SOMMAIRE

INTRODUCTION

Les maladies cardiaques demeurent un enjeu majeur de santé publique, représentant une cause significative de morbidité et de mortalité à l'échelle mondiale.

Selon l'OMS1, il meurt chaque année plus de personnes en raison de maladies cardio-vasculaires que de toute autre cause. On estime à 17,7 millions le nombre de décès imputables aux maladies cardio-vasculaires, soit 31% de la mortalité mondiale totale. Parmi ces décès, on estime que 7,4 millions sont dus à une cardiopathie coronarienne et 6,7 millions à un AVC (chiffres 2015). (OMS 2017)

Comprendre les facteurs qui contribuent au diagnostic précis de ces affections revêt une importance capitale pour améliorer les stratégies de prévention et de prise en charge.

Alors face à la complexité des maladies cardiaques, il devient impératif d'identifier des marqueurs fiables et des indicateurs précis pour un diagnostic précoce. La question centrale qui guide notre démarche est la suivante : quels sont les facteurs déterminants qui contribuent de manière significative au diagnostic des maladies cardiaques ?

Notre étude se propose donc d'explorer en profondeur ces aspects, en mettant en lumière des éclairages significatifs obtenus à partir d'une analyse approfondie des données.

Pour ce faire, notre démarche s'articulera autour de trois grandes parties. Tout d'abord, nous aborderons le prétraitement des données, où nous mettrons en œuvre des méthodes rigoureuses pour garantir la qualité et la cohérence des informations recueillies. Ensuite, nous procéderons à une analyse univariée, examinant chaque variable de manière isolée pour déceler des tendances ou des relations potentielles avec les maladies cardiaques. Enfin, la troisième partie portera sur l'analyse bivariée, permettant d'explorer les interactions entre différentes variables et d'identifier des associations significatives.

[1] Organisation Mondiale de la Santé

Cette approche structurée nous permettra d'obtenir des éclairages pertinents et d'enrichir la base de connaissances sur les maladies cardiaques, contribuant ainsi à l'amélioration des stratégies de prévention et de prise en charge.

PREMIERE PARTIE: PRETRAITEMENT DES DONNEES

Introduction

Selon une étude menée par IBM, la mauvaise qualité des données coûte désormais 3 100 milliards de dollars par an aux États-Unis. Et ce coût augmente de façon exponentielle. Cette étude met en exergue l'importance d'avoir des données de tres bonnes qualités afin d'avoir des résultats pertinents et fiables. Le nettoyage des données ou data cleansing est donc une étape essentielle de l'analyse statistique et de la Data Science. Il s'agit de corriger ou supprimer des enregistrements inexacts dans des jeux de données afin pouvoir les exploiter par la suite. C'est un processus qui vise, de manière générale, à améliorer la qualité des données. En pratique, il consiste à importer les données dans un logiciel statistique, dans notre cas l'IDE Rstudio du logiciel R. Ensuite, faire une première visualisation pour avoir une idée succincte de la structure de nos données et enfin, identifier et corriger les valeurs manquantes, valeurs aberrantes, valeurs extrêmes afin de qu'elles puissent être plus cohérentes et sans erreurs.

I. DESCRIPTION DES DONNEES

1. Dictionnaire des données

Le dictionnaire de donnée ou spécification fonctionnelle des données est un document qui renseigne sur le contexte d'une base de données et qui fournit les informations nécessaires pour pouvoir l'interpréter. C'est donc un outil tres important qui facilite la compréhension de la structure des données pour les administrateurs et les utilisateurs. Dans le cadre de notre étude ce document se présente comme suit :

Tableau 1: Dictionnaire des données

VARIABLE	NATURE	DESCRIPTION	MODALITES
Age	Quantitative	âge du patient	Numérique Entier en annees

Sex	Qualitative	sexe/genre du patient	M : Masculin ou F : Feminin
ChestPainType	Qualitative	Type de douleur thoracique	TA : Angine de poitrine typique, ATA : Angine de poitrine atypique, NAP: Douleur autre que l'angine, ASY: Asymptomatique
RestingBP	Quantitative	Tension artérielle au repos (en mm Hg)	Numerique Decimal
Cholesterol	Quantitative	Taux de cholestérol (en mm/dl)	Numerique Decimal
FastingBS	Qualitative	glycémie à jeun	Qualitative Binaire (1 si FastingBS > 120 mg/dl, 0 sinon)
RestingECG	Qualitative	électrocardiogramme au repos	LVH, Normal et ST
MaxHR	Quantitative	Fréquence cardiaque maximale (en battements par minutes)	Numerique Decimal
ExerciseAngina	Qualitative	angine induite par l'exercice	Non.angine (Présence d'angine apres exercice) et Oui.angine(Absence d'angine apres exercice)
Oldpeak	Quantitative	Dépression du segment ST induite par l'exercice	Numerique Decimal
ST_Slope	Qualitative	pente du segment ST	Up: croissante ; Flat : constante ; Down : décroissante
HeartDisease	Qualitative	Présence ou absence de maladie cardiaque	Non ou Oui

2. Visualisation du jeu de donnée brute et de sa structure

Avant de débuter tout analyse, il est primordial de faire une première visualisation des données dans le but de se faire une idée succincte de la répartition des données, des noms des différentes colonnes et de l'agencement des différentes lignes. Par ailleurs, cette

première phase est appuyée par la visualisation de la structure et de la statistique descriptive du dataframe.

Tableau 2: Aperçu du jeu de données

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

(Source: Kaggle, nos calculs 2021)

Ce premier aperçu met à nu quelques anomalies au niveau des types des variables FastingBS et HeartDisease qui sont considérées comme des variables quantitatives alors qu'elles sont qualitatives. Nous allons donc les recoder en variables qualitatives.

Tableau 3: Structure avant recodage

'data.frame': 918 obs. of 12 variables:
$ Age : int 40 49 37 48 54 39 45 54 37 48 ...
$ Sex : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 1 2 2 1 ...
$ ChestPainType : Factor w/ 4 levels "ASY","ATA","NAP",..: 2 3 2 1 3 3 2 2 1 2
$ RestingBP : int 140 160 130 138 150 120 130 110 140 120 ...
$ Cholesterol : int 289 180 283 214 195 339 237 208 207 284 ...
$ FastingBS : int 0 0 0 0 0 0 0 0 0 0 ...
$ RestingECG : Factor w/ 3 levels "LVH","Normal",..: 2 2 3 2 2 2 2 2 2 2
$ MaxHR : int 172 156 98 108 122 170 170 142 130 120 ...
$ ExerciseAngina: Factor w/ 2 levels "N","Y": 1 1 1 2 1 1 1 1 2 1 ...
$ Oldpeak : num 0 1 0 1.5 0 0 0 0 1.5 0 ...

$ ST_Slope : Factor w/ 3 levels "Down","Flat",..: 3 2 3 2 3 3 3 3 2 3 ...
$ HeartDisease : int 0 1 0 1 0 0 0 0 1 0 ...