Projet_R_2_Medical Research

Marie Thomassin , Sewa Fumey et Zakaria Maanane

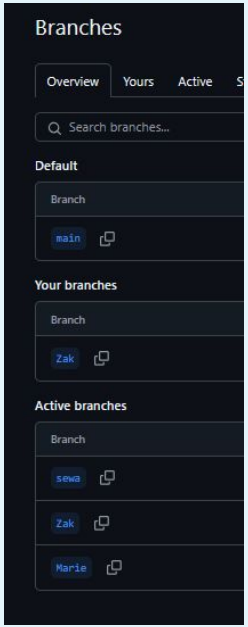
Le Sujet

Les **maladies cardiovasculaires**, et en particulier les crises cardiaques, représentent aujourd'hui l'une des principales causes de mortalité dans le monde. **Comprendre les facteurs** qui influencent la survenue d'un arrêt cardiaque est un enjeu majeur en santé publique, tant pour **améliorer la prévention** que pour optimiser la prise en charge des patients à risque.

Ce projet a pour objectif d'**analyser un jeu de données médicales** comprenant diverses informations cliniques (âge, sexe, rythme cardiaque, tension artérielle, taux de sucre, troponine, etc.) afin de mieux **comprendre les caractéristiques associées à la survenue d'une crise cardiaque.** À travers des **visualisations interactives et une modélisation prédictive**, l'ambition est d'**identifier les variables les plus déterminantes** dans l'apparition de ce type d'événement critique.

L'objectif est donc d'identifier le facteur permettant de différencier un patient sain d'un patient ayant déjà été victime d'une crise cardiaque, ou plus largement d'un événement cardiaque ayant causé des lésions détectables, afin de mieux prévenir d'éventuelles rechutes.

Méthode et répartition des tâches

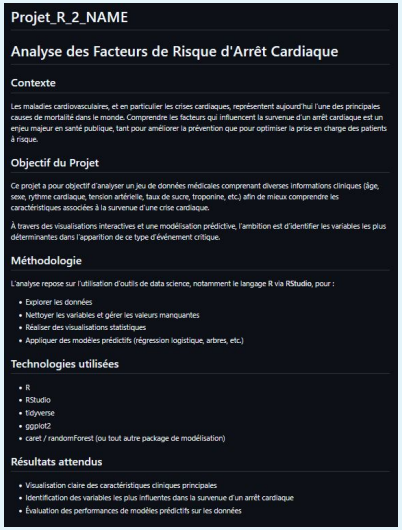


Étape 1 : Sélection et Exploration du Dataset
Objectif : Choisir un dataset pertinent et comprendre sa structure.

Étape 2 : Nettoyage et Transformation des Données
Objectif : Corriger les erreurs et structurer les données pour l'analyse.

Étape 3 : Développement du Dashboard Interactif sous R Shiny
Objectif : Créer une interface fluide permettant une exploration dynamique des données.

Étape 4 : Synthèse et Communication des Résultats
Objectif : Présenter une analyse claire et exploitable.


Projet_R_2_NAME

Analyse des Facteurs de Risque d'Arrêt Cardiaque

Contexte

Les maladies cardiovasculaires, et en particulier les crises cardiaques, représentent aujourd'hui l'une des principales causes de mortalité dans le monde. Comprendre les facteurs qui influencent la survenue d'un arrêt cardiaque est un enjeu majeur en santé publique, tant pour améliorer la prévention que pour optimiser la prise en charge des patients à risque.

Objectif du Projet

Ce projet a pour objectif d'analyser un jeu de données médicales comprenant diverses informations cliniques (âge, sexe, rythme cardiaque, tension artérielle, taux de sucre, troponine, etc.) afin de mieux comprendre les caractéristiques associées à la survenue d'une crise cardiaque.

À travers des visualisations interactives et une modélisation prédictive, l'ambition est d'identifier les variables les plus déterminantes dans l'apparition de ce type d'événement critique.

Méthodologie

L'analyse repose sur l'utilisation d'outils de data science, notamment le langage R via RStudio, pour :

- Explorer les données
- Nettoyer les variables et gérer les valeurs manquantes
- Réaliser des visualisations statistiques
- Appliquer des modèles prédictifs (régression logistique, arbres, etc.)

Technologies utilisées

- R
- RStudio
- tidyverse
- ggplot2
- caret / randomForest (ou tout autre package de modélisation)

Résultats attendus

- Visualisation claire des caractéristiques cliniques principales
- Identification des variables les plus influentes dans la survenue d'un arrêt cardiaque
- Évaluation des performances de modèles prédictifs sur les données

Choix un dataset pertinent et compréhension de sa structure.

Jeu de données : liste de patients ayant subi une crise cardiaque, accompagnée de leurs caractéristiques (âge, sexe, fréquence cardiaque, etc.).

- Les relevées d'informations sont des mesures en chiffre donc on évite de basculer nos données en binaires et on évite de basculer des données qui ne sont pas adaptées à être passée en binaire (ex carburant automobile)
- Possibilité de visualisation graphique large et de croiser les chiffres des résultats (1 en abscisse et 2 en ordonnées)

	A	B	C	D	E	F	G	H	I
1	Age	Gender	Heart rate	Systolic blood pressu	Diastolic blood press	Blood sugar	CK-MB	Troponin	Result
2	64	1	66	160	83	160	1.8	0.012	negative
3	21	1	94	98	46	296	6.75	1.06	positive
4	55	1	64	160	77	270	1.99	0.003	negative
5	64	1	70	120	55	270	13.87	0.122	positive
6	55	1	64	112	65	300	1.08	0.003	negative
7	58	0	61	112	58	87	1.83	0.004	negative
8	32	0	40	179	68	102	0.71	0.003	negative
9	63	1	60	214	82	87	300	2.37	positive
10	44	0	60	154	81	135	2.35	0.004	negative
11	67	1	61	160	95	100	2.84	0.011	negative
12	44	0	60	166	90	102	2.39	0.006	negative
13	63	0	60	150	83	198	2.39	0.013	negative
14	64	1	60	199	99	92	3.43	6.37	positive

age	age du patient
Sexe	homme/femme
Fréquence cardiaque	nombre de bâtiment cardiaque par minutes
Pression artérielle sytolique	la pression dans les artères lorsque le coeur se contracte
Pression artérielle diastolique	la pression artérielle entre les battements cardiaque
glycémie	taux de glucose dans le sang
CK-mb	une enzyme cardiaque libérée lors des lésions du muscle cardiaque
Troponine	Un biomarqueur de protéines très spécifique pour les lésions du muscle cardiaque
Résultat	L'étiquette de résultat indiquant si le patient a eu ou non une crise cardiaque

IDENTIFICATION DES PROBLEME DU DATASET

➔ Quels sont les principaux problèmes du dataset ?

Supprimer les doublons, nettoyer les types de données,

supprimer ou recoder les valeurs aberrantes,

et nettoyer les colonnes de textes.

➔ Quelles variables nécessitent un nettoyage en priorité ?

Vérification des valeurs aberrante de toutes les variables

Passage de Gender en binaire

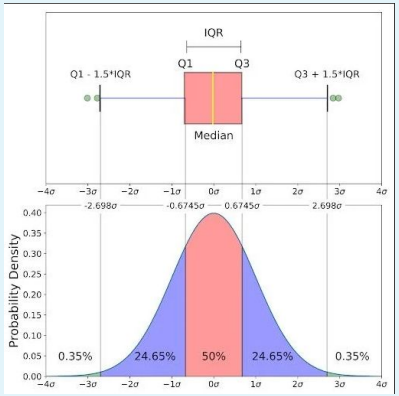
Renommer les résultats positif ou négatif en Sain – Pathologique (malade) pour plus de clarté

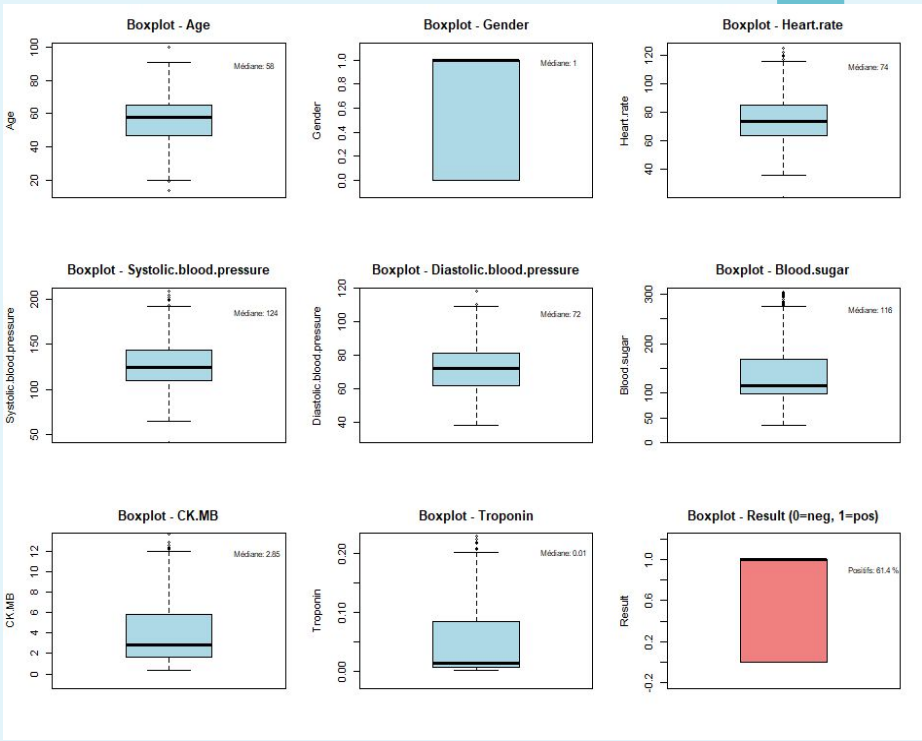
Détecter les valeurs aberrantes avec un boxplot

Interprétations :

➔ Quels sont les principaux problèmes du dataset ?

➔ Quelles variables nécessitent un nettoyage en priorité ?





Nettoyage et Transformation des Données

→ Comment avez-vous corrigé les problèmes de qualité des données ?

Dans notre code R , nous avons utilisée df$gender et df$result

```r
# 🔄 Transformation des données
df$gender <- factor(df$gender, levels = c(0, 1), labels = c("Femme", "Homme"))
df$result <- factor(df$result, levels = c("negative", "positive"), labels = c("Sain", "Pathologique"))
df <- df %>% distinct()
```

Ici nous avons le nétoyage des noms de colonnes

```r
# 🧽 Nettoyage des noms de colonnes
df <- clean_names(df)
```

Ici nous avons la détéction des doublons

```r
# 🧹 Détection des doublons
df <- df %>% distinct()
```

Détéction avec Boxplot des valeurs abérrantes

```r
# 📊 Détection des valeurs aberrantes pour CK-MB
boxplot(df$ck_mb, main = "Boxplot CK-MB", col = "lightblue")

# 🚫 Filtrage des valeurs CK-MB extrêmes (facultatif, à ajuster)
df <- df %>% filter(ck_mb < 50)
```

Face à l'impossibilité de mettre de côté toutes les données aberrantes , car dans notre cas, ce sont justement les valeurs qui se distinguent qui nous permettent de différencier les patients sains des patients pathologiques — nous avons mis en place un filtre qui nous permet d'ajuster le maximum ou le minimum que l'on souhaite conserver.

➔ Quels traitements ont eu le plus d'impact sur la structure des données ?

Le traitement qui a eu le plus d'impact est la détection et gestion des valeures aberrantes , cela nous a permis de pouvoir comparer sur une même échelle les patients sain et patient pathologique de manière visible .

```r

# ================  Filtrage des valeurs aberrantes ==========================================
df <- df %>%
  filter(!(result == "Pathologique" & ck_mb > 20)) %>%  # CK-MB > 20 pour pathologiques
  filter(troponin <= 2.5)  # Troponine > 2.5

# 🚨 Filtrage des valeurs selon critères médicaux spécifiques
print(paste("Nombre de lignes avant filtrage:", nrow(df)))

df <- df %>%
 filter(
   troponin <= 0.35,                  # Troponine ≤ 1.0 ng/mL
   ck_mb <= 10,                      # CK-MB ≤ 10
   diastolic_blood_pressure <= 100,    # Pression diastolique ≤ 100 mmHg
   blood_sugar <= 300,                # Glycémie ≤ 300 mg/dL
   systolic_blood_pressure <= 170,     # Pression systolique ≤ 170 mmHg
   heart_rate <= 110                  # Fréquence cardiaque ≤ 110 bpm
 )

print(paste("Nombre de lignes après filtrage:", nrow(df)))

#==========================================================================================

```

DEMO : ÉTAPE 3 : Développement du Dashboard Interactif sous R Shiny

Conclusion

Parmi tous les facteurs mesurés, la troponine est le seul à présenter une élévation marquée chez les patients pathologiques. Ce biomarqueur s'impose donc comme l'indicateur le plus fiable pour confirmer un infarctus.

Pour aller plus loin, il serait pertinent d'utiliser un jeu de données plus complet, incluant des variables comme le cholestérol, le tabagisme ou les antécédents familiaux.