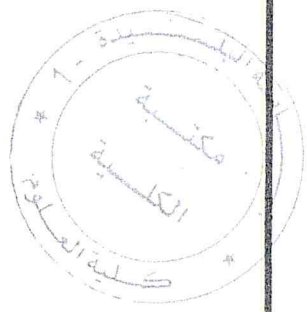
RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE SCIENTIFIQUE



UNIVERSITÉ SAAD DAHLAB BLIDA FACULTÉ DES SCIENCES
DÉPARTEMENT INFORMATIQUE

Mémoire de fin d'étude
Pour l'obtention d'un diplôme de Master en Informatique
Option : Systèmes Informatiques et Réseaux

Thème

Système de prédiction pour les services d'admission des hôpitaux dans un contexte Big Data

Réalisé par :

- ELFARROUDJI Zakaria Abderakib
- SIDI YAKHLEF Ayoub

Promoteur :

- Mme ZAHRA Fatma Zohra

Encadreur :

- Mme BOULKRINAT Nour El Houda

Organisme d'accueil :Centre de Recherche sur l'Information Scientifique et Technique

Jury :

Président : Mr DERRAR Hacen

Examinateur : Mme FAREH Messaouda

Date de Soutenance : le 30 /06/2018

Promotion : 2017/2018

Remerciements

Nos premiers remerciements vont à ALLAH le Tout Puissant.

C'est avec grand plaisir que nous réservons cette page, en signe de gratitude et de reconnaissance à tous ceux qui nous ont aidés à la réalisation de ce travail.

Nous remercions Mme Boulkrinat notre encadreur pour sa grande disponibilité, sa rigueur et professionnalisme qui n'a eu de cesse de nous inspirer et aussi pour la confiance qu'elle nous a accordée en proposant ce travail.

Nous remercions, notre promoteur Mme Zahra pour sa rigueur et la pertinence de ses jugements qui ont été très constructifs et nous ont permis de faire ce travail.

Nous remercions vivement les membres de jury pour nous avoir fait l'honneur d'accepter d'examiner notre travail.

Nous n'oublions pas non plus de remercier l'université de nous avoir donné cette opportunité de mettre en œuvre nos connaissance acquise durant notre parcours éducatif. Nous somme reconnaissants aussi à tous nos amis pour leurs dévouement et leurs amitié sans faille, ils nous ont beaucoup soutenus moralement.

Nous voudrons exprimer à nos proches toute notre gratitude : nos chers grands parents, nos très chers parents, nos frères et nos sœurs, nos oncles et nos tantes. Sans leur amour, leur soutien, leur confiance et leurs encouragements, nous n'y serons peut à être pas arrivés.

Résumé

L'analyse prédictive est devenue un domaine d'étude important tant pour les praticiens que pour les chercheurs, reflétant l'ampleur et l'impact des problèmes liés aux données à résoudre dans les organisations professionnelles contemporaines.

Notre travail consiste donc à utiliser l'analyse prédictive pour faire des prédiction pour les services d'admission des hôpitaux dans un contexte Big Data en utilisant les données de l'hôpital Mustapha Bacha.

Pour se faire nous avons utiliser deux méthodes de la technique des séries chronologiques à savoir la méthode ARIMA et la méthode de Holt-Winter pour prédire le nombre d'admission et la durée moyenne de séjour des patients.

Mots clés : Big Data, analyse prédictive, séries chronologiques , ARIMA, Holt-Winter.

Abstract

Predictive analytics has become an important area of study for both practitioners and researchers, reflecting the scale and impact of data-related problems in contemporary professional organizations.

Our job is therefore to use predictive analytics to make predictions for hospital admission services in a Big Data context using data from Mustapha Bacha Hospital.

To do so we used two methods of the time series technique namely the ARIMA method and the Holt-Winter method to predict the number of admission and the average length of stay of patients.

Keywords : Big Data, predictive data analysis, Time serie, ARIMA, Holt-Winter.

Table des matières

Table des figures

Liste des tableaux

1 Introduction générale

La donnée est la pierre angulaire de chaque organisation, sans quoi, une organisation ne peut pas fonctionner. Ces dernières années, des volumes considérables de données sont devenus de plus en plus varié (structuré, semi structuré et non structuré) sont créés tous les jours à partir d'une variété de source à savoir des données utilisateur générées automatiquement sur Internet. Réseaux sociaux, appareils mobiles, messagerie électronique, blogs, vidéos, transactions bancaires et autres interactions utilisateur pilotent désormais les campagnes Marketing, les études sociodémographiques, les enquêtes de polices et les intentions électorales. Ce tsunami de données est nommé Big Data.
Le Big Data a cessé d'être une tendance technologique pour devenir une réalité employée chaque jour par des millions d'entreprises dans le monde pour permettre la captation et l'exploitation avancées de ces données volumineuses et riches. Cela est devenu possible grâce à la révolution dans le domaine de la gestion des données. Les moteurs de base de données basés sur le standard SQL sont créés dans les années 1970 ont de bonnes performances lors du traitement de petites quantités de données relationnelles, mais ces outils sont très limités face à l'expansion des données en volume et en complexité. Le traitement MPP (Massively Parallel Processor) créé initialement au début des années 1980 a amélioré légèrement les indicateurs de performance pour les volumes de données complexes. Cependant, ce traitement n'a pas pu être utilisé pour le traitement des données non-relationnelles à expansion permanente. La nécessité de combler ce problème a guidé vers la naissance d'une nouvelle technologie capable de gérer cette importante masse de données, à savoir les outils du Big Data comme : hadoop, spark... etc
Ces dernières années, les méthodes d'analyse de données se sont développées de telle façon qu'elles permettent de synthétiser, de traiter et d'extraire des connaissances à partir d'une masse de données, ces méthodes facilitent aux usagers la lecture de ces données à travers plusieurs moyens comme les représentations graphiques, et les aident et les accompagnent à la prise de décision. L'analyse prédictive est l'une des méthodes les plus efficaces de l'analyse des données qui porte essentiellement sur l'analyse des données historiques et actuelles disponibles afin de prédire ce qui va se passer dans le futur, et ainsi, permettre aux usagers de contrôler le présent. Les entreprises utilisent les analyses prédictives de différentes façons en l'occurrence, le marketing prédictif, le data mining, les algorithmes de Machine Learning ou d'intelligence artificielle sont autant de manières d'optimiser les processus et de découvrir de nouvelles patterns statistiques. Pour faire simple, elles permettent aux ordinateurs d'apprendre du passé pour mieux effectuer certains business processes et délivrer de nouvelles informations sur le fonctionnement d'une entreprise.
Basée sur l'analyse d'énormes volumes de données, l'analyse prédictive est particulièrement prometteuse dans le secteur de la santé. Cette dernière est passée d'une technique avant-gardiste peu répandue à une arme concurrentielle dont la portée se développe rapidement.

Le service d'admission hospitalier est le premier point de contact du patient avec l'hôpital. Toute information nécessaire à la prise en charge du patient et à la gestion de l'hôpital transite par cette structure. Actuellement, le service d'admission au niveau de l'hôpital de Mustapha Bacha rencontre des difficultés dans la prise de décision relative au pilotage de son activité quotidienne. Dans ce travail, nous présentons une solution qui permet au service d'admission de l'hôpital (Bureau des entrées) de remédier à ses problèmes.

Conscient de l'enjeu majeur de l'augmentation continue de la masse de données au niveau des hôpitaux et de l'importance d'avoir des analyses prédictives permettant une meilleure prise de décision, nous nous intéressons dans ce projet à l'utilisation des technologies du Big Data afin

d'analyser le volume immense de données, pour prédire de nouvelles connaissances qui peuvent aider à mieux prendre en charge les patients au niveau des services d'admission.
Le principal objectif de ce travail est de mieux utiliser les données générées dans l'hôpital (données du service d'admission du Centre Hospitalo-universitaire Mustapha) dans le but d'améliorer le parcours et le traitement des patients, et d'optimiser l'utilisation des ressources hospitalières (humaines et matérielles) et permettre une meilleure prise en charge des malades, en fournissant des prédictions sur le nombre future des patients qui seront admis (pour la planification des ressources humaines et matérielles nécessaires) et la durée moyenne de séjour (permet aux décideurs de bien gérer la quantité du stock pharmaceutique ainsi que le manque en matière d'équipements sanitaires).

1.1 Problématique

Le service d'admission assure l'accueil, l'orientation des usagers, la prise en charge administrative des patients hospitalisés et la réalisation de la facturation des frais de séjour à l'hôpital. Le bureau des entrées est le premier lieu de gestion de la performance et de l'attractivité au niveau des établissements de santé, il est donc nécessaire d'optimiser l'organisation de ce bureau tout en répondant à l'enjeu stratégique d'amélioration de la qualité de la prise en charge des patients. Le service d'admission se sert actuellement des systèmes opérationnels pour le suivi des patients. Ces systèmes fournissent en continu des flux de données non traitées, non consolidées et mal comprises donc, couvre uniquement les besoins opérationnels et connaît un déficit incontestable en matière d'aide à la décision qui se présente comme suit : - L'hôpital dispose d'un volume important de données issues de différentes sources, avec manque de cohérence entre ces données, d'où le besoin d'un référentiel unique de données. - Les données sont généralement présentées sous format tabulaire et ne permettent pas de synthétiser l'activité hospitalière ou détecter les lacunes que rencontre le patient durant son parcours. - Retard dans l'établissement des rapports d'activité hospitalière. Ceci est dû à l'utilisation des outils non pertinents dans l'élaboration des statistiques. - Les services d'admission sont souvent mis à rude épreuve du fait qu'ils n'ont pas à l'avance des données ou des prévisions sur le nombre de cas d'admission pour une période donnée, ni sur les causes probantes et fréquentes de ces admissions.

1.2 Objectifs

Le principal objectif de ce travail est de mieux utiliser les données générées dans un hôpital dans le but d'améliorer le parcours et le traitement des patients dans un contexte d'optimisation des coûts. Plus particulièrement nous voulons analyser ces données pour améliorer la qualité des soins, de réduire les coûts et de permettre un meilleur suivi des patients. En conséquence, prédire par exemple : le nombre de patients potentiels pour : une admission un jour donné, une hospitalisation un jour donné, les causes d'admission (et donc les moyens à mettre en œuvre). L'objectif de notre travail est la mise en œuvre d'une plateforme Big Data pour la prise de décision, cette plateforme doit assurer : la collecte de grandes données, l'analyse via les techniques de prédiction (ARIMA), et enfin la visualisation des résultats pour le décideur. L'intérêt de cette plateforme est non seulement amélioré les coûts et les délais, mais aussi de bénéficier des avancées considérables du domaine Big Data (concepts, techniques et outils) et de les mettre au service des hôpitaux algériennes.

Organisation du mémoire

Ce mémoire est subdivisé en quatre chapitres : Le premier est consacré pour la présentation de Big Data et ses principe fondamentale ainsi que les différentes plateforme, quant au deuxième

chapitre il présente l'analyse prédictive ainsi que les différentes méthodes utilisées.
Dans le troisième chapitre nous présentons notre démarche conceptuelle abordée en décrivons les choix architecturaux et méthodologiques, ensuite nous entamons l'architecture fonctionnelle conçue et les différentes parties qui harmonisent son fonctionnement.
Le quatrième chapitre et le dernier est la phase finale de ce projet, il consiste en la concrétisation de la solution en projetant l'architecture fonctionnelle déjà développée sur notre cas d'étude.

Chapitre 1

Big Data

1 Introduction

L'utilisation des données a connu une croissance rapide de nos jours. En commençant par les choses très faciles que nous pouvons profiter tous les jours jusqu'à ce que nous l'utilisions pour analyser les processus d'affaires. Le développement de la technologie provoque l'effet des utilisations de données, par exemple, nous avions l'habitude d'utiliser le disque flottant qui a seulement 5Mb d'espace, parce qu'à ce moment-là, 5Mb était suffisant. Mais de nos jours, 5 Mo seulement peuvent conserver 1 fichier de musique, c'est pourquoi le développement de la technologie provoque l'effet des usages de données. Le montant de la transaction d'aujourd'hui a augmenté plus que les années précédentes. Par Internet, tout le monde peut être connecté et transférer des données, sans penser au temps et à la distance. La fin de ce développement est une très grande quantité de données recueillies au cas où nous l'appellerions Big Data.

2 Définition

Plusieurs définitions ont été proposées, citons deux parmi eux :

1. **Mike Barlow :** Big Data est lorsque la taille des données devient une partie du problème [1].
2. **Gartner :** Big Data sont des ressources d'information volumineuses, à haute vélocité et très variées qui exigent des formes innovantes et rentables de traitement de l'information pour améliorer la compréhension et la prise de décision [2].

3 Caractéristiques du Big Data

Les 5V du big data font référence à cinq éléments clés à prendre en compte et à optimiser dans le cadre d'une démarche d'optimisation de la gestion du big data. Ces 5V sont (figure 1.1) :

- **Volume** : Les volumes de données à collecter et analyser sont considérables et en augmentation constante[3].
- **Variété** : Les données peuvent prendre des formes très variées et très hétérogènes (voix, données faciales, données transactionnelles, web analytics, textes, images, etc.)[4].
- **Vitesse** : De plus en plus souvent, les données doivent être collectées et traitées en temps réel, comme par exemple Healthcare monitoring : des capteurs surveillent l'activité du corps humain,toute mesure anormale exige une réaction immédiate [5].

— **Véracité** : La véracité ou fiabilité des données est notamment menacée par les comportements déclaratifs (sur formulaires), par les diversités des points de collecte, par la multiplication des formats de données et par l'activité des robots et faux profils innombrables sévissant sur Internet[6].

— **Valeur** : Dans un contexte d'infobésité, il s'agit d'être capable de se concentrer sur les données ayant une réelle valeur et étant actionnables [7].

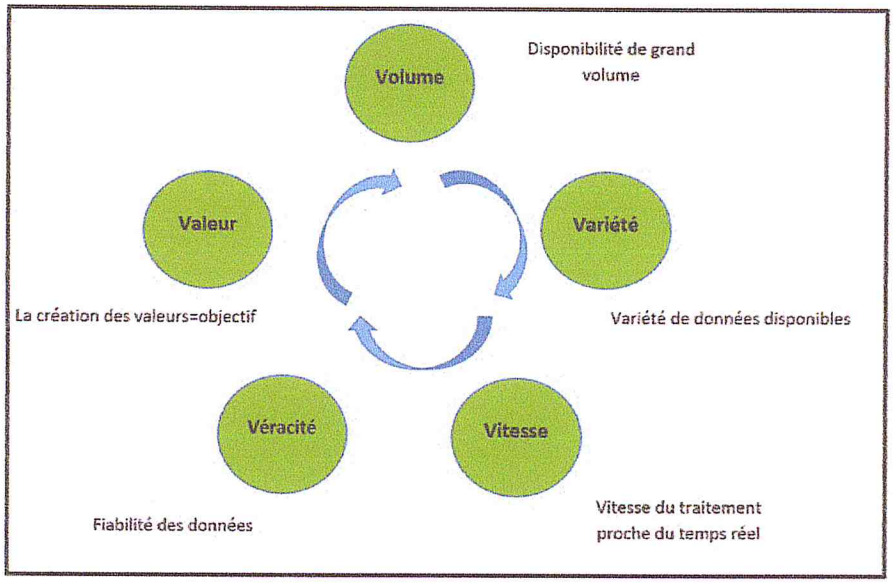


FIGURE 1.1: Caractéristiques du Big Data [3]

4 Domaines d'application du Big Data

Le Big Data couvre de nombreux domaines d'applications telles que l'industrie, les banques, l'assurance, le transport, loisirs et le télécom. Des exemples sont cités ci-dessous

4.1 Marketing

Les responsables du marketing ont toujours consacré une part importante de leur budget à la recherche de nouvelles données sur le comportement du consommateur et sur les moyens de renforcer l'image de marque tout en affinant leurs modèles de segmentation client. Par conséquent, ils sont parmi les premiers au sein de l'entreprise à considérer les données internes de l'entreprise comme une mine d'or potentielle, plutôt que comme une ressource contraignante à gérer. Ainsi, le marketing a vocation à devenir une force motrice au sein de l'entreprise dans l'émergence et le développement d'outils permettant de gérer les Big Data afin d'en dégager un avantage concurrentiel notable [8].

4.2 Surveillance

Depuis la fin des années 90, nous sommes entrés dans l'ère du renseignement. Satellites, drones, espionnage, intrusion sur des serveurs gouvernementaux, fichage de la population, écoutes téléphoniques, surveillance des mails et transactions bancaires sont autant de moyens mis en œuvre par les états au nom de la défense du territoire et de la protection des citoyens contre

toute menace ou attaque. De milliards de données non structurées sont ainsi collectées sous forme d'images (photos, empreintes), d'enregistrements audio ou vidéo, de schémas ou de cartes, de documents cryptés qu'il faut pouvoir stocker, trier en fonction de la pertinence et analyser afin d'en faire ressortir des informations critiques .
Le Big Data aide à résoudre efficacement des enquêtes policières (analyser des indices, trouver une corrélation entre plusieurs affaires), ou prévenir un attentat (suivre les déplacements d'un suspect, reconnaissance faciale sur des vidéos, interception de correspondance par mails entre terroristes). Il permet de réduire le temps de résolution des affaires et d'en augmenter le taux de résolution. Alors nous pouvons prédire jusqu'à 90 pourcent des attaques des terroristes[9].

4.3 Santé

La quantité de données issue de la prise en charge d'individus dans un cadre sanitaire, médical, médicosocial ne cesse d'augmenter, de même que le nombre de sources de données disponibles. Si l'on associée constat aux évolution technologiques, chaque individu peut ainsi espérer bénéficier d'une médecine prédictive, préventive, personnalisée et participative.
Cette révolution place le 'Big Data' au coeur du système de santé tant l'origine de ces données est variée [10].

- Dossiers médicaux électroniques dans les cabines médicaux, les hopitaux, les centres d'imagerie, les laboratoires, les pharmacies et d'autres structures productrices de soins.
- Données micro-économiques en rapport avec les dépenses de santé, la consommation de soins, de produit de santé (SNIIRAM[1], PMSI[2]...) et plus généralement les bases de données publiques médico-administratives.
- Données de saisies par les individus : préférences personnelles, niveau de satisfaction, historique de consommation, données d'auto-mesure, etc.

La gestion de ces données massives est un important levier pour une meilleurs compréhension des maladies, du développement des médicaments et du traitement des patients.
Une étude de l'institut McKinsey a ainsi mis en avant en 2013 les principaux secteurs de santé pour lesquels le 'Big Data' apporterait de réels bénéfices [11] :

- La prévention ciblée permet d'évoluer vers un mode de vie toujours favorable à la préservation de l'état de santé des individus.
- L'aide au diagnostiques et à la mise en place de soins adaptés à chaque patient tout en assurant sa sécurité, en tendant ainsi vers la médecine personnalisée.
- L'optimisation du médicament pour obtenir l'impact clinique attendu, et l'optimisation des ressources médicales par la mise à disposition de professionnels adaptés au cas du patient.
- La maîtrise des coûts pour une qualité de soin égale ou meilleure, en automatisant les remboursements et la détection de fraude.

5 Architecture du Big Data (Lambda)

Il existe aujourd'hui un nombre important d'architectures big data, l'architecture Lambda, l'architecture kappa ou l'architecture Zeta, regroupées sous le nom de traitement polyglotte (Polyglot Processing). Dans ce qui suit, nous allons nous intéresser à l'architecture Lambda qui est

1. SNIIRAM=Le Système National d'Information Interrégimes de l'Assurance Maladie.
2. PMSI=Programme de Médicalisation des Systèmes d'Information

la plus répandue en ce moment.

Le but de l'architecture Lambda est de fournir un modèle de traitement presque temps réel sur des volumes importants de données, en proposant un nouveau modèle de calcul. Ce modèle essaie de trouver l'équilibre entre la tolérance aux pannes, les contraintes de latence (latence très faible pour les lectures/écritures) et le débit des disques durs en se basant à la fois sur les traitements batch qui fournissent des vues batch et les traitements temps réel qui fournissent des vues, puis les joint avant leur présentation [12].
L'architecture Lambda est indépendante de la technologie, et se base sur le précalcul des résultats, puis à les récupérer dans une base et les envoyant au demandeur. Elle est composée de trois couches (Figure 1.2).

1) Couche Batch : mode de fonctionnement classique des applications big data type Hadoop, cette couche est responsable de deux choses : récupérer les données et les stocker en format brut (AS-IS : pour pouvoir répondre à de nouveaux besoins métier sans impacter les données initiales) dans des puits de données (Data Lake en anglais), et lancer périodiquement des traitements sur les données, pour précalculer les résultats sous forme de vue logique. Le résultat est ensuite stocké typiquement dans des bases en lecture seule et les mises à jour remplacent les vues logiques précalculées. Cette couche peut être implémentée à l'aide de Apache Hadoop, MapReduce et Spark.

2) Couche de vitesse en temps réel (Speeding) : traite les nouveaux flux de données en temps réel, sans aucun prétraitement (correction des jeux des données). Cette couche minimise la latence et fournit en temps réel des vues avec les données les plus récentes. C'est un fonctionnement en mode continu unitairement pour chaque nouvelle donnée. Les résultats (les vues temps réel) fournis par cette couche ne sont pas aussi fiables, que ceux de la couche batch. Cette couche peut être implémentée à l'aide de Apache Storm ou Spark Streaming.

3) Couche de service (Serving) : rend exploitables les résultats précalculés par la couche batch et la couche temps réel, pour effectuer des requêtes à la volée (ad hoc). Cette couche peut être implémentée à l'aide des technologies NoSQL Apache HBase, Cassandra, et ElasticSearch qui permettent de merger les vues batch et les vues temps réel.

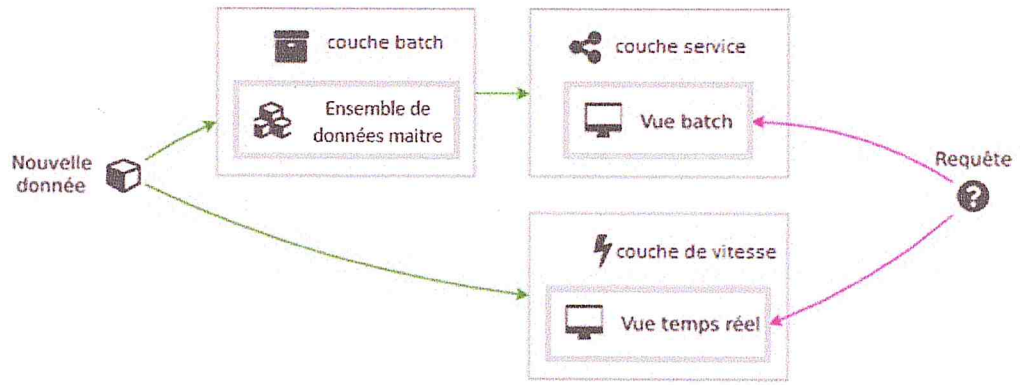


FIGURE 1.2: Architecture Lambda [12]

6 Avantages de l'architecture Big Data

Plusieurs avantages peuvent être associés à une architecture Big Data, nous pouvons citer par exemple :

- **Evolutivité (scalabilité) :** Lorsque l'on parle de scalabilité, il est courant d'opposer deux approches :
  - **La scalabilité horizontale :** qui consiste en l'ajout de ressources semblables à celles déjà en place dans le pool utilisé ou en la suppression d'unités de ressources au sein de ce pool.
    - ajouter de nouveaux volumes de stockage sur une machine virtuelle est de la scalabilité horizontale
  - **la scalabilité verticale :** qui consiste en la modification des caractéristiques de la ressource.
    - remplacer un volume de stockage par un volume plus gros est de la scalabilité horizontale
- **Performance :** Grâce au traitement parallèle des données et à son système de fichiers distribués, le concept Big Data est hautement performant en diminuant la latence des requêtes.
- **Coût faible :** Le principal outil Big Data à savoir Hadoop est en Open Source, en plus on n'aura plus besoin de centraliser les données dans des baies de stockage souvent excessivement chère, avec le Big Data et grâce au système de fichiers distribués les disques internes des serveurs suffiront.
- **Disponibilité :** On a plus besoin des RAID disques, souvent coûteux. L'architecture Big Data apporte ses propres mécanismes de haute disponibilité.

7 Big Data et DataWarehouse

Fondamentalement, le Big Data et le Data Warehouse ont le même concept, les Big Data consistent en une sorte d'information traitée à plus grande échelle. Le Big Data se concentre sur le volume et les données non structurées, l'entreposage de données se concentre sur les données structurées et la traçabilité. La technologie qui conviendra le mieux à l'organisation dépend

de nombreux facteurs. L'utilisation de la technologie traditionnelle d'entrepôt de données pour analyser les données générées par les capteurs n'est probablement pas une bonne idée en raison du volume élevé de données, tout comme l'utilisation de la technologie Big Data pour effectuer des rapports réglementaires n'est pas une bonne idée. Big Data est un terme appliqué à des ensembles de données dont la taille dépasse la capacité des outils couramment utilisés pour capturer, gérer et traiter les données dans un délai écoulé tolérable. cependant Data-warehouse est une collection de data marts[3] représentant des données historiques provenant de différentes opérations de l'entreprise. Ce qui signifie que Big Data est une collection de données volumineuses d'une manière particulière mais que Data-warehouse collecte des données provenant de différents départements d'une organisation. Cependant Data-warehouse nécessite une technique de gestion efficace. Conceptuellement, ils ne sont identiques qu'à un seul facteur : ils collectent une grande quantité d'informations.

8 Solutions de Big Data

Dans le domaine du Big Data il existe trois types de solution qui sont : les solutions Open Source, les solutions Hybride et les solutions Propriétaires. Nous allons présenter dans ce qui suit ces trois types en donnant des exemples de chaque type.

8.1 Solutions Open Source

Des entreprises innovante comme Google, Yahoo, Facebook ont décidé de rendre des projets qui ont été interne open source. Le tableau suivant présente quelques exemples de technologies open source utilisées pour la gestion des données massives et dont l'origine est un développement interne :

3. Data mart= un sous-ensemble d'un data warehouse destiné à fournir des données aux utilisateurs

Société	Technologie développée	Type de technologie
Google	BigTable	Système de base de données distribuée propriétaire reposant sur GFS(Google File System). Technologie non open source mais qui intègre HBase qui est open source
	MapReduce	plate-forme de développement pour traitement distribués
Yahoo	Hadoop	Plate-forme Java destinée aux application distribuée et à la gestion intensive des données. Issue à l'origine de Google BigTable, MapReduce et Google File System
	S4	Plate-forme de développement dédiée aux application de traitement continu des flux de données
Facebook	Cassandra	Base de données de type NoSQL et distribuée
	Hive	Logiciel d'analyse de données utilisant Hadoop
Twitter	Storm	Plate-forme de traitement de données massives
	FlockDB	Base de données distribuée de type graphe
LinkedIn	Kafka	Système distribué de gestion des messages
	SenseiDB	Base de données temps réel distribuée et semi-structurée
	Voldemort	Base de données distribuée destinée aux très grosses volumétries

TABLE 1.1: Solutions open source du Big Data

Dans ce qui suit nous allons détailler quelques solutions open sources les plus répondus :

8.1.1 Hadoop

Hadoop est un framework open source utilisé pour le stockage distribué et le traitement de grands ensembles de données en utilisant le modèle de programmation MapReduce. Il se compose de grappes d'ordinateurs construites en utilisant du matériel de base [13].
Hadoop est composé des éléments suivants (figure 1.3) :

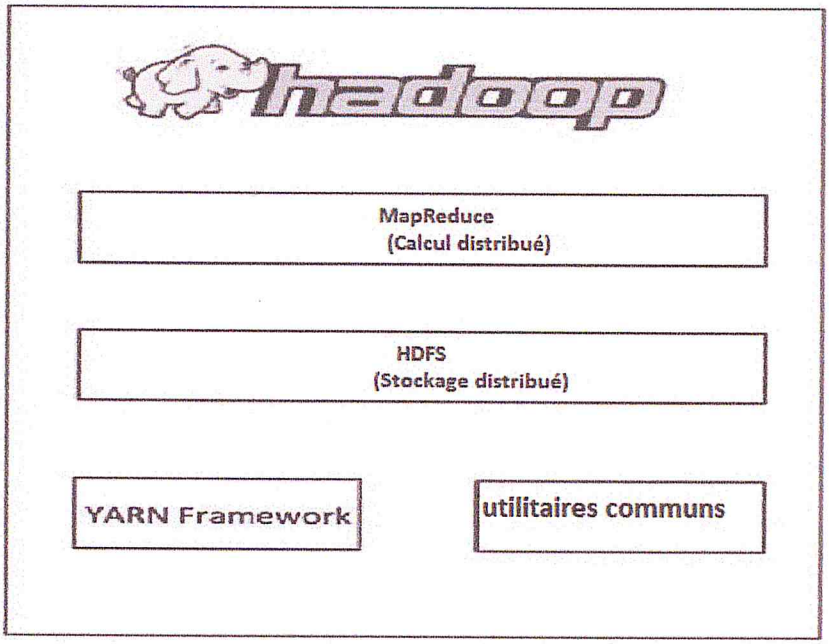


FIGURE 1.3: Architecture d'Hadoop [13]

— **HDFS** : Le système de fichiers distribués Hadoop (HDFS) est basé sur le système de fichiers Google (GFS) et fournit un système de fichiers distribué conçu pour fonctionner avec du matériel de base. Il présente de nombreuses similitudes avec les systèmes de fichiers distribués existants. Cependant, les différences par rapport aux autres systèmes de fichiers distribués sont significatives. Il est très tolérant aux pannes et conçu pour être déployé sur du matériel à faible coût. Il fournit un accès à haut débit aux données d'application et convient aux applications ayant de grands ensembles de données.

— **MapReduce** :c'est un modèle de programmation parallèle pour l'écriture d'applications distribuées conçues par Google pour un traitement efficace de grandes quantités de données (ensembles de données de plusieurs téraoctets), sur une grande échelle.
Le framework Hadoop inclut également les deux modules suivants :

— **Hadoop Common :** il s'agit des bibliothèques Java et des utilitaires requis par d'autres modules Hadoop.

— **Hadoop YARN :** il s'agit d'un cadre pour la planification des travaux et la gestion des ressources de cluster.

8.1.2 Cassandra

Apache Cassandra est une base de données NoSQL massivement évolutive. Les racines techniques de Cassandra peuvent être trouvées aux entreprises reconnues pour leur capacité d'efficacement gérer des Big Data comme Google, Amazone et Facebook. La figure 1.4 représente l'architecture de Cassandra

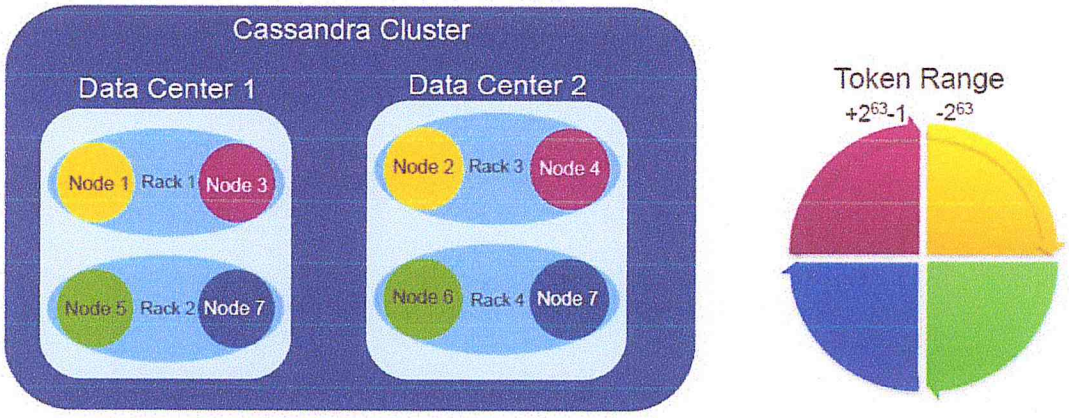


FIGURE 1.4: Architecture de Cassandra [15]

Architecture Cassandra

L'architecture Cassandra est composée de :

- **Node :** une instance de Cassandra (processus Java).
- **Partition :**unité de données ordonnée et réplicable sur un noeud identifié par un jeton.
- **Partitioner :**il s'occupe de distribuer les données entre les nœuds.
- **Rack :** ensemble logique de noeuds.
- **Data Center :** ensemble logique de racks.
- **Cluster :**ensemble complet de nœuds qui correspond à un seul anneau à jeton complet.

8.1.3 KNIME

Aussi appelé Konstanz Information Miner, il s'agit d'une plate-forme d'analyse de données open source. Il intègre divers composants pour l'exploration de données et l'apprentissage automatique grâce à son concept modulaire de pipeline de données [15].

Architecture KNIME

L'architecture de KNIME a été conçue avec trois principes en tête.

- **Cadre visuel et interactif :** Les flux de données doivent être combinés par simple glisser-déposer à partir de diverses unités de traitement. Les applications personnalisées peuvent être modélisées via des pipelines de données individuels.
- **Modularité :** Les unités de traitement et les conteneurs de données ne doivent pas dépendre l'un de l'autre pour permettre une distribution facile du calcul et permettre le développement indépendant de différents algorithmes. Les types de données sont encapsulés, c'est-à-dire qu'aucun type n'est prédéfini, de nouveaux types peuvent facilement être ajoutés avec des moteurs de rendu et des comparateurs spécifiques aux types. Les nouveaux types peuvent être déclarés compatibles avec les types existants.
- **Extensibilité facile :** Il devrait être facile d'ajouter de nouveaux nœuds ou vues de traitement et de les distribuer via un simple mécanisme de plugin sans nécessiter de procédures compliquées d'installation / désinstallation.

8.1.4 Spark

Apache Spark est largement considéré comme le successeur de MapReduce pour le traitement de données à usage général sur les clusters Apache Hadoop.
Comme les applications MapReduce, chaque application Spark est un calcul autonome qui exécute du code fourni par l'utilisateur pour calculer un résultat.
Comme pour les travaux MapReduce, les applications Spark peuvent utiliser les ressources de plusieurs hôtes. Cependant, Spark présente de nombreux avantages par rapport à MapReduce.
Dans MapReduce, l'unité de calcul de plus haut niveau est un job. Un job charge des données, applique une fonction de carte, la mélange, applique une fonction de réduction et écrit les données dans un stockage persistant.
Dans Spark, l'unité de calcul de plus haut niveau est une application. Une application Spark peut être utilisée pour un job en batch unique, une session interactive avec plusieurs jobs ou un serveur à longue durée de vie satisfaisant continuellement des demandes. Un job Spark peut consister plus q'une seule Map and Reduce.
MapReduce démarre un processus pour chaque tâche. En revanche, une application Spark peut exécuter des processus en son nom même si elle n'exécute pas de job.
En outre, plusieurs tâches peuvent s'exécuter dans le même exécuteur. Les deux se combinent pour permettre un temps de démarrage des tâches extrêmement rapide et un stockage de données en mémoire, ce qui permet d'obtenir des performances plus rapides que MapReduce [13].

L'exécution d'une application Spark implique des concepts d'exécution tels que le pilote (driver), l'exécuteur, la tâche (task), le job et la scène (stage). Comprendre ces concepts est essentiel pour l'écriture rapide et efficace des programmes Spark (figure 1.5).
Lors de l'exécution, une application Spark est mappée à un seul processus de pilote et à un ensemble de processus exécuteurs répartis sur les hôtes d'un cluster.
Le processus du pilote gère le flux de jobs et les tâches planifiées, il est disponible pendant toute la durée de l'application. Généralement, ce processus de pilote est le même que le processus client utilisé pour lancer le job, mais lorsqu'il est exécuté sur YARN, le pilote peut s'exécuter dans le cluster. En mode interactif, le shell lui-même est le processus du pilote.
Les exécuteurs sont responsables de l'exécution du job, sous la forme de tâches, ainsi que du stockage des données que vous mettez en cache. La durée de vie de l'exécuteur dépend de l'activation de l'allocation dynamique. Un exécuteur a un certain nombre d'emplacements pour

les tâches en cours d'exécution, et exécutera plusieurs simultanément tout au long de sa durée de vie.

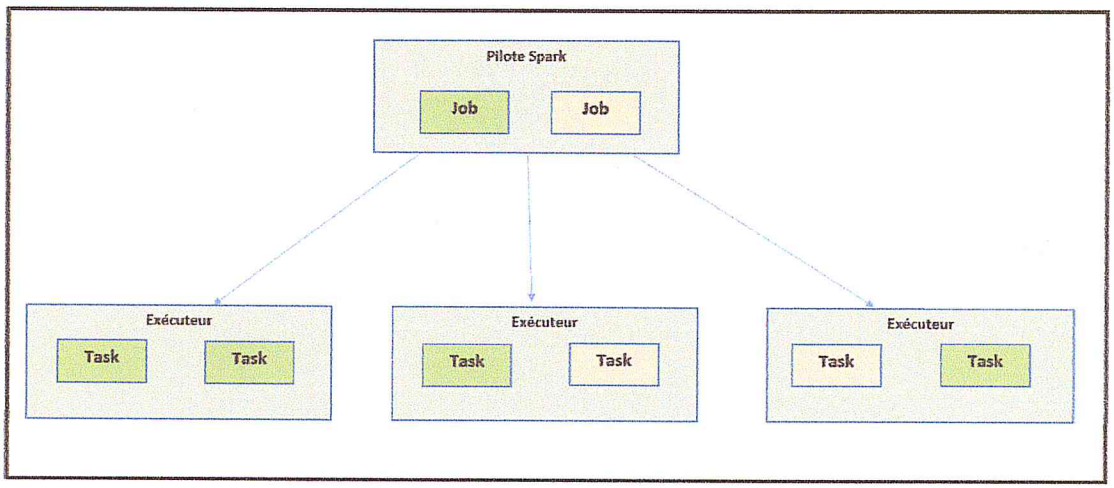


FIGURE 1.5: Architecture de Spark [13]

L'appel d'une action dans une application Spark déclenche le lancement d'un job pour l'exécuter. Spark examine l'ensemble de données dont dépend cette action et formule un plan d'exécution. Le plan d'exécution assemble les transferts de données en étapes. Une étape est une collection de tâches qui exécutent le même code, chacune sur un sous-ensemble différent des données [13].

8.1.5 Caractéristiques / avantages et inconvénients de chaque solution

Le tableau suivant présenter les avantages et inconvénients de chaque solution avec leur caractéristiques.

	Caractersistiques	avantages	inconvénients
Hadoop	• Code en java et c • Grande partie de stokage HDFS • Utilisation des nœuds dans un cluster. • Il transfère le code emballé dans différents nœuds pour traiter les données en parallèle. • Composé de : Hadoop Common, HDFS, Hadoop YARN and Hadoop MapReduce.	[+] Simple. [+] Accessible. [+] Disponible. [+] Adapté à tous les cas d'usages.	[-] Plus lent que les solutions MapR.
Cassandra	• Tolérant aux pannes • Décentralisé • Cohérence réglable. • Support MapReduce. • La réplication de plusieurs centres de données .	[+] Tolérance aux panne. [+] Décentraliser. [+] Élastique. [+] Haute disponibilité.	[-] Lenteur. [-] Pas d'interface graphique. [-] Difficultés à l'utilisation. [-] Limitation de la taille des données.
Spark	• structure spécifiques • Gestion des Graphes. • Machine Learning(Apprentissage automatique) .	[+] Rapide. [+] Simplifie le développement. [+] Plusieurs modes de déploiement. [+] Différents modes de stockage.	[-] Coûteux : nécessite beaucoup de RAM. [-] Latence plus élevée.

KMINE	• Code en java Eclipse. • Crées flux de données,exécuter sélectivement tout ou partie. • Inspecter les modèles, les résultats et les vues interactives du flux.	[+]Présentation graphique.	[-] Lenteur. [-] Compliquer.

TABLE 1.2: Caractéristiques, avantages et inconvénients des solutions open source

8.2 Solutions Hybrides et propriétaires

Bien qu'il existe des solutions open sources, il existe d'autre solutions propriétaires, citons parmi eux :

8.2.1 IBM

L'entreprise IBM (International Business Machines) a développé une plateforme Big data qui permettra de relever tous les défis métier du Big Data. Cette plateforme associe les technologies classiques bien adaptées aux tâches structurées et répétitives aux nouvelles technologies complémentaires tournées vers la vitesse et la flexibilité et donc idéales pour l'exploration, la reconnaissance de données et l'analyse de données non structurées [18].

Les fonctionnalités de la plateforme IBM

— Analyse Hadoop : traite et analyse tout type de données sur différents clusters de serveurs courants.

— Stream computing : permet l'analyse continue de volumes massifs de flux de données en continu avec des temps de réponse inférieurs à la milliseconde.

— Entrepôts de données : fournit des connaissances opérationnelles approfondies grâce à l'analyse intégrée à la base de données.

— Intégration des informations et la gouvernance : permet de comprendre, nettoyer, régir et distribuer des informations pour vos initiatives métier critiques.

8.2.2 Pivotal HD

la première plate-forme combinant la technologie de framework d'analyse open source Hadoop avec sa base de données massivement parallèle (de GreenPlum) et la technologie de base de données SQL en mémoire [19].
Pivotal HD permet de :

— Analyser les données de tous types - données structurées, semi-structurées et non structurées.

— Tirer parti des compétences SQL existantes et des outils associés pour effectuer des analyses complexes et des requêtes interactives sur des données à l'échelle du pétaoctet.

— Construire, exécuter et itérer des algorithmes d'apprentissage automatique et de science des données à l'échelle pour glaner des idées prédictives.

— Développer et soutenir des applications intelligentes pilotées par les données qui opérationnalisent les informations Big Data en fournissant des informations dans leur contexte.

8.2.3 MapR

La plate-forme de données convergées MapR intègre Hadoop, Spark et Apache Drill avec des fonctionnalités de base de données en temps réel, un streaming d'événements global et un stockage d'entreprise évolutif pour alimenter une nouvelle génération d'applications Big Data. La plate-forme MapR offre une sécurité, une fiabilité et des performances en temps réel de niveau professionnel tout en réduisant considérablement les coûts matériels et opérationnels des applications et de données les plus importantes [20].
MapR offre le meilleur compromis entre haute performance et scalabilité, tout en maximisant la facilité d'usage [21].

8.2.4 Caractéristiques/avantages et inconvénients

Le tableau suivant contiens les principaux caractéristiques, avantages et inconvénients de chaque solution.

	Caractersistiques	avantages	inconvénients
IBM	• analyse hadoop. • stream computing. • entrepôt de donnée. • Intégration des informations et la gouvernance.	[+] Intégration parfaite. [+] Distribution de la version Hadoop standard.	[-] Plus cher que ses concurrents.
Pivotal HD	• analyse hadoop. • stream computing.	[+] Des ajouts de nombreuses fonctionnalités depuis sa marketplace. [+] Distribution standard de Horton-Works [4].	[-] le manque de documentations.

MAPR	• workers : liste de nœuds Hadoop capables de traiter des taches MapReduce. • master : la gestion des taches. • client : lance le traitement MapReduce (souvent nommé driver).	[+] La plus rapide comparatif Hadoop. [+] Prend en charge les opérations en temps réel. [+] Intégration aisée et fiabilité.	[-] Interface console moins facile.

TABLE 1.3: Caractéristiques, avantages et inconvénients des solutions propriétaires

9 Conclusion

Le Big Data est un domaine qui évolue très rapidement, et qui a plusieurs domaines d'application où le domaine de santé ne fait pas l'exception.
Dans ce chapitre, nous avons mis l'accent sur les différentes notions relatives au concept du Big Data, nous avons défini les Big Data comme étant un problème à gérer. Dans ce qui suit nous allons nous intéresser au méthodes de gestion de cette masse de données qui sera l'axe principal du chapitre suivant.

Chapitre 2

Analyse prédictive

1 Introduction

Les entreprises du monde entier génèrent des ensembles de données gigantesques, notamment des transactions de vente, des enregistrements de transactions boursières, des descriptions de produits, des promotions des ventes, des profils d'entreprise et des performances,et les commentaires des clients. Cette croissance explosive du volume de données disponible est le résultat de l'informatisation de notre société et du développement rapide de puissants outils de collecte et de stockage de données.
Ce corpus de données explosif, largement disponible et gigantesque fait de notre temps l'ère des données. Des outils puissants et polyvalents sont indispensables pour découvrir automatiquement des informations précieuses à partir des énormes quantités de données et transformer ces données en connaissances organisées. Cette nécessité a conduit à la naissance de l'analyse de données et du Data Data mining.

2 Définitions de Data Mining

Le Data Mining, littéralement "forage ou fouille de données", est l'application des techniques de statistiques, d'analyse des données et d'intelligence artificielle à l'exploration et l'analyse de grandes banques de données informatiques, en vue d'en extraire des informations nouvelles et utiles [22].
Gardarin donne la définition suivante [23] : par analogie à la recherche des pépites d'or dans un gisement, la fouille de données vise à extraire des informations cachées par analyse globale et à découvrir des modèles, appelés motifs, difficiles à percevoir directement du fait du volume important des données, du nombre de variables à considérer et enfin du fait qu'il y ait des hypothèses imprévisibles.

3 Données, informations et savoir dans le Data Mining

— **Données :** Les données sont tout type d'entrées pouvant être traités par un ordinateur. Nous distinguons trois types de données :

1. Les données opérationnelles ou transactionnelles : ce sont les données ou il y a quotidiennement des changements
2. Les données non opérationnelles, telles que les ventes industrielles, les données prévisionnelles, les données macro-économiques [24].

3. Les métadonnées, à savoir les données concernant les données elles-mêmes, telles que les définitions d'un dictionnaire de données [24].

— **Informations :** C'est la transformation de ces données non structurées en utilisant des patterns, associations et relations en quelques choses qui a un sens. Donc l'information est une donnée interprétée.

— **Savoir :** Les informations peuvent être converties en savoir à propos de patterns historiques ou des tendances futures.

4 Les méthodes de data mining

Il existe différentes techniques et méthodes d'analyse de données citons :

4.1 Classification

La classification consiste à étudier les caractéristiques d'un nouvel objet pour lui attribuer une classe prédéfinie. Les objets à classifiés sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jour chaque enregistrement en déterminant un champ de classe. La tâche de classification est caractérisée par une définition de classes bien précise et un ensemble d'exemples classés auparavant. L'objectif est de créer un modèle qui peut être appliqué aux données non classifiées dans le but de les classifiées [25]. Types de modèles de classification :

— Classification par induction de l'arbre de décision.

— Classification bayésienne.

— Réseaux de neurones.

— Machines vectorielles de support (SVM).

— Classification basée sur les associations.

4.2 Clustering

Le clustering est un processus de partitionnement d'un ensemble de données (ou d'objets) en un ensemble de sous-classes significatives, appelées clusters [26].
La Classification non supervisée appelée clustering peut être considéré comme l'identification de classes d'objets similaires. En utilisant des techniques de regroupement, nous pouvons identifier davantage les régions denses et clairsemées dans l'espace objet et découvrir le schéma de distribution global et les corrélations entre les attributs de données [27].

Les types de méthodes de clustering sont[28].

— Méthodes de partitionnement.

— Méthodes hiérarchiques d'agglomération (séparatives).

— Méthodes basées sur la densité.

— Méthodes basées sur la grille.

— Méthodes basées sur un modèle.

La différence entre la classification supervisée et la classification non supervisée (le clustering) et que la première consiste à apprendre une méthode de prédiction de la classe d'instance à partir d'instances pré-étiquetées (classifiées),comme le montre la figure 2.1.

Par contre la deuxième consiste à trouver le regroupement naturel d'instances ayant reçu des données non étiquetées comme le montre la figure 2.2.

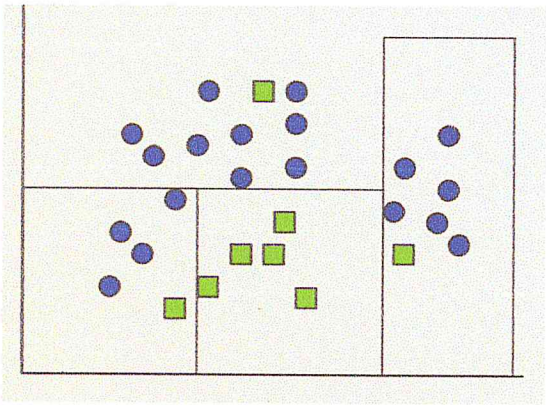


FIGURE 2.1: Classification supervisée

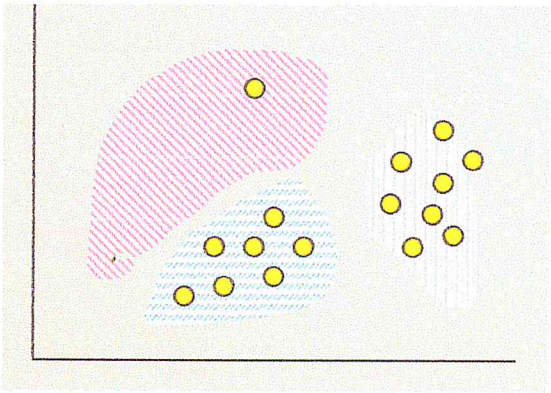


FIGURE 2.2: Classification nom supervisée

4.3 Régression

L'analyse de régression est une méthodologie statistique qui est la plus souvent utilisée pour la prédiction numérique, bien que d'autres méthodes existent également. La régression englobe également l'identification des tendances de distribution à partir des données disponibles. La classification et la régression doivent être précédées d'une analyse de la pertinence, qui tente d'identifier les attributs significativement pertinents pour le processus de classification et de régression. Ces attributs seront sélectionnés pour le processus de classification et de régression. D'autres attributs, non pertinents, peuvent alors être exclus de la considération [22].

4.3.1 Régression linéaire

C'est la forme la plus simple de la régression,elle utilise la formule d'une ligne droite (y = mx + b) et détermine les valeurs appropriées pour m et b pour prédire la valeur de y sur la base d'une valeur donnée de x (figure 2.3)[22].
Fondamentalement, un modèle de régression linéaire est utilisé pour montrer ou prédire la relation entre deux variables ou facteurs. Le facteur qui est prédit (le facteur pour lequel l'équation résout) s'appelle la variable dépendante. Les facteurs qui sont utilisés pour prédire la valeur de la variable dépendante sont appelés les variables indépendantes [29].

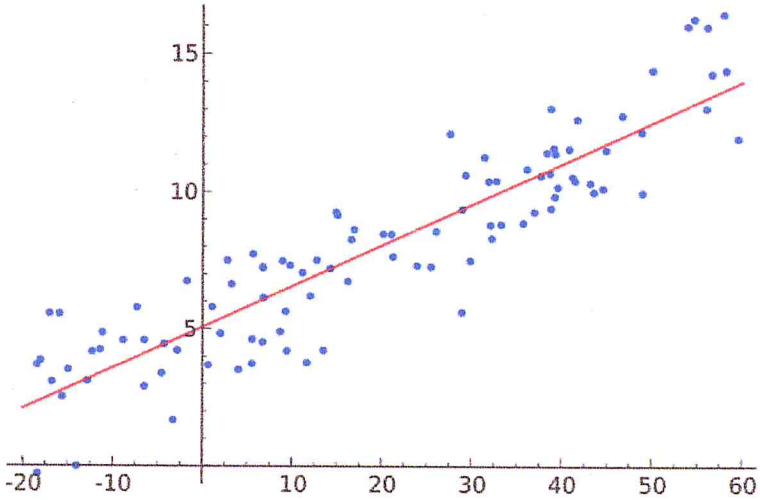


FIGURE 2.3: La régression linéaire [?]

4.3.2 Régression logistique

La régression logistique est une méthode statistique pour effectuer des classifications binaires. Elle prend en entrée des variables prédictives qualitatives et/ou ordinales et mesure la probabilité de la valeur de sortie en utilisant la fonction sigmoïdal (représentée dans la figure 2.4).

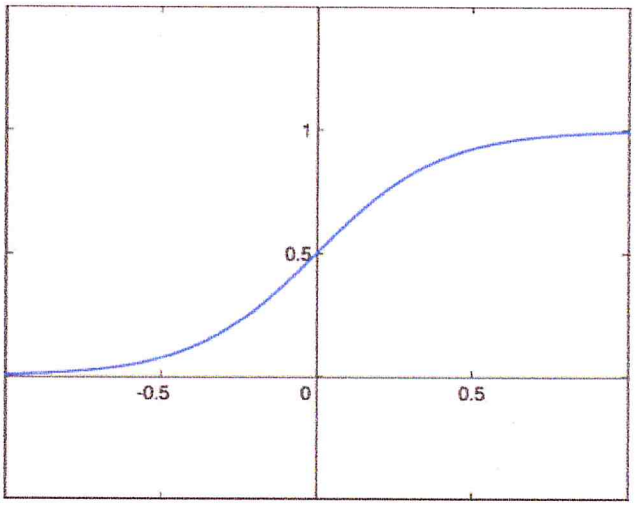


FIGURE 2.4: la régression logistique [?]

5 Analyse de données

L'analyse des données est un processus qui permet de transformer une masse d'informations en information structurée permettant la prise de décision. Nous distinguons trois types d'analyse de données qui sont :

1. Descriptive : En réalité, ce type a (presque) toujours existé. Il s'agit de transformer des données en connaissance, pas à pas, en suivant des principes statistiques [30].
2. Prédictive : ce type est beaucoup plus récent, parfois appelée analyse avancée, il consiste à utiliser des technique analytiques et statistiques permettant la prise de décision, son but est de réduire les risque ainsi que d'identifier des opportunités.
3. Prescriptive : ce type a pour but d'optimiser le résultat de ces prédictions en orientant les décideurs vers le meilleur scénario [30].

6 Le choix de la technique

Les caractéristiques des données peuvent souvent nous aider à déterminer les techniques de modélisation prédictive les mieux appropriées aux besoins de l'analyste des données.
Voici un certain nombre de points à prendre en compte lors du choix de la technique à utiliser, en fonction de nos données et du problème à résoudre [31].

— Lorsque les données sont groupées par observations, les outils tels que l'analyse de clusters, les règles d'association et les k plus proches voisins fournissent habituellement les meilleurs résultats.

— Utiliser la classification pour séparer les données en classes en fonction de la variable réponse – classes binaires comme Vrai ou Faux, ainsi que situations multi-classes.

— Utiliser la régression unique, multiple et polynomiale lorsque vous tentez d'effectuer une estimation plutôt qu'une classification.

7 Analyse prédictive

L'analyse prédictive, parfois appelée analyse avancée, est un terme utilisé pour décrire une série de techniques analytiques et statistiques permettant de prédire des actions ou des comportements futurs. Dans les entreprises, l'analyse prédictive est utilisée pour prendre des décisions proactives et déterminer des actions, au moyen de modèles statistiques permettant de découvrir des schémas dans des données historiques et transactionnelles, dans le but d'identifier des risques potentiels et des opportunités [31].

7.1 Utilisations métier de l'analyse prédictive

Vu la grande concurrence dans le marché, la prise des bonnes décisions est devenu plus que nécessaire, c'est pour cela que l'utilisation de l'analyse prédictive est devenu indispensable au sein des entreprises parce qu'elle offre une analyse plus intelligente, qui permet d'optimiser les prises de décision, elle permet aussi d'anticiper des événements, de comprendre des comportements et de les modéliser pour établir des prévisions.

— **Ventes et marketing direct :** L'analyse prédictive permet en effet d'anticiper les besoins personnelles des individus, ce qui permet de prédire leurs réactions, et proposer des offres pour chaque consommateur.

— **Fraude à l'assurance :** Plusieurs types de fraudes ont une approche prévisible et peuvent être identifiés à l'aide de modèles statistiques, ce qui aide la prévention, les enquêtes après fraude et le recouvrement [32].

— **Résultats en matière de santé :** L'apport de l'analyse prédictive dans le domaine de la santé est que les médecins pouvaient avoir un autre outil pour faire des choix plus éclairés sur le traitement d'un patient en fonction des cohortes des patients, ainsi que déterminer

rapidement et précisément le meilleur plan de traitement pour un patient donné en fonction de leurs antécédents médicaux.

7.2 La modélisation prédictive

L'utilisation de l'analyse prédictive implique la compréhension et la préparation des données, la définition du modèle prédictif et le respect du processus prédictif. Les modèles prédictifs peuvent prendre diverses formes et tailles, selon leur complexité et l'application pour laquelle ils sont conçus. La première étape consiste à comprendre les questions auxquelles vous essayez d'apporter une réponse pour votre entreprise. Le niveau de détail et de complexité de vos questions va augmenter tandis que vous progresserez dans le processus d'analyse [31].

7.3 Processus de l'analyse prédictive

Le processus de l'analyse prédictive est composé des étapes suivantes (figure 2.5) :

1. **Définition et compréhension du problème :** Avec la compréhension du problème, on peut préparer les données nécessaires à l'exploration, établir les objectifs métiers et interpréter correctement les résultats obtenus.
2. **Collecte des données :** La définition du problème permet d'avoir une idée sur les données qui doivent être utilisées.
3. **Pré-traitement :** Cette étape consiste à normaliser ou éliminer les données incohérentes c.-à-d. qui sortent des intervalles permis, (par exemple inférieur à 0) ainsi que les données omises à cause des erreurs de frappe ou à causes des erreurs dues au système lui-même, dans ce cas il faut remplacer ces données ou éliminer complètement leurs enregistrements.
4. **Estimation du modèle :** après avoir fixé l'objectif et les données utiliser il faut choisir la technique la mieux adapter pour l'extraction des données.
5. **Interprétation du modèle et établissement des conclusions :** Cette étape consiste à résumer les résultats obtenue d'une façon qu'elle soit compréhensible par l'utilisateur final

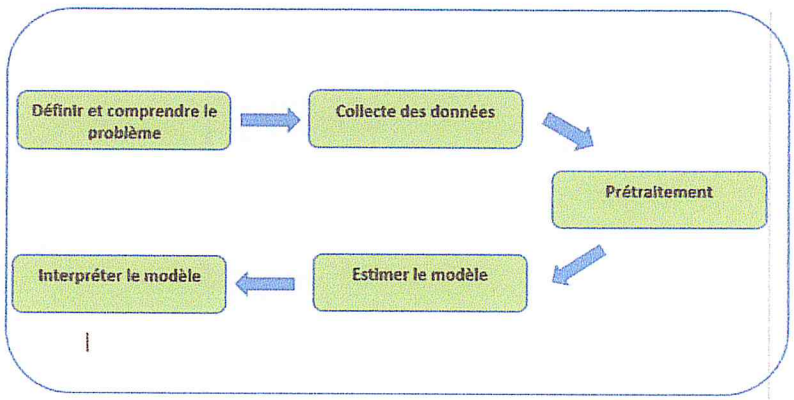


FIGURE 2.5: Processus de l'analyse prédictive

8 Les techniques de prédiction

Il existe plusieurs techniques de prédiction, dans ce qui suit nous allons citer quelques-unes afin de donner une description générale.

8.1 Le raisonnement base sur la mémoire

Le raisonnement basé sur la mémoire (RBM) est une technique de prédiction et de classification utilisée dans le cadre de la découverte de connaissances dirigée. Elle peut être également utilisée pour l'estimation. Pour chaque nouvelle instance présentée, le système recherche le(s) voisin(s) le(s) plus proche(s) et procède ainsi à l'affectation ou estimation. L'avantage du RBM est qu'il est facile à mettre en œuvre, très stable et supporte tout type de données [33].

8.2 Les arbres de décision

Les arbres de décision sont utilisés dans le cadre de la découverte de connaissances dirigée. Ce sont des outils très puissants principalement utilisés pour la classification, la description ou l'estimation[33].
L'arbre de décision est constitué de nœuds qui forment un arbre dirigé avec un nœud appelé "racine" qui n'a pas d'entrée. Tous les autres nœuds ont exactement un arc entrant. Un nœud avec des arcs sortants est appelé un nœud interne ou un nœud de test. Tous les autres nœuds sont appelés feuilles (également appelés nœuds terminaux ou de décision). la figure suivante représente l'architecture d'un arbre de décision :

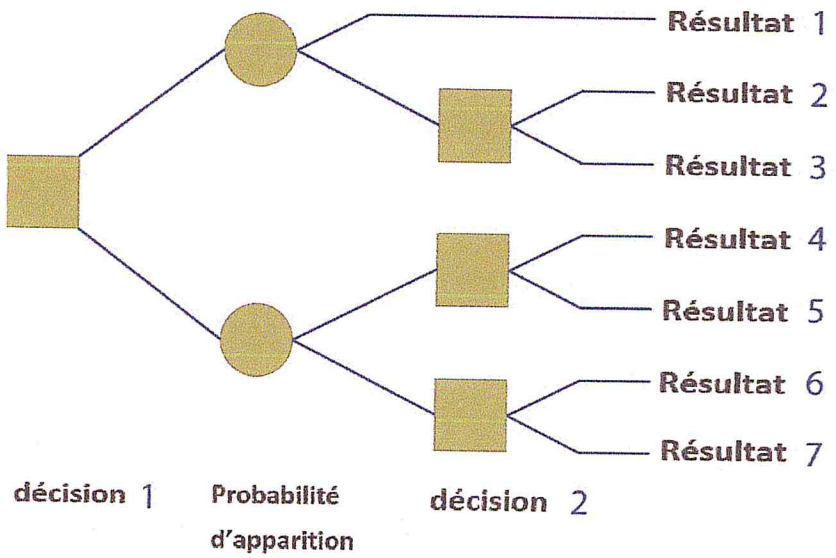


FIGURE 2.6: Architecture d'un arbre de décision

8.3 Les réseaux de neurones

Un réseau de neurones est un modèle de calcul dont le fonctionnement schématique est inspiré du fonctionnement des neurones biologique. Chaque neurone fait une somme pondérée de ses entrées (ou synapses) et retourne une valeur en fonction de sa fonction d'activation. Cette valeur peut être utilisée soit comme une des entrées d'une nouvelle couche de neurones, soit comme un résultat qu'il appartient à l'utilisateur d'interpréter (classe, résultat d'un calcul, etc.). La structure d'un réseau de neurone est la suivante :

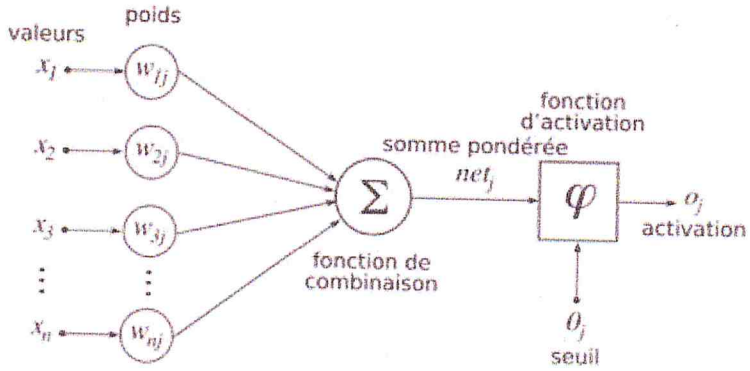


FIGURE 2.7: Schéma général d'un réseau de neurone [34]

La phase d'apprentissage d'un réseau de neurones permet de régler le poids associé à chaque synapse d'entrée (on parle également de coefficient synaptique). C'est un processus long qui doit être réitéré à chaque modification structurelle de la base de données traitée [34].

8.4 Les séries chronologiques

Une série temporelle (ou encore une série chronologique) est une suite finie (x1, · · · , xn) de données indexées par le temps. L'indice temps peut être selon les cas la minute, l'heure, le jour, l'année etc.... Le nombre n est appelé la longueur de la série. Il est la plupart du temps bien utile de représenter la série temporelle sur un graphe construit de la manière suivante : en abscisse le temps, en ordonnée la valeur de l'observation à chaque instant. Pour des questions de lisibilité, les points ainsi obtenus sont reliés par des segments de droite. Le graphe apparaît donc comme une ligne brisée [40].

8.4.1 Les composantes d'une série chronologique :

- **La tendance :** La tendance correspond à l'évolution à long terme de la série, l'évolution fondamentale de la série.
- **Les variations saisonnières :** Les variations saisonnières sont des fluctuations périodiques à l'intérieur d'une année, et qui se reproduisent de façon plus ou moins permanente d'une année sur l'autre.
- **Les variations accidentelles ou résiduelles :** Les variations accidentelles sont des fluctuations irrégulières et imprévisibles. Elles sont supposées en général de faible amplitude.

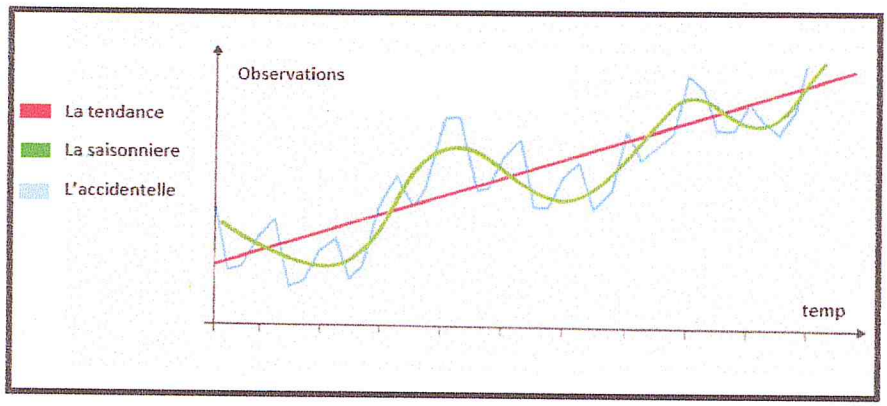


FIGURE 2.8: Les composants d'une série chronologique [40]

8.4.2 Les modèles de série chronologique [45] :

Les trois modèles de séries chronologiques univariées qui sont largement appliqués dans la littérature pour modéliser différents phénomènes sont :

1. **le processus autorégressif (AR) :** Les processus autorégressifs construits à partir de l'idée que l'observation à l'instant t s'explique linéairement par les observations précédentes. Le modèle AR est intuitivement intéressant car il décrit comment une observation dépend directement d'une ou plusieurs mesures précédentes, plus un bruit blanc. On dit que (Yt) est un processus autorégressif d'ordre p (centré) s'il s'écrit :

$$Y_t = \sum_{i=1}^{P} a_i Y_(t-i) + \epsilon_t \tag{2.1}$$

2. **Le processus de moyenne mobile(MA) :** Le modèle de moyenne mobile MA(q) est une fonction linaire où les valeurs de la série (Yt) s'expriment par la combinaison linéaire de l'erreur aléatoire qui a entaché les q essais précédents. Le modèle moyenne mobile (Moving Average) d'ordre q répond à l'équation :

$$Y_t = \epsilon_t + \sum_{j=1}^{q} b_j \epsilon_{t-j} \tag{2.2}$$

où b-j est un paramètre non saisonnier du modèle MA(q).

3. **le modèle ARMA (autorégressif et moyenne mobile) qui mixte les deux processus.** Le modèle mixte ou ARMA(p,q) définit des processus sous la forme d'une récurrence autorégressive avec un second membre de type moyenne mobile. Le modèle ARMA a d'abord été présenté par Box et Jenkins ([15]. Un processus autorégressif moyenne mobile d'ordres p et q, ARMA(p,q), est de la forme :

$$Y_t = \sum_{i=1}^{p} a_i Yt - i + \sum_{j=1}^{q} b_j \epsilon_{t-j} + \epsilon_t \tag{2.3}$$

4. **La méthode ARIMA [47]**
Il existe deux catégories de modèles pour rendre compte d'une série temporelle. Les premiers considèrent que les données sont une fonction du temps (y = f(t)).
Une seconde catégorie de modèles cherche à déterminer chaque valeur de la série en fonction des valeurs qui la précède (yt = f(yt-1, yt-2, ...)). C'est le cas des modèles ARIMA ("AutoRegressive – Integrated – Moving Average"). Cette catégorie de modèles a été popularisée et formalisée par Box et Jenkins (1976).

   - Les processus autorégressifs supposent que chaque point peut être prédit par la somme pondérée d'un ensemble de points précédents, plus un terme aléatoire d'erreur.
   - Le processus d'intégration suppose que chaque point présente une différence constante avec le point précédent.
   - Les processus de moyenne mobile supposent que chaque point est fonction des erreurs entachant les points précédant, plus sa propre erreur.
   - Un modèle ARIMA est étiqueté comme modèle ARIMA (p,d,q), dans lequel :
     - p est le nombre de termes auto-régressifs
     - d est le nombre de différences
     - q est le nombre de moyennes mobiles.

Différentiation :
L'estimation des modèles ARIMA suppose que l'on travaille sur une série stationnaire. Ceci signifie que la moyenne de la série est constante dans le temps, ainsi que la variance. La meilleure méthode pour éliminer toute tendance est de différencier, c'est-à-dire de remplacer la série originale par la série des différences adjacentes. Une série temporelle qui a besoin d'être différenciée pour atteindre la stationnarité est considérée comme une version intégrée d'une série stationnaire (d'où le terme Integrated).

Signification des paramètres des modèles ARIMA :
L'objectif essentiel des modèles ARIMA est de permettre une prédiction de l'évolution future d'un phénomène. Son développement dans le domaine de l'économétrie est basé sur ce principe.

   - Un processus non différencié à bruit blanc (ARIMA(0,0,0) suggère des fluctuations aléatoires autour d'une valeur de référence. Cette valeur de référence peut être considérée comme une caractéristique stable du système étudié (trait de personnalité, mémoire, capacité stabilisée, etc..)
   - Un processus de moyenne mobile suggère que la valeur de référence évolue d'une mesure à l'autre. Plus précisément, la valeur de référence est fonction de la valeur de référence précédente et de l'erreur ayant entaché la mesure précédente.
   - Un processus auto-régressif suggère que le phénomène étudié n'est pas déterminé par une valeur de référence. C'est la performance précédente (ou les performances précédentes) qui déterminent entièrement la performance présente.

Détermination des termes AR et MA :
Après que la série ait été stationnarisée, l'étape suivante consiste à identifier les termes AR et MA nécessaires pour corriger les auto- corrélations résiduelles. Cette analyse est basée sur l'examen des fonctions d'auto-corrélation (ACF) et d'auto-corrélation partielle

(PACF). Rappelons que l'autocorrélation est la corrélation d'une série avec elle-même, selon un décalage (lag) défini. L'auto-corrélation de décalage 0 est par définition égale à 1. La fonction d'auto-corrélation fait correspondre à chaque décalage l'auto-corrélation correspondante.
D'une manière générale, un corrélation partielle entre deux variables est la quantité de corrélations qui n'est pas expliquée par les relations de ces variables avec un ensemble spécifié d'autres variables.

5. **La méthode de Holt-Winters :** La méthode de Holt et Winters permet en effet d'effectuer des prévisions sur des séries chronologiques assez irrégulières et soumises ou non à des variations saisonnières qui sont des variations dues à un effet momentané se reproduisent régulièrement dans le temps suivant non seulement un modèle additif qui est le plus simple dans lequel la variation saisonnière s'ajoute simplement à la tendance , Mais aussi avec un modèle multiplicatif qui introduit la composante saisonnière de manière multiplicative.
L'approche de Holt et Winters consiste en trois lissages exponentiels simultanés. On définit donc trois paramètre notés. A chaque instant, elle donne une estimation :
   - De la tendance
   - Du coefficient saisonnier correspondant
   - De la valeur observée

   On peut choisir les coefficients arbitrairement : faible si l'on considère que la valeur à l'instant « t » dépend d'un grand nombre d'observations antérieures, élever dans le cas contraire. On peut aussi calculer les valeurs optimales en minimisant la somme des carrés des différences entre les valeurs observées et estimées. On procède ensuite aux prévisions, en considérant que la tendance suit un modèle linéaire additif ou multiplicatif à très court terme.

9 Solutions de l'analyse prédictive pour les services d'admission hospitaliers

Au cours des dernières années, en raison de la demande croissante aux soins médicaux, des ressources matérielles et humaines importantes sont nécessaire pour faire face à cet afflux de patients. Malheureusement ces ressources sont très limitées dans la plupart des cas.
C'est pour cela qu'une étude prédictive dans les services d'admission des hôpitaux peut être très utile pour l'optimisation de l'utilisation de ces ressources afin de réduire le nombre de patients dans les files d'attentes et en améliorer la qualité de soin.
citons quelques travaux existants :

9.1 La durée du séjour

Cet indicateur mesure le temps qui s'écoule entre le moment où le patient s'enregistre et le moment où il quitte les services. On peut aussi mesurer la durée de séjour dans le service d'urgence et la durée de séjour dans un autre service à part.
Cela permettra d'une part d'améliorer la gestion de flux des patients et de réduire le temps d'attente des patients aux urgences pour obtenir un diagnostic, un traitement ou un transfert à un lit d'hôpital.
D'autre part on peut mesurer le pourcentage de journées d'hospitalisation des patients qui occupent des lits de soins actifs après leurs congés. la technique de data mining la mieux placé pour la prédiction de la durée de séjour est la technique de réseaux de neurones

Pei-Fang et al. [35] ont fait une étude comparative entre la méthode de réseaux de neurone et la technique de régression linéaire. les résultats ont montré que les réseaux de neurones ont un avantage dans le cas ou la durée de séjour est comprise entre 16 et 22 jours.
Dans le travail de Mobley et al. [36] les auteurs ont prédit la durée de séjour exacts pour les patients dans une unité de soins post-coronariens. Avec 629 admissions dans le fichier de formation et 127 admissions dans le fichier de test, un total de 74 variables d'entrée ont été utilisées pour prédire 1 à 20 jours de séjour dans les réseaux de neurones. La durée moyenne de séjour était de 3,84 jours dans le fichier de formation et de 3,49 jours dans le fichier de test. Ils n'ont montré aucune différence significative dans la distribution à partir de la durée de séjour prédit et la durée de séjour réel.

9.2 Le taux de réadmission au bout d'un mois

Le taux de réadmission mesure le pourcentage de réadmissions non facultatives dans un hôpital dans les 30 jours (pour certains troubles et selon les groupes de maladies analogues)[37]. Une réadmission inattendue d'un patient peut entrainer une frustration pour lui, mais aussi elle peut nuire à la réputation de l'hôpital, elle constitue également une utilisation inefficace des ressources de hôpital.
Les hôpitaux peuvent réduire leurs taux en cernant les patients les plus susceptibles de retourner à l'hôpital dans un court délai et en améliorant les processus de congé [37].
Pour cela une recherche a été faite en Amérique en 2015 [44] Pour réduire le nombre de réadmissions aux hôpitaux. Cette recherche a proposé diverses approches d'exploration de données pour identifier le groupe de risques d'un patient particulier, y compris le modèle de réseau neuronal, l'algorithme de forêt aléatoire et le modèle hybride d'heuristique d'intelligence d'essaim et de SVM. L'algorithme de réseau neuronal proposé, les classificateurs RF et SVM sont utilisés pour modéliser les caractéristiques des patients, tels que leur âge, les assureurs, les risques de médicaments, etc. Des expériences sont menées pour comparer les performances des modèles proposés à des recherches antérieures. Les résultats expérimentaux indiquent que le modèle SVM de prédiction proposé avec réglage des paramètres d'essaim de particules surpasse les autres algorithmes et atteint 78,4% de précision de prédiction globale, 97,3% de sensibilité. La haute sensibilité montre sa force à identifier correctement les patients réadmis.

9.3 Le nombre de patients

La prévision de l'afflux des patients d'une structure d'urgences au sein d'un centre hospitalier est un enjeu crucial limitant l'attente des patients et améliorant la qualité des soins. Dans ce contexte, l'exploitation de l'ensemble des données composant l'historique des consultations du service permet de modéliser et prévoir ce flux [39].
Les séries chronologiques se sont révélées être un outil efficace pour prévoir la fréquentation en utilisant des données historiques.

- deux méthodes d'analyse de séries chronologiques (ARIMA et l'approche ad hoc) ont été utilisées pour modéliser les admissions mensuelles à la clinique de Roi Fayçal, Al -Khobar, en Arabie Saoudite [42].
- la méthode de Box-Jenkins (ARMA) a été utilisée pour prévoir le nombre d'admissions quotidiennes et le nombre de lits occupés aux urgences de l'hôpital de Bromley au Royaume-Uni [41].
- deux méthodes statistiques (un lissage exponentiel et la méthode de BoxJenkins)ont été utilisées pour prévoir le nombre d'admission chaque mois pour la période 2000 à 2005 aux urgences de l'hôpital régional de Victoria en Australie [43].

9.4 Discussion

D'après les travaux existants on peut conclure que la meilleur méthodes pour la prédiction de la durée moyenne de séjour pour les périodes coutes est bien la méthode de réseaux de neurones, par contre si il s'agit de long durée il vaut mieux d'utiliser les séries chronologique. En ce qui concerne le nombre de patients, deux méthode ont été utilisées qui sont les séries chronologiques et plus précisément la méthode ARIMA et la méthode de lissage exponentiel. Les résultats ont montré que la méthode ARIMA donne des prédiction plus précises.

10 Conclusion

Dans ce chapitre nous avons présenté une revue autour des concepts de Data Mining et de l'analyse prédictive. Nous avons abordé dans un premier temps le Data Mining pour montrer l'utilité de ce dernier dans l'extraction de l'information depuis de grande quantité de données. Nous avons ensuite présenté l'analyse prédictive comme étant un outil très important d'aide à la décision et qui constitue une véritable opportunité pour les décideur.

Chapitre 3

Conception

1 Introduction

Le défi de répondre aux challenges liés à ce projet nous a mené à concevoir notre solution en se basant sur une architecture fonctionnelle et technique qui s'aligne avec le challenge de la facilitation de la prise de décision au saint des hôpitaux.
Le système que nous allons concevoir a pour objectif de guider les décideurs afin de prendre les bonne décisions qui devra en fin de compte leur permettre d'optimiser la gestion des ressources disponibles.
Pour pouvoir réaliser ce système, nous avons tout d'abord passé par une étude conceptuelle permettant de formaliser les étapes préliminaires du développement de ce système.

1.1 Architecture du système

Comme c'est présenté dans la partie état de l'art, le choix d'une architecture fonctionnelle et technique est une partie très critique dans l'implémentation d'une solution Big Data. Dans notre cas nous allons utiliser l'architecture Lambda (voir chapitre 1 titre 5) pour réaliser les traitements souhaités en mode Batch. pour se faire nous allons tout d'abord faire une projection de notre architecture conceptuelle sur l'architecture Lambda :

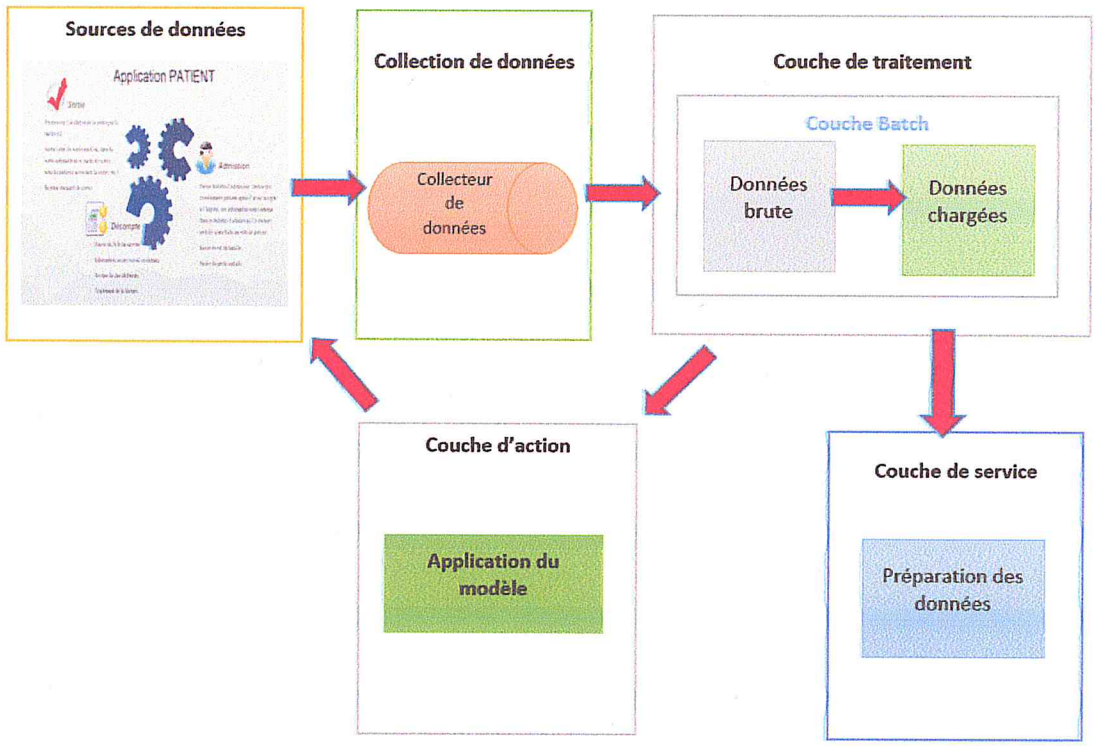


FIGURE 3.1: Architecture du système

— **Sources de données (Data source)du service d'admission du CHU Mustapha Bacha :** dans notre système, les données proviennent essentiellement de l'application PATIENT qui est la source principale des données interne (données sur les Patient, données sur les médecins...), en plus des données de l'application PATIENT nous avons pensé à ajouter des données externes par rapport à l'hôpital, comme les données de météo, les données de mariage etc... qui influencent sur les données et leurs traitements. Comme le montre la figure suivante, ces données vont être ajouté au système de fichier distribuer (HDFS) pour être ensuite utilisé pour l'extraction des informations.

— **Collection de données :** représente la partie intermédiaire qui se charge de la planification et la répartition des données entre la couche temps réel et Batch. Dans notre cas, c'est la couche Batch qui sera utilisée,(la couche temps réel ne sera pas utilisée vu que les données serons traités à la demande des décideurs, et non pas lors de leurs intégration dans la base de données).

— **Couche de traitements (Processing layer) :** Dans l'architecture de base de Lambda nous intégrons les données via des jobs de types Batch, qui ont le rôle de charger les données dans le modèle de données d'une façon graduelle, de nettoyer les données et réaliser des fonctions d'enrichissement supplémentaire et restructuration des données.

— **Couche de service (Service layer) :**cette partie représente la BDD et les interfaces d'accès pour réaliser les opérations d'analyses approfondies.

— **Couche d'actions :** il permet de réaliser des actions en fonction des données reçues qui interagissent directement sur le système afin de rendre le système plus réactif et fiable, ainsi d'appliquer les modèles choisis.

2 Analyse prédictive

Nous avons présenté dans le chapitre précédent de l'analyse prédictive et son processus (voir chapitre 2, titre 7.3), a cet effetn nous construisons notre système en se basons sur ce processus.

2.1 Définition et compréhension du problème :

Comme c'est déjà mentionné dans le chapitre précédent, le processus de l'analyse prédictive se débute par l'étape de définition et compréhension du problème, cette étape représente la pierre angulaire du projet.
Le problème qui se pose dans notre cas est l'absence des outils d'aide à la décision dans le service d'admission de l'hopital de Mustapha Bacha, c'est pourquoi nous avons défini comme but de notre projet les point suivants :

1. La prédiction du nombre de patients.
2. La prédiction de la durée moyenne de séjour.

C'est deux indicateurs doivent permettre aux décideurs d'avoir une estimation approximative de l'état future du service, qui leurs permettra de répondre aux besoins des patients de la bonne maniéré.

2.2 Compréhension des données

2.2.1 Collecte des données

Pour la bonne analyse et conception de l'outil prédictif, la phase de récupération de données et d'identification de leurs structures est très importante, voir critique. Effectivement, l'acquisition des données et de l'historique du patient représente un atout majeur dans la conception de notre système, car toute méthode d'analyse prédictive se base sur un historique pour pouvoir prédire le comportement futur. Il est très important de se doter de données significatives, complètes et récentes. Dans le cadre de notre projet, pour la réalisation de prévisions sur les patients nous avons besoin d'une source de données réelles sur lesquelles nous effectuons notre travail. Le Bureau des Entrées (service d'admission) au niveau de l'hôpital Mustapha Bacha assure l'accueil, l'orientation des usagers, la prise en charge administrative des patients hospitalisés et la réalisation de la facturation des frais de séjour à l'hôpital. Le bureau des entrées est le premier lieu de gestion de la performance et de l'attractivité au niveau des établissements de santé, il est donc nécessaire d'optimiser l'organisation de ce bureau tout en répondant à l'enjeu stratégique d'amélioration de la qualité de la prise en charge des patients. Le service d'admission dispose d'un outil informatique qui est l'application PATIENT : le rôle de cette application est

la gestion et le suivi du malade depuis son admission à l'hôpital jusqu'à sa sortie, sur la plan administratif (identification) et en prenant en considération toutes les prestations qui lui ont été fournies durant son séjour. L'application PATIENT est la source principale des données du service d'admission. Elle présente plusieurs fonctionnalités présentées dans la figure qui suit.

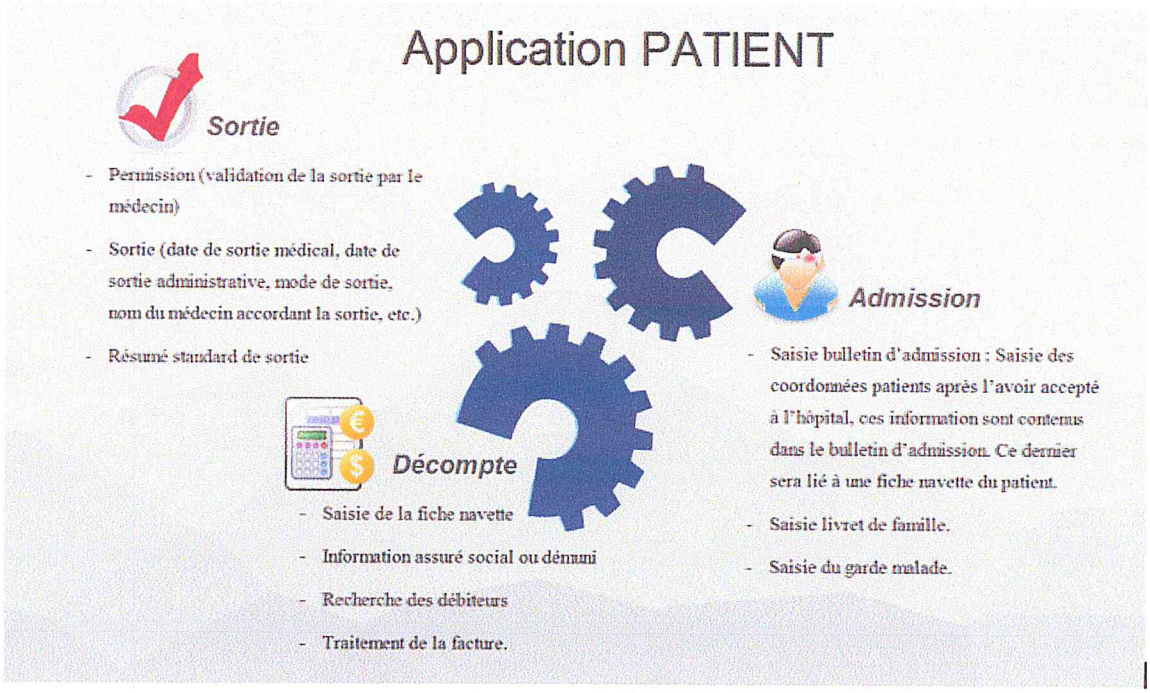


FIGURE 3.2: Fonctionnalités de l'application Patient

2.2.2 Description des données

Actuellement, la Division Systèmes d'Information et Systèmes Multimédia (DSISM) [1] du CERIST dispose des donnés issus de l'application patient, nous exploitons cette source de données dans notre travail. La description des différentes table de cette base de données sont détaillées dans l'annexe et le schéma suivant.

1. http ://www.cerist.dz/index.php/fr/rechercheetdevelop/114-divisions-de-recherche/524-systemes-d-information-et-systemes-multimedia-fr

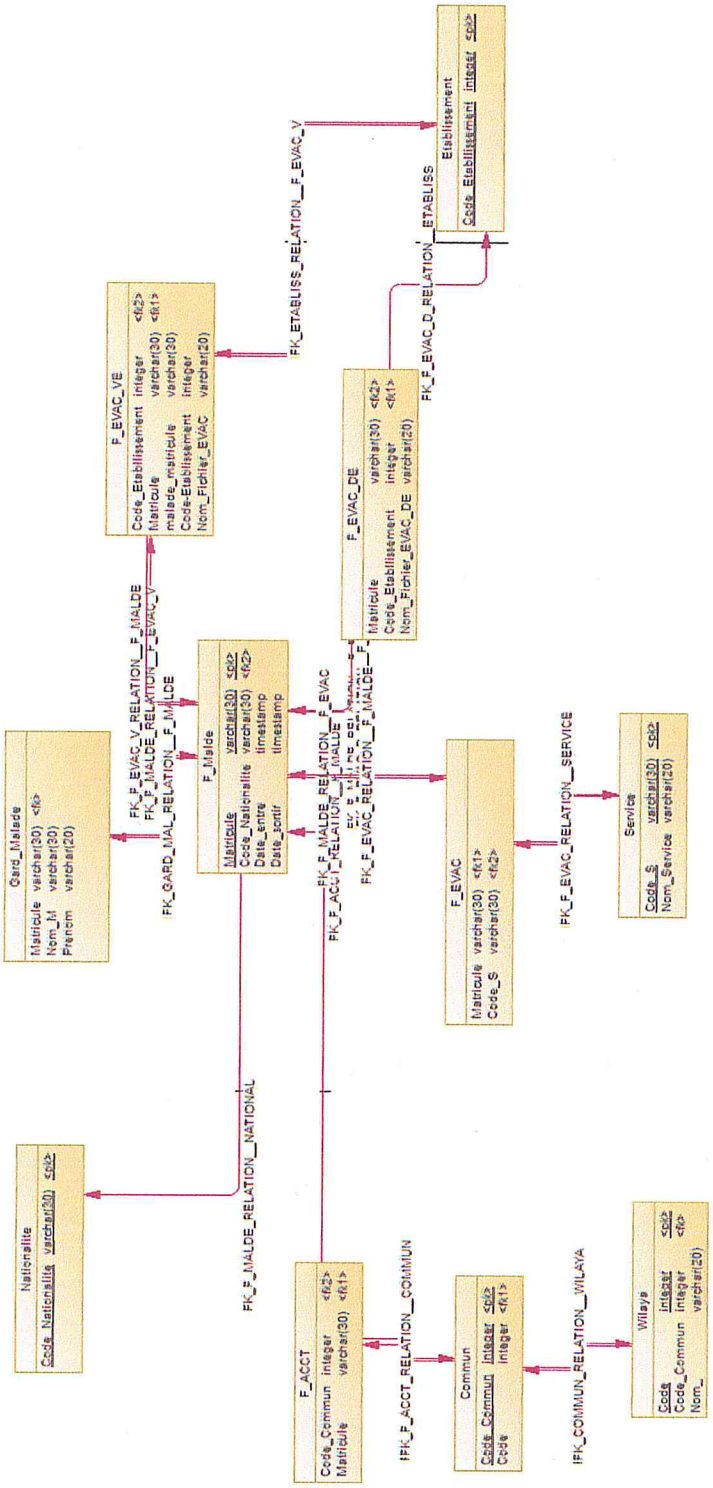


FIGURE 3.3: Schéma de la base de données

2.3 Pré-traitement

L'étape de pré-traitement ou l'étape de préparation des données consiste a nettoyer les données en éliminant les données contenants des anomalies, cette étape inclut aussi la recherche des corrélations cachées dans les données et la détermination des données qui sembles utiles pour la prédiction. Dans notre cas nous allons utiliser sur tout la table F-MALADE pour la prédiction du nombre de patients et de la durée moyenne de séjour, nous allons utiliser l'attribut **FER-DAT-EN** qui représente la date d'entrée du patient pour prédire le nombre de patients et les deux attributs **FER-DAT-EN** et **FER-DT-SOR** qui représentent la date d'entrée et de sortie pour la prédiction de la durée moyenne de séjour.

- **La prédiction du nombre de patients :** Le taux d'admission des patients à l'hôpital varie d'une saison à une autre, d'un service à un autre et selon différents paramètres (la météo, les jours fériés, le weekend end, etc.). Les décideurs ont donc besoin d'avoir une information préalable sur le nombre de patients selon différents axes :
  - Le service d'hospitalisation, le temps et le sexe.
  - Le service d'hospitalisation, le temps et la tranche d'âge.
  - Le service d'hospitalisation, le temps et la zone géographique.
  - Le service d'hospitalisation, le temps et la condition d'entrée.

  Ces informations agissent sur la planification des ressources humaines et matérielles nécessaires, donc elles permettent une meilleure prise en charge des patients.
- **La prédiction du la durée moyenne de séjour :** Fait référence au nombre moyen de jours que les patients passent à l'hôpital. Elle est fréquemment utilisée comme indicateur de l'efficience. Les hôpitaux cherchent généralement à réduire la durée moyenne des séjours tout en maintenant, voire en améliorant, la qualité des soins. Il est donc nécessaire d'évaluer cette mesure pour chaque service d'hospitalisation donné. Cette information permet aux décideurs de bien gérer la quantité du stock pharmaceutique ainsi que le manque en matière d'équipements sanitaires.

A cet effet nous allons utiliser les données de la période de septembre 2009 jusqu'à mars 2017, nous avons remarquer que les données des premières années contiennes beaucoup d'anomalies et beaucoup de manque d'informations (par exemple l'année 2009 ne contient qu'un seul enregistrement), pour cela nous avons éliminé une grande partie de ces données (nous avons supprimé les données de 2009 jusqu'à décembre 2012).

2.4 Estimation du modèle

Lors de la conception d'un système prédictif, on se trouve généralement face à une multitude de techniques et méthodes prédictives, qui semblent à première vue toutes rapides et efficaces. Cependant, si l'on regarde de plus près, chaque méthode présente des avantages qui la rendent imbattable sur une catégorie de prédictions, et des inconvénients la rendant inutilisable dans d'autres catégories. Cela veut dire que chaque problématique a sa propre technique de prédiction optimale :

- Algorithmes de régression : est utilisé pour mettre en relation une valeur prédite et un élément de données.
- Algorithmes de classification : relie les données dans des groupes prédéfinis (les catégories ou les classes).

— Algorithme « séries chronologiques » : dans ce cas les variables de modélisation évoluent avec le temps.

— Algorithmes de Clustering : similaire à la classification à ceci près que les classes ne sont pas prédéfinies, mais seront déterminées à partir de l'ensemble de données.

— Algorithmes d'association : se réfère à la tâche de Data Mining qui consiste à découvrir les relations entre les données.

Comme nous disposons de données d'une masse assez importante et que ces données représentent les dates d'entrée et sotie des patients (des intervalles de temps : une journée, des jours, des semaines, des mois). Il s'agit donc de variables continues. Les données représentent une série du nombre d'admission historique sur une période, ou sur une période de plusieurs mois. La méthode de série chronologique (voir chapitre 2, titre 8.4) semble être la mieux adaptée.

.

2.5 Interprétation du modèle et établissement des conclusions

Nous avons opté pour la création d'une application avec des interfaces graphiques simples pour facilité l'utilisation du système et la visualisations des résultats pour les utilisateurs finaux que ce soit décideur ou administrateur. Pour cela nous allons créer un espace pour chaqu'un d'eux.

3 Processus d'application du modèle ARIMA

Pour pouvoir mettre en place la solution de la prédiction du nombre d'admission dans les services d'admission des hôpitaux, nous allons passer par les étapes suivantes :

3.1 Sources de données

Dans les hôpitaux les sources de données peuvent être très diverses : des rangées de nombres structurées aux textes non structurés tels que les notes. Rien que cette dernière catégorie peut s'avérer une source très utile d'information mais malheureusement nous avons pu avoir que les données de l'application patient (données structurés). Les données de l'application PATIENT seront importé par l'administrateur dans HDFS, les images concernant le dossier du patient seront alimentés et archivés avec les données de ce dernier dans Hadoop.

3.2 Pré-traitement des données

Le prétraitement des données est d'une importance capitale pour la suite du processus. La réalisation de cette étape a nécessité la prise en charge de diverses difficultés concernant : les sources, la qualité et le format des données.

1. **Identification et sélection du jeu de données :**Cette étape se concrétise via la sélection des données qui vont être utilisées pour implémenter la solution, pour cela nous avons utilisé sur tout la table Fmalade et plus précisément la colonne qui représente la date d'entrée des patients pour la prédiction du nombre d'admission.
2. **Traitement de valeurs manquantes :** nous avons écarter les enregistrement ou il y avait des données manquantes(figure4.5 et figure 4.6).

3	2010-07-29	<NA>
4	2011-02-23	2015-05-01
5	2010-10-07	2014-09-23
6	2011-03-13	2011-03-13
7	2011-04-15	<NA>
8	2011-04-15	<NA>
9	2011-05-06	<NA>
10	2011-05-07	<NA>
11	2011-06-02	2011-06-02
12	2011-06-11	2011-06-11
13	2011-07-07	<NA>
14	2011-07-23	<NA>
15	2011-09-18	2011-09-18
16	2011-11-21	<NA>

FIGURE 3.4: Les données avant la suppression

	FER_DAT_EN	FER_DT_SOR
1	2009-09-23	2015-08-12
2	2010-04-19	2015-06-26
3	2011-02-23	2015-05-01
4	2010-10-07	2014-09-23
5	2011-03-13	2011-03-13
6	2011-06-02	2011-06-02
7	2011-06-11	2011-06-11
8	2011-09-18	2011-09-18
9	2012-01-09	2012-01-09
10	2012-02-12	2012-02-12
11	2012-03-01	2012-03-01
12	2012-06-14	2012-06-14
13	2012-08-26	2014-08-01
14	2012-11-07	2012-11-07
15	2012-11-25	2013-01-06

FIGURE 3.5: Les données aprés la suppression

La figure suivante montre les commandes nécessaire pour effectuer les pré-traitements :

```
18 x2<-x1%>%select(FER_DAT_EN,FER_DT_SOR)
19 x3<-filter(x2,FER_DT_SOR !="",FER_DAT_EN!="" )
20 x3$FER_DAT_EN <-as.Date(x3$FER_DAT_EN,format = "%d-%m-%y")
```

FIGURE 3.6: Script de pré-traitement

La ligne 19 permet de supprimer les enregistrement qui contiennes un manque au niveau des données essentiellement celles qui seront utilisées pour l'analyse, tandis que la ligne 20 permet de convertir le format de la date en format 'Date' pour qu'elle soit prête pour l'utilisation.

3.3 Analyse prédictive

Afin de réaliser la prédiction du nombre de patients nous avons décidé de laisser le choix au décideur pour décider la durée de sa prédiction (prédiction à cours terme ou prédiction à long terme), pour cela nous avons comparer deux méthodes de séries chronologique a savoir la méthodes ARIMA (la méthodes de Box et Jenkins) et la méthode de Holt-winter pour décider quel méthodes utiliser pour chaque type de prédiction.

3.3.1 Prédiction à court terme

Pour réaliser la prédiction à court terme nous allons comparer entre ARIMA et HOLT-WINTER pour choisir la méthodes qui donne les meilleurs résultats. Le graphe suivant représente les données que nous allons utiliser pour le test :

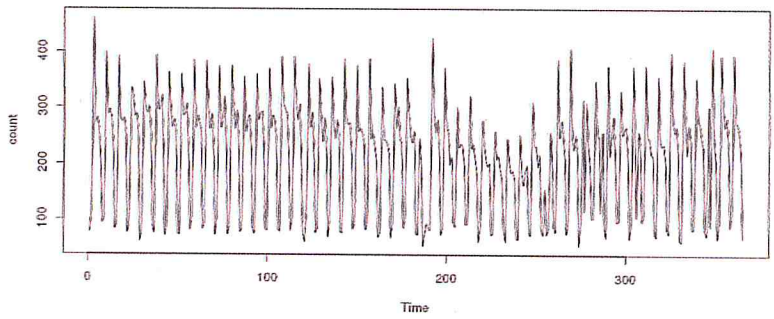


FIGURE 3.7: Graphe de la série (court terme)

(i) **La méthode ARIMA :**

— **Préparation :** Nous allons en premier temps importer les données dans R puis tracer le graphe pour avoir une idée sur les données. Tous d'abord nous allons faire le test de Dickey-Fuller pour vérifier la stationnarité de notre série. Si la valeur de (P-value inférieur à 0.05) on dit que la série est stationnaire, sinon nous devons différencier la série et refaire le test.

```
        Augmented Dickey-Fuller Test

data:  tes16
Dickey-Fuller = -4.1334, Lag order = 7, p-value = 0.01
alternative hypothesis: stationary
```

Comme la valeur P-value est inférieur à 0.05 nous pouvons conclure que notre série est stationnaire. D'après le graphe de la série (figure 4.8) et le test de Dickey-Fuller notre série est de famille ARIMA(p,1,q) ; apres le test de stationnarité il faut définir les valeur de p et q on utilisant la fonction d'autocorrélation simpe (ACF) et la fonction d'autocorrélation partiel (PACF)

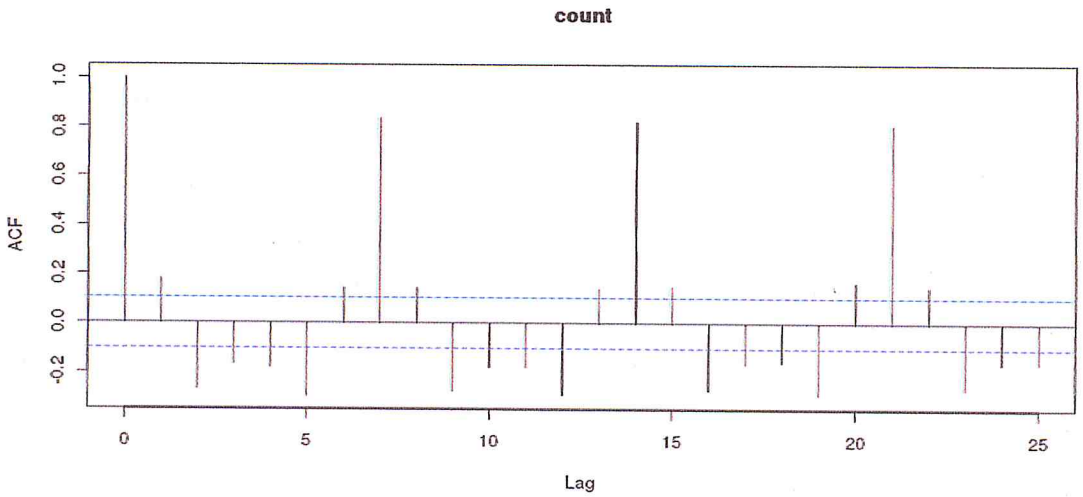


FIGURE 3.8: Graphe de l'ACF

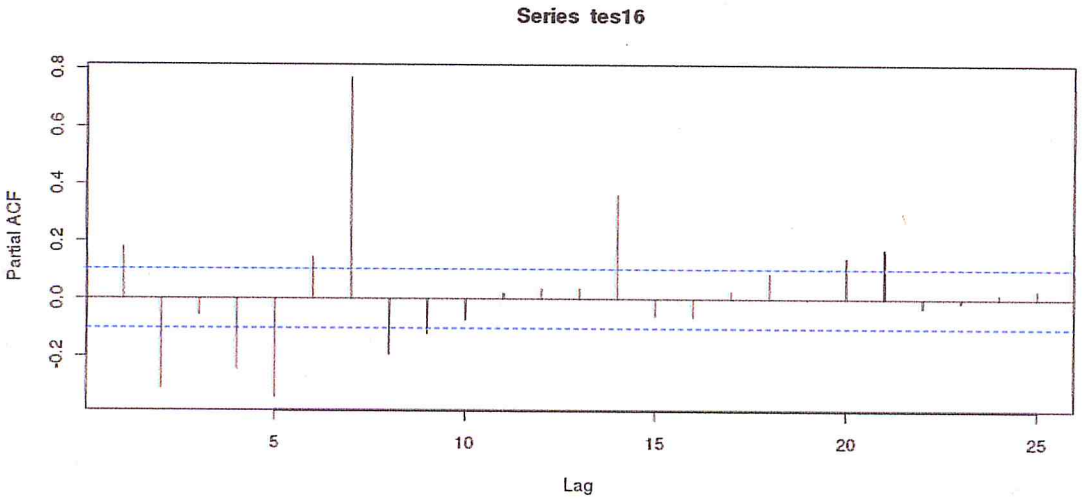


FIGURE 3.9: Graphe de PACF

D'après l'ACF et PACF on conclu que le meilleure modèle est ARIMA(3 ;1 ;4).

— **Validation :** Dans cette étape, nous allons utiliser les trois principaux testes (Shapiro-Wilck, Box-lujung et t-test) pour valider notre modèle.

```
> t.test(fit16s,mu = 0,conf.level = 0.95)

        One Sample t-test

data:  fit16s
t = -0.06012, df = 364, p-value = 0.9521
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -8.548409  8.041231
sample estimates:
 mean of x
-0.2535892
```

```
> Box.test(fit16R,lag = 7,type = "Ljung-Box",fitdf = 1)

        Box-Ljung test

data:  fit16R
X-squared = 289.78, df = 6, p-value < 2.2e-16
```

```
> fit16s<-residuals(fit16)
> shapiro.test(fit16s)

        Shapiro-Wilk normality test

data:  fit16s
W = 0.98676, p-value = 0.002047

>
```

— **Discussion :** D'après le test Shapiro-Wilk et le test Box-Ljung et le test t-test on remarque que p-value inférieur à 0.05 donc la série est bien modélisée par le modèle ARIMA(3,1,4).

(ii) **La méthode HOLT-WINTER :**

Comme la méthode de HOLT-WINTER ne nécessite pas que la série soit stationnaire, nous allons passer directement à l'étape de validation.

— **validation :** Nous allons utiliser les mêmes test pour que la comparaison soit juste .

```
> shapiro.test(fchwshort$residuals)

        Shapiro-Wilk normality test

data:  fchwshort$residuals
W = 0.79826, p-value < 2.2e-16

> Box.test(fchwshort$residuals,lag = 7,type = "Ljung-Box",fitdf = 1)

        Box-Ljung test

data:  fchwshort$residuals
X-squared = 14.001, df = 6, p-value = 0.02963

> t.test(fchwshort$residuals,mu=0,conf.level = 0.95)

        One Sample t-test

data:  fchwshort$residuals
t = 0.90092, df = 364, p-value = 0.3682
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2.438880  6.562863
sample estimates:
mean of x
 2.061991
```

— **Discussion :** D'après le test Shapiro-Wilk et le test Box-Ljung et le test t-test on remarque que p-value est inférieur 0.05 donc la série est bien modélisée.

3.3.2 Prédiction à long terme

Pour réaliser noter prédiction a long terme nous allons procéder de la même façons que dans la prédiction a court terme en comparant les deux méthodes ARIMA et HOLT-WINTER. Le graphe de données que nous allons utiliser pour le teste de prédiction à long terme est le suivant :

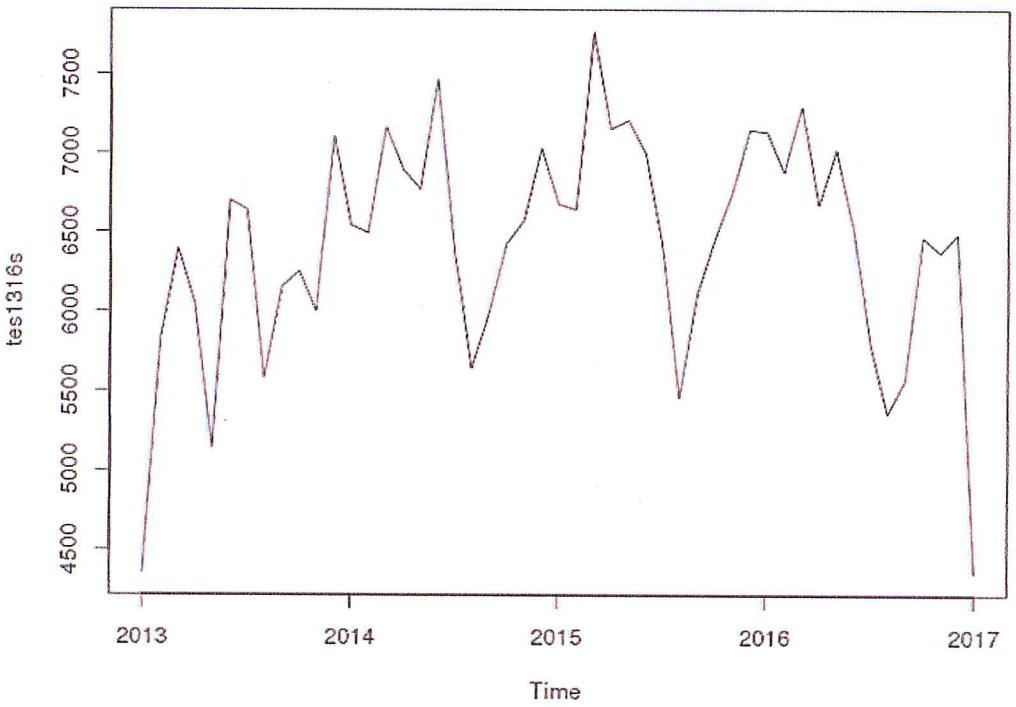


FIGURE 3.10: Graphe de données (long terme)

(i) **La méthode ARIMA :**

— **Préparation :**
Test de Dickey-Fuller :

```
> adf.test(tes1316s)

        Augmented Dickey-Fuller Test

data:  tes1316s
Dickey-Fuller = -2.0612, Lag order = 3, p-value = 0.5497
alternative hypothesis: stationary
```

On remarque que la valeur de p-value est supérieur à 0.05 donc nous pouvons conclure que la série n'est pas stationnaire, donc nous devons différencier la série. pour cela nous avons utiliser la fonction auto-ARIMA et nous avons eu les résultats suivants :

— **Validation :**

```
> shapiro.test(residuals(fit1316))

        Shapiro-Wilk normality test

data:  residuals(fit1316)
W = 0.86417, p-value = 4.497e-05

> Box.test(residuals(fit1316),lag = 3,type = "Ljung-Box",fitdf = 1)

        Box-Ljung test

data:  residuals(fit1316)
X-squared = 5.1308, df = 2, p-value = 0.07689


> t.test(residuals(fit1316),mu=0,conf.level = 0.95)

        One Sample t-test

data:  residuals(fit1316)
t = -1.3505, df = 48, p-value = 0.1832
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -252.64247   49.62162
sample estimates:
mean of x
-101.5104
```

— **Discussion :**

On remarque que la valeur de p-value du test t.test est très proche du zéro ce qui signifie que les résultats obtenues ne sont pas fiables, et que ce modèle ne donne pas des prédictions précises.

(ii) **La méthode Holt-Winter :**

— **Validation :**

```
> shapiro.test(fchwlong$residuals)

        Shapiro-Wilk normality test

data:  fchwlong$residuals
W = 0.87292, p-value = 0.0005659

> Box.test(fchwlong$residuals,lag = 3,type = "Ljung-Box",fitdf = 1)

        Box-Ljung test

data:  fchwlong$residuals
X-squared = 2.5604, df = 2, p-value = 0.278

> t.test(fchwlong$residuals,mu=0,conf.level = 0.95)

        One Sample t-test

data:  fchwlong$residuals
t = -1.2315, df = 36, p-value = 0.2261
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -257.16240   62.84787
sample estimates:
mean of x
-97.15727
```

— **Discussion :** On remarque que la valeur de p-value du test Shapiro-wilk est très proche du zéro et la valeur de p-value du test t.test n'est pas très proche du zéro donc nous pouvons conclure que ce modèle convient pour cette situation.

3.3.3 Choix de méthodes

Afin de décider quelle méthode utiliser pour chaque cas de prédiction nous effectuons une comparaison entre les résultats obtenus par chaque méthode dans les deux cas en montrant les résultats de prédiction graphiquement.
Nous avons mis les résultats des tests dans le tableau récapitulatif suivant :

Type	Test	P-value (ARIMA)	P-value(Holt-Winter)
Court terme	Shapiro.test	2.646 e06	2.2e16
	Box-Ljung test	2.2e16	0.02963
	t.test	0.951	0.3682
Long terme	Shapiro.test	4.497e05	0.005659
	Box-Ljung test	0.07689	0.278
	t.test	0.1832	0.2261

TABLE 3.1: Comparaison des résultats des tests

Résultats de prédiction :
Après avoir appliquer les modèles sur les séries que nous avons étudié nous avons eu les résultats suivant :

1. **Prédiction à court terme**

   (i) **Par la méthode ARIMA :**

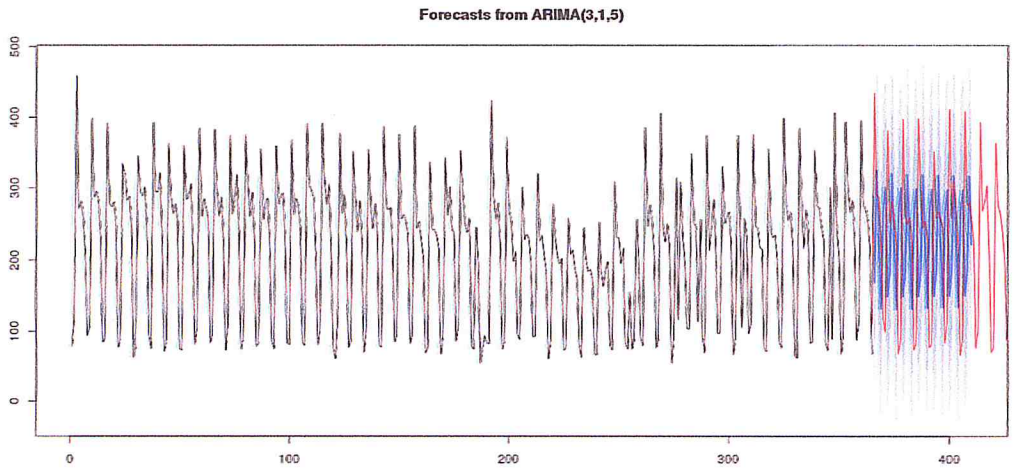


FIGURE 3.11: Prédiction à court terme (ARIMA)

   (ii) **Par la méthode Holt-Winter**

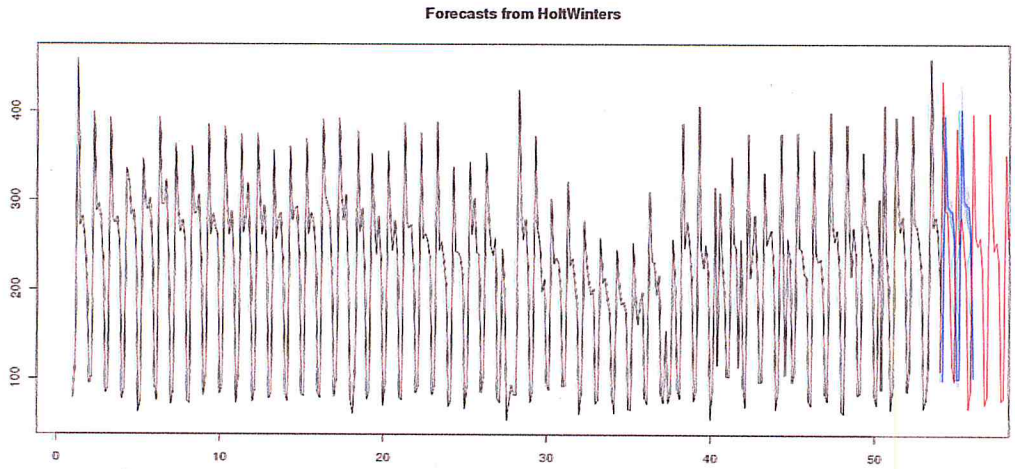


FIGURE 3.12: Prédiction à court terme (ARIMA)

Discussion : Les deux graphe précédents représentent la série dont nous avons appliqué les test avec en bleu les valeurs prédites en gris l'intervalle de confiance, et en rouge les donnée réel pour vérifier si le modèle appliqué donne des bonnes prédictions ou non.

Nous constatons que dans le graphe de Holt-Winter les valeurs sortent de l'intervalle de confiance,mais ce n'est pas le cas pour le graphe de ARIMA donc nous pouvons conclure que le modèle ARIMA s'adapte mieux à ce type de prédiction par rapport au modèle de Holt-Winter.

2. **Prédiction à long terme**

   (i) **Par la méthode ARIMA :**

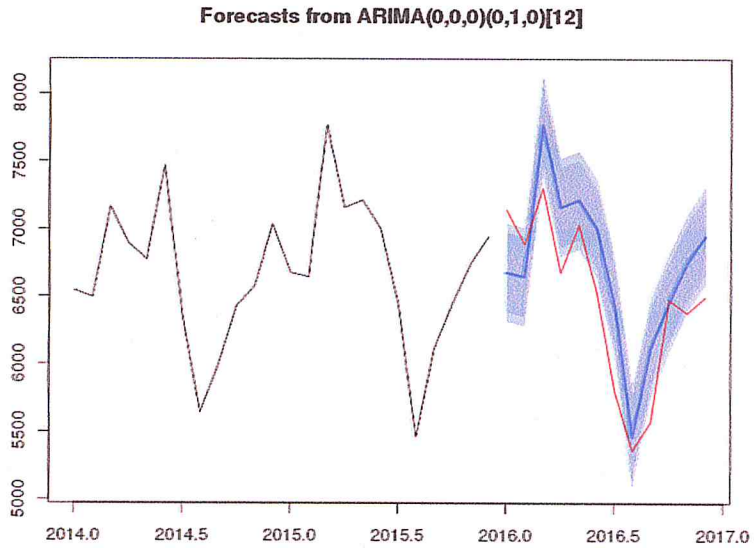


FIGURE 3.13: Prédiction à long terme (ARIMA)

(ii) **Par la méthode Holt-Winter :**

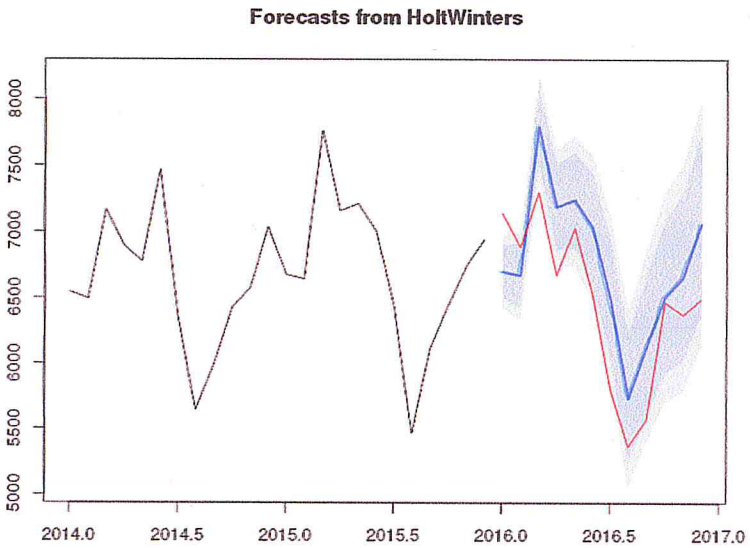


FIGURE 3.14: Prédiction à long terme (Holt-Winter)

Discussion Tout comme dans les graphe de prédiction à court terme, les deux graphes précédents contiennent la série utilisé avec les résultats de prédiction et les données de test. On remarque que les deux modèles donnent presque les mêmes résultats dans la plus part du temps, sauf dans la deuxième moitié de l'année 2016. On remarque que dans le graphe de prédiction du modèle ARIMA il y a un pic tandis que dans le graphe de test il y a une chute, et on remarque aussi que le graphe de prédiction de la méthode Holt-Winter est plus proche de graphe de taste par rapport au graphe de ARIMA. Il est à noter que la méthode de Holt-Winter s'adapte parfaitement au séries qui possède un comportement saisonnier, et c'est le cas pour notre série comme le montre la figure suivante :

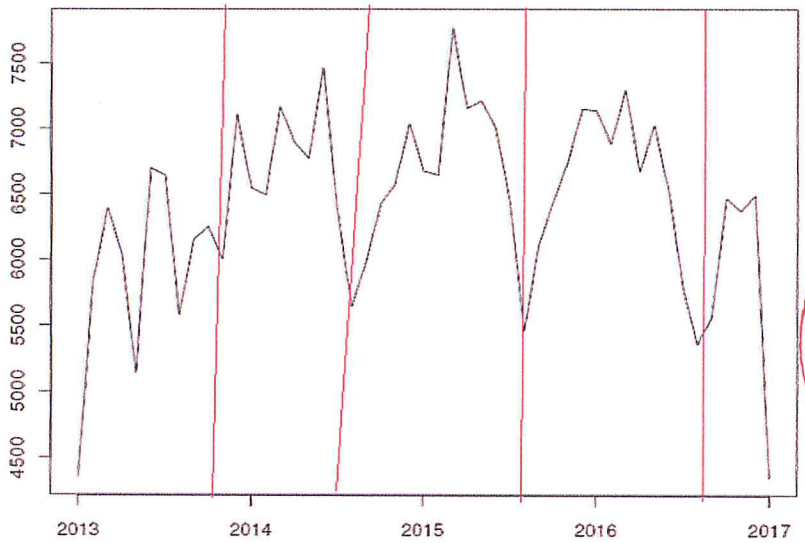


FIGURE 3.15: Série saisonnière

Donc nous pouvons conclure que la meilleure méthode pour ce genre de prédiction est bien la méthode Holt-Winter.

4 matériel et temps d'exécution

Pour réaliser les test précédents nous avons utilisé un ordinateur DELL i5 qui a les caractéristiques suivants.

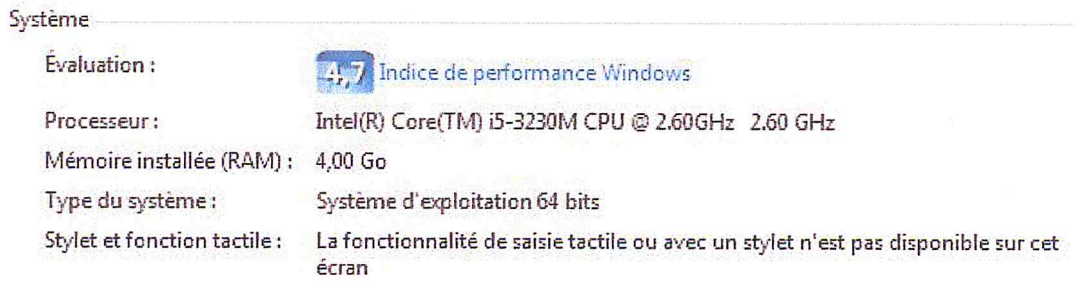


FIGURE 3.16: Caractéristiques du matériel utilisé

En utilisant ce matériel et le système d'exploitation Kali linux nous avons remarqué que temps d'exécution de l'analyse varie entre 10 et 20 secondes selon la quantité de données utilisé.

5 conception UML

UML permet de construire plusieurs modèles d'un système : certains montrent le système du point de vue des utilisateurs, d'autres montrent sa structure interne, d'autres encore en donnent une vision globale ou détaillée. Les modèles se complètent et peuvent être assemblés.

Ils sont élaborés tout au long du cycle de vie du développement d'un système (depuis le recueil des besoins jusqu'à la phase de conception) [46].

Identification des acteurs

Un acteur est l'idéalisation d'un rôle joué par une personne externe, un processus ou une chose qui interagit avec un système, dans notre système les acteur serons :

1. Le décideur : c'est l'utilisateur final principal du système, il doit être capable d'effectuer des analyses et visualiser les résultats et les données.
2. L'administrateur : c'est l'utilisateur qui sera chargé de gérer le système (gérer les données et les utilisateurs)

5.0.1 Diagramme de cas d'utilisation général

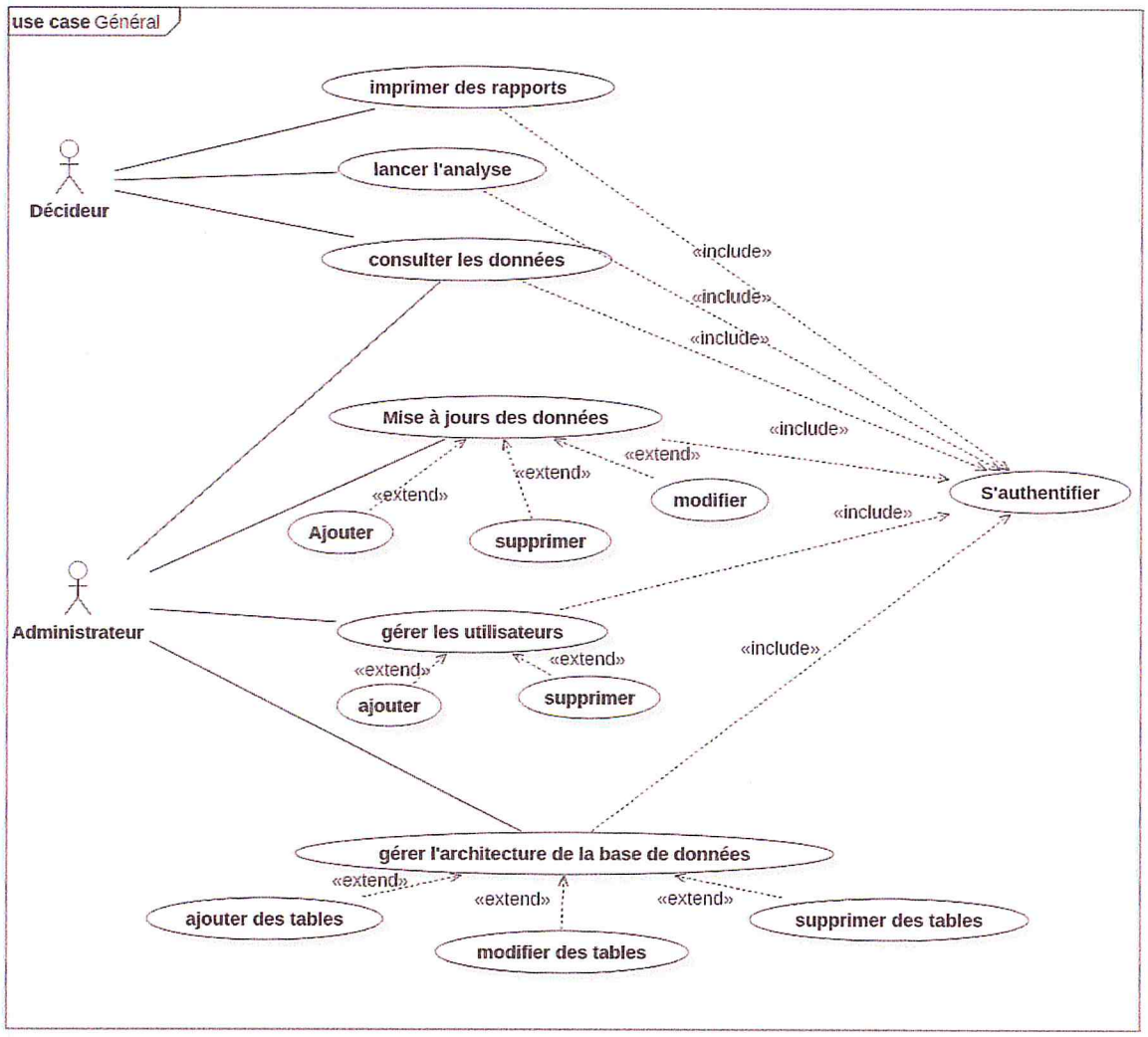


FIGURE 3.17: Diagramme de cas d'utilisation générale

5.1 Diagramme de classes

Alors que le diagramme de cas d'utilisation montre un système du point de vue des acteurs, le diagramme de classes en montre la structure interne. Il permet de fournir une représentation abstraite des objets du système qui vont interagir pour réaliser les cas d'utilisation.
Le diagramme suivant représente le diagramme de classes ne notre system :

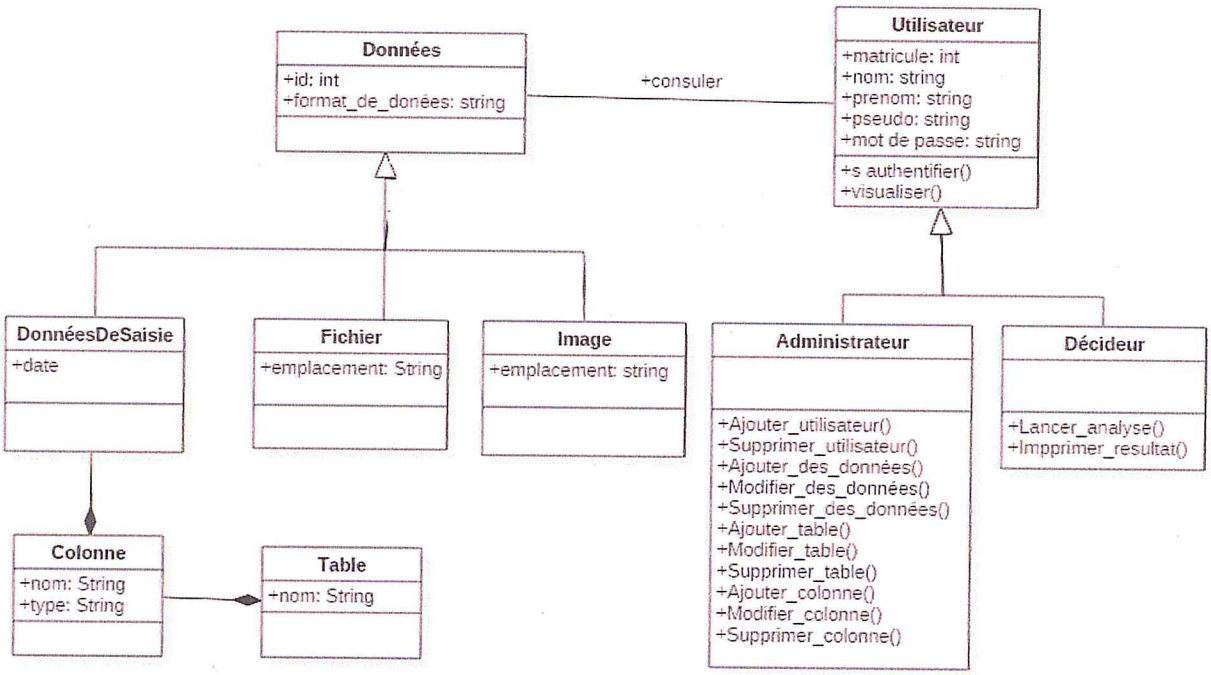


FIGURE 3.18: Diagramme de classes

5.2 Diagramme de cas d'utilisation

Un cas d'utilisation est une manière spécifique d'utiliser un système. Les acteurs sont à l'extérieur du système ; ils modélisent tout ce qui interagit avec lui.
Un cas d'utilisation réalise un service de bout en bout, avec un déclenchement, un déroulement et une fi n, pour l'acteur qui l'initie [46].

5.3 Diagramme de séquence systèmes

Les diagrammes de séquences permettent de décrire COMMENT les éléments du système interagissent entre eux et avec les acteurs :

— Les objets au coeur d'un système interagissent en s'échangent des messages.

— Les acteurs interagissent avec le système au moyen d'IHM (Interfaces Homme-Machine).

5.3.1 Diagramme de séquence authentification

L'authentification est une étape importante pour la vérification des utilisateur et pour gestion des droits d'accès. La figure 3.4 montre le diagramme de sequence d'authentification.

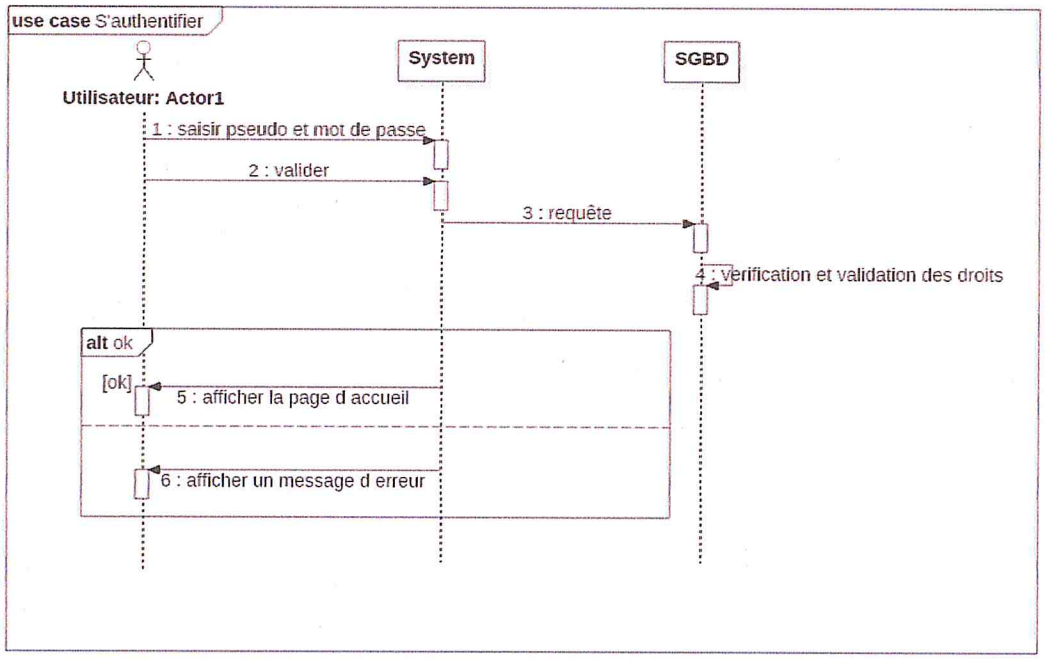


FIGURE 3.19: Diagramme de séquence authentification

5.3.2 Diagramme de séquence consulter les données

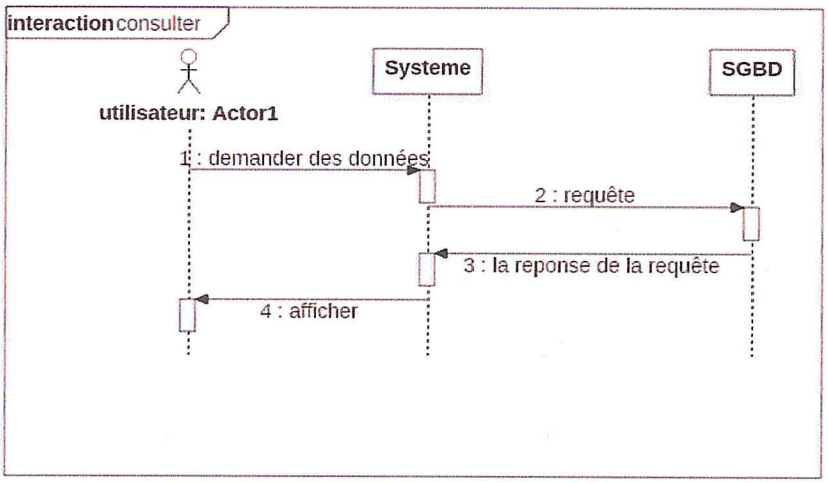


FIGURE 3.20: Diagramme de séquence consulter les données

5.3.3 Diagramme de séquence mise à jours des données

La mise à jour des données est la responsabilité de l'administrateur du système, elle a pour objectif d'assurer la fiabilité du système.

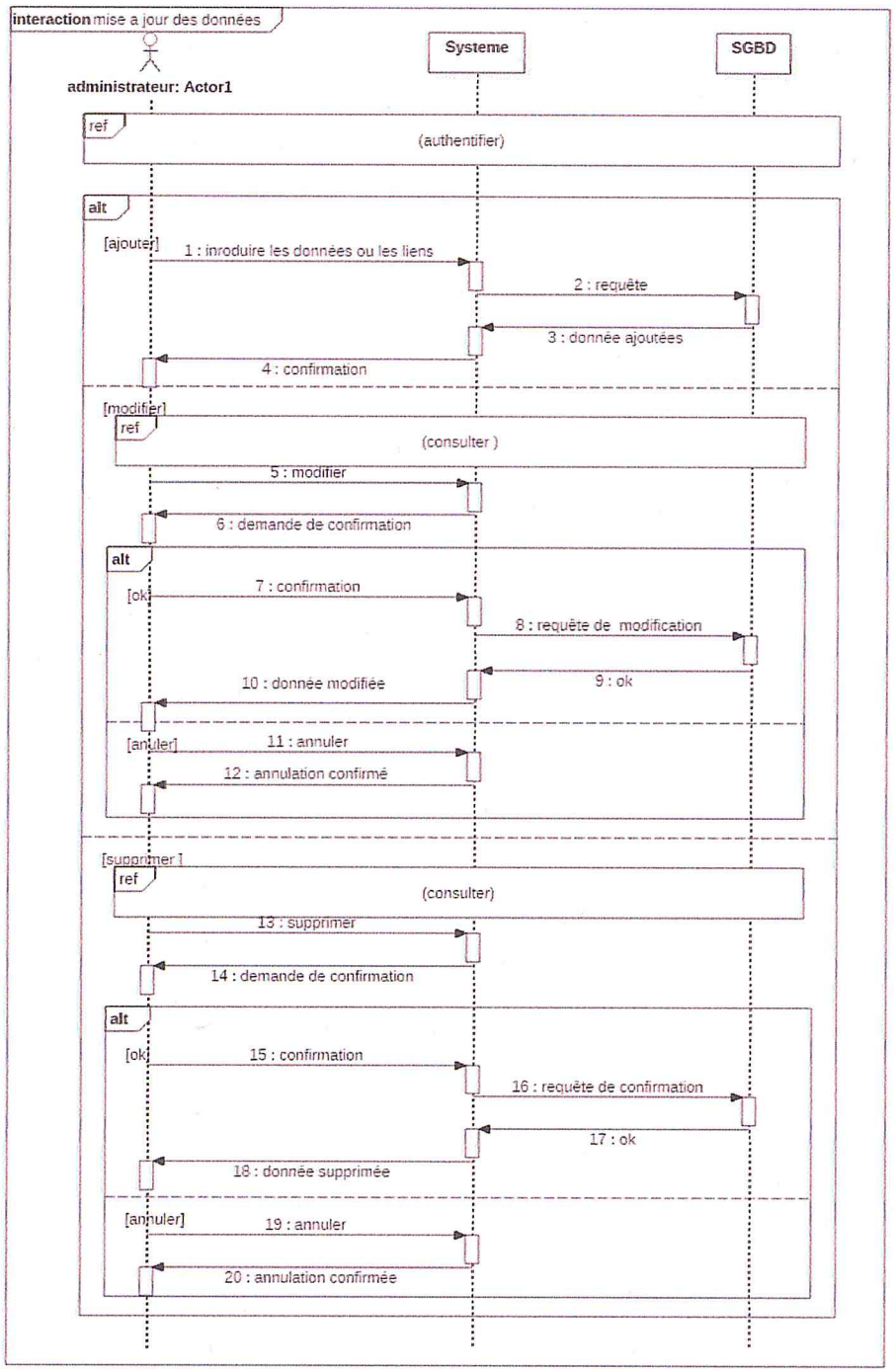


FIGURE 3.21: Diagramme de séquence mise à jours des données

5.3.4 Diagramme de séquence lancer l'analyse

Effectuer l'analyse est le but principale de notre système, c'est elle qui permet l'obtention des résultats. La figure suivante montre les étapes a suivre pour effectuer une analyse :

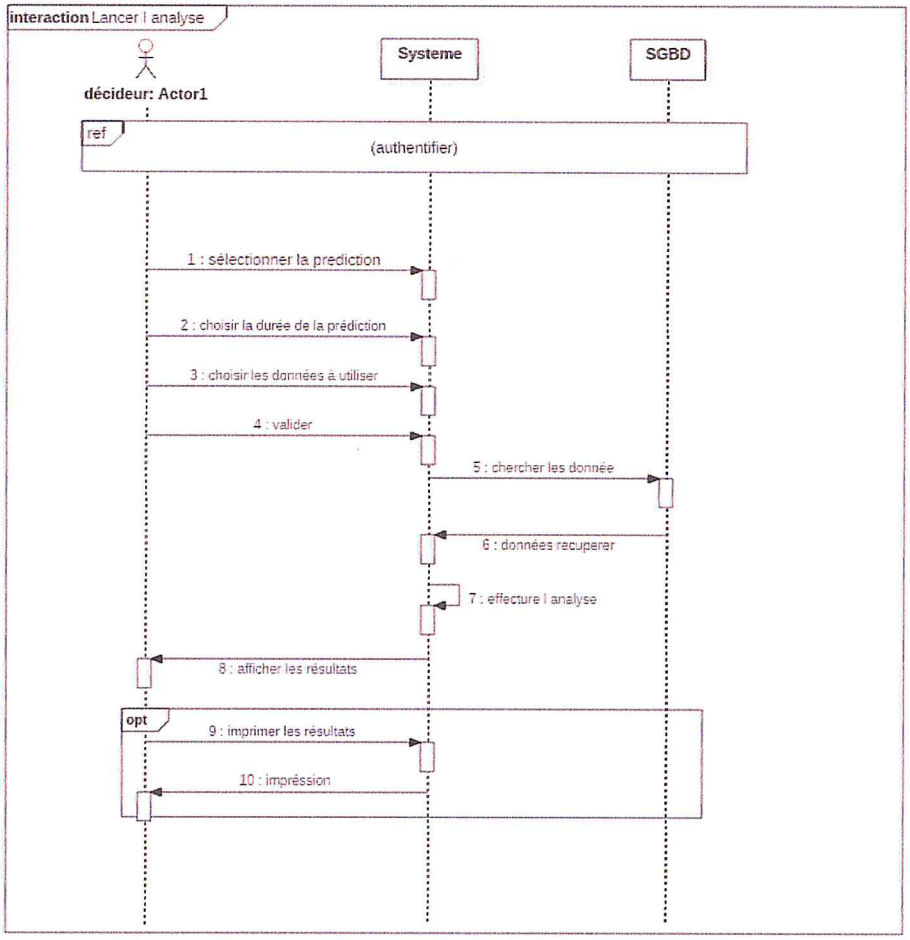


FIGURE 3.22: Diagramme de séquence lancer l'analyse

6 conclusion

La conception proposée comporte toutes les étapes nécessaire pour la réalisation du système envisagé (la collection des données, la pré-traitement, l'estimation du modèle et l'interprétation du modèle), nous avons donner aussi une idée globale du système en utilisant les différents diagrammes UML. Dans le chapitre suivant nous présentons l'implémentation de la solution que nous venons d'expliquer.

Chapitre 4

Implémentation et mise en œuvre

1 Introduction

Après avoir conçu le système, Nous sommes passé à l'étape de déploiement et réalisation de système. Dans ce chapitre le projet prend ses dimensions techniques où nous allons présenter l'architecture technique choisie et la justification associée à chaque choix.

2 Choix de plateformes

Etant donné que Hadoop et spark sont actuellement les solutions prédominantes dans le marché, et que notre projet s'inscrit dans un cadre de Big Data, nous avons donc opté pour ces deux technologies, il existe plusieurs ressources qui expliquent Hadoop et Spark et leurs architectures en détails. Dans ce qui suis nous allons détailler les parties les plus importantes de notre solution et le choix des composants qui font partie de cet écosystème de Hadoop. Ainsi que le choix de la distribution que nous avons effectué.
En fait Hadoop et Spark sont tous les deux des frameworks big data, mais ils n'ont pas vraiment le même usage. Hadoop est essentiellement une infrastructure de données distribuées : ce framework Java libre distribue les grandes quantités de données collectées à travers plusieurs nœuds (un cluster de serveurs x86), et il n'est donc pas nécessaire d'acquérir et de maintenir un hardware spécifique et coûteux. Hadoop est également capable d'indexer et de suivre ces données Big Data, ce qui facilite grandement leur traitement et leur analyse par rapport à ce qui était possible auparavant. Comparativement, Spark sait travailler avec des données distribuées. Mais il ne sait pas faire du stockage distribué. Il a donc besoin de s'appuyer sur un système de stockage distribué.

3 Installation de Hadoop

Le téléchargement de Hadoop [1] est simple, il suffit de se rendre sur la plateforme officielle de Apache Hadoop et d'accéder à la dernière version stable du framework qui est 2.7.2. Une fois connecté avec l'utilisateur Hadoop **huser** il suffit de décompresser l'archive téléchargée dans le répertoire de huser.

1. **Connection avec l'utilisateur Hadoop huser**

1. http ://hadoop.apache.org/releases.html

```
[root@namenode1 ~]# su - huser
[huser@namenode1 ~]$
```

2. **décompression de l'archive**

```
[huser@namenode1 ~]$ tar -xzf hadoop-2.7.2.tar.gz
```

4 Spark

Quand on parle d'un écosystème Hadoop, tout le monde pense à HDFS et MapReduce. Durant l'adolescence de Hadoop, il existait un seul moyen pour exécuter les Jobs et les transformations sur les données et ceci en passant par MapReduce. Aujourd'hui, Yarn a été introduit, il permet l'accès à HDFS sans passer par MapReduce, cela a donné naissance au Spark ou les statistiques montrent qu'il est en train de remplacer MapReduce, ce dernier va disparaître avant 2020 selon certaines estimations.

5 Le logiciel R

Le logiciel R est un logiciel de statistique créé par Ross Ihaka et Robert Gentleman . Il est à la fois un langage informatique et un environnement de travail : les commandes sont exécutées grâce à des instructions codées dans un langage relativement simple, les résultats sont cachés sous forme de texte et les graphiques sont visualisés directement dans une fenêtre qui leur est propre. Pour pouvoir profiter pleinement de ce logiciel quand il s'agit des séries chronologiques, il faut importer le package forecast.

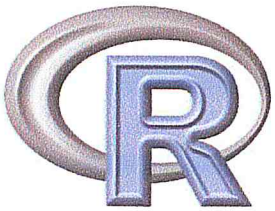


FIGURE 4.1: Logo du logiciel R

6 RStudio

RStudio [2] est un outil apparu récemment et qui vient combler un manque dans la collection des outils associés à R : il s'agit d'un environnement de développement intégré fonctionnel, libre, gratuit et multiplateforme. C'est un environnement facilitant la saisie, l'exécution de code, la visualisation des résultats, etc. RStudio est multiplateforme, donc nous pouvons le télécharger et le faire fonctionner aussi bien sous Windows, Mac OS X ou Linux .

FIGURE 4.2: Logo du logiciel RStudio

7 NetBeans et Sun Java 8

NetBeans

Netbeans [3] est un environnement de développement intégré (EDI), placé en open source. En plus de Java, NetBeans permet la prise en charge native de divers langages tels le C, le C++, le JavaScript, le XML, le Groovy, le PHP et le HTML, ou d'autres (dont Python et Ruby) par l'ajout de greffons. Il offre toutes les facilités d'un IDE moderne (éditeur en couleurs, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web). Compilé en Java, NetBeans est disponible sous Windows, Linux, Solaris (sur x86 et SPARC), Mac OS X ou sous une version indépendante des systèmes d'exploitation (requérant une machine virtuelle Java). Un environnement Java Development Kit JDK est requis pour les développements en Java.

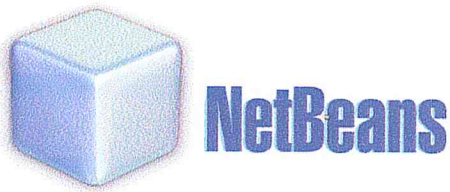


FIGURE 4.3: Logo de NetBeans

2. https ://www.rstudio.com/products/rstudio/download/
3. https ://netbeans.org/downloads/

Sun Java

Le Java Development Kit [4] (JDK) désigne un ensemble de bibliothèques logicielles de base du langage de programmation Java, ainsi que les outils avec lesquels le code Java peut être compilé.

FIGURE 4.4: Logo du Java

8 Réalisation de la solution

Dans le but d'implémenter la solution conçue, une multitude de techniques et méthodologies devront être mises en place et cela pour couvrir tous les besoins et les exigences du déploiement de telles solutions. Dans cette partie nous présentons la mise en œuvre des différentes parties du système de prédiction pour l'aide à la décision.
Dans le but de donner un apport considérable nous allons nous baser sur la force analytique qu'offre la plateforme R via R studio. Pour profiter pleinement de la solution de stockage, la mise en place une solution analytique compatible avec Spark doit être déployée. C'est dans ce sens que tout au long de la partie analytique nous allons nous baser sur la distribution SparkR, permettant de profiter des caractéristiques de R dans l'environnement Spark.

9 Interfaces de l'application

Pour assuré le fonctionnement et l'utilisation de système, il est important d'avoir une interface qui permet de monitorer le système et effectuer les tâches demandées. Dans ce qui suit nous allons démontrer les interface de notre système.

4. https ://www.java.com/fr/download/

9.1 Interface d'authentification

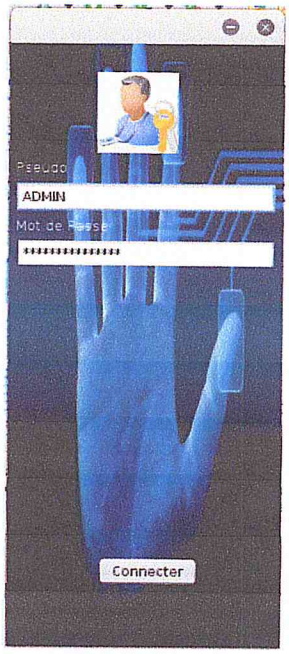


FIGURE 4.5: Interface d'authentification

L'interface d'authentification est la première interface qu'on voit une fois lancer l'application, dans cette interface l'utilisateur doit entrer son pseudo et son mot de passe pour pouvoir accéder à son espace que se soit décideur ou administrateur.

9.2 Espace décideur

Le décideur est l'utilisateur principale du système, il doit être capable de réaliser les opération suivante :

— Visualiser les données (1) .

— Effectuer des prédiction (2).

La figure suivante donne un aperçu sur la page d'accueil du décideur :

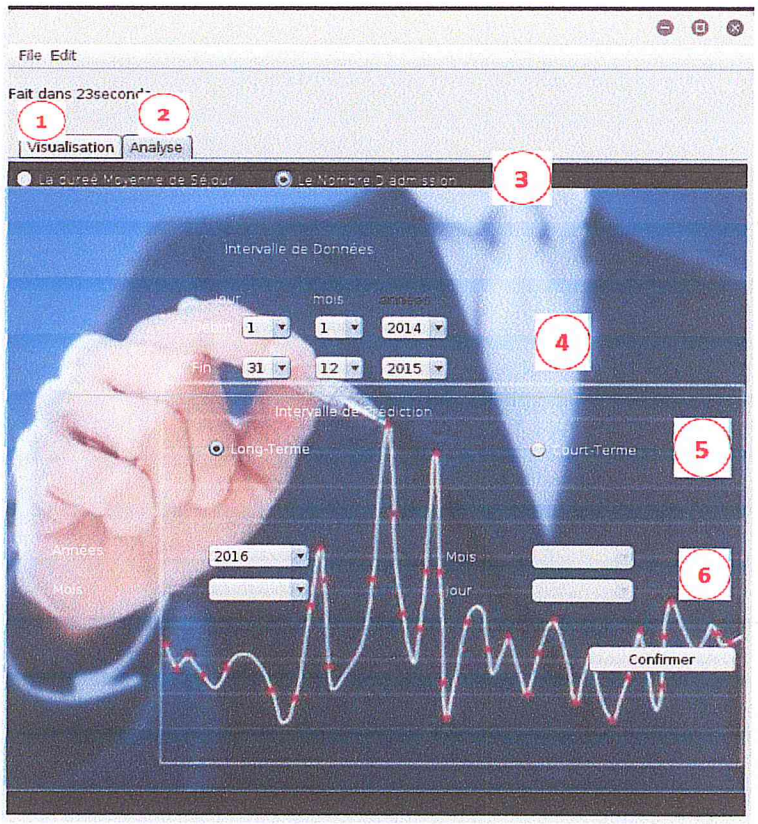


FIGURE 4.6: Interface décideur

1. Pour effectuer une analyse le décideur doit remplir les champs dans la figure, tous d'abord il doit choisir le type de prédiction (nombre de patient ou la durée moyenne de séjour)(3), ensuite il doit choisir les données à utiliser (4), ensuite il choisi les durée de prédiction (court ou long terme) (5) et la durée exacte(6).
   Après la validation, l'affichage sera le suivant :

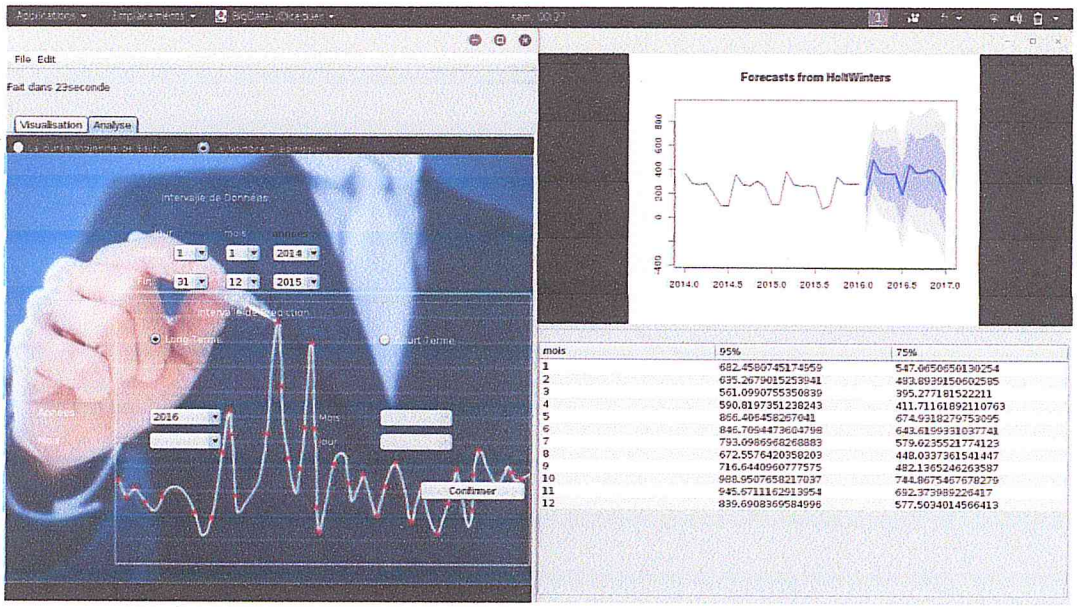


FIGURE 4.7: Interface d'analyse

2. Pour la visualisation des données le décideur doit choisir la période de données à affiché (1), deux type de présentation sont possible : soit par tableaux (2), soit par graphe (3), le décideur a la possibilité d'afficher le graphe d'une grande ou petite taille (4).

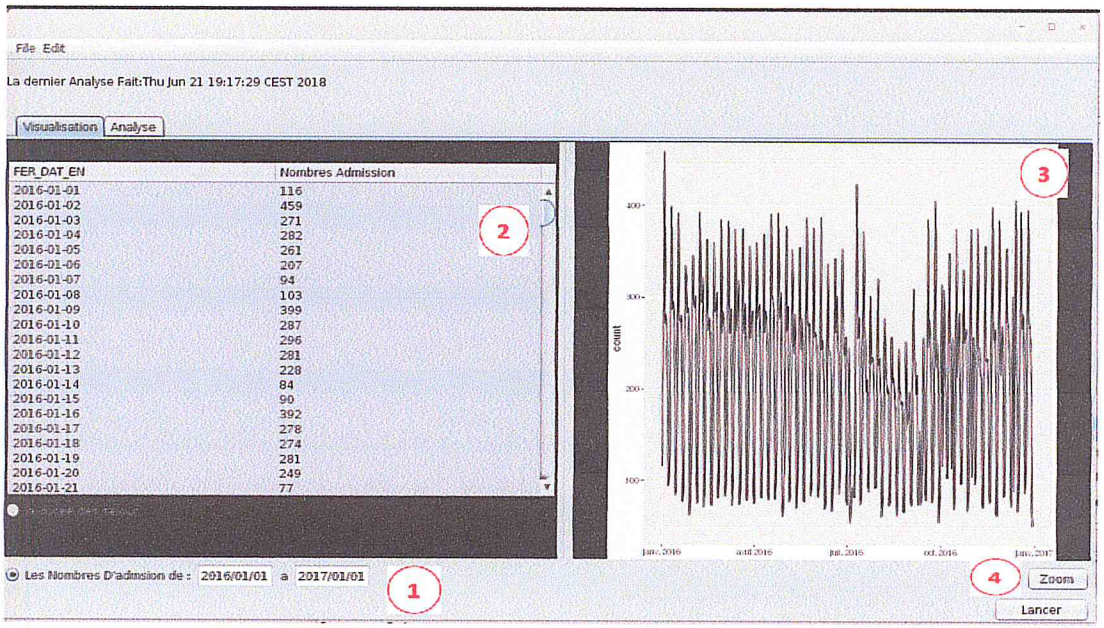


FIGURE 4.8: Interface de visualisation

9.3 Espace administrateur

L'administrateur est responsable de la gestion de l'application, parmi les actions qu'il peut faire on trouve :

— Gérer les données (ajouter, modifier, supprimer).

— Gérer les utilisateur (ajouter, supprimer).

— **Gérer les données :**

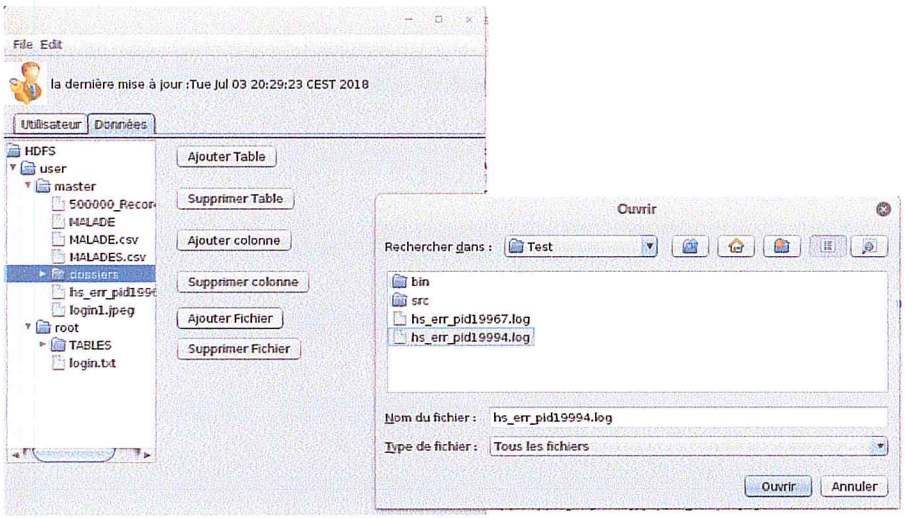


FIGURE 4.9: Interface gérer les données

— **Gérer les utilisateur :**

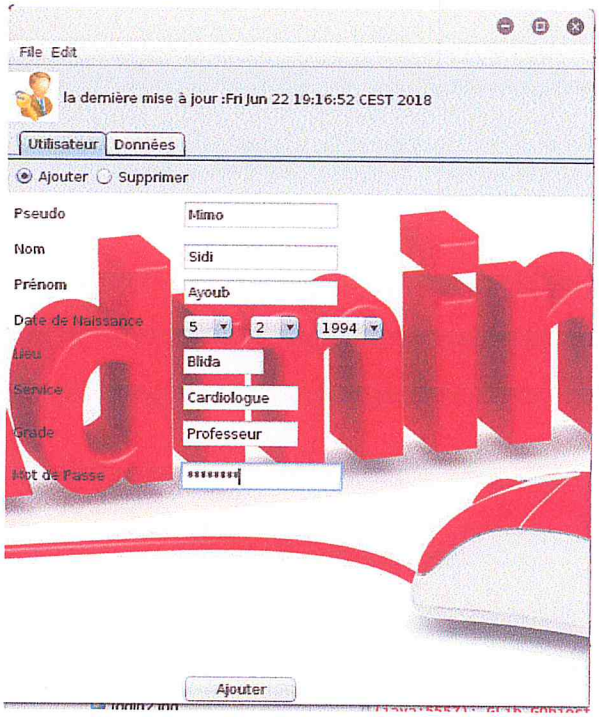


FIGURE 4.10: Interface gérer utilisateurs

— **Ajouter**

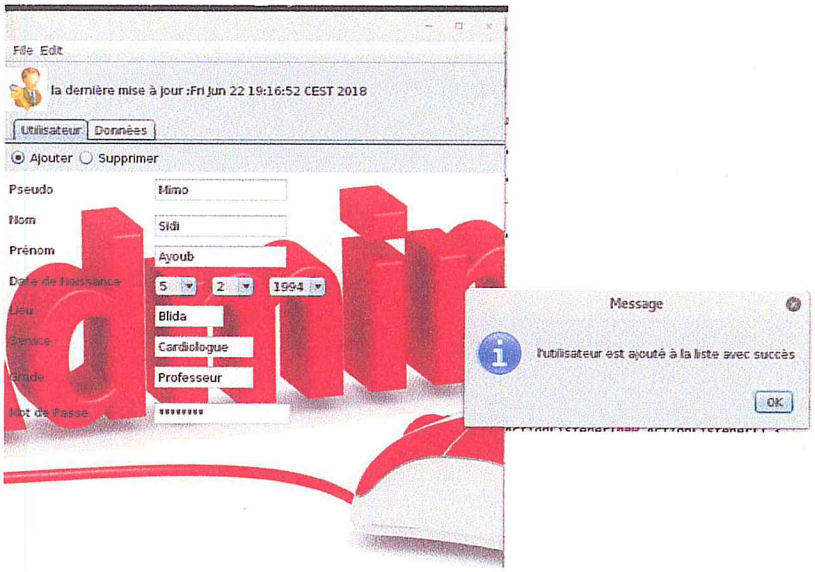


FIGURE 4.11: Interface ajouter utilisateur

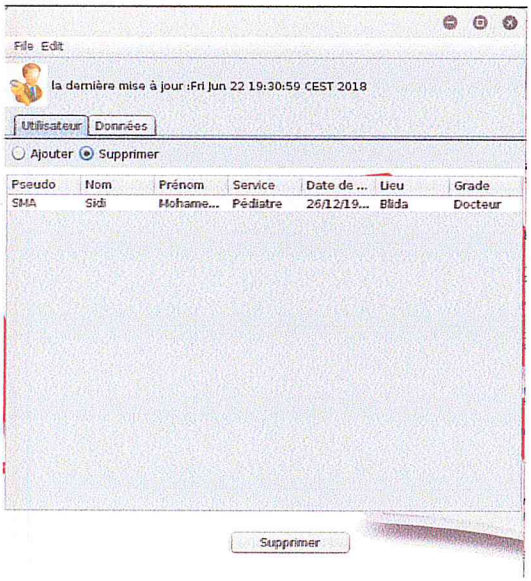


FIGURE 4.12: Interface ajouter utilisateur

— **Supprimer**

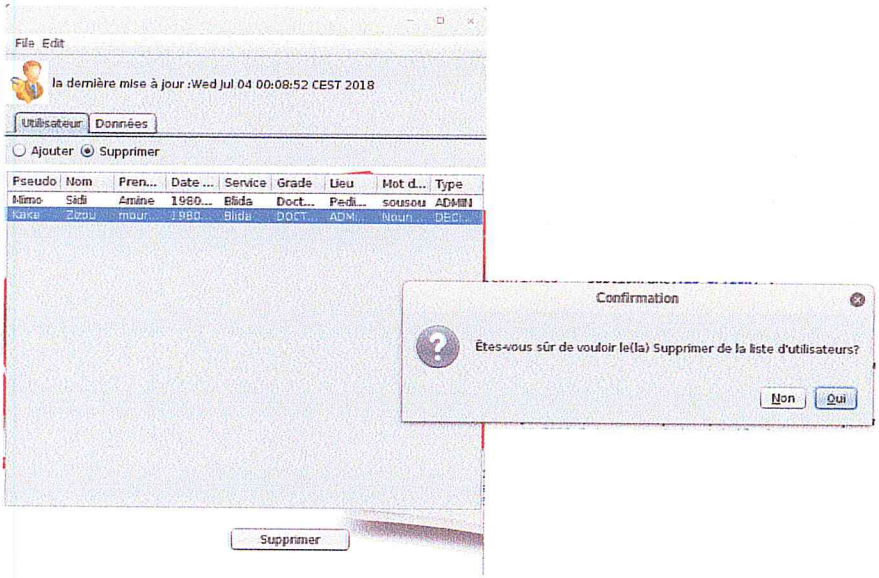


FIGURE 4.13: Interface supprimer utilisateur

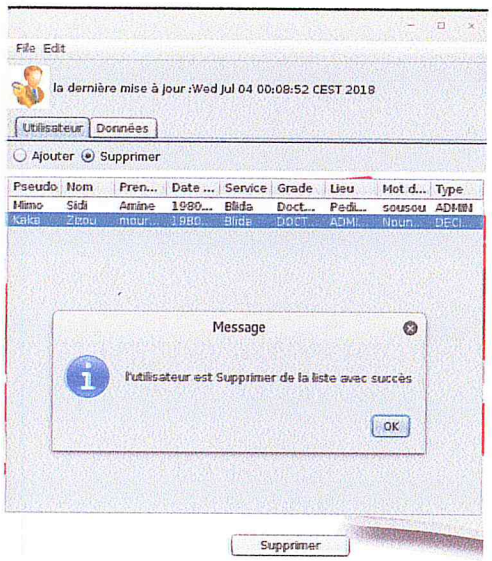


FIGURE 4.14: Interface supprimer utilisateur

10 Conclusion

Dans ce chapitre, nous avons décrit la démarche suivie ainsi que les éléments techniques pour concrétiser l'architecture proposée. Nous avons tout d'abord présenté l'environnement de développement ainsi que les différents outils utilisés, ensuite nous avons validé les modèles utilisés en faisant un exemple pour chaque type de prédiction. Nous avons ensuite donné une description détailler de notre système à travers des fenêtre de capture qui représente les interfaces de ce dernier, qui sont conçues de manière à etre conviviales et simples d'utilisation.

Chapitre 5

Conclusion générale

Ce projet nous a permis principalement de toucher les technologies Big Data et le domaine de l'analyse prédictive, où nous étions en face d'un ensemble de choix architecturale. Le système complet que nous avons conçu peut-être au fait généralisé en un système Big Data qui permet de facilité la prise de décision dans les services d'admission des hôpitaux en appliquant les techniques d'analyse prédictive.
Le développement d'un système de prédiction pour les service d'admission des hôpitaux est d'une complexité majeur, comme il implique l'utilisation des techniques de Big Data qui restes toujours nouvelles, l'application du processus de data mining et l'utilisation des techniques mathématiques de l'analyse prédictive.
Le système que nous avons réalisé permet principalement de prédire deux facteurs critiques qui devront aider à la prise de décision afin de garantir l'optimisation de l'utilisation de ressources, et la bonne prise en charge des patients à savoir le nombre de patient et la durée moyenne de séjours.
Les prédictions que fait notre système peuvent être de long ou de court terme selon le choix du décideur. Pour effectuer ces prédictions nous avons jugé que la technique de série chronologique s'adapte parfaitement à notre système.
Nous avons comparé les résultats de la méthode ARIMA et la méthode Holt-Winter qui sont deux modèles de séries chronologique pour choisir la méthode qui donne des prédictions plus fiable pour chaque cas. Après avoir comparé les résultats nous avons décidé d'utiliser la méthode ARIMA pour les prédictions à court terme et la méthode de Holt-Winter pour les prédictions à long terme.
Pour rendre ce système utilisable nous avons développé une interface simple pour facilité l'utilisation de ce dernier.

Dans ce mémoire composé de quatre chapitre nous avons commencer par une introduction générale pour mettre le lecteur au cœur du sujet, le premier chapitre a été consacré au Big Data où nous avons défini les différents concepts liés au Big Data,les caractéristiques de Big Data ainsi que les domaines où il est utilisé. Dans le deuxième chapitre nous avons présenter le domaine de l'analyse prédictive, ses méthodes et quelques exemples de solutions de l'analyse prédictive dans les services d'admission hospitalières. quant au troisième chapitre, il a été consacré pour la conception de notre système, il avait pour but de donner une idée de quoi il va ressembler ce système. Dans le quatrième chapitre qui est le dernier nous avons justifier les choix des techniques utilisées, ensuite nous avons montré les interface et comment utiliser le système.

Perspective et travaux Future :

Nous envisageons enrichir notre travail par :

— L'utilisation d'autres source de données telle que les données de météo, les données de mariage ...etc ces données peuvent influencer sur le nombre de patients.

— La réalisation d'autres prédictions comme le nombre d'accouchement.

Bibliographie

[1] M. Barlow,"The Culture of Big Data",Gravenstein Highway North Sebastopol CA,United States of America,2013

[2] M. A. Beyer and D. Laney, "The importance of big data : A definition", Stamford, CT : Gartner, 2012.

[3] S.Amir, M.Bilasco Ioan, T.Urruty and C.Djeraba,"MuMIE : Une Approche Automatique pour l'Interopérabilité des Métadonnées", dans les actes des 11èmes Journées Francophone sur l'Extraction et la Gestion des Connaissances (EGC-2011), Brest, France, 25-28 Janvier, Cepadues Editions,2011.

[4] M.DEY, "Big Data : usages et perspectives",dans les actes de la conférence internationale,Le potentiel et les défis du Big Data UIMM, PARIS, juillet 2013.

[5] Global Pulse, "Big Data for Development : Challenges and Opportunities". Global Pulse, 2012.

[6] L.Bentley and J.Whitten , "Systems Analysis and Design for the Global Enterprise SEVENTH EDITION", New York : McGraw-Hill Companies, 2007.

[7] C.Eaton, D.Dirk, D.Tom, L.George and Z.Paul, "Understanding Big Data". Mc Graw Hill.2007.

[8] Les applications du Big Data - La finance pour tous ,https ://www.lafinancepourtous.com/decryptages/finance-et-societe/nouvelles-economies/big-data/les-applications-du-big-data/ ,[consulter le 02/03/2018].

[9] Terrorisme et Big Data – Le Big Data pour lutter contre le risque d'attentats, https ://www.lebigdata.fr/terrorisme-et-big-data-0607,[consulter le 23/03/2018].

[10] Cartographie des bases de données publiques en santé https ://www.data.gouv.fr/dataset/cartographie-des-bases-de-donnees-publiques-en-sante. [Consulter le 23/03/2018].

[11] Le plan européen "eHealth action plan 2012-2020"prévoit la parution en 2014 d'un livre VERT sur la " santé mobile " : http ://epractice.eu/en/library/5362646 [consulter le 24/03/2018].

[12] N. Marz, J. Warren, "Big Data : Principles and best practices of scalable realtime data systems". Manning Publications, 2013.

[13] C.Erwann,"Hadoop : Optimisation et Ordonnancement",Université Francois-Rabelais, Tours,2014.

[14] R.Bourtembourg, "HDB++ : HIGH AVAILABILITY WITH CASSANDRA,TANGO Meeting",ESRF,May 2015 cassandra-client February, 2010.

[15] R. MichaelBerthold, N.Cebron, F.Dill, Thomas R. Gabriel, Tobias K otter, Thorsten Meinl, Peter Ohl, Kilian Thiel and Bernd Wiswedel, "KNIME – The Konstanz Information Miner Version 2.0 and Beyond",University of Konstanz,Germany,2007.

[16] O.Dridi,"Machine Learning With Spark",Centre d'Excellence en Technologies de l'Information et de la Communication,Novembre 2015.

[17] Comparatif Hadoop : top 7 des vendeurs commerciaux de distributions,https ://www.lebigdata.fr / Analytics / Data Analytics,[consulter le 16/03/2018].

[18] N.Mahé, "Les dernières actualités IBM ", interview, France, 2014.

[19] EMC Introduces World's Most Powerful Hadoop Distribution : Pivotal HD,https ://hk.emc.com/about/news/press/2013/20130225-04.htm,[consulter le 02/04/2018].

[20] J.Sun, A.Dholakia, W.Yang, D.Kangas, Lenovo Big Data Reference ,Architecture for the MapR, Converged Data, Platform Configuration reference number : BDAMAPRXX53 ,30 mars 2017.

[21] MapR : une solution Big Data pour les entreprises dynamiques ,https ://www.lebigdata.fr/mapr, [consulter le 01/04/2018].

[22] S. Tufféry, Data Mining et statistiques décisionnelle : L'intelligence des données, TECHNIP ed., France, 2007.

[23] G. Gardarin, Bases de données. Objet et relationnel Eyrolles ed., 2011.

[24] https ://www.lebigdata.fr/data-mining-definition-exemples,qu'est ce que l'exploration de données ?, consulté le 02/04/2018.

[25] M. J. BERRY, G. S. LINOFF, Mastering Data Mining : The Art and Science of Customer Relationship Management, 2000.

[26] J.STEFANOWSKI , Data Mining – Clustering ,mémoire de master à Poznan University of Technology, 2008/2009.

[27] M.Gera, S.Goel, Data Mining - Techniques, Methods and Algorithms : A Review on Tools and their Validity, International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 18, March 2015.

[28] M. Ramageri Bharati. DATA MINING TECHNIQUES AND APPLICATIONS , Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305.

[29] S.Gupta,a Regression Modeling Technique on Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 9, April 2015.

[30] J.Lyon,Predicting the future with our robot overlords. article, 2016.

[31] D. Gutierrez, BIGDATA de l'analyse prédictive,Guide Inside,

[32] GROUPE CGI INC, Analyse prédictive. L'essor et la valeur de l'analyse prédictive dans la prise de décisions. JUIN 2013.

[33] B.HOUMADI , " ÉTUDE EXPLORATOIRE D'OUTILS POUR LE DATA MINING",Mémoire ,UNIVERSITÉ DU QUÉBEC ,Avril 2007.

[34] G.CALAS, Études des principaux algorithmes de data mining,France.

[35] P.Fang et all,"Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network ",Hindawi Publishing Corporation, Journal of Healthcare Engineering, Volume 2016.

[36] B. A. Mobley, R. Leasure, and L. Davidson, "Artificial neural network predictions of lengths of stay on a post-coronary care unit," Heart and Lung, vol. 24, no. 3, pp. 251–256, 1995.

[37] Qualité des services de santé Ontario, Plans d'amélioration de la qualité des hôpitaux ,rapport,2014-2015.

[38] L.Ouellette, Analyse des scénarios pour acheminer lrd patients Monteréalais en attente de service orthopédique, mémoire universitaire, université de Monteréal 2014.

[39] M.Afilal, F.Yalaoui et all , Modèles de prévision robuste de l'afflux des patients des urgences, Revue d'Épidémiologie et de Santé Publique ,2017.

[40] M.-C. Viano , A. Philippe, Cours de Séries Temporelles, Université des Sciences et Technologies de Lille,2004.

[41] J. Pearson,Jones et Joy, " Forecasting demand of emergency care ", Health Care Manag. Sci., vol. 5, no 4, p. 297-305, 2002.

[42] A. M. Mangoud et Abdel-Aal, " Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis ", Comput. Methods Programs Biomed., vol. 56, no 3, p. 235-247, 1998.

[43] R. Champion et al " Forecasting emergency department presentations ", Aust. Health Rev. Publ. Aust. Hosp. Assoc., vol. 31, no 1, p. 83-90, 2007.

[44] B.Zheng et al, "Predictive modeling of hospital readmissions using metaheuristics and data mining",article dans Expert Systems with Applications, 2015.

[45] F.Kadri, S.Chaabane, F.Harrou, C.Tahon. Modélisation et prévision des flux quotidiens des patients aux urgences hospitalières en utilisant l'analyse de séries chronologiques. 7ème conférence de Gestion et Ingénierie des Systèmes Hospitaliers (GISEH), Jul 2014, Liège, Belgique. pp.1-8, 2014.

[46] B.Charroux, A.Osmani, Y.Thierry Mieg, Pearson Education France UML2, 3e édition, 2010

[47] D.Delignières, Séries temporelles – Modèles ARIMA.Séminaire EA "Sport – Performance – Santé". Mars 2000

Chapitre 6

Annexe

Les tableaux suivants pressente les détails de chaque table de la base de données

Attribut	**Désignation**
NA-ENT-MER	Matricule de la mère
NA-DOS-ENF	Matricule du nouveau né
NA-DT-ACCO	Date d'accouchement
NA-HEURE	Heure d'accouchement
NA-ETA-ENF	État d'enfant : vivant ou mot ne
NA-SEX-ENF	Sexe d'enfant
NA-POIDS	Poids
NA-LIT	Num lit ou berceau
NA-PRE-ENF	Prénom d'enfant
NA-NUM-ENR	Num d'enregistrement de la naissance
NA-PARTICU	Particularités :enfant abandonne, mère célibataire,...etc

TABLE 6.1: La table F-naissance

Attribut	Désignation
EV-ENTREE	Matricule du malade
EV-COD-ETA	Code d'établissement qui a évacué le malade
EV-NUM-CHA	Num de la fiche d'évacuation
EV-NOM-MED	Nom du médecin traitant dans l'établissement d'origine
EV-ACCOMP	Information sur l'accompagnateur du malade (Infirmier,...)

TABLE 6.2: La table FEVAC-DE

Attribut	Désignation
EV-ENTREE	Matricule du malade
EV-COD-ETA	Code de l'établissement ou le malade est évacué (établissement récepteur)
EV-NUM-CHA	Num de la fiche d'évacuation
EV-NOM-MED	Nom du médecin traitant
EV-MOTIF	Motif de l'évacuation (voir la table T-MOT-EV)

TABLE 6.3: La table FEVAC-VE

Attribut	Désignation
COD-WILAYA	Code de la wilaya
COD-DAIRA	Code de la daira
CODCOMMUN	Code de la commun
COD-SECT-S	Code du secteur
LIB-COMMUN	Lib de la commun

TABLE 6.4: La table COMMUNE

Attribut	Désignation
COD-TYP	Code du type
COD-WILAYA	Code de la wilaya
COD-SECT-S	Code du secteur
COD-ETAB	Code d'établissement
COD-ANEXE	Code d'annexe
NV-CODE	Nouveau code
ANC-CODE	Ancien code
CODCOMMUN	Code de la commune
LIB-ETAB	Lieu d'établissement
LIB-ABREG	Nom d'établissement
ADR-ETAB	Adresse d'établissement

TABLE 6.5: La table ÉTABLISSEMENT

Attribut	Désignation
LIEN-COD	Code du lien
LIEN-PA	Signification : père(1), mère(2), g.père(3), g.mère(4), frère(5), soeur(6), tante(8), ami(9), voisin(11), autre(99), oncle(7), époux(13), amie(10) voisine(12), épouse(14), fils(15), fille(16), p.fils(17), p.fille(18)
LIEN-SEXE	masculin(1), féminin(2)

TABLE 6.6: La table LIEN DE PARENTÉ

Attribut	Désignation
GM-ENTREE	Matricule du malade
GM-NOM	Nom du garde malade
GM-PRE	Prénom du garde malade
GM-AGE	Age du garde malade
GM-SALLE	Code service-unité du garde malade
GM-LIT	Num lit du garde malade
GM-SEXE	Sexe du garde malade
GM-TYP-PID	Type pièce identité du garde malade : carte identité nationale (1) ; permis de conduire(2) ; passeport(3)
GM-NUM-PID	Num pièce identité du garde malade
GM-DT-DLV	Date délivrance de la pièce d'identité
GM-LIEU-D1	Code wilaya-commune du lieu délivrance de la pièce d'identité
GM-LIEU-D2	Lieu délivrance de la pièce d'identité (détaillé)
GM-LIEN-PA	Code lien de parente du garde malade avec le malade (voir la table T-LIEN-P)
GM-DT-ENT	Date d'entrée du garde malade
GM-HEURE-E	Heure d'entrée du garde malade
GM-CHARGE	Garde Malade pris en charge pour la restauration par l'établissement oui (O) ou non (N)
GM-DT-SORT	Date de sortie du garde malade
GM-HEURE-S	Heure de sortie du garde malade
GM-MOTP	Mot de passe de l'agent qui a saisi le garde malade

TABLE 6.7: La table F-GARD-MA

Attribut	Désignation
FER-ENTREE	Matricule du malade
FER-MOTP	Mot de passe de l'agent qui saisi l'entrée
FER-DAT-EN	Date d'entrée
FER-HEURE	Heure d'entrée
FER-COND	Mode d'admission : voir table MOD-ADM
FER-SERV	code service : voir table FSERVICE
FER-SALLE	Code unite : voir table FSERVICE
FER-LIT	Num lit
FER-SEXE	Sexe
FER-TYP-DT	Type date
FER-DAT-NA	Date de naissance du malade
FER-LIEU-N	Lieu de naissance : code wilaya+code
FER-ADR-MA	Adresse complete du malade
FER-POS-MA	Code wilaya+commune de l'adresse du malade
FER-SIT-FA	Situation de famille : célibataire(c) ;marie(m) ; divorce(d) ; veuf(ve)
FER-NATION	Code nationalité du malade : voir table T-NATION
FER-TELM	Num telephone du malade
FER-CSP	Code de la catégorie voir table : T-CSP
FER-PROF	Detail de la professiondu malade
FER-AS-DEM	A = Malade assure ; D= Demune
FER-PA-CN	Nom de la personne a contacter
FER-PA-CP	Prenom de la personne a contacter
FER-ADR-C	Adresse de contacte
FER-POS-C	Code wilaya+commune de l'adresse de contacte
FER-ACCOMP	Renseignement sur l'accompagnateur du malade

FER-ACOMP2	Suite renseignement sur l'accompagnateur du malade
FER-NOM-ME	Code du médecin qui a accorde l'admission du malade
FER-MED-TR	Nom du médecin traitant qui a oriente le malade
FER-ETAB	Code établissement du médecin qui a oriente le malade (voir table T-ETABLI)
FER-G-SANG	Groupe sanguin
FER-ANCNUM	Matricule du malade de l'admission précédente
M-TYP-PID	Type de la pièce d'identité : carte d'identité,permis de conduire...
M-NUM-PID	Num Dde la pièce d'identité
M-DT-DLV	Date de délivrance de la pièce d'identité
M-LIEU-D1	Lieu de délivrance de la pièce d'identité : code wilayacommune
M-LIEU-D2	Lieu de délivrance de la pièce d'identité détaillé
DT-SOR-MED	Date de sortie médicale
FER-DT-SOR	Date de sortie administrative
HEURE-SORT	Heure de sortie
FER-MOD-SO	Mode de sortie (voir table MOD-SORT)
FER-TYP-DC	Type décès : naturel, accidentel, suspect
NUM-DEC	Num du décès
HEURE-DEC	Heure du décès
FER-DIAG-E	CODE diagnostic d'entrée
FER-DIAG-S	Code diagnostic de sortie
NOM-MED-SO	Nom du médecin accordant la sortie
FER-MOTP-S	Mot de passe de l'agent qui a saisi la sortie

TABLE 6.8: La table F-MALADE

Attribut	Désignation
AT-ENTREE	Matricule du malade
AT-EVACPAR	Évacue par : protection civil ; ambulance ; citoyentaxi ;autre
AT-DT-ACCT	Date accident
AT-HEURE	Heure accident
AT-ADR-ACT	Lieu accident
AT-LIEUACT	Code wilaya-commune du lieu d'accident
AT-CIRCONS	Type d'accident : accident de travail(1) ; accident de circulation(2) ; coups e blessure(3) ; Divers(4)
AT-CIR-MES	Description des circonstance de l'accidente
AT-AUT-C-E	Autorité qui charge de l'enquête :police ; gendarmerie ;autres
AT-REFEREN	Information concernant l'accompagnateur de l'accidente

TABLE 6.9: La table FACCT

Attribut	Désignation
EVAC-ENTRE	Matricule du malade
EVAC-DT-EV	Date d'évacuation vers autre service
EVAC-HEURE	Heure d'évacuation
EVAC-SERV	Service destinataire
EVAC-SALLE	Unité destinataire
EVAC-LIT	Lit destinataire
EVAC-SERVG	Service du garde malade
EVAC-SALG	Unité du garde malade
EVAC-MOTP	Mot de passe

TABLE 6.10: La table FEVAC

Attribut	Désignation
COD-MODS	Code d'admission
LIB-MODS	Mode d'admission :normale(1),évacuation(2),accidente (3),maternité (4),naissance(5), urgence(6),hopital du jour(7),victime événement(8)

TABLE 6.11: La table MOD ADMISSION

Attribut	Désignation
COD-COND	Code de sortie
LIB-MODS	Modedesortie :sortienormale(1),pardécès(2),évacuation (3), évasion(4), contre avis médical (5), transfert étranger (6),mort-né(7)

TABLE 6.12: La table MOD SORTIE

Attribut	Désignation
CODE-CAUSE	Code d'évacuation
MOTIF	Signification du Code d'évacuation

TABLE 6.13: La table MOTIF EVAC

Attribut	Désignation
NA-NATN	Code de la nationalité
NA-DESIGN	La nationalité
NA-PAYS	Pays

TABLE 6.14: La table NATIONALITÉ

Attribut	Désignation
CSP-COD	Code de la profession
CSP-LIB	Nom de la profession
CSP-DETAIL	Description de la profession

TABLE 6.15: La table PROFESSION

Attribut	Désignation
SER-COD	Code du service
SER-DESIGN	Nom du service
SER-RESPON	Nom du responsable

TABLE 6.16: La table SERVICE

Attribut	Désignation
COD-WILAYA	Code de la wilaya
LIB-WILAYA	Le nom de la wilaya
COD-REG-S	Code de la région de la wilaya

TABLE 6.17: La table WILAYA