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 : Biologie Appliquée

# *MEMOIRE de fin d'étude*

Présenté en vue de l'obtention du diplôme de Master

**Domaine :** Sciences de la nature et de la vie

**Filière :** Sciences biologiques

**Option :** Pharmacotoxicologie

# Prédiction de faisabilité de médicament en utilisant l'apprentissage automatique

**Présenté par :**

- **Mlle. Ikram BOUSBA**

**Devant le jury**

| | | |
|---|---|---|
| **Dr. Ben Lakhel Ammar** | **M.A.A Université de Tébessa** | **Président** |
| **Dr. Sahraoui Abdelatif** | **M.C.A. Université de Tébessa** | **Examinateur** |
| **Dr. Yahiaoui Ayoub** | **M.C.B. Université de Tébessa** | **Promoteur** |

***Année universitaire 2021/2022***

# Remerciements

Tous mes remerciements s’adressent tout d’abord à tout puissant d’Allah d’avoir guider mes pas vers le chemin de savoir.

A travers ce modeste travail, je tiens à remercier vivement mon encadreur **Dr. Ayoub YAHIAOUI** pour ses conseils.

Merci à monsieur **Bilel GHRISSI** pour leurs encouragements.

Je remercie les membres du jury d’avoir accepté d’examiner et d’évaluer mon travail

# Dédicace

**PAPA** personne ne mérite plus que toi que je lui dédie ce travail.

A **Maman** pour ta tendresse et tes prières tout au long de mes études.

A mes frères et sœurs pour leurs soutiens.

A mes amis pour les beaux moments que nous avons passés ensemble.

Je dédie ce travail à tous les jours difficiles que j'ai vécus au cours de ces cinq années à l'université et à la résidence.

A ma jolie **Bascuta** que je l'aime trop

Et aujourd'hui je peux dire que je dédie ce travail à moi-même.

**IKRAM**

## Résumé

Cette étude comparative a l'objectif pour prédire de faisabilité d'un médicament (streptomycine) et choisir le meilleur algorithme (classifier) parmi ces quatres (Naïve bayes ; oneR ; tree J48 ; random forest) selon les fonctions de prédiction ; le score de rappel et F1 score, j'ai utilisé le logiciel de weka et le data set extrait de « British Médical Journal de 1948, intitulé Streptomycin Treatment of Pulmonary Tuberculosis »

La streptomycine est le premier antibiotique ayant eu une action sur mycobactérie de TB et maintenant est un traitement de deuxième ligne pour le traitement de cette maladie qui est asymptomatique parfois sur le plan général particulièrement la Tb pulmonaire, ses anomalies biologiques une accélération de la vitesse de sédimentation, un leucocyte minime, une anémie peuvent s'observer.



- **Mots clés**

IA, ML, Prédiction De Faisabilité De Médicament, Streptomycine, Tuberculose, Cross Validation, Naïve Bayes, OneR, Treej48, Random Forest.

# Abstract

This comparative study aims to predict the feasibility of a drug (streptomycin) and choose the best algorithm (classify) among these four (Naïve Bayes; oneR; tree J48; random forest) according to the prediction functions; the recall score and F1 score, I used the weka software and the data set extracted from "British Medical Journal of 1948, entitled Streptomycin Treatment of Pulmonary Tuberculosis"

Streptomycin is the first antibiotic having had an action on TB mycobacteria and is now a second-line treatment for the treatment of this disease, which is sometimes asymptomatic on the general level, particularly pulmonary TB, its biological abnormalities an acceleration of the speed of sedimentation, a minimal leucocyte, anemia can be observed.



- **Keywords**

IA, ML, Drug Feasibility Prediction, Streptomycin, Tuberculosis, Cross Validation, Naive Bayes, OneR, Treej48, Random Forest

**ملخص**

تهدف هذه الدراسة المقارنة إلى التنبؤ بجدوى عقار (الستربتومايسين) واختيار أفضل خوارزمية وفقًا لوظائف التنبؤ؛ (naïve bayes ; one R ;J48 ;random forest) (تصنيف) من بين هذه الاربعة ومجموعة البيانات المستخرجة من "المجلة weka ، استخدمت برنامجF1 درجة الاستدعاء ودرجة Streptomycin Treatment of Pulmonary Tuberculosis" الطبية البريطانية لعام 1948، بعنوان الستربتوميسين هو أول مضاد حيوي له تأثير على بكتيريا السل وهو الآن علاج الخط الثاني لعلاج هذا المرض الذي يكون أحيانًا بدون أعراض على المستوى العام، وخاصة السل الرئوي، وتشوهاته البيولوجية وتسريع سرعة الترسيب، الحد الأدنى من الكريات البيض، يمكن ملاحظة فقر الدم.

كلمات مفتاحية: الذكاء الاصطناعي ; التعليم الاوتوماتيكي ; التنبؤ بمفعول الدواء ; الستريبتوميسين ;السل ;خوارزميات.

# Liste Des abréviations

- **IA :** Intelligence Artificielle
- **ML :** Machine Learning
- **Weka :** Waikato Environnement For Knowledge Analysis
- **Strep g :** Streptomycine Par Gramme
- **PAS g :** Para Amino Salycilate
- **ESR :** Erythrocytes Sédimentation Rate

- # Liste des Figures

- # Liste des Tableaux

# • Table des matières

# INTRODUCTION

# Introduction

L'intelligence artificielle (IA) est devenue le nouveau terme que l'on entend tous les jours ces dernières années, l'IA en général définit la capacité d'une machine capable d'agir par elle-même et qui n'est pas explicitement programmée pour reproduire des actions ou des fonctions qui sont généralement celles des êtres humains. Aujourd'hui, on la retrouve dans nos machines informatiques, les réseaux sociaux, les transports et dans le secteur médical.

L'application de l'IA en médecine permettant à la machine d'analyser les données par elle-même et de fournir des estimations, dans le but de prédire de nombreuses maladies afin que les médecins puissent intervenir le plus rapidement possible pour réduire le risque de complications des maladies sur la santé du patient lutter contre la mort prématurée.

L'apprentissage automatique est une discipline de l'intelligence artificielle qui cherche à trouver un moyen de créer des programmes informatiques qui s'améliorent automatiquement avec l'expérience.

A travers ce mémoire de Master, j'intéresserais à l'utilisation des algorithmes d'apprentissage automatique pour la prédiction de faisabilité de médicament.

La méthode utilisée dans ce travail est l'application les différents algorithmes de classification d'apprentissage supervisé aux données extrait de British Médical Journal de 1948, intitulé Streptomycin Treatment of Pulmonary Tuberculosis et déduire le meilleur algorithme qui donnera comme résultat une classification des patients en termes de taux de la précision et de la sensibilité du modèle. Ce travail est organisé en trois principaux chapitres comme suit:

- Le 1$^{er}$ chapitre présente un aperçu général sur la tuberculose et la streptomycine.
- Le 2 $^{eme}$ chapitre donne un aperçu sur l'apprentissage automatique, les algorithmes d'apprentissage supervisé.
- Le dernier chapitre présente d'abord une étude technique dans laquelle je définis l'environnement logiciel utilisé, puis une définition détaillée de la base du donnée utilisée. Ensuite, les résultats sont présentés, comparés et interprètes.

A la fin, ce travail est clôture par une conclusion générale résumant les idées fondamentales que j'ai apportée.

# CHAPITRE I

## 1. La tuberculose

### 1.1.Définition

La tuberculose est une maladie contagieuse résultant des effets pathogènes sur l'organisme d'un bacille tuberculeux qui appartient au genre Mycobacterium tuberculosis (Bacille de koch), beaucoup plus rarement que Mycobacterium bovis ou Mycobacterium africanum **(Chrétien et *al.*, 1990).**

La variété la plus répandue est le bacille de type humain: Mycobacterium tuberculosis. Dans les régions d'élevage, les bovidés peuvent être infectés par une autre variété Mycobacterium bovis transmissible à l'homme **(Bouref, 1987).**

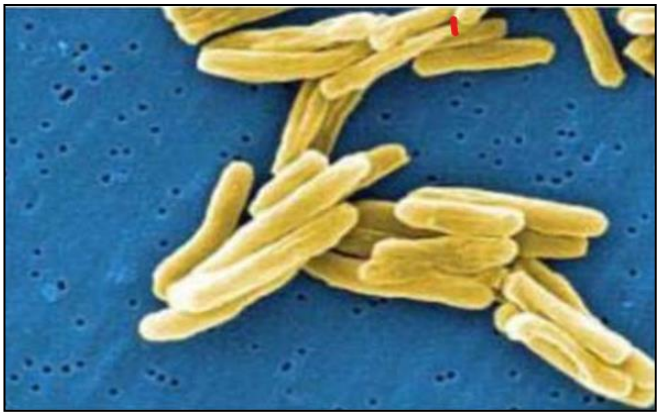


**Figure 1**: Le bacille de koch au microscope électronique **(Marcel et *al.*, 2000).**

La tuberculose primaire est en général asymptomatique, mais peut se manifester par un état fébrile, une perte pondérale et une baisse de l'état général, parfois aussi d'adénopathies hilaires unilatérales, d'un infiltrat parenchymateux et/ou d'un épanchement pleural. La tuberculose primaire peut s'accompagner d'un érythème noueux, sous forme de nodules rouges et douloureux sur la face antérieure des jambes. De telles manifestations s'observent plus souvent chez les enfants en bas âge ou les personnes immunodéprimées **(Brandli, 1998).**

## 2. Etude clinique

### 2.1.Signes généraux

Il existe une altération de l'état général (asthénie, anorexie, amaigrissement pouvant dépasser 10 kg). La fièvre est d'aspect variable, allant d'une fébricule vespérale (cas le plus

fréquent) à une fièvre élevée, oscillante, accompagnée de frissons et de sueurs nocturnes. Ces symptômes peuvent persister plusieurs semaines avant que le patient ne consulte.

**2.2.Signes fonctionnels et physiques**

Ils sont dominés par la toux ne cédant pas au traitement symptomatique. Elle peut être sèche ou productive, parfois hémoptoïque (10 % des cas). Elle s'accompagne parfois d'une dyspnée ou d'une douleur thoracique.

L'auscultation pulmonaire est souvent peu parlante, contrastant avec l'importance des signes fonctionnels.

À ce stade de la démarche diagnostique, il convient d'imposer au patient un isolement respiratoire, tant que la recherche de BK à l'examen direct n'a pas permis d'écarter un risque de contagiosité.

## 3. La tuberculose pulmonaire

La tuberculose de réactivation est habituellement caractérisée par une toux lentement progressive sur des semaines ou des mois. Cette toux échappe facilement à l'attention si le malade est tabagique **(Brandli, 1998 ; Janssens et *al.*, 1999).**

Elle est due à la dissémination par voie bronchique des bacilles à partir du nodule de primo-infection ; elle touche préférentiellement les lobes supérieurs et les segments postérieurs, qui sont les mieux ventilés. ; dans les cas d'atteinte pulmonaire, l'examen physique apporte peu d'indices. La fièvre est présente chez deux tiers environ des malades. Les anomalies biologiques, par exemple une accélération de la vitesse de sédimentation ou une augmentation du taux de l a protéine C-réactive, une leucocytose minime, une lymphopénie ou une anémie peuvent s'observer mais ne peuvent pas être utilisé pour établir le diagnostic. Le médecin doit penser à la possibilité d'une tuberculose chez les malades qui accusent des symptô mes suspects (toux persistante depuis plusieurs semaines, amaigrissement, sudations nocturnes) **(Cohen et *al.*, 1996 ; Tattevin et *al.*, 1999).**

## 4. Tuberculose extra-pulmonaire

Elle est souvent asymptomatique sur le plan général. Les manifestations extra pulmonaires les plus connues sont les lymphadénites tuberculeuses, la tuberculose pleurale, la

tuberculose uro-urinaire, osseuse, méningée et la tuberculose miliaire qui résulte d'une dissémination hématogène diffuse des mycobactéries en un ou plusieurs points de l'organisme et surtout aux poumons d'éléments nodulaires de petite taille d'origine tuberculeuse. **(Sharma et *al.*, 2005).**

## 5. Traitement de la tuberculose

### 5.1. Traitement curatif

La tuberculose est une maladie guérissable. Le seul traitement efficace est la polychimiothérapie. La durée du traitement est de 6 à 8 mois, répartie en deux phases **(Boucherit, 2012).**

Phase initiale intensive : pendant les deux premiers mois on utilise l'association de quatre molécules : l'éthambutol, la rifampicine, l'isoniazide et le pyrazinamide.

Phase de continuation de 6 mois : comprend deux molécules : l'éthambutol et l'isoniazide.

Dans le cas de retraitement on utilise l'association de quatre molécules pendant 8 mois (L'éthambutol, la rifampicine, l'isoniazide, le pyrazinamide) avec la streptomycine pendant les deux premiers mois de traitement. Aucun de ces médicaments essentiels n'est suffisamment efficace pour détruire tous les bacilles tuberculeux se trouvant chez un malade ; c'est pourquoi l'association de plusieurs médicaments antituberculeux est indispensable pour obtenir la guérison définitive d'un malade. Les médicaments antituberculeux essentiels sont au nombre de cinq **(Leclercq, 2006)**: La streptomycine (SM), L'isoniazide (INH), Le pyrazinamide (PZA), La rifampicine (RIF), L'éthambutol (EMB).

### 5.2. Traitement préventif

La prévention de la tuberculose passe par les mesures d'hygiène préventive et surtout par la vaccination au BCG qui est pratiquée à la naissance, avec un rappel à 6 ans en cas d'absence de cicatrice vaccinale et à tout âge devant un test à la tuberculine négatif **(Bouziani, 2002)**

## 6. La streptomycine

La streptomycine est le premier antibiotique cytostatique et cytotoxique de la classe des aminosides (ou aminoglycosides) découvert. C'est un antibiotique à spectre large pouvant

réagir avec les bacilles gram négatifs, avec certaines cocci gram positifs ou avec certaines mycobactéries.

La streptomycine est isolée en 1943 à partir d'actinobactérie Streptomyces griseus par un Américain, Albert Schatz, à l'époque étudiant. Mais le mérite de cette découverte rejaillit sur son professeur, Selman Waksman, qui obtient le prix Nobel de physiologie ou médecine en 1952. Cet antibiotique fait partie de la liste des médicaments essentiels de l'Organisation mondiale de la santé. Il agit en perturbant la synthèse protéique en se fixant dans les ribosomes bactériens. **(Boucherit, 2012).**

Elle est administrée traditionnellement par voie intramusculaire et dans certains pays par voie intraveineuse4. Historiquement cet antibiotique est le premier ayant eu une action sur Mycobacterium tuberculosis dans le traitement de la tuberculose.

La streptomycine est maintenant un traitement de deuxième ligne pour le traitement de cette maladie et ne doit être employée que pour des formes multi résistantes ou particulières de la tuberculose.

## 7. Usage

La streptomycine est également utilisée dans le traitement des endocardites infectieuses occasionnées par les entérocoques insensibles à la gentamicine et dans le traitement de la tularémie.

Pour réduire le développement de bactéries résistantes et maintenir l'efficacité de la streptomycine et d'autres médicaments antibactériens, la streptomycine ne doit être utilisée que pour traiter ou prévenir des infections bactériennes.

La streptomycine, comme beaucoup d'aminoglycosides, présente des risques de néphrotoxicité accentués lorsque le patient souffre de dysfonctionnements rénaux13. Cet antibiotique ne doit pas être utilisé chez la femme enceinte, il peut en effet entrainer une lésion du nerf auditif et une néphrotoxicité chez le fœtus.

Il est recommandé de ne pas associer à la streptomycine d'autres médicaments ototoxiques ou néphrotoxiques comme d'autres antibiotiques de la classe des aminosides. La streptomycine potentialise l'effet des inhibiteurs neuromusculaires comme le curare employé lors des anesthésies.

## 8. Pharmacocinétique

La streptomycine s'administre par injection intramusculaire qui peut entrainer l'apparition d'abcès stérile au point d'injection.

La concentration sérique est atteinte au bout d'une heure. Pour une injection de 500mg, la concentration maximale est de 20µg/mL ; pour une injection de 1g, la concentration maximale passe à 40µg/mL.

La streptomycine dispose d'un bon taux de diffusion au niveau des poumons, des reins. En revanche, la molécule diffuse assez mal dans le liquide céphalo-rachidien et les tissus, elle ne franchit les méninges qu'en cas d'inflammation. Cet antibiotique est présent dans le placenta et le lait maternel.

# CHAPITRE II

## 1. Historique

L'apprentissage artificiel est une discipline jeune, à l'instar de l'informatique et de l'intelligence artificielle. Il se situe au carrefour d'autres disciplines : philosophie, psychologie, biologie, logique, mathématique .les premières études remontent à des travaux de statistique dans les années 1920. C'est après la seconde guerre mondiale que les premières expériences deviennent possibles. Se développent ensuite dans les années 1960 les approches connexionnistes avec des perceptrons, et la reconnaissance des formes.

La mise en évidence des limites du perceptron simple arrête toutes les recherches dans ce domaine jusqu'à la renaissance dans les années 1980. Les années 1970 sont dominées par des systèmes mettant l'accent sur les connaissances, les systèmes experts, Les limites de tels systèmes se font sentir dans les années 1980, pendant lesquelles a lieu le retour du connexionnisme avec un nouvel algorithme d'apprentissage. **(A. Cornuéjols, L. Miclet, et Y.Kodratoff, 2002).**

## 2. L'apprentissage automatique (ou artificiel) (*machin e-learning* en anglais)

Est un des champs d'étude de l'intelligence artificielle.

L'apprentissage artificiel fait référence à la capacité d'un système à acquérir et intégrer de façon autonome des connaissances. Cette notion englobe toute méthode permettant de construire un modèle de la réalité à partir de données, soit en améliorant un modèle partiel ou moins général, soit en créant complètement le modèle.

L'apprentissage automatique fait référence au développement, l'analyse et l'implémentation de méthodes qui permettent à une machine (au sens large) d'évoluer et de remplir des tâches associées à une intelligence artificielle grâce à un processus d'apprentissage. Cet apprentissage permet d'avoir un système qui s'optimise en fonction de l'environnement, les expériences et les résultats observés Voyons quelques exemples. La capacité d'apprentissage est une caractéristique des êtres vivants. De la naissance à l'âge adulte, les êtres vivants acquièrent de nombreuses capacités qui leur permettent de survivre dans leur environnement. L'apprentissage d'un langage, de l'écriture et de la lecture sont de bons exemples des capacités humaines, et des phénomènes mis en jeu : apprentissage par cœur, apprentissage supervisé par d'autres êtres humains, apprentissage par généralisation.

**(Cornuéjols, L. Miclet, et Y.Kodratoff, 2002).**

### 3. Les types d'apprentissage automatique

Les algorithmes d'apprentissage peuvent se catégoriser selon le mode d'apprentissage qu'ils emploient :

#### 3.1.Apprentissage supervisée

L'algorithme est entrainé en utilisant une base de données d'apprentissage contenant des exemples de cas réels traités et validés. L'objectif est de trouver des corrélations entre les données d'entrée (variables explicatives) et les données de sorties (variables à prédire), pour ensuite inférer la connaissance extraite sur des entrées avec des sorties inconnues. **(Bachelier, L. 1900).**

Il existe deux modèles d'apprentissage supervisé :

**a). Classification**

Dans l'apprentissage automatique et les statistiques, la classification est le problème qui consiste à identifier à quel groupe de catégories (sous-populations) une nouvelle observation appartient, à partir d'un ensemble d'apprentissages contenant des données (ou instances) dont l'appartenance à une catégorie est connue. Par exemple, attribuer un courrier électronique donné à la classe "spam" ou "non-spam" et attribuer un diagnostic à un patient donné en fonction des caractéristiques observées du patient (sexe, pression artérielle, présence ou non de certains symptômes, etc.). La classification est considérée comme un exemple d'apprentissage supervisé et de reconnaissance de formes, c'est-à-dire un apprentissage dans lequel un ensemble de formations d'observations correctement identifiées est disponible. La procédure non supervisée correspondante est connue sous le nom de clustering. Elle consiste à regrouper des données en catégories en fonction d'une mesure de la similarité inhérente ou de la distance.

**b). Régression**

L'analyse de régression est largement utilisée pour la prévision, pour comprendre quelles variables indépendantes sont liées à la variable dépendante et pour explorer les formes de ces relations. Dans des circonstances restreintes, une analyse de régression peut être utilisée pour déduire des relations de cause à effet entre les variables indépendantes et dépendantes. **(Bachelier, L. (1900).**

### 3.2.apprentissage non supervisée

Pour ce type d'apprentissage la base de données d'apprentissage ne contient pas de variable cible (comme on l'a vu en apprentissage supervisé). Il y a seulement un ensemble de données collectées en entrée. L'algorithme doit découvrir par lui-même la structure en fonction des données. On utilise cette technique pour partitionner les données en groupes d'éléments homogènes. La distance est souvent la plus utilisée comme mesure de similarité entre les groupes. **(Bachelier, L. (1900).**

#### a). Clustering

En clustering, la catégorie de l'objet est inconnue. Cependant, nous connaissons la règle à classer (généralement basée sur la distance) et nous connaissons également les caractéristiques (variables indépendantes) pouvant décrire la classification de l'objet. Il n'y a pas d'exemple de formation pour examiner si la classification est correcte ou non. Ainsi, les objets sont assignés à des groupes simplement basés sur la règle donnée. **(N. Rakesh, B. U. Maheswari, et S. K. Srivatsa 2014 ).**

### 4. L'apprentissage semi supervisé

Il s'agit d'un mixe entre l'apprentissage supervisé et non supervisé en utilisant des données étiquetées et non-étiquetées pour le même ensemble de données. L'avantage d'utiliser cette approche réside dans le fait que l'étiquetage de données peut être couteux et prend souvent beaucoup de temps. En plus, il pourra entrainer un biais humain dans les données étiquetées. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, est très pratique. Et le fait d'inclure un grand nombre de données non étiquetées au cours du processus d'entrainement a tendance à améliorer la performance du modèle final tout en réduisant le temps et les coûts consacrés à sa construction. **(Bachelier, L. 1900).**

### 4.1.L'apprentissage par renforcement

L'apprentissage se fait sans supervision, par interaction avec l'environnement (principe d'essai / erreur) et, en observant le résultat des actions prises. Chaque action de la séquence est associée à une récompense. Le but est de déterminer la stratégie comportementale optimale afin de maximiser la récompense totale. Pour cela, un simple retour des résultats est nécessaire pour apprendre comment la machine doit agir. Ceci est appelé le signal de renforcement. Il peut être très avantageux pour la prévision financière à haute fréquence où

l'environnement est dynamique et en conséquence, il est difficile de trouver ou d'automatiser manuellement des stratégies efficaces. **(Bachelier, L. 1900).**

## 5. Les algorithmes utilisés

### 5.1.Naïve bayes

Naïve bayésienne fait partie des algorithmes d'apprentissage automatique supervisé qui sont principalement utilisés pour la classification. C'est un classificateur probabiliste simple basé sur l'application de théorème de bayes et qui aide à construire des modèles d'apprentissage automatique rapides qui peuvent faire des prédictions rapides.

Naïve dans l'algorithme se réfère à l'hypothèse naïve que l'algorithme fait, qui est que chaque fonctionnalité est indépendante des autres fonctionnalités

### 5.2.Théorème de bayes

Le théorème de Bayes (alternativement la loi de Bayes ou la règle de Bayes) d´écrit la probabilité d'un évènement, basée sur la connaissance préalable des conditions qui pourraient être liées a` l'´évènement. La formule est comme suit :

- **P (AB) = P (B/A)P (A)**
- **P (B)**

Où :

- **P (A/B):** la probabilité conditionnelle que l'évènement A se produise, étant donné que B s'est produit. Ceci est également connu comme la probabilité postérieure.
- **P (B /A):** la probabilité conditionnelle que l'´évènement B se produise, étant donné que A s'est produit.
- **P (A) et P (B):** probabilité de A et B sans égard l'un a l'autre.

### 5.3.Avantage de Naïves Bayes

Fonctionne également bien dans la prédiction multi-classes.

Lorsque l'hypothèse d'indépendance est vérifiée, un classificateur Naïve Bayésienne fonctionne mieux que d'autres modèles.

Fonctionne mieux que les modèles plus compliqués lorsque l'ensemble de données est petit.

### 5.4.Inconvénient de Naïves Bayes

Limitation de Naïve Bayésienne est l'hypothèse de fonctionnalités indépendantes. Dans la vraie vie, il est presque impossible d'obtenir un ensemble de fonctionnalités complétement indépendantes

## 6. Radom Forest (forêts aléatoires)

Radom Forest ou forêts aléatoires est un algorithme d'apprentissage supervisé très populaire Il est également utilisée pour les problèmes de régression ou de classification. Basé sur un ensemble des algorithmes d'apprentissage, qui est un processus de combinaison de plusieurs algorithmes pour résoudre un problème complexe et améliorer les performances du modèle. C'est un algorithme qui créer de nombreux arbres de décision (c'est la raison pour laquelle il est appelé une forêt) sur divers sous-ensembles de l'en- semble de données. Elle prend la prédiction de chaque arbre et sur la base des votes majoritaires des prédictions, et elle prédit le résultat final.

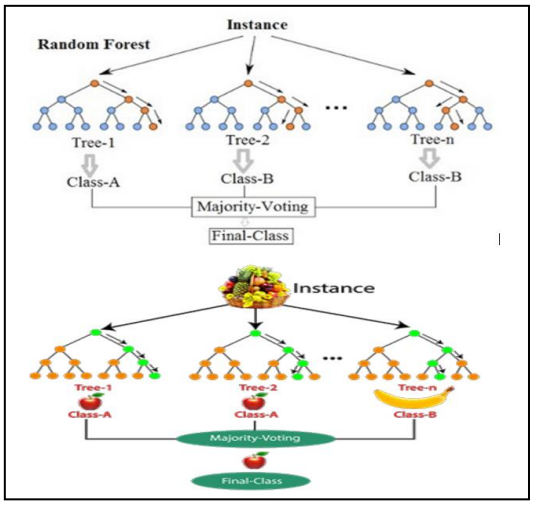



**Figure 02 :** Structure de l'algorithme Radom Forest.

### 6.1.L'interprétation d'exemple

Dans cet exemple l'ensemble de données contenant un ensemble d'images de fruits classifiées par l'algorithme Radom Forest, cette ensemble est divisé en sous-ensembles et donné à chaque arbre de d´décision et dans la phase d'apprentissage chaque arbre produit un résultat de prédiction, et lorsqu'un nouveau point de données se produit, puis sur la base de la

majorité des résultats, Random Forest prédit la décision finale (comme l'exemple dans l'image).

### 6.2.Algorithme de construction de Random Forest

- S´sélectionnez des échantillons aléatoires à partir d'un ensemble de données d'entrainement.
- Créer des arbres de décision pour chaque échantillon (sous-ensembles).
- Ensuite on obtient le résultat de prédiction de chaque arbre de d´décision.
- Pour les nouveaux points le vote sera effectué pour chaque résultat prédit
- Sélectionnez le résultat de prédiction le plus vote comme résultat de prédiction final.

### 6.3.Avantage de Random Forest

- Il s'agit de l'un des algorithmes d'apprentissage les plus précis disponibles. Pour de nombreux ensembles de données, il produit un classificateur très précis.
- Il fonctionne efficacement sur de grandes bases de données.
- Il dispose d'une m´méthode efficace pour estimer les données manquantes et maintient la précision lorsqu'une grande partie des données sont manquantes.

### 6.4.Inconvénients de Random Forest

Le principal inconvénient de l'algorithme Random Forest est qu'un grand nombre d'arbres peut rendre l'algorithme trop lent et inefficace pour les prédictions en temps réel. En général, ces algorithmes sont rapides a` entrainer, mais assez lents à créer des prédictions une fois qu'ils sont formes. Une prévision plus précise n´nécessite plus d'arbres, ce qui entraine un modèle plus lent

- **OneR :** abréviation de "One Rule", est un algorithme de classification simple mais précis qui génère une règle pour chaque prédicteur dans les données, puis sélectionne la règle avec la plus petite erreur totale comme "une règle". Pour créer une règle pour un prédicteur, nous construisons un tableau de fréquence pour chaque prédicteur par rapport à la cible. Il a été démontré que OneR produit des règles à peine moins précises que les algorithmes de classification de pointe tout en produisant des règles simples à interpréter pour les humains.
- **Algorithme OneR**

Pour chaque prédicteur,Pour chaque valeur de ce prédicteur, établissez une règle comme suit ;

- Compter la fréquence d'apparition de chaque valeur de cible (classe)
- Trouver la classe la plus fréquente
- Faire en sorte que la règle affecte cette classe à cette valeur du prédicteur
- Calculer l'erreur totale des règles de chaque prédicteur
- Choisissez le prédicteur avec la plus petite erreur totale.

**Exemple :**

Trouver le meilleur prédicteur avec la plus petite erreur totale à l'aide de l'algorithme OneR basé sur des tables de fréquences associées.

**Tableau 1 :** prédicteur de oneR

Which one is the best predictor ?

| Outlook | Temp | Humidity | Windy | Play Golf |
|---|---|---|---|---|
| Rainy | Hot | High | False | No |
| Rainy | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Sunny | Mild | High | False | Yes |
| Sunny | Cool | Normal | False | Yes |
| Sunny | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Rainy | Mild | High | False | No |
| Rainy | Cool | Normal | False | Yes |
| Sunny | Mild | Normal | False | Yes |
| Rainy | Mild | Normal | True | Yes |
| Overcast | Mild | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Sunny | Mild | High | True | No |

**Tableau 2 :** Les tableaux de fréquence

| ★ | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Outlook | Sunny | 3 | 2 |
| | Overcast | 4 | 0 |
| | Rainy | 2 | 3 |

| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Temp. | Hot | 2 | 2 |
| | Mild | 4 | 2 |
| | Cool | 3 | 1 |

| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Humidity | High | 3 | 4 |
| | Normal | 6 | 1 |

| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Windy | False | 6 | 2 |
| | True | 3 | 3 |

**Tableau 3 :** Meilleur prédicteur c'est

| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Outlook | Sunny | 3 | 2 |
| | Overcast | 4 | 0 |
| | Rainy | 2 | 3 |

IF Outlook = Sunny THEN PlayGolf = Yes
IF Outlook = Overcast THEN PlayGolf = Yes
IF Outlook = Rainy THEN PlayGolf = No

## 7. Contribution des prédicteurs

Simplement, l'erreur totale calculée à partir des tableaux de fréquences est la mesure de la contribution de chaque prédicteur. Une erreur totale faible signifie une contribution plus élevée à la prévisibilité du modèle.

### 7.1.Évaluation du modèle

La matrice de confusion suivante montre un pouvoir de prévisibilité significatif. OneR ne génère pas de score ni de probabilité, ce qui signifie que les grilles d'évaluation (Gain, Lift, K-S et ROC) ne sont pas applicables.

**Tableau 4 :** évaluation de modèle OneR

| Confusion Matrix | | Play Golf | | | |
|---|---|---|---|---|---|
| | | Yes | No | | |
| OneR | Yes | 7 | 2 | *Positive Predictive Value* | 0.78 |
| | No | 2 | 3 | *Negative Predictive Value* | 0.60 |
| | | *Sensitivity* | *Specificity* | **Accuracy** = 0.71 | |
| | | 0.78 | 0.60 | | |

## 8. Tree J48

L'algorithme C4.5 de Quinlan actualise J48 pour créer un arbre de décision C4.5 ajusté. Tous les aspects de l'information consiste à diviser en sous-ensembles mineurs pour fonder une décision. J48 regardez le gain de données standardisé qui donne vraiment les résultats diviser les informations en choisissant un attribut. Pour résumer, les données standardisées extrêmes d'attribut obtenues sont utilisé. Les sous-ensembles mineurs sont renvoyés par l'algorithme. Les stratégies de fractionnement s'arrêtent si un sous-ensemble a

une place avec un classe similaire dans toutes les instances. J48 développe un nœud de décision utilisant les estimations attendues de la classe. J48

L'arbre de décision peut traiter des caractéristiques particulières, des estimations d'attributs perdus ou manquants des données et des variations coûts d'attribut. Ici, la précision peut être augmentée par l'élagage **(Venkatesan, 2015).**

### 8.1.Limites de l'algorithme J48

Malgré le fait que J48 est l'un des algorithmes les plus connus, cet algorithme présente quelques lacunes. Quelques les limitations de J48 sont discutées ci-dessous. **(Prerna Kapoor, 2015).**

### 8.2.Branches vides

La construction d'un arbre avec une valeur significative est l'une des étapes importantes pour la génération de règles par l'algorithme J48.

Dans notre recherche, nous avons sorti de nombreux nœuds avec des valeurs nulles ou très proches de cela. Mais ces valeurs ne contribuer à créer ou aider à créer une classe pour une tâche de classification. Au lieu de cela, cela rend l'arbre plus large et immobile compliquer. **(Prerna Kapoor, 2015).**

### 8.3.Succursales non significatives

Le nombre d'attributs distincts choisis produit le même nombre de divisions potentielles pour construire un arbre de décision. Mais le fait est qu'ils ne sont pas tous significatifs pour la tâche de classification. Ces branches les moins importantes diminuent non seulement la convivialité des arbres de décision, mais aussi poser le problème de sur-ajustement. **(Srishti Taneja, 2014)**

### 8.4.Sur ajustement

Le sur ajustement se produit lorsque l'affichage de l'algorithme obtient des informations avec des attributs exceptionnels. Cela provoque de nombreux fragmentations dans la distribution des processus. Les nœuds statistiquement sans importance avec le moins d'exemples sont appelés fragmentations. Habituellement, l'algorithme J48 construit des arbres et développe ses branches "juste assez profondément pour classer parfaitement l'exemples de formation.

Cette approche fonctionne mieux avec des données sans bruit. Mais la plupart du temps, cette stratégie dépasse les limites les exemples de formation avec des données bruitées. À l'heure actuelle, il existe deux stratégies largement utilisées pour contourner ce problème.

S’inscrivant dans l'apprentissage de l'arbre de décision. **(Sagar, 2015)** Ce sont :

Si l'arbre grandit, arrêtez-le avant qu'il n'atteigne le point maximum de classification précise des données d'entraînement.

Laisser l'arbre s'adapter aux données d'apprentissage puis post-élaguer l'arbre.

Pourtant, rien de tout cela n'est une solution parfaite à ce problème. Nous avons donc proposé deux outils pour minimiser l'espace d'entrée de données dans cette recherche. Le premier outil est l'entropie de la théorie de l'information et le second est le coefficient de corrélation. Dans

Cette expérimentation, nous avons examiné les données médicales sur la dengue. Les détails de l'explication des ensembles de données sont fourni l'outil d'apprentissage automatique basé sur Java WEKA qui est utilisé pour effectuer la recherche. **(Sagar, 2015)**

# PARTIE EXPERIMENTALE

Dans ce dernier chapitre, je présente d'abord une étude technique dans laquelle je définis l'environnement logiciel utilisé, puis je définirai mon data set avec une description de ses caractéristiques et les étapes de prétraitement des données pour corriger les valeurs aberrantes et choisir le meilleur modèle à suivre.

Outils et librairies utilisés

1. **Langage java**

Le langage Java est un langage généraliste de programmation synthétisant les principaux langages existants lors de sa création en 1995 par Sun Microsystems. Il permet une programmation orientée-objet (à l'instar de Small Talk et, dans une moindre mesure, C++), modulaire (langage ADA) et reprend une syntaxe très proche de celle du langage C. Outre son orientation objet, le langage Java a l'avantage d'être modulaire (on peut écrire des portions de code génériques, c-à-d utilisables par plusieurs applications), rigoureux (la plupart des erreurs se produisent à la compilation et non à l'exécution) et portable (un même programme compilé peut s'exécuter sur différents environnements). En contrepartie, les applications Java ont le défaut d'être plus lentes à l'exécution que des applications programmées en C par exemple. **(Eyrolles, 2011).**

2. **Weka**

Weka (Waikato Environment for Knowledge Analysis) est un ensemble d'outils permettant de manipuler et d'analyser des fichiers de données, implémentant la plupart des algorithmes d'intelligence artificielle, entre autres, les arbres de d´décision et les réseaux de neurones. Il est écrit en java.

Il se compose principalement :

- De classes Java permettant de charger et de manipuler les données.
- De classes pour les principaux algorithmes de classification supervisée ou non supervisée.
- D'outils de s'élection d'attributs, de statistiques sur ces attributs.
- De classes permettant de visualiser les résultats. On peut l'utiliser `a trois niveaux :
- Via l'interface graphique, pour charger un fichier de données, lui appliquer un algorithme, vérifier son efficacité.
- Invoquer un algorithme sur la ligne de commande.

- Utiliser les classes d´définies dans ses propres programmes pour créer d'autres Méthodes, implémenter d'autres algorithmes, comparer ou combiner plusieurs m´méthodes.

C'est cette troisième possibilité qui sera utilisée en travaux pratiques.µ

### 2.1.Description des données utilisées

Résultats d'un essai prospectif randomisé, contrôlé par placebo, à 2 arm de streptomycine 2 grammes par jour ; (Arm A2) vs placebo (arm A1) pour traiter la tuberculose chez 107 jeunes patients, tel que rapporté par le Streptomycine dans le Tuberculosis Trials Commette en 1948 dans le British Médical Journal.

### 2.2.Usage

- **Strep_tb**
- **Forme**
- Une base de données avec 107 observations et 13 variables
- **Patient id :** numéro d'identification inventé pour chaque participant
- **Arm :** groupe de traitement assigné, Streptomycine ou Contrôle
- **Grp de streptomycine** =55
- **Grp contrôlé** = 52
- **Dose_strep_g :** dose de streptomycine (gramme) : numérique, 0, 1 ou 2 grammes
- **Dose_PAS_g :** dose de PAS (Para-Amino-Salicylate) : numérique, 5, 10 ou 20 grammes. Noter que personne dans cette étude initiale (étude A) n'a reçu de PAS. Cela a été ajouté pour la thérapie combinée dans études B et C.
- **Genre :** dichotomique avec niveaux : M = Masculin, F= Féminin
- **Condition :** Condition du patient au départ, 3 niveaux, 1_Good, 2_Fair, 3_Poor
- **Temp :** température à la ligne de base en degrés Fahrenheit ou Celsius, mais classée en 4 niveaux(Le niveau fébrile était apparemment des cas non mesurés avec un thermomètre) avec des niveaux :
- 1_Fébrile
- 2_<99F/<37.2C,
- 3_99-99.9F/37.2-37.75C,
- 4_100F+/37.7C+

**Esr (érythrocytes sédimentation rate)** : Taux de sédimentation des érythrocytes en mm par heure, classé en 4 niveaux, de 0 à51+ mm par heure, avec niveaux :

- 1_0-10
- 2_11-20
- 3_21-50
- 4_51+

- **Cavitation :** présence dichotomique de cavitation sur la radiographie thoracique de référence
  - 0_non.
  - 1_oui.
- **Strep_resistance :** résistance à la streptomycine après 6 mois de traitement, mesurée sur une échelle de 0 à 100+,

Classés en 3 niveaux - sensible, modéré et résistant :

- 1_sens_0-8
- 2_mod_8-99
- 3_resist_100+

Radiologic_6m : Score de Likert de la réponse radiologique sur la radiographie pulmonaire à 6 mois.

- 1_Décès.
- 2_Détérioration_considérable.
- 3_Détérioration_modérée,
- 4_Aucun changement.
- 5_Modéré_amélioration.
- 6_Amélioration_considérable.

- **Rad_num :** Score de Likert évaluation numérique de la réponse radiologique sur la radiographie pulmonaire à 6 mois : Numérique : 1-6, de la mort à une amélioration considérable
- **Amélioré :** Résultat dichotomique d'amélioration , Vrai ou fausse. 55 des 107 participants ont été améliorés.

**3. Des détails**

L'essai Streptomycine pour Tuberculoses en 1948 a été considéré comme le premier essai moderne randomisé, essai clinique contrôlé par placebo, qui a pu être réalisé en partie parce que les stocks étaient très limités de la streptomycine au Royaume-Uni après la Seconde Guerre mondiale.

Cette publication semble un peu primitive aujourd'hui, sans fonctionnalités standard comme un véritable, et une utilisation créative des graphiques pour afficher les caractéristiques de base de l'échantillon de l'étude Plus frappant encore, il n'y a pas d'approbation ou de consentement du comité d'éthique.

Il s'agissait du premier d'une série de 3 essais, dans lesquels l'efficacité initiale de la streptomycine a été établie, mais une résistance rapide s'est développée et des effets secondaires importants sont survenus à une dose de 2 grammes de streptomycine. Ce type de résistance s'est également produit avec un autre nouveau traitement antituberculeux à le temps, PAS (Para-Amino-Salicylate). Les essais ultérieurs B et C ont évalué différentes doses et combinaisons de streptomycine et PAS, et ont été publiés ensemble.

**4. La source**

Cet ensemble de données est reconstruit au mieux de mes capacités à partir de l'article du British Médical Journal de 1948, intitulé Streptomycin Treatment of Pulmonary Tuberculosis, pages 769-782 dans le Édition du 30 octobre 1948, rédigée par le Streptomycin in Tuberculosis Trials Committee.

- ✓ **Les étapes de prétraitement**
- ✓ **Exploration et visualisation**

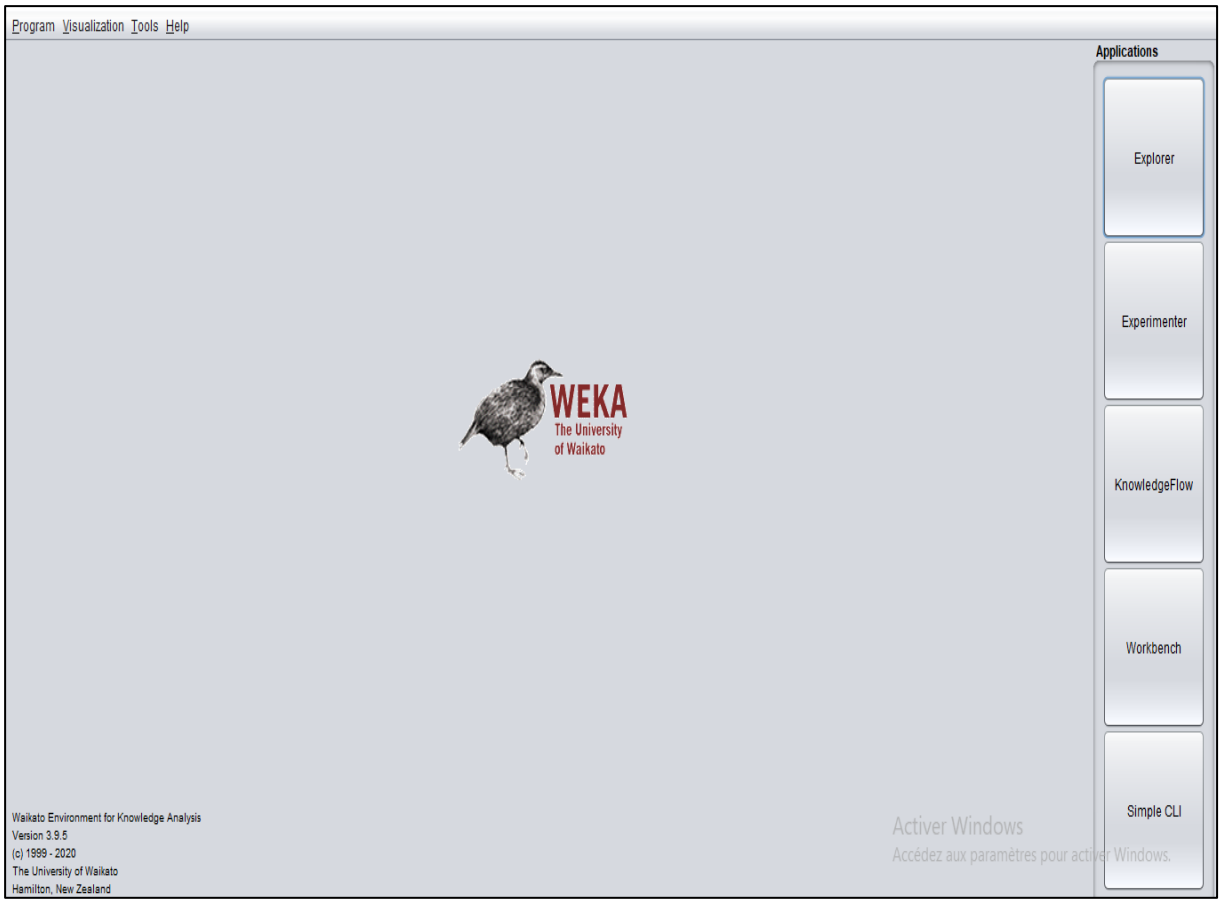
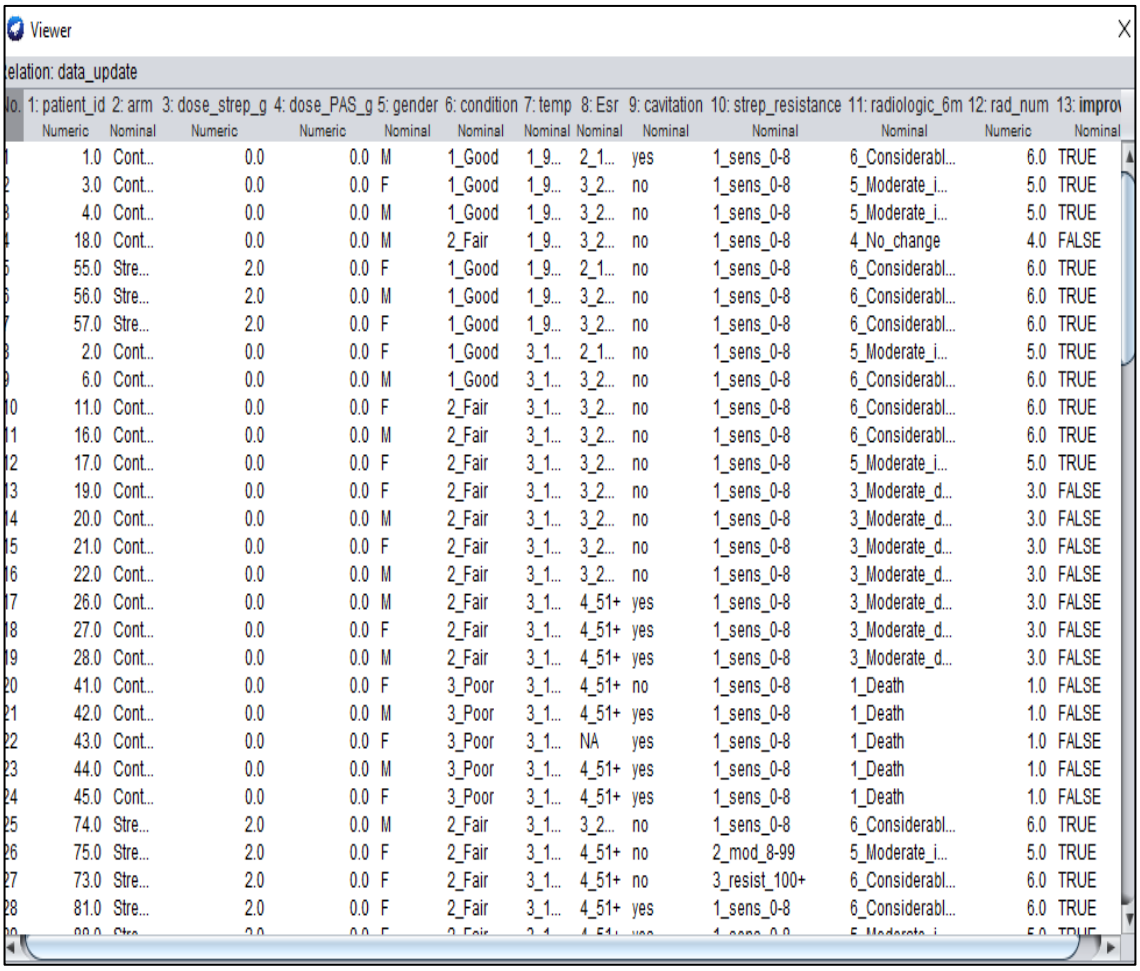



**Figure 03 :** exploration weka

La visualisation des données est définie comme l'exploration visuelle et interactive des données de toutes volumétries. Qui aident a` voir des choses n'étaient pas évidentes auparavant.

La visualisation facilite la transmission des informations de façon universelle et facilite le partage d'idées avec les autres. L'ensemble de données ressemble a` :



Viewer

elation: data_update

| No. | 1: patient_id | 2: arm | 3: dose_strep_g | 4: dose_PAS_g | 5: gender | 6: condition | 7: temp | 8: Esr | 9: cavitation | 10: strep_resistance | 11: radiologic_6m | 12: rad_num | 13: improv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Numeric | Nominal | Numeric | Numeric | Nominal | Nominal | Nominal | Nominal | Nominal | Nominal | Nominal | Numeric | Nominal |
| 1 | 1.0 | Cont... | 0.0 | 0.0 | M | 1_Good | 1_9... | 2_1... | yes | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |
| 2 | 3.0 | Cont... | 0.0 | 0.0 | F | 1_Good | 1_9... | 3_2... | no | 1_sens_0-8 | 5_Moderate_i... | 5.0 | TRUE |
| 3 | 4.0 | Cont... | 0.0 | 0.0 | M | 1_Good | 1_9... | 3_2... | no | 1_sens_0-8 | 5_Moderate_i... | 5.0 | TRUE |
| 4 | 18.0 | Cont... | 0.0 | 0.0 | M | 2_Fair | 1_9... | 3_2... | no | 1_sens_0-8 | 4_No_change | 4.0 | FALSE |
| 5 | 55.0 | Stre... | 2.0 | 0.0 | F | 1_Good | 1_9... | 2_1... | no | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |
| 6 | 56.0 | Stre... | 2.0 | 0.0 | M | 1_Good | 1_9... | 3_2... | no | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |
| 7 | 57.0 | Stre... | 2.0 | 0.0 | F | 1_Good | 1_9... | 3_2... | no | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |
| 8 | 2.0 | Cont... | 0.0 | 0.0 | F | 1_Good | 3_1... | 2_1... | no | 1_sens_0-8 | 5_Moderate_i... | 5.0 | TRUE |
| 9 | 6.0 | Cont... | 0.0 | 0.0 | M | 1_Good | 3_1... | 3_2... | no | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |
| 10 | 11.0 | Cont... | 0.0 | 0.0 | F | 2_Fair | 3_1... | 3_2... | no | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |
| 11 | 16.0 | Cont... | 0.0 | 0.0 | M | 2_Fair | 3_1... | 3_2... | no | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |
| 12 | 17.0 | Cont... | 0.0 | 0.0 | F | 2_Fair | 3_1... | 3_2... | no | 1_sens_0-8 | 5_Moderate_i... | 5.0 | TRUE |
| 13 | 19.0 | Cont... | 0.0 | 0.0 | F | 2_Fair | 3_1... | 3_2... | no | 1_sens_0-8 | 3_Moderate_d... | 3.0 | FALSE |
| 14 | 20.0 | Cont... | 0.0 | 0.0 | M | 2_Fair | 3_1... | 3_2... | no | 1_sens_0-8 | 3_Moderate_d... | 3.0 | FALSE |
| 15 | 21.0 | Cont... | 0.0 | 0.0 | F | 2_Fair | 3_1... | 3_2... | no | 1_sens_0-8 | 3_Moderate_d... | 3.0 | FALSE |
| 16 | 22.0 | Cont... | 0.0 | 0.0 | M | 2_Fair | 3_1... | 3_2... | no | 1_sens_0-8 | 3_Moderate_d... | 3.0 | FALSE |
| 17 | 26.0 | Cont... | 0.0 | 0.0 | M | 2_Fair | 3_1... | 4_51+ | yes | 1_sens_0-8 | 3_Moderate_d... | 3.0 | FALSE |
| 18 | 27.0 | Cont... | 0.0 | 0.0 | F | 2_Fair | 3_1... | 4_51+ | yes | 1_sens_0-8 | 3_Moderate_d... | 3.0 | FALSE |
| 19 | 28.0 | Cont... | 0.0 | 0.0 | M | 2_Fair | 3_1... | 4_51+ | yes | 1_sens_0-8 | 3_Moderate_d... | 3.0 | FALSE |
| 20 | 41.0 | Cont... | 0.0 | 0.0 | F | 3_Poor | 3_1... | 4_51+ | no | 1_sens_0-8 | 1_Death | 1.0 | FALSE |
| 21 | 42.0 | Cont... | 0.0 | 0.0 | M | 3_Poor | 3_1... | 4_51+ | yes | 1_sens_0-8 | 1_Death | 1.0 | FALSE |
| 22 | 43.0 | Cont... | 0.0 | 0.0 | F | 3_Poor | 3_1... | NA | yes | 1_sens_0-8 | 1_Death | 1.0 | FALSE |
| 23 | 44.0 | Cont... | 0.0 | 0.0 | M | 3_Poor | 3_1... | 4_51+ | yes | 1_sens_0-8 | 1_Death | 1.0 | FALSE |
| 24 | 45.0 | Cont... | 0.0 | 0.0 | F | 3_Poor | 3_1... | 4_51+ | yes | 1_sens_0-8 | 1_Death | 1.0 | FALSE |
| 25 | 74.0 | Stre... | 2.0 | 0.0 | M | 2_Fair | 3_1... | 3_2... | no | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |
| 26 | 75.0 | Stre... | 2.0 | 0.0 | F | 2_Fair | 3_1... | 4_51+ | no | 2_mod_8-99 | 5_Moderate_i... | 5.0 | TRUE |
| 27 | 73.0 | Stre... | 2.0 | 0.0 | F | 2_Fair | 3_1... | 4_51+ | no | 3_resist_100+ | 6_Considerabl... | 6.0 | TRUE |
| 28 | 81.0 | Stre... | 2.0 | 0.0 | F | 2_Fair | 3_1... | 4_51+ | yes | 1_sens_0-8 | 6_Considerabl... | 6.0 | TRUE |

**Figure 04 :** visualisation de données


**Figure 05 :** Visualisation de variable patient id

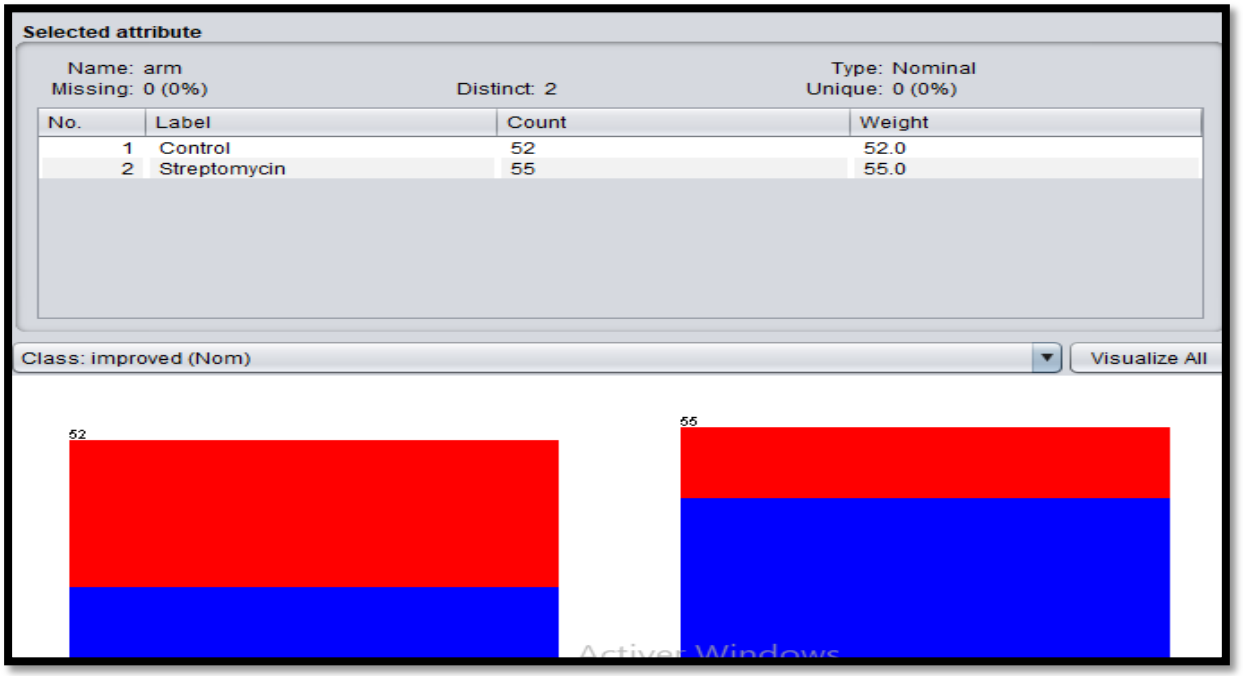



**Figure 06 :** Visualisation de variable arm

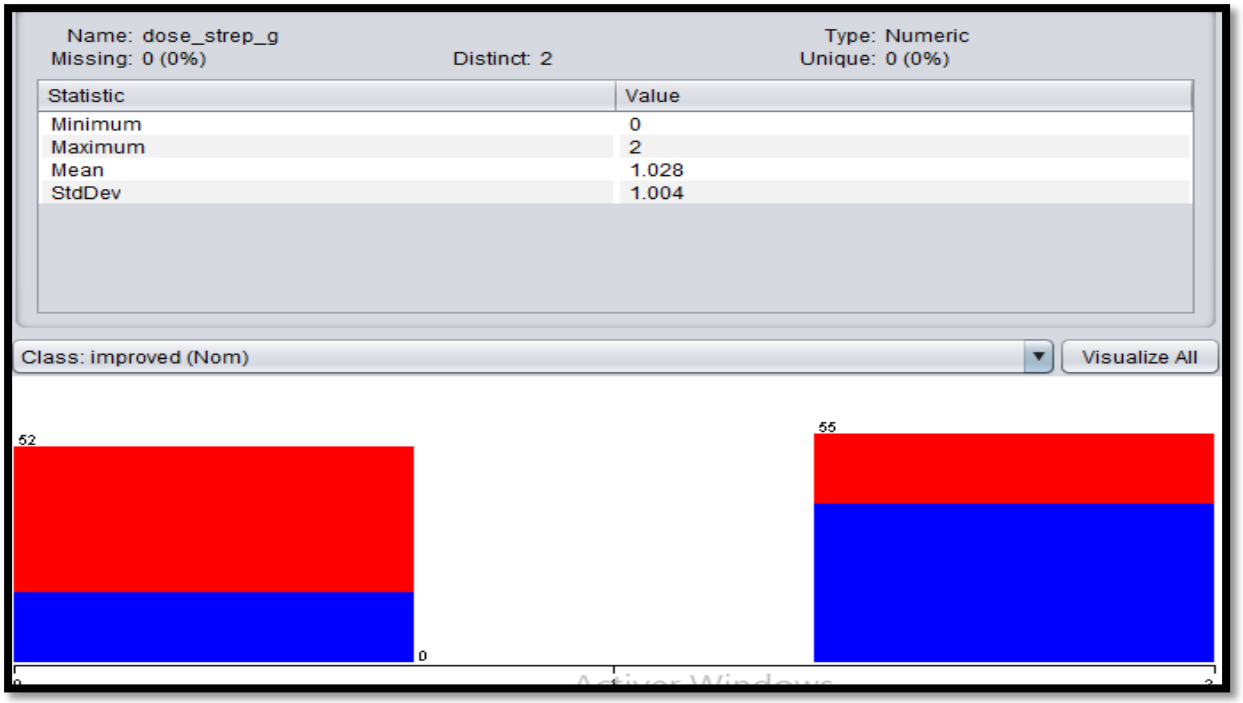



**Figure 07 :** Visualisation de variable strep dose g

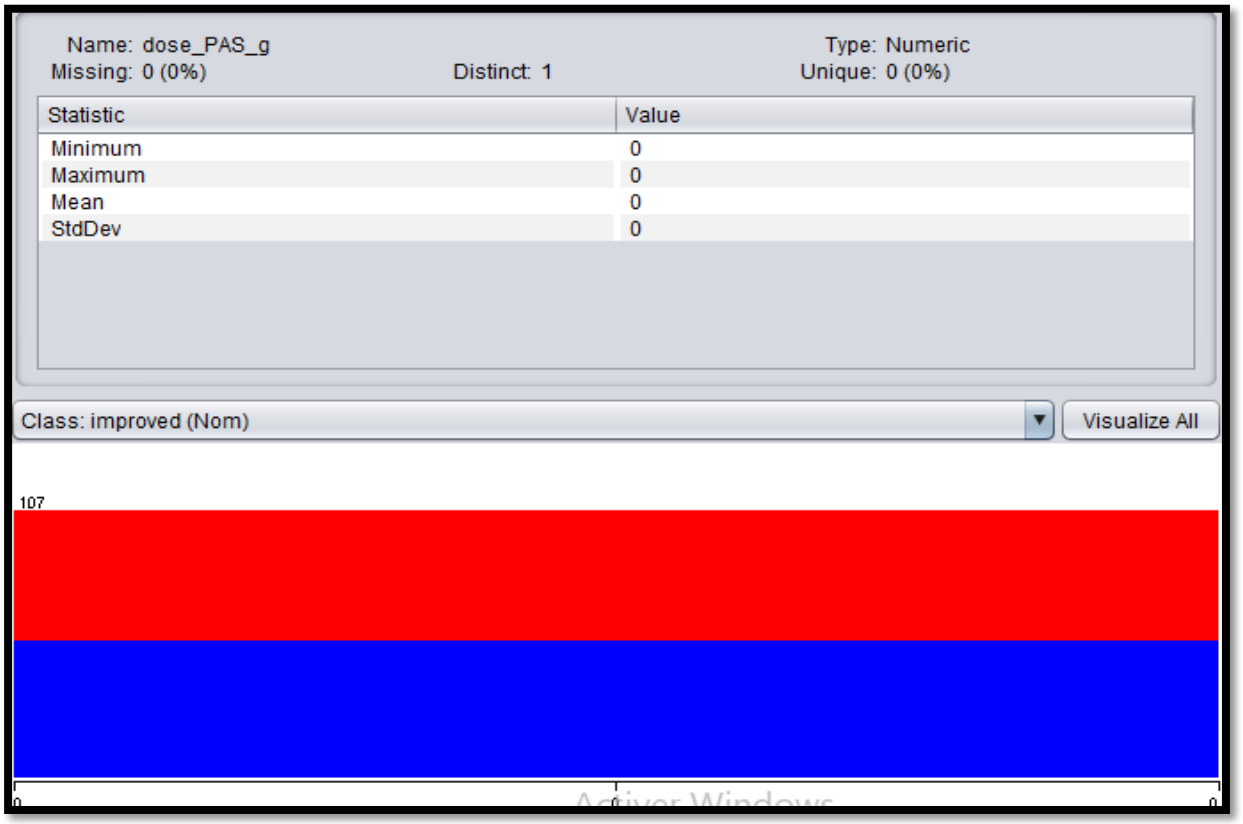



**Figure 008 :** Visualisation de variable Dose PAS g

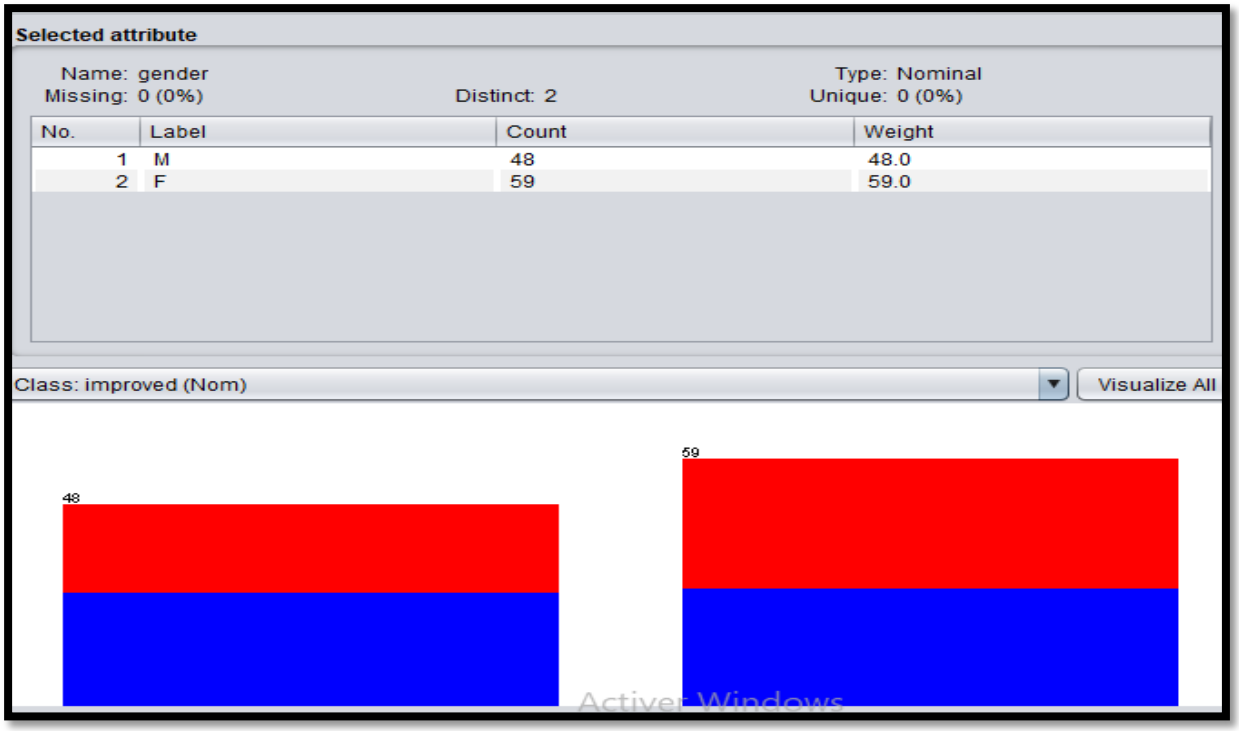



**Figure 09 :** Visualisation de variable gender


**Figure 10 :** Visualisation de variable condition

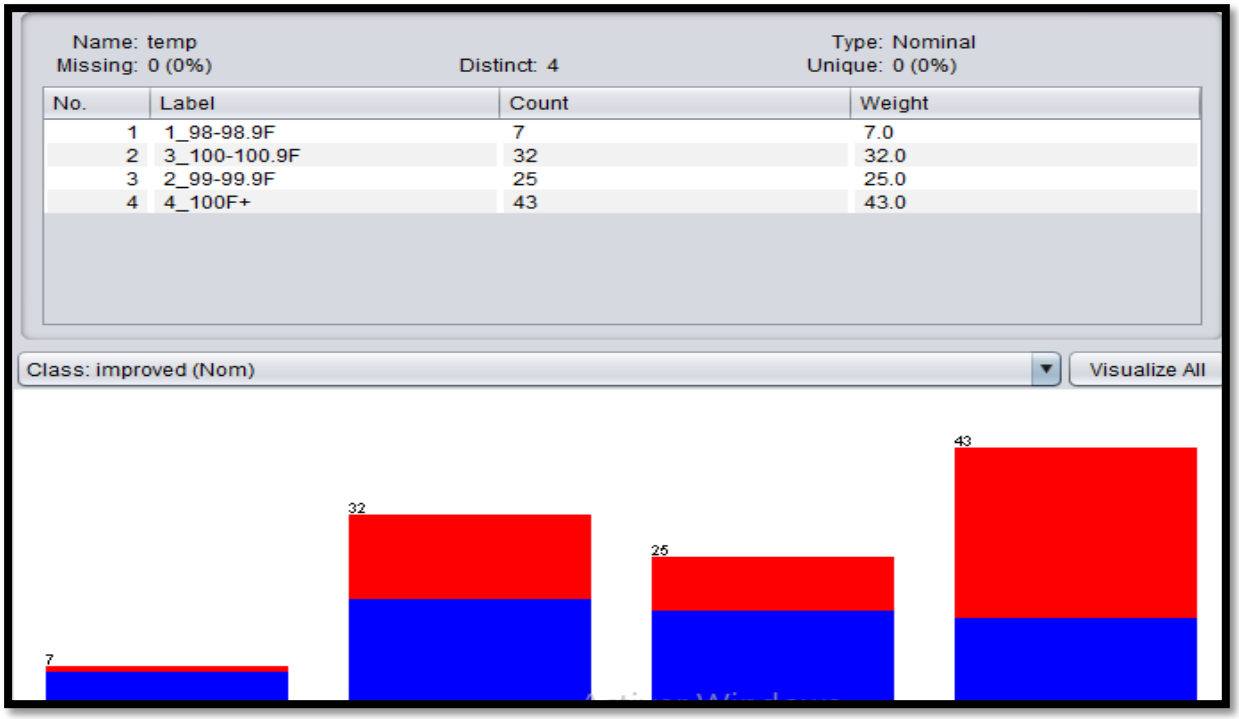



**Figure 11 :** Visualisation de variable temp

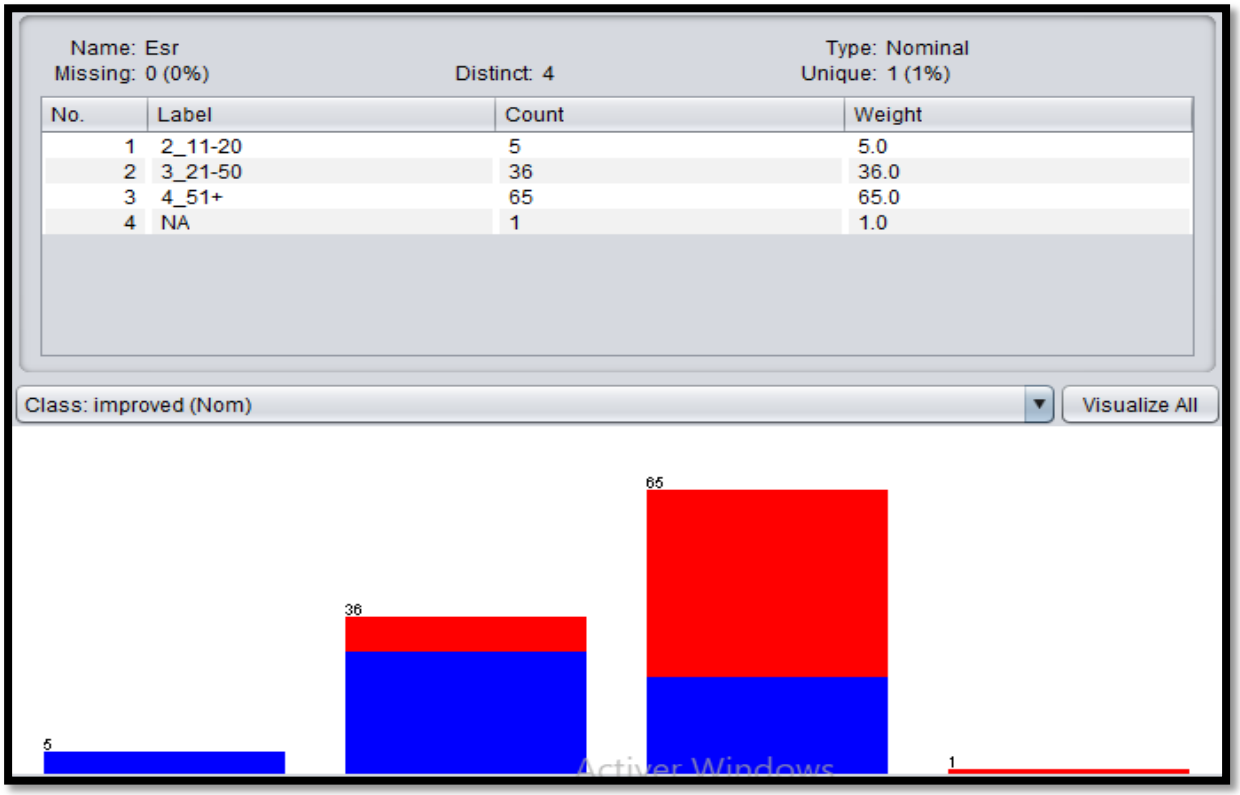



**Figure 12 :** Visualisation de variable Esr

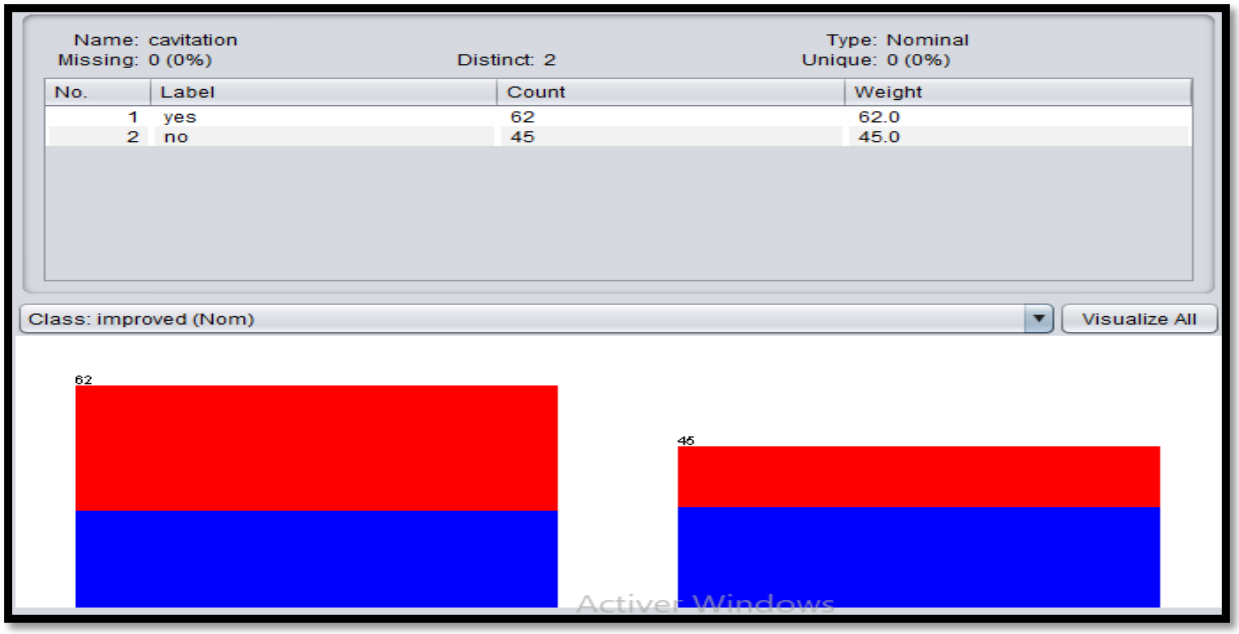



**Figure 13 :** Visualisation de variable cavitation


**Figure 14 :** Visualisation de variable strep résistance

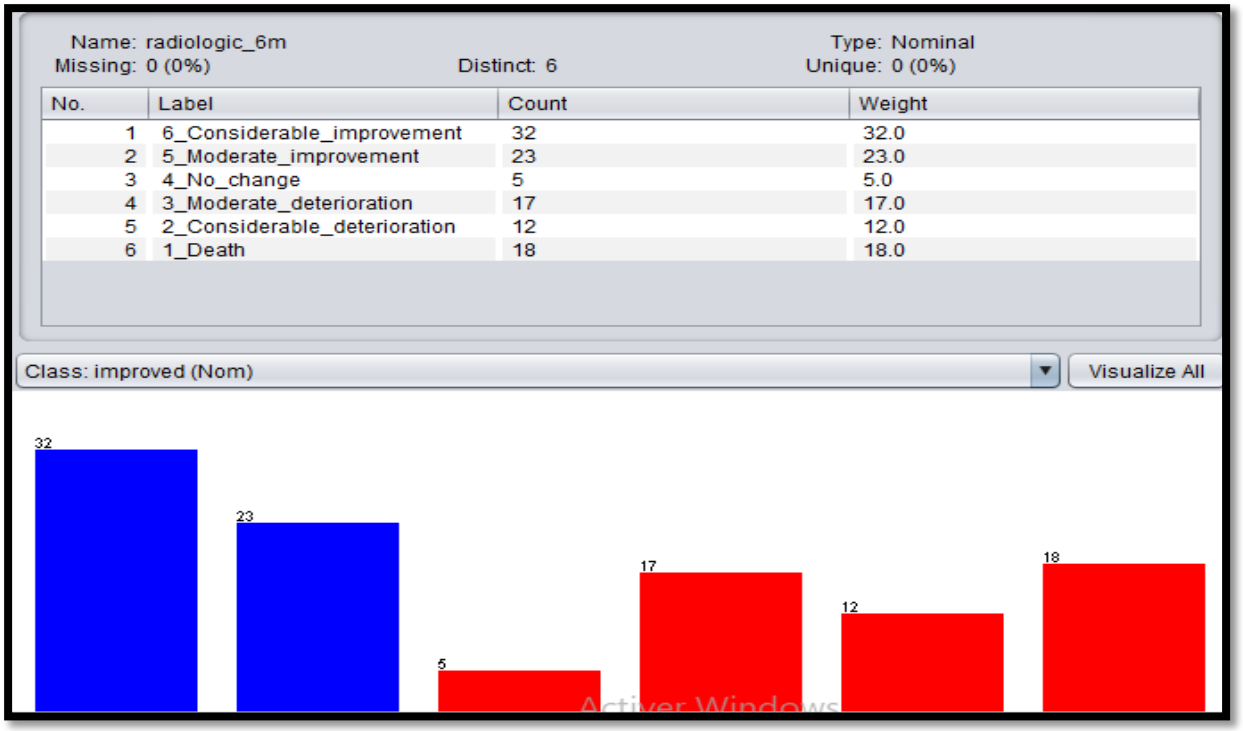



**Figure 15 :** Visualisation de variable radiologic_6m

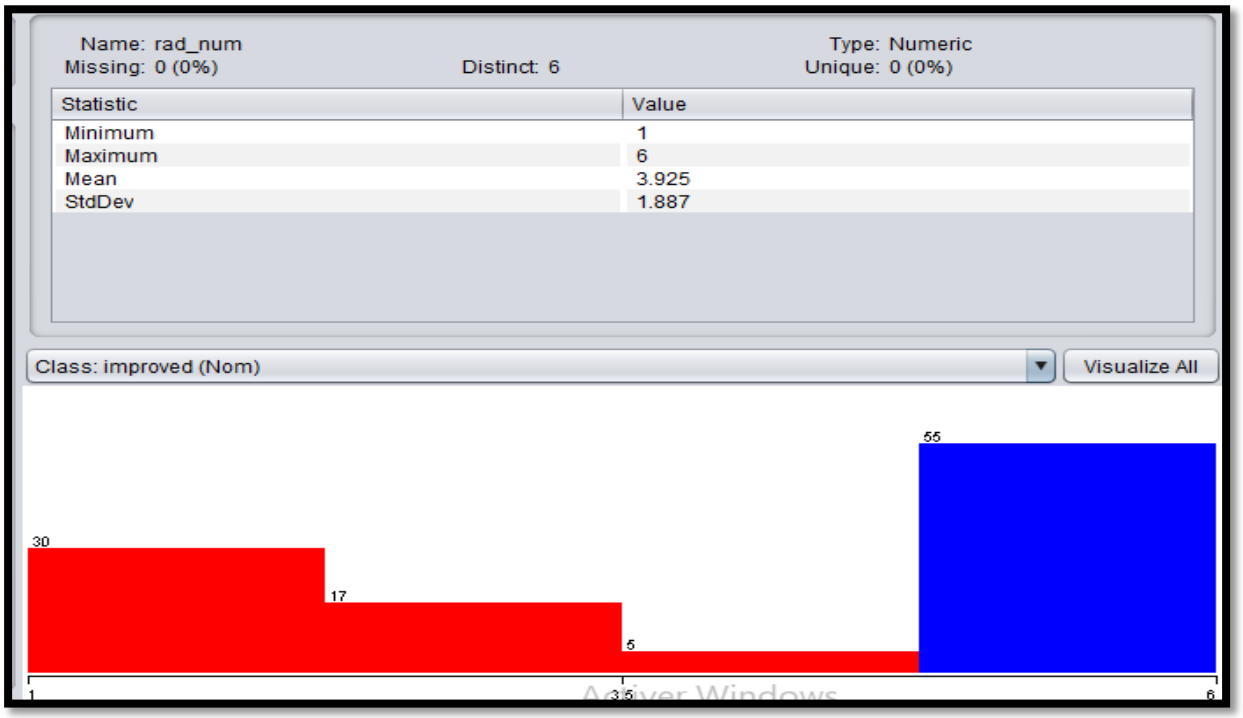



**Figure 16 :** Visualisation de variable rad num


**Figure 17 :** Visualisation de variable improved

## 5. Interprétation des figures

Pour chaque variable on constate :

- Le type d'attribut numérique ou nominal
- Le nombre des valeurs distinctes.
- Le pourcentage des valeurs unique.
- Le nombre et le pourcentage (%) des valeurs manquantes.
- La moyenne, minimum et maximum.
- La distribution de données en graphe.

## 6. Le nettoyage des données

Le nettoyage des données est un processus visant à identifier et à corriger les données altérées, inexactes ou non pertinentes. Cette étape fondamentale de prétraitement des données améliore la cohérence, la fiabilité et la valeur des données, et se traduit par de meilleures données qui fournissent de meilleurs modèles résultants.

Lors de la visualisation des données j'ai constaté qu'il existe des valeurs aberrantes dans certains colonnes comme :

Dose PAS g : toutes égales À zéro au cours de cette étape (A) donc la suppression de ce variable c mieux que l'utilise.

**7. La sélectionne des modèles**

C'est une phase très importante et le cœur de l'apprentissage automatique où on sélectionne le modèle qui fonctionne mieux pour l'ensemble des données parmi une collection des modèles d'apprentissage automatique candidats.

Les modèles utilisés pour la prédiction de faisabilité de médicament (strept Tb) sont :

- Bayes Naive Bayes
- Rules OneR
- Tree J48
- Tree Random Forest

**8. La sélectionne de La méthode d'évaluation**

**8.1.Training set**

Cette méthode consiste à divisé l'ensemble des données en deux partie : partie d'entrainement sur lequel le modèle fait son apprentissage et partie de test sur lequel on a teste le modèle et évaluer sa performance.

**8.2.Validation croisé**

Ou cross validation en anglais cette méthode consiste à diviser l'ensemble de données en k sous-ensembles (ou plis) différents puis il l'utilise l'union de k-1 sous-ensemble pour l'entrainement et le dernier sous-ensemble pour le test. Le processus est répète pour chaque sous-ensemble et la précision moyenne des tests est la précision de test.

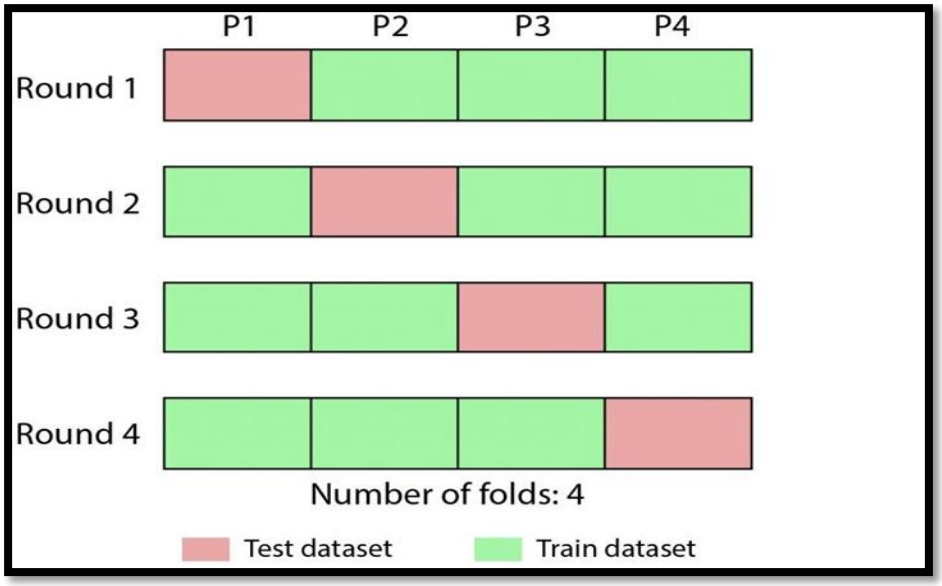



**Figure 18 :** Processus de validation croise en 4 iterations

La méthode validation croisé avec n split=10 qui indique le nombre de fois l'ensemble de données est subdivisé en sous- ensembles. La précision est calculée avec cross val score pour chaque itération.

## 9. Les résultats

Les résultats de validation croisée 10 fois Les instances correctement et incorrectement classées montrent le pourcentage d'instances de test. Les nombres bruts sont indiqués dans la matrice de confusion, avec a et b représentant les étiquettes de classe.

- **On trouve aussi**

✓ Taux de TP taux de vrais positifs (instances correctement classées comme une classe donnée)

✓ Taux de FP taux de faux positifs (instances faussement classées comme une classe donnée)

✓ Précision : proportion d'instances qui appartiennent vraiment à une classe divisée par le total des instances classées dans cette classe

✓ Rappel : proportion d'instances classées dans une classe donnée divisée par le total réel dans cette classe (équivalent au taux de TP)

✓ Mesure F : Une mesure combinée de la précision et du rappel

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          81               75.7009 %
Incorrectly Classified Instances        26               24.2991 %
Total Number of Instances              107

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                 0,769    0,255    0,741      0,769    0,755      FALSE
                 0,745    0,231    0,774      0,745    0,759      TRUE
Weighted Avg.    0,757    0,242    0,758      0,757    0,757

=== Confusion Matrix ===

  a  b   <-- classified as
 40 12 |  a = FALSE
 14 41 |  b = TRUE
```

**Figure 19 :** Résultats de bayes naïve

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          85               79.4393 %
Incorrectly Classified Instances        22               20.5607 %
Total Number of Instances              107

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                 0,788    0,200    0,788      0,788    0,788      FALSE
                 0,800    0,212    0,800      0,800    0,800      TRUE
Weighted Avg.    0,794    0,206    0,794      0,794    0,794

=== Confusion Matrix ===

  a  b   <-- classified as
 41 11 |  a = FALSE
 11 44 |  b = TRUE
```

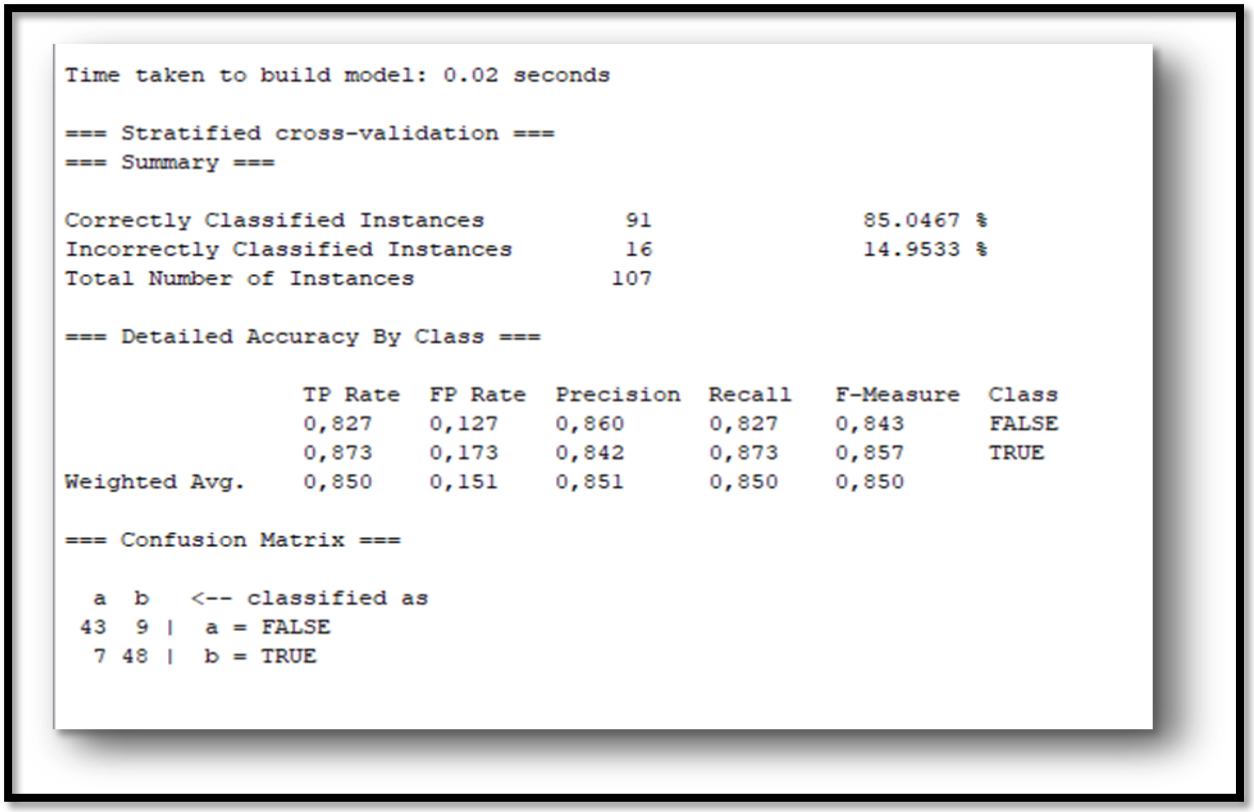
**Figure 20 :** Résultats de oneR

```
Number of Leaves  :     11

Size of the tree :      17


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          86               80.3738 %
Incorrectly Classified Instances        21               19.6262 %
Total Number of Instances              107

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                 0,769    0,164    0,816      0,769    0,792      FALSE
                 0,836    0,231    0,793      0,836    0,814      TRUE
Weighted Avg.    0,804    0,198    0,804      0,804    0,803

=== Confusion Matrix ===

  a  b   <-- classified as
 40 12 |  a = FALSE
  9 46 |  b = TRUE
```

**Figure 21 :** Résultats de Tree j48

```
Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          90               84.1121 %
Incorrectly Classified Instances        17               15.8879 %
Total Number of Instances              107

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                 0,827    0,145    0,843      0,827    0,835      FALSE
                 0,855    0,173    0,839      0,855    0,847      TRUE
Weighted Avg.    0,841    0,160    0,841      0,841    0,841

=== Confusion Matrix ===

  a  b   <-- classified as
 43  9 |  a = FALSE
  8 47 |  b = TRUE
```

**Figue 22 :** Résultats de Randomforest



```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          91               85.0467 %
Incorrectly Classified Instances        16               14.9533 %
Total Number of Instances              107

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  Class
                 0,827    0,127    0,860      0,827    0,843      FALSE
                 0,873    0,173    0,842      0,873    0,857      TRUE
Weighted Avg.    0,850    0,151    0,851      0,850    0,850

=== Confusion Matrix ===

  a  b   <-- classified as
 43  9 |  a = FALSE
  7 48 |  b = TRUE
```

**Figure 23** : résultats d'Amélioration de précision du modèle Randomforest

### 9.1.Evaluation des modèles

Pour comparer les différents modèles et évaluer ces performances on utilise 03 mesures : La précision ; recall score ; F1 score

**Tableau 5 :** Les résultats d'évaluations pour les différents modèles

| algorithmes | Précision | Score de rappel (recall) | F1 score |
|---|---|---|---|
| Naïve bayes | 75,8 | 75,7 | 75 ,7 |
| oneR | 79,4 | 79,4 | 79,4 |
| J48 | 80,4 | 80,4 | 80,3 |
| Random forest | 84,1 | 84,1 | 84,1 |
| Random forest amélioré | 85,1% | 85% | 85% |

D’après le tableau ci-dessus le modèle Random forest obtenu la meilleure précision qui égal à 84% et le meilleure score de rappel égal à 84% c’est-`a-dire que sur toutes les patients 84% d’entre eux sont correctement classés.

Je sélectionne le modèle Random forest comme le modèle le plus optimale et qui fonctionne mieux pour mon ensemble des données en raison de sa grande précision et score de rappel.

D’après les réglages manuels des paramètres j’ai observé que la précision de modèle est Augmenté ; j’ai remplacé 10 par 15 dans le paramètre de cross validation.

### 9.2.Sauvegarde le modèle

L’enregistrement du modèle finalise fait gagne beaucoup de temps car on n’a pas besoin d’entrainer le modèle à chaque exécution de l’application.

# CONCLUSION

# *Conclusion*

Au cours de ce mémoire j'ai défini l'effet de médicament (strep Tb ) comme un processus de classification tant que l'utilisation de l'informatique et de l'intelligence artificielle devient le plus en plus fréquente pour mettre en œuvre cette classification , bien que la décision des spécialistes et des experts soit le facteur le plus important ; donc j'ai mené à faire une comparaison entre quatres algorithmes d'apprentissage automatique à savoir : naïve bayes , oneR , tree J48, random forest .

Les résultats expérimentaux obtenu pour l'ensemble des données montrent que random forest est meilleur que les autres algorithmes en terme de sa grande précision dans les deux méthodes d'évaluation et son grande score pour les attributs d'évaluation ; après cela j'ai amélioré la performance de ce classificateur par un simple réglage dans les paramètres (kfolds) de cross validation et on peut le répète jusqu'à obtenir les résultats satisfaisants

Comme une perspective je souhaite pour les prochains projets de fin d'étude de créer ou développer une application web dans le but d'aide les personnes de prédire d'un effet de médicament.

# REFERENCES BIBLIOGRAPHIQUES

## *Références bibliographique*

### *A*


- **Aboutayeb R., (2009*)*** Technologie du lait et dérivés laitiers http://www.azaquar.com.
- **Acha PN., szyfres B. (2005)** Tuberculose zoonotique *In* : Zoonoses et maladies transmissibles communes à l'homme et aux animaux, Editions OIE (Organisation Mondialede la Santé Animale), Paris, 261-278.
- **Adams, L. G.** 2001. In vivo and in vitro diagnosis of Mycobacterium bovis infection.Rev.Sci.Tech. **20**:304-324.
- **Ait abdessalem A. (**1970). Microbiologie. Edition : Institut des sciences médicales ;INESSM. Alger. pp 1-31.
- **Ajmi TH, Tarmiz H, Bougmiza I, GATAA R, KNANI H, MITRAOUI A. ( 2010).**Epidimiogical profile of tuberculosis in the region of health Sousse from 1995 to2005.Revue Tunisienne d'infectiologie .4:18-22.
- **Aksu., K, Kurt., E, Parspour., S, Orman., A, Gülbaş., Z, Toraks., T.D. 2012**. Lymphocyte Subgroups in Different Forms of Tuberculosis. *Tuberculosis* **13**:5p.
- **Alais C. (1975).** Sciences du lait. Principes des techniques laitières. Edition Sepaic, Paris.
- **Alais C. (1984).** Sciences du lait. Principes de techniques laitières. 3éme édition, édition Publicité France.
- **Andrejak., Bonnaud., Cadranel., Chinet. , Marquette. 2010.** Tuberculose. *Item 106*. 26p.
- **Antoine D, Che D. (2008).** Epidemiology of tuberculosis in France in 2005. Med Mal Infect. 37: 45-52


### *B*


- **Bachelier, L. (1900).** Théorie de la spéculation. Dans Annales scientifiques de l'École normale supérieure (Vol. 17, pp. 21-86))

## *Références bibliographique*


*C*

CMIT. Tuberculose. In : E Pilly. Edition : Paris. Vivactis Plus, 2006 : p427-434.

- **COFER :** Collège Français des enseignants en Rhumatologie. Connaissance et pratique. Arthrite septique à germes banals 2004 ; 3 :262- 286.
- **Cornuéjols, L. Miclet, Y.Kodratoff,** « Apprentissage Artificiel, Concepts et algorithmes » ISBN 2-212-11020-0, 2002.

*D*

- **Dutronc H, Dauchy F-A, Dupon M.** Tuberculose. Rev Prat 2009 ; 59 :405-414.

*G*

- **Gupta, P.to wards data science.(2017).**Naive Bayes in Machine Learning.
- with python classification algorithms random forest.htm

*H*

- **Huchon G.** tuberculose et mycobactérioses non tuberculeuses. Ency Med Chir pneumologie 1997 ; 6-019-A-33: 20p.

*M*

- **M.-C. Dombret.** Tuberculose pulmonaire de l'adulte. Encycl Méd Chir, Traité de Médecine Akos 2004 ; 6-0740 : 7 p.

*N*

- **N. Rakesh, B. U. Maheswari, et S. K. Srivatsa,** "Performance analysis of propagation models in different terrain conditions for IEEE standard 802.16e WiMAX," Dans 2014 International Conference on Communication and Signal Processing, pp. 142-146, 2014.

## *Références bibliographique*


*P*

- **Pertuiset E.** Tuberculose vertébrale de l'adulte. Ency Med Chir Traité d'Appareil locomoteur 1998 ;15-852-A-10 : 19p.

*T*

- **Thaler J S, Marguire J H.Arthrites infectieuses.In : Braunwald E, Fauci A, Kasper D, Hauser S, Longo D, Jameson J, eds.** Harrison Principe de Med Interne. 14ème édition. Paris : Flammarion, 1998 ; p. 2239-2244.

*Web site*

- https ://towardsdatascience.com/naive-bayes-in-machine- learning-f49cc8f831b4
- https ://www.javatpoint.com/machine-learning-random-forest-algorithm)
- https ://www.javatpoint.com/machine-learning-random-forest-algorithm)

https ://www.tutorialspoint.com/machine learning with python /ma- chine learning