République Algérienne Démocratique et Populaire

Ministère de l'enseignement Supérieur et de la Recherche Scientifique

Université de Larbi Tébessi – Tébessa –

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

Département : Matématiques et informatique

MÉMOIRE DE MASTER

Domaine : Mathématiques et Informatique

Filière : Informatique

Spécialité : Sécurité et Réseau Informatique

Titre :

Architecture Réseau Basée sur Big Data pour EBM

Présenté par : - Abderrahmane yasmina

- Laissaoui Radia

Devant le jury :

M.C.A	Mr GAHMOUS ABDELLATIF	Université de Tébessa	Président
M.A.A	Mr BRADJI LOUARDI	Université de Tébessa	Encadreur
M.C.A	Mr ZEGGARI	Université de Tébessa	Examinateur

Date de soutenance :25/05/2017

Note : …… **Mention : …….**

Dédicace

MOI Mm ABDERRAHMANE YASMINA

Je dédie mon travaille a ma chère mère , a l'âme de ma chère sœur , mon marie , mes enfants NADJO et SAMAR

MES CHERES AMIES

Je remercie tous qui m'en aider et qui m'en soutenu de loin ou de prés

MERCI

Remerciement

MERCI A VOUS MON ENCADREUR Mr Bradji el ouardi

MERCI A LES NOMBRES DE JURY

Résumé

La médecine basée sur les preuves ou médecine factuelle et devenue actuellement un thème de recherche de haute importance pour les médecins vue les avantages qu'elle offre.

L'informatisation de l'EBM est concentrée uniquement sur la construction des bases de données comportent les articles aident la société médicale a lire et extraire les connaissances dans ont besoins.

Malgré , l'évolution rapide des données vers les Big Data , l'EBM n'a pas profité de cette évolution pour construire et proposer des études (Théoriques et Pratique) pour le soin du patient .

C'est pour cela dans notre travail de fin d'étude , nous proposons une approche théorique d'intégration Big Data et EBM pour arriver a des diagnostics et traitements optimaux pour les patients en intégrant tous les acteurs .

Pour démarche et objectif, le présent manuscrit comporte quatre chapitres.

Chapitre 1 : *Définition de BigData et traitement des donnée non structuré avec NOQGL.*

Chapitre 2 : *Définition d' EBM et les outiles de fouille de donnée avec comparaison des algorithmes les plus utilisé dans le domaine médical*

Chapitre 3 : *Des propositions sur les milieux de travails*

Chapitre 4 : *Définition sur les outilles utilisé pour l'implémentation :*

Cloudera- Weka –algorithme knn

ABSTRACT

Evidence-based medicine or factual medicine and now become a research topic of high importance to doctors view the benefits it offers.

The computerization of the EBM is concentrated solely on the construction of the databases feature articles help the medical company to read and extract knowledge in need.

In spite of the rapid evolution of the data to the Big Data, the EBM did not take advantage of this evolution to construct and propose studies (theoretical and practical) for the care of the patient.

This is why, in our final work, we propose a theoretical approach of integration Big Data and EBM to arrive at optimal diagnoses and treatments for the patients by integrating all the actors.

ملخص

الطب القائم على الأدلة أصبح الآن موضوع بحث مهم للغاية للأطباء من أجل المنافع التي يوفرها.

وقد ركزت حوسبة EBM فقط على بناء قواعد البيانات تحتوي على عناصر تساعد المجتمع الطبي على قراءة واستخراج المعرفة التي نحتاجها .

وعلى الرغم من التطور السريع للبيانات إلى بيانات كبيرة، لم تستفد EBM من هذا التطور لبناء واقتراح دراسات (النظرية والعملية) لرعاية المريض.

هذا هو هدفنا و اقتراحنا في نهاية الدراسة ، فلقد عملنا على تقديم بعض التعريفات ذات صلة بالموضوع وقدمنا مقارنة بين اللوغاريتمات المستعملة في المجال الطبي و استخلصنا الأفضل و ذلك بهدف الوصول إلى التشخيص والعلاج الأمثل للمرضى .

و من أجل هذا الهدف, قدمنا هذا العمل المتواضع والذي يحتوي على أربع فصول

الفصل الأول : يحتوي على تعريف Big Data و اشكالية المعطيات غير المنظمة .

الفصل الثاني : يحتوي على طرح الاشكالية العامة الخاصة بالكم الهائل من المعلومات الطبية المخزنة في قواعد معطيات مختلفة و كيفية التعامل معها من قبل Big Data

الفصل الثالث يحتوي على بعض الاقتراحات العرضية و الرسومات التوضيحية لمراحل العمل.

الفصل الرابع يحتوي على بعض تعريفات حول الأدوات الوسيطية المستعملة لانجاز هذا العمل مثل

تثبيت الجهاز الظاهري

Liste des figures

Liste des Tableaux

INTRODUCTION :

La commodité des systèmes d'information dans les entreprises n'est plus à prouver. Pratiquement aucune entreprise ne peut aujourd'hui s'en passer. L'activité quotidienne des banques, des assurances, des sociétés industrielles, des entreprises de distribution ou des services publics est lié du bon fonctionnement du système d'information.

Celui- ci facilite et optimise les processus métier de l'entreprise, et pratiquement toutes les fonctions (contrôle de gestion, comptabilité, vente, marketing, achats, production, ressources humaines, qualité, maintenance, recherche) sont impliqué.

Parfois on oublie que la matière première du SI est constituée des données devenues indispensables à la vie et au développement des organisations. Il est clair que les entreprises qui en font bon usage gagnent en compétitivité et que les données sont un facteur clé de succès. Des données fiables permettent notamment de prendre de bonnes décisions. [**Christophe Brasseur, 2016]**

Notre époque produit quantité de données. Toutes ces données, utiles pour la conduite de machines, notre vie sociale ou pour la surveillance d'ouvrages d'art, sentimentale, économique... laissent des traces. Ces traces maillent le monde et sont conservées de manière croissante. Le Big Data (ou méga données) y trouve des modèles pouvant améliorer les décisions ou opérations et transformer les (entreprises) firmes.

Cet ouvrage est une invitation à penser ce qu'une approche par les méga données modifie dans la recherche scientifique médicale, et comment peut être utilisé dans notre vie quotidienne.

[Pierre Delort, 2015]

1. BIG DATA:

1.1. DEFINITION DU BIG DATA:

Souvent considéré comme un déluge d'informations produites à chaque seconde dans le monde, la définition du Big Data (ou métadonnée) est encore vague pour les acteurs de la société actuelle.

La définition donnée par le **[MERCATOR ,2014]** est la suivante : **« Flux de données considérable collectés par le suivi automatique des comportements en ligne : visites de sites, achats de produits, clics sur les bannières et les liens, etc. »**. Réellement, le Big Data désigne non seulement le volume considérable de données produites pour chaque seconde, mais également son traitement, sa collecte et son stockage. En plus, ces données ne sont pas uniquement issues de la navigation sur des sites Internet par les individus, mais également d'autres canaux et supports utilisés comme les applications sur Smartphones, les transactions bancaires, la géolocalisation etc. [**Tiffany ETCHEGORRY, 1016]**

Le concept de "Big data" se trouve l'explosion du volume de données informatiques, conséquence de l'augmentation de l'usage d'Internet, au travers des réseaux sociaux, des appareils mobiles, des objets connectés, etc.

Selon **le Centre d'Expertise des Progiciels**1, Les Big Data désignent des méthodes et des technologies (pas seulement des outils) pour des environnements évolutifs (augmentation du volume de données, augmentation du nombre d'utilisateurs, augmentation de la complexité des analyses, disponibilité rapide des données) pour l'intégration, le stockage et l'analyse des données multi-structurées (structurées, semi structurées et non structurées). **[Angeline KONE, 2016]**

1.2. DIMENTIONS DU BIG DATA:

Pour pouvoir s'en faire une idée un peu plus précise, Meta Group présenta en 20012 un rapport faisant état de 3 dimensions à travers lesquelles l'amoncellement et le traitement de données diverses pouvaient être représentés : le volume, la vitesse et la variété

1.2.1.Volume :

Cette dimension est celle qui présente la plus grande difficulté et le plus grand enjeu, pour les entreprises notamment, car elle pose la question du choix et du tri des données les plus pertinentes à récolter en vue de les exploiter. Face à la quantité colossale d'informations disponible, il est impératif pour les acteurs économiques de les sélectionner de manière judicieuse pour les aider dans leur prise de décision.

La numérisation des données a augmenté d'une manière significative : en 2000, seulement 20% des informations collectées étaient numérisées, les 80 autres pourcents étant archivés de manière analogique (documents papiers, photographies, cassettes audio ou VHS). Aujourd'hui, on estime que 98% de nos informations sont stockées de manière non tangible à travers des disques durs, clé USB, ou encore le Cloud en raison de la diminution considérable des coûts liés au stockage**. [Bruno TEBOUL + Jean-Marie BOUCHER,2013]**

1.2.2.Vélocité:

L'avantage majeur de la dématérialisation des informations réside dans le fait qu'elles peuvent transiter de manière quasi instantanée en s'affranchissant de toute contrainte d'espace et de temps. Lors de l'apparition d'Internet et de sa diffusion au grand public à travers des modems, certes lents comparés à ceux disponibles de nos jours, mais d'une rapidité incroyable pour l'époque,

1le Centre d'Expertise des Progiciels : *identifie de manière exhaustive les fournisseurs de progiciels et évalue en détail leurs fonctionnalités, les caractéristiques de leurs éditeurs et des réseaux de distribution et les intégrateurs*.

Disponible sur http://www.techinfrance.fr/partenaire/le-cxp.

2Meta Group. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. [en ligne]. Disponible sur https://lc.cx/4m6X.

La récupération de nombreuses informations a alors été possible sans délais et sans supports physiques autres qu'un ordinateur. L'apparition de l'ADSL, du haut-débit puis de la fibre n'ont cessé de faire diminuer le temps nécessaire au transit des données, améliorant et accélérant grandement la connaissance des individus. Dans une société évoluant à grande vitesse, cet aspect technologique constitue un atout majeur pour les structures privées comme publiques, désormais capables de prendre des décisions en temps réel.

1.2.3. Variété :

La numérisation des informations engendrée par l'essor des nouvelles technologies (appareils photo numériques, GPS, Smartphones, bornes…) permet aujourd'hui de stocker de nombreux éléments de manière intangible quelle que soit la nature de celui-ci (texte, fichier audio, vidéo, emails…). Cette variété de supports représente un véritable souci pour les entreprises qui doivent analyser les données qu'ils produisent afin d'en retirer des éléments pertinents et parvenir à les croiser pour leur attribuer un sens. **[Bruno TEBOUL + Jean-Marie BOUCHER,2013]**

1.3. L'ANALYSE : LE POINT CLE DU BIG DATA :

Le Big Data répond à de nombreux objectifs précis parmi lesquels on trouve :

1- L'**extraction** d'informations utiles des données stockées.

2- L'**analyse** de ces données.

3- La **restitution** efficace des résultats d'analyse.

4- L'accroissement de **l'interactivité** entre utilisateurs et données.

L'analyse est le point clé de l'utilisation du Big Data. Elle permet de mieux connaître sa clientèle, d'optimiser son marketing, de détecter et prévenir des fraudes, d'analyser son image sur les réseaux sociaux et d'optimiser ses processus métiers.

1.4. PROCESSUS DE CHARGEMENT ET DE COLLECTE DE DONNEE DANS BIG DATA:

Les données traitées par le Big Data proviennent de :

- ✓ Du **Web** : journaux d'accès, réseaux sociaux, e-commerce, indexation, stockage de documents, de photos, de vidéos, linked data.
- ✓ Plus généralement, de l'**internet** et des **objets communicants** : réseaux de capteurs, journaux, des appels en téléphonie ;
- ✓ Des **sciences** : génomique, astronomie, climatologie (ex : le centre de recherche allemand sur le climat gère une base de données de 60 petaoctets).
- ✓ Des données **commerciales** (ex : historique des transactions dans une chaîne d'hypermarchés);

- ✓ Des données **personnelles** (ex : dossiers médicaux);
- ✓ Des données **publiques** (open data qui sont des donnée numérique, peut être publique ou privé dont l'accès sont laissés libres aux usagers).

Maintenant que nous avons vu globalement comment s'articule une architecture de ce type, nous allons voir le concept SGBD liés aux différents types de données que nous voulons les présenter par la suite.

1.5. No SQL :

On ne peut pas parler de Big Data sans citer le NoSQL, Not Only SQL. Il est venu pour solutionner les difficultés rencontrées pendant la gestion des données classées « Big Data » avec les systèmes SGBD relationnelles**. [Matteo Di Maglie, 2012]**

1.5. 1. HIRISTIQUE DU MOUVEMENT NoSQL:

En 1998, le monde entend pour la première fois le terme NoSQL. Terme inventé et employé par Carlo Strozzi pour nommer son SGBD relationnel Open Source léger qui n'utilisait pas le langage SQL. Ironiquement, le trouvaille de M. Strozzi n'a rien à voir avec la mouvance NoSQL que l'on connaît aujourd'hui, vu que son SGBD est de type relationnel. En effet, c'est en 2009, lors d'un rassemblement de la communauté des développeurs des SGBD non-relationnels, que le terme NoSQL a été mis au goût du jour pour englober tous les SGBD de type non-relationnel.

1.5. 2. DEFINITION:

Le NoSQL est un type de base de données, c'est une manière de stocker et de récupérer des données de façon rapide, un peu comme une base de données relationnelle, sauf qu'il n'est pas basé sur des relations mathématiques entre les tables comme dans une base de données relationnelle traditionnelle.

1.5. 3.FONCTIONNEMENT NoSQL :

Scalabilité est l'aptitude d'un système à conserver, maintenir son niveau de performance par augmentation des ressources matérielles On distingue deux types de scalabilité :

- ✓ Verticale ou interne : ajout de RAM, processeur au sein d'une machine ou remplacement par une de plus grand gabarit.
- ✓ Horizontale ou externe : Le principe consiste à simplement rajouter des serveurs identiques en parallèle afin de répondre à l'augmentation de la charge.

Le but des systèmes NoSQL est de renforcer la scalabilité horizontale. **[Khaled Tannir, 2015]**

1.5. 4. TYPE DE BASE DE DONNEE NoSQL :

Il en existe 4 types distincts qui s'utilisent différemment et qui se prêtent mieux selon le type de données que l'on souhaite y stocker :

a- LES BASES DE DONNEES CLE-VALEUR:

La base de données de type clé-valeur est considérée comme la plus élémentaire. Son principe est très simple, chaque valeur stockée est associée à une clé unique. C'est uniquement par cette clé qu'il sera possible d'exécuter des requêtes sur la valeur.

Les caractéristiques principales du système associatif:

- ✓ Aucun SQL
- ✓ Meilleure scalability horizontale du marché.
- ✓ Peut ne pas fournir un support des propriétés ACID.
- ✓ Peut offrir une architecture distribuée et tolérante aux pannes.

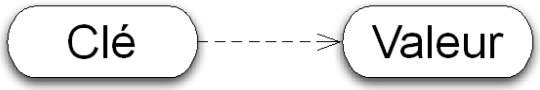

Figure 1.1 : Le principe de base de données Clé-Valeur [1]

b- LES BASES DE DONNEES ORIENTEES DOCUMENT:

La représentation en document est particulièrement adaptée au monde du Web. Il s'agit d'une extension du concept de clé-valeur qui représente la valeur sous la forme d'un document contient des données organisées de manière hiérarchique à l'image de ce que permettent XML ou JSON. Étant consciente du contenu qu'elle stocke, la base de données peut alors effectuer des indexations de différents champs et offrir des requêtes plus élaborées. **[Khaled Tannir, 2015]**

Les principaux avantages de ce type de système sont donc :

- ✓ Il est plus performant d'extraire des données pour une densité importante d'informations.
- ✓ Améliore grandement les performances sur les tris ou agrégations de données car ces opérations sont réalisées via des clés de lignes déjà triées. **[Khaled Tannir, 2015]**

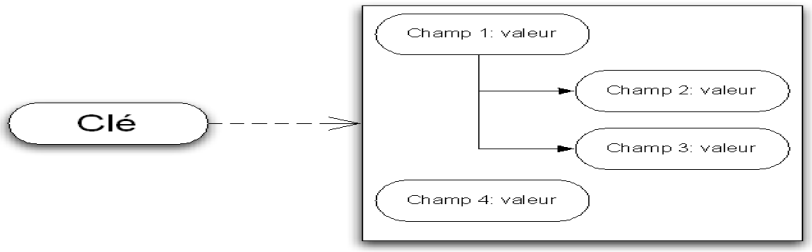

Figure 1.2. : Le principe de bases de donnée de type orientées document [1]

c- LES BASES DE DONNEES ORIENTEES COLONNE:

Le concept de colonnes est le plus simple à saisir, car l'analogie avec les bases relationnelles est proche. Dans les concepts à appréhender il existe des tables, ce qui permet de bien comprendre comment les données sont organisées. **[Khaled Tannir, 2015]**

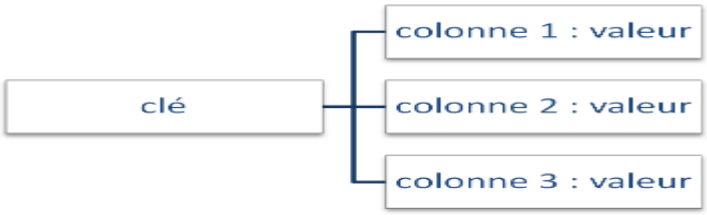

Figure 1.3. : Le principe de base de données- orientées colonnes [1]

d- LES BASES DE DONNEES ORIENTEES GRAPHE :

Les bases orientée graphe sont les moins connues de la mouvance NOSQL. Ces bases permettent la modélisation, le stockage ainsi que le traitement de données complexes reliées par des relations **[Khaled Tannir, 2015]**

Ce modèle est composé d'un :

- Moteur de stockage pour les objets : c'est une base ou chaque entité est nommée nœud
- Mécanisme qui décrit les arcs : c'est les relations entre les objets, elles contiennent des propriétés de type simple (integer, string, date, ...). **[Khaled Tannir, 2015]**

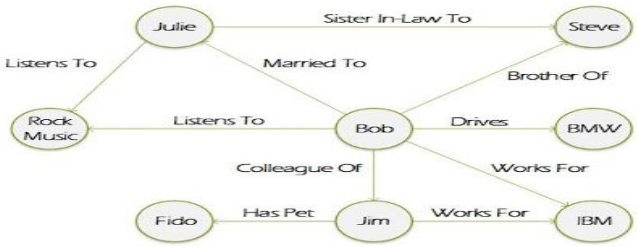

Figure 1.4. : Le principe de base de donnée orientées graphe [2]

1.6. LES BASES DE DONNEES ORIENTEES GRAPHE:

Il faut présenter les disciplines liées au Big Data que sont l'économie et le Droit.

1.6.1. APPROCHE ECONOMIQUE:

Le Big Data c'est le pétrole du 21ème siècle, la collecte et la revente des données sa conte dans l'économie mondiale et selon le Boston Consulting Group, représenter près de 1 000 milliards d'euros d'ici 2024, offrant une multitude d'opportunités à tous les professionnels

Grâce au développement des objets connectés. A peut prés 3.5 zéttaoctets ont été collecté en 2014 cette masse sera augmenté à l'horizon 2020**. [3]**

1. 6.2. APPROCHE ETHIQUE ET JURIDIQUE:

En 2012, les révélations d'Edward Snowden à propos de la NSA ont profondément indigné les citoyens du monde entier, en particulier ceux résidant aux États-Unis**. [SZADKOWSKI Michaël et al, 2013].**

Cet ancien informaticien de la CIA, ayant collaboré avec l'organisme de sécurité du pays, avait confié à la presse américaine et britannique les pratiques des deux entités en matière d'espionnage des organismes politiques mondiaux et des particuliers par le biais de PRISM, un logiciel permettant d'intercepter les flux de conversations et d'actions se déroulant la Toile.

Ce scandale a eu pour effet d'alerter les individus de l'utilisation faite des traces numériques produites et soulever de nombreuses questions en matière de droit à la vie privée.

Car si la liberté et la vie privée relèvent des droits fondamentaux des citoyens, il n'en demeure pas moins que chaque mouvement, chaque clic, chaque conversation est enregistré, dupliqué et stocké de manière quasi définitive.

Pour Céline Castets-Renard **,** professeure de Droit privé et membre d'Institut de recherche en droit européen, international et comparé, le problème majeur relève du fait que la technologie progresse bien plus vite que le Droit. Les définitions et l'encadrement de la législation concernant l'exploitation de données sont basés sur des textes obsolètes qui peinent à évoluer en même temps que la société.

Le texte de référence, la Loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, définit l'usage de l'informatique comme de suit **[legifrance, 2013] :**
« L'informatique doit être au service de chaque citoyen. Son développement doit s'opérer dans le cadre de la coopération internationale. Elle ne doit porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques. » (Article 1)
Ce texte, bien que modifié et enrichi en août 2004 et janvier 2016, ne concerne que la France.

L'Organe national de prévention et de lutte contre les infractions liées aux TIC algérien, dont le décret de mise en place a été édicté par la loi 09-04 du 5 août 2009 qui régit ces infractions, Cette loi définit, les infractions liées aux TIC, celles "portant atteinte aux systèmes de traitement automatisés de données telles que définis par le code pénal ainsi que toute autre infraction commise ou dont la commission est facilitée par un système informatique ou un système de communication électronique".

Le Big Data est une solution de comment gérer les grandes mases de données qui concerne différentes domaines, donc en va baser sur le domaine médical qui est un domaine vaste et intéressant dans notre vie quotidienne.

2. L'INFORMATIQUE MEDICALE :

2.1. DEFINITION:

L'informatique médicale concerne l'application du matériel et des logiciels informatiques ou de techniques provenant de l'informatique à la médecine. Le développement de nouvelles technologies médicales de plus en plus élaborées et douées d'une plus grande complexité nécessite souvent d'enregistrer et de traiter des données ou informations complexes sur un support informatique.

En ce sens, l'informatique médicale tend à assurer la compatibilité des systèmes et formats de données entre les sites et les outils médicaux, la transcriptibilité parfaite des données, le respect des spécifications quant aux paramètres d'enregistrement des informations, la traçabilité des données et l'anonymat des informations. Par rapport à ces contraintes, le rôle de l'informatique médicale est de rendre fonctionnel les logiciels et matériels à même de servir dans le monde médical.

L'informatique prend une place de plus en plus importante du sein du monde médical. On considère généralement que l'application de l'informatique au domaine de la santé est restreinte ou cloisonnée à un ensemble de techniques et d'outils mais il ne faut pas oublier que l'apport de l'informatique permet aux médecins de simplifier de nombreuses tâches (comptabilité, gestion des données…) et d'accéder facilement au dossier médical de leurs patients, en cabinet ou en centre hospitalier.

En conséquence, l'informatique médicale est aussi considérée comme une discipline scientifique qui contribue à la compréhension des mécanismes d'interprétation et de raisonnement médical, d'abstraction, de mémorisation, d'apprentissage et de l'élaboration des connaissances. Au centre du sujet, la notion d'information médicale donne tout son sens à l'informatique appliquée au monde de la médecine. Elle aide à recueillir les faits, à les mémoriser tout comme elle pourra servir à les interpréter.

L'informatique médicale permet d'une part de pallier aux limitations de l'être humain en termes de capacité de stockage de l'information et de son traitement mais d'autre part, de mettre en place des réseaux de communication, d'assurer un rapprochement entre la médecine et le patient (télémédecine) et de faciliter l'accès aux connaissances nécessaires à une prise en charge optimisée du malade (accès à des banques de données, utilisation de systèmes experts…). **[Agenda-medical, 2015]**

2.2. UNE SCIENCE DE L'INFORMATIQUE:

Comment mieux stocker et réutiliser l'information biomédicale, épidémiologique et médicale ? Comment modéliser les savoirs et savoir-faire médicaux ? Peut-on imaginer un système de traitement de l'information capable de gérer toutes les spécialités médicales (génétique, médecine clinique, imagerie, biologie moléculaire, biologie, etc.)? Comment partager l'information médicale tout en protégeant le secret médical ? Quel est le processus qui permet de passer du signe au diagnostic puis à la décision médicale ? Peut-on définir une bioéthique du traitement de l'information médicale ? Existe-t-il une connaissance de référence toujours valide ? Comment

intégrer des mécanismes biologiques et leurs interactions alors que nous n'en connaissons pas toute la nature ? Comment faire évoluer les connaissances des systèmes d'information ? Voilà quelques questions auxquelles les sciences de l'information médicale tentent de répondre afin de proposer et de mettre en œuvre des solutions informatiques innovantes qui répondent aux exigences des personnels de la santé en termes d'utilisabilité, aux populations en termes de soin et de service rendu, et qui permettent aux chercheurs de découvrir de nouvelles connaissances. **[Le Nouvelliste, 2016]**

2.3 PARTAGE DE L'INFORMATION MEDICALE:

Nombreux sont les patients et professionnels de santé qui utilisent les moteurs de recherche pour accéder à une information médicale sur internet. Or, trop de fausses informations médicales circulent sur le web. Ainsi, selon une récente enquête menée par l'institut IPSOS, alors que 70% des français ont déjà recherché des informations relatives à la santé sur le web, 40% d'entre eux restent insatisfaits de la qualité de l'information obtenue.

Partant du principe que les patients ont droit à une information médicale de qualité, validée et utile, nous avons développé doctideo.com, la première plateforme de partage de connaissances médicales créée par des médecins.

Qui mieux qu'un professionnel de santé pour informer les patients et le grand public sur les questions de santé ?

Les professionnels de santé aussi n'ont pas toujours le temps de faire le tour du web pour trouver la bonne information médicale, la nouvelle recommandation, ou le nouveau traitement. Ils peuvent être également victimes du succès des forums et du manque de fiabilité des informations qui y sont partagées**. [CHEIKH NGOM, 2016]**

2.4. TERMINOLOGIE ET LES CLASSIFICATION :

"La terminologie, considérée comme une science, s'intéresse au recensement des concepts d'un domaine et des termes qui le désignent pour faciliter l'échange de connaissances dans une langue et d'une langue à l'autre." Une terminologie implique la normalisation des termes d'un domaine afin de pouvoir les organiser les uns par rapport aux autres. Le principal intérêt d'une terminologie est de réduire l'ambiguïté qui existe entre les termes d'un domaine et de pouvoir donc, mieux partager de l'information.

Une classification se définit comme l'action de distribuer par classes et par catégories les concepts d'un domaine. C'est une répartition systématique par classes ou catégories ayant des caractères communs dans un contexte précis. La structure et la profondeur de la classification dépend de l'objectif du concepteur. **[Jollois.Francois_Xavier, 2003]**

3. SYSTEME D'AIDE A LA DECISION MEDICALE :

3.1. NOTION DE DECISION MEDICALE :

En médecine, la décision est considérée comme étant le centre de l'acte médical. Le processus de la décision médicale consiste entre autres à poser un diagnostic, proposer un traitement ou le différer, etc. Ainsi, de très nombreuses applications d'aide à la décision ont été développées dans ce domaine. Ces applications sont destinées à soutenir le personnel de santé dans leurs prises de décisions. Cela implique l'utilisation de divers outils d'aide à la décision. **[A BOUZIDI, 2015].**

3.2. SYSTEME D'AIDE À LA DECISION MEDICALE :

Les systèmes d'aide à la décision médicale (SADM) sont définis de manière très générale comme des outils informatiques « dont le but est de fournir aux cliniciens en temps et lieux utiles les informations décrivant la situation clinique d'un patient ainsi que les connaissances appropriées à cette situation, correctement filtrées et présentées afin d'améliorer la qualité des soins et la santé des patients. » (Berner, 2009). Il existe ainsi des SADM pour l'ensemble des activités médicales (prévention, dépistage, diagnostic, traitement) et la majorité des spécialités médicales (maladies chroniques ou affections aiguës). Ces systèmes proposent des services pour les différentes catégories de médecins (généralistes, spécialistes, étudiants) et les différents modes d'exercice (cabinets médicaux, hôpitaux, services d'urgence ou de réanimation). Plus récemment, des SADM ont été développés à destination des patients afin qu'ils soient mieux informés sur leur maladie et les soins qui pourraient leur être proposés dans un objectif de décision partagée**. [Brigitte Séroussi et al, 2015]**

Jusqu'au début des années 90, leur utilisation s'est limitée à quelques Institutions pionnières US1 qui ont développé, régulièrement évalué et progressivement amélioré leurs propres systèmes d'information clinique comportant des fonctions de gestion du dossier patient, de prescription des actes diagnostiques et des médicaments, et d'aide à la décision clinique.

La démonstration par ces Institutions du potentiel des technologies de l'information et de la communication en santé (TICS) et des SADM en termes d'amélioration de la qualité, de la sécurité et de l'efficience des soins jointe à la publication dans les années 2000 de constats sur la qualité insuffisante des soins par l'US Institute Of Médicine, ont conduit à partir des années 90 le gouvernement fédéral des Etats Unis, les systèmes d'assurance publics ou privés et les organisations de soins intégrées à des initiatives ayant pour objectifs :

- D'améliorer la qualité des soins au moyen de systèmes d'information cliniques comportant un dossier patient associé à des fonctions de prescription informatisés et d'aide à la décision
- De favoriser le développement d'une offre commerciale comportant des services d'aide à la décision par les fournisseurs de systèmes d'information clinique. **[Jean-louis RENUD-SALIS et al, 2010]**

3.3. LA TYPOLOGIE DES SYSTEMES D'AIDE À LA DECISION MEDICALE :

Il existe trois grandes catégories de systèmes d'aide à la décision médicale.

3.3.1. LES SYSTÈMES D'AIDE À LA PRISE DE DÉCISIONS OU D'ASSISTANCE DOCUMENTAIRE:

L'objectif des systèmes d'assistance documentaire est de faciliter l'accès aux informations pertinentes en un temps record mais ces systèmes n'ont pas de méthode de raisonnement à proprement parler**.**

L'accès aux résultats de laboratoire, la consultation des éléments importants du dossier médical du patient, les références bibliographiques (par exemple Medline) et les systèmes de bases de données concernant les médicaments constituent des aides indirectes à la décision. Cette aide intervient pour faciliter l'appréciation d'une situation par le médecin. **[A BOUZIDI, 2015]**

3.3.2. LES SYSTÈMES D'ALERTE OU DE RAPPELS AUTOMATIQUES :

Certains systèmes permettent de rappeler au médecin des erreurs à ne pas commettre ou des éléments importants à prendre en compte pour la décision. Ils sont plus actifs et plus directement impliqués dans la décision médicale. L'assistance fournie n'est pas une aide au raisonnement ou à l'appréhension globale du cas du patient, mais plutôt un aide-mémoire fournissant une information utile et pertinente dans une situation facile à définir a priori.

Ces systèmes, comme les précédents, ne raisonnent pas véritablement. Par exemple, le rappel des valeurs normales des résultats d'examens de biologie et l'utilisation d'une typographie permettant d'attirer l'attention sur les valeurs anormales ou encore l'émission d'un message de mise en garde devant une association de médicaments déconseillée constituent une aide simple dont l'utilité est primordiale. **[A BOUZIDI, 2015].**

3.3.3. LES SYSTÈMES CONSULTANTS:

Face à une situation médicale bien définie telle que : un diagnostic, une thérapie ou un pronostic, les systèmes consultants tentent d'émettre un avis de spécialiste.

Ces systèmes fournissent à l'utilisateur des conclusions argumentées selon les méthodes de raisonnement employées. La conception est intellectuellement plus satisfaisante que celles des systèmes n'utilisant pas de véritables processus de raisonnement. Donc les développeurs s'intéressent principalement à ce type de système où l'on note le plus de réalisations en matière de système d'aide à la décision. **[A BOUZIDI, 2015]**

CONCLUSION :

Ces dernières années, un énorme buzz autour du Big Data s'est développé à tel point qu'un certain nombre d'analystes n'y ont vu qu'un phénomène marketing de plus, visant à favoriser les ventes des fournisseurs de technologie. Les protagonistes des systèmes d'information, à commencer par les entreprises utilisatrices, s'aperçoivent à présent que le phénomène est bien réel. Et les enjeux sont de taille, car le Big Data n'est pas qu'une question technique de volumétrie et de stockage. Il constitue au contraire l'opportunité de comprendre le contenu de ces nouvelles sources et d'entirer profit. **[Christophe Brasseur, 2016]**

Le Big Data ne constitue sans doute pas une révolution de l'ampleur de celles de l'agriculture ou de l'industrie, et à même d'engendrer un nouvel âge d'or. Cependant, en ouvrant à l'induction des données, il permet des transformations de nombreux champs d'activité : politique, social, éducatif, judiciaire, sportif, médicale avec des aspects juridiques, éthiques ou encore psychologiques à ne pas oublier. **[Pierre Delort, 2015]**

L'informatique décisionnelle en médecine ce sont des outils informatiques capables de traiter l'ensemble des caractéristiques d'un patient donné afin de générer les diagnostics probables de son état clinique (aide au diagnostic) et pronostique. **[Brigitte Séroussi, et al, 2015]**

INTRODUCTION :

Dans ce chapitre nous allons présenter quelques définitions et historique de bon pratique sur la médecine basée sur les preuves ou médecine factuelle (EBM), ainsi les techniques utilisé par big data pour extraire et traité les données médicales et les passé par une fouille de donnée qui est basée sur des algorithmes d'apprentissage automatique ou en va citer quelques une .Mais avant tous en va parlé sur quelques problèmes qui sont les cause de l'apparition de l'EBM lié au BIG DATA .

1. PROBLEMATIQUE :

La notion de qualité est présente dans le monde entier que ce soit pour le domaine de l'industrie du commerce ou médicale . mais la gestion et les techniques utilisé pour l'obtention de bon qualité c'est le facteur de réussite car par exemple mettre un service de qualité aux clients est la clé de la satisfaction et par la suite on peut assurer qu' un grand nombre des clients vont utilisé ce service. aussi , donner un bon diagnostique ou traitement a un patient le résultat le patient sera plus attaché a cette source de guérison la plus sure et aussi on peut garantir la fidélité des utilisateurs .

C'est pour cela plusieurs payés , institues et labos ont investie sur ce facteur dans tous les domaines et précisément le domaine Médicale .

D'après les statistiques des dépenses des payés dans le domaine de santé , Les États-Unis dépensent plus de 16 % de ses dépenses globales equivalent a 2,3 billions de dollars dans les soins de santé .

En 2014, chaque Français dépense en moyenne 1 346 € en soins d'hospitalisation, 759 € en soins de ville (soins de médecins, d'auxiliaires…) et 515 € en médicaments, soit au total 2 621 €.

Ces dépenses ont Pas amélioré la qualité de soins. Environ 98 000 Des personnes meurent chaque année à l'hôpital en raison de l'erreur médicale ou problème des mauvaise diagnostiques.

Hier , La complexité croissante des connaissances médicales, telles que les moyens diagnostiques et thérapeutiques, obligent le médecin à gérer toujours plus d'informations pour soigner un patient , Mais cette spécialisation peut entraîner des pertes de temps , le médecin doit prendre toute les signes et des symptômes, en écoutant le patient pour effectuer une série de décisions. Il tente ensuite de prévoir l'évolution de la maladie.

Les symptômes et les maladies se sont spécialisés, de nombreuses nouvelles virus sont arrivées.

Ces médecins, ne peuvent plus maîtriser l'ensemble du savoir médical qui permet de reconnaître les maladies ou de déterminer la meilleure prise en charge thérapeutique. Aussi, ils ont souvent recours à des sources d'information externes .

Mais le probléme c'est que environ 3000 nouveaux articles sont indexés chaque jour sur les moteurs de recherches. Un Médecin , s'il réussit à lire un article par jour, ne lira qu'un millième de ce qui est publié.

Normalement , Les bases de données basés sur le standard SQL ont de bonnes performances lors du traitement de petites quantités de données relationnelles **mais** ces résultats sont très limités sur des données volumineuses et complexes et non –structurés !

Aujourd'hui, Le développement de systèmes d'aide à la décision, simulant le raisonnement du médecin, Pour cela, il importe de retracer la démarche du médecin face aux symptômes d'un malade et de faire une analyse de la décision médicale .

Mais comment ce faire ? quelles sont ces mécanismes ? les algorithmes utilisées ?

Les patients veulent de plus en plus être inclus dans les soins médicaux Processus décisionnel en cas de maladie. Un patient veut a toute moment savoir des nouvelles sur sa santé .puisque il est vivant il est dans un état de santé instable.
Il veut a chaque changement de santé savoir son diagnostique a moindre cout , a temps réduit
Qu'elle meilleur traitement a prendre ? Traitement
Qu'elle meilleur traitement qu'il lui semble le meilleur ?
es-ce –qu'il va être guérie ou son état va être amélioré ? Pronostique

Les recherches bibliographique que nous avons réalisé , nous a permis de constater que les travaux de recherches sur l'intégration de Big Data et EBM sont rares et ceux qui existent s'utilisent uniquement aux données structurées et ne tiennent pas compte de leur nettoyage .

I. MEDECINE BASEE SUR LES PREUVES (EVIDENCE BASED MEDCINE)

1. 1. INTRODUCTION:

EBM se définit comme l'utilisation consciencieuse et judicieuse des meilleures données actuelles de la recherche clinique, dans la prise en charge personnalisée de chaque patient. Née dans les années 1980 à la faculté de médecine McMaster (en Ontario, Canada), et utilisée comme méthode pédagogique, l'EBM gagne en notoriété dans les années 1990 face à la remise en question des données traditionnelles de la médecine, le but étant la promotion des données reconnues comme étant de qualité et fiables.

L'EBM tente de diminuer l'effet de la variabilité individuelle par la pratique de grands essais cliniques, et apporte des réponses qui seront applicables à un patient moyen. La notion de recommandation « individualisée » semble donc paradoxale.

Ce chapitre a pour but de comprendre comment les médecins perçoivent l'évolution actuelle du concept des recommandations. Pour ceci, il nous a paru indispensable d'étudier comment ces praticiens prennent une décision médicale et comment ils intègrent l'EBM dans leur pratique. **(Anonymous, 1992)**

1.2. EBM : UN NOUVEAU MODELE POUR LA MEDECINE :

EBM (Médecine basé sur les preuves en englais (Evidence based medecine). C'est une nouvelle façon de faire de la médecine, de poser des questions et résoudre des problèmes cliniques, c'est une nouvel manière d'apprendre .Ils évoquent également un long débat, où les idées se confronte en taux opinions, les argumentations défensives aux invectives de contre- attaque et les belles déclarations à la réalité de la

pratique et des comportements humains. Un débat ouvert depuis 10 ans dans les conférences et à travers des publications. **[DL Sackett - 1996]**

La "Médecine Factuelle" ou EBM qui désignait, au départ, une stratégie d'apprentissage des connaissances cliniques, fait maintenant partie intégrante de la pratique médicale. Elle consiste à baser les décisions cliniques, non seulement sur les connaissances théoriques, le jugement et l'expérience qui sont les principales composantes de la médecine traditionnelle, mais également sur des "preuves" scientifiques, tout en tenant compte des préférences des patients .Par "preuves", on entend les connaissances qui sont déduites de recherches cliniques systématiques, réalisées principalement dans le domaine du pronostic, du diagnostic et du traitement des maladies et qui se basent sur des résultats valides et applicables dans la pratique médicale courante .Les études cliniques considérées sont des essais contrôlés randomisés, des méta-analyses, mais aussi des études transversales ou de suivi bien construites lorsqu'il s'agit d'évaluer un test diagnostique ou de pronostiquer l'évolution d'une maladie **(Anonymous, 1992)**

Le concept d'EBM a été développé par des épidémiologistes canadiens de la McMaster Medical School au début des années 1980 puis adopté par la *Cochrane Collaboration* qui défend, depuis son origine, les essais contrôlés randomisés ainsi qu'une méthodologie rigoureuse en recherche clinique

L'EBM est une démarche qui nécessite les étapes suivantes :

1. **Formuler** clairement le problème clinique à résoudre pour chaque malade considéré.
2. **Rechercher** les articles pertinents dans lal ittérature scientifique en excluant les articles critiquables d'un point de vue méthodologique.
3. **Evaluer** la validité et l'utilité des conclusions des articles sur le plan pratique.
4. **Intégrer** ces preuves dans la pratique médicale courante afin de répondre à la question posée au départ

A chacune des conclusions des articles retenus, est attaché un **niveau de preuve** qui dépend de la méthodologie utilisée et de sa qualité :

1. **Niveau1** : essais contrôlés randomisés avec des résultats indiscutables d'un point de vue méthodologique.
2. **Niveau 2** : essais contrôlés non randomisés bien conduits.
3. **Niveau 3** : essais prospectifs non contrôlés bien menés (étude de cohortes par exemple)
4. **Niveau 4** : études de cas-témoins et essais contrôlés présentant des biais.
5. **Niveau 5** : études rétrospectives, cas cliniques et toute étude fortement biaisée**.**

Il faut noter que la pratique de l'EBM a soulevé dans la communauté scientifique un certain nombre d'objections qui peuvent être résumées comme suit **[ALAN R.Feinstein et al]**

- Manque d'études et de données scientifiques pour un certain nombre d'actes cliniques qui ne seront jamais évalués en utilisant l'approche EBM ou études non représentatives de malades auxquelles elles prétendent s'appliquer.

- Problèmes à résoudre en médecine de "premier contact" (notamment en médecine générale) où les problèmes sont le plus souvent liés à plusieurs pathologies, où se mêlent des dimensions sociales, culturelles, familiales, sanitaires.
- Informations valides et exactes d'aujourd'hui seront-elles utilisables demain? **[Davidoff F,1999]**

1.4. OBJECTIFS DE L'EBM

L'objectif principal est d'aider les patients à se soigner en leur donnant accès aux meilleures données actuelles de la science. Les objectifs secondaires sont de souligner que la durée de vie des données actuelles est parfois courte et qu'une mise à jour régulière des connaissances est donc indispensable, de donner aux médecins plus d'assurance dans leurs décisions, de compléter les connaissances (car le raisonnement biologique et l'expérience sont une base insuffisante pour prendre des décisions),d'harmoniser les pratiques des différents médecins et d'améliorer le sens critique de ces derniers lors de la lecture d'articles scientifiques.

1.5. FREINS DE L'UTILISATION DE L'EBM EN PRATIQUE

les nombreux obstacles sont a apparu . En premier lieu, les preuves en elles-mêmes, qui peuvent manquer et qui représentent une limite non-négligeable. En effet, dans certains cas, aucune étude sur le sujet n'a été réalisée et le recours à l'EBM s'avère impossible. D'autre part, les résultats des études étant parfois difficiles à comprendre et à interpréter, les médecins ne se sentent pas assez qualifiés et formés, notamment pour identifier les articles importants. Les détracteurs de l'EBM évoquent aussi le manque de temps. La lecture de la littérature médicale doit se faire en dehors du temps de travail. Les recommandations proposées sont parfois trop nombreuses et trop longues, elles changent aussi régulièrement et parfois pour des raisons procédurales. Les médecins redoutent qu'il existe des liens trop importants entre les investigateurs et les laboratoires qui financent ces études. Ils pensent également perdre leur liberté de prescripteur. Après tout, la médecine est toujours vécue comme étant un art pour certains19-20-21. Les critères d'inclusion et surtout d'exclusion permettent d'étudier un « patient moyen », ce qui limite la généralisation des résultats (on peut citer les personnes âgées souvent exclues des études, la liste de symptômes non-exhaustive, les comorbidités rarement prises en compte, la tolérance au traitement qui n'est pas toujours étudiée). Sur le plan éthique, les préférences des patients et leurs valeurs ne sont pas intégrées aux essais cliniques.

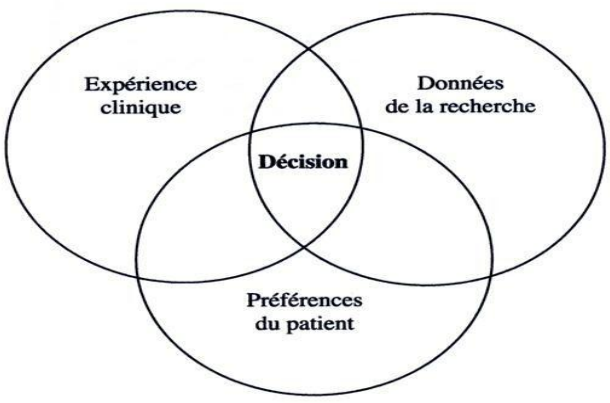

Figure 2.1 Éléments clés dans la décision médicale de type EBM

1.6. DEUX COMPOANTES DE L'EBM

Il existe deux types d'EBM : l'EBM épistémique (ou « Regulatory EBM ») et l'EBM pratique. Cette distinction est souvent méconnue ou oubliée des praticiens, bien qu'il soit important de la comprendre.

L'EBM épistémique ou « Regulatory EBM » a pour fonction de produire des connaissances. Il s'agit de hiérarchiser, évaluer et classer les connaissances répondant à la problématique qui justifie toute l'EBM : la variabilité des phénomènes vivants, à la foisen ce qui concerne les tests diagnostiques et les traitements. Ce que fait ce type d'EBM, c'est produire des connaissances concernant la sensibilité et la spécificité des tests et l'efficacité des traitements (par exemple, nombre de patients à traiter pour éviter un événement indésirable).

Ensuite, l'EBM pratique, qui correspond à la création de recommandations de bonne pratique ou « guidelines ». Il s'agit de décrire la pratique de la médecine optimale. On comprend bien que dans cette deuxième composante de l'EBM, le médecin doit continuer à exercer son expertise, lui permettant de juger si les recommandations peuvent être appliquées. [ReachG, 2012]

1.7. PRATIQUE DE L'EBM PAR LA PRETIQUE

1.7.1. FORMULER LA QUESTION CLINIQUE

La première étape de la pratique de la "Médecine Factuelle" est la formulation claire et précise de la question clinique qui doit être en relation directe avec le problème médical posé .

Problème médical	**Question**
Diagnostic	Comment sélectionner et interpréter un test diagnostique?
Etiologie	Comment identifier les causes d'une maladie?
Traitement	Comment choisir le meilleur traitement pour le patient?
Pronostic	Comment anticiper l'évolution et les complications probables d'une maladie?
Education des patients	Comment fournir aux patients et à leur famille les informations qui leur sont nécessaires?

Sur base d'un scénario clinique donné, il convient de décomposer la question en fonction des critères du tableau 1, également appelés "critères "PICO" (un exemple est présenté au tableau 2).

Critère P	Critère "P" pour "Patient characteristics or problem being addressed"	Caractéristiques du patient (âge, sexe,...) et/ou le problème qu'il pose (diagnostic,...)
Critère I	Critère "I" pour "Intervention(s) or exposure(s) being considered"	Intervention
Critère C	Critère "C" pour "Comparison intervention or exposure, when relevant"	Comparaison par rapport à une autre intervention (si elle est appropriée)
Critère O	Critère "O" pour "clinicalOutcome of interest"	Issue clinique recherchée

Tableau 1 : Critères PICO

En conclusion, le modèle "PICO" aide à diviser la question clinique en différents concepts qui serviront à construire la stratégie de recherche.

1.7.2. RECHERCHE D'ARTICLES PERTINENTS DANS LES BDs

La deuxième étape de la pratique de la "Médecine Factuelle" est la recherche d'articles pertinents dans la littérature.

1) BASES DE DONNEES BIBLIOGRAPHIQUES

L'accès à l'information est facilité depuis un certain nombre d'années par l'utilisation de bases de données bibliographiques qui permettent d'identifier rapidement les articles pertinents dans la littérature. Ces bases de données sont constituées de citations bibliographiques présentées sous la forme de différents champs:

1. Champ du ou des auteurs
2. Champ du titre
3. Champ source (= nom du journal)
4. Champ de description
5. Champ du résumé (facultatif)

Il est possible de réaliser une recherche dans chacun de ces champs. La recherche par sujet peut être réalisée en interrogeant soit le champ du titre et du résumé de l'article (recherche en langage naturel) soit le champ des descripteurs (recherche en vocabulaire contrôlé). Ce dernier champ contient un nombre limité de termes qui permettent de décrire l'article. Le processus d'attribution de descripteurs (également appelé indexation), qui définit la qualité d'une base de données, demande un certain délai et explique le retard que peuvent avoir certaines bases de données comme Medline et Embase. Les bases de données

sont ensuite stockées sur des serveurs (accès en ligne) ou déchargées sur des CD-ROMs qui doivent encore être livrés aux utilisateurs (délai supplémentaire pour ce type de support).

Les principales bases de données bibliographiques sont reprises ci-dessous:

a. Medline

Medline , qui correspond à l'Index Medicus et à une partie de l'International Nursing Index et de Index to Dental Literature, est produite depuis 1966 par la NLM. Elle indexe 4600 journaux internationaux dans la plupart des domaines de la médecine. Sa couverture est essentiellement anglo-saxonne. Elle repose sur l'utilisation du thésaurus MeSH. Chaque article est indexé manuellement en moyenne avec 15 descripteurs dans les jours ou les semaines qui suivent sa parution. Le MeSH est réactualisé tous les ans.

b. Embase

c. Current Contents (CC)

d. HEALTHSTAR / Internet Grateful Med / NLM Gateway

e. BiosisPreviews .

2)-BASES DE DONNEES ANALYTIQUES

Il existe également une série de bases de données qui sont accessibles via Internet et qui fournissent directement aux utilisateurs des données revues par des experts :

a. Cochrane Library

Qui contient les bases de données suivantes:

- **Cochrane Database of SystematicReviews (CDSR)** (accès payant) qui rassemble des revues systématiques en texte intégral qui sont produites et régulièrement remises à jour par les membres de la Cochrane Collaboration. La revue systématique implique l'utilisation d'une démarche scientifique rigoureuse pour rassembler.
- **Database of Abstracts of Reviews of Effectiveness (DARE)** est produite par le 'Service national de santé' de l'Université de York (UK). Elle contient les "résumés structurés" (résumé présenté dans un format standardisé, en utilisant un vocabulaire partiellement contrôlé) .
- **Cochrane Controlled Trials Register (CCTR)** est un registre contenant les références de toutes les études en cours sur un sujet précis .CCTR inclut également des rapports qui sont publiés dans les congrès et dans beaucoup d'autres sources non indexées dans Medline ou d'autres bases de données bibliographiques.
- **Cochrane MethodologyDatabase** contient les références de livres et articles parus dans les journaux sur les aspects méthodologiques des revues de la littérature et des méta-analyses.
- Cochrane MethodologyRegister: bibliographie d'articles et d'ouvrages sur la manière de synthétiser la recherche.
- **NHS EED [NHS Economic Evaluation Database]**: résumés correspondant à des évaluations économiques des différentes pratiques médicales estimées en terme d'analyses coût-bénéfice et coût-efficacité **[4]**
- HealthTechnologyAssessmentDatabase contient des informations sur des évaluations de technologies utilisées dans le cadre des soins de santé.

b. Cancer Library

- Cochrane Cancer Network **,** consumer support groupn CANCER BACUP et European Organisation for the Research and Treatment of Cancer (EORTC) **[5]**
- Nouvelle ressource électronique reprenant des informations basées sur les meilleures données actuelles de la science dans le domaine de la cancérologie. **[6]** (payant).

c. ACP Journal Club

- [American College of Physicians-American Society of Internal Medicine]. ***Accès payant***.
- Sélection d'études et de revues de médecine interne qui requièrent une attention immédiate des cliniciens qui veulent s'informer des développements récents de la médecine. Présentation de ces articles sous forme de **résumés à valeur ajoutée commentés par des experts cliniques.** Accessible depuis le site Ovid pour les membres de l'ULg uniquement (restriction d'accès).
- URL: http://www.acponline.org/journals/acpjc/jcmenu.htm

d. Evidence-Based Medicine

[BMJ Publishing Group et American College of Physicians-American Society of Internal Medicine]. ***Accès payant***.**Même vocation que l'ACP Journal Club** mais touchant un plus grand nombre de disciplines, comme la médecine générale, la chirurgie, la gynécologie, l'obstétrique, la pédiatrie, la psychiatrie. Présentation de ces articles sous forme de **résumés à valeur ajoutée commentés par des experts cliniques.** Les critères de sélection et d'analyse des articles sont repris sur le site d'EBM Journal. Accessible depuis le site Ovid pour les membres de l'ULg uniquement (restriction d'accès). **[7]**

e. EBM Journal

- Edition française du journal Evidence-BasedMedicine. ***Accès payant***.

- URL: http://www.ebm-journal.presse.fr/[Ovid Technologies, Inc.]. ***Accès payant***.

- Base de données reprenant les publications de la Cochrane Database of Systematic Reviews (CDSR), du Cochrane Controlled Trials Register (CCTR) (registre d'essais cliniques compilés par la Cochrane Collaboration), de l'ACP Journal Club, d'Evidence-Based Medicine et de la Database of Abstracts of Reviews of Effectiveness (DARE). Accessible depuis le site Ovid pour les membres de l'ULg uniquement (restriction d'accès). **[8]**.

f. CRD Databases

- 'National Health Service Centre for Reviews and Dissemination (CDR)' [Université de York (UK)]. ***Accès gratuit***. **[9]**
- **DARE [Database of Abstracts of Reviews of Effectiveness] :** Résumés structurés (résumé présenté dans un format standardisé, en utilisant un vocabulaire partiellement contrôlé) de revues systématiques obtenues de différentes sources et appréciées de manière critique par les spécialistes du NHS Centre for Reviews and Dissemination.
- Et aussi d'autres comme **NHS EED [NHS Economic Evaluation Database]** et **HTA [HealthTechnologyAssessmentDatabase]**

3)- BASES DE DONNEES PERSONNELLES

Il est important de pouvoir conserver les références bibliographiques intéressantes dans une base de données personnelles. Il existe différents logiciels de gestion de la bibliographie, par exemple "Reference Manager" et "EndNote" qui existent à la fois en version PC et MAC, qui restent faciles à utiliser et qui présentent l'avantage majeur d'être couplés à des programmes de traitement de texte.

1.7.3. DEFINIR LA QUALITE D'UNE RECHERCHE (SENSIBILITE ET SPECIFICITE) :

La qualité d'une recherche dépend du meilleur compromis entre deux critères: la sensibilité et la spécificité qui agissent de manière inverse.

- Une recherche "sensible" permet de retrouver la plupart des documents pertinents présents dans la base de données au risque d'inclure des documents moins pertinents (recherche large ou "broad search").
- Une recherche "spécifique" permet de retrouver les documents les plus pertinents de la base de données, au risque d'en rater certains (recherche étroite ou "narrow search").

1.7.4. DEVELOPPER UNE STRATEGIE DE RECHERCHE

Il s'agit de développer correctement la stratégie de recherche à partir de la question clinique posée. Il est conseillé de procéder comme suit :

a. Formuler correctement la question clinique

- Définir le type de question clinique:

  - ✓ Diagnostic?
  - ✓ Etiologie / causalité?
  - ✓ Intervention thérapeutique?
  - ✓ Pronostic?
  - ✓ Autre?

- Décomposer la question en fonction des critères PICO

  - ✓ Critère P (Patient or Problém) = concept 1
  - ✓ Critère I (Intervention) = concept 2
  - ✓ Critère C (Comparison intervention) = concept 3
  - ✓ Critère O (Outcomes) = concept 4

- Combiner les différents concepts au moyen des opérateurs booléens

b. Décider dans quelle base de données chercher:

- Chercher dans Medline, y compris Pre-Medline
- Chercher dans d'autres bases de données que Medline **[10]**

Pour améliorer la qualité de la stratégie de recherche, il est conseillé de procéder comme suit :

a. Elargir la question [11] : Plusieurs options sont possibles:

- Parcourir les citations pertinentes à la recherche d'autres termes à incorporer dans la recherche: des mots du langage naturel (avec troncature) à combiner éventuellement avec NEAR ou ADJ pour les retrouver dans une même phrase et dans n'importe quel ordre ou des descripteurs (vocabulaire contrôlé)
- Essayer différentes combinaisons de termes
- Ajouter et combiner des termes apparentés en utilisant or
- Utiliser une combinaison du langage naturel et des descripteurs en utilisant le or.
- Sélectionner tous les sous-descripteurs
- Remonter dans le temps pour la recherche

b. Préciser la question [11]

Plusieurs options sont possibles:

- Prendre des termes plus spécifiques dans la recherche en langage naturel ou des descripteurs plus spécifiques
- Utiliser des descripteurs plutôt que des termes du langage naturel
- Sélectionner des sous-descripteurs spécifiques
- Combiner avec l'opérateur booléen and tous les aspects de la question
- Appliquer des limites: langue, être humain, type de publication, pays ou année de publication

Il existe également des "filtres méthodologiques" développées par Anne McKibbon qui sont des stratégies de recherche qui permettent de retrouver des publications pertinentes de haute qualité dans le domaine du diagnostic, du traitement, de l'étiologie et de la causalité, du pronostic, des essais contrôlés randomisés et des revues systématiques. Sur base de ce principe, le site Pub Med propose des recherches automatisées soit plus sensibles soit plus spécifiques sur un sujet donné dans le domaine de l'étiologie, du pronostic, du diagnostic et du traitement ("clinical queries using research methodology filters"). Les guides présentés plus haut et développés dans le domaine du diagnostic, de l'étiologie / causalité, l'intervention thérapeutique, du pronostic et des revues systématiques ont également été préparés en suivant ces filtres.

2.7.5. EVALUATION CRITIQUE DE LA VALIDITE ET DE L'INTERET DES RESULTATS :

La troisième étape de la "Médecine Factuelle" consiste à évaluer de manière systématique la validité et l'intérêt des résultats et d'extraire les preuves qui sont à la base des décisions cliniques.

Malheureusement, une large proportion de recherches médicales publiées manque soit de pertinence soit d'une rigueur méthodologique suffisante pour être utilisées comme bases de décisions cliniques **[EJMullen, DL Streiner, 2006]**

De plus, il existe des sources de variabilité et d'erreur dans les essais cliniques. Elles sont de plusieurs types:

1)- ERREURS ALEATOIRE LIEE AUX FLUCTUATIONS DE L'ECHANTILLONNAGE:

- La répétition d'une même étude sur le même échantillon de patients et dans des conditions rigoureusement identiques n'arrive jamais exactement au même résultat par le simple fait du hasard.

2)- ERREURS SYSTEMATIQUES ? APPELEES BIAIS :

Ce type d'erreur dépend de la façon dont est organisée l'étude et d'interférences inhérentes au problème de santé étudié. On distingue:

- Les biais de sélection dus à la constitution des échantillons
- Les biais d'exécution liés à des différences dans la mise en oeuvre du protocole dans le groupe traité et dans le groupe contrôle
- Les biais dans le suivi de l'échantillon liés à des différences systématiques au niveau des pertes du groupe traité et du groupe témoin
- Les biais de mesure dus aux recueils des mesures et à des erreurs d'observation
- Les biais de confusion dus à la non prise en compte de variables liées en même temps à l'intervention et à la maladie

1.7.6 INTEFRATION DES RESULTATS DE L'EVALUATION DANS LA PRATIQUE CLINIQUE COURANTE :

La quatrième étape de la "Médecine Factuelle" est l'intégration des résultats de l'évaluation dans la pratique clinique courante.

Les résultats de l'évaluation ne sont applicables à un patient donné que pour autant qu'il ne soit pas différent des patients inclus dans l'étude. Il convient ensuite d'évaluer les risques et les bénéfices potentiels du traitement sur le patient en question **[Sackett DL,2000]**

La "Médecine Factuelle" est une approche qui vise à améliorer la prise en charge des patients. Elle ne se limite pas à rechercher des données prouvées scientifiquement mais elle fait également appel à l'expertise clinique individuelle qui déterminera la mise en application ou non des preuves dans la prise de décision clinique concernant un patient particulier.

D'après Rosenberg [Rosenberg et Donald, 1995] les réunions au sein des différentes équipes et les discussions entre cliniciens favorisent cette étape d'intégration des preuves dans la pratique médicale courante.

3. TRAITEMENT DES DONNEES MEDICALE EN BIG DATA PAR LA PRISE DE DECISION (DATA MINING)

La prise de decision est amélioré en utilisent les techniques associé a big data ,Selon Charles Huot **,** cinq familles de technologies sont clés pour le secteur des Big Data : text-mining, graph-mining, machine learning, data-vizualisation, ontologies. Toutes convergent vers un même objectif : simplifier l'analyse de vastes ensembles de données et permettre la découverte de nouvelles connaissances.

Dans ce chapitre on va parlé sur les techniques et les algorithmes les plus utilises en Data Mining .

3.1. DEFINITION DE LE FOUILLE DE DONNEE

Le « fouille de donnee" que l'on peut traduire par Le "Data mining" apparaît au milieu des années 1990 aux États-Unis, c'est de trouvé une réponse à la question : « Comment trouver un diamant dans un tas de charbon sans se salir les mains ? » **[12]**

On pourrait définir le DATA MINING aussi comme un Processus inductif , itératif et interactif de découverte dans les larges bases de données , de modèles de données valides , nouveaux , utiles et compréhensibles.

- ✓ Inductif : méthodes de raisonnement utilisé par les outils de Data Mining .
- ✓ Itératif : nécessite plusieurs passes.
- ✓ Interactif : l'utilisation est dans la boucle du processus.
- ✓ Valides : valable dans le futur.
- ✓ Nouveau : non prévisibles
- ✓ Utiles :permettent a l'utilisateur de prendre de décision .
- ✓ Compréhensibles : présentation simple.

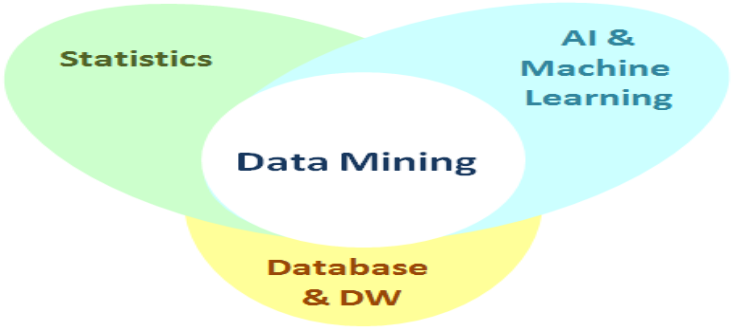

Figure n°2.2 :Data Mining **[13]**

3.2. COMPOSANTS DE DATA MAINING

3.2.1.STATISTIQUES :

a.DEFINITION ET HISTORIQUE :

La science de la collecte, de la classification, du résumé, de l'organisation, de l'analyse et de l'interprétation des données. pour cela Il faut fournir des bases statistiques pour comprendre les résultats obtenus par les algorithmes de Data mining , pour manipuler les données avec bon interprétation .

1993	arbre C4.5 de J. Ross Quinlan
1996	bagging de L. Breiman
1996	boosting de Freund et Shapire
1998	arcing de L. Breiman
1998	support vector machines de Vladimir Vapnik
2000	régression logistique PLS de Michel Tenenhaus
2001	forêts aléatoires de L. Breiman

Tableau n° 2 : Quelques historiques [14]

b. Types de techniques de Data Mining

On distingue principalement deux types de techniques en Data Mning :

- ✓ **les techniques descriptives** : projection sur des information présentes mais cachés .elles réduisent , résument et synthétisent les données.

- ✓ **les techniques prédictives :** déduire de nouvelles information a partir des informations présentes pour la résolution des problemes **.**

Le tableau ci – dessous présente les méthodes utilisées

Techniques descriptives	Techniques prédictives
- L'analyse factorielle - Classification automatique (typologie , segmentation , clustering , apprentissage non supervisé) - Recherche d'associations.	Classement/ discrimination (variable qualitative) - Analyse discrrimante / régréssion logistique. - Arbre de décision - Réseaux de neurones. Prédiction (variable quantitative) - Régression linéaire (simple et multiple) - Arbre de décision - Reseaux de neurones.

Tableau n° 3 : Méthodes descriptives et prédictives

3.2.2.BASE DE DONNEES :

Une base de données (BD, en anglais DB) , est une entité de stockage des données de façon structurée SGBDR ou Non structuré SGBD Chaque un ces caractéristiques et orienté vers un objectifs précise .

La notion de base de données est généralement couplée à celle de réseau, afin de pouvoir mettre en commun ces informations, deux type de base de données

- **Locale** : utilisable sur une machine par un utilisateur
- **Répartie** : les informations sont stockées sur des machines distantes et accessibles par réseau.

Et si on parle on terme système d'information il faut désigner toute la structure regroupant les moyens mis en place pour pouvoir partager des données.

L'avantage majeur de l'utilisation de bases de données est que plusieurs utilisateurs peut accédées simultanément.

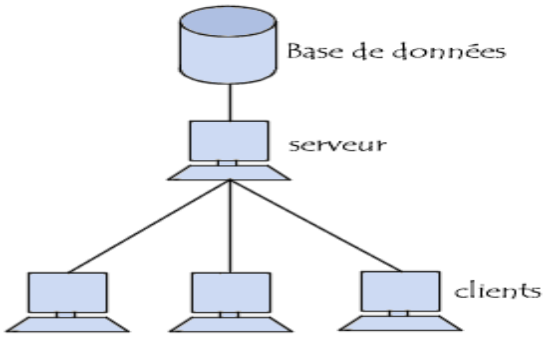

Figure n°2.3 : Base de donnée [15]

3.2.3. ENTROPOSAGE DE DONNEES (DATA WAREHOUSING) :

D'après BILL Inmon: *"Un ED est une collection de données thématiques , intégrées, non volatiles et historisées, organisées pour la prise de décision."*

• **Thématiques:** thèmes par activités;
• **Intégrées:** sources de données diverses ;

• **Non volatiles:** durable non modifiables ou effaçables;
• **Historisées:** trace des données. [D19]

Un entrepôt de données **est une structure contenant des données provenant de plusieurs sources et concernant l'ensemble des processus de l'entreprise (ventes, production, produits, etc).** [16]

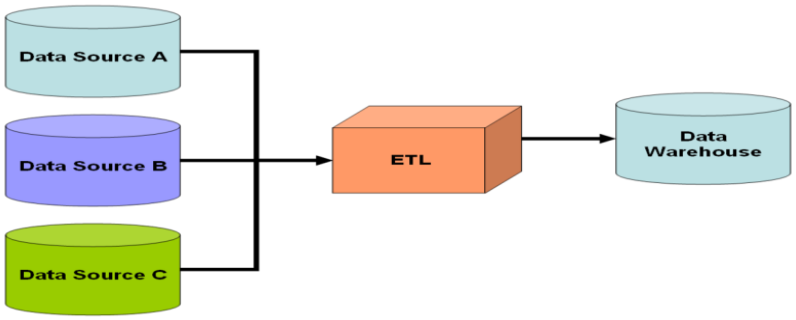

Figure n° 2.4 : Sschéma simple sur Data warehouse [16]

a. ETL : Extract –Transform - Load (Extraction, Transformation et chargement)

L'ETL ce sont plusieurs fonctionnalités combinées dans une seul solution, pour **extraire** des données d'un grand nombre de bases de données, **les transformer** en fonction des besoins et **les charger** dans une autre base de données, un data mart ou un entrepôt de données pour les analyser.

b. Les modeles d' entrepôt de données :

Basée sur des modèles de données qui sont :

- Le **modèle en étoile** basé sur un objet central la table de faits qui contient les faits.

Les tables de dimensions contiennent les attributs qui définissent chacun des membres des dimensions (pas de lien entre elles).

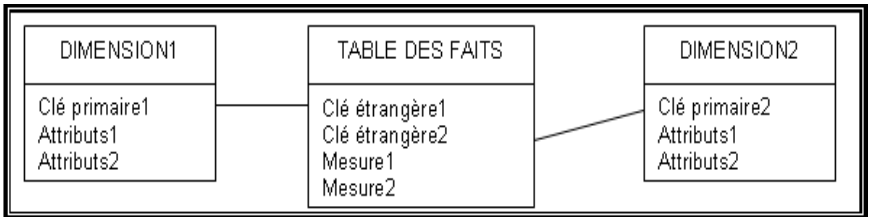

Figure n°2.5 : Le modèle en étoile [17]

- Le **modèle en flocon** : semblable au modèle en étoile la difference chacune des dimensions est décomposée selon sa (ou ses) hiérarchie(s).

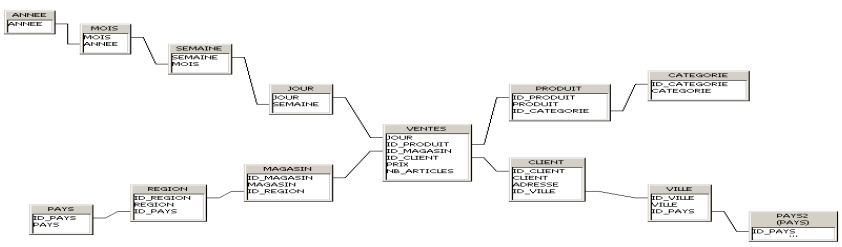

Figure n°2.6 : Le modèle en Flocon [18]

- **Le modèle en constellation** : fusionne plusieurs modèles en étoile qui utilisent des dimensions communes, il comporte donc plusieurs tables de faits et des tables de dimensions communes ou non**.[19]**

L'architecture d'un entrepôt de données se compose de trois services :

- **Base de données** : support de données résumées possédant une structure multidimensionnelle (S.G.B.D. multidimensionnel ou relationnel).
- **Serveur OLAP** : gestion de la structure multidimensionnelle pour accès aux données par des usagers.
- **Module client** : Exploration des données et affichage des données sous différentes formes (statistiques, tableaux, etc.).

Il existe trois architectures d'entrepôts de données :

- **Relational OLAP (ROLAP) :** la base de données est structurée selon un modèle étoile ou flocon. Le serveur extrait les données par les requêtes SQL et interprète les données selon une vue multidimensionnelle pour les présenter au module client sous forme statistique ou tableau**.[19]**
- **Multidimensional OLAP (MOLAP) :** Les données détaillées sont stockées dans une base de données multidimensionnelle (hypercube). elle utilise généralement une structure propriétaire (dépendant du logiciel utilisé). Le serveur OLAP extrait les données de l'hypercube et les présente directement au module client**.[W49]**
- **Hybrid OLAP (HOLAP) :** Utilise l'architecture ROLAP et Molap **[20]**

3.2. 4 . INTELLIGENCE ARTIFICIELLE :

a. DEFINITION ET HISTORIQUE :

L'origine de l'intelligence artificielle se trouve dans l'article d'Alan Turing *« Computing Machinery and Intelligence »* (Mind, octobre 1950) , où Turing propose une expérience maintenant connue sous le nom de test de Turing .

L'intelligence artificielle désigne un ensemble de théories et des techniques mises en œuvre en vue de fabriquer des machines capables de simuler l'intelligence humaine .

Vers la fin de la seconde guerre mondiale. En 1956, dix pionniers de la recherche américaine en théorie des automates, des réseaux de neurones et de l'intelligence se sont réunis pour un workshop de deux mois à Dartmouth College : J. McCarthy (Dartmouth), Minsky (Princeton), C. Shannon (Bell Labs/MIT), N. Rochester (IBM), T. More (Princeton), A. Newell (Carnegie Tech), H. Simon (Carnegie Tech), A. Samuel (IBM), R. Solomonoff (MIT), O. Selfridge (MIT). C'est lors de ce workshop, que le terme d'IA est apparu . **[21]**

Concrètement, en médecine, Watson (IBM) peut analyser toutes les données d'un patient: ses symptômes, les consultations médicales, ses antécédents familiaux, ses résultats d'examen, ses données comportementales, etc. Et appliquer le savoir scientifique à un individu particulier. «Il peut ainsi engager avec le professionnel une discussion collaborative dans le but de déterminer le diagnostic le plus vraisemblable et les options de traitement, Comme il peut comparer un patient particulier, sa situation, et son pronostic en fonction de l'effet de tous les traitements déjà appliqués à tous les patients similaires avant lui**.[22]**

Cas d'exemple En 2015, une patiente japonaise de 60 ans se présente à l'hôpital de Tokyo avec une forme grave de leucémie. Pourtant, le cancer résiste à la chimiothérapie préconisée. L'équipe décide alors de faire appel à Watson, le super ordinateur d'IBM, pour résoudre ce cas étrange. Un monceau d'informations est rentré dans la machine : profil génétique de la patiente, y compris les mutations possibles, revues d'essais cliniques, et plus de 20 millions d'études d'oncologie. Bingo : en moins de 10 minutes, Watson découvre la pathologie dont souffre la malade (une anomalie dans la moelle osseuse). **[23]**

b. Langage de L'IA :

- **CLISP (programmation symbolique).**
- **PROLOG (développé par l'Université de Marseille).**

3.2.4.1. LES SYSTÈMES EXPERTS POUR L'AIDE À LA DÉCISION MÉDICALE

l'apparition des systèmes experts a été apparu dans les années 70. **[24],** Le principe de cette approche est de demander à un ou plusieurs experts d'établir des règles qui décrivent leur façon de prendre leurs décisions **(Främling, 1992**).

Les systèmes experts tentent de simuler le savoir-faire, la façon de raisonner des experts dans un domaine donné bien délimité et précis et la mise des connaissances apprises par le processus à la disposition des utilisateurs ou experts du domaine. Ils doivent réaliser à la fois un traitement d'information et un raisonnement, ce qui signifie en termes médicaux, une stratégie de diagnostique et/ou thérapeutique **(Darmoni, 2003).**

Donc le rôle d'un système expert médical est de fournir une aide médicale sous une forme appropriée à partir de symptômes préalablement établis. **[Garg AX et al , 2005]**

a. Définition :

On peut définir le système expert comme étant un système informatique qui imite la démarche de la personne compétente dans un domaine donné, quelle que soit la méthode de raisonnement qu'elle utilise.

De plus, il doit être interactif, capable de dialoguer avec ses utilisateurs et d'expliquer ses raisonnements. **[Kawamoto Ket al , 2005]**

3.2.4.2. LES COMPOSANTS D'UN SYSTÈME EXPERT

Traditionnellement, un système expert se compose d'une base de connaissances, d'un moteur d'inférences et de différentes interfaces qui lui permettent de communiquer avec son environnement.

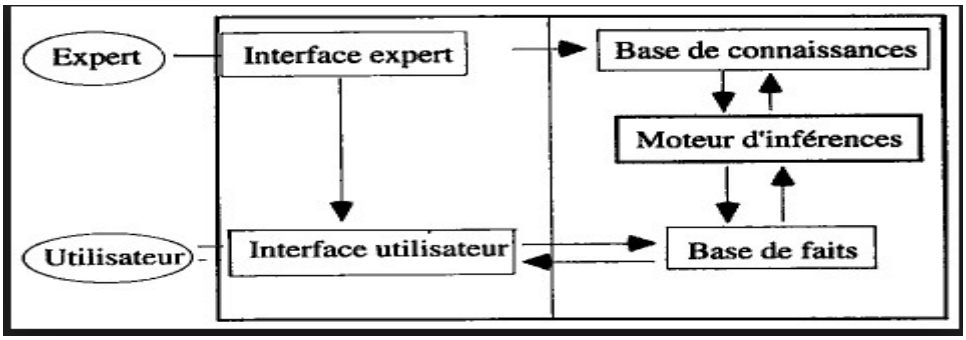

Figure n° 2.7 : Système d'aide a la decision par le Système Expert pour medecine [25]

a. La base de connaissances :

La base de connaissances est élaborée à partir de l'expertise d'un spécialiste. L'expertise elle-même s'obtient au cours d'un processus cognitif. C'est certainement à ce niveau que l'on peut situer la phase de création du savoir. La base de connaissances se compose d'une base de faits et d'une base de règles.

- **Base de règles :** La base de règles contient les connaissances expertes (règles de l'expert) qui sont représentées généralement par des règles de production s'écrivant sous la forme :

  Si Condition Alors Action

- **Base de faits :** La base de faits est l'ensemble des propositions connues du système à un moment donné. C'est la mémoire de travail du système expert. Son contenu dépend du problème traité. La base de fait s'intègre deux types de faits : les faits permanents du domaine et les faits déduits par le moteur d'inférences qui sont propres au cas traité **[Negrello, 1991].**

b. Le moteur d'inférences :

Le moteur d'inférences est un programme qui utilise les règles définies dans la base de connaissances pour résoudre un problème particulier décrit par des faits.

c. L'interface :

L'interface pour l'aide à l'acquisition des connaissances fournies par l'expert peut être plus ou moins sophistiquée, l'accent étant souvent mis sur une syntaxe des règles la plus proche possible du langage naturel

3.2.4.3.CARACTÉRISTIQUES D'UN SYSTÈME EXPERT

Un système expert peut être conçu pour qu'il ait les caractéristiques générales suivantes :

- **Haut rendement** : Le système doit avoir la capacité de répondre à un niveau de compétence égal ou supérieur à celui d'un spécialiste du domaine. Cela signifie que la qualité de conseil donné par un système doit être très haute.
- **Temps de réponse adéquat** : Le système doit agir en un temps raisonnable, comparable ou meilleur au temps exigé par un spécialiste, pour prendre une décision.
- **Fiabilité** : le système expert doit être fiable et ne doit pas connaître des "failles" sinon il ne sera pas utilisé.
- **Compréhensible** : le système doit être capable d'expliquer les étapes de son raisonnement pendant qu'elles s'exécutent, au lieu d'être seulement une boîte noire qui produit une réponse miraculeuse.
- **Flexibilité** : Vu la grande quantité de connaissance qu'un système expert peut avoir, il est important d'avoir un mécanisme efficient pour ajouter, modifier, et éliminer la connaissance. Une raison de la popularité des systèmes experts basés sur les règles est la capacité de stockage efficiente et modulaire des règles.

a. Avantages : [Schadrac KANDE KANUMAMBIDI ,2009]]

Les systèmes experts ont plusieurs caractéristiques :

- **Grande disponibilité :** production et disponibilité massive de l'éxpérience.
- **Coût réduit :** Le coût de mettre l'expérience à la disposition de l'utilisateur est réduit énormément.
- **Danger réduit :** les humains travaillons parfois dans des endrois dangereux , en peut les remplacés

par l'utilisation du systéme expert .

- **Permanence :** la connaissance d'un système expert durera indéfiniment .
- **Expérience multiple :** La connaissance des plusieurs specialistes sont disponibles 24/24h et 7/7 j
- **Explication :** il peut expliquer clairement et en détail le raisonnement qui conduit à une conclusion.
- **Réponse rapide :** il peut nous donnée des reponses en temps réelle , plus rapidement qu'un spécialiste humain .
- **Réponses solides, complètes et sans emotions :** un spécialiste humain ne fonctionnera pas avec toute sa capacité à cause de la pression et de la fatigue.
- **Enseignement intelligent :** il peut agir comme un enseignant intelligent (l'étudiant exécute des programmes qui explique le raisonnement du système).
- **Base de données intelligente :** accès à une base de données de manière intelligente.

b.INCONVENIENS: [Schadrac KANDE KANUMAMBIDI ,

2009]]

Les deux principaux inconvénients des systèmes experts sont :

- Ils créent le chômage (émulent les humains).
- on fait inférence à des connaissances même si elles sont dépassées.

3.2.4.4 .QUELQUES SYSTÈMES EXPERTS DANS LE DOMAINE MÉDICAL

Les systèmes experts ont fait leur apparition dans certains domaines, particulièrement en médecine et ce dans plusieurs spécialités. Pourtant ces systèmes n'ont eu qu'un faible impact car ils étaient sous-utilisés par les praticiens qui leur reprochaient entre autres l'absence de méthodologie valide dans la constitution des bases de connaissances.

Le tableau suivant résume quelques systèmes experts médicaux en précisant leur domaine et leur but.

SYSTÈME EXPERT	DOMAINE	BUT
MYCIN	Maladies infectieuses	Identification des microorganismes responsables des infections, conseil sur le choix d'un antibiotique
INTERNIST-I	Médecine interne	Diagnostic des problèmes complexes en médecine interne
SPHINX	Endocrinology	Aide au traitement du diabète nid, aide au diagnostique des ictères, proposition d'un modèle de simulation d'une consultation médicale
PATHFINDER	Chirurgie des ganglions lymphatiques	Diagnostic des maladies des ganglions lymphatiques
SETH	Intoxications médicamenteuses	Diagnostic et traitement des intoxications médicamenteuses

Tableau 4 : Quelques systèmes experts médicaux **[26]**

Evaluation de système de IA

L'expérience a prouvé que la maintenance et l'évolution des systèmes experts était une tâche difficile.

D'après ICML *(International Conference on Machine Learning) 6 au 11 juillet, Lille accueille .* Colin de la Higuera *ET Thierry Viéville et Sylvie Boldo* ils ont dis que l'intelligence artificielle n'a pas tenu ses promesses , Pourtant ces algorithmes d'apprentissage automatique, sont bel et bien présents. Et ceci, au delà, du fait que ces idées sont à la source de nombreuses pages de science-fiction. ***[27]***

« L'intelligence artificielle est la science (comment nous rendons les machines intelligentes), tandis que le machine learning est l'exécution des méthodes qui la soutiennent, comme les algorithmes », explique à Wired Nidhi Chappell, responsable du machine learning chez Intel.
« L'intelligence artificielle est une branche de l'informatique qui vise à développer des appareils capables de se comporter intelligemment, alors que l'Université de Stanford définit le machine learning comme « la science permettant aux ordinateurs d'agir d'eux-même sans être explicitement programmés pour ce faire » », résume pour sa part Wired. Rappelant ainsi que le domaine de l'intelligence artificielle, s'il comprend bien le machine learning, est néanmoins bien plus vaste que ce seul dernier.**[28]**

3.24.6. Case base resoning (Résonnement a base de cas) :

Le CBR est une méthode de raisonnement qui signifie qu'il résulte d'anciens cas ou d'expériences pour résoudre des problèmes ou interpréter des situations anormales .

Dans CBR, le raisonnement est basé sur le rappel des expériences passées, comme l'ont expliqué Althoff et al.**[W61]** "Pour résoudre un problème, rappelez-vous un problème similaire que vous avez résolu dans le passé et adaptez l'ancienne solution pour résoudre le nouveau problème".

CBR est issu de la recherche en sciences cognitives. Les premières contributions dans ce domaine proviennent de Roger Schank et de ses collègues de l'Université de Yale [K. D. Althoff ET AL , 1995]
.

3.3. APPRENTISSAGE AUTOMATIQUE (MACHINE LEARNING):

3.3.1.DEFINITION ET HISTORIQUE

D'apres **[Arthur Samuel ,2016]** *Le Machine Learning est le champ d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés*

l'apprentissage automatique est devenu comme une matière enseignée dans les universités également c'est un sujet de recherche très actif basant sur l'etude des familles d'algorithmes informatiques les plus efficace afin d'améliorer automatiquement l'expérience . Les algorithmes, deviennent d'autant plus performants quand la quantité de données augmente. c'est a dire travaillons avec les donnée immense des *big data* .

Les aventures de Machine learning dans le domaine de médecine

- **APPRENTISSAGE AUTOMATIQUE ET LA NEURO-SCIENCE**

Depuis les années 1990, Plusieurs équipes de chercheurs se tournent non plus seulement vers la neuro-imagerie, mais aussi vers le machine learning présente des méthodes de neuro-imagerie étudiant notre activité cérébrale et mentale.

OBJECTIF : prédire les crises d'épilepsie, contribuer au diagnostic et au traitement de désordres mentaux **– [29]**

- **APPRENTISSAGE AUTOMATIQUE ET L'OPHTALMOLOGIE**

Une fondation médicale indienne en collaboration avec Microsoft et le L V Prasad Eye Institute qui lutte contre les maladies des yeux , utilisent le machine learning et le cloud, les chercheurs en extrait des patterns afin de prévoir les résultats d'opérations de chirurgie oculaire. Le travail de ces équipes permet par exemple aux médecins d'évaluer le meilleur moment pour effectuer une opération, ou de suivre (et prévenir) la propagation de maladies des yeux, voire de déterminer les conditions dans lesquelles la vue des enfants se détériore.Face au succès des premiers tests, le programme va maintenant s'étendre à d'autres universités et instituts de recherche aux Etats-Unis, en Australie et au Brésil.

OBJECTIF : affiner les modèles prédictifs en élargissant le jeu de données**.[30]**

- **APPRENTISSAGE AUTOMATIQUE ET L'ANALYSE VOCALE**

Des chercheurs new –yorkais obtiennent des premiers résultats dans le diagnostique de troubles de stress pos-traumatique et de maladies cardiaque en utilisent des technique d'analyse vocale assistée par machine learning ,Captées des vois par smartphone, puis analysées et comparées à des masses de données d'autres patients ou de personnes en bonne santé psychique, et a travers les variations de ton, de volume sonore , autant de signes contribuant à indiquer une TSPT (toubles de stress pos-traumatique), mais aussi une dépression ou un traumatisme crânien, De telles applications, que les chercheurs espèrent a etre téléchargeables à moindre coût sur un simple smartphone, pourraient être utilisées pour identifier en amont des patients à risque ou pour surveiller, à distance, des patients après une chirurgie du cœur.[31]

3.3.2.Les grandes familles d'algorithmes : Les algorithmes construisent un modèle à partir de données dans le but d'émettre des prédictions ou des décisions , Beaucoup d'algorithmes Regroupés en grandes familles , Nous avons schématisé les classes d'algorithmes comme dans le schéma qui suit :

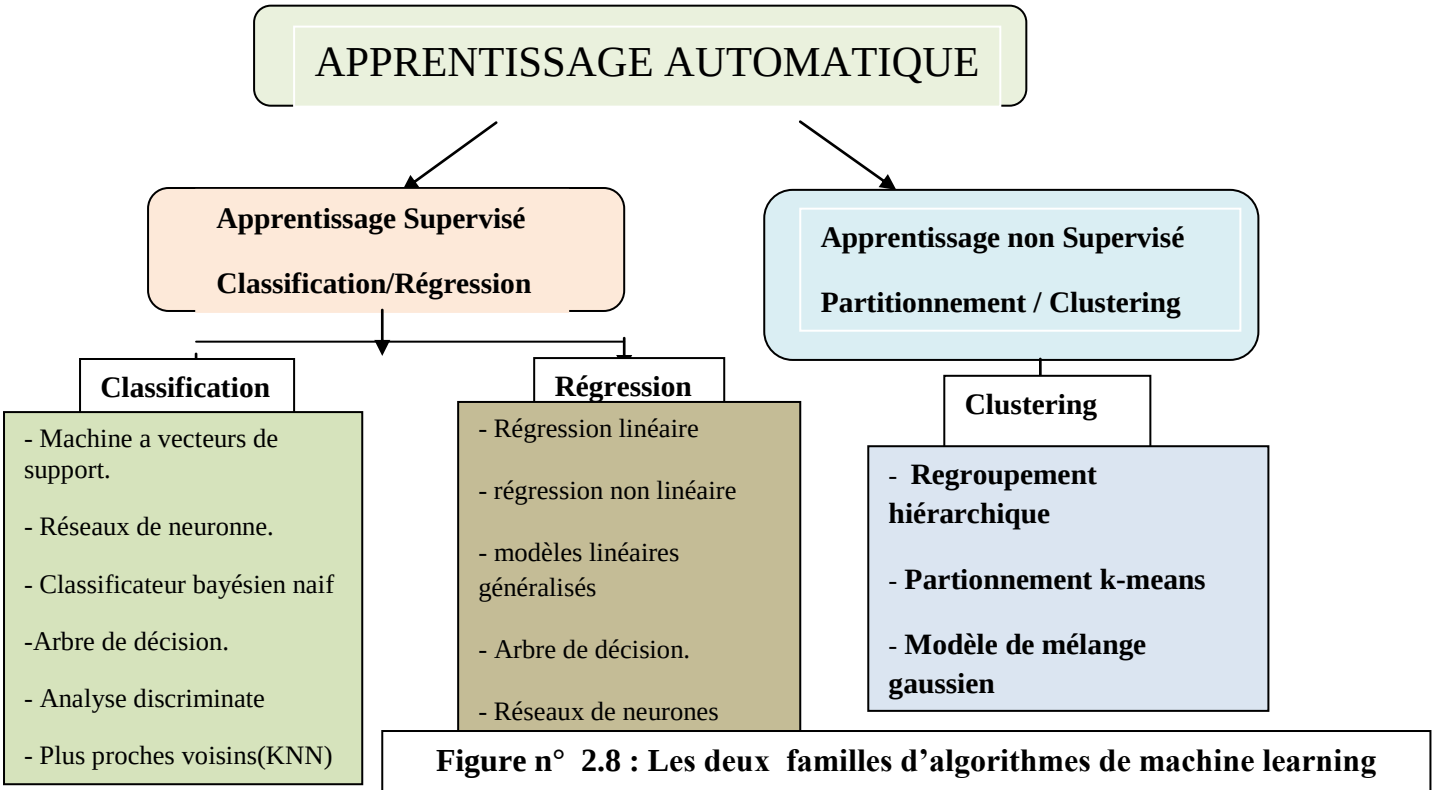

Figure n° 2.8 : Les deux familles d'algorithmes de machine learning

3.3.2.1.Apprentissage Supervisé (Classification / Régression) :

a.Definition

Utilise un ensemble de données d'apprentissage pour faire des prévisions. Et prend un ensemble connu de données d'entrée et des réponses connues aux données de sortie et forme un modèle pour générer des prédictions raisonnables .

Exemple Médicale :

vous voulez prédire si quelqu'un aura une crise cardiaque dans l'année. Vous avez un ensemble de données sur les patients précédents, l'âge, le poids, la taille, la tension artérielle,Vous savez si les patients précédents ont eu des crises cardiaques dans l'année suivant leurs mesures. Donc, le problème est de combiner toutes les données existantes en un modèle qui peut prédire si une nouvelle personne aura une crise cardiaque dans un an. **[32]**

b.Étapes de l'apprentissage supervisé :

Il existe de nombreux algorithmes pour l'apprentissage supervisé, la plupart utilisent les même étapes qui sont:

1- Préparé les données :

Toutes les méthodes d'apprentissage supervisées commencent par une matrice de données d'entrée, appelée X ici. Les rangée de X représente une observation. Et les colonne de X représente une variable ou un prédicteur ,ng peuvent ignorer les valeurs NaN . Chaque élément de Y représente la réponse correspondante de X. Les observations avec les données Y manquantes sont ignorées.

2- Choisir l'algorithme :

Il existe des caractéristiques dans des algorithmes, tels que:

- Vitesse d'apprentissage
- Utilisation de la mémoire
- Précision prédictive sur les nouvelles données
- Transparence ou interprétation, ce qui signifie vous pouvez facilement comprendre les raisons qu'un algorithme fait ses prédictions

3- Adapter un Model :

La fonction d'adaptation que vous utilisez dépend de l'algorithme que vous choisissez.

4- Choisir une Méthode de validation :

5- Examiner l'adaptation et la mise à jour jusqu'à ce que soit satisfait :

Après avoir validé le modèle, vous pouvez le modifier pour une meilleure précision, une meilleure vitesse ou moins de mémoire.

- Changez les paramètres d'ajustement pour essayer d'obtenir un modèle plus précis.
- Modifiez les paramètres d'ajustement pour essayer d'obtenir un modèle plus petit. Cela donne parfois un modèle avec plus de précision
- Essayez un algorithme différent. Pour les choix applicables

6- Utiliser un modèle adapté pour les prévisions :

Pour prédire la classification ou la réponse de régression pour les modèles les plus adaptés, utilisez la méthode de prédiction citer dans le shéma .

3.3.2.2.Apprentissage Non Supervisé :

L'apprentissage non supervisé est utilisé pour tirer des conclusions à partir de jeux de données composés de données d'entrée sans réponses catégorisées. L'apprentissage non-supervisé concerne les modèles descriptifs où il n'existe pas de cible explicite**.[33]**

3.3.3.Algorithmes les plus utilisés en medecine

a.Algorithme génétique :

Les algorithme génétique représentent une des méthodes d'optimisation les plus utilisés . ils ont été proposé Par John Holland de l'université de Michigan , son recherche avait pour objectif la modélisation des processus d'adaptation des systèmes naturels et la conception des systémes artificiels dotés des mêmes propriété .

Les travaux connexes :

Auteur	Année	Titre du livre	Objectif
Iman Beheshti , Hasan Demirel, Hiroshi Matsudaa.	**2017**	**Classification de la maladie d'Alzheimer et prédiction de la déficience cognitive légère à la transformation d'Alzheimer à partir de l'imagerie de ressource magnétique structurale en utilisant le classement des caractéristiques et un algorithme génétique**	**Ils ont développé un nouveau système de diagnostic assisté par ordinateur (CAD) qui utilise le classement des fonctionnalités et un algorithme génétique pour analyser les données structurelles d'imagerie par résonance magnétique; En utilisant ce système, ils pu prédire la conversion de la maladie cognitive modérée (MCI) à la maladie d'Alzheimer (DA) entre un et trois ans avant le diagnostic clinique**

b. Algorithme de l'arbre de décision :

Les premiers algorithmes de classification par arbres de décision sont anciens. Les deux travaux les plus marquants sont la création de CART, par Breiman en 1984 et la création de C4.5 par Quinlan en 1993.Uniquement les méthodes CHAID (Chi-squared Automatic Intteraction Detection) et CART (Classification And Regression Trees) sont utilisées de manière récurrente en médecine

Les arbres de décision sont des arbres qui permettent, la plupart du temps, de donner une réponse binaire du type oui ou non.Ils pourraient donner en sortie des valeurs réelles, mais ce n'est pas beaucoup utilisé**.[34]**

Les travaux connexes :

Auteur	Année	Titre du livre	Objectif
V.V Satyanarayana Tallpragada ,all	2015	A NOVEL MEDICAL IMAGE SEGMENTATION AND CLASSIFICATION USING COMBINED FEATURE SET AND DECISION TREE CLASSIFER	Le diagnostic automatisé de l'image médicale On observe que le système a entraîné une segmentation précise de 70% et l'utilisation des caractéristiques extraites de l'image de la tumeur segmentée, les résultats des tests sont prometteurs et une précision de 94% est atteinte

c.Algorithme de régression :

Les techniques de régression linéaire sont utilisées pour créer un modèle linéaire.

Le modèle décrit la relation entre une variable dépendante y (également appelé la réponse) en fonction d'une ou plusieurs variables indépendantes Xi (appelé prédicteurs). La variable décisionnelle est quantitative . Cette méthode peut nous aider à comprendre et a prédire le comportement de systèmes complexes ou bien à analyser des données expérimentales, financières ou **biologiques**.

travaux connexes :

Auteur	Année	Titre du livre	Objectif
Anirah Ahmad and Hasimah Hj. Mohamed	2016	L'AMÉLIORATION DE L'ALGORITHME DE REGRESSION LINEAIRE DANS MANIPULATION DES DONNÉES MANQUANTES POUR LES DONNÉES MÉDICALES	Les problèmes de données manquants sont très fréquents et Se produit principalement dans les domaines médicaux. Mackinnon, A conclu quelques données manquantes sont un problème omniprésent Presque tous les domaines de la recherche médicale. L'observation de Mackinnon, a montré une augmentation radicale des articles Sur l'application de MI aux analyses de données
J Wei, M Chao - Medical Physics	2016	SU-G-BRA-08: Diaphragm Motion Tracking Based On KV CBCT Projections with a Constrained Linear Regression Optimization	Le nouvel algorithme fournira une solution potentielle pour rendre le mouvement du diaphragme et finalement améliorer la gestion du mouvement tumoral pour la radiothérapie des patients cancéreux
P Wang, R Ge, X Xiao, M Zhou	2016	hMuLab: a Biomedical Hybrid MUlti-LABel Classifier Based on Multiple Linear Regression	Développer une nouvelle stratégie pour extraire le mouvement respiratoire du diaphragme thoracique des projections de tomodensitométrie du faisceau à cône kilovoltage (CBCT) par une technique d'optimisation de régression linéaire contrainte
CK Fisher, P Mehta	2014	Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression	LIMITS utilise une régression linéaire clairsemée avec agrégation de boostrap pour inférer un modèle de Lotka-Volterra à temps discret pour la dynamique microbienne , Sur la base de leur résultats, ils ont mis l'hypothèse que l'abondance de certaines espèces clés peut être responsable de l'individualité dans le microbiome intestinal humain

d. Algorithme de Clustering :

Le Clustering est la classification non supervisée: pas de classes prédéfinies .La classification se déroule séquentiellement en regroupant les observations les plus semblables

Les travaux connexes :

Auteur	Année	Titre du livre	Objectif
M Latha, R Surya - Brain,	2016	Détection de tumeur cérébrale à l'aide du classificateur de réseaux neuronaux et algorithme de clustering de k-Means pour la classification et la segmentation	dans un système Ils ont utilise Clustering qui été est très utile pour diagnostiquer les images cérébrales de résonance magnétique pour le diagnostic de tumeur. Ils l on appliquée avec plus d'une tranche d'IRM cérébrale afin d'obtenir une meilleure précision.

e.Les réseaux de neurones

Un réseau de neurones est un modèle de calcul dont le fonctionnement vise à simuler le fonctionnement des neurones biologiques, il est constitué d'un grand nombre d'unités (neurones) ayant chacune une petite mémoire locale et interconnectées par des canaux de communication qui transportent des données numériques. Les réseaux de neurones sont capables de prédire de nouvelles observations (sur des variables spécifiques) à partir d'autres observations (soit les même ou d'autres variables) après avoir exécuté un processus d'apprentissage sur des données existantes.
La phase d'apprentissage d'un réseau de neurones est un processus itératif permettant de régler les poids du réseau pour optimiser la prédiction des échantillons de données sur lesquelles l'apprentissage été fait. Après la phase d'apprentissage le réseau de neurones devient capable de généraliser

- ### Traveaux connexes :

Auteur	Année	Titre du livre	Objectif
S Raith, EP Vogel, N Anees, C Keul, JF Güth	2017	**Réseaux de neurones a**rtificiels comme outil numérique puissant pour classer les caractéristiques spécifiques d'une dent en fonction des données de balayage 3D	Réseaux de neurones artificiels comme outil numérique puissant pour classer les caractéristiques spécifiques d'une dent en fonction des données de balayage 3D
V Gulshan , L Peng, M Coram, MC Stumpe , D Wu	2016	Développement et validation d'un **algorithme** d' **apprentissage** en profondeur pour la détection de la rétinopathie diabétique dans les photographies du fond de la rétine	utilisée l'algorithme de réseaux de neurones pour la détection automatique de la rétinopathie diabétique diabétique dans les photographies du fond de la rétine.et l'œdème maculaire

f.Algorithme de KNN :

L'algorithme de plus proche voisin et un algorithme supervisé utilisé pour la classification KNN en englais K—nearest neighbor, est classé parmi les algorithmes les plus simples qui sont utilisé par l'apprentissage automatique .

L'objectif : comparé les facteurs pertinents d'une question par un patient avec la base de connaissance qui est notre base de donnée , est a la fin classé les patients d'après leurs maladies .

Nous avons contruit un tableau des avantages et inconvenient de chaque algorithme

3.3.4.Avantages et inconvénients de chaque algorithmes :

ALGORITHME	AVANTAGE	INCONVENIANT
Génétique	- trouver de bonnes solutions sur des problèmes très complexes, et trop éloignés des problèmes combinatoires classiques	- Sont coûteux en temps de calcul, puisqu'ils manipulent plusieurs solutions simultanément. - L'ajustement d'un algorithme génétique est délicat - lorsque les différents individus se mettent à avoir des performances similaires : les bons éléments ne sont alors plus sélectionnés, et l'algorithme ne progresse plus
Arbre de décision	– La structure hiérarchique des arbres permet d'extraire des règles compréhensibles pour l'utilisateur. – Les arbres permettent de décomposer des problèmes complexes en une union de problèmes simples. –les modèles sont robustes et fonctionnent bien sur de grandes bases de données, pour des temps de calculs faibles.	- Aucune raison qu'un tel arbre de décision ait de bonnes qualités en généralisation , parce qu'il y a un grand nombre d'arbres de décision différents ayant cette propriété et qu'un arbre choisi aléatoirement parmi eux n'a aucune chance de faire partie des arbres ayant les meilleures performances prédictives. - lors de l'utilisation d'un petit ensemble de données. Ce problème peut limiter la généralisation et la robustesse des modèles résultants.
Clustering	- k-means est très facile à comprendre et à mettre en oeuvre. - Sa simplicité conceptuelle et sa rapidité . - Applicable à des données de grandes tailles, et à tout type de données .	- si la mesure de la distance n'existe pas, nous devons la «définir», ce qui n'est pas toujours facile, surtout dans des espaces multidimensionnels. - le résultat de l'algorithme de clustering peut être interprété de différentes manières. - Elles sont coûteuses en temps de calcul et sont de plus sensibles à la dimension des données. - ID3 et C4.5 moins bons (d´etection des maladies cardiaques). - C4.5 produit plus d'erreurs de prédiction que ID3, mais ces taux restent néanmoins très faibles **[35]**
Regression lineaire	- le temps pour la prédiction est Rapide .	Un des plus gros inconvénient de la méthode de régression linéaire est qu'elle marche toujours ! C'est en effet un inconvénient car l'utilisateur qui a calculé les deux coefficients $\hat{a}$ et $\hat{b}$ de la droite de régression ne sait pas si celle-ci est un modéle acceptable de ses données ou si ce n'est pas le cas du tout**.[36]**
Reseau de Neuronne	- Les réseaux de neurones sont théoriquement capables d'approximer n'importe quelle fonction continue et ainsi le chercheur n'a pas besoin d'avoir aucunes hypothèses du modèle sous-jacent .	- Généralement les réseaux de neurones ne sont pas souvent utilisées dans les tâches du data mining parce qu'ils produisent des modèles souvent incompréhensibles et demande un longtemps d'apprentissage **[37]**
KNN	- Apprentissage rapide . - Méthode facile a comprendre	- Méthode complexe

Tableau n° 6 :Avantages et inconvénients de chaque algorithmes

3.3.5.CHOISIR L'ALGORITHME

Nous avons construits un tableau comparatif d'après la notion critères des algorithmes

Algorithme	Précision Prédictive	Vitesse d'apprentissage	Utilisation Mémoire	Interprétation
Arbre de décision	Oui	Rapide	petite	Facile , mais il ne faut pas qu'il y en ait trop c a dire plusieurs arbres de décision
Régression Linéaire	Oui , surtout pour des problèmes complexe	Rapide	Petite	**Cette méthode peut nous aider à comprendre et a prédire le comportement de systèmes complexes ou bien à analyser des données expérimentales, financières ou biologiques, mais très complexe.**
Clustering	Oui	Elles sont coûteuses en temps de calcul		le résultat de l'algorithme de clustering peut être interprété de différentes manières.
Génétique	Oui , surtout pour des problèmes complexe	Rapide , couteux en temps de calcule	Grand	Si différents individus se mettent à avoir des performances similaires : les bons éléments ne sont alors plus sélectionnés, et l'algorithme ne progresse plus.
KNN	oui	Apprentissage rapide	Grand	Facile , a comprendre

Tableau n° 7 : Les résultats de ce tableau sont basés sur une analyse sur les avantages et les désavantages de chaque algorithmes .

La vitesse:

Rapide - 0.01 seconde
Moyenne - 1 seconde
Lentement - 100 secondes

Mémoire

Petit - 1 Mo
Moyen - 4 Mo
Large - 100 Mo

D'âpres les résultats de la comparaison on va choisir l'algorithme de KNN pour sa simplicité

CONCLUSION :

Et a la fin on va choisir l'algorithme KNN pour continuer notre contribution parce que c'est un algorithme évolutif et peut aider à comprendre et a prédire le comportement de systèmes complexes et analyser des données Médicale aussi composé d'une méthode facile a comprendre avec apprentissage rapide

INTRODUCTION :

Dans ce chapitre nous allons présenté des architectures qui représente les processus de construction de la base de donnée ainsi la mise a jour , la gestion de la qualité de données médicale .

1. ARCHITECTURE PROPOSE

1.1 .CONSTRUCTION DE LA BASE DE DONNEE

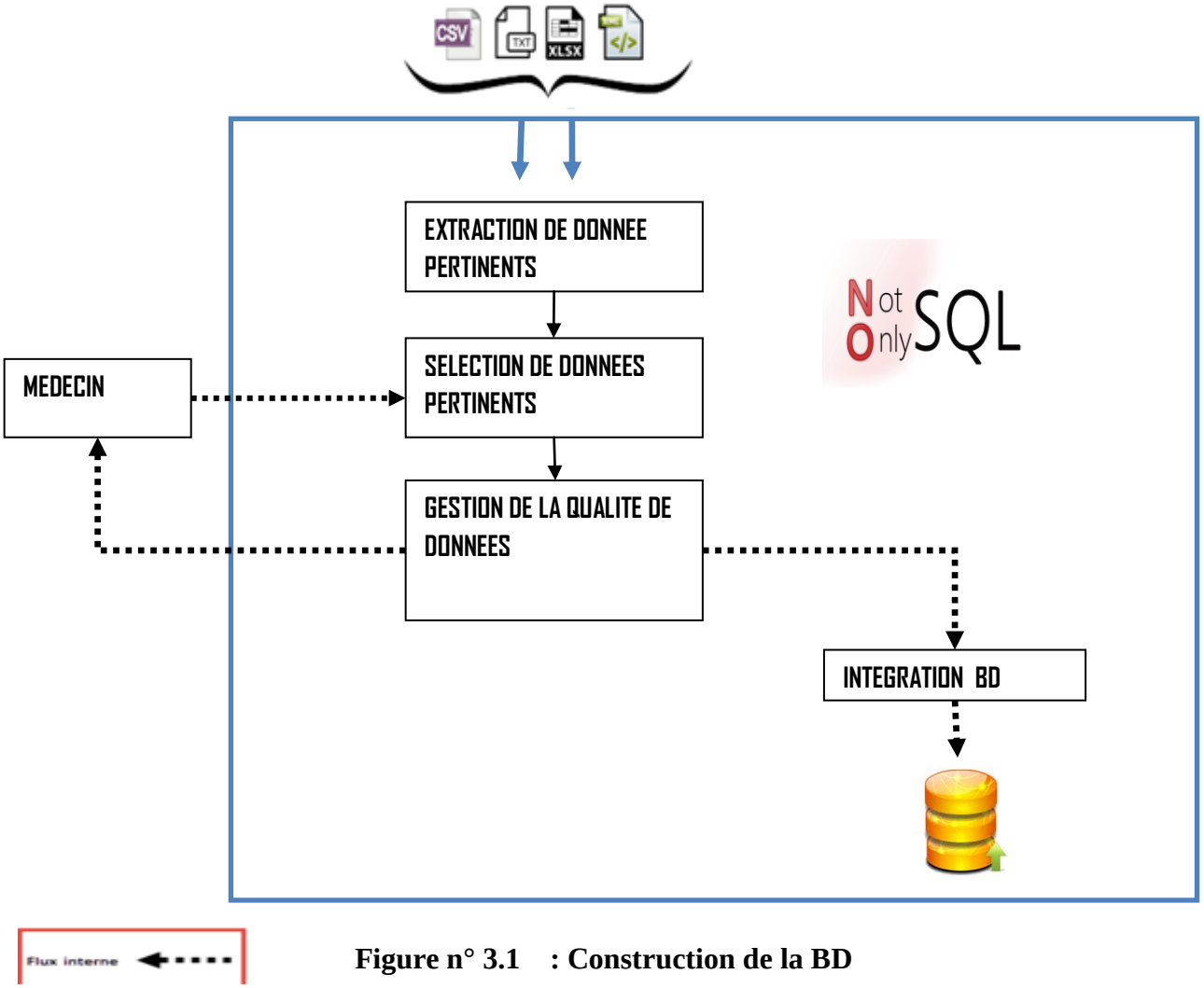

Figure n° 3.1 : Construction de la BD

Description :

Après la collecte des donnees par differnts sources, ils serons passé par les etapes suivants :

- **ETAPE 1 :** EXTRACTIONDE DONNEE PERTINENTS (basée sur les facteurs interessents ici les symptomes de la maladie a comparé).
- **ETAPE 2 :** SELECTION DE DONNEE PERTINENTS (a travers l'etape precedents + l'avis des medecins)
- **ETAPE 3 :** GESTION DE LA QUALITE DE DONNEES (les donnees collecter seront passé par des processus de gestion ou traitements).
- **ETAPE 4 :** INTEGRATION DE LA BASE DE DONNEE :Stockage des données traité dans une base de donnée

1.2.MISE A JOUR DE LA BASE DE DONNEE

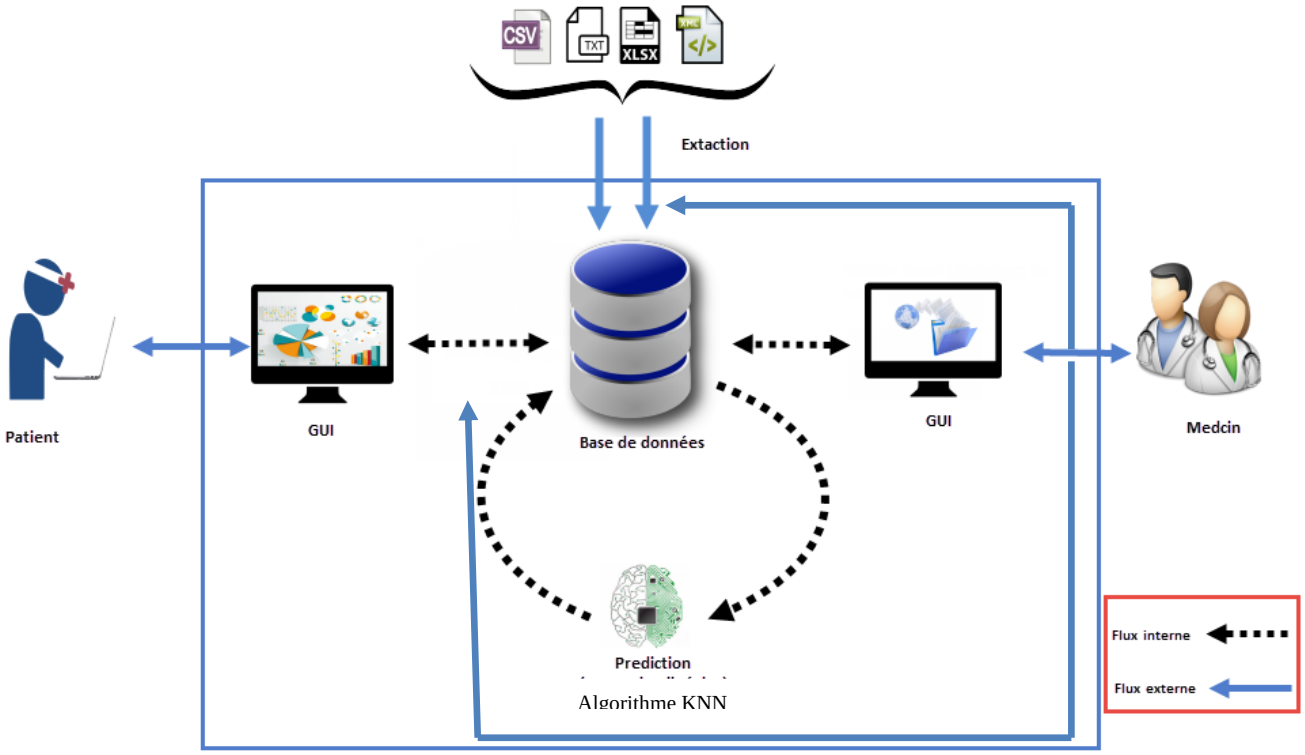

Figure n ° 3.2 : Mise a jour de la BD

Description :

Après l' intégration de la base de donnée

- **INTERFACE UTILISATEUR :** le patient pose la question a travers l'interface utilisateur Aussi le médecin donne son avis a travers son interface il recoit la question du patient ainsi le pourcentage réalisé par les algorithmes spécialisé et reste son avis sur la maladie

- **BASE DE DONNEE :** contient les données collecter et traiter

- **PREDICTION :** a travers l'apprentissage automatique qui utilise des algorithmes ici nous avons choisie l'algorithme de REGRESSION LINEAIRE (une comparaison qui a été faite dans le chapitre précédent)

- **Médecin** : il reçoit le questionnaire du patient avec le pourcentage réalisé par l'algorithme et la décision et rester au médecin qui va donner son avis **.**

DIGRAMME DE FLUX DE DONNEE : est un type de représentation graphique du flux de données à travers un système d'information. Cet outil est souvent utilisé comme étape préliminaire dans la conception

d'un système afin de créer un aperçu de ce système. Mais il n'indique ni la temporalité des transmissions de données, ni l'ordre dans lequel les données circulent.

- **Flux de données interne :** Ensemble des informations utiles à une activité précise circulant d'un point interne dans un système à un autre point interne dans le même systeme.ici nous avons :

L'avis du médecin va être intégrer dans la base de donnée pour qu'elle sera utilisé a l'aide de décision a la prochaine fois .

L'avis du médecin sera transmis comme résultat ou diagnostic au patient .

- **Flux de données externe :**Ensemble des informations utiles à une activité précise circulant d'un point externe hors du système à un autre point interne dans le système ou vis versa.

1.3. GESTION DE LA QUALITE DES DONNEES MEDICALES :

Dans notre première phase , nous proposons quelques indicateurs pour mesurer la qualité des données utilisables par les acteurs medicaux .

Les indicateurs sont :

1.3.1. Utilisabilité : ce facteur nous permet de mesurer l'utilisation d'une information a la date (1-)si une information n'est pas utilisé pour une durée par les experts , alors elle serra archivé(en supprime pas les informations médicales).

1.3.2. Pertinence : sur tout pour les avis médicaux du médecin et du patient nous avons introduits un paramètre de pertinence

Pour chaque avis , a la demande de l'intéressé de mesuré sa qualité (1 si elle a de bonne qualité , -1 sinon)

Puis : si Représente de l'avis et si P_+ est les cas positifs alors :

P_+/P est le taux de pertinence

Si $P_+/P < ع$ (ع: seuil d'avis par les experts)

Alors l'avis sera archivé.

$$P = P_+ + P_- + N$$

N : le nombre des utilisateurs des avis qui ne répond pas a la question

Selon l'algorithme suivant :

Algorithme Pertinence

P : entier

ε: reel

Début

Lire (avis)

Si avis alors

P_+ ← P_+ +1

Sinon

Si avis alors

P_- ← P_- +1

Fsi

Fsi

Si P_+ / P < = ε alors archivé (avis)

Fin

Fin

1.4.PROCESSUS DE QUESTION :

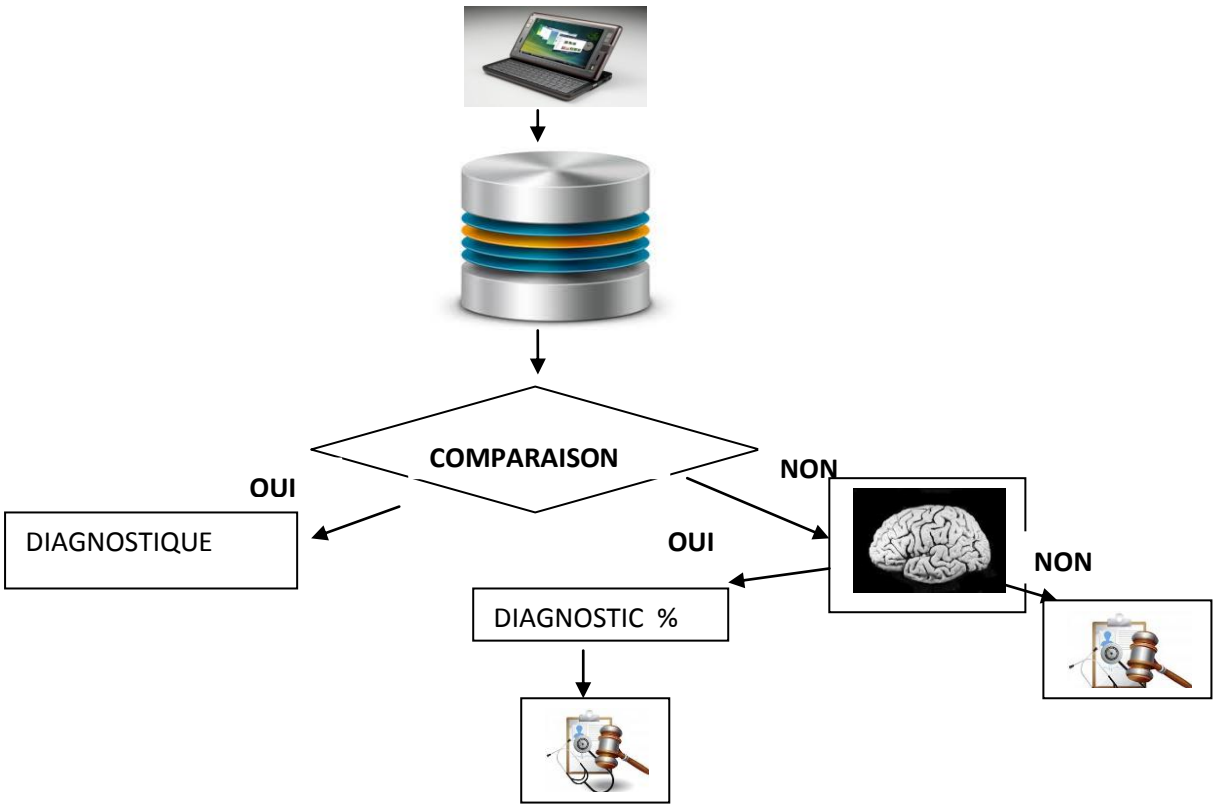

Figure n ° 3.3 : PROCESSUS DE QUESTION

Description :

Apres le processus de construction de base de donnée et de son intégration ,L'utilisateur pose sa question a travers l'interface utilisateur .

COMPARAISON **:** entre les données de la question et les donnes de la base de donnée .

SI des cas similaires alors un diagnostic sera affiché

Sinon le processus de l'apprentissage automatique sera déclenché pour traité les données .

Si des cas a peut prés simillaire alors pourcentage et par la suite passé a l'interface du medecin pour donné son avis .

Si non directement l'avis du patient .

2. PROPOSITION D'UN MECANISME DE DEPLOIEMENT

2. 1. INTERFACE UTILISATEUR

Le dialogue entre l'usager et le Moteur de recherche est réalisé par l'intermédiaire de l'interface utilisateur , C'est une application écrit en java qui contient des champs ou l'utilisateur va remplir

2. 2. BASE DE DONNEE MEDICALE

a.definition

C'est une base de donnée qui contient Les **données de santé** qui sont toutes les données médicales sur une personne, d'un groupe de personnes. Ces données sont utilisées pour le suivi et l'évaluation des systèmes d'aide a la decision et par la suite la prédiction .

b.proposition

Notre proposition de la base de donnée se compose de plusieurs champs plusieurs classe de maladies et chaque maladie a ses propres symptômes , aussi un nombres de patients .

Ces donnees sont collecter par plusieurs sources comme fichiers EXEL , fichier TEXTE , ACCESS...

2. 3 . APPRENTISSAGE AUTOMATIQUE

a.definion : l'apprentissage automatique est basée sur la capacité des ordinateurs a apprendre sans être programmés .

b.proposition :

On va utilisé le logiciel WEKA decrit dans le chapitre 4

2.4.ALGORITHME UTILISE (KNN)

KNN est un algorithme qui stocke tous les cas disponibles et prédit la cible numérique en fonction d'une mesure de similarité (par exemple, les fonctions de distance). Historique de l'utilisation c'été pour l'estimation statistique et la reconnaissance de formes au début des années 1970 comme technique non paramétrique.

2.4.1.Algorithme

Une simple mise en œuvre de la régression KNN est de calculer la moyenne de la cible numérique des voisins les plus proches de K. Une autre approche utilise une moyenne pondérée par distance inverse des voisins les plus proches de K. La régression KNN utilise les mêmes fonctions de distance que la classification KNN.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Les trois mesures à distance ci-dessus ne sont valables que pour les variables continues. Dans le cas des variables catégoriques, vous devez utiliser la distance Hamming, qui est une mesure du nombre d'instances dans lesquelles les symboles correspondants sont différents en deux chaînes de même longueur.

Hamming Distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

Choisir la valeur optimale pour K est mieux fait en inspectant d'abord les données. En général, une grande valeur K est plus précise car elle réduit le bruit global; Cependant, le compromis est que les limites

distinctes dans l'espace caractéristique sont floues. La validation croisée est une autre façon de déterminer rétrospectivement une bonne valeur K en utilisant un ensemble de données indépendant pour valider votre valeur K. Le K optimal pour la plupart des jeux de données est de 10 ou plus. Cela produit de bien meilleurs résultats que 1-NN.

Exemple :

Considérez les données suivantes concernant l'indice des prix de la maison ou l'IPH. L'âge et le prêt sont deux variables numériques (prédicteurs) et HPI est la cible numérique.

Age	Loan	House Price Index	Distance	
25	$40,000	135	102000	
35	$60,000	256	82000	
45	$80,000	231	62000	
20	$20,000	267	122000	
35	$120,000	139	22000	2
52	$18,000	150	124000	
23	$95,000	127	47000	
40	$62,000	216	80000	
60	$100,000	139	42000	3
48	$220,000	250	78000	
33	$150,000	264	8000	1
48	**$142,000**	**?**		

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Nous pouvons maintenant utiliser l'ensemble de formation pour classer un cas inconnu (Âge = 33 et Prêt = 150 000 $) en utilisant la distance Euclidienne. Si K = 1, le voisin le plus proche est le dernier cas dans l'ensemble d'entraînement avec HPI = 264.

D = Sqrt [(48-33) ^ 2 + (142000-150000) ^ 2] = 8000.01 >> HPI = 264

En ayant K = 3, la prédiction pour HPI est égale à la moyenne de HPI pour les trois premiers voisins.

CONCLUSION :

Dans ce chapitre, nous avons proposé quelques architectures sur le déroulement de travail lié a notre thème de recherche, avec quelques propositions de mécanismes de déploiement .

Introduction :

Concevoir un logiciel est un processus créatif qui demande un certain savoir-faire. Une bonne conception est la clé du développement d'un logiciel efficace. Un système bien conçu est facile à réaliser et à maintenir, facile à comprendre et fiable, il est de qualité. La phase de conception et la phase la plus cruciale du processus du développement d'un logiciel.

Dans ce chapitre, nous présenterons et décrirons les différents outils de développement (langages) que nous avons utilisés dans la réalisation de notre application.

1. Apache Hadoop

1.1.Définition :

Hadoop fait référence à un écosystème de logiciels open source qui compose une infrastructure de traitement, de stockage et d'analyse distribués des jeux de données volumineuses sur des clusters d'ordinateurs.

Il était le projet open source d'origine pour le traitement des données volumineuses. Il fut suivi du développement de logiciels et d'utilitaires connexes considérés comme faisant partie intégrante de la pile de technologies Hadoop, notamment ***Apache Hive***, ***Apache HBase***, ***Apache Spark***, ***Apache Kafka*** et bien davantage **[38]** .

Composé des Fon.ctions et des procédures écrit en Java (Multiplateformes pour les environnements de systèmes d'exploitation Unix- Windows -.) , il est destiné à faciliter la création d'applications distribuées (au niveau du stockage des données et de leur traitement) permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données.

Hadoop a 3 grandes caractéristiques **:[39]**

- **Économique** : Hadoop permet aux entreprises de libérer toute la valeur de leurs données en utilisant des serveurs peu onéreux.
- **Flexible** : Hadoop permet de stocker de manière extensible tous types de données. Les données peuvent être non structurées et ne suivre aucun schéma structurées (PDF, MP3, base de données, etc.) grâce à son système de fichier HTDFS « HadoopDistributed File System ». Les utilisateurs peuvent transférer leurs données vers Hadoop sans avoir besoin de les reformater.
- **Tolère les pannes**: les données sont répliquées à travers le cluster[1] afin qu'elles soient facilement récupérables suite à une défaillance du disque, du nœud ou du bloc.

[1]Il représente un ensemble d'ordinateurs effectuant la même tâche pour assurer la disponibilité, la répartition de la charge de travail et une meilleure gestion des ressources.

1. 2.Avantages et inconvénients

Le principal atout des clusters Hadoop est leur adéquation à l'analyse de gros volumes de données. Le Big Data est le plus souvent non structuré et largement distribué**.[39]**

- Il garantit la disponibilité et la durabilité des données, par réplication. C'est une approche logicielle à contre-courant des solutions matérielles traditionnelles (raid, san, …)

- Il garantit une scalabilité linéaire des capacités de stockage et de traitement par simple ajout de machine. Stockage et traitement sont distribués et co-localisés

- Il apporte une capacité à traiter des données peu ou pas structurées

Malgré ses nombreux avantages, la solution des clusters Hadoop ne convient pas aux besoins d'analyse de données de toutes les organisations**.[39]**

- Pour les petits volumes de données, ses bénéfices seront inexistants, même s'il est besoin d'une analyse poussée de ces dernières.

- Un autre inconvénient des clusters Hadoop réside dans le principe même de la solution de clustering, qui suppose que les données puissent être fractionnées et analysées par des processus parallèles exécutés sur des noeuds de cluster distinct. Si l'analyse ne peut pas se dérouler dans un environnement de traitement parallèle, un cluster Hadoop n'est tout simplement pas le bon outil pour ce travail.

- Mais le plus grand obstacle à l'utilisation d'un cluster Hadoop reste probablement la courbe d'apprentissage considérable associée à la construction, au fonctionnement et à la prise en charge de la solution. À moins que votre service informatique ne compte un spécialiste de cette technologie dans ses rangs, il vous faudra quelque temps pour apprendre à construire le cluster et réaliser l'analyse requise des données.

1. 3.Architecture Apache Hadoop

Hadoop est une collection d'applications logicielles constitué de plusieurs composants : **[**Brien Posey , 2015**]**

1.3.1. Système de fichiers distribué HDFS

HDFS est le système de fichiers par défaut utilisé par Hadoop pour découper les données et répliquer ces pièces de données en **triple** sur plusieurs nœuds.

Ce système dispose de deux modes d'implémentations :

- **Distribué** : permettant la triple réplication.

- **Pseudo-distribué** : utilisé comme moyen de test, ce mode est implémenté sur une seule machine d'un seul nœud.

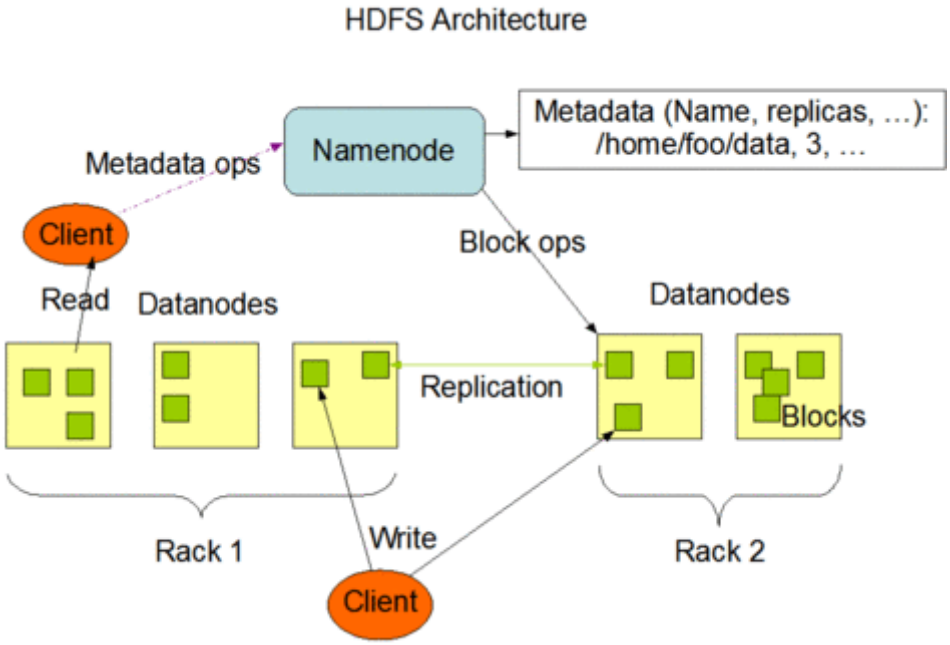

Figure 4.1 : Architecture HDFS **[**Brien Posey , 2015**]**

HDFS utilise une architecture maître/esclave. Dans un cluster HDFS, on retrouve un NameNode, qui est le serveur maitre qui gère le Namespace du système de fichiers et qui régule les accès aux fichiers par les clients.

Cependant, On retrouve seulement un DataNode par nœud dans un Cluster, qui eux gèrent le stockage des données dans le nœud dans lesquels ils fonctionnent. En interne, HDFS découpe un fichier en un ou plusieurs blocs qui sont stockés dans un set de Datanodes.

Le NameNode exécute les opérations telles que l'ouverture, fermeture, renommage des fichiers et dossiers, il est également en charge de l'attribution des blocs de fichiers aux Datanodes.

Le DataNode, lui est en charge de répondre aux requêtes de lecture et écriture des clients. Suite à la demande du NameNode, il peut aussi effectuer des opérations de réplications entre les nœuds.

1.3.2 .MapReduce

MapReduce est un Framework Java qui fournit une API pour écrire des applications qui vont pouvoir traiter de larges quantités de données en parallèle sur des clusters**.[40]**

La fonction **Map** permet de décomposer une requête importante en un ensemble de requêtes plus petites qui produisent chacune un sous ensemble du résultat final.

La fonction **Reduce** permet d'assembler tous ces résultats en un résultat final.

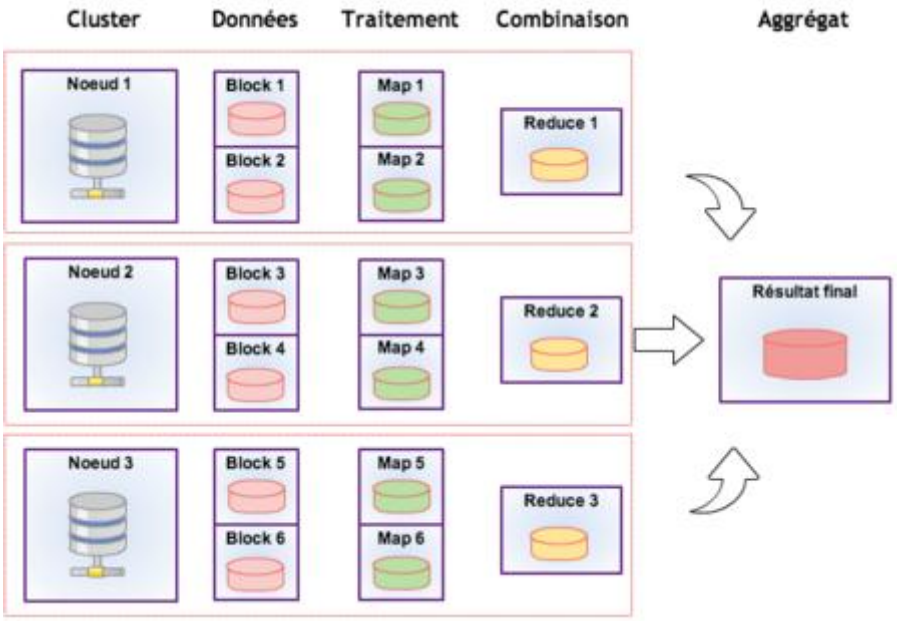

Figure 4.2 : Schéma explicatif MapReduce [40]

1.3.3. Présentation d'une architecture distribuée dans le cadre d'Hadoop

Il faut tout d'abord savoir qu'une architecture Hadoop est basée sur les deux principaux rôles maître / esclave. Des sous-rôles relatifs au système de fichiers et à l'exécution de tâhces distribuées sont associés à chaque machine.

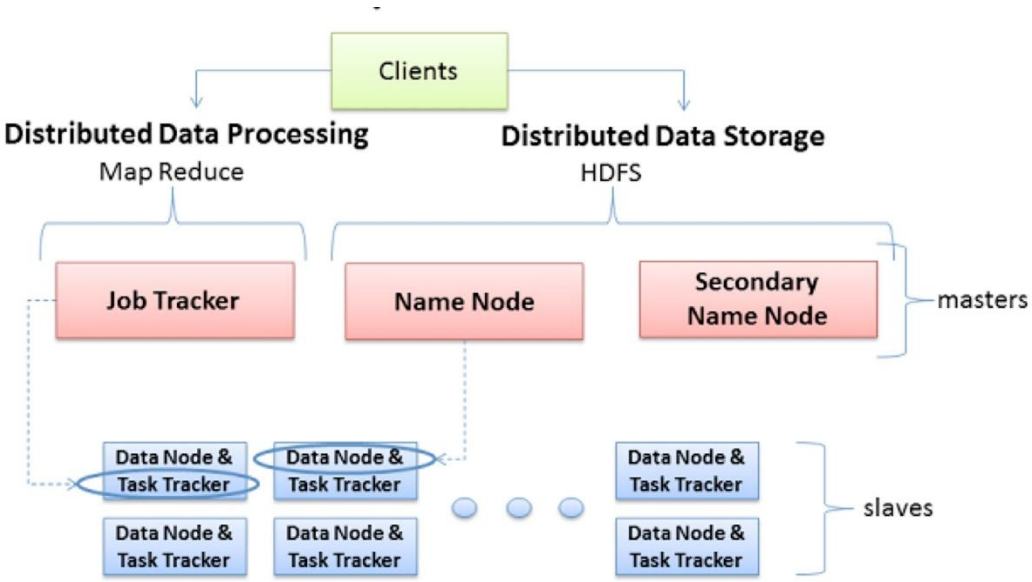

Figure 4.3 : Schéma d'architecture Hadoop représentant les principaux rôles des machines

Dans le cadre des machines maîtres, trois principaux sous-rôles sont associés :

- **JobTracker** : Il s'agit de la responsabilité pour une machine maître de lancer des tâches distribuées, en coordonnant les esclaves. Le JobTracker, planifie les exécutions, gère l'état des machines esclaves et aggrège les résultats de calculs dans le cadre de MapReduce.

- **NameNode** : Le rôle de NameNode intervient dans le fonctionnement du système de fichiers distribué HDFS. Une machine maître associée à ce rôle a pour responsabilité de répartir les données sur les machines esclaves et de gérer l'espace de noms du cluster. Elle contient les meta-datas lu permettant de savoir sur quelle machine chauqe fichier est hébergé.

- **SecondaryNameNode** : Ce rôle intervient dans le cadre de la redondance du NameNode. Généralement assumé par une autre machine physique que le NameNode, il permet, en cas de panne de celui-ci, la continuité de fonctionnement du cluster Hadoop.

Les machines esclaves ont pour chacune d'elles deux sous-rôles qui leurs sont associées :

- **TaskTracker** : Il s'agit du rôle permettant à un esclave d'exécuter une tâche MapReduce sur les données qu'elle héberge. Le TaskTracker est piloté par le JobTracker d'une machine maître qui lui envoie la tâche à exécuter.

- **DataNode** : Comme son nom l'indique, il s'agit d'un noeud de l'architecture hébergeant une partie des données du cluster. Les noeuds de données sont généralement répliqués dans le cadre d'une architecture Hadoop afin d'assurer la haute disponibilité des données.

Lorsqu'un client à besoin d'accéder à une donnée ou d'exécuter une tâche distribuée, elle passe par la machine maître jouant les rôles de JobTracker et de NameNode.

Maintenant que nous avons vu globalement comment s'articule une architecture de ce type, nous allons citer quelques autres outilles de Hadoop qui sont :

- **HBase :** est un système de gestion de base de données non relationnelle, distribuée et orientée colonnes, prenant pour modèle Big Table de Google
- **Hive :** est un système d'entrepôt de données facilitant l'agrégation des données, les requêtes ad hoc, et l'analyse de grands ensembles de données stockées dans les systèmes de fichiers compatibles Hadoop. Hive dispose d'un langage de type SQL appelé HiveQL.
- **HCatalog :**est une couche de métalangage permettant d'attaquer les données HDFS via des schémas de type tables de données en lecture/écriture.
- **Pig :** est une plate-forme d'analyse de vastes ensembles de données. Pig comprend un langage de haut niveau gérant la parallélisassions des traitements d'analyse.
- **Oozie :** est un outil de workflow dont l'objectif est de simplifier la coordination et la séquence de différents traitements. Le système permet aux utilisateurs de définir des actions et les dépendances entre ces actions.
- **ZooKeeper :**est un service centralisé pour gérer les informations de configuration, de nommage, et assurer la synchronisation des différents serveurs via un cluster. Tous les services pris en charge par ZooKeeper peuvent être utilisés sous une forme ou une autre par les applications distribuées. [**41**]

Et aussi d'autres outilles comme Ambari , Avro , Cassandra …

1.4.HBase

Appache HBase est un open source , version distribuée , non relationnel , orienté colonne . Notre Base de donnée médicale qui apparaitre dans notre architecture , c'est bien HBase qui lire /écrire données reçu par/dans HDFS.Le but de ce projet si veut dieu c'est l'hébergement de tableaux de taille large (bilions de lignes et millions de colonnes .

Le principe le meme que celui d'Hadoop . que les noms sont différent

Architecture	**Hadoop**	**HBase**
Serveur- maitre	**Namenode**	**regionserver**
Serveur - esclave	**Datanode**	**Region**
Fichier de stockage	**Datailef**	**storefile**

- Table : collection de rangés
- Rangé: collection de famille de colonnes
- Famille de Colonne : collection de colonnes (un ou plusieurs)
- Colonne : collection de valeurs.

2. Environnement de développement :

2.1. Cloudera Virtual Machine

Ou nommé aussi Cloudera distribution de apache Hadoop et les outiles reliée en Open source , comme Cloudera Impala et cloudera de recherche .

Cloudera est passé par plusieurs version(le dernier CDH 5.7 ,April 2016), on a utilisé CDH 5.3.0 qui travaille avec CentOS Operating system , l'environnement UNIX

2.2. Langage de programmation choisi : JAVA

Java est un langage de programmation et une plate-forme informatique qui ont été créés par Sun Microsystems en 1995. Beaucoup d'applications et de sites Web ne fonctionnent pas si Java n'est pas installé et leur nombre ne cesse de croître chaque jour. Java est rapide, sécurisé et fiable. Des ordinateurs portables aux centres de données, des consoles de jeux aux superordinateurs scientifiques, des téléphones portables à Internet, la technologie Java est présente sur tous les fronts .

Il permet de créer des logiciels compatibles avec de nombreux systèmes d'exploitation (Windows, Linux, Macintosh, Solaris). Java donne aussi la possibilité de développer des programmes pour téléphones portables et assistants personnels. Enfin, ce langage peut-être utilisé sur internet pour des petites applications intégrées à la page web (applet) ou encore comme langage serveur (jsp).

2. 2.1. L'environnement IDE Eclips :

Les IDE sont des programmes qui regroupent un ensemble d'outils pour le développement de logiciels.

De façon générale, un IDE contient un éditeur de texte, un compilateur, des outils automatiques de fabrication, et très souvent un débogueur.

Il existe des IDE pour de nombreux langages, cependant il est très courant qu'un IDE soit conçu pour un seul langage de programmation.

Il est également possible qu'un IDE dispose d'un système de gestion de versions et différents outils pour faciliter la création des interfaces graphiques.

1.2.2. Fonctionnalités requises et besoins :

Afin d'être le plus efficace possible lors du développement de l'application SubJects, il est nécessaire d'énumérer les différents critères à partir desquels sera réalisé le choix de l'IDE.

La solution utilisée devra, au mieux, répondre aux besoins suivants :

- Possibilité de déploiement d'applications Web ;
- Rapidité de fonctionnement ;
- Léger au lancement ;
- Compilation possible du projet ;
- Gestion de plusieurs projets ;
- Débogueur précis ;
- Visualisation aisée de la JavaDoc ;
- Interfaçage avec un gestionnaire de versions ;
- Logiciel simple et facilité d'utilisation.

2.2.3.Eclips :

Nous avons utilisé l'environnement de développement intégré IDEE, en peut lui ajouter le plugins , d'abord conçu pour le langage Java .mais peut etre un environnement de développement pour de nombreux autres langages de programmation (C/C++, Python, PHP, Ruby, …).Ici notre langage est java .

2.3.weka :

Weka est une collection d'algorithmes d'apprentissage machine pour les tâches d'exploration de données.
On a passé par une etude comparatif entres plusieurs algorithmes qui sont utilisé dans le domaine médicale et on a choisie algorithme KNN.

2.4.Base de donnée médicale :

La base de donnée se compose de plusieurs lignes et plusieurs colonnes sur des maladies et ses symptômes approprié .
Ces données sont extrait de différents sources comme EXCEL ,fichier TEXTE , ACCESS …..

Les données peuvent être structuré ou non structuré comme des images des radio scanner ou URM qui sont traiter en futur .

3. Présentation de quelques captures :

3.1. Bureau de Cloudera

Figure n° 4.4 : Bureau de Cloudera

3.2. Interface utilisateur

3.2.1.Créer un nouveau utilisateur :

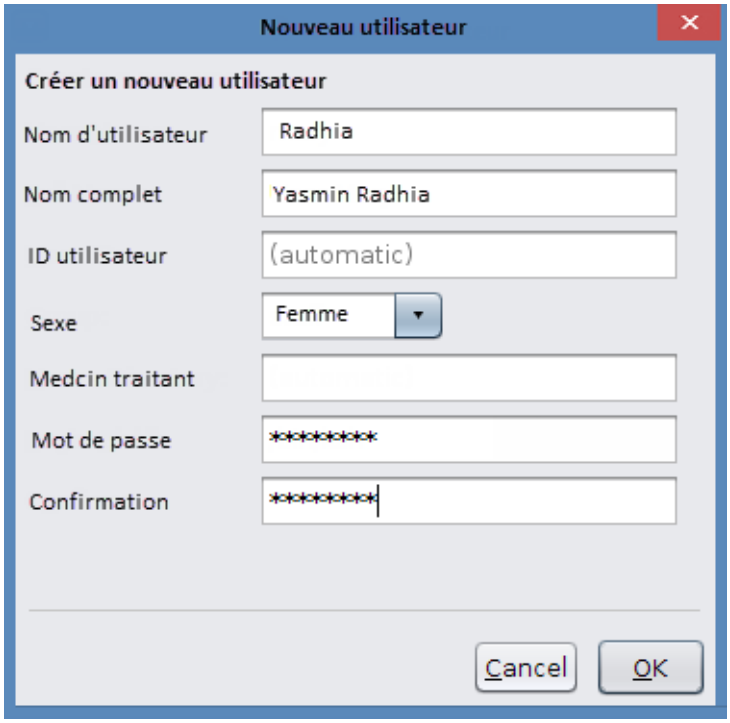

Figure 4.5 : Créer un nouveau utilisateur

3.2.2. Menu d'authentification

Figure 4.6 : Menu d'authentification

3.3.Menu de questionnaire

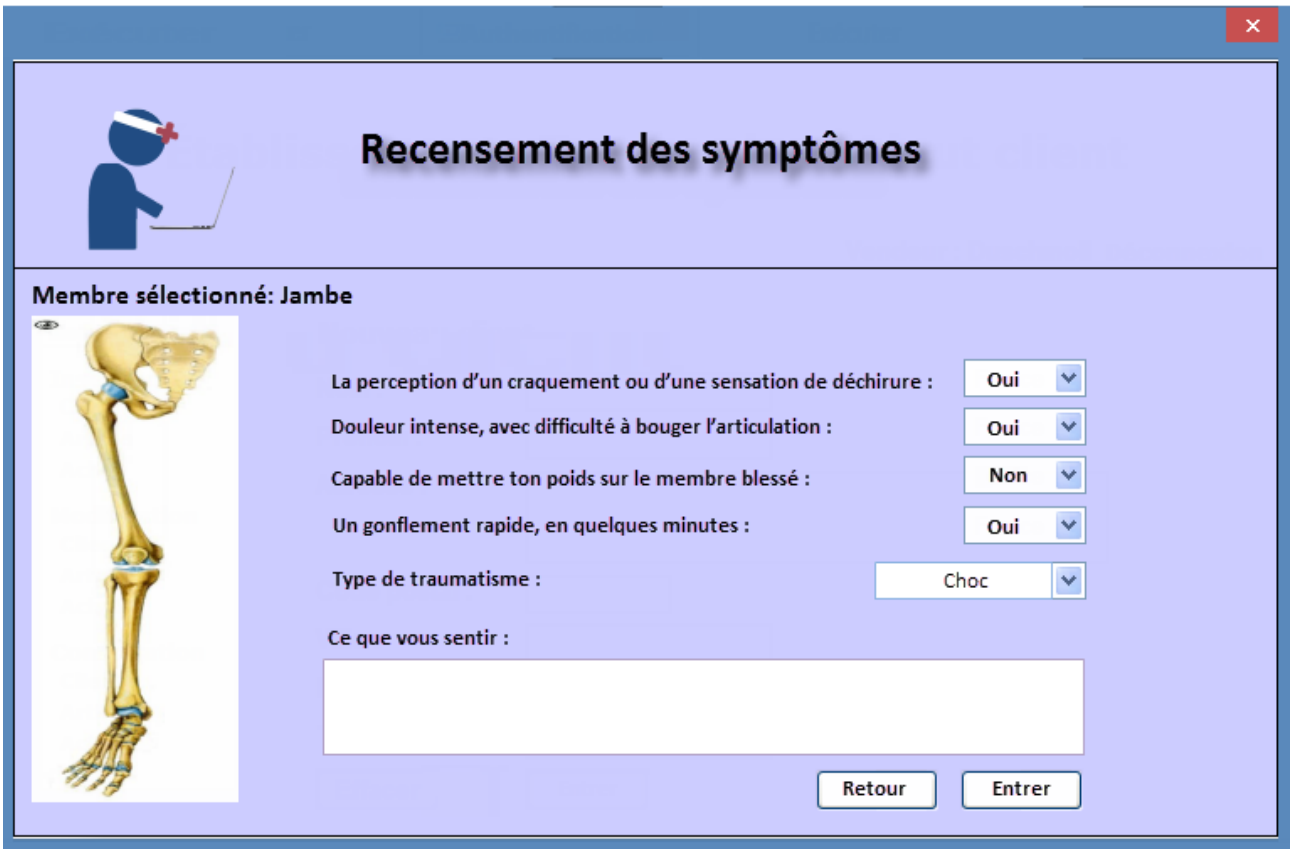

Figure n° 4.7 : Questionnaire

Menu 3 : DIAGNOSTIQUE

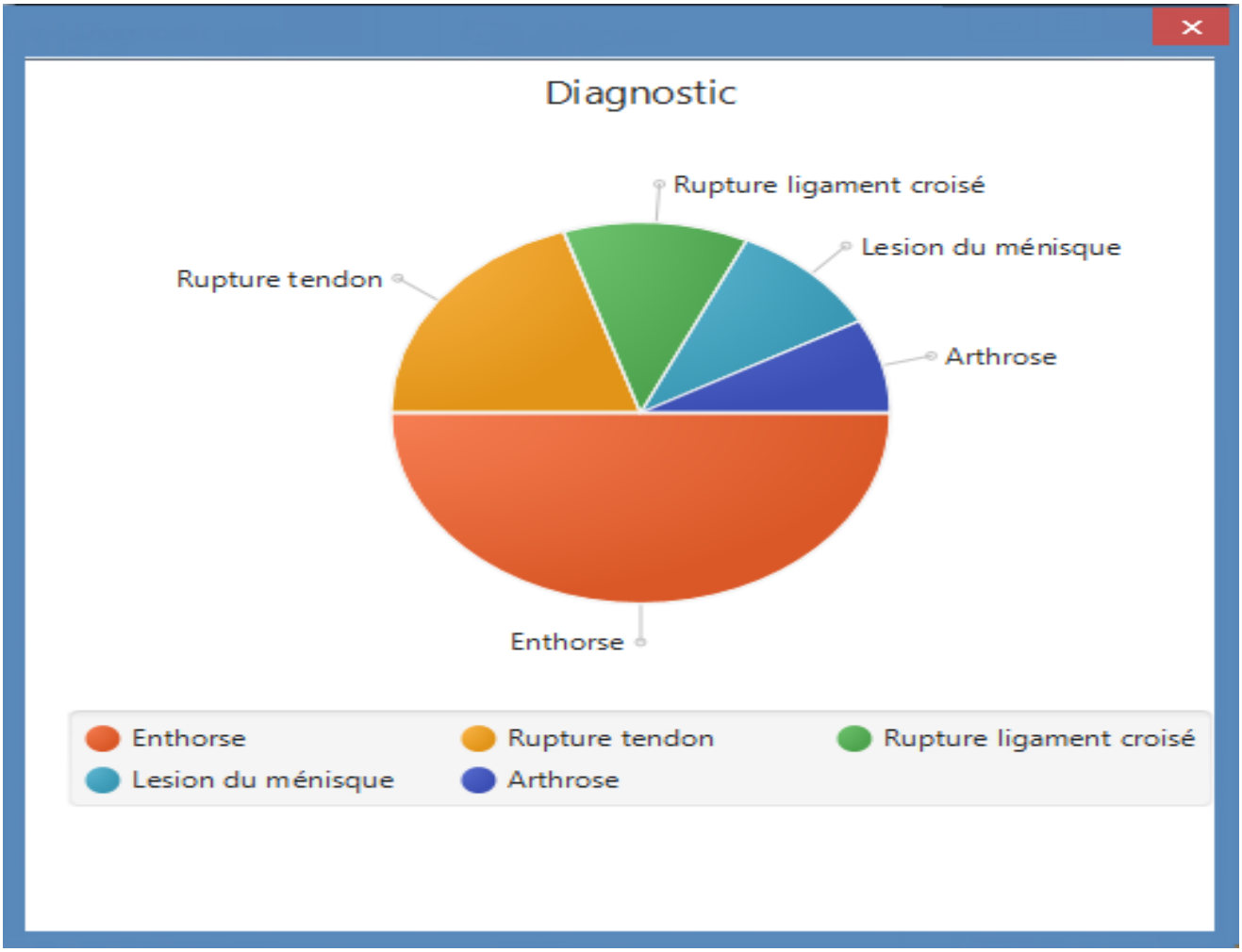

Figure 4.8 : Diagnostique

Conclusion :

Dans ce chapitre, nous avons présenté l'environnement de développement matériel et l'environnement logiciel avec lesquels ce projet a été réalisé. On a installer cloudera sous unix pour présenté le Hadoop car c’ est la distribution la plus populaire,et on propose de travailler avec eclipse IDE sous cloudera dont il offre les principales fonctionnalités nécessaires du langage JAVA pour realiser l’interface utilisateur .

Références bibliographiques

N°	
[Christophe Brasseur, 2016]	Enjeux et usages du Big Data, 2e edition,2016 La voisier, Paris. https://www.lavoisier.fr/.../9782746247581_enjeux-et-usages-du-big-data-2e-ed-colle
[Pierre Delort, 2015]	Le Big Data 2015, https://www.puf.com/content/Le_Big_Data
[Tiffany ETCHEGORRY, 2016]	MASTER TOURISME ET HÔTELLERIE, Parcours « Management en Hôtellerie-Restauration», MÉMOIRE DE PREMIÈRE ANNÉE, LE BIG DATA DANS L'HÔTELLERIE FRANÇAISE
[Angeline KONE, 2013]	: Memoire Online > Rapports de stage,Big data (rapport de stage),par Angeline KONE ,INSA LYON - Mastère spécialisé SI 2013 http://www.memoireonline.com/05/14/8890/m_Big-data-rapport-de-stage9.html
[MERCATOR ,2014]	lexique-publicite-definition-big-data,www.mercator-publicitor.fr/lexique-publicite-definition-big-data,2014
[Bruno TEBOUL et Jean-Marie BOUCHER ,2013]	Bruno TEBOUL + Jean-Marie BOUCHER, L'Absolu Marketing: Web 3.0, Big Data, Neuromarketing, https://www.amazon.fr/Tout-savoir-LAbsolu-Marketing-Neuromarketing/dp/2918866628, 28 mars 2013
[Matteo Di Maglie, 2012]	Matteo Di Maglie, Adoption d'une solution NoSQL dans l'entreprise, Travail de Bachelor réalisé en vue de l'obtention du Bachelor HES, Carouge, 12 septembre 2012 Haute École de Gestion de Genève (HEG-GE), https://doc.rero.ch/record/31286/files/TDIG_82.pdf , *12 septembre 2012*
[Khaled Tannir, 2015]	Khaled Tannir . NoSQL. (Montréal), https://www.linkedin.com/in/tannir?ppe=1, 01 Décembre 2015.
[Brigitte Séroussi et al, 2015]	Systèmes informatiques d'aide à la décision en médecine: panorama des approches utilisant les données et les connaissances, 2015, https://hal.inria.fr/hal-01100249/document
[1]	Orienté colonne, https://basesdedonnees.wordpress.com/nosql/oriente-colonne/
[2]	Margaret Rouse , Base de données orientée graphes, http://www.lemagit.fr/definition/Base-de-donnees-orientee-graphes,
[3]	RICHARDIN Anaïs. La *data*, pétrole de l'économie numérique. Stratégies, 13 mai 2014,. Disponible sur http://urlz.fr/3iVo. http://www.strategies.fr/etudes-tendances/tendances/235848W/la-data-petrole-de-l-economie-numerique.html
[SZADKOWSKI Michaël et al, 2013]	SZADKOWSKI Michaël, LELOUP Damien. Prism, Snowden, surveillance : 7 questions pour tout comprendre. Le Monde Technologies 2013, [en ligne]. Disponible sur https://lc.cx/4mEm
[Jérôme Béranger, 2016]	Lors de la conférence *Les enjeux juridiques qu'impliquent le partage et les usages des Big data* dans le cadre de la Grande Conf' Sciences humaines et sociales et *Big Data*, Toulouse, le 21-02-2016.

[legifrance, 2013]	**Legifrance. *Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*, 2013, 76 p. [en ligne). Disponible sur https://lc.cx/4mrB. https://www.cnil.fr/fr/loi-78-17-du-6-janvier-1978-modifiee**
[Agenda-medical, 2015]	**2015 Agenda-medical.fr, http://www.agenda-medical.fr/informatique-medicale-l59.php**
[CHEIKH NGOM, 2016]	**Doctideo : plateforme de partage d'informations médicales créée par des médecins, 2016, https://bonjouridee.com/doctideo/**
N°	
(Anonymous, 1992)	**Introduction à l' 'Evidence-Based Medicine' (EBM) (Anonymous, 1992) , http://www.ebm.lib.ulg.ac.be/prostate/ebm.htm**
DL Sackett - 1996	**Evidence based medicine: what it is and what it isn't. - NCBI https://www.ncbi.nlm.nih.gov/pubmed/8555924 - Traduire cette page de DL Sackett - 1996**
(Sackett- 1994)	**Introduction à l' 'Evidence-Based Medicine' (EBM) (Sackett- 1994) , http://www.ebm.lib.ulg.ac.be/prostate/ebm.htm**
[ALAN R.Feinstein et al]	**Problems in the « evidence » of « evidence based medecine » ALAN R.Feinstein , MD ,Ralph I.Horwitz,MD ,New Haven,Connecticut, https://www.researchgate.net/publication/13801263_Problems_in_the_Evidence_of_Evidence-Based_Medicine**
[Davidoff F,1999]	**Davidoff F. ; *In the teeth of the evidence: the curious case of evidence-based medicine* ; Mt Sinai J Med, mars 1999 ; 66(2) : 75-83 , https://books.google.dz/books?id=T5fmUbge2kUC&pg=PA113&lpg=PA113&dq=Feinstein+et+Horwitz,+1997&source=bl&ots=nru-jpfr-3&sig=MUwWH16P3TA2-HJuIicynGawZ0A&hl=fr&sa=X&ved=0ahUKEwiL8ovb4t3TAhWKuBoKHSINDecQ6AEIMzAD#v=onepage&q=Feinstein%20et%20Horwitz%2C%201997&f=false**
[ReachG, 2012]	**ReachG."L'inertie Clinique.Une critique de la raison médicale". Springer, 2012 www.springer.com/br/book/9782817803128 , https://www.google.dz/search?tbm=bks&hl=fr&q=ReachG.%22L%E2%80%99inertie+Clinique.Une+critique+de+la+raison+m%C3%A9dicale%22.+Springer%2C+2012.**
[4]	**URL: http://www.update-software.com/cochrane/**
[5]	**(http://www.eortc.be/)**
[6]	**URL: http://www.update-software.com/cancer/default.htm**
[7]	**URL: http://www.acponline.org/journals/ebm/pastiss.htm?idx**
[8]	**URL: http://gateway.ovid.com/autologin.html**

[9]	**URL: http://nhscrd.york.ac.uk/welcome.html**
[10]	**Evidence-Based Medicine (ULg) ,www.ebm.lib.ulg.ac.be/prostate/ebm.htm , de C Delvenne - 1999 , Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC . Developing optimal search strategies for detecting clinically sound studies in MEDLINE. Journal of the American Medical Informatics Association. 1994;1:447-58.**
[11]	**The development of PubMed search strategies for patient preferences for treatment outcomes ,Ralph van Hoorn,[1] Wietske Kievit,[1] Andrew Booth,[2] Kati Mozygemba,[3,7] Kristin Bakke Lysdahl,[4] Pietro Refolo,[5] Dario Sacchini,[5] Ansgar Gerhardus,[3,7] Gert Jan van der Wilt,[6] and Marcia Tummers[1]** **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4966584/**
[EJMullen, DL Streiner, 2006]	**The evidence for and against evidence-based practice** **EJ Mullen, DL Streiner - … of evidence-based social work practice, 2006 - books.google.com**
[Sackett DL,2000]	**Evidence based medicine: what it is and what it isn't ,Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. Evidence-Based Medicine: how to practice and teach EBM. Second ed. London: Churchill Livingstone; 2000.** **, https://www.cebma.org/wp-content/uploads/Sackett-Evidence-Based-Medicine.pdf**
[(Rosenberg et Donald, 1995]	**(Rosenberg et Donald, 1995) , Foundations of Evidence-Based Social Work Practice** **https://books.google.dz/books?id=UWs2Vq1Jd-sC&pg=PA68&lpg=PA68&dq=rosenberg+%26+donald+1995&source=bl&ots=A1vSO1y46p&sig=8QozZTgo6zCMBeBJoCqjNnfuj-c&hl=fr&sa=X&ved=0ahUKEwjvrPjK1N3TAhWG2hoKHQlxC-YQ6AEIQzAE#v=onepage&q=rosenberg%20%26%20donald%201995&f=false**
[12]	http://cybertim.timone.univ-mrs.fr/enseignement/doc-enseignement/statistiques/OutilsStatistiquesDM-RG/docpeda_fichier
[13]	http://www.saedsayad.com/data_mining.htm
[14]	**Data mining et statistique décisionnelle: l'intelligence dans les bases de données, https://books.google.dz/books?id=XHSPydK5Qr0C&pg=PA1&lpg=PA1&dq=les+statistiques+en+data+mining+en+medecine&source=bl&ots=KHr3TSiHVp&sig=cTcJDdtEy94FYCPYwrKkcgYQQb4&hl=fr&sa=X&ved=0ahUKEwjOt_3Y38_TAhWiJcAKHXegDhEQ6AEIOzAE#v=onepage&q=les%20statistiques%20en%20data%20mining%20en%20medecine&f=false? EDITION 2005**
[15]	http://www.commentcamarche.net/contents/104-bases-de-donnees-introduction
[16]	**(Stair Ralph M.) , Principles of Information Systems, http://www.christian.braesch.fr/biblio-livre/principles-of-information-systems , 2011**
[17]	Les outils d'ETL (extraction, transformation et chargement) automatisent totalement la création, la maintenance et l'extension des entrepôts de données, data marts, micro marts et magasins de données opérationnelles. - See more at : http://www.informationbuilders.fr/outils-etl#sthash.o6PTd3Th.dpuf

[18]	https://blog.developpez.com/jmalkovich/p8718/modelisation/modele_en_etoile_ou_en_flocons
[19]	**christian.braesch ,Les modèles d'un entrepôt de données, http://www.christian.braesch.fr/page/les-modeles-dun-entrepot-de-donnees**
[20]	Architecture d'un entrepôt de données, http://www.christian.braesch.fr/page/architecture-dun-entrepot-de-donnees
[21]	Une petite histoire du Machine Learning **October 28, 2015 , site https://www.quantmetry.com/single-post/2015/10/28/Une-petite-histoire-du-Machine-Learning**
[22]	Intelligence artificielle: la médecine en mutation - Planete sante https://www.planetesante.ch/...et.../Intelligence-artificielle-la-medecine-en-mutation ,5 janv. 2016
[23]	Japon : une intelligence artificielle diagnostique des maladies rares fr.ubergizmo.com › Life ,3 août 2016
[24]	Les systèmes d'aide à la décision médicale | Cairn.info https://www.cairn.info/revue-les-cahiers-du-numerique-2001-2-page-125.htm ,de M Cléret - 2001
[Garg AX et al , 2005**]**	Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerizedclinicaldecision support systems on practitioner performance and patient outcomes. JAMA 2005;293(10) : 1223-38.
[Kawamoto K, 2005**]**	Kawamoto K, Houlihan CA, Balas EA, et al. Improvingclinical practice usingclinical decision support systems: asystematicreview of trials to identifyfeaturescritical to success. BMJ 2005 Apr;330(7494) : 765.
[25]	**https://www.cairn.info/revue-les-cahiers-du-numerique-2001-2-page-125.htm**
[SchadracKANDE KANUMAMBIDI ,2009]	Réalisation d'un système expert pour la thérapeutique et le diagnostic des maladies de la tuberculose *par* Schadrac KANDE KANUMUAMBIDI ,Université de Notre Dame du Kasayi - Licence 2009 ,http://www.memoireonline.com/09/10/3851/m_Realisation-dun-systeme-expert-pour-la-therapeutique-et-le-diagnostic-des-maladies-de-la-tuber19.html
[26]	chapitre i systeme d'aide a la decision medicale - dspace dspace.univ-tlemcen.dz/bitstream/112/8301/3/CHAPITRE1.pdf
[27]	L'apprentissage automatique : pas à pas ! | binaire , naire.blog.lemonde.fr/2015/06/23/pasapas/ , 23 juin 2015
[28]	Intelligence artificielle : Quelles différences entre le machine learning ... www.lemondeinformatique.fr › Toute l'actualité › Logiciel › Machine Learning ,29 juin 2016
[K. D. Althoff ET AL , 1995]	K. D. Althoff, R. Barletta, M. Manago, and E. Auriol, "A Review of Industrial Case-Based Reasoning Tools," AI Intelligence, Oxford, 1995.

[29]	Institut des Neurosciences Paris-Saclay - I2BM - Neurospin neuro-psi.cnrs.fr/spip.php?article784
[30]	Comment mieux soigner les maladies des yeux grâce au machine https://rslnmag.fr/innovation/yeux-machine-learning-cloud/6 déc. 2016
[31]	L'analyse vocale et les algorithmes au chevet des patients https://rslnmag.fr/innovation/analyse-vocale-algorithmes-patients/ , 3 févr. 2017
[32]	Voici comment reconnaître une crise cardiaque un mois avant qu'elle arrive , http://sain-et-naturel.com/voici-comment-reconnaitre-une-crise-cardiaque-un-mois-avant-quelle-arrive.html 19 juin 2015
[33]	Apprentissage automatique pour trouver des modèles cachés ou des structures intrinsèques dans les données , https://www.mathworks.com/discovery/apprentissage-non-supervise.html?action=changeCountry&domain_lang=fr&s_tid=gn_loc_drop
[34]	**Arbres de Décision Ricco RAKOTOMALALA Laboratoire ERIC Université Lumière Lyon 2 , https://www.rocq.inria.fr/axis/modulad/archives/numero-33/tutorial-rakotomalala-33/rakotomalala-33-tutorial.pdf, 2005**
[35]	mémoire fin d'étude ***dspace.univ-tlemcen. dz/bitstream /112 /1045/4/Memoire.pdf* , de M KOUDRI - 2011** - ID3 et C4.5 moins bons (d´etection des maladies cardiaques). - C4.5 produit plus d'erreurs de prédiction que ID3, mais ces taux restent néanmoins très faibles . Dataminig - C4.5 - DBSCAN **devezeb.free.fr/downloads/ecrits/datamining.pdf**
[36]	**]**Régression linéaire math.unice.fr/~diener/MpB2009-2010/REG.pdf
[37]	**CHAMI Djazia , Une plate forme orientée agent pour le data mining, djazia. Chami .pdf , 2010**
[38]	Santiago Cortez ,Présentation de l'écosystème Hadoop sur Azure HDInsight , https://docs.microsoft.com/fr-fr/azure/hdinsight/hdinsight-hadoop-introduction#overview**,** 14/12/2016
[39]	*Mathieu Despriee* , votre-premier-projet-hadoop , http://blog.octo.com/votre-premier-projet-hadoop/, 01/03/2013
[Brien Posey , 2015**]**	Brien Posey , Clusters Hadoop : avantages et limites pour l'analyse des Big Data, http://www.lemagit.fr/article/Clusters-Hadoop-avantages-et-limites-pour-le-Big-Data , avril 2015
[40]	HDFS Architecture Guide , https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
[41]	Les composants Hadoop au crible http://www.journaldunet.com/developpeur/outils/les-solutions-du-big-data/les-composants-hadoop-au-crible.shtml