

République Algérienne Démocratique et Populaire  
وزارة التعليم العالي والبحث العلمي  
Ministère de l'enseignement supérieur et de la recherche scientifique  
جامعة برج بوعريريج  
Université de Bordj Bou Arréridj

Faculté de Mathématiques et d'Informatique

Département Informatique

**MEMOIRE**

**Master Informatique**

*Spécialité : Technologies de l'Information et de la Communication*



**Thème :**

**Application pour la classification et la migration entre  
classes de la maladie cardiaque**

Présent par :

- **Ghebouli Farida**
- **Lakroum Amel**

Devant le jury composé :

Présidente Mme : **Hammoudi Sara** MCA à L'U.EI Bachir El Ibrahimi-bba  
Examinatrice Mme : **Zouaoui Hakima** MCA à L'U.EI Bachir El Ibrahimi-bba

**Promotion : 2020/2021**



République Algérienne Démocratique et Populaire  
وزارة التعليم العالي والبحث العلمي  
Ministère de l'enseignement supérieur et de la recherche scientifique  
جامعة برج بوعريريج  
Université de Bordj Bou Arréridj

Faculté de Mathématiques et d'Informatique

Département Informatique

**MEMOIRE**

**Master Informatique**

*Spécialité : Technologies de l'Information et de la Communication*



**Thème :**

**Application pour la classification et la migration entre  
classes de la maladie cardiaque**

Présent par :

- **Ghebouli Farida**
- **Lakroum Amel**

Devant le jury composé :

Présidente Mme : **Hammoudi Sara** MCA à L'U.EI Bachir El Ibrahimi-bba  
Examinatrice Mme : **Zouaoui Hakima** MCA à L'U.EI Bachir El Ibrahimi-bba

**Promotion : 2020/2021**

## *Remerciement*

*En premier lieu, nous remercions DIEU le tout puissant de nous avoir donné la volonté, le courage et la patience pour réaliser ce*

*travail.*

*Nous adressons nos sincères remerciements à Mme BOUTOUHAMI Sara notre encadreur pour son suivi, son aide et ses précieux conseils et ce malgré les contraintes de la situation sanitaire elle n'a pas cessé de déployer les efforts nécessaire pour l'accomplissement de la réussite de ce travail.*



# *Dédicace*

*C'est avec une grande gratitude et des mots sincères, que  
Je dédie ce modeste travail à mes parents qui sont pour moi le  
symbole de la bonté par excellence et la source de la tendresse.*

*Rien au monde ne vaut les efforts fournis jour et nuit  
pour mon éducation et mon bien être. Hommage à mon père  
(rabi yarahmou) et que dieu le tout puissant préserve et accorde  
santé, bonheur et longue vie à ma mère mon frère et mes sœurs,  
ainsi qu'à toute ma belle-famille. A celui que j'aime beaucoup  
et qui m'a soutenue tout au long de ce projet, présent tous les  
moments de ma vie mon mari Hakim et ma princesse mon cœur et  
ma puce ma fille Racha que Dieu la béni et la protège.  
Merci à ma collègue Farid avec qui j'ai partagé les meilleurs  
moments durant la réalisation de ce travail.*



*Amel*

## *Table des matières*

<b>Liste des tableaux</b> .....	
<b>Liste des figures</b> .....	
<b>INTRODUCTION GÉNÉRALE</b>	
I. Contexte et problématique .....	1
II. Objectif et contribution.....	1
III. Plan de mémoire .....	3
<b>Chapitre 1 : État de l'art</b>	
1.1. Introduction .....	4
1.2. Définition de la fouille de données .....	4
1.3. Classification des méthodes de fouille de données .....	6
1.4. La classification supervisée .....	8
1.4.1 La méthode de classification « K plus proches voisins K-PP ».....	9
1.4.1.1 Les mesures de similarités .....	11
1.4.1.2 Comment choisir la valeur K ?.....	13
1.4.1.3 Avantages de K-NN .....	13
1.4.1.4 Inconvénients de K-NN .....	13
1.4.2. La méthode de classification «Arbre de décision » .....	14
1.4.2.1 Comment faire un arbre de décision ?.....	15
1.4.2.2 Classification .....	16
1.4.2.3 Choix de la bonne taille de l'arbre.....	17
1.4.2.4 Avantages et inconvénients .....	17
1.5. Conclusion .....	18

## **Chapitre 2 : Architecture et modélisation**

2.1. Introduction .....	19
2.2. Partie 1 : La base de données médicale cardiovasculaire .....	21
2.3. Partie 2 : Application des méthodes de classification sur des maladies cardiaque.....	26
2.3.1 Application de la méthode KPP .....	27
2.3.1.2 Classification du nouvel patient P par le KPP.....	27
2.3.2 Application de la méthode des Arbres de Décision .....	29
2.3.1.2 Classification du nouvel patient P par AD .....	29
2.3.1.2.1 Principe de la construction de l'arbre de décision .....	29
2.3.1.2.2 Prédiction par Arbre de décision du nouvel patient P .....	31
2. 4. Calcul des valeurs pour la Migration entre classes .....	33
2.4.1 Migration entre classes en utilisant le KNN .....	34
2.4.1.1 Exemple de Migration par KNN .....	36
2.4.2 Migration entre classes en utilisant l'AD .....	38
2.4.2.1 Exemple de Migration par AD .....	40
2.5. Conclusion : .....	43

## **Chapitre 3 : Implémentation et Bilan**

3.1. Introduction : .....	44
3.2. Evaluation des résultats .....	44
3.2.1 Critères et mesures d'évaluation .....	44
3.2.2 Le classifieur KPP.....	48
3.2.2.1 Choix de la valeur de K .....	48

3.2.2.2 Matrices de confusion du classifieur KPP.....	49
3.2.3 Le classifieur AD .....	51
3.2.3.1 Choix de la bonne taille de l'arbre .....	51
3.2.3.2 Matrices de confusion du classifieur AD .....	52
3.2.4 La migration entre classes .....	54
3.2.4.1 La migration entre classes par KPP .....	54
3.2.4.2 La migration entre classes par AD .....	55
3.3. Outils et langage utilises .....	58
3.3.1. Les outils .....	58
3.3.2. Les langages .....	59
3.4. Présentation de l'application .....	60
3.5. Conclusion .....	61
<b>Conclusion générale et perspectives</b>	
Conclusion générale.....	62
Perspectives .....	63



### *Liste des tableaux*

<b>Tableau</b>	<b>page</b>
<b>Tableau 1.1 : Quelques méthodes de fouille de données</b>	<b>07</b>
<b>Tableau 2.1 : Informations d'attributs de la base de données.</b>	<b>25</b>
<b>Tableau 2.2 : Echantillon de la base de données</b>	<b>26</b>
<b>Tableau 2.3 : Les données d'un nouvel patient</b>	<b>26</b>
<b>Tableau 2.4 : Normalisation de l'échantillon1 de la base médicale.</b>	<b>28</b>
<b>Tableau 2.5 : Normalisation des caractéristiques du patient p</b>	<b>28</b>
<b>Tableau 2.6: les distances des instances de l'échantillon par rapport au nouvel patient.</b>	<b>28</b>
<b>Tableau 2.7 : La forme textuelle de l'arbre de décision (base heart modifiée)</b>	<b>31</b>
<b>Tableau 2.8 : Les données d'un nouvel patient</b>	<b>31</b>
<b>Tableau 2.9 : La classification du nouvel patient par AD</b>	<b>32</b>
<b>Tableau 2.10 : Les données d'un patient qui souhaite faire des migrations</b>	<b>36</b>
<b>Tableau 2.10 : Table des chemins de l'arbre de décision (base heart modifiée)</b>	<b>38</b>
<b>Tableau 2.11 : Table des chemins de l'arbre de décision (base heart originale)</b>	<b>39</b>
<b>Tableau 2.12 : Les données d'un patient qui souhaite faire des migrations</b>	<b>40</b>
<b>Tableau 3.1 : Matrice de Confusion</b>	<b>45</b>
<b>Tableau 3.2 : Les différents critères d'évaluation d'un modèle de classification</b>	<b>46</b>
<b>Tableau 3.3: La matrice de confusion de notre modèles</b>	<b>47</b>
<b>Tableau 3.4 : Choix de la valeur de K pour le classifieur KPP</b>	<b>49</b>
<b>Tableau 3.5 : Matrice de confusion de « KPP» base heart originale</b>	<b>49</b>

<b>Tableau 3.6 : Rapport de classification pour l’algorithme « KPP » base heart originale</b>	<b>50</b>
<b>Tableau 3.7 : Matrice de confusion de « KPP» base heart modifiée</b>	<b>50</b>
<b>Tableau 3.8 : Rapport de classification pour l’algorithme « KPP » base heart originale</b>	<b>50</b>
<b>Tableau 3.9 : Choix de la bonne taille de l’AD</b>	<b>52</b>
<b>Tableau 3.10 : Matrice de confusion de « AD» base heart originale</b>	<b>52</b>
<b>Tableau 3.11 : Rapport de classification pour l’algorithme « AD » base heart originale</b>	<b>53</b>
<b>Tableau 3.12 : Matrice de confusion de « AD» base heart modifiée</b>	<b>53</b>
<b>Tableau 3.13 : Rapport de classification pour l’algorithme « AD » base heart originale</b>	<b>53</b>
<b>Tableau 3.14 : Pourcentage des possibilités de migration en utilisant KPP pour la base heart originale</b>	<b>55</b>
<b>Tableau 3.15 : Pourcentage des possibilités de migration en utilisant KPP pour la base heart modifiée</b>	<b>55</b>
<b>Tableau 3.16 : Pourcentage des possibilités de migration en utilisant AD pour la base heart originale</b>	<b>55</b>
<b>Tableau 3.17 : Pourcentage des possibilités de migration en utilisant AD pour la base heart modifiée</b>	<b>56</b>
<b>Tableau 3.18 : Pourcentage des possibilités de migration en utilisant AD pour la base heart originale</b>	<b>56</b>
<b>Tableau 3.19 : Pourcentage des possibilités de migration en utilisant AD pour la base heart modifiée</b>	<b>56</b>
<b>Tableau 3.20 : Pourcentage des possibilités de migration en utilisant AD pour la base heart originale</b>	<b>57</b>
<b>Tableau 3.19 : Pourcentage des possibilités de migration en utilisant AD pour la base heart modifiée</b>	<b>57</b>

### *Liste des figures*

<b>Figure</b>	<b>page</b>
<b>Figure 1.1 : Classification par KPP</b>	<b>10</b>
<b>Figure 1.2 : Modèle arbre de décision.</b>	<b>15</b>
<b>Figure 2.1 : Schéma global de notre architecture</b>	<b>20</b>
<b>Figure 2.2 : Arbre de décision</b>	<b>30</b>
<b>Figure 2.3 : Schéma explicatif de la prédiction</b>	<b>33</b>
<b>Figure 2.4 : Schéma explicatif de la migration entre classe</b>	<b>34</b>
<b>Figure 2.5 : Méta-algorithme de la migration entre classe pour KNN</b>	<b>35</b>
<b>Figure 2.6 : Méta-algorithme de la migration entre classe pour AD</b>	<b>40</b>
<b>Figure 3.1 : Formulaire de test des maladies cardiaques.</b>	<b>60</b>

# INTRODUCTION

# Introduction

---

## I. Contexte et problématique

Une maladie chronique est une maladie de longue durée, évolutive, avec un retentissement sur la vie quotidienne. Elle peut avoir une incidence sur la capacité d'une personne à effectuer ses activités quotidiennes et sur sa qualité de vie. Elle peut générer des incapacités, voire des complications graves. Toute personne risque d'être atteinte d'une maladie chronique, mais il existe des façons pour réduire les risques grâce à l'identification des facteurs de risques et des facteurs d'aggravations.

Les maladies cardiaques sont l'une des principales causes de morbidité et de mortalité au sein de la population mondiale. La prédiction des maladies cardiovasculaires est considérée comme l'un des sujets les plus importants dans la section de l'analyse des données cliniques.

Depuis des dizaines d'année les enseignes de santé ont opté pour la collecte et le stockage de données médicales des patients sur de longues périodes. Ces données représentent d'immense nombre de facteurs et d'indices sur les malades ainsi que leurs environnements. L'objectif de ces enseignes est d'appliquer les techniques d'apprentissage afin de pouvoir exploiter ces collections pour trouver des relations ou des corrélations entre les symptômes, les maladies et leurs traitements pour plusieurs raisons : identification de facteurs de risque de maladie, aide au diagnostic, prévention de maladie, au choix et au suivi de l'efficacité des traitements, surveillance épidémiologique,...

De nombreuses maladies chroniques peuvent être contrôlées ou leurs effets diminués par la prévention ou la gestion des facteurs de risque. Identifier les facteurs qui peuvent influencer sur l'évolution de la maladie et connaître le rôle qu'ils jouent dans la maladie peut aider à en tenir compte et à être vigilant. L'apprentissage automatique s'avère efficace pour aider à prendre des décisions et des prévisions à partir de la grande quantité de données produites par l'industrie des soins de santé.

## II. Objectif et contribution

La *prévention*, la *prédiction* et la gestion des maladies chroniques est une démarche en matière de soins de santé qui vise à aider les personnes qui en sont atteintes à maintenir leur autonomie et à demeurer en aussi bonne santé que possible grâce à la détection précoce de ces maladies ainsi qu'à leur prévention et à leur gestion.

## Introduction

---

La problématique que nous abordons dans notre travail est dans un premier temps l'application de techniques d'apprentissage automatique pour la *prédiction* de la maladie cardiaque. Les techniques que nous allons utiliser sont : le Plus Proche voisin (KPP) et les Arbres de Décision (AD). Nous avons utilisé les données médicales « Cleveland Heart Disease dataset from the UCI Repository ». C'est une base de données sur les maladies cardiovasculaire que nous avons utilisé pour la construction de trois modèles d'apprentissages capables de prédire (classer) si une personne souffre d'une maladie cardiaque (classée en 4 types) ou non.

Comme mentionnée plus haut, la base de données que nous avons utilisée permet de prédire si une personne souffre d'une maladie cardiovasculaire de type1, typ2, type3 ou type4 ou bien si elle est en bonne santé.

Cette variété de type de maladie nous à pousser à réfléchir à la possibilité de migration (passage) d'une classe à une autre. Autrement dit, avec les informations réelles que nous disposons d'une personne donnée notre système permet de prédire le type de sa maladie cardiovasculaire dont elle souffre : par exemple de type3. Généralement les gens ont un autre souci est de savoir quels changements doivent-ils faire pour être ou ne pas être dans une autre classe (type de maladie). Cette migration (passage) entre classe peut être vue comme une perspective d'amélioration de l'état d'un malade ou malheureusement par une dégradation (complication) de son état de santé. L'objectif principal est toujours ralentissement de l'évolution vers une forme invalidante de la maladie. L'idée est donc *d'intervenir* le plus tôt possible et de manière soutenue afin de prévenir l'aggravation et la sévérité de la maladie.

Pour pouvoir effectuer la migration entre classes, nous avons proposé un algorithme de calcul de nouveaux paramètres pour la techniques KPP, basé sur la recherche parmi les plus proches voisins, les plus proches voisins satisfaisant le passage vers classe souhaitée en indiquant les paramètres que le patient souhaite de maintenir ou qu'il accepté leur modification.

Pour la technique des arbres de décision, nous avons proposé de sauvegarder l'arbre de décision (tous les chemins) dans une table et d'effectuer des requêtes pour satisfaire la migration vers la classe souhaitée sous les contraintes imposées par le patient à savoir maintenir les valeurs initiales ou accepter leurs modifications.

# Introduction

---

Nous avons opté pour la réalisation d'une application web, ce qui est très pratique pour les patients qui souhaitent suivre leurs états de santé ainsi que de voir l'impact des changements des paramètres aussi minimaux qu'ils soient sur leur état de santé.

## III. Plan du mémoire

Le mémoire est organisé comme suit :

Le chapitre 1 présente brièvement le processus de fouille de données (Data Mining DM), les méthodes et les techniques utilisées dans ce domaine. Nous n'allons présenter que les techniques d'apprentissages supervisés qui nous intéressent à savoir : le Plus Proche voisin (KPP) et les Arbres de Décision (AD).

Le chapitre 2 est organisé en trois parties. Dans la première partie nous présentons la base de données cardiovasculaire que nous avons utilisée. Dans la deuxième partie, nous entrons dans le vif de notre sujet : la construction du modèle de classification pour chacune des techniques citées dans le précédent chapitre ainsi que leur application en vue de la prédiction du type de la maladie cardiovasculaire des nouveaux patients. Dans la troisième partie nous présentons le principe de la migration entre classes et nous présentons les algorithmes que nous avons proposés pour chaque technique. Nous terminons le chapitre par une discussion des résultats obtenus.

Dans le chapitre 3, nous présentons les différents outils et langages de programmation que nous avons utilisés pour réaliser notre application web ainsi que ces différentes interfaces.

Finalement, nous terminons le mémoire par une conclusion générale qui résume les contributions de notre travail ainsi que certaines perspectives futures.

I-

# ÉTAT DE L'ART



# Chapitre1 : État de l'art

---

## 1. Introduction

Ce chapitre a pour objectif de présenter le processus de l'extraction de connaissances à partir des données, la fouille de données, ses tâches et plus précisément la classification supervisée. La fouille de données désigne en réalité un ensemble de traitements différents menant à la découverte de connaissances variées. Ces traitements sont la classification, la segmentation et les règles d'association.

La classification est une tâche centrale de l'étape de fouille de données dans le processus d'extraction de connaissances à partir de données. Elle consiste à construire un classifieur à partir d'un ensemble d'exemples étiquetés par leur classe (phase d'apprentissage) et ensuite à prédire la classe de nouveaux exemples avec le classifieur (phase de classement) [1].

## 2. Définition de la fouille de données

La fouille de données est un domaine qui est apparu avec l'explosion des quantités d'informations stockées avec le progrès important des vitesses de traitement et des supports de stockage. La fouille de données vise à découvrir, dans les grandes quantités de données, les informations précieuses qui peuvent aider à comprendre les données ou à prédire le comportement des données futures. Le datamining utilise depuis son apparition plusieurs outils de statistiques et d'intelligence artificielle pour atteindre ses objectifs. La fouille de données s'intègre dans le processus d'extraction des connaissances à partir des données ECD ou (KDD : Knowledge Discovery from Data en anglais). Ce domaine en pleine expansion est souvent appelé le data mining [2].

La fouille de données est souvent définie comme étant le processus de découverte des nouvelles connaissances en examinant de larges quantités de données (stockées dans des entrepôts) en utilisant les technologies de reconnaissance de formes de même que les techniques statistiques et mathématiques. Ces connaissances, qu'on ignore au début, peuvent être des corrélations, des patterns ou des tendances générales de ces données. La science et l'ingénierie modernes sont basées sur l'idée d'analyser les problèmes pour comprendre leurs principes et leur développer les modèles mathématiques adéquats. Les données expérimentales sont utilisées par la suite pour vérifier la correction du système ou l'estimation de quelques paramètres difficiles à la modélisation

## Chapitre1 : État de l'art

---

mathématiques. Cependant, dans la majorité des cas, les systèmes n'ont pas de principes compris ou qui sont trop complexes pour la modélisation mathématique. Avec le développement des ordinateurs, on a pu rassembler une très grande quantité de données à propos de ces systèmes. La fouille de données vise à exploiter ces données pour extraire des modèles en estimant les relations entre les variables (entrées et sorties) de ses systèmes. En effet, chaque jour nos banques, nos hôpitaux, nos institutions scientifiques, nos magasins, ... produisent et enregistrent des milliards et des milliards de données. La fouille de données représente tout le processus utilisant les techniques informatiques (y compris les plus récentes) pour extraire les connaissances utiles dans ces données. Actuellement, La fouille de données utilise divers outils manuels et automatiques : on commence par la description des données, résumer leurs attributs statistiques (moyennes, variances, covariance,...), les visualiser en utilisant les courbes, les graphes, les diagrammes, et enfin rechercher les liens significatifs potentiels entre les variables (tel que les valeurs qui se répètent ensemble). Mais la description des données toute seule ne fournit pas un plan d'action. On doit bâtir un modèle de prédiction basé sur les informations découvertes, puis tester ce modèle sur des données autres que celles originales. La fouille de données a aujourd'hui une grande importance économique du fait qu'elle permet d'optimiser la gestion des ressources (humaines et matérielles) [3]. Elle est utilisée par exemple dans :

- organisme de crédit : pour décider d'accorder ou non un crédit en fonction du profil du demandeur de crédit, de sa demande, et des expériences passées de prêts ;
- optimisation du nombre de places dans les avions, hôtels, ... ) surréservation
- organisation des rayonnages dans les supermarchés en regroupant les produits qui sont généralement achetés ensemble (pour que les clients n'oublient pas bêtement d'acheter un produit parce qu'il est situé à l'autre bout du magasin). Par exemple, on extraira une règle du genre : "les clients qui achètent le produit X en fin de semaine, pendant l'été, achètent généralement également le produit Y";
- organisation de campagne de publicité, promotions, ... (ciblage des offres)
- diagnostic médical : "les patients ayant tels et tels symptômes et demeurant dans des agglomérations de plus de 104 habitants développent couramment telle pathologie";

## Chapitre1 : État de l'art

---

- analyse du génome
- classification d'objets (astronomie, ...)
- commerce électronique
- analyser les pratiques et stratégies commerciales et leurs impacts sur les ventes
- moteur de recherche sur internet : fouille du web
- extraction d'information depuis des textes : fouille de textes
- évolution dans le temps de données : fouille de séquences [5].

### 3. Classification des méthodes de fouille de données

Les méthodes de la fouille de données peuvent être classées selon plusieurs paramètres.

#### A. Classification selon le type d'apprentissage

- **Fouille supervisée** : comprenant à la fois des données d'entrée et de sortie dont le but est de classer correctement un nouvel exemple.
- **Fouille non supervisée**: c'est un processus dans lequel l'apprenant reçoit des exemples d'apprentissage (pas notion de classe) ne comprenant que des données d'entrées dont le but est de regrouper les exemples en segment (cluster) d'exemples similaires.

#### B. Classification selon les objectifs

- Classification : elle permet de prédire si une instance de donnée est membre d'un groupe ou d'une classe prédéfinie.
- Segmentation : elle consiste à partitionner logiquement la base de données en clusters (former des groupes homogènes à l'intérieur d'une population).
- Règles d'associations : elle consiste à déterminer les corrélations (ou relations) entre les attributs.

Des exemples des méthodes de fouille de données classées selon cette organisation sont présentes dans le Tableau 1.1

## Chapitre 1 : État de l'art

---

	supervisées	Non supervisées
Classification	<ul style="list-style-type: none"><li>• Arbre de décision</li><li>• Réseaux de Neurones avec perceptron</li><li>• Modèles/Réseau Bayésiens</li><li>• Machines à vecteur support « SVM »</li></ul>	<ul style="list-style-type: none"><li>• K plus proche voisin (ppv – raisonnement à partir de ce cas)</li><li>• Règles temporelles</li><li>Reconnaissance des formes</li></ul>
Segmentation		<ul style="list-style-type: none"><li>• K-means</li><li>• K plus proche voisin (ppv – raisonnement à partir de ce cas)</li><li>• Réseaux de Neurones avec cartes de Kohonen</li></ul>
Association		<ul style="list-style-type: none"><li>• Règles d'association</li></ul>

Tableau 1.1 : Quelques méthodes de fouille de données

### C. Classification selon la méthode

Méthodes prédictives et méthodes descriptives [6].

- Les méthodes **descriptives** : elles permettent de décrire la situation actuelle, elles caractérisent les propriétés générales des données dans la base de données et mettent l'accent sur la compréhension et l'interprétation de ces dernières.
- Les méthodes **prédictives** : qui, en apprenant sur le passé, simulent le futur. Elles utilisent les données avec des résultats connus pour développer des modèles permettant de prédire les valeurs des autres données.

### 4. La classification supervisée

La classification supervisée est une tâche de fouille de données qui utilise la connaissance à priori sur l'appartenance d'un exemple à une classe. Elle permet d'apprendre à l'aide d'un ensemble d'entraînement, une procédure de classification qui permet de prédire l'appartenance d'un nouvel exemple à une classe [7].

## Chapitre1 : État de l'art

---

Dans la classification supervisée, le nombre ainsi que l'identité des classes sont connus à l'avance. La classification a donc pour objectif d'identifier les classes auxquelles appartiennent des objets à partir de traits descriptifs.

Le fonctionnement de la classification supervisée se décompose en deux points :

1. Le premier est la phase d'apprentissage, tout ce qui est appris par l'algorithme est représenté sous la forme des règles de classification que l'on appelle le modèle d'apprentissage.
2. Le second point est la phase de la classification proprement dite, dans laquelle les données tests vont être utilisées pour estimer la précision des règles de classification générées pendant la première phase. Si la précision du modèle est considérée comme acceptable, les règles de classification peuvent être appliquées à des nouvelles données.

La construction d'un modèle prédictif se fait généralement en trois phases :

- Une phase d'entraînement : On utilise l'échantillon d'entraînement pour créer le modèle.
- Une phase de validation : On utilise l'échantillon de validation pour évaluer la performance du modèle sur des données qui n'ont pas servi à l'entraînement, de façon à éviter le sur-apprentissage. La performance du modèle peut se baser sur différents indicateurs.
- Une phase de test : On utilise l'échantillon de test pour évaluer la performance finale du modèle. Elle est utile lorsque l'on souhaite une évaluation rigoureuse de la performance finale du modèle.

Généralement on sépare aléatoirement l'échantillon en trois (un échantillon pour chaque phase).

L'échantillon d'entraînement comprend généralement entre 50% et 80% des données. L'échantillon de validation comprend entre 20% et 40% des données et l'échantillon de test utilise entre 5% et 10% des données (il est fréquent, en pratique, que cette étape soit omise).

Il existe de nombreuses méthodes de classification supervisée :

- K plus proches voisins « KPP »

## Chapitre1 : État de l'art

---

- Analyse (factorielle) discriminante
- Régression logistique « RL »
- Arbres de décision « AD »
- Forêts aléatoires « RF »
- Réseaux de neurones « RN »
- Support Vector machines « SVM »
- Réseau bayésien « RB »

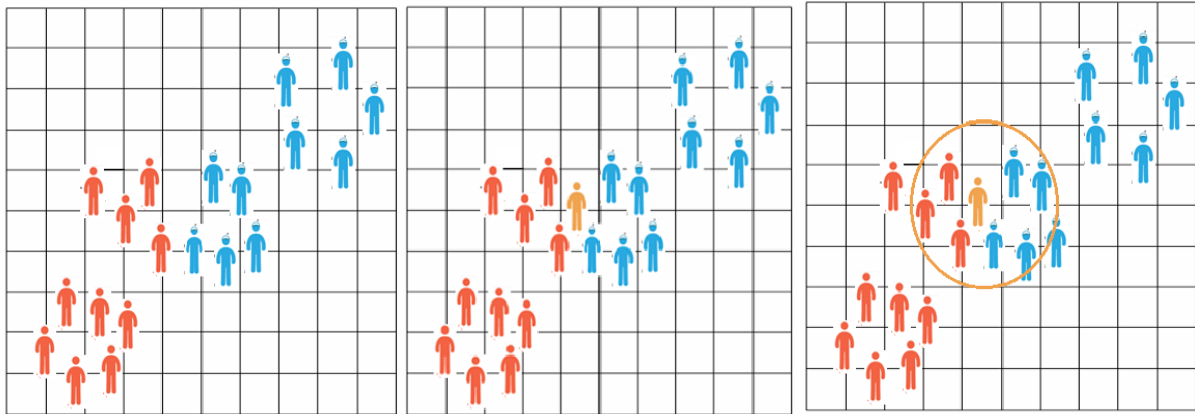
### **4.1. La méthode de classification « K plus proches voisins K-PP » :**

Cette méthode est basée sur l'algorithme K plus proche voisin (K-Nearest Neighbors K-NN), c'est l'un des algorithmes de classification les plus simples et les plus utilisés. K-NN est un algorithme d'apprentissage paresseux. Pour effectuer une prédiction, l'algorithme K-NN ne va pas calculer un modèle prédictif à partir d'un échantillon d'apprentissage. Pour K-NN il n'existe pas de phase d'apprentissage proprement dite.

Pour pouvoir effectuer une prédiction, K-NN se base sur la base de données pour produire un résultat. Principe de K-NN : dis-moi qui sont tes voisins, je te dirais qui tu es ! K est le nombre de voisin à vérifier. L'algorithme K-NN se base sur la base de données en entier. Pour une observation, qui ne fait pas parti de la base de données, qu'on souhaite prédire, l'algorithme va chercher les K instances du jeu de données les plus proches (selon une mesure de similarité) de notre observation. Ensuite pour ces voisins, l'algorithme se basera sur leurs variables de sortie (output variable) pour calculer la valeur de la variable de l'observation qu'on souhaite prédire.

## Chapitre 1 : État de l'art

---



**Figure 1.1 : Classification par KPP**

La figure précédente explique mieux le principe du k plus proche voisin. Dans cet exemple les instances sont classés en deux classes soit les hommes bleu ou bien les rouge .Si on souhaite classer une nouvelle instance l'homme en orange. L'idée est voir ces voisins (il ressemble à ses voisins). Nous allons voir l'ensemble des voisins (S'il parle avec plus des rouges que des bleus, on pourra supposer avec un certain degré de certitude qu'il est rouge. Au contraire, s'il converse avec plusieurs bleus et peu des rouges, on peut assumer qu'il est bleu. dont la distance qui les sépare est égale à k, la classe du nouvel élément est la classe majoritaire de ses voisins.

Dans cet exemple, on choisit  $K = 7$  voisins proches du sujet mystérieux. Si on trace un cercle autour de lui, on peut voir que de ces 7 individus, il y a trois hommes rouges et quatre bleus. Donc, on peut assumer que la probabilité selon laquelle l'invité mystérieux est rouge est de  $3/7$  tandis que la probabilité qu'il soit bleu est de  $4/7$ .

On peut schématiser le fonctionnement de K-NN en l'écrivant en pseudo-code suivant :

## Chapitre1 : État de l'art

---

### Début Algorithme

Données en entrée :

- Un ensemble de données **D**.
- Une fonction de définition distance **d**.
- Un nombre entier **K**.

Pour une nouvelle observation **X** dont on veut prédire sa variable de sortie **y** Faire :

1. Calculer toutes les distances de cette observation **X** avec les autres observations du jeu de données **D**
2. Retenir les **K** observations du jeu de données **D** les proches de **X** en utilisant la fonction de calcul de distance **d**
3. Prendre les valeurs de **y** des **K** observations retenues :
  - Si on effectue une régression, calculer la moyenne (ou la médiane) de **y** retenues
  - Si on effectue une classification, calculer le mode de **y** retenues
4. Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par KNN pour l'observation **X**.

### Fin Algorithme

### 4.1.1 Les mesures de similarités

Comme on vient de le voir l'algorithme K-NN a besoin d'une fonction de calcul de distance entre deux observations. Il existe plusieurs fonctions de calcul de distance : la distance euclidienne, la distance de Manhattan, la distance de Minkowski, celle de Jaccard, la distance de Hamming...etc. La fonction de distance est choisie en fonction des types de données manipulées. Ainsi pour les données quantitatives (exemple : poids, salaires, taille, montant de panier électronique etc...) et du même type, la distance euclidienne est un bon candidat. Quant à la



## Chapitre1 : État de l'art

---

distance de Manhattan, elle est une bonne mesure à utiliser quand les données (input variables) ne sont pas du même type (exemple : Age, sexe, longueur, poids etc...) [8].

Il est inutile de coder, soi-même ces distances, généralement, les bibliothèques de Machine Learning comme Scikit Learn, effectue ces calculs en interne. Il suffit juste d'indiquer la mesure de distance qu'on souhaite utiliser.

Pour les curieux, voici les définitions mathématiques des distances qu'on vient d'évoquer.

### ➤ La distance euclidienne :

Distance qui calcule la racine carrée de la somme des différences carrées entre les coordonnées de deux points :

$$D_e(x, y) = \sqrt{\sum_{j=1}^n (x_j - y_j)^2}$$

### ➤ Distance Manhattan :

La distance de Manhattan: calcule la somme des valeurs absolues des différences entre les coordonnées de deux points :

$$D_m(x, y) = \sum_{i=1}^k |x_i - y_i|$$

### ➤ Distance Minkowski :

La distance entre deux points donnés est la différence maximale entre leurs coordonnées sur une dimension :

$$d(p, q) = \left( \sum_{i=1}^n |p_i - q_i|^c \right)^{\frac{1}{c}}$$

### ➤ Distance Hamming :

La distance entre deux points donnés est la différence maximale entre leurs coordonnées sur une dimension.

$$D_h(x, y) = \sum_{i=1}^k |x_i - y_i|$$

- Avec :  $x = y \Rightarrow D = 0$  /  $x \neq y \Rightarrow D = 1$

## Chapitre1 : État de l'art

---

Notez bien qu'il existe d'autres distances selon le cas d'utilisation de l'algorithme, mais la distance euclidienne reste la plus utilisée.

### 4.1.2 Comment choisir la valeur K ?

Le choix de la valeur K à utiliser pour effectuer une prédiction avec K-NN, varie en fonction du jeu de données. En règle générale, moins on utilisera de voisins (un nombre K petit) plus on sera sujette au problème du sous apprentissage (underfitting). Par ailleurs, plus on utilise de voisins (un nombre K grand) plus, sera fiable dans notre prédiction. Toutefois, si on utilise K nombre de voisins avec  $K=N$  et N étant le nombre d'observations, on risque d'avoir le problème de sur-apprentissage (overfitting) et par conséquent un modèle qui se généralise mal sur des observations qu'il n'a pas encore vu.

Pour sélectionner la valeur de K qui convient aux données, il faut exécuter plusieurs fois l'algorithme KPP avec différents valeur de K. Puis choisir le K qui réduit le nombre d'erreurs rencontrées tout en maintenant la capacité de l'algorithme à effectuer des prédictions avec précision lorsqu'il reçoit des données nouvelles

### 4.1.3 Avantages de K-NN

- Algorithme simple pour expliquer et comprendre / interpréter,
- Haute précision (relativement),
- L'algorithme est polyvalent. Il peut être utilisé pour la classification, la régression et la recherche d'informations.

### 4.1.4 Inconvénients de K-NN

- Stocke toutes (ou presque toutes) les données d'entraînement,
- L'étape de prédiction peut être lente (avec un grand N),
- Sensible aux fonctionnalités non pertinentes et à l'échelle des données,
- L'algorithme ralentit considérablement à mesure que le nombre d'observations et/ou de variables dépendantes/indépendantes augmente. En effet, l'algorithme parcourt l'ensemble des observations pour calculer chaque distance,

## Chapitre1 : État de l'art

---

- Pas efficace pour des bases des données larges,
- L'estimation de ce modèle devient de mauvaise qualité quand le nombre de variables explicatives est grand.

### 4.2.1 La méthode de classification «Arbre de décision »

Les arbres de décision représentent une méthode très efficace d'apprentissage supervisé. Il s'agit de partitionner un ensemble de données en des groupes les plus homogènes possible du point de vue de la variable à prédire. On prend en entrée un ensemble de données classées, et on fournit en sortie un arbre qui ressemble beaucoup à un diagramme d'orientation où chaque nœud final (feuille) représente une décision (une classe) et chaque nœud non-final (interne) représente un test. Chaque feuille représente la décision d'appartenance à une classe des données vérifiant tous les tests du chemin menant de la racine à cette feuille [9].

Un arbre de décision est un schéma représentant les résultats possibles d'une série de choix interconnectés. Il permet à une personne ou une organisation d'évaluer différentes actions possibles en fonction de leur coût, leur probabilité et leurs bénéfices. Il peut être utilisé pour alimenter une discussion informelle ou pour générer un algorithme qui détermine le meilleur choix de façon mathématique.

Un arbre de décision commence généralement par un nœud d'où découlent plusieurs résultats possibles. Chacun de ces résultats mène à d'autres nœuds, d'où émanent d'autres possibilités. Le schéma ainsi obtenu rappelle la forme d'un arbre.

### 4.2.1 Construction

Dans ces structures d'arbre, les feuilles représentent les valeurs de la variable-cible et les embranchements correspondent à des combinaisons de variables d'entrée qui mènent à ces valeurs.

Principe de la construction : Au départ, les points de la base d'apprentissage sont tous placés dans le nœud racine. Une des variables de description des points est la classe du point, cette variable est dite « variable cible ». La variable cible peut être catégorielle (problème de classement) ou valeur réelle (problème de régression). Chaque nœud est coupé (opération *split*) donnant naissance à plusieurs nœuds descendants. Un élément de la base d'apprentissage situé dans un

## Chapitre1 : État de l'art

---

noeud se retrouvera dans un seul de ses descendants. L'arbre est construit par partition récursive de chaque noeud en fonction de la valeur de l'attribut testé à chaque itération (*top-down induction*). Le processus s'arrête quand les éléments d'un noeud ont la même valeur pour la variable cible.

L'idée de construction de l'arbre de décision est simple : il faut commencer par diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis à l'aide des attributs jusqu'à ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples appartenant à une même classe. Cette idée débouche sur des méthodes de construction *Top-Down*, c'est-à-dire construisant l'arbre de la racine vers les feuilles récursives [10].

Dans toutes les méthodes, on trouve les trois opérateurs suivants :

- Décider si un noeud est terminal, c'est-à-dire décider si un noeud doit être étiqueté comme une feuille ou porter un test.
- Si un noeud n'est pas terminal, sélectionner un test à lui associer.
- Si un noeud est terminal, lui affecter une classe.

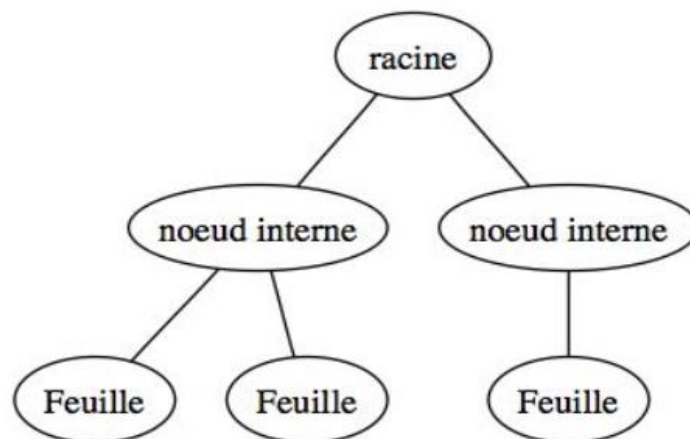


Figure1-2 : Modèle arbre de décision.

## Chapitre1 : État de l'art

---

Les mesures de sélection d'attributs :

1. Gain :

$$Gain(p, t) = i(p) - \sum_{j=1}^n P_j \cdot i(p_j)$$

2. Le critère de Gini :

$$Gini(p) = 1 - \sum_{k=1}^c (P(k|p))^2$$

3. L'entropie :

$$Entropie(Tr) = - \sum_{y \in \mathcal{Y}} P(Y = y) \log_2 P(Y = y)$$

La figure suivante illustre le fonctionnement de l'algorithme de l'arbre de décision :

**Début:**

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

**Répéter :**

Décider si le nœud courant est terminal.

*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.

*Fin Si*

Passer au nœud suivant non exploré s'il en existe

Jusqu'à obtenir un arbre de décision

**Fin.**

## Chapitre1 : État de l'art

---

Un arbre de décision est considéré comme optimal lorsqu'il représente la plus grande quantité de données possible avec un nombre minimal de niveaux ou de questions. Les algorithmes conçus pour créer des arbres de décision optimisés incluent notamment CART, ASSISTANT, CLS et ID3/4/5. Il est également possible de créer un arbre de décision en générant des règles d'associations, en plaçant la variable cible sur la droite.

Chaque méthode doit déterminer quelle est la meilleure façon de répartir les données à chaque niveau. Les méthodes courantes pour ce faire comprennent l'indice d'impureté de Gini, le gain d'information et la réduction de la variance.

### 4.2.3 Choix de la bonne taille de l'arbre

Il n'est pas toujours souhaitable en pratique de construire un arbre dont les feuilles correspondent à des sous-ensembles parfaitement homogènes du point de vue de la variable-cible. Plus le modèle est complexe (plus l'arbre est grand, plus il a de branches, plus il a de feuilles), plus l'on court le risque de voir ce modèle incapable d'être extrapolé à de nouvelles données, c'est-à-dire de rendre compte de la réalité que l'on cherche à appréhender.

### 4.2.4 Avantages et inconvénients des AD

La popularité des arbres de décision se justifie par les raisons suivantes :

- Ils sont faciles à comprendre.
- Ils peuvent être utiles avec ou sans données concrètes, et les données — quelles qu'elles soient — nécessitent une préparation minimale.
- De nouvelles options peuvent être ajoutées aux arbres existants.
- Ils permettent de sélectionner l'option la plus appropriée parmi plusieurs.
- Il est facile de les associer à d'autres outils de prise de décision.

L'utilisation des arbres de décision dans l'apprentissage automatique présente plusieurs avantages :

- Le coût d'utilisation de l'arbre pour prédire des données diminue à chaque point de donnée supplémentaire.
- Ils fonctionnent aussi bien pour les données de catégorie que numériques.
- La modélisation des problèmes est possible avec plusieurs données de sortie.

## Chapitre1 : État de l'art

---

- Ils utilisent un modèle de boîte blanche, ce qui rend les résultats faciles à expliquer.
- La fiabilité d'un arbre peut être testée et quantifiée.
- Ils tendent à être précis, même si les hypothèses des données source ne sont pas respectées.

Mais ils présentent aussi quelques inconvénients :

Les arbres de décision peuvent toutefois devenir extrêmement complexes. Dans ce cas, un diagramme d'influence, plus compact, peut représenter une bonne alternative. Les diagrammes d'influence se focalisent sur les décisions, données et objectifs critiques.

- Lors de la gestion de données de catégorie comportant plusieurs niveaux, le gain d'information est biaisé en faveur des attributs disposant du plus de niveaux.
- Les calculs peuvent devenir compliqués lorsqu'une certaine incertitude est de mise et que de nombreux résultats sont liés entre eux.
- Les conjonctions entre les nœuds sont limitées à l'opérateur « ET », alors que les graphiques décisionnels permettent de connecter des nœuds avec l'opérateur « OU » .

### 5. Conclusion

Les quantités de données collectées dans différents domaines deviennent de plus en plus importantes et de plus en plus exigeantes à analyser. Nous avons abordé dans ce chapitre le problème de fouille de données et des techniques utilisées pour l'extraction de connaissances. Nous nous sommes focalisé uniquement sur deux techniques d'apprentissage supervisée que nous allons utiliser dans notre application de prédiction dans le diagnostic médical à savoir : K plus proche voisins et les arbres de décision. Pour chacune de ces techniques nous avons essayé de présenter brièvement le principe de fonctionnement ainsi que quelques avantages et inconvénients. Le prochain suivant contiendra plus de détails sur l'application de ces techniques sur des données provenant d'une base médicale sur la maladie cardiovasculaire.

**II-**

# **Architecture et modélisation**



## Chapitre 2 : Architecture et modélisation

---

### 1. Introduction

L'apprentissage automatique s'avère efficace pour aider à prendre des décisions et des prévisions à partir de la grande quantité de données produites par l'industrie des soins de santé. Les maladies cardiaques sont l'une des principales causes de morbidité et de mortalité au sein de la population mondiale. La prédiction des maladies cardiovasculaires est considérée comme l'un des sujets les plus importants dans la section de l'analyse des données cliniques.

La problématique nous nous abordons dans notre travail est d'appliquer deux techniques d'apprentissage automatique à savoir : le plus proche voisin et les arbres de décision pour la construction de modèles capables de classer si une personne souffre d'une maladie cardiaque (classée en 4 types) ou non, en utilisant l'un des ensembles de données les plus utilisés - Cleveland Heart Disease dataset from the UCI Repository. La base de données que nous avons utilisée permet de prédire si une personne souffre d'une maladie cardiovasculaire de type1, typ2, type3 ou type4 ou bien si elle est en bonne santé. Cette variété de type de maladie nous à pousser à réfléchir à la possibilité de migration (passage) d'une classe à une autre. Nous avons proposé pour chaque technique un algorithme de calcul de nouvelle valeur des paramètres pour la migration d'une classe à une autre.

## Chapitre 2 : Architecture et modélisation

Le schéma suivant illustre de façon exhaustive notre travail.

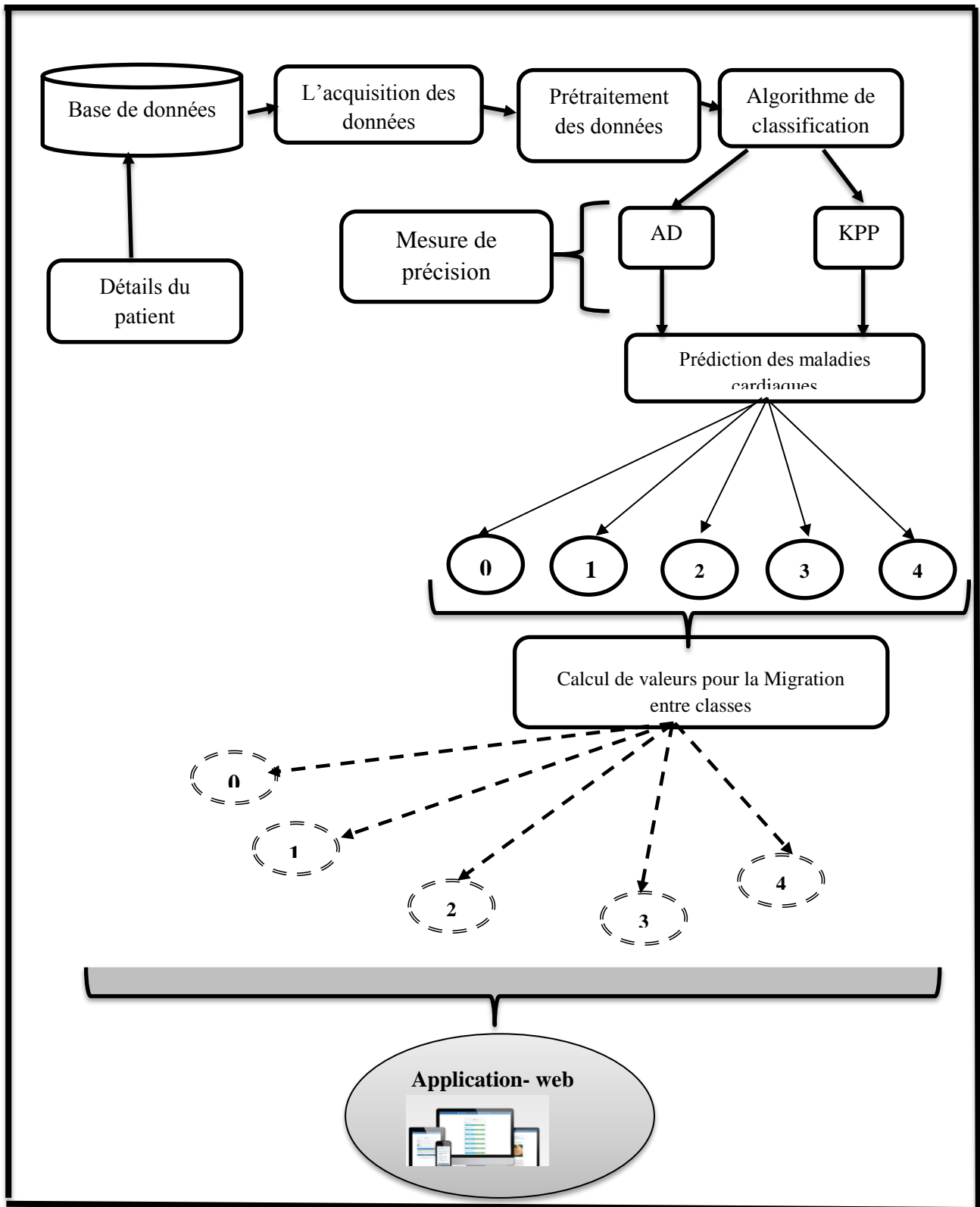


Figure 2.1 : Schéma global de notre architecture.

## Chapitre 2 : Architecture et modélisation

---

### 2. Partie 1 : La base de données médicale cardiovasculaire

Afin d'évaluer correctement notre système proposé nous avons utilisé une base de données provenant de véritables sources médicales. Nous avons opté pour le choix de cette base de données étant donné que le résultat de la classification n'est pas binaire mais plutôt cinq classes. Ce qui nous offre plus de choix et plus de perspectives futures pour notre travail.

Cette base de données contient 76 attributs mais toutes les expériences publiées font référence à l'utilisation d'un sous-ensemble de 14 d'entre eux. En particulier, cette base de données est la seule qui ait été utilisée par les chercheurs du Machine Learning (ML) à ce jour. Le champ "objectif" représente l'étiquette de classe, il fait référence à la présence d'une maladie cardiaque chez le patient. C'est un entier allant de 0 (pas de présence) à 4 [10, 11].

#### **Remarque très importante :**

Nous nous sommes confrontés à un problème lors de l'utilisation de notre base de données pour le calcul de valeurs pour la migration (passage) entre classe que le champ « objectif » ne contient que deux valeurs 0 et 1. Ils ne se sont pas intéressés au type de la maladie et ont confondu tout type de maladie en 1. Nous avons donc décidé d'appliquer notre travail sur deux variantes de la base de données Heart. La première version : est la version originale avec les vraies valeurs. La deuxième est une version dans laquelle modifiée la colonne « objectif » pour faire apparaître les valeurs entre 0 (pas de présence de la maladie) à 4. Cette modification va influencer drastiquement sur la première partie du travail de la prédiction. Mais ce qui nous intéresse est le principe de calcul des nouvelles valeurs des paramètres pour permettre la migration entre classe, il nous a semblé très intéressant d'avoir plusieurs classes. Nous sommes conscients que l'interprétation des résultats est faussée étant donné que les données utilisées sont faussées initialement.

## Chapitre 2 : Architecture et modélisation

---

Dans l'ensemble de données réel, nous avons 76 fonctionnalités mais pour notre étude, nous avons choisi uniquement les 14 ci-dessous :

**Âge:** l'âge est le facteur de risque le plus important dans le développement de maladies cardiovasculaires ou cardiaques, avec environ un triplement du risque à chaque décennie de la vie. Des stries graisseuses coronaires peuvent commencer à se former à l'adolescence. On estime que 82% des personnes décédées d'une maladie coronarienne ont 65 ans et plus. Simultanément, le risque d'AVC double chaque décennie après l'âge de 55 ans.

**Sexe:** les hommes sont plus à risque de maladie cardiaque que les femmes préménopausées. Une fois la ménopause passée, il a été avancé que le risque pour une femme était similaire à celui d'un homme, bien que des données plus récentes de l'OMS et des Nations Unies le contestent. Si une femme est diabétique, elle est plus susceptible de développer une maladie cardiaque qu'un homme diabétique [12].

**Angine** (douleur thoracique): l'angine de poitrine est une douleur thoracique ou une gêne provoquée lorsque votre muscle cardiaque ne reçoit pas suffisamment de sang riche en oxygène. Cela peut ressembler à une pression ou à une compression dans votre poitrine. L'inconfort peut également se produire dans vos épaules, vos bras, votre cou, votre mâchoire ou votre dos. La douleur de l'angine peut même ressembler à une indigestion [13].

**Pression artérielle au repos:** au fil du temps, l'hypertension artérielle peut endommager les artères qui alimentent votre cœur. L'hypertension artérielle qui se produit avec d'autres conditions, telles que l'obésité, l'hypercholestérolémie ou le diabète, augmente encore plus votre risque [14].

**Cholestérol sérique :** Un niveau élevé de cholestérol à lipoprotéines de basse densité (LDL) (le «mauvais» cholestérol) est le plus susceptible de rétrécir les artères. Un taux élevé de triglycérides, un type de graisse sanguine lié à votre alimentation, augmente également le risque de crise cardiaque. Cependant, un taux élevé de cholestérol à lipoprotéines de haute densité (HDL) (le «bon» cholestérol) réduit le risque de crise cardiaque [15].

**Glycémie à jeun :** ne pas produire suffisamment d'hormone sécrétée par votre pancréas (insuline) ou ne pas répondre correctement à l'insuline fait augmenter la glycémie de votre corps, augmentant ainsi le risque de crise cardiaque.

## Chapitre 2 : Architecture et modélisation

---

**ECG au repos:** pour les personnes à faible risque de maladie cardiovasculaire, l'USPSTF conclut avec une certitude modérée que les dommages potentiels du dépistage au repos ou à l'effort ECG égalent ou dépassent les avantages potentiels. Pour les personnes à risque intermédiaire à risque élevé, les preuves actuelles sont insuffisantes pour évaluer l'équilibre des avantages et des inconvénients du dépistage [16].

**Fréquence cardiaque maximale atteinte :** l'augmentation du risque cardiovasculaire, associée à l'accélération de la fréquence cardiaque, était comparable à l'augmentation du risque observée avec l'hypertension artérielle. Il a été démontré qu'une augmentation de la fréquence cardiaque de 10 battements par minute était associée à une augmentation du risque de décès cardiaque d'au moins 20%, et cette augmentation du risque est similaire à celle observée avec une augmentation du sang systolique pression de 10 mm Hg.

**Angine de poitrine induite par l'exercice :** la douleur ou l'inconfort associés à l'angine de poitrine est généralement resserré, saisissant ou serrant, et peut varier de légère à sévère. L'angine de poitrine est généralement ressentie au centre de votre poitrine, mais peut se propager à l'une ou aux deux épaules, ou à votre dos, votre cou, votre mâchoire ou votre bras. Cela peut même être ressenti dans vos mains. Types d'angine de poitrine a. Angine de poitrine stable / Angine de poitrine b. Angine instable v. Variante (Prinzmetal) Angine de poitrine d. Angine microvasculaire.

**Segment ST d'exercice maximal :** Un test d'effort ECG sur tapis roulant est considéré comme anormal lorsqu'il y a une dépression du segment ST horizontale ou en pente descendante  $\geq 1$  mm à 60–80 ms après le point J. Les ECG d'effort avec des dépressions du segment ST en pente ascendante sont généralement signalés comme un test «équivoque». En général, la survenue d'une dépression du segment ST horizontale ou en pente descendante à une charge de travail plus faible (calculée en MET) ou à une fréquence cardiaque indique un pronostic plus mauvais et une probabilité plus élevée de maladie multi-vaisseaux. La durée de la dépression du segment ST est également importante, car une récupération prolongée après un pic de stress est compatible avec un test de stress ECG sur tapis roulant positif. Une autre constatation qui indique fortement une CAD importante est la survenue d'une élévation du segment ST  $> 1$  mm (suggérant souvent une ischémie transmurale); ces patients sont fréquemment référés en urgence pour une angiographie coronarienne.

## Chapitre 2 : Architecture et modélisation

**Thal** : La thalassémie majeure est caractérisée par une érythropoïèse chronique inefficace et une anémie problèmes primaires. Ceux-ci, à leur tour, produisent des adaptations physiologiques dans le système cardiovasculaire ainsi que des processus pathologiques / iatrogènes tels que surcharge en fer, splénectomie, nutrition carences, stress oxydatif chronique et maladie pulmonaire. Cet article traite de la physiopathologie de la thalassémie en ce qui concerne le système cardiovasculaire, les mécanismes et surveillance de la cardiomyopathie ferrique, de l'hypertension pulmonaire et du vieillissement vasculaire patients thalassémiques. [17].

L'ensemble de données comprend 500 données individuelles. L'ensemble de données comprend 14 colonnes, décrites ci-dessous :

N° de paramètre	Abréviation	signification
1	Age	l'âge de l'individu.
2	Sexe	le sexe de la personne au format suivant: 1= mâle 0 = femme
3	Cp	Type de douleur thoracique ( <i>Chest-Pain Type</i> ): affiche le type de douleur thoracique ressentie par l'individu au format suivant: <ul style="list-style-type: none"><li>• 1 = angine de poitrine typique</li><li>• 2 = angine atypique</li><li>• 3 = douleur non angorale</li><li>• 4 = asymptotique</li></ul>
4	Trestbps	Pression artérielle au repos ( <i>Resting Blood Pressure</i> ): affiche la valeur de la pression artérielle au repos d'un individu en mmHg (unité)
5	Chol	Cholestrol sérique ( <i>Serum Cholestrol</i> ): affiche le cholestérol sérique en mg / dl (unité)
6	Fbs	Glycémie à jeun ( <i>Fasting Blood Sugar</i> ): compare la valeur de la glycémie à jeun d'un individu à 120 mg / dl. <ul style="list-style-type: none"><li>• 1 (vrai) = Si la glycémie à jeun &gt; 120 mg / dl</li><li>• 0 (faux) = sinon:</li></ul>
7	Restecg	ECG au repos ( <i>Resting ECG</i> ): affiche les résultats électrocardiographiques au repos <ul style="list-style-type: none"><li>• 0 = normal</li></ul>

## Chapitre 2 : Architecture et modélisation

		<ul style="list-style-type: none"> <li>• 1 = présentant une anomalie de l'onde STT</li> <li>• 2 = hypertrophie ventriculaire gauche</li> </ul>
8	Thalach	Fréquence cardiaque maximale atteinte ( <i>Max heart rate achieved</i> ): affiche la fréquence cardiaque maximale atteinte par un individu
9	Exang	Angine induite par l'exercice ( <i>Exercise induced angina</i> ): <ul style="list-style-type: none"> <li>• 1 = oui</li> <li>• 0 = non</li> </ul>
10	Oldpeak	Dépression ST induite par l'exercice par rapport au repos ( <i>ST depression induced by exercise relative to rest</i> ): affiche la valeur qui est un entier ou un flottant.
11	Slope	Segment ST d'exercice maximal ( <i>Peak exercise ST segment</i> ): <ul style="list-style-type: none"> <li>• 1 = ascendant</li> <li>• 2 = plat</li> <li>• 3 = descente</li> </ul>
12	Ca	Nombre de vaisseaux principaux (0–3) colorés par fluoroscopie ( <i>Number of major vessels (0–3) colored by flourosopy</i> ): affiche la valeur sous forme d'entier ou de flottant.
13	Thal	Thal: affiche la thalassémie: 3 = normal 6= défaut fixe 7 = défaut réversible
14	Num(pred_attribute)	Diagnostic des maladies cardiaques: Indique si la personne souffre de maladies cardiaques ou non: <ul style="list-style-type: none"> <li>• 0 = absence</li> <li>• 1, 2, 3, 4 = présent.</li> </ul>

**Tableau 2.1. Informations d'attributs de la base de données.**

### Types d'attributs :

- **Attributs catégoriels** (à deux catégories ou plus et chaque valeur de cette caractéristique peut être classés par eux): sexe, douleur thoracique(Angine).
- **Attributs ordinaux** (Variable ayant un ordre relatif ou un tri entre les valeurs): Glycémie à jeun, électrocardiographie, angor induit, pas de vaisseaux, thalassémie, diagnostic.

## Chapitre 2 : Architecture et modélisation

- **Attributs continus** (variable prenant des valeurs entre deux points quelconques ou entre le minimum ou valeurs maximales dans la colonne de caractéristiques): âge, tension artérielle, cholestérol sérique, max fréquence cardiaque, dépression ST.

### 3. Partie 2 : Application des méthodes de classification sur des maladies cardiaque

Les problèmes d'apprentissage sont énoncés sous forme de données. Ces séries caractérisant une série d'instances du phénomène à apprendre, que l'on nomme patients.

Chaque patient **P** est constitué d'une description **D** et d'une sortie **S**

**P = (D, S)**

Où :

- $D \in X = \{\text{age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, olbpeak, slope, ca, thal.}\}$
- $S \in Y = \{0, 1, 2, 3, 4\}$

Voici un échantillon de la base de d'apprentissage que nous avons utilisé

Age	sex	Cp	trestbps	chol	fbs	restecg	thalach	exang	olbpeak	Slope	ca	thal	Pred-attribute
63	1	1	145	233	1	2	150	0	2.3	3	0	6	<b>0</b>
67	1	4	160	286	0	2	108	1	1.5	2	3	3	<b>2</b>
67	1	4	120	229	0	2	129	1	2.6	2	2	7	<b>1</b>
37	1	3	130	250	0	0	187	0	3.5	3	0	3	<b>0</b>
41	0	2	130	204	0	2	172	0	1.4	1	0	3	<b>0</b>
56	1	2	120	236	0	0	178	0	0.8	1	0	3	<b>0</b>
62	0	4	140	268	0	2	160	0	3.6	3	2	3	<b>3</b>
57	0	4	120	354	0	0	163	1	0.6	1	0	3	<b>0</b>
63	1	4	130	254	0	2	147	0	1.4	2	1	7	<b>2</b>
53	1	4	140	203	1	2	155	1	3.1	3	0	7	<b>1</b>

**Tableau 2.2 : Echantillon de la base de données**

Soit le nouvel patient P ayant les caractéristiques présentées dans le tableau suivant, que l'on souhaite classifier.

age	Sex	cp	Trestbps	chol	fbs	restecg	thalach	exang	olbpeak	slope	ca	thal	Pred-attribute
57	1	4	140	192	0	0	148	0	0.4	2	0	6	<b>?</b>

**Tableau 2.3 : Les données d'un nouvel patient**



## Chapitre 2 : Architecture et modélisation

---

### 3.1 Application de la méthode KPP

#### 3.1.2 Classification du nouvel patient P par le KPP

Les étapes pour classer ce patient P par le KPP sont :

##### 1. *Etape normalisation :*

En pratique certains critères auront des valeurs maximales et minimales nettement supérieures à celles d'autres critères. Il est alors nécessaire de d'ajuster (normaliser) les valeurs afin d'éviter que certains critères n'aient un poids trop important. Pour standardiser nous pouvons utiliser la distance « Min-Max Scaling » peut-être appliqué quand les données varient dans des échelles différentes. A l'issue de cette transformation, les valeurs seront comprises dans un intervalle fixe [0,1]. Le but d'avoir un intervalle restreint est de réduire l'espace de variation des valeurs d'un attribut (feature) et par conséquent réduire l'effet des valeurs aberrantes.

- La transformation se fait grâce à la formule suivante :

$$\text{Valeur normalisé} = \frac{\text{valeur} - \text{valeur min}}{\text{valeur max} - \text{valeur min}}$$

- ❖ *valeur min* : La plus petite valeur observée pour l'attribut X.
- ❖ *valeur max* : La plus grande valeur observée pour l'attribut X.
- ❖ *valeur* : La valeur de l'attribut qu'on cherche à normaliser.

##### 2. *Calculer toutes les distances de ce nouveau patient P avec les autres observations* (patients) de la base d'apprentissage **D**. Pour calculer la distance on utilisant **La**

**distance euclidienne** :  $De(x, y) = \sqrt{\sum_{j=1}^n (x_j - y_j)^2}$

Avec :

- $X_j$  : la valeur de l'attribut de la base d'apprentissage
- $y_j$  : la valeur de l'attribut D'un patient p

##### 3. *Retenir les K observations de la base d'apprentissage D les proches de P* en utilisation la fonction de calcul de distance **d**.

##### 4. *Retourner la valeur calculée S* comme étant la valeur qui a été prédite par KNN pour l'observation **P**.

## Chapitre 2 : Architecture et modélisation

On Applique *ces* étapes à ce patient **P** :

- Normalisation toutes des données : Après avoir effectué le processus normalisation, nous obtenons les valeurs suivantes :

Age	sex	Cp	trestbps	chol	fbs	restecg	thalach	exang	olbpeak	slope	ca	thal	Pred-attribue
0.86	1	0	0.625	0.25309	1	1	0.53165	0	0.593	1	0	0.75	0
1	1	1	1	0.58025	0	1	0	1	0.343	0.5	1	0	2
1	1	1	0	0.2284	0	1	0.26582	1	0.68	0.5	0.66	1	1
0	1	0.666	0.25	0.35803	0	0	1	0	0.968	1	0	0	0
0.13	0	0.333	0.25	0.07407	0	1	0.81013	0	0.31	0	0	0	0
0.63	1	0.333	0	0.27161	0	0	0.88608	0	0.1	0	0	0	0
0.83	0	1	0.5	0.46914	0	1	0.65823	0	1	1	0.66	0	3
0.66	0	1	0	1	0	0	0.6962	1	0.06	0	0	0	0
0.86	1	1	0.25	0.38272	0	1	0.49367	0	0.31	0.5	0.33	1	2
0.53	1	1	0.5	0.0679	1	1	0.59494	1	0.843	1	0	1	1

**Tableau 2.4: Normalisation de l'échantillon1 de la base médicale.**

Les valeurs normalisées de patient **p** :

Age	sex	cp	Trestbps	chol	fbs	restecg	Thalach	exang	olbpeak	slope	ca	thal	Pred-attribute
0.66	1	1	0.5	0	0	0	0.50633	0	0	0.5	0	0.75	?

**Tableau 2.5: Normalisation des caractéristiques du patient p**

- Calculer toutes les distances de ce patient **P** avec les autres observations de la base d'apprentissage **D** : pour calculer la distance on utilisant La distance euclidienne

Après calcul, nous obtenons les valeurs suivantes :

age	Sex	cp	Trestbps	chol	fbs	restecg	thalach	exang	olbpeak	slope	ca	thal	Pred-attribute	distance
0.86	1	0	0.625	0.25309	1	1	0.53165	0	0.593	1	0	0.75	0	3.7227
1	1	1	1	0.58025	0	1	0	1	0.343	0.5	1	0	2	4.6348
1	1	1	0	0.2284	0	1	0.26582	1	0.68	0.5	0.66	1	0	<b>3.4507</b>
0	1	0.666	0.25	0.35803	0	0	1	0	0.968	1	0	0	1	<b>2.5520</b>
0.13	0	0.333	0.25	0.07407	0	1	0.81013	0	0.31	0	0	0	0	3.5988
0.63	1	0.333	0	0.27161	0	0	0.88608	0	0.1	0	0	0	1	<b>1.73</b>
0.83	0	1	0.5	0.46914	0	1	0.65823	0	1	1	0.66	0	3	4.5190
0.66	0	1	0	1	0	0	0.6962	1	0.06	0	0	0	0	4.1021
0.86	1	1	0.25	0.38272	0	1	0.49367	0	0.31	0.5	0.33	1	2	<b>1.51</b>
0.53	1	1	0.5	0.0679	1	1	0.59494	1	0.843	1	0	1	1	4.0555

**Tableau 2.6: les distances des instances de l'échantillon par rapport au nouvel patient.**

## Chapitre 2 : Architecture et modélisation

---

- Pour  $k=4$ , nous avons trois patients classés comme suit :
- Deux classés dans la **1**(4<sup>ème</sup> et 6<sup>ème</sup> lignes du tableau précédent).
  - Un classé dans la classe **2**(9<sup>ème</sup> ligne du tableau précédent).
  - Un classé dans la classe **0**(3 ligne du tableau précédent).

La classe majoritaire est la classe **1**,  Le patient **P** est dans la classe **1**

### 3.2 Application de la méthode des Arbres de Décision

#### 3.1.2 Classification du nouvel patient P par AD

##### 3.1.2.1 Principe de la construction de l'arbre de décision

L'idée de construction de l'arbre de décision est simple : il faut commencer par diviser récursivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage par des tests définis à l'aide des attributs jusqu'à ce que l'on obtienne des sous-ensembles d'exemples ne contenant (presque) que des exemples appartenant à une même classe. Cette idée débouche sur des méthodes de construction *Top-Down*, c'est-à-dire construisant l'arbre de la racine vers les feuilles récursives.

On commence généralement par le choix d'un attribut puis le choix d'un nombre de critères pour son nœud. On crée pour chaque critère un nœud concernant les données vérifiant ce critère. L'algorithme continue d'une façon récursive jusqu'à obtenir des nœuds concernant les données de chaque même classe.

# Chapitre 2 : Architecture et modélisation

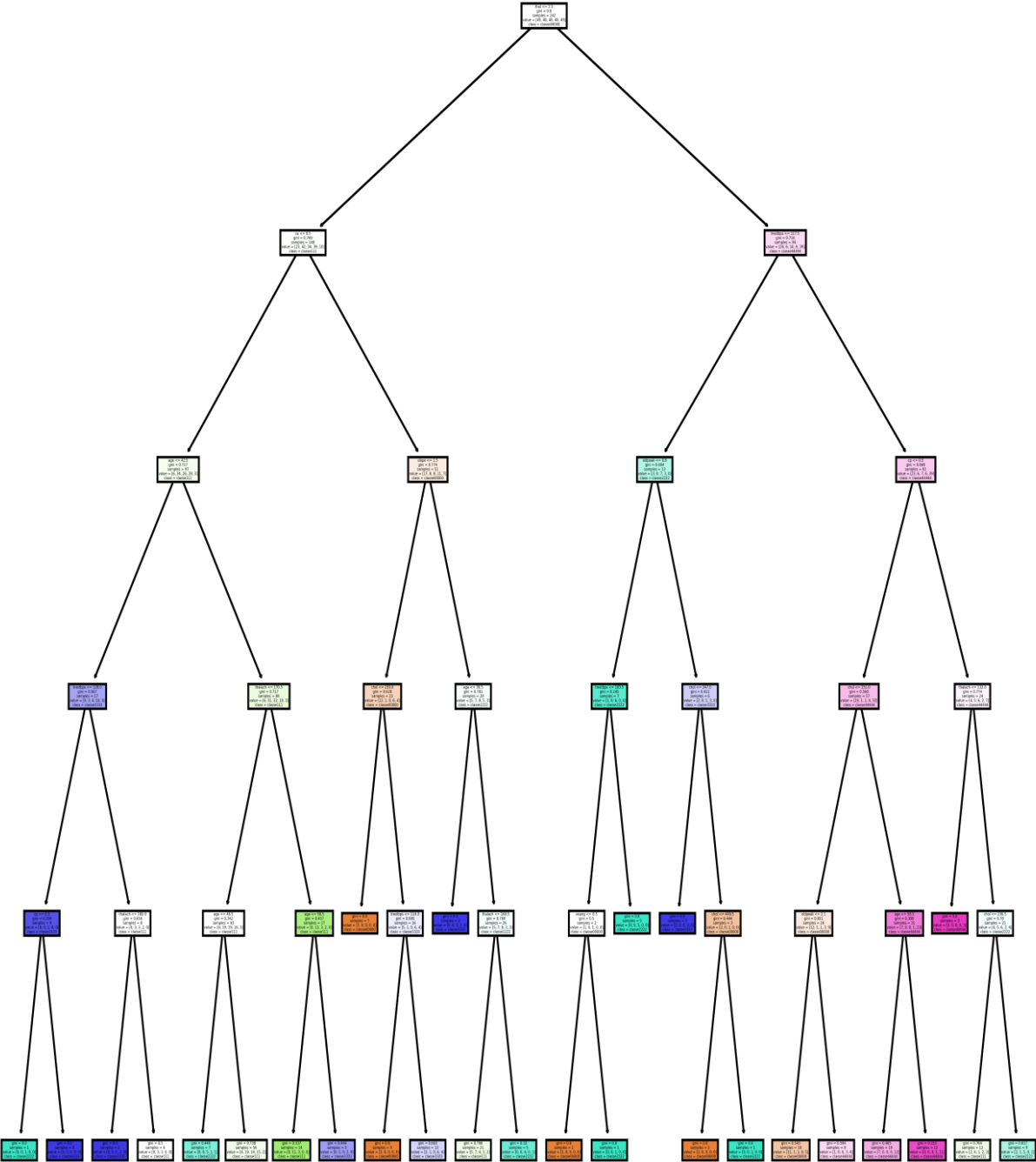


Figure 2.2 : Arbre de décision

## Chapitre 2 : Architecture et modélisation

La figure de l'arbre de décision n'est pas lisible nous avons retranscrit sous forme textuelle

<pre>  --- thal &lt;= 2.50    --- ca &lt;= 0.50      --- age &lt;= 42.50        --- trestbps &lt;= 128.00          --- cp &lt;= 0.50            --- class: 2          --- cp &gt; 0.50            --- class: 3        --- trestbps &gt; 128.00          --- thalach &lt;= 160.00          --- class: 3          --- thalach &gt; 160.00            --- class: 1      --- age &gt; 42.50        --- thalach &lt;= 170.50          --- age &lt;= 46.50          --- class: 2          --- age &gt; 46.50          --- class: 1        --- thalach &gt; 170.50          --- age &lt;= 58.50          --- class: 1          --- age &gt; 58.50          --- class: 3    --- ca &gt; 0.50      --- slope &lt;= 1.50        --- chol &lt;= 239.00          --- class: 0        --- chol &gt; 239.00          --- trestbps &lt;= 114.00          --- class: 0          --- trestbps &gt; 114.00          --- class: 3      --- slope &gt; 1.50        --- age &lt;= 39.50          --- class: 3        --- age &gt; 39.50          --- thalach &lt;= 168.50          --- class: 1          --- thalach &gt; 168.50          --- class: 2 </pre>	<pre>  --- thal &gt; 2.50    --- trestbps &lt;= 117.50      --- oldpeak &lt;= 0.80        --- trestbps &lt;= 100.50          --- exang &lt;= 0.50          --- class: 0          --- exang &gt; 0.50          --- class: 2        --- trestbps &gt; 100.50          --- class: 2        --- oldpeak &gt; 0.80          --- chol &lt;= 247.00          --- class: 3          --- chol &gt; 247.00          --- chol &lt;= 449.50          --- class: 0          --- chol &gt; 449.50          --- class: 2      --- trestbps &gt; 117.50        --- cp &lt;= 0.50          --- chol &lt;= 251.00          --- oldpeak &lt;= 2.10          --- class: 0          --- oldpeak &gt; 2.10          --- class: 4          --- chol &gt; 251.00          --- age &lt;= 58.50          --- class: 4          --- age &gt; 58.50          --- class: 4      --- cp &gt; 0.50        --- thalach &lt;= 132.00          --- class: 4        --- thalach &gt; 132.00          --- chol &lt;= 236.50          --- class: 1          --- chol &gt; 236.50          --- class: 2 </pre>
--	--

**Tableau 2.7 : La forme textuelle de l'arbre de décision (base heart modifiée)**

### 3.1.2.2 Prédiction par Arbre de décision du nouvel patient P

Le processus de décision est équivalent à une « descente » dans l'arbre (de la racine vers une des feuilles) : à chaque étape un attribut est testé et un sous-arbre est choisi, la parcours s'arrête dans une feuille (une décision est prise).

Soit le nouvel patient P que l'on souhaite classifier.

age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	olbpeak	slope	ca	thal	Pred-attribue
57	1	4	140	192	0	0	148	0	0.4	2	0	6	?

**Tableau 2.8 : Les données d'un nouvel patient**

## Chapitre 2 : Architecture et modélisation

--- thal <= 2.50	--- <b>thal &gt; 2.50</b>
--- ca <= 0.50	--- trestbps <= 117.50
--- age <= 42.50	--- oldpeak <= 0.80
--- trestbps <= 128.00	--- trestbps <= 100.50
--- cp <= 0.50	--- exang <= 0.50
--- class: 2	--- class: 0
--- cp > 0.50	--- exang > 0.50
--- class: 3	--- class: 2
--- trestbps > 128.00	--- trestbps > 100.50
--- thalach <= 160.00	--- class: 2
--- class: 3	--- oldpeak > 0.80
--- thalach > 160.00	--- chol <= 247.00
--- class: 1	--- class: 3
--- age > 42.50	--- chol > 247.00
--- thalach <= 170.50	--- chol <= 449.50
--- age <= 46.50	--- class: 0
--- class: 2	--- chol > 449.50
--- age > 46.50	--- class: 2
--- class: 1	--- <b>trestbps &gt; 117.50</b>
--- thalach > 170.50	--- cp <= 0.50
--- age <= 58.50	--- chol <= 251.00
--- class: 1	--- oldpeak <= 2.10
--- age > 58.50	--- class: 0
--- class: 3	--- oldpeak > 2.10
--- ca > 0.50	--- class: 4
--- slope <= 1.50	--- chol > 251.00
--- chol <= 239.00	--- age <= 58.50
--- class: 0	--- class: 4
--- chol > 239.00	--- age > 58.50
--- trestbps <= 114.00	--- class: 4
--- class: 0	--- <b>cp &gt; 0.50</b>
--- trestbps > 114.00	--- thalach <= 132.00
--- class: 3	--- class: 4
--- slope > 1.50	--- <b>thalach &gt; 132.00</b>
--- age <= 39.50	--- <b>chol &lt;= 236.50</b>
--- class: 3	--- <b>class: 1</b>
--- age > 39.50	--- chol > 236.50
--- thalach <= 168.50	--- class: 2
--- class: 1	
--- thalach > 168.50	
--- class: 2	

Tableau 2.9 : La classification du nouvel patient par AD

*Le patient P est dans la classe 1*

## Chapitre 2 : Architecture et modélisation

### 4. Calcul des valeurs pour la Migration entre classes

L'étape de la prédiction de la maladie cardiaque déclenche chez les gens généralement d'autres préoccupations, l'envie de contrôler la maladie pour améliorer l'état de santé de la personne ou d'éviter une aggravation possible. L'objectif principal est toujours ralentissement de l'évolution vers une forme invalidante de la maladie. L'idée est donc *d'intervenir* le plus tôt possible et de manière soutenue afin de prévenir l'aggravation et la sévérité de la maladie.

Ceci se traduit simplement par un changement de certaines valeurs de certains paramètres. La maladie cardiaque peut être classée en cinq catégories. Cette variété de type de maladie nous à pousser à réfléchir à la possibilité de migration (passage) d'une classe à une autre.

Pour un patient donné P, en fonction de ses informations (valeurs des treize paramètres) notre système permet la prédiction du type de sa maladie par exemple *classe i*.

P (*age* = age\_valeur1 ; *sex* = sex\_valeur1 ; *cp* = cp\_valeur1 ; *trestbps* = trestbps\_valeur1 ; *chol* = chol\_valeur1 ; *fbs* = fbs\_valeur1 ; *restecg* = restecg\_valeur1 ; *thalach* = thalach\_valeur1 ; *exang* = exang\_valeur1 ; *oldpeak* = oldpeak\_valeur1 ; *slope* = slope\_valeur1 ; *ca* = ca\_valeur1 ; *thal* = thal\_valeur1).

Ceci peut être schématisé par le schéma explicatif suivant :

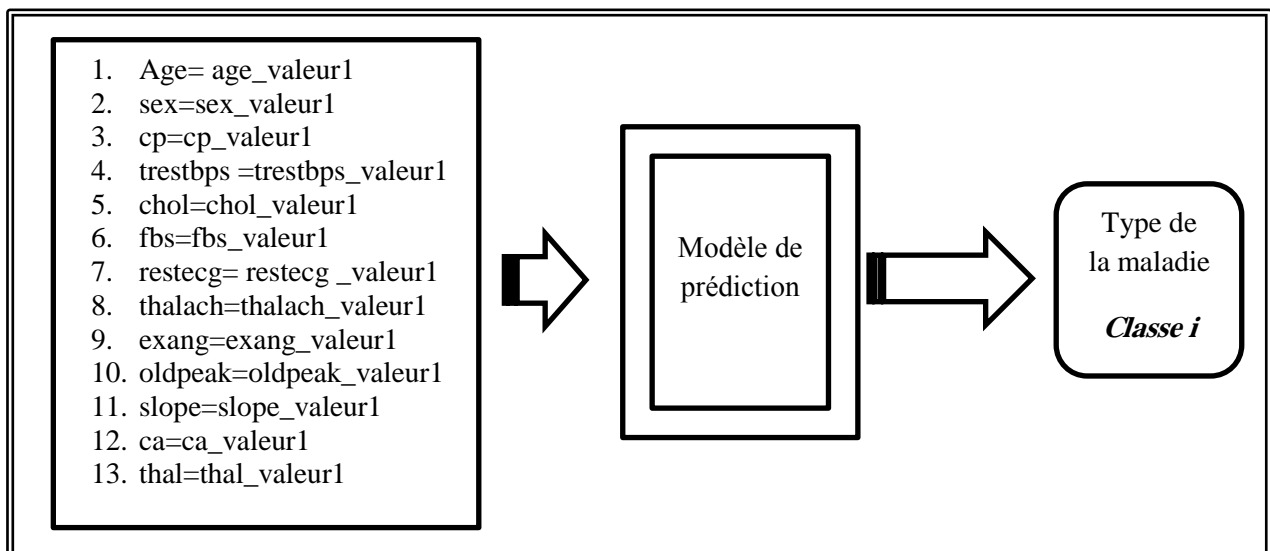


Figure 2. 3 : Schéma explicatif de la prédiction

## Chapitre 2 : Architecture et modélisation

La migration entre la classe revient à trouver de nouvelles valeurs de certains paramètres ou bien de tous les paramètres. Ces nouvelles valeurs une fois réinjectées dans le système de prédiction permettent de prédire la classe souhaitée par le patient *classe-j*.

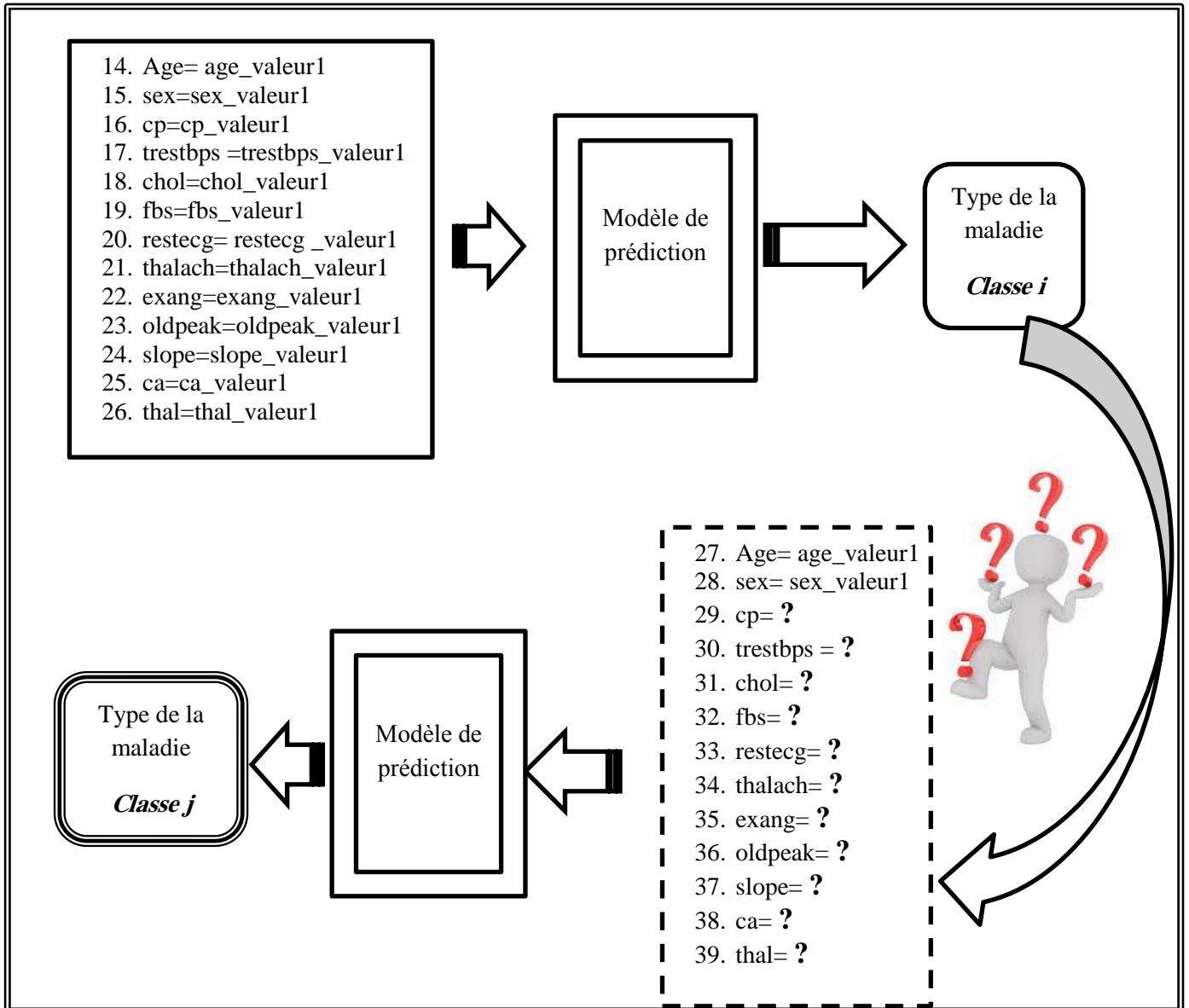


Figure 2.4 : Schéma explicatif de la migration entre classe

### 4.1 Migration entre classes en utilisant le KNN

Pour pouvoir effectuer la migration entre classes, nous avons proposé un algorithme de calcul de nouveaux paramètres pour la techniques KPP, basé sur la recherche parmi les plus proches voisins, les plus proches voisins satisfaisant le passage vers classe souhaitée en indiquant les paramètres que le patient souhaite de maintenir ou qu'il accepté leur modification.



## Chapitre 2 : Architecture et modélisation

La démarche consiste à trouver parmi les plus proches voisins les voisins, les voisins satisfaisant deux types de contraintes (contrainte\_stricte = *la classe souhaitée*, et la liste des paramètres inchangés «contraintes de valeurs»). Certains paramètres sont inchangés dans ce cas, il ne faut sélectionner parmi les plus proches voisins que ceux qui ont les mêmes valeurs *identiques* avec les valeurs du patient.

### Partie 1

- **Entrée** : paramètres du patient P
- **Prédiction** de la classe du patient P
- **Sortie** :
  - **Classe-i** = La classe du patient P
  - **Voisins** : Tous les voisins utilisés pour la prédiction du patient P **ordonné** selon la distance.

### Partie 2

- **Entrée** : **Voisins**
- Prédiction de la classe de tous les voisins
- **Sortie** : **Voisins** : Tous les voisins utilisés pour la prédiction du patient P **ordonné** selon la distance en plus de la colonne *Classe\_prédite* par le système pour chaque voisin.

### Partie 3

- **Entrée** :
  - **Classe-j**= la classe souhaitée
  - **Voisins**
  - **Liste des paramètres inchangés L**
- **Etape 1** :
  - *Sélection des voisins dont la classe prédite par KNN est la classe\_j*
  - *Voisins = select \* from Voisins where classe\_predite=classe\_j*
- **Etape 2**:
  - *Pour chaque paramètre li dans la liste L faire une sélection dans les Voisins de ceux ayant la même valeur pour le paramètre li du patient P*
  - *Voisins = select \* from Voisins where li=li\_val\_patient\_P*
- **Sortie** :
  - *Les voisins satisfaisant les contraintes*
  - *Les valeurs du premier voisins satisfaisant l'ensemble des contraintes*

Figure 2.5 : Méta-algorithme de la migration entre classe pour KNN

## Chapitre 2 : Architecture et modélisation

### 4.1.1 Exemple de Migration par KNN

Soit le nouvel patient P ayant les caractéristiques présentées dans le tableau suivant

age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Pred-attribute
57	1	4	140	192	0	0	148	0	0.4	2	0	6	1

Tableau 2.10 : Les données d'un patient qui souhaite faire des migrations

- Migration vers la classe0

- **Attribut inchangés (Age, sex, fbs, restecg, exang, oldpeak)**

- **Résultat :** Nous avons trouvé une seule combinaison

Voisins

```
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal Classe_Predite
57 1 2 128 229 0 0 150 0 0.4 1 1 3 0
```

- **Attribut inchangés (Age et sex)**

- **Résultat :** Nous avons trouvés 5 combinaisons possibles

Voisins

```
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal Classe_Predite
57 1 0 132 207 0 1 168 1 0.0 2 0 3 0
57 1 0 110 201 0 1 126 1 1.5 1 0 1 0
57 1 2 128 229 0 0 150 0 0.4 1 1 3 0
57 1 1 154 232 0 0 164 0 0.0 2 1 2 0
57 1 0 130 131 0 1 115 1 1.2 1 1 3 0
```

- **Attribut inchangés (Age, sex et cp)**

- **Résultat :** Nous n'avons pas trouvé de combinaison
- *Impossible de migrer vers la classe 0 avec ces conditions*

- Migration vers la classe2

- **Attribut inchangés (Age, sex, fbs, restecg, exang, oldpeak)**

- **Résultat :** Nous n'avons pas trouvé de combinaison
- *Impossible de migrer vers la classe 2 avec ces conditions*

- **Attribut inchangés (Age et sex)**

- **Résultat :** Nous n'avons pas trouvé de combinaison
- *Impossible de migrer vers la classe 2 avec ces conditions*

- **Attribut inchangés (Sex, fbs, restecg, exang, slope, ca)**

- **Résultat :** Nous avons trouvé une seule combinaison

Voisins

```
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal Classe_Predite
45 1 1 128 308 0 0 170 0 0 2 0 2 2
```

## Chapitre 2 : Architecture et modélisation

- Migration vers la classe3

- Attribut inchangés (Age et sex)

- **Résultat** : Nous avons trouvé 3 combinaisons

```
Voisins
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal Classe_Predite
57 1 2 150 168 0 1 174 0 1.6 2 0 2 3
57 1 0 150 276 0 0 112 1 0.6 1 1 1 3
57 1 0 152 274 0 1 88 1 1.2 1 1 3 3
```

- Attribut inchangés (Age, sex et cp)

- **Résultat** : Nous n'avons pas trouvé de combinaison
- *Impossible de migrer vers la classe 2 avec ces conditions*

- Attribut inchangés (Sex, fbs, restecg, exang, slope, ca)

- **Résultat** : Nous avons trouvé 2 combinaisons

```
Voisins
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal Classe_Predite
66 1 0 160 228 0 0 138 0 2.3 2 0 1 3
70 1 1 156 245 0 0 143 0 0.0 2 0 2 3
```

- Migration vers la classe4

- Attribut inchangés (Age et sex)

- **Résultat** : Nous avons trouvé 2 combinaisons

```
Voisins
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal Classe_Predite
57 1 0 165 289 1 0 124 0 1 1 3 3 4
57 1 0 110 335 0 1 143 1 3 1 1 3 4
```

- Attribut inchangés (Age, sex et cp)

- **Résultat** : Nous n'avons pas trouvé de combinaison
- *Impossible de migrer vers la classe 2 avec ces conditions*

- Attribut inchangés (Sex, fbs, restecg, exang, slope, ca)

- **Résultat** : Nous avons trouvé 1 combinaison

```
Voisins
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal Classe_Predite
59 1 3 160 273 0 0 125 0 0 2 0 2 4
```

## Chapitre 2 : Architecture et modélisation

### 4.2 Migration entre classes en utilisant l'AD

Pour la technique des arbres de décision, nous avons proposé de sauvegarder l'arbre de décision (tous les chemins) dans une table et d'effectuer des requêtes pour satisfaire la migration vers la classe souhaitée sous les contraintes imposées par le patient à savoir maintenir les valeurs initiales ou accepter leurs modifications.

La démarche consiste donc à rechercher les chemins satisfaisant les contraintes imposés par le patient. Il faut savoir qu'un chemin ne représenté pas une seule combinaison de valeur mais plutôt toute une sous population.

Nous avons sauvegardé l'arbre de décision dans le tableau suivant :

comp	age	comp	sex	comp	cp	comp	trestbps	comp	chol	comp	lbs	comp	restecg	comp	thalach	comp	exang	comp	oldpeak	comp	slope	comp	ca	comp	thal	Classe_Predite
<=	42.5	Ind	-	Ind	-	<=	128	Ind	-	Ind	-	Ind	-	<=	142	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	2
<=	42.5	Ind	-	Ind	-	<=	128	Ind	-	Ind	-	Ind	-	>	142	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	3
<=	42.5	Ind	-	Ind	-	>	128	Ind	-	Ind	-	Ind	-	<=	160	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	3
<=	42.5	Ind	-	Ind	-	>	128	Ind	-	Ind	-	Ind	-	>	160	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	1
>	42.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	170.5	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	2
>	42.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	170.5	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	1
>	42.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	170.5	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	1
>	42.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	170.5	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	3
Ind	-	Ind	-	Ind	-	Ind	-	<=	239	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	1.5	>	0.5	<=	2.5	0
Ind	-	Ind	-	Ind	-	<=	114	>	239	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	0.5	<=	2.5	0
Ind	-	Ind	-	Ind	-	>	114	>	239	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	0.5	<=	2.5	3
Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	>	1.5	>	0.5	<=	2.5	4
Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	>	0.5	Ind	-	Ind	-	Ind	-	>	1.5	>	0.5	<=	2.5	0
Ind	-	Ind	-	>	0.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	0.3	>	1.5	>	0.5	<=	2.5	2
Ind	-	Ind	-	>	0.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	0.3	>	1.5	>	0.5	<=	2.5	1
<=	56	Ind	-	Ind	-	<=	117.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	0.8	Ind	-	Ind	-	>	2.5	2
>	56	Ind	-	Ind	-	<=	117.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	0.8	Ind	-	<=	0.5	>	2.5	2
>	56	Ind	-	Ind	-	<=	117.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	0.8	Ind	-	>	0.5	>	2.5	0
Ind	-	Ind	-	Ind	-	<=	117.5	<=	247	Ind	-	Ind	-	Ind	-	Ind	-	>	0.8	Ind	-	Ind	-	>	2.5	3
<=	62	Ind	-	Ind	-	<=	117.5	>	247	Ind	-	Ind	-	Ind	-	Ind	-	>	0.8	Ind	-	Ind	-	>	2.5	0
>	62	Ind	-	Ind	-	<=	117.5	>	247	Ind	-	Ind	-	Ind	-	Ind	-	>	0.8	Ind	-	Ind	-	>	2.5	2
Ind	-	Ind	-	<=	0.5	>	117.5	<=	251	Ind	-	Ind	-	Ind	-	Ind	-	<=	2.1	Ind	-	Ind	-	>	2.5	0
Ind	-	Ind	-	<=	0.5	>	117.5	<=	251	Ind	-	Ind	-	Ind	-	Ind	-	>	2.1	Ind	-	Ind	-	>	2.5	4
Ind	-	Ind	-	<=	0.5	>	117.5	>	251	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	2.5	4
Ind	-	Ind	-	>	0.5	>	117.5	Ind	-	Ind	-	Ind	-	<=	132	Ind	-	Ind	-	Ind	-	Ind	-	>	2.5	4
Ind	-	Ind	-	>	0.5	>	117.5	<=	236.5	Ind	-	Ind	-	>	132	Ind	-	Ind	-	Ind	-	Ind	-	>	2.5	1
Ind	-	Ind	-	>	0.5	>	117.5	>	236.5	Ind	-	Ind	-	>	132	Ind	-	Ind	-	Ind	-	Ind	-	>	2.5	2

Tableau 2.10 : Table des chemins de l'arbre de décision (base heart modifiée)

## Chapitre 2 : Architecture et modélisation

comp	age	comp	sex	comp	cp	comp	trestbps	comp	chol	comp	fbs	comp	restecg	comp	thalach	comp	exang	comp	oldpeak	comp	slope	comp	ca	comp	thal	Classe_Predite
Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	Ind	-	<=	96.5	<=	0.5	Ind	-	Ind	-	<=	0.5	<=	2.5	0
Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	Ind	-	>	96.5	<=	0.5	Ind	-	Ind	-	<=	0.5	<=	2.5	1
Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	<=	0.5	Ind	-	<=	0.5	Ind	-	Ind	-	<=	0.5	>	2.5	0
Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	>	0.5	Ind	-	<=	0.5	Ind	-	Ind	-	<=	0.5	>	2.5	1
Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	Ind	-	<=	162	>	0.5	Ind	-	Ind	-	<=	0.5	Ind	-	0
Ind	-	Ind	-	<=	0.5	Ind	-	Ind	-	Ind	-	Ind	-	<=	162	>	0.5	Ind	-	Ind	-	<=	0.5	Ind	-	1
Ind	-	Ind	-	<=	0.5	<=	109	<=	233.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	0.5	Ind	-	1
Ind	-	Ind	-	<=	0.5	<=	109	<=	233.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	0.5	Ind	-	0
Ind	-	<=	0.5	<=	0.5	>	109	<=	285.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	0.5	Ind	-	0
Ind	-	<=	0.5	<=	0.5	>	109	>	285.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	0.5	Ind	-	1
Ind	-	>	0.5	<=	0.5	>	109	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	0.5	Ind	-	0
<=	55.5	Ind	-	>	0.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	3.55	Ind	-	Ind	-	<=	2.5	1
<=	55.5	Ind	-	>	0.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	>	3.55	Ind	-	Ind	-	<=	2.5	0
<=	59.5	Ind	-	>	0.5	<=	138	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	2.5	0
<=	59.5	Ind	-	>	0.5	>	138	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	2.5	1
>	59.5	Ind	-	>	0.5	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	Ind	-	<=	2.5	1
Ind	-	Ind	-	>	0.5	Ind	-	Ind	-	Ind	-	Ind	-	<=	142.5	Ind	-	Ind	-	Ind	-	Ind	-	>	2.5	0
Ind	-	Ind	-	>	0.5	<=	176	Ind	-	Ind	-	Ind	-	>	142.5	Ind	-	<=	1.95	Ind	-	Ind	-	>	2.5	1
Ind	-	Ind	-	>	0.5	>	176	Ind	-	Ind	-	Ind	-	>	142.5	Ind	-	<=	1.95	Ind	-	Ind	-	>	2.5	0
Ind	-	Ind	-	>	0.5	<=	159	Ind	-	Ind	-	Ind	-	>	142.5	Ind	-	>	1.95	Ind	-	Ind	-	>	2.5	0
Ind	-	Ind	-	>	0.5	>	159	Ind	-	Ind	-	Ind	-	>	142.5	Ind	-	>	1.95	Ind	-	Ind	-	>	2.5	1

Tableau 2.11 : Table des chemins de l'arbre de décision (base heart originale)

### Explication d'une ligne du tableau :

Nous avons défini pour chaque attribut deux colonnes, la première colonne pour sauvegarder un opérateur de comparaison et l'autre colonne pour sauvegarder la valeur de cet attribut.

Nous avons utilisé trois valeurs possibles dans la colonne **comp** :

- Ind : dans le cas où l'attribut en question ne figure pas dans le chemin
- <= : inférieur ou égale à la valeur qui est dans la colonne de l'attribut en question
- > : inférieur ou égale à la valeur qui est dans la colonne de l'attribut en question

Le chemin suivi est sauvegardé dans la première ligne du tableau 2.10.

```

|--- thal <= 2.50
| |--- ca <= 0.50
| | |--- age <= 42.50
| | | |--- trestbps <= 128.00
| | | | |--- cp <= 0.50
| | | | | |--- class: 2
    
```

Comp	Age	comp	sex	comp	cp	comp	trestbps	comp	chol	comp	fbs	comp	restecg
<=	42.5	Ind	-	Ind	-	<=	128	Ind	-	Ind	-	Ind	-

Comp	thalach	comp	exang	comp	oldpeak	comp	slope	comp	ca	comp	thal	Classe_Predite
<=	142	Ind	-	Ind	-	Ind	-	<=	0.5	<=	2.5	2

## Chapitre 2 : Architecture et modélisation

### Partie 1

- **Entrée** : paramètres du patient P
- **Prédiction** de la classe du patient P
- **Sortie** :
  - **Classe-i** = La classe du patient P

### Partie 2

- **Entrée** :
  - **Classe-j**= la classe souhaitée
  - **Table des chemins de l'AD**
  - **Liste des paramètres inchangés L**
- **Etape 1** :
  - *Sélection des chemins dont la classe prédite par AD est la classe\_j*
  - **Chemins** = *select \* from Table-des-chemins-AD where classe\_predite=classe\_j*
- **Etape 2**:
  - *Pour chaque paramètre li dans la liste L faire une sélection dans les chemins de ceux ayant la même valeur pour le paramètre li du patient P*
  - **Chemins** = *select \* from Table-des-chemins-AD where li=li\_val\_patient\_P*
- **Sortie** :
  - **Les chemins satisfaisant les contraintes**

Figure 2.6 : Méta-algorithme de la migration entre classe pour AD

### 4.2.1 Exemple de Migration par AD

Soit le nouvel patient P ayant les caractéristiques présentées dans le tableau suivant

age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	olbpeak	slope	ca	thal	Pred-attribute
57	1	4	140	192	0	0	148	0	0.4	2	0	6	1

Tableau 2.12 : Les données d'un patient qui souhaite faire des migrations

## Chapitre 2 : Architecture et modélisation

- Migration vers la classe0
  - **Attribut inchangés (Age, sex, cp, fbs, restcg, exang, oldpeak)**
    - **Résultat :** Nous avons trouvé trois chemins possibles

```
> Voisins
  age_comp age  sex_comp sex  cp_comp cp  trestbps_comp trestbps  chol_comp chol  fbs_comp fbs
1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 239 QuelqueSoit -1
2 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 114.0 Superieur 239 QuelqueSoit -1
3 Superieur 56 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 117.5 QuelqueSoit -1 QuelqueSoit -1
  restecg_comp restecg thalach_comp thalach  exang_comp exang  oldpeak_comp oldpeak  slope_comp slope  ca_comp
1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 1.5 Superieur
2 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 QuelqueSoit -1.0 Superieur
3 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 0.8 QuelqueSoit -1.0 Superieur
  ca  thal_comp thal Classe_Predite
1 0.5 InferieurOuEgale 2.5 0
2 0.5 InferieurOuEgale 2.5 0
3 0.5 Superieur 2.5 0
```

- **Attribut inchangés (Age et sex)**
  - **Résultat :** Nous avons trouvés 6 chemins possibles

```
> Voisins
  age_comp age  sex_comp sex  cp_comp cp  trestbps_comp trestbps  chol_comp chol  fbs_comp
1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 QuelqueSoit -1.0 InferieurOuEgale 239 QuelqueSoit
2 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 114.0 Superieur 239 QuelqueSoit
3 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 0.5 QuelqueSoit -1.0 QuelqueSoit -1 QuelqueSoit
4 Superieur 56 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 117.5 QuelqueSoit -1 QuelqueSoit
5 InferieurOuEgale 62 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 117.5 Superieur 247 QuelqueSoit
6 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 0.5 Superieur 117.5 InferieurOuEgale 251 QuelqueSoit
  fbs_comp fbs  restecg_comp restecg thalach_comp thalach  exang_comp exang  oldpeak_comp oldpeak  slope_comp slope
1 -1 QuelqueSoit -1.0 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 1.5
2 -1 QuelqueSoit -1.0 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 QuelqueSoit -1.0
3 -1 Superieur 0.5 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 Superieur 1.5
4 -1 QuelqueSoit -1.0 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 0.8 QuelqueSoit -1.0
5 -1 QuelqueSoit -1.0 QuelqueSoit -1 QuelqueSoit -1 Superieur 0.8 QuelqueSoit -1.0
6 -1 QuelqueSoit -1.0 QuelqueSoit -1 InferieurOuEgale 2.1 QuelqueSoit -1.0
  ca_comp ca  thal_comp thal Classe_Predite
1 Superieur 0.5 InferieurOuEgale 2.5 0
2 Superieur 0.5 InferieurOuEgale 2.5 0
3 Superieur 0.5 InferieurOuEgale 2.5 0
4 Superieur 0.5 Superieur 2.5 0
5 QuelqueSoit -1.0 Superieur 2.5 0
6 QuelqueSoit -1.0 Superieur 2.5 0
```

- **Attribut inchangés (Age, sex et cp)**
  - **Résultat :** Nous avons trouvé 4 chemins

```
> Voisins
  age_comp age  sex_comp sex  cp_comp cp  trestbps_comp trestbps  chol_comp chol  fbs_comp fbs
1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 239 QuelqueSoit -1
2 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 114.0 Superieur 239 QuelqueSoit -1
3 Superieur 56 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 117.5 QuelqueSoit -1 QuelqueSoit -1
4 InferieurOuEgale 62 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 117.5 Superieur 247 QuelqueSoit -1
  restecg_comp restecg thalach_comp thalach  exang_comp exang  oldpeak_comp oldpeak  slope_comp slope
1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 1.5
2 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1.0 QuelqueSoit -1.0
3 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 0.8 QuelqueSoit -1.0
4 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit -1 Superieur 0.8 QuelqueSoit -1.0
  ca_comp ca  thal_comp thal Classe_Predite
1 Superieur 0.5 InferieurOuEgale 2.5 0
2 Superieur 0.5 InferieurOuEgale 2.5 0
3 Superieur 0.5 Superieur 2.5 0
4 QuelqueSoit -1.0 Superieur 2.5 0
```



## Chapitre 2 : Architecture et modélisation

- Migration vers la classe2

- **Attribut inchangés (Age, sex, fbs, restcg, exang, oldpeak, ca)**

- **Résultat : Nous avons trouvé 3 chemins**

```
> Voisins
  age_comp age  sex_comp sex  cp_comp cp  trestbps_comp trestbps  chol_comp chol  fbs_comp fbs
1  Supérieur 42.5 QuelqueSoit -1 QuelqueSoit -1.0  QuelqueSoit -1.0 QuelqueSoit -1.0 QuelqueSoit -1
2  Supérieur 56.0 QuelqueSoit -1 QuelqueSoit -1.0 InferieurOuEgale 117.5 QuelqueSoit -1.0 QuelqueSoit -1
3  QuelqueSoit -1.0 QuelqueSoit -1 Supérieur 0.5  Supérieur 117.5 Supérieur 236.5 QuelqueSoit -1
  restcg_comp restcg  thalach_comp thalach  exang_comp exang  oldpeak_comp oldpeak  slope_comp slope
1  QuelqueSoit -1 InferieurOuEgale 170.5 QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1
2  QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1 InferieurOuEgale 0.8 QuelqueSoit -1
3  QuelqueSoit -1  Supérieur 132.0 QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1
  ca_comp ca  thal_comp thal Classe_Predite
1 InferieurOuEgale 0.5 InferieurOuEgale 2.5 2
2 InferieurOuEgale 0.5  Supérieur 2.5 2
3  QuelqueSoit -1.0  Supérieur 2.5 2
```

- Migration vers la classe3

- **Attribut inchangés (Age, sex et cp)**

- **Résultat : Nous avons trouvé 3 chemins**

```
> Voisins
  age_comp age  sex_comp sex  cp_comp cp  trestbps_comp trestbps  chol_comp chol  fbs_comp fbs
1  Supérieur 42.5 QuelqueSoit -1 QuelqueSoit -1  QuelqueSoit -1.0  QuelqueSoit -1 QuelqueSoit -1
2  QuelqueSoit -1.0 QuelqueSoit -1 QuelqueSoit -1  Supérieur 114.0  Supérieur 239 QuelqueSoit -1
3  QuelqueSoit -1.0 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 117.5 InferieurOuEgale 247 QuelqueSoit -1
  restcg_comp restcg  thalach_comp thalach  exang_comp exang  oldpeak_comp oldpeak  slope_comp slope  ca_comp
1  QuelqueSoit -1  Supérieur 170.5 QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1 InferieurOuEgale
2  QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1  Supérieur
3  QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1  Supérieur 0.8 QuelqueSoit -1  QuelqueSoit
  ca  thal_comp thal Classe_Predite
1 0.5 InferieurOuEgale 2.5 3
2 0.5 InferieurOuEgale 2.5 3
3 -1.0  Supérieur 2.5 3
```

- **Attribut inchangés (Age, sex, cp, fbs, restcg, exang, slope, ca)**

- **Résultat : Nous avons trouvé 2 chemins**

```
> Voisins
  age_comp age  sex_comp sex  cp_comp cp  trestbps_comp trestbps  chol_comp chol  fbs_comp fbs
1  Supérieur 42.5 QuelqueSoit -1 QuelqueSoit -1  QuelqueSoit -1.0  QuelqueSoit -1 QuelqueSoit -1
2  QuelqueSoit -1.0 QuelqueSoit -1 QuelqueSoit -1 InferieurOuEgale 117.5 InferieurOuEgale 247 QuelqueSoit -1
  restcg_comp restcg  thalach_comp thalach  exang_comp exang  oldpeak_comp oldpeak  slope_comp slope  ca_comp
1  QuelqueSoit -1  Supérieur 170.5 QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1 InferieurOuEgale
2  QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1  QuelqueSoit -1.0 QuelqueSoit -1  Supérieur
  ca  thal_comp thal Classe_Predite
1 0.5 InferieurOuEgale 2.5 3
2 -1.0  Supérieur 2.5 3
```

- Migration vers la classe4

- **Attribut inchangés (Age, sex et cp)**

- **Résultat : Nous avons trouvé 1 chemin**

```
> Voisins
  age_comp age  sex_comp sex  cp_comp cp  trestbps_comp trestbps  chol_comp chol  fbs_comp fbs restcg_comp
1  QuelqueSoit -1 QuelqueSoit -1 Supérieur 0.5  Supérieur 117.5 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit
  restcg  thalach_comp thalach  exang_comp exang  oldpeak_comp oldpeak  slope_comp slope  ca_comp ca thal_comp
1  -1 InferieurOuEgale 132 QuelqueSoit -1  QuelqueSoit -1  QuelqueSoit -1  QuelqueSoit -1 Supérieur
  thal Classe_Predite
1 2.5 4
```



## Chapitre 2 : Architecture et modélisation

---

- **Attribut inchangés (Age, sex, cp, fbs, restecg, exang, slope, ca)**
  - **Résultat :** Nous avons trouvé 1 chemin

```
> Voisins
  age_comp age      sex_comp sex      cp_comp cp      trestbps_comp trestbps      chol_comp chol      fbs_comp fbs      restecg_comp
1 QuelqueSoit -1 QuelqueSoit -1 Supérieur 0.5      Supérieur 117.5 QuelqueSoit -1 QuelqueSoit -1 QuelqueSoit
  restecg      thalach_comp thalach      exang_comp exang      oldpeak_comp oldpeak      slope_comp slope      ca_comp ca      thal_comp
1      -1 InferieurOuEgale      132 QuelqueSoit      -1 QuelqueSoit      -1 QuelqueSoit      -1 QuelqueSoit -1 Supérieur
  thal Classe_Predite
1 2.5 4
```

## 5. Conclusion

Dans notre travail, nous nous sommes intéressés à deux techniques d'apprentissages supervisés : K plus proche voisins et les arbres de décisions pour la prédiction précoce des maladies cardiaques peut réduire le risque pour la santé d'un patient.

L'objectif de prédiction des différentes classes de maladie est d'intervenir le plus tôt possible et de manière soutenue afin de prévenir l'aggravation et la sévérité de la maladie. Le patient doit prendre conscience de la réalité de son état médical afin de percevoir pleinement ses enjeux pour suivre activement son traitement et ceux à venir. L'objectif principal est un ralentissement de l'évolution vers une forme invalidante de la maladie.

Généralement les gens ont un autre souci est de savoir quels changements doivent-ils faire pour être ou ne pas être dans une autre classe (type de maladie). Nous avons donc pensé à l'idée de la migration entre classes. Cette migration (passage) entre classe peut être vue comme une perspective d'amélioration de l'état d'un malade ou malheureusement par une dégradation (complication) de son état de santé. Nous avons proposé pour chaque technique un algorithme de calcul de nouvelle valeur des paramètres pour la migration d'une classe à une autre.

Nous avons opté pour la réalisation d'une application web-mobile, ce qui est très pratique pour les patients qui souhaitent suivre leurs états de santé ainsi que de voir l'impact des changements des paramètres aussi minimaux qu'ils soient sur leur état de santé.

Le chapitre suivant est dédié à la présentation de la réalisation et l'implémentation de ces techniques dans une application web.

III

# Implémentation et Bilan

## 1. Introduction

Ce chapitre est dédié à présentation de notre application web-mobile et à l'implémentation des différentes techniques de classification que nous avons utilisées. Le chapitre est organisé comme suit :

- Dans la première partie du chapitre, nous présentons une évaluation des résultats des algorithmes utilisés et comparerons les résultats.
- Dans la seconde partie, nous présenterons les outils et langages utilisés dans notre projet.
- Enfin, nous présenterons notre application Web-mobile pour la classification et la migration entre classes de la maladie cardiaque. Notre application porte le nom « Prédiction et Migration pour le diagnostic précoce de la maladie cardiaque ».

## 2. Evaluation des résultats

Dans cette section nous allons présenter les résultats obtenus par nos différentes techniques utilisées. Diverses métriques (indices, critères) sont utilisées dans le cadre de l'apprentissage-machine pour mesurer la précision prédictive d'un modèle. La validation croisée est une mesure qui permet d'évaluer la performance d'un modèle. Le principe est de construire un modèle prédictif et de le tester sur des données qu'il n'a jamais vues. Ceci imite le monde réel où le modèle sera déployé et utilisé comme outil prédictif. Les données sont divisées en deux ensembles d'observations : l'échantillon d'apprentissage (= entraînement), de taille généralement plus importante, et l'échantillon test. Les performances prédictives du modèle sont mises à l'épreuve avec l'échantillon test. Le modèle prédit les résultats  $Y$  sur la base des données de l'échantillon test. Si les prédictions et les données  $Y$  réelles (= observées) sont proches, alors l'algorithme a une bonne performance. Il existe des indices (critères) qui mesurent cette performance.

### 2.1 Critères et mesures d'évaluation

#### ➤ Matrice de confusion

Dans les problématiques de classification, la plupart des indices de performance sont calculés à partir d'une **matrice de confusion**. Cette matrice affiche le nombre de succès et

## Chapitre 3 : Implémentation et Bilan

d'échecs de prédiction pour chaque catégorie de la variable à prédire. La matrice de confusion est une table qui montre chaque classe dans les données d'évaluation, ainsi que le nombre ou le pourcentage de prédictions correctes et incorrectes.

Dans le cas d'une tâche de classification supervisée binaire, où la modalité de la variable à prédire correspond à la classe «positive» et l'autre à la classe «négative», on nomme les coefficients de la matrice de confusion de la manière suivante :

- VN : Nombre de *vrais négatifs* (True Negatif TN)
- FN : Nombre de faux négatifs (False Negatif FN)
- FP : Nombre de faux positifs (False Positif FP)
- VP : Nombre de *vrais positifs* (True Positif TP)

		Y prédit par le modèle	
		Y=1	Y=0
Y réel(Y')	Y'=1	Nombre de 1 prédits correctement Vrai Positifs (VP) True Positif (TP)	Nombre de 1 prédits en 0 Faux Négatif (FN) False Negatif (FN)
	Y'=0	Nombre de 0 prédits en 1 Faux Positifs (FP) False Positif (FP)	Nombre de 0 prédits correctement Vrai Négatif (VN) True Negatif (TN)

**Tableau 3.1 Matrice de Confusion**

- **Accuracy** (*Exactitude*, justesse) (la proportion de prédictions correctes): il s'agit d'une description d'erreurs systématiques, d'une mesure du biais statistique; faible précision provoque une différence entre un résultat et une valeur "vraie".

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

- **Précision** : Proportion de solutions trouvées qui sont pertinentes. A quel point les prédictions positives sont précises.

$$\text{Précision} = \frac{TP}{TP+FP} * 100\%$$

## Chapitre 3 : Implémentation et Bilan

- **Rappel (sensitivity, reccal) :** Proportion des solutions pertinentes qui sont trouvées. Mesure la capacité du système à donner toutes les solutions pertinentes. Couverture des observations vraiment positives.

$$\text{Rappel (sensitivity, reccal)} = \frac{TP}{TP+FN} * 100\%$$

- **F-mesure (F-score) :** La F-mesure correspond à un compromis de la précision et du rappel donnant la performance du modèle. Moyenne harmonique de la précision et du rappel. Mesure la capacité du modèle à donner toutes les solutions pertinentes et à refuser les autres.

$$\text{F1 score} = 2 * \frac{\text{Rappel} * \text{Précision}}{\text{Rappel} + \text{Précision}} * 100\%$$

Nous pouvons résumer ces indicateurs principaux dans le tableau suivant:

Indicateur	Formule	Interprétation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Perforance globale du modèle
Précision	$\frac{TP}{TP+FP}$	À quel point les predictions positives sont précises
Rappel "Sensibilité"	$\frac{TP}{TP+FN}$	Couverture des observations vraiment positifs
Spécificité	$\frac{TN}{TN+FP}$	Couverture des observations vraiment négatives
F-mesure	$\frac{2TP}{2TP+FP+FN}$	Indicateur hybride utilisé pour les classes non-balancées

**Tableau 3.2 Les différents critères d'évaluation d'un modèle de classification**

- **Matrice de confusion pour la prédiction de maladie cardiaque**

Notre variable de prédiction n'est pas binaire mais prend ses valeurs dans l'ensemble {0, 1, 2, 3,4}

		Classe prédite par le modèle				
		Y=0	Y=1	Y=2	Y=3	Y=4
Classe réel	Y'=0	<b>m<sub>00</sub></b>	m <sub>01</sub>	m <sub>02</sub>	m <sub>03</sub>	m <sub>04</sub>
	Y'=1	m <sub>10</sub>	<b>m<sub>11</sub></b>	m <sub>12</sub>	m <sub>13</sub>	m <sub>14</sub>
	Y'=2	m <sub>20</sub>	m <sub>21</sub>	<b>m<sub>22</sub></b>	m <sub>23</sub>	m <sub>24</sub>
	Y'=3	m <sub>30</sub>	m <sub>31</sub>	m <sub>32</sub>	<b>m<sub>33</sub></b>	m <sub>34</sub>
	Y'=4	m <sub>40</sub>	m <sub>41</sub>	m <sub>42</sub>	m <sub>43</sub>	<b>m<sub>44</sub></b>

**Tableau 3.3 : La matrice de confusion de notre modèle.**

- **m<sub>ij</sub>** : représente le nombre de patients de **classe i** prédit **classe j** par le modèle.
- **m<sub>cc</sub>** : Le nombre de patients de la **classe c** prédits correctement par le modèle (classe i)  
(représente les true positif de la **classe c**)

❖ **Accuracy**

Correspond à la proportion d'observations bien classées.

$$Accuracy = \frac{\sum_i m_{ii}}{\sum_{i,j} m_{ij}}$$

❖ **Taux d'erreur global**

Le taux d'erreur global, correspond à la proportion d'observations mal classées, qui dépend du ratio entre la trace de la matrice de confusion (c'est-à-dire la somme des coefficients diagonaux, donc le nombre de bonnes prédictions), et la somme de tous les coefficients (autrement dit le nombre total de prédictions) :

$$E = 1 - \frac{\sum_{i=0}^4 m_{ii}}{\sum_{i,j} m_{ij}}$$

❖ **Précision par rapport à une classe**

La précision d'un classifieur par rapport à une certaine classe (autrement dit, par rapport à une certaine modalité de la variable à prédire), se mesure comme la proportion d'individus,

parmi tous ceux pour lesquels le classifieur a prédit cette classe, qui appartiennent réellement à celle-ci.

$$\text{Précision}_{\text{classe } c}(P_c) = \frac{m_{cc}}{\sum_i m_{ic}}$$

### ❖ Rappel par rapport à une classe

Le rappel d'un classifieur par rapport à une certaine classe se mesure, quant à lui, comme la proportion d'individus, parmi tous ceux qui appartiennent réellement à cette classe, pour lesquels le classifieur a prédit cette classe  $c$ .

$$\text{Rappel}_{\text{classe } c}(R_c) = \frac{m_{cc}}{\sum_i m_{ci}}$$

### ❖ F-mesure par rapport à une classe

On peut résumer les mesures de précision de rappel par rapport à une classe  $c$  en un seul indicateur, en calculant la moyenne harmonique :

$$F_{\text{classe } c} = \frac{P_c \times R_c}{P_c + R_c}$$

## 2.2 Le classifieur KPP

### 2.2.1 Choix de la valeur de K

Quelques règles sur le choix de  $k$  : Le paramètre  $k$  doit être déterminé par l'utilisateur :  $k \in \mathbb{N}$ . En classification binaire, il est utile de choisir  $k$  impair pour éviter les votes égalitaires. Le meilleur choix de  $k$  dépend du jeu de donnée. En général, les grandes valeurs de  $k$  réduisent l'effet du bruit sur la classification et donc le risque de sur-apprentissage, mais rendent les frontières entre classes moins distinctes. Il convient donc de faire un choix de compromis entre la variabilité associée à une faible valeur de  $k$  contre un 'oversmoothing' ou surlissage (i.e gommage des détails) pour une forte valeur de  $k$ . Un bon  $k$  peut être sélectionné par diverses techniques heuristiques, par exemple, de validation-croisée. Nous choisirons la valeur de  $k$  qui minimise l'erreur de classification. [18]

Nous avons appliqué l'algorithme KNN avec plusieurs valeurs du paramètre  $K$  et nous avons conclu que la valeur de  $K=4$  est le meilleur choix à faire pour notre base de données heart originale et modifiée.

## Chapitre 3 : Implémentation et Bilan

• Heart originale	• Heart Modifie
<ul style="list-style-type: none"> <li>Classification KNeighborsClassifier la valeur de k = 1                             <ul style="list-style-type: none"> <li>KNN Training Score: 1.0</li> <li>KNN Testing Score: 0.6557377049180327</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 2                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.78099173553719</li> <li>KNN Testing Score: 0.639344262295082</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 3                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.7727272727272727</li> <li>KNN Testing Score: 0.7049180327868853</li> </ul> </li> <li><b>Classification KNeighborsClassifier la valeur de k = 4</b> <ul style="list-style-type: none"> <li><b>KNN Training Score: 0.743801652892562</b></li> <li><b>KNN Testing Score: 0.7540983606557377</b></li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 5                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.743801652892562</li> <li>KNN Testing Score: 0.7049180327868853</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 6                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.7396694214876033</li> <li>KNN Testing Score: 0.6885245901639344</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 7                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.7355371900826446</li> <li>KNN Testing Score: 0.7213114754098361</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 8                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.7355371900826446</li> <li>KNN Testing Score: 0.7377049180327869</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 9                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.71900826446281</li> <li>KNN Testing Score: 0.7213114754098361</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>Classification KNeighborsClassifier la valeur de k = 1                             <ul style="list-style-type: none"> <li>KNN Training Score: 1.0</li> <li>KNN Testing Score: 0.22950819672131148</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 2                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.6446280991735537</li> <li>KNN Testing Score: 0.19672131147540983</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 3                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.5785123966942148</li> <li>KNN Testing Score: 0.2786885245901639</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 4                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.5165289256198347</li> <li>KNN Testing Score: 0.32786885245901637</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 5                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.4834710743801653</li> <li>KNN Testing Score: 0.21311475409836064</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 6                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.4793388429752066</li> <li>KNN Testing Score: 0.21311475409836064</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 7                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.44214876033057854</li> <li>KNN Testing Score: 0.2786885245901639</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 8                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.4380165289256198</li> <li>KNN Testing Score: 0.22950819672131148</li> </ul> </li> <li>Classification KNeighborsClassifier la valeur de k = 9                             <ul style="list-style-type: none"> <li>KNN Training Score: 0.4090909090909091</li> <li>KNN Testing Score: 0.21311475409836064</li> </ul> </li> </ul>

**Tableau 3.4 : Choix de la valeur de K pour le classifieur KPP**

### 2.2.2 Matrices de confusion du classifieur KPP

#### A. Base heart originale

		Classe prédite par le modèle	
		Classe 0	Classe 1
Classe reel	classe 0	<b>23</b>	5
	classe 1	10	<b>23</b>

**Tableau 3.5 : Matrice de confusion de « KPP » base heart originale**



	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
<b>0</b>	0.70	0.82	0.75	28
<b>1</b>	0.82	0.70	0.75	33
<b>accuracy</b>			0.75	61
<b>macro avg</b>	0.76	0.76	0.75	61
<b>weighted avg</b>	0.76	0.75	0.75	61

**Tableau 3.6 : Rapport de classification pour l'algorithme « KPP » base heart originale**

### B. Base heart modifiée

		Classe prédite par le modèle				
		Classe 0	Classe 1	Classe 2	Classe 3	Classe 4
Classe réel	classe 0	<b>7</b>	3	0	3	0
	classe 1	3	<b>3</b>	2	4	0
	classe 2	1	6	<b>4</b>	1	0
	classe 3	3	2	1	<b>4</b>	2
	classe 4	3	4	3	0	<b>2</b>

**Tableau 3.7 : Matrice de confusion de « KPP » base heart modifiée**

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
<b>0</b>	0.41	0.54	0.47	13
<b>1</b>	0.17	0.25	0.20	12
<b>2</b>	0.40	0.33	0.36	12
<b>3</b>	0.33	0.33	0.33	12
<b>4</b>	0.50	0.17	0.25	12
<b>accuracy</b>			0.33	61
<b>macro avg</b>	0.36	0.32	0.32	61
<b>weighted avg</b>	0.36	0.33	0.33	61

**Tableau 3.8 : Rapport de classification pour l'algorithme « KPP » base heart originale**

### Discussion :

- 1) **Base heart originale** : Les résultats de la précision du classifieur KPP sont : 70% pour la classe 0 et 82% pour la classe 1. La précision globale du classification KPP est 75% ce qui suggère que le modèle permet une bonne prédiction de la maladie.
- 2) **Base heart modifiée** : Les résultats de la précision du classifieur KPP sont : 41% pour la classe 0, et de 17% pour la classe 1, 40% pour la classe 2, 33% pour la classe 3 et de 50% pour la classe 4. La précision globale du classification KPP est 33%.
  - Malgré que ces résultats ne semblent pas satisfaisants pour un classifieur. Nous nous n'intéressons pas à l'interprétation de ces résultats car nous sommes conscients que nous avons appliqué un changement aléatoire sur les classes sans étude préalable uniquement pour avoir une variété de classes pour l'étape de la migration entre classes. Avoir une bonne prédiction pour ces données c'est un problème.

## 2.3 Le classifieur AD

### 2.3.1 Choix de la bonne taille de l'arbre

Il n'est pas toujours souhaitable en pratique de construire un arbre dont les feuilles correspondent à des sous-ensembles parfaitement homogènes du point de vue de la variable-cible. Plus le modèle est complexe (plus l'arbre est grand, plus il a de branches, plus il a de feuilles), plus l'on court le risque de voir ce modèle incapable d'être extrapolé à de nouvelles données, c'est-à-dire de rendre compte de la réalité que l'on cherche à appréhender.

Pour notre base heart nous avons appliqué l'algorithme AD en variant la profondeur de l'arbre de 1 jusqu'à 13 et nous avons trouvé que la profondeur 5 est la meilleure.

# Chapitre 3 : Implémentation et Bilan

Heart Originale	Heart modifiée
<ul style="list-style-type: none"> <li>La classification par Arbre de Decision : max-depth = 2               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.756198347107438</li> <li>Arbre de Decision Testing Score: 0.7704918032786885</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 3               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.8471074380165289</li> <li>Arbre de Decision Testing Score: 0.7868852459016393</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 4               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.8966942148760331</li> <li>Arbre de Decision Testing Score: 0.7377049180327869</li> </ul> </li> <li><b>La classification par Arbre de Decision : max-depth = 5</b> <ul style="list-style-type: none"> <li><b>Arbre de Decision Training Score: 0.9338842975206612</b></li> <li><b>Arbre de Decision Testing Score: 0.7377049180327869</b></li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 6               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.9710743801652892</li> <li>Arbre de Decision Testing Score: 0.7377049180327869</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 7               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.987603305785124</li> <li>Arbre de Decision Testing Score: 0.7704918032786885</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 8               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.9958677685950413</li> <li>Arbre de Decision Testing Score: 0.7868852459016393</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 9               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.9958677685950413</li> <li>Arbre de Decision Testing Score: 0.8032786885245902</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 10               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 1.0</li> <li>Arbre de Decision Testing Score: 0.7213114754098361</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 11               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 1.0</li> <li>Arbre de Decision Testing Score: 0.7704918032786885</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 12               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 1.0</li> <li>Arbre de Decision Testing Score: 0.8032786885245902</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>La classification par Arbre de Decision : max-depth = 2               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.40082644628099173</li> <li>Arbre de Decision Testing Score: 0.3442622950819672</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 3               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.45041322314049587</li> <li>Arbre de Decision Testing Score: 0.32786885245901637</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 4               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.5</li> <li>Arbre de Decision Testing Score: 0.3770491803278688</li> </ul> </li> <li><b>La classification par Arbre de Decision : max-depth = 5</b> <ul style="list-style-type: none"> <li><b>Arbre de Decision Training Score: 0.5909090909090909</b></li> <li><b>Arbre de Decision Testing Score: 0.36065573770491804</b></li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 6               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.6652892561983471</li> <li>Arbre de Decision Testing Score: 0.36065573770491804</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 7               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.7396694214876033</li> <li>Arbre de Decision Testing Score: 0.36065573770491804</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 8               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.8388429752066116</li> <li>Arbre de Decision Testing Score: 0.2786885245901639</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 9               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.8925619834710744</li> <li>Arbre de Decision Testing Score: 0.32786885245901637</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 10               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.9504132231404959</li> <li>Arbre de Decision Testing Score: 0.32786885245901637</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 11               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.9793388429752066</li> <li>Arbre de Decision Testing Score: 0.29508196721311475</li> </ul> </li> <li>La classification par Arbre de Decision : max-depth = 12               <ul style="list-style-type: none"> <li>Arbre de Decision Training Score: 0.987603305785124</li> <li>Arbre de Decision Testing Score: 0.29508196721311475</li> </ul> </li> </ul>

**Tableau 3.9 : Choix de la bonne taille de l'AD**

## 2.3.2 Matrices de confusion du classifieur AD

### A. Base heart originale

		Classe prédite par le modèle	
		Classe 0	Classe 1
Classe réel	classe 0	<b>19</b>	9
	classe 1	9	<b>24</b>

**Tableau 3.10 : Matrice de confusion de « AD » base heart originale**

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
<b>0</b>	0.68	0.68	0.68	28
<b>1</b>	0.73	0.73	0.73	33
<b>accuracy</b>			0.70	61
<b>macro avg</b>	0.70	0.70	0.70	61
<b>weighted avg</b>	0.70	0.70	0.70	61

**Tableau 3.11 : Rapport de classification pour l'algorithme « AD » base heart originale**

### B. Base heart modifiée

		Classe prédite par le modèle				
		Classe 0	Classe 1	Classe 2	Classe 3	Classe 4
Classe réel	classe 0	<b>4</b>	2	2	2	3
	classe 1	1	<b>8</b>	2	1	0
	classe 2	2	4	<b>3</b>	3	0
	classe 3	0	7	1	<b>2</b>	2
	classe 4	0	2	2	3	<b>5</b>

**Tableau 3.12 : Matrice de confusion de « AD » base heart modifiée**

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
<b>0</b>	0.57	0.31	0.40	13
<b>1</b>	0.35	0.67	0.46	12
<b>2</b>	0.30	0.25	0.27	12
<b>3</b>	0.18	0.17	0.17	12
<b>4</b>	0.50	0.42	0.45	12
<b>accuracy</b>			0.36	61
<b>macro avg</b>	0.38	0.36	0.35	61
<b>weighted avg</b>	0.38	0.36	0.35	61

**Tableau 3.13 : Rapport de classification pour l'algorithme « AD » base heart originale**

### Discussion :

- 1) **Base heart originale** : Les résultats de la précision du classifieur AD sont : 68% pour la classe 0 et 73% pour la classe 1. La précision globale du classification AD est 70% ce qui suggère que le modèle permet une bonne prédiction de la maladie.
- 2) **Base heart modifiée** : Les résultats de la précision du classifieur AD sont : 57% pour la classe 0, et de 35% pour la classe 1, 30% pour la classe 2, 18% pour la classe 3 et de 50% pour la classe 4. La précision globale du classification AD est 36%.
  - Pareil que pour le KPP, malgré ces résultats ne semble pas satisfaisant pour un classifieur. Nous nous n'intéressons pas à l'interprétation de ces résultats car nous sommes conscients que nous avons appliqué un changement aléatoire sur les classes sans étude préalable uniquement pour avoir une variété de classes pour l'étape de la migration entre classes. Avoir une bonne prédiction pour ces données c'est un problème.

## 2.4 La migration entre classes

Avant de commencer cette partie, il est très important de signaler que l'avis du corps médical dans cette étape aurait pu donner une meilleure explication à l'idée de la migration entre classes. Certains paramètres ne peuvent pas être changés alors que dans notre application nous permettons ces changements. L'application à un domaine moins sensible comme le domaine commercial permet une meilleure valorisation. C'est l'idée d'avoir plusieurs classes qui nous a poussé à choisir cette base de données et l'inexistence de base de données avec plusieurs classes dans les benchmark disponibles.

### 2.4.1 La migration entre classes par KPP

Nous avons décidé d'appliquer notre algorithme de migration sur l'ensemble des données d'apprentissage et calculer le pourcentage de passage entre toutes les classes.

Nous avons décidé de ne maintenir que les deux paramètres *age* et *sex*.

		Migration vers la classe	
		Classe 0	Classe 1
Classe prédite	classe 0	-	65%
	classe 1	73%	-

**Tableau 3.14 : Pourcentage des possibilités de migration en utilisant KPP pour la base heart originale**

		Migration vers la classe				
		Classe 0	Classe 1	Classe 2	Classe 3	Classe 4
Classe prédite	classe 0	-	57%	38%	46%	51%
	classe 1	46%	-	41%	61%	34%
	classe 2	46%	54%	-	49%	28%
	classe 3	55%	64%	31%	-	50%
	classe 4	76%	54%	30%	68%	-

**Tableau 3.15 : Pourcentage des possibilités de migration en utilisant KPP pour la base heart modifiée**

### 2.4.2 La migration entre classes par AD

Nous avons décidé d'appliquer notre algorithme de migration sur l'ensemble des données d'apprentissage et calculer le pourcentage de passage entre toutes les classes.

Nous avons décidé de ne maintenir que les deux paramètres *age* et *sex*.

		Migration vers la classe	
		Classe 0	Classe 1
Classe prédite	classe 0	-	9%
	classe 1	14%	-

**Tableau 3.16 : Pourcentage des possibilités de migration en utilisant AD pour la base heart originale**

## Chapitre 3 : Implémentation et Bilan

		Migration vers la classe				
		Classe 0	Classe 1	Classe 2	Classe 3	Classe 4
Classe rprédite	classe 0	-	8%	2%	2%	8%
	classe 1	7%	-	8%	5%	7%
	classe 2	3%	13%	-	13%	3%
	classe 3	2%	7%	10%	-	10%
	classe 4	14%	11%	3%	11%	-

**Tableau 3.17 : Pourcentage des possibilités de migration en utilisant AD pour la base heart modifiée**

		Migration vers la classe	
		Classe 0	Classe 1
Classe prédite	classe 0	-	100%
	classe 1	100%	-

**Tableau 3.18 : Pourcentage des possibilités de migration en utilisant AD pour la base heart originale**

		Migration vers la classe				
		Classe 0	Classe 1	Classe 2	Classe 3	Classe 4
Classe rprédite	classe 0	-	100%	100%	100%	100%
	classe 1	100%	-	100%	100%	100%
	classe 2	100%	100%	-	100%	100%
	classe 3	100%	100%	100%	-	100%
	classe 4	100%	100%	100%	100%	-

**Tableau 3.19 : Pourcentage des possibilités de migration en utilisant AD pour la base heart modifiée**

## Chapitre 3 : Implémentation et Bilan

Nous avons décidé de ne maintenir les cinq paramètres *age, sex, cp, ca et thal* :

		Migration vers la classe	
		Classe 0	Classe 1
Classe prédite	classe 0	-	100%
	classe 1	83%	-

**Tableau 3.20 : Pourcentage des possibilités de migration en utilisant AD pour la base heart originale**

		Migration vers la classe				
		Classe 0	Classe 1	Classe 2	Classe 3	Classe 4
Classe rprédite	classe 0	-	35%	100%	100%	68%
	classe 1	23%	-	32%	52%	12%
	classe 2	78%	89%	-	81%	43%
	classe 3	55%	93%	55%	-	10%
	classe 4	100%	15%	100%	100%	-

**Tableau 3.19 : Pourcentage des possibilités de migration en utilisant AD pour la base heart modifiée**

### Discussion :

On constate que la technique des arbres de décisions donne des pourcentages plus élevés que la technique du plu proche voisin. Ceci peut être expliqué par le fait que les arbres de décision donne des prédictions basées sur des intervalles de valeurs alors que la technique du KPP est basée sur l'égalité des paramètres.

Le choix de la méthode à utiliser pour la migration entre classes va dépendre du domaine d'application et à quel point le changement d'un paramètre est critique par rapport à la décision finale.



### 3. Outils et langage utilisés

#### 3.1. Les outils

- Google Colab : ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur « Jupyter Notebook » et destiné à la formation et à la recherche dans l'apprentissage automatique.

Cette plateforme permet d'entraîner des modèles de Machine

Learning directement dans le cloud. Sans donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur. Cool, n'est-ce pas ? Avant de présenter ce magnifique service, nous rappellerons ce qu'est un Jupyter Notebook



```
from google.colab import drive
```

- Jupyter Notebook: l'outil incontournable du scientifique des données Jupyter Notebook est une application Web Open Source permettant de créer et de partager des documents contenant du code (exécutable



directement dans le document), des équations, des images et du texte. Avec cette application il est possible de faire du traitement de données, de la modélisation statistique, de la visualisation de données, du Machine Learning, etc. Elle est disponible par défaut dans la distribution Anaconda (suivre ce lien pour savoir comment l'installer). [19].

- PyCharm : est un environnement de développement intégré utilisé pour programmer en Python. Il permet l'analyse de code et contient un débogueur Graphique. Il permet également la gestion des tests unitaires, l'intégration de logiciel de gestion de versions, et supporte le développement web avec Django.



Développé par l'entreprise tchèque JetBrains, c'est un logiciel multiplateforme qui fonctionne sous Windows, Mac OS X et Linux. Il est décliné en édition professionnelle, diffusé sous licence propriétaire, et en édition communautaire diffusé sous licence Apache [20].

### 3.2. Les langages

- **Python :**

Est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est facile à interpréter, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications dans de nombreux domaines et sur la plupart des plateformes.

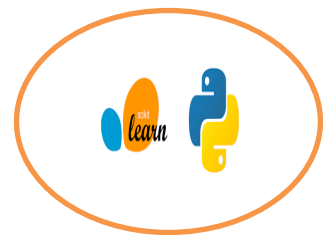


L'interpréteur Python et sa vaste bibliothèque standard sont disponibles librement, sous forme de sources ou de binaires, pour toutes les plateformes majeures depuis le site Internet <https://www.python.org/> et peuvent être librement redistribués. Ce même site distribue et pointe vers des modules, des programmes et des outils tiers. Enfin, il constitue une source de documentation. Avec Python, je me suis appuyé sur un Framework et les bibliothèques Python :

- ❖ **Flask :** Ce Framework permet de développer des serveurs Web backend basés sur Python. Ce Framework est considéré comme le plus populaire et le meilleur pour les débutants.

```
from flask import Flask, render_template,
```

- ❖ **Scikit-Learn :** Est une bibliothèque qui fournit une gamme d'algorithmes D'apprentissage supervisés et non supervisés via une interface cohérente en Python. La vision de la bibliothèque est un niveau de robustesse et de support requis pour une utilisation dans les Systèmes de production.



Cela signifie qu'il faut se concentrer sur des préoccupations telles que la simplicité d'utilisation, la qualité du code, la collaboration, la documentation et les performances [21].

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import confusion_matrix
from sklearn import metrics
```

```
from sklearn.externals import joblib
```

## Chapitre 3 : Implémentation et Bilan

- ❖ **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. Pandas est un logiciel libre sous licence.

Ce ne sont pas toutes les bibliothèques utilisées dans notre projet. Nous avons utilisé

```
import pandas as pd
```

D'autres bibliothèques comme : Numpy, Matplotlib, scipy, Seaborn [22].

- ❖ **Numpy** : est une extension du langage de programmation Python, Destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que des fonctions mathématiques opérant sur ces tableaux.



```
import numpy as np
```

### 4. Présentation de l'application :

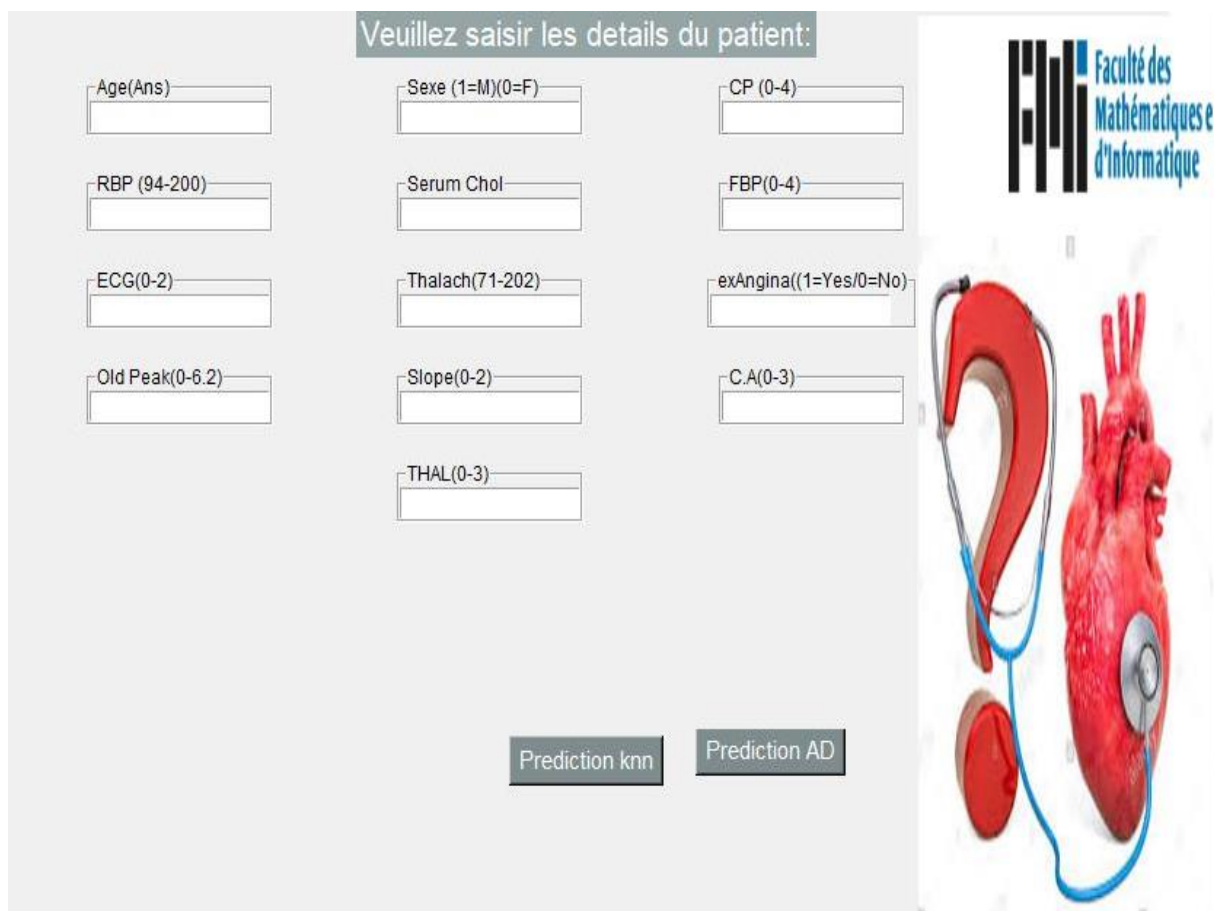
The screenshot shows a web form titled 'Veillez saisir les détails du patient:'. It contains 13 input fields arranged in a grid. The fields are: Age(Ans), Sexe (1=M)(0=F), CP (0-4), RBP (94-200), Serum Chol, FBP(0-4), ECG(0-2), Thalach(71-202), exAngina((1=Yes/0=No)), Old Peak(0-6.2), Slope(0-2), C.A(0-3), and THAL(0-3). At the bottom, there are two buttons: 'Prediction knn' and 'Prediction AD'. On the right side, there is a logo for 'Faculté des Mathématiques et d'Informatique' and an image of a human heart with a stethoscope.

Figure 3.1 : Formulaire de test des maladies cardiaques.

## Chapitre 3 : Implémentation et Bilan

---

Cette figure (figure 3.1) montre le menu principal (la première fenêtre) de l'application. Elle se compose de trois boutons : le premier (**accueil**) Contient des informations sur les maladies cardiaques. et le deuxième (**services**) Ici, nous faisons une prédiction des maladies cardiaques en utilisant les méthodes mentionnées précédemment et le troisième bouton est (à propos de).

Le bouton (prediction) donne à l'utilisateur la possibilité de tester de nouveaux individus.

Le formulaire Au-dessus donne la main à l'utilisateur pour remplir les informations d'un patient et prédire à quelle classe est-il associé (malade {1.2.3.4}, ou sain {0}). Selon les 4 méthodes le résultat s'affiche sur l'écran.

Ensuite, nous remplissons les champs d'attributs et cliquons sur l'une des méthodes utilisées pour la prédiction.

### 5. Conclusion

Ce dernier chapitre présente une étude liée à la mise en œuvre des différents algorithmes utilisés en plus des algorithmes proposés pour la migration entre classes.

Cette étude nous a permis de faire une validation et une évaluation des performances de chacune des méthodes présentées.

# Conclusion Générale

## Conclusion générale et perspectives

---

### Conclusion générale :

De nombreuses maladies chroniques peuvent être contrôlées ou leurs effets diminués par la prévention ou la gestion des facteurs de risque. Identifier les facteurs qui peuvent influencer sur l'évolution de la maladie et connaître le rôle qu'ils jouent dans la maladie peut aider à en tenir compte et à être vigilant. L'apprentissage automatique s'avère efficace pour aider à prendre des décisions et des prévisions à partir de la grande quantité de données produites par l'industrie des soins de santé.

Le secteur de la santé a principalement profité de l'utilisation des techniques de prédiction d'apprentissage automatique. L'objectif principal est la proposition d'outils d'aide à la décision visant à améliorer la qualité des soins aux patients et la prédiction de maladies.

Il est difficile de déterminer manuellement les chances de contracter une maladie cardiaque en fonction des facteurs de risque. L'apprentissage automatique s'avère efficace pour aider à prendre des décisions et des prévisions à partir de la grande quantité de données produites par l'industrie des soins de santé.

Dans notre travail, nous nous sommes intéressés à l'application de techniques d'apprentissages et la comparaison de leurs efficacités. Nous avons donc appliqué ces techniques d'apprentissages à savoir : KPP et AD sur la base de données Cleveland Heart Disease dataset, dont l'objectif de ces techniques et la prédiction pour un nouvel patient (en fonction de certains paramètres) si ce patient souffre d'une maladie cardiaque (classée en 4 types) ou non.

Cette variété de type de maladie nous à pousser à réfléchir à la possibilité de migration (passage) d'une classe à une autre. Autrement dit, avec les informations réelles que nous disposons d'une personne donnée notre système permet de prédire le type de sa maladie cardiovasculaire dont elle souffre : par exemple de type3. Généralement les gens ont un autre souci est de savoir quels changements doivent-ils faire pour être ou ne pas être dans une autre classe (type de maladie). Cette migration (passage) entre classe peut être vue comme une perspective d'amélioration de l'état d'un malade ou malheureusement par une dégradation (complication) de son état de santé. L'objectif principal est toujours ralentissement de l'évolution vers une forme invalidante de la maladie. L'idée est donc *d'intervenir* le plus tôt possible et de manière soutenue afin de prévenir l'aggravation et la sévérité de la maladie.

## Conclusion générale et perspectives

---

Nous avons proposé pour chaque technique un algorithme de calcul de nouvelle valeur des paramètres pour la migration d'une classe à une autre.

Nous avons réalisé une application -web-, ce qui est très pratique pour les médecins ainsi que les patients qui souhaite suivre leurs états de santé ainsi que de voir l'impact des changements des paramètres aussi minimaux qu'ils soient sur leur état de santé (intervenir tôt pour prévenir l'aggravation).

### Perspectives

Ce projet nous permis d'aborder le domaine d'apprentissage et son application au domaine médical « prédiction de maladies cardiaques ». Nous sommes très motivés au futures perspectives de notre travail qui sont nombreux :

- Introduire d'autres techniques de classification et comparer pour ne garder que ceux qui donnent de meilleures précisions
- Introduire plus de paramètres. D'autres informations peuvent être utilisées pour obtenir des résultats plus précis, tels que des données liées à l'analyse des maladies cardiovasculaires, telles qu'un ECG, une échocardiographie ou une coronarographie pour des recherches ultérieures.
- Collaboration avec des médecins (experts) pour valider pratiquement ces résultats.

Notre projet vise à fournir une plate-forme Web pour prédire l'apparition de maladies en fonction de divers symptômes. L'utilisateur peut identifier divers symptômes et maladies peuvent être trouvées par leurs nombres de probabilité. Notre Application-web peut être améliorée :

- En mettant en œuvre une proposition de médicament aux patients en fonction des résultats.
- Une option de chat en direct peut également être mise en œuvre où le patient peut discuter avec le médecin disponible concernant Médicament.
- Notre projet peut être utilisé comme une trousse d'outils de formation pour les infirmières et les médecins nouvellement introduits dans le domaine de la cardiologie.
- Le patient peut choisir les médicaments à prendre pour une vie plus saine. De plus, s'il est mis en œuvre à grande échelle, il peut être utilisé dans des établissements médicaux tels que des hôpitaux et des cliniques où le patient n'aura pas à attendre

## Conclusion générale et perspectives

---

longtemps pour se faire soigner. S'il présente des symptômes liés à une maladie cardiaque.

- À l'avenir, nous pouvons également ajouter des comptes pour chaque utilisateur, puis l'enregistrement de contrôle précédent de l'état cardiaque de l'utilisateur peut être surveillé pour voir s'il y a une amélioration ou si l'état s'est détérioré.



## Résumé :

Les maladies cardiaques ont un pacte d'attention abondant dans la recherche médicale en raison de leur impact sur la santé de l'homme. Les maladies cardiaques sont parmi les principales causes de décès. La *prévention*, la *prédiction* et la gestion des maladies chroniques est une démarche en matière de soins de santé qui vise à aider les personnes qui en sont atteintes à maintenir leur autonomie et à demeurer en aussi bonne santé que possible grâce à la détection précoce de ces maladies ainsi qu'à leur prévention et à leur gestion. Généralement les gens ont un autre souci est de savoir quels changements doivent-ils faire pour être ou ne pas être dans une autre classe (type de maladie). Cette migration (passage) entre classe peut être vue comme une perspective d'amélioration de l'état d'un malade ou malheureusement par une dégradation (complication) de son état de santé.

Dans ce travail, des algorithmes d'apprentissage automatique supervisé, à savoir KPP et les AD sont utilisés pour prédire les maladies cardiaques. Nous avons proposé pour chaque technique un algorithme de calcul de nouvelle valeur des paramètres pour la migration d'une classe à une autre.

**Mots clés :** KPP, AD, maladies cardiaques, fouille de données, migration entre classes.

### الملخص:

لأمراض القلب ميثاق اهتمام وفير في البحث الطبي لما لها من تأثير على صحة الإنسان. مرض القلب هو أحد الأسباب الرئيسية للوفاة. الوقاية والتنبؤ وإدارة الأمراض المزمنة هي نهج رعاية صحية يهدف إلى مساعدة الأشخاص المصابين بأمراض مزمنة على الحفاظ على استقلاليتهم والبقاء بصحة جيدة قدر الإمكان من خلال الكشف المبكر عن هذه الأمراض والوقاية منها ومعالجتها. عادة ما يكون لدى الناس قلق آخر بشأن التغييرات التي يجب عليهم إجراؤها أو عدم التواجد في فئة أخرى (نوع المرض). يمكن النظر إلى هذه الهجرة (العبور) بين الطبقات على أنها احتمالية لتحسين حالة المريض أو لسوء الحظ من خلال تدهور (مضاعفات) حالته الصحية.

في هذا العمل ، يتم استخدام خوارزميات الإعلام الآلي الخاضعة للإشراف ، وبالتحديد KPP و AD للتنبؤ بأمراض القلب. لقد اقترحنا لكل تقنية خوارزمية لحساب قيم المعلمات الجديدة للترحيل من فئة إلى أخرى.

**الكلمات المفتاحية:** KPP - ، AD، أمراض القلب ، التقيب في البيانات ، الهجرة من فئة إلى أخرى.

### Abstract :

Heart disease has an abundant attention pact in medical research because of its impact on human health. Heart disease is one of the leading causes of death. The prevention, prediction and management of chronic disease is a health care approach that aims to help people with chronic disease maintain their independence and stay as healthy as possible through the early detection of these diseases. diseases and their prevention and management. Usually people have another worry about what changes should they make to be or not to be in another class (type of disease). This migration (passage) between classes can be seen as a prospect of improving the condition of a patient or unfortunately by a deterioration (complication) of his state of health.

In this work, supervised machine learning algorithms, namely KPP and ADs are used to predict heart disease. We have proposed for each technique an algorithm for calculating new parameter values for the migration from one class to another.

**Keywords :** KPP, AD, heart disease, data mining, class-to-class migration.

## ***Référence :***

- [1] : <https://www.microstrategy.com/us/resources/introductory-guides/data-mining-explained>.
- [2] J. Han, M. Kamber, and J. Pei. Data mining : concepts and techniques. Morgan Kaufmann Pub, 2011.
- [3] M.Kantardzic. Data mining : concepts, models, methods, and algorithms. WileyInterscience, 2003. [3] P. Preux. Fouille de données, notes de cours. Disponible sur internet, 2006
- [4] G. Huang, S. Song, J. Gupta, and C. Wu, Semi-supervised and unsupervised extreme learning machines, Cybernetics, IEEE Transactions, 44(12): 24052417, Dec 2014.
- [5] Parsaye K: Surveying Decision Support: New Realms of Analysis. Information Discovery, 1996.
- [6] J.A. Michael, Gordon ET S. Linoff: Data Mining: Techniques appliquées au marketing, à la vente et aux services clients, 1997.
- [7] Trevor Hastie, Robert Tibshirani, Jérôme Friedman : the éléments of statistical Learning (data mining.inference and prediction) 2ème Edition.
- [8] M.Usama, Fayyad, P.S. Gregory et S. Padhraic: From data mining to knowledge discovery: An overview. In Advances in Knowledge Discovery and Data Mining, pages 1–34. 1996.
- [9] Dr. Abdelhamid DJEFFAL Site web : [www.abdelhamid-djeffal.net](http://www.abdelhamid-djeffal.net) explained. Pierre-Louis, Gonzalez, Méthodes de classification <https://fr.scribd.com/doc/263239855/Classification-2008-2>
- [10] W. Frawley, J. Piatetsky-Shapiro, Gregory et M. Christopher: Knowledge Discovery in Databases: An Overview, 1992.
- [11] Parsaye K : Surveying Decision Support: New Realms of Analysis. Information Discovery, 1996.
- [12] J.A. Michael, Gordon ET S. Linoff: Data Mining: Techniques appliquées au marketing, à la vente et aux services clients, 1997.
- [13] “Heart Disease Data Set, UCI Machine Learning Repository”.
- URL: [http://archive.ics.uci.edu/ml/datasets/Heart Disease](http://archive.ics.uci.edu/ml/datasets/Heart+Disease)”

- [14] “Heart Disease Data Set of kaggle“.
- URL: <https://www.kaggle.com/ronitf/heart-disease-uci>, Consulté Mai 2020.
- [15] URL:<http://apiacoa.org/publications/teaching/data-mining/m2p6/supervised-slides.pdf>.
- URL:<http://www.bioinfo-biostats-etudiants.universite-parisaclay.fr/Ressources/Cours/Master%202/ECT/>.
- [16] Angina (Chest Pain) | American Heart Association.” [Online]. Available:
- URL: <https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain>.
- [17] J. Stamler and R. Stamler, “Intervention for the prevention and control of hypertension And atherosclerotic diseases: United States and international experience,” Am. J. Med., 1984.
- [18] N. Sarwar et al. “Diabetes mellitus, fasting blood glucose concentration, and risk of Vascular disease: A collaborative meta-analysis of 102 prospective studies,” Lancet, 2010.
- [19] M. A. Westwood et al. “Normalized left ventricular volumes and function in Thalassemia major patients with normal myocardial iron,” J. Magn. Reson. Imaging, 2007.
- [20] Prof. Priya R. Patil, Prof. S. A. Kinariwala” Marathwada Institute of “Technology. Automated Diagnosis of Heart Disease using Random Forest Algorithm. 2017.
- URL : <https://www.ijariit.com/manuscripts/v3i2/V3I2-1197.pdf>.
- [21] Eve Mathieu-Dupas. Algorithme des k plus proches voisins pondérés et application en diagnostic.42èmes Journées de Statistique, 2010, Marseille, France, France.
- [22] URL : <https://ledatascientist.com/google-colab-le-guide-ultime/>.