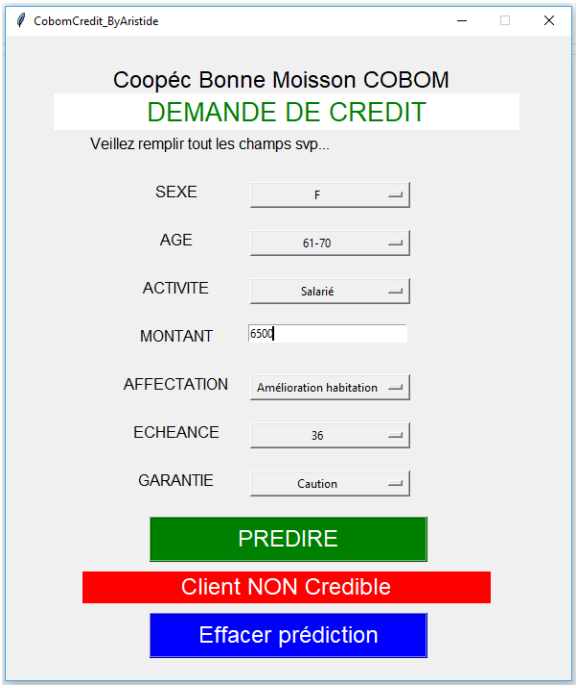

Figure 4-5: Interface: Client non crédible

IV.5. CONCLUSION PARTIELLE

Dans cette partie, nous avons entrainé 4 modèles de prédiction avec nos données. Pour un bon apprentissage, nous avons préparé les données en faisant le Label Encoding, la normalisation avec le StandardScaler et en subdivisant notre data set en deux dont 70% pour le training set et 30% pour le test set. Notons qu'après évaluation avec nos métriques, le modèle des Arbres de décision a été meilleur et c'est lui que nous avons utilisé lors de la prédiction individuelle des clients. Nous avons ensuite mis en place une interface graphique GUI permettant à l'utilisateur de fournir des données à notre modèle et avoir les résultats.

CONCLUSION GENERALE

Au cours de ce travail, nous avons montré comment utiliser les techniques de Data Mining afin de prédire la crédibilité d'un client dans une coopérative d'épargne et de crédit, cela en utilisant les données issues de la Coopéc Bonne Moisson à Goma.

Pour y arriver, nous nous sommes posé les questions suivantes :

Y-a-t-il moyen d'utiliser une technique permettant de minimiser le risque associé à l'octroi de crédits dans une coopérative pour ainsi se pérenniser et rester viable ? Est-il possible d'utiliser ces techniques pour doter la coopérative d'un outil de prédiction de crédibilité d'un client ? Comment mettre en place une solution efficace contre les risques que court une coopérative lors de l'octroi de crédit ?

Diverses techniques de Data Mining et des statistiques nous ont permis de nettoyer (gérer les données manquantes, les données aberrantes,…) et analyser (uni-variée et bi-variée) les données que nous avons reçus de la part de la coopérative d'épargne et de crédit Bonne Moisson. Nous avons ensuite étiqueté les données avec un nouvel attribut qui a été créé en fonction de l'attribut Reste que nous avons mis en % de par le montant prêter et le montant remboursé, de l'échéance et du nombre de jours de retard. Dans la suite, nous avons effectué l'analyse bi-variée de l'étiquette et d'autres attributs, et nous avons constaté qu'elles ne sont en relation qu'avec seulement quelques attributs comme le montant de prêt, l'échéance et l'activité du client. Cependant, d'autres tests de sélections d'attributs utiles ont montré que les autres attributs ont une influence sur l'étiquette excepté l'attribut Catégorie dont l'influence est négligeable.

Nous avons entrainé 4 Modèles de prédiction dont les arbres de décision, la régression logistique, les réseaux bayésiens et les forêts aléatoires. À la fin nous avons choisi une de ces techniques en fonction des métriques d'évaluation et c'est avec le modèle choisi que nous avons prédit la crédibilité d'un nouveau client.

Nous avons en fin remarqué que les techniques de Data Mining ont pu apporter une contribution dans le domaine de l'octroi de crédit et nous avons implémenté une interface graphique à mettre à la disposition des décideurs des Coopéc pour leurs permettre de prendre des décisions plus ou

moins fiables et données des crédits aux clients qui ont plus de probabilité de remboursé à l'échéance donnée.

Toutes les matières ne pouvait pas être épuisées en ce qui concerne ce travail, nombreux points restent ouverts aux chercheurs parmi lesquels l'utilisation des réseaux des neurones si la Coopéc fournit une grande quantité des données.

BIBLIOGRAPHIE

[1]. «Machine Learning », [En ligne]. Avalable : https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning/4011858-quest-ce-que-le-machine-learning. [Accès le 15 Février 2019]

[2]. MULENDA K. "cours d'intelligence artificielle et systèmes experts". Inédit FSTA ULPGL. juin 2016.

[3]. « Techniques de DM pour la GRC dans les banques» [En ligne]. Avalable : https://docplayer.fr/9030881-Methodes-de-dm-pour-la-grc-dans-les-banques.html [Accès le 20 Février 2019]

[4]. SAERENS M. DECAESTECKER C. Les arbres de décision (Decision Trees). Belgique: inédit ULB, 2006, p. 10.

[5]. ANDREW Ng.Stanford «CS229 -Machine Learning -Ng» [En ligne]. Avalable : https://www.coursera.org/learn/machine-learning. [Accès le 25 Mars 2019]

[6]. Alexandre KOWALCZYK. «SVM Tutorial. » [En ligne]. Avalable : http://ww.svm-tutorial.com/. [Accès le 10 Avril 2019]

[7]. Eric Biernat. Michel Lutz. Data science: Fondamentaux et études des cas, Machine Learning avec python et R.

[8]. Eric KIM. «Everything you wanted to know about the Kernel Trick» [En ligne]. Avalable : http: / /www.eric- kim.net/eric- kim- net/posts/1/kernel_trick.html. [Accès le 10 Avril 2019]

[9]. «Processus data Mining » [En ligne]. Avalable : https://barnraisersllc.com/2018/10/data-mining-process-essential-steps/ [Accès le 11 Avril 2019]

[10]. «Les applications de data Mining » [En ligne]. Avalable : https://bigdata-madesimple.com/14-useful-applications-of-data-mining/ [Accès le 11 Avril 2019]

[11]. «Apprentissage supervisé vs non supervisé » [En ligne]. Avalable : https://www.actuia.com/vulgarisation/difference-entre-apprentissage-supervise-apprentissage-non-supervise/ [Accès le 16 Avril 2019]

[12]. «Intelligence artificielle et les crédits » [En ligne]. Avalable : https://www.prologia.fr/intelligence-artificielle-et-credit/ [Accès le 20 Avril 2019]

[13]. Khadidiatou NDIAYE. Le scoring en microfinance : un outil de gestion du risque de crédit. Janvier 2012

[14]. Little, R.J.A., and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: John Wiley & Sons, Inc., pp. 11 -13

[15]. Becker WILLIAM. "Uncertainty propagation through large nonlinear models". Thèse de doct. University of Sheffield, 2011.

[16]. Bernard Clément, Phd. "Modele d'analyse de variance avec STATISTICA"

[17]. «Le test d'indépendance du Khi-carré de PEARSON » [En ligne]. Avalable : http://www.info.univ-angers.fr/~gh/wstat/Perfectionnement_R/mazerolle-khi-carre.pdf [Accès le 10 Mai 2019]

[18]. Sunny Srinidhi, «Encodage Label Encoding vs one hot » [En ligne]. Avalable : http://blog.contactsunny.com/data-science/label-encoder-vs-one-hot-encoder-in-machine-learning [Accès le 20 Mai 2019]

[19]. Raheel Shaikh, «Choosing the right encoding method-Label vs OneHot Encoder» [En ligne]. Avalable: https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b [Accès le 20 Mai 2019]

[20]. Jeff Hale, «Scale, Standardize, or Normalize with Scikit-Learn» [En ligne]. Avalable : https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02 [Accès le 05 Juin 2019]

[21]. « Matrice de Confusion » [En ligne]. Avalable : https://www.lebigdata.fr/confusion-matrix-definition [Accès le 20 Juin 2019]

[22]. MICROSOFT AZUE : « Évaluation des performances d'un modèle dans Azure Machine Learning Studio » [En ligne]. Avalable : https://docs.microsoft.com/fr-fr/azure/machine-learning/studio/evaluate-model-performance. [Accès le 20 Juillet 2019]

[23]. «ISN-Documentation de Tkinter » [En ligne]. Avalable : http://tkinter.fdex.eu [Accès le 15 Aout 2019]

[24]. «Python et Jupyter Notebooks » [En ligne]. Avalable : https://fr.wikipedia.org/wiki/ [Accès le 20 Aout 2019]

[25]. Département des crédits : Coopérative d'Épargne et des Crédit Bonne Moisson Goma