MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE
LA RECHERCHE SCIENTIFIQUE
UNIVERSITE ABDELHAMID IBN BADIS MOSTAGANEM

Faculté des Sciences Exactes et de l'Informatique
Département de Mathématiques et d'Informatique

L'extraction des connaissances Pour Le suivi épidémiologique

Etudiants : AMMOUR Zahia
BENGUEDDA Nour el yakine kheira

Encadreur : M. HAMAMI Dalila

Deuxième année Master Ingénierie des Systèmes d'Information

Année Universitaire 2012/ 2013

Sommaire

Listes des figures

Introduction générale

L'Épidémiologie en informatique occupe une place importante. Intiment liée à des disciplines telles que les statistiques ou la génomique, elle contribue à la mise en œuvre des projets de recherche et, de façon générale accompagne les différentes thématiques et activités scientifiques de l'unité. La collecte, le stockage, le traitement et l'analyse d'un grand nombre de données hétérogènes sont des étapes nécessaires à toute étude épidémiologique.

L'épidémiologie est indispensable pour comprendre la physiopathologie d'une maladie chez l'homme : l'omniprésence de la variabilité ne peut être surmontée ni par le recours à des modèles animaux ou *in vitro*, par construction réductionniste, ni par l'accès à des informations jusque-là inconnues grâce à la découverte de nouvelles technologies, telles que celles de la génomique : **[Tilstone, 2003]** la cause d'une maladie, même simple, ne peut être trouvée en lisant quelques puces ADN : on retrouva rapidement qu'en génomique aussi, il fallait employer la méthodologie de l'épidémiologie : mesurer les cofacteurs, travailler sur des groupes de taille suffisante constitués sans biais, etc., à tel point que des recommandations explicites furent données aux chercheurs pour qu'ils tiennent compte de cette variabilité incontournable d'où la naissance de la modélisation épidémiologique. **[Alain 2006]**

Les techniques de modélisation épidémiologique (les équations différentielles, les automates cellulaires, les systèmes multi-agents et les algèbres des processus), ont contribué à la compréhension du comportement des maladies infectieuses transmissible, de ses impacts possibles et de prévisions futures au sujet de sa propagation. Elles consistent en gros à construire des modèles qui sont utilisés dans la comparaison, la planification, la mise en œuvre, l'évaluation et l'optimisation de divers programmes de prévention, de thérapie et de contrôle, mais elles s'avèrent insuffisantes pour l'aide a décision du fait de la grandeur des bases de données, d'où l'orientation vers le les techniques du Data Minig.

Le Data Mining est un domaine pluridisciplinaire permettant, à partir d'une très importante quantité de données brutes, d'en extraire de façon automatique ou semi-automatique des informations cachées, pertinentes et inconnues auparavant en vue d'une utilisation industrielle ou opérationnelle de ce savoir. Il peut également mettre en avant les

associations et les tendances et donc servir d'outil de prévisions au service de l'organe décisionnel. **[Silvi, 2002]**

L'objectif de ce mémoire est d'une part comprendre le monde des systèmes épidémiologique et faire ressortir les plus importantes méthodes de modélisation et simulation, et d'autre part ; mettre en avant l'utilité du Data Mining pour l'extraction des informations pertinentes à cette étude, dans le but commun est de simplifier à la fois l'interaction entre développeur et expert, et la reproduction quasi-parfaite du système pour une meilleure prédiction et prise de décision.

Le présent manuscrit est structuré en quatre chapitres, organisé comme suit :

Le premier chapitre sera consacré à l'étude épidémiologique avec ses domaines d'applications, ses facteurs et son suivi. Nous commencerons par la définir, aborderons ensuite les types d'épidémiologie, ses domain d'applications, les indicateurs de performances , . Nous terminerons par définir la tuberculose qui fera l'objet de notre base de données de tests.

Le second chapitre étudiera ensuite les différentes techniques de modélisation de l'épidémiologie que l'expert utilise souvent pour modéliser ce problème et l'analyser.

Le troisième chapitre sera consacré au Data Mining, définition, domaine d'application, et d'autre part à sa relation avec l'épidémiologie, les tâches qu'il exécute en l'épidémiologie et les méthodes existantes.

Enfin, au quatrième chapitre nous présenterons les étapes d'implémentation de notre approche. Nous y détaillerons la réalisation de certaines fonctionnalités, les résultats d'expérimentation et leurs interprétations.
Et nous finirons par conclure et présenter des perspectives pour de futurs travaux.

Chapitre 1 L’épidémiologie humaine

Introduction

L'épidémiologie est au cœur des sciences fondamentales et appliquées qui traitent de la sante publique. D'emblée, elle envisage les données médicales par des méthodes mathématiques, en particulier les statistiques progressivement mettre en place un panel d'outils et les infrastructures nécessaires à la gestion de divers systèmes d'information complexes. Afin de suivre le développement des différentes thématiques, ce système d'information a connu une phase d'évolution progressive permettant aux projets successifs de bénéficier de nouveau outils et matériels. **[Alain, 2006**]

La démarche épidémiologique consiste à mesurer la fréquence d'un phénomène de santé, telle qu'une maladie, d'en faire la distribution selon les caractéristiques de personne, de lieu et de temps afin d'émettre des hypothèses sur les déterminants de cette fréquence. Une fois les hypothèses vérifiées, des actions appropriées seront menées pour contrôler voire éliminer le phénomène en question. L'impact des actions implantées sur la fréquence du phénomène de santé est ensuite évalué.

L'épidémiologie humaine est non seulement une science qui, a ce titre, mérite d'entrer dans la culture générale de chacun, mais aussi parce que c'est la science en amont de la sante publique. Pour l'instant des méthodes et résultats de l'épidémiologie, en particulier en ce qui concerne la détermination et la quantification des facteurs de risque biologiques. La recherche méthodologique couvre des aspects diversifies : amélioration des protocoles d'étude, des méthodes de mesure, d'analyse statistique, de modélisation prévisionnelle. Elle est presque toujours intégrée au sein des structures d'épidémiologie généralistes ou spécialisées par problème. **[Alain, 2006]**

L'épidémiologie est une science récente qui a été individualisée au cours du 19ème siècle, mais elle n'est vraiment devenue une discipline de base de la santé publique qu'à partir de la deuxième moitié du 20ème siècle.

Depuis 1964, le terme *épidémiologie* est devenu l'un des mots les plus répandu du domaine médical. Beaucoup de chercheurs tentent de définir l'épidémiologie à partir des différents aspects et domaines d'applications. Parmi de nombreuses définitions, nous en choisissons une des plus citées.

1. Définition

Selon **l'OMS [SW]**: l'épidémiologie étudie les méthodes des fréquences et des répartitions dans le temps et dans l'espace des problèmes de santé dans des populations humaines, et le rôle des facteurs qui les détermine. L'épidémiologie clinique est l'application de ces méthodes à l'activité clinique, cette fréquence et cette répartition se situe au sein de populations humaines.

L'épidémiologie d'intervention a comme objectif l'action sur le terrain dans un but de contrôle et de prévention. C'est une des disciplines de base de la santé publique, mais elle est en relation avec de nombreux autres domaines : les sciences sociales, l'économie de la santé, mais également d'autres disciplines comme la démographie, l'histoire, le droit, la géographie, les bio-statistiques... etc. **[SW1]**

L'épidémiologie concerne l'ensemble des maladies et situations pathologiques, et pas seulement les seules maladies transmissibles. La démarche épidémiologique se diffuse dans l'ensemble de la recherche médicale.

2. Les différents types de l'épidémiologie

Sachant que l'épidémiologie est une discipline scientifique qui vise la cognitive et la science appliquée destinée à aider à la décision clinique et de santé publique, l'épidémiologie s'exerce dans des contextes institutionnels très diversifiés on peut regrouper en trois grands types que nous citons ci-dessous ; dans lesquels on rencontre des épidémiologistes « spécialistes ». **[TIBICHE 2012]**

2.1. L'épidémiologie descriptive

L'épidémiologie descriptive étudie la fréquence et la répartition des problèmes de santé dans la population, en fonction des caractéristiques des personnes (âge, sexe, profession), de la répartition géographique et de leur évolution dans le temps. Ceci dit, elle permet d'élaborer des hypothèses étiologiques, et se repose actuellement sur des données recueillies en dehors de la recherche académiques, dans le cadre d'administration centrales ou hospitalières et d'agences sanitaires qui produisent des « statistiques sanitaires ».

2.2. L'épidémiologie explicative ou analytique

L'épidémiologie analytique mesure les risques, recherche les causes des problèmes de santé et quantifie leurs importances, elle étudie le rôle de l'exposition à des facteurs pouvant favoriser l'apparition de pathologies.

2.3. L'épidémiologie évaluative

L'épidémiologie évaluative apprécie les résultats d'une action de santé dans la collectivité. Elle regroupe l'évaluation des stratégies, des pratiques, des programmes de santé et des thérapeutiques. Elle fait appel si nécessaire à l'évaluation médico-économique qui associe la mesure des coûts et des conséquences des actions de santé. Elle fait l'objet d'une polycopie spécifique.

3. Domaines d'investigation de l'épidémiologie

Généralement pour une étude épidémiologique, certains critères qui ont trait à la santé d'une population sont pris en considération : Décès (mortalité), maladie (morbidité), chroniques (maladies cardio-vasculaires, cancer, VIH, tuberculose..), d'autres ont trais à l'environnement et son rôle pathologies ainsi qu'aux comportements: milieu socioprofessionnel, Nutrition, Toxique.

4. Les indicateurs

Quelque soit la conception de la santé et de la maladie, il importe d'avoir une définition opérationnelle, c'est à dire permettant des mesures. Les instruments de mesure sont le plus souvent basés sur des indicateurs.

4.1. Définition

D'après **Kistemaker [Bouyer, 2003]:** il les a décrits comme des instruments qui mesurent un aspect quantifiable des soins, pour guider les professionnels dans le suivi et dans l'évaluation de la qualité. Il en fait des repaires pour décider des futures études d'évaluation. Ces indicateurs sont :

- **les indicateurs sentinelles** qui mesurent des évènements sérieux, indésirables et pour lesquels il n'est pas possible à priori de fixer de normes.

- **les indicateurs basés sur des taux** qui mesurent un évènement pour lequel un certain pourcentage de survenue est acceptable.

4.2. Les différents types d'indicateurs

La mesure des indicateurs nécessite une définition rigoureuse à l'aide de critères précis et de questionnaires standardisés. Différents indicateurs sont calculés à partir de ces données :

- **le ratio** est le rapport d'un numérateur et d'un dénominateur de nature différente. Il est statique et n'a pas d'unité : exemple sex-ratio (homme/femme),
- **la proportion** est le rapport d'un nombre de personnes atteintes d'un problème de santé à l'effectif de la population correspondante. Elle est statique et sans unité. C'est en général un pourcentage,
- **le taux** est le rapport du nombre de nouveaux cas d'un problème de santé apparu pendant une période à la population moyenne pendant cette période. Il permet de comparer les populations de taille différente. C'est une mesure des évolutions,
- **le quotient** est le rapport du nombre de personnes touchées par un problème de santé dans une période à la population concernée au début de la période. Une mesure de la probabilité de survenue du problème dans la population au cours de la période.

5. Surveillance épidémiologique

Les données collectées dans le cadre de la surveillance épidémiologique permettent de surveiller l'évolution des maladies, d'identifier les facteurs de risque et ainsi mettre en place des mesures de prévention et de lutte pour réduire l'incidence et la prévalence de ces maladies, donc de faire le diagnostic de l'état de santé de la population.

5.1. Définition [Tibiche 2012]

La surveillance épidémiologique consiste en la collecte systématique continue, l'analyse, l'interprétation des données sanitaires, afin d'élaborer, de mettre en place et d'évaluer les programmes de santé publique ainsi que la diffusion rapide des données de santé. Cette action est essentielle à la pratique de la santé publique.

Dans le cadre des maladies transmissibles, le but de cette surveillance est :

- Connaître l'incidence et les caractéristiques d'une maladie infectieuse ; d'étudier la dynamique de diffusion sociale, temporelle et spatiale d'une maladie et d'en prédire l'extension.
- Disposer de système et d'indicateurs d'alerte d'épidémie; d'intervenir lors d'une épidémie pour interrompre la chaîne de transmission;
- Connaître les facteurs de risque des infections afin de proposer les mesures de prévention et des recommandations les plus adaptées;
- Evaluer les actions de prévention.

5.2. Objectives de la surveillance épidémiologique

- Apprécier l'ampleur d'un phénomène de santé et de suivre ses tendances selon les caractéristiques de temps, de personne et de lieu = *décrire*,
- Détecter les épidémies = *alerter*,
- Evaluer l'impact des mesures de prévention et de contrôle = *évaluer*,
- Emettre des hypothèses,
- Identifier des phénomènes pour la recherche épidémiologique,
- Faire les projections des besoins en soins de santé.

6. Les épidémies importantes

Afin de déceler les plus importantes épidémies, d'un point de vue épidémiologique deux facteurs sont à considérer : la fréquence (incidence ou prévalence ou maladie au potentiel épidémique), et la sévérité (mortalité importante ou cause d'incapacité). Par exemple le choléra est importante par ce que l'incidence et la mortalité peuvent être élevé, de même d'autre maladies sont importantes comme le SIDA, tuberculose…etc. cette dernière fera l'objet de notre base d'expérimentation.

7. La Tuberculose [SW05]

La tuberculose est l'une des maladies dues à un agent infectieux unique les plus meurtrières au monde; elle se situe en seconde position juste après le VIH/sida.

En 2011, 8,7 millions de personnes ont développé la tuberculose et 1,4 million en sont mortes. Plus de 95% des décès par tuberculose se produisent dans les pays à revenu

faible et intermédiaire, et la maladie est l'une des trois principales causes de décès chez les femmes âgées de 15 à 44 ans.

En 2010, on comptait environ 10 millions d'enfants orphelins dont les parents étaient décédés de la tuberculose.

Le nombre de personnes développant la tuberculose chaque année est, selon les estimations, en diminution – bien que très lente – ce qui signifie que le monde est sur la bonne voie pour atteindre l'objectif du Millénaire pour le développement consistant à inverser la tendance de la maladie d'ici à 2015.

7.1 Définition

La tuberculose (TP) est une maladie infectieuse causée par une bactérie appelée mycobactérie tuberculosis. Elle affecte habituellement les poumons, mais peut aussi toucher d'autres parties du corps comme les reins, la colonne vertébrale et le cerveau.

Lorsqu'une personne développe une tuberculose active (maladie), les symptômes (toux, fièvre, sueurs nocturnes, perte de poids, etc.) peuvent rester modérés pendant de nombreux mois. Cela peut inciter le malade à repousser le moment de consulter, et se traduire par la transmission de la bactérie à d'autres personnes. Les personnes atteintes de tuberculose évolutive peuvent infecter jusqu'à 10 à 15 autres personnes avec lesquelles elles sont en contact étroit en l'espace d'une année. Sans un traitement approprié, jusqu'à deux tiers des personnes atteintes de tuberculose évolutive en mourront.

La tuberculose est contagieuse. Les personne qui ont la maladie tuberculose (tuberculose active) propagent des germe de TP dans l'air.il est important que les personnes atteintes de maladie tuberculose se fassent traiter sans tarder. Les traitements contre la maladie tuberculose peuvent guérir l'infection et empêcher sa transmission à d'autres personnes.

Conclusion

L'épidémiologie est devenue la science de référence de la santé publique, et de fondement de la « médecine fondée sur les preuve scientifique », les agences de santé publique, internationales, ont mis au point des « échelles de niveau de preuve » dans lesquelles elles proposent une hiérarchie des études épidémiologique au sommet desquelles on trouve l'approche expérimentale, puis les études de cohorte, puis les études cas témoins, enfin les études épidémiologiques descriptives et les opinions d'experts. **[SW1]**

Un meilleur contrôle des maladies infectieuses passe inévitablement par une meilleure Compréhension de la manière dont elles se propagent. Par conséquent, les modèles informatiques constituent des outils précieux au niveau décisionnel grâce au data minig pour les autorités en matière de santé publique.

Ces classifications sont utiles pour interpréter l'information produite par les travaux scientifiques visant à évaluer l'efficacité de traitement.

Pour mieux comprendre et voir l'utilité de représenter la propagation d'une maladie sous la forme d'un modèle simulable, le chapitre suivant permet de revenir sur les principales méthodes de modélisation épidémique.

Chapitre

2 Les techniques de modélisation

Introduction

La modélisation épidémiologique a contribué à la compréhension du comportement des maladies infectieuses, de ses impacts possibles et de prévisions futures au sujet de la propagation de la maladie. Elle consiste en gros à construire des modèles qui sont utilisés dans la comparaison, la planification, la mise en œuvre, l'évaluation et l'optimisation de divers programmes de prévention, de thérapie et de contrôle.

Certains domaines d'études tel que l'épidémiologie, qui sont en général abordé uniquement par le corps médical (épidémiologistes), nécessitent de la part des développeurs, la mise en œuvre de raisonnement complexes. On leur demande de se « mettre dans la peau » des médecins et cela n'est pas sans leur poser de nombreux problèmes. Développer une interface ou coupler des outils permettant à l'homme d'interagir avec le modèle en le considérant comme une boite noire. **[BO-OU, 2012]**

1. La modélisation épidémiologique

La modélisation épidémiologique des maladies infectieuses est caractérisée par la transmission des maladies infectieuse dans la population d'accueil qui est un processus fondamental.

Pour la modélisation de notre problème épidémiologique nous avons effectué quelque recherche qui nous ont permis de sélectionner parmi une gamme de méthodes de modélisation, dont La majorité des modèles épidémiques actuelles utilisent des équations différentielles, des automates cellulaires et des systèmes multi-agents pour mieux comprendre et voir l'utilité de représenter la propagation d'une maladie sous la forme d'un modèle simulable.

1.1 Les équations différentielles

L'équation différentielle est une équation qui définit une relation entre une fonction et un ou plusieurs de ses dérivés. Donc, une équation différentielle est une équation contenant des dérivés d'une ou plusieurs variables dépendantes, par rapport à une ou plusieurs variables indépendantes.

De manière générale, une équation différentielle est une équation :

- dont l'inconnue est une fonction y dépendant d'une variable x (ou t),
- qui fait intervenir y et certaines de ses dérivées y′, y″, etc., et éventuellement la variable x (ou t).

1.1.1 Les avantages et les inconvénients des équations différentielles

La modélisation de systèmes épidémiologiques par équations différentielles est actuellement largement répandue. Elle comporte deux avantages essentiels [site 3].

- l'approche est formalisée donc une équation mathématique est universellement compréhensible, des solutions analytiques peuvent être trouvées et si ce n'est pas le cas, des simulations numériques peuvent être effectuées.
- un système d'équations différentielles permet de décrire l'évolution d'une population de cellules ou de nombreux types d'interactions entre plusieurs populations de cellules.

Ce formalisme se heurte toutefois à plusieurs types de problèmes :

- Les réseaux de taille réelle - de l'ordre d'une centaine de réactions - sont difficiles à modéliser par un jeu d'équations différentielles.
- L'introduction de nouvelles populations et l'amélioration du modèle nécessitent la modification de la plupart des équations du modèle.
- La modélisation par équation différentielle nécessite un haut niveau d'abstraction.
- Les équations différentielles **[Bagni, 2002]** ne tiennent pas compte des facteurs spatiaux tels que la variable de la densité de population et la dynamique de la population **[Roy, 1991]**

1.2 Les automates cellulaires

Les automates cellulaires (AC) se caractérisent par leur discrétisation de l'espace et du temps **[Codd, 1967].** Ou il consiste à un diagramme où chaque nœud est un automate à états finis ou une cellule. Ce graphique est généralement sous la forme d'un réseau bidimensionnel dont les cellules évoluent selon une fonction de mise à jour globale appliquée uniformément sur toutes les cellules. Comme arguments, cette fonction de mise à jour prend l'état actuel de la cellule et les états des cellules dans son voisinage d'interaction

Les AC offrent la possibilité d'introduire des paramètres et de les estimer en fonctions des données récoltés sur le terrain. Parmi les auteurs qui ont travaillé avec les AC on a :

Mikler et all **[Saporta, 2000]**ont proposé un paradigme cellulaire stochastique global des automates, qui a incorporé des contraintes géographiques, démographiques et migratrices.

Zhang et all **[Silvi, 2002]** ont étudié les impacts que la su-urbanisation avait portés à la transmission de maladies infectieuses par le modèle d'automate cellulaire.

Martín del Rey **[White et al, 2007],** propose un modèle SEIR mis en œuvre par l'intermédiaire d'un automate cellulaire où chaque cellule représente une population particulière comme le cœur rural ou urbain. Les voisins de chaque cellule sont ceux entre lesquels il existe un canal de communication qui permet la circulation de population de l'un à l'autre.

Liu et al **[NLI, 2006]** Met en œuvre un modèle classique basé sur SEIR équations différentielle ordinaire. Le document explore le comportement spatial des maladies épidémiques qui sont saisonnières. Pour simuler le mouvement spatio-temporel associé aux différentes vagues d'épidémie ils misent en œuvre un modèle appelé «dépondant des voisins», qui est une modification du modèle classique, de sorte que les équations n'évoluent pas seulement à l'instant t, mais elles dépendent aussi de l'espace, ainsi résultant vers un système d'équations aux dérivées partielles discrétisées dans le temps et l'espace.

1.2.1.Avantages et limites des automates cellulaires

L'avantage d'automate cellulaire par rapport aux autres approches tel que les modèles mathématiques est d'ajouter une composante spatiale. Cependant, il y a deux limites à l'utilisation des automates cellulaires. En effet, la grille est généralement artificielle (non liée au phénomène étudié). Cet inconvénient a été contourné par la mise en œuvre des automates cellulaires en utilisant la grille irrégulière **[Shi, 2000]**. La seconde limite est que les automates cellulaires ne peuvent pas gérer les individus et leur mobilité dans l'environnement géographique. Cela semble être une contrainte importante lorsque l'on considère les phénomènes sociaux dans lesquels la mobilité des personnes doit être simulée.

Les automates cellulaires (AC) ont un intérêt évident pour la modélisation épidémiologique. La représentation du territoire qu'ils introduisent correspond à un ensemble de cellules de forme identique, en général, ayant chacune des caractéristiques de

milieu, d'individu, de population ; des fonctions d'état permettent de caractériser l'évolution de chaque cellule en fonction de son état et de celui de ses voisines ; les aléas sont introduits par les interactions entre éléments.

Cependant, cette formalisation ne permet pas aisément de prendre en compte des entités décisionnelles qui influencent le comportement des automates cellulaires. C'est pourquoi les systèmes multi-agents (SMA) trouvent un réel engouement dans le domaine de la modélisation épidémiologique. La conception d'un SMA se fonde sur la formalisation d'interactions locales entre des agents et avec leur environnement **[SW04].**

1.3. Les systèmes multi-agent (SMA)

Un système multi-agents (SMA) est une intégration de la théorie des systèmes adaptatifs complexes, l'intelligence artificielle distribuée et les techniques de vie artificielle. Il a été un moyen important dans l'analyse et la simulation des systèmes complexes et a été largement utilisé dans la simulation épidémique, économique, politique, sociale, écologique, …etc. **[Oechslein, 2001] [Parunak, 1998]**.

L'utilisation de l'approche système multi-agents pour la modélisation du système épidémique est devenue populaire pour de nombreux chercheurs. Développer un modèle de simulation à l'aide de ces approches consiste à modéliser les entités par des agents.

1.3.1. Définition de système multi-agent

Un système multi-agents est constitué de composants (entités) qui représentent les caractéristiques du système. Les entités communiquent les unes avec les autres et avec l'environnement dans lequel elles vivent et sont modélisées et mises en œuvre en utilisant des agents **[Davidsson, 2000]**. Les agents ont des comportements et des caractéristiques et ils représentent les différentes composantes qui forment le modèle.

Ils ont aussi un protocole de communication qui les aide à comprendre les messages et l'échange d'informations entre eux. Le modèle est le résultat des caractéristiques et des comportements des agents, de leur interaction et avec le milieu dans lequel ils vivent **[Parunak, 1998].**

Les caractéristiques et les comportements des agents dans un système multi-agent peuvent être considéré selon deux aspects: interne et externe. L'aspect interne correspond aux caractéristiques et aux comportements internes des agents tandis que l'aspect externe

comprend les comportements et les caractéristiques des agents lors de l'interaction avec d'autres agents et l'environnement auquel ils appartiennent **[Omicini, 2000]**.

D'après Alain-Jérôme **[Alain, 2003]**, le système multi-agent englobe 3 sous-systèmes : un SMA pour la simulation des épidémies, un SMA pour la détection d'éventuelles épidémies et un système d'aide à la décision exploitant des connaissances médicales pour diagnostiquer des maladies et des épidémies. Chaque agent du système de détection d'épidémies possède ses propres connaissances pour établir les diagnostiques en relation avec les rôles qu'ils jouent au sein de l'organisation et de l'environnement. La prise de décision est ainsi répartie.

1.3.2. Les avantages et les limites des SMA

Chen et al. **[Chen, 2004]** et Deng et al **[Deng, 2004]** ont utilisé la méthode SMA pour simuler les systèmes épidémiques. Les SMA sont l'intégration de la théorie des systèmes complexes adaptatifs, de l'intelligence artificielle distribuée et des techniques de la vie artificielle. À l'heure actuelle, ils ont été un moyen important dans l'analyse et la simulation de systèmes complexes. Ils ont pu pallier les inconvénients des ODE et des AC, cependant ils présentent les inconvénients suivants : [**Bubniakvà, 2007]**

- l'épidémie généralement reconnue comme survenant au fil du temps, il est évidemment un processus impliquant l'espace, Mais la plupart des modèles épidémiques SMA existants sont principalement porté sur les impacts des paramètres d'infection tels que le taux d'infection et le taux de guérison.
- la plupart de ces modèles épidémique basée sur SMA ne prennent pas toujours en compte l'influence des relations sociales entre les agents.

1.4. Algèbres de processus

Malgré l'interprétation des méthodes de modélisation dans la dynamique des épidémies comprenant principalement les équations différentielles, les automates cellulaires, et système multi-agents, ils ne sont pas satisfaisants pour expliquer les phénomènes épidémiologiques. Pour cette raison les experts s'appuient habituellement sur un nouvel outil de modélisation «les algèbres des processus», ainsi que sur la simulation pour essayer de comprendre le comportement de ces systèmes.

L'algèbre de processus est un formalisme algébrique qui représente une abstraction, utile pour décrire les systèmes qui peuvent être considérés comme des compositions de divers composants individuels, chacun présente des comportements

spécifiques. Un système peut être considéré comme l'interaction de plusieurs comportements. Le comportement lui-même peut être considéré comme les changements dynamiques dans le système tel que les événements, les actions ou les évolutions. **[Ciocchetta, 2010**]

1.4.1. PEPA (Performance Evaluation Process Algebra)

Est un formalisme, développé par Hillston en 1994, qui étend l'algèbre des processus classiques en associant, à chaque action, une variable aléatoire, représentant la durée. Ces variables aléatoires sont supposées être exponentiellement distribuées, ce qui établit clairement une relation entre le modèle de l'algèbre des processus et un processus de Markov en temps continu. Via ce processus de Markov, des mesures de performances peuvent être extraites à partir du modèle.

il a été à l'origine défini pour la modélisation d'exécution des systèmes concurrents. Les systèmes sont représentés comme composition des composants ou des agents qui entreprennent des actions. La syntaxe du langage est présentée ci-dessous **[Ciocchetta, 1996].**

Nous décrivons quelques dispositifs des algèbres de processus qui sont utiles dans le cadre des systèmes biologiques:

- **Compositionalité *:*** Le système entier peut être défini à partir de la définition de ses composants secondaires.
- **Signification formelle *:*** Les algèbres de processus sont des formalismes mathématiques, elles sont basées sur la syntaxe précise et leur sémantique sont clairement définie. Par conséquent, les modèles peuvent être définis sans ambiguïté.
- **Abstraction *:*** Des algèbres de processus peuvent être employées pour modéliser à divers niveaux d'abstraction.
- **Analyse *:*** les algèbres des processus soutiennent divers genres d'analyse, permettant de vérifier, valider le modèle et corriger toutes les inexactitudes **[Alain, 2003]**.

1.4.1.1. La syntaxe de PEPA

Un modèle PEPA est décrit comme l'interaction d'un ensemble de composantes. Chaque composante peut exécuter un ensemble d'actions : une action $\alpha \in$ Act est décrite comme une paire (α, r), où $\alpha \in A$ est le type de l'action et $r \in R+$ est le paramètre de la distribution exponentielle négative régissant sa durée.

1.5. Bio-PEPA

Bio-PEPA est un langage défini récemment pour la modélisation et l'analyse des réseaux biochimiques **[Ciocchetta, 2008]**. Un réseau biochimique se compose d'ensemble d'espèces moléculaires, telles que des protéines, de petites molécules, et des gènes, qui agissent les uns sur les autres au travers de réactions. Les espèces moléculaires sont situées en compartiments, tels que le noyau et le cytosol, ou sur les membranes qui les enferment.

Bio-PEPA soutient la définition des compartiments statiques comme noms : les compartiments sont des récipients pour les espèces moléculaires et ne sont impliqués dans aucune réaction qui change leur taille ou la structure. Un volume constant (ou la taille) peut être associé à eux.

1.5.1 La Syntaxe Bio-PEPA

- La syntaxe de Bio-PEPA **[Ciocchetta, 2010]** est constitué de composants et le modèle séquentiel
  composante. Ce sont définis comme suit:

$$S ::= (\alpha, k)\ op\ S \mid S\ +\ S \mid C$$

$$P ::= \mathrm{P}\ \underset{L}{\bowtie}\ \mathrm{P} \mid S[l]$$

Le composant S est la composante séquentielle et représente une espèce biochimique, P est la composante de modèle et définit les interactions entre composants séquentiels **[Deng, 2004]**. Les différents opérateurs sont expliqués ci-dessous:

- *Préfix* : est représenté comme suit :

$$S ::= (\alpha, k)\ op\ S$$

Où :

- α représente le type d'action (c'est à dire un nom unique associé à chaque réaction dans le système).

- k représente la stœchiométrie de l'espèce dans la réaction.
- Le combinateur « op » représente le rôle des composants.

(↓ Représente réactif, ↑ Signifie produit, ⊕ signifie un activateur, ⊖ et un inhibiteur, Θ un modifié ⊙ générique respectivement.)

- ***Choix:*** L'opérateur choix est représenté par S + P. Cela signifie un composante qui peut se comporter soit comme S ou P. Les deux les processus commence leurs activités simultanément. Toutefois, le composant qui se termine la première activité est choisi et l'autre est éliminé.
- ***Constant*:** Les constantes sont des composants dont la signification est donnée par une équation tels que $A \overset{def}{=} P$, Constantes. tirent leur sens de l'équation de planification. Par exemple, le comportement de A peut être attribué à P. Il nous permet d'attribuer noms des composants.
- ***Coopération:*** Le terme $P \bowtie_{L} Q$ est utilisé pour représenter lorsque P et Q interagir un avec l'autre pour effectuer une activité ensemble. Les activités communes sont répertoriées dans L l'ensemble des activités. Si une activité est répertoriée dans cet ensemble, puis les composants sont forcés de se synchroniser sur cette activité. Là activités à l'extérieur ensemble L procéder de façon indépendante. Notez que Bio-PEPA supporte la synchronisation à plusieurs voies, dans lequel plus de 2 éléments peuvent coopérer sur une activité donnée.
- Les niveaux de concentration: Le terme S [l] désigne le nombre de niveaux de chaque espèce, où l est la concentration de niveaux discrets. **[Alain, 2003] [car, 1978]**

1.5.2 Avantage de Bio-PEPA

La description de la syntaxe de Bio-PEPA permet de visualiser son importance par rapport aux autres méthodes de modélisation, du fait qu'elle regroupe les avantages de chacune d'elles tout en facilitant au développeur son implémentation.

A la découverte de Bio-PEPA, le monde de la modélisation a révolutionné car il est possible maintenant de réaliser un modèle quasi-parfait. Il a le grand avantage d'offrir des outils d'analyse, tels que, la simulation stochastique, la vérification de modèle et l'extraction des ODE à partir du modèle. En particulier, le modélisateur peut choisir l'approche la plus appropriée pour l'étude des techniques de modélisation et d'analyse pouvant être utilisés ensemble pour une meilleure compréhension du comportement du système.

Conclusion

Dans ce chapitre nous avons présenté la particularité des méthodes de modélisation épidémiologique, une principale variante de celle-ci : Bio-PEPA pour la modélisation et l'analyse des modèles épidémiologiques.

La description de la syntaxe de Bio-PEPA permet de visualiser son importance par rapport aux autres méthodes de modélisation, du fait qu'elle regroupe les avantages de chacune d'elles tout en facilitant au développeur son implémentation.

Toutes fois, quel qu'il soit l'outil de modélisation, l'interaction entre développeur et expert du domaine reste une lourde tâche à surmonter lors de la réalisation du modèle et parfois même ils se retrouvent devant une panoplie d'informations qu'ils ne peuvent gérer. Donc il devient important pour le bon suivi épidémiologique de se retourner vers les techniques du Data Mining, qui semblent être le moyen le plus performant à l'extraction d'informations pertinentes à l'étude épidémiologique, le chapitre suivant en fera une large description.

Introduction

Le Data Mining est au cœur de toutes les préoccupations du monde économique. C'est un processus qui permet de découvrir, dans de grosse base de données consolidées, des informations jusque la inconnus mais qui peuvent être utiles et lucrative et d'utiliser ces informations pour soutenir des décisions commerciales tactiques et stratégiques.

Les techniques statistiques du Data Mining sont bien connues. Il s'agit notamment de la régression linéaire et logistique, de l'analyse multi variée, de l'analyse des composantes principale, des arbres décisionnels et des réseaux de neurones. Cependant, les approches traditionnelles de l'inférence statistique échouent avec les grosses bases de données, car en présence de milliers ou de millions de cas et de centaines ou de milliers de variables, on trouvera forcément un niveau élevé de redondance parmi les variables, certaines relation seront fausses, et même les relations les plus faibles paraitront statistiquement importantes dans tout test statistique **[Gong, 2006].**

1. Data Mining

De manière générale, on peut définir le Data Mining (ou exploitation des gisements de donnée) comme l'extraction, à partir de gros volumes de données, d'informations ou de connaissances originales, auparavant inconnues, potentiellement utiles.

Le Data Mining correspond donc à l'ensemble des techniques et des méthodes qui à partir de gros volumes de données, permettent d'obtenir des connaissances exploitables pour l'aide à la décision.

Il existe une distinction précise entre le concept de « découverte de connaissances dans les bases de données » et celui de Data Mining. En effet, ce dernier n'est que l'une des étapes du processus de découverte de connaissances correspondant à l'extraction des connaissances à partir des données. **[Didier, déce1998] [Didier, Juin1998]**

2. Naissance du Data Mining

Selon Hebrail et Lechevallier [**Hebrail, 2003]** : « La naissance du Data Mining est essentiellement due à la conjonction des deux facteurs suivants :

- l'accroissement exponentiel dans les entreprises, de données liées à leur activité (données sur la clientèle, les stocks, la fabrication, la comptabilité...) qu'il serait dommage de jeter car elles contiennent des informations clés sur leur fonctionnement stratégique pour la prise de décision.
- les progrès très rapides des matériels et des logiciels.

L'objectif poursuivi par le Data Mining est donc celui de la valorisation des données contenues dans les systèmes d'information des entreprises. »

Les premières applications se sont faites dans le domaine de la gestion de la relation client qui consiste à analyser le comportement de la clientèle pour mieux la fidéliser et lui proposer des produits adaptés. Ce qui caractérise la fouille de données (et choque souvent certains statisticiens) est qu'il s'agit d'une analyse dite ≪ *secondaire* » de données recueillies à d'autres fins (souvent de gestion) sans qu'un protocole expérimental ou une méthode de sondage ait été mis en œuvre. **[Hand, 1998]**

Quand elle est bien menée, la fouille de données apporte des succès certains, à tel point que l'engouement qu'elle suscite a pu entrainer la transformation (au moins nominale) de services statistiques de grandes entreprises en services de Data Mining.

La recherche d'information dans les grandes bases de données médicales ou de santé (enquêtes, données hospitalières, etc.) par des techniques de Data Mining est encore relativement peu développée, mais devrait se développer très vite à partir du moment ou les outils existent. **[Lavrac, 1999]**

3. Domaines d'applications [Rémi, 2004]

Le Data Mining touche aujourd'hui pratiquement tous les domaines tels que :

- Marketing direct: population à cibler (âge, sexe, profession, habitation, région, …) pour un publipostage.
- Gestion et analyse des marchés : Ex. Grande distribution : profils des consommateurs, modèle d'achat, effet des périodes de solde ou de publicité, «panier de la ménagère»
- Détection de fraudes: Télécommunications, ...

- Gestion de stocks: quand commander un produit, quelle quantité demander, …
- Analyse financière: maximiser l'investissement de portefeuilles d'actions.
- Gestion et analyse de risque: Assurances, Banques (crédit accordé ou non)
- Bioinformatique et Génome: ADN Mining, …
- Médecine et pharmacie :
    - Diagnostic : découvrir d'après les symptômes du patient sa maladie
    - Choix du médicament le plus approprié pour guérir une maladie donné.

4. Les tâches du Data Mining [Rémi, 2004]

La liste suivante indique les tâches les plus courantes que le Data Mining est amené à accomplir :

- Classification
- Estimation
- Prédiction
- Groupement par similitudes
- Segmentation (ou clusterisation)
- Description
- Optimisation

4.1. La classification: « La classification consiste à examiner des caractéristique d'un élément nouvellement présenté afin de l'affecter à une classe d'un ensemble prédéfini » **[Jiawei, 2000].** Donc la classification consiste à examiner des caractéristiques d'un objet et lui attribuer une classe, la classe est un champ particulier à valeurs discrètes.

4.2. L'estimation: consiste a estimé la valeur d'un champ à partir des caractéristiques d'un objet. Contrairement à la classification, le champ est un champ à valeur continue.

4.3. La prédiction: elle ressemble à la classification et à l'estimation mais dans une échelle temporelle différente. Tout comme les tâches précédentes, elle s'appuis sur le passé et le présent mais son résultat se situe dans un futur généralement précisé.

4.4. Le regroupement par similitude: Le regroupement par similitudes consiste à grouper les éléments qui vont naturellement ensembles.

4.5. La segmentation: La segmentation ou l'analyse des clusters consiste à former des groupes (cluster) homogènes à l'intérieur d'une population. Pour cette tâche il n'y a pas de classe à expliquer ou de valeur à prédire définie à priori, il s'agit de créer des groupes homogènes dans la population (l'ensemble des enregistrements). Il appartient ensuite à un expert du domaine de déterminer l'intérêt et la signification des groupes ainsi constitués. Cette tâche est souvent effectuée avant les précédentes pour construire des groupes sur lesquels on applique des tâches de classification ou d'estimation.

4.6. La description: C'est souvent l'une des premières tâches demandées à un outil de data Mining. On lui demande de décrire les données d'une base complexe. Cela engendre souvent une exploitation supplémentaire en vue de fournir des explications.

4.7. L'optimisation: pour résoudre de nombreux problèmes, il est courant pour chaque solution potentielle d'y associer une fonction d'évaluation. Le but de l'optimisation est de maximiser ou minimiser cette fonction, quelques spécialiste considèrent que ce type de problème ne relève pas du data Mining.

5. Data Mining en épidémiologie

L'utilisation des méthodes du Data Mining en épidémiologie et santé publique est en forte croissance. Comme dans d'autres domaines, c'est la disponibilité de vastes bases de données historiques qui incite à les valoriser, alors qu'au dire de beaucoup de spécialistes elles sont actuellement sous-utilisées. **[Lavrac, 1999]**

Après une large recherche bibliographique, dans le monde du Data Mining appliqué aux épidémiologies, nous avons pu recenser les techniques suivantes :

5.1. k-means: Proposé en 1967 par MacQueen **[MacQueen, 1967],** L'algorithme k-means est une des techniques les plus simples de clustering et il est couramment utilisé en imagerie médicale, la biométrie et des domaines connexes.

Thangavel et all [**Thangavel, 2006**] utilise l'algorithme de clustering K-means pour analyser les patients atteints de cancer du col utérin et a constaté qu'avec cette technique, ils trouvaient de meilleurs résultats prédictifs que sur un avis médical existant. Ils ont trouvé un ensemble d'attributs intéressants qui pourraient être utilisés par les médecins comme une aide supplémentaire sur l'opportunité ou non de recommander une biopsie pour un patient

suspecté d'avoir le cancer du col. Canlas **[Canlas, 2009]** a découvert que le cancer du col utérin chez les femmes utilisant la biopsie est une tâche difficile. Ils ont même découvert quelques facteurs qui peuvent assister les médecins pour prendre la décision de savoir si les patients souffrant d'un cancer du col de l'utérus doivent être proposés pour une biopsie ou non **[Govardhan, 2011].**

Asha [**Asha, 2011]** ont présenté une méthode pour la détection automatique et la classification de la tuberculose (TB) qui se propage dans l'air et attaque facilement les organes immunitaires faibles. Leur méthodologie est basée sur le regroupement et la classification en particulier K-means qui classe la tuberculose dans les deux catégories de clusters, la tuberculose pulmonaire (TBP) et PTB rétroviral (RPTB) due au VIH. Ils ont utilisé pour cela une base de données sur la tuberculose répertoriant 700 instances obtenues à partir d'un hôpital de la ville. Selon eux, la meilleure précision obtenue est de 98,7% par rapport à la machine à support vectorielle (SVM) par rapport à d'autres classificateurs. L'approche proposée permet aux médecins dans leurs diagnostique et aussi dans leurs procédures de planification de traitement pour les différentes catégories.

5.2. Les règle d'association: Cette méthode est une des innovations du Data Mining, introduite en 1993 par des chercheurs en base de données d'IBM **[Brossette, 2000]**, elle a pour but de rechercher des conjonctions significatives d'évènements. Typiquement une règle de décision s'exprime sous la forme : si (A et B) alors C mais il s'agit d'une règle probabiliste et non déterministe. On définit le support de la règle comme la probabilité d'observer à la fois la prémisse X et la conclusion Y : $P(X \cap Y)$ et la confiance comme $P(Y/X)$. Les premières applications ont concerné les achats dans les grandes surfaces : parmi les milliers de références disponibles et les millions de croisements, identifier les achats concomitants qui correspondent à des fréquences importantes. Cette méthode s'étend bien au-delà de ce type d'application. L'originalité tient essentiellement à la complexité algorithmique du problème.

Gupta et Agrawal [**Gupta, 2009**] ont appliquée d'une minière quantitative des règles d'association dans le domaine médical pour détecter la nature des associations entre les différents acides aminés qui sont présents dans une protéine. Les règles d'association ont amélioré la compréhension de la composition des protéines et ont le potentiel de donner des indices concernant les interactions entre les ensembles particuliers d'acides aminés qui se

produisent dans les protéines. Ils ont découvert des règles basées non seulement sur la présence d'acides aminés, mais aussi de leur absence.

Ramesh kumar a utilisé l'algorithme Apriori dans les bases de données du SIDA / VIH afin d'extraire les règles de contamination, ce qui a pu aider à comprendre la relation entre les informations sur le traitement et les patients tels que l'impact du nombre de jours entre chaque traitement. **[Rameshkumar 2011]**

Kumar et all **[Kumar, 2012],** utilisé différentes techniques d'exploration de données médicales pour la détection de la maladie de la tuberculose. Ils ont appliqué trois algorithmes différents: les algorithmes Apriori, une IRM priori, un PT Priori. Ils pouvaient ainsi diagnostiquer les patients à tuberculose pulmonaire et ceux à tuberculose extra-pulmonaire.

5.3. Les arbres de décision: La popularité de la méthode repose en grande partie sur sa simplicité. Il s'agit de trouver un partitionnement des individus que l'on représente sous la forme d'un arbre de décision. L'objectif est de produire des groupes d'individus les plus homogènes possibles du point de vue de la variable à prédire. Il est d'usage de représenter la distribution empirique de l'attribut à prédire sur chaque sommet (nœud) de l'arbre. **[Ricco, 2010].**

Smitha et all**,** ont appliqué la méthode d'arbre de décision pour prédire les chances de présence des cas d'une maladie dans une zone, en particulier les bidonvilles. Ce modèle a permis la prédiction de l'insolvabilité d'habitants bien à l'avance et a également identifié deux types de patients : habitants devenant patients (insolvable) en raison des facteurs de risque climatique tels que le climat saisonniers, les données pluviométriques, la propagation de maladies mortelles, la température de surface de l'eau, la température et mesure de la perception ... etc. Et les habitants devenant malades à cause de facteurs de risque non climatiques tels que l'immunité de la population et des activités de contrôle, l'abondance des vecteurs, des antécédents familiaux [**Smitha, 2012].**

Attaluri et all explorer l'intégration de l'arbre de décision (DT) et le modèle de Markov caché (HMM) pour la prédiction de sous-type de virus grippaux humain. Ils ont développé un système Web pour la détection de séquences de virus grippaux humains. Leur

expérience préliminaire a montré que ce système est facile à utiliser et puissant pour identifier les sous-types de la grippe humaine **[Attaluri, 2009].**

Conclusion

La recherche bibliographique que nous avons fait nous a permis de faire ressortir une série de travaux qui ont soulevés l'importance de l'utilisation du data Mining pour l'épidémiologie humaine.

Le chapitre suivant, fera l'objet d'une conception et implémentation d'une application mettant en œuvre les concepts acquis dans les trois premiers chapitres.

Introduction

Dans ce chapitre nous allons aborder les outils et l'algorithme que nous avons choisi pour la modélisation notre solution. Nous présentons une interface d'interaction intégrée au modèle Bio-PEPA qui faciliterait à l'expert d'une part de valider le modèle et d'autre part d'effectuer toute modification lui semblant adéquate sans faire appel à l'intervention du développeur.

Le but de notre application est l'automatisation des modèles épidémique dans Bio_PEPA après l'extraction des connaissances avec Weka.

1. Spécification et conception

1.1. Architecture de l'application

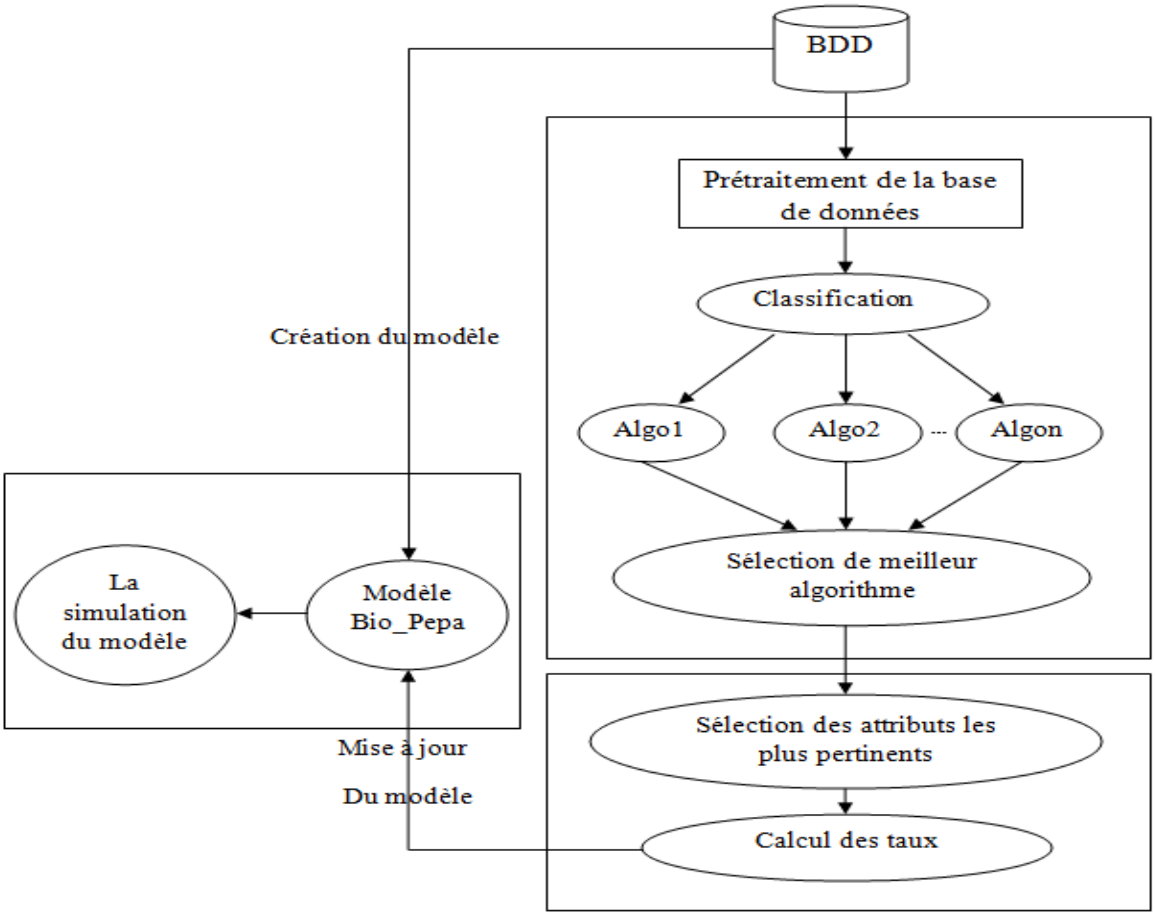


Figure 1 : Architecture de l'application

La figure 1 illustre l'architecture de notre application, détaillée comme suit :

- **Base de données**

Afin d'atteindre notre objectif, nous nous sommes intéressés à une base de données médicale de la tuberculose concernant la population de la wilaya de Mostaganem qui date de 2007, que nous avons ramené du centre de SEMEP (Service Epidémiologique de la MEdecine Préventive). Elle contient 336 individus atteints de la tuberculose pulmonaire (TP) et extra pulmonaire (TEP) décrits par 23 attributs. Notre travail se limite à l'étude de ceux atteint d'une tuberculose pulmonaires.

- **Création du modèle Bio-PEPA initial**

D'après l'analyse visuelle de la base de données ainsi que la conversation avec l'expert, la figure suivante illustre le modèle épidémique du suivi de la tuberculose que nous avons créé dans Bio-PEPA comme modèle initial.

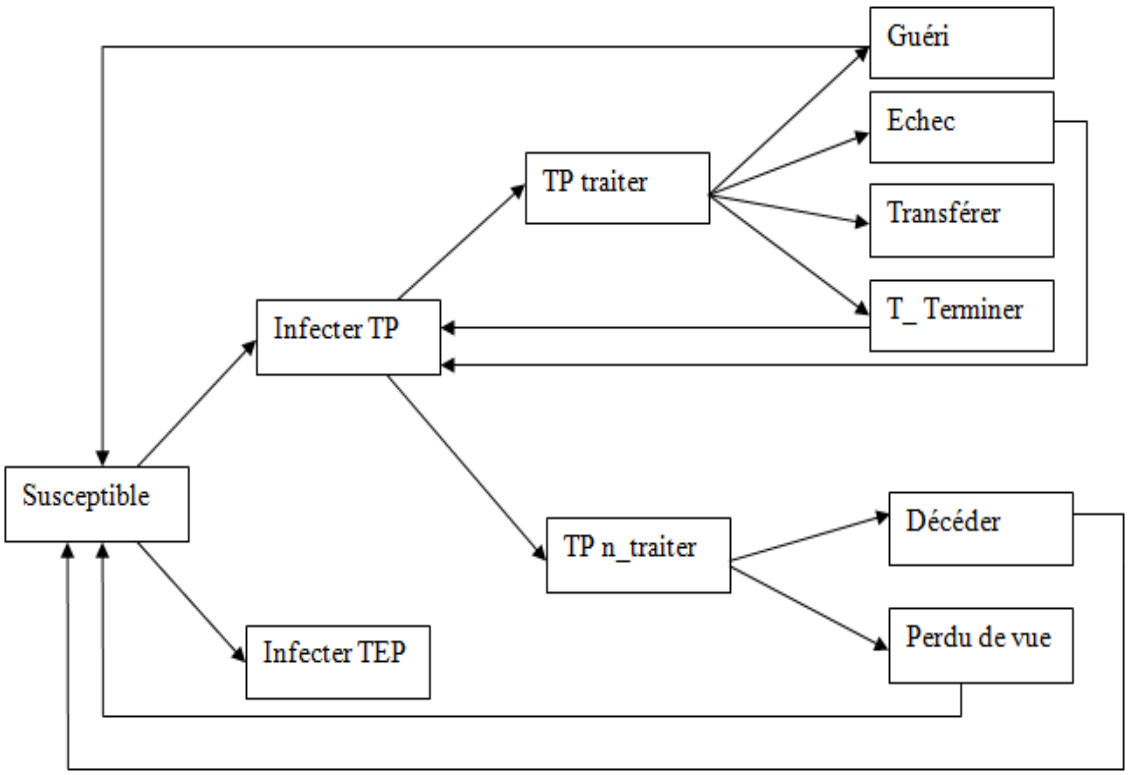


Figure 2: Le modèle épidémique de la tuberculose.

La figure 2 décrit clairement comment un état tuberculeux évolue au cours du traitement Tel qu'un susceptible peut passer à l'état infecté (TP ou TEP), et un TP peut également évoluer en d'autres états (TP traité ou non traité…etc).

Le modèle ainsi réalisé fera l'objet d'une simulation et par la suite une série d'analyse par l'expert qui le validera ou non.

- **Extraction des connaissances**
  - ✓ **Prétraitement de la base de données**

Le prétraitement de la base de données, consiste à nettoyer la base, des attributs (instances) inutiles à l'étude. Pour cela nous avons fait appel à l'instinct de l'expert.

  - ✓ **La classification**

Afin d'aboutir à l'ensemble d'attributs pertinents qui est l'objectif clé de notre modélisation, nous avons utilisé un ensemble d'algorithmes de classifications implantés dans Weka**[Bridgite, 2011]**.

- **Mise à jour du modèle Bio-PEPA**

Une fois les attributs pertinents sélectionnés, l'expert devra valider parmi eux celui qui a le plus d'impact sur l'étude, et au développeur d'optimiser le modèle initial.

1.2. Conception UML

Pour la specification et la conception de notre solution, nous avons opté pour le langage universel UML (Unified Modeling Language) **[LAU, 2008].**

UML est un langage de modélisation au sens de la théorie des langages. Il contient de ce fait les éléments constituants de tout langage, à savoir : des concepts, une syntaxe et une sémantique. De plus, UML a choisi une notation supplémentaire ; il s'agit d'une forme visuelle graphique fondée sur des diagrammes. Si l'unification est secondaire par rapport aux éléments constituants du langage, elle reste cependant primordiale pour la communication et la compréhension humaine.

UML a connu un vrai succès, ce dernier s'explique par la réussite de la normalisation des concepts objets qui ont des avantages indéniables au niveau des applications informatiques, ses principaux avantages sont la réutilisabilité des composants logiciels, la facilité de maintenance, de prototypage et d'extension des applications **[PAS, 2008].**

UML nous propose 11 diagrammes, nous en choisissons quatre :

- le diagramme de cas d'utilisation pour définir les différentes fonctionnalités de notre application.
- le diagramme de classes pour avoir un aperçu des classes constituant le système.
- le diagramme de séquences pour décrire le séquencemnt des procédures les plus importantes.
- Le diagramme d'activités pour spécifier le comportement interne des opérations.

1.2.1. Diagramme de cas d'utilisation : la figure 3 illustre le diagramme de cas d'utilisation qui montre toutes les fonctionnalités qu'offre le système à l'utilisateur.

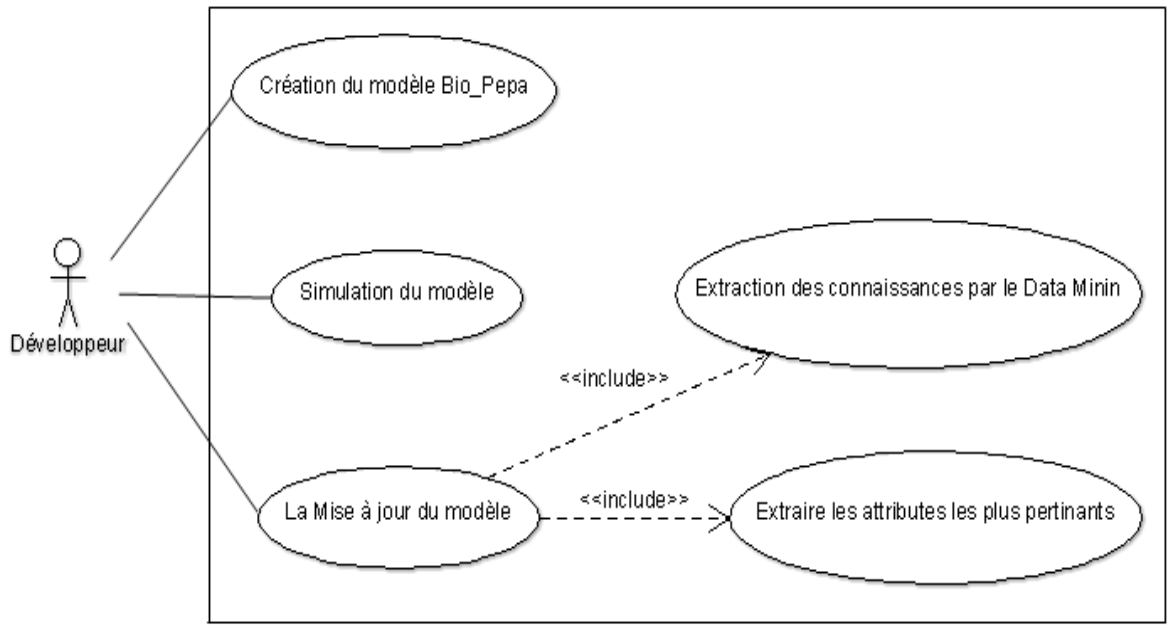


Figure 3: diagramme de cas d'utilisation.

❖ **Diagramme de séquence**

Un diagramme de séquence est un diagramme d'interaction qui se concentre sur l'ordre temporel des messages entre objets. Il peut servir à illustrer un cas d'utilisation. On n'y décrit pas le contexte ou l'état des objets. La représentation se concentre sur l'expression des interactions. L'ordre d'envoi d'un message est déterminé par sa position sur l'axe vertical du diagramme ; le temps s'écoule "de haut en bas" de cet axe **[ZHA, 2010].** Nous

avons décrit les procedures de la création et la mise à jour du modèle épidémique à l'aide des diagrammes de séquence illustrés par les **Figures 4 et 5**.

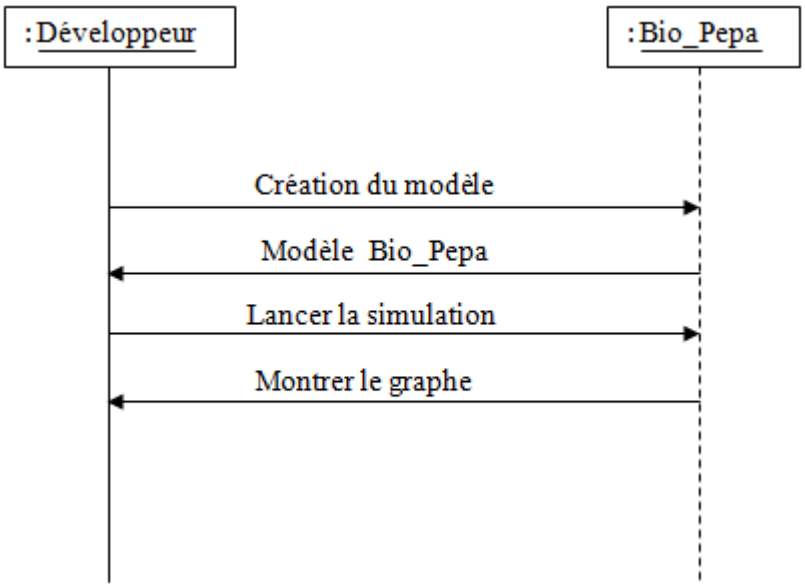


Figure 4: Diagramme de séquence de la création du modèle Bio-PEPA.

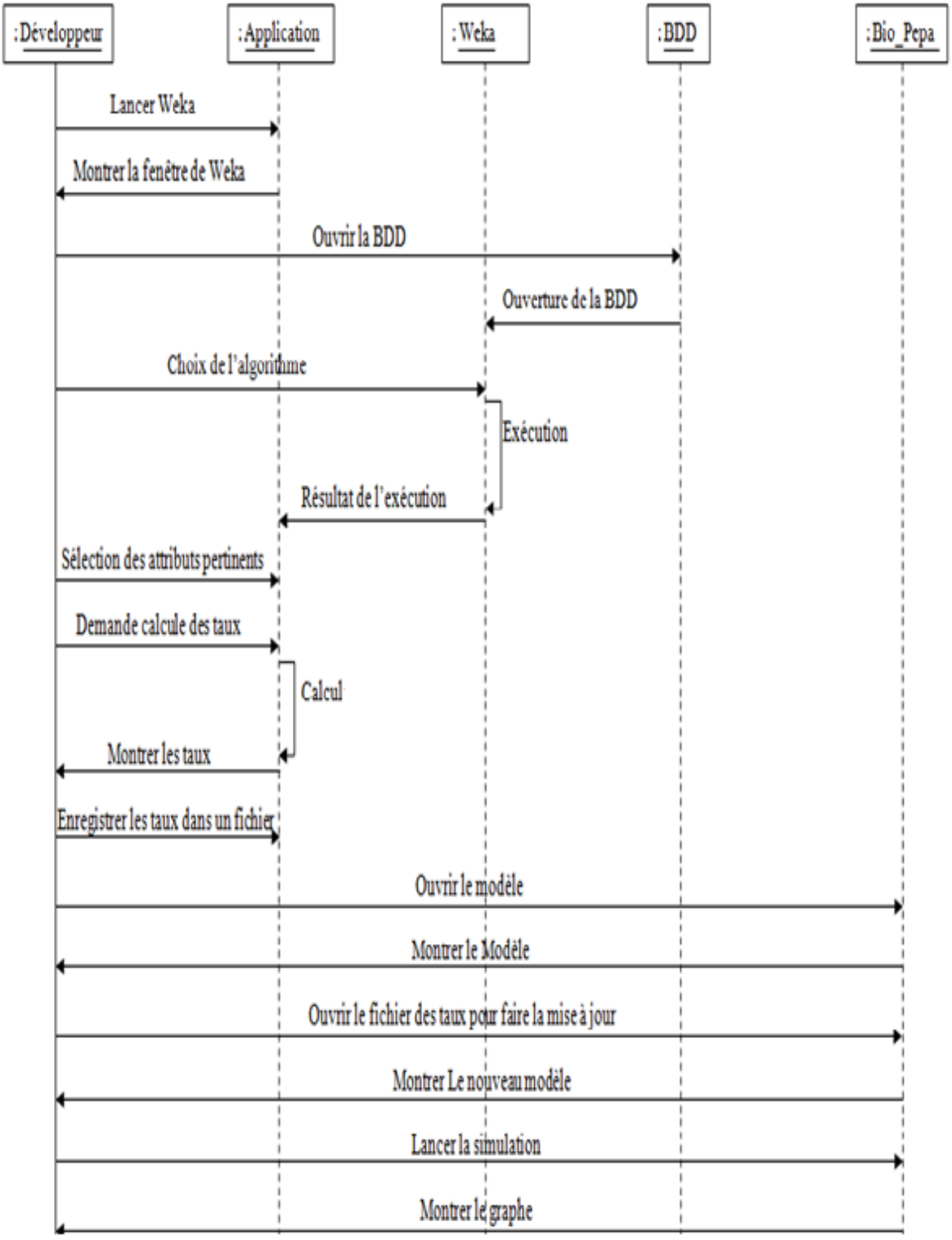


Figure 5: Diagramme de séquence de la mise à jour du modèle.

❖ **Diagramme d'activité**

Le diagramme d'activité représente la dynamique du système. Il montre l'enchaînement des activités d'un système ou même d'une opération. Le diagramme d'activité représente le flot de contrôle qui retrace le fil d'exécution et qui transite d'une activité à l'autre dans le système. La **Figure 6** , présente le diagramme d'activité de la création du modèle épidémique et la **Figure 7**, le diagramme d'activité de la mise à jour du modèle épidémique.

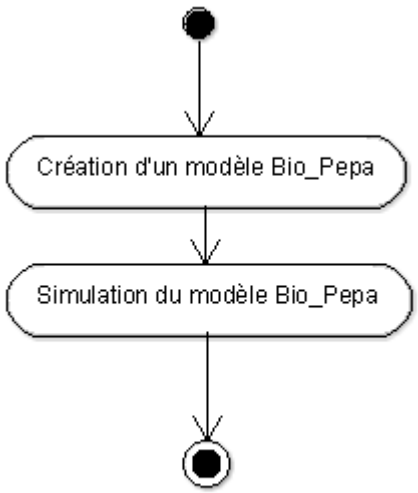


Figure 6: Diagramme d'activité de la création du model Bio-PEPA.

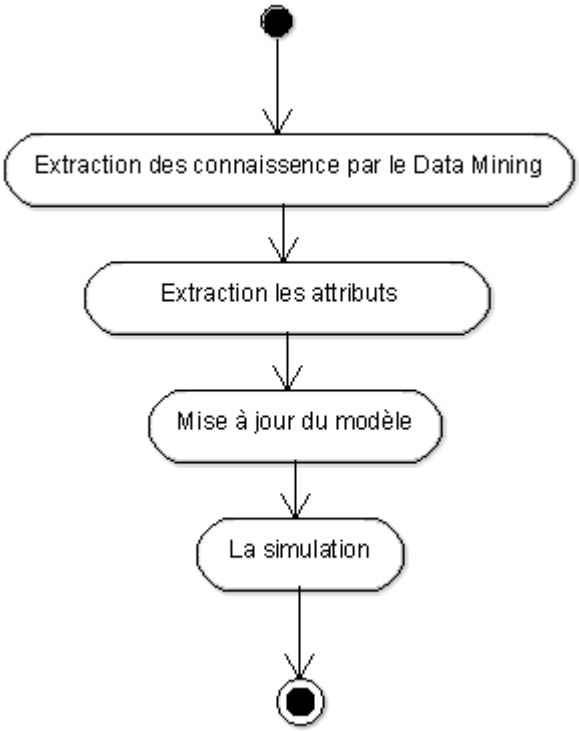


Figure 7 : Diagramme d'activité de la mise à jour du modèle.

❖ **Diagramme de classe**

Le diagramme de classe constitue l'un des pivots essentiels de la modélisation avec UML. Ce diagramme permet de donner la représentation du système à développer. Cette représentation est centrée sur les concepts de classes et d'associations. Chaque classe se décrit par ses attributs et ses méthodes **[PAS, 2008],** tel que c'est décrit dans la **Figure 8**.

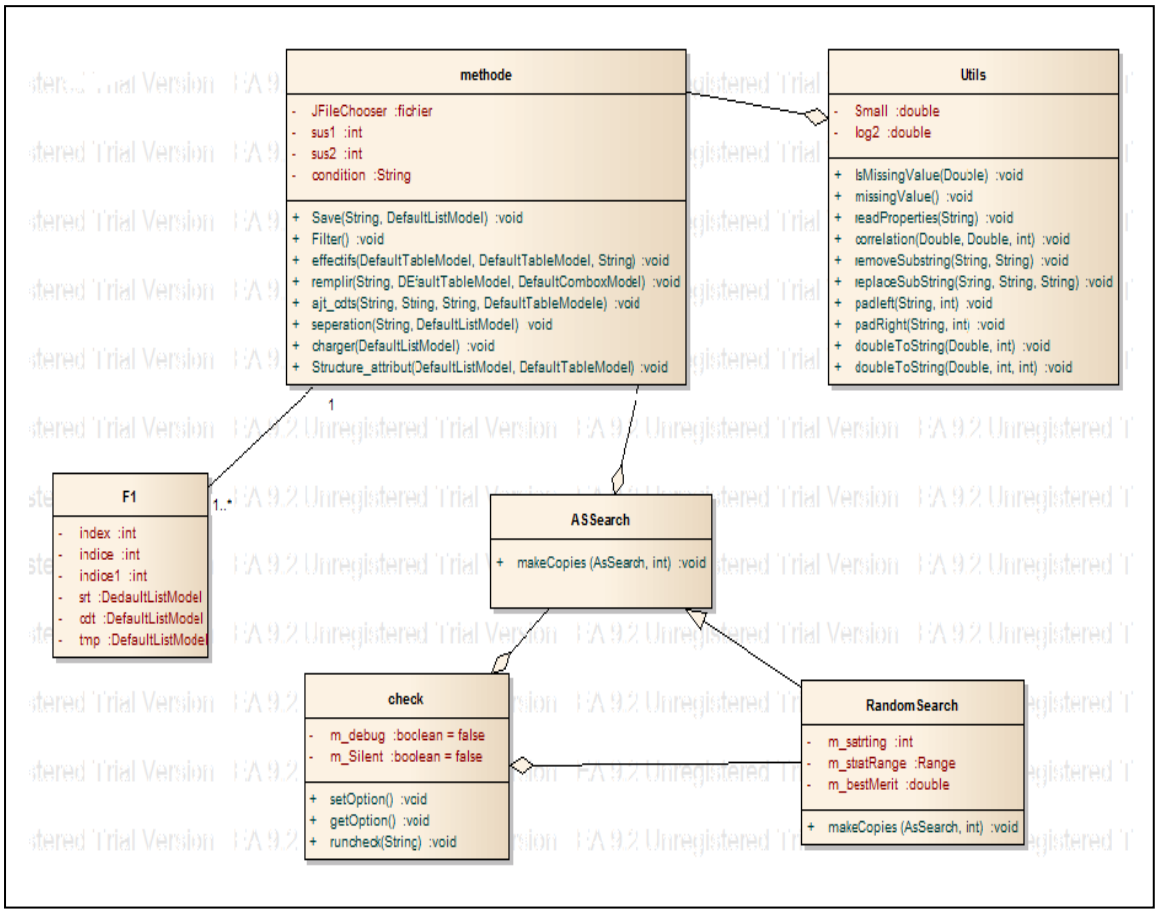


Figure 8 : Diagramme de classe.

2. Implémentation

2.1. Les outils utilisés

- **NetBeans**

C'est un projet open source fondé par Sun Microsystems. L'IDE NetBeans est un environnement de développement permettant d'écrire, compiler, déboguer et déployer des programmes. [SW04]

- **Bio-PEPA et Eclipse Plug-in**

Eclipse est un environnement de modélisation riche pour Bio-PEPA, qui vise à fournir un appui solide pour étudier le comportement dynamique des systèmes. Le Bio-PEPA Eclipse

Plug-in prend en charge le processus expérimental en permettant aux utilisateurs d'exécuter des ensembles de simulations stochastiques indépendantes.

Cela illustre la force d'un langage de modélisation de haut niveau comme un calcul des processus. **[Micro, 2007]**

- **Weka (Waikato Environment for Knowledge Analysis):**

Weka est un logiciel libre qui propose un ensemble d'algorithmes d'apprentissage automatique. Il possède également toute une palette d'outils pour le traitement de données, la sélection d'attributs, la visualisation de distributions, de modèles et de résultats. Il permet à l'utilisateur de créer, exécuter, modifier, et d'analyser les expériences d'une manière plus pratique que possible **[Bridgite, 2011]**.

2.2. Mise en œuvre de l'application

Cette mise en œuvre reprend les étapes décrites dans l'architecture du système de la figure1.

- **Création du modèle initiale**

Développé un modèle avec Bio-PEPA est la première étape de notre travail, tel que c'est illustré dans la figure 9, cette dernière n'est qu'une partielle reproduction de la figure 02 (pour la clarté du code), où le nombre de personnes sont définis selon leur statut par les variables suivantes: Susceptible X, infecté (tuberculeux pulmonaire) TP, tuberculeux pulmonaire traiter TPT, tuberculeux pulmonaire non traiter TPNT, décéder DCD, les guéri GUERI, traitement échouer ECHEC, les perdus de vus PV, les transférés TRANSFERE, traitement terminer TRTfini. Cette étape démontre l'importance de l'utilisation d'un tel outil.

```
p= 0.494;
Beta = 0.897;
      .
      .
      .
nt= 1;
sizeLocal=336;

location world : size =336, type = compartment;
location Local in world : size = sizeLocal, type = compartment;


kineticLawOf TP_Traite :   Beta * TP@Local;
      .
      .
      .
kineticLawOf PV_X :  Mu2 * PV@Local;

X = (susceptible_infected,1) << X + (Gueri_X,1) >> X + (DCD_X,1) >> X+(PV_X,1) >> X;
      .
      .
      .
PV = (PV_Non_Traite,1) >> PV + (PV_X,1) << PV;


X@Local[336]<*> ... <*> PV@Local[16]
```

Figure 9: le modèle tuberculose par la syntaxe Bio-PEPA.

La figure 10 décrit la première partie du code qui consiste à déclarer les paramètres qui sont les taux de passage d'un état vers un autre comme suit.

```
p= 0.494;
Beta = 0.897;
alpha = 0.102;
teta1= 0.034;
teta2 = 0.376;
teta3= 0.013;
teta4 = 0.577;
teta5 = 0.059;
teta6 = 0.941;
MuT= 1;
Mu= 1;
Mu1 = 1;
Mu2 = 1;
nt= 1;
sizeLocal=336;
```

Figure 10 : déclaration des paramètres.

La figure 11, représente l'environnement où les individus sont situés, il est à remarqué qu'il y a un seul compartiment où se situent tous les individus.

```
location world : size =336, type = compartment;
location Local in world : size = sizeLocal, type = compartment;
```

Figure 11 : Location.

La figure 12, décrit les fonctions de transition des individus d'un état vers un autre qui sont en fonction des paramètres définis dans la figure 10.

```
kineticLawOf susceptible_infected : p * X@Local * TP@Loca
kineticLawOf TP_Traite :   Beta * TP@Local;
kineticLawOf TP_Non_Traite :   alpha * TP@Local;
kineticLawOf Echec_Traite :  teta1 * TPT@Local;
kineticLawOf Gueri_Traite : teta2 * TPT@Local;
kineticLawOf Transfere_Traite : teta3 * TPT@Local;
kineticLawOf Trait_termine_Traite : teta4 * TPT@Local;
kineticLawOf DCD_Non_Traite : teta5 * TPNT@Local;
kineticLawOf PV_Non_Traite : teta6 * TPNT@Local;
kineticLawOf Echec_TP :  MuT * Echec@Local;
kineticLawOf Gueri_X :  nt * Gueri@Local;
kineticLawOf TRTfini_TP :  Mu * TRTfini@Local;
kineticLawOf DCD_X :  Mu1 * DCD@Local;
kineticLawOf PV_X :  Mu2 * PV@Local;
```

Figure 12 : Les taux fonctionnels.

La figure 13, représente l'évolution des espèces selon les fonctions exécutées, où, par exemple (susceptible_infected) est exécutée par un susceptible (X) et qui exprime le contact entre un susceptible et un infecté, suite à cette exécution le nombre des susceptibles va diminuer.

```
X = (susceptible_infected,1) << X + (Gueri_X,1) >> X + (DCD_X,1) >> X+(PV_X,1) >> X;
TP = (susceptible_infected,1) >> TP +(TP_Traite,1) << TP+ (Echec_TP,1) >> TP +
     (TRTfini_TP,1) >> TP + (TP_Non_Traite,1) << TP;
TPT= (TP_Traite,1) >> TPT + (Echec_Traite,1) << TPT+ (Gueri_Traite,1) << TPT +
     (Transfere_Traite,1) << TPT +(Trait_termine_Traite,1) << TPT;
TPNT = (TP_Non_Traite,1) >> TPNT +(DCD_Non_Traite,1) << TPNT+(PV_Non_Traite,1) << TPNT ;
Echec = (Echec_Traite,1) >> Echec + (Echec_TP,1) << Echec;
Gueri= (Gueri_Traite,1) >> Gueri + (Gueri_X,1) << Gueri;
Transfere=(Transfere_Traite,1) >> Transfere;
TRTfini=(Trait_termine_Traite,1) >> TRTfini + (TRTfini_TP,1) << TRTfini;
DCD = ( DCD_Non_Traite,1) >> DCD + (DCD_X,1) << DCD;
PV = (PV_Non_Traite,1) >> PV + (PV_X,1) << PV;
```

Figure 13 : Les espèces.

La figure 14, illustre l'initialisation des différentes espèces ainsi que le mode d'interaction.

```
X@Local[336]<*> TP@Local[166] <*> TPT@Local[149] <*> TPNT@Local[17]<*> Echec@Local[5]<*>
Gueri@Local[56]<*> Transfere@Local[2]<*> TRTfini@Local[86]<*> DCD@Local[1]<*> PV@Local[16]
```

Figure 14 : Le modèle.

- **Simulation du modèle initial**

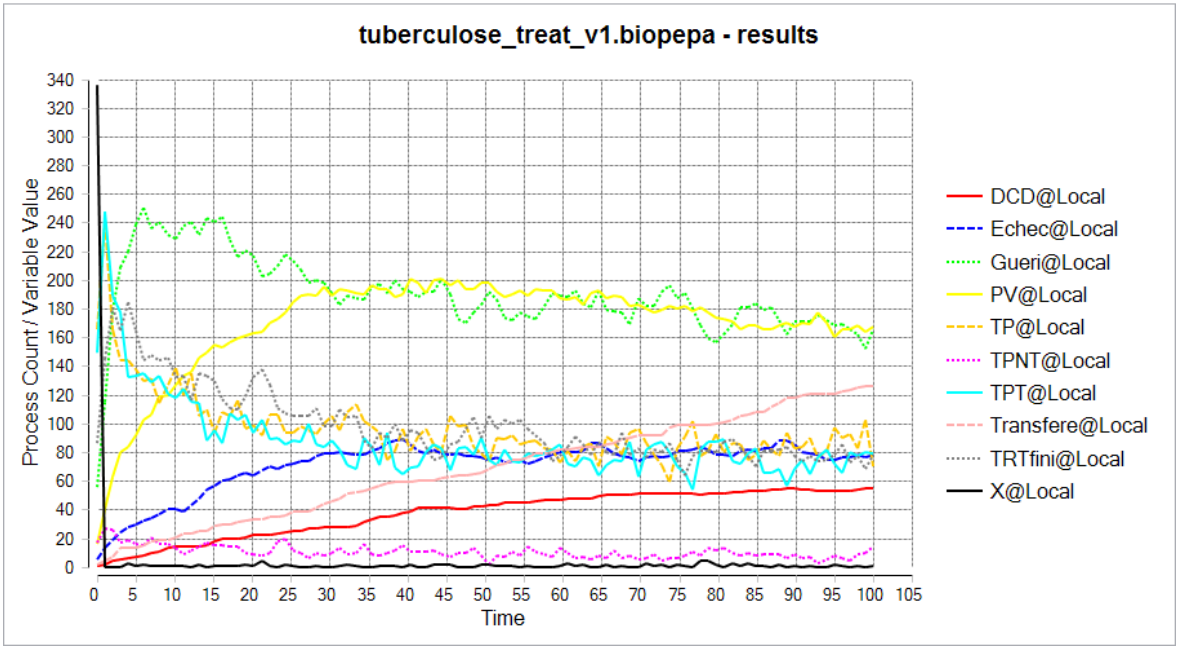


Figure 15: Simulation du modèle.

Selon le modèle dans la figure 2 nous allons interpréter le graphe comme suit :

On voit que les susceptible peuvent passés soit à l'état tuberculeux pulmonaire ou l'état tuberculeux extra pulmonaire mais l'état extra pulmonaire n'est pas repris dans notre étude, la raison pour la quel on le voit pas dans le graphe.

Par exemple au fur et à mesure que les susceptibles diminuent (noir), les tuberculeux pulmonaires (orange) augmentent.

- **Analyse de l'expert**

Selon l'analyse de l'expert, on observe que le nombre des transférés (résistant au traitement) n'arrête pas de s'accroître et que cette observation est belle et bien confirmée par la réalité, sauf que l'expert ne peut prendre de décision du fait il ne sait à quel agir.

- **Extraction des connaissances**

Après l'analyse de l'expert il a été quasi impossible de détecter le facteur déterminant à l'augmentation des transférés, la raison pour la quel nous avons acheminé notre travail vers le Data Minig ce dernier montre la deuxième partie de notre travail sous weka.

- **Prétraitement de la base de données**

Afin d'entamer notre travail nous avions besoin de nettoyer et traiter la base de donnée de la tuberculose (figure 16), ensuite la convertir en un fichier .arff (figure 17) pour qu'on puisse l'exploiter sous Weka.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	ID1	AGE	AGE1	SEXE	MOIS	TRIM	ADRES	QART	COM	SS	LOC	SIEG	SIEG1	REGIM	TYPMAL	PREVTEP
1		30	30	M		1	1	10 RUE AYACHI BEL	ARSA	MOST	2	TP	PUL		1	NOV
2		38	40	M		1	1	BT 4 Nø2 400 LGT	ROUTE O	MOST	1	TEP	CEREBRAL	6	3	NOV
3		14	10	F		1	1	BLOC C Nø100	ARSA	MOST	2	TEP	ADENITE	2	3	NOV
4		40	40	M		1	1	RUE 1 Nø26	MONT PL	MOST	3	TEP	OSSEUSE	4	1	NOV
5		20	20	F		1	1	HAY SI NOUREDDI	HAY SI NO	A NOUI	5	TEP	ADENITE	2	3	NOV
6		26	30	M		1	1	BELLE VUE Nø14	BLLE VUE	A NOUI	5	TP	PUL		1	NOV
7		70	70	F		1	1	HAY DJEBLI MED N	MONT PL	MOST	3	TP	PUL		1	NOV
8		34	30	F		1	1	BT A2 Nø8 15 348	348 LGT	MOST	2	TEP	CUTANEE	7	3	NOV
9		26	30	M		1	1	BT A Nø11 60 LGT	TIDJDITT	MOST	2	TP	PUL		1	NOV

Figure 16: Base de données brute.

Viewer

Relation: zahiaaaa11

No.	AGE1 Numeric	SEXE Nominal	TRIM Numeric	QART Nominal	COM Nominal	LOC Nominal	SIEG Nominal	REGIM Numeric	TYPMAL Nominal	EXAMTEP Nominal	EXAMTP Nominal	BACIL1 Nominal	BACIL2 Nominal	BACIL3 Nominal	**ARETTRT** Nominal
1	30.0	M	1.0	ARSA	MOST	TP	PUL	1.0	NOV	NULL	M+	M-	NULL	NULL	TRT TERM
2	40.0	M	1.0	ROUT...	MOST	TEP	CEREB...	3.0	NOV	SCANNER	NULL	NULL	NULL	NULL	DCD
3	10.0	F	1.0	ARSA	MOST	TEP	ADENITE	3.0	NOV	CYTOPO...	NULL	NULL	NULL	NULL	TRT TERM
4	40.0	M	1.0	MONT ...	MOST	TEP	OSSE...	1.0	NOV	CULT PUS	NULL	NULL	NULL	NULL	PV
5	20.0	F	1.0	HAY S...	A NOUI	TEP	ADENITE	3.0	NOV	CYTOPO...	NULL	NULL	NULL	NULL	GUERI
6	30.0	M	1.0	BLLE VUE	A NOUI	TP	PUL	1.0	NOV	NULL	M+	M-	M-	M-	GUERI
7	70.0	F	1.0	MONT ...	MOST	TP	PUL	1.0	NOV	NULL	NF	NULL	NULL	NULL	GUERI
8	30.0	F	1.0	348 LGT	MOST	TEP	CUTA...	3.0	NOV	IDRT	NULL	NULL	NULL	NULL	GUERI

Undo OK Cancel

Figure 17: Base de données après le prétraitement.

- **Classification**

D'après la figure 18 on remarque que les résultats d'algorithme des règles d'associations (Apriori) ne nous aident pas dans notre travail. C'est pour cela que nous nous sommes dirigé vers la classification avec les arbres de décision.

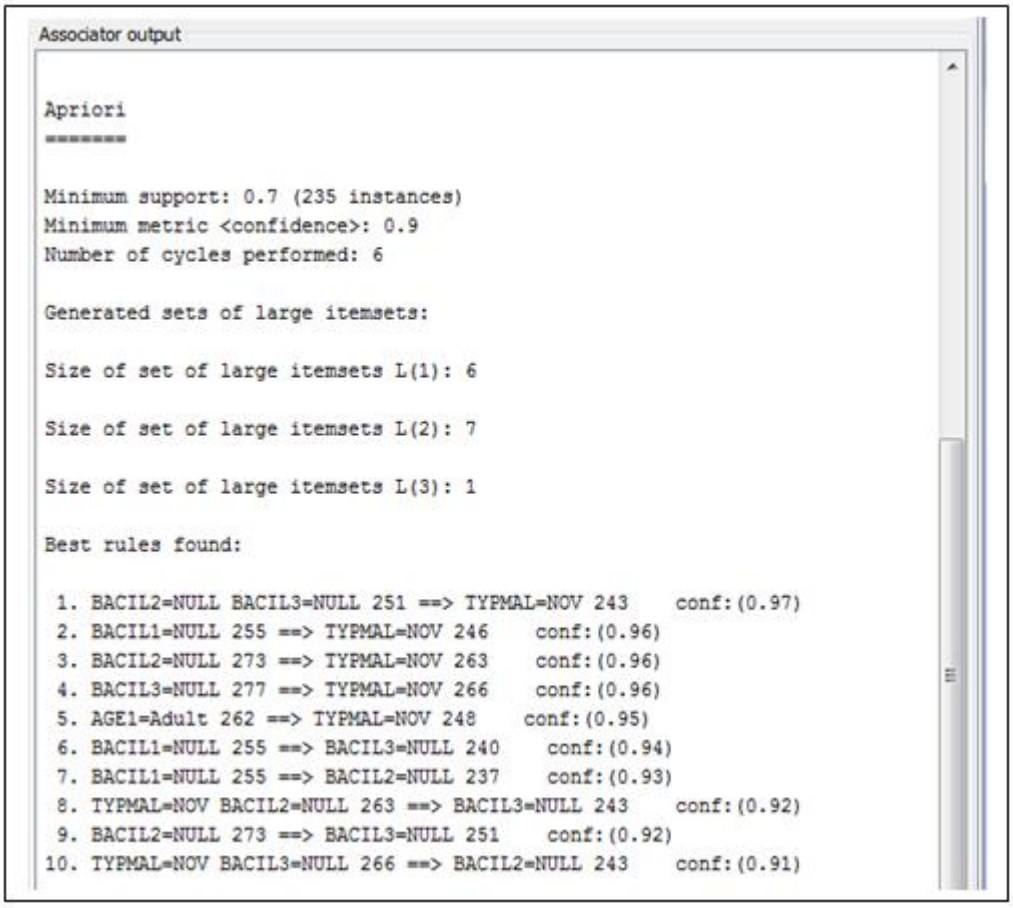


Figure 18 : Résultat de l'algorithme des règles d'association.

Apres lancement de plusieurs algorithmes des arbres de décision on a choisi SimpleCart selon le taux de succès de la classification, tel que le montre le tableau 2.

Algorithme	**Tous les attributs de la BDD**	**Après le filtrage de la BDD**
DecisionStump	59.8214 %	53.5714 %
J48	59.2262 %	53.5714 %
J48 graft	59.2262 %	53.5714 %
Ladtree	60.7143 %	44.0476 %
REPTree	60.7143 %	53.5714 %
SimpleCart	58.9286 %	75.5952 %

Tableau 1: Résultats des classifications.

- **Développement et application**

Pour aboutir a notre but nous avons développé une application que nous allons la démontrer par un exemple pour mieux comprendre son fonctionnement.

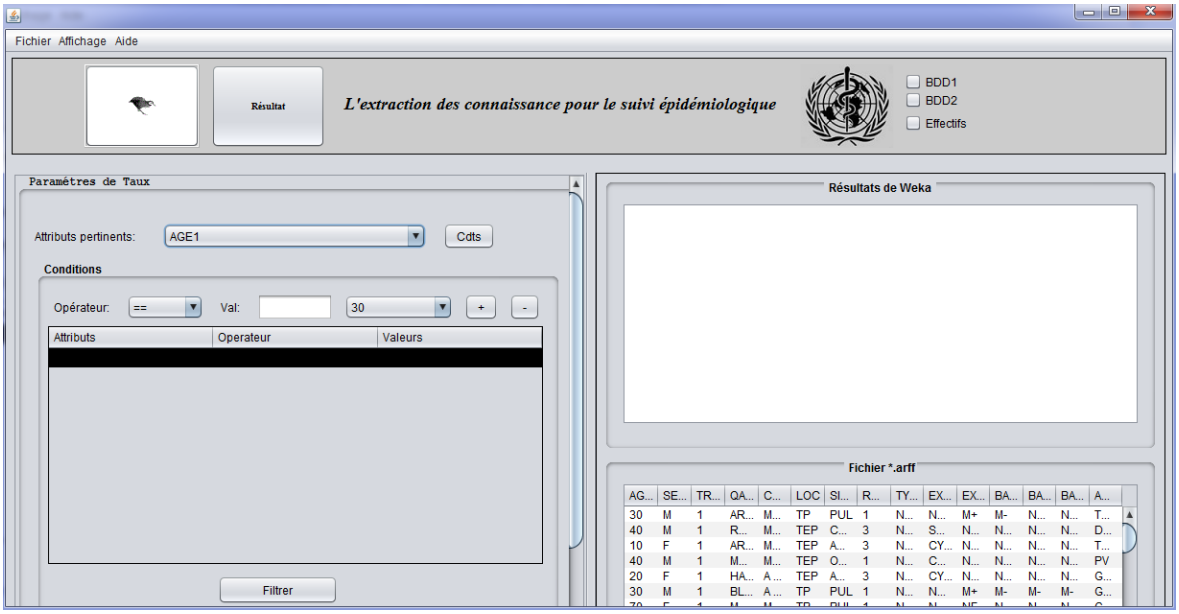


Figure 19 : l'application.

La figure 19, schématise l'interface de notre application intitulée « extraction des connaissances pour le suivi épidémiologique ».

Les figures suivantes décrivent le déroulement de notre application.

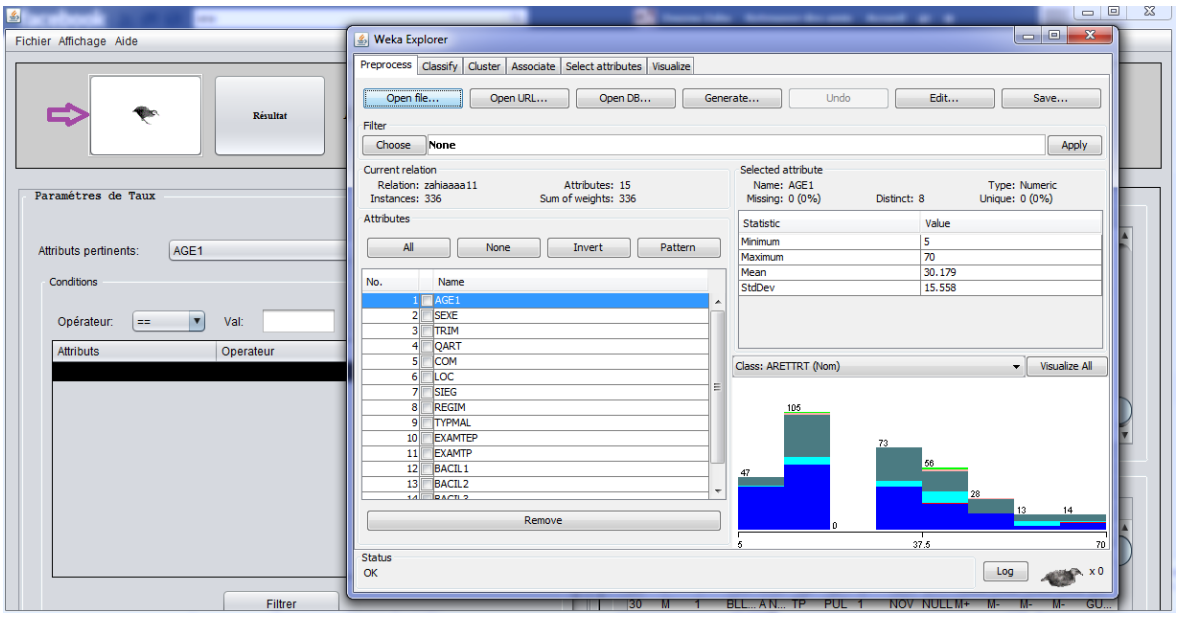


Figure 20 : Lancement de Weka.

1. Nous avons crée un bouton (figure 20), qui fait l'appel au logiciel Weka qui est intégré dans notre application.

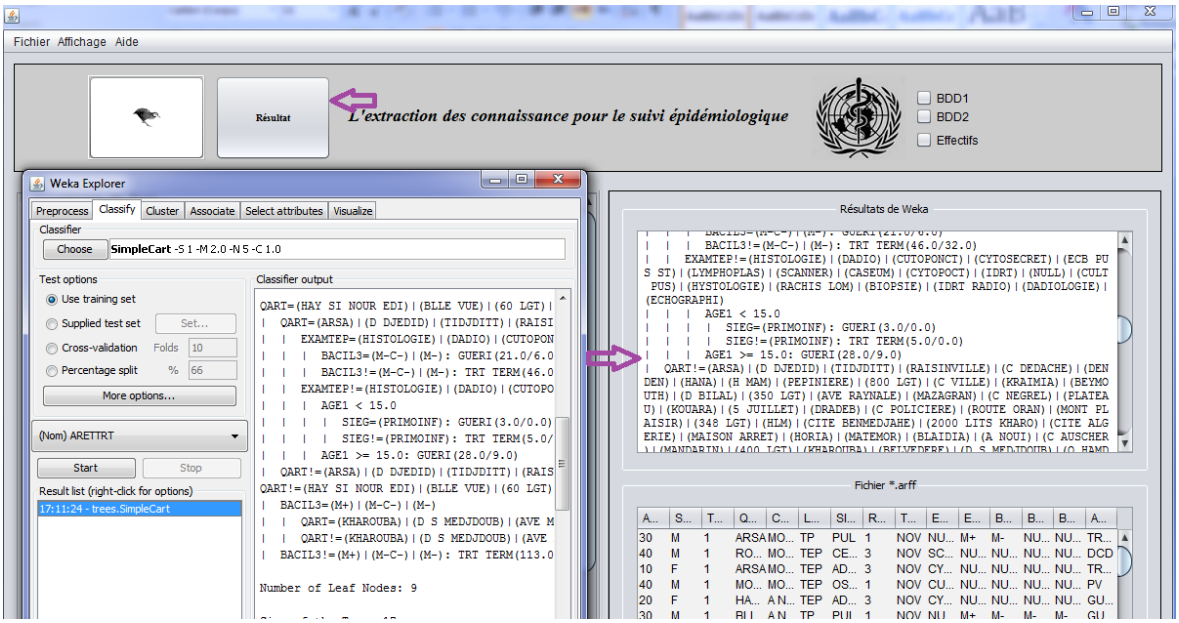


Figure 21: Enregistrement du résultat de Weka.

2. Une fois la classification terminée par Weka (figure 21), on clique sur le bouton Résultat pour enregistrer le résultat dans notre interface.

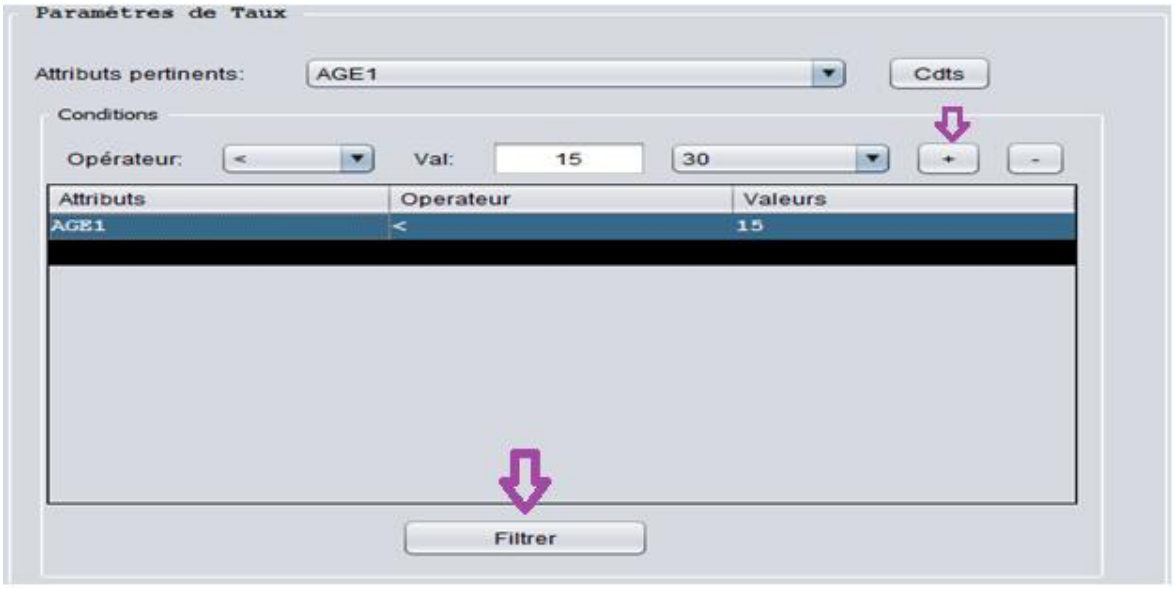


Figure 22: sélection d'attribut pertinent.

3. Après l'analyse du résultat on peut sélectionner les attributs les plus pertinents (AGE1<15 ans) pour filtrer notre base de données (figure 22) et les visualiser (figure 24).

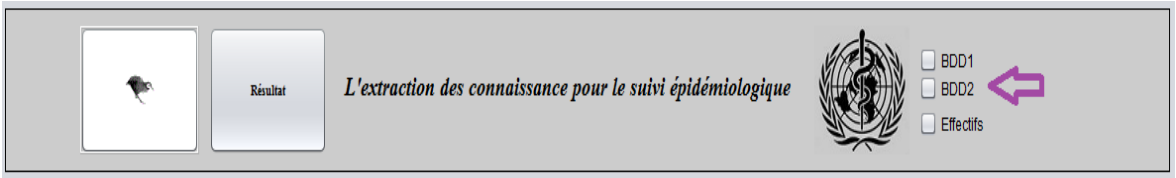
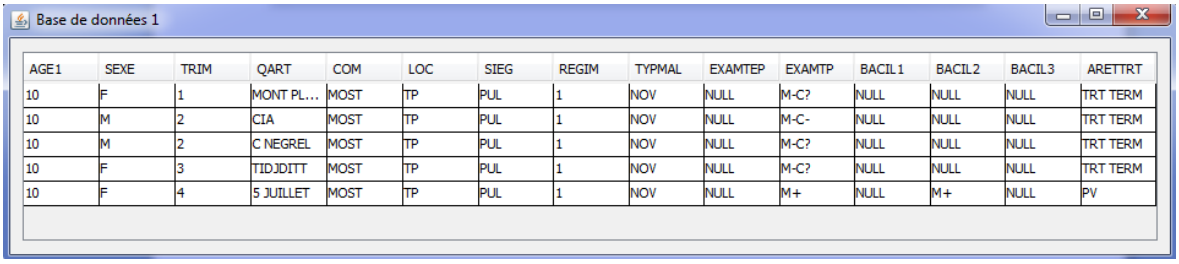
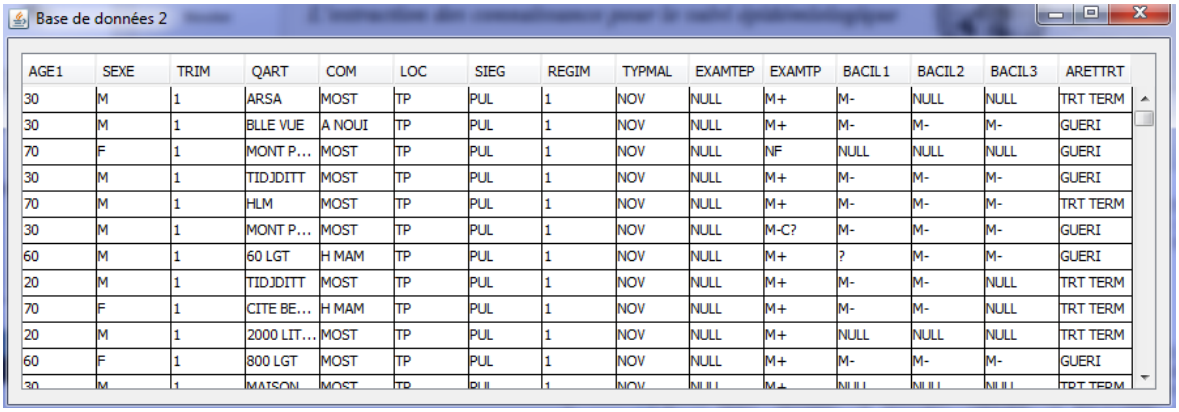


Figure 23 : visualisation.


Base de données 1

AGE1	SEXE	TRIM	QART	COM	LOC	SIEG	REGIM	TYPMAL	EXAMTEP	EXAMTP	BACIL1	BACIL2	BACIL3	ARETTRT
10	F	1	MONT PL…	MOST	TP	PUL	1	NOV	NULL	M-C?	NULL	NULL	NULL	TRT TERM
10	M	2	CIA	MOST	TP	PUL	1	NOV	NULL	M-C-	NULL	NULL	NULL	TRT TERM
10	M	2	C NEGREL	MOST	TP	PUL	1	NOV	NULL	M-C?	NULL	NULL	NULL	TRT TERM
10	F	3	TIDJDITT	MOST	TP	PUL	1	NOV	NULL	M-C?	NULL	NULL	NULL	TRT TERM
10	F	4	5 JUILLET	MOST	TP	PUL	1	NOV	NULL	M+	NULL	M+	NULL	PV

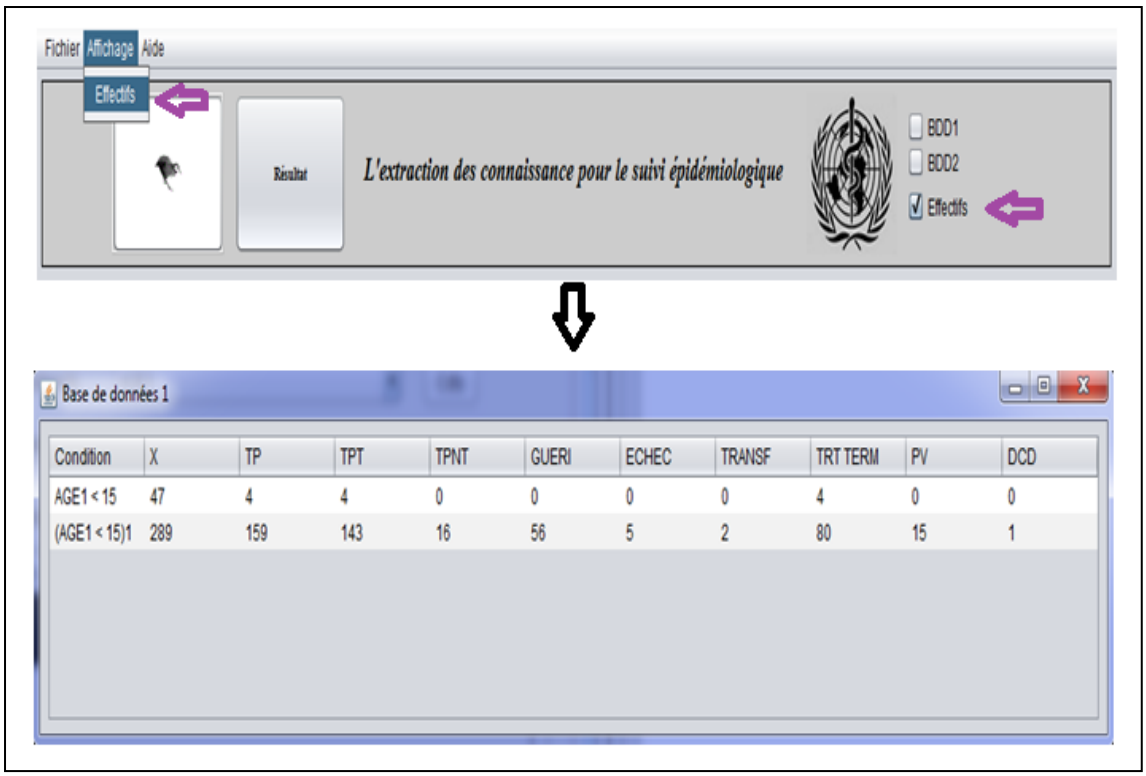
Figure 24 : BDD filtrer par rapport a la condition.


Base de données 2

AGE1	SEXE	TRIM	QART	COM	LOC	SIEG	REGIM	TYPMAL	EXAMTEP	EXAMTP	BACIL1	BACIL2	BACIL3	ARETTRT
30	M	1	ARSA	MOST	TP	PUL	1	NOV	NULL	M+	M-	NULL	NULL	TRT TERM
30	M	1	BLLE VUE	A NOUI	TP	PUL	1	NOV	NULL	M+	M-	M-	M-	GUERI
70	F	1	MONT P…	MOST	TP	PUL	1	NOV	NULL	NF	NULL	NULL	NULL	GUERI
30	M	1	TIDJDITT	MOST	TP	PUL	1	NOV	NULL	M+	M-	M-	M-	GUERI
70	M	1	HLM	MOST	TP	PUL	1	NOV	NULL	M+	M-	M-	M-	TRT TERM
30	M	1	MONT P…	MOST	TP	PUL	1	NOV	NULL	M-C?	M-	M-	M-	GUERI
60	M	1	60 LGT	H MAM	TP	PUL	1	NOV	NULL	M+	?	M-	M-	GUERI
20	M	1	TIDJDITT	MOST	TP	PUL	1	NOV	NULL	M+	M-	M-	NULL	TRT TERM
70	F	1	CITE BE…	H MAM	TP	PUL	1	NOV	NULL	M+	M-	M-	NULL	TRT TERM
20	M	1	2000 LIT…	MOST	TP	PUL	1	NOV	NULL	M+	NULL	NULL	NULL	TRT TERM
60	F	1	800 LGT	MOST	TP	PUL	1	NOV	NULL	M+	M-	M-	M-	GUERI
30	M	1	MAISON	MOST	TP	PUL	1	NOV	NULL	M+	NULL	NULL	NULL	TRT TERM

Figure 25: Le reste de la BDD filtré.

4. Notre base de données s'est partagée en deux bases (figure 25 et 26), la première base contient les instances qui vérifier la condition (AGE1<15 ans) et la deuxième contient le reste des instances (AGE1>= 15ans).



Condition	X	TP	TPT	TPNT	GUERI	ECHEC	TRANSF	TRT TERM	PV	DCD
AGE1 < 15	47	4	4	0	0	0	0	4	0	0
(AGE1 < 15)1	289	159	143	16	56	5	2	80	15	1

Figure 26 : les effectifs.

La figure 26, affiche le tableau des effectifs qui sont calculés à partir des deux base de données filtrées. Il contient le nombre des instances de chaque état.

Figure 27 : L'enregistrement des résultats dans un fichier.txt.

Notre application pour l'extraction des connaissances par le Data Mining ce termine à l'enregistrement du résultat final des effectifs dans un fichier.txt (figure 27), l'utilisation de ce fichier a pour but de calculer les taux pour mettre à jour notre modèle Bio-PEPA.

- **Mise à jour du modèle initiale**

Afin de pouvoir mettre à jour le modèle existant, nous avions dû ajouter un item dans le menu de Bio-PEPA plugin (figure 28).

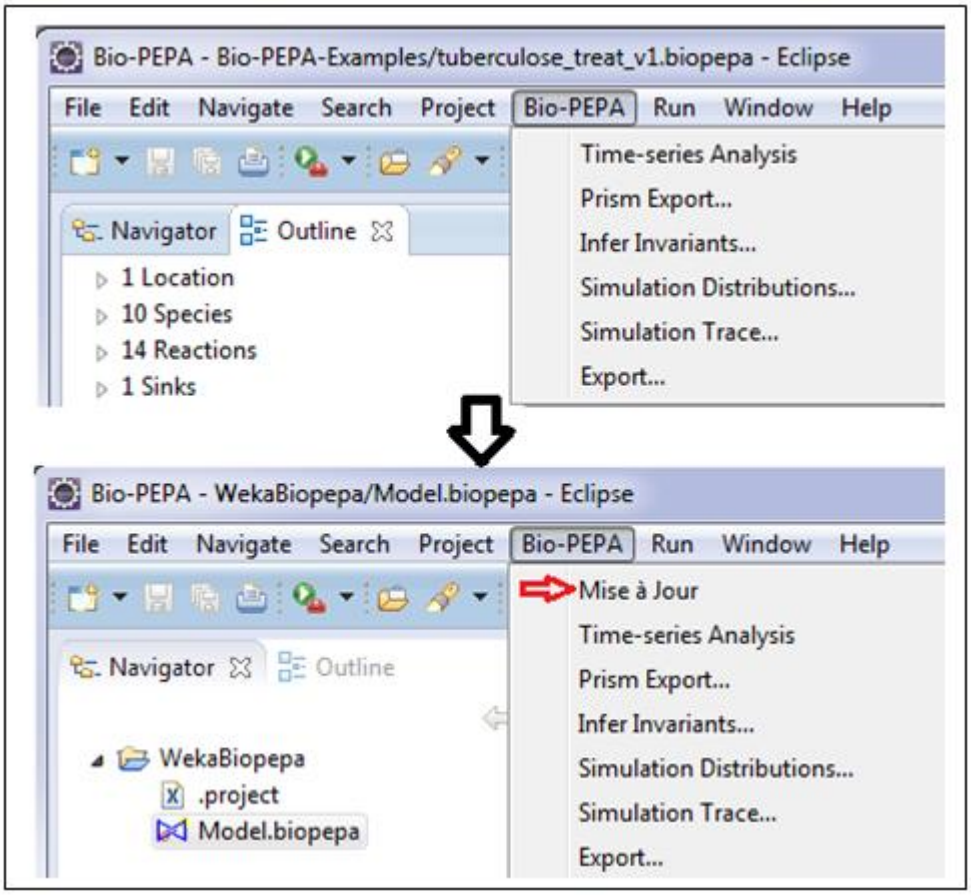


Figure 28 : l'ajout d'un item de mise à jour.

Une fois que l'item « Mise à jour » sélectionné, une fenêtre permettant de sélectionner le fichier (figure 27) résultant de WEKA apparait (figure 29).

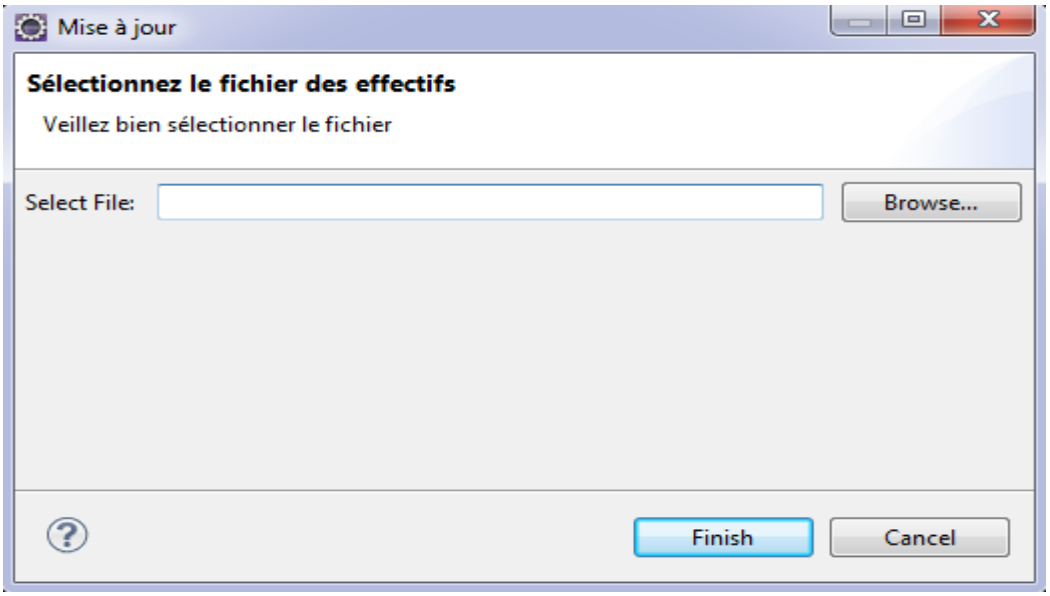


Figure 30 : Sélection du fichier.

La figure 30 permet de produire le nouveau code dans Bio-PEPA adéquat à l'expérimentation faite auparavant.

```
kineticLawOf susceptible_infected1 : p1 * X@Local1 * TP@Local1;
kineticLawOf TP_Traite1 : beta1 * TP@Local1;
kineticLawOf TP_Non_Traite1 : alpha1 * TP@Local1;
     .
     .
     .
kineticLawOf PV_X1 : Mu12 * PV@Local1;

//------------------------------

kineticLawOf susceptible_infected2 : p2 * X@Local2 * TP@Local2;
kineticLawOf TP_Traite2 : beta2 * TP@Local2;
kineticLawOf TP_Non_Traite2 : alpha2 * TP@Local2;
   .
   .
   .
kineticLawOf PV_X2 : Mu22 * PV@Local2;
```

Figure 30 : Le nouveau modèle après la mise à jour.

La figure 30 démontre clairement que par exemple la fonction « susceptible_infected » préconisée pour un seul compartiment a été dupliquée pour les deux compartiments (selon la condition) avec des taux spécifiques.

- **Simulation**

La figure 31 présente le graphe de simulation après la mise à jour.

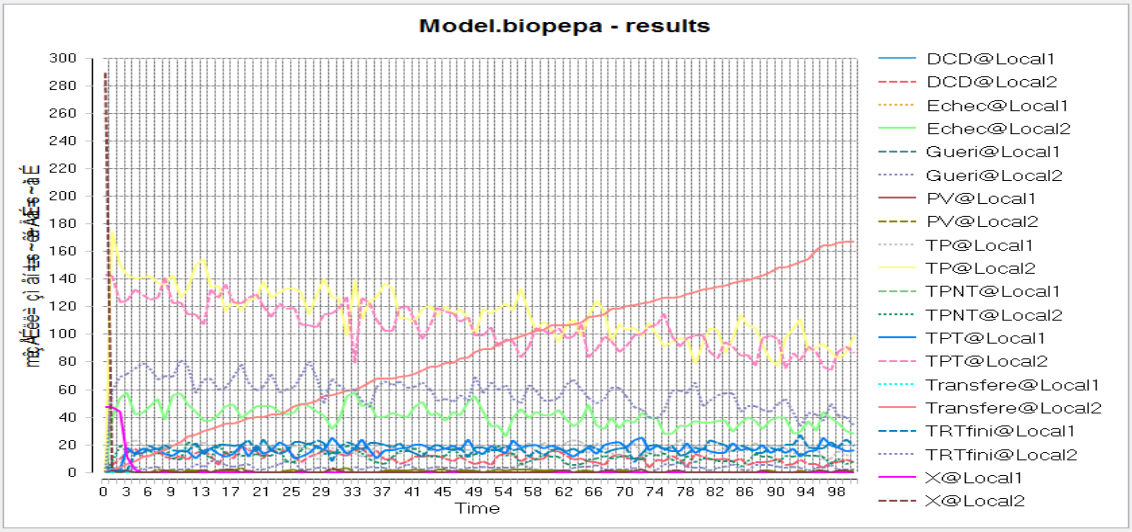


Figure 31 : Graphe de simulation.

Le graphe de la figure 32, permet à l'expert d'aboutir à la conclusion que la croissance des transférés est particulièrement dû à ceux dont l'âge est supérieur ou égal à quinze ans (>=15 ans), et donc il sera primordiale de prendre des décisions afin de protéger ou mettre en quarantaine les concernés.

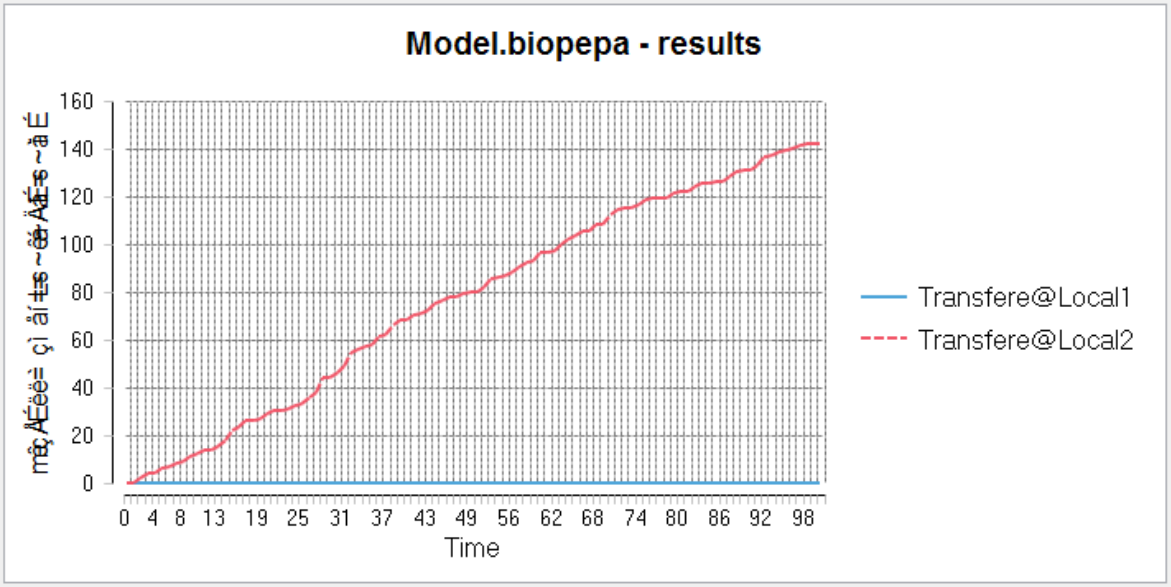


Figure 32 : graphe de simulation des transférés.

Conclusion

Nous avons présenté dans ce chapitre, l'architecture de notre application basée sur la modélisation de la tuberculose par Bio-PEPA, la fouille de données épidémiques par Weka et la mise à jour du modèle Bio-PEPA, en se basant sur les résultats obtenus des différentes expérimentations.

L'application ainsi réalisée, le développeur pourra maintenant, facilement réaliser des modèles épidémiques et les optimiser, sans trop se soucier de la justesse de la récolte d'informations pertinentes ni trop interagir avec l'expert, ce dernier pourra lui aussi prendre les décisions, prévenir et même détecter à temps les facteurs influant la propagation des épidémies dans une population.

Conclusion générale

Nous avons détaillé dans ce manuscrit l'évolution de l'épidémiologie dans le domaine informatique qui s'est imposée par la découverte de facteurs de risque de plusieurs maladies très répandues (notamment le tabac pour le cancer des bronches et les maladies cardiovasculaires.et pulmonaires et extra-pulmonaires . .), par son rôle dans l'identification et la description de maladies émergentes (sida, hépatite C, . . .).

La métaphore du Data Mining signifie qu'il y a des trésors ou pépites cachés sous des montagnes de données que l'on peut découvrir avec des outils spécialisés. Notre études dans ce rapport s'est focalisée sur la découverte de toutes les méthodes et techniques de ce dernier qui ont contribué à l'aide de décision en épidémiologie.

D'une part L'état de l'art ainsi établi nous a permis de conclure que les techniques de data Mining sont utile au suivit épidémiologique. Et d'une autre part les résultats ainsi obtenue ont belle et bien confirmés l'utilité du data Mining en épidémiologie.

Comme perspectives nous proposons d'appliquer les autre techniques et tâches (clusterig, et association) du data Mining afin d'opter pour celles qui sont les plus adéquates à l'étude épidémiologique.

Bibliographie

[A]

[Alain, 2006] Alain-Jacques Valleron RAPPORT SUR LA SCIENCE ET LA TECHNOLOGIE N°23.2006

[Alain, 2008] Alain-Jacques Valleron L'épidémiologie humaine Conditions de son développement en France, 2008.

[Alain, 2003] Alain-Jérôme Fougères , Modele De Communication Pour Des Agents D'assistance Dans Les Systemes Complexes, M3M - Université de Technologie de Belfort-Montbéliard Rue du château Sévenans, 90010 BELFORT – France, (2003).

[Asha, 2011] T. Asha and S. Natarajan and K. N. Balasubramanya Murthy, A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification, CoRR, abs/1108.1045, 2011.

[Attaluri, 2009] Attaluri,, P.K. and Chen, Z. and Weerakoon, A.M. and Lu, G.Integrating decision tree and Hidden Markov Model (HMM) for subtype prediction of human influenza A virus, Cutting-Edge Research Topics on Multiple Criteria Decision Making,P52-58, Springer, 2009 .

[B]

[Bagn, 2002] R. Bagni, R. Berchi, and P. Cariello. A comparison of simulation models applied to epidemics. Journal of Artificial Societies and Social Simulation, 5(3), June 2002.

[Bekri, 2011] Bekri, E. and A Govardhan, F., Association of Data Mining and healthcare domain: Issues and current state of the art, Global Journal of Computer Science and Technology, 11(21), 2011.

[Bouyer, 2003] Bouyer J,cordier S,Levallois épidémiologie. In : environnement et santé publiques-fondement et pratique, pp.89-118, 2003.

[Bridgite, 2011] Brigitte Bigi (LPL - _Equipe C3I) WEKA : c'est quoi ? 15 fevrier 2011

[Brossette, 2000] Brossette SE. et al. Association rules and data mining in hospital infection control and public health surveillance. J Am Med Inform Assoc, **5** (4) :373-81. (1998. 2000).

[Bergstra, 1986] JA Bergstra and JW Klop. Algebra of Communicating Processes. InMathematics and computer science: proceedings of the CWI symposium, November 1983, page 89. North Holland, (1986).

[C]

[**Canlas, 2009**] Canlas Jr, R.D.Data Mining in Healthcare: Current Applications and Issues, MS in Information Technology thesis, 2009.

[Car, 1978] CAR Hoare. Communicating sequential processes. (1978).

[Ciocchetta, 1996] (Federica Ciocchetta _ and Jane Hillston) Bio-PEPA: a framework for the modelling and analysis of biological systems J. Hillston. A Compositional Approach to Performance Modelling, Cambridge University Press, 1996.

[Ciocchetta, 2008] An Automatic Mapping from the Systems Biology Markup Language to the Bio-PEPA Process Algebra par Kanimozhi Ellavarason UNIVERSIT A DEGLI STUDI DI TRENTO, Anno Accademico 2007/2008

[Ciocchetta, 2010] Ciocchetta.F, Hillston..J "Bio-PEPA for Epidemiological Models". Electr. Notes Theor. Comput. Sci. 261: 43-69 (2010).

[Codd, 1967] F. Codd. Cellular Automata. Academic Press, (1967).

[Calder, 2006]M. Calder, J. Hillston, and S. Gilmore. Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA. Lecture Notes in Computer Science, 4230:1–23, (2006).

[D]

[Deepak, 2012] Deepak Dangwal, Papender Kumar, Nidhi puri, ISSN: 0976–7754 & E-ISSN: 0976–7762 , Volume 3, Issue 1, 2012, pp-133-135, 2012.

[Deng, 2004] Deng H Z, Chi Y, Tan Y J. Multiagent-based simulation of disease infection (in Chinese). Comput Sim, 21: 167–170, (2004).

[Didier, juin1998] Didier Nackache « data werehouse et data mining », conservatoire national des arts et métiers de Lille, juin 1998.

[Didier, décembre1998] Didier Nackache « data mining sur internet », conservatoire national des arts et métiers de Lille, décembre 1998.

[G]

[**Gupta, 2009**] R.K.Gupta, D.P.Agrawal, "Improving the performance of association rule mining algorithms by filtering insignificant transactions dynamically. Asian J.Inform.Manage., 3:7-17, 2009.

[H]

[Hand, 1998] Hand DJ. Data mining : statistics and more ? The American Statistician, **52** : 112-118**,** 1998.

[Hebrail, 2003] Hebrail G., Lechevallier Y. Data Mining et Analyse des donnees. In :Analyse des données, G. Govaert (ed.). Hermes, 323-355. (2003).

[J]

[Jiawei, 2000] Jiawei Han and Micheline Kamber, Morgan Kaufmann Publishers « Data Mining: Concepts and Techniques » 550 pages. ISBN 1-55860-489-8, August 2000.

[K]

[Kumar, 2012] Kumar.P, Dangwal.D, & Pur.N DIAGNOSIS OF TUBERCULOSIS USING ASSOCIATION RULE METHOD, *3*(1), 133–135, 2012.

[Keeling, 2008]Keeling, M. J. & Rohani, P. "Modeling infectious diseases in humans and animals", Princeton University Press, 2008.

[Kuttler, 2006]C.N.J. Kuttler, J. Niehren, and R. Blossey. Gene regulation in the pi calculus: Simulating cooperativity at the lambda switch.Lecture notes in computer science, 4230:24–55, (2006).

[L]

[Lavrac, 1999] Lavrac. N Selected techniques for data mining in medicine. Artificial Intelligence in Medicine, **16** : 3-33, 1999.

[LAU, 2008]Laurent Debrauwer, Fien Van Der Heyde. " UML2, initiation exemples et exercices corrigés', édition ENI, 2008

[Liu, 2004]Liu Y, Chen Y. Simulation and analysis on the control of SARS by complexity adaptive system theory (in Chin ese). Complex Sys Com-plexity Sci, 1: 74–79 (2004).

[M]

[MacQueen, 1967] J. MacQueen. Some methods for classification and analysis of multivariate observations. In 5th Berkeley symposium on Mathematical statistics and probability, pages 281–297, 1967.

[M-C, 2006] Dr M-C Picot Unité Recherche Clinique et Épidémiologie Département de l'Information Médicale à l'épidémiologie – Montpellier 2006_2007(Décembre 2006).

[Micro, 2007] Mirco Tribastone. The PEPA Plug-in Project. In Mor Harchol-Balter, Marta Kwiatkowska, and Miklos Telek, editors,Pro-ceedings of the 4th International Conference on the Quan-titative Evaluation of SysTems (QEST), pages 53–54. IEEE, September (2007).

[O]

[Oechslein, 2001]C. Oechslein, A. Hornlein, and F. Klugl. Evolutionary Optimization of Societies In Simulated Multi-Agent Systems. In Modelling Artificial Societies and Hybrid Organizations; workshop on the ECAI, (2001).

[P]

[PAS, 2008] Pascal Roques. « UML2 par la pratique », édition ERYROLLES, 2008.

[Parunak, 1998]H.Parunak, R. Savit, and R. Riolo. Agent-Based Modeling vs. Equation-Based Modeling: A Case Study and Users' Guide. In Proceedings of Multi-agent systems and Agent- based Simulation (MABS 98), (1998).

[R]

[Rameshkumar, 2011] K.Rameshkumar (2011),” extracting association rules from hiv infected patients' treatment dataset”, Trends in Bioinformatics 4 (1):35-46, ISSN 1994-7941 / DOI: 10.3923/tb.2011.35.46, Asian Network for Scientific Information.

[Rémi, 2004] Rémi Gilleron, Marc Tommasi, « découverte de connaissances à partir de données », Université de lille3, 2004.

[Ricco, 2010] Ricco RAKOTOMALALA Laboratoire ERIC Université Lumière Lyon2.

[Roy, 1991] Roy M. Anderson and Robert M. May. Infectious Diseases of Humans: Dynamics And Control. Oxford University Press, (1991).

[S]

[Saporta, 2000] Saporta G. Data Mining and Official Statistics, Quinta Conferenza Nationale di Statistica, ISTAT, Rome, 2000.

[Silvi, 2002] Silvi.s, warembourg.P « DATAMINING : Etude et analyse des ventes d’une chaine de magasin », Université Paris Dauphine, 2002.

[Smitha, 2012] T.Smitha, Sundaram, V. Classification Rules by Decision Tree for Disease Prediction. International Journal of Computer Applications, 43(8), 6–12. doi:10.5120/6121-8323, 2012.

[T]

[**Thangavel, 2006**] Thangavel, K., Jaganathan, P.P. and Easmi, P.O. Data Mining Approach to Cervical Can- cer Patients Analysis Using Clustering Tech- nique. Asian Journal of Information Technol- ogy (5) 4, 413-417, 2006.

[TIBICHE, 2012] Dr Arézki TIBICHE, Maître Assistant en Epidémiologie, Faculté de Médecine, Université Mouloud Mammeri de Tizi Ouzou, Service d’Epidémiologie et de Médecine Préventive, CHU de Tizi Ouzou 2012

[Z]

[ZHA, 10] Qi Zhang , Lu Cheng and Raouf Boutaba. “Cloud computing: state-of-the-art and research challenges”. The Brazilian Computer Society , pages 1-10, 2010.

Les sites

[WS01] http://perso.limsi.fr/jps/enseignement/examsma/2005/3.simulation_1/carriere.html

[WS02] http://www.infirmiers.com/pdf/cours-en-vrac/tuberculose.html

[WS03] http://www.who.int/mediacentre/factsheets/fs104/fr vu le 03_2012

[WS04] http://www.techno-science.net/?onglet=glossaire&definition=5346

[WS05] http://www.who.int/fr/

[WS06]http://wwwA.VERGNENEGRESIME/AV/Courssp/4èannée/polycop2010-2011/épidémio.html