République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Abderahmane Mira de Béjaia
Faculté des Sciences Exactes
Département de Recherche Opérationnelle

Mémoire de fin de cycle

En vue de l'obtention d'un diplôme de Master en Mathématiques
Spécialité : Mathématiques financières

Realisé par
DJAFRI Razik

Thème

DATA MINING

Classification par apprentissage supervisé pour prédire le remboursement d'un crédit

---

Soutenu publiquement, le 04/07/2019 devant le jury composé de :

M.	**S. OUAZINE**	MCB	U.A.M. Béjaia	Président
M.	**K. ABBAS**	Professeur	U.A.M. Béjaia	Encadreur
M.	**M. SOUFIT**	Docteur	U.A.M. Béjaia	Examinateur
M.	**A. BEY**	Doctorant	U.A.M. Béjaia	Examinateur

2018/2019

Dédicaces

C'est avec un grand amour et gratitude que je dédie mon mémoire de fin d'études aux personnes qui ont marqués ma vie, et sans lesquelles je n'aurais connu le sens du bonheur.

A ma chère mère qui a toujours été là pour m'offrir beaucoup de tendresse, d'écoute, d'encouragement et tout ce qu'une bonne maman peut donner à son enfant.

A mon cher père qui a toujours fait de son mieux pour m'assurer toute condition de vie stable et honorable.

A mes chers grands parents, mes chères sœurs, qui m'ont toujours assisté par leurs conseils, par leurs affections et qui m'encourageaient en toute situation.

A mon enseignante de l'école primaire Moussaoui Nora qui a semis en moi les grains de réussite.

A mes meilleurs amis, qui sont la source de mes meilleurs souvenirs.

Une spéciale dédicace à tous le personnel de l'ANSEJ, qui mon faciliter l'accès à leur base de données.

Et à toute personne m'ayant encouragé et soutenu durant mon cursus universitaire.
Je vous aime.

RAZIK

Remerciement

Au terme de ce travail, je tiens à remercier tous ceux qui ont apporté leur contribution à la bonne marche de mon travail.

Je tiens à présenter mes reconnaissances et mes remerciements à mon tuteur encadrant Mr. Abbas, pour le temps consacré à la lecture et aux réunions qui ont rythmées les différentes étapes de mon mémoire. Les discussions que nous avons partagées ont permis d'orienter mon travail d'une manière pertinente. Je le remercie aussi pour sa disponibilité à encadrer ce travail à travers ses critiques et ses propositions d'amélioration.

Je remercié ma chère sœur Tifithene ainsi que mon cher papa pour leur aide.

Je tiens à présenter mes remerciements les plus sincères à tout le personnel de L'ANSEG , pour son accueil et son appui durant toute la période de stage.

Que mes dames, messieurs les membres du jury trouvent ici l'expression de ma reconnaissance pour avoir accepté de juger mon travail.

Que tout le corps professoral et administratif de l'université A. Mira Bejaia trouve ici le témoignage de ma profonde reconnaissance pour leur contribution à ma formation.

Enfin je remercie toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce mémoire, dont les noms ne figurent pas dans ce document.

Résumé

Un projet datamining commence par un besoin métier, dans notre cas une entreprise possède des données relatives à ses clients ces dernier sont classifier comme ayant un bon ou un mauvais dossier de prêt. Cette classification est basée sur l'historique des clients dans cette entreprise et indique si le dossier de prêt a été correctement rembourser ou non.

L'entreprise envisage d'utiliser ces données pour prédire si un futur client présente un bon dossier de prêt, ce qui permettra enfin de réduire le nombre de prêts à risque qui ne seraient pas remboursés.

L'objectif de notre travail est de concevoir un modèle pour prédire le remboursement ou pas d'un crédit octroyé à un client et pour atteindre cet objectif on opte pour la classification par apprentissage supervisé.

Mots clés : Datamining, Classification, Apprentissage supervisé, Prédiction, modèle

Abstract

A datamining project starts with a business need, in our case a company has data relating to his customers which are classified as having a good or a bad loan file. This classification is based on customer history in this company and indicates whether the loan file has been properly repaid or not.

The company plans to use this data to predict if a future client presents a good loan file which will finally reduce the number of loans at risk who would not be reimbursed.

The purpose of our work is to design a model to predict the repayment or not of a credit granted to a customer and to achieve that goal we opt for a classification by supervised learning.

Keywords : Data Mining, Classification, Supervised Learning, Prediction, Model

Table des matières

Table des figures

Introduction générale

Afin de garantir la réussite des projets accompagnés par les organismes financiers tel que l'Agence Nationale de Soutien à l'Emploi des Jeunes (ANSEJ) ils doivent d'abord garantir le remboursement des crédits accordés à ses clients, en faisant recours aux études de datamining.

Le datamining, dans sa forme et compréhension actuelle, à la fois comme champ scientifique et industriel, est apparu au début des années 90. Cette émergence n'est pas le fruit du hasard mais le résultat de la combinaison de nombreux facteurs à la fois technologiques, économiques et même sociopolitiques. On peut voir le datamining comme une nécessité imposée par le besoin des entreprises de valoriser les données qu'elles accumulent dans leurs bases.
En effet, le développement des capacités de stockage et les vitesses de transmission des réseaux ont conduit les utilisateurs à accumuler de plus en plus de données. Certains experts estiment que le volume des données double tous les ans. C'est avec ces données que les dataminers répondent aux problématiques posés par les différents organismes.

Ce mémoire vise dans un premier temps à donner un large aperçu sur le datamining ainsi que les connaissances théoriques, pratiques (outils et techniques) et processus pour réaliser une étude dans ce domaine.
Ce travail m'a également apporté une vraie première expérience dans le datamining au sein duquel j'ai évolué en répondant à la question : **Comment développer un modèle qui va nous permettre de savoir si un nouveau client peut OUI ou NON rembourser son crédit ?**

Notre travail est organisé comme suit :

- Le premier chapitre est dédié à la présentation de l'entreprise,
- Le deuxième chapitre est consacré au Datamining,
- Dans le troisième chapitre, nous allons répondre à la problématique posé,
- Nous terminons avec une conclusion générale où nous exposons quelques perspectives de ce modeste travail.

Chapitre 1

Présentation de l'entreprise

1.1 Présentation de L'Agence Nationale de Soutien a l'emploi des jeunes

L'Agence Nationale de Soutien à l'Emploi des Jeunes, par abréviation ANSEJ, créée en 1996, est un organisme à caractère spécifique, doté de la personnalité morale et de l'autonomie financière, placé sous la tutelle du Ministre chargé de l'emploi.

L'ANSEJ est créée pour accompagner les porteurs de projets pour la création et l'extension de micro-entreprises de production de biens et de services, elle est fondée sur une approche économique, de création de richesse et d'emploi.

L'agence dispose d'un réseau de 51 antennes implantées dans toutes les wilayas du pays ainsi que d'annexes situées dans certaines localités.

1.1.1 Missions de l'ANSEJ

- Soutenir, conseiller et accompagner les jeunes promoteurs la création d'activités.
- Mettre à la disposition des jeunes promoteurs toute information économique, technique, législative et réglementaire relative à leurs activités.
- Développer des relations avec les différents partenaires du dispositif (banques, impôts, cnas et casnos. . .).
- Développer un partenariat intersectoriel pour l'identification des opportunités d'investissement.
- Assurer une formation en relation avec l'entreprise au profit des jeunes promoteurs.
- Encourager toute autre forme d'actions et de mesures pour la promotion de la création et l'extension d'activités.

1.1.2 Objectifs principaux

- Favoriser la création d'activités de biens et de services par les jeunes promoteurs.
- Encourager toutes formes d'actions et de mesures tendant promouvoir l'entrepreneuriat.

1.1.3 Conditions d'éligibilité

1. Être âgé (e) de 19 à 35 ans. Dans des cas exceptionnels, lorsque l'investissement génère au moins trois (3) emplois permanents (y compris les jeunes promoteurs associés dans l'entreprise) l'âge limite du gérant de l'entreprise créée pourra être porté quarante (40) ans.
2. Être titulaire d'un diplôme, d'une qualification professionnelle et/ou posséder un savoir-faire reconnu ;
3. Mobiliser un apport personnel sous forme de fonds propres qui varie selon le type de financement et le niveau de l'investissement.
4. Ne pas occuper un emploi rémunéré au moment de l'introduction du formulaire d'inscription pour bénéficier de l'aide.
5. Être inscrit auprès des services de l'agence nationale de l'emploi comme chômeur demandeur d'emploi.
6. Ne pas être inscrit au niveau d'un centre de formation, institut ou université au moment de l'introduction de la demande d'aide, sauf s'il s'agit d'un perfectionnement dans son activité.
7. Ne pas avoir bénéficié d'une mesure d'aide au titre de la création d'activité.

1.1.4 Montant maximum de l'investissement

Le montant maximum de l'investissement est de dix millions (10.000.000) de dinars, pour chacune des phases : Création ou extension.
Les prêts non rémunérés supplémentaires sont octroyés en sus du montant de l'investissement.

1.1.5 Inscription

L'inscription des promoteurs au niveau de l'Agence Nationale de Soutien à l'Emploi des Jeunes ANSEJ, se fait exclusivement sur la base d'un imprimé appelé
« formulaire d'inscription » à télécharger du site « www.ansej.org.dz » ou avec l'inscription online sur le site « promoteur.ansej.org.dz ».

1.2 Parcours de création de la micro-entreprise

1.2.1 Sensibilisation et information

La participation du jeune aux diverses manifestations organisées périodiquement par l'agence, soit par l'accès au site internet ou bien par son rapprochement au niveau des antennes et annexes de l'agence qui couvrent tout le territoire national, lui permet d'être informé sur les opportunités d'investissement et les avantages accordés par le dispositif.

1.2.2 Formulation de l'idée du projet

L'idée du projet doit être le résultat d'une étude et d'une recherche efficace sur les opportunités d'investissement et en cohérence avec les qualifications du jeune futur promoteur (diplômante /qualifiante) et ses capacités pour sa réalisation.

1.2.3 Inscription via le portail

Une fois le choix du projet est fait et les équipements acquis, le jeune peut accéder au site internet de l'agence pour initier le processus d'inscription électronique, en insérant toutes les données relatives à sa personne, ses associés le cas échéant, et sa micro-entreprise.

1.2.4 Etude du projet et plan d'affaires

Après l'achèvement de la phase d'enregistrement, l'étude du projet et l'élaboration du plan d'affaire débuteront en profondeur, avec l'appui du cadre chargé d'accompagnent, en recueillant toutes les informations nécessaires concernant :

- Les équipements à acquérir.
- L'implantation du projet, notamment 'environnement de la future micro- entreprise.
- L'étude de marché.
- Les choix techniques.
- La ressource humaine.
- L'étude financière.

1.2.5 Présentation du projet au Comité de Sélection, de Validation et de Financement des projets (CSVF)

A cette étape le jeune promoteur doit présenter son projet au niveau du CSVF pour étude et prise de décision par une validation, ajournement ou un rejet.

►**Cas de validation** : dépôt du dossier administratif et financier.
► **Cas d'ajournement** : levée des réserves émises par le comité et représenter le projet au comité.
► **Cas de rejet** : possibilité de présenter un recours dans un délai de quinze (15) jours après notification de la décision de rejet du comité.

1.2.6 Accord bancaire et création juridique de la micro-entreprise

1. Le dossier est déposé au niveau de la banque (financement triangulaire) par le représentant de l'ANSEJ pour l'obtention de l'accord bancaire.
2. Dès notification de l'accord bancaire, le promoteur est tenu de procéder à la création juridique de sa micro-entreprise.

1.2.7 Formation du promoteur

Avant le financement du projet, le promoteur doit obligatoirement suivre une formation sur les techniques de gestion de sa micro- entreprise cette formation est assurée, en interne, par les formateurs de l'ANSEJ.

1.2.8 Financement du projet

Après la création juridique de la micro-entreprise, et la finalisation des procédures, l'ANSEJ procède au financement du projet.

1.2.9 La réalisation du projet et l'entrée en exploitation

Après le financement du projet et la finalisation des procédures concernant cette étape, il y a lieu d'acquérir et d'installer les équipements pour le démarrage de l'activité. Enfin Ce que vous attendez depuis longtemps est arrivé : vous êtes chefs d'entreprise

1.3 Les modes de financement

Le dispositif ANSEJ prévoit trois modes de financement

► Le financement triangulaire.

► Le financement mixte.

► L'autofinancement.

1.3.1 Création de micro-entreprises dans le cadre du financement triangulaire

► Le montage financier

Le financement triangulaire est constitué comme suit

- Apport personnel du jeune promoteur ;
- Prêt non rémunéré de l'ANSEJ (PNR) ;
- Crédit bancaire bonifié 100% pour tous les secteurs d'activités, et garanti par le Fonds de Caution Mutuelle de Garantie Risques/Crédits Jeunes Promoteurs.

► La structure du financement triangulaire

Niveau 1			
Montant de l'investissement	**Prix non rémunéré (ANSEJ)**	**Apport Personnel**	**Crédit bancaire**
jusqu'à 5.000.000DA	29%	01%	70%

Niveau 2			
Montant de l'investissement	**Prix non rémunéré (ANSEJ)**	**Apport Personnel**	**Crédit bancaire**
jusqu'à 10.000.000DA	28%	02%	70%

1.3.2 Création de micro-entreprise dans le cadre financement mixte

► **Le montage financier**

Le financement mixte est constitué comme suit

- Apport personnel du jeune promoteur ;
- Prêt non rémunéré de l'ANSEJ (PNR) ;

► **La structure du financement triangulaire**

Niveau 1		
Montant de l'investissement	**Prix non rémunéré (ANSEJ)**	**Apport Personnel**
jusqu'à 5.000.000DA	29%	71%

Niveau 2		
Montant de l'investissement	**Prix non rémunéré (ANSEJ)**	**Apport Personnel**
jusqu'à 10.000.000DA	28%	72%

1.3.3 Création de micro-entreprise en type de l'autofinancement

Niveau 1	
Montant de l'investissement	**Apport Personnel**
jusqu'à 10.000.000DA	100%

1.4 Aides financières et avantages fiscaux accordes par le dispositif ANSEJ

Le jeune promoteur bénéficie d'avantages fiscaux et aides financières au moment de la réalisation, et d'exonération lors de l'exploitation de son projet. Ces avantages sont accordés tant en phase de création que lors de l'extension des capacités de production.
Les avantages fiscaux accordés à la micro-entreprise, en phase d'extension, concernent uniquement les nouveaux apports. Le prorata est déterminé par rapport.

1.4.1 Aides financières

- Un prêt non rémunéré
- Un prêt non rémunéré supplémentaire si nécessaire. (cas financement triangulaire)
- Une bonification du taux d'intérêt bancaire à 100%. (cas financement triangulaire)

1.4.2 Avantages fiscaux

La micro-entreprise bénéficie des avantages fiscaux suivants

1. **Phase réalisation**

   - Exemption du droit de mutation à titre onéreux pour les acquisitions immobilières effectuées dans le cadre de la création d'une activité industrielle.
   - Exonération des droits en matière d'enregistrement pour les actes constitutifs de sociétés.
   - Application du taux réduit de 5% en matière de droits de douane pour les équipements entrant directement dans la réalisation de l'investissement.

2. **Phase exploitation**

   - Exonération de la taxe foncière sur les constructions et additions de constructions pour une période de 3 ans, 06 ans ou 10 ans selon le lieu de l'implantation du projet, à compter de la date de sa réalisation
   - Exonération totale pour une période de 3 ans, 06 ans ou 10 ans selon l'implantation du projet, à de la date de mise en exploitation de l'impôt Forfaitaire Unique (lFU) ou de l'impôt d'après le régime du bénéfice réel.
   - A l'expiration de la période d'exonération citée dans le tiret $n^o 2$, cette dernière peut être prorogée de deux (2) années, lorsque le promoteur d'investissement s'engage à recruter au moins trois (3) employés à durée indéterminée.
     Le non-respect des engagements liés au nombre d'emplois crées entraîne le retrait des avantages et le rappel des droits et taxes qui auraient dus être acquittés.

Toutefois, les investisseurs - les personnes physiques au titre de I 'impôt forfaitaire unique- demeurent assujettis au payement d'un minimum d'imposition correspondant à du montant (10000 DA), prévu dans le code des impôts soit, pour chaque exercice, et quel que soit le chiffre d'affaires réalisé.
Un abattement d'impôt Sur le revenu global (IRG) ou l'impôt sur les bénéfices des sociétés (IBS), selon le cas, ainsi que sur la taxe sur l'activité professionnelle (TAP) à l'issue de la période d'exonération, pendant les trois premières années d'imposition comme suit :

- 70 % durant la première année d'imposition
- 50 % durant la deuxième année d'imposition
- 25 % durant la troisième année d'imposition

1.5 les prêts non rémunères supplémentaires

En plus du prêt non rémunéré (PNR) classique, les jeunes porteurs de projets peuvent bénéficier d'une aide sous forme d'un prêt non rémunéré supplémentaire, selon les trois formules suivantes

1.5.1 Prêt location

Le Prêt Non Rémunéré « Location » (PNR-LO) est une aide supplémentaire accordée aux promoteurs, d'un montant à hauteur de cinq cent mille dinars $(500000DA)$ remboursable, pour la prise en charge du loyer du local ou du poste à quai au niveau des ports, destiné à abriter l'activité projetée. Il est accordé exclusivement aux promoteurs sollicitant un financement triangulaire et en phase de création d'activité, à l'exception des activités non sédentaires et des activités dédiées aux cabinets groupés.
Ne peuvent bénéficier de ce prêt :

- Les jeunes porteurs de projets d'activités non sédentaires.
- Les jeunes porteurs de projets d'activités à créer dans le cadre des cabinets groupés.
- Lorsque le propriétaire du local est un ascendant ou conjoint du promoteur

1.5.2 Cabinets Groupés

Le Prêt Non Rémunéré « Cabinet Groupé » (PNR-CG) est une aide supplémentaire, accordée aux diplômés de l'enseignement supérieur, d'un montant qui ne saurait dépasser un (1) million de dinars remboursable, pour la prise en charge du loyer des locaux destinés à la création de cabinets groupés, et il est accordé exclusivement aux promoteurs sollicitant un financement triangulaire en phase de création d'activité.
Lorsque le propriétaire du local est un ascendant ou conjoint du promoteur, il ne peut bénéficier de ce prêt.

On entend par cabinet groupé l'association de deux (02) projets minimums, occupant le même local, présentés par des jeunes promoteurs, exerçant dans le même domaine d'activité, relevant des domaines médical, auxiliaires de justice, expertise comptable, commissariat aux comptes, comptables agréés, bureaux d'études et de suivi relevant des secteurs du bâtiment, des travaux publics et de l'hydraulique.

1.5.3 Véhicules-Ateliers

Le Prêt Non Rémunéré « Véhicule Atelier » (PNR-VA) est une aide d'un montant de cinq cent mille dinars $(500000DA)$ remboursable, destiné à l'acquisition d'un véhicule atelier, Il est accordé exclusivement aux jeunes promoteurs diplômés du système de la formation professionnelle, sollicitant un financement triangulaire en phase de création, pour l'exercice des activités non sédentaires de plomberie, bâtiment, chauffage, climatisation, Vitrerie, peinture-bâtiment et mécanique automobile.

Chapitre 2

Data Mining

2.1 Introduction

Le datamining est un domaine qui est apparu avec l'explosion des quantités d'informations stockées avec le progrès important des vitesses de traitement et des supports de stockage. En effet, chaque jour nos banques, nos hôpitaux, nos institutions scientifiques, nos magasins, produisent et enregistrent des milliards et des milliards de données.

En français datamining se traduit fouille de données, Mais on utilise aussi l'expression extraction de connaissances à partir de données. (knowledge discovery in databases,en anglais) Ainsi le datamining est un ensemble de méthodes qui permettent d'extraire de l'information de grandes bases de données, il vise à découvrir, dans les grandes quantités de données, les informations précieuses qui peuvent aider à comprendre les données ou à prédire le comportement des données futures.
Les informations découvertes sont des règles, des associations, des tendance inconnues, des structures particulières qui doivent aider à prendre des décisions.

Le terme datamining est apparu au départ dans le secteur tertiaire : banque, assurance, vente par correspondance. Dans ce secteur, l'objectif d'une étude de datamining est en général, de mieux connaître sa clientèle pour la fidéliser ou pour attirer de nouveaux clients. Cependant le datamining est aujourd'hui utilisé dans un grand nombre d'applications de tout type et dans des domaines variés (vente, industrie, médecine, environnement, etc.).

Le datamining utilise depuis son apparition plusieurs outils de statistiques et d'intelligence artificielle pour atteindre ses objectifs.

2.2 Domaines d'application du Data Mining

Les domaines d'application du datamining sont très nombreux c'est un outil puissant d'analyse dès que les données sont nombreuses.[19]

2.2.1 Le datamining dans la banque

C'est en 1941 que le scoring de risque est né dans le secteur bancaire, a une époque où les moyens de calcul étaient très rudimentaires. Depuis, de multiples techniques de datamining (scoring, classification, associations de produits. . .) ont envahi la banque de détail comme la banque d'entreprise, mais c'est surtout à la banque de particuliers que le datamining s'applique, du fait des montants unitaires modérés, du grand nombre de dossiers et de leur caractère relativement standard. Les problèmes de scoring ne sont généralement pas très complexes sur un plan théorique et les techniques classiques d'analyse discriminante et de régression logistique y font merveille .cet essor du datamining dans l'activité bancaire s'explique par la conjonction de plusieurs éléments : le développement des nouvelles technologies de communication (internet ,téléphonie mobile . . .) et de traitement de l'information (entrepôts de données), les attentes accrues de qualité de service des clients, la concurrence exercée sur les banques à réseau par les sociétés de crédit et les « nouveaux entrants » (banques étrangères ,grande distribution et compagnies d'assurance ,lesquelles développent parfois une activité bancaire au travers d'un partenariat avec une banque traditionnelle), la pression économique internationale pour une plus grande rentabilité et productivité des banques ,sans oublier l'aspect réglementaire avec le grand dossier bancaire du moment : la réforme du ratio de solvabilité qui a donné un grand essor au développement des modèles de risque.

2.2.2 Le datamining dans l'assurance IARD

I.A.R.D est une abréviation utilisée dans le monde de l'assurance pour incendie, accidents et risques divers, en opposition avec l'assurance de personnes.

C'est avec le développement de la concurrence des nouveaux entrants que le besoin de datamining dans l'assurance s'est exacerbé, que sont les banques qui, pratiquant ce que l'on nomme la bancassurance, possèdent l'avantage de réseaux étendus, de contacts fréquents avec le client et de bases de données riches. Cet avantage est surtout tangible face aux assureurs « traditionnels » non mutualistes, qui éprouvent parfois des difficultés à fédérer dans des bases de données marketing des informations disséminées et jalousement détenues par leurs agents généraux. De surcroît, les bases clients de ces assureurs, quand elles ne sont pas compartimentées par agent général, sont encore souvent structurées par contrat et non par client. Pourtant. Ces réseaux, avec leurs taux de fidélisation inférieurs à ceux des mutuelles, ont besoin d'améliorer leur gestion de la relation client, et donc leur connaissance globale de leur clientèle. Si les études d'appétence de l'assurance ressemblent à celles de la banque, les études de sinistralité présentent quelques particularités avec par exemple l'intervention de la loi de Poisson dans le modèle linéaire généralisé pour modéliser le nombre de sinistres

2.2.3 Le datamining dans la téléphonie

L'ouverture à la concurrence du marché de la téléphonie, et l'arrivée à maturité puis à saturation du marché de la téléphonie mobile, ont avivé les problèmes de churn appelé également taux d'attrition, (départ pour la concurrence) des clients, particuliers, professionnels ou entreprises. On imagine l'importance de la fidélisation dans ce secteur, quand on prend en considération les coûts d'acquisition

d'un client en téléphonie mobile et qu'un nombre important d'utilisateurs changent chaque année d'opérateur dans certains pays. C'est donc tout naturellement le score de churn qui tient la vedette du datamining dans la téléphonie. Pour les mêmes raisons, des opérateurs utilisent des outils de text mining afin d'analyser automatiquement le contenu des lettres de réclamation des clients.

2.2.4 Le datamining dans L'industrie automobile

Le datamining est utilisé assez couramment dans l'industrie automobile, un thème classique est le score de réachat d'un véhicule d'une marque. Certaine marque ont ainsi construit un modèle prédisant les clients susceptibles d'acheter un de leur nouveau véhicule dans les six mois à venir. Ces clients sont identifiés à partir des données des concessionnaires, lesquels reçoivent en retour une liste de clients au score élevé, qu'ils peuvent alors contacter. Dans le domaine de la production, le datamining est utilisé pour rechercher l'origine des défauts de construction, de façon à pouvoir les minimiser. Des études de satisfaction sont aussi réalisées à partir d'enquêtes auprès des clients, afin d'améliorer la conception des véhicules (qualité, confort...). Dans les laboratoires des constructeurs automobiles, les accidents sont étudiés afin de les classer en profils-type et d'identifier leurs causes, un grand nombre de données sont analysées, portant sur le véhicule, le conducteur et les conditions externes (état de la route, circulation, heure, météo...), Une dernière application du datamining dans le domaine automobile : I'UTAC (L'Union Technique de l'Automobile du motocycle et du Cycle) a présenté une étude dont le but était de déterminer, parmi les véhicules recalés au contrôle technique et soumis à une Obligation de contre-visite, lesquels ne se présenteraient pas à la contre-visite.

2.2.5 Le datamining dans le commerce

Des cartes de crédit privatives sont développées par la grande distribution pour qu'elle puisse constituer de grandes bases de données (parfois de plusieurs millions de porteurs) enrichies par les informations comportementales provenant des tickets de caisses, et lui permettent de concurrencer les banques dans la connaissance du client. En outre, les services associés à ces cartes (caisses réservées, promotions exclusives...) sont facteurs de fidélisation. La détection des associations de produits sur les tickets de caisse permet d'identifier les profils de clients, de mieux choisir les produits et de mieux les disposer dans les rayons, en tenant compte du facteur « régional » dans les analyses. Les résultats les plus intéressants sont obtenus quand les paiements sont effectués avec une carte de fidélité, non seulement car il est alors possible de croiser les associations détectées sur le ticket de caisse avec les informations sociodémographiques (âge, situation familiale, catégorie socioprofessionnelle) fournies par le client à la souscription de la carte, mais aussi parce que l'utilisation de cette carte permet de suivre les paiements d'un client au cours du temps et de mettre en place des plans d'animation des clients, en revenant vers eux selon un tempo et des sujets éventuellement recommandés par le modèle.

La vente par correspondance (VPC) pratique depuis longtemps l'analyse des données sur ses clients afin d'optimiser ses ciblages et d'en réduire les coûts, qui peuvent être énormes lorsqu'un catalogue de mille pages en couleurs est adressé à plusieurs dizaines de millions de clients. Si la banque est à l'origine du score de risque, la VPC fait partie des premiers secteurs à avoir eu recours au score d'appétence.

2.2.6 Le datamining dans la médecine

Le datamining est très répandu dans le secteur médical autant grand utilisateur de statistique, tant dans les applications descriptives que prédictives.

Parmi les applications descriptives, on rencontre la détermination de groupes de patients susceptibles d'être soumis à des protocoles thérapeutiques déterminés, chaque groupe rassemblant tous les patients réagissant de la même façon. On a aussi les études sur les associations de médicaments, en vue notamment de détecter des anomalies de prescription.

Parmi les applications prédictives, on trouve la recherche des facteurs de décès ou de survie dans certaines pathologies (infarctus, cancers...), à partir des données recueillies lors des essais cliniques, afin de choisir le traitement le plus approprié en fonction de la pathologie et de l'individu. On pratique bien sûr la technique prédictive connue sous le nom d'analyse de survie, dans laquelle la variable à prédire est une durée. Les données de survie sont dites «censurées» car la durée est parfaitement connue pour les individus «partis» , tandis que pour ceux qui restent. On ne connaît que leur durée minimale de survie. On peut chercher à prédire le temps de rétablissement après une opération, en fonction des données concernant le patient (âge, poids, taille, fumeur ou non, métier, antécédents médicaux, etc.) et le praticien (nombre d'opérations pratiquées, nombre d'années d'expérience, etc.).

2.2.7 Le datamining dans l'industrie agro-alimentaire

L'industrie agro-alimentaire est également grande consommatrice de statistique. Elle est utilisée dans les « analyses sensorielles », qui croisent les données sensorielles (goût, saveur, texture...) perçues par les consommateurs avec les mesures instrumentales physico-chimiques, ainsi qu'avec les préférences en faveur de tel ou tel produit. Ce sont par ailleurs des modèles prédictifs d'analyse discriminante et de régression logistique qui ont permis de distinguer des spiritueux de leurs contrefaçons, à partir de l'analyse d'une dizaine de molécules présentes dans le breuvage. La chimiométrie est l'extraction d'information à partir de mesures physiques et de données recueillies en chimie analytique. Comme en génomique, le nombre de variables explicatives devient Vite très grand et peut justifier l'utilisation de la régression PLS (« Partial Least Squares regression » et/ou « Projection to Latent Structure »). L'analyse des risques sanitaires est spécifique à l'industrie agro-alimentaire : il s'agit de comprendre et maîtriser l'évolution des micro-organismes, de prévenir les risques liés à leur développement dans les industries agro-alimentaires, et de gérer la date limite de conservation. Enfin, comme dans toute industrie, il faut apprendre à mieux maîtriser les processus pour améliorer la qualité des produits.

2.2.8 Le datamining dans la biologie

De façon générale, la biologie utilise beaucoup la statistique. On la rencontre depuis longtemps dans la classification des espèces vivantes et nous reparlerons de l'exemple classique du classement de trois espèces d'iris par Fisher grâce à son analyse discriminante linéaire. L'agronomie demande à la statistique d'évaluer rigoureusement l'effet d'engrais ou de pesticides.

2.2.9 Autre utilisation du datamining

- **Organisme de crédit**, pour décider d'accorder ou non un crédit en fonction du profil du demandeur de crédit, de sa demande, et des expériences passées de prêts ;
- **L'image mining** est utilisé dans l'imagerie médicale pour déceler automatiquement une échographie anormale, reconnaître une tumeur.
- **La cosmétique** a recours au datamining par exemple pour prédire les effets de nouveaux produits sur la peau humaine en limitant le nombre de tests sur les animaux Dans la prévention des crimes plusieurs expériences ont été menées. Une utilisation aux USA a par exemple été d'identifier les associations de lieu et de plages horaires auxquelles les crimes se produisaient le plus, afin de renforcer la présence policière en conséquence.
- **Détection de fraude** dans les systèmes complexes gérant un nombre d'utilisateurs importants (les administrations par exemple), le datamining, utilise la classification sur les données. Ce mécanisme peut notamment permettre de détecter les données qui vont sortir de l'ordinaire, qui n'auront pas la même empreinte que les comportements "normaux". Certains comportements "normaux" peuvent également sortir de l'ordinaire et constitueront des faux positifs dans le cas de la détection de la fraude, mais c'est une méthode qui permettra de faire ressortir les cas à surveiller.
- **Détection des facteurs** expliquant la pollution de l'air.
- **reconnaissance vocale**
- **Google**, très tôt, a été utilisateur des techniques de datamining, ce que l'on comprend aisément étant donné les volumes de données traités (rappel : 2000000 recherches/minute). Quelques outils utilisant le datamining : Google spell checker (le dictionnaire est en fait constitué en fonction des recherches des utilisateurs) , Autocomplétion, Recherche local.
  Google nous apprend donc que l'enregistrement et l'analyse des logs des recherches des utilisateurs est ce qui permet à Google d'améliorer ses résultats. Tout comme la disponibilité des données a été source d'avancées par le passé, Google anticipe que ce qui sera fait avec ces logs de recherche le sera à l'avenir.
- **Poker !** Trois personnes ont utilisé en 2009 le datamining à l'encontre un joueur en ligne. Ils avaient non seulement utilisé les données des parties qu'ils avaient jouées contre ce joueur, mais étaient également allés jusqu'à acheter l'historique d'un autre joueur. Grâce aux données de plusieurs dizaines de milliers de mains, ils ont pu établir un profil extrêmement précis de leur adversaire et élaborer un plan, qui s'est avéré juteux, puisqu'en à peine 5 heures de jeu, plus de 4 millions de dollars ont été emportés.

Ce dernier exemple montre bien l'étendue des domaines d'application du datamining, dès que les données sont nombreuses, c'est un outil puissant d'analyse.

2.3 Méthodes Data Mining

Dans les méthodes utilisées par le datamining, on distingue deux grandes familles de techniques. Chacune de ces familles de méthodes comporte plusieurs techniques appropriées aux différents types de tableaux de données. Certaines sont mieux adaptées à des données quantitatives alors que d'autres

sont plus généralement dédiées aux traitements de données qualitatives. Nous allons donner à présent un aperçu général sur les principales méthodes.

2.3.1 Apprentissage non supervisé

L'apprentissage non supervisé autrement dit le clustering sont des méthodes de classification automatique qui permettent d'organiser, de simplifier et d'aider à comprendre l'information à partir des sources de données, elles visent à mettre en évidence des informations présentes mais cachées par le volume des données (c'est le cas des segmentations de clientèle et des recherches d'associations de produits sur les tickets de caisse). L'apprentissage non supervisé réduit, résume, synthétise les données et il faut retenir qu'il n'y a pas de variable « cible » à prédire.[21]

L'objectif de la classification automatique c'est d'identifier des groupes d'observations ayant des caractéristiques similaires c'est à dire on veut que les individus dans un même groupe se ressemblent le plus possible et les individus dans des groupes différents se démarquent le plus possible, pour identifier des structures sous-jacentes dans les données, résumer des comportements, affecter de nouveaux individus à des catégories et Identifier les cas totalement atypiques.[18]

parmi les méthodes les plus utilisées de l'apprentissage non supervisé on cite les suivantes.

Classification ascendante hiérarchique – CAH

La Classification ascendante hiérarchique (cah) c'est une technique très populaire utilisables dès que l'on dispose d'une distance (dans un espace des individus ou des variables).
L'objectif de cet algorithme est de rassembler des objets dans des classes de plus en plus larges, en utilisant certaines mesures de similarité ou de distance. Les résultats de ce type de classification sont habituellement représentés sous la forme d'un dendrogramme (arbre de la classification hiérarchique).[8]

► Algorithme CAH

L'idée de l'algorithme de Classification Ascendante Hiérarchique (CAH) est de créer, à chaque étape, une partition de $\Gamma = \{ w_1,...,w_n \}$ en regroupant les deux éléments les plus proches. Le terme "élément" désigne aussi bien un individu qu'un groupe d'individus.
L'algorithme de CAH est décrit ci-dessous :

1. On choisit un écart. On construit le tableau des écarts pour la partition initiale des n individus de Γ :

   $$P_0 = (\{w_1\}, \ldots, \{w_n\}).$$

   Chaque individu constitue un élément.

2. On parcours le tableau des écarts pour identifier le couple d'individus ayant l'écart le plus petit. Le regroupement de ces deux individus forme un groupe A. On a donc une partition de Γ de $n-1$ éléments : A et les $n-2$ individus restants.

3. On calcule le tableau des écarts entre les $n-1$ éléments obtenus à l'étape précédente et on regroupe les deux éléments ayant l'écart le plus petit (cela peut être deux des $n-2$ individus, ou un individu des $n-2$ individus restants avec A). On a donc une partition de Γ de $n-2$ éléments.
4. On itère la procédure précédente jusqu'à ce qu'il ne reste que deux éléments.

► **Exemple**
On considère la matrice de données X dans $\mathbf{R}^2$ définie par

$$X = \begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{pmatrix} 2 & 2 \\ 7.5 & 4 \\ 3 & 3 \\ 0.5 & 5 \\ 6 & 4 \end{pmatrix} \end{array}$$

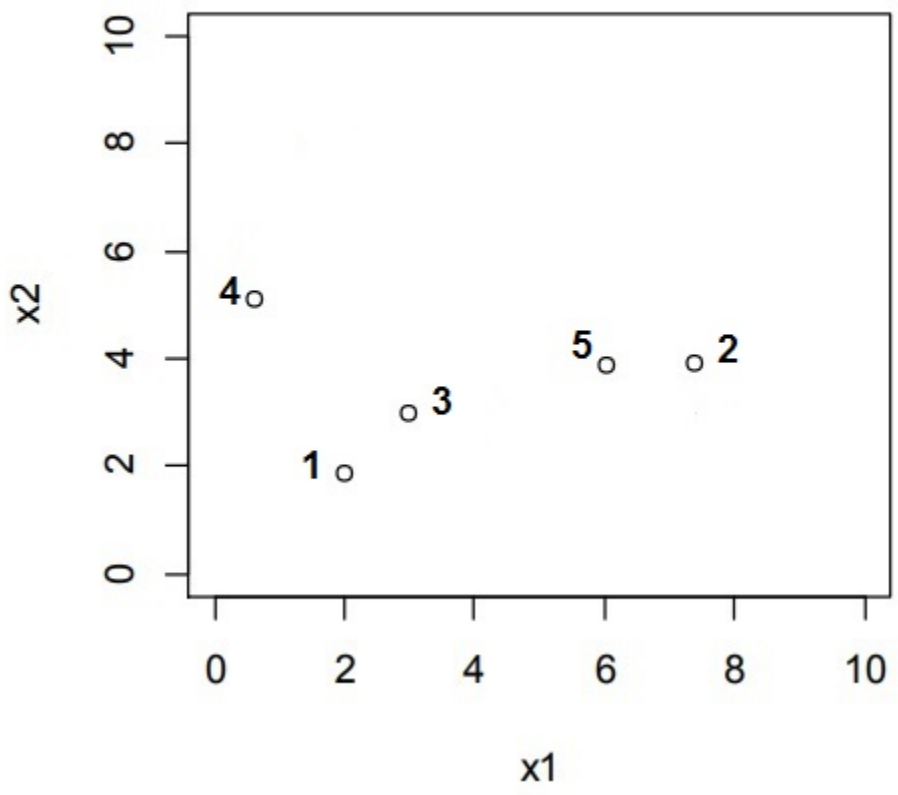

FIGURE 2.1: Présentation graphique X

On va regrouper les individus avec l'algorithme CAH et la méthode du voisin le plus éloigné munie de la distance euclidienne.

1- Le tableau des écarts associé à $P_0 = (\{w_1\}, ..., \{w_5\})$ est

	w_1	w_2	w_3	w_4	w_5
w_1	0	/	/	/	/
w_2	5.85	0	/	/	/
w_3	**1.41**	4.60	0	/	/
w_4	3.35	7.07	3.20	0	/
w_5	4.47	1.50	3.16	5.59	0

Les éléments (individus) w_1 et w_3 ont l'écart le plus petit : ce sont les éléments les plus proches. On les rassemble pour former le groupe : $A = \{w_1, w_3\}$. On a une nouvelle partition de Γ :

$$P_1 = (\{w_2\}, \{w_4\}, \{w_5\}, A).$$

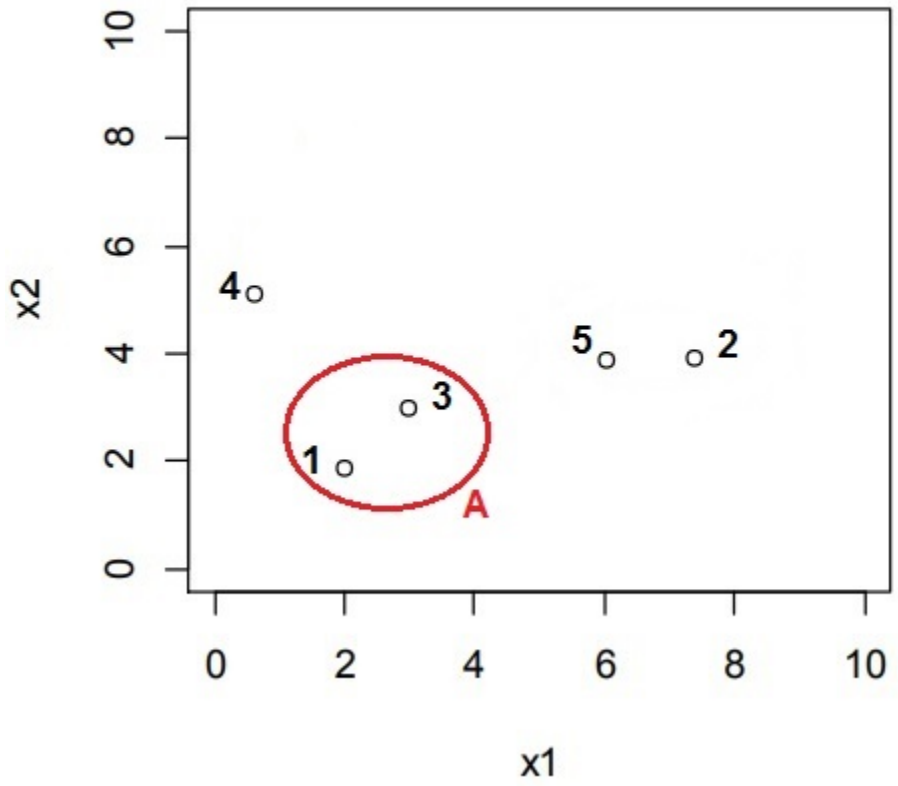

FIGURE 2.2: Présentation graphique 1

2- Le tableau des écarts associé à P_1 est

	w_2	w_4	w_5	A
w_2	0	/	/	/
w_4	7.07	0	/	/
w_5	**1.50**	5.59	0	/
A	5.85	3.35	4.47	0

On a

$e(w_2, A) = max(e(w_2, w_1), e(w_2, w_3)) = max(5.85, 4.60) = 5.85$

$e(w_4,A) = max(e(w_4,w_1),e(w_4,w_3)) = max(3.35,3.16) = 3.35$
et
$e(w_5,A) = max(e(w_5,w_1),e(w_5,w_3)) = max(4.47,3.16) = 4.47.$

Les éléments (individus) w_2 et w_5 sont les plus proches. On les rassemble pour former le groupe : $B = \{w_2,w_2\}$. On a une nouvelle partition de Γ :

$$P2 = (\{w_4\},A,B).$$

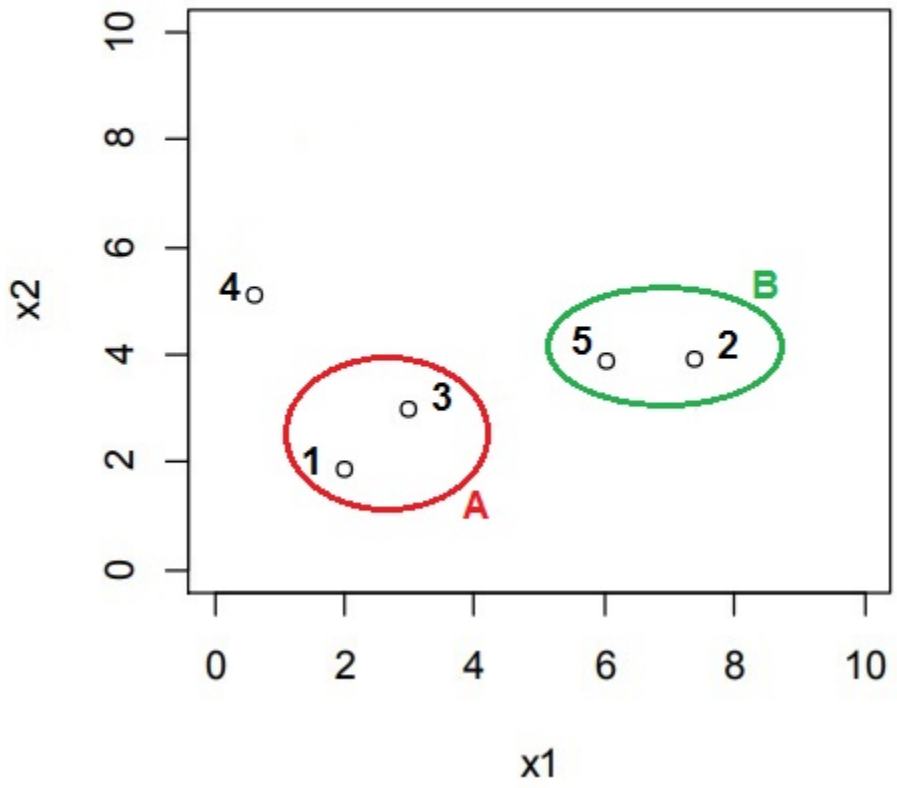

FIGURE 2.3: Présentation graphique 2

3- Le tableau des écarts associé à P_2 est

	w_4	A	B
w_4	0	/	/
A	**3.35**	0	/
B	7.07	5.85	0

On a
$e(B,w_4) = max(e(w_2,w_4),e(w_5,w_4)) = max(7.07,5.59) = 7.07$
et
$e(B,A) = max(e(w_2,A),e(w_5,A)) = max(3.35,3.16) = 3.35$

Les éléments (individus) w_4 et A sont les plus proches. On les rassemble pour former le groupe : $C = \{w_4, A\}$. On a une nouvelle partition de Γ :

$$P3 = (B, C).$$

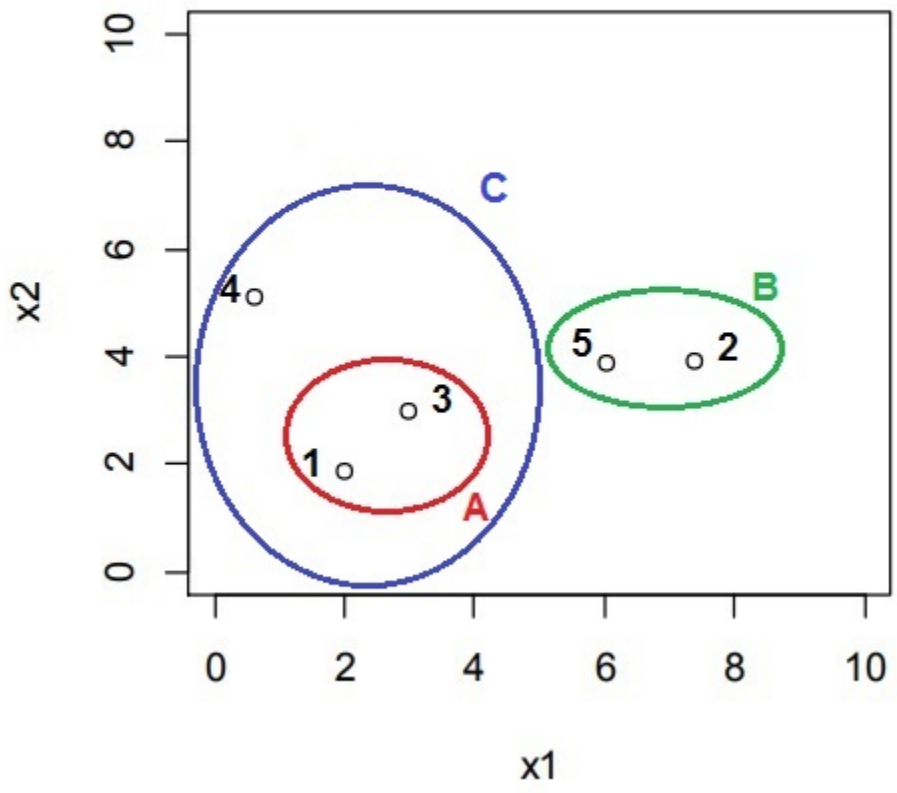

FIGURE 2.4: Présentation graphique 3

4- Le tableau des écarts associé à P_3 est

	B	C
B	0	/
C	**7.07**	0

On a
$e(C,B) = max(e(w_4,B), e(A,B)) = max(7.07, 5.85) = 7.07$

Il ne reste que 2 éléments, B et C ; on les regroupe. On obtient la partition

$$P4 = \{w1, ..., w5\} = \Gamma.$$

Cela termine l'algorithme de CAH.

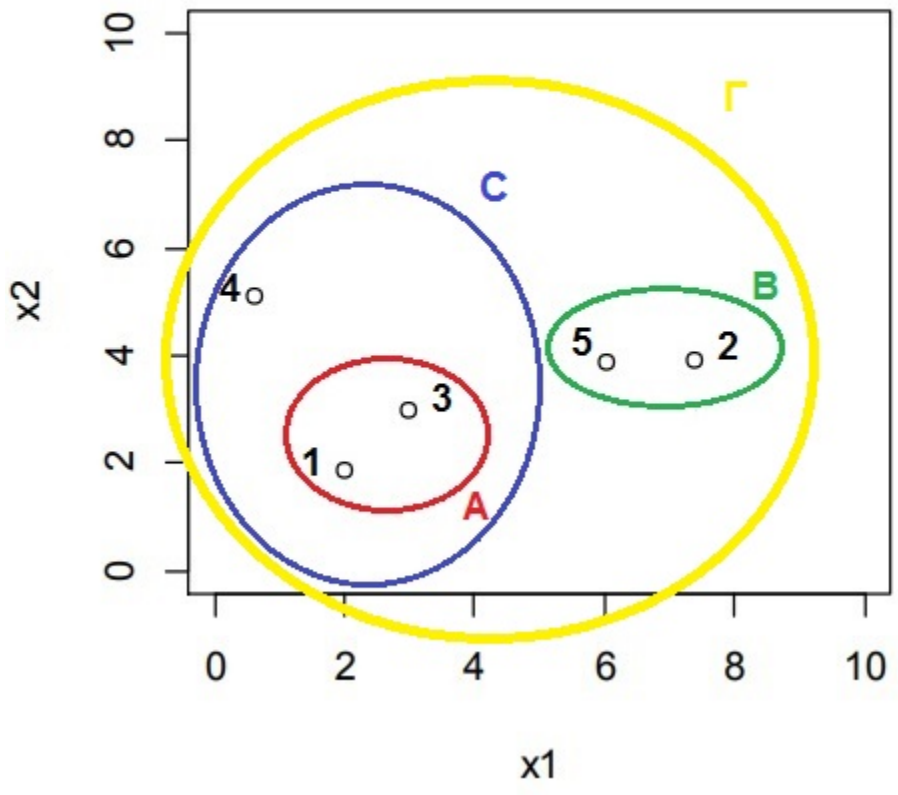

FIGURE 2.5: Présentation graphique 4

- Au final, les éléments $\{w_1\}$ et $\{w_3\}$ ont été regroupés avec un écart de 1.41,
- les éléments $\{w_2\}$ et $\{w_5\}$ ont été regroupés avec un écart de 1.50,
- les éléments $A = \{w_1, w_3\}$ et $\{w_4\}$ ont été regroupés avec un écart de 3.35,
- les éléments $C = \{w_4, A\}$ et $B = \{w_2, w_5\}$ ont été regroupés avec un écart de 7.07.

On peut donc construire le dendrogramme associé :

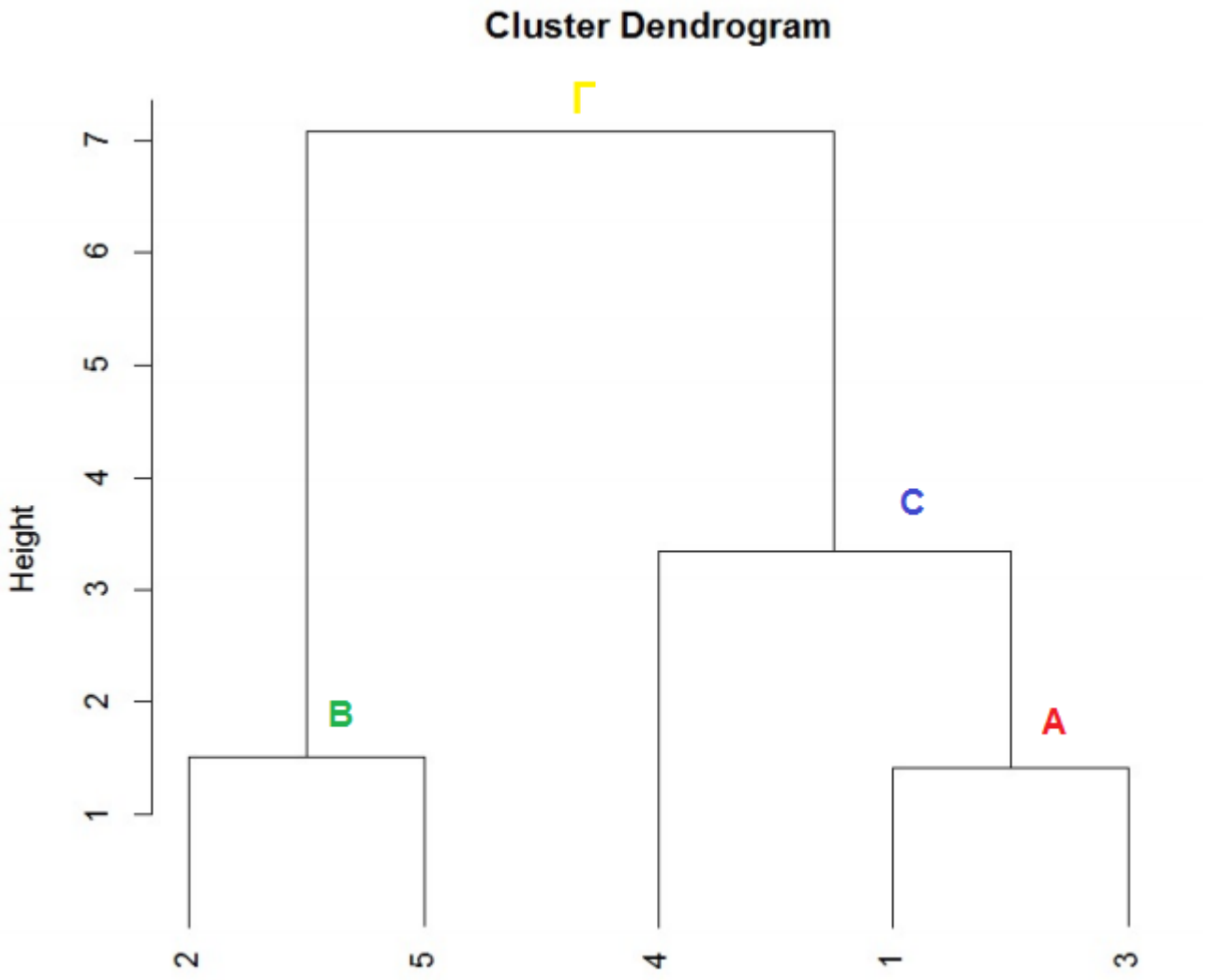

FIGURE 2.6: Dendrogramme

Comme le plus grand saut se situe entre les éléments B et C (on a $7.07 - 3.35 = 3.72$), on propose les deux groupes : B et C.

Méthode des centres mobiles - K-Means

La méthode des k-means a été introduite par J. McQueen en 1971 et mise en œuvre sous sa forme actuelle par E. Forgy . De nombreuses variantes se sont succédées depuis afin d'étendre ses capacités de classification (séparations non linéaires) : kernel k-means (k-means basée sur des méthodes à noyaux), améliorer ses performances : global-k-means , K-Harmonic means, automatiser le choix du nombre de clusters : Gaussian-means, X-means .[16] [6] [22] [14] [23] [9] [17]

L'objectif de la méthode est de partitionner en différentes classes des individus pour lesquels on dispose de mesures. On représente les individus comme des points de l'espace ayant pour coordonnées ces mesures. On cherche à regrouper les individus les plus semblables du point de vue des mesures que l'on possède tout en séparant le mieux possible les classes les unes des autres.[5]
Comme dans la classification hiérarchique ascendante on choisit de procéder de façon automatique, c'est à dire qu'on ne cherche pas à utiliser l'expertise que l'on aurait des individus pour trouver des regroupements mais plutôt à faire apparaître, uniquement à partir des mesures, des ressemblances et des différences à priori peu visibles.

► **Algorithme K-means**

La méthode de k-means s'applique lorsque l'on sait à l'avance combien de classes on veut obtenir. Appelons k ce nombre de classes. L'algorithme est le suivant :

- Etape 0 : Pour initialiser l'algorithme, on tire au hasard k individus appartenant à la population, C_1^0 , C_2^0 , ..., C_k^0 : ce sont les k centres initiaux. On notera que l'indice numérote les différents centres et l'exposant indique qu'il s'agit des k centres initiaux. On choisit aussi une distance entre individus.

- Etape 1 : Constitution de classes : On répartit l'ensemble des individus en k classes $\Gamma_1^0, \Gamma_2^0, ..., \Gamma_k^0$ en regroupant autour de chaque centre C_1^0 pour $i = 1, ..., n$ l'ensemble des individus qui sont plus proches du centre C_i^0 que des autres centres C_j^0 pour $j \neq i$ (au sens de la distance choisie).

- Etape 2 : Calcul des nouveaux centres : On détermine les centres de gravité $G_1, G_2, ..., G_k$ des k classes ainsi obtenues et on désigne ces points comme les nouveaux centres $C_1^1 = G_1, C_2^1 = G_2, ..., C_k^1 = G_k$

- Répétition des étapes 1 et 2 : on répète ces deux étapes jusqu' à la stabilisation de l'algorithme, c'est-à-dire jusqu'à ce que le découpage en classes obtenu ne soit (presque) plus modifié par une itération supplémentaire.

► **Exemple**

Dans cet exemple nous allons travailler avec k = 2 classes . On choisit aussi une distance entre individus,soit disant la distance choisie est la distance euclidienne

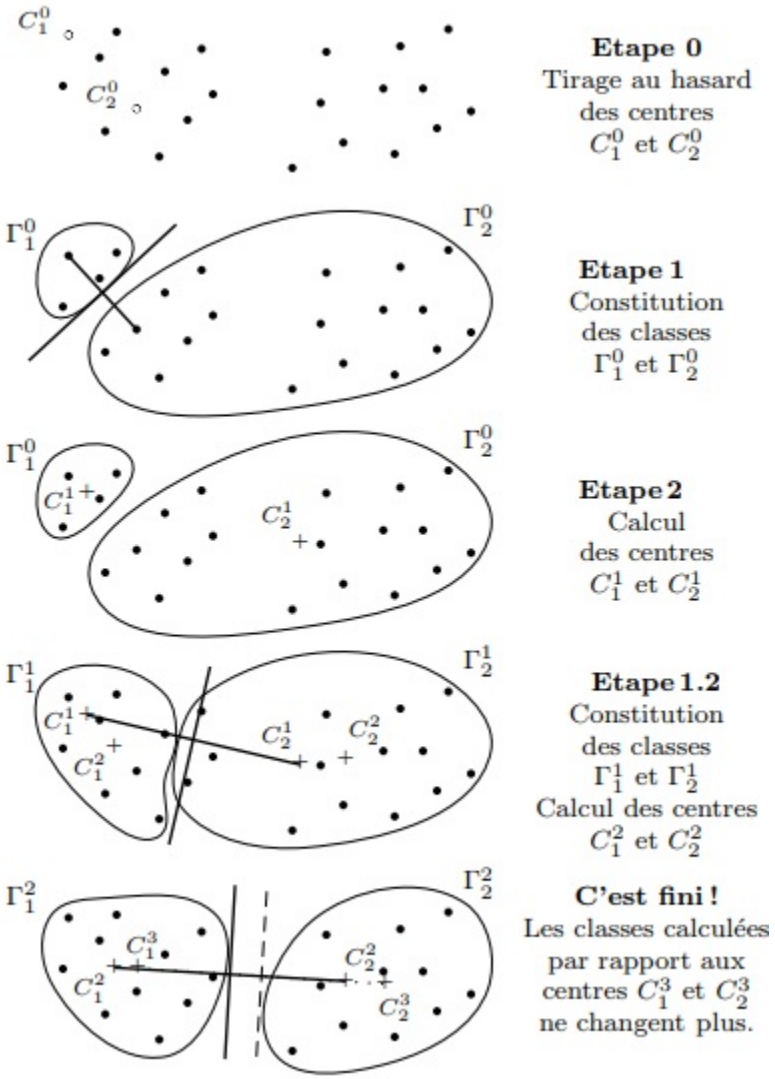

FIGURE 2.7: Exemple k-means

2.3.2 Apprentissage supervisé

L'apprentissage supervisé vise à expliquer ou prévoir plusieurs phénomènes observables et effectivement mesurés. On cherche à prédire la valeur d'une variable cible à partir des valeurs de prédicteurs. Autrement dit on cherche à anticiper la valeur de quelque chose (par exemple, si un client risque de ne pas pouvoir rembourser un prêt, c'est la variable cible) en fonction de ses caractéristiques connues (âge, emploi, salaire... ce sont les prédicteurs), en se basant pour cela sur les données dont on dispose (les précédents clients et les valeurs des prédicteurs et des variables cibles).[20]

Les méthodes d'apprentissage supervisé visent à extrapoler de nouvelles informations à partir des informations présentes, expliquent les données et il y a une variable « cible » à prédire.
La figure 2.8 représente quelques méthodes d'apprentissage supervisé et pour notre étude nous retenons la méthode de régression logistique

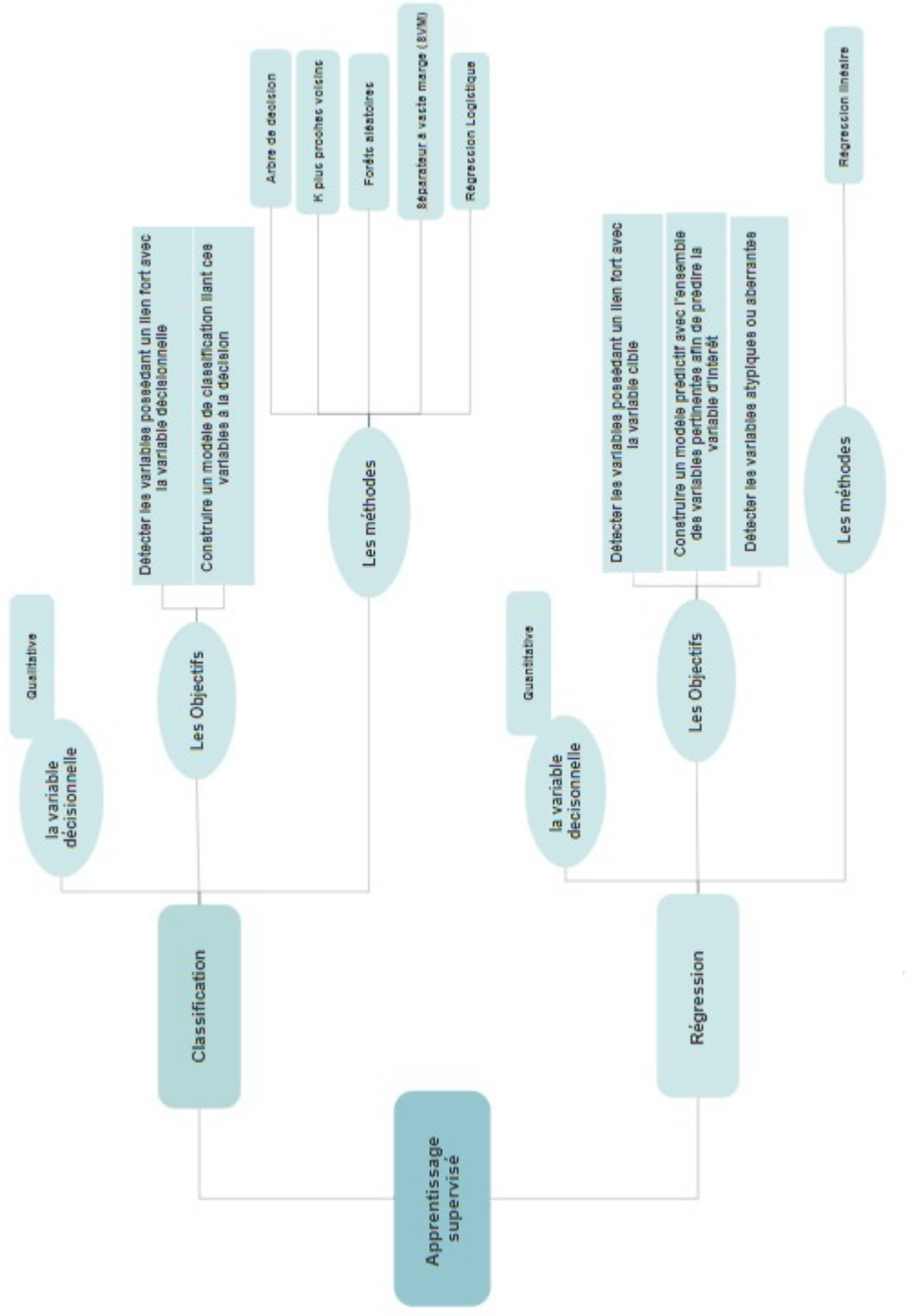

FIGURE 2.8: Les méthodes d'apprentissage supervisé

Régression Logistique

[13] [1] [15] [11]

La régression logistique, cherche à décrire la liaison entre une variable nominale y (variable à expliquer) et un ensemble de p variables $(x_l, x_2, ..., x_p)$. On veut également connaître l'effet d'une variable sur la variable à expliquer en tenant compte des liaisons qu'elle entretient avec les autres variables du modèle. Le plus souvent la variable à expliquer est dichotomique (c'est-à-dire de type « oui / non » ou « vrai / faux » ou « 1 / 0 »). Les variables explicatives qui seront introduites dans le modèle peuvent être quantitatives ou qualitatives.

Un modèle de régression logistique permet de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression. Ce résultat varie toujours entre 0 et 1. Lorsque la valeur prédite est supérieure à $0,5$, l'événement est susceptible de se produire, alors que lorsque cette valeur est inférieure à $0,5$, il ne l'est pas.

► Le modèle logistique

On suppose que la probabilité qu'un individu a d'appartenir au premier groupe I_1 $(y = 1)$ dépend des valeurs des variables explicatives $(x_l, x_2, ..., x_p)$ observées sur cet individu.
On note x le vecteur dont les p composantes sont les valeurs des variables explicatives.
Le modèle logistique se propose de fournir une estimation de cette probabilité notée $\pi(x)$:

$$\pi(x) = P(I_1 \mid x) = P(y = 1 \mid x) \tag{2.1}$$

Théorème de Bayes :

$$P(I_k \mid x) = \frac{P(x \mid I_k)P(I_k)}{P(x)}$$

$$P(I_k \mid x) = \frac{P(x \mid I_k)P(I_k)}{\sum_{k=1}^{q} P(x \mid I_k)P(I_k)} \tag{2.2}$$

Le théorème de Bayes (2.2) nous permet d'écrire dans le cas de deux groupes I_1 et I_2 :

$$P(I_1 \mid x) = \frac{P(x \mid I_1)P(I_1)}{P(x \mid I_1)P(I_1) + P(x \mid I_2)P(I_2)} \tag{2.3}$$

qui s'écrit encore :

$$P(I_1 \mid x) = \frac{\frac{P(x|I_1)P(I_1)}{P(x|I_2)P(I_2)}}{1 + \frac{P(x|I_1)P(I_1)}{P(x|I_2)P(I_2)}} \tag{2.4}$$

Dans le cas multinomiale avec matrices des covariances $\sum$ égales dans les deux groupes, chacune des deux probabilités conditionnelles s'écrit, pour $k = 1,2$

$$P(I_k \mid x) = (2\pi)^{-p/2} |\sum|^{-1/2} \exp\{-\frac{1}{2}(x-\mu_k)' \sum^{-1} (x-\mu_k)'\} \tag{2.5}$$

Le quotient des probabilités pondérées fait disparaître les termes du second degré en x et s'écrit comme l'exponentielle d'une forme linéaire en x avec terme constant (fonction affine de x) :

$$\frac{P(x \mid I_1)P(I_1)}{P(x \mid I_2)P(I_2)} = \exp\{\beta' x + b\} \tag{2.6}$$

Pour alléger les notations, le vecteur x désignera désormais un vecteur à $p+1$ composantes (avec $x_0 = 1$ et les autres composantes égales à celles de l'ancien x) et le nouveau vecteur de coefficients sera désigné par α, de sorte que $\beta' x + b$ s'écrit maintenant $\alpha' x$.

Ceci permet de réécrire la formule (2.4) et conduit à l'expression du modèle logistique :

$$\pi(x) = \frac{\exp\{\alpha' x\}}{1 + \exp\{\alpha' x\}} = \frac{\exp\{\sum_{j=0}^{p} \alpha_j x_j\}}{1 + \exp\{\sum_{j=0}^{p} \alpha_j x_j\}} \tag{2.7}$$

où les α_j, composantes du vecteur α, sont les coefficients inconnus du modèle. Il s'agit d'un modèle qui ne fait pas intervenir de termes d'interaction entre les variables explicatives.

On peut écrire (2.7) sous la forme :

$$\frac{\pi(x)}{1-\pi(x)} = \exp\{\alpha' x\} \tag{2.8}$$

ou encore :

$$\log \frac{\pi(x)}{1-\pi(x)} = \alpha' x = \sum_{j=0}^{p} \alpha_j x_j \tag{2.9}$$

la fonction :

$$F(\pi(x)) = \log \frac{\pi(x)}{1-\pi(x)} \tag{2.10}$$

est appelée fonction **Logit**.
La quantité $\frac{\pi(x)}{1-\pi(x)} = \frac{P(Y=1|X)}{P(Y=0|X)}$ exprime un **odds** c.-à-d. un rapport de chances. Par exemple, si un individu présente un odds de 2, cela veut dire qu'il a 2 fois plus de chances d'être 1 que d'être 0.
Lorsque $\pi(x)$ varie dans $]0;1[$, la fonction Logit prend ses valeurs dans l'intervalle $]-\infty;+\infty[$ tout entier.

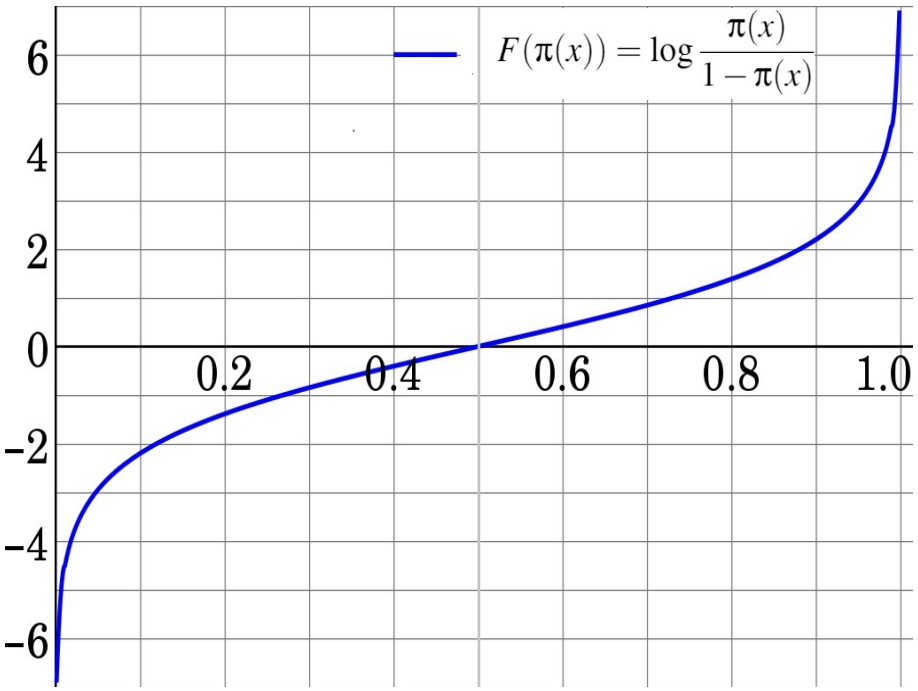

FIGURE 2.9: Représentation graphique de la fonction logit

►Estimation et tests des coefficients

Pour estimer les coefficients α_j du modèle, on utilise le plus souvent la méthode du maximum de vraisemblance.
Les n observations (y_i, x_i) où $x_i = (x_{i1}, x_{i2}, ..., x_{ip})^{'}$ sont indépendantes et les y_i sont des variables de Bernoulli.
La vraisemblance $\pounds(\alpha, y_i)$ pour une observation s'écrit :

$$\pounds(\alpha, y_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{2.11}$$

et pour l'ensemble des observations, on a :

$$\pounds(\alpha, y) = \prod_{i=1}^{n} \pounds(\alpha, y_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{2.12}$$

La procédure d'estimation revient à rechercher la valeur $\hat{\alpha}$ de α qui maximise le logarithme de la vraisemblance :

$$\log [\pounds(\alpha, y)] = \sum_i \left[y_i \log \frac{\pi(x_i)}{1 - \pi(x_i)} + \log [1 - \pi(x_i)] \right] \tag{2.13}$$

Soit encore en exprimant $\pi(x_i)$ par la fonction Logit (cf. (2.8), (2.9)) :

$$\log [\pounds(\alpha, y)] = \sum_i y_i \alpha^{'} x_i + \sum_i \log \left[1 + \exp (\alpha^{'} x_i) \right] \tag{2.14}$$

Pour apprécier l'éventuelle non-influence d'une variable ou d'une modalité x_j sur la variable y, on teste l'hypothèse nulle H_0 :

$$H_0 : \alpha_j = 0 \tag{2.15}$$

On considère alors la statistique de Student :

$$t = \frac{\hat{\alpha}_j}{\sqrt{Var(\hat{\alpha}_j)}} \tag{2.16}$$

où $\hat{\alpha}_j$ est la j^{eme} composante de l'estimateur $\hat{\alpha}$ et $Var(\hat{\alpha}_j)$ est la variance estimée associée à cette composante.
Pour tester l'influence d'une variable nominale à q modalités, on procède à un test de nullité des q coefficients α_j affectés à ses modalités. D'une manière générale, l'hypothèse H_0 stipulant une éventuelle non-influence d'un ensemble de q variables $(x_1, x2, ..., x_p)$ sur y, s'exprime par la nullité des q coefficients associés :

$$H_0 : \alpha_1 = \alpha_2 ... = \alpha_q = 0 \tag{2.17}$$

Notons $\hat{\alpha}_0$ l'estimateur des α_j sous l'hypothèse H_0 et â l'estimateur des coefficients du modèle alternatif.

On teste l'hypothèse nulle en calculant la statistique du rapport de vraisemblance

$$\wedge = 2(\pounds(\hat{\alpha}, y) - \pounds(\hat{\alpha_0}, y)) \tag{2.18}$$

On démontre qu'elle suit une distribution du χ^2 à q degrés de liberté sous des hypothèses de travail convenables. Si l'hypothèse nulle est rejetée, on en déduit qu'au moins une des q variables (ou une modalité de la variable nominale) influe sur la variable y.

2.4 Processus du datamining

Il est très important de comprendre que le datamining n'est pas seulement le problème de découverte de modèles dans un ensemble de données. Ce n'est qu'une seule étape dans tout un processus suivi par les scientifiques, les ingénieurs ou toute autre personne qui cherche à extraire les connaissances à partir des données. En 1996 un groupe d'analystes définit le datamining comme étant un processus composé de cinq étapes sous le standard CRISP-DM (Cross-lndustry Standard Process for Data Mining) [10], comme schématisé ci-dessous

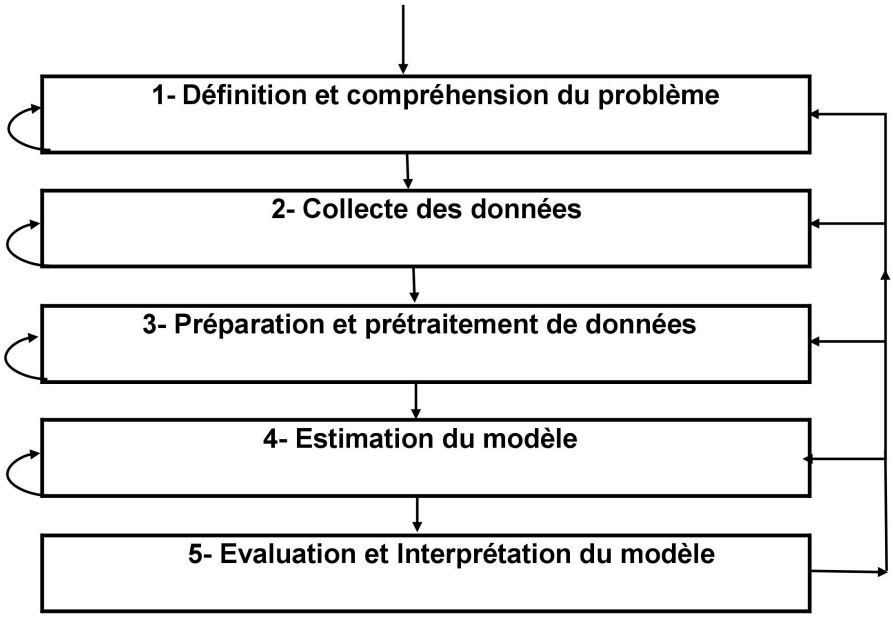

FIGURE 2.10: Processus du datamining

Ce processus, composé de cinq étapes, n'est pas linéaire, on peut avoir besoin de revenir à des étapes précédentes pour corriger ou ajouter des données. Par exemple, on peut découvrir l'étape d'exploration (5) de nouvelles données qui nécessitent d'être ajoutées aux données initiales à l'étape de collection (2). Décrivons maintenant ces étapes.

2.4.1 Définition et compréhension du problème

Dans la plupart des 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. En effet, Avec la compréhension du problème, on peut préparer les données nécessaires à l'exploration et interpréter correctement les résultats obtenus. Généralement, le datamining est effectué dans un domaine particulier (banques, médecine, biologie, marketing, ...etc) où la connaissance et L'expérience dans ce domaine jouent un rôle très important dans la définition du problème, l'orientation de l'exploration et l'explication des résultats obtenus. Une bonne compréhension du problème comporte une mesure des résultats de l'exploration, et éventuellement une justification de son coût. C'est-à-dire, pouvoir évaluer les résultats obtenus et convaincre l'utilisateur de leur rentabilité.

2.4.2 Collecte des données

dans cette étape, on s'intéresse à la manière dont les données sont générées et collectées. D'après la définition du problème et des objectifs du datamining, 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 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, etc.).

2.4.3 Préparation et prétraitement de données

L'objectif de la préparation des données est de constituer un jeu de données de qualité homogène, dont la structure et les formats sont cohérents et bien définis.
Avant de traiter les données, il faut vérifier leur qualité. Les données peuvent être : manquantes, aberrantes ou en double.
Les valeurs manquantes ou aberrantes sont présentes pratiquement dans toutes les bases de données des applications réelles. Elles peuvent correspondre aux erreurs de saisie ou à la naïveté de l'enquêteur. La mauvaise gestion de ces valeurs peut conduire à l'induction de modèles erronés et à des analyses fallacieuses. Le traitement des valeurs manquantes et/ou aberrantes est souvent une tâche exigeante, tant du point de vue méthodologique qu'en termes de calcul.
Le nettoyage des données consiste à éliminer toutes les informations que l'on ne souhaite pas conserver telles quelles. Ces informations peuvent être erronées, inexactes, induire des erreurs ou simplement être sans intérêt pour la suite de l'analyse ou la modélisation.
La phase de préparation des données peut également consister à manipuler, modifier voire créer de nouvelles informations à partir des informations disponibles.[3]

Valeurs manquantes

Les valeurs qui n'ont pas pu être observées, perdues ou incohérentes sont considérer comme des valeurs manquantes.

► Méthodes d'imputation

L'imputation consiste à produire une « valeur artificielle » pour remplacer la valeur manquante, avec pour objectif de produire des estimations approximativement sans biais. Les deux méthodes les plus courantes de l'imputation sont les suivantes

a. **Imputation par la moyenne ou le mode**

   • **Variables quantitatives :** On remplace chacune des valeurs manquantes par la valeur moyenne de l'ensemble de réponses obtenues c'est-à-dire, si les valeurs manquantes sont absentes pour des raisons vraiment aléatoires, on peut sans gros problème les remplacer par la moyenne ou la médiane des variables correspondantes.

   • **Variables qualitatives :** On remplace chacune des valeurs manquantes par le mode

b. **Imputation par le ratio / régression**

   • **Imputation par le ratio :** chaque valeur manquante y_i est remplacée par la valeur prévue $\hat{y}_i$ obtenue par régression de y sur x.

   • **Imputation par régression :** c'est une extension naturelle de l'imputation par la méthode du ratio où l'on se sert de q variables auxiliaires $x_1, \ldots, x_q$.

Valeurs aberrantes

Avant d'entreprendre l'imputation des données manquantes, on doit chercher s'il n'y a pas des valeurs aberrantes.

► Définition

Une valeur aberrante est une valeur qui diffère de façon significative de la tendance globale des autres observations quand on observe un ensemble de données ayant des caractéristiques communes. Le traitement des valeurs aberrantes est complexe, Il y a trois possibilités pour traiter ses valeurs :

a. Les valeurs aberrantes pouvant provenir d'erreurs de saisie. Si c'est le cas, on retourne au questionnaire papier quand c'est possible et on corrige. Si on ne retrouve pas le questionnaire, on les supprime et on applique ensuite une des méthodes d'imputation (moyenne, médiane…).

b. Si la valeur a été bien saisie (erreur d'échantillonnage ou due par la méthode d'échantillonnage), on la laisse comme ça et on fait les analyses avec.

c. Normaliser les variables par une ACP (analyse en composantes principales.[2]

2.4.4 Estimation du modèle

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

Méthodes de rééchantillonnages

À partir d'un jeu de données initial On crée un échantillon d'entraînement, sur lequel on va constituer le modèle, et un échantillon de test, sur lequel on va tester le modèle. En pratique, on a l'habitude de prendre 75% des données pour l'échantillon d'entraînement et 25% des données pour l'échantillon de test. En effet, on pourrait avoir envie d'utiliser cette séparation des données pour faire le meilleur modèle possible. On pourrait ainsi essayer différents choix de variables, plusieurs paramétrages d'un modèle sur l'échantillon d'apprentissage et voir lequel performe le mieux sur l'échantillon test. C'est une idée effectivement perspicace, puisqu'elle nous permettrait de trouver le modèle optimal.

2.4.5 Evaluation et Interprétation du modèle

Il est souvent très facile de construire un modèle qui restitue très bien les données utilisées pour son estimation. Il est néanmoins bien plus difficile de faire en sorte que ce modèle puisse se généraliser, c'est-à-dire qu'il soit capable de prédire de façon satisfaisante de nouvelles observations, non utilisées lors du calcul du modèle. Pour trouver un juste équilibre entre apprentissage du modèle et capacité prédictive, il est indispensable de mettre en place un dispositif qui permette d'évaluer globalement la qualité d'un modèle.

Autrement dit Les modèles extraits, notamment par les méthodes d'apprentissage supervisé, ne peuvent être utilisés directement en toute fiabilité. Nous devons les évaluer, c'est-à-dire les soumettre à l'épreuve de la réalité et apprécier leur justesse. Le procédé habituel consiste à estimer au mieux le taux d'erreur du modèle. Ainsi, l'utilisateur décidera d'appliquer ou non le modèle de prédiction en connaissance des risques qu'il prend.

Matrice de confusion

Une matrice de confusion [12] contient des informations sur les classifications réelles et prédites effectuées par un système de classification. Les performances de tels systèmes sont généralement évaluées à l'aide des données de la matrice.
Le tableau suivant montre la matrice de confusion pour un classifieur à deux classes :

		Valeur prédite	
		Négatif	Positif
Valeur réel	Négatif	VN	FP
	Positif	FN	VP

Les entrées de la matrice de confusion ont la signification suivante :

- VN est le nombre de prédictions correctes qu'une instance est négative (Vrais négatif)
- FP est le nombre de prédictions incorrectes selon lesquelles une instance (Faux positif) est positive ,
- FN est le nombre de prédictions incorrectes qu'une instance est négative (Faux négatif)
- VP est le nombre de prédictions correctes qu'une instance est positive (Vrais positif)

  Plusieurs termes standard ont été définis pour la matrice à 2 classes :

  (a) La précision totale (PT) est la proportion du nombre total de prédictions correctes. Il est déterminé en utilisant l'équation :

  $$\text{PT} = \frac{VN+VP}{VN+FP+FN+VP}$$

  (b) L'erreur est la proportion du nombre total de prédictions incorrectes. Il est déterminé en utilisant l'équation :

  $$\text{Taux d'erreur} = \frac{FP+FN}{VN+FP+FN+VP}$$

  (c) Le rappel ou le véritable ratio positif autrement dit la sensibilité est la proportion de cas positifs correctement identifiés, calculée à l'aide de l'équation :

  $$\text{La sensibilité} = \frac{VP}{FN+VP}$$

  (d) Le véritable taux négatif autrement dit la spécifité est défini comme la proportion de cas négatifs correctement classés, calculée à l'aide de l'équation :

  $$\text{La spécifité} = \frac{VN}{VN+FP}$$

  (e) Le taux de faux positifs est la proportion de cas négatifs classés incorrectement comme positifs, calculés à l'aide de l'équation :

  $$\text{Le taux de faux positifs} = \frac{FP}{VN+FP}$$

  (f) Le taux de faux négatifs est la proportion de cas positifs mal classés comme négatifs, calculée à l'aide de l'équation :

  $$\text{Le taux de faux négatifs} = \frac{FN}{FN+VP}$$

  (g) Enfin, la précision (P) est la proportion des cas positifs prédits qui étaient corrects, calculée à l'aide de l'équation :

  $$\text{P} = \frac{VP}{FP+VP}$$

Validation croisée

Le principe de la validation croisée c'est de diviser l'échantillon original en k échantillons, puis on sélectionne un des k échantillons comme ensemble de validation et les $k-1$ autres échantillons constitueront l'ensemble d'apprentissage. On calcule l'erreur quadratique moyenne , puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les $k-1$ échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs quadratiques moyennes est enfin calculée pour estimer l'erreur de prédiction.[24]

La courbe ROC

Les graphes ROC sont un autre moyen, outre les matrices de confusion, d'examiner les performances des classificateurs (Swets, 1988). Un graphique ROC est un graphique avec le taux de faux positifs sur l'axe des X et le taux de vrais positifs sur l'axe des Y. Le point (0,1) est le classificateur parfait : il classe correctement tous les cas positifs et négatifs. C'est $(0,1)$ parce que le taux de faux positifs est 0 (aucun) et le taux de vrais positifs est 1 (tous). Le point $(0,0)$ représente un classificateur qui prédit que tous les cas sont négatifs, tandis que le point $(1,1)$ correspond à un classificateur qui prédit que chaque cas est positif. Le point $(1,0)$ est le classificateur incorrect pour toutes les classifications.[11]

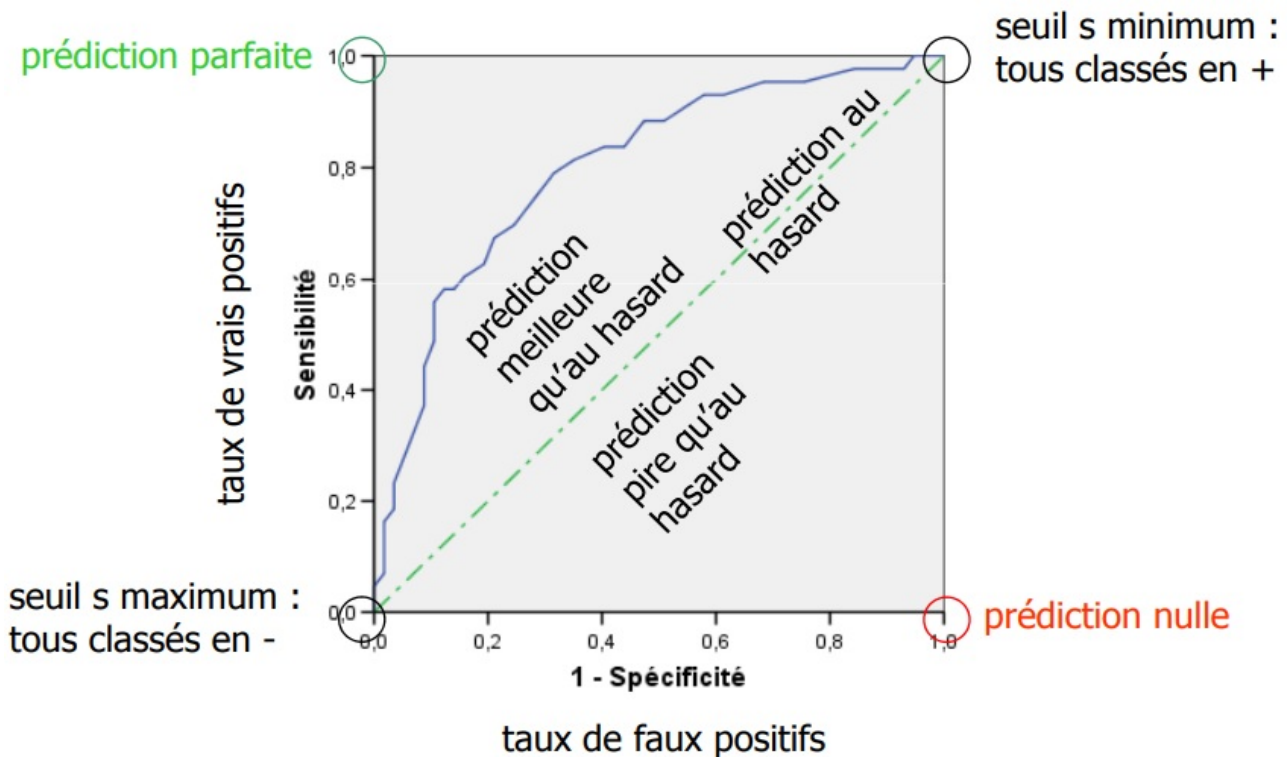

FIGURE 2.11: La courbe ROC

Chapitre 3

Cas pratique

3.1 choix de logiciel

R est un logiciel complet et open source qui répondait efficacement à mon problème. L'avantage de R est qu'il possède de nombreuses librairies et qu'en plus d'être bien documentée, les différents algorithmes implémentés sont entièrement personnalisables et paramétrables, donnant une grande liberté d'action dans l'optimisation des modèles.
Etant donné que mon projet a évolué progressivement vers une méthode de classification par apprentissage supervisé, R s'est imposé comme le choix le plus judicieux et celui qui me correspondait le plus.

3.2 Description des variables

Nom de la variable	Nom en base	Type	Modalités
Dossier de prêt	CIBLE	Qualitative	- 1. Bon (remboursé) - 0. Mauvais (non remboursé)
Situation familiale	Code_Sit_Fam	Qualitative	- 1. Célibataire - 2. Marié
Sexe du client	CODE_SEXE	Qualitative	- M. Male - F. Femelle
Age du client	AGE	Qualitative	- 18-21 - 21-24 - 25-28 - 29-32 - 33-36 - 37-40
Niveau d'instruction	CODE_NIV	Qualitative	- 0.Non défini - 1. Primaire - 2. Moyen - 3. Secondaire - 4. Universitaire - 5. Formation professionnelle

Expérience professionnelle	EXPERIANCE_PRO	Qualitative	- VRAI - FAUX
Client décidé	Personne.DecedŽe	Qualitative	- VRAI - FAUX
Client handicapé	Handicape	Qualitative	- VRAI - FAUX
Relation entre la formation du client et la nature de projet	Correlation_Proff_Projet	Qualitative	- 1. Oui - 0. Non
Montant prêt bancaire effectif	Mont_PretBancaire_Effectif	Quantitative	————————
Montant apport personnel prévu	Mont_Apport_Personnel _Prevu	Quantitative	————————
Numéro d'identification statistique	NIS	Qualitative	————————
Dossier financé	«Dossier.FinancŽ	Qualitative	- VRAI - FAUX
Dossier annulé	Dossierannule	Qualitative	- VRAI - FAUX
Dossier rejeté	dossier.RejetŽ	Qualitative	- VRAI - FAUX
Dossier transféré	XTransfertSortant	Qualitative	- VRAI - FAUX

3.3 Description du jeu de données

Le jeu de données se compose de 39 variables décrivant 2755 clients de l'ANSEJ. Cette base de données retrace l'historique des clients.
Nous cherchons à modéliser une variable qualitative binaire (remboursement /Non remboursement) d'un crédit à l'aide d'un ensemble déterministe de variables explicatives qualitatives et quantitatives et à déterminer les relations d'interactions possibles entre ces variables explicatives. Pour ce faire, il existe de nombreuses méthodes dites de « Apprentissage supervisée » parmi lesquelles la régression logistique, les arbres de décision, etc. Nous allons présenter une méthode appropriée la Régression Logistique .

3.3.1 Présentation de la régression logistique

La régression logistique est un type de modèle d'apprentissage supervisé qui appartient à la famille des Modèles Linéaires Généralisés. Elle s'utilise lorsque la variable à expliquer est qualitative, le plus souvent binaire. Les variables explicatives peuvent être par contre soit qualitatives ou soit quantitatives. La variable dépendante est habituellement la survenue ou non d'un évènement et les variables explicatives sont celles susceptibles d'expliquer la survenue de cet évènement. Contrairement à la régression linéaire multiple et l'analyse discriminant, la régression logistique n'exige pas une distribution normale des prédicteurs, ni l'homogénéité des variances. Par ses nombreuses qualités donc, cette technique est de plus en plus préférée à l'analyse discriminante par le data miner.

3.3.2 Explication du choix de la méthode

La régression logistique convient pour la modélisation d'événements. On cherche à prédire la probabilité d'un remboursement /Non remboursement d'un crédit en expliquant la variable binaire «cible» qui suit une loi binomiale. Dans pareil cas, la nécessité d'utiliser des modèles particuliers se justifie par le fait que :

- L'utilisation d'un modèle de régression linéaire classique n'est plus adéquat ;
- La mise en œuvre d'une régression linéaire va produire des valeurs continues qui ne s'interprètent pas comme des probabilités, or on veut uniquement des probabilités sur [0,1] ;
- La variable de réponse « cible » suit une loi binomiale. Ainsi, la régression logistique peut alors être utilisée pour prédire cette variable en fonction de variables quantitatives, continues, binaires et qualitatives.

3.4 Analyse préliminaire du jeu de données

Ces données nécessitent un travail préliminaire d'exploration et de nettoyage afin de construire le meilleur modèle possible. Cette étape nous a conduit à corriger, transformer et supprimer des variables ou individus. Ces traitements sont pris en compte dans le programme R.

3.4.1 Nettoyage des données

Pour afficher la structure du jeu de données sous R, on utilise la commande (str(data)).
Après visualisation de la structure des données (str(data)) , nous avons procédé à :

La Suppression des variables inutiles

Après importation des données sous R, il a été décidé de supprimer les variables comme date_dépôt de dossier car leur étude n'est pas pertinente.

Recodage des classes des variables

Certaines variables qualitatives sont décrites comme quantitatives comme par exemple la variable « CIBLE» , donc il est important de la transformer en variable qualitative. Il est nécessaire de faire cela en premier car les données manquantes sont remplacées selon le type des variables.

Suppression des variables manquantes

La suppression des variables dont les valeurs manquantes représentent plus que les valeurs renseignées comme «NIS».

Traitement des données manquantes

Le jeu de données compte 5419 données manquantes. Pour procéder au remplacement de ces valeurs, nous avons utilisé une fonction nommée
« traitN » qui remplace la donnée manquante dans chaque variable qualitative par son mode et par la moyenne chaque donnée manquante dans la variable quantitative. Après vérification, il ne reste pas de valeurs manquantes, on peut continuer la phase de préparation des données.

3.5 Analyse descriptive des données

Afin d'étudier la nature des variables présentes , nous les séparons en deux data frame : (data.quanti) et (data.quali) qui contiennent respectivement les variables quantitatives et les variables qualitatives.

3.5.1 Représentation graphique des données

Représentation graphique des variables qualitatives

La figure 3.1 démontre la présence de plusieurs variables comme «Dossier.FinancŽ », « Dossier annule», « dossier.RejetŽ», «XTransfertSortant», sont supprimées car elles n'apportent aucune information.

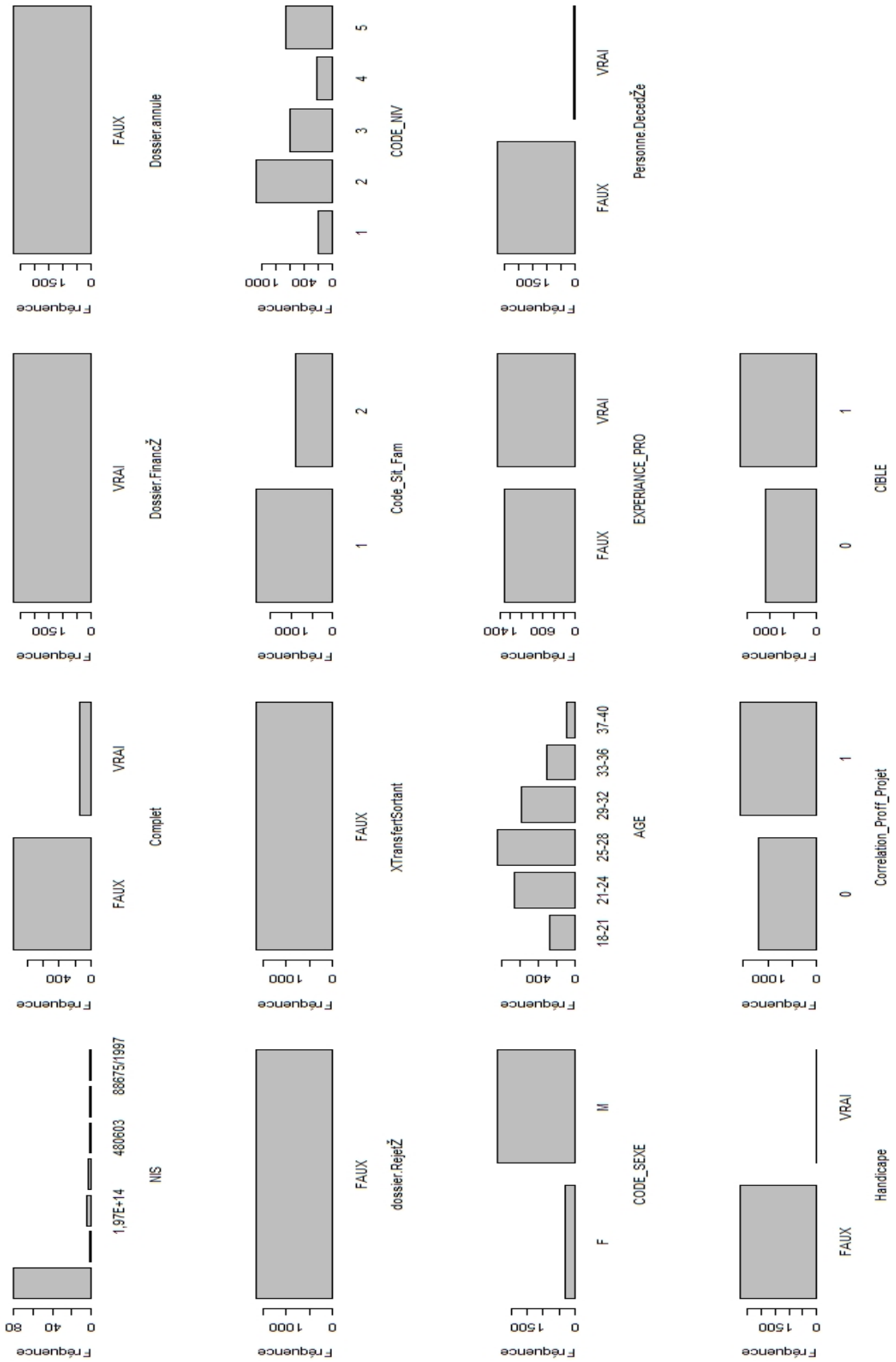

FIGURE 3.1: Représentation graphique des données qualitatives par des diagrammes en barres

3.5.2 Etude de la colinéarité entre les variables quantitatives

La figure montre les coefficients de corrélation entre les variables explicatives quantitatives. Un nombre substantiel de variables s'avèrent fortement colinéaires, par exemple Mont_PretBancaire_Prevu et Mont_PretBancair_effectif. Le but de la corrélation est de déterminer les variables quantitatives qui sont fortement liées entre elles ,dans notre cas on ne garde que les variables qui ont un coefficient de corrélation qui est au-dessous de 0,75

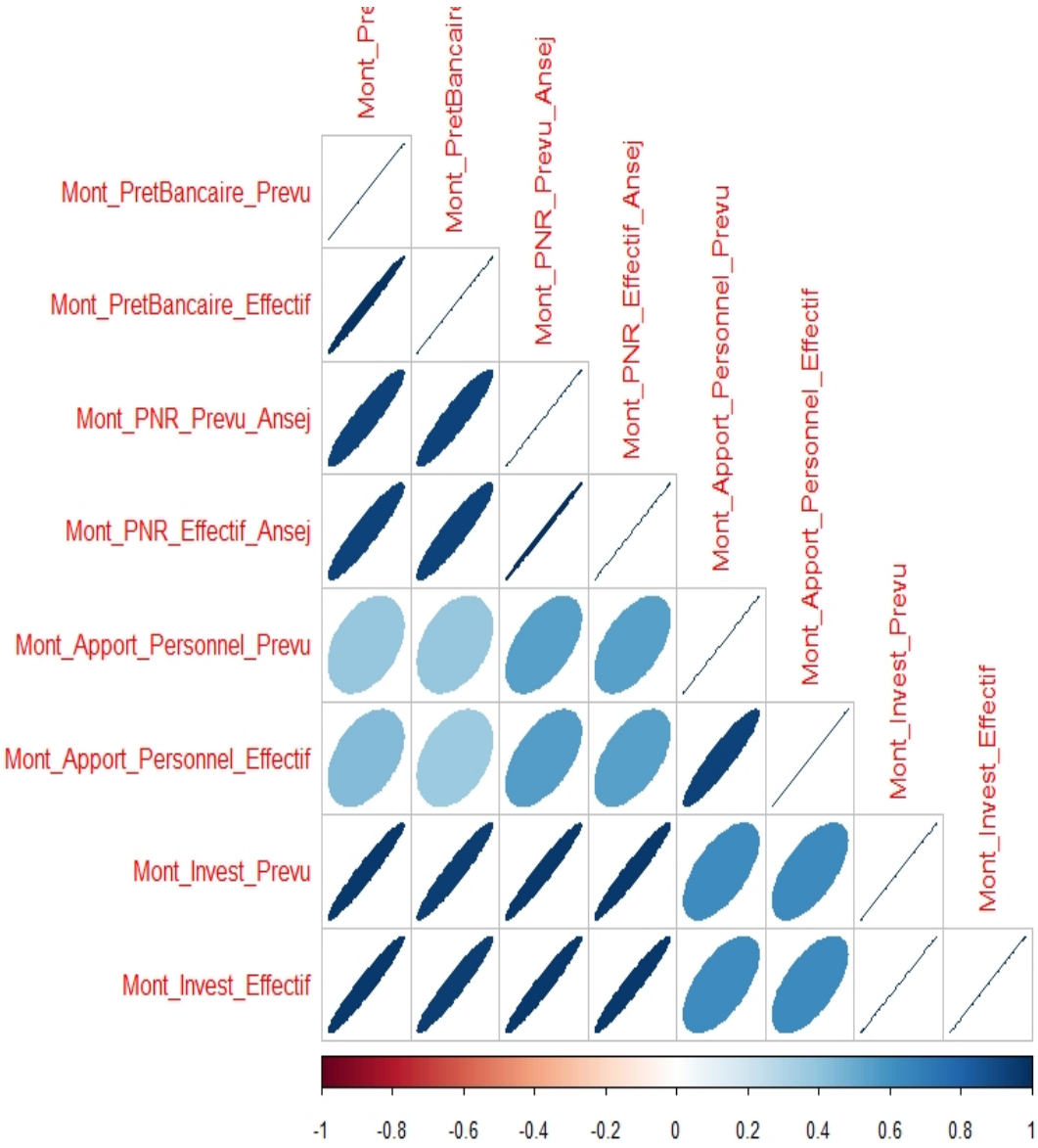

FIGURE 3.2: Matrice des corrélations entre les variables quantitatives

Les corrélations positives sont affichées en bleu et les corrélations négatives en rouge. L'intensité de la couleur et la taille des cercles sont proportionnelles aux coefficients de corrélation. A droite du corrélogramme, la légende de couleurs montre les coefficients de corrélation et les couleurs correspondantes.

3.6 Analyse en ACP des clients pour la détection d'outliers

L'Analyse en Composantes Principales est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables corrélées en nouvelles variables décorrélées les unes des autres. Ces nouvelles variables sont nommées composantes principales. Elle permet de réduire le nombre de variables et de rendre l'information moins redondante.
Il y a un double intérêt à l'utiliser : travailler avec des variables non corrélées et repérer les indices aberrants grâce aux graphiques.
Pour faire l'ACP, on utilise le package « PCAmixdata ».
Le graphe suivant représente les individus qui ont remboursé ou pas le crédit. sur les deux premières composantes principales .(40% NON ;60% OUI)

Dans notre cas on élimine les individus aberrants (167,124,1864,1748,78,88,140,45,1884, 116,1909, 1676, 1908,1845,1938,1889,1771,1821,919,659,1084,1283,604,1787,223,2094,1531,862, 1614,2108,2371,2518,2442,2472,2309)

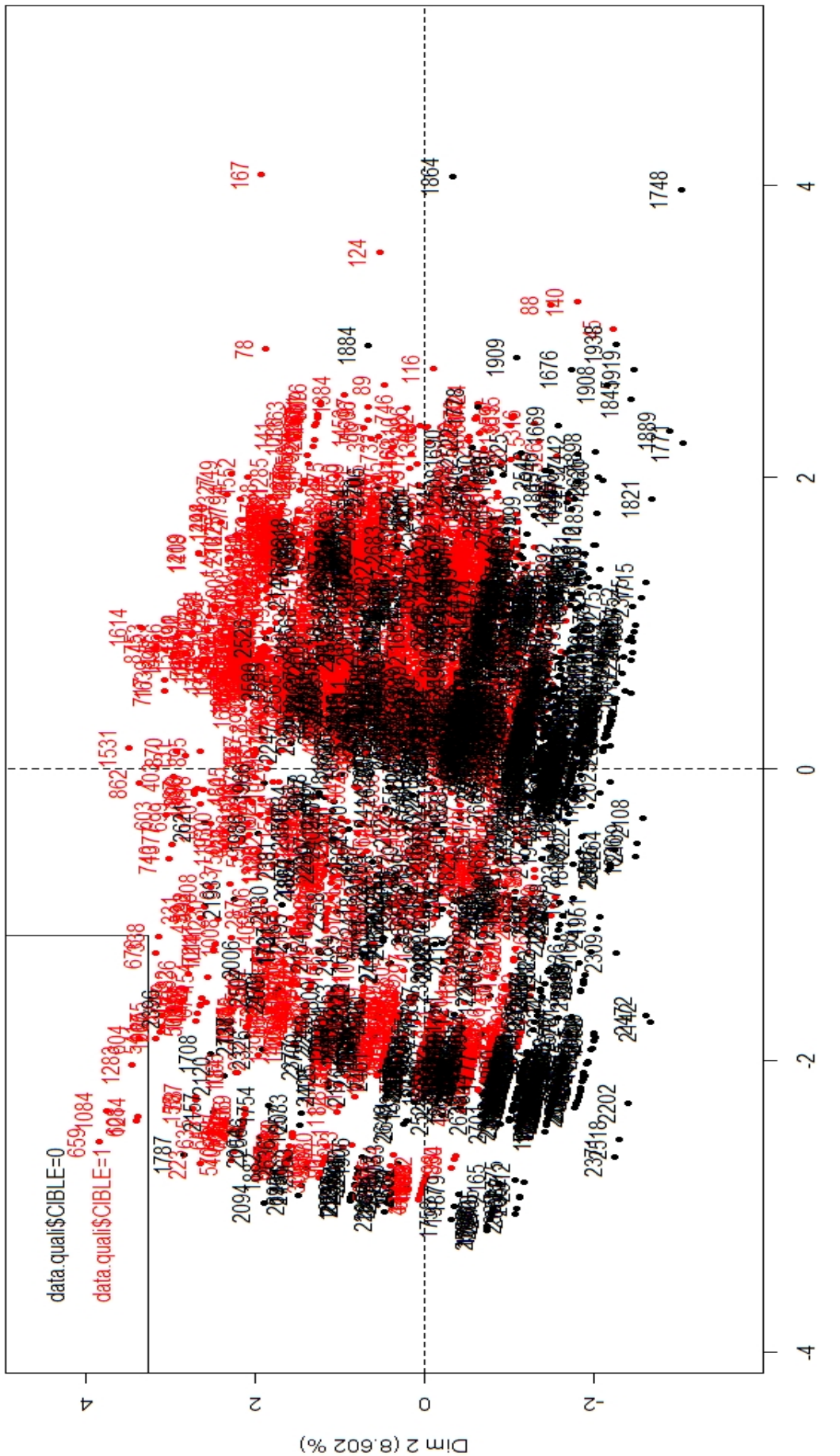

FIGURE 3.3: Représentation graphique des individus de l'ACP

3.7 Confection de modèle

maintenant notre jeu de données qui se compose de 2720 individus et de 11 variables (2 variables quantitatives et 9 variables qualitatives). est prêt pour la confection de modèle.

3.7.1 Régression logistique avec GLM selon les méthodes de rééchantillonnages

L'intérêt de cette technique est de construire le modèle sur un échantillon d'apprentissage et le prédire sur l'échantillon test.
nous séparons notre échantillon en deux échantillons, un échantillon d'apprentissage de 75% (2040 individus) nommé «Dtrain» et un échantillon test de 25% (680 individus) nommé «Dtest».

```
> print(mc.glm1) # Matrice de confusion
   class.fit.glm1
      0   1
  0 140 103
  1  98 339
> # Taux d'erreur
> err.glm1<- 1-sum(diag(mc.glm1))/sum(mc.glm1)
> print(err.glm1) # 0.29 55882
[1] 0.2955882
```

3.7.2 Régression logistique avec GLM en validation croisée

Le modèle doit être construit selon le critère de validation croisée qui consiste à découper nos données en K échantillons, on sélectionne un premier échantillon sur lequel on va tester le modèle construit sur les autres échantillons, et ainsi de suite jusqu'à ce que tous les échantillons aient étés utilisés. On obtient finalement K erreurs, l'erreur finale est la moyenne de ces erreurs. L'itération du découpage par validation croisée consiste à estimer les erreurs de classement avec des variances plus faibles. Dans notre cas, nous divisions nos données en 10 échantillons de manière aléatoire. Nous obtenons après échantillonnage 10 échantillons de 272 individus pris aléatoirement dans nos données, sans remise (c'est-à-dire qu'un individu appartient à un seul échantillon et chaque individu est représenté une seule fois).
Le vecteur d'erreur est donné par

```
> # Les erreurs récupérées :
> print(err.cv)
 [1] 0.2904412 0.3161765 0.2977941 0.3345588 0.3419118 0.3198529 0.3419118 0.3198529
 [9] 0.3308824 0.3566176
> err.glm <- mean(err.cv) # Taux d'erreur  0.325
> print(err.glm)
[1] 0.325
```

3.7.3 La courbe Roc

L'observation de la courbe de ROC vient confirmer les bonnes performances de notre modèle.

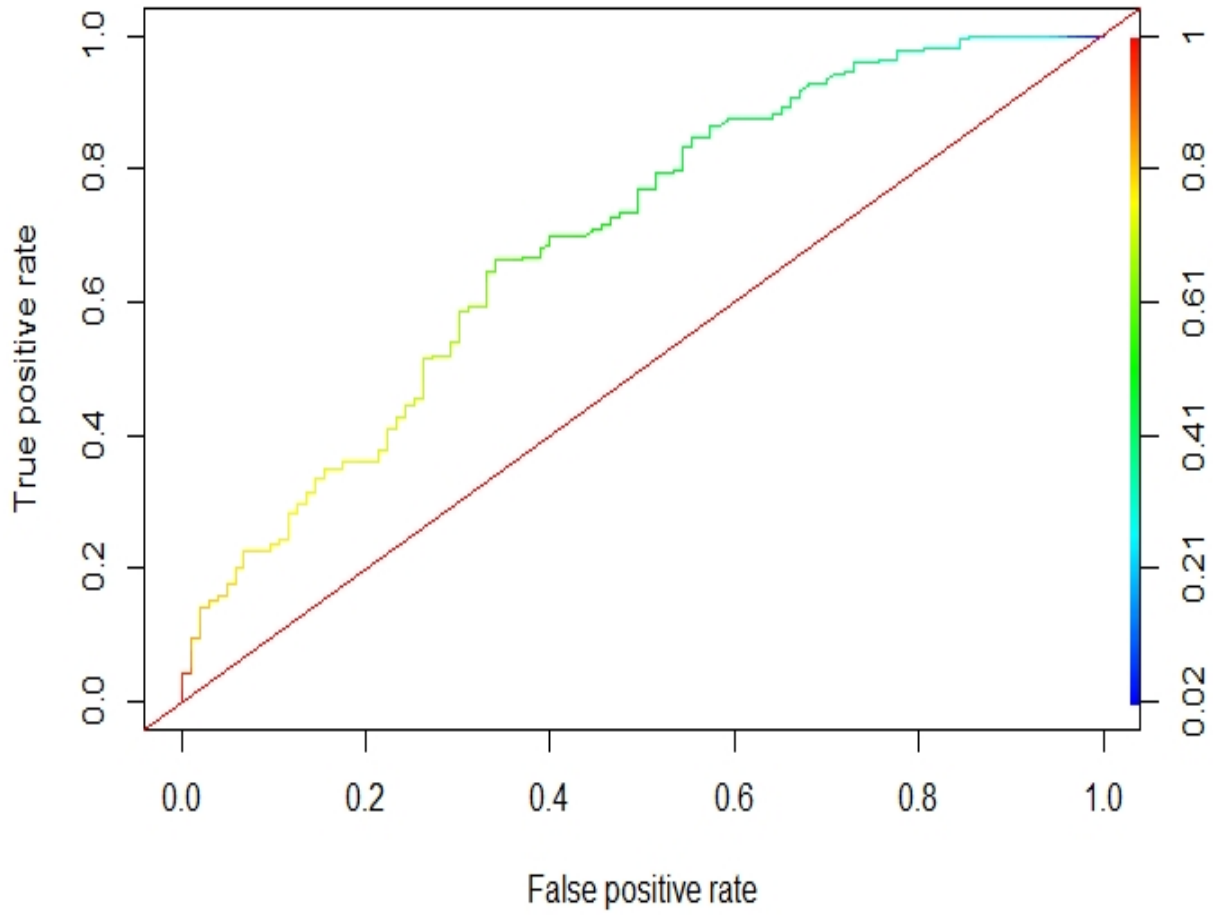

FIGURE 3.4: Courbe ROC

3.8 Conclusion de l'étude

Ce que nous pouvons retenir de cette étude est que les méthodes de datamining sont vraiment intéressantes, performantes si notre modèle est constitué via une base bien construite avec un choix de variables pertinentes. Le modèle retenu est donné comme suit :

1.807 - 1.042 $\times 10^{-6}\times$ **Montant prêt bancaire effectif + 2.349**$\times 10^{-6}\times$
Montant apport personnel prévu + 4.238$\times 10^{-1}\times$ **Situation familiale (Marié)**
+ 7.334$\times 10^{-1}\times$ **Niveau d'instruction (Universitaire) - 1.933**$\times 10^{-1}\times$
Expérience professionnelle (Vrai) - 9.208$\times 10^{-1}\times$ **Client décidé (Vrai)**
- 1.254 $\times$ **Relation entre la formation du client et la nature de projet (Oui)**

Le modèle est d'une capacité de bonne prédiction qui est de 70%. Nous avons 70% de chance de bien prédire si un nouveau client est capable de rembourser son crédit. Ce modèle est une aide a la décision il va permettre à ses utilisateurs d'accorder oui ou non un crédit aux nouveaux clients, il suffit d'introduire les valeurs des variables dans le modèle retenu. Si la valeur obtenue est supérieure à 0.5 on code par 1 (crédit accordé) sinon par 0 (crédit refusé).

Il n'y a pas de meilleur modèle qu'un autre et chaque modèle est adapté à une sorte de données. Un modèle peut être fiable sur un certain échantillon et pas un autre.

Conclusion générale

Généralement, l'objectif du data mining est d'aider à la prise de décision en fournissant des modèles compréhensibles aux utilisateurs. En effet, les utilisateurs ne demandent pas des pages et des pages de chiffres, mais des modèles et leurs interprétations. Les expériences montrent que les modèles simples sont plus compréhensibles mais moins précis, alors que les modelés complexes sont plus précis mais difficiles à interpréter.

A travers ce mémoire nous avons pu répondre à notre problématique qui est de **développer un modèle qui va nous permettre de savoir si un nouveau client peut OUI ou NON rembourser son crédit.**

Nous concluions par donner certains perspectives à ce modeste travail.
Cette étude de datamining pourra séduire les grandes entreprises. Cela me motive d'améliorer mes compétences dans le domaine de datamining, pour développer des modèles fiable à partir de grandes base de données (BIG DATA) on utilisons des logiciels plus performant et des langages de programmation plus adéquat comme Python.

Annexe

```r
#################################################################
#########  Classification par apprentissage supervisée ##########
######### pour Prédire le remboursement ou pas d'un crédit ansej#
#################################################################


##Vider la mémoire R
rm(list =ls())

# Chargement des packages nécessaires

library(MASS)
library(PCAmixdata) # ACP
library(rpart)# arbre de decision
library(class)
library(caret)#CORRELATION
library(corrplot)
library(ROCR) #### la courbe ROC

#####Selectionner le repertoire de travail
setwd("E:/MEMOIRE RAZIK/données")

##### Lecture des données (avec détection des valeurs manquantes)
data <- read.csv( "Echantillon.csv", header = TRUE,
                          dec = ",",sep= ";",na.strings="")



# Dimension du jeu de données
dim(data) # Le jeu de donnée possède 39 variables et 2755 individus

### Statistiques descriptives
head(data) # Affiche les 6 premiers individus des données

str(data) # Affiche la structure du jeu de données


summary(data) # Statistiques descriptives des variables

##### Elimination de la variable dates,adresses...
      ##(inutile pour l'analyse)
data <- data[,-c(1,3,14,18,19,20,21,22,27,28,29,32,34,36,38)]
```

```r
data <- data[,-c(1,3,14,18,19,20,21,22,27,28,29,32,34,36,38)]

####### Préparation des données

### Recodage des classes des données qualitatives enregistrées
######comme quantitatives
data$CIBLE <- as.factor(data$CIBLE) # la variable a prédire

data$Correlation_Proff_Projet <- as.factor(data$Correlation_Proff_Projet)


data$CODE_NIV <- as.factor(data$CODE_NIV)

data$Code_Sit_Fam  <- as.factor(data$Code_Sit_Fam )


# Affichage des types des variables
sapply(data,class)

# Séparation des variables qualitatives et quantitatives
data.quanti <- data[,sapply(data,is.factor)==F]
sapply(data.quanti,class) # Vérification 9

data.quali <- data[,sapply(data,is.factor)==T]
sapply(data.quali,class) #15

# Représentation graphique des données qualitatives
#par des diagrammes en barres
X11()
par(mfrow=c(4,4)) # Partage de l'espace graphique
for (i in 1:15){
  barplot(table(data.quali[,i]),xlab=colnames(data.quali)[i],
          ylab="Fréquence")
}

# Représentation graphique des données quantitatives
##par des boîtes à moustache
X11()
par(mfrow=c(1,1)) # Espace graphique 1*1
boxplot(scale(data.quanti), col = terrain.colors(9),
   names = c("Mont_PretBancaire_Prevu", "Mont_PretBancaire_Effectif",
        "Mont_PNR_Prevu_Ansej", "Mont_PNR_Effectif_Ansej",
```

```r
        "Mont_PNR_Prevu_Ansej", "Mont_PNR_Effectif_Ansej",
        "Mont_Apport_Personnel_Prevu", "Mont_Apport_Personnel_Effectif",
        "Mont_Invest_Prevu","Mont_Invest_Effectif","MONT_ECHEANCE"),
        cex.axis = 0.4)

########Supprimer les valeurs constantes
# Supprimer les variables constantes
data.quali<- data.quali[,-c(3,4,5,6)]

# Regroupement des deux types de variables
data.clean <- cbind.data.frame(data.quanti,data.quali)


# Nombre de valeurs manquantes ?
sum(is.na(data.clean)==T) # 5 419 données manquantes à traiter


##### Traitement des valeurs manquantes
## determiner le nombre de valeurs manquantes
som_manq <- function(x){
  return(sum(is.na(x)==T))
} ##

lapply(data.quali,som_manq)## $NIS [1]  2664;$Complet [1] 1648
lapply(data.quanti,som_manq) ## $MONT_ECHEANCE 1107


#Traitement des valeurs manquantes
traitN <- function(x){
y <- na.omit(x) # Vecteur qui contient les données non manquantes
if (length(y) == length(x)){ # Si pas de données
                             ###manquantes on ne change rien
  return(x)
  } else { # Sinon si il y a des données manquantes alors
    if (is.factor(x) == T){ # Si la variable est qualitative
      dist.freq <- table(y) # Distribution des fréquences
      id.mode <- which.max(dist.freq) # Identification du mode
                                      # de la variable
      z <- x
      z[is.na(z)] <- levels(z)[id.mode] # Imputation par le mode
      return(z)
```

```
      return(z)
    } else { # Sinon si la variable est quantitative
      moyenne <- mean(y) # Calcul des moyennes des variables
      z <- x
      z[is.na(z)] <- moyenne # Imputation par la moyenne
      return(z)
    }
  }
}


# Application de la fonction sur l'ensemble de données (data frame)
traitement.md <- function(X){
  return(as.data.frame(lapply(X,traitN)))
}

# Application de la fonction sur notre jeu de données (data)
data.clean <- traitement.md(data.clean )

# Vérification
sum(is.na(data.clean)==T) # 0 données manquantes


#SUPPRIMER LES VARIABLES AVEC DES VALEURS MANQUANTES
##PLUS QUE DES VALEURS RENSEIGNEES
data.quanti<-data.quanti[,-9]#
data.quali<-data.quali[,-c(1,2)]

##### Elimination des variables quantitatives constantes

##### Elimination des variables quantitatives constantes
ecart_type <- sapply(data.quanti,sd)# pas de variables constantes

################suppression des variables quantitatives corréllées

#### Calcul de la matrice de corrélations
corr.matrix <- cor(data.quanti)

#### Tracé des corrélations
x11()
corrplot(corr.matrix,type = "lower", method = "ellipse")
```

```r
corrplot(corr.matrix,type = "lower", method = "ellipse")
set.seed(7)
corr <- cor(data.quanti);
highlyCorr<- findCorrelation(corr, cutoff=0.75,names = FALSE );
print(highlyCorr)
data.quanti<-data.quanti[,-highlyCorr];

colnames(data.clean[highlyCorr]) #les variables
                        ##quantitatives corréllées



# Vérification

# Regroupement des deux types de variables

data.clean <- cbind.data.frame(data.quanti,data.quali)


##### Réalisation d'une ACP sur données mixtes
acp <- PCAmix(X.quanti = data.quanti, X.quali = data.quali,
              ndim = 5, rename.level = T, graph=TRUE)
x11()
plot(acp,axes=c(1,2), choice = "ind", coloring.ind = data.quali$CIBLE)



##suppression des individus 167,124,1864,1748,78,88,140,45,1884,116,
################################1909,1676,1908.......
## supprimer certains membres

data.quali <- data.quali[-c(167,124,1864,1748,78,88,140,45,1884,
                            116,1909,1676, 1908,1845,1938,1889,1771
                            ,1821,919,659,1084,1283,604,1787
                            ,223,2094,1531,862,1614,2108,2371,
                            2518,2442,2472,2309),]
data.quanti <- data.quanti[-c(167,124,1864,1748,78,88,140,45,
                              1884,116,1909,1676,1908,1845,1938,1889,
                              1771,1821,919,659,1084
                              ,1283,604,1787,223,2094,1531,862,
                              1614,2108,2371,
                              2518,2442,2472,2309),]
```

81:54 (Untitled)

```
                                    2518,2442,2472,2309),]
# Regroupement des deux types de variables
data.clean <- cbind.data.frame(data.quanti,data.quali)

### Etude de la variable cible (CIBLE)
freq <- table(data.clean$CIBLE)
print(freq/sum(freq)) #  40% % mauvais dossier


#####################################################
### Régression logistique########"
##########selon les methodes de rééchantillonnages###
#####################################################

#########methodes de rééchantillonnages##########
#################################################
# Echantillon Dtrain et Dtest
# 75% des points pour Ntrain et 25% pour Ntest
n=2720
p=0.75;
ntrain=floor(n*p);
ntest=n-ntrain;
ntrain;
ntest
ind=sample(n);# Mélanger le jeu de données aléatoirement
length(ind);
head(ind)
ind_train=ind[1:ntrain];
length(ind_train);
head(ind_train)
ind_test=ind[(ntrain+1):n];
length(ind_test);head(ind_test)
Dtrain=data.clean[ind_train,];
dim(Dtrain);
head(Dtrain)
Dtest=data.clean[ind_test,];
dim(Dtest);

```

```r

######## Construction et prédiction ########
fit.glm1 <- glm(CIBLE ~ .,data=Dtrain,family=binomial)
summary(fit.glm1) # la formule de modele
pred.fit.glm1 <- predict(fit.glm1,newdata=Dtest, type="response")
class.fit.glm1<- as.numeric(pred.fit.glm1>=0.5)# permet d'affecter
############################  0 ou 1 à la place  d'une probabilité
############################afin de comparer avec CIBLE
mc.glm1 <- table(Dtest$CIBLE,class.fit.glm1) # Matrice de confusion
print(mc.glm1) # Matrice de confusion

# Taux d'erreur
err.glm1<- 1-sum(diag(mc.glm1))/sum(mc.glm1)
print(err.glm1) # 0.29 55882



####################################
#### VALIDATION CROISEE - glm  ####
####################################
##processus de la validation croisée!!!
# Subdivision aléatoire en K = 10 partitions
K <- 10
n <- nrow(data.clean)
# Génération de nombre aléatoire
alea <- runif(n)
# Découpage en K portions de fréquence égales
blocs <- cut(alea,breaks=quantile(alea,probs=seq(0,1,1/K)),
                                        include.lowest=T)
levels(blocs) <- 1:K
# Nombre d'observations par blocs
print(table(blocs))

err.cv <- numeric(K)

for (n in 1:K){
  data.app <- data.clean[(blocs!=n),]
  data.test <- data.clean[blocs==n,]
  # Apprentissage
  fit.glm <- glm(CIBLE ~ .,data=data.app,family =binomial)
```

```r
  # Apprentissage
  fit.glm <- glm(CIBLE ~ .,data=data.app,family =binomial)
  # Test
  pred.fit <- predict(fit.glm,newdata=data.test, type="response")
  pred <- as.numeric(pred.fit>=0.5) # permet d'affecter 0 ou 1
  ############################ à la place  d'une probabilité afin
  ############################de comparer avec CIBLE
  # Matrice de confusion
  mc <- table(data.test$CIBLE,pred)
  print(mc)
  # Taux d'erreur
  err <- 1-sum(diag(mc))/sum(mc)
  # Récupération de la valeur
  err.cv[n] <- err
}

# Les erreurs récupérées :
print(err.cv)
err.glm <- mean(err.cv) # Taux d'erreur  0.325
print(err.glm)
summary(fit.glm) ### la formule d modele



#### la courbe ROC
roclogit=predict(fit.glm1, newdata=data.test,type="response")
predlogistic=prediction(roclogit,data.test$CIBLE)
perflogistic=performance(predlogistic, "tpr","fpr")
X11()
plot(perflogistic,col=1, colorize=TRUE)
abline(a=0.0,b=1.0,col="red")

perf <- performance(predlogistic, "auc")
perf@y.values[[1]]
####
###  on obtient un modele de 70 % de bonnes prédictions globales
```

Bibliographie

[1] Biernat. E and Lutz. M. *Data science : fondamentaux et etudes de ca - Machine learning avec python et R.*

[2] Besse. P. *Exploration Statistique Multidimensionnelle Data Mining*, Équipe de Statistique et Probabilités, Institut de Mathématiques de Toulouse — UMR CNRS C5219 Département Génie Mathématique et Modélisation, ,Institut National des Sciences Appliquées de Toulouse — 31077 – Toulouse cedex 4.

[3] Bahouayila. B. *Cours de traitement des données*, https ://hal.archives-ouvertes.fr/cel-01317637, 2016.

[4] Chesneau. C. *Eléments de classification*,Université de Caen.

[5] Cours *de Classification automatique de données par la méthode des centres mobiles.*, Mathématiques pour la Biologie,Université de Nice, Département de Mathématiques, 2010-2011.

[6] Forgy. E. *Cluster analysis of multivariate data : Efficiency vs. interpretability of classifications. Biometrics, page 21 :768,* 1965.

[7] Hosmer .D.W and Lemeshow. S, *Applied Logistic Regression , Wiley,* 2003.

[8] Husson. F. *Classification ascendante hiérarchique (CAH),Laboratoire de mathématiques appliquées - Agrocampus Rennes.*

[9] Hamerly. G and Elkan. C. *Learning the k in k-means. In NIPS,* 2003.

[10] Han. J. *Data mining : concepts and techniques* / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed. ISBN 978-0-12-381479-1, 2009.

[11] Hamilton. HJ. *http ://www2.cs.uregina.ca/∼ dbd/cs*831/*notes*/*ROC*/*ROC.html*

[12] Kohavi. R. and Provost. F. *(1998) Glossary of terms. Machine Learning—Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. Machine Learning, 30, 271-274.*

[13] Lebart. L, Piron M. and Morineau. A. *Statistique exploratoire multidimensionnelle*2000.

[14] Likas. A, Vlassis. N and Verbeek. JJ.*The global k-means clustering algorithm. Pattern Recognition, 36(2) :451-461,* 2003.

[15] Lemberger .P, Batty. M, Morel. M and Raffaëlli. JL. *Big Data et Machine Learning*,2015

[16] McQueen. J. *Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* 1 :281-297, 1967.

[17] Pelleg. D and Moore. AW. *X-means : Extending K-means with efficient estimation of the number of clusters. In Proc. 17th International Conf. on Machine Learning, pages 727-734. Morgan Kaufmann, San Francisco, CA,* 2000.

[18] Rakotomalala. R . *Méthode des centres mobiles - classification par partition-Les méthode de réallocation,Université Lumière Lyon2.*

[19] Tufféry. S .*Data Mining et statistique décisionnelle - L'intelligence des données, Quatrième édition ,*2012.

[20] Tufféry. S .*Techniques prédictives de data mining,*2009.

[21] Tufféry. S .*Techniques descriptives de data mining,*2009.

[22] Zhang. R and Rudnicky. AL. *A large scale clustering scheme for kernel k-means. In ICPR (4), pages 289-292,* 2002.

[23] Zhang. B and Hsu. M. *K-harmonic means - a data clustering algorithm,* 1999.

[24] Zighed. D.A and Rakotomalala. R. *Data Mining,2002.*