EPIGRAPHE

Celui qui n'a pas défini son objectif n'est pas près de l'atteindre.

DEDICACE

A mes très chers parents, Jean-Marie KALAMBAYI et Julie SAKADI ;
Qui, des manières constantes, se sont donnés beaucoup de peines, en me soutenant totalement. Qu'ils en soient remerciés et trouvent par ici la gratitude de mon cœur sincère.

Je dédie ce travail.

KALAMBAYI MULUMA-MUNTUNTU Jonathan

REMERCIEMENTS

A vous monsieur le professeur Pierre KAFUNDA, je vous remercie pour m'avoir assisté à la réalisation de ce travail malgré vos multiples occupations.

Nous tenons à remercier le corps académique de l'université Révérend KIM en général et en particulier la Faculté des informatiques de gestion pour le savoir et opportunités offertes au profit de notre formation.

Nous pensons modestement à toute la famille qui n'a pas arrêté de nous donner toute l'affection et soutien nécessaires : Me Benjamin KALAMAYI, Mme Miriam KALAMBAYI, Me Jojo KALAMBAYI, Me Daniel KALAMBAYI, Me Billy Paul KALAMBAYI, Mme Julie KALAMBAYI et Me Peter KALAMBAYI.

A mes cousins et cousines ; nièces et neveux ; tantes et oncles ainsi qu'à mes grands-parents. Très reconnaissant de votre soutien tant financier, moral que matériel.

Nos pensées à toutes les autres personnes qui nous sont d'une portée particulière : Mme Rebecca MUSHIYA, Me Elie KABASELE, Me Jérémie KINANO, et ceux qui ne sont pas cités en cet instant, qu'ils savent que nous gardons un très bon souvenir de leur apport.

Nous tenons particulièrement à remercier l'Ir Armel pour ses conseils et son aide tant précieuse qui nous a été d'un apport considérable.

A tous mes amis et connaissances, sans oubliés les compagnons de situations difficiles, pour de raison de modestie, je me réserve de citer les noms. Très reconnaissant de vos conseils, encouragement, et encadrement.

0. INTRODUCTION GENERALE

Le churn ou attrition de client est devenue aujourd'hui très importante avec l'accroissement de la concurrence et la diversité des offres sur le marché. En Europe, dans l'industrie des télécommunications, le départ d'un client coûte environ 500 euros. Le taux mensuel moyen de churn chez les opérateurs de téléphonie mobile varie entre 8 et 12% [1]. Le taux d'attrition annuel varie de 20% à 40 % chez la plupart des opérateurs de téléphonie mobile [2]. Dans le secteur bancaire, la compétition entre les organismes de crédit a augmenté avec l'accroissement du marché, ce qui a accentué le churn. De faibles taux sont proposés aux nouveaux clients au détriment des anciens, ce qui cause le départ de ces derniers chez de nouveaux organismes de prêt [3].

Dans un marché d'une telle compétitivité, une stratégie de marketing défensive revêt beaucoup d'importance. Au lieu de tenter d'acquérir de nouveaux clients ou d'attirer les abonnés loin de la concurrence, il est important de s'intéresse plutôt à la réduction des départs de ses clients [2], surtout qu'il est 5 fois plus coûteux d'acquérir un nouveau client que d'en garder un [4].

Dans un contexte économique difficile, où les organisations sont à la fois confrontées au départ de leurs clients ainsi qu'à celui de leurs employés, le défi majeur est de détecter les personnes ayant l'intention de les quitter afin d'anticiper leur départ et de les retenir via des actions adéquates.

Certes, il est arrivé alors la nécessité de fouiller, torturer les données des clients provenant des différentes sources de production pour en dégager les corrélations, relations entre les clients pour une prise de décision.

Face à de telles constations, il est évident de constituer un support d'aide à la décision pour s'imprégner de toutes les données de clients en se basant sur l'exploitation de bases de données évoluées à l'aide des techniques de datamining qui mettent en œuvre de puissants outils d'extraction des connaissances à partir des données.

0.1 PROBLEMATIQUE

La perte des clients d'une entreprise appelée communément « *attrition (churn)* » constitue un vrai problème pour les entreprises évoluant dans les différents secteurs d'activité surtout en situation de concurrence. Nul n'ignore que ce phénomène n'a pas épargné le secteur de la télécommunication. Orange RDC évoluant dans ce secteur est buttée aussi à ce phénomène pour ses abonnés, très surtout en situation de concurrence.

A cet effet, la prise de décision pour la Direction des Marketing pour la réduction de ce phénomène suscite trop d'interrogations pour l'éradiquer:

- Quels sont les abonnés fragiles au vu de leurs profils d'utilisation du réseau ?

- Sur quel facteur agir pour modifier les comportements des abonnés ?

- Quels sont les abonnés fidèles ?

0.2 HYPOTHESES

Ce travail s'inscrit dans le cadre de la fouille de données et des méthodes de traitement de l'information de l'entreprise. Basé sur des études récentes sur les comportements des abonnés afin de maitriser l'attrition de la clientèle.

Du fait que dans le secteur de télécommunication, les clients ne sont pas engagés contractuellement et peuvent cessez leur activité sans préavis. Afin d'estimer l'effort de la fidélisation qui peut être engagé au cas par cas, l'opérateur doit donc distinguer les clients fidèles et fragiles et sur quels facteurs ajuster pour modifier leurs comportements.

Pour y parvenir, nous pensons mettre sur pied un outil d'extraction de connaissance caché dans les données en utilisant les techniques de datamining qui met en œuvre des outils pointus permettant de maitriser ce phénomène c'est-à-dire les profiler afin

de dégager les tendances, relations inconnues a priori. *La méthode prédictive* répond à cette problématique d'exploitation de grande volumineuses de données.

0.3. CHOIX ET INTERET DU SUJET

Notre travail s'intitule ***"mise en place d'un outil d'extraction de connaissance basé sur les technique de data mining appliqué sur l'analyse de churn, étude menée au sein de Orange RDC***"

En effet, L'heure est la gestion de la relation avec les clients pour favoriser leur fidélisation à long terme. Les opérations de marketing étant très couteuses, les décideurs ont besoin d'avoir la clarté sur les abonnés afin de savoir sur quels facteurs agir pour les fidéliser. Combattre le coût élevé de la perte de la clientèle, il est possible d'employer des techniques de plus en plus sophistiquées pour analyser les raisons de la perte de la clientèle et quels clients sont les plus fragiles et fidèles. Ces informations peuvent être utilisées par les services de marketing d'une entreprise de télécommunication (notamment de Orange RDC) pour mieux cibler les campagnes de recrutement et permettre une surveillance active de la base d'appels des abonnés afin de repérer leurs comportements.

0.4. DELIMITATION DU SUJET

Nous nous sommes donné une mission de :

- Profiler les clients avec objectif d'avoir une idée sur leurs caractéristiques (comportements) afin de les cibler;

- Réduire le taux d'attrition au sein de l'entreprise dans une période de 6 mois ;

- Appliquer la classification automatique hiérarchique qui consiste à opérer un regroupement des abonnés par rapport à critère. Regrouper les abonnés par rapport à la zone géographique, nombre d'appels sortants, nombre d'appels entrants, nombre de messages sortants et entrants... afin d'avoir une idée sur le facteur à agir pour les fidéliser.

0.5 METHODES ET TECHNIQUES UTILISEES

Dans le cadre de ce travail, nous avons utilisé les méthodes suivantes :

- **Technique documentaire :** Elle nous a permis d'élaborer notre approche théorique en consultant les ouvrages, les mémoires, les travaux de fin de cycle et les notes de cours qui cadrent avec notre sujet.
- **Technique d'interview :** Cette technique nous a permis d'obtenir les informations fiables auprès des personnes et agents qui travaillent dans les entreprises commerciales.

- **Méthode historique :** cette méthode nous a permis à connaitre les activités des années les activités des clients.

- **Méthode statistique** : elle nous a aidés à réaliser divers calculs (classification des différentes variables comparatives d'une période à l'autre) de toutes les activités de vente possibles. cette l'application de cette méthode a été possible grâce à quelques techniques dont nous avons fait usage :

- **Internet :** C'est une bibliothèque universelle, elle nous a procuré des informations nécessaires à la réalisation du présent travail.

0.6 CANEVAS DU TRAVAIL

Hormis l'introduction générale et la conclusion générale, notre mémoire comporte quatre chapitres, à savoir :

- Chapitre I : Généralités sur le système décisionnel;

- Chapitre II : Les techniques de datamining ;

- Chapitre III : Problématique de churn ;

- Chapitre IV : Application.

CHAPITRE I : LES SYSTEMES DECISIONNELS

I.1. INTRODUCTION

Dans l'entreprise le système d'information (SI) a pour objectif de faciliter l'établissement et la mise en œuvre de la stratégie, en particulier de concrètement supporter la réalisation des activités. Il est construit à partir des exigences des métiers, des processus définis par l'entreprise.
Pour des raisons techniques, les systèmes d'information de gestion ont été historiquement structurés en deux sous-systèmes : l'un dit opérationnel qui prend en charge la réalisation des opérations au jour le jour, et l'autre dit décisionnel qui fournit des informations pour définir la stratégie, piloter les opérations et analyser les résultats.

Un système décisionnel est donc avant tout un moyen qui a pour but de faciliter la définition et la mise en œuvre de stratégies gagnantes. Mais il ne s'agit pas de définir une stratégie une fois pour toute, mais d'être à même continuellement de s'adapter à son environnement, et de le faire plus vite que ses concurrents. Pour cela il convient de bien comprendre son environnement, d'ajuster ses interactions avec lui en faisant les meilleurs choix de cibles d'actions[1]. Concrètement le chemin à suivre peut être caractérisé par les quatre objectifs suivants : comprendre son environnement, se focaliser sur des cibles, aligner son organisation et mettre en œuvre les plans d'actions nécessaires.

Un système décisionnel va en particulier aider au pilotage des plans d'actions (prévision, planification, suivi), à l'apprentissage (acquisition de savoir-faire, de connaissances, de compétences) et à la réalisation d'innovations incrémentales. Les systèmes décisionnels traditionnels permettent de faire l'analyse des activités déjà réalisées et d'en tirer des enseignements pour les activités futures, pour cela ils utilisent des données plus ou moins récentes (au mieux mises à jour quotidiennement). Les systèmes décisionnels plus avancés gèrent des données plus fraîches (certaines sont mises à jour en quasi temps réel),

[1] MICHEL BRULEY, système d'information décisionnel : à quoi cela sert-il ? 2014 Page 1

automatisent des décisions et supportent en temps réel des opérations (centre d'appels, web par exemple)[2].

I.2. DEFINITION ET ORIGINE DES SYSTEMES D'AIDE A LA DECISION (SAD)

I.2.1 Définition

Le concept de système d'aide à la décision(SAD) a été définit de plusieurs façon selon différents auteurs, notamment :

- A. Gorry et M. Scott Morton (1971), l'ont défini de manière formelle comme suit : *« système informatisé interactif aidant le décideur à manipuler des données et des modèles pour résoudre des problèmes mal structurés ».*
- P. Keen et M. Scott Morton (1978) ont proposé la définition suivante : *«Les SAD réunissent les ressources intellectuelles des individus avec les potentialités des ordinateurs dans le but d'améliorer les décisions prises ».*

A la lumière de ce qui précède nous pouvons définir le système d'aide à la décision comme un ensemble des moyens et techniques permettant au décideur de prendre une décision stratégique. Une décision stratégique étant une décision entreprise par les décideurs de l'organisation et qui vise à améliorer, qualitativement ou quantitativement les performances de l'entreprise afin de garantir sa croissance.

En d'autres termes un système décisionnel représente un ensemble de moyens, d'outils et de méthodes permettant de collecter, consolider, modéliser et de restituer les données de l'entreprise dans le but d'apporter une aide à la prise de décision. Dans la vue globale du système décisionnel il y a cinq questions importantes auxquelles il faudrait chaque fois trouver solution :

1. Que s'est-il passé ?
2. Pour quoi ?
3. Que va-t-il se passer ?
4. Qu'est-il en train de se passer ?

[2] MICHEL BRULEY, système d'information décisionnel : à quoi cela sert-il ? 2014 Page 2

5. Que faire ?

I.2.2. Origine

Le concept de **S**ystèmes d'**A**ide à la **D**écision (**SAD**), en Sciences de Gestion, a initialement été défini de manière formelle par A. Gorry et M. Scott Morton (1971). Leur démarche de raisonnement a été d'intégrer les deux taxinomies suivantes :

- Types d'activités de management décrivant les activités de management sous la forme de trois niveaux : stratégique, intermédiaire et opérationnel ;
- Types de décision proposant, premièrement, une analyse des problèmes sous l'angle de la possibilité de les formaliser ou non (du programmable ou non programmable) et deuxièmement, un modèle décrivant le processus de prise de décision individuel (le modèle « *Intelligence – Design – Choice* »).

Sur ces bases, A. Gorry et M. Scott Morton ont défini les SAD de la manière suivante : « *système informatisé interactif aidant le décideur à manipuler des données et des modèles pour résoudre des problèmes mal structurés* ». Dès lors, de nombreux chercheurs ont décrit et proposé des angles d'étude du concept de SAD. La fonction d'aide à la décision a été intégrée dans le concept plus vaste de management des systèmes d'information.

Puis, P. Keen et M. Scott Morton (1978) ont pris en compte la dimension cognitive du décideur en proposant alors la définition suivante : « *Les SAD réunissent les ressources intellectuelles des individus avec les potentialités des ordinateurs dans le but d'améliorer les décisions prises* ». S. Alter (1977) a proposé une taxinomie des différents types de SAD selon leurs fonctionnalités. Notamment, il a mis en avant deux grands types de SAD (orientés modèle et orienté données) comme l'indique le schéma suivant:

La taxinomie des SAD de S. Alter

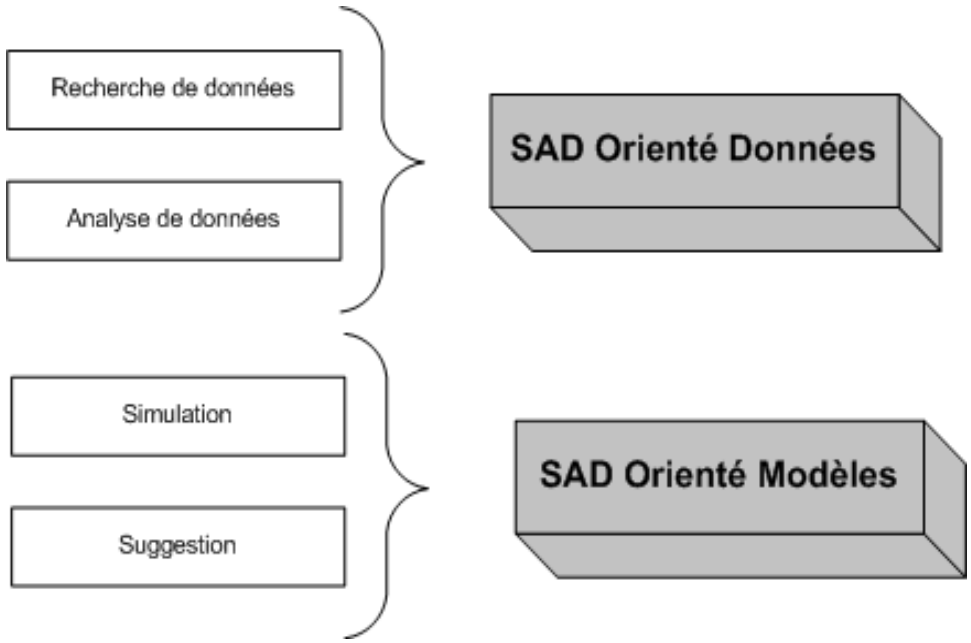


Figure I.1 : la taxinomie des SAD de S. Alter

De nombreuses autres recherches ont conduit à façonner le concept de SAD. Cependant, il s'agissait toujours d'aide à des décisions individuelles. Un nouveau champ d'étude a, alors, été envisagé avec la prise en compte des décisions de groupe et donc la création des **S**ystèmes d'**A**ide à la **D**écision de **G**roupe (**SADG**). G. De Sanctis et R. Gallupe (1987) ont posé les bases de systèmes concernant les décisions de groupe, c'est-à-dire, les décisions dans lesquelles la responsabilité est partagée entre plusieurs membres. L'objectif demeure l'amélioration de la qualité des décisions prises par le groupe. Dans les SADG, une attention particulière est portée sur les relations de communication entre les décideurs et, l'objectif fondamental des SADG réside dans l'assistance à l'échange d'idées, d'opinions et de préférences dans un groupe. Ainsi, l'aide apporté aux décideurs par les SADG est double : poser les bases de la communication entre décideurs (SADG de niveau 1) et fournir des modèles décisionnels (SADG de niveau 2).

En résumé, les recherches sur les SAD ont débuté au début des années 1970 et ont conduit à envisager la manière avec laquelle les technologies peuvent assister un décideur, ou un

groupe de décideur, afin d'améliorer la qualité des décisions prises dans une organisation[3].

I.3. L'ENTREPRISE

Une entreprise est une organisation dotée d'une mission et d'un objectif métier. Elle doit gérer sa raison d'être et/ou sa pérennité au travers de différents objectifs (sécurité, développement, rentabilité). Par voie de conséquence, cette organisation humaine est dotée d'un centre de décision.

I.3.1 La Problématique des Entreprises

Une entreprise se doit en permanence de pouvoir se situer par rapport à la concurrence, mais également par rapport à la demande et à ce qu'elle peut offrir. C'est sur ces points qu'un système décisionnel intervient.

I.3.2. Le Rôle du Décideur

Le décideur peut-être le responsable de l'entreprise, d'une fonction ou d'un secteur. Il est donc celui qui engage la pérennité ou la raison d'être de l'entreprise.

Pour ces raisons, il doit s'entourer de différents moyens lui permettant une prise de décision la plus pertinente. Parmi ces moyens, nous avons le système décisionnel. En effet, les systèmes décisionnels contiennent les données de toute l'activité de l'entreprise, mais la difficulté principale réside dans l'exploitation de ces informations. Pour cela, il est primordial de bien penser aux techniques de datamining.

En fait, un système décisionnel bien conçu doit donc :

- Fournir un accès à des données fiables.
- Aider l'utilisateur à comprendre ces données.
- Faciliter la prise de décision. Connaître la signification d'une information c'est bien. Savoir quoi en faire c'est mieux.
- Aider à la diffusion de l'information et à la mise en œuvre des actions.

[3] LEBRATY JEAN – FABRICE, les systèmes décisionnels. 2014 Page 1

I.4. LE PROCESSUS DECISIONNEL DANS UNE ENTREPRISE

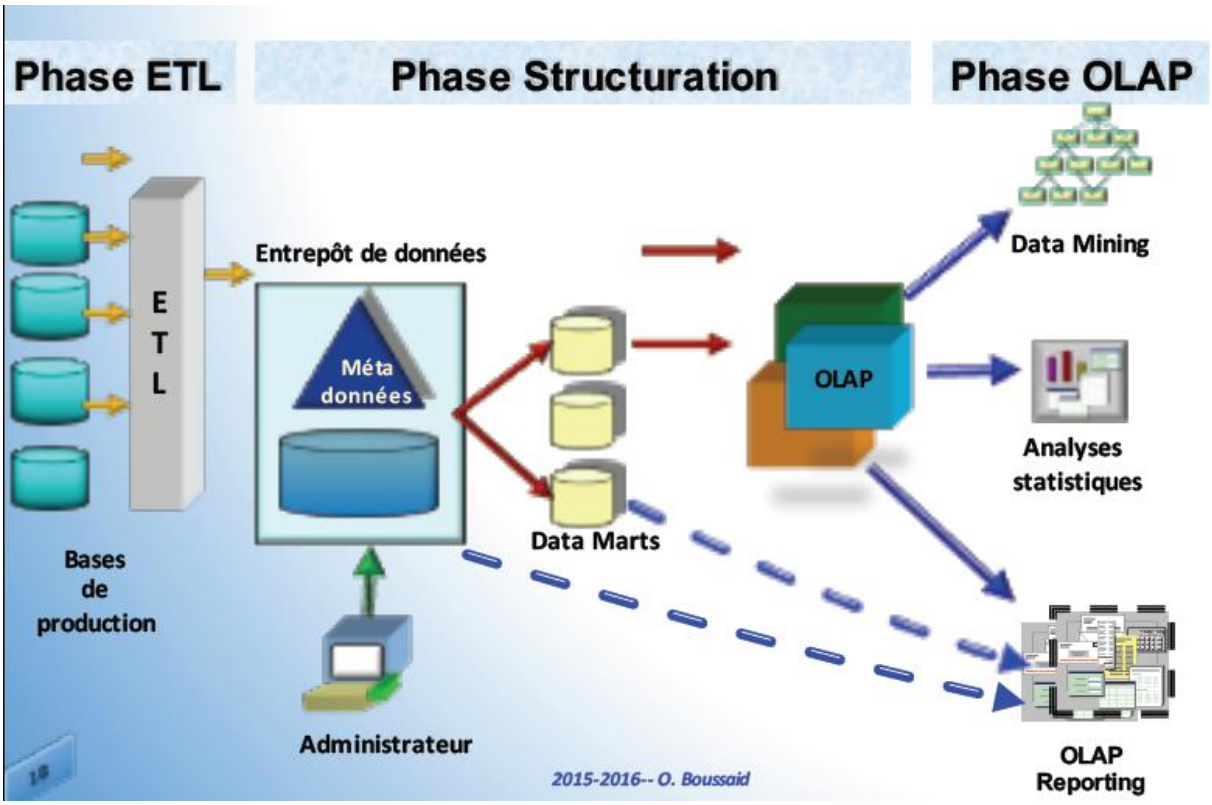
Une décision est le résultat d'un processus comportant le choix conscient entre plusieurs solutions en vue d'atteindre un objectif.
L'efficacité des services d'une entreprise dépend de la qualité de ses décisions, donc améliorer l'habilité à prendre des décisions c'est optimiser l'usage des ressources dont dispose l'entreprise. La technique de processus décisionnel comporte sept étapes :

1. **Définir le problème** : le problème à résoudre est souvent présenté dans des termes vagues et peut être difficile à identifier. Il arrive que ce que nous croyons être le problème ne soit en réalité qu'un symptôme et si l'on s'attaque au symptôme, on n'arrivera jamais au fond du problème. Avant de pouvoir être précis, il faut donc que la définition initiale parle du problème tel qu'il se présente.
2. **Rassembler les faits et les données :** il faut ici rassembler tous les renseignements pertinents, car le manque de faits probant diminue grandement la qualité et l'efficacité des décisions.
3. **Evaluer ou interpréter les faits et les données** : pour que la décision soit efficace, il est souvent nécessaire de traiter une quantité considérable de renseignement en utilisant une méthode de classification de données recueillies en les regroupant en catégories avec des critères communs.
4. **Etablir plusieurs solutions** : il faut tâcher d'envisager toutes les solutions possibles visant à atteindre les objectifs, sans juger leur valeur objective. Au cours de cette étape, l'imagination doit avoir libre cours en provoquant des réactions en chaine, car d'une idée peut en jaillir une autre.
5. **Choisir une solution :** c'est qu'il faut évaluer, juger, critiquer. Il faut tenir compte du pour et du contre pour chaque solution ; évaluer leurs avantages et les désavantages.
6. **Délais de prises de décision :** en utilisant le processus de précision et la méthodologie qui lui sert de support (si nécessaire), le décideur a une idée consciente et claire de façon dont il faut utiliser systématiquement les informations

en respectant les délais qui lui sont impartis pour prendre une décision raisonnable et cohérente.

7. **Décision regrettées** : une décision regrettée vient en générale de ce que :
   - Les renseignements manquaient ;
   - Les conséquences n'avaient pas été pesées ;
   - Le moment avait été mal choisi.

1.5. ARCHITECTURE DE SYSTEME DECISIONNEL



I.6. LES FONCTIONS ESSENTIELLES D'UN SYSTEME DECISIONNEL

Un système d'information décisionnel (SID) assure quatre fonctions fondamentales, à savoir la collecte, l'intégration, la diffusion et la présentation des données. A ces quatre fonctions s'ajoute une fonction de contrôle système d'information lui-même, l'administration.

- **Collecte des données** :
  La collecte est l'ensemble des tâches consistant à détecter, à sélectionner, à extraire et à filtrer les données brutes issues des environnements pertinents compte tenu du périmètre du système d'information. Les sources de données internes ou externes étant souvent hétérogènes tant sur le plan technique que sur le plan sémantique, cette fonction est la plus délicate à mettre en place dans un système décisionnel complexe. Elle s'appuie notamment sur des outils d'ETL (extract-transform-load pour extraction-transformation-chargement).

Cette alimentation utilise les données sources issues des systèmes transactionnels de production, le plus souvent sous forme de :

- Compte-rendu d'événement ou compte-rendu d'opération : c'est le constat au fil du temps des opérations (achats, ventes, écritures comptables,…), le film de l'activité de l'entreprise ;
- Compte-rendu d'inventaire ou compte-rendu de stock : c'est l'image prise à un instant donné (à une fin de période, etc.) de l'ensemble du stock (les clients, les contrats, les commandes, les encours etc.).

La fonction de collecte joue également, au besoin, un rôle de recodage. Une donnée représentée différemment d'une source à une autre impose le choix d'une représentation unique pour les futures analyses.

- **Intégration des Données :**
  L'intégration consiste à concentrer les données collectées dans un espace unifié, dont le socle informatique essentiel est l'entrepôt de données. Elément central du dispositif, il permet aux applications décisionnelles de bénéficier d'une source d'information commune, homogène, normalisée et fiable, susceptible de masquer la diversité de l'origine de données. Au passage les données sont transformées par :

- ❖ Un filtrage et une validation des données en vue du maintien de la cohérence d'ensemble ;
- ❖ Une synchronisation ;
- ❖ Une certification pour rapprocher les données de l'entrepôt des autres systèmes « légaux » de l'entreprise.

C'est également dans cette fonction que sont effectués éventuellement les calculs et les agrégations communes à l'ensemble du projet.
La fonction d'intégration est généralement assurée par la gestion des métadonnées, qui assurent l'interopérabilité entre toutes les ressources informatiques, que ce soit des données structurées (bases de données accédées par des progiciels ou application), ou des données non structurées (documents et autres ressources non structurées, manipulés par les systèmes de gestion de contenu).

- **Diffusion :**
  La diffusion met les données à la disposition des utilisateurs, selon des schémas correspondants au profil ou au métier de chacun, sachant que l'accès direct à l'entrepôt de données ne correspondrait généralement pas aux besoins d'un décideur ou d'un analyste.
  L'objectif prioritaire est de segmenter les données en contextes informationnels fortement cohérents, simples à utiliser et correspondant à une activité décisionnelle particulière. Chaque contexte peut correspondre à un Data Mart.

- **Présentation des Données :**
  Cette quatrième fonction, la plus visible pour l'utilisateur, régit les conditions d'accès de l'utilisateur aux informations. Elle assure le fonctionnement du poste de travail, le contrôle d'accès, la prise en charge des requêtes, la visualisation des résultats sous une forme ou une autre. Elle utilise toutes les techniques de communication possible (outils bureautiques, requêteurs et générateurs d'états spécialisés, infrastructure web, etc.).

- **Administration des données :**
  C'est la fonction transversale qui supervise la bonne exécution de toutes les autres. Elle pilote le processus de mise à jour des données, la documentation sur les données (les métadonnées), la sécurité, les sauvegarde, la gestion des incidents.

I.7. LES APPORTS DES SYSTEMES DECISIONNELS

Il est toujours difficile d'expliquer à des dirigeants que l'on doit dépenser de l'argent, parfois beaucoup trop, pour analyser et manipuler des données existant dans les systèmes d'information de l'entreprise. Les apports des systèmes décisionnels sont néanmoins réels.
Ils peuvent être classés en deux catégories :

- L'amélioration de l'efficacité de la communication et de la distribution des informations de pilotage ;
- L'amélioration du pilotage des entreprises résultant de meilleures décisions, prises plus rapidement.

Si le premier point est aisément compréhensible, présente peu de risque de mise en œuvre et pose peu de problème d'évaluation ce n'est clairement pas en revanche une source de gains significative.
Il sera très difficile, le plus souvent, de justifier les coûts d'un projet sur cette seule promesse.
La seconde catégorie à nettement plus de potentiel de gains mais il faut bien reconnaître que les risques de ne pas atteindre les objectifs initiaux sont réels, sans parler des énormes difficultés d'évaluation des bénéfices escomptés.
Les bénéfices de ce type les plus souvent cités sont les suivants :

- Unicité des chiffres, une seule vérité acceptée par tous ;
- Meilleure planification ;
- Amélioration de la prise de décision ;
- Amélioration de l'efficacité des processus ;
- Amélioration de la satisfaction des clients et des fournisseurs ;
- Amélioration de la satisfaction des employés.

I.8. LES FACTEURS CLES DE SUCCES D'UN PROJET DECISIONNEL

Les points clés pour la réussite d'un projet décisionnel sont :

- Obtenir l'engagement des managers opérationnels concernés. Leur soutien sera crucial pour réussir la mise en œuvre des changements de l'organisation et des modes de fonctionnement nécessaires pour tirer parti des apports du projet décisionnel,
- Bâtir une compréhension partagée des objectifs et des orientations du projet décisionnel,
- Prioriser les domaines fonctionnels d'application, ne pas tout faire en même temps mais privilégier une succession de petits projets à succès.
- Donner au projet les moyens financiers et humains suffisants pour bâtir et faire vivre la solution décisionnelle.
- Redonner toute sa dimension au business. Les projets décisionnels ne doivent plus être uniquement des projets techniques centrés sur les outils. Une réelle conception des logiques d'analyse et des indicateurs est nécessaire, quitte à enrichir les données sources si l'on souhaite aller au-delà de la stricte amélioration de la diffusion des informations de pilotage.
- Assurer la fiabilité des données.

D'après tous ceux qui précédent, nous dirons tout simplement qu'au cours de ces dernières années, la puissance et les performances des outils décisionnels se sont grandement améliorés.

Les outils sont plus simples à mettre en place, à utiliser et exploiter. Leur intégration avec les applications transactionnelles est plus facile et moins coûteuse.

L'accroissement de la performance des outils change la polarité des projets décisionnels. Hier très focalisés sur les aspects techniques, ils sont aujourd'hui principalement centrés sur des questions fonctionnelles.

L'offre de solutions décisionnelles couvre un large spectre fonctionnel et regroupe de nombreux acteurs du marché proposant des produits de bonne qualité mais très différents.

Les apports, quoique difficiles à évaluer financièrement, sont réels et le plus souvent constatés par les entreprises utilisatrices.
Les risques de ne pas tirer tout le profit de sa solution décisionnelle et de se limiter aux gains de productivité administrative liés à l'amélioration de la diffusion des informations de pilotage sont réels.

Sans détailler à nouveaux les facteurs clés de succès, maximiser les apports de son projet décisionnel implique :

- De définir et de partager des objectifs précis ;
- De réfléchir à ses besoins et de choisir l'offre progicielle adaptée ;
- De donner toute sa dimension fonctionnelle au projet (Quels indicateurs ? Quelles logiques d'analyses souhaitons-nous mettre en place ? …).

D'être prêt à faire évoluer ses modes de fonctionnement et son organisation. Un nouveau tableau de chiffre n'améliora pas, en lui-même, sensiblement le fonctionnement d'une entreprise. En revanche, de nouveaux modes de fonctionnement rendus possibles par la disponibilité fréquente, d'informations de pilotage fiables et pertinentes peuvent améliorer sensiblement les performances d'une entreprise.

I.9. LES ENJEUX D'UN SYSTEME DECISIONNEL

De nos jours, les données applicatives sont stockées dans une base (ou plusieurs) base(s) de données relationnelle(s) ou non relationnelle(s). Ces données sont extraites, transformées et chargées dans un entrepôt de données généralement par un outil de type ETC (Extraction-Transformation-Chargement).

I.10. COMPARAISON ENTRE SYSTEME DECISIONNEL ET LE SYSTEME OPERATIONNEL

Nous illustrerons la nette démarcation existant entre ces deux systèmes à partir de questions suivantes :

- ❖ Quels sont les noms et prénoms des personnels de mon entreprise qui reçoivent des primes d'un montant supérieur à 20% de leur salaire ?
- ❖ Est-ce que le nombre de salariés, qui reçoivent des primes d'un montant supérieur à 20% de leur salaire, est en augmentation sur les cinq dernières années ?

En effet, la première question se réfère directement à des données issues d'une base de données servant à enregistrer les transactions de l'organisation. Par contre la seconde question nécessite de mettre en œuvre un système permettant, notamment, de conserver d'une manière structurée un historique des données.

Données décisionnelles	**Données opérationnelles**
Orientée activité (thème, sujet), condensées, représentent des données historique	Orientées application, détaillées, précises au moment de l'accès
Pas de mise à jour interactive de la part des utilisateurs	Mise à jour interactive possible de la part des utilisateurs
Utilisées par l'ensemble des analystes, gérés par sous-ensemble	Accédées de façon unitaire par une personne à la fois
Exigence différente, haute disponibilité ponctuelle	Haute disponibilité en continu
Peuvent être redondantes	Unique (pas de redondance en théorie)
Grande quantité de données utilisées par les traitements	Petite quantité des données utilisées par un traitement
Cycle de vie différent	Réalisation des opérations au jour le jour
Faible probabilité d'accès	Forte probabilité d'accès
Utilisée de façon aléatoire	Utilisée de façon répétitive

I.11. L'INFORMATIQUE DECISIONNELLE

L'informatique décisionnelle (ou Business Intelligence en anglais) désigne les moyens, les outils et les méthodes qui permettent de collecter, consolider, modéliser et restituer les données, matérielles ou immatérielles d'une entreprise en vue d'offrir une aide à la décision et de permettre aux responsables de prendre des stratégies pour l'entreprise et d'avoir une vue d'ensemble de l'activité traitée au sein de l'entreprise.

En générale, ce type d'application utilise un Data warehouse pour stocker des données provenant de plusieurs sources hétérogènes et fait appel à des traitements par lots pour la collecte de ces informations. L'informatique décisionnelle s'insère dans l'architecture plus large d'un système d'information.

I.12. Entrepôt de données (Data Warehouse)

Constituant principale d'un système informatique décisionnel, un Data Warehouse, est une vision centralisée et universelle de toutes les informations de l'entreprise. C'est une structure (comme une base de données) qui a pour but, contrairement aux bases de données, de regrouper les données de l'entreprise pour des fins analytiques et pour aider à la décision stratégique. Ils apportent une aide bien précieuse aux dirigeants des entreprises en leur fournissant une vue synthétique de leur entreprise. Ou encore, le data warehouse peut se définir comme étant un ensemble de données intégrées, orientées sujet, non volatiles, gérées dans un environnement de stockage particulier, historisées, résumées, disponibles pour l'interrogation, l'analyse et organisées pour le support d'un processus d'aide à la décision. Lesquelles données provenant des sources externes, des applications de productions, d'Internet…, et est alimentées par le biais des outils informatique appelés **ETL** « **Extract, Transform, Load**».

D'après toutes les définitions citées ci-dessus, nous voyons donc les données d'un Entrepôt possèdent les caractéristiques ci-après :

- **Intégrées :**

  C'est-à-dire que les données de l'entrepôt proviennent de différentes sources éventuellement hétérogènes. L'intégration consiste à résoudre les problèmes d'hétérogénéité des systèmes de stockage, des modèles de données, de sémantique de données etc.

- **Orientées sujet :**

  Après leur intégration dans une sorte de source globale, les données sont réorganisées autour de thèmes tels que : client, vendeur, produit…etc.

- **Non volatile :**

  Tout se conserve, rien ne se perd : cette caractéristique est primordiale dans les entrepôts de données. En effet, et contrairement aux bases de données classiques, un entrepôt de données est accessible en ajout ou en consultation uniquement. Les modifications ne sont autorisées que pour des cas particuliers (correction d'erreurs…etc.).

- **Historisées :**

  La conservation de l'évolution des données dans le temps, constitue une caractéristique majeure des entrepôts de données. Elle consiste à s'appuyer sur les résultats passés pour la prise de décision et faire des prédictions ; autrement dit, la conservation des données afin de mieux appréhender le présent et d'anticiper le futur.

- **Résumés :**
  Les informations issues des sources de données doivent être agrégées et réorganisées afin de faciliter le processus de prise de décision.

I.12.1. Particularité De L'entrepôt De Donnés Par Rapport Aux Bases De Données

Un Data Warehouse est une base de données conçue pour l'interrogation et l'analyse plutôt que le traitement de transactions. Il contient généralement des données historiques dérivées de données transactionnelles, mais il peut comprendre des données d'autres origines.
Les Data Warehouse séparent la charge d'analyse de la charge transactionnelle. Ils permettent aux entreprises de consolider des données de différentes origines. Au sein d'une même entité fonctionnelle, le Data Warehouse joue le rôle d'outil analytique.

En complément d'une base de données, un Data Warehouse inclut une solution d'extraction, de transformation et de chargement (ETL), des fonctionnalités de traitement analytique en ligne (OLAP) et de Data mining, des outils d'analyse client et d'autres applications qui gèrent le processus de collecte et de mise à la disposition de données. Si nous pouvons encore aller très loin, nous constaterons que l'entrepôt de données est constitué de ces deux tables :

- **Table de faits :** c'est une table qui contient les données à analyser. Elle est souvent reconnaissable par sa taille ; en effet, lorsqu'on visualise un schéma, c'est celle qui est au centre et qui est la plus grande. Ce type de table est aussi facilement reconnaissable, car elle comporte un grand nombre de clés étrangères afin de la lier avec des tables de dimensions.
- **Tables de dimensions :** ces sont les tables qui entourent la table de fait. Elles sont composés des attributs et ces derniers permettent de stocker la description des dimensions et sont utilisés comme source de contrainte.

I.13. MODELE MULTIDIMENSIONNEL

Les modèles basés sur le concept multidimensionnel, sont les plus appropriés, à capturer les caractéristiques de Data warehouse. Ils permettent en effet, de donner une vision simple, et facilement interprétable par des non informaticiens, et de visualiser les données selon différentes dimensions.

Le modèle multidimensionnel contient deux types d'attributs : les dimensions et les mesures. Les dimensions sont les valeurs numériques que l'on compare, les dimensions sont les points de vue depuis lesquels les mesures peuvent être observées. La modélisation multidimensionnelle est illustrée par des cubes de données ou des hyper cubes

Le cube de donnée

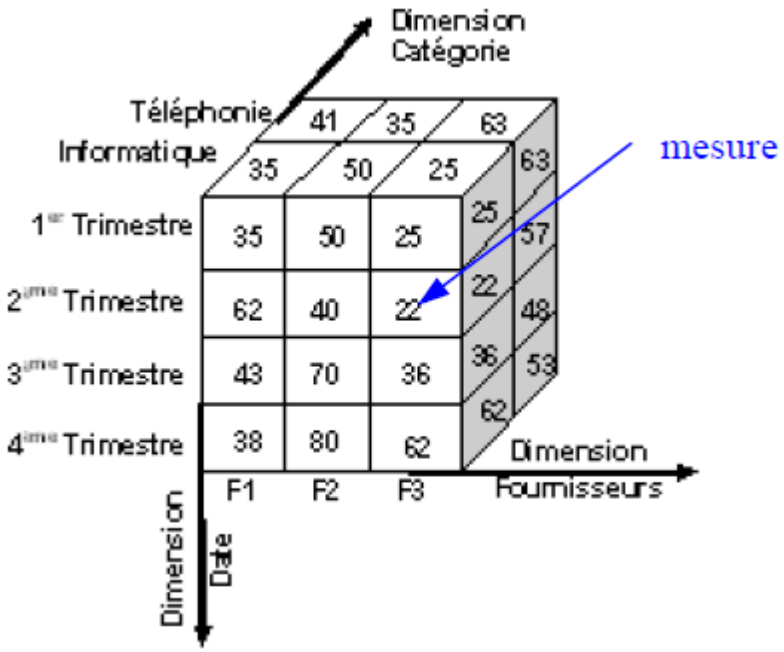


***Fig I.2**. Exemple d'un cube de données.*

De part sa nature, un data warehouse est alimenté via les informations de l'entreprise. Or ces informatique sont stockées sous les formes les plus hétérogènes. On peut retrouver ainsi plusieurs types de base de données (acces, DB2, MySQL,…), des tableurs, des fichiers à plats… Il existe un quasi infinité de format de source.
En générale, on retrouve trois types de contraintes à la mise en œuvre d'un data warehouse :

- ❖ Alimentation des données de production aux normes des données du référentiel.
- ❖ Organisation du stockage des informations.
- ❖ Sur le plan fonctionnel, garantir l'intégrité des données par des définitions uniques et réutilisables par tous les utilisateurs. Avant de se lance dans la mise sa place, il est impératif de maîtrises les principes et les contraintes de fonctionnement du data warehouse.

La mise en place d'une base multidimensionnelle correspond donc à un certain nombre de critères :

- **L'utilité des données** : Inutile de s'encombrer avec données superflues. Le choix des données à transfert dans le cube d'analyse doit être dicté par la direction de l'analyse elle-même. Ainsi dans un data mart de type commercial, contenant l'ensemble des factures clients émises, sera inutile de transférer les numéros de factures vers le cube d'analyse. Cet indicateur, non significatif, ne peut être rattaché à aucune dimension. Il faut donc l'écarter du transfert. Sa présence dans le cube n'apporterait rien à l'analyse mais ralentirait les temps de réponse.

- **Le référencement** : A l'inverse, toutes les données utiles seront inscrites dans les tables via un référencement strict. Cette nomenclature sera définie pour les besoins d'analyses. Ainsi les dates de factures seront rattachées à des périodes, les comptes clients vont être rattachés à des groupes de clients.

- **La pertinence** : toujours garder a l'esprit le résultat attendu. Avant de se lancer dans la constitution d'un cube, il faut vérifier que les informations qui y seront produites auront un sens. Dans une société où les clients sont de passage et n'acquièrent le plus souvent qu'un seul produit, le couple, c'est à dire le croisement de la dimension clients et de la dimension produits, n'aura aucun intérêt et n'apportera donc pas d'information, tout en générant un cube très encombrant en espace mémoire. Il est donc impératif de penser à l'organisation des dimensions du cube d'analyse en

fonction de l'activité de l'entreprise. Il n'existe pas de modèle standard de données multidimensionnelles. Le modèle de données et sa structure restent en effet uniques pour chaque entreprise.

I.14. LES MAGASINS DE DONNEES (DATA MART)

Les magasins de données consistent à extraire une partie de l'information décisionnelle contenue dans l'entrepôt ; il s'agit de la partie des données utile pour une classe d'utilisateurs ou pour un besoin d'analyse spécifique, en ce sens ils sont orientés sujet.
L'objectif du magasin est de supporter efficacement des processus d'analyse de type OLAP.

Ainsi, les données stockées dans un magasin doivent correspondre à une structuration adaptée des données (selon plusieurs axes d'analyses) reflétant la vision des analystes. Cette représentation des données est basée sur une modélisation multidimensionnelle.

L'approche multidimensionnelle vise à modéliser les données conformément à la perception des analystes, c'est-à-dire les données sont représentées suivant les différents axes d'analyses possibles. Le modèle multidimensionnel comprend un fait contenant les mesures à analyser et des dimensions contenant les paramètres de l'analyse. Dans chaque dimension, les paramètres sont hiérarchisés selon des niveaux de détail.

Un magasin de données ressemble en fait à un Data Warehouse sauf qu'il est moins générique. Une approche courante consiste à maintenir des informations détaillées au niveau du Data warehouse et à les synthétiser dans un Data mart ou magasin de données pour chaque groupe ou département fonctionnel.

Une autre manière de conception consiste à créer des Data marts pour chaque département puis à fusionner ultérieurement ces données dans l'entrepôt global. Chacune de ces méthodes présente l'avantage de centraliser les informations pour les utilisateurs finaux.

- Les magasins de données consistent à extraire une partie de l'information décisionnelle contenue dans l'entrepôt ; il s'agit de la partie des données utile pour une classe d'utilisateurs ou pour un besoin d'analyse spécifique, en ce sens ils sont orientés sujet. L'objectif du magasin est de supporter efficacement des processus d'analyse de type OLAP.
- Ainsi, les données stockées dans un magasin doivent correspondre à une structuration adaptée des données (selon plusieurs axes d'analyses) reflétant la vision des analystes. Cette représentation des données est basée sur une modélisation multidimensionnelle.
- L'approche multidimensionnelle vise à modéliser les données conformément à la perception des analystes, c'est-à-dire les données sont représentées suivant les différents axes d'analyses possibles. Le modèle multidimensionnel comprend un fait contenant les mesures à analyser et des dimensions contenant les paramètres de l'analyse. Dans chaque dimension, les paramètres sont hiérarchisés selon des niveaux de détail.

I.15. OLAP(Online Analytical Processus)

OLAP signifie « Online Analytical Processus » repose sur une base de données multidimensionnelle, destinée à exploiter rapidement les dimensions d'une population de données. Le modèle OLAP sera celui du Data Warehouse, il sera construit pour sélectionner et croiser plusieurs données provenant des sources diverses afin d'en tirer une information implicite.

Ceci a évolué pour aboutir à une méthode d'analyse permettent aux décideurs un accès rapide et de manière pertinente présentée sous divers angles, dimensions sous forme de cube. L'outil OLAP repose sur la restructuration et le stockage des données dans un format multidimensionnel issues de fichiers plats ou de bases relationnelles.

Ce format multidimensionnel est connu sous le nom d'hyper cube, organise les données le long de dimensions. Ainsi, les utilisateurs analysent les données suivant les axes propres à leur métier.

I.16. ETL (EXTRACT, TRANSFORM, LOAD)

Outil informatique destiné à extraire des données de diverses sources (bases de données de production, fichiers, Internet, etc.), à les transformer et à les charger dans un entrepôt de données.

CHAPITRE II : LES TECHNIQUE DE DATAMININQ

II.1. INTRODUCTION

Le terme Data mining est souvent employé pour désigner un ensemble d'outils permettant aux utilisateurs d'accéder aux données de l'entreprise et des analyses. Les outils d'aide à la décision, qu'ils soient relationnels ou OLAP, laissent l'initiative à l'utilisateur de choisir les éléments qu'il veut observer ou analyser. Au contraire, dans le cas du data mining, le système a l'initiative et découvre lui-même les associations entre les données, sans que l'utilisateur ait à lui dire de rechercher plutôt dans telle ou telle direction ou à poser des hypothèses. Les modèles classiques de recherche d'informations ne sont pas adaptés pour traiter des masses gigantesques de données, souvent hétérogènes. C'est ce constat qui a permis au data mining d'émerger et vulgariser les méthodes d'analyse.

Le data mining (ou la fouille de données) a pour objet l'extraction d'un savoir à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques. La fouille de données repose sur un ensemble de fonctions mais aussi sur une méthodologie de travail.

Le terme de data mining signifie littéralement exploitation des données. Comme dans toute exploitation, le but recherché est de pouvoir extraire de la richesse. Ici, la richesse est la connaissance de l'entreprise. Fort du constat qu'il existe au sein des bases de données de chaque entreprise une ressource de données cachées et surtout inexploitée, le data Mining permet de faire les apparaître, et cela grâce à un certain nombre de techniques spécifiques. Nous appellerons data mining l'ensemble des techniques qui permettent de transformer les données en connaissances. Le périmètre d'exploitation du data mining ne se limite pas à l'exploitation des Data warehouse. Il veut d'être capable d'exploiter toutes bases de données contenant de grandes quantités de données telles que des bases relationnelles, des entrepôts de données mais également des sources plus ou moins structurées comme internet. Dans ces cas, il faut néanmoins

construire une base de données ou un entrepôt de données qui sera dédié à l'analyse.

Le data mining est un processus itératif qui met en œuvre un ensemble de techniques hétéroclites tel que le data warehouse, de la statistique, de l'intelligence artificielle, de l'analyse des données et des interfaces de communication homme - machine. Le résultat du datamining peut se présenter sous différent format : texte plat, tableau, graphique...

Le datamining est un ensemble d'outils d'analyse d'entrepôt de données et de cube apportant aux décideurs des éléments supplémentaire de prise de décisions qui ne sont pas forcément visible aux premiers abords.

II.2 DEFINITION

Le Data mining est un ensemble de techniques et de méthodes du domaine des statistiques, des mathématiques et de l'informatique permettant d'explorer et d'extraire de données brutes, des connaissances originales auparavant inconnues, afin de : décrire le comportement actuel des données et/ou prédire le comportement futur des données.

Autrement dit, le data mining fait donc référence à un ensemble de techniques d'exploration et d'analyse, par des moyens automatiques ou semi-automatiques, d'une masse importante de données dans le but de découvrir des tendances cachées ou des règles significatives (non triviales, implicites et potentiellement utiles). Ces outils reposent en général, sur des techniques basées sur les statistiques, la classification ou l'extraction de règles associatives.

Partant de là, on comprend mieux pourquoi le data mining est annoncé comme étant « un des développements technologiques les plus révolutionnaires des dix prochaines décennies » selon le magazine en ligne ZDNET News. (Rachel Konrad, février 2001). En effet, cette technologie est perçue comme étant réellement indispensable de nos jours à l'analyse de

la quantité toujours plus vaste d'informations produites par tous les systèmes d'information de l'entreprise.

Enfin, le data mining, en tant que processus, il est pertinent de souligner ici qu'il ne se réfère pas seulement à des outils et à une technologie informatique très développée. Effectivement, il faut également relever le rôle fondamental de l'humain dont l'implication se doit d'être totale dans chaque phase du processus. Il est erroné de penser que le data mining est une entité qui fonctionne de manière autonome.

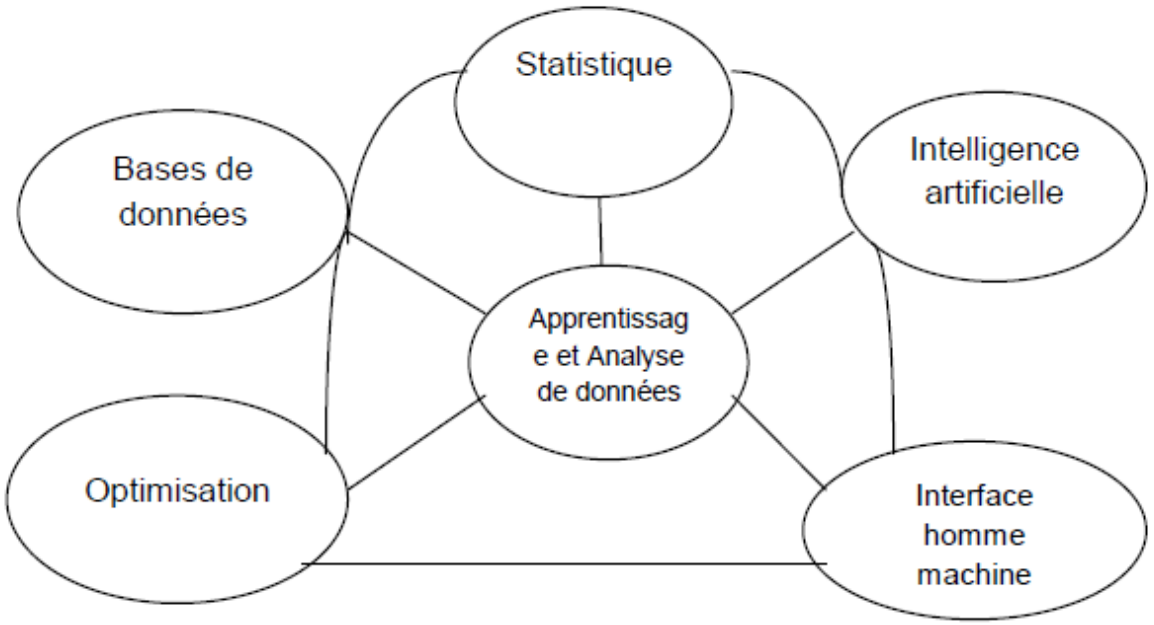


Fig II.1. *Les disciplines qui fonctionnent avec le data mining*

Comme le montre la figure précédente, le datamining est aperçu comme un processus itératif qui met en œuvre un ensemble de techniques hétéroclites tel que des bases de données (data warehouse), de la statistique, de l'intelligence artificielle, de l'analyse des données, des interfaces de communication homme-machine. Le résultat du datamining peut se présenter sous différent format : texte plat, tableau, graphique…

II.3. OBJECTIF DU DATA MINING

- **Expliquer : L**e data mining pourra tenter d'expliquer un événement ou un incident indiscernable. Par la consultation des informations contenues dans la base de données de l'entreprise, on peut être en mesure de formuler la question suivante : pour qu'elle raison perd-t-on des clients pour tel produit spécifique dans telle région ? tout en se basant sur des données collectées ou des mises en signification de paramètres liés, le data mining va essayer de trouver un certain nombre d'explication à cette question. Le Data Mining va aider à trouver des hypothèses d'explications.
- **Confirmer :** Le data Mining aidera à confirmer un comportement ou une hypothèse. Dans le cas où le décideur aurait un doute concernant une hypothèse, le data Mining pourra tenter de confirmer cette hypothèse en la vérifiant en appliquant des méthodes statistiques ou d'intelligence artificielle.
- **Explorer** : Enfin, le data mining peut explorer les données pour découvrir un lien "inconnu" jusque-là. Quand le décideur n'as pas d'hypothèse ou d'idée sur un fait précis, il peut demander au système de proposer des associations ou des corrélations qui pourront aboutir à une explication. Il est utopique de croire que le data mining pourrait remplacer la réflexion humaine. Le data mining ne doit être vu et utiliser uniquement en tant qu'aide à la prise de décision. Par contre, l'informatique décisionnelle dans son ensemble, et plus particulièrement le data mining permet de suggérer des hypothèses. La décision finale appartiendra toujours au décideur.

II.4. LES PARTICULARITES DU DATA MINING

1. Les techniques ne sont pas dans la culture des staticiens, car les données proviennent de l'apprentissage automatique (intelligence artificielle) et des bases de données.

2. Le data mining est intégré dans le schéma organisationnel de l'entreprise. Ainsi, les données ne sont plus issues d'enquêtes ou des sondages, mais proviennent d'entrepôts construits sciemment pour une

exploitation aux fins d'analyse. D'une part, une réorganisation du flux de données au sein de l'entreprise devient nécessaire (l'enchainement des bases de production, le data warehouse et les data mart) ;

3. Le traitement des données se développe en traitant, non plus seulement des fichiers plats « individus x variables », mais également des données sous forme non structurée, le texte, les images et les vidéos. Cette orientation attribue une place primordiale à l'appréhension et la préparation de données.

II.5. Les techniques du data mining

Le data mining met en œuvre un **ensemble de techniques** issues des statistiques, de l'analyse de données et de l'informatique pour explorer les données. On distingue deux grandes catégories de techniques : les techniques descriptives et les techniques prédictives.

II.5.1. Les techniques descriptives

Elles consistent à Décrire, Résumer, synthétiser, réduire, classer, Mettre en évidence des informations présentes mais cachées par le volume des données. Il n'y a pas de variable cible à prédire. On les appelle aussi : technique non supervisées. Elles produisent des modèles de classement : typologie, méta-typologie.

II.5.2. Les techniques prédictives

Elles consistent à Prédire, Extrapoler de nouvelles informations à partir des informations présentes. Les techniques prédictives présentent une variable cible à prédire. L'objectif est de prévoir la variable cible mais aussi de classer à partir de la variable cible. On les appelle aussi : techniques supervisées. Elles sont plus délicates à mettre en œuvre que les techniques descriptives. Elles demandent plus d'historique que les techniques descriptives. Elles produisent des modèles de prédiction.

II.6. LES OUTILS UTILISES DANS DATA MANING

Selon le type de données disponibles et le type de connaissances recherchées, la méthode d'obtention des règles finales va varier grandement. Chacune des techniques décrites ci-dessous possède certains avantages et certains inconvénients. Face à une problématique il convient de connaitre chacune d'entre elle pour apporter la solution la plus efficace :

- ✓ **Les arbres de décision** :

C'est une technique de classification automatisée. L'analyse d'un ensemble d'enregistrements préalablement classifies permet de générer une structure arborescente optimale pour la classification des autres enregistrements disponibles.
Ceci est fait en partitionnant successivement la population initiale des enregistrements de sorte à isoler au mieux les enregistrements d'une même classe. Ce sont pour la plupart des algorithmes légers, performants, et la forme arborescente des résultats permet une grande lisibilité. Mathématiquement, on se donne un ensemble X de N exemples notés xi dont les P attributs sont quantitatifs ou qualitatifs. Chaque exemple x est étiqueté, c'est-à-dire qu'il lui est associée une « classe » ou un « attribut cible » que l'on note y appartient à Y. A partir de ces exemples, on construit un arbre dit « de décision » tel que :

- chaque nœud correspond à un test sur la valeur d'un ou plusieurs attributs ;
- chaque branche partant d'un nœud correspond à une ou plusieurs valeurs de ce test ;
- A chaque feuille est associée une valeur de l'attribut cible.

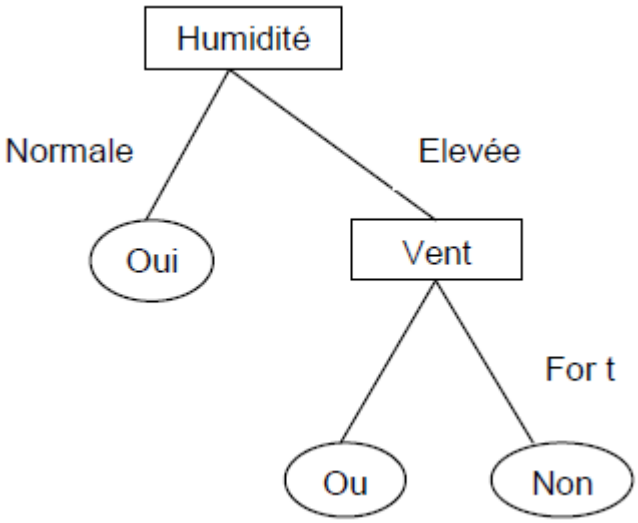


Fig II.2. *Exemple sur le jeu de données « jouer au tennis ? »*

- ✓ **Les réseaux neuronaux :**

C'est une technique de classification automatisée. De même que pour les arbres de décision, le principe consiste à apprendre classifier correctement les données à partir d'un jeu d'exemples déjà classifiés.
Présentement, les champs d'information des enregistrements forment les entrées d'un réseau dont la sortie correspond à la classe de l'enregistrement. L'apprentissage consiste alors à faire passer les enregistrements classés en entrée du réseau, et à corriger un petit peu l'erreur fatalement obtenue en sortie en modifiant les nœuds internes du réseau. Au fur et à mesure que celui-ci s'adapte, et finit par classer correctement les enregistrements. Si ces algorithmes sont puissants, ils nécessitent bien évidemment plus de travaux de mise en œuvre.

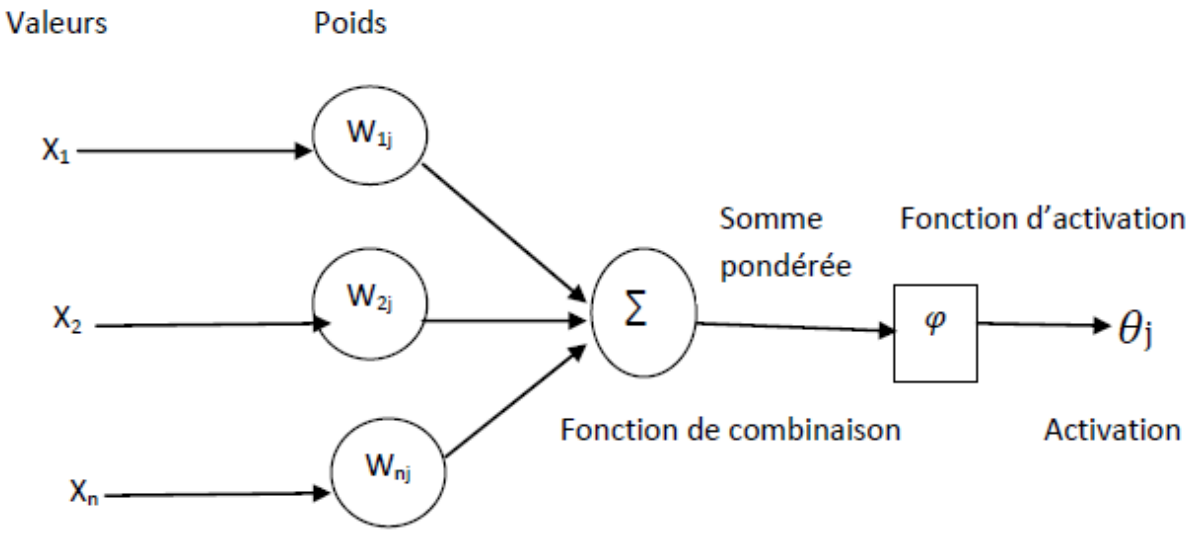


Fig II.3. *Structure d'un neurone artificiel.*

✓ La régression logistique

Historiquement, la régression logistique ou régression binomiale fut la première méthode utilisée, notamment en marketing pour le scoring et en épidémiologie, pour aborder la modélisation d'une variable binaire binomiale (nombre de succès pour ni essais) ou de Bernoulli (avec ni = 1) : possession on non d'un produit, bon ou mauvais client, décès ou survie d'un patient, absence ou présence d'une pathologie... Bien connue dans ces types d'application et largement répandue, la régression logistique conduit à des interprétations pouvant être complexes mais rentrées dans les usages pour quantifier, par exemple, des facteurs de risque liés à une pathologie, une faillite... Cette méthode reste donc celle la plus utilisée même si, en terme de qualité prévisionnelle, d'autres approches sont susceptibles, en fonction des données étudiées, d'apporter de bien meilleurs résultats.
Il est donc important de bien maîtriser les différents aspects de la régression logistiques dont l'interprétation des paramètres, la sélection de modèle par sélection de variables ou par régularisation (LASSO).

II.6. NOTIONS DE BASE

Le data mining travaille sur des tableaux de données pour lesquels :

- Le nom du tableau est une abstraction, c'est-à-dire une représentation mentale générale et abstraite d'un objet qui se manifeste comme un résultat de l'opération de l'esprit permettant de placer tel objet dans telle catégorie et non dans telle autre. Par exemple un tableau de clients, de malades etc.

- Chaque colonne du tableau a un nom qui est un attribut du tableau. On parle aussi de « propriété» ou de « champ ». Le nom de la colonne est aussi une abstraction. Pour un objet concret, la colonne a une valeur particulière qui est la valeur particulière de l'attribut pour l'objet concret. En data mining (et en statistique), les attributs des objets sont appelés : « variables ».

- Chaque ligne du tableau est un élément du tableau, c'est-à-dire un objet concret correspondant au concept abstrait dont on parle. En data mining un objet concret est appelé : «individu », la valeur d'un attribut pour un individu est appelé : « donnée » et enfin, l'ensemble des individus est appelé : « population ».

- Un sous-ensemble de valeurs pour un ou plusieurs attributs donnés peut être appelé : « type », « classe », « catégorie », « segment » ou encore « modalité ». Par exemple, « grand » et « petit » sont deux types (ou classe, ou catégorie, ou segment) de l'attribut « taille ». On parle de « variable catégorielle » par opposition aux « variables numériques ». Par exemple, si la variable (attribut) « taille » peut prendre deux valeurs possibles : « grand » et « petit », c'est une variable catégorielle. Si les valeurs de la variable « taille » sont données en cm, c'est une variable numérique.

- Quand on fait de la prévision, on travaille sur une variable particulière appelée : « variable cible » et sur un ensemble d'autres variables utiles pour la prédiction appelées : « prédicteurs ». Le principe général de la prédiction sera : si le ou les prédicteurs valent tant, alors la variable cible vaut tant.

Ainsi donc une entreprise pour mieux connaître sa clientèle, peut décider d'effectuer une classification basée sur le comportement des clients. Ceci implique la nécessité de mettre en place des outils de datamining ou fouille de données. Ces outils reposent en général sur des techniques basées sur les chaque décideur d'une entreprise doit disposer d'une vue sur les informations qui lui sont pertinentes, et qui peuvent influer dans ses décisions pour une meilleure exploitation de ces données.

II.6.2. Le Modèle

En fouille des données, un modèle est un résumé global des relations entre variables, exprimable sous forme algorithmique ou analytique, et permettant de comprendre des phénomènes (description) et émettre des prévisions (prédiction, raisonnement).

La démarche de Data Mining, dont nous nous attelons, ne traite pas d'estimation et de tests de modèles pré spécifiés, mais de la découverte de modèles à l'aide d'un processus de recherche algorithmique d'exploration de modèles:

- linéaires ou non,
- explicites ou implicites: réseaux de neurones, arbres de décision, SVM, régression logistique, réseaux bayésiens…

Un modèle n'est donc pas issu d'une théorie, mais de l'exploration des données ou apprentissage sur les données.

II.6.3. La Prédiction

On dispose d'un ensemble X de N données étiquetées. Chaque donnée xi est caractérisée par P attributs et par sa classe

yi ∈ Y. Dans un problème de classification (prédiction), la classe prend sa valeur parmi un ensemble fini et par la suite, on considérera qu'elle ne prend que deux valeurs, soit Y= {-1,1}. La démarche consiste alors à prédire la classe de toute nouvelle donnée x ∈ D, en s'appuyant sur l'ensemble d'exemples $X = \{(X_i, Yi)_{i\{1,...,N\}}\}$. Cette prédiction se fait à l'aide d'une procédure ou algorithme appelé ***« Classeur ou Classifieur ».*** On distingue généralement deux types de classifieurs :

- Ceux qui prédisent directement la classe de la nouvelle donnée, sur base des exemples ou données d'apprentissage ;
- Ceux qui construisent d'abord un modèle à partir des exemples (*modèle de prédiction*), et l'utilisent pour prédire la classe d'une nouvelle donnée.

II.7. LES DIFFERENTS TYPES DE DONNEES RENCONTRES

Classiquement, les données utilisées pour la prédiction ou classification sont décrites dans un tableau individus-variables par une valeur unique. On parlera alors de « *tableau de descriptions univaluées ou classiques* ». Dans les applications réelles, où le grand souci est de prendre en compte la variabilité et la richesse d'informations au sein des données, il est courant d'avoir affaire à des données complexes et hétérogènes (ou mixtes). Ce qui se traduit par le fait que chaque case du tableau de descriptions peut tenir non seulement une valeur unique mais également de valeurs, un intervalle de valeurs ou une distribution sur un ensemble de valeurs. On dira alors que la prédiction ou classification va porter sur un « *tableau de descriptions symboliques* ».

II.7.1. Les Variables

II.7.1.1. Description Classique

Classiquement, une variable y_h est définie par une application :

$$Y_h : X \rightarrow 0_h$$

$$X_i \in X \rightarrow Y_h(X_i)$$

Où : $X=\{X_1, X2,\ldots, Xn\}$ est l'ensemble des individus. L'ensemble d'arrivée 0_n est appelé domaine d'observation de la variable y_h. Un individu est alors décrit sur une variable y_h par une valeur unique de 0_h. La distinction de différentes variables se fait, dans notre cas d'étude, sur base de l'ensemble de leurs valeurs et de leur rôle dans le modèle de prédiction.

- Du point de vue de l'ensemble des valeurs, on distingue deux types de variables : *variables quantitatives* et *variables qualitatives*.

1. Variable Quantitative

Une variable quantitative ou numérique est une variable qui reflète une notion de grandeur, c'est-à-dire ses valeurs sont prises dans un ensemble numérique.

Une Variable Quantitative peut être discrète ou continue selon que l'ensemble de ses valeurs est fini ou infini.

2. Variable Qualitative

Une variable qualitative ou catégorielle est une variable pour laquelle la valeur mesurée sur chaque individu ne représente pas une quantité. Les différentes valeurs que peut prendre cette variables sont appelées les *catégories*, *modalités* ou *niveaux*.

Une Variable Qualitative peut être ordinal ou nominal selon que l'ensemble de ses valeurs est ordonné ou pas.

Nota : *On ne peut effectuer des opérations arithmétiques sur des variables qualitatives.*

- Du point du modèle de prédiction, on distingue les variables prédictives de la Variable cible.

1. Variables Prédictives

   Les Variables prédictives, discriminantes, d'entrée ou explicatives sont des variables dont les valeurs sont obtenues par observation et qui permettent d'expliquer ou de prédire les valeurs d'autres variables.
2. Variable Cible

   La Variable Cible, expliquée ou d'intérêt est une variable dont la valeur dépend d'autres variables.

II.8. CONSTRUCTION DU MODELE DE PREDICTION

Plusieurs Algorithmes ou Classifieurs sont utilisés pour construire un modèle de prédiction de la classe d'une nouvelle donnée, notamment : les Arbres de décision, les Réseaux de Neurones, Les SVM,…

Nous allons, dans le cadre de notre travail, nous atteler sur Les Arbres de décision.

II.8.1. Les arbres de décision

II.8.1.1. Présentation

Un arbre est une structure composée de noeuds et de feuilles (classes ou noeuds terminaux) reliés par des branches. On le représente généralement par mise de la racine en haut et les feuilles en bas (contrairement à un arbre réel). Pour bien comprendre cette définition, nous nous permettons ici de se

donner un ensemble X de N noeuds notés xi dont les P attributs sont qualitatif ou quantitatif ; chaque noeud est étiqueté, c'est-à-dire qu'il lui est associée une « classe » ou un attribut « cible » que l'on note y appartenant à Y. A partir de ces exemples, on construit un arbre dit « de décision » tel que :

- ✓ Chaque noeud correspond à un test sur la valeur d'un ou des plusieurs attributs ;
- ✓ Chaque branche partant d'un noeud correspond à une ou plusieurs valeurs de ce test ;
- ✓ A chaque feuille est associée une valeur de l'attribut cible.

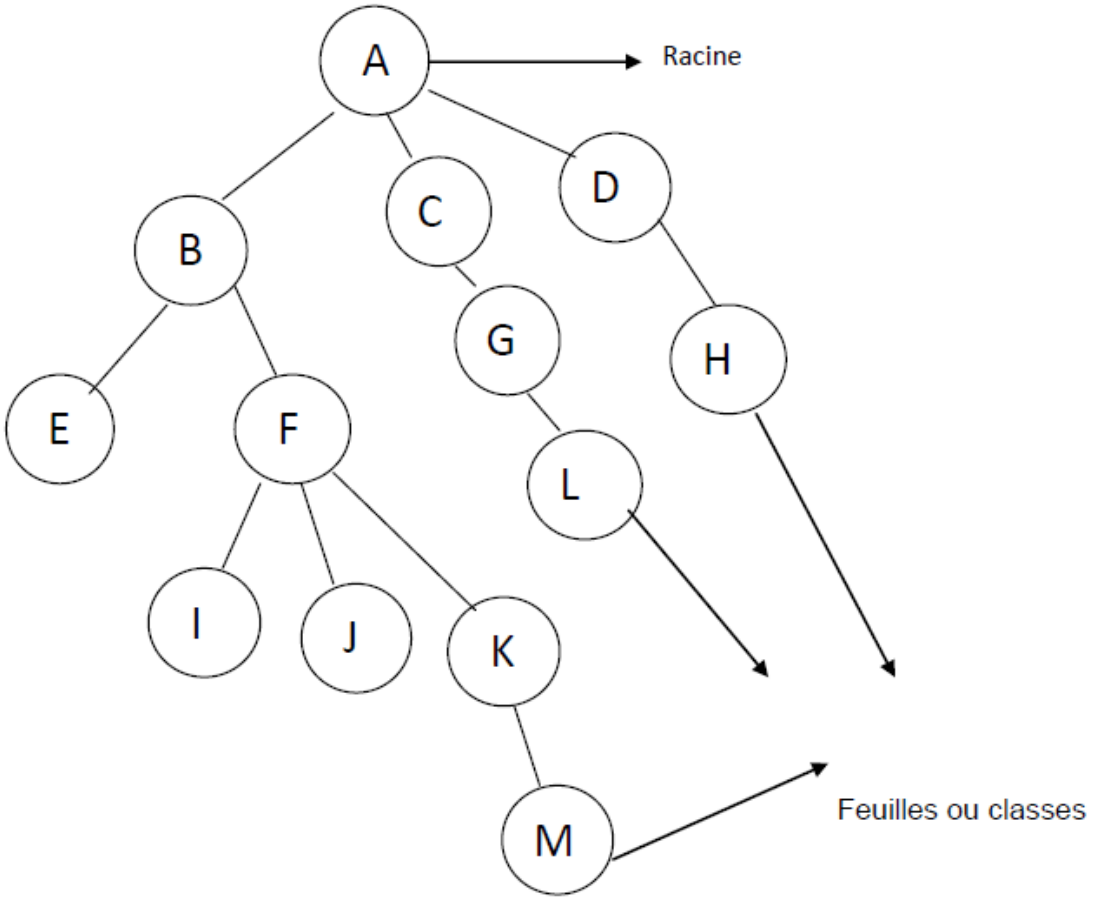


Fig III.1. *Exemple d'un arbre binaire de décision*

I. ARBRES BINAIRES

Un arbre binaire T est un ensemble fini d'éléments appelés noeuds tels que :

- T est vide (arbre nul ou arbre vide) ;
- T comporte un noeud particulier R, appelé racine de T et les autres noeuds de T forment une paire ordonnée T_1 et T_2 d'arbres binaires disjoints.

Si T comporte une racine R, les deux arbres T1 et T2 sont appelés respectivement sous - arbres de gauche et de droite de R.
Si T1 est non vide, sa racine est appelée successeur de gauche R ;
Si par contre T2 est non vide, sa racine est le successeur de droite de R. un arbre binaire T est très souvent représenté sous forme d'un arbre.

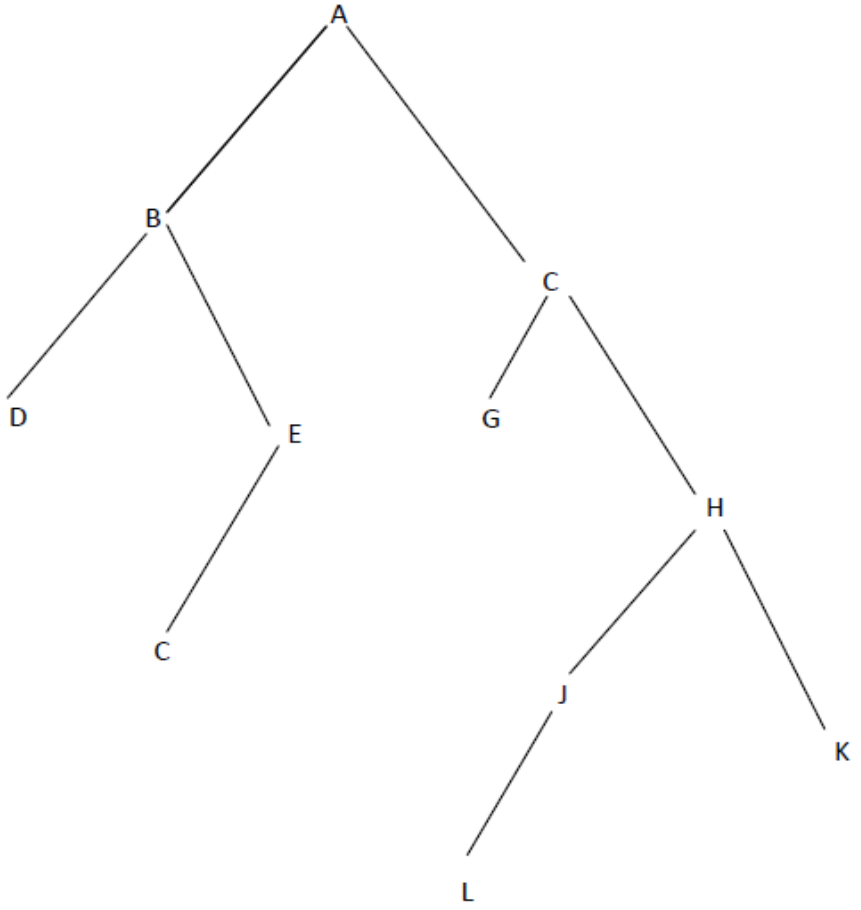


Fig III.2. *Exemple d'un arbre binaire*

2. Interprétation De Cet Arbre

1. La racine de T est le noeud ;
2. B est un successeur de gauche et C un successeur de droite du noeud;
3. Le sous arbres de gauche de la racine A comprend les noeuds B, D, E et F et celui de droite, les noeuds C, G, H, J, K et L.
4. Les noeuds A, B, C et H possèdent chacun deux successeurs, les noeuds E et J un seul, et les noeuds D, F, G, L et K aucun.

Dans la mesure où l'arbre est définit en fonction des sous - arbres T1 et T2, sa définition est récursive, c'est-à-dire chaque noeud N de T comporte un sous arbre de gauche et un sous arbres de droite. De plus si N est un noeud terminal, les deux sous - arbres, gauche et droite, sont vides. Deux arbres binaires T et T' sont similaires s'ils ont la même structure ou, en d'autres termes, s'ils ont la même forme.

3. Terminologie Des Arbres

Soit N un noeud d'un arbre T, de successeur de gauche G_1 et de successeur de droite G_2. N est appelé parent de G_1 et G_2. De plus G1 est appelé enfant de gauche de N, G_2 enfant de droite de N. G_1 et G_2 sont dits frères. Chaque noeud d'un arbre binaire T possède un parent unique, appelé prédécesseur de N.

Un noeud D est dit descendant d'un noeud N s'il existe une succession d'enfants de N à D. N étant alors appelé ancêtre de D. D est appelé descendant de gauche ou de droite de N selon que D appartient au sous - arbre de gauche ou de droite de N. Le segment de droite qui va d'un noeud N de T à son successeur est appelé arête, une séquence d'arêtes consécutives constitue un chemin. Un noeud terminal est appelé feuille ou classe et un chemin qui se termine par une feuille est une branche. A chaque noeud d'un arbre T est affecté un numéro de niveau. La hauteur d'un arbre binaire est le nombre de noeuds qui constituent le plus long chemin de la racine à une feuille. Un arbre vide a une hauteur nulle.

4. Défilement Dans Un Arbre Binaire

Il existe trois manières de faire défiler un arbre binaire T de racine R :

- Pré-ordre (noeud-gauche-droite) : ici le principe est de traiter tout d'abord la racine R, puis faire défiler le sous - arbre de gauche R et enfin, faire défiler le sous - arbre de droite R c'est-à-dire on lit l'information la première fois.
- In-ordre (gauche-noeud-droite) : le principe est tel qu'il faire défiler le sous - arbre de gauche R, puis traiter le sous - arbre de droite de R enfin, faire défiler le sous - arbre de droite R c'est-à-dire on lit cette information la deuxième fois (si on passe deux fois par ce noeud).
- Post-ordre (gauche-droit-noeud) : le principe est de faire défiler le sous - arbre de gauche, puis faire défiler le sous arbre de droite de R et enfin, traiter la racine de R. La différence entre ces trois algorithme réside dans le choix du moment où la racine R est traitée (on lit cette information la dernière fois.).

5. Validation D'un Arbre

Une fois un arbre de décision construit, il est essentiel de le valider en essayant d'estimer les erreurs de classification qu'il fait, autrement dit, la probabilité que la classe prédite pour une donnée quelconque soit correcte. Dépendant de l'ensemble de données qui est utilisé pour la mesurer, cette quantité est donc une variable aléatoire dont il faut estimer la valeur.

- Le noeud A est la racine de l'arbre.
- Les noeuds E, I, J, M, L et H sont des feuilles.
- Les noeuds B, C, D, F, G et K sont des noeuds intermédiaires.
- Si une branche relie un noeud n_i à un noeud n_j situe plus bas, on dit que n_i est un ancêtre de n_j.
- Dans un arbre, un noeud n'a qu'un seul père (ancêtre direct).
- Un noeud peut contenir une ou plusieurs valeurs.
- La hauteur (ou profondeur) d'un noeud est la longueur du chemin qui le lie à la racine.

II.9. VALIDATION D'UN MODELE DE PREDICTION

Après avoir construit un modèle de prédiction, il est important de le valider en tentant d'estimer les erreurs de classification qu'il fait, ou autrement, la probabilité que la classe prédite pour une donnée quelconque soit correcte.

L'erreur de classification d'un modèle de prédiction représente donc la probabilité que ce modèle ne prédise pas correctement la classe d'une donnée de l'espace de données.

Cette démarche de validation exige qu'il y ait deux types de jeu des données lors de la construction d'un modèle de prédiction :

- ✓ *Le Jeu de Données d'Apprentissage ou d'Entrainement* qui permet de construire le modèle de prédiction ;

- ✓ *Le Jeu de Données de Test* qui permet d'estimer les erreurs de classification : pour ces exemples, on suppose que leur classe n'est pas connue. On les classe avec le modèle de prédiction construit sur base des données d'apprentissage, puis on regarde s'ils sont classés correctement. Bien entendu, idéalement, l'intersection entre jeu d'apprentissage et jeu de test doit être vide.

II.9.1. Les Techniques de Validation Croisée

La Validation Croisée (Cross Validation, en Anglais) est une méthode qui permet de tester la fiabilité (précision) d'un modèle de prédiction sur base d'une technique d'échantillonnage. Parmi les techniques de validation croisée, on peut en énumérer trois :

1. *La Technique de test set validation ou holdout method ;*
2. *La Technique de k-fold cross-validation ou validation croisée a k-plis;*
3. *La Technique de leave-one-out cross-validation*.

II.9.1.1. Technique de Test Set Validation

Cette technique consiste à diviser l'échantillon de taille n en échantillon d'apprentissage (> 60 % de l'échantillon) et échantillon de test. Le modèle est bâti sur l'échantillon d'apprentissage et validé sur l'échantillon de test. L'erreur est estimée en calculant un test, une mesure ou un score de performance du modèle sur l'échantillon de test.

II.9.1.2. Technique de k-fold Cross-Validation

Cette technique consiste à diviser l'échantillon original en k échantillons, puis on sélectionne un des k échantillons comme ensemble de validation et les (k-1) autres échantillons constitueront l'ensemble d'apprentissage. On calcule ensuite l'erreur quadratique moyenne. Puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les (k-1) échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs est enfin calculée pour estimer l'erreur de prédiction.

II.9.1.3. La Technique de Leave-One-Out Cross-Validation

Cette technique est un cas particulier de la deuxième où k=M, c'est-à-dire que l'on apprend sur (M-1) observations puis on valide le modèle sur la énième observation et l'on répète cette opération M fois.

II.9.2. La Technique de Bootstrap

Le jeu des données d'apprentissage est constitué en effectuant M tirages avec remise sur l'ensemble des données. Ce qui implique que certaines données soient sélectionnés plusieurs fois, et d'autres jamais.

II.9.3. Evaluation de la qualité d'un modèle de prédiction

L'évaluation de la qualité d'un modèle de prédiction permet de tester son efficacité et aussi de le comparer aux modèles construits par d'autres techniques (algorithmes).

II.10. L'APPORT DE DATA MANING DANS LE DOMAINE DE MARKETING

Actuellement, nous constatons que l'accroissance des technologies, la mondialisation des marchés et le raccourcissement du cycle de vie des produits rendent dans une certaine mesure la concurrence plus rude.
C'est pourquoi le fort déclin de la publicité de masse illustre cette difficulté à gagner et à conserver des parts de marché en se focalisant uniquement sur le produit.

Dès les années 80, les techniques de marketing de masse n'apportant guère de résultats positifs pour les entreprises, elles cèdent leurs places à un marketing direct, orienté vers le client, qu'il faut comprendre, satisfaire et avec qui il faut communiquer « directement » afin d'optimiser le succès futur de l'entreprise. Le CRM s'inscrit clairement dans cette évolution et représente, d'une certaine manière, le dernier marketing direct.

C'est dans ce cadre que les applications du data mining permettent de réaliser :

- La grande distribution ;
- La vente par correspondance ;
- Le marketing direct ;
- La gestion de la relation client etc.

1. **Marketing direct** :

La gestion de la relation client (GRC) appelée également Customer Relationship Management (CRM) est un processus qui

consiste à gagner, à conserver, à élargir une clientèle. Sa stratégie est de placer le client au centre de préoccupations de l'entreprise en instaurant un dialogue, une relation de confiance et un respect mutuel avec les clients.

L'utilisation du data mining dans ce domaine permet de sélectionner parmi les clients principaux, ceux pour lesquels il est utile de leur envoyer un catalogue spécialisé en terme de ciblage. Cela augmente le taux de retour. Pour y arriver, il important de considérer l'historique des achats qui va permettre au décideur de déterminer quel client est susceptible d'acheter un article sur un catalogue spécialisé ?

2. **Centre d'appels :**

Un centre d'appel permet de mettre en place une opération de phoning pour relancer les clients suite à l'envoi d'un message. C'est en fait grâce au data mining que l'on parvient à déterminer vers quel clients / prospects il peut être utile d'effectuer une relance téléphonique en étudiant son comportement face aux relances.

L'entreprise s'apercevra ainsi que pour un client, une seule relance est inutile alors que pour d'autres, elle peut aller jusqu'à 5 relances. Le tout étant de ne plus généraliser les relances mais de les cibler.

3. **Fidélisation des clients :**

En mettant en place une solution de data mining, les entreprises vont essayer d'allonger la durée de vie d'un client en repérant les risques de son départ.

4. **Comportement des clients de grandes surfaces spécialisées :**

Les grandes boutiques utilisent les techniques d'associations de produits pour anticiper le comportement futur de ses clients. Un client qui achète une baignoire va probablement envisager l'achat de robinets. Par conséquent, les outils de data mining peuvent permettre de

sélectionner selon les achats antérieurs des clients vers qui l'envoi d'un mailing sera efficace.

5. **Comportement d'achat du client de grandes et moyennes surfaces :**

Les premières applications de Data Mining concernaient l'étude des tickets de caisse des clients de grande surface. Cela a permis de montrer que pour certaines catégories de clients les promotions mises en place pour des produits qu'ils avaient l'habitude d'acheter simultanément n'étaient pas efficaces et n'engendraient pas d'augmentation de chiffre d'affaires.

Le Data Mining semble aujourd'hui prendre de l'essor au sein de la fonction marketing. En effet, la mise en place des sites Web d'entreprises permettent de collecter de plus en plus d'informations qu'il va falloir exploiter au maximum.

Cependant, le data mining ne doit pas être considéré comme une solution miracle à l'ensemble des problèmes des entreprises. Il correspond à une avancée technologique qui permet de faire face au volume croissant des données collectées.

Mais, les entreprises devront instaurer un climat de confiance afin de ne pas porter atteinte à la vie privée des clients / prospects en exploitant les données collectées. Bref les objectifs de data mining peuvent être synthétisé en trois points ci-dessous :

- ✓ **Expliquer** : Le datamining pourra tenter d'expliquer un événement ou un incident indiscernable. Par la consultation des informations contenues dans l'entrepôt de données de l'entreprise, on peut être en mesure de formuler la question suivante : Pour qu'elle raison perd-t-on des clients pour un produit spécifique dans une région précise ? Tout en se basant sur des données connectées ou des mises en signification de paramètres liés, le datamining va essayer de trouver un certain nombre d'explication à cette question. Le Datamining va aider à trouver des hypothèses d'explications.
- ✓ **Confirmer** : Le datamining aidera à confirmer un comportement ou une hypothèse. Dans le cas où le

décisionnaire aurait un doute concernant une hypothèse, le datamining pourra tenter de confirmer cette hypothèse en la vérifiant en appliquant des méthodes statistiques ou d'intelligence artificielle.

- ✓ **Explorer** : Le datamining peut explorer les données pour découvrir un lien inconnu jusque-là. Quand le décisionnaire n'as pas d'hypothèse ou d'idée sur un fait précis, il peut demander au système de proposer des associations ou des corrélations qui pourront aboutir à une explication. Il est utopique de croire que le datamining pourrait remplacer la réflexion humaine. Le datamining ne doit être vu et utiliser uniquement en tant qu'aide à la prise de décision. Par contre, l'informatique décisionnelle dans son ensemble, et plus particulièrement le datamining permet de suggérer des hypothèses. La décision finale appartiendra toujours au décideur.

CHAPITRE III : PROBLEMATIQUE DE CHURN

III.1. RAPPEL SUR LA NOTION DE GESTION DE LA RELATION CLIENT

III.1.1. Historique & Définition

L'évolution actuelle des technologies, la mondialisation des marchés et le raccourcissement du cycle de vie des produits rendent la concurrence toujours plus rude. Il devient très difficile pour une entreprise de conserver sa part de marché en se basant uniquement sur les prix et les produits.

Le fort déclin de la publicité de masse illustre cette difficulté à gagner et à conserver des parts de marché en se focalisant uniquement sur le produit.

Dès les années 80, les techniques de marketing de masse (orienté produit) n'apportant guère de résultats satisfaisants pour les entreprises, elles cèdent leurs places à un marketing relationnel (direct), orienté vers le client, qu'il faut comprendre, satisfaire et avec qui il faut communiquer directement afin d'optimiser le succès futur de l'entreprise.

Afin de construire une relation avec le client dans le but d'aboutir à une fidélité de ce dernier, les entreprises s'orientent donc actuellement et de plus en plus vers la gestion de la relation client.

La Gestion de la Relation Client ou Customer Relationship Management (CRM) en anglais, se définit donc comme un ensemble d'outils et techniques permettant de renforcer la communication entre l'entreprise et ses clients afin d'améliorer la relation avec la clientèle, en automatisant les différentes composantes de la relation client. C'est donc une stratégie par laquelle une entreprise essaie de comprendre, anticiper et gérer les besoins de ses clients actuels et potentiels.

III.1.2. Objectifs du CRM

La mise en place d'un CRM permet d'intégrer le client dans son organisation, de connaître ses interlocuteurs afin de leur fournir une relation personnalisée et optimiser le contact client tout au long du cycle de vente. Les objectifs poursuivis par la gestion de la relation client sont essentiellement liés à :

- La Fidélisation des clients existants ;
- L'acquisition de nouveaux clients ;
- La Capitalisation sur les clients les plus profitables ;
- La Réduction des coûts.

III.1.3. Différents Types de CRM

Une gestion de la relation client peut être perçue sous trois dimensions interdépendantes : CRM Collaboratif, CRM Analytique et CRM Organisationnel.

1. CRM Collaboratif

Le CRM collaboratif fait référence aux moyens de communication au travers desquels l'entreprise entre directement en contact avec le client. Un des objectifs du CRM collaboratif est de donner une image positive et uniforme de l'entreprise au client et d'établir une relation de confiance avec lui en proposant un service personnalisé et de qualité. C'est également, dans cette même idée, l'occasion de récolter de très précieuses informations sur le client, qui seront stockées et permettront de mieux cibler ses besoins et ses attentes.

2. CRM Analytique

Ce type de CRM Analyse des données contenues dans le Data Warehouse ; et à partir de ces données, utilise différents outils d'extraction des connaissances servant de support à la prise de décision.

3. CRM Organisationnel

Le CRM opérationnel peut être défini comme l'automatisation et l'amélioration constante des processus de vente, de marketing et de service client.

III.1.4. Bénéfices Attendus du CRM

Les bénéfices attendus de la mise en place d'une solution CRM dans une entreprise peuvent être évalués au niveau du client, de la force de vente et de l'entreprise dans sa globalité.

- Au Niveau du client, on peut noter les apports suivants :
  - L'amélioration de la qualité des contacts ;
  - L'amélioration la fidélisation ;
  - La transformation du client en ambassadeur.

- Au niveau de la force de vente, on peut noter :
  - L'aide à la vente ;
  - L'accélération de l'intégration de nouveaux vendeurs ;
  - L'accélération du cycle de vente ;
  - L'augmentation du taux de transformation.

- Au niveau de l'entreprise, le CRM facilite :
  - La réduction des coûts ;
  - L'accroissement des résultats ;
  - La Réduction de l'attrition ;
  - L'amélioration de la qualité de l'information ;
  - L'augmentation de la valeur de l'entreprise.

De tous les bénéfices énumérés ci-haut, seule la réduction de l'attrition ou de churn sera abordé dans les lignes qui suivent, car étant l'objectif même du présent travail.

III.2. LE CHURN OU ATTRITION

III.2.1. Définitions

Le churn qui est né de la contraction en anglais des mots « change » et « turn », décrit le phénomène de perte d'un client. Ce phénomène peut subvenir de manière brutale, lorsque que le client choisit de rompre son contrat (attrition définitive) ou de manière discrète, lorsqu'il détourne progressivement (attrition relative).

Le taux d'attrition, quant à lui, désigne la proportion de clients perdus ou ayant changé de produit et service de la même entreprise, au cours d'une période donnée.
Ce terme est principalement utilisé dans les secteurs des télécommunications et bancaire, notamment autour de la fidélisation aux offres, mesurée par le taux de fidélité.

Le taux d'attrition concerne trois formes de changement :

- ✓ *abandon et résiliation* : le client n'utilise plus le type de produit ou de service ;
- ✓ *passage à la concurrence* : le client se tourne vers un produit directement concurrent ;
- ✓ *passage à une autre offre de l'entreprise* : le client passe à une offre différente, commercialisée par la même entreprise recouvrant aussi ses besoins (ex. passer de la téléphonie en présélection à la VoIP).

III.2.2. Types de Churn

De ce qui précède, on peut distinguer les types de churn ou d'attrition sur base de l'intention et de la destination du client.

- ➢ Selon la destination du client, on distingue :

    - **Le churn interne** : lorsque le client change de produit ou d'offre recouvrant aussi ses besoins tout en restant au sein de la même organisation ;
    - **Le churn externe** : lorsque le client quitte pour partir chez le concurrent.

- Selon l'intention du client, on note :

    - **churn actif ou délibéré :** lorsque le client décide de résilier délibérément son contrat pour passer à la concurrence. Ce qui peut être dû à : une insatisfaction de la qualité de service, des coûts trop élevés , une absence des plans de prix concurrentiels et des récompenses pour la fidélité des clients , un mauvais support , une manque d'information sur les raisons et prédiction de temps de résolution des problèmes de service , sa vie privée , etc…

    - **churn rotationnel ou accessoires** : lorsque le client résilie son contrat sans le but de passer à un concurrent. cette situation intervient lorsque le client se trouve dans l'incapacité d'exiger d'avantage le service, suite à des problèmes financiers, conduisant à l'impossibilité de paiement; ou à un changement de son emplacement géographique vers un endroit où le service de la société est indisponible.

    - **churn passive ou involontaire :** lorsque le client quitte le produit ou le service involontairement, par exemple en cas de décès ou de résiliation du contrat pour impayés.

Note

Le churn volontaire (actif ou rotationnel) est parfois compliqué à prédire. Et comme le churn accessoire ne représente qu'une petite fraction de l'ensemble des départs, il est particulièrement intéressant de prévoir et réagir en prenant des mesures appropriées pour empêcher le churn délibéré . Cependant, pour empêcher les départs volontaires des clients, l'entreprise a besoin de connaître les possibles churners et les éventuelles causes de leurs départs, afin de mieux appliquer les techniques de fidélisation.

III.2.3. Le Scoring de Churn

Le scoring de Churn ou d'attrition est une méthode de Data Mining qui permet d'observer les caractéristiques des clients afin de détecter ou prédire ceux qui vont partir à la concurrence, cesser d'utiliser une offre ou résilier un abonnement.

Cette méthode s'avère généralement utile lors qu'on souhaite :

- Expliquer les raisons de l'attrition ;
- Anticiper les départs à la concurrence et déclencher des actions pro actives ;
- En réaction lors d'un contact client, détecter une fragilité et agir en conséquence ;
- Fidéliser le client dans tous les cas.

III.2.3.1. Types de Scoring de Churn

Il existe deux types de scoring de churn :

- ✓ **Le scoring d'attrition descriptif** : lorsqu'on essaie de décrire le profil des clients qui vont arrêter de consommer un produit.
- ✓ **Le scoring d'attrition prédictif** : dans le cas où on prédit le profil va partir à la concurrence dans les X périodes qui suivent une date précise ou un évènement.

Généralement la forme du scoring dépend donc des objectifs de l'entreprise. Pour le cas de l'attrition, le scoring prédictif est plus adapté, car l'aspect descriptif peut être réglé avec pertinence par une analyse de la satisfaction client.

III.2.3.2. Processus de Scoring (prédiction) de Churn

Le processus de scoring de churn peut se résumer aux étapes suivantes :

1. Définition ou préparation des données de churn anciens ;
2. Séparation des données d'apprentissage et données de test ;
3. Construction du modèle de prédiction sur base des données d'apprentissage;

4. Utilisation des données de test sur le modèle construit ;
5. Evaluation de la performance du modèle ;
6. Utilisation de nouvelles données sur le modèle ;
7. Prédiction de possibles churners ;
8. Utilisation des prédictions pour des campagnes marketing

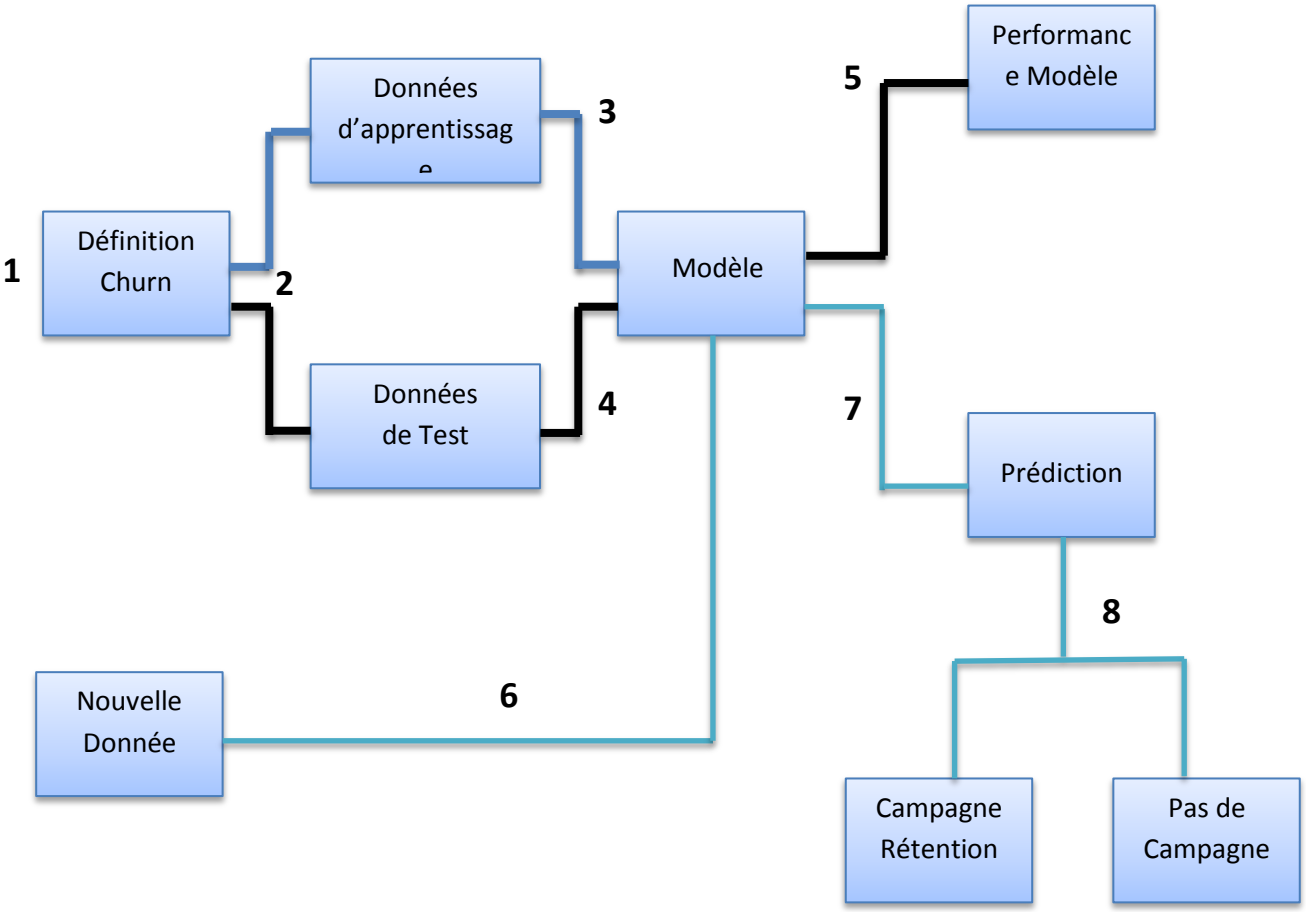


Figure 13: *Processus de Scoring de Churn*

CHAPITRE IV : DEVELOPPEMENT DU SYSTEME

Pour trouver solution au problème posé dans notre cahier de charge, il était important pour nous de commencer tout d'abord à faire la présentation de l'entreprise dans laquelle nous avions effectué nos recherches, et finalement nous passerons au développement du système

SECTION .1. PRESENTATION DE L'ENTREPRISE

IV.1.1. SITUATION GEOGRAPHIQUE

L'entreprise Orange RDC est située sur le boulevard du 30 juin à une centaine de mètre de l'ambassade de la France.

IV.1.2. HISTORIQUE

L'entreprise Orange RDC est une filiale à 100% du groupe Français Télécom Orange. Apres avoir racheté la société du groupe chinois Congo chine Télécom (CCT en sigle), il lance ses offres et services 3G en République Démocratique du Congo le 5 décembre 2012.

IV.1.3. OBJECTIF

Orange en RDC a pour est d'être l'opérateur préféré des congolais en leur apportant les services et les offres dont ils ont besoin et envie, en leur permettant d'aller plus loin et de communiquer mieux en toute simplicité. La promesse d'Orange RDC est de proposer au grand public, aux jeunes et aux professionnels une offre pour chacun, une offre qui leurs ressemble.

Orange vise à : démocratiser l'accès au mobile, avec des offres simples et proches, et à l'internet mobile 3G grâce à la connectivité internationale, favoriser et augmenter les usages de l'internet haut débit, dynamiser le secteur des technologies de

l'information et de la communication. Avec Orange, vous avez la possibilité de communiquer simplement, accéder à vos services préférés, partager des images et des musiques et s'ouvrir au monde.

IV.1.4. SERVICES

Être Orange, c'est vivre nos valeurs, d'audace, de dynamisme, de transparence, de simplicité et de proximité qui guident à chaque instant nos actions et nourrissent la relation avec nos clients.

Les services réseau de meilleure qualité offert par Orange, c'est grâce aux infrastructures modernes et évolutives. Il propose également une expérience client différente grâce à un réseau de distribution renforcée. Des boutiques inédites à Kinshasa et Lubumbashi offrant un accueil et une écoute de qualité avec une découverte sur-mesure des produits et des offres grâce à un conseil personnalisé.

Des points de ventes Orange proches des clients avec des partenaires distributeurs sur l'ensemble du territoire. Une nouvelle approche de la relation client avec 130 conseillers clients à votre écoute et à votre service 7j/7.Orange se charge d'accompagner le développement de nouveaux usages via une gamme d'offres et de services, adaptée pour chacun (voix, SMS, internet mobile haut débit 3G)

Orange et l'internet mobile haut débit pour tous : plus de liberté : un accès facilité et fiable plus de sensations : un partage d'images, de musique plus d'efficacité : une qualité aux standards Orange L'innovation pour vous Orange est au plus proche des préoccupations de ses clients.

L'offre de mobile payement Orange Money, qui arrivera prochainement, illustre le savoir-faire technologique du Groupe au service des besoins de chacun. 5 millions de clients dans 11 pays

peuvent, avec Orange Money, envoyer de l'argent à leur famille, payer leurs factures, percevoir leur salaire ou encore acheter avec leur mobile Orange.

IV.1.5. LE SOCIAL

Orange partenaire du football africain Orange est le principal partenaire des six compétitions majeures de la Confédération Africaine de Football (CAF) et ce jusqu'en 2016. Orange est aux côtés des supporters pour leur permettre de vivre une expérience enrichie et renouvelée de leur passion, à travers des offres et des services innovants.

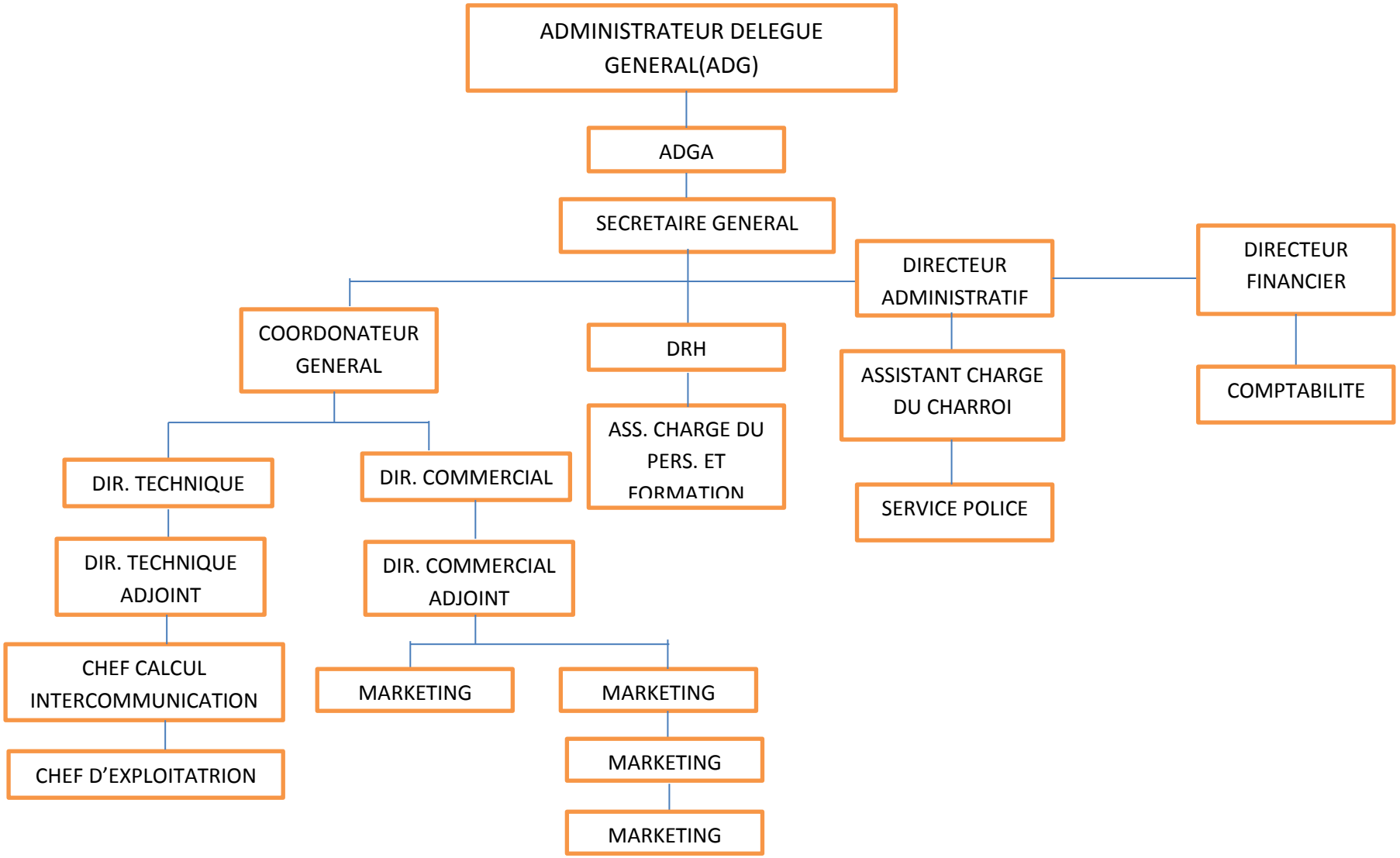
En RDC, Orange va accompagner pendant deux ans toutes les équipes nationales de football dont Les Léopards. Devenir un employeur de référence Orange RDC a aussi pour objectif de devenir un employeur de référence en mettant les femmes et les hommes au cœur de notre projet d'entreprise. La qualité sociale et de l'environnement de travail sont le préalable à toute autre performance.

Notre politique de formation, levier majeur de développement, nous a déjà permis de réaliser plus de 18 000 heures de formation en 2012. Un engagement pérenne et responsable Nous sommes une entreprise soucieuse de participer au développement économique et social dans les pays où nous opérons. Chaque nouvelle implantation est pour nous une réelle opportunité pour se rapprocher de notre cœur de métier : utiliser les nouveaux outils numériques pour faire avancer la solidarité, aider les personnes en difficulté pour entrer dans ce nouveau monde sont les objectifs d'Orange solidarité numérique.

La Fondation Orange agit au plus près des populations locales, partout où nous sommes présents, en s'adaptant au contexte de chaque pays, pour apporter plus de lien social et de solidarité, dans les domaines de la santé et du handicap, de l'éducation et de la culture. A ce jour, 15 filiales Orange ont créé une fondation locale qui pilote les activités de mécénat en cohérence avec le

groupe. La République Démocratique du Congo verra la naissance de la 16ème Fondation Orange au cours de l'année 2013.

IV.1.6. ORGANIGRAMME GENERAL D'ORANGE RDC



SECTION .2. DEVELOPPEMENT DU SYSTEME

Cette section, basée sur le développement de notre système décisionnel permettant de prédire le phénomène de churn de client. La méthodologie utilisée dans cette partie est le CRISP-DM, la quelle méthode qui va nous amener étape par étape à la réalisation dudit système.

IV.2.1. Préparation Des Donnés

L'ensemble de données utilisées pour la construction de notre modèle de prédiction a été simulé sur base d'un échantillon des données d'étude de churn Telecom repris dans l'ouvrage « ***Discovering Knowledge in Data:An Introduction to Data Mining.***

En effet, cet ensemble simulé de données contient un échantillon de quelques milieu d'enregistrements décrivant le comportement des abonnés pendant une période de temps, avec pour chaque abonné, 9 variables explicatives ou prédictive et une variable cible, tel que décrit ci-dessous :

Sortes	**Variables**	**Type de Valeur**
Variable Cible	Statut_churn	Booléen
Variables Prédictives	Longueur_compte	Entier
	Message_vocal	Entier
	Appel_jour	Entier
	Appel_soir	Entier
	Appel_nuit	Entier
	Appel_International	Entier
	Appel_service_client	Entiel
	Appel_servclient_disc	Intervalle codifié
	Inactivité	Entier

Tableau IV.2.1. variables d'analyse

Identification du type d'application de data mining

Bien que notre apprentissage porte un nom spécifique, nous sommes ici dans une problématique de classification automatique. De ce fait, les méthodologies de préparation de données comme entre autre le traitement des valeurs manquantes ou aberrantes, la sélection des meilleures variables pour la modélisation, ou en encore échantillonnage, peuvent être appliqué à toutes les problématiques de classification.

Objectif à atteindre

Le but d'un algorithme d'arbre de décision est de créer un ensemble de nœuds feuilles qui soient le plus pures possible avec des branches les plus courtes et les moins nombreuses possibles.

Problème de la construction d'un arbre : la scission

Deux problèmes vont intervenir pour construire l'arbre de décision :

- Le problème du nœud : quelle variable choisit-on à chaque nœud ?
- Le problème de la branche : quelles branches définit-on sous chaque nœud. Autrement dit quelles catégories choisit-on pour les prédicteurs ? Ces deux problèmes seront finalement liés : c'est **le problème de la scission**.

Visualisation de données

Les données utilisées pour la construction de notre modèle de prédiction a été simulé sur base d'un échantillon des données d'étude sur le turn-over des entreprises que nous avons téléchargé sur le portail du ***centre for Machine Learning and Intelligent Systems*** au lien ***http://archive.ics.uci.edu/ml/.***

Cette simulation contient un échantillon de quelques centaines d'enregistrements décrivant le comportement des abonnés pendant une période donnée. Pour chaque abonné.

Ci-dessous, nous présentons les données d'apprentissage enregistrées dans un classeur Excel.

Tableau de données

Statut_churn	Longueur_compte	Message_vocal	Appel_jour	Appel_soir	Appel_nuit	Appel_International	Appel_service_client	Inactivité	Tota
OUI	<20	0	0	0	0	0	0	95	17
NON	>20	0	1	0	2	0	3	6	21
NON	>20	4	0	4	2	1	1	1	19
NON	>20	0	0	2	0	4	3	11	17
NON	>20	3	4	3	0	1	4	3	16
NON	>20	1	1	2	3	0	4	25	17
NON	>20	0	1	0	2	3	1	24	25
NON	>20	1	4	2	4	2	2	24	14
NON	>20	1	0	1	0	2	4	21	14
NON	>20	1	2	3	1	0	3	22	23
NON	>20	1	2	0	0	0	2	19	15
NON	>20	0	2	4	3	0	2	6	23
NON	>20	3	4	2	0	2	2	10	25
NON	>20	3	1	2	2	1	1	19	19
NON	>20	3	3	3	3	0	2	21	21
NON	>20	0	4	1	3	3	3	3	15
NON	>20	3	1	1	0	4	1	7	25
NON	>20	0	2	2	1	0	1	17	17
NON	>20	2	1	3	3	0	0	1	18
NON	>20	1	3	2	1	1	2	24	18
OUI	<20	1	1	1	1	1	1	104	19
OUI	<20	1	1	1	1	1	1	105	15
OUI	<20	1	1	1	1	1	1	106	15
OUI	<20	1	1	1	1	1	1	107	19
OUI	<20	0	3	1	3	1	1	6	18
NON	<20	2	2	0	0	0	1	20	21

OUI	<20	4	2	3	2	2	4	8	19
OUI	<20	3	3	1	2	1	2	2	22
NON	<20	2	4	0	4	2	3	22	25
NON	>20	4	3	3	2	3	3	11	14
NON	>20	4	4	2	4	0	3	21	19
OUI	>20	0	0	0	0	0	0	115	22
OUI	>20	0	0	0	0	0	0	115	24
OUI	<20	0	0	0	0	0	0	115	20
OUI	<20	0	0	0	0	0	0	115	18
OUI	<20	0	0	0	0	0	0	115	14
OUI	<20	0	0	0	0	0	0	115	15
OUI	<20	0	0	0	0	0	0	115	16
OUI	<20	0	0	0	0	0	0	115	24
OUI	>20	0	0	0	0	0	0	115	19
NON	>20	2	2	1	0	3	2	25	25
NON	>20	0	0	4	1	4	4	25	19
NON	>20	1	1	1	2	3	4	12	22
NON	>20	2	1	3	4	3	3	21	25
NON	>20	3	4	1	4	3	0	14	22
NON	>20	4	4	1	1	3	2	9	16
NON	>20	4	3	2	3	4	2	3	25
NON	>20	1	2	4	1	0	3	19	14
NON	>20	1	4	4	3	0	0	2	17
NON	>20	1	1	4	3	2	0	2	22
NON	>20	2	1	4	1	3	4	1	25
NON	>20	3	1	0	3	3	0	24	15
NON	>20	3	1	4	2	0	3	17	24
OUI	>20	2	3	2	0	1	2	25	14
NON	>20	4	1	3	3	1	4	5	24

NON	>20	0	0	2	2	0	4	12	15
NON	>20	0	0	2	2	4	4	21	15
NON	>20	2	4	1	2	3	1	25	23
NON	>20	3	1	4	4	0	4	23	22
OUI	>20	2	4	0	0	0	4	22	22
NON	>20	3	1	0	1	3	0	9	20
NON	>20	1	2	1	4	1	0	13	17
NON	>20	3	4	3	3	3	4	19	20
NON	>20	1	4	4	3	4	2	12	19
NON	>20	1	2	1	1	3	4	6	23
OUI	>20	0	0	0	0	0	0	114	22
NON	>20	3	4	1	3	2	1	10	17
NON	>20	0	0	3	4	1	1	19	25
NON	>20	4	4	3	1	2	3	17	24
NON	>20	4	2	4	0	1	0	18	15
NON	>20	2	0	2	1	1	2	6	24
NON	>20	4	0	4	2	0	1	16	15
NON	>20	0	4	1	0	4	0	16	24
OUI	>20	1	0	0	0	0	0	114	25
NON	>20	1	4	2	3	3	3	23	25
NON	>20	2	2	2	1	3	0	9	17
NON	>20	4	0	2	2	3	2	7	16
NON	>20	3	3	1	4	3	3	15	14

Tableau IV.2.2. Tableau de données à traiter

APPLICATION

Outils Utilisés

L'application dédiée à la construction et évaluation du modèle de prédiction basé sur la technique des arbres de décision avec l'algorithme ID3 est développée en langage C#, sous l'environnement Visual Studio 2012 avec un Framework 4.5.

Outre le langage de programmation, l'utilisation de notre application sera rendu simple grâce à l'utilisation du logiciel open source Waikato Environment for Knowledge Analysis en sigle WEKA version 3.8.0, pour la visualisation graphique du modèle de prédiction qui est basé sur les techniques des arbres de décisions, développée en langage Java.

WEKA se compose principalement :

- ✓ de classes Java permettant de charger et de manipuler les données.
- ✓ de classes pour les principaux algorithmes de classification supervisée et non supervisée.
- ✓ d'outils de sélection d'attributs et de statistiques sur ces attributs.
- ✓ de classes permettant de visualiser les résultats.

Cet environnement peut être utilisé de 3 manières :

1. via l'interface graphique, pour charger un fichier de données, lui appliquer un algorithme, vérifier son efficacité.
2. invoquer un algorithme par ligne de commande.
3. utiliser les classes définies dans ses propres programmes pour créer d'autres méthodes, implémenter d'autres algorithmes, comparer ou combiner plusieurs méthodes.

Dans le cadre de ce travail, nous étudierons les possibilités 1 et 3.

IV.2.2. Fonctionnalités de l'application

Les principales fonctionnalités de l'application ne sont rien d'autres que les réalisations de tâches du processus de scoring de Churn décrit ci-haut :

On peut noter :

- Importation des données de Churn anciens (Fichiers Plats) ;
- Construction du modèle de prédiction avec les données d'apprentissage ;
- Utilisation de nouvelles données sur le modèle pour une éventuelle prédiction.
- Exportation des résultats dans un autre fichier, pour des analyses supplémentaire
- Construction de l'arbre de décision

Diagramme de Cas d'utilisation

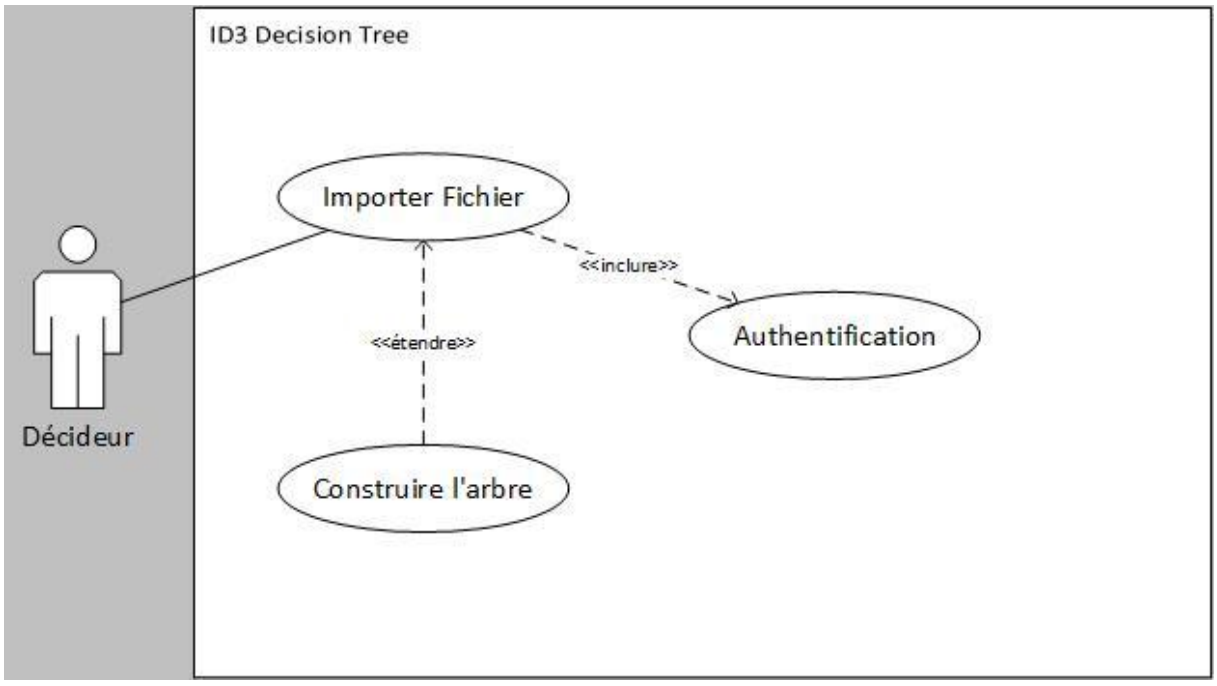


Figure IV.2.1 Cas d'utilisation

IV.2.3. Présentation de Quelques Interfaces

Importation de données à partir d'un fichier plat ;
Présentation du fichier plat ;

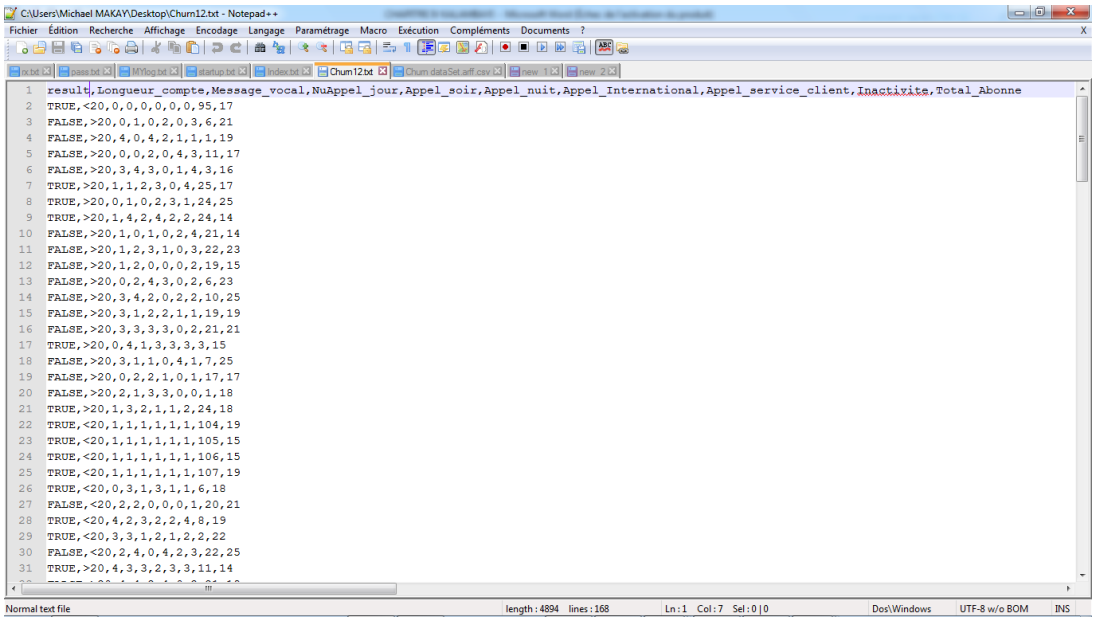


Figure. IV.2.2. fichier plat

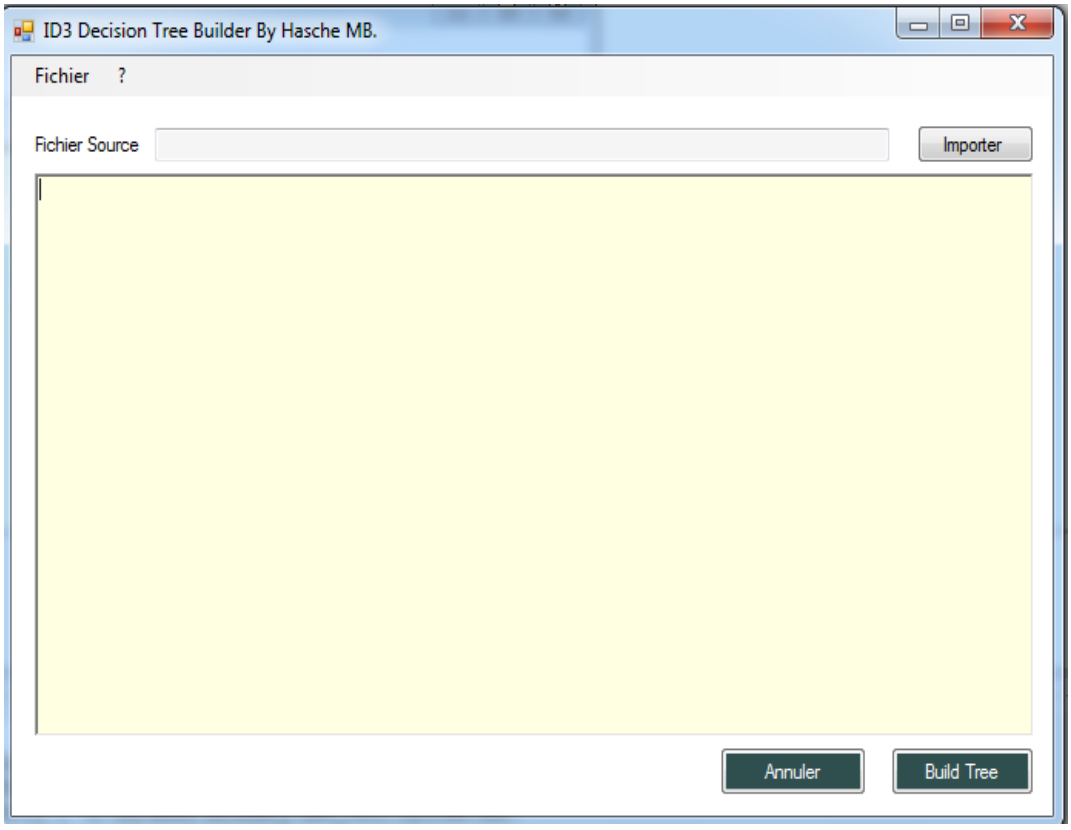


Figure. IV.2.3. page d'importation de données

Construction de l'arbre de décision avec ID3

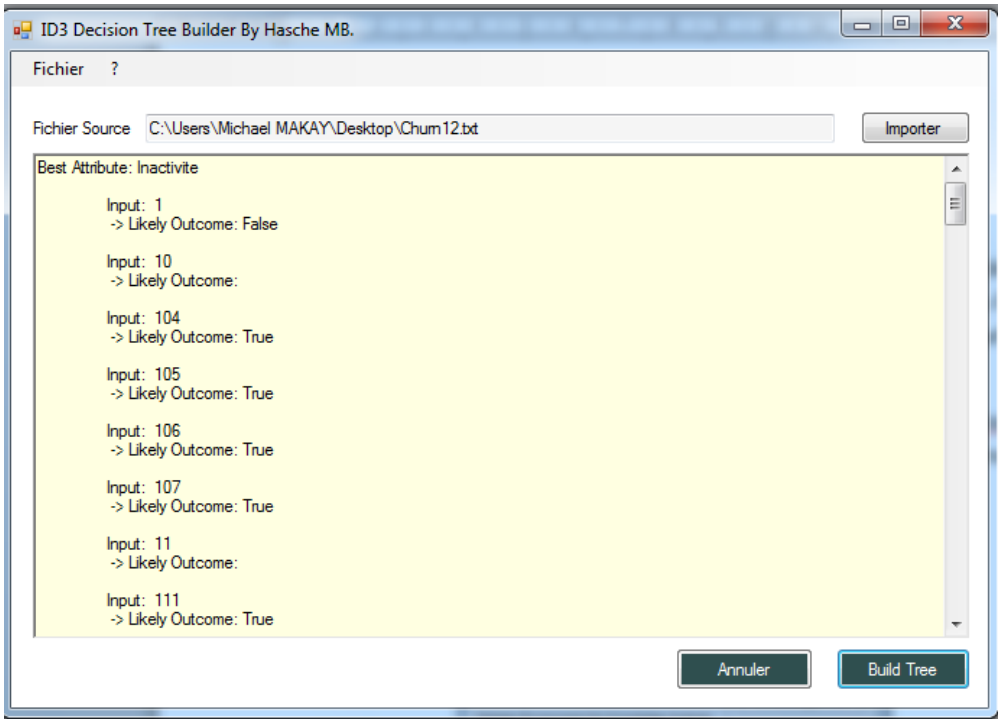


Figure *IV*.2.4. Construction de l'arbre de décision avec ID3

Comme présenté dans la figure 4, nous constatons que notre arbre de décision est créé mais reste difficile à interpréter. D'où nous faisons recours au logiciel Weka 3.8.0 présenté ci-haut, pour l'affichage graphique de notre arbre de décision.
Nous commençons par ajouter notre module dans le logiciel Weka, pour ensuite avoir une représentation graphique de l'arbre de décision. Ainsi nous commençons par présenter Weka3.8.0.

Figure *IV*.2.5. Interface utilisateur Weka

Après l'ajout de notre module ID3 decision Tree dans Weka, nous cliquons sur le bouton ***Explorer***, la fenêtre ci-dessous nous donne la possibilité d'importer le fichier .arff, et ainsi appliquer l'algorithme.

Fichier arff

WEKA utilise le fichier de format ***arff*** pour enregistrer les données. Un fichier *arff* est composé d'une liste d'exemples définis par leurs valeurs d'attributs.

Ci-dessous, nous présentons le fichier *arff* correspondant à notre fichier Excel présenté au tableau x.

```
@RELATION Churn

@ATTRIBUTE Statut_churn {TRUE,FALSE}
@ATTRIBUTE Longueur_compte {<20,>20}
@ATTRIBUTE Message_vocal NUMERIC
@ATTRIBUTE NuAppel_jour  NUMERIC
@ATTRIBUTE  NUMERIC
@ATTRIBUTE  NUMERIC
@ATTRIBUTE Appel_International NUMERIC
@ATTRIBUTE Appel_service_client NUMERIC
@ATTRIBUTE Inactivite NUMERIC
@ATTRIBUTE Total_Abonne NUMERIC

@Data
TRUE,<20,0,0,0,0,0,0,95,17
FALSE,>20,0,1,0,2,0,3,6,21
FALSE,>20,4,0,4,2,1,1,1,19
FALSE,>20,0,0,2,0,4,3,11,17
FALSE,>20,3,4,3,0,1,4,3,16
TRUE,>20,1,1,2,3,0,4,25,17
TRUE,>20,0,1,0,2,3,1,24,25
TRUE,>20,1,4,2,4,2,2,24,14
FALSE,>20,1,0,1,0,2,4,21,14
FALSE,>20,1,2,3,1,0,3,22,23
FALSE,>20,1,2,0,0,0,2,19,15
FALSE,>20,0,2,4,3,0,2,6,23
FALSE,>20,3,4,2,0,2,2,10,25
FALSE,>20,3,1,2,2,1,1,19,19
FALSE,>20,3,3,3,3,0,2,21,21
TRUE,>20,0,4,1,3,3,3,3,15
FALSE,>20,3,1,1,0,4,1,7,25
FALSE,>20,0,2,2,1,0,1,17,17
FALSE,>20,2,1,3,3,0,0,1,18
TRUE,>20,1,3,2,1,1,2,24,18
TRUE,<20,1,1,1,1,1,1,104,19
TRUE,<20,1,1,1,1,1,1,105,15
TRUE,<20,1,1,1,1,1,1,106,15
TRUE,<20,1,1,1,1,1,1,107,19
TRUE,<20,0,3,1,3,1,1,6,18
FALSE,<20,2,2,0,0,0,1,20,21
TRUE,<20,4,2,3,2,2,4,8,19
TRUE,<20,3,3,1,2,1,2,2,22
FALSE,<20,2,4,0,4,2,3,22,25
TRUE,>20,4,3,3,2,3,3,11,14
FALSE,>20,4,4,2,4,0,3,21,19
TRUE,>20,0,0,0,0,0,0,115,22
TRUE,>20,0,0,0,0,0,0,115,24
TRUE,<20,0,0,0,0,0,0,115,20
TRUE,<20,0,0,0,0,0,0,115,18
TRUE,<20,0,0,0,0,0,0,115,14
TRUE,<20,0,0,0,0,0,0,115,15
TRUE,<20,0,0,0,0,0,0,115,16
TRUE,<20,0,0,0,0,0,0,115,24
TRUE,>20,0,0,0,0,0,0,115,19
FALSE,>20,2,2,1,0,3,2,25,25
FALSE,>20,0,0,4,1,4,4,25,19
FALSE,>20,1,1,1,2,3,4,12,22
FALSE,>20,2,1,3,4,3,3,21,25
```

Figure. IV.2.6 Fichier arff Triaté

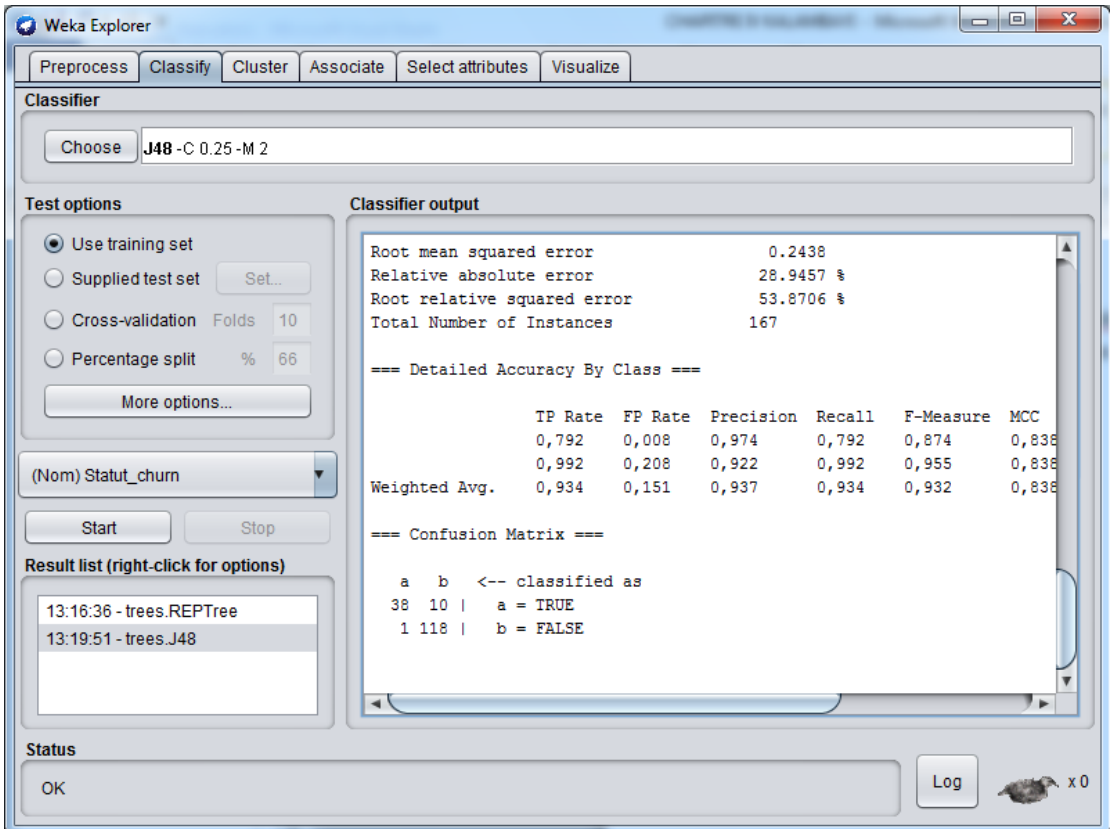


Figure IV.2.7. la fenêtre Explorer

L'affichage de l'arbre de décision

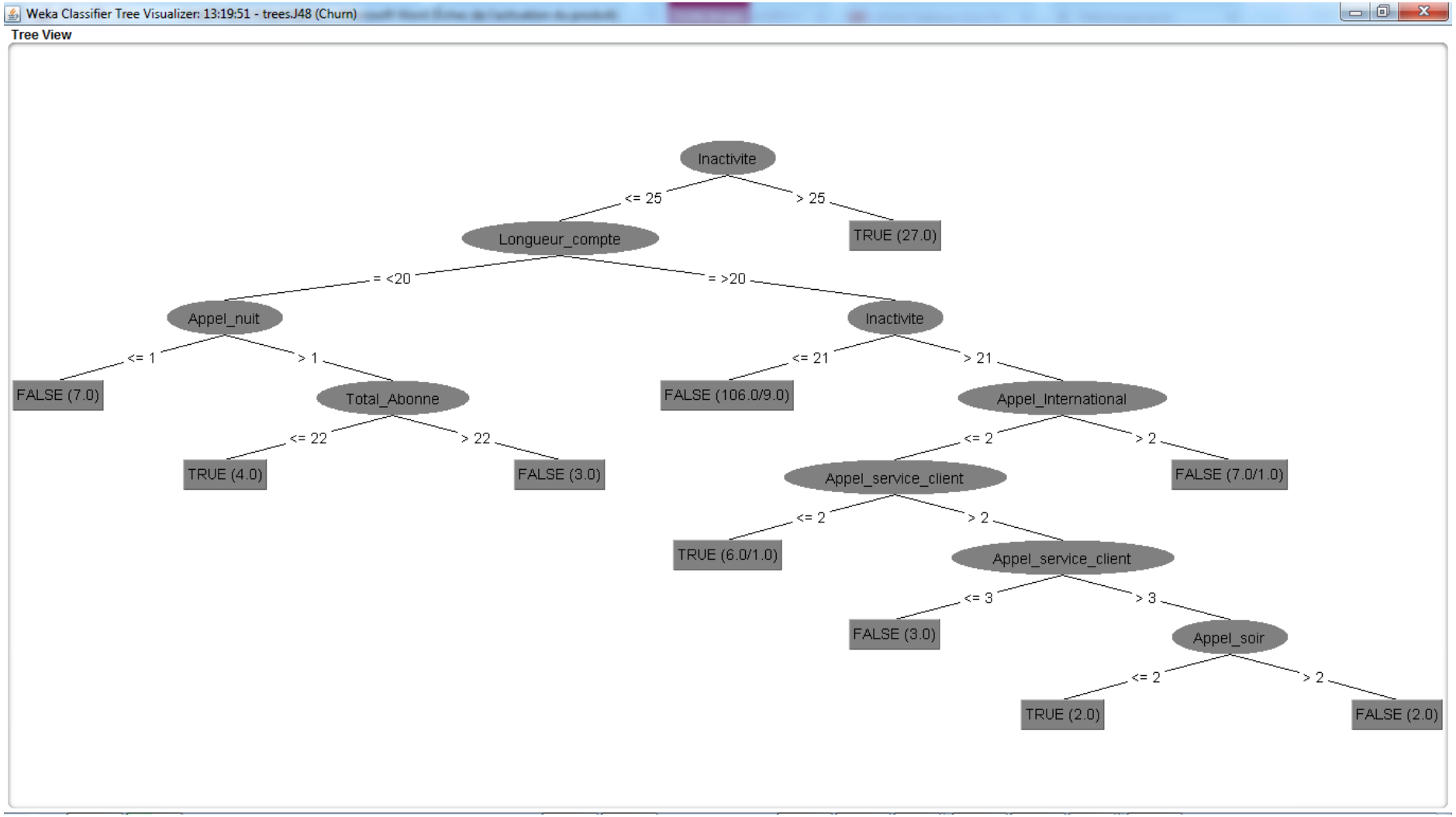
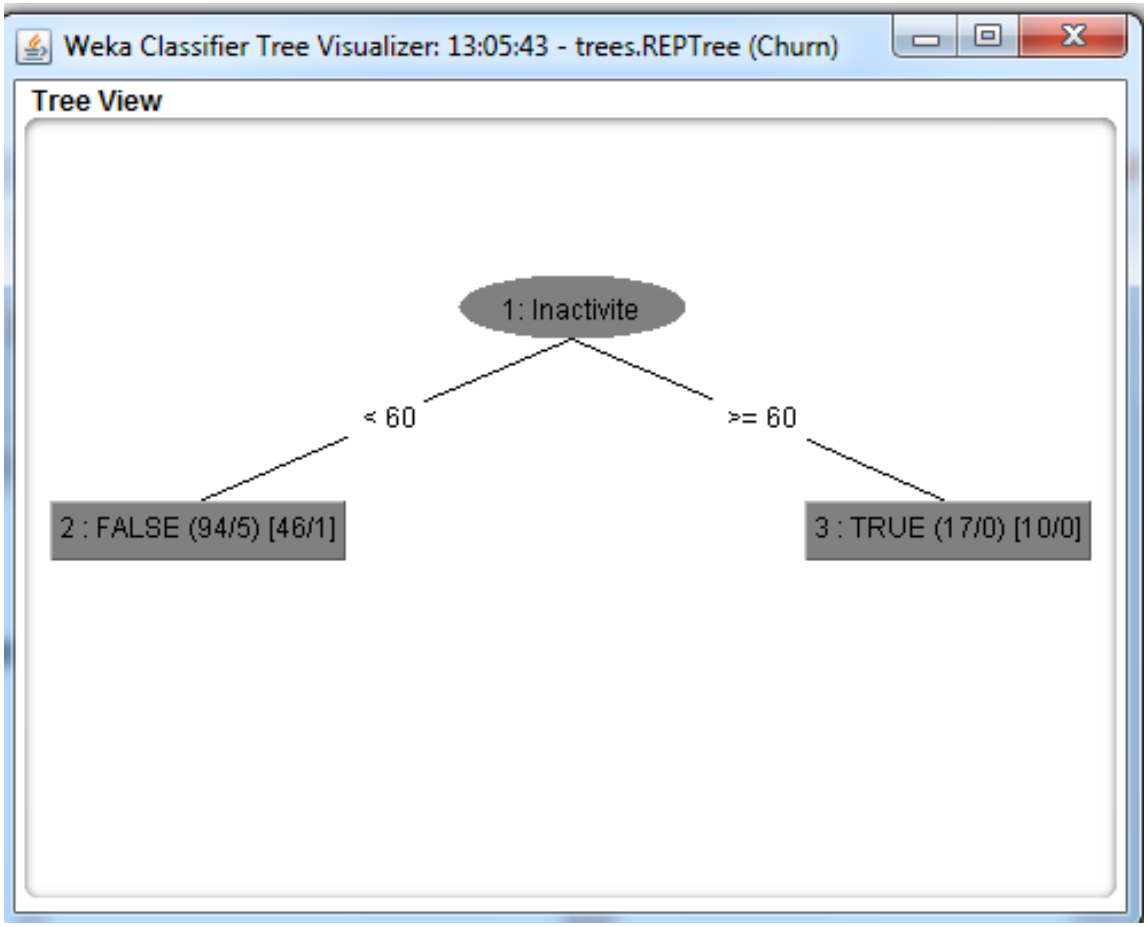
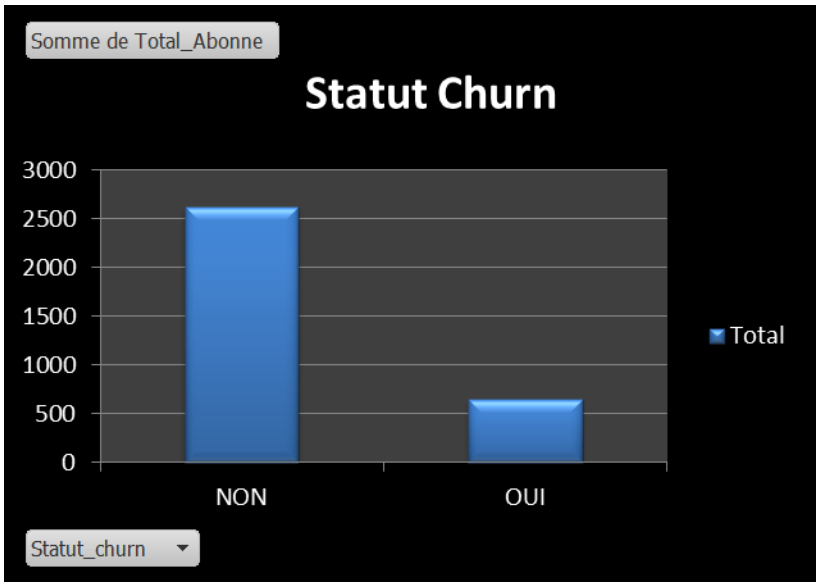


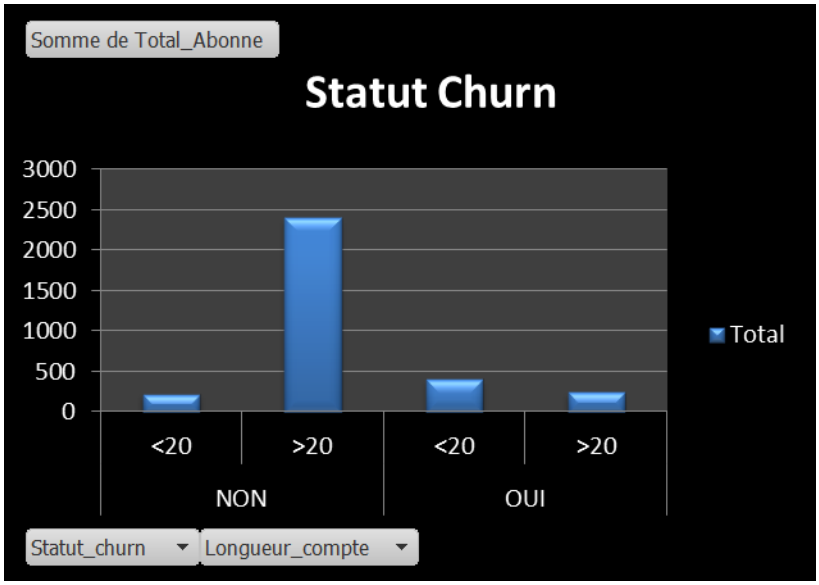
Figure IV.2.8.Arbre de décision



Interprétation

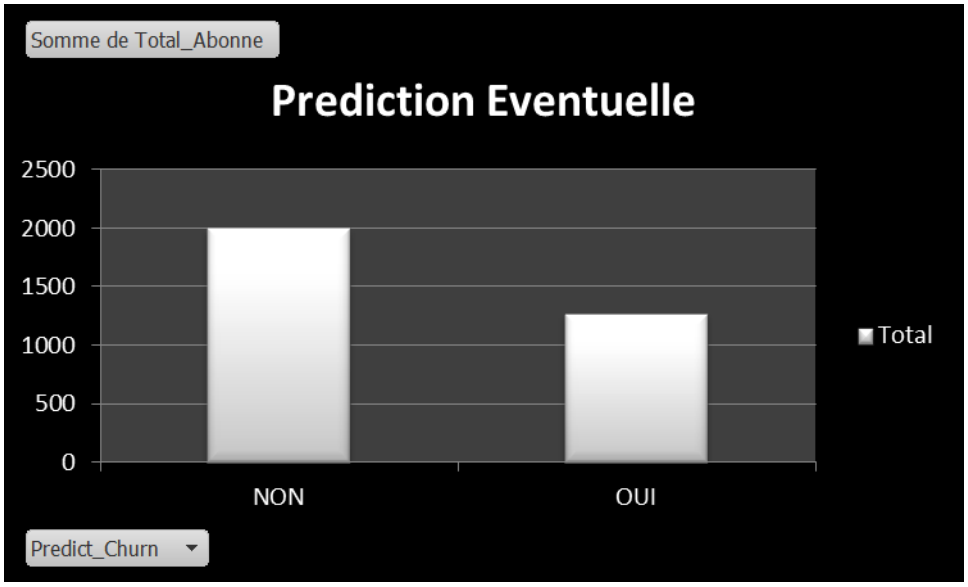


Nous voyons dans ce graphique que la situation est telle que sur 3000 abonnés dont dispose entreprise, il plus de 2500 clients fidèle et environ 500 churners.



Dans ce graphique, lors que nous observons le churn par rapport à la variable Longueur_compte nous constatons que le taux de client fidèle etant de 2500. Le client ayant le compte inférieur à 20 est environ 200 donc c'est très faible. Les taux de churners est d'environ 500. Les taux des churners ayant un compte supérieur 20 est inferieur 250.

Prediction Eventuelle



Nous voyons dans ce graphique que la prédiction pour un avenir proche nous donne une situation telle que sur 2500 abonnés restant le taux des abonnés fidèle serra en – dessous de 2000. Le taux de churners éventuelles serra supérieur à 100.

Ainsi nous estimons qu'il serra nécessaire de mener de campagne de rétention pour fidéliser les clients en anticipation de churn éventuel avenir en faisant recours au marketing direct.

CONCLUSION

Nous avons montré dans ce travail, les spécificités de l'attrition de la clientèle dans une entreprise de télécommunication et nous avons mise en place un outil d'extraction de connaissance permettant de maitriser le phénomène. Notre étude qui a porté sur « la mise en place d'un outil d'extraction de connaissance basé sur le technique de data mining appliqué sur l'analyse de churn (cas d'Orange RDC) ».Celle-ci a été menée dans l'objectif d'étudier les comportements des abonnés afin de prédire l'attrition de clients ; réduire le cout de la perte de la clientèle ; les opérations de marketing étant très couteuses, les décideurs ont besoin d'avoir la clarté sur les abonnés afin de savoir sur quels facteurs agir pour les fidéliser permettant ainsi une bonne prise de décision.

Pour y parvenir, nous avons subdivisé notre travail a quatre chapitres ; dont le premier donne bien évidement une idée sur les systèmes décisionnels. Le deuxième sur les différentes techniques de datamining ; celui-ci détaille les panoramas des techniques de datamining de résolution. Suivi du troisième sur la problématique de churn ; qui est en fait, une image nous aidant à comprendre le problème. et l'application qui présente les résultats trouvés par notre expérience.

Le logiciel coheris SPAD qui nous a aidé de faire une classification automatique hiérarchique. La classification est la technique de datamining retenue par notre étude afin de regrouper les abonnés par rapport à un critère de similarité dans une période de 6 mois pour s'imprégner sur quel facteur agir pour maitriser ce phénomène.

Notre réflexion se « termine » ainsi sur une ouverture, une enquête à poursuivre et à approfondir par des études ultérieures.