École Supérieure Privée d'Ingénierie et de Technologies

Projet Statistiques Semestre : 1

Classes : 4DS Nombre de pages : 3

AU : 2023 - 2024

Description de base des donnees

Les maladies cardiovasculaires (MCV) sont la principale cause de décès à l'échelle mondiale, causant environ 17,9 millions de vies chaque année, ce qui représente 31 % de tous les décès dans le monde. Quatre décès sur cinq liés aux MCV sont dus à des crises cardiaques et des accidents vasculaires cérébraux (AVC), et un tiers de ces décès surviennent prématurément chez des personnes de moins de 70 ans. L'insuffisance cardiaque est un événement courant causé par les MCV, et ce jeu de données contient 12 variables qui peuvent être utilisées pour prédire une éventuelle maladie cardiaque. Les personnes atteintes de maladies cardiovasculaires ou présentant un risque élevé de maladies cardiovasculaires (en raison de la présence d'un ou de plusieurs facteurs de risque tels que l'hypertension, le diabète, l'hyperlipidémie ou une maladie déjà établie) ont besoin d'une détection précoce et d'une prise en charge.
Les variables sont données dans ce tableau.

Nom de la Variable	Type	Description de la Variable
Age	numérique	Âge du patient
Sexe	nominale	Sexe du patient (M : Masculin, F : Féminin)
TypeDouleurThoracique	nominale	Type de douleur thoracique (TA : Angine typique, ATA : Angine atypique, NAP : Douleur non angineuse, ASY : Asymptomatique)
TensionArterielleRepos	nominale	Tension artérielle au repos
Cholesterol	numérique	Cholestérol sérique
GlycemieJeune	numérique	Glycémie à jeun (1 : si GlycémieJeûne ¿ 120 mg/dL, 0 : sinon)
ECGRepos	nominale	Résultats de l'électrocardiogramme au repos (Normal : Normal, ST : Anomalie de l'onde ST-T (inversions de l'onde T et/ou élévation ou dépression de ¿ 0,05 mV), LVH : Hypertrophie ventriculaire gauche probable ou certaine selon les critères d'Estes)
FrequenceCardiaqueMax	numérique	Fréquence cardiaque maximale atteinte (Valeur entre 60 et 202)
AngineExercice	nominale	Angine induite par l'exercice (Y : Oui, N : Non)
DepressionAncienne	numérique	Dépression de l'onde ST
PenteSTExercice	nominale	Pente du segment ST d'exercice maximal (Up : montante, Flat : plate, Down : descendante)
MaladieCardiaque	numerique	Présence de maladie cardiaque (1 : maladie cardiaque, 0 : Normal)

Table 1 – Description des Variables de la Base de Données

Vous devez importer le jeu de données `HeartDisease.csv` dans R et afficher les 10 premières lignes pour examiner les premières observations.

Variables cibles

— La variable qualitative cible est "`MaladieCardiaque`"

— La variable quantitative cible est "`Cholesterol`"

Partie 1 : Préparation et Exploration des Données

1. **Préparation des Données**
   Effectuez une vérification de la qualité des données, traitez les valeurs manquantes, et gérez les données aberrantes si nécessaire.

   Effectuez un prétraitement des variables, notamment le codage des variables catégorielles et la mise à l'échelle des variables si nécessaire.
2. **Analyse univariée**
   Réalisez une analyse univariée pour chaque variable du jeu de données, en calculant des statistiques descriptives et en créant des graphiques pertinents pour visualiser la distribution des données.
3. **Analyse bivariée**
   Explorez les relations entre les variables à l'aide de graphiques et de tests statistiques, en mettant l'accent sur la relation entre "MaladieCardiaque" et les autres variables du jeu de données.

4. **Régression linéaire**
   Effectuez des analyses de régression linéaire (simple/multiple) pour examiner la relation entre "Cholestérol" (variable quantitative) et d'autres variables potentiellement liées.

Partie 2 : Analyse Multivariée

1. **Analyse de variance (ANOVA)**
   Si nécessaire, réalisez une analyse de variance pour comparer plusieurs groupes en fonction de certaines variables.
2. **Régression logistique**
   Effectuez une étude bibliographique sur les modèles de régression logistique et les cas d'utilisations.
   Mettez en place un modèle de régression logistique pour prédire la présence de maladie cardiaque en utilisant des variables explicatives pertinentes.

Partie 3 : Modélisation avancée

1. **Analyse discriminante linéaire (ADL)**
   Effectuez une étude bibliographique sur les modèles de l'analyse en composantes principales et les cas d'utilisations.

Si nécessaire, utilisez l'analyse discriminante linéaire pour évaluer la capacité de discrimination des variables explicatives entre les groupes de patients atteints de maladie cardiaque et ceux sans maladie cardiaque.

Livrables

— Un **script R** et/ou ficher **RMarkdown**, le script doit etre fonctionnel et exécutable, contenant toutes les taches demandées.

— Une présentation (max. 30 diapositives) pour expliquer vos analyses et conclusions.

Références

1. Madsen, Henrik ; Thyregod, Poul (2011). Introduction to General and Generalized Linear Models. Chapman & Hall/CRCC.
2. :Jolliffe Ian T. and Cadima Jorge 2016 Principal component analysis : a review and recent developments Phil. Trans. R. Soc.
3. Izenman, A.J. (2013). Linear Discriminant Analysis. In : Modern Multivariate Statistical Techniques. Springer Texts in Statistics. Springer, New York, NY.