République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abbas Laghrour Khenchela
Faculté des Sciences et de la Technologie
Département de Mathématique et Informatique

Domaine : Informatiques
Filière : Informatiques
Spécialité : sécurité et technologie web

# Thème

## Une approche basée IA pour la prédiction précoce de la maladie de la tension Artérielle

Mémoire de fin d'étude pour l'obtention de master académique

**Encadré par :**

Mme. Hioual Ouassila

**Réalisé par :**

M. Hamzaoui Badis

M. Redah Moussa

Année Universitaire : 2020/2021

# Remerciement

Nos remerciements vont à l'adresse de tous ceux qui, de près ou de loin, ont aidé à la concrétisation de ce travail.

Nous remercions Dieu pour nous avoir permis d'aller jusqu'au bout de ce mémoire, nos parents pour leurs honnêtes et infaillibles sacrifices.

Nous vaudrions remercier particulièrement notre encadreur,

***Hioual Ouassila*** ,

Pour le

*tout d' abord pour son aide et sa compagnie pendant toutes les années où était notre professeur , pour sa méthode , pour son travail dure  en classe et en dehors , et surtout pour sa gentillesse*

deuxièmement, pour le  soutien prodigué et lui exprimer notre reconnaissance et toute notre gratitude pour avoir encadré ce travail et pour la confiance qu'elle nous a accordée, ses conseils et ses encouragements.

Nous remercions également le président du jury et ses membres pour avoir accepté de présider et prendre part à l'évaluation de ce travail.

Que tous ceux que nous n'avons pas nommément cités trouvent ici l'expression de notre profonde gratitude et notre salut éternel

Hamzaoui, Redah

# Table des matières

## Chapitre 2 :L'apprentissage automatique dans le domaine médical : avec la particularité de la prédiction de de maladie de tension artérielle

## Chapitre 3

## Prétraitement des données

## Chapitre 4

## Implémentation et réalisation

Liste des figures et tableaux

Liste des figures et tableaux

# Introduction général

## Introduction général

### Contexte

L'hypertension est un problème de santé courant qui est devenu un problème dans le monde moderne ; cela fait partie du syndrome métabolique et d'une affection multifactorielle dans laquelle un individu est diagnostiqué avec une pression artérielle systolique ≥140 mmHg et/ou une pression diastolique ≥90 mmHg. Ses causes exactes sont inconnues, mais une mutation génétique, un apport accru en sodium, une diminution de l'activité physique et l'obésité contribuent à sa progression. Dans certains cas, l'hypertension agit comme un « tueur silencieux » ; seulement remarqué quand il atteint un niveau dangereux.

Selon l'Organisation mondiale de la santé (OMS), l'hypertension dans le monde contribue à 12,8% du total des décès et cause environ 7,5 millions de décès. Une enquête menée au Algérie en 2008 a montré que 23.6% de la population âgée de 18 à 64 ans était hypertendue ou prenait des médicaments pour l'hypertension ; 17,0 % des femmes et 11,1 % des hommes, tandis que 71,9 % ne prenaient pas de médicaments pour l'hypertension.

Pour certaines personnes, le traitement de l'hypertension peut inclure uniquement des ajustements de style de vie sans l'utilisation de médicaments. Intervenir dans le mode de vie comprend, mais sans s'y limiter, la réduction de la consommation de sel, l'adoption d'un régime pauvre en graisses, la consommation plus de fruits et légumes, l'adoption d'un mode de vie actif et l'arrêt du tabac. Une approche de prévention réussie consiste à spécifier les personnes à haut risque et à les cibler. La recherche a montré que le développement de l'hypertension n'est pas seulement influencé par le statut de pré hypertension, mais également par d'autres facteurs tels que l'âge, le sexe, l'alimentation, l'indice de masse corporelle, le stress, l'historique de famille, apport de sel, niveau d'hémoglobine, mais surtout le fait qu'une personne a déjà d'autre maladie chronique.

En raison des coûts énormes des maladies chroniques, des études ont été menées pour estimer le risque d'hypertension, afin d'éviter une prise en charge et un traitement coûteux des complications. Les modèles prédictifs sont utiles pour prédire l'hypertension, étant essentiels dans la pratique médicale en raison de leur valeur dans la prise en charge des patients. Utilisé en milieu clinique, le score de risque d'hypertension de Framingham, un algorithme spécifique au sexe, est utilisé pour prédire le risque de développer des maladies cardiovasculaires à 10

ans[1]. C'est l'un des principaux scores utilisés pour indiquer l'hypertension. De nombreuses méthodes utilisant des techniques d'apprentissage automatique (Machine Learning, en anglais (ML)) sont utilisées dans les modèles de risque d'hypertension, par exemple : les réseaux de neurones artificiels, la machine à vecteurs de support, les forêts aléatoires, le classificateur naïf de Bayes, la machines à gradient, les arbres de décision et la régression logistique.

À l'aide de plusieurs techniques de ML, un certain nombre de facteurs prédictifs ont été identifiés pour prédire l'hypertension, par exemple : comorbidité, antécédents médicamenteux, âge > 60 ans, sexe, tabagisme, antécédents familiaux d'hypertension, indice de masse corporelle, niveau d'éducation, régime salé, légumes, fruits et consommation de viande, exercice physique régulier, lipides de faible densité, statut professionnel, état dépressif et anxieux.

Selon des études à travers le monde entier, les coûts par épisode de soins pour l'hypertension dans les pays à revenu faible et intermédiaire coutent chers. De plus, les gens développent souvent une hypertension à l'âge mûr, ce qui entraîne une perte d'années productives, entraînant des dépenses supplémentaires pour le système de santé.

Des données limitées sont disponibles sur le fardeau économique de l'hypertension en Algérie. L'identification précoce des patients hypertendus contribuera à réduire les coûts et les charges économiques de l'hypertension sur tout système de santé. Ces personnes à haut risque peuvent être identifiées à l'aide de modèles prédictifs basés sur des données de base facilement obtenues à partir de procédures non intrusives. Ces personnes peuvent être envoyées dans des établissements de santé pour des mesures préventives. En plus de servir de pilote pour de futures études.

C’est dans cette optique qu’entre le travail de ce projet d fin d’étude. Notre objectif est de de comparer puis de construire des modèles prédictifs de et des techniques d’aide à la décision, en utilisant les techniques d’apprentissage automatique, pour identifier les personnes à haut risque de développer une hypertension sans avoir besoin de procédures cliniques invasives.

---

[1] Kivimäki M, B. G.-M. (2009). Validating the Framingham hypertension risk score: results from the Whitehall II Study: hypertesnion . *HYPERTENSIONAHA.109.132373*, 54(3):496–501.

Afin de répondre à notre objectif, ce manuscrit est subdivisé en quatre chapitres et une conclusion. Le premier est introductif, où nous allons détailler les principes de base concernant les concepts liés à notre travail. Le deuxième chapitre présente l'apport de l'IA au domaine de la santé et plus particulière la prévention des maladies. Le troisième chapitre présente un prétraitement des données et une comparaison expérimentale des différents techniques de classification afin qu'on puisse en choisir une qui donnera plus de précision lors de la prise de décision. Dans le dernier chapitre, nous allons présenter l'implémentation de notre système. Nous terminerons par une conclusion dans laquelle nous discutons quelques perspectives.

# Chapitre 1

# Introduction à l'apprentissage automatique

# Et son enjeu dans les problèmes de la classification

## Introduction

Au cours des deux dernières décennies, l'apprentissage automatique est devenu l'un des piliers des technologies de l'information, et avec les quantités toujours croissantes de données disponibles, il y a de bonnes raisons de croire que l'analyse intelligente des données deviendra encore plus vaste en tant qu'ingrédient nécessaire au progrès technologique. Le but de ce chapitre est de fournir au lecteur une vue d'ensemble sur la vaste gamme d'applications qui ont pour cœur un problème d'apprentissage automatique. Et puisque notre projet est basé, en premier lieu, sur la notion de la classification, nous discuterons certains algorithmes utilisés pour cette dernière, et nous expliquerons également pourquoi nous avons choisi d'utiliser des algorithmes de classification plutôt que des algorithmes de régression. Enfin, nous présenterons un ensemble d'algorithmes assez basiques mais efficaces pour résoudre un tel problème, à savoir celui de la classification.

## 1- Apprentissage automatique : concepts et enjeux

Une définition qui s'applique à un programme informatique comme à un robot, un animal de compagnie ou un être humain est celle proposée par Fabien Benureau (2015) : «L'apprentissage est une modification d'un comportement sur la base d'une expérience ».

Dans le cas d'un programme informatique, à quoi nous nous intéressons, on parle d'apprentissage automatique, ou machine Learning, quand ce programme a la capacité d'apprendre sans être programmé. Cette définition est celle donnée par Arthur Samuel (1959). On peut ainsi opposer un programme classique, qui utilise une procédure et des données qu'il reçoit en entrée pour produire en sortie des réponses, à un programme d'apprentissage automatique, qui utilise des données et des réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières.[1]

**Exemple :**

Supposons qu'une entreprise a besoin de connaître le montant total dépensé par un client ou une cliente à partir de ses factures. Il suffit, pour cela, d'appliquer un algorithme classique, à savoir une simple addition : un algorithme d'apprentissage n'est pas nécessaire.

Supposons maintenant qu'on veut utiliser ces factures pour déterminer quels produits le client est le plus susceptible d'acheter dans un mois. Bien que cela soit vraisemblablement lié, nous n'avons manifestement pas toutes les informations nécessaires pour ce faire. Cependant, si nous disposons de l'historique d'achat d'un grand nombre d'individus, il devient possible d'utiliser un algorithme de machine Learning pour qu'il en tire un modèle prédictif nous permettant d'apporter une réponse à notre question.[2]

### 1.1 Apprentissage automatique et résolution des problèmes :

L'apprentissage automatique est là pour trouver des solutions à ce :

• que l'on ne sait pas résoudre (comme dans l'exemple de la prédiction d'achats ci-dessus).
• que l'on sait résoudre, mais dont on ne sait formaliser en termes algorithmiques comment nous les résolvons (c'est le cas par exemple de la reconnaissance d'images ou de la compréhension du langage naturel) ;

[1] Azencott, C.-A. (2019). Introduction au Machine Learning. Paris: Dunod , pp. 1-2.
[2] *Ibid.*, p. 2.

• que l'on sait résoudre, mais avec des procédures beaucoup trop gourmandes en ressources informatiques (c'est le cas par exemple de la prédiction d'interactions entre molécules de grande taille, pour lesquelles les simulations sont très lourdes).[1]

### 1.2 Piliers fondamentaux de l'apprentissage automatique :

Les deux piliers fondamentaux de l'apprentissage automatique sont :
• d'une part, les données, qui sont les exemples à partir duquel l'algorithme va apprendre.
• d'autre part, l'algorithme d'apprentissage, qui est la procédure que l'on fait tourner sur ces données pour produire un modèle. On appelle entraînement le fait de faire tourner un algorithme d'apprentissage sur un jeu de données.[2]

## 2- Types de problèmes d'apprentissage automatique

### 2.1 Apprentissage supervisé

L'apprentissage **supervisé** commence généralement par un ensemble de données bien défini et une certaine compréhension de la façon dont ces données sont classifiées. L'apprentissage supervisé a pour but de déceler des modèles au sein des données et de les appliquer à un processus analytique. Ces données comportent des caractéristiques associées à des libellés qui définissent leur signification. Nous pouvons, par exemple, créer une application d'apprentissage automatique capable de faire la distinction entre plusieurs millions d'animaux, en se basant sur des images et des descriptions écrites.[3]

#### 2.1.1. Définition formelle :

A partir de la base d'apprentissage $\mathbf{S = (X_i, U_i)}_{1 \le i \le N}$, On va chercher une loi de dépendance entre **x** et ***u***. Par exemple : une fonction ***h*** aussi proche de ***f*** (fonction cible) que possible tel que $\mathbf{U_i = f(x_i)}$ ou une distribution de probabilité $\mathbf{P(X_i, U_i)}$. [4]

**Remarque importante :**

---

[1] Azencott, C.-A. (2019), op. *cit*, p. 2.
[2] Azencott, C.-A. (2019), op. *cit*, p. 2.
[3] Hurwitz, J. (s.d.). *Le machine learning et la science des données*. Consulté le 05/06/2021, sur site officiel de IBM: https://www.ibm.com/fr-fr/analytics/machine-learning
[4] Teytaud, F. (2020, août 25). *coursApprentissage*. Consulté le 06 06, 2021, sur site officiel de l'Université du Littoral Cote d'Opale: https://www-lisic.univ littoral.fr/~teytaud/files/Cours/Apprentissage/coursApprentissage.pdf

• Si ***f*** est une fonction continue on parle alors de régression.

• Si **f** est une fonction discrète on parle alors de classification.

### 2.2 Apprentissage non supervisé

L'apprentissage non supervisé est utilisé lorsque le problème nécessite une quantité massive de données non étiquetées. Par exemple, les applications de réseaux sociaux, telles que Twitter, Instagram et Snapchat, exploitent toutes de très grandes quantités de données non étiquetées. Pour comprendre le sens de ces données, il est nécessaire d'utiliser des algorithmes qui classifient les données en fonction des tendances ou des clusters qu'ils décèlent. L'apprentissage non supervisé mène un processus itératif, analysant les données sans intervention humaine. Il est utilisé, par exemple, avec la technologie de détection de spam envoyé par e-mail. Les e-mails normaux et les spams comportent un nombre de variables beaucoup trop élevé pour qu'un analyste puisse étiqueter les e-mails indésirables envoyés en masse. En revanche, les discriminants d'apprentissage automatique, basés sur la mise en cluster et l'association, sont appliqués pour identifier les courriers électroniques non désirés.[1]

#### 2.2.1. Définition formelle :

Dans ce cadre aucun expert n'est disponible. A partir de la base d'apprentissage $\mathbf{S = (X_i)_{1 \le i \le N}}$, On va chercher des régularités. Par exemple, sous forme de fonctions : régression, OU sous forme de nuages de points.[2]

### 2.3. Apprentissage par renforcement

L'apprentissage **par renforcement** est un modèle d'apprentissage comportemental. L'algorithme reçoit un feedback de l'analyse des données et guide l'utilisateur vers le meilleur résultat. L'apprentissage par renforcement diffère des autres types d'apprentissage supervisé, car le système n'est pas formé avec un ensemble de données exemple. Au lieu de cela, le système apprend plutôt par le biais d'une méthode d'essais et d'erreurs. Par conséquent, une séquence de décisions fructueuses aboutit au renforcement du processus, car c'est lui qui résout le plus efficacement le problème posé.[3]

### 2.4. Apprentissage en profondeur

---

[1] Hurwitz, J. (s.d.), op. *cit*.
[2] Teytaud, F, op. *cit*, p7.
[3] Hurwitz, J. (s.d.), op. *cit*.

L'apprentissage **en profondeur** est une méthode spécifique d'apprentissage automatique qui intègre des réseaux neuronaux en couches successives afin d'apprendre des données de manière itérative. L'apprentissage en profondeur est particulièrement utile lorsqu'on tente de détecter des tendances à partir de données non structurées. Les réseaux neuronaux complexes d'apprentissage en profondeur sont conçus pour émuler le fonctionnement du cerveau humain, de sorte que les ordinateurs peuvent être entraînés pour faire face à des abstractions et des problèmes mal définis. Les réseaux neuronaux et l'apprentissage en profondeur sont souvent utilisés dans les applications de reconnaissance d'image, de communication orale et de vision numérique.[1]

### 3- Classification et Régression

L'apprentissage supervisé est généralement effectué dans le contexte de la classification et de la régression.[2]

**3.1. Classification :** Un problème de classification survient lorsque la variable de sortie est une catégorie, telle que « rouge », « bleu » ou « maladie » et « pas de maladie ».

***Exemples* :**

- En finance et dans le secteur bancaire pour la détection de la fraude par carte de crédit (fraude, pas fraude).
- Détection de courrier électronique indésirable (spam, pas spam).
- Dans le domaine du marketing utilisé pour l'analyse du sentiment de texte (heureux, pas heureux).
- En médecine, pour prédire si un patient a une maladie particulière ou non.[3]

**3.2. Régression :** Un problème de régression se pose lorsque la variable de sortie est une valeur réelle, telle que « dollars » ou « poids ».

***Exemples :***

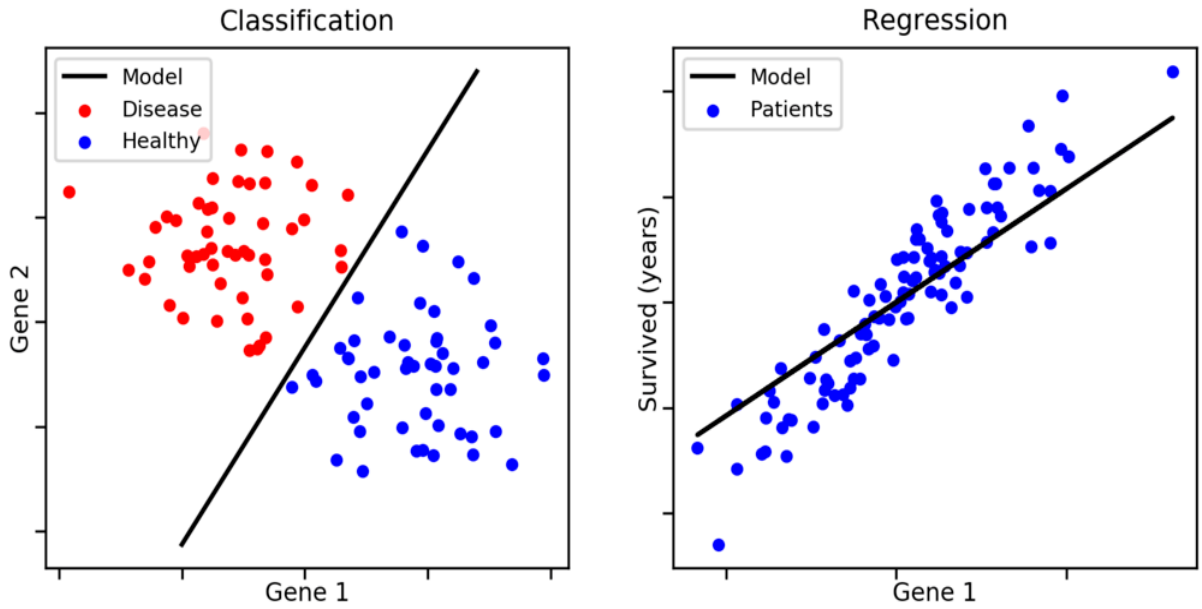
- Prédire le cours de bourse.
- Prédire le prix de l'immobilier. [4]

---

[1] Hurwitz, J. (s.d.), op. *cit*.
[2] Hurwitz, J. (s.d.), op. *cit*.
[3] Ismaili, Z. (s.d.). *apprentissage-supervise-vs-non-supervise.* Consulté le 06 06, 2021, sur analytics and insights: https://analyticsinsights.io/apprentissage-supervise-vs-non-supervise/
[4] Ibid.

La figure ci-dessous illustre la différence entre les deux méthodes citées ci-dessus.

**Figure 1 :** Différence entre la classification et la régression.[1]




### 4- Algorithmes d'apprentissage :

Dans cette section, nous allons parler uniquement des algorithmes d'apprentissage qui touchent le domaine de recherche de notre projet, généralement ces algorithmes sont les plus utilisé dans les problèmes de **classification**.

#### 4-1- Arbres de décision

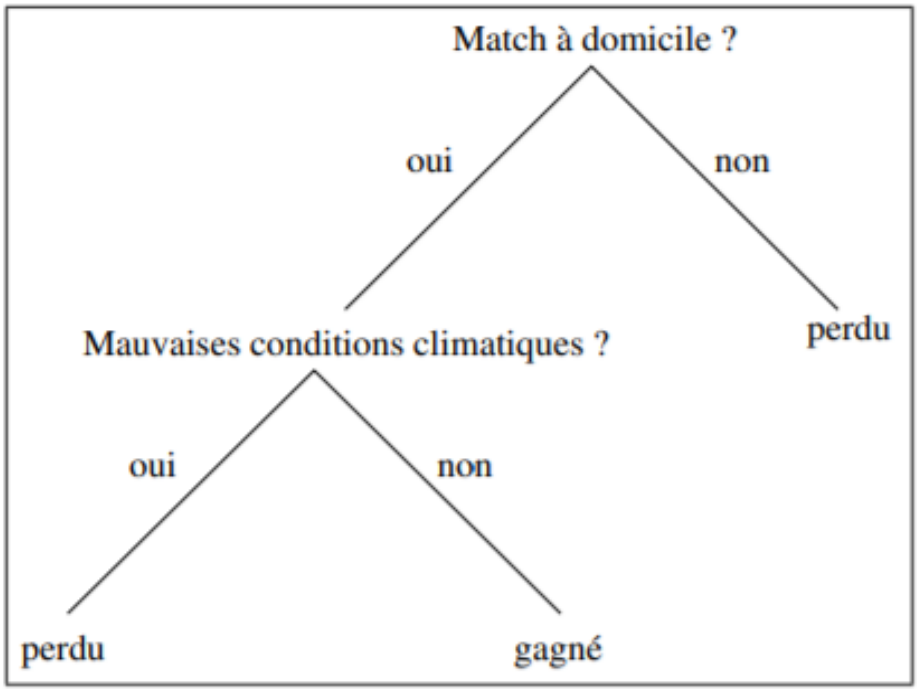
#### 4.1.1. Définition

Les arbres de décision sont des règles de classification qui se basent, dans leur décision, sur une suite de tests associés aux attributs, les tests étant organisés de manière arborescente. [2]

***Exemple*** : La situation en début d'un match de football est décrite par deux attributs binaires : *Match à domicile* et *Mauvaises conditions climatiques*. L'attribut classe prend deux valeurs : *gagné* et *perdu (voir Figure 2)*.

[1]Ismaili, Z. op. *cit*..
[2] Denis, F. (s.d.). chapitre 2:Les arbres de décision. Consulté le 06 2021, 06, sur http://pageperso.lif.univ-mrs.fr/~francois.denis/IAAM1/chap2.pdf , p 13 .

**Figure 2 :** Un exemple d'arbre de décision, source : [1]




**4.1.2. Définition formelle :** Etant donnés *n* attributs ***A1, . . . , An,*** l'espace de description ***X*** est le produit cartésien des domaines ***Xi*** de chaque attribut ***Ai*** :[2]

$$X = \prod_{i=1}^{n} X_i \; ou \; X_i = Dom(A_i)$$

Les attributs peuvent être :

– binaires,

– *n*-aires,

– réels

**4.1.3. Algorithmes d'apprentissage par arbres de décision** [3]

**entrée :** échantillon *S*

**début**

Initialiser l'arbre courant à l'arbre vide ; la racine est le nœud courant

**répéter**

Décider si le nœud courant est terminal

[1] Ibid.p 13.
[2] Denis, F. op. *cit*. p 13 .
[3] Denis, F. op. *cit*. pp 15-16 .

Si le nœud est terminal alors

Lui affecter une classe

**sinon**

Sélectionner un test et créer autant de nouveaux nœuds fils qu'il y a de réponses possibles au test

**FinSi**

Passer au nœud suivant non exploré s'il en existe

Jusqu`à obtenir un arbre de décision

**fin**

En général, on décide qu'un nœud est terminal lorsque tous les exemples associés à ce nœud, ou du moins la plupart d'entre eux sont dans la même classe, ou encore, s'il n'y a plus d'attributs non utilisés dans la branche correspondante.[1]

En général, on attribue au nœud la classe majoritaire (éventuellement calculée à l'aide d'une fonction de coût lorsque les erreurs de prédiction ne sont pas équivalentes). Lorsque plusieurs classes sont en concurrence, on peut choisir la classe la plus représentée dans l'ensemble de l'échantillon, ou en choisir une au hasard.[2]

La sélection d'un test à associer à un nœud est plus délicate, puisqu'on cherche à construire un arbre de décision le plus petit possible rendant compte au mieux des données. Une idée naturelle consiste à chercher un test qui fait le plus progresser dans la tache de classification des données d'apprentissage. Comment mesurer cette progression ? on utilise l'*indice de Gini* ou on utilise la notion d'*entropie.*[3]

Pratiquement, si l'on suppose que la classe prend une valeur **1,2,3…m** dans l'ensemble, et si $f_i$ désigne la fraction des éléments de l'ensemble avec l'étiquette **i** dans l'ensemble, on aura :

$$Gini(f) = 1 - \sum_{i=1}^{m} f_i^2$$

L'entropie permet de mesurer le désordre dans un ensemble de données et est utilisée pour choisir la valeur permettant de maximiser le gain d'information. En utilisant les mêmes notations que pour l'indice de diversité de Gini, on obtient la formule suivante :

[1] Denis, F. op. *cit*. pp 15-16 .
[2] Ibid.
[3] Ibid.

$$entropy(f) = -\sum_{i=1}^{m} f_i \times \log(f_i)$$

**4.1.4. Avantages et les inconvénients des arbres de décision :**[1]

**A. Avantages** :

• Par rapport à d'autres algorithmes, les arbres de décision nécessitent moins d'efforts pour la préparation des données lors du prétraitement.

• Un arbre de décision ne nécessite pas de normalisation des données.

• Un arbre de décision ne nécessite pas non plus de mise à l'échelle des données.

• Les valeurs manquantes dans les données n'affectent pas non plus le processus de construction d'un arbre de décision dans une mesure considérable.

• Un modèle d'arbre de décision est très intuitif et facile à expliquer aux équipes techniques ainsi qu'aux parties prenantes.

**B. Inconvénients :**[2]

• Un petit changement dans les données peut provoquer un grand changement dans la structure de l'arbre de décision provoquant une instabilité.

• Pour un arbre de décision, le calcul peut parfois devenir beaucoup plus complexe par rapport à d'autres algorithmes.

• L'arbre de décision implique souvent plus de temps pour former le modèle.

• La formation à l'arbre de décision est relativement coûteuse car la complexité et le temps requis sont plus importants.

• L'algorithme de l'arbre de décision est inadéquat pour appliquer une régression et prédire des valeurs continues.

### 4-2- Forêts aléatoires

**4.2.1. Définition**

Les forêts aléatoires sont composées (comme le terme "forêt" l'indique) d'un ensemble d'arbres décisionnels binaires dans lequel a été introduit de l'aléatoire. Ces arbres se distinguent les uns des autres par le sous-échantillon de données sur lequel ils sont

[1] Dhiraj, K. (2019, 05 26). Top 5 advantages and disadvantages of Decision Tree Algorithm. Consulté le 06 2021, 06, sur dhirajkumarblog: https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

[2] Ibid.

entraînés. Ces sous-échantillons sont tirés au hasard (d'où le terme "aléatoire") dans un jeu de données.[1]

Alors les forets utilisent un classificateur d'ensemble utilisant de nombreux modèles d'arbre de décision, les forets peuvent être utilisés pour la classification ou la régression. Par ailleurs, des informations précises et d'importance variable sont fournies avec les résultats.[2]

**4.2.2. Principe de fonctionnement des forets** :

• Un sous-ensemble différent des données d'apprentissage est sélectionné, avec remplacement, pour entraîner chaque arbre.

• Les données d'entraînement restantes sont utilisées pour estimer l'erreur et l'importance des variables.

• L'attribution des classes est faite par le nombre de votes de tous les arbres et pour la régression la moyenne des résultats est utilisée.[3]

**A. Utilisation d'un sous-ensemble de variables :**[4]

- Un sous-ensemble de variables sélectionné au hasard est utilisé pour diviser chaque nœud.
- Le nombre de variables utilisées est décidé par l'utilisateur.
- Un sous-ensemble plus petit produit moins de corrélation (taux d'erreur inférieur) mais un pouvoir prédictif inférieur (taux d'erreur élevé)
- La plage de valeurs optimale est souvent assez large.

**B. Variables communes pour les forêts aléatoires** :[5]

- Données d'entrée (prédicteur et cible)
- Nombre d'arbres
- Nombre de variables à utiliser à chaque division
- Options pour calculer les informations d'erreur et de signification variable

**C. Informations provenant des forêts aléatoires** :[6]

- Précision de la classification

---

[1] Moudachirou, M. K. (2017, Juillet). classification et forêts aléatoires: application à l'aide à la décision chirurgicale du genou par arthroplastie. *mémoire de maîtrise en technologie de l'information*. Québec, Canada: Télé-université, p 66 .

[2] Horning, N. (s.d.). *Introduction to decision trees and random forests.* (A. M. History's, Éd.), p 18 .

[3] Ibid. p 19.

[4] Ibid. p 20.

[5] Ibid. p 21.

[6] Ibid. p 22.

- Importance variable
- Estimation des données manquantes
- Taux d'erreur pour les objets forestiers aléatoires

**4.2.3. Algorithme de forêt aléatoire** :

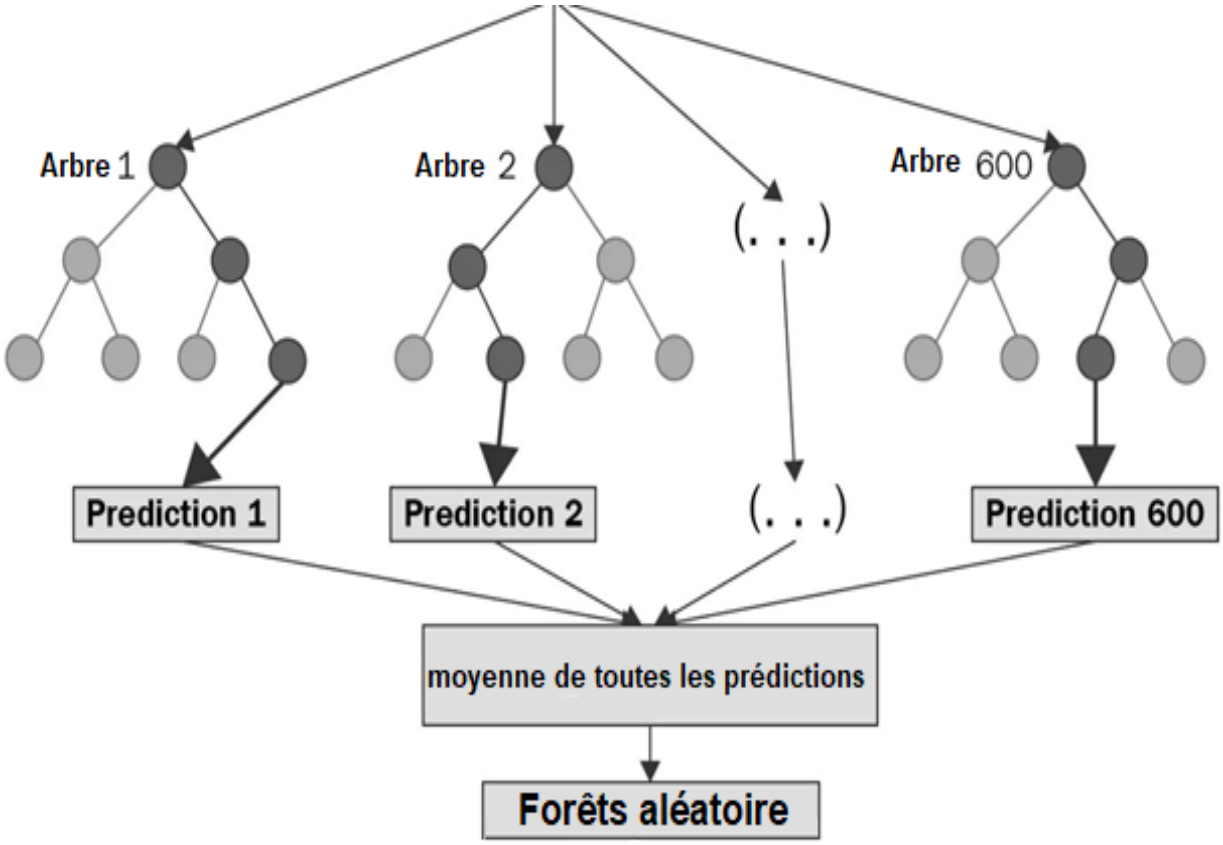
L'algorithme de forêt aléatoire fonctionne en deux phases, il s'agit d'abord de créer la forêt aléatoire en combinant **N** arbres de décision, et ensuite de faire des prédictions pour chaque arbre créé dans la première phase.[1]

Le processus de fonctionnement peut être expliqué par les étapes suivantes:

**Étape 1** : sélectionner **K** points de données aléatoires dans l'ensemble d'apprentissage.
**Étape 2** : Construire les arbres de décision associés aux points de données sélectionnés (sous-ensembles).
**Étape 3** : choisir le nombre **N** pour les arbres de décision qu'on souhaite créer.
**Étape 4** : Répéter l'étape 1 et l'étape 2.
**Étape 5** : Pour les nouveaux points de données, rechercher les prédictions de chaque arbre de décision et attribuer les nouveaux points de données à la catégorie qui remporte la majorité des votes.

[1] javatpoint. (s.d.). Random Forest Algorithm. Récupéré sur javatpoint: https://www.javatpoint.com/machine-learning-random-forest-algorithm

**Figure 3 :** Principe de fonctionnement de foret aléatoire.[1]




**Remarque importante :**

Un grand nombre d'arbres dans la forêt conduit à une plus grande précision et évite le problème de sur-apprentissage (Overfitting).

**4.2.4. Avantages des forêts aléatoires** :[2]

• Pas besoin d'élaguer les arbres.

• Précision et importance variable générées automatiquement.

• Le sur-apprentissage n'est pas un problème.

• Peu sensible aux valeurs aberrantes dans les données d'entraînement.

• Paramètres faciles à régler.

**4.2.5. Limites des forêts aléatoires :**[3]

[1] javatpoint. (s.d.),op. *cit* .
[2] Horning, N. ,op. *cit* .p 23 .
[3] Horning, N. ,op. *cit* .p 24 .

• La régression ne peut pas prédire au-delà de la plage dans les données d'entraînement
• Dans la régression, les valeurs extrêmes ne sont souvent pas prédites avec précision.

### 4-3- Machine à vecteurs de support (Support Vector Machine (SVM))

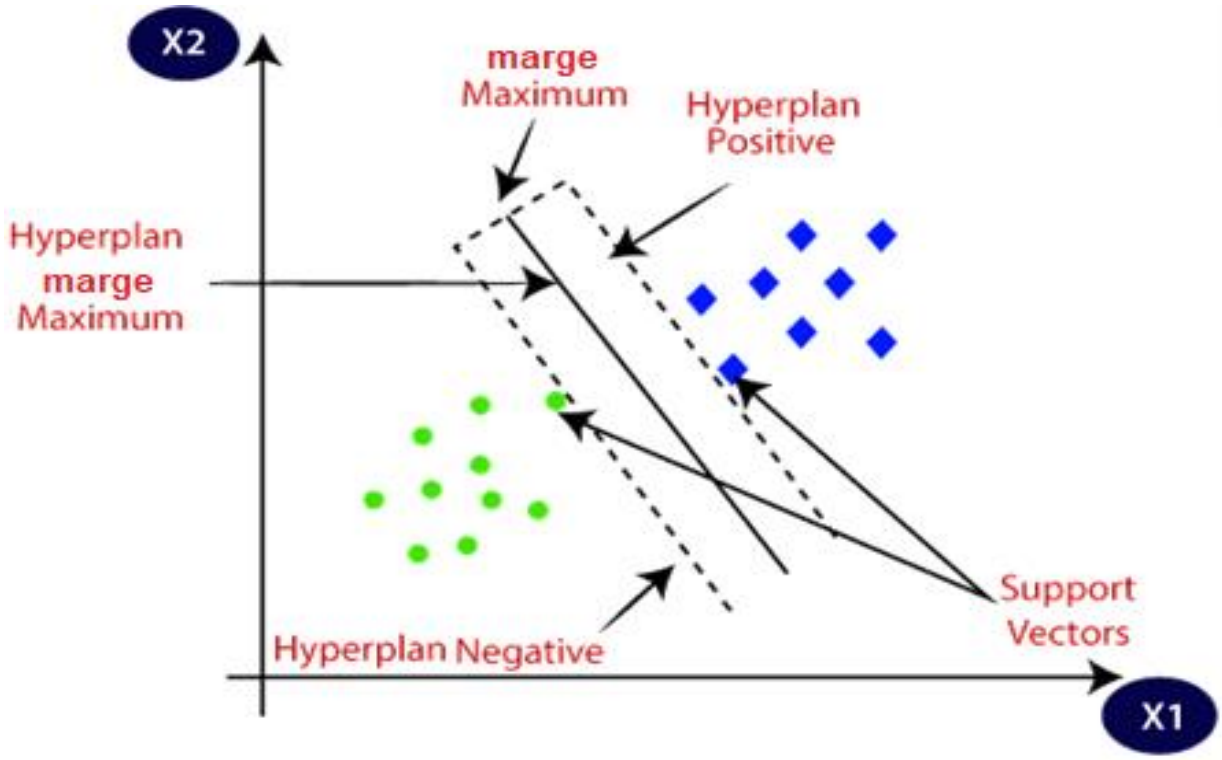
La machine à vecteurs de support (SVM) est l'un des algorithmes d'apprentissage supervisé les plus populaires, qui est utilisé pour les problèmes de classification ainsi que de régression. Cependant, il est principalement utilisé pour les problèmes de classification en apprentissage automatique.[1]

L'objectif de l'algorithme SVM est de créer la meilleure ligne ou limite de décision qui puisse séparer l'espace à **n** dimensions en classes afin que nous puissions facilement placer le nouveau point de données dans la bonne catégorie à l'avenir. Cette limite de meilleure décision est appelée ***hyperplan***.[2]

La SVM choisit les points/vecteurs extrêmes qui aident à créer l'hyperplan. Ces cas extrêmes sont appelés vecteurs de support et, par conséquent, l'algorithme est appelé machine à vecteurs de support. Pour mieux illustrer le fonctionnement du SVM, considérons le diagramme ci-dessous (voir Figure 4) dans lequel il existe deux catégories différentes qui sont classées à l'aide d'une limite de décision ou d'un hyperplan. [3]

[1] javatpoint. (s.d.),op. *cit* .
[2] javatpoint. (s.d.),op. *cit* .
[3] javatpoint. (s.d.),op. *cit* .

**Figure 4 :** Principe de fonctionnement de l'SVM [1]




En approfondissant l'aspect mathématique, les Support Vector Machines relèvent d'une catégorie d'algorithmes de Machine Learning appelés méthodes à noyau, où les caractéristiques peuvent être transformées à l'aide d'une fonction noyau. Les fonctions noyau mappent les données sur un espace dimensionnel différent (souvent plus grand) dans l'espoir que les classes soient plus faciles à séparer après cette transformation. Ceci permet potentiellement de simplifier les frontières de décisions complexes non linéaires en frontières linéaires dans un espace mappé de dimension supérieur. Dans ce processus, il n'est pas nécessaire de transformer explicitement les données, ce qui serait coûteux en calcul. C'est ce qu'on appelle communément l'astuce des noyaux.[2]

### 4.3.1. Avantages et inconvénients des machines à vecteurs de support

#### A. Avantages [3]

• Sa grande précision de prédiction

---

[1] Javatpoint. (s.d.),op. cit .

[2] Mathworks. (s.d.). Support Vector Machine (SVM). Consulté le 06 06, 2021, sur mathworks: https://fr.mathworks.com/discovery/support-vector-machine.html

[3] Issarane, H. (s.d.). Support Vector Machines. Récupéré sur analytic and sinsights.: https://analyticsinsights.io/les-svm-support-vector-machine/

• Fonctionne bien sûr de plus petits data sets

• Ils peuvent être plus efficaces car ils utilisent un sous-ensemble de points d'entraînement.

**B. Inconvénients** [1]

• Ne convient pas à des jeux de données plus volumineux, car le temps d'entraînement avec les SVM peut être long

• Moins efficace sur les jeux de données contenant du bruit.

## 5- Les réseaux de neurones

### 5.1. Définitions :

Un réseau de neurones est une série d'algorithmes qui s'efforce de reconnaître les relations sous-jacentes dans un ensemble de données grâce à un processus qui imite le fonctionnement du cerveau humain. En ce sens, les réseaux de neurones font référence à des systèmes de neurones, de nature organique ou artificielle. Les réseaux de neurones peuvent s'adapter aux changements d'entrée ; ainsi le réseau génère le meilleur résultat possible sans avoir besoin de reconcevoir les critères de sortie. [2]

En termes simples, un réseau de neurones est un graphique connecté avec des neurones d'entrée, des neurones de sortie et des bords pondérés (Figure 5).[3]

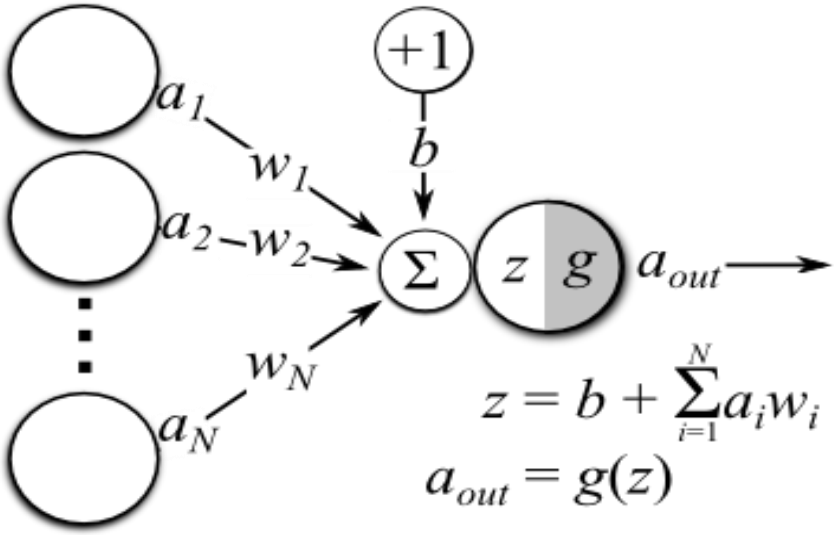
### 5.2. Composants de réseau de neurones :

**5.2.1. Neurone** : C'est l'unité de base d'un réseau de neurones. Il obtient un certain nombre d'entrées et une valeur de biais. Lorsqu'un signal (valeur) arrive, il est multiplié par une valeur de poids. Si un neurone a 4 entrées, il a 4 valeurs de poids qui peuvent être ajustées pendant le temps d'entraînement.[4]

$$z = x_1 * w_1 + x_2 * w_2 + \ldots\ldots + x_n * w_n + b * 1$$

[1] Issarane, H.op.cit .
[2] Touzet, C. (1992). Les réseaux de neurones artificiels, introduction au connexionnisme. HAL.
[3] Ibid.
[4] Ahirwar, K. (2017). Everything you need to know about Neural Networks. Consulté le 06 06, 2021, sur hackernoon : https://hackernoon.com/everything-you-need-to-know-about-neural-networks-8988c3ee4491

**Figure 5 :** Structure de perceptron simple, source (Ahirwar, 2017)

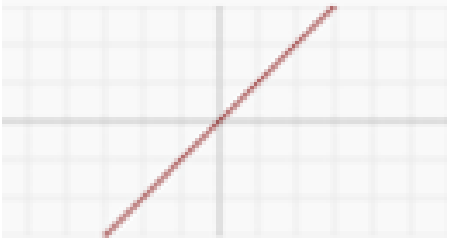



**5.2.2. Connexions** : Il connecte un neurone d'une couche à un autre neurone d'une autre couche ou de la même couche. Une connexion a toujours une valeur de poids qui lui est associée. Le but de l'entraînement est de mettre à jour cette valeur de poids pour diminuer la perte.[1]

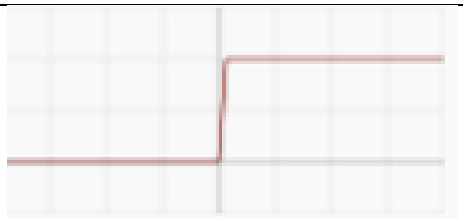
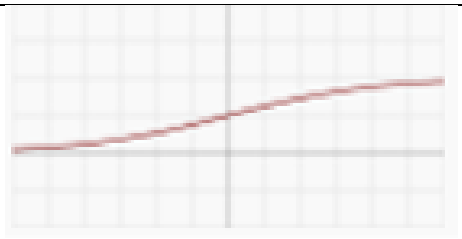
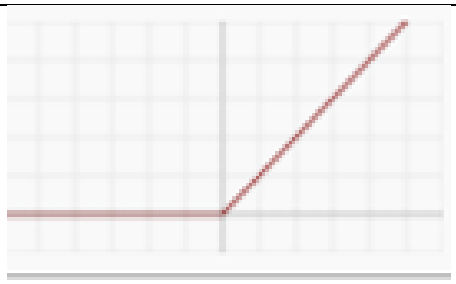
**5.2.3. Bais** : C'est une entrée supplémentaire pour les neurones et c'est toujours 1, et a son propre poids de connexion. Cela garantit que même lorsque toutes les entrées sont nulles, il y aura une activation dans le neurone.[2]

**5.2.4. Fonction d'activation (fonction de transfert)** : Les fonctions d'activation sont utilisées pour introduire la non-linéarité dans les réseaux de neurones. Il écrase les valeurs dans une plage plus petite, à savoir. Une fonction d'activation sigmoïde écrase les valeurs comprises entre 0 et 1. Il existe de nombreuses fonctions d'activation comme illustré dans le tableau 1.[3]

**Tableau 1 :** Quelques fonctions d'activation.[4]

| fonction | Courbe | Equation | Dérivé |
|---|---|---|---|
| Identité |  | $f(x) = x$ | $f'(x) = 1$ |

[1] Ahirwar, K. (2017). op. *cit*.
[2] Ibid.
[3] Ibid.
[4] Ibid.

| | | | |
|---|---|---|---|
| Binaire |  | $f(x) = \begin{cases} 0\ si\ \ x < 0 \\ 1\ si\ x \geq 0 \end{cases}$ | $f'(x) = \begin{cases} 0\ si\ \ x \neq 0 \\ ?\ \ si\ x = 0 \end{cases}$ |
| Logistique |  | $f(x) = \frac{1}{1 + e^{-x}}$ | $f'(x) = f(x)(1 - f(x)$ |
| ReLu |  | $f(x) = \begin{cases} 0\ si\ \ x < 0 \\ x\ si\ x \geq 0 \end{cases}$ | $f'(x) = \begin{cases} 0\ si\ \ x < 0 \\ 1\ si\ x \geq 0 \end{cases}$ |

### 5.3. Dispositions de base du réseau de neurones

**5.3.1. Couche d'entrée :** Il s'agit de la première couche du réseau de neurones. Il prend les signaux d'entrée (valeurs) et les transmet à la couche suivante. Il n'applique aucune opération sur les signaux d'entrée (valeurs) et n'a pas de valeurs de poids et de biais associées. [1]

**5.3.2. Couches cachées :** Les couches cachées ont des neurones (nœuds) qui appliquent différentes transformations aux données d'entrée. Une couche cachée est une collection de neurones empilés verticalement (représentation). La dernière couche cachée transmet les valeurs à la couche de sortie. Tous les neurones d'une couche cachée sont connectés à chaque neurone de la couche suivante. [2]

**5.3.3. Couche de sortie :** Cette couche est la dernière couche du réseau et reçoit l'entrée de la dernière couche cachée. Avec cette couche, nous pouvons obtenir le nombre de valeurs souhaité et dans une plage souhaitée. [3]

---

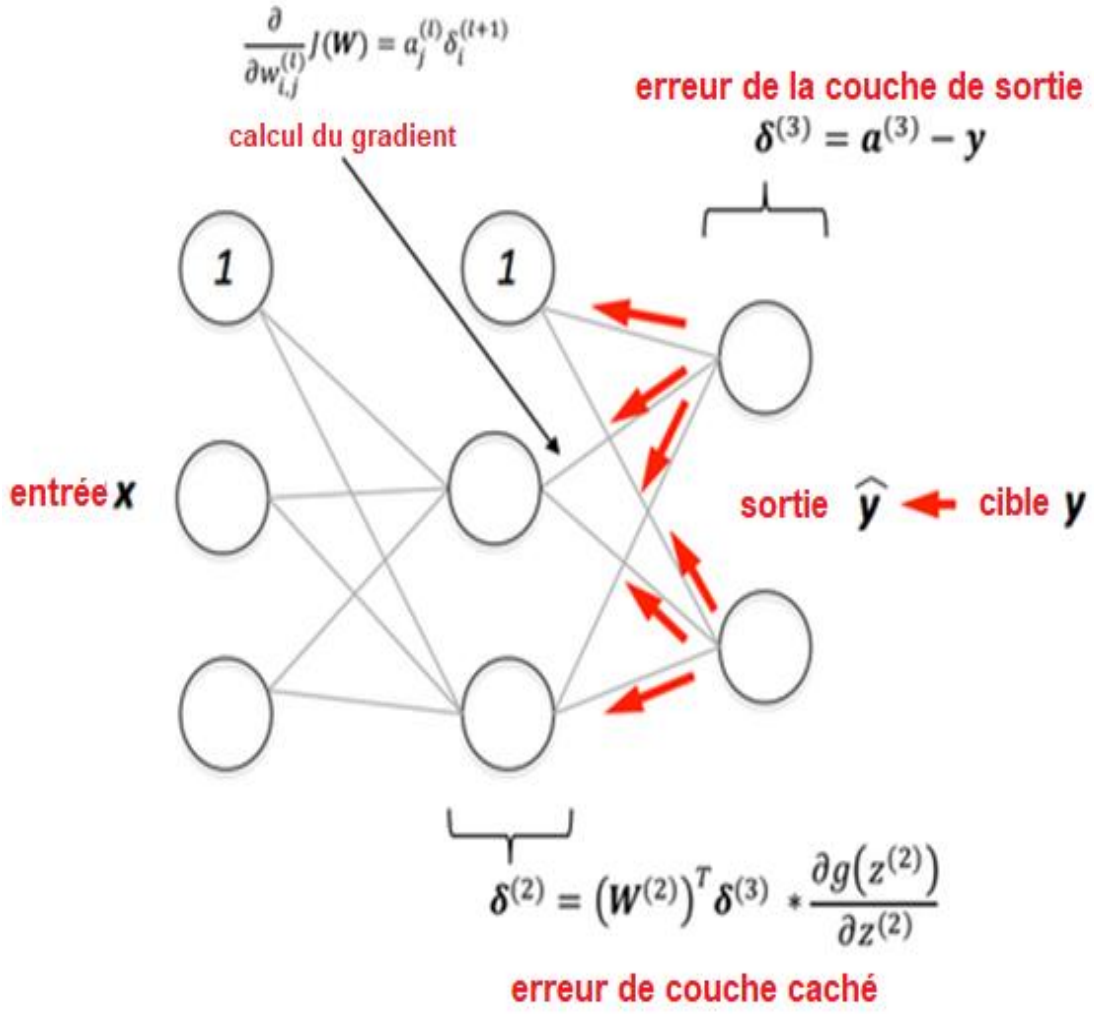
[1] Ahirwar, K. (2017). op. *cit*.
[2] Ibid.
[3] Ibid.

**5.3.4. Poids (Paramètres)** : Un poids représente la force de la connexion entre les unités. Si le poids du nœud 1 au nœud 2 a une plus grande amplitude, cela signifie que le neurone 1 a une plus grande influence sur le neurone 2. Un poids réduit l'importance de la valeur d'entrée. Des poids proches de zéro signifient que la modification de cette entrée ne modifiera pas la sortie. Les poids négatifs signifient que l'augmentation de cette entrée diminuera la sortie. Un poids décide de l'influence que l'entrée aura sur la sortie. [1]

**5.3.5. Propagation vers l'avant :** La propagation vers l'avant est un processus qui consiste à fournir des valeurs d'entrée au réseau de neurones et à obtenir une sortie que nous appelons valeur prédite (Voir Figure 6). Parfois, nous appelons la propagation vers l'avant l'inférence. Lorsque nous transmettons les valeurs d'entrée à la première couche du réseau de neurones, cela se passe sans aucune opération. La deuxième couche prend les valeurs de la première couche et applique les opérations de multiplication, d'addition et d'activation et passe cette valeur à la couche suivante. Le même processus se répète pour les couches suivantes et finalement nous obtenons une valeur de sortie de la dernière couche.[2]

[1] Ahirwar, K. (2017). op. *cit*.
[2] Ibid.

**Figure 6 :** Propagation vers l'avant (Feed Forward).[1]




**5.3.6. Rétro-propagation** : Après la propagation vers l'avant, nous obtenons une valeur de sortie qui est la *valeur prédite*. Pour calculer l'erreur, nous comparons la valeur prédite avec la *valeur de sortie réelle*. Nous utilisons une *fonction de perte* (mentionnée ci-dessous) pour calculer la *valeur d'erreur.* Ensuite, nous calculons la dérivée de la *valeur d'erreur* par rapport à chaque poids dans le réseau de neurones. La rétro propagation utilise la règle de la chaîne du calcul différentiel. Dans la règle de la chaîne, nous calculons d'abord les dérivées de *la valeur d'erreur* par rapport aux *valeurs* de *poids* de la dernière couche. Nous appelons ces dérivés des *gradients* et utilisons ces *gradients* valeurs pour calculer les *gradients* de l'avant-dernière couche. Nous répétons ce processus jusqu'à ce que nous obtenions des *gradients* pour chaque poids de notre réseau de neurones. Ensuite, nous soustrayons cette *valeur* de *gradient de la*

[1] Ahirwar, K. (2017). op. *cit.*

*valeur* de *poids* pour réduire la *valeur d'erreur*. De cette façon, nous nous rapprochons (descente) *des minima locaux* (signifie la perte minimale). [1]

## 6- Les paramètres usuels dans les algorithme d'apprentissage

**6.1. Taux d'apprentissage :** Le taux d'apprentissage détermine la rapidité ou la lenteur avec laquelle nous souhaitons mettre à jour nos valeurs de poids (paramètres). Le taux d'apprentissage doit être suffisamment élevé pour qu'il ne prenne pas des années à converger, et il doit être suffisamment faible pour qu'il trouve le minima local. [2]

**6.2. Précision :** La précision fait référence à la proximité de deux ou plusieurs mesures les unes par rapport aux autres. C'est la répétabilité ou la reproductibilité de la mesure. [3]

**6.3. Rappel (Sensibilité) :** Le rappel fait référence à la fraction d'instances pertinentes qui ont été récupérées par rapport au nombre total d'instances pertinentes

$$Precision = \frac{tp}{tp + fp}$$

$$Sensibilité\ = \frac{tp}{tp + fn}$$

$tp$ = true positive (vrai positif) , $fp$ = false positive (faux positif) , $fn$= false negative (faux négatif) [4]

**6.4. Matrice de corrélation**

Une matrice de corrélation est un tableau montrant les coefficients de corrélation entre les variables. Chaque cellule du tableau montre la corrélation entre deux variables. Une matrice de corrélation est utilisée pour résumer les données, comme entrée dans une analyse plus avancée et comme diagnostic pour les analyses avancées. [5]

Les décisions clés à prendre lors de la création d'une matrice de corrélation comprennent : le choix de la statistique de corrélation, le codage des variables, le traitement des données manquantes et la présentation. [6]

---

[1] Ahirwar, K. (2017). op. *cit*.
[2] Ibid.
[3] Ibid.
[4] Ibid.
[5] Ibid.
[6] Ibid.

**6.5. Matrice de confusion :**

Dans le domaine de l'apprentissage automatique et plus précisément du problème de la classification statistique , une matrice de confusion , également appelée matrice d'erreur, est une disposition de tableau spécifique qui permet de visualiser les performances d'un algorithme, typiquement un apprentissage supervisé (en apprentissage non supervisé, il est généralement appelée matrice d'appariement ). Chaque ligne de la matrice représente les instances d'une classe prédite tandis que chaque colonne représente les instances d'une classe réelle (ou vice versa) (Voir Tableau 2). Le nom vient du fait qu'il permet de voir facilement si le système confond deux classes (c'est-à-dire en étiquetant couramment l'une comme l'autre). [1]

**Tableau 2 :** Exemple d'une matrice de confusion

| | | classe réelle | | |
|---|---|---|---|---|
| | | chatte | Chien | Lapine |
| classe prévue | chatte | 5 | 2 | 0 |
| | Chien | 3 | 3 | 2 |
| | Lapine | 0 | 1 | 11 |

**6.6. Fonction de perte/Fonction de coût** : La fonction de perte calcule l'erreur pour un seul exemple d'apprentissage. La fonction de coût est la moyenne des fonctions de perte de l'ensemble d'apprentissage complet. **[2]**

**6.7. Optimiseurs de modèle** : L'optimiseur est une technique de recherche utilisée pour mettre à jour les pondérations dans le modèle. [3]

**6.8. Métriques de performances** : Les métriques de performances sont utilisées pour mesurer les performances des algorithmes d'apprentissage. La précision, la perte, la précision de validation, la perte de validation, l'erreur absolue moyenne, la précision, le rappel et le score ***f1*** sont quelques mesures de performance.[4]

[1] Ahirwar, K. (2017). op. *cit*.
[2] Ibid.
[3] Ibid.
[4] Ibid.

## 7- Conclusion

Dans ce chapitre, nous avons présenté un aperçu de l'apprentissage automatique. Nous avons, ensuite, discuté la différence entre la régression et la classification et nous avons décrit les paramètres les plus importants de l'apprentissage automatique.

Nous avons, également, expliqué les outils les plus importants et nécessaires pour accomplir notre projet, à savoir : les arbres de décision, les forêts aléatoires, les SVM et les réseaux de neurones

# Chapitre 2

# L'apprentissage automatique dans le domaine médical

# (Cas particulier : la prédiction de la tension artérielle)

## Introduction

Le domaine de la santé est l'un des domaines de recherche les plus importants dans le scénario actuel avec l'amélioration rapide de la technologie et des données. Il est difficile de gérer l'énorme quantité de données des patients. Il est plus facile de gérer ces données grâce à l’analyse de données volumineuses. Il existe de nombreuses procédures pour le traitement de plusieurs maladies à travers le monde. L'apprentissage automatique est une approche émergente qui aide à la prédiction et au diagnostic d'une maladie. Dans ce chapitre, on va aborder le problème de la prédiction d’une maladie à l'aide de l'apprentissage automatique et nous définirons également la maladie de l'hypertension et expliquerons sa propagation dans le monde et en Algérie en particulier.

### 1- L’intelligence artificielle et la médecine

Avant que l'IA ne commence à être appliquée aux informations médicales dans les années 2000, les modèles prédictifs dans le domaine de la santé ne pouvaient prendre en compte que des variables limitées dans des données de santé propres et bien organisées. Aujourd'hui, on constate que c’est l’air des outils sophistiqués d'apprentissage automatique qui utilisent des algorithmes comme les réseaux de neurones artificiels, les arbres de décision et les forets aléatoires pour apprendre des relations extrêmement complexes ou un apprentissage en profondeur. Il a été démontré que les technologies soutiennent - et parfois dépassent- les capacités humaines dans l'exécution de certaines tâches médicales. Les systèmes d'IA sont conçus pour traiter les données complexes générées par les soins cliniques modernes.[65]

Les outils basés sur l'IA peuvent identifier des relations significatives dans les données brutes et ont le potentiel d'être appliqués dans presque tous les domaines de la médecine, y compris le développement de médicaments, les décisions de traitement, les soins aux patients et les décisions financières et opérationnelles.[66]

Avec l'IA, les professionnels de la santé pourraient s'attaquer à des problèmes complexes qu'il serait difficile, chronophage ou inefficace de résoudre seuls. L'IA pourrait être une ressource précieuse pour les professionnels de la santé, leur permettant de mieux utiliser leur expertise et de fournir de la valeur à l'ensemble de l'écosystème de la santé.[67]

### 2- Apports de l'intelligence artificielle en médecine

Les outils compatibles avec l'IA peuvent extraire des informations pertinentes à partir de grandes quantités de données et générer des informations exploitables qui pourraient être appliquées à de nombreuses applications :[68]

#### 2.1. Aperçu des traitements de surface

Avec les technologies d'IA, les médecins pourraient trouver des informations dans la littérature médicale non structurée pour soutenir les décisions de soins.

---

[65] Ibm. (2021). *Artificial intelligence in medicine.* Récupéré sur ibm: https://www.ibm.com/watson-health/learn/artificial-intelligence-medicine

[66] Ibid.

[67] Ibid.

[68] Ibid.

### 2.2. Soutenir les besoins des utilisateurs

L'IA peut rechercher et présenter des données pour aider les gens à utiliser des informations complètes sur la santé, ce qui pourrait conduire à des utilisateurs plus informés.

### 2.3. Identifier les informations à partir des données des patients

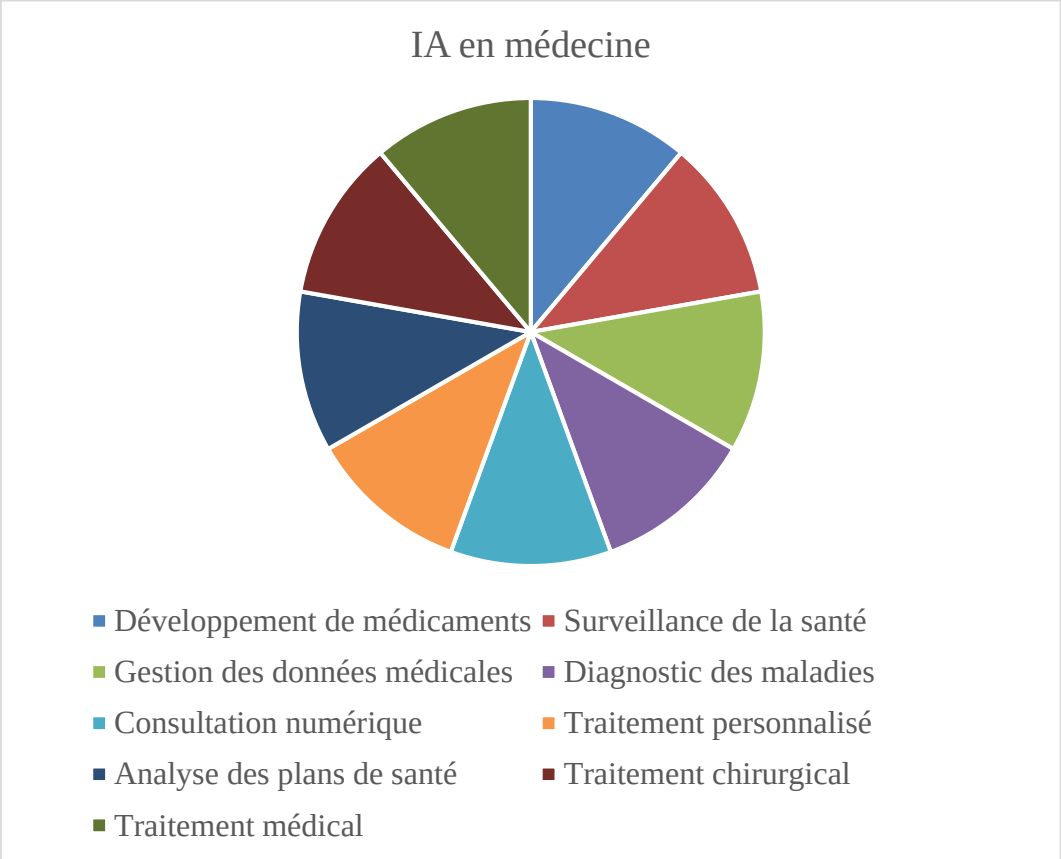
Les outils d'IA pourraient rechercher des dossiers médicaux structurés et non structurés pour fournir des antécédents de patients pertinents.

### 2.4. Cibler les similitudes et les modèles

L'IA pourrait identifier des modèles et aider les chercheurs à créer des cohortes de patients dynamiques pour des études et des essais cliniques.

Le diagramme ci-dessous résume les grandes applications de l'IA en médecine.

**Figure 7 :** Domaines d'application de L'IA en médecine.[69]




[69] Amisha, P. M. (2019 , juillet). Overview of artificial intelligence in medicine. journal of medecine and primary care (PMCID: PMC6691444). Récupéré sur : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6691444/

### 3- Types d'IA en médecine

L'IA en médecine peut être dichotomisée en deux sous-types : ***virtuel*** et ***physique***. [70]

- La ***partie virtuelle*** va des applications telles que les systèmes de dossiers de santé électroniques aux conseils basés sur les réseaux neuronaux dans les décisions de traitement.
- La ***partie physique*** traite des robots d'assistance aux interventions chirurgicales, des prothèses intelligentes pour les personnes handicapées et des soins aux personnes âgées.

### 4- L'apprentissage automatique et la prédiction des maladies

L'apprentissage automatique, comme déjà mentionné dans le chapitre précédent, est un domaine qui relie le problème de l'apprentissage à partir d'échantillons de données au concept général d'inférence. Chaque processus d'apprentissage se compose de deux phases : (i) l'estimation des dépendances inconnues dans un système à partir d'un ensemble de données donné et (ii) l'utilisation des dépendances estimées pour prédire de nouvelles sorties du système. Ainsi, l'apprentissage automatique joue un rôle très intéressant dans la recherche biomédicale avec de nombreuses applications, où une généralisation acceptable est obtenue en recherchant dans un espace n-dimensionnel pour un ensemble donné d'échantillons biologiques, en utilisant différentes techniques et algorithmes.[71]

**Exemple :**

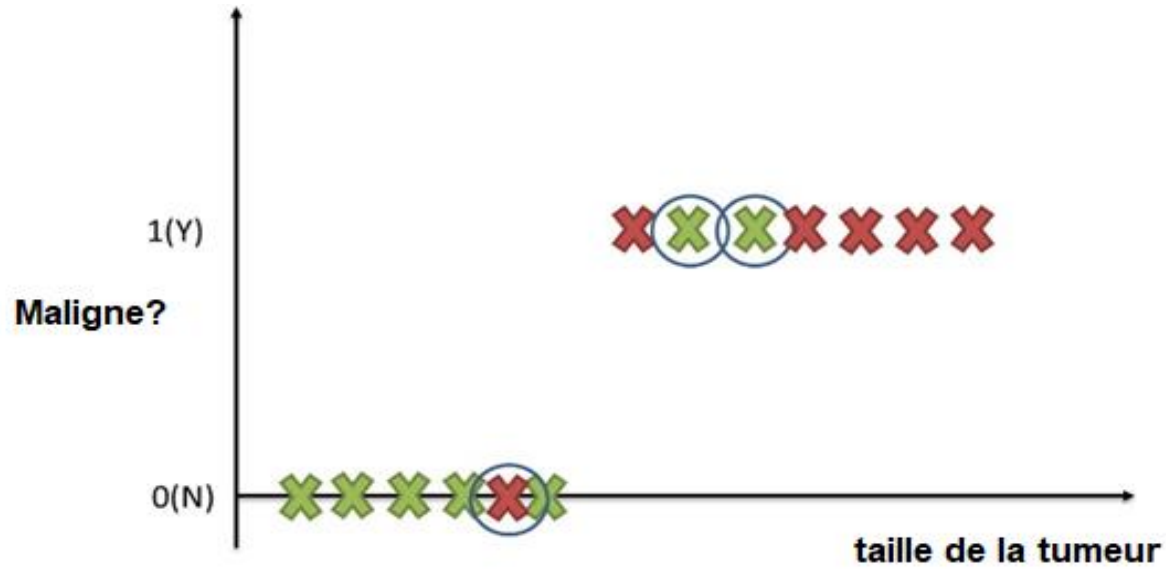
Supposons par exemple que nous ayons collecté des dossiers médicaux relatifs au cancer du sein et que nous essayions de prédire si une tumeur est maligne ou bénigne en fonction de sa taille. La question serait renvoyée à l'estimation de la probabilité que la tumeur soit maligne ou non (1 = Oui, 0 = Non). Dans la Figure 8, les enregistrements encerclés décrivent toute erreur de classification du type de tumeur produite par la procédure.

**Figure 8. :** Illustration du processus de classification d'une tumeur maligne ou non.

---

[70] Amisha, P. M.op.*cit*.

[71] Konstantina Kourou, T. P. (2015, 8 17). Machine learning applications in cancer prognosis and prediction. (Elsevier, Éd.) *Computational and Structural Biotechnology Journal, 13*, 9. Récupéré sur www.elsevier.com/locate/c sbj



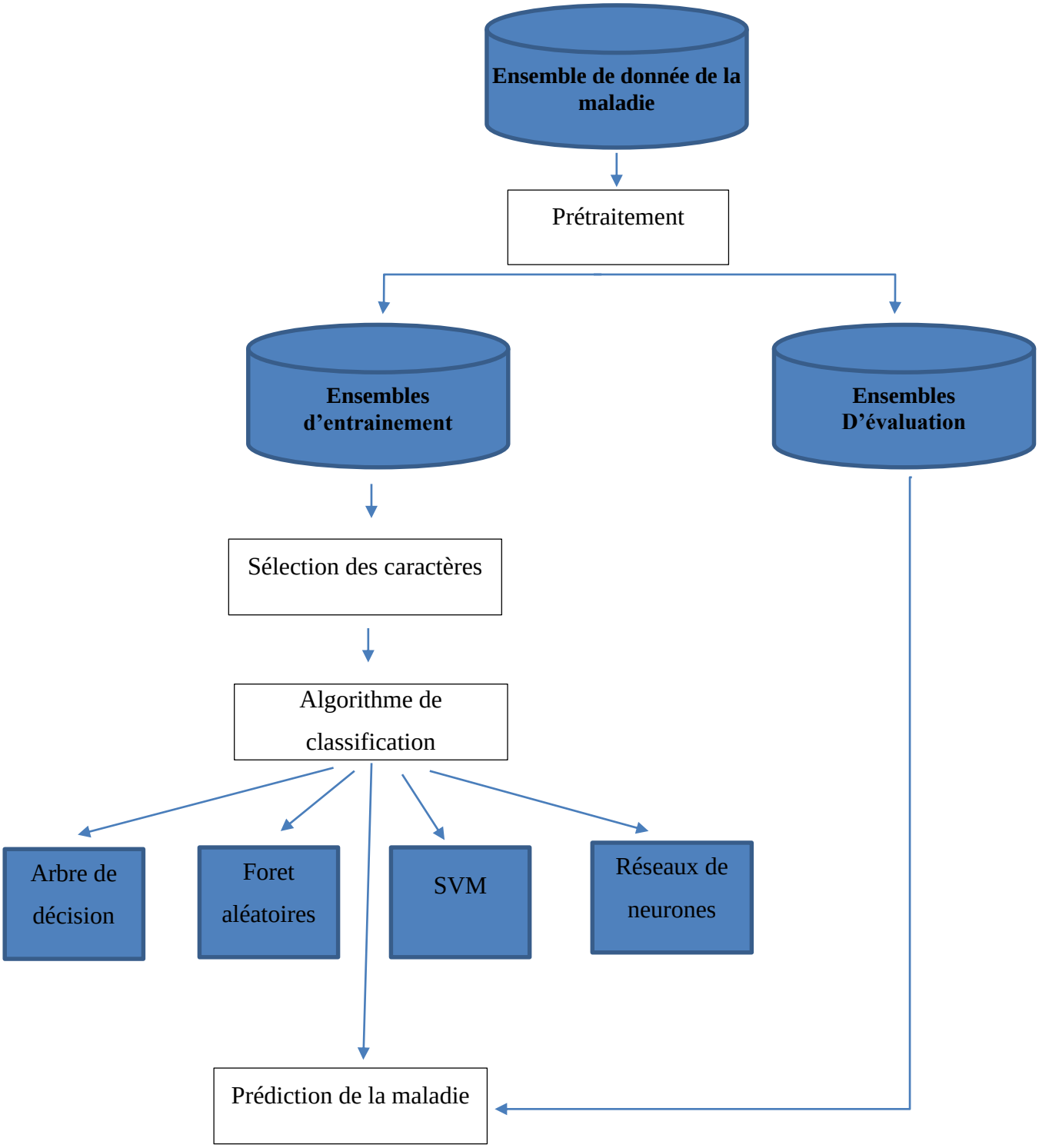

### 5- Procédure d'application d'algorithme d'apprentissage sur un ensemble de donnée lors de prédiction d'une maladie

Lors de l'application d'une méthode d'apprentissage automatique, les échantillons de données constituent les composants de base. Chaque échantillon est décrit avec plusieurs caractéristiques et chaque caractéristique se compose de différents types de valeurs. De plus, connaître à l'avance le type spécifique de données utilisées permet de sélectionner correctement les outils et les techniques pouvant être utilisés pour leur analyse. Certains problèmes liés aux données font référence à la qualité des données et aux étapes de prétraitement pour les rendre plus adaptées. Les problèmes de qualité des données incluent la présence de bruit, de valeurs aberrantes, de données manquantes ou en double et de données biaisées ou non représentatives. Lors de l'amélioration de la qualité des données, la qualité de l'analyse résultante est généralement également améliorée.[72]

Les différentes étapes de prédiction d'une maladie en utilisant un algorithme d'apprentissage automatique sont illustrées dans la Figure 9.

[72] Konstantina Kourou, T. P.op.*cit* .

**Figure 9 :** Etapes de prédiction d'une maladie avec les algorithmes d'apprentissage




### 5.1. Le Prétraitement des données

Afin de rendre les données brutes plus adaptées à une analyse plus approfondie, des étapes de prétraitement doivent être appliquées, axées sur la modification des données. Il existe un certain nombre de techniques et de stratégies différentes, pertinentes pour le prétraitement

des données, qui se concentrent sur la modification des données pour un meilleur ajustement dans une méthode de ML spécifique. Parmi ces techniques, certaines des approches les plus importantes incluent (i) la réduction de la dimensionnalité (ii) la sélection de caractéristiques et (iii) l'extraction de caractéristiques. Il existe de nombreux avantages concernant la réduction de la dimensionnalité lorsque les jeux de données ont un grand nombre de fonctionnalités. Les algorithmes d'apprentissage automatique fonctionnent mieux lorsque la dimensionnalité est plus faible. De plus, la réduction de la dimensionnalité peut éliminer les caractéristiques non pertinentes, réduire le bruit et produire des modèles d'apprentissage plus robustes en raison de l'implication de moins de caractéristiques. En général, la réduction de la dimensionnalité en sélectionnant de nouvelles caractéristiques qui sont un sous-ensemble des anciennes est connu sous le nom de sélection de caractéristiques. [73]

### 5.2. Sélection des caractéristiques :

Il existe trois approches principales pour la sélection de caractéristiques : ***les approches intégrées***, ***par filtre*** et ***par enveloppe***. Dans le cas de l'extraction d'entités, un nouvel ensemble d'entités peut être créé à partir de l'ensemble initial qui capture toutes les informations importantes dans un jeu de données. La création de nouveaux ensembles de fonctionnalités permet de rassembler les avantages décrits de la réduction de la dimensionnalité.

Cependant, l'application de techniques de sélection de caractéristiques peut entraîner des fluctuations spécifiques concernant la création de listes de caractéristiques prédictives. [74]

### 5.3. Production d'un model

L'objectif principal des techniques d'apprentissage automatique est de produire un modèle pouvant être utilisé pour effectuer une classification, une prédiction, une estimation ou toute autre tâche similaire. La tâche la plus courante dans le processus d'apprentissage est la classification. Comme mentionné précédemment, cette fonction d'apprentissage classe l'élément de données dans l'une de plusieurs classes prédéfinies. Un bon modèle de classification doit bien s'adapter à l'ensemble d'apprentissage et classer avec précision toutes les instances. Si les taux d'erreur de test d'un modèle commencent à augmenter même si les taux d'erreur d'apprentissage diminuent, le phénomène de sur ajustement du modèle se produit. Cette situation est liée à la complexité du modèle, ce qui signifie que les erreurs d'apprentissage d'un

[73] Konstantina Kourou, T. P.op.*cit.*
[74] Ibid.

modèle peuvent être réduites si la complexité du modèle augmente. Évidemment, la complexité idéale d'un modèle non susceptible de sur ajustement est celle qui produit l'erreur de généralisation la plus faible. [75]

### 5.4. L'analyse des erreurs

Une méthode formelle pour analyser l'erreur de généralisation attendue d'un algorithme d'apprentissage est la décomposition biais-variance. La composante de biais d'un algorithme d'apprentissage particulier mesure le taux d'erreur de cet algorithme. De plus, une deuxième source d'erreur sur tous les ensembles d'apprentissage possibles de taille donnée et tous les ensembles de test possibles est appelée variance de la méthode d'apprentissage. L'erreur globale attendue d'un modèle de classification est constituée de la somme du biais et de la variance, à savoir la décomposition biais-variance.[76]

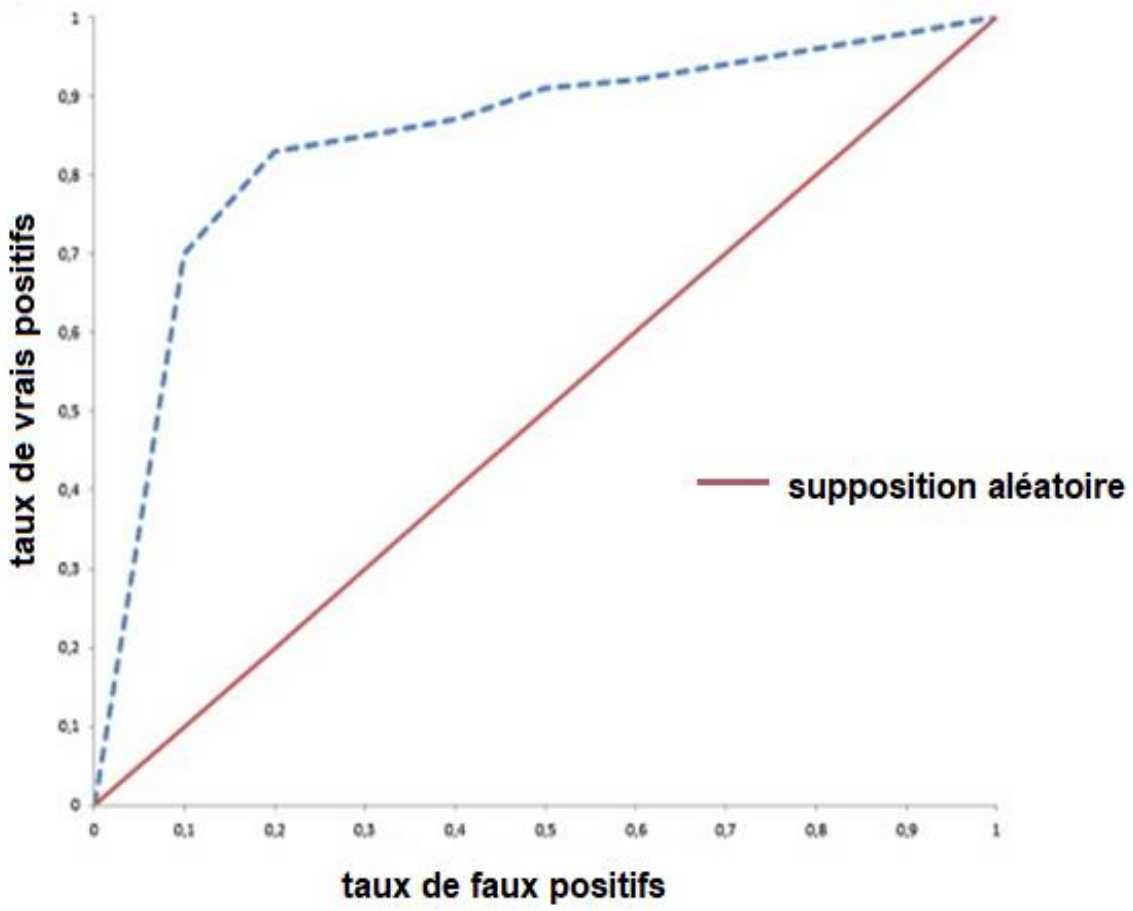
### 5.5. Estimation de performance de model

Une fois qu'un modèle de classification est obtenu à l'aide d'une ou plusieurs techniques d'apprentissage, il est important d'estimer les performances. L'analyse des performances de chaque modèle proposé est mesurée en termes de sensibilité, de spécificité, de précision et d'aire sous la courbe. La sensibilité est définie comme la proportion de vrais positifs qui sont correctement observés par le classificateur, tandis que la spécificité est donnée par la proportion de vrais négatifs qui sont correctement identifiés. Les mesures quantitatives de précision et d'aire sous la courbe sont utilisées pour évaluer la performance globale. Plus précisément, la précision est une mesure liée au nombre total de prédictions correctes. Au contraire, d'aire sous la courbe est une mesure de la performance du modèle qui est basée sur la courbe Classificateur aléatoire.[77]

---

[75] Konstantina Kourou, T. P.op.*cit.*
[76] Ibid.
[77] Ibid.

**Figure 10 :** Une courbe indicative de deux classificateurs : (a) un classificateur Random Guess (courbe rouge) et (b) un classificateur fournissant des prédictions plus robustes (courbe en pointillé bleu). [78]




### 6- Les systèmes d'aide à la décision clinique

Un système d'aide à la décision clinique (SADC) est destiné à améliorer la prestation des soins de santé en améliorant les décisions médicales avec des connaissances cliniques ciblées, des informations sur les patients et d'autres informations sur la santé. Un SADC traditionnel est composée d'un logiciel conçu pour être une aide directe à la prise de décision clinique. Dans un tel système, les caractéristiques d'un patient individuel sont comparées à une base de connaissances cliniques informatisée et des évaluations ou des recommandations spécifiques au patient qui sont ensuite présentées au clinicien pour une décision. Les SADC sont aujourd'hui

[78] Konstantina Kourou, T. P.op.*cit.*

principalement utilisées au point de service, pour le clinicien à combiner ses connaissances avec les informations ou suggestions fournies par le SADC. Cependant, de plus en plus, des SADC sont développées avec la capacité d'exploiter des données et des observations autrement impossibles à obtenir ou à interpréter par les humains. [79]

Les SADC ont été classés et subdivisés en diverses catégories et types, y compris le moment de l'intervention, et s'ils ont une prestation active ou passive. Les SADC sont fréquemment classées comme ***basés sur la connaissance*** ou **non basés sur la connaissance**. [80]

- Dans les systèmes basés sur la connaissance, des règles (instructions IF-THEN) sont créées, le système récupérant les données pour évaluer la règle et produire une action ou une sortie, Les règles peuvent être établies en utilisant des preuves basées sur la littérature, la pratique ou dirigées par le patient.
- Les SADC qui ne sont pas basés sur les connaissances nécessitent toujours une source de données, mais la décision s'appuie sur l'intelligence artificielle (IA), l'apprentissage automatique (ML) ou la reconnaissance statistique des modèles, plutôt que d'être programmé pour suivre les connaissances médicales d'experts.

Les SADC non basés sur les connaissances sont composés de :[81]

1) une base de règles qui sont programmées dans le système (basées sur les connaissances), l'algorithme utilisé pour modéliser la décision (basées sur les connaissances), ainsi que les données disponibles,

2) un moteur d'inférence qui prend les règles programmées ou déterminées par l'IA et les structures de données, et les applique aux données cliniques du patient pour générer une sortie ou une action, qui est présentée à l'utilisateur final (par exemple, un médecin).

3) Et, un mécanisme de communication le site Web qui consiste en une application ou interface frontale, avec laquelle l'utilisateur final interagit avec le système.

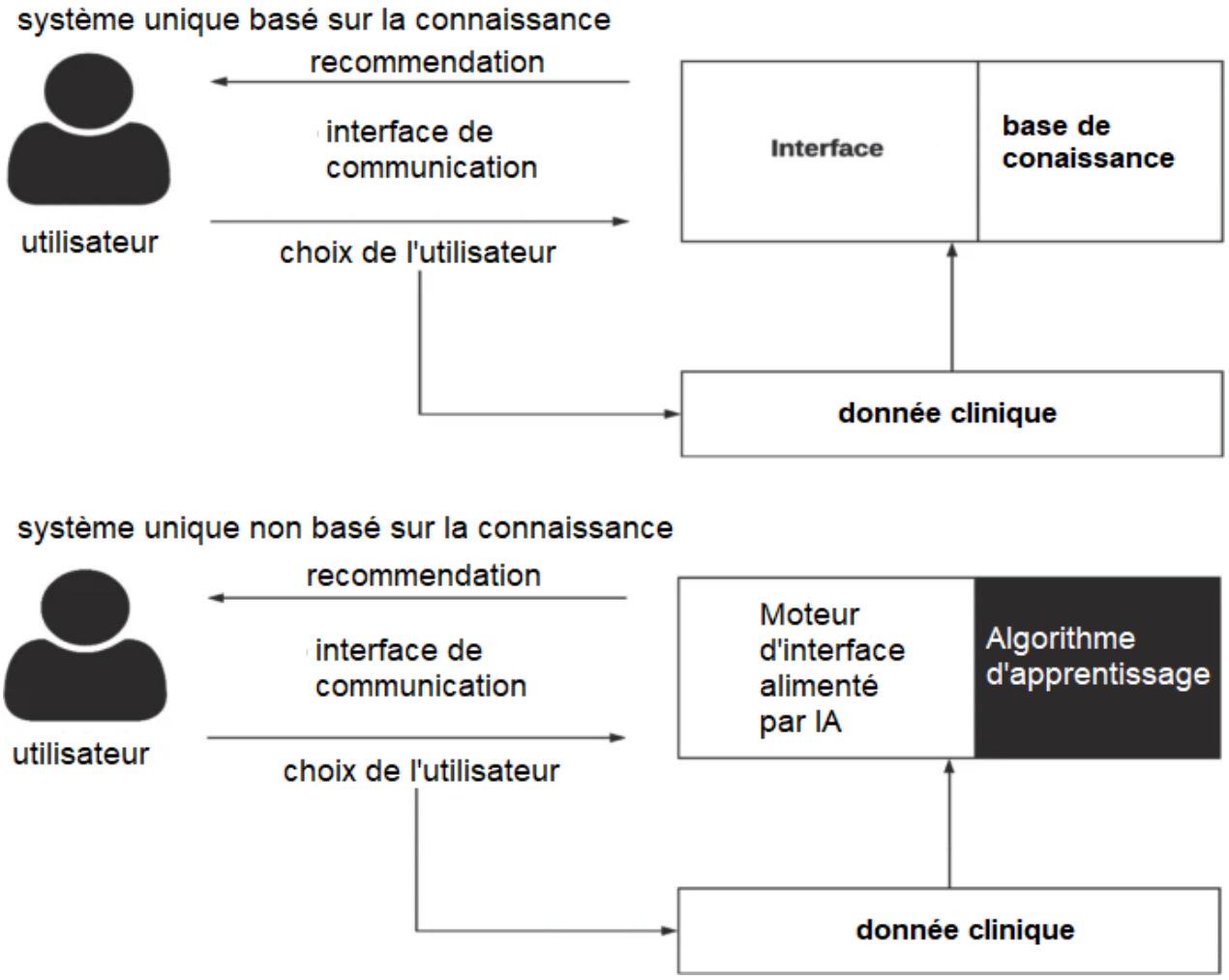
---

[79] Reed T. Sutton, D. P. (2020, 2 6). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*(17). Consulté le 06 07, 2021, sur https://www.nature.com/articles/s41746-020-0221-y , p 1.

[80] Ibid.

[81] Ibid. p 3.

**Figure 11 :** Diagramme des interactions des SADC.[82]




## 7- Maladie de l'hypertension artérielle :

### 7.1. Définition

L'hypertension artérielle (hypertension) est une affection courante dans laquelle la force à long terme du sang contre les parois des artères est suffisamment élevée pour éventuellement causer des problèmes de santé, comme une maladie cardiaque. [83]

La pression artérielle est déterminée, à la fois, par la quantité de sang que le cœur pompe et par la résistance au flux sanguin dans des artères. Plus le cœur pompe de sang et plus les

[82] Reed T. Sutton, op .cit .p 2 .
[83] Mayo, p. d. (2021, 07 01). High blood pressure (hypertension). Consulté le 07 03, 2021, sur mayo clinic: https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-

artères sont étroites, plus la tension artérielle est élevée. Une lecture de la pression artérielle est donnée en millimètres de mercure (mm Hg). Il y a deux nombres :[84]

- **Numéro supérieur (pression systolique).** Le premier chiffre, ou chiffre supérieur, mesure la pression dans les artères lorsque le cœur bat.
- **Nombre inférieur (pression diastolique).** Le deuxième nombre, ou inférieur, mesure la pression dans les artères entre les battements.

### 7.2. Symptômes

La plupart des personnes souffrant d'hypertension artérielle ne présentent aucun signe ni symptôme, même si les lectures de tension artérielle atteignent des niveaux dangereusement élevés.

Quelques personnes souffrant d'hypertension peuvent avoir des maux de tête, un essoufflement ou des saignements de nez, mais ces signes et symptômes ne sont pas spécifiques et n'apparaissent généralement que lorsque l'hypertension a atteint un stade grave ou potentiellement mortel.[85]

### 7.3. Facteurs de risque

L'hypertension artérielle comporte de nombreux facteurs de risque, notamment :[86]

- Âge.
- Course.
- Histoire de famille.
- Être en surpoids ou obèse.
- Ne pas être physiquement actif.
- Utiliser du tabac.
- Trop de sel (sodium) dans l’alimentation.
- Trop peu de potassium dans l’alimentation.
- Boire trop d'alcool.
- Stress
- Certaines maladies chroniques.

---

[84] Ibid.
[85] Ibid.
[86] Ibid.

***Remarque.*** Parfois, la grossesse contribue également à l'hypertension artérielle.

### 7.4. Complications

Une hypertension artérielle non contrôlée peut entraîner des complications, notamment :[87]

- **Crise cardiaque ou accident vasculaire cérébral.** L'hypertension artérielle peut provoquer un durcissement et un épaississement des artères (athérosclérose), ce qui peut entraîner une crise cardiaque, un accident vasculaire cérébral ou d'autres complications.
- **Anévrisme.** L'augmentation de la pression artérielle peut affaiblir et gonfler vos vaisseaux sanguins, formant un anévrisme. Si un anévrisme se rompt, il peut mettre la vie en danger.
- **Insuffisance cardiaque.** Pour pomper le sang contre la pression plus élevée dans les vaisseaux, le cœur doit travailler plus fort. Cela provoque un épaississement des parois de la chambre de pompage du cœur (hypertrophie ventriculaire gauche).
- **Vaisseaux sanguins affaiblis et rétrécis dans les reins.** Cela peut empêcher ces organes de fonctionner normalement.
- **Vaisseaux sanguins épaissis, rétrécis ou déchirés dans les yeux.** Cela peut entraîner une perte de vision.
- **Syndrome métabolique.** Ce syndrome est un groupe de troubles du métabolisme de corps, notamment une augmentation du tour de taille, des triglycérides élevés, une diminution du cholestérol à lipoprotéines de haute densité (HDL).
- **Problème de mémoire ou de compréhension.** Une hypertension artérielle non contrôlée peut également affecter la capacité à penser, à se souvenir et à apprendre.

### 7.5. La maladie de l'hypertension artérielle dans le monde

La prévalence de l'hypertension varie selon les régions et les groupes de revenus des pays. La Région africaine de l'OMS a la prévalence d'hypertension la plus élevée (27 %) tandis que la Région OMS des Amériques a la prévalence d'hypertension la plus faible (18 %).

Le nombre d'adultes souffrant d'hypertension est passé de 594 millions en 1975 à 1,13 milliard en 2015, cette augmentation étant principalement observée dans les pays à revenu faible et intermédiaire. Cette augmentation est principalement due à une augmentation des facteurs de risque d'hypertension dans ces populations.[88]

---

[87] Mayo, p. d , op .cit.

[88] Who. (2021, May 17). Hypertension. Consulté le 06 08, 2021, sur world Heath organisation : https://www.who.int/news-room/fact-sheets/detail/hypertension#:~:text=An%20estimated%201.13%20billion%20people,cause%20of%20premature%20death%20worldwide.

**Figure 12 :** Prévalence de l'hypertension artérielle dans le monde, 25 ans et plus, standardisé selon l'âge, les deux sexes, 2008, source (organisation mondial de la santé, 2011)

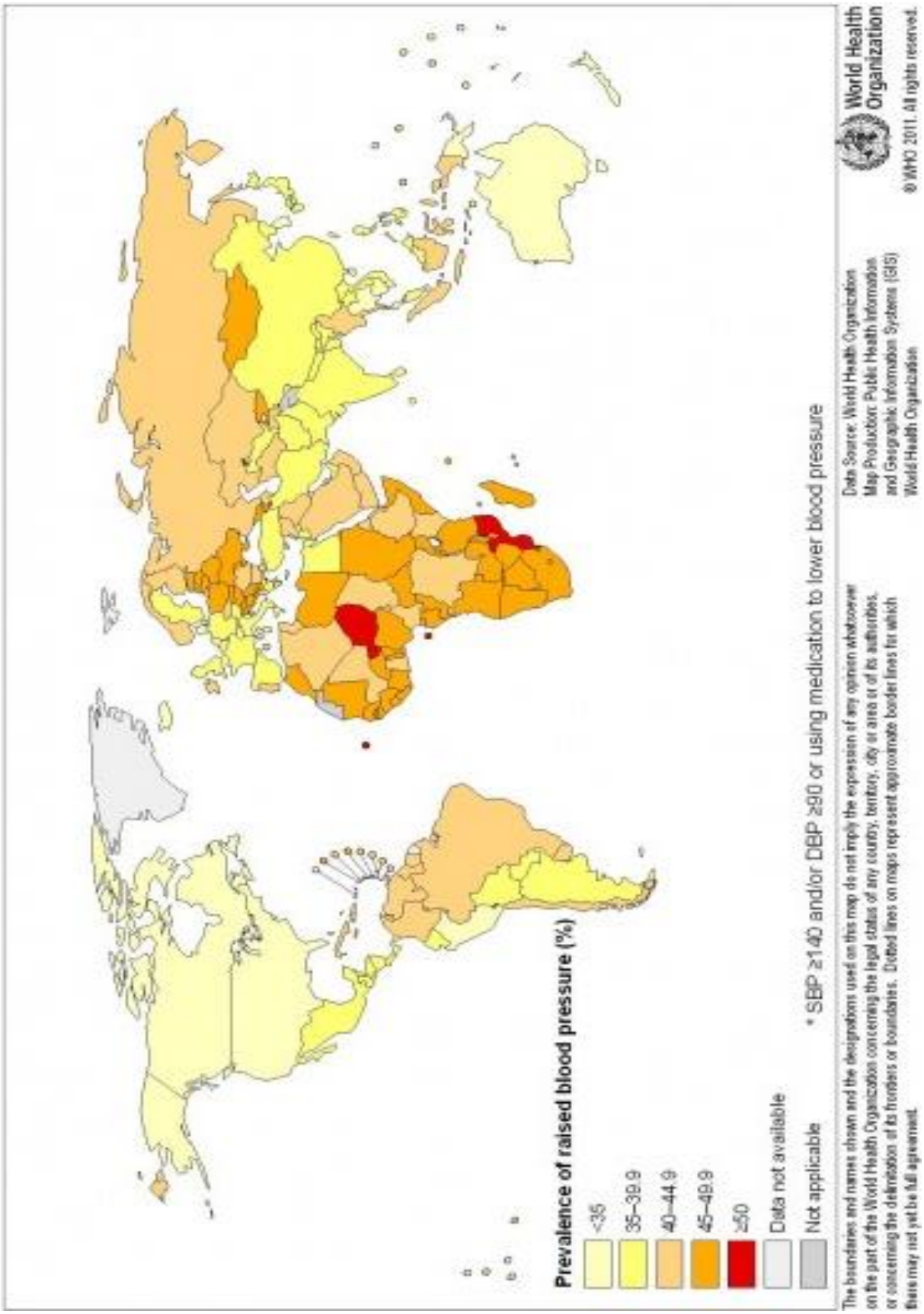
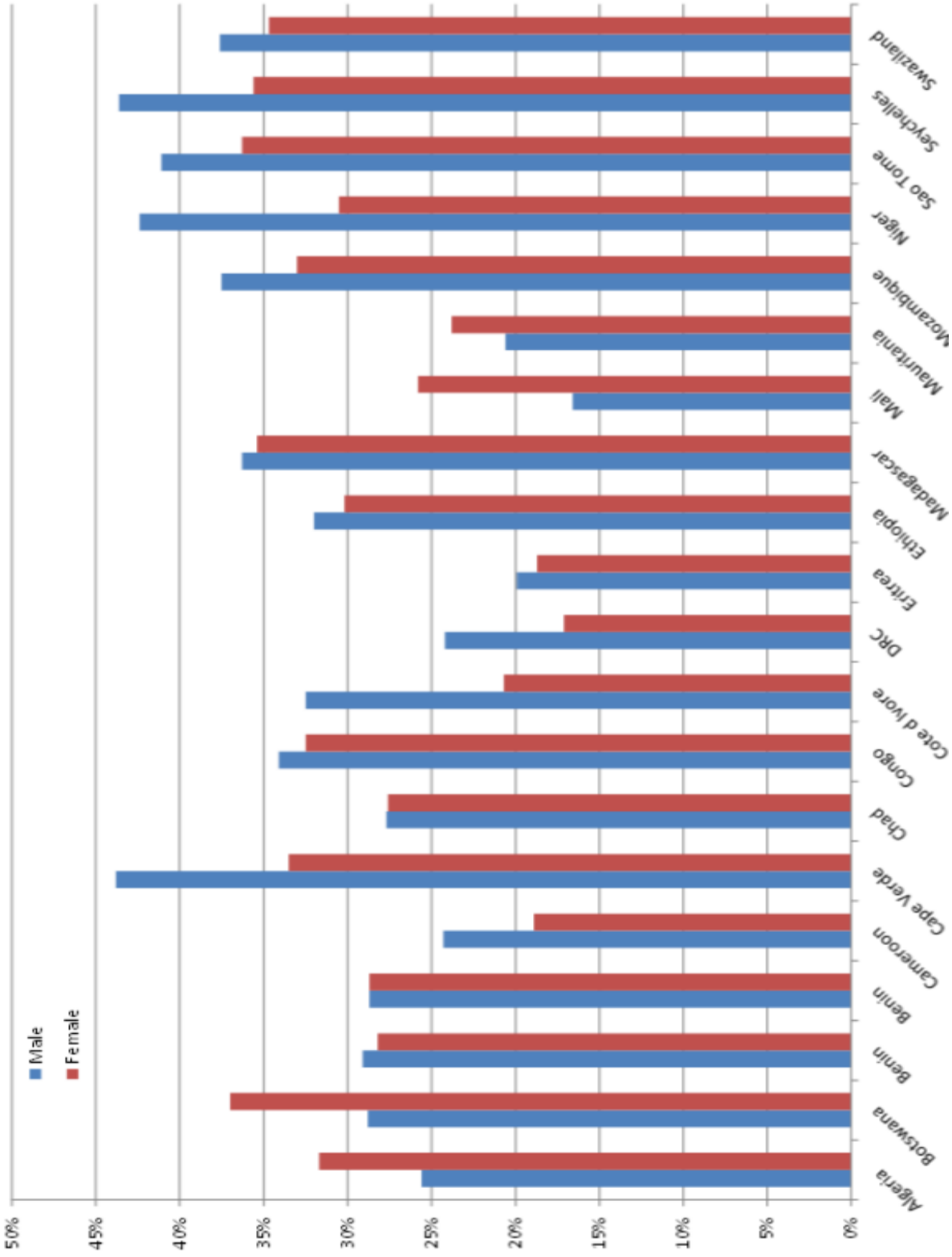


**Figure 13 :** Prévalence de l'hypertension dans certains pays africains ayant participé aux enquêtes OMS-STEPS (2003 à 2009), source (organisation mondial de la santé, 2011)



### 7.6. La maladie de l'hypertension artérielle en Algérie :

L'Algérie compte 23,6% de la population souffrant d'hypertension artérielle, qui est plus de 8 millions de personnes infectées, tandis que 71,9% d'entre eux ne reçoivent pas de traitement. Ainsi, cette maladie mortelle fait partie des maladies mortelles en Algérie, surtout à la lumière de la terrible et croissante propagation de celui-ci, et les complications de cette maladie qui se développe en une crise cardiaque et des artères cérébrales et constitue un « lourd fardeau » pour l'État.[89]

## 8. Conclusion

Dans ce chapitre, nous avons discuté comment le secteur de la santé a été développé par l'introduction de l'intelligence artificielle à de différents stades, et nous avons constaté que l'un des principaux usages de l'IA dans ce domaine est la prédiction des maladies. Nous avons expliqué le processus d'apprentissage automatique dans celui-ci. Nous avons, également, parlé du système d'aide à la décision médicale. Enfin, nous avons parlé de la maladie de l'hypertension, ses causes et sa propagation dans le monde et en Algérie en particulier.

[89] Rezzaki, D. (2019, 05 18). Artice de journal : « Plus de 8 millions de personnes souffrant d'hypertension artérielle en Algérie ». *eldjazaironline*. Consulté le 06 08, 2021, sur http://eldjazaironline.dz/

# Chapitre 3

# Prétraitement des données

## Introduction

Dans ce chapitre, nous analyserons les facteurs impliqués dans la prédiction de l'hypertension. Ensuite, nous étudions l’ensemble de données collectées à l'aide des techniques d'exploration de données. Ces dernières nous permettront de faire des calculs et des comparaisons et de faire de bonnes hypothèses pour identifier l’ensemble de données qui est le plus approprié pour notre projet.

## 1- Facteurs impliqués dans l'hypertension artérielle :

### 1.1. L'âge :

Le risque d'hypertension augmente avec l'âge. La tension normale d'un sujet est au plus de 14 pour le chiffre le plus haut et 9 pour le chiffre le plus bas. Elle a tendance à s'élever avec l'âge, mais il est maintenant admis qu'il ne faut pas la laisser dépasser certaines valeurs, même chez le sujet âgé.[90] Le tableau 3 illustre la relation entre la tension artérielle et l'âge.

**Tableau 3 :** Intervalle de tension artérielle par âge.[91]

| | Age | SBP | DBP |
|---|---|---|---|
| **Male** | | | |
| | 21-25 | 120.5 | 78.5 |
| | 26-30 | 119.5 | 76.5 |
| | 31-35 | 114.5 | 75.5 |
| | 36-40 | 120.5 | 75.5 |
| | 41-45 | 115.5 | 78.5 |
| | 46-50 | 119.5 | 80.5 |
| | 51-55 | 125.5 | 80.5 |
| | 56-60 | 129.5 | 79.5 |
| | 61-65 | 143.5 | 76.5 |
| **Female** | | | |
| | 21-25 | 115.5 | 70.5 |
| | 26-30 | 113.5 | 71.5 |
| | 31-35 | 110.5 | 72.5 |
| | 36-40 | 112.5 | 74.5 |
| | 41-45 | 116.5 | 73.5 |
| | 46-50 | 124 | 78.5 |
| | 51-55 | 122.55 | 74.5 |
| | 56-60 | 132.5 | 78.5 |
| | 61-65 | 130.5 | 77.5 |

### 1.2.Le genre (le sexe) :

On ne sait pas encore si l'hypertension est influencée par le sexe, mais des études récentes ont montré que les hommes sont plus susceptibles de développer de l'hypertension que les femmes.

---

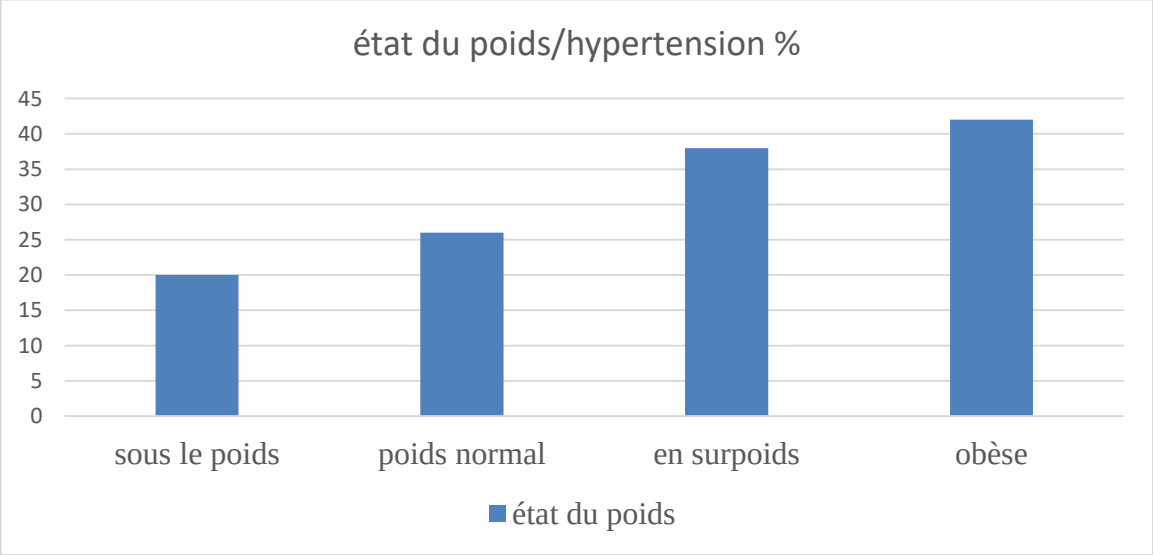
[90] Valle, A.-C. D. (2021, 02 03). Tension artérielle : normale, élevée, basse, mesure, âge. journal des femmes. Consulté le 06 08, 2021, sur https://sante.journaldesfemmes.fr/fiches-maladies/2488800-tension-arterielle-definition-normale-basse-haute-diastolique-systolique-age-mesure-tableau/

[91] Ibid .

### 1.3. Le poids :

De nombreuses études scientifiques ont révélé une corrélation directe entre l'augmentation de la pression artérielle et la prise de poids. En effet, la graisse viscérale peut contribuer à l'augmentation de la pression artérielle car elle est associée à une production accrue de cytokines inflammatoires (telles que l'interleukine-1-β, le facteur de nécrose tumorale-α et l'interleukine-6) et à des facteurs inflammatoires (tels que C -protéine réactive), induisant un dysfonctionnement endothélial et par conséquent une hypertension artérielle (AH). [92]La figure 15 montre la relation de corrélation entre le poids et la tension artérielle.

**Figure 14 :** Relation entre poids et la tension artérielle. (Organisation mondial de la santé,2020)




### 1.4. Facteur génétique :

L'hypertension a tendance à fonctionner dans les familles. Les personnes dont les parents souffrent d'hypertension ont un risque élevé de développer la maladie, en particulier si les deux parents sont touchés. Cependant, le modèle d'héritage est inconnu. Les formes génétiques rares d'hypertension suivent le modèle d'hérédité de la maladie individuelle.[93]

---

[92] Daniele, A. N. (2019). *The "Weight" of Obesity on Arterial Hypertension* . Open access peer-reviewed chapter. doi:10.5772/intechopen.87774.p 1 .

[93] Hunt, P. N. (2003). Genetics of hypertension . *Genetics in Medicine, 5*, 413–429. Récupéré sur https://www.nature.com/articles/gim2003368.

### 1.5. Niveau d'hémoglobine :

Un mécanisme évident d'augmentation de la pression artérielle avec une augmentation des taux d'Hb serait une augmentation de la viscosité sanguine. Il a été rapporté qu'une élévation de l'hématocrite et des taux d'Hb augmente la viscosité du sang et qu'une viscosité accrue en partie par un effet sur la pression artérielle peut aggraver la fonction cardiovasculaire.

### 1.6. Le tabagisme :

Le tabagisme affectant la rigidité artérielle et la réflexion des ondes, pourrait avoir un effet néfaste plus important sur la pression artérielle centrale, qui est plus étroitement liée aux lésions des organes cibles que la pression artérielle brachiale. [94]

### 1.7. Activité physique :

Les exercices physiques ont été associés à des réductions significatives immédiates de la pression artérielle systolique. Cette réduction immédiate de la pression artérielle après l'effort peut persister pendant près de 24 heures et est appelée hypotension post-exercice avec les effets les plus prononcés observés chez les personnes ayant une tension artérielle initiale plus élevée.[95]

### 1.8. Apport de Sodium

La relation entre l'hypertension et l'apport alimentaire en sodium est largement reconnue et étayée par plusieurs études. Une réduction du sodium alimentaire diminue non seulement la tension artérielle et l'incidence de l'hypertension, mais est également associée à une réduction de la morbidité et de la mortalité dues aux maladies cardiovasculaires. [96]

### 1.9. Stress

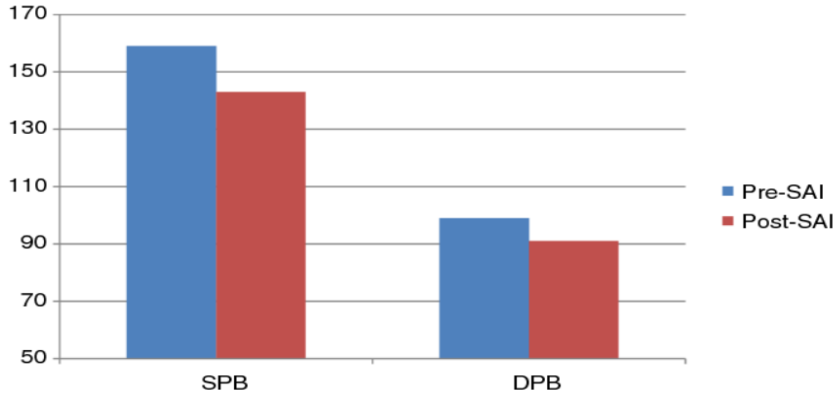
Le stress peut provoquer une hypertension par des élévations répétées de la pression artérielle ainsi que par la stimulation du système nerveux pour produire de grandes quantités d'hormones vasoconstrictrices qui augmentent la pression artérielle. Lorsqu'un facteur de risque est associé à d'autres facteurs de stress, l'effet sur la pression artérielle est multiplié. Le

---

[94] A Virdis, C. G. (s.d.). Cigarette smoking and hypertension PMID: 20550499. *National Library of Medecine*. doi:10.2174/138161210792062920.
[95] Solomon, S. M. (2015, 08 16). Influence of Physical Activity on Hypertension and Cardiac Structure and Function. *Current Hypertension Reports*. doi:10.1007/s11906-015-0588-3
[96] Andrea Grillo, L. S. (2019). Sodium Intake and Hypertension. Nutrients. doi:10.3390/nu11091970

graphique ci-dessous (Voir Figure 16) montre la SBP et la DBP moyenne avant et après SAI (Stress Alleviating Intervention). [97]

**Figure 15 :** Illustration de relation entre stresse/tension artérielle .[98]




### 1.10. Consommation d'alcool

Des études épidémiologiques, précliniques et cliniques ont établi l'association entre une consommation élevée d'alcool et l'hypertension. Cependant, le mécanisme par lequel l'alcool augmente la pression artérielle reste insaisissable. Plusieurs mécanismes possibles ont été proposés tels qu'un déséquilibre du système nerveux central, une altération des barorécepteurs, une augmentation de l'activité sympathique. [99]

### 1.11. Maladie chronique :

Dans environ 1 cas sur 20, l'hypertension artérielle survient à la suite d'un problème de santé sous-jacent ou de la prise d'un certain médicament.

Les problèmes de santé qui peuvent causer une hypertension artérielle comprennent :[100]

- Maladie du rein
- Diabète

---

[97] S Kulkarni, I. O. (1998). Stress and hypertension. Medical College of Wisconsin. Récupéré sur https://pubmed.ncbi.nlm.nih.gov/9894438/
[98] Ibid.
[99] Kazim Husain, R. A. (2014, 05 26). Alcohol-induced hypertension: Mechanism and prevention . World J Cardiol, 245–252. doi:10.4330/wjc.v6.i5.245
[100] NHS. (2019). High blood pressure (hypertension). Consulté le 06 09, 2021, sur Le Service national de santé britanique (NHS): https://www.nhs.uk/conditions/high-blood-pressure-hypertension/causes/

- Hépatite
- Infections rénales à long terme
- Glomérulonéphrite - dommages aux minuscules filtres à l'intérieur des reins
- Rétrécissement des artères alimentant les reins
- Problèmes hormonaux - tels qu'une thyroïde sous-active, une thyroïde hyperactive, le syndrome de Cushing, l'acromégalie, des niveaux accrus de l'hormone aldostérone.

### 1.12. Facteur socioéconomique :

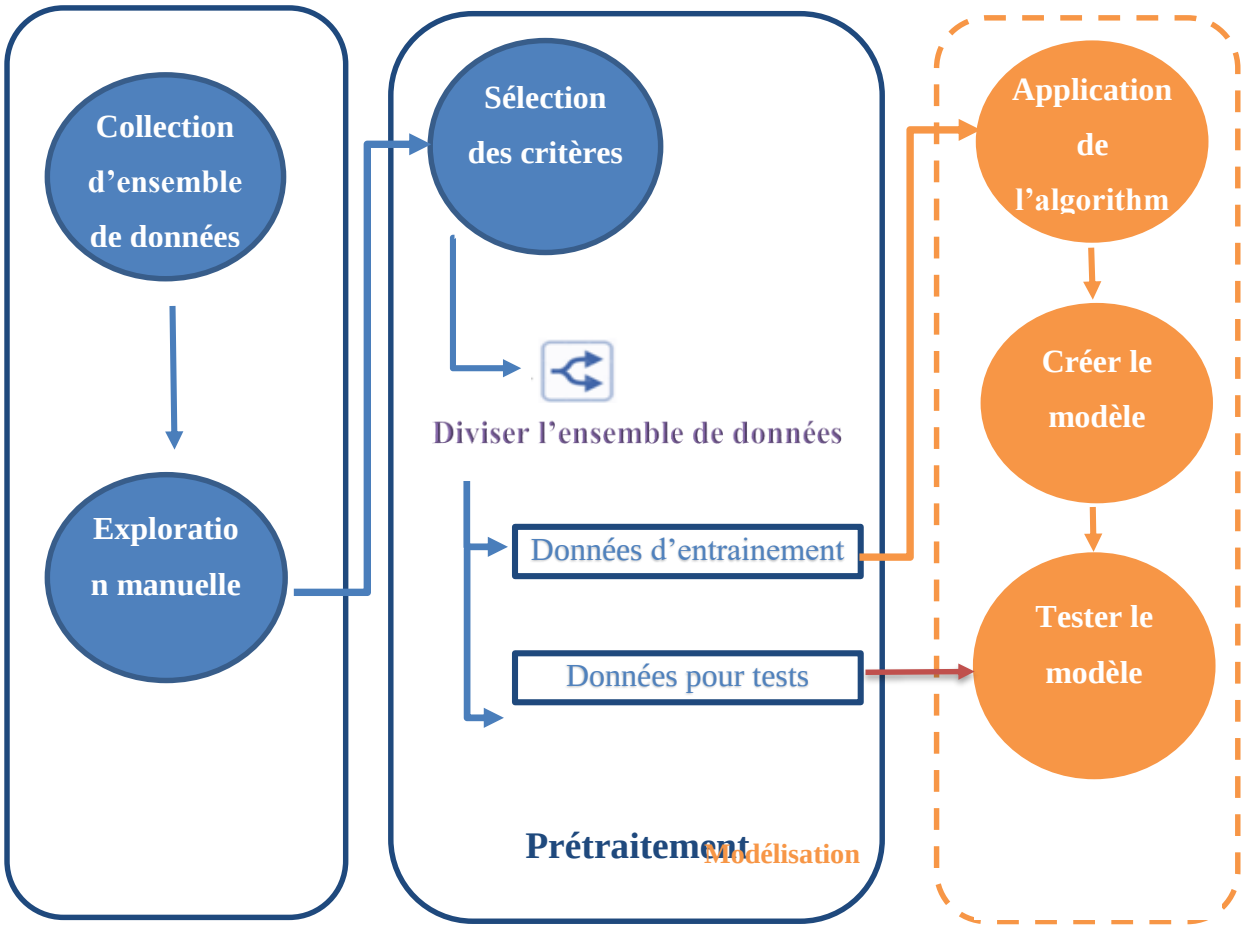
Parmi ces facteurs, on parle essentiellement du revenu, de la profession et de l'éducation. Un faible SSE (Status socio-économique) est associé à une pression artérielle plus élevée, et cette association est particulièrement évidente dans le niveau d'éducation. Il est important d'identifier et de surveiller l'hypertension pour réduire le risque de cette maladie parmi les groupes les plus vulnérables dans différents pays et entre différentes sociétés. [101]

## 2- Etapes de notre étude :

Notre étude est basée sur deux grandes parties : la phase de prétraitement, où nous avons choisi les attributs les plus pertinents, et la seconde applique des algorithmes de Machine Learning afin de sélectionner l'algorithme qui donne une meilleure précision. Notre proposition est divisée en plusieurs phases, l'approche est expliquée en détail dans la Figure suivante :

[101] Bing Leng, Y. J. (2015, 02). Socioeconomic status and hypertension : a meta-analysis. Journal of Hypertension, v 33,pp 221-229. doi:10.1097/HJH.0000000000000428

**Figure 16 : Procédure Analyse de données / prétraitement .**




### 3- Collection de données

Pour réaliser ce projet, malheureusement, nous n'avons pas pu mettre la main sur des ensembles de données produits par les hôpitaux et les établissements de santé algériens car la plupart des hôpitaux ne disposent pas d'ensembles de données structurés spécialement conçus pour la prédiction de l'hypertension, nous avons donc dû nous appuyer sur les ressources déployées sur Internet.

Nous avons trouvé trois (3) ensembles de données sur les sites web **kaggle,** d'autres ensembles de données trouvés dans d'autres sites sont mal structurés et nécessitent un travail énorme pour les reconstituer (nous ne pouvions pas les utiliser). Nous avons éliminé un parmi les trois ensembles des données car il manque des variables et n'avait que quatre (4) variables : ***âge***, ***sexe***, ***poids*** et ***éducation***. Par conséquent, nous avons décidé de travailler sur les deux jeux de données et déterminer lequel est le plus adapté à notre application

Nous avons utilisé deux (2) ensembles de donnée différents, pour mieux comprendre l'impact de différents facteurs (prédicteurs) dans la prédiction de tension artérielle chacun de ces ensembles de donnée possède sa propre taille et ses propres ensembles de prédicteurs. Tous les ensembles de donnée ont été téléchargés à partir du site web reconnu internationalement **kaggle** qui contient des milliers des base d'information utilisé dans des laboratoires de recherche autour du monde.

### 3.1. Ensembles de donnée 1 :

#### 3.1.1. Exploration manuelle

L'exploration des données ou l'exploration manuelle est l'étape initiale de l'analyse des données, où les utilisateurs explorent un grand ensemble de données de manière non structurée pour découvrir les modèles initiaux, les caractéristiques et les points d'intérêt. Ce processus n'est pas destiné à révéler toutes les informations contenues dans un ensemble de données, mais plutôt à aider à créer une vue d'ensemble des tendances importantes et des principaux points à étudier plus en détail. Dans notre étude, l'ensemble des données que nous allons explorer contient 21613 enregistrements, initialement 22 colonnes et on a supprimé les colonnes 13 à 19 puisqu'elles contiennent que des zéros (0). La première colonne HY_YN est la valeur cible qui signifie : ***HYpertension Yes/No***. Elle est de type binaire.

Le tableau suivant présente un aperçu de l'ensemble de données obtenu à partir du fichier ***MGE303_Term Project_Classification_Data (Disease).csv*** [102]Ici, le nombre de colonnes est de 15 et le nombre d'échantillons est de 21613. Certaines colonnes sont catégoriques et, par conséquent, le nombre de variables est supérieur au nombre de colonnes, à savoir 19.

[102] Télécharger depuis Kaggle, lien : https://www.kaggle.com/kmukhammadjon/hypertension-incidence-in-two-and-half-years

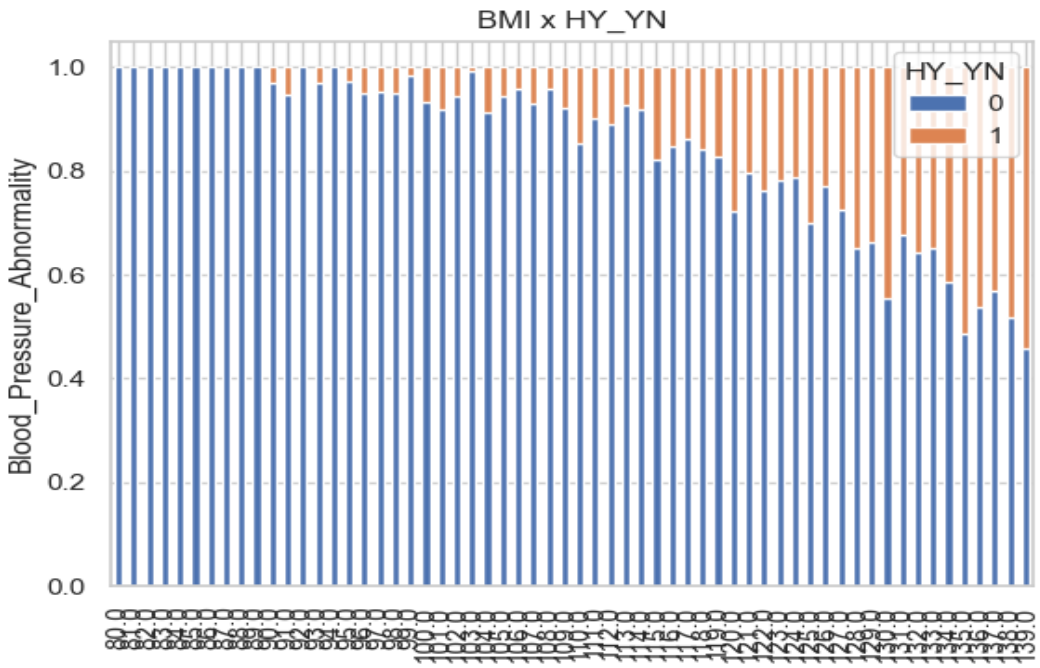
**Tableau 4** : aperçu sur l'ensemble de données numéro 1

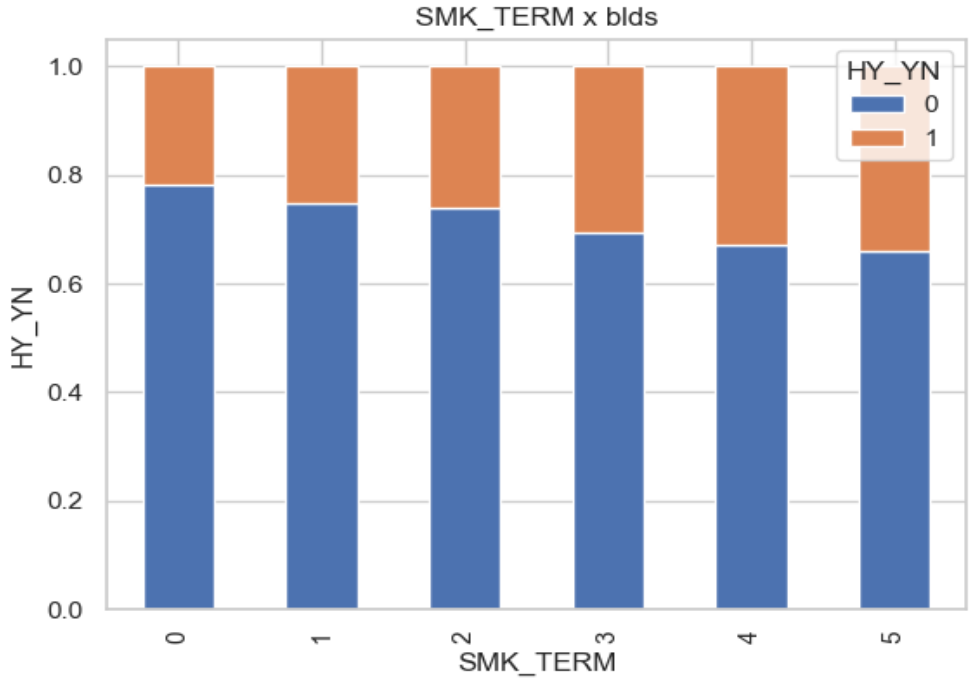
| HY_YN | AGE | gender | WAIST | BP_HIGH | BP_LWST | BLDS | TOT_CHOLE | SGOT_ASL | SGPT_ALT | GAMMA_GTP | HMG | DSQTY | SMK_TERM | EXER_M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 66 | 1 | 76.0 | 120.0 | 70.0 | 87.0 | 2.0 | 22.0 | 13.0 | 1.0 | 15.0 | 0 | 5 | Once a week |
| 1 | 55 | 1 | 84.0 | 110.0 | 80.0 | 85.0 | 2.0 | 40.0 | 35.0 | 3.0 | 14.9 | 0 | 0 | Never |
| 1 | 54 | 0 | 84.0 | 117.0 | 76.0 | 92.0 | 2.0 | 20.0 | 20.0 | 1.0 | 13.3 | 0 | 0 | Once a week |
| 1 | 29 | 1 | 80.0 | 126.0 | 82.0 | 86.0 | 2.0 | 25.0 | 27.0 | 2.0 | 17.0 | 1 | 3 | Once a week |
| 1 | 43 | 1 | 82.0 | 130.0 | 85.0 | 104.0 | 1.0 | 31.0 | 23.0 | 2.0 | 14.1 | 2 | 3 | Once a week |
| 1 | 66 | 1 | 78.0 | 130.0 | 76.0 | 123.0 | 2.0 | 26.0 | 13.0 | 3.0 | 14.1 | 1 | 5 | Once a week |
| 1 | 53 | 1 | 77.0 | 115.0 | 70.0 | 101.0 | 3.0 | 23.0 | 18.0 | 2.0 | 13.1 | 3 | 1 | 2-3 a week |
| 1 | 47 | 1 | 81.0 | 120.0 | 80.0 | 97.0 | 2.0 | 43.0 | 37.0 | 4.0 | 16.6 | 2 | 1 | Once a week |
| 1 | 50 | 0 | 86.0 | 130.0 | 80.0 | 101.0 | 3.0 | 22.0 | 25.0 | 1.0 | 11.6 | 0 | 1 | Never |
| 1 | 56 | 1 | 105.0 | 138.0 | 85.0 | 79.0 | 2.0 | 19.0 | 22.0 | 1.0 | 14.3 | 0 | 1 | Once a week |
| 1 | 51 | 1 | 85.0 | 130.0 | 80.0 | 121.0 | 2.0 | 16.0 | 10.0 | 1.0 | 13.1 | 2 | 1 | >5 a week |
| 1 | 50 | 0 | 69.0 | 110.0 | 76.0 | 89.0 | 3.0 | 16.0 | 10.0 | 1.0 | 13.8 | 0 | 1 | Once a week |
| 1 | 32 | 1 | 75.0 | 135.0 | 78.0 | 111.0 | 2.0 | 25.0 | 28.0 | 3.0 | 14.6 | 1 | 1 | Never |
| 1 | 61 | 1 | 88.0 | 130.0 | 80.0 | 93.0 | 2.0 | 29.0 | 34.0 | 2.0 | 15.5 | 0 | 0 | 4-5 a week |
| 1 | 48 | 1 | 86.0 | 139.0 | 89.0 | 92.0 | 2.0 | 22.0 | 28.0 | 2.0 | 14.4 | 1 | 2 | Never |
| 1 | 64 | 1 | 85.0 | 128.0 | 80.0 | 81.0 | 2.0 | 15.0 | 12.0 | 1.0 | 14.3 | 0 | 0 | Never |
| 1 | 46 | 1 | 79.0 | 130.0 | 80.0 | 83.0 | 2.0 | 53.0 | 55.0 | 4.0 | 16.3 | 2 | 3 | Once a week |
| 1 | 48 | 0 | 96.0 | 130.0 | 70.0 | 94.0 | 2.0 | 19.0 | 11.0 | 1.0 | 11.3 | 0 | 0 | Never |
| 1 | 58 | 1 | 92.0 | 120.0 | 80.0 | 97.0 | 2.0 | 29.0 | 23.0 | 1.0 | 14.9 | 0 | 0 | Never |
| 1 | 53 | 1 | 91.0 | 130.0 | 80.0 | 97.0 | 2.0 | 22.0 | 21.0 | 1.0 | 15.9 | 1 | 0 | Never |
| 1 | 48 | 1 | 86.0 | 139.0 | 89.0 | 92.0 | 2.0 | 22.0 | 28.0 | 2.0 | 14.4 | 1 | 2 | Never |

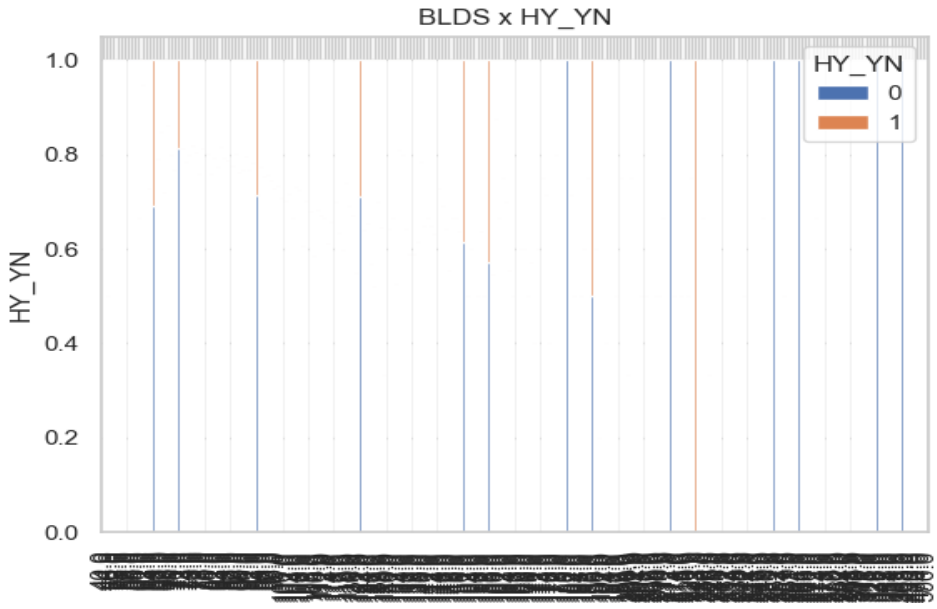
### 3.1.2. Sélection des caractéristiques :

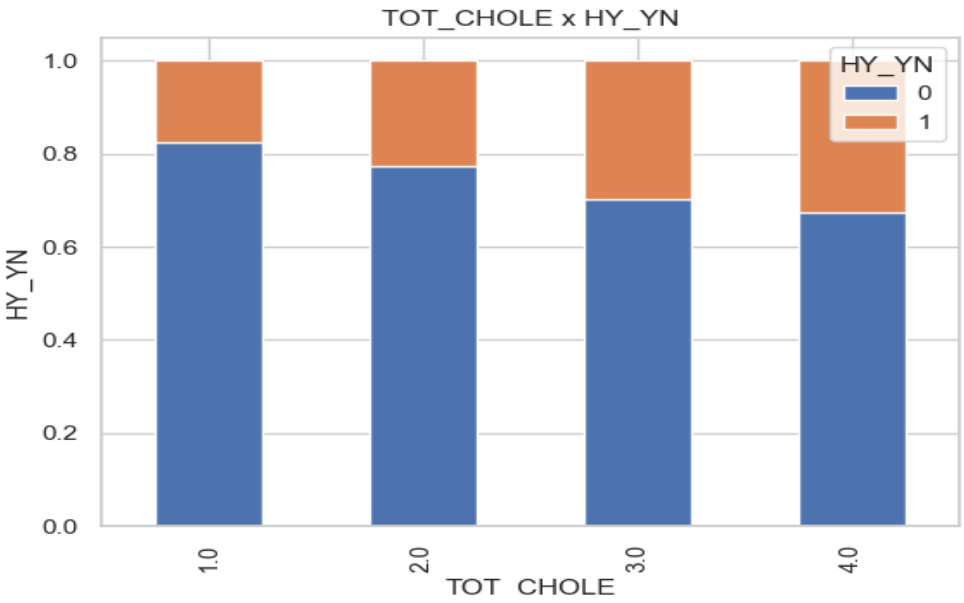
Cette étape est basée sur la matrice de corrélation et les diagrammes de fréquence entre les variable prédicteur et la variable cible On s'est concentré sur la relation de chaque variable prédicteur avec la variable cible :

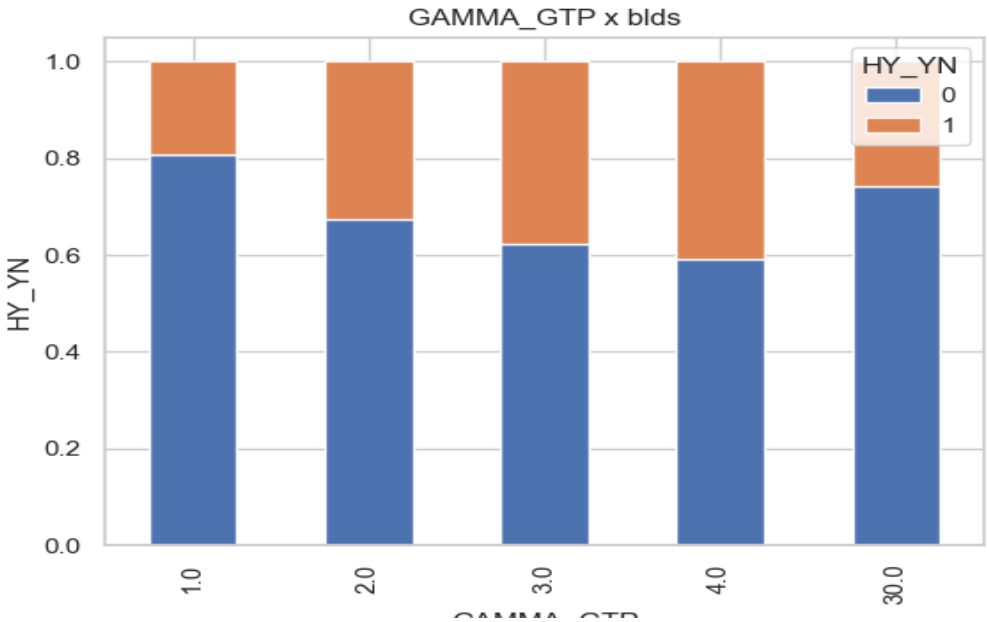
Grace à l'utilisation de la bibliothèque ***python Panda*** et la bibliothèque ***matplotlib. Pyplot,*** nous avons pu générer les tracés de la fréquence entre chaque variable et la variable cible :

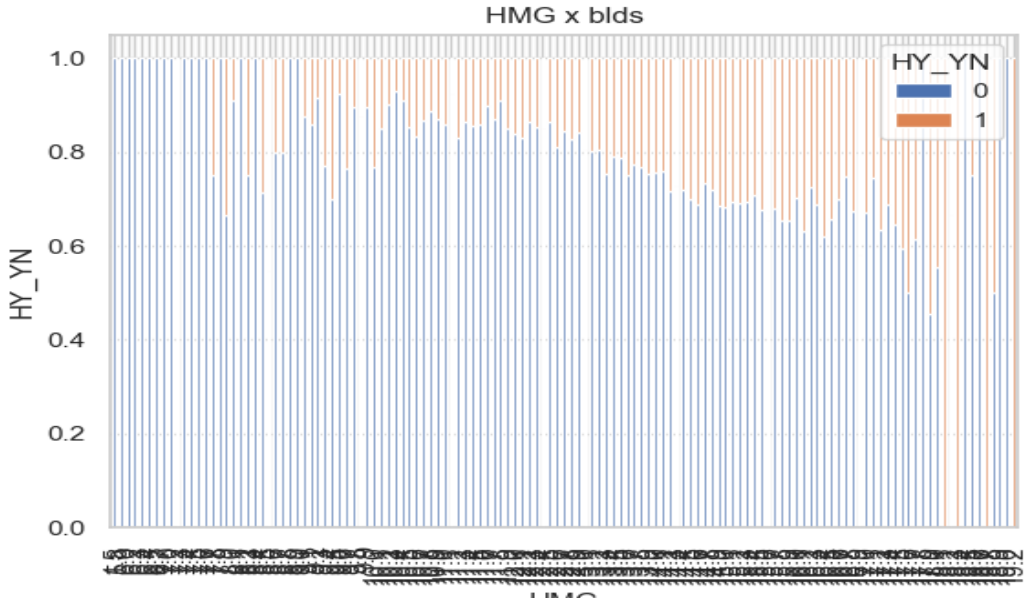


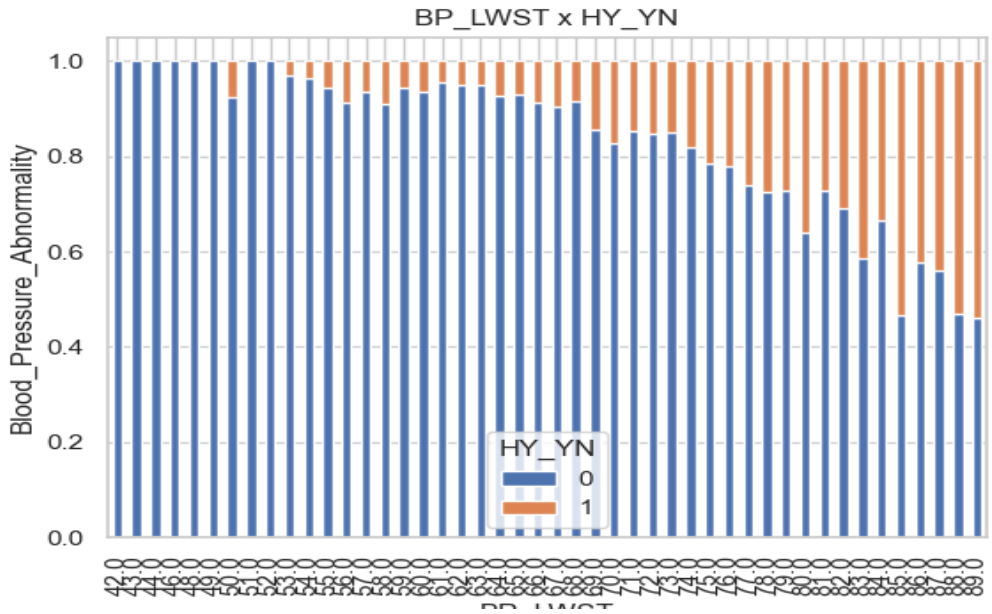



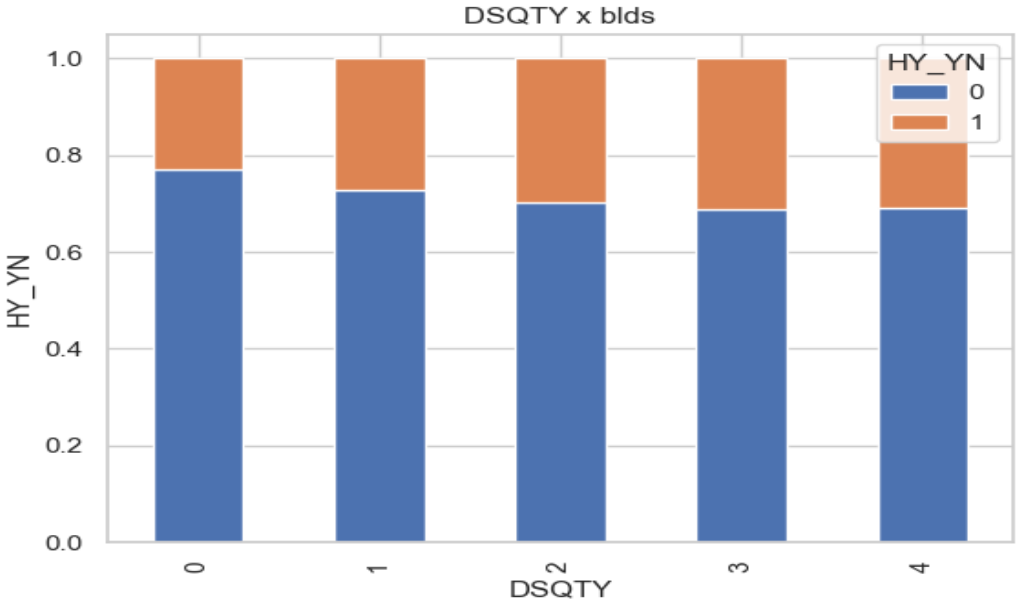

BLDS x HY_YN
HY_YN
1.0
0.8
0.6
0.4
0.2
0.0
HY_YN
0
1

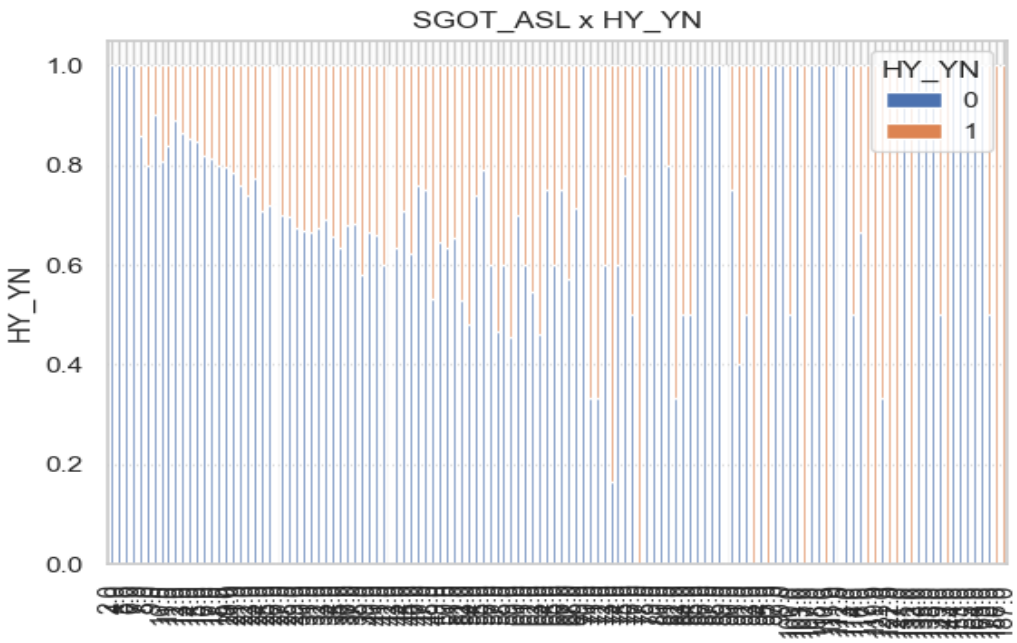

TOT_CHOLE x HY_YN
HY_YN
1.0
0.8
0.6
0.4
0.2
0.0
HY_YN
0
1
1.0
2.0
3.0
4.0
TOT_CHOLE

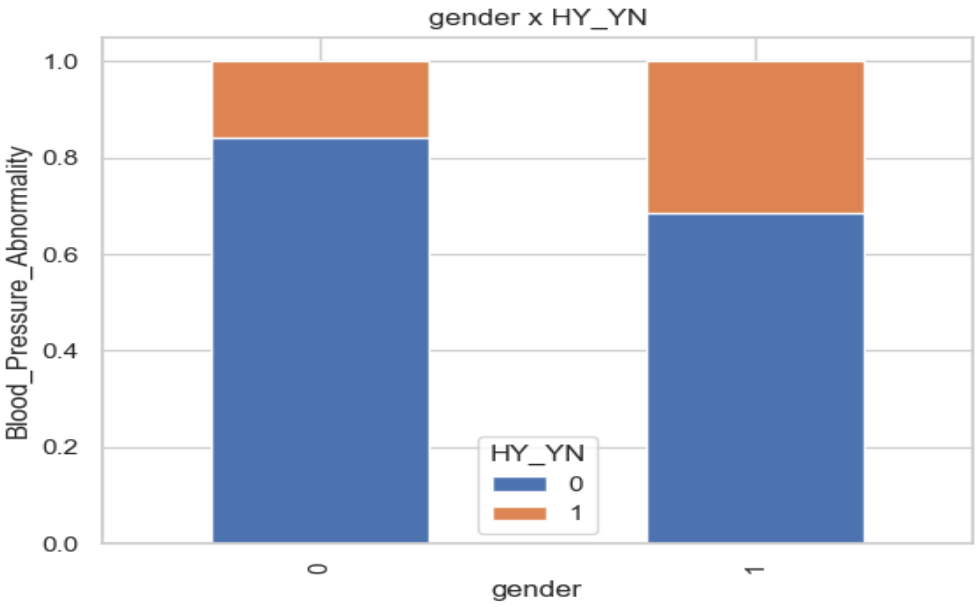

GAMMA_GTP x blds
HY_YN
1.0
0.8
0.6
0.4
0.2
0.0
HY_YN
0
1
1.0
2.0
3.0
4.0
30.0

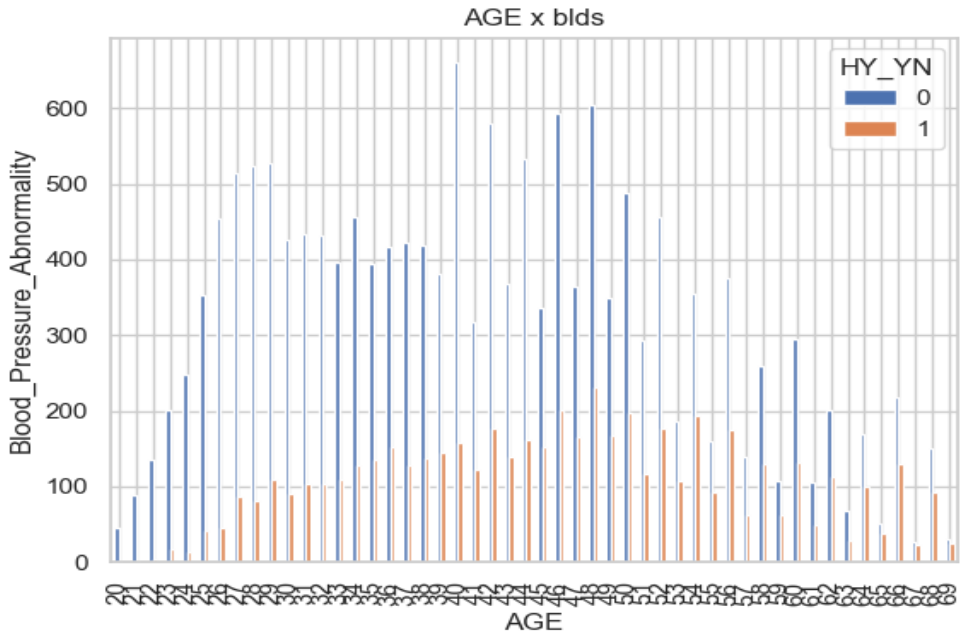

HMG x blds
HY_YN
1.0
0.8
0.6
0.4
0.2
0.0
HY_YN
0
1

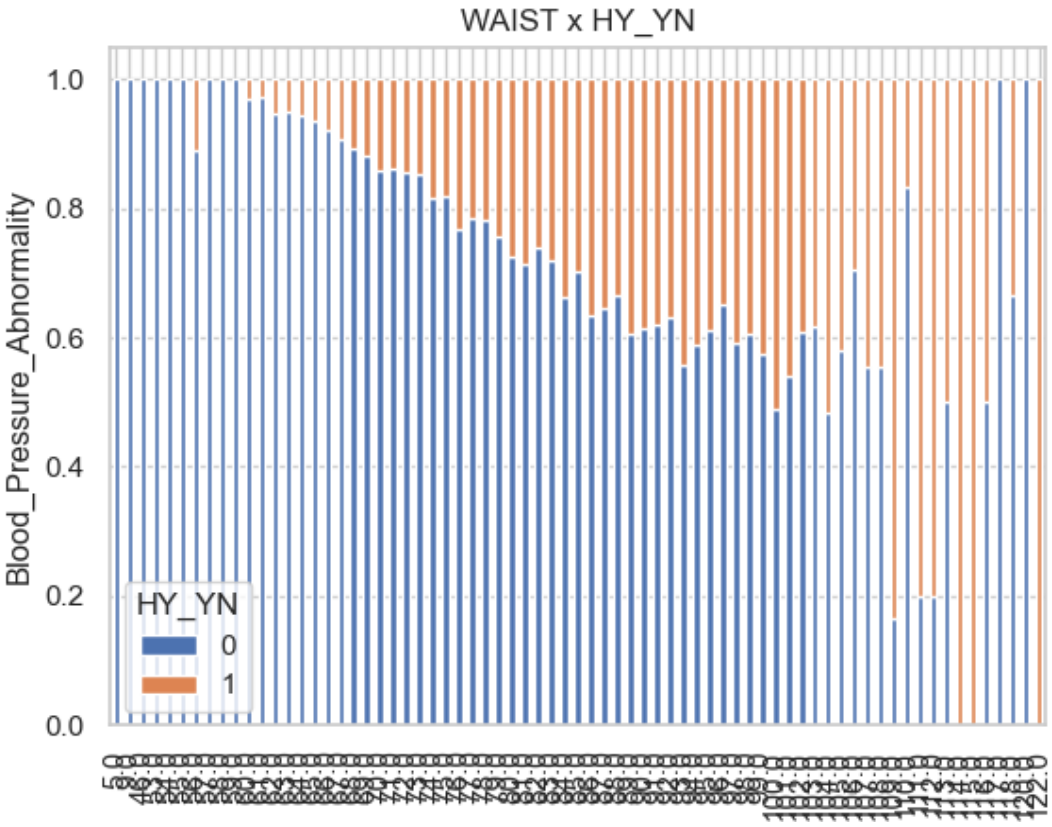

BP_LWST x HY_YN
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
HY_YN
0
1


DSQTY x blds
HY_YN
1.0
0.8
0.6
0.4
0.2
0.0
HY_YN
0
1
0
1
2
3
4
DSQTY

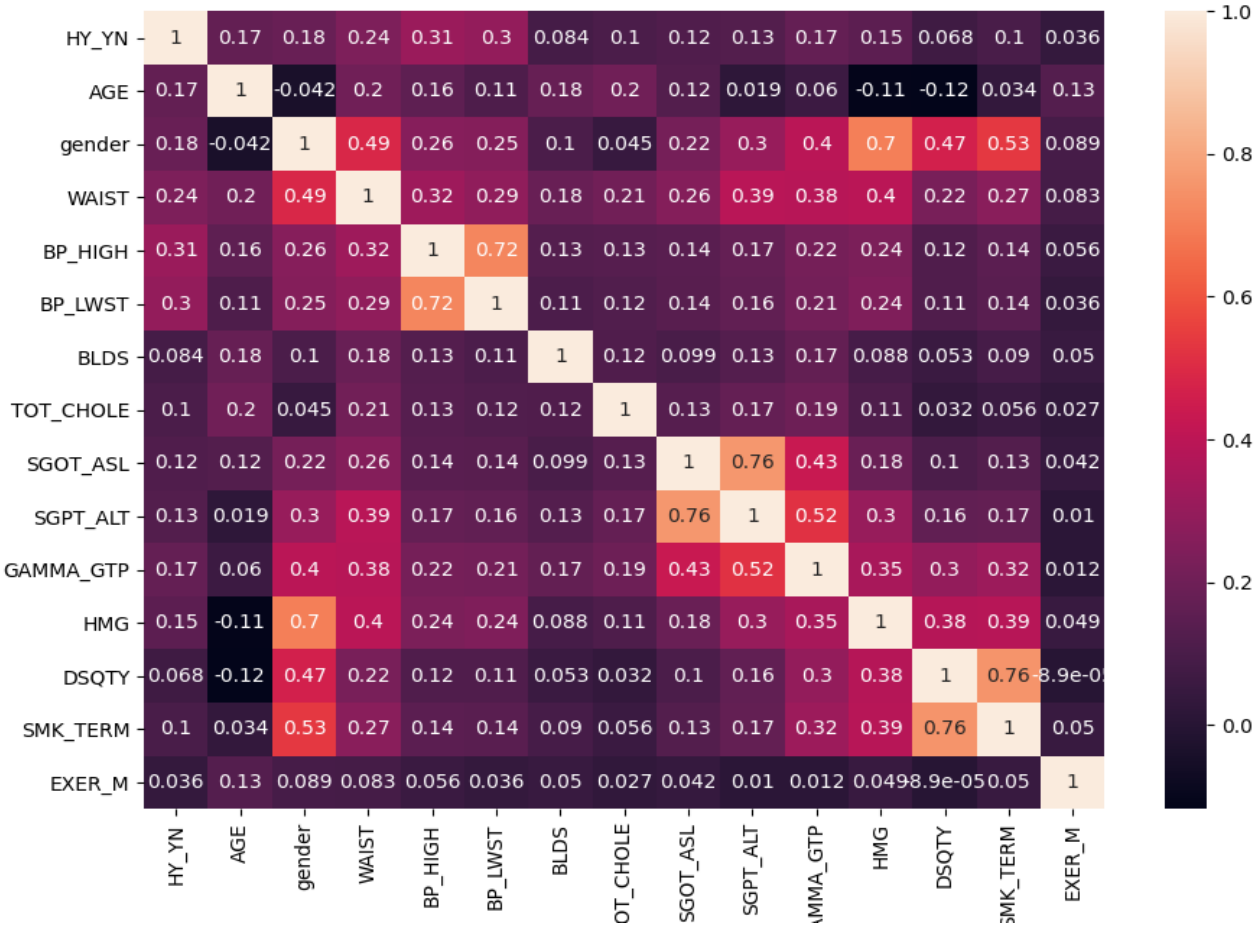

SGOT_ASL x HY_YN
HY_YN
1.0
0.8
0.6
0.4
0.2
0.0
HY_YN
0
1


gender x HY_YN
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
HY_YN
0
1
0
1
gender


AGE x blds
Blood_Pressure_Abnormality
600
500
400
300
200
100
0
HY_YN
0
1
AGE




### 3.1.3. Matrice de corrélation



À partir des diagrammes de fréquence et de la matrice de corrélation, nous remarquons que trois variables sont très mal corrélées avec les variables cibles (BLDS, DSQTY, Exer_M)

Par conséquent, la variable cible ne dépend pas de ces variables, elles sont donc éliminées de l'ensemble de données

Le tableau suivant résume chaque variable et ses caractéristiques : (listes de prédicteurs), les colonnes en rouges sont éliminées.

| Variable | Type | Min/max | Valeur normale | Remarque |
|---|---|---|---|---|
| AGE | numérique | 20 /69 | / | / |
| gender | Binaire | 0/1 | / | |
| WAIST | numérique | 70-122 | < 90 cm | tour de taille |
| BP_HIGH | numérique | 80-139 | Voir tableau 1 | la pression artérielle systolique |
| BP_LWST | numérique | 42-89 | Voir tableau 1 | La pression artérielle diastolique |
| BLDS | numérique | 40-150 | Voir tableau 1 | pression artérielle moyenne |
| TOT_CHOLE | numérique | 1-4 | < 200 mg/dL | On a divisé cette variable en 4 catégories :<br>1 : < 150<br>2 : entre 150 et 200<br>3 : entre 200 et 240<br>4 : > 240 |
| SGOT_ASL | numérique | 2-196 | Entre 8 et 45 u/ls unités par litre de sérum | Le test SGOT est un test sanguin. Il aide à déterminer le bon fonctionnement du foie, ce test est associé à la consommation d'alcool et à d'autres troubles hépatiques. |
| SGPT_ALT | numérique | 2-198 | Entre 7 et 56 u/ls unités par litre de sérum | Un taux très élevé de SGPT dans le sang peut être une indication de dommages ou de problèmes liés au foie. Certaines maladies comme la cirrhose et l'hépatite. |
| GAMMA_GTP | numérique | 1-30 | < 45 UI/L | La gamma-GT est une enzyme qui existe au niveau de la membrane cellulaire de nombreux organes comme les reins ou le pancréas. |
| HMG | numérique | 1.5-19.2 | 13.5 et 17.5 g/dL<br>Pour les hommes<br>12 et 15 g/dL<br>Pour les femmes | niveau d'hémoglobine |

| | | | | |
|---|---|---|---|---|
| DSQTY | numérique | 0-4 | 135 et 145 milliéquivalents par litre (mEq/L) | On a divisé cette variable en 4 catégories :<br>1 : < 135<br>2 : entre 135 et 145<br>3 : entre 145 et 150<br>4 : > 150 |
| SMK_TERM | numérique | 0-5 | / | niveau de tabagisme |
| EXER_M | Catégorial | 1 / >5 | | activité physique par semaine |

### 3.1.4. Fractionnement de l'ensemble de données

Nous avons divisé notre ensemble de données en deux parties : entraînement et test avec respectivement 80% et 20% de la population.

### 3.1.5. Application de l'algorithme d'apprentissage :

Dans cette phase, on a implémenté quatre (4) algorithmes d'apprentissage les plus connus pour les problèmes de classification, à savoir : l'arbre de décision, le réseau neurone, les forêts aléatoires et la SVM

#### A. Arbre de décision (arbre de classification) :

Dans ce type d'algorithme, on a utilisé les bibliothèques Suivantes :

- **Panda** : Pour la manipulation et l'analyse de données.
- **Scikit-Lean** : elle comporte divers algorithmes de classification, de régression et de clustering.

- On a divisé les données en deux parties : une partie pour l'entrainement et une partie pour les tests :

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

- Pour la création de l'arbre de classification : on a utilisé la fonction d'entropie puisqu'elle donne des meilleurs résultats relativement à la fonction Gini (fonction par défaut).

```
clf = DecisionTreeClassifier(criterion="entropy" ,splitter="random")
```

- Phase de l'entrainement :

```
clf = clf.fit(X_train,y_train)
```

- Phase de prédiction :

```
y_pred = clf.predict(X_test)
```

- Évaluation de la précision :

```
print ("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

- Résultat :

```
Le résultat de précision été : 74.49 %
```

- La matrice de confusion :

```
print(confusion_matrix(y_test, y_pred))
```

- Sortie :

```
[[1591    0]

[ 538    0]]
```

**B. Réseau neurone :**

Les bibliothèques utilisées sont :

- **Keras :** bibliothèque qui fournit une interface Python pour les réseaux de neurones artificiels.
- **Tensorflow :** est une bibliothèque logicielle gratuite et open source pour l'apprentissage automatique.
- **Pandas :** pour la manipulation et l'analyse de données.

**- Structure de réseaux neurone :**

On a implémenté une structure simple de deux couches :

1 – Première couche : fonction d'activation est : ***relu***

```
x = layers.Dense(32, activation="relu")(all_features)

x = layers.Dropout(0.5)(x)
```

2- Deuxième couche : fonction d'activation est : ***sigmoïde***

```
output = layers.Dense(1, activation="sigmoid")(x)

model = keras.Model(all_inputs, output)
```

- Taux d'apprentissage Learning rate : on a fixé sa valeur à 0.01

```
opt = keras.optimizers.Adam(learning_rate=0.01)
```

- Compilation de model :

```
model.compile("adam", "binary_crossentropy", metrics=["accuracy"])
```

- Nombre d'arrêts : on a implémenté avec les valeurs de 10 à 500 pour atteindre la valeur optimale :

```
keras.utils.plot_model(model, show_shapes=True, rankdir="LR")
```

- Phase de l'entrainement :

```
model.fit(train_ds, epochs=1, validation_data=val_ds)
```

**Remarque *:*** Le choix du nombre d'arrêts et de valeur du taux d'apprentissage dépend de la puissance de la machine.

- Entrainer le model :

```
model.fit(train_ds, epochs=1, validation_data=val_ds)
```

- Résultat :

```
Précision : 0.7510
```

### C. Foret aléatoire

Les bibliothèques utilisées :

**Panda** : pour la manipulation et l'analyse de données

**Scikit-Lean** : Il comporte divers algorithmes de classification, de régression et de clustering.

On a suivi la même procédure que l'arbre de décision mais dans le cas des forêts aléatoires, nous utilisons **RandomForestClassifier** et en divisant le model en deux parties :

```
X_train, X_test, y_train, y_test =train_test_split(X, y, test_size =0.20)
```

- Nous obtenons la :

Précision : 0.98

**C. SVM (Support vector machine)**

On utilise la même bibliothèque **Sklearn** et on suit une procédure similaire que les autres algorithmes mais cette fois utilisant algorithme SVM :

```
from sklearn import svm
#Create a svm Classifier
clf = svm.SVC(kernel='linear') # Linear Kernel
```

- calcule de precision :

Le résultat été : 75,26 %

### 3.2. Ensemble de donnée 2 :

#### 3.2.1. Exploration manuelle

Initialement, cet ensemble de données contenait 2001 enregistrements et 15 colonnes, mais on a supprimé la colonne qui contient la variable (***Pregnancy***) qui signifie grossesse, car cette variable va créer beaucoup de confusion pour le résultat final parce que nous ne savons pas si la grossesse est un bon prédicteur de l'hypertension artérielle à long terme. Aussi, on a supprimé la colonne ***Patient_Number*** qui signifie numéro de malade.

On a aussi éliminé quelques autres enregistrements qui contiennent des informations erronées ou male saisie. Donc, finalement cet ensemble de données contient 1538 enregistrements, et 13 variables.

Le tableau suivant présente un aperçu de l'ensemble de données obtenu à partir du fichier **data.csv.**[103] Ici, le nombre de colonnes est de 13 et le nombre d'échantillons est de 1538.

Ici ‘***Blood_Pressure_Abnormality*****''** de type binaire est la variable cible, "1" si la personne a une haut tension artérielle et ‘'0" sinon.

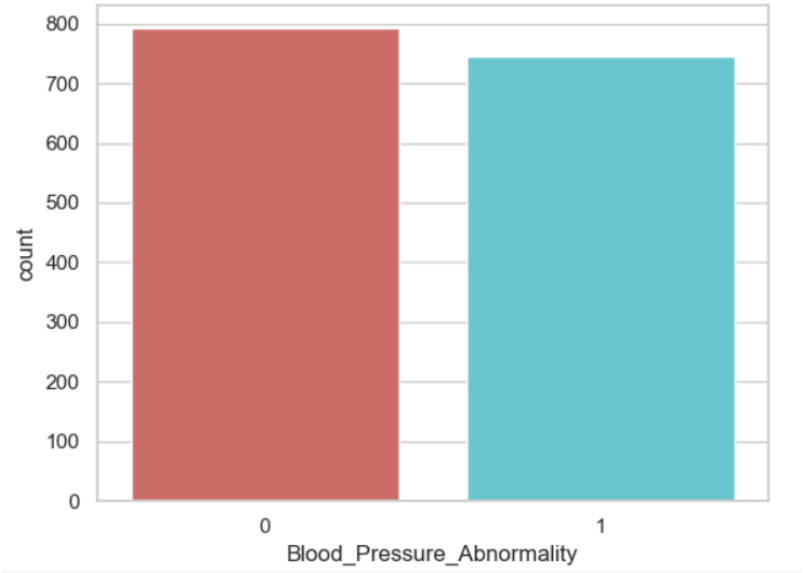
[103]=télécharger depuis Kaggle, lien https://www.kaggle.com/pavanbodanki/blood-press

**Tableau 4** : aperçu sur l'ensemble de données numéro 2

| Blood_Pressure_Abnormality | Level_of_Hemoglobin | Genetic_Pedigree_Coefficient | Age | BMI | Sex | Smoking | Physical_activity | salt_content_in_the_diet | alcohol_consumption_per_day | Level_of_Stress | Chronic_kidney_disease | Adrenal_and_thyroid_disorders |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 9.75 | 0.23 | 54 | 33 | 1 | 0 | 3 | 3 | 3.0 | 3 | 0 | 0 |
| 1 | 10.79 | 0.91 | 70 | 49 | 0 | 0 | 1 | 3 | 1.0 | 2 | 1 | 0 |
| 0 | 11.00 | 0.43 | 71 | 50 | 0 | 0 | 2 | 1 | 3.0 | 1 | 1 | 0 |
| 1 | 14.17 | 0.83 | 52 | 19 | 0 | 0 | 2 | 5 | 4.0 | 2 | 0 | 0 |
| 0 | 12.70 | 0.41 | 48 | 20 | 0 | 0 | 3 | 3 | 2.0 | 2 | 0 | 0 |
| 0 | 10.88 | 0.68 | 72 | 44 | 0 | 0 | 1 | 1 | 1.0 | 3 | 0 | 0 |
| 1 | 14.56 | 0.61 | 40 | 44 | 0 | 0 | 1 | 2 | 1.0 | 2 | 0 | 0 |
| 1 | 8.58 | 0.13 | 70 | 28 | 1 | 0 | 5 | 3 | 1.0 | 1 | 1 | 1 |
| 1 | 12.77 | 0.10 | 35 | 17 | 0 | 0 | 3 | 5 | 4.0 | 3 | 1 | 0 |
| 1 | 16.40 | 0.45 | 31 | 50 | 0 | 1 | 3 | 5 | 1.0 | 3 | 0 | 1 |
| 0 | 16.42 | 0.13 | 75 | 40 | 0 | 1 | 2 | 5 | 4.0 | 3 | 0 | 1 |
| 1 | 11.98 | 0.06 | 65 | 28 | 1 | 0 | 3 | 5 | 4.0 | 2 | 0 | 0 |
| 1 | 11.60 | 0.09 | 61 | 44 | 0 | 0 | 5 | 4 | 2.0 | 2 | 0 | 0 |
| 1 | 16.95 | 0.98 | 40 | 49 | 0 | 1 | 2 | 2 | 4.0 | 1 | 1 | 1 |
| 0 | 10.85 | 0.57 | 61 | 28 | 1 | 1 | 5 | 2 | 1.0 | 3 | 0 | 1 |
| 0 | 11.19 | 0.75 | 71 | 41 | 1 | 1 | 2 | 3 | 3.0 | 3 | 0 | 0 |
| 1 | 13.29 | 0.87 | 27 | 17 | 0 | 0 | 5 | 2 | 1.0 | 2 | 1 | 0 |
| 0 | 10.40 | 0.69 | 66 | 40 | 1 | 0 | 3 | 2 | 3.0 | 1 | 0 | 1 |
| 0 | 10.40 | 0.69 | 66 | 40 | 1 | 0 | 3 | 2 | 3.0 | 1 | 0 | 1 |
| 1 | 12.77 | 0.10 | 35 | 17 | 0 | 0 | 3 | 5 | 4.0 | 3 | 1 | 0 |
| 1 | 16.40 | 0.45 | 31 | 50 | 0 | 1 | 3 | 5 | 1.0 | 3 | 0 | 1 |
| 0 | 16.42 | 0.13 | 75 | 40 | 0 | 1 | 2 | 5 | 4.0 | 3 | 0 | 1 |
| 1 | 11.98 | 0.06 | 65 | 28 | 1 | 0 | 3 | 5 | 4.0 | 2 | 0 | 0 |
| 1 | 11.60 | 0.09 | 61 | 44 | 0 | 0 | 5 | 4 | 2.0 | 2 | 0 | 0 |
| 1 | 16.95 | 0.98 | 40 | 49 | 0 | 1 | 2 | 2 | 4.0 | 1 | 1 | 1 |

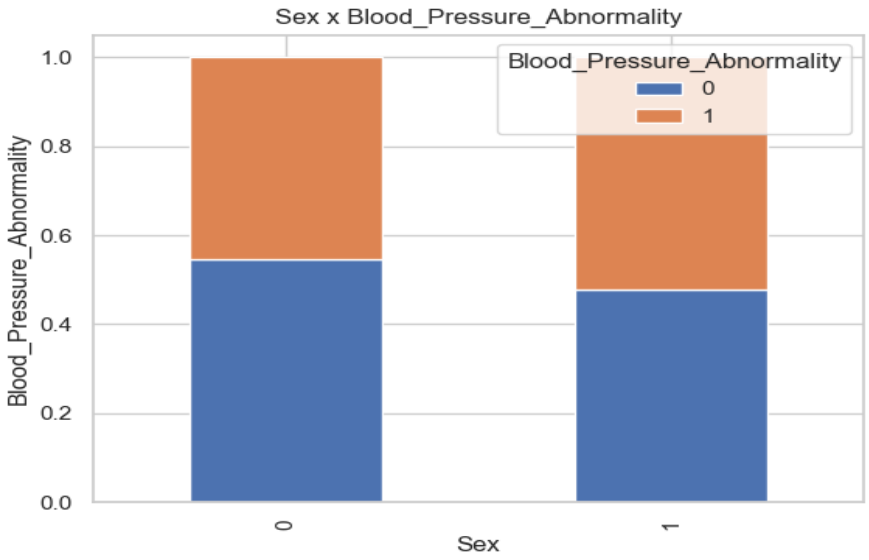
### 3.2.2. Sélection des caractéristiques :

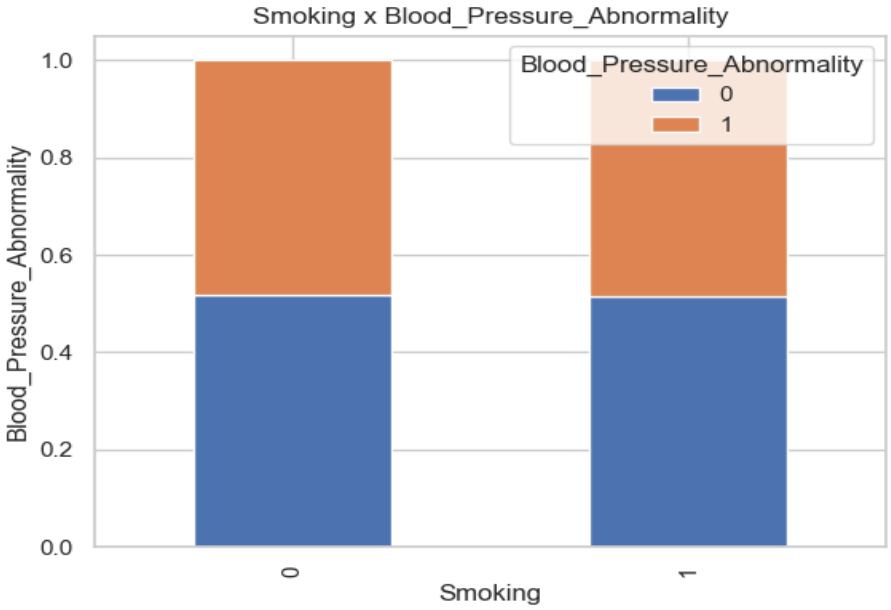
Même procédure que le premier ensemble de données :

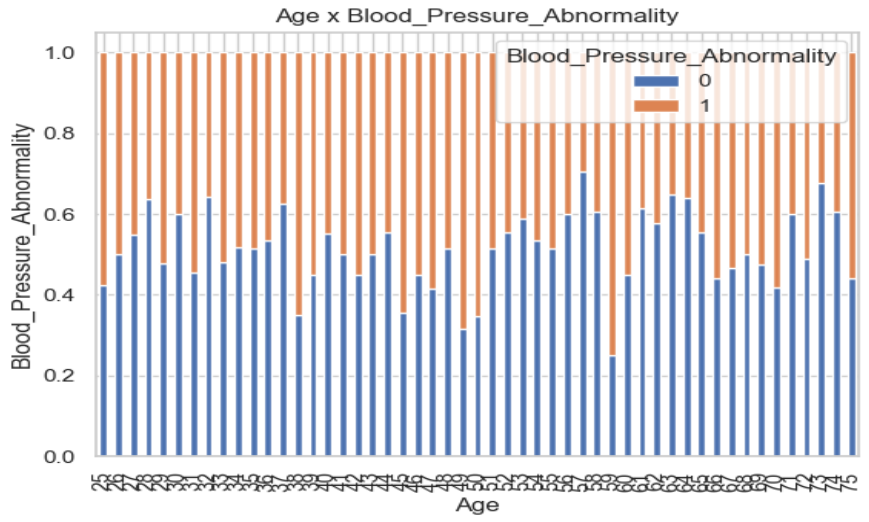


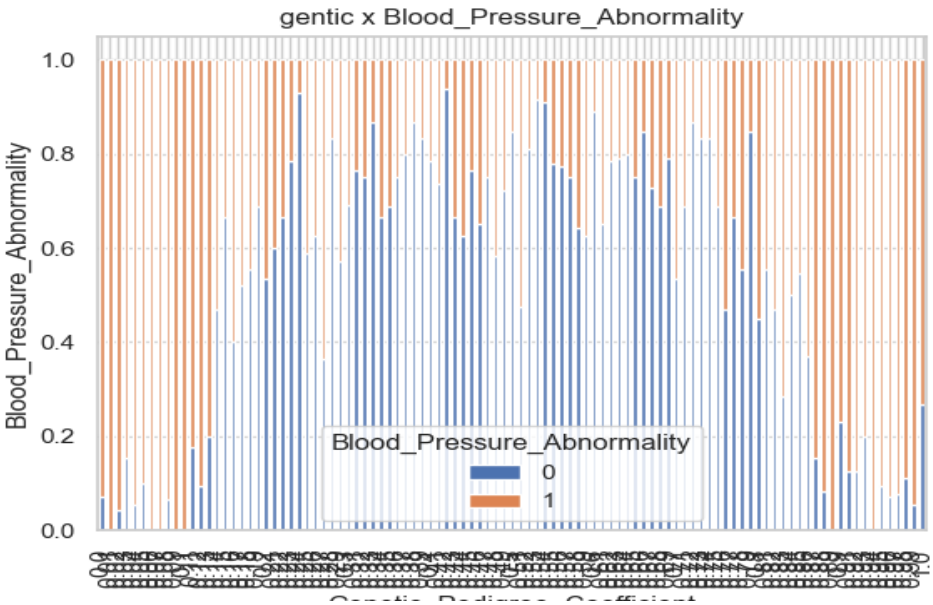

Pourcentage de personne qui n'ont pas hypertension 51.52895250487963

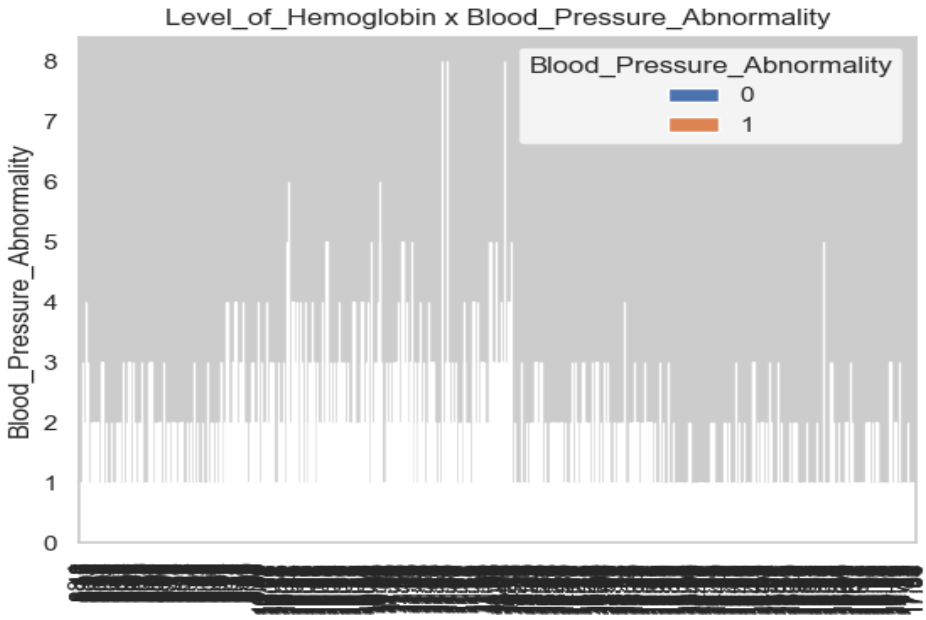
Pourcentage de personne qui ont une hypertension 48.47104749512036

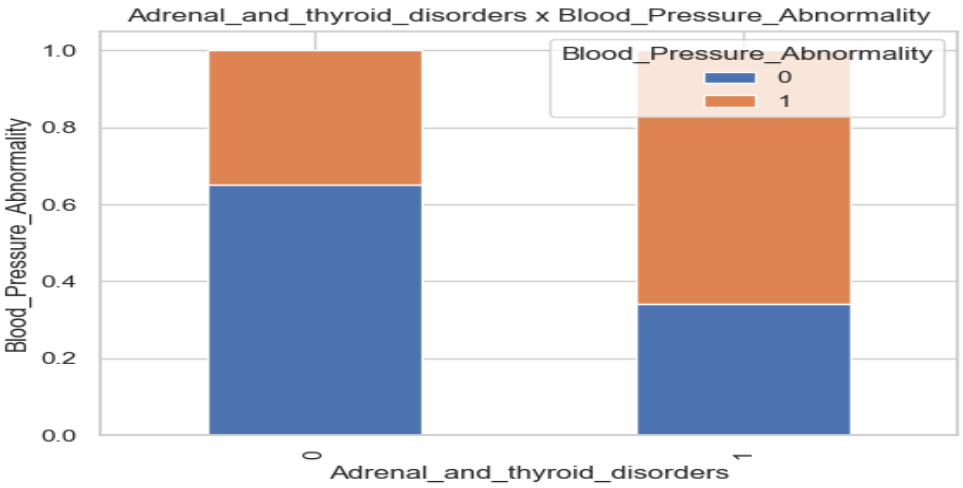

Sex x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
0
1
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Sex

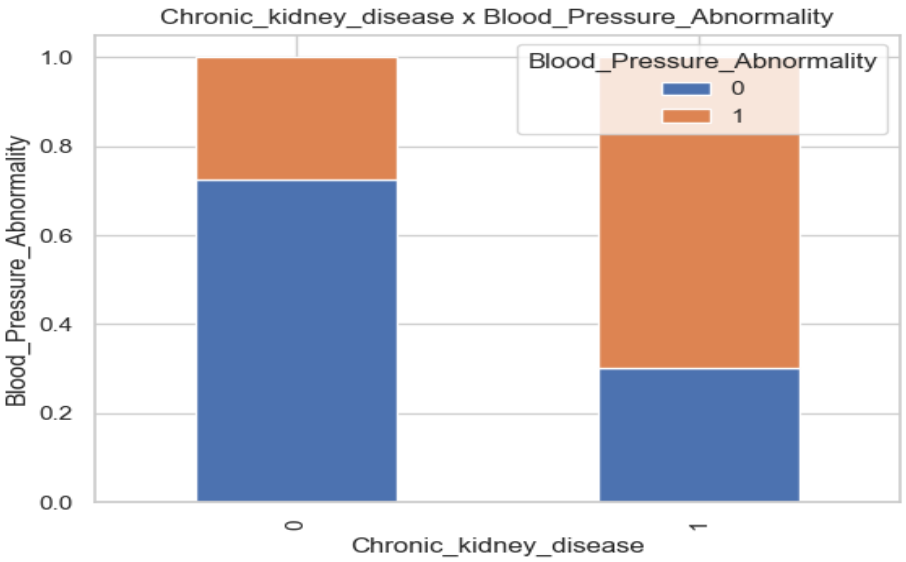

Smoking x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
0
1
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Smoking

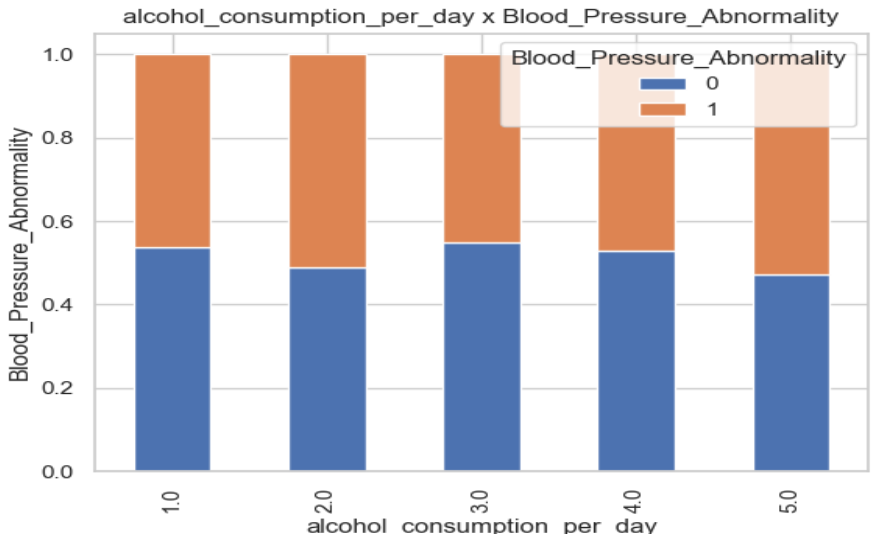

Age x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
0
1
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
Age

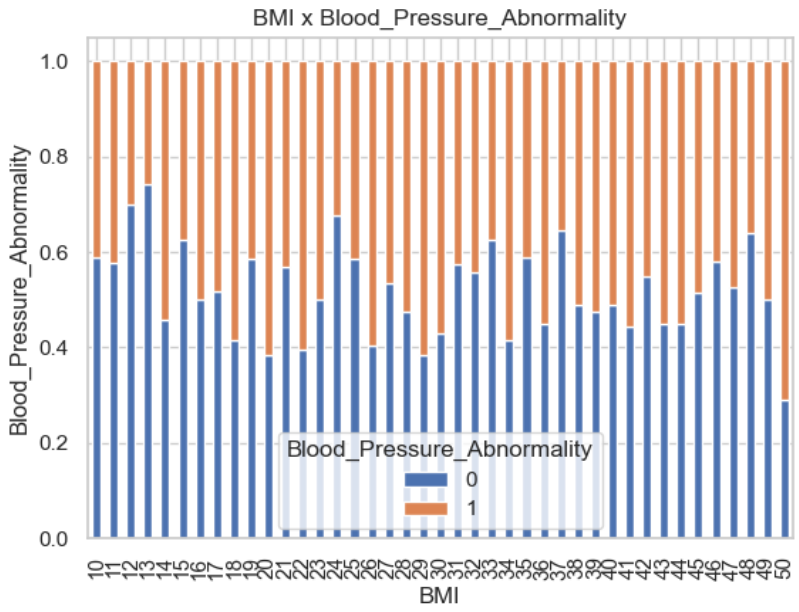

gentic x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
Blood_Pressure_Abnormality
0
1

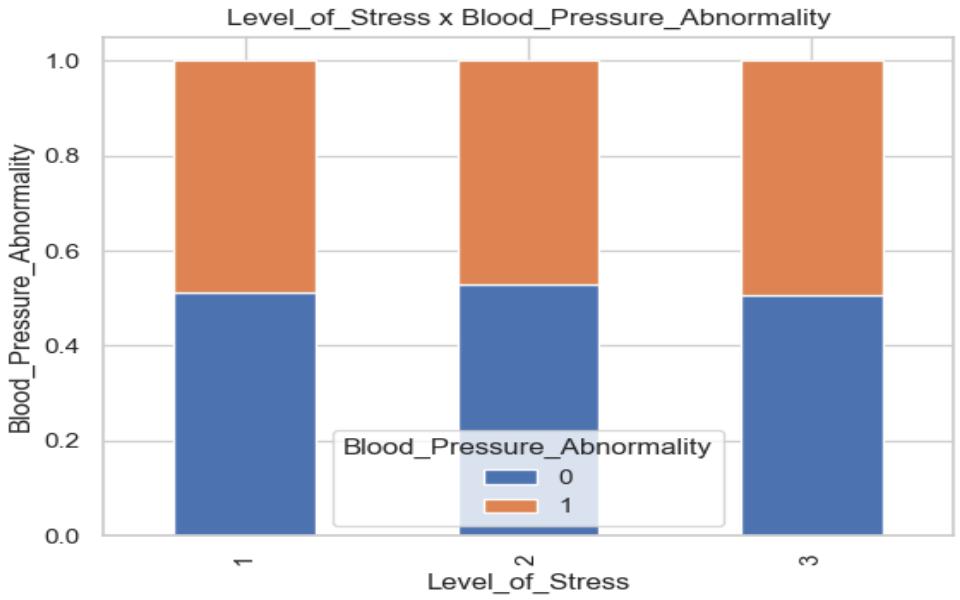

Level_of_Hemoglobin x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
8
7
6
5
4
3
2
1
0
Blood_Pressure_Abnormality
0
1

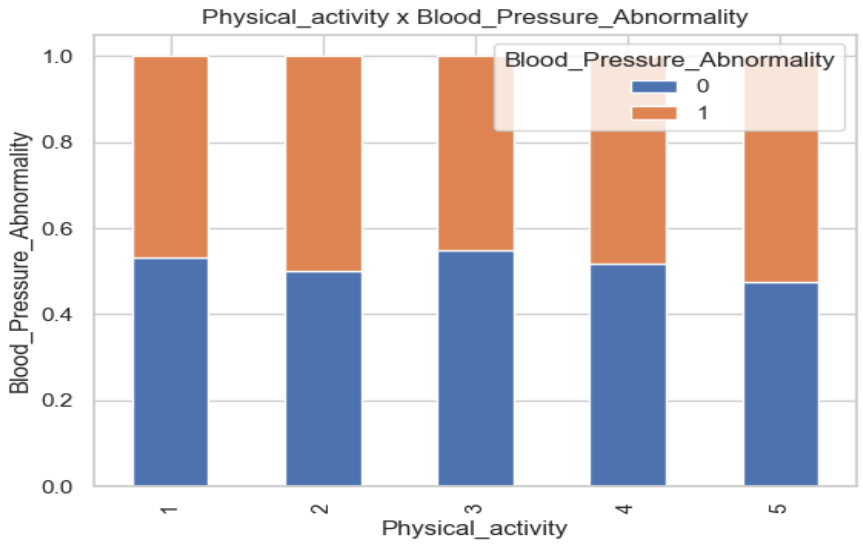

Adrenal_and_thyroid_disorders x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
Blood_Pressure_Abnormality
0
1
0
1
Adrenal_and_thyroid_disorders

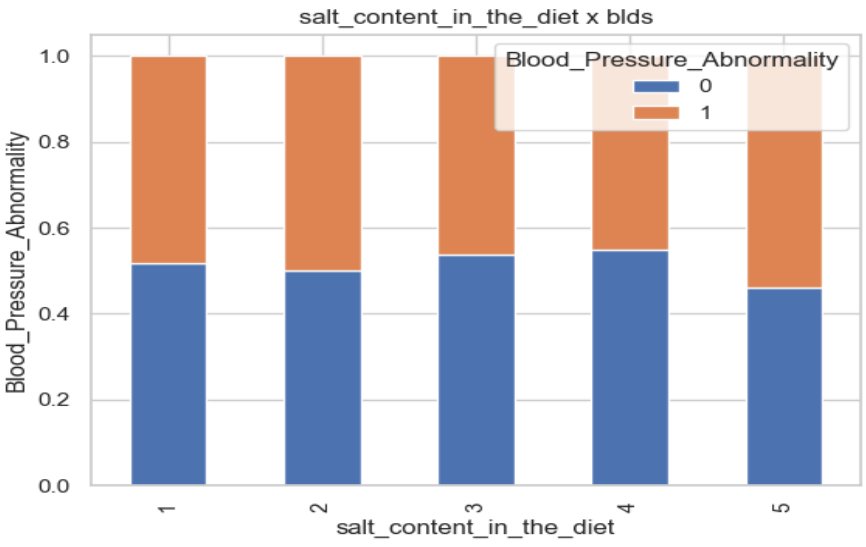

Chronic_kidney_disease x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
0
1
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
0
1
Chronic_kidney_disease


alcohol_consumption_per_day x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
0
1
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
1.0
2.0
3.0
4.0
5.0
alcohol_consumption_per_day

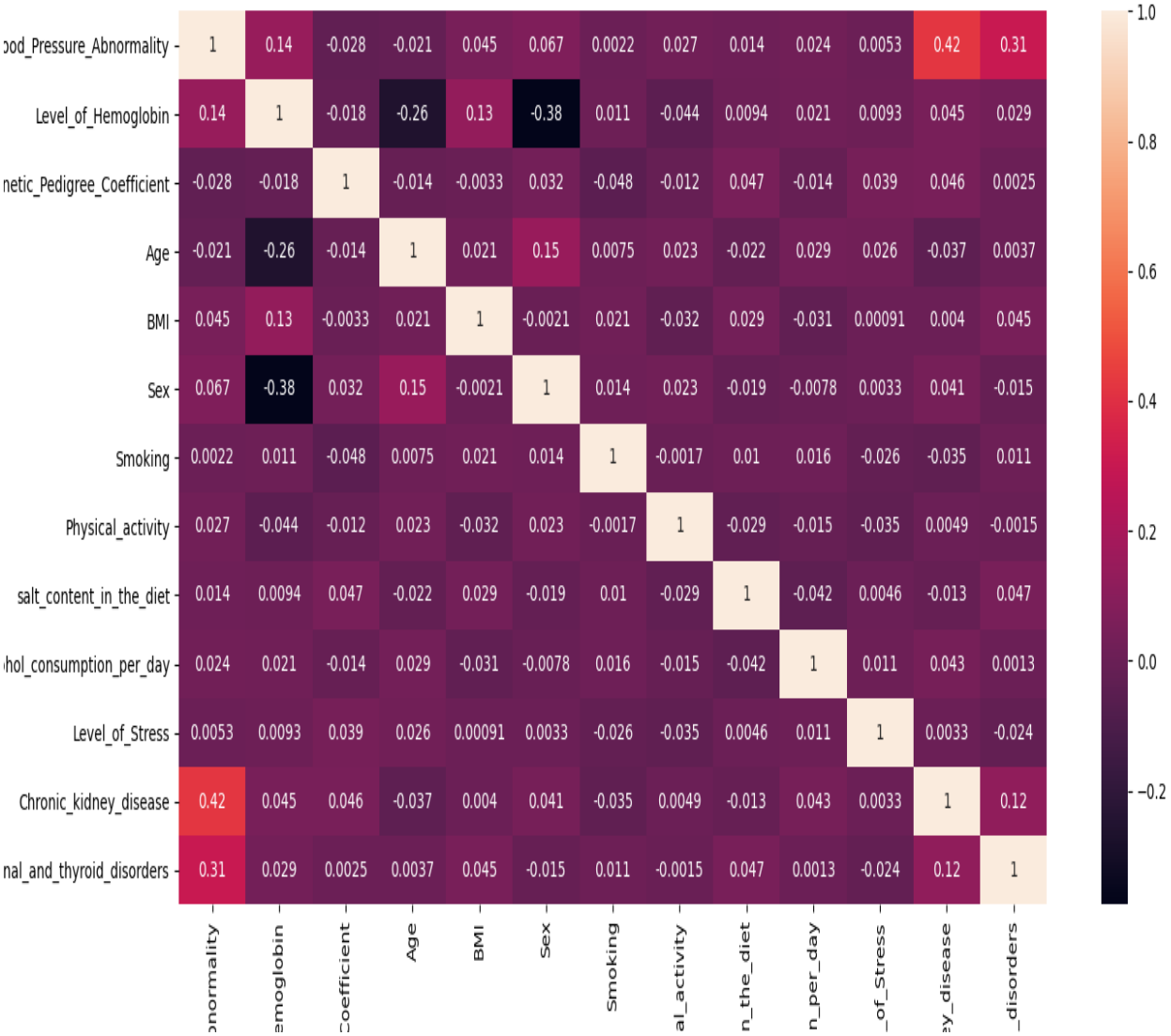

BMI x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
0
1
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
BMI


Level_of_Stress x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
Blood_Pressure_Abnormality
0
1
1
2
3
Level_of_Stress


Physical_activity x Blood_Pressure_Abnormality
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
Blood_Pressure_Abnormality
0
1
1
2
3
4
5
Physical_activity


salt_content_in_the_diet x blds
Blood_Pressure_Abnormality
1.0
0.8
0.6
0.4
0.2
0.0
Blood_Pressure_Abnormality
0
1
1
2
3
4
5
salt_content_in_the_diet

### 3.2.3. Matrice de corrélation




Après avoir analysé la matrice de corrélation et les graphiques de fréquence, nous avons remarqué que les deux variables Smoking et ***Level_of_stress*** ont une très faible corrélation avec la variable cible, donc ces variables ont éliminé.

Le tableau suivant résume chaque variable et ses caractéristiques : (listes de prédicteurs)

| **Variable** | **Type** | **Min/max** | **Valeur normale** | **Remarque** |
|---|---|---|---|---|
| ,Level_of_Hemoglobin | numérique | 8.1/17.6 | 13.5 et 17.5 | niveau d'hémoglobine |
| Genetic_Pedigree_Coefficient | numérique | 0 /1 | / | une représentation génétique qui schématise l'héritage d'une maladie sur plusieurs générations |
| Age | numérique | 25-75 | / | / |
| BMI | numérique | 10-50 | 18.5 / 24.9 kg/m² | L'indice de masse corporelle est une valeur dérivée de la masse et de la taille d'une personne. L'IMC est défini comme la masse corporelle divisée par le carré de la taille du corps, et est exprimé en unités de kg/m² |
| Sex | Binaire | 0/1 | / | Dans cet ensemble de donnée :<br>1 : femme<br>0 : homme |
| Smoking | Binaire | 0/1 | / | 1 : fumeur<br>2 : non-fumeur |
| Physical_activity | numérique | 0/5 | / | Pour des raisons de simplification On a divisé cette variable on 5 catégorie :<br>1 :0- 10000<br>2 : 10000 -20000<br>3 : 20000-30000<br>4 : 30000-40000<br>5 :> 40000<br>Cette variable a été mesurée par la quantité de calories brûlées par semaine pendant l'exercice physique |
| salt_content_in_the_diet | numérique | 0/5 | 3,400 mg par jour (23800 par semaine) | Consommation de sel par semaine Pour des raisons de simplification On a divisé cette variable on 5 catégorie :<br>1 :0- 10000<br>2 : 10000 -20000 |

| | | | | |
|---|---|---|---|---|
| | | | | 3 : 20000-30000<br>4 : 30000-40000<br>5 :> 40000 |
| alcohol_consumption_per_day | numérique | 0-499 | < 200 ml | Consommation d'alcool par jour<br>Pour des raisons de simplification On a divisé cette variable on 5 catégorie :<br>1 :< 100<br>2 : entre 100 et 200<br>3 : entre 200 et 300<br>4 : entre 300 et 400<br>5 :>400 |
| Level_of_Stress | numérique | 1-3 | / | Niveaux de stress<br>Ici on a 3 catégorie :<br>1 : normal<br>2 : moyen<br>3 : élevé |
| Chronic_kidney_disease | Binaire | 1/0 | 0 | Absence /présence d'une maladie rénale chronique |
| Adrenal_and_thyroid_disorder | Binaire | 1/0 | 0 | Absence /présence d'une Troubles surrénaliens et thyroïdiens |

### 3.2.4. Application d'algorithme d'apprentissage :

**A. Arbre de décision :**

Précision : 67,87 %

**B. Réseau de neurone :**

Précision : 88.60 %

**C. Forêt aléatoire :**

Précision : 100 %

**D. SVM :**

Précision : 75,32%

## 4. Discussion :

Nous remarquons que dans les deux ensembles de données, l'algorithme de forêt aléatoire est très puissant avec une précision de 98 % à 100% respectivement. En plus, les performances des autres algorithmes sont nettement supérieures dans le deuxième ensemble de donnée avec 74,49 % et 67,87% dans l'ensemble de données 1 pour les arbres de décision, pour

les réseaux de neurones 75% pour l'ensemble 1 et 88.60% pour l'ensemble 2. Pour l'algorithme SVM, le deuxième ensemble est plus précis avec 75.32%.

## 5. Conclusion

Dans ce chapitre, nous avons décrit les facteurs les plus importants dans la prévention de l'hypertension. Puis, nous avons exploré nos ensembles de données en les prétraitant chacun séparément. Ensuite, nous avons appliqué des algorithmes d'apprentissage automatique. Par conséquent, les résultats ont montré que l'algorithme le plus approprié pour notre application est la forêt aléatoire avec à la fois une précision remarquable de plus de 98%. Nous avons, également, choisi de travailler avec le deuxième ensemble de données car il avait plus de corrélation entre les variables que le premier et l'autre algorithme y fonctionnait mieux, surtout car il avait des variables plus importantes comme le coefficient de pedigree génétique, l'apport en sel et l'activité physique qui sont des facteurs très courants dans l'hypertension en Algérie.

# Chapitre 4 :

# Implémentation et réalisation

## Introduction

Après la phase de prétraitement, nous sommes maintenant prêts à créer une application basée sur l'ensemble de données choisi comme indiqué dans le chapitre précédent. Dans ce chapitre, nous expliquerons les composants du système et son architecture et ses principales fonctionnalités.

## 1-Environnement de travail et outils utilisés

L'environnement de travail est constitué de deux parties nommées environnement matériel et environnement logiciel.

### 1- 1-Environnement matériel

L'environnement matériel utilisé pour accomplir ce travail est caractérisé par :

a. un système d'exploitation : Windows 10 professionnel 64-bit

b. un CPU : Intel(R) Core (TM) i7-8700k CPU 3.70GHz  3.70 GHz

c. une mémoire : 32 go ddr4

d. Et, une carte graphique : NVidia Gtx 1080 8 go gdd5

### 1-2- Environnement logiciel

**a- python :**

Python est un langage de programmation populaire. Il a été créé par Guido van Rossum et sorti en 1991.

- Il est utilisé pour le /les :

  - Développement web (côté serveur),
  - Développement de logiciels,
  - Mathématiques,
  - Scripts système.[104]

- Que peut faire Python ?

  - Python peut être utilisé sur un serveur pour créer des applications Web.
  - Python peut être utilisé avec un logiciel pour créer des flux de travail.
  - Python peut se connecter aux systèmes de base de données. Il peut également lire et modifier des fichiers.

[104] W3schools. Récupéré sur https://www.w3schools.com/python/python_intro.asp

- Python peut être utilisé pour gérer le Big Data et effectuer des mathématiques complexes.
- Python peut être utilisé pour le prototypage rapide ou pour le développement de logiciels prêts pour la production.[105]

- Pourquoi Python ?

- Python fonctionne sur différentes plateformes (Windows, Mac, Linux, Raspberry Pi, etc.).
- Python a une syntaxe simple similaire à la langue anglaise.
- Python a une syntaxe qui permet aux développeurs d'écrire des programmes avec moins de lignes que certains autres langages de programmation.
- Python s'exécute sur un système d'interprétation, ce qui signifie que le code peut être exécuté dès qu'il est écrit. Cela signifie que le prototypage peut être très rapide. (w3schools, s.d.)
- Python peut être traité de manière procédurale, orientée objet ou fonctionnelle

**b- Les bibliothèques utilisées :**

**b1. Sklearn :**

Sklearn est un module Python intégrant des algorithmes d'apprentissage automatique classiques dans le monde très soudé des packages scientifiques Python (numpy, scipy, matplotlib).

Il vise à apporter des solutions simples et efficaces aux problèmes d'apprentissage, accessibles à tous et réutilisables dans divers contextes : l'apprentissage automatique comme outil polyvalent pour la science et l'ingénierie.[106]

**b2. Pandas :** est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. (Kite, s.d.)

**b3. PyDotPlus** : est une version améliorée de l'ancien projet ***pydot*** qui fournit une interface Python au langage Dot de Graphviz.[107]

**2- Architecture du système**

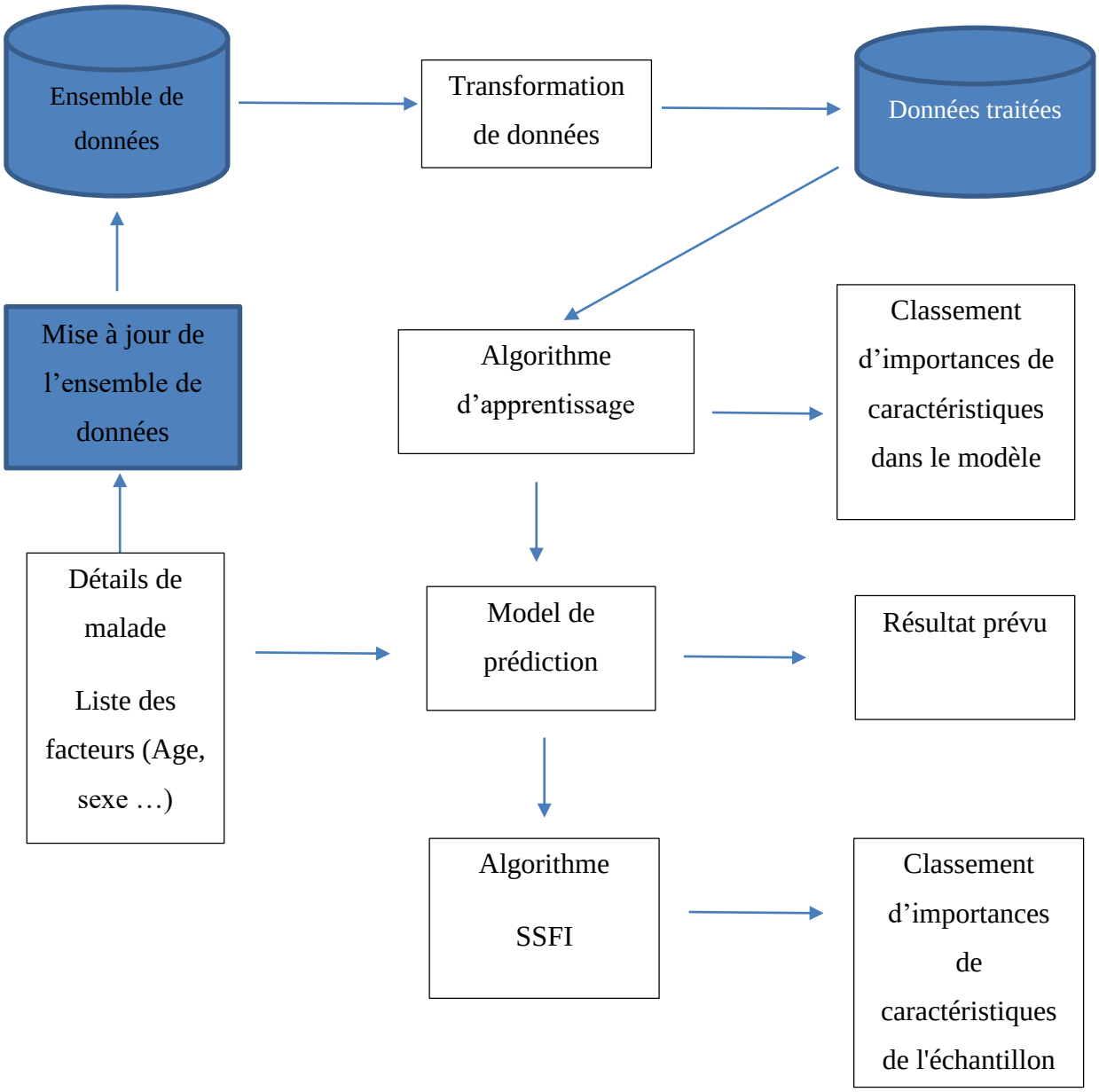
La prédiction de la maladie à l'aide de l'apprentissage automatique prédit la présence de la maladie pour l'utilisateur en fonction de divers symptômes et des informations fournies par

[105] W3schools, op , *cit* .
[106] Kite. (s.d.). Récupéré sur https://www.kite.com/python/docs/sklearn
[107] Ibid.

l'utilisateur, telles que le taux de sucre, le taux d'hémoglobine et bien d'autres informations générales à travers les symptômes. L'architecture de notre système, de prédiction de l'hypertension, se compose de divers ensembles de données à travers lesquels nous comparerons les données de l'utilisateur et les prédisons, puis les ensembles de données sont transformés en des ensembles plus grands. A partir de ces ensembles, les données sont classées en fonction des algorithmes de classification. Ensuite, les données classifiées sont traitées par les algorithmes d'apprentissage automatique, grâce auxquels les données sont traitées et transmises au modèle de prédiction de la maladie en utilisant toutes les entrées de l'utilisateur mentionnées ci-dessus. Ainsi, une fois que l'utilisateur a saisi les informations ci-dessus et les données sont globalement traitées, se combinent et se comparent dans le modèle de prédiction du système et prédisent enfin la maladie.

Le diagramme ci-dessous est un diagramme d'architecture qui est une représentation graphique d'un ensemble de concepts faisant partie de notre architecture, y compris leurs principes, éléments et composants. Le diagramme explique le fonctionnement du système dans la perception de la vue d'ensemble du système.

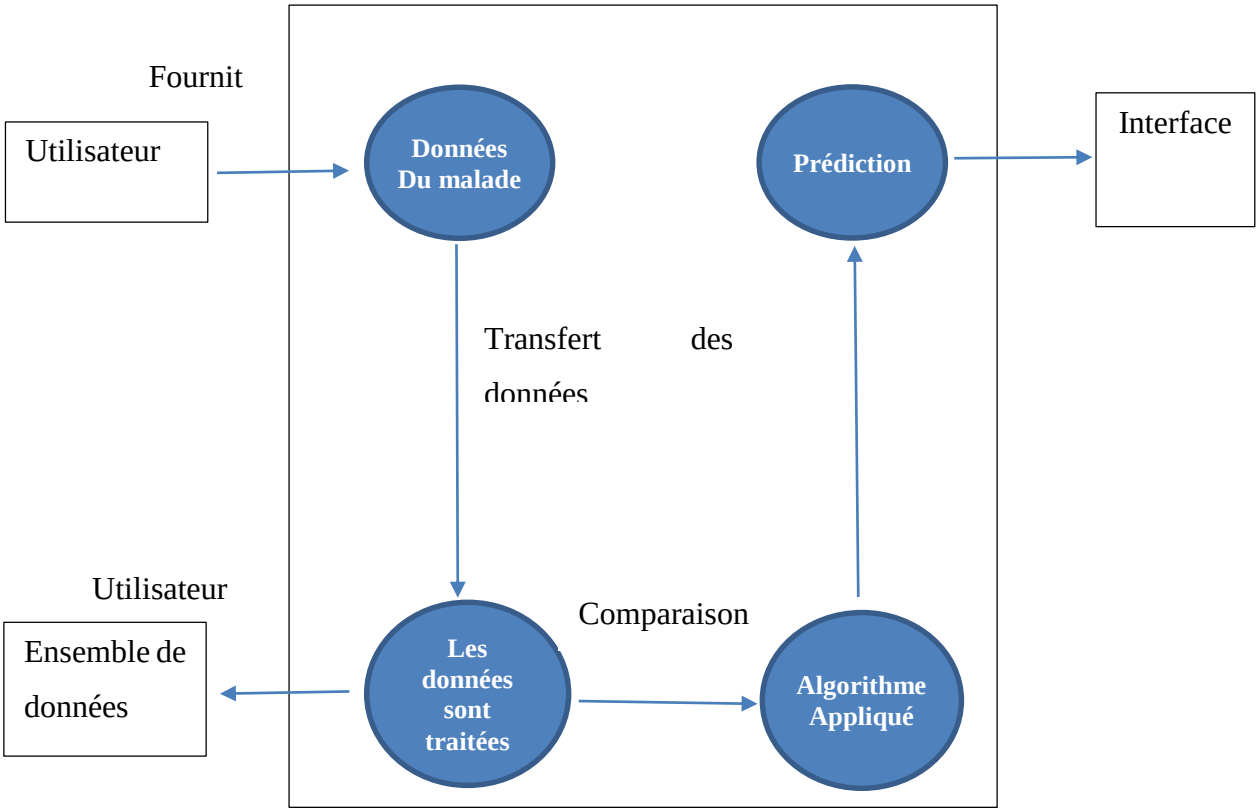
**Figure 17** : Architecture de système réalisé




### 3- Diagramme de flux de données :

Le diagramme de flux de données du projet de prédiction de la maladie de tension à l'aide de l'apprentissage automatique comprend tous les divers aspects qu'un diagramme de flux normal requiert. Le diagramme de flux de données montre comment, à partir du démarrage, le modèle passe d'une étape à l'autre (Voir Figure 19). Lorsque l'utilisateur ouvre le système, il fait entrer toutes les informations nécessaires ainsi que les facteurs qui peuvent aider à prédire le taux d'estimation de tomber malade ou non. A partir du modèle de prédiction, le système prédit les résultats appropriés, sinon il montre les détails où l'utilisateur s'est trompé lors de la saisie des informations.

**Figure 18** : Diagramme de flux de données




**4- Implémentation :**

L'objectif de cette application est de fournir aux médecins ou aux travailleurs d'un laboratoire un système d'aide à la décision qui pourrait aider à comprendre les facteurs impliqués dans l'hypertension artérielle. Ce dernier utilise l'un des algorithmes d'apprentissage automatique qui est celui des forêts aléatoires à appliquer sur l'ensemble de données prétraité dans le chapitre précédent, en prenant tous les facteurs pris en compte et fournissant à l'utilisateur une représentation visuelle de l'importance de chaque facteur pour chaque instance donnant à tout moment. Il pourrait, donc, être un bon outil pour comprendre si une personne risque de développer une hypertension. Notre application fournit, également, une forte estimation sur lequel des facteurs aura plus d'influence que l'autre sur un malade donné. Cette application permet, également, à un médecin de faire son propre jeu de données pour faire des prédictions dans le futur.

**4-1-Les fonctions de base :**

L'application est basée sur quatre (4) fonctions principales :

**Fonction 1 :** Elle met à jour l'ensemble de données en téléchargeant de nouvelles données chaque fois que l'utilisateur les fait entrer et entraîne à nouveau l'ensemble de données. Les étapes de cette procédure ont été déjà expliquées dans le chapitre 3.

**Code de la fonction 1 :**

```
def train():

global col_names

global pima

global X

global y

global clf

global feature_cols

col_names =
['Blood_Pressure_Abnormality','Level_of_Hemoglobin','Genetic_Pedigree_Coefficient','Age','
BMI','Sex','Smoking'
,'Physical_activity','salt_content_in_the_diet','alcohol_consumption_per_day'
,'Level_of_Stress','Chronic_kidney_disease','Adrenal_and_thyroid_disorders']

  # chargement de l'ensemble de données

  pima = pd.read_csv("Dataset2.csv", header=None, names=col_names)

  pima.head()

feature_cols=['Level_of_Hemoglobin','Genetic_Pedigree_Coefficient','Age','Chronic_kidney_
disease','Adrenal_and_thyroid_disorders']

X = pima[feature_cols] # facteur

y= pima.Blood_Pressure_Abnormality # variable cible

# fractionner l'ensemble de données : 80 % pour l'entrainement et 20 % pour tester

  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.70)
```

```
clf = RandomForestClassifier(n_estimators = 5  )

# Entraîner le modèle sur l'ensemble de données avec l'algorithme forêt aléatoire .

clf.fit(X_train, y_train)

print("done")
```

**Fonction 2 :** elle classe les facteurs dans le modèle des plus influant aux moins influant. Pour le faire, on fait appelé à la fonction prédéfinie (***feature_importances***_), puis on l'implémente dans un graphe qui soit un diagramme circulaire ou un histogramme …

**Code de la fonction 2 :**

```
def fearturesinmodel():
import matplotlib.pyplot as plt2 # bibliothèque nécessaire  pour génération de graph
y = np.array(clf.feature_importances_)
mylabels = ["niveau hemogolobin","facteur génétique","age","trouble rénal","trouble surrénalien"]
plt2.pie(y ,labels = mylabels)
plt2.xlabel("Random Forest Feature Importance")
plt2.show()
```

**Fonction 3 :** elle classe les facteurs dans l'échantillon des plus influant aux moins influant.

Cette fonction ne doit pas être confondue avec la seconde fonction qui ne classe que les caractéristiques dans le modèle. Cette fonction classe les caractéristiques pour chaque instance, par exemple si l'utilisateur entre les facteurs (âge = 30, taux d'hémoglobine : 13,5, facteur génétique : 0.5, ...etc.) la fonction va classer les caractéristiques de cette instance. Par conséquent, elle permettra à l'utilisateur d'observer les facteurs les plus importants pour chaque cas.

Cette fonction est construite sur la base d'un algorithme appelé SSFI (single sample feature importance).[108]

[108] Joseph Gatto, R. L. (2019, 11 27). Single Sample Feature Importance: An Interpretable Algorithm for Low-Level Feature Analysis. Jet Propulsion Propulsion Laboratory, California Institute of Technology 2 Columbia University. Récupéré sur https://arxiv.org/pdf/1911.11901.pdf

**Algorithme 1:** SSFI .[109]

```
        SSFI Results = {}
for sample in Dataset do
        Feature Importances = {}
        Train Data = Dataset – sample
        Test Sample = sample
        RF Model = RF.train(Train Data)
        while RF Model.predict(Test Sample) do
        for tree in RF Model do
                Predict Path = tree.predict(Test Sample)
                for (node, feature name) in Predict Path do
                        Feature Importance = node importance(node)
                        Feature Importances[feature name] += Feature Importance
SSFI Results[sample] = Feature Importances
```

**Code de la fonction 3 :**

```
def samplepredict ():
 global Level_of_Hemoglobin
 Level_of_Hemoglobin = float(entry1.get())
 global Genetic_Pedigree_Coefficient
 Genetic_Pedigree_Coefficient = float(entry2.get())
 for x in range(5):
  for node in graph.get_node_list():
     if node.get_attributes().get('label') is None:
       continue
     if 'samples = ' in node.get_attributes()['label']:
       labels = node.get_attributes()['label'].split('<br/>')
       for i, label in enumerate(labels):
          if label.startswith('samples = '):
             labels[i] = 'samples = 0'
```

[109] Joseph Gatto. op. cit . p 5 .

```
      node.set('label', '<br/>'.join(labels))
      node.set_fillcolor('white');
samples = [[13.5 ,0.5 ,30 , 0, 0]]
decision_paths = clf.estimators_[x].decision_path(samples)
for decision_path in decision_paths:
   for n, node_value in enumerate(decision_path.toarray()[0]):
      if node_value == 0:
         continue
      node = graph.get_node(str(n))[0]
      labels = node.get_attributes()['label'].split('<br/>')
      for i, label in enumerate(labels):
         if re.findall("Genet",label):
           n = i+1
           for n, label in enumerate(labels):
              if re.findall("gini",label):
                 ginigene=float(label[7:12])+ginigene
         if re.findall("Level",label):
           r = i+1
           for r, label in enumerate(labels):
              if re.findall("gini",label):
                 ginihmg=float(label[7:12])+ginihmg
         if re.findall("Age",label):
           t = i+1
           for t, label in enumerate(labels):
              if re.findall("gini",label):
                 giniage=float(label[7:12])+giniage
         if re.findall("Chroni",label):
           o = i+1
           for o, label in enumerate(labels):
              if re.findall("gini",label):
                 ginikidney=float(label[7:12])+ginikidney
         if re.findall("Adrenal",label):
           p = i+1
           for p, label in enumerate(labels):
```

```
                if re.findall("gini",label):
                    giniadrenal=float(label[7:12])+giniadrenal
                    print(giniadrenal)
        # node.set('label', '<br/>'.join(labels))
tablx = [  ginihmg , ginigene , giniage  , ginikidney , giniadrenal]
```

**Fonction 4 :** qui donne une prédiction sous une forme d'un nombre réel qui définit le pourcentage de la probabilité de développer une maladie hypertension basée sur le modèle formé.

Après l'entrainement d'un nouvel ensemble de donnée, l'utilisateur sera capable de faire une prédiction. Grace à l'utilisation de bibliothèque ***sklearn***, la fonction ***predict()*** a été simple à implémenter pour faire une prédiction. La sortie de cette fonction est une valeur numérique (pourcentage).

**Code de la fonction 4 :**

```
        def predict ():
            Prediction_result= ('Predicted Result: ',
            clf.predict([[Level_of_Hemoglobin,Genetic_Pedigree_Coefficient,
            Age,Chronic_kidney_disease, Adrenal_and_thyroid_disorders]]))
            label_Prediction = Label(root, text= Prediction result, bg='sky blue')
            canvas1.create_window(270, 280, window=label Prediction)
```

### 4-2- Interface graphique :

On explique dans ce qui suit la structure que nous utiliserons dans notre interface :

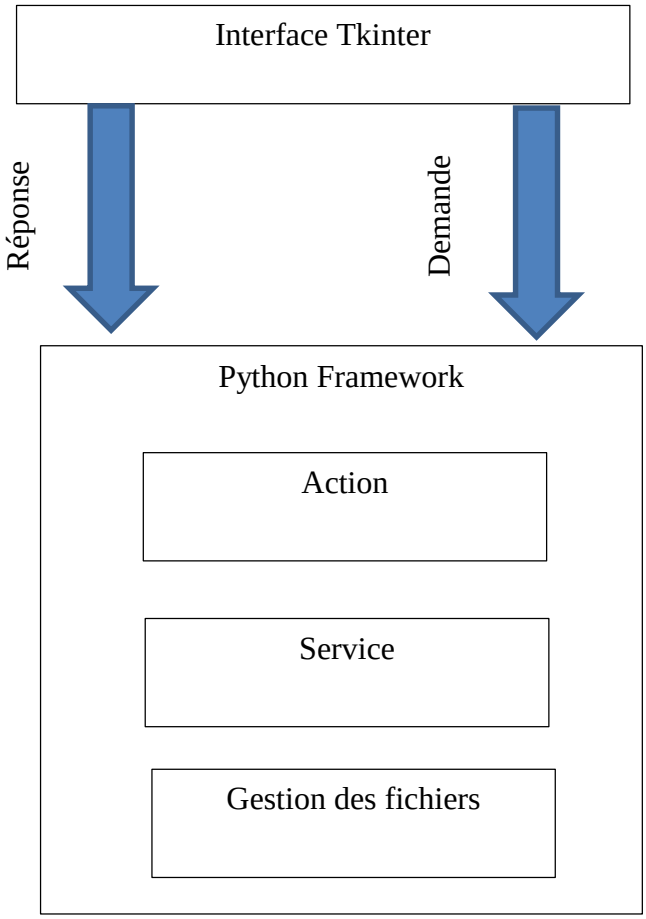
• L'interface utilisateur de ce système se compose de l'interface de la bibliothèque Python appelée ***tkinter.***

• Ensuite, il entre dans le modèle de Framework où toutes les actions et tous les services sont combinés, puis le résultat est traité.

• Il se compose également d'un système de fichiers où toutes les informations relatives à l'utilisateur sont stockées, telles que le nom d'utilisateur, le mot de passe, l'âge, etc.

On explique ci-dessous la structure de l'interface utilisateur ainsi que les implémentations nécessaires.

Diagramme d'interface




• Après l'interface utilisateur, il s'agit d'expliquer le cadre dans lequel le système fonctionne en utilisant toutes les technologies, algorithmes et divers outils dans lesquels le projet fonctionne en conséquence.

• Le cadre se compose de tous les modules à partir de la préparation des données, de la construction des données et de l'étape d'évaluation.

• Ces trois facteurs entrent ensuite dans la phase de collecte de données, où les données sont classées, en conséquence, à l'aide des modèles et algorithmes appropriés tels que l'arbre de décision, la forêt aléatoire…

• Ensuite, tous ces algorithmes utilisent les ensembles de données et forment les ensembles où toutes les données précédentes sont stockées. Puis, en utilisant ces données, ils les comparent aux nouvelles données et le résultat est généré.

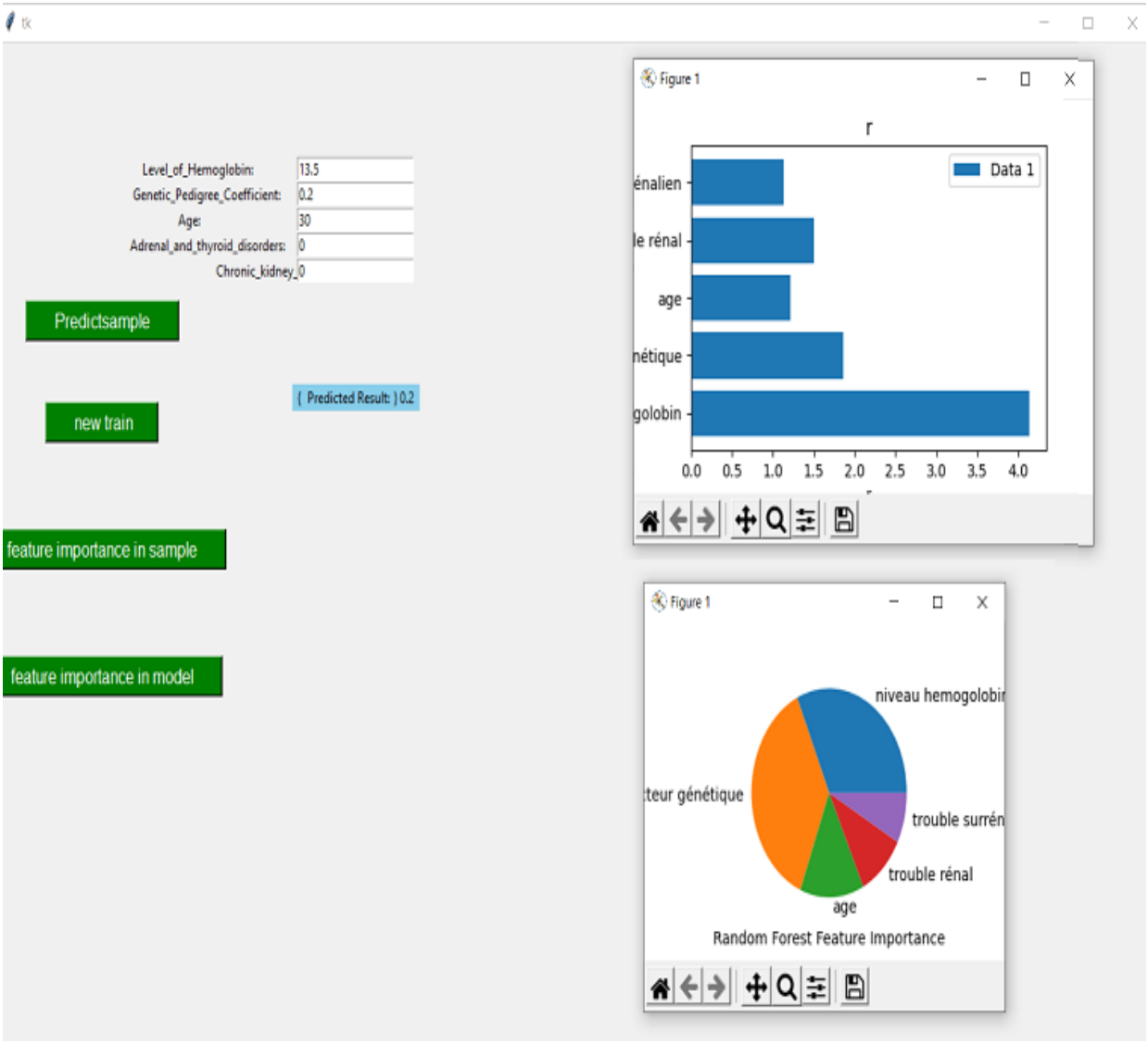
• Ensuite, un travail de prétraitement aura lieu pour réduire et analyser les données présentes dans le système.

• Ensuite, à l'aide de l'interface utilisateur, les données sont transférées dans l'écran principal.

• Plus tard, toutes ces données sont analysées et validées puis le résultat final est généré.

• Enfin, une fois que l'utilisateur a saisi les facteurs, tous les mécanismes principaux fonctionnent et le résultat prévu est affiché dans l'interface utilisateur.

**Figure 19 :** l'interface finale de notre application



## 6. Conclusion

Dans ce dernier chapitre, nous avons expliqué l'architecture du système et tous ses composants en soulignant toutes les fonctions principales. Ce système est simple et il pourrait avoir plus de développements à l'avenir en ajoutant plus de fonctionnalités par exemple en ajoutant un nouveau bouton pour choisir quel algorithme d'apprentissage à utiliser et également en créant une interface plus avancée pour gérer les données de chaque personne par elle-même.

# Conclusion générale

**Conclusion général :**

Nous pouvons conclure en disant que, ce projet « La prédiction de maladie de la tension artérielle à l'aide de l'apprentissage automatique est très utile » dans la vie quotidienne de chacun et il est principalement plus important pour le secteur de la santé, car c'est ce dernier qui utilise quotidiennement ces systèmes pour prédire les maladies des patients en fonction de leurs informations générales et des facteurs impliqués dans la prédiction. De nos jours, la santé intelligente joue un rôle majeur dans la guérison des maladies, c'est donc aussi une sorte d'aide de l'industrie de la santé, et c'est également utile pour l'utilisateur au cas où il ne voudrait pas y aller à l'hôpital. Ainsi, en saisissant simplement les facteurs et toutes autres informations utiles, l'utilisateur peut connaître sa maladie. Si l'industrie de la santé adopte ce projet, les efforts des médecins peuvent être réduits et ils peuvent facilement prédire la maladie du patient. La prédiction de la maladie est de fournir une prédiction pour les maladies diverses qui, lorsqu'elles ne sont pas contrôlées et parfois ignorées, peuvent se transformer en maladie mortelle et causent de nombreux problèmes au patient et aux membres de sa famille.

L’apprentissage automatique a été utilisé pour générer une forêt d’arbres aléatoires qui est utile dans la prédiction de l'hypertension. L'utilisation de modèles prédictifs pour identifier les personnes potentiellement hypertendues à plusieurs implications dans le monde réel, notamment l'adaptation de solutions préventives aux personnes à haut risque de développer une hypertension. Grâce à une communication précise des risques, les modèles prédictifs peuvent aider à améliorer la prise de décision partagée en matière de santé concernant les personnes les plus à risque de développer la maladie. Les modèles prédictifs de l'hypertension peuvent également aider à décider du niveau d'interventions nécessaires au sein de la communauté et ainsi assurer un impact positif. Les recherches futures devraient envisager d'améliorer la précision prédictive des modèles en utilisant les algorithmes dans des populations générales plus larges pour éviter l'effet de population en bonne santé. Cette recherche peut être étendue en évaluant d'autres prédicteurs et en utilisant différents algorithmes de prédiction tels qu'un réseau de neurones artificiels, une machine à vecteurs de support, un classificateur naïf de Bayes et des machines d'amplification de gradient.

**Amélioration future**

- Interface utilisateur plus interactive.

• Peut être intégré comme une page Web et créer un réseau d'échange entre les médecines ou les utilisateurs généraux de cette application pour échanger de nouvelles bases de données.

• Peut être transformée en une application mobile.

# Bibliographie


A Virdis, C. G. (s.d.). Cigarette smoking and hypertension PMID: 20550499. *National Library of Medecine*. doi:10.2174/138161210792062920.

Ahirwar, K. (2017). *Everything you need to know about Neural Networks.* Consulté le 06 06, 2021, sur hackernoon: https://hackernoon.com/everything-you-need-to-know-about-neural-networks-8988c3ee4491

Amisha, P. M. (2019 , juillet). Overview of artificial intelligence in medicine. *journal of medecine and primary care* (PMCID: PMC6691444). Récupéré sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6691444/

Andrea Grillo, L. S. (2019). Sodium Intake and Hypertension. *Nutrients*. doi:10.3390/nu11091970

Azencott, C.-A. (2019). *Introduction au Machine Learning.* Paris: Dunod.

Bing Leng, Y. J. (2015, 02). Socioeconomic status and hypertension: a meta-analysis. *Journal of Hypertension, 33*, 221-229. doi:10.1097/HJH.0000000000000428

Daniele, A. N. (2019). *The "Weight" of Obesity on Arterial Hypertension* . Open access peer-reviewed chapter. doi:10.5772/intechopen.87774

Denis, F. (s.d.). *chapitre 2:Les arbres de décision.* Consulté le 06 2021, 06, sur http://pageperso.lif.univ-mrs.fr/~francois.denis/IAAM1/chap2.pdf

Dhiraj, K. (2019, 05 26). *Top 5 advantages and disadvantages of Decision Tree Algorithm.* Consulté le 06 2021, 06, sur dhirajkumarblog: https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

Ferguson, J. (s.d.). *neural networks in healthcare*. Consulté le 06 06, 2021, sur royal jay: https://royaljay.com/healthcare/neural-networks-in-healthcare/

Horning, N. (s.d.). *Introduction to decision trees and random forests.* (A. M. History's, Éd.)

Hunt, P. N. (2003). Genetics of hypertension . *Genetics in Medicine, 5*, 413–429. Récupéré sur https://www.nature.com/articles/gim2003368

Hurwitz, J. (s.d.). *Le machine learning et la science des données.* Consulté le 05/06/2021, sur IBM: https://www.ibm.com/fr-fr/analytics/machine-learning

ibm. (2021). *Artificial intelligence in medicine.* Récupéré sur ibm: https://www.ibm.com/watson-health/learn/artificial-intelligence-medicine

Ismaili, Z. (s.d.). *apprentissage-supervise-vs-non-supervise.* Consulté le 06 06, 2021, sur analytics and insights: https://analyticsinsights.io/apprentissage-supervise-vs-non-supervise/#:~:text=La%20principale%20diff%C3%A9rence%20entre%20les,la%20base%20d'une%20v%C3%A9rit%C3%A9.&text=En%20revanche%2C%20l'apprentissage%20non,ensemble%20de%20points%20de%20donn%C3%A9es.

Issarane, H. (s.d.). *Support Vector Machines.* Récupéré sur analytic and sinsights.: https://analyticsinsights.io/les-svm-support-vector-machine/

javatpoint. (s.d.). *Random Forest Algorithm.* Récupéré sur javatpoint: https://www.javatpoint.com/machine-learning-random-forest-algorithm

Joseph Gatto, R. L. (2019, 11 27). Single Sample Feature Importance: An Interpretable Algorithm for Low-Level Feature Analysis. Jet Propulsion Propulsion Laboratory, California Institute of Technology 2 Columbia University. Récupéré sur https://arxiv.org/pdf/1911.11901.pdf

Kazim Husain, R. A. (2014, 05 26). Alcohol-induced hypertension: Mechanism and prevention . *World J Cardiol*, 245–252. doi:10.4330/wjc.v6.i5.245

Kite. (s.d.). Récupéré sur https://www.kite.com/python/docs/sklearn

Kivimäki M, B. G.-M. (2009). Validating the Framingham hypertension risk score: results from the Whitehall II Study: hypertesnion . *HYPERTENSIONAHA.109.132373*, 54(3):496–501.

Konstantina Kourou, T. P. (2015, 8 17). Machine learning applications in cancer prognosis and prediction. (Elsevier, Éd.) *Computational and Structural Biotechnology Journal, 13*, 9. Récupéré sur www.elsevier.com/locate/c sbj

mathworks. (s.d.). *Support Vector Machine (SVM).* Consulté le 06 06, 2021, sur mathworks: https://fr.mathworks.com/discovery/support-vector-machine.html

Mayo, p. d. (2021, 07 01). *High blood pressure (hypertension).* Consulté le 07 03, 2021, sur mayo clinic: https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410#:~:text=High%20blood%20pressure%20(hypertension)%20is,problems%2C%20such%20as%20heart%20disease.

Moudachirou, M. K. (2017, Juillet). CLASSIFICATION ET FORÊTS ALÉATOIRES: APPLICATION À L'AIDE À LA DÉCISION CHIRURGICALE DU GENOU PAR ARTHROPLASTIE. *mémoire de maîtrise en technologie de l'information*. Québec, Canada: Télé-université.

NHS. (2019). *High blood pressure (hypertension)*. Consulté le 06 09, 2021, sur Le Service national de santé britanique (NHS): https://www.nhs.uk/conditions/high-blood-pressure-hypertension/causes/#:~:text=Known%20causes%20of%20high%20blood%20pressure&text=kidney%20disease,during%20sleep%2C%20interrupting%20normal%20breathing

Reed T. Sutton, D. P. (2020, 2 6). An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*(17). Consulté le 06 07, 2021, sur https://www.nature.com/articles/s41746-020-0221-y

Rezzaki, D. (2019, 05 18). Plus de 8 millions de personnes souffrant d'hypertension artérielle en Algérie. *eldjazaironline*. Consulté le 06 08, 2021, sur http://eldjazaironline.dz/

S Kulkarni, I. O. (1998). Stress and hypertension. *Medical College of Wisconsin*. Récupéré sur https://pubmed.ncbi.nlm.nih.gov/9894438/

Solomon, S. M. (2015, 08 16). Influence of Physical Activity on Hypertension and Cardiac Structure and Function. *Current Hypertension Reports*. doi:10.1007/s11906-015-0588-3

Teytaud, F. (2020, août 25). *coursApprentissage.* Consulté le 06 06, 2021, sur site officiel de l'Université du Littoral Cote d'Opale: https://www-lisic.univ-littoral.fr/~teytaud/files/Cours/Apprentissage/coursApprentissage.pdf

Touzet, C. (1992). *Les réseaux de neurones artificiels, introduction au connexionnisme.* HAL.

Valle, A.-C. D. (2021, 02 03). Tension artérielle : normale, élevée, basse, mesure, âge. *journal des femmes*. Consulté le 06 08, 2021, sur https://sante.journaldesfemmes.fr/fiches-maladies/2488800-tension-arterielle-definition-normale-basse-haute-diastolique-systolique-age-mesure-tableau/

w3schools. (s.d.). Récupéré sur https://www.w3schools.com/python/python_intro.asp

who. (2021, May 17). *Hypertension.* Consulté le 06 08, 2021, sur world health organization: https://www.who.int/news-room/fact-sheets/detail/hypertension#:~:text=An%20estimated%201.13%20billion%20people,cause%20of%20premature%20death%20worldwide.