2023-02

Mise en place d’un système automatise d’analyse et de traitement des données d’une entreprise

Kenguruka, Eliphase

UB

République du Burundi

Ministère de l'Education Nationale et de la Recherche Scientifique

Université du Burundi

Faculté des Sciences de l'Ingénieur

Master en Génie Informatique

Année Académique :

2021-2022

MISE EN PLACE D'UN SYSTEME AUTOMATISE D'ANALYSE ET DE TRAITEMENT DES DONNEES D'UNE ENTREPRISE

MEMOIRE

Présenté par

KENGURUKA Eliphase

à la

FACULTE DES SCIENCES DE L'INGENIEUR (FSI)

En vue de l'obtention du grade de

MASTER

en

Génie Informatique

Soutenu le 28 /02/2023, devant le jury composé de :

Dr. SAHINGUVU	**William**	Président
Pr. NDIKUMAGENGE	**Jérémie**	Vice-Président
Dr. MUKESHIMANA	**Michele**	Secrétaire
Dr. NDAYISABA	**Longin**	Directeur
Dr. NIBITANGA	**Roméo**	Membre

IDENTIFICATION DES MEMBRES DU JURY

Dr. SAHINGUVU	**William**	: Président
Pr. NDIKUMAGENGE	**Jérémie**	: Vice-Président
Dr. MUKESHIMANA	**Michele**	: Secrétaire
Dr. NDAYISABA	**Longin**	: Directeur
Dr. NIBITANGA	**Roméo**	: Membre

DEDICACES

A Dieu tout puissant ;

A mon regret Père qui nous a quittés sitôt ;

A ma très chère maman ;

A mes frères et sœurs ;

A mes neveux ;

A mes Cousins et Cousines ;

A tous mes camarades de classe ;

A tous mes amis et connaissances.

REMERCIEMENTS

Au terme de ce travail, je tiens tout d'abord à remercier Dieu le tout puissant et miséricordieux qui m'a donné la force et la patience durant ces longues années d'étude. Je tiens à exprimer ma profonde gratitude et mes sincères remerciements à mon directeur de mémoire Dr Longin NDAYISABA pour son soutien, sa patience, ses précieux conseils, son aide, sa disponibilité tout au long de mes études et sans qui ce mémoire n'aurait jamais vu le jour. Qu'il trouve dans ce travail un hommage vivant à son grand dévouement et à sa haute personnalité. Je désire remercier sincèrement mes chers parents, frères et sœurs, pour le soutien constant tant matériel et morale qu'ils m'ont témoigné jusqu'à l'aboutissement de ce travail. Je tiens tout particulièrement à remercier les enseignants du département d'informatique pour leur disponibilité et encouragement, ainsi que tous les enseignants qui ont contribué à notre formation. Ma reconnaissance va aussi aux membres de jury, pour l'honneur qu'ils auront fait en acceptant de juger ce travail.

Je remercie, enfin tous ceux qui, d'une manière ou d'une autre, ont contribué à la réussite de ce travail et qui n'ont pas pu être cités ici.

RESUME

Les entreprises d'aujourd'hui sont appelées à développer leurs compétences. Celles-ci représentent la source d'un avantage concurrentiel déterminant pour la continuité du fonctionnement des organisations. Avec l'internalisation des marchés, l'entreprise doit s'adapter, si possible anticiper, parfois influencer, en tout cas réagir avec agilité. Pour y parvenir dans de bonnes conditions, les gestionnaires d'entreprises ont besoin de l'information appropriée, au moment opportun, pour la prise des décisions. La place centrale qu'occupe l'information dans le processus de prise de décision n'est plus à démontrer. Ainsi, le traitement de l'information décisionnelle constitue l'élément moteur du processus de création de leurs avantages. Dans un contexte de compétitivité économique fondée sur l'usage de l'information, les entreprises et administrations sont de plus en plus fréquemment amenées à entreprendre une démarche stratégique de traitement automatique de l'information. Par contre, beaucoup des entreprises jusqu'aujourd'hui utilisent encore des outils ou systèmes archaïques tels que les outils calculatoires ou les logiciels non spécifiques au traitement de l'information tel que Microsoft Office Excel et ainsi que d'autres du genre. Ceci devient une source de ralentissement dans la maîtrise de gestion et dans le développement.

Dans ce projet, nous allons voir comment procèdent ces algorithmes de traitement de données de la discipline Data Mining. Le premier chapitre est consacré à la présentation du projet de recherche. Nous y présentons premièrement la contextualisation et actualités du projet de recherche, l'objectif global, les objectif spécifiques, le problématique lié à notre travail, les solutions proposées, les résultats attendus ainsi que les apports scientifiques et technologique seront également abordées dans ce chapitre. Dans les chapitres suivants, on va parler sur l'importance des données dans une entreprise ainsi que les techniques de leurs explorations dans le but de prendre une décision.

Mots-clés : **système**, **automatisé, analyse**, **traitement**, **données**, **exploration**, **fouille**, **Data Mining**, **modélisation**, **optimisation**, **prédiction**, **décision**.

ABSTRACT

Companies today are called upon to develop their skills. These represent the source of a decisive competitive advantage for the continuity of the functioning of organizations. With the internalization of markets, the company must adapt, if possible, anticipate, sometimes influence, in any case react with agility. To achieve this under the right conditions, business managers need the right information, at the right time, for decision-making. The central place occupied by information in the decision-making process no longer needs to be demonstrated. Thus, the processing of decisional information is the driving force in the process of creating their advantages. In a context of economic competitiveness based on the use of information, companies and administrations are more and more frequently called upon to undertake a strategic approach to the automatic processing of information. On the other hand, many companies until today still use archaic tools or systems such as calculators or non-specific information processing software such as Microsoft Office Excel and others of the genre. This becomes a source of slowdown in management control and in development.

In this project, we will see how these data processing algorithms of the Data Mining discipline proceed. The first chapter is devoted to the presentation of the research project. We first present the context and news of the research project, the global objective, the specific objectives, the problematic related to our work, the proposed solutions, the expected results as well as the scientific and technological contributions will also be discussed in this chapter. In the following chapters, we will talk about the importance of data in a company as well as the techniques of their explorations in order to make a decision.

Keywords: **system, automated, analysis, processing, data, exploration, mining, data mining, modeling, optimization, prediction, decision**.

TABLES DES MATIERES

LISTE DES FIGURES :

LISTE DES TABLEAUX

SIGLES ET ABREVIATIONS :

DAF : Direction Administrative et Financière

ECD : Extraction des Connaissances à partir des données

ETL : Extract Transform Load

INSEE : Institut National de la Statistique et des Études Économiques

MERISE : Méthode d'Etude et de Réalisation Informatique pour les Systèmes d'Entreprise

OMT : Object Modeling Technique

PC : Composantes Principales

PL : Programmation Linéaire

PWC : Price Waterhouse Coopers

ROI : Return On Investment

UML : Unified Modeling Language

UP : Unified Process

AVANT-PROPOS

Ce mémoire s'inscrit dans le cadre d'un projet de fin d'études du deuxième cycle universitaire afin d'obtenir un diplôme de Master en Génie Informatique. Il se focalise sur automatisation du processus de recherche d'une solution optimale à un problème linéaire et du traitement des données afin de prendre des décisions par la « mise en place d'un système automatisé d'analyse et de traitement des données d'une entreprise ». Un problème majeur se remarque dans le fait que beaucoup des entreprises jusqu'aujourd'hui utilisent encore des outils ou systèmes archaïques tels que les outils calculatoires ou les logiciels non spécifiques au traitement de l'information tel que Microsoft Office Excel et ainsi que d'autres du genre. Ceci devient une source de ralentissement dans la maîtrise de gestion et dans le développement.

En effet, nous avons commencé par identifier la problématique liée au traitement manuel ou au traitement en utilisant des logiciels non spécifiques précisément au traitement des données dans le domaine entrepreneurial. Par après, nous avons proposé des solutions aboutissant à la mise en place d'un système automatisé d'analyse et de traitement des données d'une entreprise implémenté avec le langage de programmation python, CSS et le gestionnaire de base de données Microsoft Office Excel.

Au cours de la réalisation de ce projet, la plus grande difficulté rencontrée est liée à l'accès aux données de l'entreprise dans laquelle nous avons choisi pour nos recherches. Quoi que les entreprises commerciales burundaises mettent en cachette leurs données plus utiles pour les projets de recherche, nous avons pu travailler avec les données facilement accessibles.

CHAP I : INTRODUCTION GENERALE

I.1. Introduction

Depuis l'antiquité, les hommes ont essayé de créer des instruments pour calculer, traiter ou stocker l'information. Les technologies de Data Mining permettent, grâce aux processus d'intelligence artificielle, de traiter des masses gigantesques de données afin d'en extraire l'information cruciale qui sera déterminante pour une prise de décision efficace. Une connaissance est définie par un ensemble de relations (règles, phénomènes, exceptions, tendances…) entre les données. Le Data Mining est apparu au début des années 1990. Cette émergence est le résultat de la combinaison de plusieurs facteurs à la fois technologiques, économiques et même sociopolitiques. Les volumes gigantesques de données constituent, dès lors, des mines d'informations stratégiques aussi bien pour les décideurs que pour les utilisateurs. Le Data Mining est l'un des maillons de la chaîne de traitement pour la découverte de connaissances à partir de données. Sous forme imagée, on pourrait dire que l'ECD est un véhicule dont le Data Mining est le moteur.

Au Burundi, beaucoup d'institutions connaissent un retard notoire dans la maîtrise et l'usage des technologies de l'information intéressantes pour le développement du pays en général et de l'entreprise en particulier. Dans le contexte de production des produits des entreprises, aucun organisme ne pourrait ignorer le rôle et la place de la prise d'une bonne décision dans la maîtrise de la concurrence et l'amélioration des services. Cela est vrai aussi dans le marketing de leurs produits pour augmenter la production suivant la demande des clients. L'usage des techniques d'exploration de données au quotidien apporte une valeur grouillée de profits pour l'entreprise. Cet usage s'avère être un moyen de rentabiliser ses ressources, augmenter ses performances et atteindre les objectifs fixés dans les délais prévus. Ce retard est devenu un handicap pour le développement socio-économique du pays en général et des entreprises en particulier.

I.2. Contextualisation

Depuis quelques années la révolution numérique, les objets connectés, les applications mobiles, les smartphones, Internet, les réseaux sociaux, et autres sources multiplient les émissions de données, sous diverses formes et formats, suscitant des convoitises en matière d'information nouvelle. En effet une fois exploitées, ces données pourraient délivrer des informations auxquelles personne n'a encore pensé. Ces données forment désormais un capital au sein de ces organisations, qui les positionnent comme la pierre angulaire de leur projet de transformation numérique. La conservation de ces données devient donc une priorité pour ces organisations, qui constituent ainsi un patrimoine de données. Dans un monde en constante évolution, où les données sont de plus en plus nombreuses, la nécessité de les regrouper s'est imposée d'elle-même. Les organisations ont donc comme défi de créer une architecture d'entreprise moderne pour organiser, gérer, exploiter ces larges volumes de données de manière opérationnelle. Le système d'information de ces organisations, et donc son architecture des données doit alors évoluer pour s'adapter à ces nouvelles attentes.

A partir de l'année 2014, l'essor de certaines technologies, telles que l'apparition des systèmes de stockage arborescents comme Apache Hadoop 1 font émerger un nouveau concept pour tirer parti de ce capital de données : les « lacs de données ». D'abord assimilés simplement à un nouveau moyen de stocker des données, puis associés à un phénomène marketing, les lacs de données créent un engouement très fort dans le monde industriel, qui les adoptent de façon massive. C'est donc sous l'impulsion, et la vision, très commerciale, du monde industriel que ce nouveau concept des lacs de données se positionne désormais comme incontournable dans le système d'information. Aujourd'hui, avec une grande taille de ces lacs de données déjà stockées dans les bases de données, il est difficile de les parcourir pour les analyser et les traiter manuellement ou avec des logiciels génériques de traitement. Des méthodes ou fonctions mathématiques ont été mises en place comme remède à ce problème. Le Data Mining dont nous allons nous focaliser sur les techniques dans le chapitre quatre est devenu une discipline plus connue et plus utilisée dans ce domaine d'analyse et de traitement de données.

I.3. Objectifs du travail

I.3.1. Objectif global du travail

L'objectif global de ce projet est de mettre en place un système automatisé d'analyse et de traitement des données qui facilite aux entreprises, usines, coopératives ou organisations d'analyser des données de production que ça soit les biens ou les services et à leurs représentants (Directeurs généraux, gestionnaires, ...) de bien prendre des décisions sûres et fiables, rapidement et facilement.

I.3.2. Objectifs spécifiques

Les principaux objectifs spécifiques sont :

1) Modéliser un problème sous forme mathématique
2) Donner une solution optimale à un ce problème par des méthodes mathématiques
3) Concevoir une application d'exploration et de représentation des données

I.4. Problématique

Étant donné que dans de beaucoup des entreprises ou organisations l'analyse et le traitement des données se font rarement en Microsoft Excel et fréquemment de manière quasi manuelle, des problèmes suivants se présentent :

1. Difficulté de parcourir toutes les données facilement et dans le délai,
2. Difficulté de détection des fraudes à l'intérieur de l'entreprise,
3. Difficulté liée à la prise de décision facilement et rapidement,
4. Difficulté d'étude des marches selon la situation de l'entreprise et le comportement des clients,
5. Service clientèle non satisfait

Tous ces problèmes ci-haut cités ont amenés à se demander si la conception et la réalisation d'une application de fouille de données ne serait-il pas un moyen de déraciner ces problèmes.

I.5. Solutions proposées

1. L'utilisation des appareils informatiques (ordinateurs, imprimantes, …) afin de minimiser le temps de raisonnement et le nombre d'erreurs dans de certaines activités.
2. La mise en place d'un serveur de base de données muni d'un système de gestion de base de données accessibles par les dirigeants,
3. La mise en place d'un système automatisé d'analyse et de traitement des données basé sur les techniques d'exploration des données.

I.6. Résultats attendus

1. Stockage et la gestion de données dans un système de base de données multidimensionnel
2. Analyser et traiter les données grâce à un logiciel conçus sur les techniques d'exploration des données.
3. Solution ETL
4. Présentation des données dans un format simplifié (tableau, graphique…)
5. Prise de décision facile et rapide et sûres

Les entreprises voient arriver des données dans de multiples formats à une vitesse et dans des volumes sans précédent. Le succès de toute structure dépend désormais de sa rapidité à exploiter les connaissances issues du lac de données et à les intégrer dans le processus décisionnel et métier afin d'identifier et conduire des actions pertinentes au sein de l'organisation. Le data mining aide les entreprises à optimiser leur avenir. Il leur permet de comprendre le passé et le présent et de faire des prédictions précises sur ce qui est susceptible d'arriver. Le data mining peut être utilisé pour répondre à de nombreux objectifs business et commerciaux comme : Augmentation de ses revenus, fidélisation des clients et augmentation du taux de rétention (fidélité), augmentation du retour sur l'investissement des campagnes marketing, amélioration de la clientèle, suivie des performances opérationnelles. Grâce au data mining, les décisions sont basées sur des véritables affaires intelligentes, plutôt que sur des intuitions ou instincts. Cela permet d'obtenir des résultats cohérents et de prendre ou conserver une avance sur votre concurrence.

I.7. Apports scientifiques et technologiques

La mise en place d'un système automatisé de réduction du lac de données en données réduites va faciliter aux gestionnaires et auditeurs des entreprises ou organisations à la rédaction facile et rapide des rapports, la conception et réalisation d'un système automatisé d'aide à la prise de décision sera un atout aux décideurs de bien décider pour un meilleur fonctionnement de leurs entreprises ou organisations.

I.8. Domaines d'application

Cette application sera utilisée principalement dans le domaine commercial, dans le domaine cadastral, dans la prise des décisions par l'administration et dans l'analyse de la situation démographique.

I.9. Méthodologie et outils mathématiques utilisés

Pour mener à bien notre travail de recherche, nous avons fait recours aux modèles du cycle de développement du logiciel et mathématiquement on a dû recourir aux techniques d'exploration de données. Le processus d'exploration de données suit six étapes principales : compréhension commerciale, compréhension des données, préparation des données, modélisation, évaluation et déploiement. La phase de compréhension commerciale consiste à comprendre de manière approfondie les paramètres et le cadre du projet puis à définir les facteurs principaux de réussite. La compréhension des données consiste à cibler les informations nécessaires qui permettront de répondre à l'objectif défini. Ensuite, il faut dresser la liste des ressources qui comportent les données utiles et regrouper ces dernières. La préparation des données consiste à préparer les données dans le format adéquat afin de répondre à la finalité puis vérifier leur qualité et corriger les problèmes de manque ou de duplication. La modélisation est l'usage d'algorithmes pour l'identification de modèles. Pour évaluer, on détermine si les résultats obtenus par un pattern permettront d'atteindre l'objectif commercial final et dans quelle mesure cela est possible. Généralement, il y a une phase itérative pour dénicher le meilleur algorithme et par conséquent le meilleur résultat. Le déploiement consiste à remettre les résultats de l'analyse aux décideurs et à se servir des informations finales pour adapter la stratégie.

I.10. Outils de développement

Dans notre projet, nous avons recouru aux outils de développement suivants :

1. Environnement de développement : Sublime Text
2. Langage de programmation : PYTHON3
3. Librairie pandas de python pour importer les données du format Excel
4. Librairie NUMPY de python pour des fonctions mathématiques
5. Librairie MATPLOTLIB lui aussi de python pour le traitement graphique
6. Une base de données sous format Excel

CHAP II : PRÉSENTATION ET FONCTIONNEMENT D'UNE INSTITUTION CIBLE

II.1. Introduction

Le monde de l'entreprise est au cœur de l'actualité et du débat public. Pourtant, nous avons souvent du mal à en prendre la mesure, voire à le définir.

D'après l'INSEE, l'entreprise est une « unité économique, juridiquement autonome dont la fonction principale est de produire des biens ou des services pour le marché ».
Autrement dit, il y a entreprise dès que des personnes mobilisent leur talent et leur énergie, rassemblent des moyens matériels et de l'argent pour apporter un produit ou un service à des clients. Les entreprises sont au cœur de nos vies, il est donc essentiel de mieux les connaître. Les entreprises rythment la vie économique et sociale et animent notre quotidien. En tant que consommateurs, nous nous appuyons à chaque pas sur ces organisations qui nous nourrissent, nous vêtissent, nous transportent, nous divertissent, nous maintiennent en bonne santé, nous fournissent les moyens de communication, les équipements et l'énergie dont nous avons besoin. En tant qu'employés ou entrepreneurs, nous trouvons dans l'entreprise l'un de nos principaux champs d'expression. Nous y investissons une grande part de notre temps, de notre énergie et de notre créativité. Nous y développons nos compétences et y affirmons notre personnalité.

II.2. Organisation et fonctionnement d'une entreprise

Toutes les activités de la firme sont réalisées au sein de différents services. Pourtant, la diversité de leurs missions et des individus qui les exécutent rend nécessaire leur coordination. Il faut donc construire une organisation pour l'entreprise. Ainsi, selon Mintzberg (1982), la structure correspond à « la somme totale des moyens employés pour diviser le travail en tâches distinctes et pour ensuite assurer la coordination nécessaire entre ces tâches ». Elle désigne donc l'architecture générale de l'entreprise, son ossature. Elle assure l'agencement et l'articulation entre les différents services et oriente les comportements des individus. Sa représentation graphique la plus courante est l'organigramme. Celui-ci permet de mettre en évidence la séparation technique des domaines de compétences (division du travail en tâches distincts), les relations entre les composants de la structure et le positionnement des membres sur l'échelle hiérarchique. L'organigramme, seul, ne permet toutefois pas de saisir la complexité des relations entre les services et les individus. En effet, il ne représente pas les liens informels que nouent les acteurs entre eux et qui jouent un rôle essentiel dans le fonctionnement organisationnel. La structure réelle est ainsi une combinaison entre un

design formel et des relations informels. La conception de la structure est donc un exercice délicat. Son influence sur la réalisation des objectifs économiques, sociaux et sociétaux et la mise en œuvre de la stratégie en font une mission relevant de la direction générale. [17]

Le mode d'organisation et de fonctionnement des entreprises reposent sur certaines caractéristiques communes. Il est influencé par la stratégie, le métier, la taille, la maturité, l'histoire et la culture de l'organisation. L'activité d'une même entreprise est répartie au sein de diverses fonctions. Toute entreprise est organisée autour de deux pôles d'activité opérationnelle : la production et le commerce. La production regroupe l'ensemble des fonctions qui produisent les biens et les services que l'entreprise commercialise. Le commerce regroupe l'ensemble des fonctions qui commercialisent les biens et les services que l'entreprise produit.

II.3. Principales fonctions d'une entreprise

Selon la logique fonctionnelle, qui complète l'approche économique de l'entreprise, celle-ci est un organe autonome doté de plusieurs fonctions, à la fois différentes et interdépendantes. De la qualité de ses fonctions et de leur synergie, dépendent la réalisation des objectifs économiques et commerciaux de l'entrepris. La métaphore (comparaison) biologique est très utile pour comprendre l'approche fonctionnelle de l'entreprise. L'entreprise ressemble au corps humain : celui-ci est doté d'un ensemble d'organes ayant chacun une mission spécifique (les fonctions de l'entreprise). Le fonctionnement optimal du corps humain nécessite la mise en commun et l'interaction de tous les organes, c'est-à-dire, la nécessaire coordination entre toutes ces fonctions. En l'absence d'interaction dynamique entre tous ces organes, le corps s'essouffle et l'organisation est en difficulté. Enfin, il y a lieu de remarquer que tous les organes poursuivent un objectif commun : celui de la survie de l'individu (performance et survie de l'entreprise). Le degré de « centralité » d'une fonction dépend des biens ou services produits par l'entreprise, c'est-à-dire, que certaines fonctions auront de l'importance alors que d'autres seront jugées comme complémentaires. En effet, il n'existe pas de modèle général de fonctions à appliquer quelle que soit l'entreprise, mais nous retrouverons certaines fonctions qui seront la plupart du temps présentes dans l'organisation. [18]

Tableau 1 : Fonctions principales *d'une entreprise*

Fonction	**Services**	**Attribution**
Direction	État-major, secrétariat général, corps d'inspection, Services généraux	Études, projets, stratégie, organisation, contrôle
Financement	Service de trésorerie, service comptable, …	Etude, stratégie, organisation et contrôle
Approvisionnement	Achat, gestion de stocks et magasins	Politique d'approvisionnement et gestion de stock
Production	Etude, méthode d'ordonnancement, fabrication et contrôle de qualité	Préparation technique du travail
Commercialisation	Vente et marketing	Estimation des besoins
Ressources humaines	Embauches, formation et relations sociales	Recrutement du personnel, gestion du personnel
Recherches & Développement	Innovation, gestion de la concurrence, adaptation du marché	Création des nouveaux produits, s'adapter aux besoins

II.4. Caractéristiques d'une entreprise

L'entreprise est une entité créée dans le but de faire des affaires, c'est-à-dire de générer des profits en vendant des biens ou des services. Pour ce faire, elle doit posséder certaines caractéristiques qui lui permettent de fonctionner de manière optimale et d'atteindre ses objectifs. Parmi ces caractéristiques, on peut citer la vision, la mission, les valeurs, les objectifs, la stratégie, le business model, les ressources, les compétences et l'expertise. La vision est la raison d'être de l'entreprise, c'est-à-dire ce qu'elle souhaite accomplir à long terme. Elle doit être claire, concise et motivante pour tous les membres de l'entreprise. La mission définit les activités principales de l'entreprise et son objectif principal, c'est-à-dire ce qu'elle souhaite réaliser à court et moyen terme. Les valeurs sont les principes fondamentaux qui guident les actions de l'entreprise et ses relations avec ses parties prenantes. Elles doivent être en accord avec la vision et la mission de l'entreprise. Les objectifs définissent ce que l'entreprise souhaite atteindre dans un avenir prévisible et mesurable. Ils doivent être SMART, c'est-à-dire spécifiques, mesurables, atteignables,

réalistes et temporellement définis. La stratégie est le plan d'action mis en place par l'entreprise pour atteindre ses objectifs. Elle doit être adaptée au contexte dans lequel l'entreprise évolue et aux ressources dont elle dispose. Le business model décrit la manière dont l'entreprise génère des revenus en proposant ses biens ou services sur le marché. Il est important qu'il soit adapté au contexte économique et aux besoins du marché. Les ressources sont les moyens dont dispose l'entreprise pour mettre en œuvre sa stratégie. Elles peuvent être financières, humaines, matérielles ou informationnelles. Les compétences sont les savoir-faire et savoir-être nécessaires pour mettre en œuvre la stratégie de l'entreprise. Elles peuvent être acquises par le biais de l'expérience ou de la formation. L'expertise est le savoir spécialisé nécessaire pour mener à bien les activités de l'entreprise. Elle peut être acquise par le biais de la recherche ou de l'expérience professionnelle. [19]

La division des tâches au sein d'une structure organisationnelle suppose que celles-ci sont ensuite reliées par un ensemble de liens qui peuvent être :

1. Des liens hiérarchiques : qui impliquent alors la définition de liens de subordination entre les différents éléments.
2. Des liens fonctionnels : les décisions d'un élément de la structure doivent pouvoir s'appliquer aux autres éléments dépendant de ce centre de compétence.
3. Des liens de conseil : un élément de la structure peut contribuer au bon fonctionnement d'un autre élément.

II.5. Structures d'une entreprise

Toutes les activités de la firme, qu'elles soient liées aux fonctions principales ou de soutien, sont réalisées au sein de différents services. La diversité de leurs missions et des individus qui les exécutent rend nécessaire la définition d'une organisation pour l'entreprise - ce que l'on nomme la structure. Selon Mintzberg (1982), la structure correspond à « la somme totale des moyens employés pour diviser le travail en tâches distinctes et pour ensuite assurer la coordination nécessaire entre ces tâches ». Elle désigne donc l'architecture générale de l'entreprise, son ossature. Elle assure l'agencement et l'articulation entre les différents services et oriente les comportements des individus. Sa représentation graphique la plus courante est l'organigramme. Celui-ci met en évidence la séparation technique des domaines de compétences (division du travail en tâches distinctes), les responsabilités respectives, les relations entre les composants de la structure et le positionnement des membres sur l'échelle hiérarchique. Bien qu'indispensable, l'organigramme

seul ne permet pas de saisir la complexité des relations entre les services et les individus. En effet, il ne représente pas les liens informels que nouent les acteurs entre eux et qui jouent un rôle essentiel dans le fonctionnement organisationnel. La structure réelle est finalement une combinaison entre un design formel et des relations informelles. [21]

Les entreprises peuvent adopter différents types de structures selon la manière dont elles organisent la division interne du travail (degré de départementalisation).

On distingue généralement deux grands types de structures qui se distinguent par le fait que l'une est centrée sur la notion de fonction alors que l'autre repose sur l'idée de produit.[17]

II.6. Structure fonctionnelle d'une entreprise

La structure fonctionnelle organise l'entreprise par fonctions. Elle est aussi appelée structure en U (Unitary). C'est la forme d'organisation la plus répandue pour les entreprises mono-activité de taille moyenne. Elle connaît 3 stades de développement : la structure en soleil, la structure fonctionnelle simple et la structure fonctionnelle évoluée (staff and line). La DG (Direction Générale) assure la cohérence des actions menées. La spécialisation des différentes fonctions correspond à un partage des tâches et des responsabilités. La structure fonctionnelle permet à l'entreprise mono-activité de se développer en améliorant sa compétitivité dans un environnement stable. La division du travail par fonctions (opérationnelles et supports) permet une spécialisation qui favorise l'efficacité de l'entreprise. La structure fonctionnelle est la première forme d'organisation d'entreprise. Elle s'inspire du modèle fonctionnel de Frederick Taylor (début du XXe siècle) basé sur la division verticale et horizontale du travail, et du modèle hiérarchique d'Henri Fayol, le pionnier français du « management ».[22]

La structure de l'entreprise repose sur les différentes fonctions exercées au sein de l'organisation (fonctions de production, commerciale, financière, de gestion des ressources humaines…). Ce type de structure repose sur deux principes essentiels :

1. Unité de commandement : la voie hiérarchique constituée se traduit par le fait que tout membre de l'entreprise ne dépend que d'un seul supérieur
2. Modes de communication : la communication entre les membres est à la fois verticale (selon la voie hiérarchique définie) et horizontale (coopération entre les niveaux hiérarchiques parallèles). [17].

II.7. Structure divisionnelle d'une entreprise

La structure divisionnelle organise l'entreprise par divisions, ou centres de profits distincts (activité spécifique, zone géographique particulière, segment de clientèle), disposant de pouvoirs étendus sur leurs produits et leurs marchés. Elle est aussi appelée structure en M (multidivisionnel). Cette structure est la plus efficace lorsque l'entreprise est diversifiée. Des fonctions supports (services centraux) peuvent faire partager leur expertise aux fonctions opérationnelles de toutes les divisions. La structure divisionnelle permet à chaque division de l'entreprise de s'organiser avec autonomie et responsabilisation. Chaque division est ensuite déclinée en structure fonctionnelle, à partir de sa propre chaîne de valeur. Cette structure est la plus souvent rencontrée parmi les grandes entreprises diversifiées.

Exemple : Saint-Gobain, leader mondial de l'habitat, se réorganise en 2019 en structure divisionnelle allégée, par pays et non plus par activités (vitrage, matériaux de construction, distribution bâtiment), afin d'être plus proche de ses marchés.

Une division peut comporter un seul DAS (domaine d'activité stratégique) ou un regroupement de DAS (autour d'un métier de l'entreprise). L'entreprise est ici organisée autour du bien ou service final qu'elle produit. Chacune des divisions de la structure organisationnelle de l'entreprise peut à son tour être structurée selon le modèle de la structure fonctionnelle.

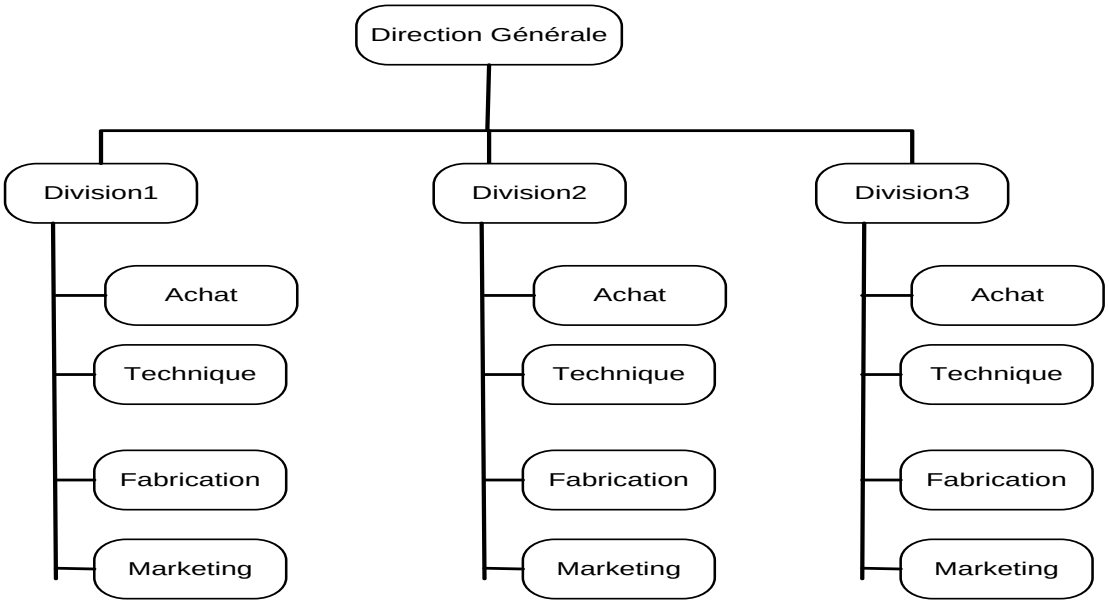

Figure 1 : Structure divisionnelle d'une entreprise

Ce type de structure est de plus en plus utilisée par les grandes entreprises qui, dans le même ordre d'idée adoptent une structure organisationnelle géographique que l'on peut assimiler à une structure divisionnelle puisque les grandes fonctions sont dupliquées dans les différentes divisions géographiques.

II.8. Informatique d'une entreprise

L'informatique est aujourd'hui très présentée à tous les niveaux de l'entreprise. Confondre l'organisation d'une entreprise à ses flux et ses traitements des données numériques est une approche qui reflète une réalité. Le concept du CIM décrit l'automatisation complète des processus de fabrication : tous les équipements de l'usine fonctionnent sous le contrôle permanent des ordinateurs, automates programmables et autres systèmes numériques. La pyramide du CIM est une représentation conceptuelle très en vogue dans le milieu industriel des années 1980, et qui reste très actuelle. Elle comporte quatre niveaux auxquels correspondent des niveaux de décision. Plus on s'élève dans cette pyramide, plus le niveau de décision/d'abstraction est important, plus la visibilité est globale et plus les horizons et cycles opérationnels s'allongent.

L'informatique est devenue indispensable pour les grandes mais également les petites entreprises comme les PME et PMI. Toute entreprise dispose d'ordinateurs pour que les salariés puissent travailler, mais également de serveurs dans le but de sauvegarder leurs données. Les serveurs permettent d'améliorer le travail de vos salariés en favorisant la collaboration entre les différents services et en facilitant le partage de données.

L'informatique d'entreprise en France est devenue un outil permettant d'améliorer la rentabilité et la productivité d'une entreprise. Cependant il suffit d'une panne informatique dans votre entreprise pour risquer de perdre toutes vos données logicielles, ou encore un simple bug informatique de votre logiciel de gestion et c'est votre entreprise qui est au point mort.

L'informatique d'entreprise est donc un véritable atout mais reste à double tranchant. De plus, la gestion en interne de votre système informatique est chronophage et coûteuse.

CHAP III : PROBLEME LINEAIRE PAR MÉTHODE DU SIMPLEXE

III.1. Introduction

L'objectif du présent chapitre est de voir comment on peut résoudre des problèmes modélisés mathématiquement où on veut maximiser ou minimiser une fonction qui dépendant de plusieurs variables qui sont soumises à plusieurs contraintes. Modéliser c'est abstraire une situation ou un objet complexe schématiquement ou mathématiquement afin de mieux comprendre.

III.2. Programmation linéaire

La programmation linéaire est une mode de résolution d'une fonction linéaire. Elle permet de déterminer l'optimum d'une fonction économique en tenant compte des contraintes.

III.2.1. Conditions de formulation d'un problème linéaire

La programmation linéaire comme étant un modèle, admet des conditions que le décideur doit valider avant de pouvoir les utiliser pour modéliser son problème. Ces conditions sont : variables de décision du problème doivent être positives, critère de sélection de la meilleure décision est décrit par une fonction linéaire de ces variables, c'est à dire, que la fonction ne peut pas contenir par exemple un produit croisé de deux de ces variables. La fonction qui représente le critère de sélection est dite fonction objective (ou fonction économique), restrictions relatives aux variables de décision (exemple : limitations des ressources) peuvent être exprimées par un ensemble d'équations linéaires. Ces équations forment l'ensemble des contraintes et paramètres du problème en dehors des variables de décisions ont une valeur connue avec certitude.

III.2.2. Etapes de formulation d'un problème linéaire

Généralement il y a trois étapes à suivre pour pouvoir construire le modèle d'un problème linéaire: Identification des variables du problème à valeur non connues ou variable de décision et les représenter sous forme symbolique (Ex : x_1, y_1), identification des contraintes du problème et les exprimer par un système d'équations linéaires et en fin identification de l'objectif ou le critère de sélection et le représenter sous une forme linéaire en fonction des variables de décision. Spécifier si le critère de sélection est à maximiser ou à minimiser.

III.2.3. Modèle mathématique d'un problème linéaire

Le problème linéaire se présente sous la forme suivante :

$$\begin{matrix} \boldsymbol{Max\ Z} \\ \begin{cases} \boldsymbol{Ax + S \leq b} \\ \boldsymbol{x \geq 0, S \geq 0} \end{cases} \end{matrix} \qquad (1.1)$$

La mise sous forme standard consiste à introduire des variables supplémentaires (une pour chaque contrainte) de manière à réécrire les inégalités (≤) sous la forme d'égalités. Chacune de ces variables représente le nombre de ressources non utilisés. On les appelle variable d'écart. La forme standard du programme linéaire est :

$$\begin{matrix} \boldsymbol{MaxZ:\ c_1x_1 + c_1x_1 + \cdots + c_Nx_N} \\ \begin{cases} \boldsymbol{a_{11}x_1 + a_{12}x_2 + \cdots + a_{1N}x_N + S_1 = b_1} \\ \boldsymbol{a_{21}x_1 + a_{22}x_2 + \cdots + a_{2N}x_N + S_2 = b_2} \\ \vdots \\ \boldsymbol{a_{M1}x_1 + a_{M2}x_2 + \cdots + a_{MN}x_N + S_M = b_M} \\ \boldsymbol{x_1 \geq 0, x_2 \geq 0, \cdots x_N \geq 0} \\ \boldsymbol{S_1 \geq 0, S_2 \geq 0, \cdots S_N \geq 0} \end{cases} \end{matrix} \qquad (1.2)$$

L'impact de ces variables d'écart sur la fonction objective est nul. Ceci explique le fait que leur existence soit tout simplement liée à une mise en forme du programme linéaire initial. Ces variables d'écart peuvent prendre des valeurs non négatives. Le fait de donner la valeur des variables d'écart à l'optimum donne une idée du nombre des ressources non utilisées.

III.2.4. Méthode des tableaux

La méthode de simplexe commence par l'identification d'une solution réalisable de base et ensuite, elle essaie de trouver d'autres solutions réalisables de base jusqu'à atteindre à la solution optimale. Ainsi, on doit tout d'abord retrouver cette solution réalisable de base.

Résumé de la procédure de la méthode du simplexe

Etapes	Justification
1. Formuler un programme linéaire pour le problème réel.	Pour obtenir une représentation mathématique du problème
2. Vérifier que le second membre du programme linéaire est positif	Ceci est nécessaire pour obtenir comme variable de base initiale l'origine
3. Ecrire le programme linéaire sous une forme standard	Mettre toutes les contraintes sous forme d'égalité
4. Construire le premier tableau de simplexe	Ce tableau correspond à la solution initiale de base
5. Choisir comme variable entrante dans la base celle qui admet le plus grand effet net positif $cj - zj$.	La valeur de $cj - zj$ indique la quantité d'augmentation de la fonction objectif si on augmente la valeur de xj d'une unité
6. Choisir la variable sortante de la base celle qui admet le plus petit ratio supérieur à zéro.	La plus petite valeur de Qi/aij indique le nombre maximal d'unité de xj qu'on peut introduire avant que la variable de base de l'$ième$ ligne ne soit égale à zéro.
7. Construire le nouveau tableau en utilisant la règle de pivot	Cette règle nous permet entre autres de calculer les valeurs des nouvelles variables de décision
8. Faire le test d'optimalité. Si $(cj - zj) \leq 0$ pour toutes les variables (hors base), la solution obtenue est donc optimale. Si non retourner à l'étape 5.	Si $(cj - zj) \leq 0$ alors on n'a pas d'intérêt à faire entrer dans la base aucune de ces variables. Une telle introduction engendra une diminution de la fonction objective.

Exemple d'application

Résoudre le programme linéaire suivant en utilisant la méthode de simplexe.

$$MAX\ Z\ =\ 25x1 + 15x2$$

$$\begin{cases} 2x1 + 2x2 \leq 240 \\ 3x1 + 1x2 \leq 140 \\ x1, x2\ \geq 0 \end{cases} \qquad (2.1)$$

- La forme standard :

$$MAX\ Z\ =\ 25x1 + 15x2 + 0e1 + 0e2$$

$$\begin{cases} 2x1 + 2x2 + e1 = 240 \\ 3x1 + 1x2 + e2 = 140 \\ x1, x2, e1, e2\ \geq 0 \end{cases} \qquad (2.2)$$

A l'origine :

$$\begin{cases} e1 = 240 \\ e2 = 140 \end{cases} \qquad (2.3) \qquad \text{et} \qquad \begin{cases} x1 = 0 \\ x2 = 0 \end{cases} \qquad (2.4)$$

→ Tableau de simplexe initial (1ère itération)

C_j			25	15	0	0	
	VB	φ	x_1	x_2	e_1	e_2	R_T
0	$e1$	240	2	2	1	1	$\frac{240}{2} = 120$
0	$e2$	140	3	1	0	0	$\frac{140}{3} = 46{,}67$ →
	Z_j		0	0	0	0	
	$C_j - Z_j$		25	15	0	0	

La variable entrante est $x1$ puisqu'elle présente le plus grand effet net positif. La variable sortante est $e2$ car elle correspond au plus petit quotient positif.

→ 2ème itération

C_j			25	15	0	0	
	VB	φ	x_1	x_2	e_1	e_2	R_T
0	$e1$	$\frac{440}{3}$	0	$\frac{4}{3}$	1	$-\frac{2}{3}$	$\frac{440}{4} = 110$
25	$x1$	$\frac{140}{3}$	$\frac{3}{3} = 1$	$\frac{1}{3}$	$\frac{0}{3}$	$\frac{1}{3}$	140
	Z_j		25	$\frac{25}{3}$	0	$\frac{25}{3}$	
	Z_j		25	$\frac{25}{3}$	0	$\frac{25}{3}$	
	$C_j - Z_j$		0	$15 - \frac{25}{3}$	0	$-\frac{25}{3}$	

La variable entrante est $x2$ et la variable sortante est $e1$

→ 3ème itération

C_j			25	15	0	0	
	VB	φ	x_1	x_2	e_1	e_2	R_T
0	$x1$	$\frac{440}{4} = 110$	0	1	$\frac{3}{4}$	$-\frac{1}{2}$	$\frac{440}{4} = 110$
25	$x2$	10	1	0	$-\frac{1}{4}$	$\frac{1}{2}$	140
	Z_j		25	15	5	5	
	$C_j - Z_j$		0	0	-5	-5	

Tous les $cj - zj \leq 0$ donc le tableau de simplexe est optimal et la solution optimale du programme linéaire est :

$$\begin{cases} \boldsymbol{x_1 = 10} \\ \boldsymbol{x_2 = 110} \end{cases} \qquad (3)$$

Z = **1900**

Remarque :

La solution optimale donnée par le dernier tableau de simplexe correspond au point φ. Le tableau du simplexe suivant est :

C_j			25	15	0	0	
	VB	φ	x_1	x_2	e_1	e_2	R_T
0	$x1$	$\frac{440}{4} = 110$	0	1	$\frac{3}{4}$	$-\frac{1}{2}$	$\frac{440}{4} = 110$
25	$x2$	10	1	0	$-\frac{1}{4}$	$\frac{1}{2}$	140
	Z_j		25	15	5	5	
	$C_j - Z_j$		0	0	-5	-5	

Le tableau est optimal et la solution correspondante est :

$x1 = 10$

$x2 = 110$

La valeur de la fonction objective est **1900**.

CHAP IV : GÉNÉRALITÉS SUR LA FOUILLE DE DONNÉES

IV.1. Introduction

Aujourd'hui les méthodes d'analyse de données sont employées dans un grand nombre de domaines qu'il est impossible d'énumérer. Actuellement ces méthodes sont beaucoup utilisées en marketing par exemple pour la gestion de la clientèle (pour proposer de nouvelles ores ciblées par exemple). Elles permettent également l'analyse d'enquêtes par exemple par l'interprétation de sondages (où de nombreuses données qualitatives doivent être prises en compte). Nous pouvons également citer la recherche documentaire qui est de plus en plus utile notamment avec internet (la difficulté porte ici sur le type de données textuelles ou autres). Le grand nombre de données en météorologie a été une des premières motivations pour le développement des méthodes d'analyse de données. En fait, tout domaine scientifique qui doit gérer de grande quantité de données de type varié ont recours à ces approches (écologie, linguistique, économie, etc.) ainsi que tout domaine industriel (assurance, banque, téléphonie, etc.). Ces approches ont également été mis à profit en traitement du signal et des images, où elles sont souvent employées comme prétraitements (qui peuvent être vus comme des filtres). En ingénierie mécanique, elles peuvent aussi permettre d'extraire des informations intéressantes sans avoir recours à des modèles parfois alourdis pour tenir compte de toutes les données.

IV.2. Valeur d'une donnée sujette à la fouille

IV.2.1. Définition d'une donnée et information

Une donnée est ce qui est connu et qui sert de point de départ à un raisonnement ayant pour objet la détermination d'une solution à un problème en relation avec cette donnée. Cela peut être une description élémentaire d'une réalité, le résultat d'une comparaison entre deux événements du même ordre (mesure) soit en d'autres termes une observation ou une mesure. La donnée brute est dépourvue de tout raisonnement, supposition, constatation, probabilité. Si elle est considérée comme indiscutable ou même si elle est indiscutée par méconnaissance, elle peut servir de base à une recherche, à un examen quelconque. Les données pouvant être de nature très différente suivant leur source, elles doivent souvent faire l'objet d'une transformation préalable avant traitement. Jusqu'à il y a quelques siècles l'être humain n'a eu connaissance du monde réel qu'à travers ses sens naturels, la vue, l'ouïe, l'odorat, Son cerveau a développé une capacité de raisonnement permettant de combler un peu les lacunes inhérentes à la faiblesse de ses capteurs. Cela lui a permis de développer son intelligence et de développer des outils permettant d'augmenter sa capacité à connaître le monde réel.

L'information est un élément de connaissance, qui peut être collecté, traité, conservé, communiqué au sein de l'organisation ou auprès de ses partenaires. L'information est constituée de deux éléments : des données et un sens qui dépend de chaque individu.

IV.2.2. Types de données et informations

En analyse de données, on distingue principalement deux modèles de données ou variables : les données quantitatives et les données qualitatives. Il existe une différence notable entre une donnée quantitative et une donnée qualitative. Ces deux modèles sont largement utilisés en analyse de données. Pour en tirer le meilleur parti, il est essentiel de les maîtriser pour réaliser une analyse pertinente des données et tirer de meilleures conclusions.

Une donnée quantitative ou numérique désigne des informations ou des caractéristiques quantifiables qui prennent des nombres comme valeur. Les données quantitatives sont structurées et parfaitement adaptées à l'analyse de données. Le nombre d'employés dans une entreprise, l'âge, le poids, la hauteur, la température, le temps, la superficie, le chiffre d'affaires d'une société sont autant d'exemples de données quantitatives. Une donnée quantitative peut être représentée à l'aide de tableaux, de diagrammes et de graphiques. On comprend ici la différence entre données structurées et non-structurées. Une donnée quantitative est dite continue lorsqu'elle prend un nombre infini de valeurs réelles à l'intérieur d'un intervalle donné. La taille d'une personne est un exemple de données quantitatives continues. Même si elle ne peut pas prendre toutes les valeurs réelles possibles, elles peuvent prendre une infinité de valeurs dans un intervalle défini selon l'objet mesuré. Le poids d'une personne, la hauteur d'un immeuble sont également des exemples de données quantitatives continues. Entre deux valeurs de poids par exemple, il y a des millions de poids possibles. En général, les données qui proviennent d'une mesure sont quantitatives. Etudes statistiques, on désigne par données quantitatives discrètes des data qui ne peuvent prendre qu'un nombre fini de valeurs réelles possibles au sein d'un intervalle donné. Elles ne peuvent donc pas être réduites en parties plus petites. C'est d'ailleurs en cela qu'une donnée quantitative discrète se distingue d'une donnée quantitative continue. Le nombre d'employés d'une entreprise ou encore la taille d'un ménage sont des exemples de données quantitatives discrètes. Le nombre d'employés d'une entreprise est également une donnée quantitative discrète. En prenant l'exemple des entreprises qui ont au plus 100 employés, le nombre de valeurs possibles prises par une telle variable ne peut bien évidemment pas excéder 100. On sait en effet qu'il est impossible pour une entreprise de disposer d'un nombre d'employés qui serait une fraction d'un nombre entier comme 60,9 par exemple.

L'analyse de données numériques prend place dans le cadre d'une étude quantitative. La première étape de celle-ci est toujours la collecte des données ou des informations. Qu'elles soient discrètes ou continues, les données quantitatives peuvent être obtenues au moyen d'une méthode ou stratégie comme l'enquête ou l'observation contrôlée. Les sondages, les études longitudinales et les entretiens téléphoniques ou face-à-face sont aussi des méthodes et techniques habituellement utilisées pour la collecte de données quantitatives. L'étape suivante est le traitement des données. À cette phase, les data récoltées sont remises en forme afin d'être analysées plus efficacement. L'analyse peut alors commencer. Les données collectées peuvent être recoupées sous forme de graphique, de tableau. Ces résultats sont ensuite analysés au moyen de logiciels et d'outils statistiques. Des conclusions sont ensuite tirées pour l'étude.

Les données qualitatives ou catégoriques font référence à une caractéristique non quantifiable le plus souvent issue d'un comptage. Contrairement à une donnée quantitative, une data qualitative ne donne pas de chiffres qui peuvent faire l'objet de représentation graphique. Ces données servent notamment au classement des réponses en fonction des propriétés et d'attributs. Les données qualitatives sont souvent interprétées dans un langage simple. Elles sont utilisées pour décrire les informations, caractériser des objets ou des observations. Leur nature descriptive les rend plus difficiles à analyser. L'utilisation de données qualitatives permet aux chercheurs et aux entreprises de mieux cerner les comportements, la personnalité et les émotions de leurs répondants. De même, dans le cadre d'une étude de marché par exemple, les données qualitatives jouent un rôle déterminant puisqu'elles aident les chercheurs à mieux comprendre leurs clients. La connaissance des motivations de ces derniers grâce aux données qualitatives aide les marques à prendre de meilleures décisions commerciales.

Une donnée qualitative nominale décrit un nom ou une catégorie sans ordre particulier. Les données qualitatives nominales servent essentiellement pour étiqueter des variables. C'est d'ailleurs pour cette raison qu'elles sont parfois appelées étiquettes. Le mode de transport utilisé par les employés d'une entreprise, le sexe, la couleur associée à une marque sont autant d'exemples de données qualitatives nominales. Une donnée qualitative ordinale est une donnée qui présente des valeurs définies par une relation d'ordre entre les différentes catégories possibles. L'appréciation des clients de la qualité des services d'une entreprise est un exemple de données qualitatives ordinales. Elle présente en effet des catégories comme « Bon », « Très bon », « Excellent » entre lesquelles une relation évidente d'ordre peut être établie. La catégorie « Très bien » est meilleure que la catégorie « Bon », mais moins intéressante que la catégorie « Excellente ». Les données qualitatives ordinales

ont par contre un défaut. Quand bien même on y trouve un ordre, on ne peut par exemple pas savoir dans quelle mesure une catégorie donnée est meilleure que l'autre.

Les données qualitatives sont très prisées dans le cadre des recherches en sciences sociales comme la sociologie. Différentes méthodes sont utilisées pour leur collecte au sein d'un échantillon d'une population donnée. C'est notamment le cas des entretiens individuels, des groupes de discussion, des études de cas, des questions d'enquête ouvertes, de la recherche observationnelle. Le sondage peut également être utilisé pour la collecte de données qualitatives. Les entretiens par exemple favorisent une meilleure analyse d'une hypothèse notamment grâce à une approche individuelle. Quant aux groupes de discussion, ils permettent à plusieurs personnes d'exprimer leurs idées et opinions sur un sujet donné. Les études de cas fournissent aux entreprises des retours de consommateurs.

IV.2.3. Caractéristiques d'une donnée et information

Une information est caractérisée par sa forme, son mode de présentation, ses qualités et son coût. Parmi les différentes formes que peut prendre une information, les plus courantes sont : les informations orales, les informations écrites, les informations visuelles, les informations audiovisuelles, les informations qualitatives et les informations quantitatives**.** On peut également relever des informations olfactives, tactiles et gustatives. L'information peut avoir un codage, être traduite dans plusieurs langues et avoir une couleur**.** Pour qu'une information soit de qualité, elle doit être fiable (la source est connue ou est clairement identifiable), pertinente, (elle doit répondre à un besoin), actualité (les renseignements sont récents et mis à jour régulièrement), non redondante (nouvelle ; elle ne doit pas être déjà dans l'organisation et accessible **(**on peut l'obtenir facilement).

L'information est un élément primordial dans l'entreprise. En effet, elle représente un outil de prise de décision. Par exemple, lorsque le vendeur fait constat, auprès de son manager, qu'il ne reste que peu de produits en stock, le manager va déclencher le processus de réapprovisionnement. Il va prendre la décision de commander de nouveaux produits. De plus, l'information est un outil de communication interne lorsqu'elle intervient, par exemple au cours d'une réunion d'équipe, mais aussi un outil de communication externe lorsqu'elle est transmise entre l'entreprise et ses partenaires.

En fin, l'information est un outil de travail collectif. Par exemple, lors d'une réunion entre les représentants et le chef régional, les informations collectées et diffusées par chaque représentant

(exemple : arrivée de nouveaux concurrents, ouverture de nouveaux points de ventes) vont permettre d'améliorer les performances de l'ensemble de l'équipe en ajustant les actions de chacun.

Remarque :

L'information est une donnée supplémentaire dans le patrimoine intellectuel de l'individu et de l'organisation. Cette information peut revêtir plusieurs formes (écrites, orale, qualitative, …) et peut être représentée par un code, une couleur.

Toutefois, pour qu'une information soit de qualité, il faut qu'elle remplisse cinq critères à savoir la fiabilité, l'actualité, l'originalité et l'accessibilité. La plupart du temps, l'obtention de cette information a un coût qui doit être raisonnable par rapport à l'objectif à atteindre. L'information remplit trois rôles principaux dans l'entreprise. Elle est perçue comme un outil d'aide à la décision, un outil de communication interne et externe et un outil de travail collectif.

IV.3. Techniques de Fouille de données

La liste suivante indique les tâches les plus courantes que le data mining est amené accomplir : la description, l'estimation, la prévision, la classification, le clustering ou groupement et l'association

IV.3.1. Technique de description

Parfois, les chercheurs et les analystes essaient simplement de trouver des façons de décrire des tendances cachées dans les données. Les descriptions des modèles et des tendances servent à expliquer ou vérifier un fait. Par exemple : « ceux qui ont le plus de diplômes sont les plus susceptibles d'avoir un poste à responsabilité. ».[1]

L'analyse en composantes principales, ou ACP, est une méthode de réduction de la dimensionnalité qui est souvent utilisée pour réduire la dimensionnalité de grands ensembles de données, en transformant un grand ensemble de variables en une plus petite qui contient encore la plupart des informations du grand ensemble. La réduction du nombre de variables d'un ensemble de données se fait naturellement au détriment de la précision, mais l'astuce de la réduction de la dimensionnalité consiste à échanger un peu de précision contre de la simplicité., Parce que les petits ensembles de données sont plus faciles à explorer et à visualiser et rendent l'analyse des données beaucoup plus facile et plus rapide pour les algorithmes d'apprentissage automatique sans variables étrangères à traiter. Donc, pour résumer, l'idée de l'ACP est simple : réduire le nombre de variables d'un ensemble de données, tout en préservant autant d'informations que possible.

L'objectif de cette étape est de normaliser la gamme des variables initiales continues afin que chacune d'entre elles contribue également à l'analyse., Plus précisément, la raison pour laquelle il est essentiel d'effectuer la normalisation avant la ACP, est que cette dernière est assez sensible en ce qui concerne les variances des variables initiales. Autrement dit, s'il existe de grandes différences entre les plages de variables initiales, les variables avec des plages plus grandes domineront sur celles avec de petites plages (par exemple, une variable comprise entre 0 et 100 dominera sur une variable comprise entre 0 et 1), ce qui entraînera des résultats biaisés. Ainsi, la transformation des données à des échelles comparables peut éviter ce problème., Mathématiquement, cela peut être fait en soustrayant la moyenne et en divisant par l'écart-type pour chaque valeur de chaque variable. Une fois la normalisation effectuée, toutes les variables seront transformées à la même échelle. Le but de cette étape est de comprendre comment les variables de l'ensemble de données d'entrée varient par rapport à la moyenne les unes par rapport aux autres, ou en d'autres termes, de voir s'il existe une relation entre elles., Parce que parfois, les variables sont fortement corrélées de telle sorte qu'elles contiennent des informations redondantes. Ainsi, afin d'identifier ces corrélations, nous calculons la matrice de covariance. La matrice de covariance est une matrice symétrique p × p (où p est le nombre de dimensions) qui a comme entrées les covariances associées à toutes les paires possibles des variables initiales.

Par exemple, pour un ensemble de données tridimensionnelles avec 3 variables x, y et z, la matrice de covariance est une matrice 3×3 de ceci à partir de : matrice de Covariance pour les données tridimensionnelles puisque la covariance d'une variable avec elle-même est sa variance ($\boldsymbol{COV(a,a) = Var(a)}$), dans la diagonale principale (de haut en bas à droite) nous avons en fait les variances de chaque variable initiale., Et puisque la covariance est commutative ($\boldsymbol{Cov(a,b) = Cov(b,a)}$), les entrées de la matrice de covariance sont symétriques par rapport à la diagonale principale, ce qui signifie que les parties triangulaires supérieure et inférieure sont égales. Qu'est-ce que les covariances que nous avons comme entrées de la matrice nous disent sur les corrélations entre les variables ? C'est en fait le signe de la covariance qui compte : si positif alors (les deux variables augmentent ou diminuent ensemble, corrélées), si négatif alors (l'une augmente quand l'autre diminue, inversement corrélée). Maintenant, que nous savons que la matrice de covariance n'est pas plus qu'une table qui résume les corrélations entre toutes les paires de variables possibles, passons à l'étape suivante.,

Les vecteurs propres et les valeurs propres sont les concepts d'algèbre linéaire que nous devons calculer à partir de la matrice de covariance afin de déterminer les composantes principales des

données. Avant d'arriver à l'explication de ces concepts, comprenons d'abord ce que nous entendons par Composants principaux. Les composantes principales sont de nouvelles variables qui sont construites comme des combinaisons linéaires ou des mélanges des variables initiales., Ces combinaisons sont faites de telle sorte que les nouvelles variables (c.-à-d. les composantes principales) ne sont pas corrélées et que la plupart des informations contenues dans les variables initiales sont compressées ou compressées dans les premières composantes. Donc, l'idée est que les données à 10 dimensions vous donnent 10 Composants principaux, mais PCA essaie de mettre le maximum d'informations possibles dans le premier composant, puis le maximum d'informations restantes dans le second et ainsi de suite, jusqu'à avoir quelque chose comme indiqué dans le tracé d'éboulis ci-dessous., Organiser les informations dans les composants principaux de cette façon, vous permettra de réduire la dimensionnalité sans perdre beaucoup d'informations, et ce en écartant les composants avec des informations faibles et en considérant les composants restants comme vos nouvelles variables., Une chose importante à réaliser ici est que les composants principaux sont moins interprétables et n'ont pas de signification réelle car ils sont construits comme des combinaisons linéaires des variables initiales. Géométriquement parlant, les composantes principales représentent les directions des données qui expliquent un montant maximum de variance, c'est-à-dire les lignes qui capture la plupart des informations de données., La relation entre la variance et de l'information ici, c'est que, plus la variance portée par une ligne, plus la dispersion des points de données le long d'elle, et plus la dispersion le long d'une ligne, plus les informations qu'il possède. Pour dire tout cela simplement, il suffit de penser aux composants principaux comme de nouveaux axes qui fournissent le meilleur angle pour voir et évaluer les données, afin que les différences entre les observations

Comme il y a autant de composantes principales que de variables dans les données, les composantes principales sont construites de telle manière que la première composante principale représente la plus grande variance possible dans l'ensemble de données. Par exemple, supposons que le nuage de points de notre ensemble de données est comme indiqué ci-dessous, pouvons-nous deviner le premier composant principal ? Oui, c'est approximativement la ligne qui correspond aux marques violettes car elle passe par l'origine et c'est la ligne dans laquelle la projection des points (points rouges) est la plus étalée. Ou mathématiquement parlant, c'est la ligne qui maximise la variance (la moyenne des distances au carré des points projetés (points rouges) à l'origine). La deuxième composante principale est calculée de la même manière, à la condition qu'elle ne soit pas corrélée avec (c'est-à-dire perpendiculaire à) la première composante principale et qu'elle tienne compte de

la variance la plus élevée suivante., Cela continue jusqu'à ce qu'un total de P composantes principales ait été calculé, égal au nombre initial de variables.

Maintenant que nous avons compris ce que nous entendons par Composants principaux, revenons aux vecteurs propres et aux valeurs propres. Ce que vous devez d'abord savoir à leur sujet, c'est qu'ils viennent toujours par paires, de sorte que chaque vecteur propre a une valeur propre. Et leur nombre est égal au nombre de dimensions des données. Par exemple, pour un ensemble de données à 3 dimensions, il y a 3 variables, donc il y a 3 vecteurs propres avec 3 valeurs propres correspondantes., Sans plus tarder, ce sont les vecteurs propres et les valeurs propres qui sont derrière toute la magie expliquée ci-dessus, car les vecteurs propres de la matrice de Covariance sont en fait les directions des axes où il y a le plus de variance (le plus d'informations) et que nous appelons composantes principales. Et les valeurs propres sont simplement les coefficients attachés aux vecteurs propres, qui donnent la quantité de variance portée dans chaque composante principale. En classant vos vecteurs propres par ordre de leurs valeurs propres, du plus haut au plus bas, vous obtenez les composants principaux par ordre de signification.,

Supposons que notre ensemble de données est à 2 dimensions avec 2 variables x, y et que les vecteurs propres et les valeurs propres de la matrice de covariance sont les suivants : V1, V2 et $\lambda 1, \lambda 2$. Si nous classons les valeurs propres par ordre décroissant, nous obtenons $\lambda 1 > \lambda 2$, ce qui signifie que le vecteur propre qui correspond à la première composante principale (**PC1**) est V1 et celui qui correspond à la deuxième composante (**PC2**) est V2., Après avoir les composantes principales, pour calculer le pourcentage de variance (information) pris en compte par chaque composante, nous divisons la valeur propre de chaque composante par la somme des valeurs propres. Si nous appliquons cela sur l'exemple ci-dessus, nous constatons que PC1 et PC2 portent respectivement 96% et 4% de la variance des données.

Comme nous l'avons vu à l'étape précédente, le calcul des vecteurs propres et leur classement par leurs valeurs propres dans l'ordre décroissant, nous permettent de trouver les composants principaux par ordre de signification., Dans cette étape, ce que nous faisons est de choisir de conserver tous ces composants ou de rejeter ceux de moindre importance (de faibles valeurs propres), et de former avec les autres une matrice de vecteurs que nous appelons vecteur D'entités. Donc, le vecteur d'entités est simplement une matrice qui a comme colonnes les vecteurs propres des composants que nous décidons de conserver. Cela en fait la première étape vers la réduction de la dimensionnalité, car si nous choisissons de ne garder que p vecteurs propres (composants) hors

de n, l'ensemble de données final n'aura que des dimensions p., En continuant avec l'exemple de l'étape précédente, nous pouvons soit former un vecteur caractéristique avec les deux vecteurs propres V1 et v2 :

Ou ignorer le vecteur propre v2, Qui est celui de moindre importance, et former un vecteur caractéristique avec v1 seulement : Le rejet du vecteur propre v2 réduira la dimensionnalité de 1 et entraînera par conséquent une perte d'informations dans l'ensemble de données final., Mais étant donné que v2 ne transportait que 4% des informations, la perte ne sera donc pas importante et nous aurons toujours 96% des informations portées par v1. Donc, comme nous l'avons vu dans l'exemple, c'est à vous de choisir de conserver tous les composants ou les jeter celles de moindre importance, en fonction de ce que vous cherchez. Parce que si vous voulez simplement décrire vos données en termes de nouvelles variables (composantes principales) qui ne sont pas corrélées sans chercher à réduire la dimensionnalité, il n'est pas nécessaire de laisser de côté les composantes moins significatives., Dans les étapes précédentes, en dehors de la normalisation, vous n'apportez aucune modification sur les données, vous sélectionnez simplement les composants principaux et formez le vecteur d'entités, mais l'ensemble de données d'entrée reste toujours en termes d'axes, Dans cette étape, qui est la dernière, l'objectif est d'utiliser le vecteur caractéristique formé à l'aide des vecteurs propres de la matrice de covariance, pour réorienter les données des axes d'origine vers celles représentées par les composantes principales (d'où le nom D'analyse des composantes principales). Cela peut être fait en multipliant la transposition de l'ensemble de données d'origine par la transposition du vecteur d'entités. [4]

IV.3.2. Technique d'estimation

L'estimation est similaire à la classification, sauf que la variable cible est numérique plutôt que catégorique. Les modèles sont construits en utilisant des données, qui fournissent la valeur de la variable cible, ainsi que les « prédicteurs ». Par exemple : « l'estimation de la pression artérielle d'un patient d'hôpital, basée sur son âge, son sexe, son indice de masse corporelle, et le taux de sodium. La relation entre la pression artérielle et le prédicteur variable de l'ensemble de formation nous donnerait un modèle d'estimation. Nous pouvons alors appliquer ce modèle à de nouveaux cas.

IV.3.3. Technique de prédiction

L'analyse prédictive est de plus en plus utilisée par les entreprises et notamment par les départements marketing, les institutions financières et mêmes les organismes médicaux. En

combinant une analyse de données historiques et nouvelles, elle aide les organisations à anticiper des tendances, à prévoir et évaluer des risques et ainsi à prendre des décisions optimales pour mener les bonnes actions auprès des bonnes personnes et au bon moment. La prédiction est semblable à la classification et l'estimation, sauf que pour la prévision, les résultats se situent dans l'avenir. Exemples de tâches de prévision appliquée au marketing : « Prédire le prix d'un stock de trois mois dans le futur ».

L'analyse prédictive (ou logique prédictive) est la technique analytique et statistique qui, en utilisant à la fois des données actuelles et historiques, permet de créer hypothèses et prédictions sur des événements futurs. Dans le monde de l'entreprise, cette logique analytique permet de dresser des schémas et modèles prédictifs afin d'anticiper des tendances et de détecter des risques et des opportunités. La création de ses modèles prédictifs (patterns) repose sur le data mining (exploration de données) qui est un processus d'analyse de volumes importants de données et notamment du Big Data. Ce procédé rapproche des data afin de déceler des relations entre des variables et de transformer les données en informations exploitables stratégiquement. Il est cependant essentiel de comprendre que la logique prédictive fournit des probabilités et des hypothèses. Elle n'apporte pas de résultat véridique ou absolu. On peut voir cette méthode statistique comme un outil mais non comme une science exacte. Il convient donc de l'utiliser comme tel, pour ensuite la rapprocher de ses pratiques et de sa connaissance pour en faire un outil stratégique, et non de suivre aveuglément les prédictions formulées à la lettre.

Ce type d'analyse est très utilisé dans le domaine marketing puisqu'il permet de réaliser des prédictions sur les comportements, les préférences et les besoins des clients. Plus précisément, l'analyse prédictive est une démarche statistique qui permet d'optimiser sa stratégie de rétention client (fidélisation) et de limiter l'attrition de la clientèle. Concrètement, en utilisant des informations historiques et récentes, une organisation va pouvoir anticiper les attentes de son client et ainsi lui proposer une solution qui réponde à son besoin (produit, offre promotionnelle, message, service, etc.) avant que ce dernier ne manifeste cette attente.

Dans le domaine sanitaire, l'analyse prédictive se révèle est un allié redoutable. En observant et combinant des données sur l'historique médical des patients et des diagnostics actuels, cette logique statistique va permettre de déceler des prédispositions à des maladies comme le diabète. Grâce à ces informations et ces prédictions, les professionnels de santé peuvent apporter une attention particulière à certains symptômes ou programmer des examens médicaux réguliers chez des patients considérés comme « à risque ». En instaurant une maintenance dite prédictive, vous assurez le fonctionnement continu de vos machines et de vos systèmes. Les données historiques et actuelles

vous permettent de prédire d'éventuelles pannes pour réparer ou remplacer le matériel avant qu'une défaillance ne se manifeste et mette à mal votre productivité. Ce processus et ces modèles de prédiction vous aide à limiter le facteur risques, à augmenter votre efficacité opérationnelle et à faire en sorte que votre chiffre d'affaires ne soit pas tributaire d'éventuelles incidents mécaniques et techniques.

Dans le domaine financier et bancaire, cette méthode analytique permet de prévenir et réduire le phénomène de fraude. En explorant les données des opérations frauduleuses précédentes et en les combinant à leurs fichiers actuels, les institutions financières vont pouvoir détecter les risques et, de ce fait, les contrer en menant des actions optimales. Mais c'est en combinant des modèles prédictifs à l'analyse Big Data et aux solutions informatiques innovantes comme apprentissage profond, apprentissage automatique et l'intelligence artificielle qu'un organisme va être en mesure d'identifier les actions les plus adéquates à mettre en œuvre. On passe alors de l'analyse descriptive et prédictive à l'analyse prescriptive.

La méthode des moindres carrés a vu le jour au début du 19e siècle et a été construite par Legendre (1805) et Gauss (1809) [10]. Cette méthode permet de faire une relation entre les données expérimentales et le modèle mathématique supposé décrire ces données, ceci réglant le problème des erreurs de mesure.

Exemple d'algorithme de la méthode des moindres carrées pour prédire les ventes :

On donne le tableau suivant concernant les informations financières d'un magasin de restauration rapide dans la banlieue parisienne. Vous devez étudier la prévision de ces données pour l'année N+1.

Années	Chiffres d'affaires
N-4	42
N-3	58
N-2	64
N-1	74

Dans cette méthode de prévision des ventes d'ajustement linéaire, on pose les éléments suivants : la période considérée soit « xi », le chiffre d'affaires correspond à « yi » et le nombre de lignes du tableau des éléments passés correspond à « n ». Avec les éléments précédents, vous devez

calculer deux moyennes : la moyenne des périodes et la moyenne des chiffres d'affaires. Le terme moyenne est identifié à l'aide d'une petite barre sur la lettre comme dans les formules suivantes. La formule de la moyenne des périodes que l'on appelle « moyenne de x » :

$$\bar{x} = \frac{\sum x_i}{n} \qquad (3.1)$$

La formule de la moyenne des chiffres d'affaires que l'on appelle « moyenne de y » :

$$\bar{y} = \frac{\sum y_i}{n} \qquad (3.2)$$

La lecture de la formule des moyennes des périodes est la suivante :

Moyenne de « x » égal somme des « xi » divisée par « n »

La lecture de la formule des moyennes des chiffres 'affaires est la suivante :

Moyenne de « y » égale somme des « yi » divisée par « n »

Dans un premier temps, vous ne pouvez pas appliquer de formules pour trouver ces deux moyennes. Vous devez passer par l'élaboration d'un tableau préparatoire. Voici un exemple de tableau préparatoire qui va vous aider à trouver les éléments utiles : La ligne la plus importante du tableau est la ligne « $Total$ ». Le « $Rang$ » correspond à un chiffre que l'on attribue à une période afin de pouvoir s'en servir pour appliquer une formule. Il faut toujours commencer par le chiffre « 1 » pour la période la plus lointaine et incrémenter ($augmenter$) de « 1 » pour chaque nouvelle période. Le total de la colonne « xi » est utile pour calculer la moyenne des périodes. Le total de la colonne « yi » est utile pour le calcul des moyennes des chiffres d'affaires. Je vais appliquer les formules avec les éléments chiffrés de notre exemple de départ.

Dans notre exemple « n » = 4 car il y a quatre lignes dans le tableau de l'énoncé. Les formules (3.1) et (3.2) de la moyenne de x est la suivante :

$$\bar{x} = \frac{10}{4} = 2{,}5 \qquad (3.1.a)$$

Et la moyenne de y est :

$$\bar{y} = \frac{241}{4} = 60{,}25 \qquad (3.2.a)$$

Ce même tableau sert également à calculer d'autres formules. Pour déterminer l'équation de la droite de la forme « $y = ax + b$ », il est nécessaire de trouver dans un premier temps les paramètres de l'équation, c'est-à-dire l'élément « a » et l'élément « b ». C'est là qu'interviennent les totaux des colonnes « $xi.yi$ » et « xi au carré ». « $xi.yi$ » signifie xi multiplié par yi. Ce calcul est fait pour chaque ligne du tableau. « xi » au carré signifie que pour chaque ligne de « xi » du tableau la valeur est mise au carré. Pour trouver les paramètres « a » et « b » il faut appliquer les formules suivantes :

$$a = \frac{\sum x_i.y_i - n.\bar{x}.\bar{y}}{\sum x_i^2 - n\bar{x}^2} \qquad (4.1)$$

$$b = \bar{y} - a\bar{x} \qquad (4.2)$$

Dans la formule du paramètre « a », nous avons les éléments suivants : la somme du produit « $xi.yi$ », le nombre de lignes du tableau « n », la somme de la colonne « xi au carré » et les moyennes de x et de y. Dans la formule du paramètre « b », nous avons les moyennes de x et de y d'une part, et d'autre part le paramètre « a ».

Voici donc ce que cela donne dans notre exemple :

Années	**Xi (rang)**	**CA y_i**	$x_i.y_i$	x_i^2
N-4	1	42	42	1
N-3	2	58	116	4
N-2	3	64	192	9
N-1	4	77	308	16
Total	**10**	**241**	**658**	**30**

Maintenant nous pouvons appliquer les formules$(4.1)\ et\ (4.2)$ des paramètres « a » et « b » :

$$a = \frac{658 - (4*2{,}5*60{,}25)}{30 - (4*2{,}5*2{,}5)}\ soit\ donc\ \frac{55{,}5}{5} = 11.10 \qquad (4.1.a)$$

$$b = 60{,}25 - (11{,}10x2{,}5)\ soit\ donc\ 32{,}5 \qquad (4.2.a)$$

Pour interpréter les valeurs trouvées, il faut remplacer « a » et « b » dans l'équation

$$y = ax + b \qquad (5)$$

Par ces mêmes valeurs :

$$y = 11{,}10x + 32{,}5 \qquad (5.a)$$

On peut donc écrire que cette équation permet de trouver n'importe quels chiffres d'affaires futurs (c'est à dire prévisionnels). Une fois l'équation déterminée, on doit remplacer « x » par le rang de la période recherchée. Dans notre exemple le rang de la période recherchée est six.

Pourquoi « six » ? Dans notre tableau initial, le dernier rang est quatre correspondants à la l'année « N-1 ».

Donc l'année N correspond au rang cinq et l'année N+1 correspond bien au rang six.

D'où l'équation suivante :

$$y = (11{,}10\ x\ 6) + 32{,}5 = 99{,}10 \qquad (5.b)$$

Nous pouvons donc écrire que le chiffre d'affaires prévisionnel pour l'année N+1 est de 99,10 K€.

Signalons que pour déterminer l'équation de la droite d'ajustement d'un nuage de points donné, on préférera utiliser une méthode basée sur la minimisation des carrés des écarts entre les points du nuage et des points de la droite d'ajustement. La méthode des moindres carrés consiste à déterminer la droite dite « de régression de y en x ». Dans la pratique, on détermine cette droite de régression de y en x, d'équation $y = ax + b$, à l'aide de la calculatrice. Les coefficients a et b de l'équation de cette droite sont définis par $a = \frac{\sigma xy}{\sigma x^2}$ et $b = \bar{y} - a\bar{x}$, où σx est l'écart-type de la série x, et σxy la covariance des séries x et y.

IV.3.4. Technique de classification

Supposons qu'un décideur veuille classer ses employés par tranches de revenu, ou n'importe quelle autre caractéristique associée à cette personne, comme l'âge, le sexe et la profession. Cette tâche est une tâche de classification. Dans ce projet prenons un exemple d'une méthode ou algorithme de K plus proches voisins. L'algorithme K-Nearest Neighbors python d'apprentissage des k plus proches voisins est utilisé pour effectuer de la classification de données. Pour prédire la classification d'une nouvelle donnée, l'algorithme se base sur les k enregistrements issus de l'ensemble de données d'apprentissage sont alors localisés les plus similaires à ce nouvel enregistrement. La similitude entre les enregistrements peut être mesurée de différentes manières. Et généralement un bon point de départ est la distance euclidienne.

L'algorithme de k plus proches voisins est l'un des algorithmes utilisés dans le domaine de l'intelligence artificielle. C'est un algorithme d'apprentissage automatique supervisé qui attribue une catégorie à un élément en fonction de la classe majoritaire de ses plus proches voisins dans l'échantillon d'entraînement. Son principe peut être résumé par cette phrase : Dis-moi qui sont tes amis et je te dirai qui tu es. Cet algorithme d'apprentissage automatique est par exemple utilisé par des entreprises d'Internet comme Amazon, Netflix, Spotify ou iTunes afin de prévoir si vous seriez ou non intéressés par un produit donné en utilisant vos données et en les comparant à celles des clients ayant acheté ce produit particulier.
Cet algorithme a été introduit en 1951 par Fix et Hodges dans un rapport de la faculté de médecine aéronautique de la US Air Force. [1]

Pour une entrée x :

1) Trouver les k entrée partie les données d'entraînement qui sont les plus "proches" de mon entrée x (ici on utilisera par exemple la distance euclidienne)

2) Faire voter chacune de ces données d'entraînement pour sa classe y.

3) Retourner la classe majoritaire

Ainsi, Le succès de l'algorithme va reposer sur la quantité de donnée d'entraînement et sur la qualité de la mesure de la distance entre 2 vecteurs x.

IV.3.5. Technique de clustering

Le Clustering désigne le regroupement des données, des observations ou des cas dans des classes d'objets similaires. Un cluster maximise la similarité des objets de du même cluster et minimise la similarité des objets de cluster différents. En effet, il n'y a pas de variable cible pour le clustering. La tâche de clustering ne cherche pas à classer, estimer, ou prédire la valeur d'une variable cible. Mais plutôt à segmenter l'ensemble des données en sous-groupes relativement homogènes à l'aide de mesures de distances. L'approche du clustering est la suivante :

1) Algorithmes de Partitionnement : Construire plusieurs partitions puis les évaluer selon certains critères
2) Algorithmes hiérarchiques : Créer une décomposition hiérarchique des objets selon certains critères
3) Algorithmes basés sur la densité : basés sur des notions de connectivité et de densité

Les caractéristiques du clustering sont les suivantes : Extensibilité, habilité à traiter différents types de données, découverte de clusters de différentes formes, connaissances requises (paramètres de l'algorithme) et habilité à traiter les données bruitées et isolées.

Les étapes de l'algorithme à partitionnement sont les suivants :

1) Construire une partition à k clusters d'une base D de n objets
2) Les k clusters doivent optimiser le critère choisi
3) Global optimal : Considérer toutes les k-partitions
4) Heuristic methods: Algorithmes k-means et k-medoids
5) k-means (MacQueen'67) : Chaque cluster est représenté par son centre
6) k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87) : Chaque cluster est représenté par un de ses objets

L'algorithme k-means est en 4 étapes :

1) Choisir k objets formant ainsi k clusters

2) (Ré)affecter chaque objet O au cluster Ci de centre Mi tel que dist (O, Mi) est minimal

3) Recalculer Mi de chaque cluster (le barycentre)

4) Aller à l'étape 2 si on vient de faire une affectation

K-Means : Exemple

A= {1,2,3,6,7,8,13,15,17}. Créer 3 clusters à partir de A

On prend 3 objets au hasard. Supposons que c'est 1, 2 et 3.

Ça donne C1= {1}, M1=1, C2= {2}, M2=2, C3= {3} et M3=3

Chaque objet O est affecté au cluster au milieu duquel, O est le plus proche. 6 est affecté à C3 car dist(M3,6) <dist(M2,6) et dist(M3,6) <dist(M1,6) ; On a :

1) C1 = {1}, M1=1,

2) C2 = {2}, M2=2

3) C3 = {3, 6,7,8,13,15,17}, M3=69/7=9.86

a. K-Means : Exemples(suite)

1) dist (3, M2) <dist(3,M3) ⟹ 3 passe dans C2. Tous les autres objets ne bougent pas. C1= {1}, M1=1, C2= {2,3}, M2=2.5, C3= {6,7,8,13,15,17} et M3= 66/6=11

2) dist(6,M2)<dist(6,M3) ⟹ 6 passe dans C2. Tous les autres objets ne bougent pas. C1= {1}, M1=1, C2= {2,3,6}, M2=11/3=3.67, C3= {7,8,13,15,17}, M3= 12

3) dist (2, M1) <dist (2, M2) ⟹ 2 passe en C1. dist (7, M2) <dist (7, M3) ⟹ 7 passe en C2. Les autres ne bougent pas. C1= {1,2}, M1=1.5, C2= {3,6,7}, M2=5.34, C3= {8,13,15,17}, M3=13.25

4) dist (3, M1) <dist (3, M2) ⟹ 3 passe en 1. dist (8, M2) <dist (8, M3) ⟹8 passe en 2

C1= {1,2,3}, M1=2, C2= {6,7,8}, M2=7, C3= {13,15,17}, M3=15

Voici alors un exemple d'une idée du processus de l'algorithme de K-moyennes sous forme graphique. Cette première étape consiste à établir ou préciser les centres (centroïdes) initiaux.

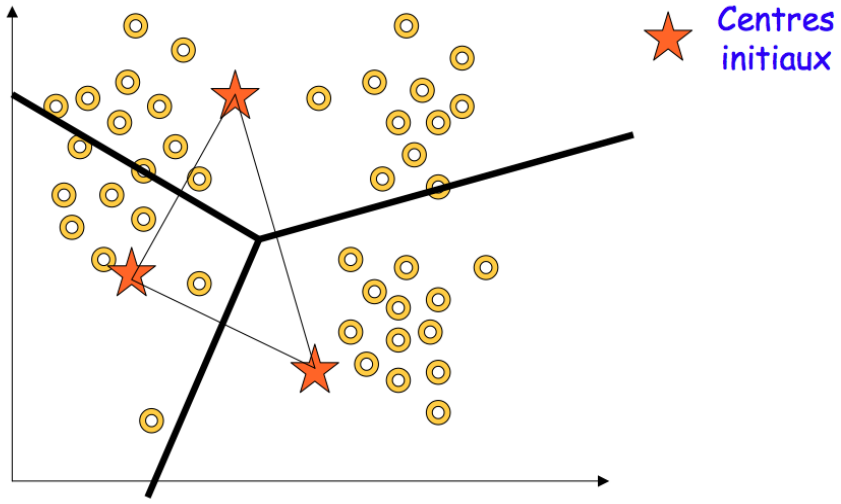

Figure 2 : Première illustration de l'algorithme de K-Means

La figure en dessous montre la seconde étape de l'algorithme de K-moyennes. Sur cette étape, les points se regroupent suivant le point centroïde plus proche et puis recherche encore de nouveau d'autres centroïdes parmi les groupes nouvellement formés.

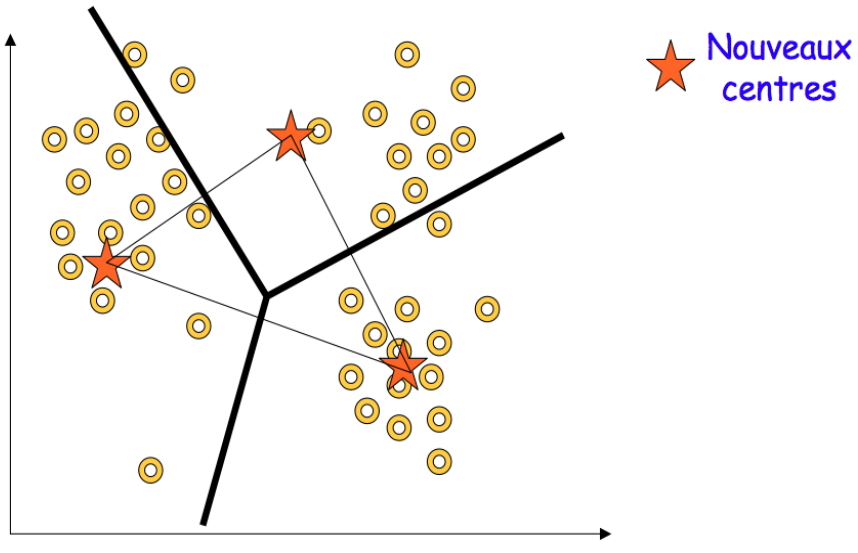

Figure 3: Deuxième illustration de l'algorithme de K-Means

Lorsque l'algorithme atteint un niveau optimal, il va sortir les centroïdes finaux et les groupes finalement formés.

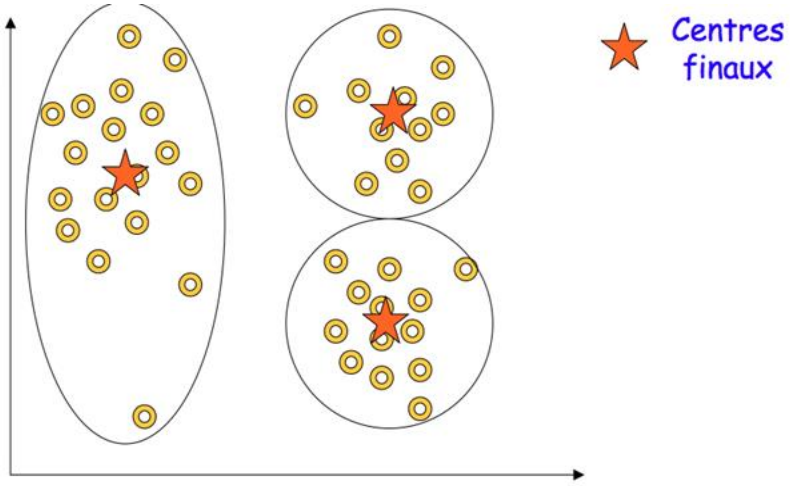

Figure 4: Troisième illustration de l'algorithme de K-Means

Exemples du tableau montrant les données de base pour l'exécution d'un algorithme de K-Moyennes

- 8 points A, …, H de l 'espace euclidéen 2D. k=2 (2 groupes)
- Tire aléatoirement 2 centres : B et D choisis.

points	Centre D(2,4), B(2,2)	Centre D(2,4), I(27/7,17/7)	Centre J(5/3,10/3), K(24/5,11/5)
A(1,3)	B	D	J
B(2,2)	B	I	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K

Tableau 2 : Exemple de la méthode de K-Moyennes

IV.3.6. Technique de règles d'association

L'essor technologique ne cesse de croître, la multitude de sources de données est de plus en plus diverse et variée. La représentation et la présentation de l'information deviennent encore plus abstraites. La nécessité de se munir d'outils d'analyse et d'extraction de ces colossaux recueils de données devient plus que vitale. L'objectif est de découvrir des associations ou des corrélations intéressantes entre des éléments dans ces grandes collections et bases de données. Ces éléments peuvent être des : attributs, objets, individus, Items... etc. Prenons par exemple la transaction effectuée par un ensemble de clients d'une grande surface commerciale. Un individu qui achète du

café, du sucre et du lait représente une règle d'association entre les attributs café, sucre et lait. Les règles d'association représentent un outil tangible, efficace et performant. Une règle d'association se présente de la forme X ⇒ Y, X en association avec Y ce qui veut dire que les transactions ou requêtes qui contiennent l'ensemble des objets X ont tendance à inclure les objets de l'ensemble Y. La recherche des règles d'association est un procédé important dans le Data Mining. Plusieurs algorithmes de recherche des règles d'association existent et permettent de découvrir des relations d'intérêt entre deux ou plusieurs variables stockées dans de très grandes bases de données. Nous avons par exemple les algorithmes Apriori, FP-growth, Eclat, GUHA, OPUS et Apriori. La plupart des algorithmes d'extraction des règles d'association mettent en œuvre deux propriétés le support et la confiance. On parlera de ces critères et des mesures, mais aussi de l'algorithme d'exploration des données permettant d'extraire les règles d'association ultérieurement. Dans ce chapitre on présentera un état de l'art sur les règles d'association, ainsi que les techniques d'extraction, les différents algorithmes et leurs variantes.

Les données issues des différents documents et bases de données transactionnelles peuvent être représentées sous la forme d'une matrice booléenne à deux dimensions. Dans une telle base, chaque tuple représente une transaction tandis que les différents champs correspondent aux objets inclus dans la transaction. On note par Ne le nombre de transactions, par p le nombre d'articles, par 0 l'évènement d'absence de chaque article et par 1 sa présence dans la transaction. De ce fait on construira une matrice binaire de la base de données.

Tableau 3 : Représentation binaire d'une base de données

Transactions	Article1	Article2	Article3	Article4	Article5	Article6	Article7
T1	0	1	1	1	0	1	1
T2	0	0	0	1	1	1	0
T3	1	0	1	0	0	0	1
T4	0	0	0	0	1	1	1
T5	0	1	1	1	0	0	0

Dans ce tableau, on retrouve la représentation binaire d'une base de données. Par exemple la transaction T1, contient les articles 2, 3, 4, 6, 7.

Ce tableau représente une matrice creuse a deux dimensions Ne* P. Avec Ne = 5 (Le nombre de transactions), et P = 7. (Le nombre d'items)

Une transaction T représente un sous-ensemble E. Exemple : soit un ensemble

E = {élément 1, élément 2, …... élément n}.

On aura T= {élément 1, élément 2}.

Notre exemple (Tableau 3. 2) donne ceci :

T1 = {Article2, Article3, Article4, Article6, Article7}

Un Item est un objet, élément ou un article d'une base de données.

Exemple 1 : Article1 représente un item.

Exemple 2 : Article3 représente un item.

Un $Itemset$ est un ensemble d'items, d'objets ou d'articles d'une base de données.

Exemple : {item2, item3, item4, item6}

Un $K - Itemset$ est un ensemble de k éléments, ou k-Items, il est aussi un $Itemset$.

Exemple 1 : {item2, item3, item4, item6} représente un-4-Itemset.

Exemple 2 : {item2, item4, item6} représente un 3-Itemset.

Le support d'un $Itemset$ représente le nombre total des transactions d'une base de données comportant cet $Itemset$ divisé par le nombre total des observations de cette base de données.

[54]. Par exemple, soit une base de données D et soit X un Itemset de n éléments. Dans une base de données transactionnelle D, le support de l'itemset X est le nombre de transactions dans D incluant X, divisé par le nombre total des transactions de D (figure 3.1)

Exemple 1 : Soit X un Itemset, avec X= {Article2, Article3}.

Soit D la base des transactions présentées précédemment dans le tableau 3.1.

Card(X) est le nombre de transactions dans D, de tel que les Items Article2 et Article3 apparaissent simultanément dans chacune de ces transactions de D. Il est égal à 2 Card(D) est le nombre total des transactions. Il est égal à cinq.

$$\text{Alors Support } (X) = 2/5.$$

On dit qu'un Itemset X est un Itemset fréquent si et seulement si le support associé à cet Itemset est supérieur à un support minimum défini par l'utilisateur

Une règle d'association est une application de la forme X ⟹ Y, qui exprime une corrélation de cooccurrence. Il existe deux mesures importantes, le support et la confiance, la robustesse d'une règle d'association est déterminée grâce à ces deux métriques [58]. Une règle d'association qui a un support faible va être observée rarement. La confiance mesure la pertinence de l'inférence dans une règle, par exemple plus grande est la mesure de confiance de la règle X ⟹ Y, plus cette règle sera pertinente.

Le support d'une règle d'association s'exprime par le nombre de transactions qui contiennent les éléments de X et les éléments de Y divisé par le nombre total des transactions de la base des transactions. Dans une base de données D, le support d'une règle d'association $X \Rightarrow Y$ est le nombre de transactions qui contiennent X et Y divisé par le nombre total des transactions

$$\boldsymbol{Support(X \Rightarrow Y) = \frac{card(X \cup Y)}{card(D)}} \qquad (6.1)$$

Exemple, la règle d'association « Lait ⇒ Pain », littéralement, Si Lait Alors Pain. Le support représente le nombre de transactions dans lesquelles on trouve les Items Lait et Pain, divisé par le nombre total des transactions.

La confiance d'une règle d'association s'exprime par le nombre de transactions qui contiennent la relation d'union entre la transaction X et la transaction Y divisé par le nombre des transactions qui contiennent la transaction X.

$X \Rightarrow Y$ Représente une règle d'association.

$X \cup Y$ Représente l'ensemble union contenant les éléments de la transaction X et les éléments de la transaction Y.

La confiance d'une règle d'association est définie comme suit :

$$\boldsymbol{Confiance(X \Rightarrow Y)} = \frac{\boldsymbol{Support(X \cup Y)}}{\boldsymbol{Support(X)}} \qquad (6.2)$$

Exemple :

La confiance d'une règle d'association " Lait ⇒ Pain", est égale au support de la règle « Lait ⇒Pain » divisé par le support de l'Item « Lait ».

Une étape importante avant de démarrer un processus d'exploration de données en data mining est la phase de préparation de données. En effet, pour pouvoir exploiter ces données il est souvent nécessaire d'appliquer un processus de nettoyage, les informations sont souvent bruitées et incomplètes. Il est alors important d'en tenir compte car la qualité des résultats est affectée fortement par les données utilisées, si on utilise une donnée bruitée on aura des relations entre des données intéressantes et des données peu significatives.

Il existe plusieurs façons d'explorer les règles d'association, l'une de ces méthodes est la méthode naïve, on utilise alors toutes les combinaisons possibles des attributs et de leurs valeurs pour créer toutes les règles d'association possibles. Ce qui pose problème sur le plan complexité computationnelle du fait de l'explosion combinatoire. En effet, le nombre de règles générées est énorme. On peut optimiser cette méthode en gardant juste les règles avec un support et une confiance minimum. Cela reste insuffisant et les résultats sont insatisfaisants. L'algorithme Apriori représente une approche révolutionnaire dans l'apprentissage et l'exploration des règles d'association. Nous allons présenter deux différents algorithmes d'extraction des règles d'association : l'algorithme Apriori et l'algorithme Fp-Growth. Ces deux algorithmes représentent

une solution efficace. Le premier est très coûteux en termes d'accès à la base de transactions, le deuxième quant à lui optimise le coût d'accès.

L'algorithme Apriori créé par $Agrawal$ et $Srikant$ en 1994, procède en deux temps. Il est basé sur le principe lié à l'approche de support et de confiance. L'algorithme parcourt le treillis des $itemsets$ pour rechercher les $itemsets$ fréquents et en déduire les règles d'association dont la confiance dépasse le seuil de confiance min conf. Le treillis des $itemsets$ permet d'utiliser plus efficacement cet algorithme d'extraction en admettant les propriétés suivantes : Tout sous-ensemble d'un $Itemset$ fréquent est fréquent. Tout sur-ensemble d'un $itemset$ non fréquent est non fréquent.

Le nombre d'itemsets fréquents qui peuvent être générés de n items est de 2n , la génération des Itemsets fréquents est de complexité exponentielle, il est alors essentiel de trouver la méthode de recherche la plus optimale. Ces Itemsets représentent un treillis d'Itemsets représenté sous la forme d'un diagramme de Hasse présenté à la figure 13.

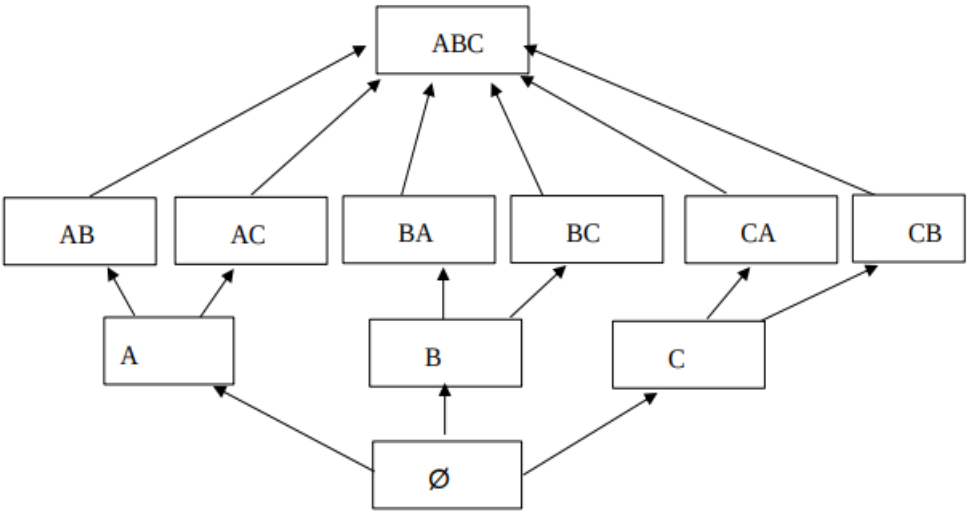

Figure5 : Exemple de treillis d'Itemsets ou diagramme de Hasse

D'une manière plus concise, le déploiement de l'algorithme Apriori se fait comme suit :

1) Générer les Règles candidates.
2) Calculer le support pour chaque règle candidate.
3) Apparier les règles dont on a calculé le $support$ avec le support choisi.
4) On rejette les candidats dont le $support$ est inférieur au $suppmin$.

On termine en sortie avec toutes les règles dont le support est supérieur au support minimal.

En résumé le déroulement se fait comme suit :

L'algorithme Apriori fonctionne en deux phases (voir la figure 3.5). La première consiste en la recherche des ensembles d'items fréquents notée EIF et la seconde utilise ces ensembles pour trouver les règles d'association dont la confiance est supérieure à un seuil prédéfini. Le processus de découverte des ensembles d'items fréquents est itératif, on commence par la construction de EIF avec un seul item, ensuite on réitère pour construire des EIF avec deux items. Ceci va déterminer la taille de chaque ensemble d'items fréquents qui sera noté Ln. L'ensemble fréquent avec un seul item fréquent sera noté L1

Les ensembles d'items candidats sont construits à partir des ensembles d'items fréquents de taille Ln-1 et seront notés Cn. Par exemple C2, l'ensemble d'items fréquents candidats de taille 2, sera produit par les ensembles fréquents de taille 1. Ln est obtenue en ne gardant que les éléments de Cn dont le support est supérieur au seuil. On continue les itérations jusqu'à ce que les Ln soient vides.

Apriori (T, t)

Calcul de L1

N ← 2

Tant que LN-1 ensemble vide

CN ← **Apriori** (LN-1)

Pour chaque transaction t ∈ T

Ct ← **Sous-ensemble** (LK, t)

Faire

Pour chaque candidate c ∈ C

Faire

$LN \leftarrow count \geq SuppMin$

$N \leftarrow N+1$

Retourner LN

Dans cet algorithme les ensembles Ln et Cn sont des enregistrements dans lesquels on stocke les informations et les valeurs des variables. La procédure count permet de calculer et de stocker la fréquence de chaque item de la base de données.

Exemple : l'utilisateur définit les seuils minimaux de support et confiance, on aura min_{supp} =0,30 et min_{conf} =0,30.

Soit la base de données D suivante :

Tableau 4 : Algorithme Apriori Etape 0

$Transitions$	$Items$
T1	1, 2 ,5
T2	2 ,4
T3	2,3
T4	1,2,4
T5	1,2,3
T6	2,3,5
T7	1,3
T8	1,2,1 3,5
T9	1,2,3
T10	2,3

Dans un premier temps on calcule la $fréquence$ d'apparitions de chaque $item$.

Tableau des $fréquences$ d'apparitions de chaque $Item$:

Tableau 5 : Algorithme Apriori Etape 1

Transitions	Fréquence
1	6
2	9
3	7
4	2
5	3

Calcul du $support$ de chaque $Item$:

Tableau 6 : Algorithme Apriori Etape 2

Transitions	Fréquence
1	6
2	9
3	7
5	3

Remarque, l'Item 4 est supprimé, car son support $Support$ (4) = 0,20 est inférieur à $MinSupp$ = 0,30. Ensuite, on calcule les Li et Ci, comme présenté dans l'algorithme Apriori de la Figure 3.5.

Génération des candidats C2 :

Tableau 7 : Algorithme Apriori Etape 3

$Itemset$	$Fréquence$
1,2	5
1,3	4
1,5	2
2,3	6
2,5	3
3,5	2

Génération des $Itemsets$ fréquents $L2$:

Tableau 8 : Algorithme Apriori Etape 4

$Itemset$	$Fréquence$
1,2	5
1,3	4
2,3	6
2,5	3

Génération des candidats $C3$:

Tableau 9 : Algorithme Apriori Etape 5

$Itemset$	$Fréquence$
1,2,3	5
1,2,5	4
2,3,5	2

Génération des $Itemsets$ fréquents $L3$:

Tableau 10 : Algorithme Apriori Etape 6

$Itemset$	$Fréquence$
1,2,3	5

L'algorithme ne génère plus de candidats. De ce fait il s'arrête. Le dernier $itemset$ candidat contient un seul élément.

Les règles retenues par l'algorithme Apriori sont celles formées par les $Itemsets\ L2$ et $L3$.

Après déroulement on aura les règles d'association suivantes :

Les combinaisons formées des Itemsets de $L2$ donnent les règles :

Si 1 $Alors$ 2

Si 1 $Alors$ 3

Si 2 $Alors$ 3

Si 2 $Alors$ 5

Les combinaisons formées des $Itemsets$ de $L3$ donnent les règles :

Si 1, 2 Alors 3

Si 1, 3 Alors 2

Si 3, 2 Alors 1

Il existe une multitude d'avantages dans l'utilisation de l'algorithme Apriori. On en énumère quelques-uns [72,73] : La découverte rapide de règles d'association pertinentes entre objets. La facilité d'interprétation des résultats lors de l'extraction des règles d'association, malgré le nombre important de ces dernières. Comme inconvénients auxquels on fait face lors d'une utilisation de l'algorithme Apriori, les algorithmes d'extraction liés à l'approche $\boldsymbol{support/confiance}$ génèrent un grand nombre de règles d'association. Un nombre important de configurations d'items ne peuvent pas engendrer de règles d'association. La recherche de règles d'association impose un temps considérable qui peut s'avérer désavantageux si l'on fait face à une énorme base de données.

CHAP V : PRÉSENTATION DU NOUVEAU SYSTÈME

V.1. Introduction

A cette étape, il correspond le développement de notre application ; donc la traduction en langage de programmation. Tout cela se fait selon les modèles du cycle de vie du logiciel dans le but de garantir l'obtention d'un produit de bonne qualité. Ces modèles de cycle de vie d'un logiciel font objet du génie logiciel. Ainsi, dans ce chapitre, nous allons présenter en bref quelques modèles du cycle de vie d'un logiciel ainsi que le fonctionnement de notre application avec les techniques du Data Mining.

V.2. Cycles de vie du développement d'un logiciel

La réalisation d'un projet informatique répondant à l'évolution du système d'information d'une organisation, repose sur les concepts fondamentaux du génie logiciel soit la production d'application réellement adaptée aux besoins des utilisateurs, la réduction des coûts de production et de maintenance, l'augmentation de la fiabilité, de la performance et de l'interopérabilité, l'optimisation de la gestion du temps de production du logiciel, l'augmentation du temps de vie du logiciel par l'intégration de l'adaptabilité et du paramétrage.

V.2.1. Etapes du cycle de vie d'un logiciel

L'étude préalable consiste à identifier les besoins, en en élaborant une version de base du cahier des charges. Pour rédiger ce document, le chef de service va mener une étude de l'existant du système d'information dont une enquête précise auprès des utilisateurs de la future application résultant du projet. A cette étape, le cahier de charges intégrera la décision de faisabilité, ainsi qu'un plan général du projet, et une estimation approchée du coût et des délais.

La spécification a pour but de la description du modèle fonctionnel de l'évolution du système d'information, l'estimation des ressources nécessaires, et une première version de la planification de la phase de développement. La démarche de spécification s'appuie sur les collectes d'informations, obtenues lors de l'étude préalable ; avec pour but final de définir un cahier des charges détaillées qui servira de contrats entre le maître d'œuvre et le maître d'ouvrage. La conception générale doit définir la description de l'architecture du logiciel, correspondant au domaine d'étude du système d'information, en un ensemble de modules fonctionnels, de maquettes, et de structures de données. On définira pour chaque module son rôle et son interaction avec les

autres modules. On mettra en œuvre un analyse conceptuelle et organisationnelle de façon à effectuer les choix de l'infrastructure technologique et des outils et méthodes de développement.
La conception détaillée consiste à définir l'organisation humaine et technique ; ainsi qu'à répartir les modules à des équipes de développement du maître d'œuvre, qui mettront en œuvre des techniques et éventuellement des outils logiciels pour préparer l'approche technologique du développement.
La codification correspond à l'étape de l'implémentation des modules constituant le projet par les équipes de développement du maître d'œuvre. Les équipes de développements seront amenées à utiliser des outils de développement leurs permettant de développer rapidement et à moindre coût. Cette étape comprend la rédaction de la documentation technique.
Les tests unitaires consistent à vérifier le bon fonctionnement individuel des modules du logiciel, ceci à partir de jeux d'essai correspondant à un sous-ensemble représentatif d'informations du domaine d'étude concernée par le projet d'évaluation du système d'information.
Les tests systèmes correspondent à la phase de tests de différents modules en interaction les uns avec les autres, et permettent ainsi de déterminer que le logiciel correspond aux besoins exprimés par le cahier de charges. Cette étape correspond aussi à la rédaction de la documentation utilisateur.
Le déploiement et l'exploitation doivent permettre au maître d'œuvre d'installer le logiciel au sein de la composante technologique du système d'information du maître d'ouvrage ; ainsi que d'en assurer la formation aux utilisateurs, ceux-ci pourront alors en assurer l'exploitation dans le cadre de leur activité dans l'organisation.
La maintenance est une étape particulière, qui doit permettre la correction, mais aussi l'évolution du logiciel en fonction des contraintes rencontrées par les acteurs du système d'information. La maintenance donc intervient à des différentes étapes du cycle de vie du logiciel. Il existe plusieurs types de maintenance tels que la maintenance corrective qui permet de corriger les erreurs qui n'ont pas été détectées lors des précédentes phases de tests, la maintenance adaptative qui doit permettre l'adaptation et l'évolution du logiciel à l'apparition et la maintenance perfective qui a pour objectif l'optimisation des performances du logiciel. [11]

V.2.2. Types d'utilisation du cycle de vie d'un logiciel

1) Cycle de vie linéaire

Le cycle de vie linéaire consiste en la gestion du projet par étape, de la plus abstraite vers la plus concrète. A la validation d'une étape, on enchaine la suivante. L'application de ce modèle est relativement simple à mettre en œuvre et permet une démarche rigoureuse, qui a été utilisée dans

de nombreuses méthodes. Cette technique correspond toute fois à des méthodes nécessitant un temps de développement relativement important entre le début du projet et sa réalisation.

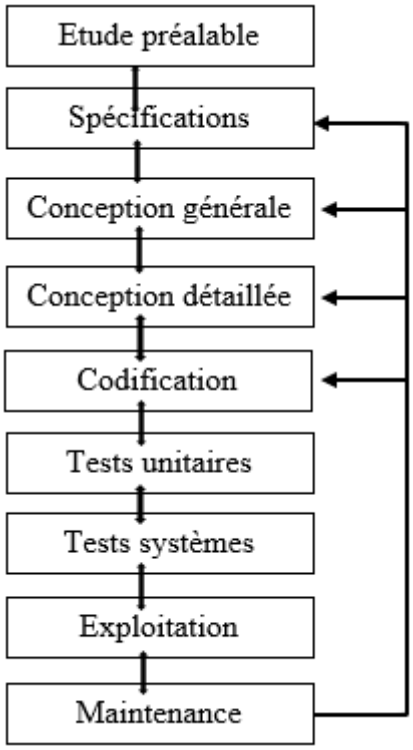

Figure 6 : Modèle de cycle de vie linéaire

2) Cycle de vie en V

Le cycle de vie en V conserve l'approche du modèle linéaire, en y ajoutant une réactivité par niveau (représenté par les flèches en pointillés). L'objectif est de permettre d'établir à chaque niveau des phases de validation conçues au départ du projet, qui vont fiabiliser le développement en ciblant précisément les problèmes qui pourront alors être corrigés.

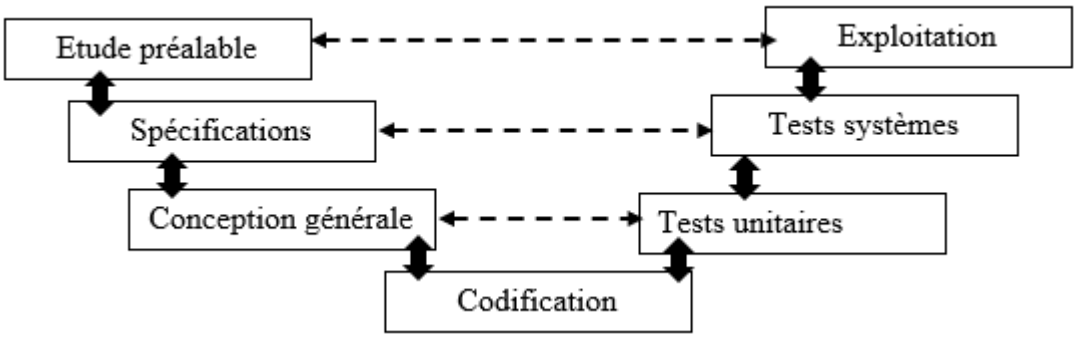

Figure 7 : Modèle de cycle de vie en V

3) Cycle de vie en spirale

Le cycle de vie en spirale consiste en un processus itératif de développement du projet, à partir des phases successives : spécification, conception, codification, tests de validation, qui vont s'enchaîner à un nouveau cycle reprenant par un complément de la phase des spécifications et qui détermine un nouveau cycle. Cette approche peut être particulièrement répondante en particulier par la validation successive de chaque cycle. Elle peut toutefois être utilisée dans le cadre de développement rapide d'application, où l'on va pouvoir livrer en exploitation aux utilisateurs une application développée progressivement module par module, qui viendront constituer au fur et à mesure le logiciel.

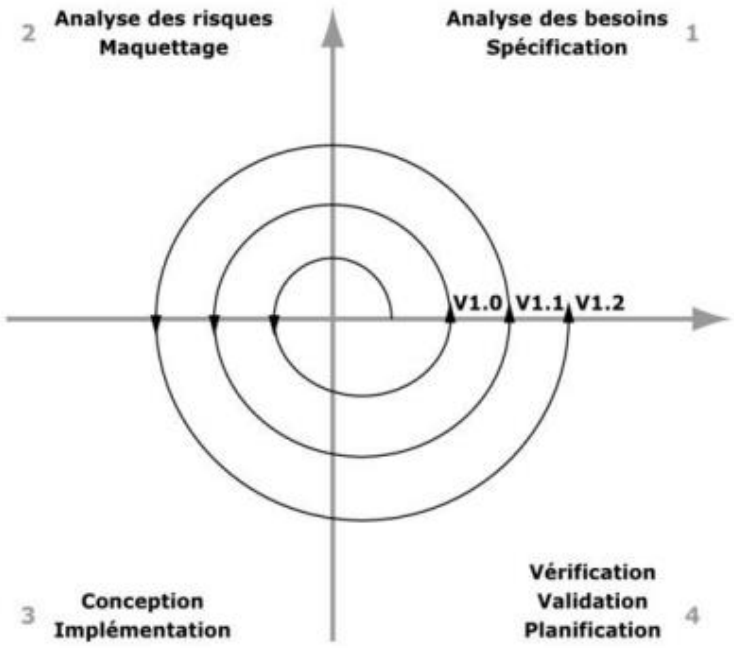

Figure 8 : Modèle de cycle de vie en spirale

V.3. Technique de description des données

Sur cette étape, après avoir importé les données sur lesquelles nous travaillons sur, nous cliquons sur le bouton « ANALYSER » pour afficher les données en composantes principales (Composantes, Variance expliquée, Rapport de variance expliquée).

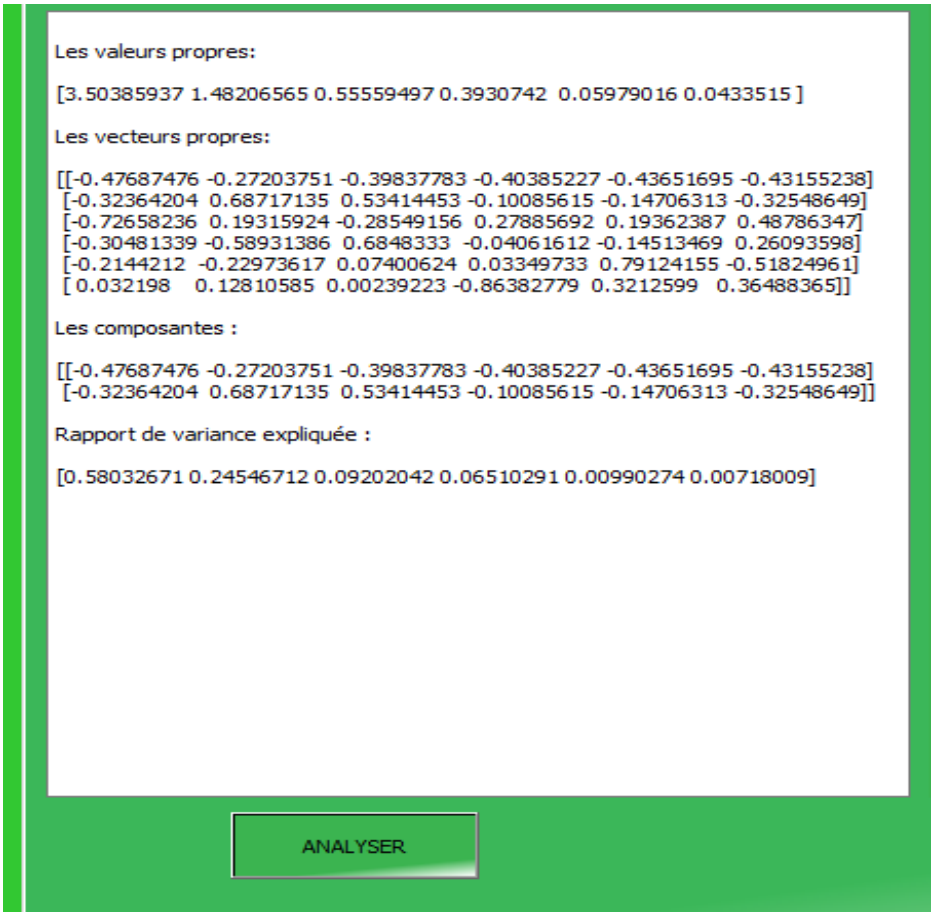

Figure 9 : Représentation des composantes principales

Lorsqu'on clique sur les boutons Variance Ration une nouvelle fenêtre s'ouvre pour afficher variance expliquée ou variance ration (figure 7).

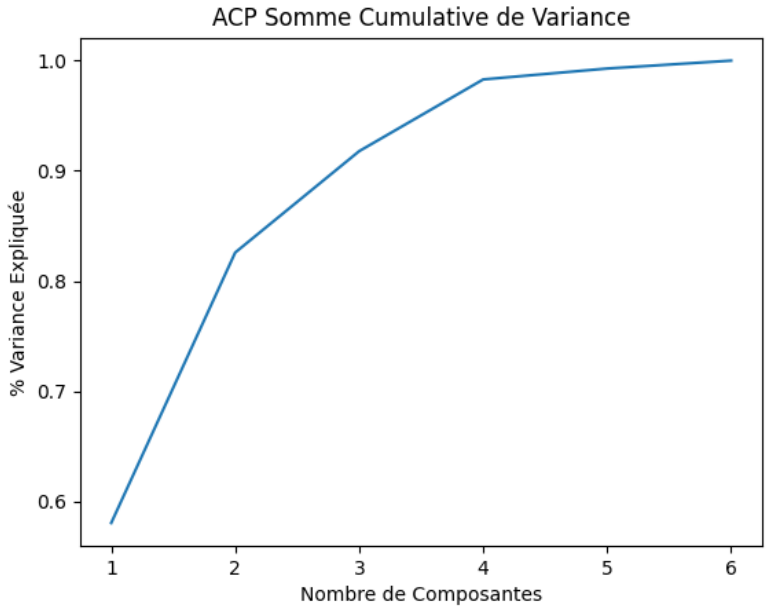

Figure 10 : Représentation de la variance expliquée

Lorsqu'on clique sur le bouton « MATRICE DE COVARIENCES » la fenêtre montrant la matrice des covariances semblable à la figure en dessous va s'ouvrir. Sur cette matrice on trouve toujours un diagonale des unités. Ici avec un langage python, nous avons trouvé des valeurs avec deux chiffres après la virgule (figure 8).

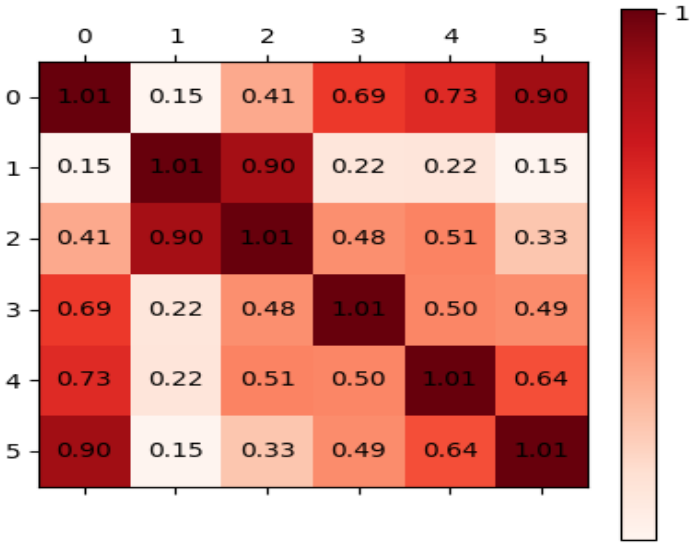

Figure 11 : Représentation de la matrice des covariances ou corrélations

Et en fin, lorsqu'on clique sur le bouton **ACP**, il s'affiche les données réduites avec une représentation bidirectionnelle (figure 9). Parmi ces deux figures, la figure à gauche montre les données originales avant la réduction. Celle à droite montre les données après la réduction.

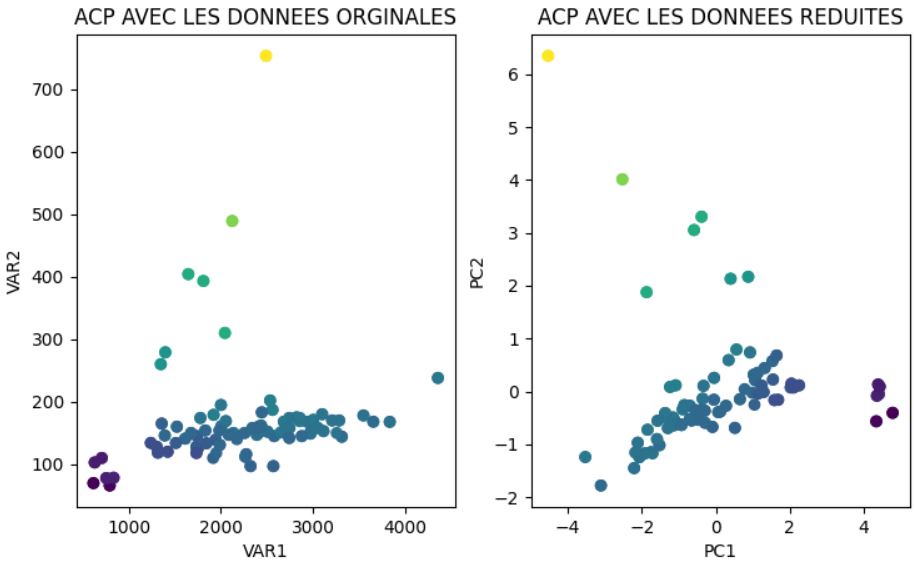

Figure 12 : Représentation des données réduites

Il s'agit de la même chose pour les autres techniques (classification, règles d'association, segmentions ainsi que la prédiction). Il s'agit d'importer les données à traiter et cliquer sur les boutons en suivant les indications des noms des boutons.

V.4. Méthode de maximisation optimale de la production

Sur cette méthode, l'utilisateur donne un modèle mathématique du problème de maximisation sous forme du système d'inéquations.

Donc, ici sur la figure suivante (figure10), il va entrer le nombre de variables contraintes trouvé ainsi que le nombre d'équations. Lorsqu'il clique sur le bouton « VALIDER »**,** une nouvelle fenêtre s'ouvre pour entrer les constantes des inéquations et de la fonction objective et puis clique sur le bouton « VALIDER » pour valider ce formulaire. La fenêtre se ferme et on clique sur le bouton « MAXIMIZER » pour afficher la solution optimale.

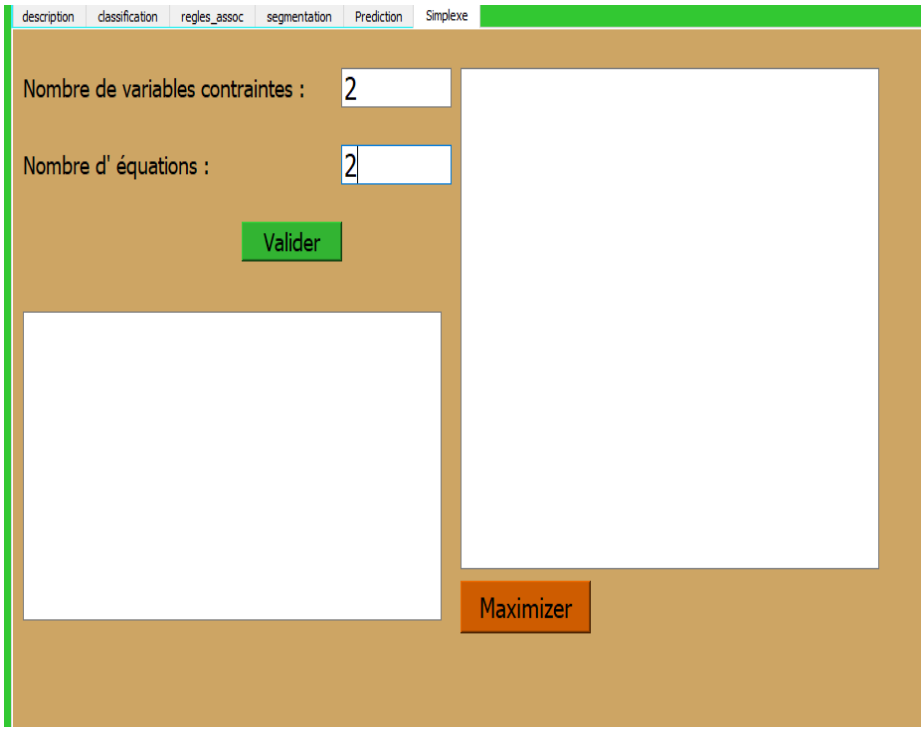

Figure 13 : Entrer le nombre de variables et le nombre d'équations

Après avoir entré le nombre de variables contraintes et le nombre d'équations, on clique sur le bouton Valider. Une nouvelle fenêtre s'ouvre (figure 11). Cette fenêtre nous permet d'entrer les constantes des variables contraintes ainsi que les constantes des variables d'une fonction objective. Et puis on clique sur le bouton **Valider**. La fenêtre disparaît et on retrouve sur l'ancienne fenêtre.

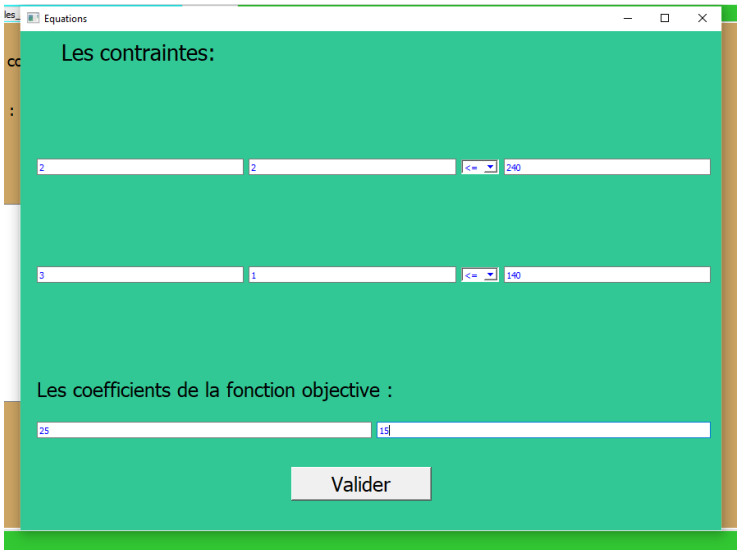

Figure 14 : Nouvelle fenêtre pour entrer les constantes du système

Après avoir Valider les données entrées sur la figure 11, On se retrouve sur la précédente fenêtre. Une fois-là, on clique sur le bouton **Maximizer.** C'est là où s'affiche les résultats (Les valeurs des variables contraintes ainsi que la valeur de la solution optimale) sur la figure 12.

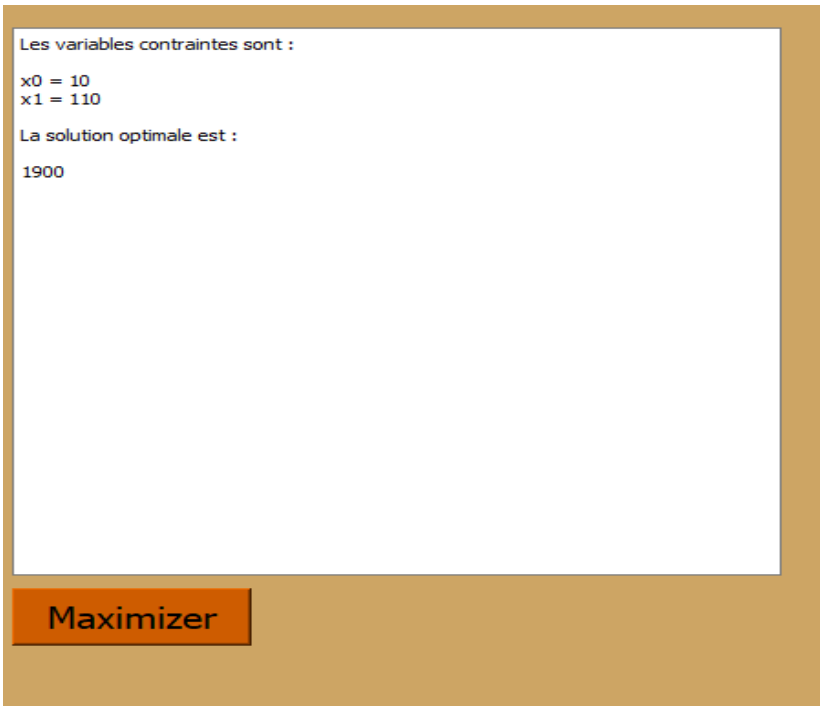

Figure 15 : Affichage de la solution optimale

V.5. Prédiction du chiffre d'affaires

Cette méthode permet à un utilisateur de prédire les données construites sous formes d'une série temporelle. Pour montrer pratiquement les étapes de cette méthode, nous allons nous référencer sur l'exemple donné dans le chapitre quatre. Après avoir ouvrir l'application, l'utilisateur va cliquer sur l'onglet « Prédiction ». Une fois là-bas, voici comment ressemble l'interface utilisateur sur la figure 16.

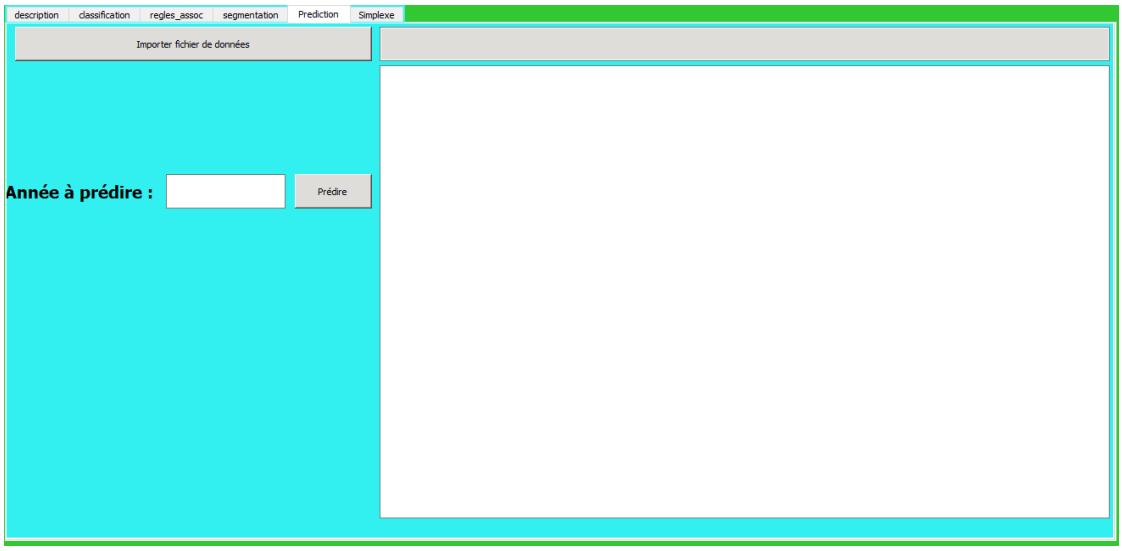

Figure 16 : Interface pour la prédiction

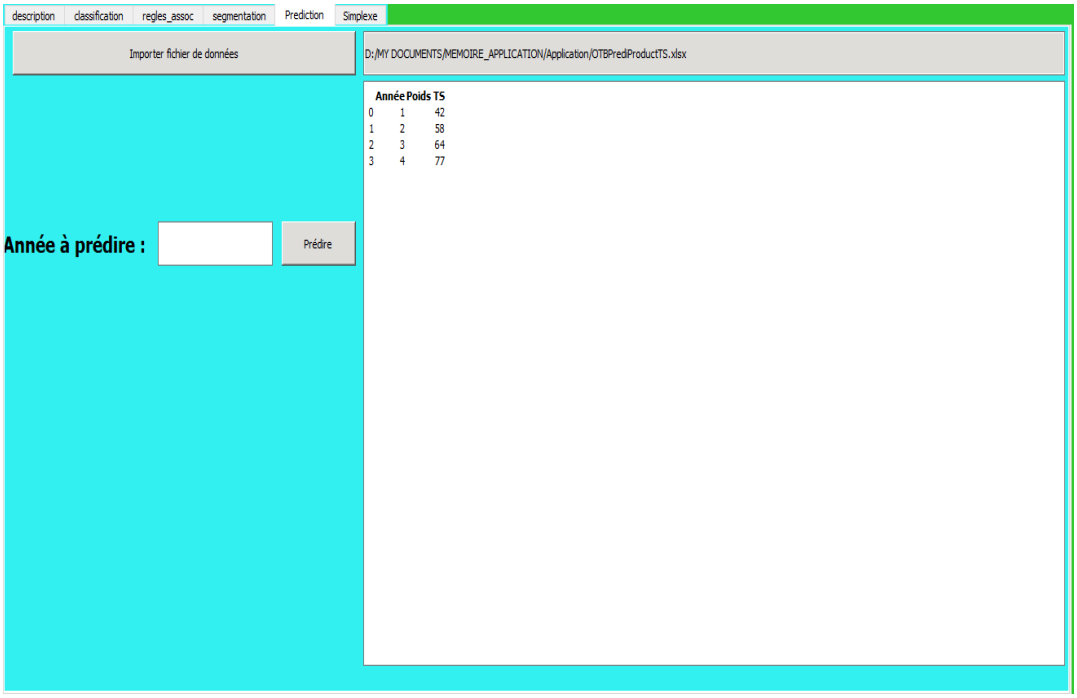

Figure 17: Représentation des données pour la prédiction

L'utilisateur clique sur le bouton « Importer fichier de données ». Voici comment ressemble l'interface avec l'affichage des données, utilisé dans l'exemple dans le chapitre quatre de notre projet, sous forme d'une série temporelle sur la figure 17.

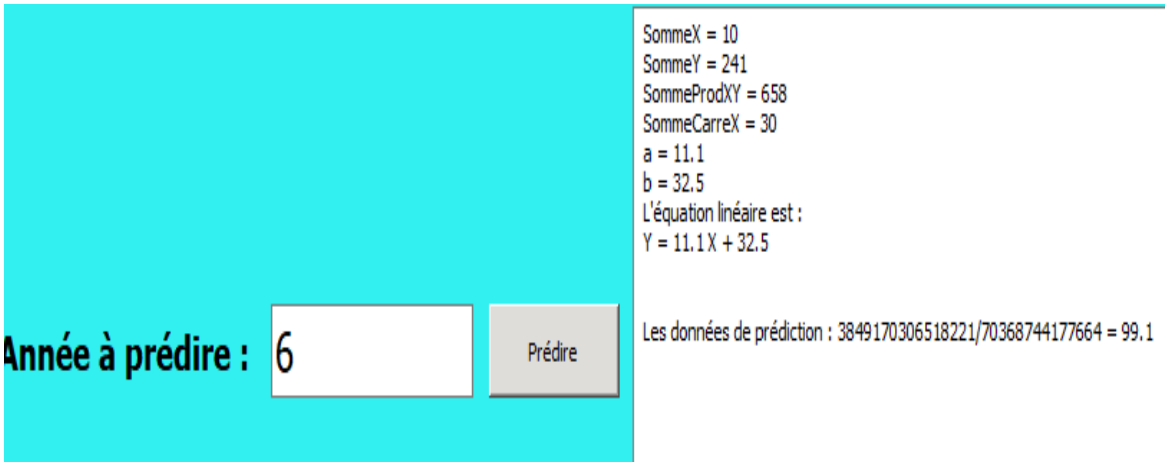

Figure 18: Interface d'affichage du résultat de prédiction

Avec une prédiction après deux ans d'expérience sur la figure 18, la solution obtenue ressemble à celle trouvée dans le chapitre quatre sur l'exemple donnée.

Dans ce chapitre, nous venons de montrer l'implémentation de notre système. Le système consiste répondre à des problèmes liés à la réduction d'un grand volume de données (figure 9) en données facile à analyser manuellement, à l'optimisation par recherche de la solution optimale, ainsi qu'à la prédiction pour projeter dans le future avec estimation. Pour tous les cas, l'utilisateur entre les données à traiter et le système fait sortir les résultats visibles sur l'interface utilisateur. Pour l'optimisation et la prédiction, nous venons de voir que en prenant les exemples utilisés dans les chapitres précédents, les résultats obtenus se ressemblent. Tenant compte des autres tests faits sur les données trouvées sur le site officiel des scientistes en sciences des données [11], nous pouvons en conclure que le système fonctionne bien suivant les spécifications citées dans les chapitres précédents.

CONCLUSION GENERALE

Ce travail de mémoire, intitulé mise en place d'un système automatisé d'analyse et de traitement des données d'une entreprise, a porté sur l'optimisation d'un problème linéaire ainsi que les techniques de l'exploration des données (Data Mining en anglais).

Dans le premier chapitre nous avons développé la raison d'être du sujet de notre travail, objectifs du sujet, problématique, solutions proposées et apport scientifique. Dans le deuxième chapitre nous avons donné une présentation générale et un fonctionnement d'une entreprise. Nous avons vu les principales fonctions, les caractéristiques et les structures d'une entreprise ainsi que l'importance de l'informatique dans une entreprise. Le troisième chapitre est dédié à la programmation linéaire par la méthode du simplexe, nous avons vu comment modéliser un problème linéaire : un système d'inéquations pour les contraintes et une équation pour une fonction objective. Quant au quatrième chapitre nous nous sommes focalisés sur les techniques d'exploration des données permettant à la réduction d'un grand volume de données pour trouver un peu de données facile à traiter. Grace à la combinaison des différentes idées modélisées et techniques ; ça nous a permis l'implémentation d'un système qui a eu lieu au cinquième chapitre et enfin la conclusion générale au sixième chapitre.

Le présent système offre à des utilisateurs (entreprises ou organisations) un gain de temps dans la prise de décision. Il va également faciliter la tâche à un gestionnaire d'une entreprise à la détection facile d'anomalies ou des succès pour mieux rédiger des rapports. Il sera aussi un support aux futurs chercheurs d'approfondir leurs recherches afin d'améliorer la discipline Data Mining. Il nous a permis également d'approfondir nos connaissances théoriques et pratiques en rapport avec le traitement des données et les bases de données gigantesques et les techniques de programmation native. Nous avons eu l'occasion aussi de renforcer nos connaissances techniques notamment les langages Python, CSS et aussi savoir manipuler le système de gestion de base de données EXCEL avec le langage de programmation python.

RECOMMANDATIONS

Après avoir implémenté cette application d'analyse et de traitement des données, nous recommandons :

Aux futurs chercheurs d'améliorer le système afin que le traitement de données soit de plus en plus facile et de chercher des entreprises qui donnent accès à ses données pour adapter l'application.

Aux entrepreneurs, d'utiliser l'application pour ne pas se casser la tête dans la gestion et pour améliorer la qualité de leur production.

Au gouvernement principalement aux entreprises qui produisent et vendent, de mettre en place l'usage des technologies de l'information et de la communication et ainsi adopter l'usage de notre application dans le traitement de données qui diminuera les différents problèmes liés à la mauvaise gestion de leurs activités.

REFERENCES BIBLIOGRAPHIQUES

Ouvrages généraux

[1] BOUROCHE J.-M., SAPORTA G. (1980), L'analyse des données, Paris, Presses ;

[2] Universitaires CASIN PH. (1999), Analyse des données et des panels de données, Bruxelles et Paris, De Boeck. ;

[3] GOURIEROUX CH., MONTFORT A. (1995), « Séries temporelles et modèles dynamiques », 2è éd., Economica ;

[4] HARRIS Y. (1997), "Principal components analysis of cointegrated time series", Econometric Theory 13, p. 529-557;

[5] G. Saint-Cirgue, Apprendre le Machine Learning, machinelearnia,2019 ;

Webographie

[1] https:/ /www.studocu.com/fr/M%C3%A9thode_des_k_plus_proches_voisins consulté le 10/09/2022 à 09h10' ;

[2] https://www.python.org/ consulté le 30/09/2022 à 10h00' ;

[3] https://archive.ics.uci.edu/ml/machine-learning-databases/00292/ consulté le 30/09/2022 à 10h55' ;

[4]https://oraprdnt.uqtr.uquebec.ca/pls/public/docs/FWG/GSC/Publication/1645/34/1918/1/99811/8/F117728571_M_moire_ABDERRAOUF_NOUASRIA__version_finale_.pdf consulté le 01/10/2022 à 07h05' ;

[5]https://www.studocu.com/fr/document/universite-paul-valery-montpellier/statistiques-et-cao/fiche-formules-statistiques-descriptives-l1/1863004 consulté le 04/10/2022 à 08h15' ;

[6] https://pageperso.lis-lab.fr/~remi.eyraud/CAD/cours_4-clustering.pdf consulté le 10/10/2022 à 11h55' ;

[7]http://patrick.monassier.free.fr/cours_entreprise/entreprise/organisation_fonctionnemet_entreprises.pdf consulté le 02/11/2022 à 14h15' ;

[8] https://www.youtube.com/watch?v=INnDjQ0ik2s consulté le 15/11/2022 à 21h15' ;

[9] https://www.espacecommercial.fr/cours/mco3/prevision-ventes.php (consulté le 5/12/2022 à 22h00') ;

[10] https://monbtsmco.com/prevision-des-ventes/#methode-des-moindres-carres consulté le 07/12/2022 à 10h00' ;

[11] https://www.maxicours.com/se/cours/fonctionnalites-et-cycle-de-vie-de-l-application/ consulté le 07/12/2022 à 17h00' ;

[12] https://www.cairn.info/organisation-et-gestion-de-l-entreprise--9782100582808-page-45.htm consulté le 13/06/2023 08h15' ;

[13]https://www.cairn.info/management-des-entreprises--9782340058453-page-103.htm consulté le 15/06/2023 08h15' ;

[14] https://www.charpentes-gross.com/94/quelles-sont-les-caracteristiques-dune-entreprise-commerciale/ consulté le 16/06/2023 08h15'

[15] https://www.amc-models.com/30/quelles-sont-les-caracteristiques-dune-entreprise/ consulté le 16/06/2023 17h15' ;

[16] https://www.cairn.info/management-des-entreprises--9782100520497-page-89.htm consulté le 16/06/2023 19h15' ;

[17] https://www.cairn.info/la-boite-a-outils-de-la-strategie-3e-ed--9782100791651-page-112.htm consulté le 16/06/2023 19h15' ;

[18] https://www.kaggle.com/datasets consulté le 08/12/2022 à 13h00' ;