الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université de 8 Mai 1945 – Guelma –

Faculté des Mathématiques, d'Informatique et des Sciences de la matière

Département d'Informatique

Mémoire de Fin d'études Master

Filière: Informatique

Option : Systèmes informatiques

Thème :

Système intelligent de prédiction des maladies cardiaques

Encadré Par :

Dr. GUERROUI NADIA

Présenté par :

BOUTARFA MOHAMMED AMIN

Juin 2022

Remerciements

En premier lieu, je rends grâce à Dieu, le Tout-Puissant à qui j'exprime ma gratitude pour m'avoir donné, le courage et la patience et la volonté de réaliser ce travail.

Je tiens à remercier sincèrement Mme GUEROUI NADIA pour son encadrement et son soutien et ses conseils précieux tout au long de ma recherche.

Mes remerciements vont également aux membres de jury pour l'intérêt qu'ils ont porté à mon projet en acceptant D'examiner mon travail et de l'enrichir de leurs propositions.

Évidemment, je n'oublie pas de remercier ma famille, mes amis et mes collègues et tous ceux qui, avec leur aide, leurs conseils, leurs encouragements, m'ont aidé à mener ce travail à terme.

Résumé

Les maladies cardiaques sont considérées comme l'une des maladies les plus dangereuses car elles menacent l'organe dont dépend la vie humaine, et puisque le diagnostic des maladies dépend de tests médicaux et de certains indicateurs, le diagnostic des maladies cardiaques est considéré comme le plus complexe et le plus difficile, en raison du grand nombre d'analyses et d'indicateurs dans ce diagnostic, qui est ce qui consomme beaucoup Cela a conduit de nombreux chercheurs à s'engager dans des techniques d'exploration de données et d'apprentissage automatique pour prédire cette maladie.

Ce qui nous a incités à créer un système qui nous permet de prédire les patients cardiaques sur la base d'une base de données de patients diagnostiqués. Et c'est en utilisant des techniques d'apprentissage automatique et en appliquant la méthode ACP pour traiter les données et les préparer pour la classification via l'algorithme KNN, ce qui nous a permis d'atteindre une exactitude de 97.83%.

Mots-clés : Apprentissage automatique, maladies cardiaques, maladie cardio-vasculaire, prédiction, classification .

Abstract

Heart diseases are considered as one of the most dangerous diseases as it threatens the organ on which human life depends, and as the diagnosis of diseases depends on medical tests and certain indicators. The cardiac disease diagnosis is considered to be the most complex and difficult, due to the large number of analyzes and indicators in this diagnosis, which has led many researchers to use data mining and machine learning techniques to be able to predict this disease.

This prompted us to create a system that allows us to predict cardiac patients from a database of diagnosed patients. This was achieved by using machine learning techniques and applying the PCA method for data processing and preparing the collected data for classification by the KNN algorithm, which allowed us to achieve an accuracy of 97.83%.

Keywords : Machine learning, heart disease, prediction, classification.

Table des matières

Liste des tableaux

Liste des figures

Introduction

De nos jours, les maladies cardiaques sont l'un des problèmes les plus importants qui menacent la vie humaine. Elles sont la principale cause de décès dans le monde, tuant environ 17,9 millions de personnes chaque année [1]. Afin de réduire le nombre de décès dus à des maladies cardiovasculaires, il faut identifier les personnes à risques de développer une maladie cardiaque le plus tôt possible et à s'assurer qu'elles reçoivent le traitement approprié, ce qui pourrait éviter des décès prématurés, et le diagnostic de la maladie nécessite de nombreux tests : la pression artérielle, le glucose, les signes vitaux, les douleurs thoraciques, les électrocardiogrammes, la fréquence cardiaque maximale...etc. En plus des analyses, il y a également des études cliniques, des antécédents de patients et des réponses à leurs questions.

Pour faire tous ces diagnostics, des erreurs se produisent souvent ou des tests de diagnostic sont retardés. De plus, il est également plus coûteux et demande beaucoup de calculs, et les évaluations prennent du temps. Dans ce contexte, de nombreuses recherches et études ont été menées dans le but de prédire les maladies cardiaques, en utilisant des techniques d'intelligence artificielle, y compris l'apprentissage automatique basées sur les données du patient, ses antécédents médicaux et un ensemble d'analyses.

Notre objectif principal dans ce projet est de proposer une approche d'apprentissage automatique pour prédire les maladies cardiaques à partir de l'utilisation des données des patients et pour atteindre notre objectif, nous avons mis en œuvre l'approche proposée à travers l'implémentation d'une application. Notre travail est présenté en trois chapitres principaux : Le premier chapitre présente les maladies cardiaques avec leurs types, symptômes et causes, et dans le même chapitre, nous indiquons le principe de l'apprentissage automatique, et nous décrivons également les recherches les plus récentes dans ce contexte. Dans le deuxième chapitre, nous détaillons la conception et l'architecture de notre système proposé qui vise à prédire les maladies cardiaques. Dans le troisième chapitre, nous évaluons les résultats de notre système proposé et nous discutons les résultats ainsi que les étapes de la mise en œuvre de l'application. Enfin, une

conclusion générale souligne l'intérêt de l'approche, met en évidence ses limites et propose quelques perspectives de recherches futures basées sur les principaux résultats.

Chapitre 1

État de l’Art

1.1 Introduction

La maladie cardiovasculaire est un ensemble de troubles qui touchent le cœur et les vaisseaux sanguin, et comme ils sont devenus une menace pour la vie de nombreuses personnes à travers le monde, les chercheurs se sont concentrés sur les façons de réduire leur risque ou de prédire leur apparition, et probablement les plus importantes de ses recherches sont celles liées à la technologie et plus particulièrement à l'apprentissage automatique.

Dans ce chapitre, nous en apprenons davantage sur les maladies cardiaques en identifiant leurs types, causes et symptômes. Ensuite nous aborderons les concepts d'apprentissage automatique, et les travaux effectués pour prédire les maladies cardiaques à l'aide de différents algorithmes de l'apprentissage automatique.

1.2 Les maladies cardiaques

Le cœur est l'organe responsable de pomper du sang vers les différentes parties du corps à travers les vaisseaux sanguins. S'il y a un problème dans le travail de pompage du cœur, les principaux organes du corps tels que le cerveau et les reins sont affectés négativement, mais si le cœur cesse de fonctionner, la mort de la personne survient en quelques minutes, ainsi, beaucoup de recherches scientifiques ont été menées pour réduire le risque de maladie cardiaque pour la vie humaine[2], incluant l'utilisation de techniques d'apprentissage automatique pour prédire la probabilité de développer ces maladies.

1.2.1 Types des maladies cardiaques

Les maladies cardiovasculaires (MCV) englobent plusieurs types de troubles de l'appareil circulatoire, soit les maladies congénitales, ils incluent dans[1]:

- *Maladie coronarienne* : une maladie des vaisseaux sanguins alimentant le muscle cardiaque.
- *Maladie cérébro-vasculaire:* une maladie des vaisseaux sanguins alimentant le cerveau.
- *Maladie artérielle périphérique* : une maladie des vaisseaux sanguins alimentant les bras et les jambes.

- *Cardiopathie rhumatismale* : dommages au muscle cardiaque et aux valves cardiaques dus au rhumatisme articulaire aigu, causé par des bactéries streptococciques.
- *Cardiopathie congénitale* : malformations congénitales qui affectent le développement et le fonctionnement normaux du cœur causé par des malformations de la structure cardiaque dès la naissance.
- *Thrombose veineuse profonde et embolie pulmonaire* : caillots sanguins dans les veines des jambes, qui peuvent se déloger et se déplacer vers le cœur et les poumons.

1.2.2 Causes des maladies cardiaques

Outre les causes congénitales, les principales causes des maladies cardiovasculaires sont l'hypertension artérielle, le diabète, les lipides sanguins, le tabagisme, une mauvaise alimentation, le manque d'activité physique et le surpoids [3].

1. **Hypertension :**le risque de MCV augmente lorsque la tension artérielle sur les parois des vaisseaux sanguins est trop haute donc le traitement de la haute pression sanguine réduit l'incidence de la plupart des MCV et le taux de mortalité toutes causes confondues.

2. **Diabète :** Le diabète accroît de manière significative l'incidence des maladies cardiovasculaires, donc les patients diabétiques sont plus susceptibles de développer une maladie cardiovasculaire parce que le diabète et les maladies cardiovasculaires sont associés à de nombreux facteurs de risque, donc le traitement du diabète joue un rôle important dans la prévention des maladies cardiovasculaires.

3. **Dyslipidémie :** La dyslipidémie se définit par un niveau anormalement faible ou élevé de graisse dans la circulation sanguine, cette maladie est associée à l'augmentation du risque de maladies cardiovasculaires et l'abaissement du taux de cholestérol est l'une des stratégies les plus efficaces pour prévenir les maladies cardiovasculaires.

4. **Tabagisme :** On sait que le tabagisme constitue un facteur de risque majeur de maladies cardiovasculaires, il existe également une proportion directe entre le nombre de cigarettes fumées par jour et le risque de maladie cardiovasculaire et l'usage du tabac double le taux de mortalité cardio-vasculaire, donc arrêter de fumer est le moyen le plus efficace de prévenir les maladies cardiovasculaires.

5. **Alimentation malsaine :** Le régime alimentaire affecte le risque de MCV en affectant le cholestérol, la tension artérielle, le poids et le diabète. Il est également interdit de consommer des boissons gazeuses ou de l'alcool de façon excessive, par ailleurs une alimentation riche en sucre simple, en sel et en gras saturés est incluse dans la plupart des directives de prévention des MCV.

6. **Sédentarité :** La sédentarité constitue un important facteur de risque de maladies cardiovasculaires, par conséquent une activité physique régulière a un effet positif sur le diabète, l'hypertension artérielle, la dyslipidémie et l'excès de poids. Ainsi, adopter un mode de vie actif réduit le risque de maladie cardiaque.

7. **Embonpoint et obésité :** l'embonpoint et l'obésité sont des troubles de santé importants, en prédisposant les personnes aux principaux facteurs de risque de maladies cardiovasculaires, notamment l'inactivité physique, l'hypertension, la dyslipidémie et le diabète, ces conditions sont associées à un risque accru de mortalité cardiovasculaire et de mortalité toutes causes confondues.

1.2.3 Symptômes des maladies cardiaques

Il y a plusieurs symptômes indiquent une augmentation de la probabilité d'avoir une maladie cardiaque, notamment [4]:

- Oppression cardiaque, pression et douleur.
- Douleurs dans la cage thoracique, les bras, le cou, la mâchoire et le dos.
- Difficulté à respirer.
- Etourdissements.
- Agitation.

Mais ils ne restent que des symptômes qui ne confirment pas la maladie, doncon a besoin de faire beaucoup de tests et d'examens médicaux, et cela nous fournit beaucoup d'attributs essentiels qui contiennent plus d'informations qui nous permettent de détecter la maladie.

1.2.4 Attributs de la maladie

Afin de diagnostiquer les maladies cardiovasculaires, il est nécessaire de recueillir autant d'informations que possibles et identifier les symptômes pour assurer un diagnostic correct. Et pour obtenir ces informations, de nombreux chercheurs ont eu recours aux informations disponibles dans la base de données de Cleveland [5], cette base de données contient 303 instances et 76 attributs, mais toutes les études publiées indiquent que 14 attributs sont utilisés, Ils sont :

1. **age :** âge en années.
2. **sex :**sexe (1 = masculin ; 0 = féminin).
3. **cp :**type de douleur thoracique(1= angine typique ; 2= angine atypique ; 3= douleur non angineuse ; 4= asymptomatique).
4. **trestbps :**tension artérielle au repos (en mm Hg à l'admission à l'hôpital).

5. **chol :** cholestérol sérique en mg/dl.
6. **fbs :** (glycémie à jeun > 120 mg/dl) (1 = vrai ; 0 = faux).
7. **restecg :** résultats électro cardiographiques au repos (0= normal ; 1= présentant une anomalie de l'onde ST-T ; 2= montrant une hypertrophie ventriculaire gauche probable ou certaine).
8. **thalach :** fréquence cardiaque maximale atteinte.
9. **exang :** angine induite par l'effort (1 = oui ; 0 = non).
10. **oldpeak :** dépression ST induite par l'exercice par rapport au repos.
11. **slope:** la pente du segment ST d'exercice de pointe (1 = montant ; 2 = plat ; 3 : = descendant)
12. **ca :** le nombre de vaisseaux principaux (0-3) colorés par fluoroscopie.
13. **thal :** 3 = normale ; 6 = défaut corrigé ; 7 = défaut réversible.
14. **num:** l'attribut prédit (diagnostic de maladie cardiaque (état de la maladie angiographique) 0 : < 50 % de rétrécissement du diamètre ;1 : > 50 % de rétrécissement du diamètre).

1.3 Apprentissage automatique

L'apprentissage automatique -Machine Learning (ML)- est un terme courant ces jours-ci en raison du développement technologique que nous avons atteint dans tous les domaines, et ici on va baser sur le domaine de la prédiction des maladies cardiovasculaires. Alors c'est quoi l'apprentissage automatique ? Et quelles sont ses techniques utilisées pour prédire les maladies cardiaques ?

1.3.1 Définition

Selon Arthur Samuel [6] : "L'apprentissage automatique est le domaine d'étude qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés".

À l'aide d'un lexique informatique. Tom Mitchell l'a présenté comme suit : « On dit qu'un programme informatique apprend de l'expérience (E) en ce qui concerne une classe de tâches (T) et une mesure de performance (P), si sa performance aux tâches dans T, telle que mesurée par P, s'améliore avec l'expérience E » [7].

Et grâce à ces différentes définitions, l'idée est de former les ordinateurs en fonction de l'exécution intelligente des tâches, en apprenant l'environnement par des exemples répétitifs, contrairement au calcul traditionnel des chiffres.

1.3.2 Processus d'apprentissage automatique

Il est difficile, si ce n'est pas impossible, pour un humain d'identifier des schémas cachés dans de grandes quantités de données. Il a donc de la difficulté à prendre des décisions ou à faire des prédictions. Il a donc recours à des algorithmes d'apprentissage automatique qui détectent des schémas cachés dans l'ensemble de données d'entrée et créent des modèles. Ensuite, à partir de ces modèles, faire des prévisions précises de tout nouveau jeu de données pour les algorithmes. De cette manière, l'apprentissage rend la machine plus intelligente [8].

Le processus d'apprentissage automatique peut être résumé dans les étapes suivantes : collection des données, prétraitement des données, le modèle qui dépend d'un ou plusieurs algorithmes et la sortie du modèle est un classificateur qui fait une prédiction lorsqu'il reçoit des valeurs d'entrée (**Figure 1.1**).

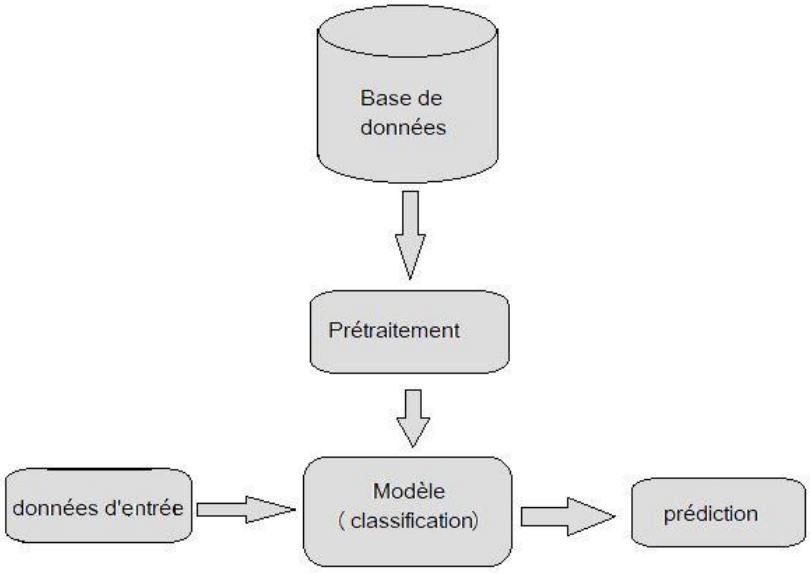

Figure 1-1- La représentation simplifiée de processus d'apprentissage automatique

- **Collection des données :** dans notre cas c'est la base de données de Cleveland [5], qui contient les données des patients.

- **Prétraitement des données:** les techniques de prétraitement de données jeu un rôle très important dans la performance et la précision du modèle [8], les étapes de prétraitement des données incluent : le nettoyage des données, la transformation des données,

l'imputation des valeurs manquantes, la normalisation des données, la sélection des caractéristiques, fractionnement des données et d'autres étapes dépendant de la nature de l'ensemble de données.

- **Modèle :**le modèle se compose d'un seul algorithme, ou il peut contenir plusieurs algorithmes travaillant ensemble dans une approche hybride, sa sortie est un classificateur, c'est là que se trouve l'intelligence, et c'est ce qui fera la prédiction. Ensuite on peut évaluer ce modèle à l'aide des mesures d'évaluation (précision, F-score…etc.).

1.3.3 Technique d'apprentissage automatique

Il existe de nombreuses techniques d'apprentissage automatique [6], mais les principales techniques peuvent se résumer en [8] :

1.3.3.1 Apprentissage non-supervisé

Dans cette technique, les réponses et les objectifs corrects ne sont pas fournis. La technique d'apprentissage non supervisée essaie de découvrir les similitudes entre les données d'entrée et, à partir de ces similitudes, la technique d'apprentissage non supervisée classe les données. Ceci est également connu sous le nom d'estimation de la densité. L'apprentissage non supervisé contient le clustering. Parmi les algorithmes de clustering : l'Analyse en Composants Principales (ACP) et K-moyennes (K-means).

1.3.3.1.1 Analyse en Composants Principales (ACP)

L'Analyse en Composantes Principales est une méthode d'analyse permettant d'explorer de vastes jeux de données multidimensionnels, reposant sur des variables quantitatives, et permet de convertir des variables corrélées en variables sans corrélation appelées « composantes principales », cette méthode a pour objectif de réduire le nombre de variables appliquées aux individus, de simplifier les observations tout en préservant un maximum d'informations.

1.3.3.2 Apprentissage par renforcement

Le renforcement de l'apprentissage diffère de l'apprentissage supervisé, car des ensembles précis d'entrées et de sorties ne sont pas offerts. Cet apprentissage est encouragé par la psychologie behavioriste. L'algorithme est informé quand la réponse est incorrecte, mais ne précise pas comment la corriger. Il doit explorer et tester différentes possibilités pour trouver la

bonne réponse. Il est également connu comme l'apprentissage avec un critique. Il ne recommande pas d'améliorations.

1.3.3.3 Apprentissage supervisé

Dans cette technique, un ensemble de données existe avec des individus et leurs résultats. L'algorithme cherche à apprendre les relations entre les données en entrées et leurs résultats via un processus d'apprentissage ; il peut alors répondre à toute nouvelle entrée en fonction de ce qu'il a appris. À titre d'exemple de technique d'apprentissage supervisé, mentionnons : la classification et la régression[9].

1.3.3.3.1 La classification

Les algorithmes classent les données en deux classes ou plus, et ils donnent la prédiction de oui ou non [9].

Le type d'apprentissage le plus courant est la technique d'apprentissage supervisé, plus particulièrement, la technique de classification qui est largement utilisée pour la prédiction. Dans cette section, nous nous concentrons principalement sur les algorithmes de classification utilisés pour diagnostiquer les maladies cardiaques.

A. Arbre de décision (DT)

Un arbre décisionnel est une structure de type organigramme qui inclut un nœud racine, des branches et des nœuds de feuilles. Les attributs d'ensemble de données sont définis via les nœuds internes. Les branches sont le résultat de chaque test contre chaque nœud [10].

- **Avantages**
  - L'algorithme est simple à comprendre, à interpréter et à visualiser car l'idée est principalement utilisée dans notre vie quotidienne. La sortie d'un arbre de décision peut être facilement interprétée par les humains.
  - Ne nécessite pas de standardisation ou de normalisation des fonctionnalités car il utilise une approche basée sur des règles au lieu d'un calcul de distance.
  - Généralement robuste aux valeurs aberrantes et aux valeurs manquantes et peut les gérer automatiquement.
- **Inconvénients**
  - Sur-ajustement : C'est le principal problème de l'arbre de décision. Afin d'adapter les données, il continue de générer de nouveaux nœuds et, finalement, l'arbre devient trop complexe à interpréter. De cette façon, il perd ses capacités de

généralisation. Cela fonctionne très bien sur les données formées mais commence à faire de nombreuses erreurs sur les nouvelles données.

- Instable : Ajouter un nouveau point de données peut entraîner la régénération de la structure arborescente globale et tous les nœuds doivent être recalculés et recréés.
- Un peu de bruit peut le rendre instable, ce qui conduit à de mauvaises prédictions.

Exemple d'un arbre de décision :

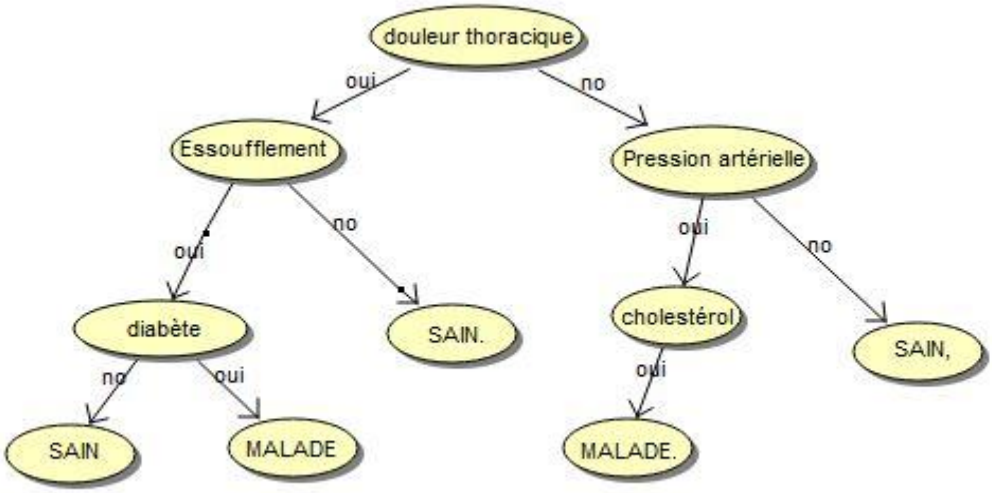

Figure 1-2- Exemple d'un arbre de décision

B. Naïve Bayes (NB)

Le classificateur Naïf Bayes est basé sur le théorème de Bayes. Dans ce classificateur, l'hypothèse principale pour faire une prédiction c'est l'indépendance entre les attributs de l'ensemble de données. Il est facile à implémenter et surtout utile pour des ensembles de données très volumineux. En plus d'être simple, ce modèle dépasse même les méthodes de classification les plus sophistiquées [10] Il est principalement utilisé pour le regroupement et la classification en fonction de la probabilité conditionnelle de se produire.

Le théorème de Bayes fournit un moyen de calculer la probabilité a posteriori P(c|x) à partir de P(c), P(x) et P(x|c). Regardez l'équation Eq(1.1) ci-dessous :

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad \text{eq (1, 1)}$$

$$P(c \mid X) = P(x_1 \mid c) * P(x_2 \mid c) * \ldots * P(x_n \mid c) * P(c)$$

- P(c|x) : représente la probabilité postérieure de la classe (c, cible) donnée au prédicteur (x, attributs).
- P(c) : correspond à la probabilité a priori pour la classe.
- P(x|c) :(likelihood) est la probabilité du prédicteur pour une classe donnée.
- P(x) : est la probabilité a priori du prédicteur.

Suivons les étapes ci-dessous pour l'exécuter :

- **Étape 1 :** Convertir l'ensemble de données en un tableau de fréquences.

Tableau 1-1 -Tableau de fréquence

Les attributs	Les classes		
	Classe 1	…..	Classe n
Attribut 1	3	…	2
Attribut 2	5	…	1
: : : :	: : : :		
Attribut n	4	…	0
Total	T1	…	Tn

- **Étape 2 :** Créer un tableau de probabilité (likelihood).

Tableau 1-2 Tableau de probabilité

Les attributs	Les classes				
	Classe 1	…..	Classe n		
Attribut 1	3	…	2	(3+….+2)/T	P_1
Attribut 2	5	…	1	(5+…+1)/T	P_2
: : : :	: : : :				
Attribut n	4	…	0	(4+…+0)/T	P_n
	T1/T	…	Tn/T		
	$P_{classe\ 1}$	…	$P_{classe\ n}$		

- **Étape 3 :** Maintenant, utilisez l'équation bayésienne (mentionné ci-dessus) pour calculer la probabilité postérieure pour chaque classe. La classe ayant la plus forte probabilité postérieure est le résultat de la prédiction.

- **Avantages**
  - Cet algorithme fonctionne rapidement et peut faire gagner beaucoup de temps.
  - Naive Bayes est adapté à la résolution de problèmes de prédiction multi-classes.
  - Naive Bayes est mieux adapté aux variables d'entrée catégorielles qu'aux variables numériques.
- **Inconvénients**
  - Naive Bayes suppose que tous les prédicateurs (ou caractéristiques) sont indépendants, ce qui se produit rarement dans la vie réelle. Cela limite l'applicabilité de cet algorithme dans des cas d'utilisation réels.
  - Cet algorithme a « problème de fréquence zéro » où il attribue une probabilité nulle à une variable catégorielle dans le test dont la catégorie n'était pas disponible dans l'ensemble de données d'apprentissage.

C. Les K plus proches voisins (KNN)

K-NN classe un objet par le vote majoritaire de ses plus proches voisins. La classe d'une nouvelle instance sera prédite à la base de certaines métriques de distance (voir la **Figure**

1.3). La métrique de distance utilisée dans les méthodes du plus proche voisin pour les attributs numériques peut être une distance euclidienne [10].

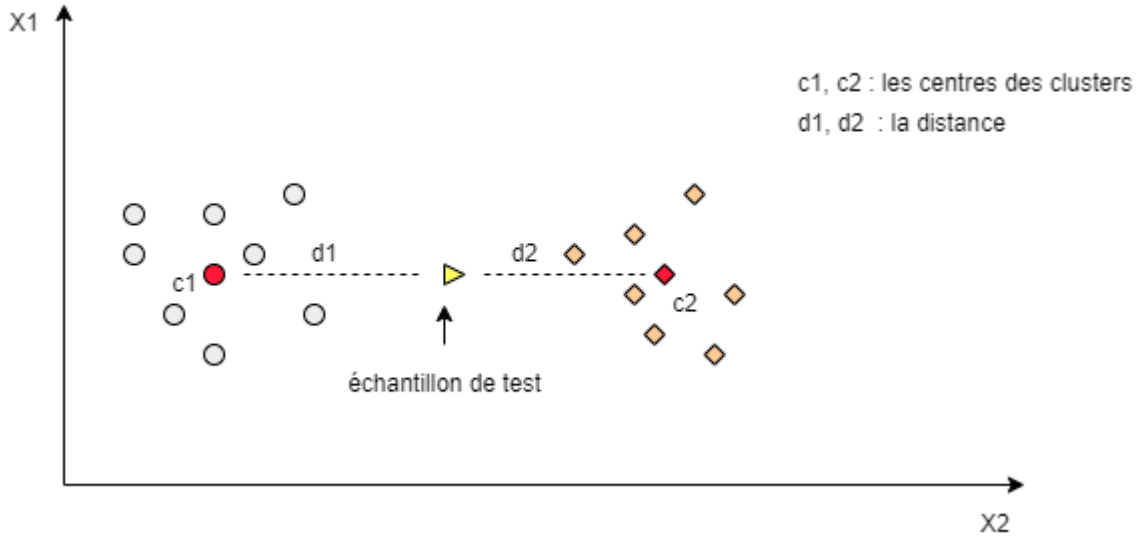

Figure 1-3- Utilisation de la distance dans l'algorithme KNN

Considérez un plan XY avec des points dans le graphe, son principe de fonctionnement est le suivant:

- Choisissez la valeur K: tracer la courbe du coude entre les diverses valeurs K et l'erreur, choisissez la valeur K si le taux d'erreur chute subitement.
- Calculez la distance entre tous les points d'entraînement et les nouveaux points de données.
- Triez la distance calculée par ordre croissant entre les points d'entraînement et les nouveaux points de données.
- Choisissez les premières K distances dans la liste triée.
- Prendre le mode/moyenne des classes associées aux distances.

Et voici un pseudo code de l'algorithme KNN[11]:

pour tous les échantillons inconnus Déséchantillonner (i)

pour tous les échantillons connus Echantillon(j)

calculer la distance entre Déséchantillonner (i) et Echantillon (j)

fin pour

trouver les k plus petites distances

localiser les échantillons correspondants Échantillon(j1),..,Échantillon(jk)

affecter Déséchantillonner (i) à la classe qui apparaît plus fréquemment

fin pour

Figure 1-4- Pseudo code de Algorithme KNN

- **Avantage**s
  - Simplicité, efficacité, intuitivité et performance de classification compétitive dans de nombreux domaines.
  - Il est robuste aux données d'entraînement bruitées et efficace si les données d'entraînement sont volumineuses.
- **Inconvénients**
  - Il a de mauvaises performances d'exécution lorsque l'ensemble d'apprentissage est grand.
  - Il est très sensible aux caractéristiques non pertinentes ou redondantes car toutes les caractéristiques contribuent à la similarité et donc à la classification.
  - L'apprentissage à distance n'est pas clair quel type de distance utiliser et quel attribut utiliser pour produire les meilleurs résultats.
  - Le coût de calcul est assez élevé car nous devons calculer la distance de chaque instance de requête à tous les échantillons d'apprentissage.

D. Machines à Vecteurs Supports (SVM)

SVM sont définis comme des espaces vectoriels de dimension finie dans lesquels chaque dimension représente une « caractéristique » d'un objet particulier. Il s'est avéré être une approche efficace dans les problèmes spatiaux de grande dimension[10].

Certains termes liés à 1 représentent des mots-clés pour comprendre le fonctionnement de l'algorithme :

- ✓ **Vecteurs de support :** ce sont les points les plus proches de l'hyperplan. Une ligne de séparation sera définie à l'aide de ces points de données.

- ✓ **Marge :** c'est la distance entre l'hyperplan et les observations les plus proches de l'hyperplan (vecteurs supports). Dans SVM, une grande marge est considérée comme une bonne marge.

Si nous traçons les points de données dans un graphe bidimensionnel, nous appelons cette frontière de décision une ligne droite, mais si nous avons plus de dimensions, nous appelons cette frontière de décision un "hyperplan", l'objectif principal de SVM est de trouver le meilleur hyperplan qui a la distance maximale des classes.

On peut résumer les étapes d'algorithme dans [12] :

- ❖ **Étape 1 :** Utilisez tous les échantillons d'apprentissage pour former un SVM initial, ce qui entraîne des vecteurs de support de la fonction de décision correspondante.
- ❖ **Étape 2 :** Excluez de l'ensemble d'apprentissage les vecteurs de support, dont les projections sur l'hyperplan ont les plus grandes courbures :
    1. Pour chaque vecteur de support, trouvez sa projection sur l'hyperplan, le long du gradient de la fonction de décision.
    2. Pour chaque vecteur support, calculer la courbure généralisée de sur l'hyper surface
    3. Triez dans l'ordre décroissant de et excluez les ***n*** premiers pourcentages de vecteurs de support de l'ensemble d'apprentissage.
- ❖ **Étape 3 :** utiliser les échantillons restants pour recycler le SVM, ce qui entraîne des vecteurs de support et la fonction de décision correspondante.
- ❖ **Étape 4 :** Utilisez les paires de points de données pour former enfin le SVM, ce qui donne des vecteurs de support et la fonction de décision correspondante.

E. Perceptron multicouche (MLP)

Le classificateur MLP contient des multiples de nœuds disposés en couches. Cela crée un graphique orienté qui couvre les couches d'entrée, cachées et de sortie et chaque couche est entièrement connectée à la suivante. Ce classificateur a une architecture claire et un algorithme simple, c'est donc l'un des modèles de réseaux de neurones les plus connus.

F. Forêt aléatoire (RF)

Forêt aléatoire peut être utilisé pour des problèmes de classification et de régression parce qu'elle repose sur l'apprentissage par arbre de décision (DT), et elle combine des décisions multiples pour prendre une décision unique à l'aide d'un vote majoritaire comme la **Figure 1-5** l'explique.

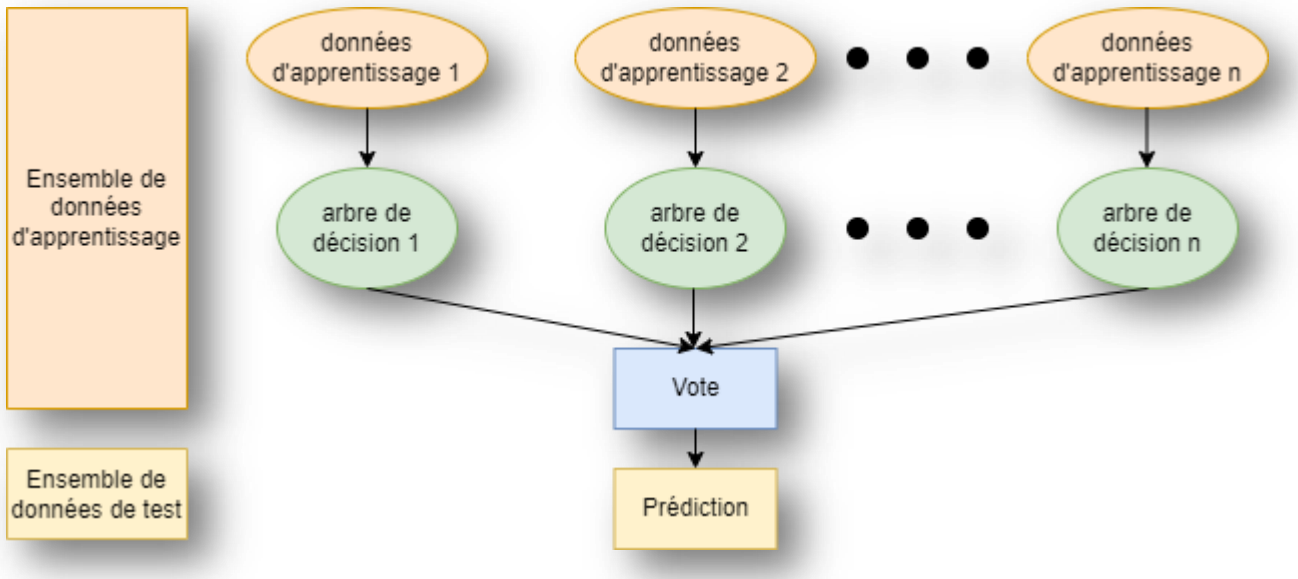

Figure 1-5-Schéma explicatif de l'algorithme de la Forêt aléatoire

1.3.3.3.2 La régression

Les algorithmes de régression donnent la réponse de "Combien"[9], donc ils sont utilisés si la prédiction est une valeur réelle.

Les algorithmes de régression peuvent aussi des algorithmes de classification et ça dépend de l'utilisation de l'algorithme comme: **SVM, DT.**

1.3.4 Méthodes de validation

Pour valider les performances d'un modèle ML consiste à former un modèle avec les données disponibles et à évaluer ses performances de classification à l'aide d'un ensemble de données distinct. Les deux méthodes de validations les plus connues sont : fractionnement de données (Train/Test Split) et la validation croisée (K-Fold Cross Validation)[13].

1.3.4.1 Fractionnement des données

Consiste à séparer une partie des données avant de développer un modèle ML et à utiliser ces données pour le teste et l'autre partie pour l'apprentissage comme la **Figure 1-6** montre.

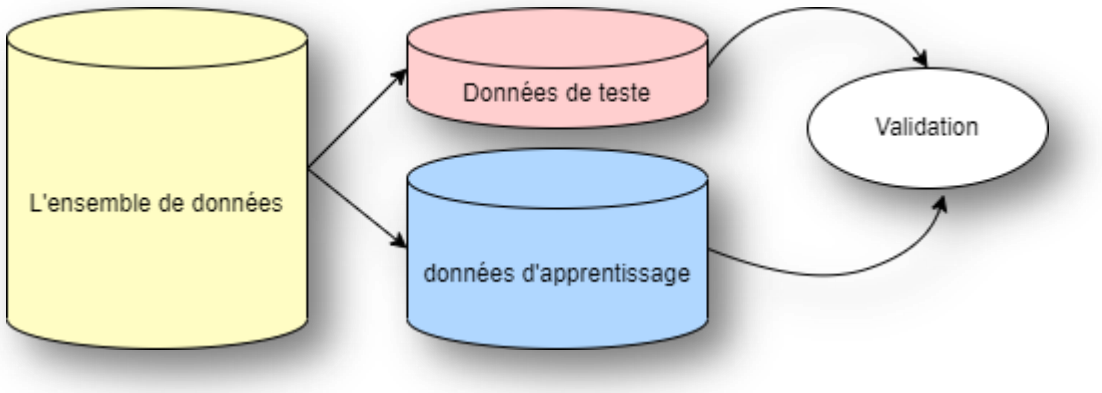

Figure 1-6-Fractionnement des données

1.3.4.2 Validation croisée

Au lieu de former un modèle fixe une seule fois comme dans la méthode précédente (Train/teste Split), avec la validation croisée plusieurs modèles sont développés de manière itérative sur différentes parties des données (k). Une partie des données est séparée pour le tester, laissant le reste (k-1) pour former un modèle et prédire les classes sur les données de test. Ce processus est répété k fois jusqu'à ce que toutes les données sont utilisées (la **Figure 1-7**).

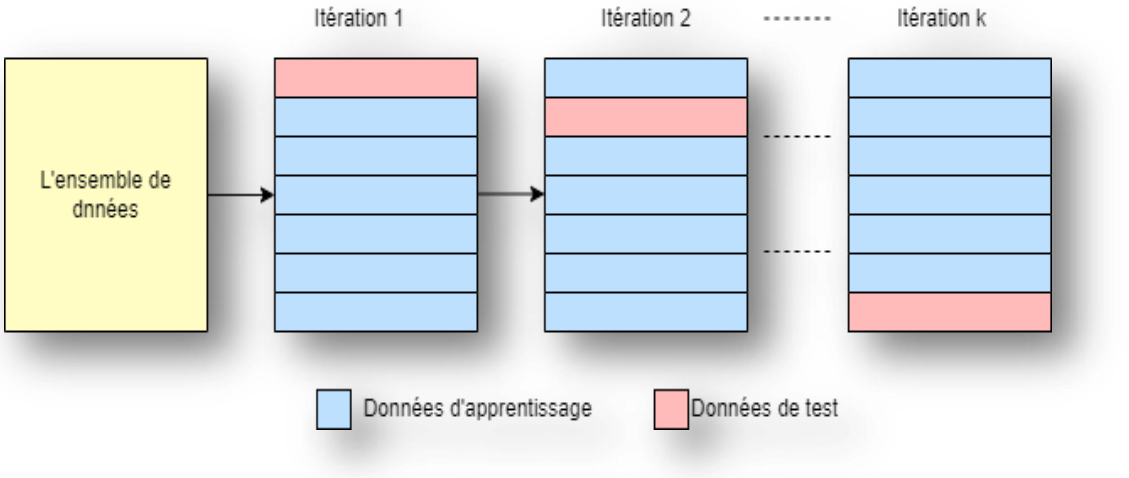

Figure 1-7-Validation croisée

1.3.5 Mesures de performance

Les mesures de performance sont des métriques indiquant la qualité de correspondance entre les valeurs prévues et les valeurs obtenues par le modèle, qui est obtenu par la matrice de confusion, si on prend l'exemple de la maladie on obtient la matrice suivante :

Tableau 1-3-Matrice de confusion

		Si le patient est malade ou non (réel)	
		Est malade	N'est pas malade
La prédiction du modèle	Est malade	**Vrai positif (VP)**	**Faux positif (FP)**
	N'est pas malade	**Faux négatif (FN)**	**Vrai négatif (VN)**

D'après la matrice de confusion on obtient les valeurs suivantes :

- **Vrai positif (VP):** les valeurs réelles et prédites sont identiques et positives.
- **Vrai négatif (VN) :** les valeurs réelles et prédites sont identiques et négatives.
- **Faux positif (FP) :** les valeurs réelles et prédites sont différentes. Le patient n'est pas malade, mais le modèle prédit qu'il est malade.
- **Faux négatif (FN) :** les valeurs réelles et prédites sont différentes. Le patient est malade, et le modèle prédit qu'il n'est pas malade.

A travers ces valeurs, on peut comprendre les mesures retenus pour comparer ou valider les modèles, et voici ces mesures :

1. **Accuracy:** La précision est l'une des principales mesures de performance pour la classification. Il est défini comme la proportion entre la classification correcte et l'échantillon total, comme indiqué dans l'équation eq(1.2) suivante :

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN} \qquad \text{eq (1, 2)}$$

2. **Rappel :** C'est la petite proportion des individus par rapport à la quantité globale des individus applicables. L'équation eq(1.3) de rappel est représentée comme suit:

$$Rappel = \frac{VP}{VP + FN} \qquad \text{eq (1, 3)}$$

3. **Précision :** C’est la proportion des individus qui sont correctement identifiées par le modèle. L’équation de précision est représentée comme suit:

$$Précision = \frac{VP}{VP + FP} \qquad \text{eq (1, 4)}$$

4. **F1-score :** C'est la moyenne entre la précision et le rappel :

$$F1 - score = 2 * \frac{Précision * Rappel}{Précision + Rappel} \qquad \text{eq (1, 5)}$$

1.4 L’apprentissage automatique et les maladies cardiovasculaires

Dans l'article[14], une prédiction des maladies cardiaques est proposée à l'aide de techniques d'apprentissage automatique supervisées. Il a utilisé les algorithmes suivants : k plus proches voisins (KNN), forêt aléatoire (RF), régression logistique (LR), machines à vecteurs supports (SVM), Naïve Bayes (NB), arbre de décision (DT). Il a été trouvé dans les résultats que la régression logistique a atteint la meilleure accuracy de 92.30 %.

[15]ils utilisent la base de données de Cleveland [5] pour prédire les possibilités de survenue d'une maladie cardiaque chez les patients, à l’aide de deux algorithmes : arbre de décision (DT) et Naïve Bayes (NB), et lorsqu'ils ont comparé ces deux algorithmes, ils ont trouvé DT le mieux adapté à une telle classification (accuracy = 91%), où les auteurs voient la raison de la grande précision de cet algorithme, c'est que ce modèle analyse l'ensemble de données dans le format de la forme de l'arbre. Par conséquent, chaque nœud de l'ensemble de données a été analysé.

[16] Ils essayent d'améliorer la précision de l'analyse de la prédiction des maladies cardiaques sur la base de la méthode d'ensemble, ils utilisent deux méthodes d'extraction des caractéristiques : l'analyse discriminante linéaire (ADL) et l'analyse en composantes principales (ACP)pour sélectionner les caractéristiques essentielles de l'ensemble de données, et puis ils appliquent les algorithmes suivants :k plus proches voisins (KNN), machines à vecteurs supports (SVM), arbre de décision (DT), forêt aléatoire (RF) et Naïve Bayes (NB). La comparaison entre les résultats des algorithmes pour les caractéristiques extraites par (ACP) et les résultats des algorithmes pour les caractéristiques extraites par(ADL) montré que la forêt aléatoire (RF) avec (ADL) était plus performante (accuracy=98.4%), Et puis lors de l'application de la première

technique « d'ensemble: bagging », l'arbre de décision (DT) avec (ACP) montre son efficacité(accuracy=98.6%). Ensuite, en appliquant la deuxième technique « d'ensemble:boosting », il se trouve que la forêt aléatoire (RF) avec (ACP) est la plus performante (accuracy=98.3%), donc la technique d'ensemble bagging montre son efficacité pour améliorer les résultats où l'arbre de décision (DT) avec (ACP) atteint la meilleur résultat accuracy 98.6%.

[17]Dans cet ouvrage, un modèle d'apprentissage automatique hybride pour prédire les maladies cardiaques est discuté, ce modèle hybride produit à l'aide d'un arbre de décision (DT) et d'un algorithme de forêt aléatoire (RF). Le modèle hybride fonctionne sur la base de probabilités de forêts aléatoires où les probabilités de la forêt aléatoire (RF) sont ajoutées aux données d'entraînement et transmises à l'algorithme de l'arbre de décision. De même, les probabilités de l'arbre de décision sont identifiées et transmises pour tester les données. Enfin, les valeurs sont prédites, et ce modèle a surpassé les modèles réguliers (DT) et (RF), car il a atteint accuracy = 88%.

Le tableau suivant montre une partie du travail qui a été fait dans ce contexte :

Tableau 1-4- Travaux existants

Année	Auteur	Objectif	Techniques	Accuracy
2018	• Pahulpreet Singh Kohli • ShriyaArora	Application de l'apprentissage automatique à la prédiction des maladies	• Adaboost • DT • LR • RF • SVM	LR=87.1%
2019	• SanthanaKrishnan.J • Geetha.S	Prédiction des maladies cardiaques à l'aide d'algorithmes d'apprentissage automatique.	• DT • NB	DT=91% NB=87%

2020	• Latifur Rahman • Rahad Arman Nabid • FarhadHossain	Sélection de département basée sur l'analyse des symptômes de la maladie à l'aide de l'apprentissage automatique pour le traitement médical	• LR • KNN • DT • NB • SVM • MLP • LDA • MCLOvR • MCLoC	MCLOvR=80%
2021	• Xiao-Yan Gao • Abdelmegeid Amin Ali • Hassan Shaban Hassan • Eman M. Anwar	Amélioration de la précision de l'analyse de la prédiction des maladies cardiaques basée sur la méthode d'ensemble	• KNN • DT • RF • SVM • NB • Bagging • Boosting • ACP • ADL	DT + ACP + Bagging = 98.6%
2021	• Qi Zhenya • Zuoru Zhang	Un ensemble hybride sensible aux coûts pour la prédiction des maladies cardiaques	• RF • LR • ELM • KNN • SVM • Relief algorithm • Cross validation	
2021	• M. Kavitha • G.Gnaneswar • R.Dinesh • Y.Rohith Sai	Prédiction des maladies cardiaques à	• DT • RF • Modèle hybride	Le modèle hybride (DT+RF)=88.7

	• R.SaiSuraj	l'aide d'un modèle d'apprentissage automatique hybride	(DT+RF)	%
2021	• HarshitJindal	Prédiction des maladies cardiaques à l'aide d'algorithmes d'apprentissage automatique	• KNN • LR • RF	LR = 88.5% KNN= 88.5%
2022	• Chiradeep Gupta • Athina Saha • N V SubbaReddy • U DineshAcharya	Prédiction des maladies cardiaques à l'aide de techniques d'apprentissage automatique supervisé	• KNN • DT • LR • NB • SVM • RF	LR = 92.3%

1.5 Conclusion

Dans ce chapitre, nous avons étudié les maladies cardiaques, leurs types, leurs causes, leurs symptômes et leurs caractéristiques. Nous avons ensuite abordé les concepts de l'apprentissage automatique, les travaux réalisés dans la littérature pour la prédiction des maladies cardiaques à l'aide de différents algorithmes d'apprentissage automatique, et nous les avons aussi comparés en termes de précision de prédiction.

Dans le prochain chapitre, nous discuterons le modèle que nous proposons pour prédire les cardiopathies à l'aide d'algorithmes d'apprentissage automatique.

Chapitre 2

Conception d'un système intelligent de prédiction des cardiopathies

2.1 Introduction

Dans le chapitre précédent, une présentation des maladies cardiaques et de l'apprentissage automatique a été faite, ainsi qu'une discussion de certains travaux réalisés dans ce contexte. Dans le présent chapitre, nous décrivons le système que nous proposons pour prédire les maladies cardiovasculaires. A partir d'un ensemble de données contenant des données provenant de nombreux patients et passant par plusieurs étapes. Tout d'abord, il s'agit du prétraitement des données et de la méthode d'analyse en composantes principales (ACP) pour extraire les caractéristiques importantes du processus de classification des entités comme malades ou non-malades. Par la suite, les données sont prêtes pour appliquer l'algorithme d'apprentissage automatique KNN. En effet, après avoir divisé les données en données d'entraînement et en données de test, nous entraînons notre algorithme pour obtenir un modèle capable de prédire l'état du patient en fonction des entrées correspondantes, la **Figure 2.1** illustre le processus de déroulement global de notre démarche.

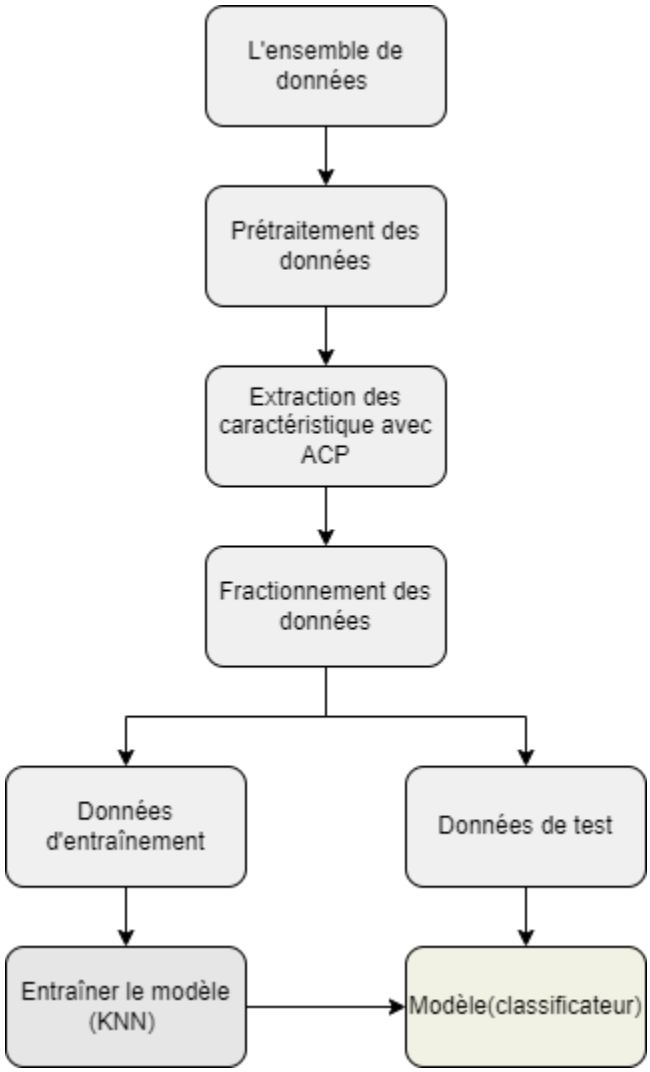

Figure 2-1- Les étapes de construction de notre système

2.2 L'ensemble de données

Le système que nous proposons utilise des données et des analyses médicales pour un groupe de personnes, toutes collectées dans **une base de données** de Cleveland[5], que nous avons déjà parlé dans le premier chapitre, et qui permettent de mieux assimiler ces informations. Il vaut mieux les explorer et les transformer en une forme plus lisible (graphiques) et qui leur assure un format et un contexte permettant de les interpréter.

2.2.1 Exploration des données

La **Figure 2-2** représente les deux classes que contient la base de données, ou **0** représente la classe non malade et **1** représente la classe non malade, et il apparaît également qu'ils sont proches en termes de nombre de patient.

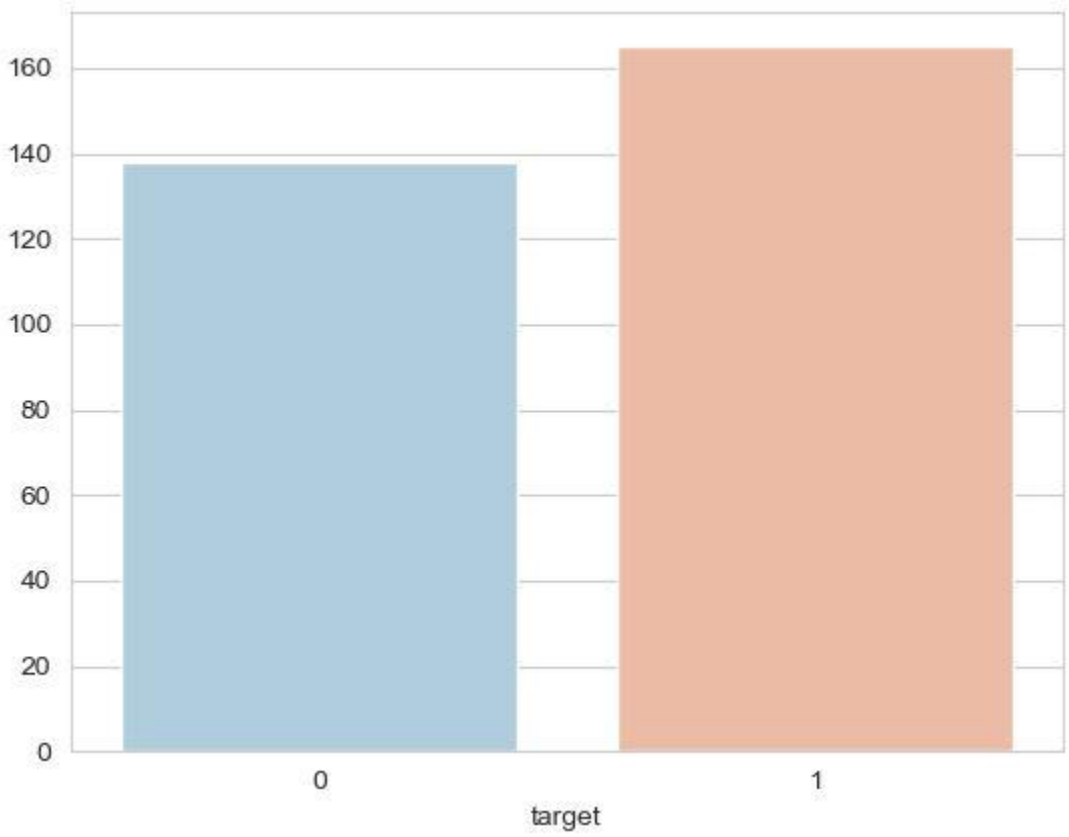

Figure 2-2- Graphe illustrant les deux classes

L'un des indicateurs porteurs d'informations qui jouent un rôle relatif pour déterminer si la maladie est identifiée ou non est l'âge. La **Figure 2-3** représente la répartition des patients par âge où les âges varient entre 29 et 77 ans, il est aussi clair que le groupe prédominant est celui qui a plus de 50 ans.

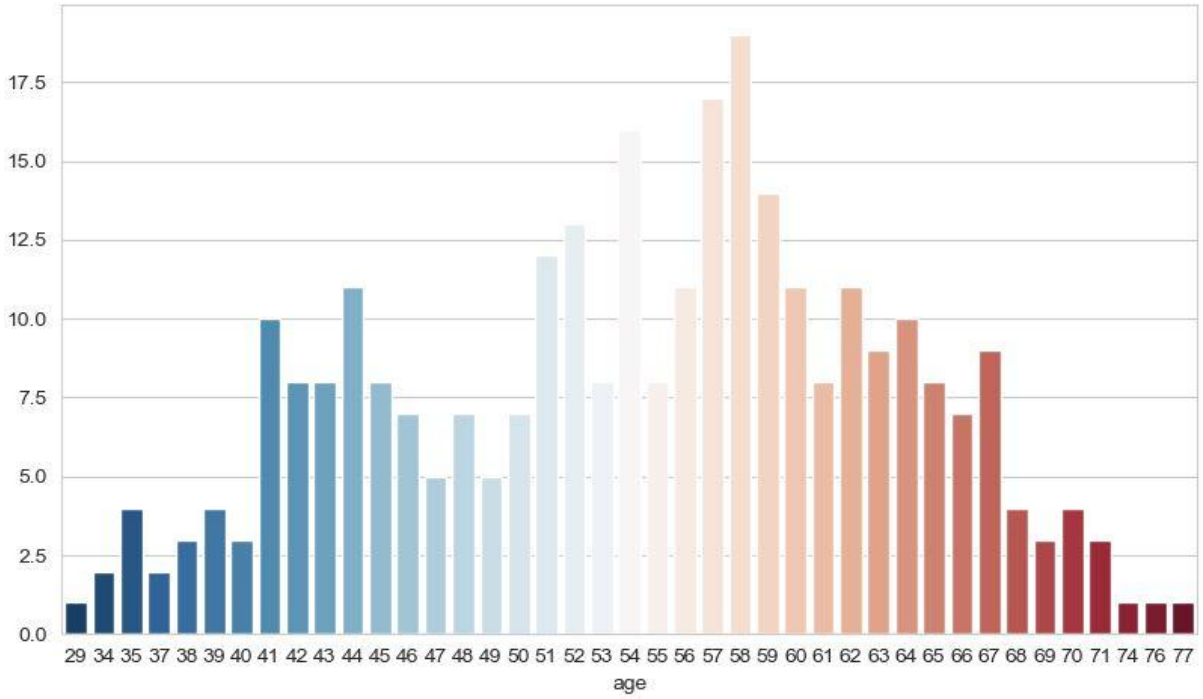

Figure 2-3 Graphe illustrant la répartition par âge

En plus de l'âge, il y a 12 autres indicateurs contiennent divers éléments d'information qui aident à déterminer l'incidence de la maladie, et pour permettre une meilleure compréhension de ces indicateurs, nous avons créé une matrice de corrélation qui nous montrerait la relation entre les indicateurs et l'étendue des relations entre eux. La **Figure 2-4** représente cette matrice où chaque indice de corrélation calculé entre deux indicateurs x et y prend la valeur [-1, +1] qui représente la significativité de l'indice x par rapport à l'indice y est lorsque le coefficient se rapproche de 1 ou de -1 (vert foncé ou rouge foncé respectivement), ce qui indique une forte corrélation entre les deux indicateurs x et y, mais quand la valeur approche de 0, cela signifie une faible corrélation entre les deux indicateurs.

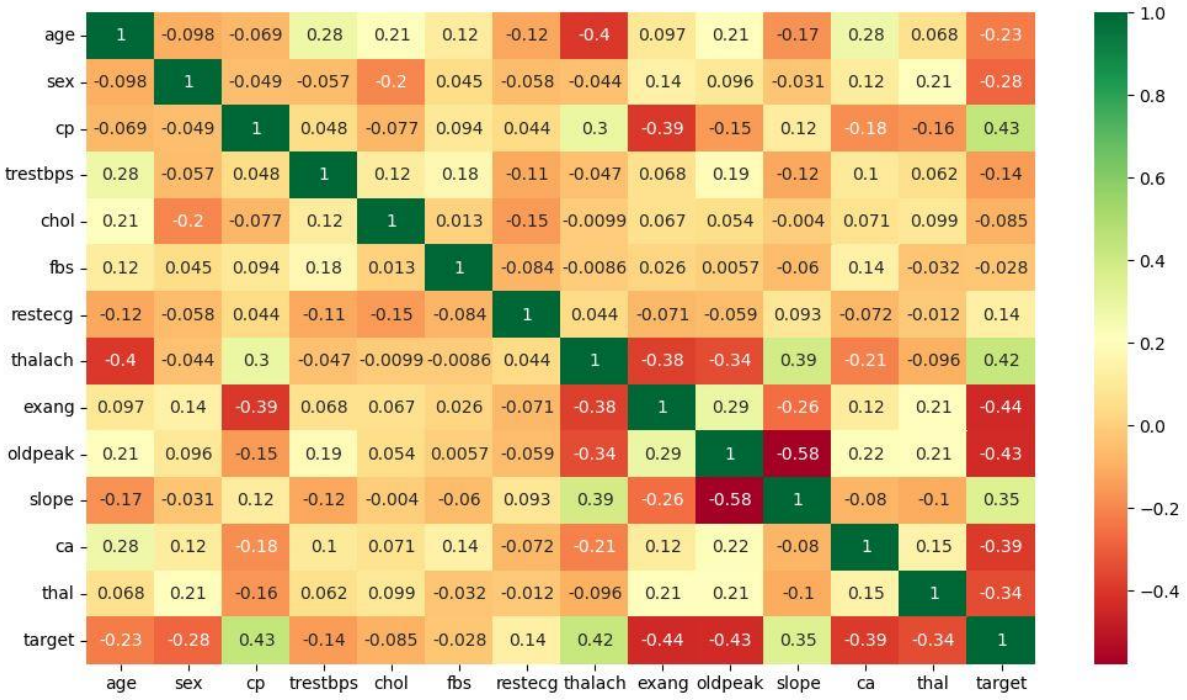

Figure 2-4 Matrice de corrélation entre les indicateurs

2.3 Le prétraitement de données

Après la collecte des données, nous allons travailler là-dessus, l'étape pour faire en sorte que sa qualité passe par l'examen de plusieurs critères de qualité liés aux données, notamment :**la cohérence**, **la précision**, **la disponibilité** afin qu'elle ne soit pas déficiente en un ou plusieurs intrants, ...etc. Dans le but de fournir cette étape, il faut passer par l'étape de Nettoyage de données, et comme le KNN est affecté par la gamme de fonctionnalités et utilise les distances entre les points de données pour déterminer leur similitude il faut aussi passer par normalisation et standardisation des données, parce que avec les données d'échelles variables peut constituer un problème dans l'analyse en ce sens qu'une variable numérique comme l'âge entre 29 et 77 sera plus pesante dans l'analyse qu'une variable dont les valeurs sont comprises entre 0 et 1 ce qui causerait un problème de biais par la suite, et très souvent, nous devons travailler à partir de données numériques, et ces données sont rarement comparables à l'état brut.

2.3.1 Nettoyage des données

Dans cette étape les données incomplètes, altérées, inexactes sont identifiées et corrigées ou supprimées pour améliorer les critères de qualités, mais dans notre nous n'avons la réalisé parce que nous avons apporté cette base de données d'une source intermédiaire [18] qui est passé par l'étape de Nettoyage de données.

2.3.2 Normalisation des données

La normalisation est une technique de mise à l'échelle dans laquelle les valeurs sont décalées et remises à l'échelle afin de se retrouver dans [0, 1], elle est aussi appelée échelle Min-Max.

Et sa formule s'écrit comme ci-dessous :

$$\grave{X} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad \text{eq (2, 1)}$$

Si X est la valeur minimale dans la colonne, le numérateur sera 0, et donc$\grave{X}$ est 0,d'autre part si X est le max donc $\grave{X}$ est 1.

Si on l'applique, par exemple, à l'indicateur d'âge, on trouve :

Exemple:X=35

$$\grave{X} = \frac{35 - 29}{77 - 29} \qquad \text{eq (2, 2)}$$

On a $\grave{X}$= 0.13 ∈ [0, 1]

2.3.3 Standardisation des données

La standardisation est une autre technique de mise à l'échelle où les valeurs sont centrées autour de la moyenne avec un écart type unitaire, cela signifie que la transformation est plus subtile que simplement ramener l'ensemble des valeurs dans un intervalle réduit (normalisation). Mais avant cela, nous devons considérer certaines variables catégorielles telles que : sexe, ca, fbs... ne sont pas comme d'autres variables considérées comme indicatives telles que : chol, trestbps... Il faut donc convertir ces variables.

Et puis nous continuons le processus de standardisation, selon la formule suivante :

$$\grave{X} = \frac{X - \mu}{\sigma} \qquad \text{eq (2, 3)}$$

Où :

- μ est la moyenne des valeurs des caractéristiques.
- σest l'écart type des valeurs des caractéristiques.

Et après avoir appliqué la formule à l'ensemble de données, nous obtenons des données standardisées prêtes pour la prochaine étape.

2.4 Extraction des caractéristiques avec l'ACP

2.4.1 Extraction des caractéristiques

L'extraction des caractéristiques fait partie du processus de prétraitement de données, mais en raison de l'importance de cette partie dans la détermination de l'efficacité de notre système, nous avons décidé de l'aborder séparément.

L'extraction de caractéristiques est une forme particulière de réduction de dimension, elle simplifie le coût des ressources nécessaires pour décrire précisément un ensemble de données important, parce que l'analyse des données complexe nécessite une grande utilisation de la mémoire et de traitement ou des algorithmes de classification qui nécessitent un seuil d'ajustement élevé avec les échantillons à tester, donc le résultat de cette étape est de déterminer un sous-ensemble des caractéristiques initiales est appelé caractéristiques extraites, ses caractéristiques-là, devraient contenir les informations pertinentes issues des données d'entrée, de sorte que la méthode ou l'algorithme à appliquer puisse être effectuée en utilisant cette représentation réduite au lieu des données initiales complètes. Et parmi les méthodes d'extraction des caractéristiques nous choisissons l'ACP (Analyse en Composantes Principales).

1. **Remarque**

   Il y a une différence entre l'extraction et la sélection, car l'extraction signifie l'obtention des nouvelles caractéristiques utiles à partir de caractéristiques existantes, alors que la sélection signifie choisir d'un sous-ensemble du groupe de caractéristiques d'origine, comme expliqué dans la **Figure 2-5**.

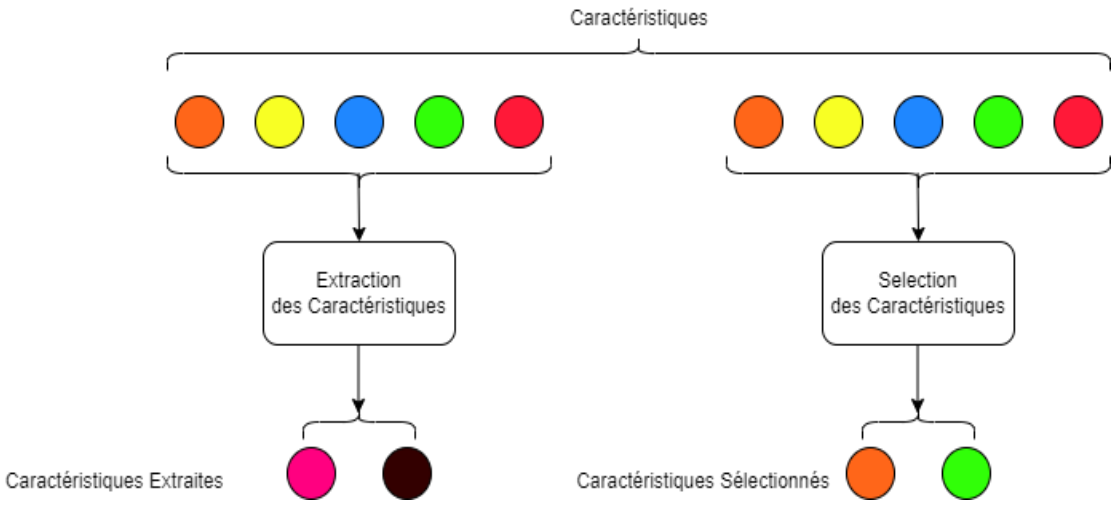

Figure 2-5- Différence entre sélection et extraction

2.4.2 Processus de l'ACP

Les variables sur lesquelles nous allons travailler est préférable qu'ils soient sous la forme de taux, de moyennes ou de proportions, pour éviter des corrélations structurelles biaisant l'analyse, c'est pourquoi nous l'avons préparé comme nous l'avons fait avant, donc on va directement commencer à :

1 **Calculer la covariance** avec la formule **eq(2.4)** suivante :

$$\mathrm{COV}(X_j, X_{j'}) = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ij'} - \bar{X}_{j'}) \qquad \text{eq (2. 4)}$$

2 **Calculer la corrélation linéaire** avec la formule **eq(2.5)** suivante :

$$R(X_j, X_{j'}) = \frac{\mathrm{COV}(X_j, X_{j'})}{\sigma_j \sigma_{j'}} \qquad \text{eq (2.5)}$$

3 **Calculer les valeurs et les vecteurs propres**

λ est dite valeur propre de la matrice As'il existe un vecteur non nul X $\in K^n$ tel que : AX $= \lambda$X.

Ou k sera $\mathbb{R}$ ou $\mathbb{C}$ et Aest une matrice carrée de taille $n * n$

Le vecteur X est alors appelé vecteur propre de$A$ associé à la valeur propre λ

Et à partir de ça et de la matrice de corrélation linéaire qu'on a eue plus tôt, on peut calculer les valeurs propres, et puis on calcul des vecteurs propres associés aux valeurs.

4 **Il reste que le nombre de propriétés que nous choisissons,** et cette action est très importante car elle déterminera la validité des indicateurs ce qui déterminent l'exactitude des résultats de notre système proposé.

2.5 Le fractionnement de données

Dans cette étape, l'ensemble de données des maladies cardiaques est divisé en un ensemble d'apprentissage à 85 % et un ensemble de test à 15 % (voir la **Figure 2-6**). L'ensemble d'apprentissage est utilisé pour entraîner le modèle, et l'ensemble de test est utilisé pour évaluer le modèle.

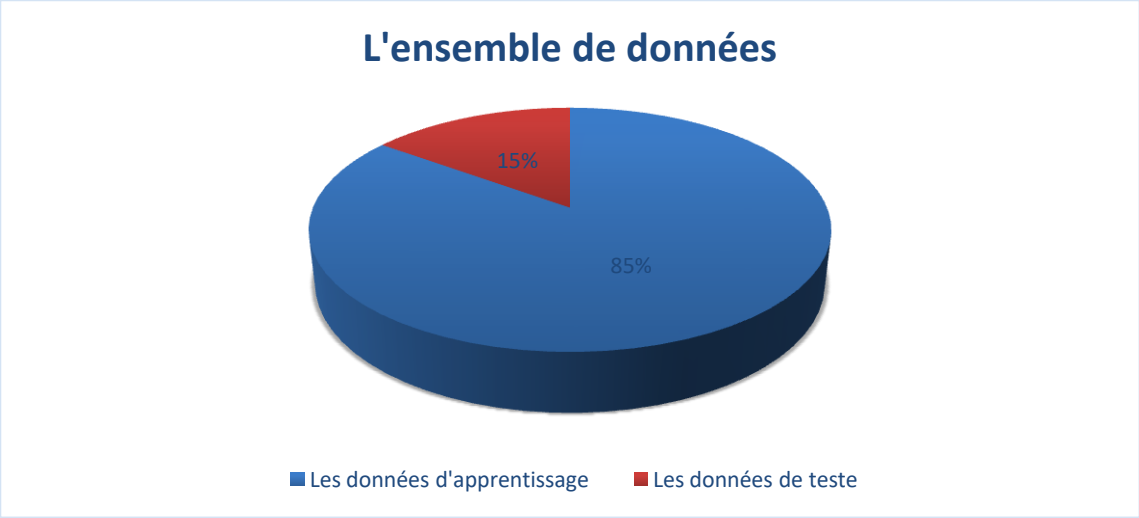

Figure 2-6 Fractionnement de données

2.6 Entraîner le modèle

Le modèle se compose d'exemples de données de sortie et des ensembles correspondants de données d'entrée qui ont une influence sur la sortie. Le modèle d'apprentissage est utilisé pour exécuter les données d'entrée dans l'algorithme KNN que nous avons déjà discuté dans le premier chapitre, afin de mettre en corrélation le résultat traité avec le résultat de l'échantillon. Et cette corrélation s'appelle évaluation.

Le résultat de cette corrélation est utilisé pour changer le modèle, et ce changement s'appelle adaptation et cela se fait en ajustant les poids du modèle qui est initialisé de manière aléatoire. De cette façon, l'algorithme continue à ajuster les poids et après quelque itération il fait un point de

contrôle qui représente un vidage intermédiaire de tout l'état interne d'un modèle (ses poids, son taux d'apprentissage actuel, etc.). Et le processus continue jusqu'à compléter toutes les données d'entrainement, une fois terminé, le modèle sera enregistré (voir la **Figure 2-7**)

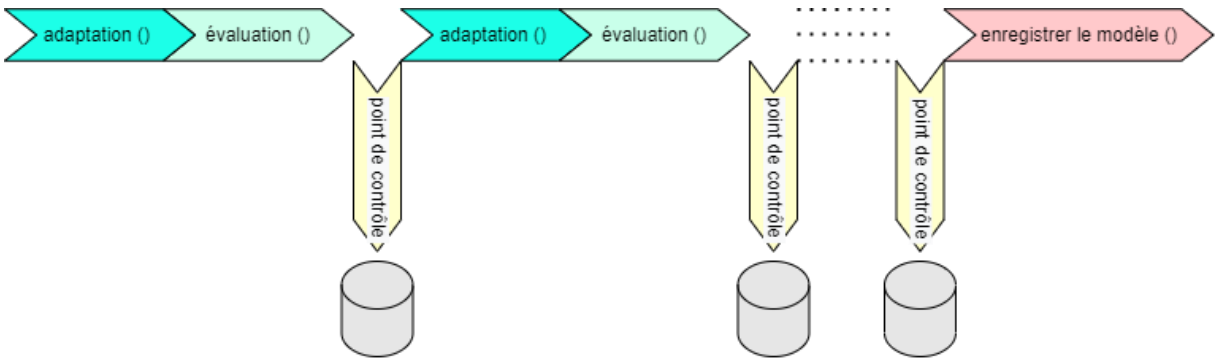

Figure 2-7- Le processus d'entrainement du modèle

Le résultat de cette étape est un modèle fonctionnel qui peut ensuite être validé, testé et déployé. Les performances du modèle pendant la phase de formation permettront de déterminer son efficacité lorsqu'il sera finalement intégré dans une application destinée aux utilisateurs finaux.

2.7 Conception d'application

A ce stade de la conception, nous utilisons la méthodologie UML pour concevoir notre application selon les règles de modélisation de cette méthode, afin d'obtenir des diagrammes permettant d'expliquer les étapes de notre projet.

En effet, nous constatons que deux types de représentations peuvent être utilisées pour expliquer notre système. Sur le plan statique, nous nous sommes satisfaits appuyés sur le diagramme de cas d'utilisation, et sur le plan dynamique, sur le diagramme de séquence.

2.7.1 Diagramme de Cas d'utilisation

Le diagramme de cas d'utilisation dans la **Figure 2-8** exprime l'interaction entre l'acteur (l'utilisateur) et le système, dans lequel l'utilisateur peut interagir avec l'application en saisissant les données du patient pour prédire, puis lance le processus de prédiction. Ensuite, le système transfère les données saisies à la phase de prétraitement, où il vérifie et normalise ces données

puis il extrait les caractéristiques et classifie cet échantillon selon le modèle qui apprenait dans la phase d'apprentissage. Finalement, l'utilisateur peut voir le résultat de ce système.

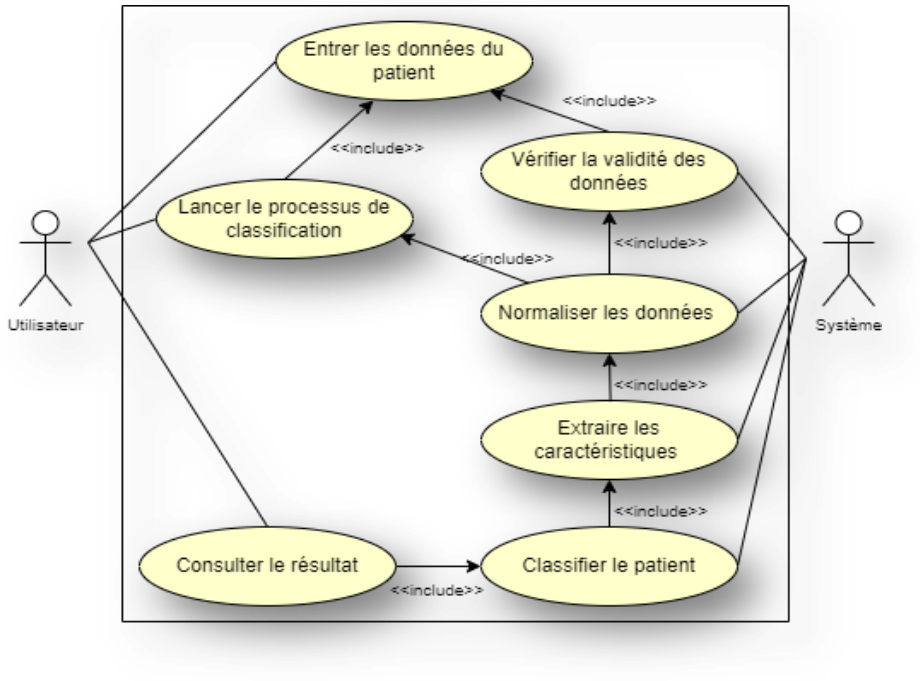

Figure 2-8- Diagramme de cas d'utilisation

2.7.2 Diagramme de Séquence

Le diagramme de séquence dans la **Figure 2-9** représente les interactions entre l'utilisateur et le système en insistant sur la chronologie des envois de messages.

L'utilisateur entre les données du patient puis le système renvoi à lui un message de validité des données, puis l'utilisateur lance le processus de classification (malade/non malade). Le système normalise les données et extrait les caractéristiques les plus signifiantes. Enfin le système classifié le patient et retourne le résultat au utilisateur.

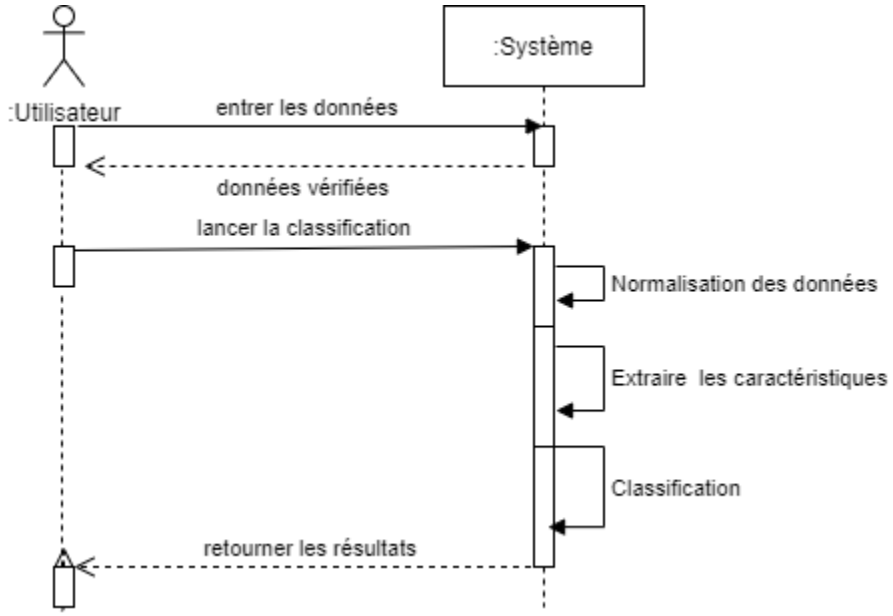

Figure 2-9- Diagramme de Séquence

2.8 Conclusion

Prédire l'état d'un patient est un processus très sérieux car il est lié à la vie des gens. Par conséquent, ce chapitre a été consacré à la présentation du système que nous proposons pour prédire les maladies cardiovasculaires. Le système suggéré utilise l'algorithme du K-plus proche voisin pour la classification, et l'analyse en composantes principales (**ACP**) en tant que méthode d'extraction de caractéristiques, puis nous avons conclu avec la mise au point de notre projet. Nous devrons ensuite réaliser des tests et en vérifier la performance, ce qui sera accompli dans le prochain chapitre.

Chapitre 3

Implémentation et résultats expérimentaux

3.1 Introduction

Dans ce chapitre, nous allons tester le système que nous avons proposé dans le chapitre précédent, en utilisant un échantillon des données sélectionnées, et en comparant les résultats obtenus avec des travaux similaires, tout en présentant la démarche et l'environnement de développement de son application et de sa mise en fonctionnement.

3.2 Validation du système

Pour valider notre Système on utilise l'ensemble de donnée de test de le même ensemble globale que on a fractionné entre apprentissage avec 85% des données et test avec 15%, la **Figure3-1** suivantes montre la structure de l'ensemble de données globale avec 14 colonnes et 303 instances.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

Figure 3-1-La structure de l'ensemble de données

Le fractionnement des données en données pour l'apprentissage et le test a affecté après l'étape d'extraction des caractéristiques, car lors de l'étape de standardisation, les valeurs maximales et minimales peuvent changer dans les deux sous-ensembles (sous-ensemble d'apprentissage, sous-ensemble de tests), ce qui affecte l'écart type et donc les changements dans les valeurs des caractéristiques.

3.2.1 Prétraitement des données

Pour commencer, nous traitons les variables catégoriques et les convertissons en variables indicatrices, pour nous aider à généraliser leur traitement au moyen des mêmes formules, on obtient la structure qui apparait dans la **Figure 3-2** avec 31 colonnes (on affiche le sommet de l'ensemble de donnée).

	age	trestbps	chol	thalach	oldpeak	target	sex_0	sex_1	cp_0	cp_1	...	slope_2	ca_0	ca_1	ca_2	ca_3	ca_4	thal_0	thal_1	thal_2	thal_3
0	63	145	233	150	2.3	1	0	1	0	0	...	0	1	0	0	0	0	0	1	0	0
1	37	130	250	187	3.5	1	0	1	0	0	...	0	1	0	0	0	0	0	0	1	0
2	41	130	204	172	1.4	1	1	0	0	1	...	1	1	0	0	0	0	0	0	1	0
3	56	120	236	178	0.8	1	0	1	0	1	...	1	1	0	0	0	0	0	0	1	0
4	57	120	354	163	0.6	1	1	0	1	0	...	1	1	0	0	0	0	0	0	1	0

5 rows × 31 columns

Figure 3-2- Conversion des variables

Ensuite on applique la formule de standardisation que nous avons présenté dans le chapitre précédent à l'ensemble globale de données, nous obtenons des données normalisées prêtes pour la prochaine étape, comme le montre la **Figure 3-3** (le sommet de l'ensemble de donnée).

	age	trestbps	chol	thalach	oldpeak	target	sex_0	sex_1	cp_0	cp_1	...	slope_2	ca_0	ca_1	ca_2	ca_3	ca_4	thal_0	thal_1	thal_2	th
0	0.952197	0.763956	-0.256334	0.015443	1.087338	1	0	1	0	0	...	0	1	0	0	0	0	0	1	0	
1	-1.915313	-0.092738	0.072199	1.633471	2.122573	1	0	1	0	0	...	0	1	0	0	0	0	0	0	1	
2	-1.474158	-0.092738	-0.816773	0.977514	0.310912	1	1	0	0	1	...	1	1	0	0	0	0	0	0	1	
3	0.180175	-0.663867	-0.198357	1.239897	-0.206705	1	0	1	0	1	...	1	1	0	0	0	0	0	0	1	
4	0.290464	-0.663867	2.082050	0.583939	-0.379244	1	1	0	1	0	...	1	1	0	0	0	0	0	0	1	

5 rows × 31 columns

Figure 3-3 Standardisation des caractéristiques

3.2.2 Extraction des caractéristiques

Nous avons appliqué l'ACP pour extraire les caractéristiques les plus pertinents, et nous avons réduit la dimension de 31 variables à seulement 7 variables, ce qui est inférieur au nombre initial de variables, qui était de 13, comme la **Figure 3-4** montre. Après cette réduction dimensionnelle, on aura économisé le temps et la mémoire nécessaires à l'exécution du programme.

	0	1	2	3	4	5	6
0	0.969801	0.494354	-1.245541	-0.285735	0.693541	-0.774336	-0.704952
1	-1.195970	-0.595284	-1.473504	2.293725	1.834882	-0.448362	0.463814
2	-1.779909	-0.176499	-0.848912	0.563050	0.911993	0.444955	-0.947963
3	-1.612381	0.172336	0.003211	-0.137820	0.413270	-0.947255	0.496781
4	-0.468681	1.429043	1.660341	0.953813	0.151164	0.375653	0.917246
...	...	...	...	...	...	...	...
298	1.051892	-0.264933	0.340272	-0.566682	-0.650461	1.470588	0.850045
299	-0.071622	-1.336476	0.830584	0.855677	0.047336	0.232151	0.657475
300	2.344296	-0.648012	-1.364241	-0.590402	0.706107	-1.374897	0.745961
301	1.392126	-2.456839	-0.426422	-1.437139	-0.744746	0.251603	0.476910
302	-0.764526	0.733665	0.266620	-0.482859	0.088981	-0.235145	-0.965860

303 rows × 7 columns

Figure 3-4 Les caractéristiques extraites

3.2.3 Test du modèle

3.2.3.1 Test sur le sous-ensemble de teste

Dans cette section, nous exécutons le modèle de prédiction sur le sous-ensemble de tests avec différentes valeurs de K, où K varie entre 1 et 20, et nous avons noté l'Accuracy (exactitude) obtenue pour chaque valeur de K.

Nous le ferons dans deux cas :

- Les données sont passées à toutes les étapes du processus de classification proposé, sauf l'extraction de leurs caractéristiques (KNN sans ACP) telle qu'elle apparaît dans la **Figure 3-5.**
- Les données sont passées à toutes les étapes du processus de classification proposé (KNN avec ACP) telle qu'elle apparaît dans la **Figure 3-6**.

Pour évaluer le modèle proposé, on se focalise sur certains critères, à savoir l'exactitude (Accuracy), le rappel, la précision et le F1-score (voir la *sous- section* **chapitre 1.3.5**). L'Accuracy est l'une des mesures de performance les plus importantes pour la classification. Cette évaluation de la performance permet de mieux estimer la qualité du système proposé.

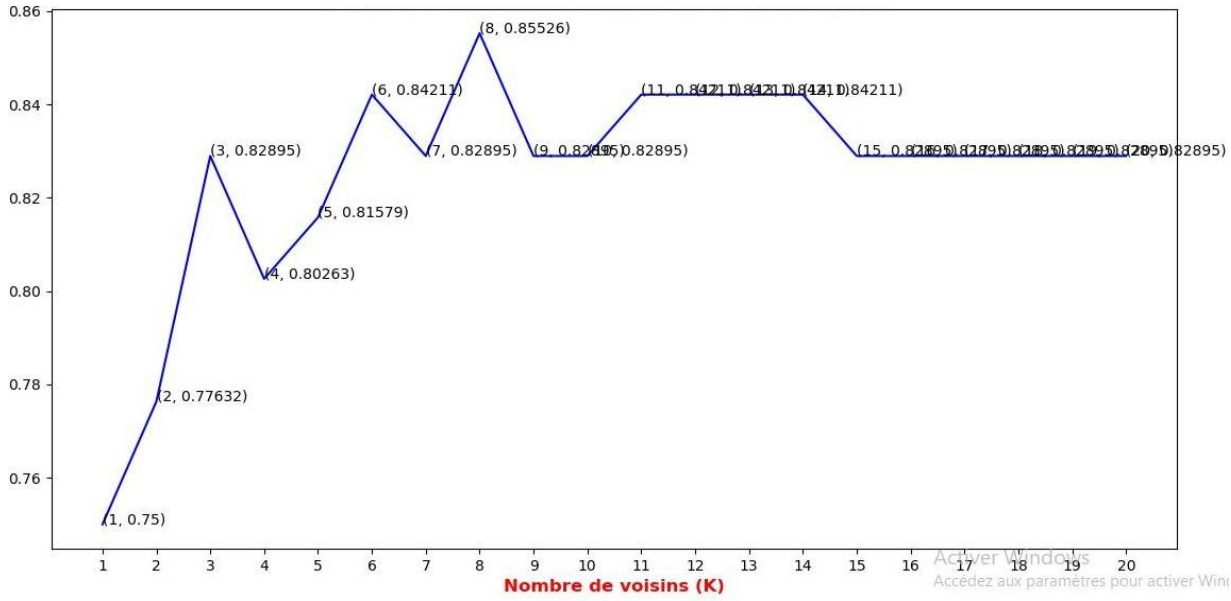

Figure 3-5- Accuracy du KNN sans ACP

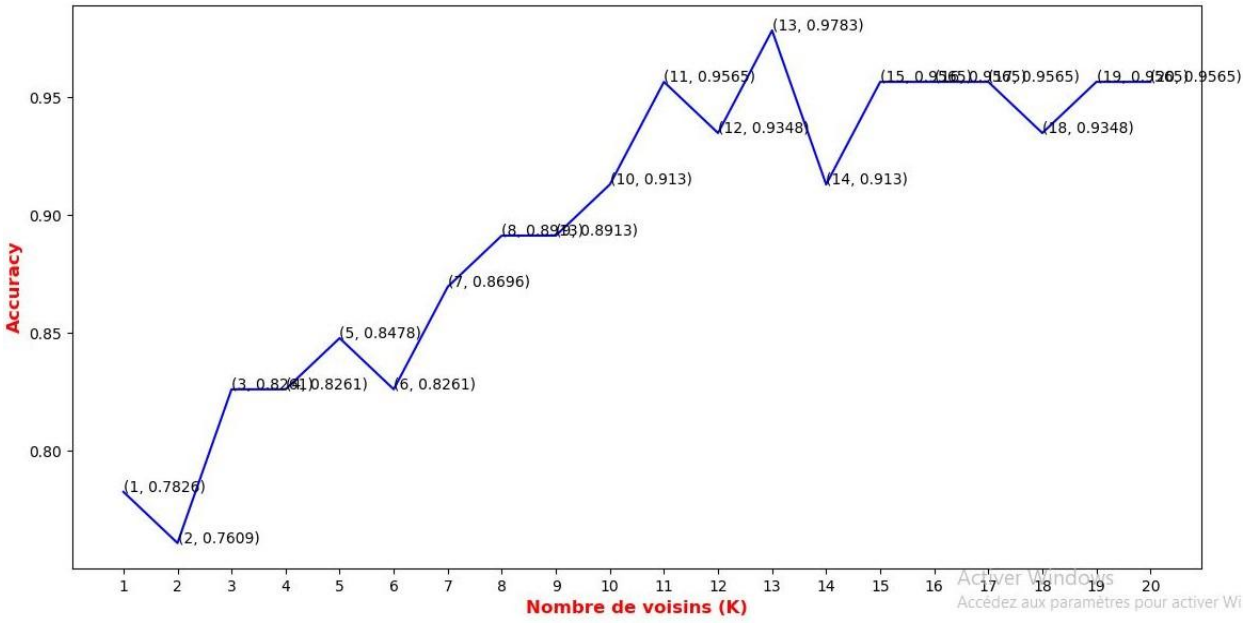

Figure 3-6- Accuracy du KNN avec ACP

3.2.3.2 Test sur un échantillon de teste

1- Pour tester notre modèle nous avons sélectionné un échantillon de patients (P1, P2…P8) de l'ensemble de données afin d'effectuer le test sur eux comme il apparaît dans le **Tableau 3-1.**

Tableau 3-1- Échantillon pour le test

Patient	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Classe
P1	63	1	3	145	233	1	0	150	0	2.3	0	0	1	malade
P2	37	1	2	130	250	0	1	187	0	3.5	0	0	2	Malade
P3	41	0	1	130	204	0	0	172	0	1.4	2	0	2	Malade
P4	56	1	1	120	236	0	1	178	0	0.8	2	0	2	Malade
P5	45	1	3	110	264	0	1	132	0	1.2	1	0	3	Non malade
P6	68	1	0	144	193	1	1	141	0	3.4	1	2	3	Non malade
P7	57	1	0	130	131	0	1	115	1	1.2	1	1	3	Non malade
P8	57	0	1	130	236	0	0	174	0	0.0	1	1	2	Non malade

2- Une fois l'échantillon passé par chacune des étapes d'initialisation et d'extraction des fonctionnalités, nous obtenons des échantillons avec 7 nouvelles caractéristiques. Ensuite, nous transférons ces échantillons à l'étape de la classification en fonction de ces nouvelles caractéristiques, et nous obtenons les résultats qui sont présentés dans le **Tableau 3-2**.

Tableau 3-2- Résultat d'échantillon

Patient	0	1	2	3	4	5	6	Prédiction
P1	0.969801	0.494354	-1.245541	-0.285735	0.693541	-0.774336	-0.704952	Non malade
P2	-1.195970	-0.595284	-1.473504	2.293725	1.834882	-0.448362	0.463814	Malade
P3	-1.779909	-0.176499	-0.848912	0.563050	0.911993	0.444955	-0.947963	Malade
P4	-1.612381	0.172336	0.003211	-0.137820	0.413270	-0.947255	0.496781	Malade
P5	-0.071622	-1.336476	0.830584	0.855677	0.047336	0.232151	0.657475	Non malade
P6	2.344296	-0.648012	-1.364241	-0.590402	0.706107	-1.374897	0.745961	Non malade
P7	1.392126	-2.456839	-0.426422	-1.437139	-0.744746	0.251603	0.476910	Non malade
P8	-0.764526	0.733665	0.266620	-0.482859	0.088981	-0.235145	-0.965860	Malade

3.2.4 Discussion des résultats

Depuis la **Figure 3-5**le modèle de KNN atteint une exactitude de 85.52 % à la valeur de K=8,etla **Figure 3-6** montre que le modèle de KNN avec la méthode ACP atteint une exactitude de 97,83 % à la valeur de K=13, Les performances ont donc augmenté de 12.31 %.

Depuis le résultat d'échantillon du test qui apparait dans le **Tableau 3-2**nous remarquons les erreurs de notre modèle, et pour bien valider notre modèle, nous devons effectuer les autres mesures de performance que nous les avons évoqués dans le premier chapitre pour la même valeur de K pour lequel notre modèle proposé a atteint la plus grande exactitude, donc nous obtenons le **Tableau 3-3.**

Tableau 3-3-Une évaluation du modèle proposé avec quelques critères de performance

	Accuracy (%)	Précision (%)	F1-score (%)	Rappel (%)
KNN avec ACP	97.83	100	97.78	95.65

Afin de déterminer l'efficacité de notre modèle, nous confrontons nos résultats à ceux d'autres travaux reconnus, comme le montre le **Tableau 3-4**.

Tableau 3-4- Comparaison des résultats avec d'autres travaux

Les algorithmes	**Accuracy (%)**
KNN avec ACP pour extraire les caractéristiques	**97.83**
DT[2]	**91**
DT+RF [17]	**88.7**
KNN [4]	**88.5**
LR [5]	**92.3**

3.3 Implémentation du système

3.3.1 Environnement du développement

3.3.1.1 Partie matérielle

Les résultats discutés sont implémentés sur une machine a les caractéristiques suivantes :

- Processeur Intel(R) : Core(TM) i5-2450M CPU @ 2.50GHz 2.50 GHz
- Mémoire installée RAM : 6,00 Go (5,78 Go utilisable)
- Disque dur : 750 Go

3.3.1.2 Partie logicielle

3.3.1.2.1 PyCharm IDE

PyCharm est un environnement de développement intégré (IDE) qui fournit une large gamme d'outils essentiels pour les développeurs Python, étroitement intégrés pour créer un environnement pratique pour le développement productif de Python, du Web et de la science des données. C'est un logiciel multiplateforme qui fonctionne sous Windows, Mac OS X et GNU/Linux qui est développé par l'entreprise **JetBrains**. Il est disponible en édition professionnelle, et en édition communautaire, et dans notre projet nous avons utilisé l'édition communautaire 2019.1.3

3.3.1.2.2 Langage de programmation Python

Python est un langage de programmation interprété de haut niveau. Il prend en charge plusieurs paradigmes de programmation : structuré, orienté objet et fonctionnel. Il a été publié

pour la première fois par Guido van Rossum en 1991 sous le nom de Python 0.9.0. Ses structures de données intégrées de haut niveau, combinées avec le typage et la liaison dynamiques et un grand nombre de bibliothèques, le rendent très intéressant pour le développement rapide des applications. Dans notre projet nous avons utilisé la version 3.7 de Python avec les bibliothèques suivantes :

i. Pandas (version 1.2.4)

Pandas est un package Python fournissant des structures de données rapides, flexibles et expressives conçues pour faciliter le travail avec des données « relationnelles » ou « étiquetées ». De plus, il est en passe de devenir l'outil d'analyse/manipulation de données open source le plus puissant et le plus flexible disponible dans n'importe quel langage. Car il gère la grande majorité des cas d'utilisation typiques dans les domaines de la finance, des statistiques, des sciences sociales et de nombreux domaines de l'ingénierie. C'est pour ces avantages que nous l'avons choisi dans notre projet pour faciliter la manipulation de l'ensemble de donnée que nous avons utilisé sous l'extension (.csv).

ii. NumPy (version 1.21.5)

NumPy, qui signifie Numerical Python, est une bibliothèque composée d'objets de tableau multidimensionnel et d'une collection de routines pour traiter ces tableaux. L'utilisation de NumPy permet de faciliter les opérations mathématiques et logiques sur les tableaux, et son rôle est similaire au package précédent, Panda.

iii. Scikit-learn (version 1.0.2)

Scikit-learn est une bibliothèque d'apprentissage automatique gratuit pour Python. Il comporte divers algorithmes avec ses documentations et tous ces fonctions comme les fonctions de calcule les mesures de performance (Accuracy, précision…), et parmi ces algorithmes : les K-plus proches voisins, les forêts aléatoires, ACP…etc., et il prend également en chargent les bibliothèques numériques et scientifiques Python telles que NumPy.

iv. Matplotlib (version 3.3.4)

Matplotlib est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python et son extension numérique NumPy. Matplotlib rend les choses difficiles faciles, parce que le script Matplotlib est structuré de sorte que quelques lignes de code suffisent dans la plupart des cas pour générer un graphique de données visuel.

v. Seaborn (version 0.11.2)

Seaborn est une bibliothèque Python de visualisation de données basées sur Matplotlib. Il fournit une interface de haut niveau pour dessiner des graphiques statistiques attrayantes et informatives.

vi. Joblib (version 1.0.1)

Joblib est un ensemble d'outils pour fournir un pipelining léger en Python, et en particulier : mise en cache disque transparente des fonctions et réévaluation paresseuse calcul parallèle simple et facile. Joblib est optimisé pour être rapide et robuste sur les données volumineuses en particulier et a des optimisations spécifiques pour les tableaux NumPy. Mais nous utilisons à partir de cet ensemble d'outils ce qui est spécial pour le stockage (Joblib.dump) et la récupération (Joblib.load) des données dans un fichier et en spécifions leur extension (.pkl), afin de stocker notre modèle formé et d'exécuter des tests sur lui.

vii. Tkinter (version 8.6.10)

Tkinter est la bibliothèque GUI standard pour Python, et elle est avec python fournissent un moyen rapide et facile de créer des applications GUI et une puissante interface orientée objet.

3.3.2 Mode d'utilisation de l'application

Notre application contient une fenêtre principale (**Figure 3-6**) avec laquelle l'utilisateur peut remplir un formulaire qui contient les informations sur le patient dont nous voudrions prédire l'état.

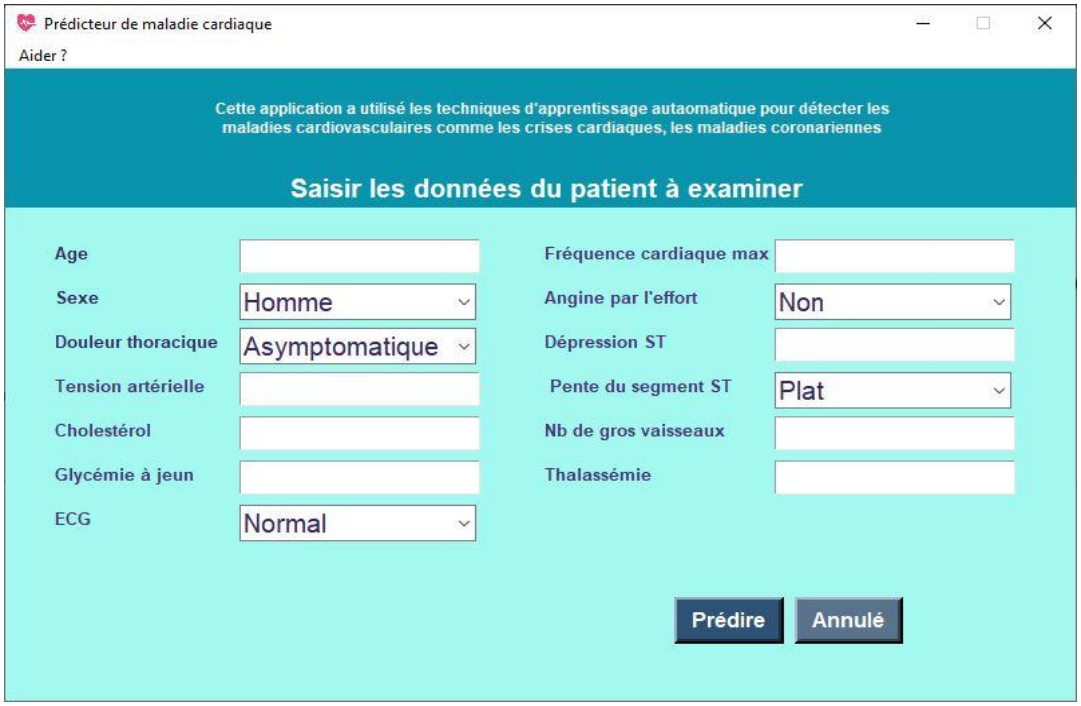

Figure 3-7- La fenêtre principale de l'application

1 Après avoir rempli le formulaire on clique sur le bouton **Prédire** pour lancer la prédiction, ou bien le bouton **Annuler** pour vider le formulaire.

2 Si le patient a une maladie cardiovasculaire un message de confirmation va afficher (**Figure 3-7**).

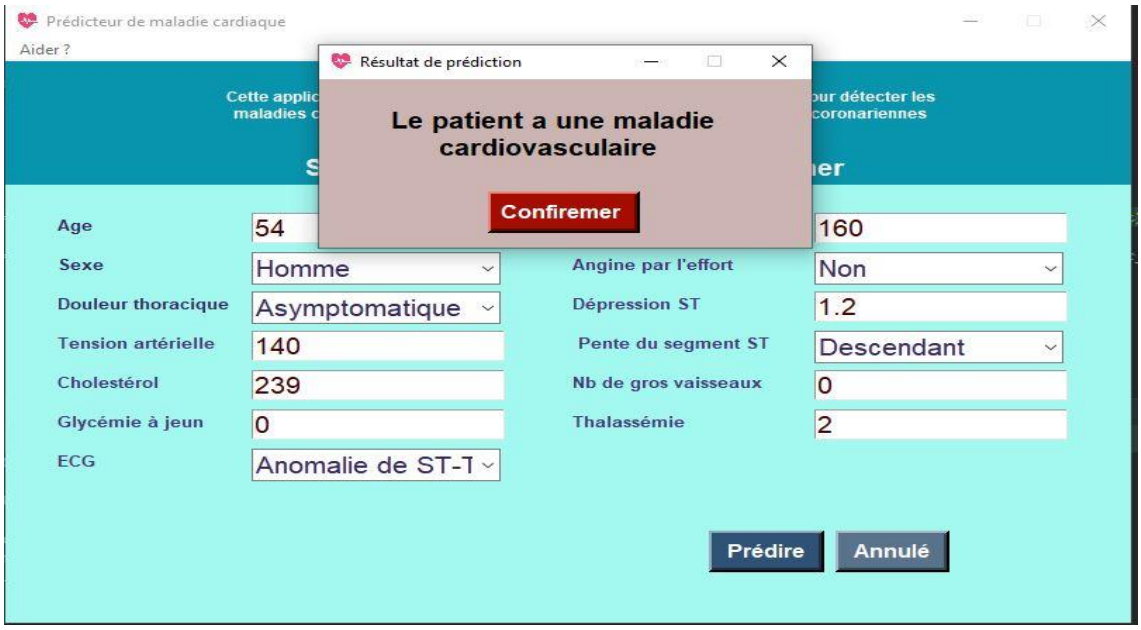

Figure 3-8- Message de confirmation de la maladie

3 Si le patient n'a pas une maladie cardiovasculaire un message de négation va afficher (**Figure 3-8).**

Figure 3-9- Message de négation de la maladie

3.4 Conclusion

Dans ce chapitre, nous avons validé et expérimenté notre système, dont nous avons comparé les résultats avec ceux des travaux déjà existants, et nous avons également présenté les outils d'implémentation de notre application en donnant un exemple de déroulement pour chaque catégorie.

Conclusion

Dans le cadre de ce projet, nous avons étudié les travaux récemment réalisés dans le domaine de la prédiction des maladies cardiaques, et à partir de l'analyse des travaux étudiés, nous avons vu qu'il existe des travaux proposés basés sur la comparaison des performances des algorithmes et le choix de l'algorithme le plus efficace, cependant cette méthode était moins efficace que les systèmes hybrides contenant plus d'un algorithme. Par conséquent, nous avons adapté la méthode PCA à l'algorithme KNN, ce qui nous a permis d’atteindre une exactitude de prédiction de 97,83 %.

Comme le manque de précision dans de tels cas est un facteur décisif car il dépend de la vie des gens, toutes les précisions restent faibles si l'on se rend compte qu'elles reflètent la vie des personnes et pas seulement des chiffres.

À l'avenir, nous espérons pouvoir accroître la précision dans la prédiction des maladies cardiaques afin de réduire autant que possible le taux d'erreur et de pouvoir garantir un taux de confiance plus important permettant d'assurer un maximum de sécurité pour la vie des gens.

Bibliographie

[1] w. h. organisation, «world health organisation,» 09 02 2022. [En ligne]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[2] S. K. M. G. M. M. Animesh Hazra, «Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques A Review,» *Advances in Computational Sciences and Technology,* pp. 2137-2159, 2017.

[3] C. Khanji, Évaluation de la qualité des soins et des services préventifs cardiovasculaires en première ligne, Montréal: Université de Montréal, 2018.

[4] S. P. Dhyan Chandra Yadav, «Prediction of Heart Disease Using Feature Selection and Random Forest Ensemble Method,» *International Journal for Pharmaceutical Research Scholars,* pp. 56-66, 2020.

[5] D. W. Aha, «dataset,» 20 01 2022. [En ligne]. Available: 2022.

[6] B. Mahesh, «Machine Learning Algorithms -A Review,» *International Journal of Science and Research (IJSR),* pp. 381-386, 2018.

[7] I. E. Naqa et M. J. Murphy, *Machine Learning in Radiation Oncology,* Springer International Publishing, 2015.

[8] A.-J. M. I., Q. M. H. et H. Mohammad, «Machine Learning Classification Techniques for Heart Disease Prediction: A Review,» *International Journal of Engineering & Technology,* pp. 5373-5379, 2018.

[9] M. P. M. Fatima, «Survey of machine learning algorithms for disease diagnostic,» *Journal of Intelligent Learning Systems and Applications,* pp. 1-16, 2017.

[10] S. V. G. S. G. D. P. H. A. J. G. Seyedamin Pouriyeh, «A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease,» *22nd IEEE Symposium on Computers and Communication (ISCC 2017),* 2017.

[11] M. B. Sadegh Bafandeh Imandoust, «Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background,» *S B Imandoust et al. Int. Journal of Engineering Research and Applications* , pp. 605-610, 2013.

[12] S. M. A. M. H. J. S. M. j. R. Sasan Karamizadeh, «Advantage and Drawback of Support Vector Machine Functionality,» *IEEE 2014 International Conference on Computer, Communication, and Control Technology* , pp. 63-65, 2014.

[13] E. G. P. J. C. Andrius Vabalas, «Machine learning algorithm validation with a limited sample size,» *journals.plos.org*, 2019.

[14] C. Gupta, «Cardiac Disease Prediction using Supervised Machine Learning Techniques,» *Journal of Physics: Conference Series*, 2022.

[15] Krishnan.J et Gretha.S, «Prediction of Heart Disease Using Machine Learning Algorithms.,» *1st international conference of innovations in information and communication technology (ICIIcT)*, 2019.

[16] X.-Y. Gao, A. A. Ali, H. S. Hassan et E. M. Anwar, «Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method,» *hindawi*, p. 10, 2021.

[17] M. Kavitha, G.Gnaneswar, R.Dinesh, Y. Sai et R. Suraj, «Heart Disease Prediction using Hybrid machine Learning Model,» chez *Proceedings of the Sixth International Conference on Inventive Computation Technologies [ICICT 2021]*, 21.

[18] D. LAPP, 2022. [En ligne]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[19] H. Jindal, «Heart disease prediction using machine learning algorithms,» chez *IOP Conference Series: Materials Science and Engineering*, 2021.