Master Web Intelligence et Science des Données

Méthode d'Analyse Factorielle pour l'Evaluation du Risque Cardiaque

Réalisées par :

BAHNAS Hajar & BOURASSI Oumaima

Encadré par :

Mr.ALJ AbdelKamel

Année Universitaire :

2024-2025

Contents

Remerciements

Nous tenons à exprimer notre profonde gratitude à **Monsieur Abdelkamel Alj**, enseignant du module *Modélisation et Statistiques*, pour son engagement et sa pédagogie.

Ses conseils précieux et son accompagnement nous ont permis de mieux appréhender les concepts statistiques et d'avancer efficacement dans notre travail.

Merci pour son soutien et sa disponibilité tout au long de ce semestre.

1 Introduction

Les maladies cardiovasculaires représentent l'une des principales causes de mortalité à l'échelle mondiale. Selon l'Organisation Mondiale de la Santé (OMS), elles sont responsables de près de 17,9 millions de décès chaque année, soit environ 32 % de tous les décès mondiaux. Parmi ces maladies, les crises cardiaques figurent parmi les formes les plus courantes et les plus graves.

L'analyse de données cliniques et biométriques joue un rôle crucial dans l'identification des principaux facteurs contribuant au risque cardiaque. Ces données incluent des indicateurs tels que l'âge, le sexe, le cholestérol, la pression artérielle, et d'autres variables médicales. Cependant, l'interprétation de ces variables devient complexe lorsque leur nombre augmente, rendant difficile l'extraction des tendances et des corrélations clés.

Pour résoudre ce problème, des techniques statistiques telles que l'Analyse en Composantes Principales (ACP) sont utilisées. L'ACP permet de réduire la dimensionnalité des données tout en conservant les informations les plus pertinentes. Cette méthode est particulièrement utile pour simplifier l'analyse des données cliniques, identifier les variables les plus influentes et améliorer la compréhension des relations entre elles.

2 Problématique

Les maladies cardiovasculaires, et en particulier les crises cardiaques, représentent un enjeu majeur de santé publique à l'échelle mondiale. Leur prévention repose sur la capacité à identifier les facteurs de risque les plus significatifs à partir de données cliniques souvent complexes et multidimensionnelles.

Cependant, l'analyse de ces données peut être entravée par le grand nombre de variables impliquées et la difficulté d'interpréter les interactions entre elles. Face à cette complexité, l'Analyse en Composantes Principales (ACP) peut-elle aider à mieux comprendre les facteurs influençant le risque cardiaque ?

Ainsi, nous nous posons les questions suivantes :

- Comment peut-on identifier les principales variables expliquant la variabilité du risque cardiaque à l'aide de l'ACP ?
- Quels sont les avantages de l'ACP pour l'analyse de données médicales ?
- Comment choisir les composantes principales les plus pertinentes ?
- Comment les résultats peuvent-ils améliorer la prédiction du risque cardiaque ?

3 Solution proposée

- Identifier les variables les plus influentes du dataset via l'ACP ;
- Réduire la dimensionnalité des données pour une meilleure interprétation ;
- Visualiser les résultats sous forme de graphiques pour une meilleure compréhension.

4 Introduction à l'Analyse Factorielle

4.1 Définition et Objectifs

L'analyse factorielle est un ensemble de techniques statistiques permettant d'explorer des relations cachées entre plusieurs variables et de réduire la complexité des données.

4.2 Les Différentes Méthodes d'Analyse Factorielle

4.2.1 Analyse en Composantes Principales (ACP)

L'ACP est utilisée pour réduire la dimensionnalité des données tout en conservant un maximum d'information.

4.2.2 Analyse Factorielle Exploratoire (AFE)

L'AFE est souvent utilisée pour détecter des structures latentes dans les données.

4.2.3 Analyse Factorielle des Correspondances (AFC)

L'AFC est adaptée aux données qualitatives sous forme de tableaux de contingence.

4.2.4 Analyse Factorielle Confirmatoire (AFC)

Méthode qui vise à vérifier une hypothèse spécifique sur la structure des données.

5 L'Analyse en Composantes Principales (ACP)

5.1 Définition et Objectifs

L'ACP est une méthode statistique permettant de transformer un ensemble de variables corrélées en un nouvel ensemble de variables non corrélées appelées **composantes principales**.

5.2 Formulation Mathématique de l'ACP

5.2.1 Standardisation des Données

Avant d'appliquer l'ACP, il est essentiel de centrer et réduire les données :

$$X_{standard} = \frac{X - \mu}{\sigma}$$

5.2.2 Calcul de la Matrice de Covariance

La matrice de covariance des données est donnée par :

$$\Sigma = \frac{1}{n} X^T X$$

5.2.3 Décomposition en Valeurs Propres

On résout le problème suivant :

$$\Sigma V = \lambda V$$

où V est la matrice des vecteurs propres et λ les valeurs propres.

5.3 Critères de Sélection du Nombre de Composantes

5.3.1 Critère du Pourcentage de Variance Expliquée

On retient les composantes qui expliquent un pourcentage significatif de la variance totale.

5.3.2 Critère de Kaiser (Valeurs Propres ¿ 1)

On garde uniquement les composantes ayant une valeur propre supérieure à 1.

5.3.3 Scree Plot (Critère de Cattell)

Le "Scree Plot" permet d'observer le point d'inflexion où la variance commence à stagner.

5.4 Interprétation des Résultats de l'ACP

5.4.1 Cercle des Corrélations

Le cercle des corrélations permet d'analyser l'influence des variables initiales sur les composantes.

5.4.2 Projection des Individus

On représente les observations projetées dans l'espace des composantes principales.

6 Modèles Associés à l'Analyse Factorielle et à l'ACP

6.1 Modèle de l'Analyse Factorielle Classique

L'idée est que chaque variable est une combinaison linéaire de facteurs latents plus un bruit aléatoire.

6.2 ACP et Réduction de Dimension

L'ACP est souvent utilisée pour réduire le nombre de variables explicatives sans perdre trop d'informations.

6.3 ACP et Régression Linéaire

Une variante de l'ACP appelée **PCR (Principal Component Regression)** est utilisée pour éviter la colinéarité.

6.4 ACP Non Linéaire et Avancées Récentes

6.4.1 ACP Kernel

Utilisée pour capturer des structures non linéaires dans les données.

6.4.2 ACP Sparse

Permet de sélectionner un sous-ensemble des variables les plus pertinentes.

6.4.3 ACP et Machine Learning

L'ACP est couramment utilisée en prétraitement dans des modèles de deep learning.

6.4.4 Conclusion

L'ACP est une méthode puissante pour l'analyse exploratoire et la réduction de dimensionnalité des données. Son application aux données médicales permet d'extraire les facteurs de risque les plus pertinents.

7 Présentation du Dataset

7.1 Description Générale

L'étude repose sur un dataset contenant **8 763 observations** et **22 variables**, qui décrivent divers facteurs de risque liés aux maladies cardiovasculaires. Ce dataset, intitulé *Heart Attack Risk Prediction*, regroupe des informations cliniques et comportementales de patients.

7.2 Description des Variables

Les principales variables incluses sont les suivantes :

7.2.1 Variables démographiques

- **Age** : Âge du patient en années.
- **Sex** : Sexe du patient (1 : Femme, 0: Homme).

7.2.2 Facteurs biologiques

- **Cholesterol** : Niveau de cholestérol sanguin (mg/dL).
- **Blood.Pressure** : Pression artérielle sous la forme "Systolique/Diastolique".
- **Heart.Rate** : Fréquence cardiaque en battements par minute (BPM).
- **BMI** : Indice de Masse Corporelle (IMC).
- **Triglycerides** : Niveau de triglycérides sanguins (mg/dL).

7.2.3 Antécédents médicaux et conditions préexistantes

- **Diabetes** : Indicateur de présence du diabète (0 : Non, 1 : Oui).

- **Family.History** : Historique familial de maladies cardiaques (0 : Non, 1 : Oui).
- **Previous.Heart.Problems** : Antécédents de maladies cardiaques.
- **Medication.Use** : Utilisation de médicaments liés à la santé cardiaque.

7.2.4 Mode de vie et habitudes

- **Smoking** : Consommation de tabac (0 : Non, 1 : Oui).
- **Obesity** : Indicateur d'obésité (0 : Non, 1 : Oui).
- **Alcohol.Consumption** : Consommation d'alcool (0 : Non, 1 : Oui).
- **Exercise.Hours.Per.Week** : Nombre d'heures d'exercice par semaine.
- **Physical.Activity.Days.Per.Week** : Nombre de jours d'activité physique par semaine.
- **Sedentary.Hours.Per.Day** : Nombre d'heures passées sans activité physique par jour.

7.2.5 Autres facteurs

- **Stress.Level** : Niveau de stress (échelle de 1 à 10).
- **Sleep.Hours.Per.Day** : Nombre d'heures de sommeil par jour.
- **Income** : Revenu annuel en dollars.

7.2.6 Variable cible

- **Heart.Attack.Risk** : Indicateur de risque de crise cardiaque (0 : Faible, 1 : Élevé).

7.3 Observation Initiale

Les données contiennent une diversité de variables quantitatives et qualitatives permettant d'évaluer l'impact des facteurs de risque sur les maladies cardiaques. Une analyse plus approfondie, notamment via l'**Analyse en Composantes Principales (ACP)**, permettra d'identifier les variables les plus influentes dans la prédiction du risque cardiaque.

8 Préparation de données

8.1 Prétraitement des Données

Avant d'appliquer l'Analyse en Composantes Principales (ACP), un prétraitement des données est nécessaire pour assurer la qualité des résultats. Cette étape comprend la suppression des variables non pertinentes et la transformation des variables qualitatives en variables quantitatives.

8.1.1 Suppression des Variables Inutiles

Certaines variables qualitatives ne sont pas pertinentes pour l'ACP et doivent être supprimées afin d'éviter qu'elles influencent les résultats. Dans notre cas, nous supprimons les colonnes **Country**, **Continent** et **Hemisphere**, car elles n'apportent pas d'information significative pour notre analyse.

En langage R, cette opération est réalisée comme suit :

```
data <- data %>%
  select(-Country, -Continent, -Hemisphere)
```

Ces variables sont qualitatives et ne sont pas directement adaptées à l'Analyse en Composantes Principales (ACP), qui repose sur des mesures de variance et de corrélation entre variables numériques. Voici les principales raisons de leur suppression :

Aucune contribution à la variance expliquée

L'ACP fonctionne en maximisant la variance des données. Les variables qualitatives, comme Country, Continent et Hemisphere, ne varient pas de manière numérique et ne contribuent donc pas à la construction des axes principaux. Incompatibilité avec les calculs de distance

L'ACP repose sur des calculs de distances et de projections qui nécessitent des valeurs continues. Les variables qualitatives sont des catégories qui ne permettent pas ces calculs directement. Solutions alternatives existent

Si l'on souhaite analyser des variables qualitatives avec une approche similaire à l'ACP, on peut utiliser l'Analyse des Correspondances Multiples (ACM), qui est spécialement conçue pour ce type de données.

8.1.2 Transformation des Variables Qualitatives

Les algorithmes statistiques comme l'ACP nécessitent des données numériques. Ainsi, les variables qualitatives doivent être converties en variables numériques.

Encodage de la variable Sex

La variable **Sex** est une variable binaire (Male/Female). Nous appliquons un encodage binaire comme suit :

- **Male** = 0
- **Female** = 1

En langage R, cette transformation est réalisée par :

```
data <- data %>%
  mutate(Sex = ifelse(Sex == "Male", 0, 1))
```

Encodage de la variable Diet

La variable **Diet** est une variable catégorielle avec trois niveaux : *Healthy*, *Average* et *Unhealthy*. Nous appliquons un encodage ordinal pour refléter l'impact potentiel du régime alimentaire :

- **Healthy** = 2
- **Average** = 1
- **Unhealthy** = 0

L'encodage en langage R est le suivant :

```
data <- data %>%
  mutate(Diet = case_when(
    Diet == "Healthy" ~ 2,
    Diet == "Average" ~ 1,
    Diet == "Unhealthy" ~ 0,
    TRUE ~ NA_real_  # Gérer les valeurs manquantes ou inconnues
  ))
```

Ces transformations permettent d'obtenir une base de données entièrement numérique et exploitable par l'ACP.

9 Méthodologie

9.1 Introduction à l'Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales (ACP) est une méthode statistique qui permet de transformer un ensemble de variables corrélées en un nouvel ensemble de variables non corrélées, appelées **composantes principales**. Ces nouvelles variables sont obtenues sous forme de combinaisons linéaires des variables originales.

L'objectif principal de l'ACP est de **réduire la complexité** des données tout en conservant un maximum d'information. Cette technique est largement utilisée en *apprentissage automatique*, en *analyse exploratoire des données* et en *traitement du signal*.

9.2 Objectifs de l'ACP

L'ACP est principalement utilisée pour :

- **Réduction de dimensionnalité** : Elle permet de diminuer le nombre de variables tout en préservant autant que possible l'information contenue dans la base de données.
- **Exploration des données** : L'ACP facilite la visualisation et l'interprétation de jeux de données complexes en les projetant dans un espace de plus faible dimension.
- **Amélioration des performances des modèles** : En réduisant la dimensionnalité, l'ACP peut limiter les problèmes de *bruit* et de *colinéarité*, ce qui améliore l'efficacité des algorithmes d'apprentissage automatique.

9.3 Principe de l'ACP

L'ACP repose sur deux concepts clés :

9.3.1 Composantes principales

Les composantes principales sont de nouvelles variables obtenues à partir des variables initiales. Elles sont ordonnées de manière à ce que :

- La **première composante principale** (CP1) capture la plus grande variance des données.
- La **deuxième composante principale** (CP2) capture la deuxième plus grande variance, sous contrainte d'être orthogonale à CP1.
- Les autres composantes suivent le même principe.

Mathématiquement, chaque composante principale Z_k est définie par une combinaison linéaire des variables originales $X_1, X_2, ..., X_p$:

$$Z_k = a_{k1}X_1 + a_{k2}X_2 + ... + a_{kp}X_p$$

où a_{ki} sont les coefficients déterminés de manière à maximiser la variance de Z_k.

9.3.2 Autovecteurs et Autovalues

L'ACP repose sur le calcul des **autovecteurs** et des **autovalues** de la matrice de covariance ou de la matrice de corrélation des données :

- Les **autovecteurs** définissent les directions des nouvelles composantes principales.
- Les **autovalues** indiquent l'importance relative de chaque composante en termes de variance expliquée.

Soit Σ la matrice de covariance des données, les autovaleurs λ_k et les autovecteurs v_k sont obtenus en résolvant :

$$\Sigma v_k = \lambda_k v_k$$

où λ_k représente la variance expliquée par la k-ième composante principale.

L'ordre des composantes principales est donné par le classement des autovalues en ordre décroissant.

9.4 Conclusion

L'ACP est un outil puissant pour analyser des jeux de données multidimensionnels en réduisant leur complexité tout en conservant l'essentiel de l'information. Elle est largement utilisée dans de nombreux domaines, notamment en science des données et en apprentissage automatique.

10 Standardisation des données

10.1 Pourquoi standardiser les données ?

Lors de l'analyse de données multivariées, notamment avec des techniques comme l'Analyse en Composantes Principales (ACP), il est important de standardiser les variables pour que toutes aient une influence égale sur l'analyse. Sans standardisation, les variables ayant une grande échelle de valeurs (par exemple, des chiffres élevés en cholestérol ou en pression artérielle) risquent de dominer les résultats de l'ACP.

Les données qui ne sont pas standardisées peuvent mener à une mauvaise interprétation, car l'ACP va donner plus de poids aux variables dont les valeurs sont plus grandes en termes d'échelle. Cela peut fausser les résultats et conduire à des conclusions erronées.

10.2 Méthodes de Standardisation

La méthode la plus courante de standardisation est la transformation des variables en scores z, aussi appelés scores standardisés. Cela consiste à soustraire la moyenne de chaque variable à chaque observation, puis à diviser par l'écart-type de cette même variable. La formule générale est la suivante :

$$Z = \frac{X - \mu}{\sigma}$$

Où :

- X est la valeur d'origine de la variable,
- μ est la moyenne de la variable,
- σ est l'écart-type de la variable,
- Z est la valeur standardisée.

Ainsi, après standardisation, chaque variable aura une moyenne de 0 et un écart-type de 1. Cela permet à chaque variable d'être sur la même échelle et évite que des différences de plage de valeurs n'affectent l'analyse.

10.3 Application de la Standardisation sur le Dataset

Dans le cadre de notre analyse des risques cardiaques, la standardisation des variables est une étape préliminaire essentielle. Nous allons appliquer cette méthode à toutes les variables quantitatives du dataset. Cela inclut les variables telles que l'âge, le cholestérol, la pression artérielle, et d'autres mesures biométriques.

Utilisation de la fonction scale() pour standardiser les données. `scale()` :

```
> scaled_data <- scale(data_numeric)
> scaled_data <- as.data.frame(scaled_data)
> head(scaled_data)
         Age        Sex Cholesterol  Heart.Rate   Diabetes Family.History
1  0.62552144 -0.6587276  -0.6415423 -0.14703371 -1.3695727     -0.9860046
2 -1.53923448 -0.6587276   1.5968038  1.11811475  0.7300714      1.0140783
3 -1.53923448  1.5179050   0.7929779 -0.14703371  0.7300714     -0.9860046
4  1.42553993 -0.6587276   1.5226045 -0.09837415  0.7300714      1.0140783
5  0.57846153 -0.6587276   0.7187786  0.87481697  0.7300714      1.0140783
6  0.01374259  1.5179050   0.4590809 -1.31486305  0.7300714      1.0140783
     Smoking    Obesity Alcohol.Consumption Exercise.Hours.Per.Week
1  0.3391375 -1.0027998          -1.2197972             -1.01078010
2  0.3391375  0.9970943           0.8197149             -1.41794656
3 -2.9483202 -1.0027998          -1.2197972             -1.37210925
4  0.3391375 -1.0027998           0.8197149             -0.03218573
5  0.3391375  0.9970943          -1.2197972             -0.72789939
6  0.3391375 -1.0027998           0.8197149             -1.62339018
         Diet Previous.Heart.Problems Medication.Use Stress.Level
1 -0.009636138              -0.9916473     -0.9966392    1.2345331
2 -1.233425687               1.0083079     -0.9966392   -1.5630396
3  1.214153411               1.0083079      1.0032576    1.2345331
4 -0.009636138               1.0083079     -0.9966392    1.2345331
5 -1.233425687               1.0083079     -0.9966392    0.1854433
6 -1.233425687               1.0083079      1.0032576   -1.2133430
  Sedentary.Hours.Per.Day      Income        BMI Triglycerides
1               0.1792403  1.28005677  0.3734324    -0.5885057
2              -0.2972085  1.58243272 -0.2684640    -0.8164405
3               1.0009741  0.95586268 -0.1131278     0.7567569
4               0.4775301 -0.40487874  1.1984556    -0.1773291
5              -1.2920961  0.02844322 -1.1207625    -0.8343178
6               0.5207372  1.03103471 -1.3838195     1.6863736
  Physical.Activity.Days.Per.Week Sleep.Hours.Per.Day Heart.Attack.Risk
1                      -1.5287562          -0.5147206        -0.7470475
2                      -1.0906761          -0.0118221        -0.7470475
3                       0.2235644          -1.5205177        -0.7470475
4                      -0.2145158          -1.5205177        -0.7470475
5                      -1.0906761          -1.0176191        -0.7470475
6                       0.6616445           1.4968734         1.3384502
>
```

10.4 Réduction de Dimensionnalité par l'Analyse en Composantes Principales (ACP)

10.4.1 Introduction

L'Analyse en Composantes Principales (ACP) est une méthode statistique qui permet de transformer un jeu de variables corrélées en un ensemble de variables non corrélées appelées **composantes principales**. L'objectif principal est de réduire la dimensionnalité tout en conservant un maximum d'information. Cette étape est cruciale dans notre étude, car elle permet d'éliminer les redondances dans les données tout en simplifiant leur interprétation.

10.4.2 Méthodologie Appliquée

Après la standardisation des données, nous avons appliqué l'ACP à l'aide du langage R pour identifier les composantes principales les plus pertinentes.

Étapes suivies :

- **Chargement des bibliothèques** : Les bibliothèques suivantes ont été utilisées pour effectuer l'ACP et interpréter les résultats :
  - `FactoMineR` : Utilisée pour réaliser l'Analyse en Composantes Principales (ACP) en fournissant des fonctions dédiées, comme `PCA()`, qui permettent de calculer les composantes principales, les variances expliquées et les coordonnées des individus et des variables.
  - `factoextra` : Complète `FactoMineR` en facilitant la visualisation des résultats de l'ACP grâce à des graphiques interprétables, tels que les nuages de points des individus, les contributions des variables et les cercles de corrélation.
  - `tidyverse` : Fournit un ensemble d'outils pour la manipulation et le nettoyage des données avant l'ACP. Par exemple, `dplyr` permet de sélectionner, filtrer ou transformer les données, et `ggplot2` peut être utilisé pour des visualisations personnalisées.
- **Réaliser l'ACP** : Cette étape consiste à appliquer l'Analyse en Composantes Principales (ACP) en utilisant la fonction `PCA()` de la bibliothèque `FactoMineR`. La commande suivante est utilisée :

```
res.pca <- PCA(data_scaled, graph = FALSE)
```

  * `data_scaled` : Ce jeu de données contient les variables quantitatives standardisées, une étape préalable essentielle pour éviter que les différences d'échelle entre les variables influencent les résultats.
  * `graph = FALSE` : Permet de désactiver l'affichage des graphiques automatiques générés par la fonction, pour privilégier une visualisation personnalisée avec des outils comme `factoextra`.
  * **Résultat** : L'objet `res.pca` contient toutes les informations nécessaires à l'analyse, telles que les variances expliquées par chaque composante, les coordonnées des individus et des variables dans l'espace factoriel, ainsi que leurs contributions respectives.

– **Résumé des résultats** : La commande `summary(res.pca)` permet d'obtenir une vue synthétique des résultats de l'ACP. Ce résumé comprend :

  * **Variances expliquées** : Les pourcentages de variance capturés par chaque composante principale, ce qui aide à déterminer combien de dimensions conserver pour une interprétation significative.
  * **Coordonnées des variables** : Les projections des variables sur les axes principaux, permettant d'identifier leur influence sur les composantes.
  * **Coordonnées des individus** : Les positions des observations dans l'espace factoriel, utiles pour repérer des regroupements ou des similarités entre individus.

  Cette étape est cruciale pour interpréter les résultats avant de procéder à des analyses plus approfondies ou des visualisations.

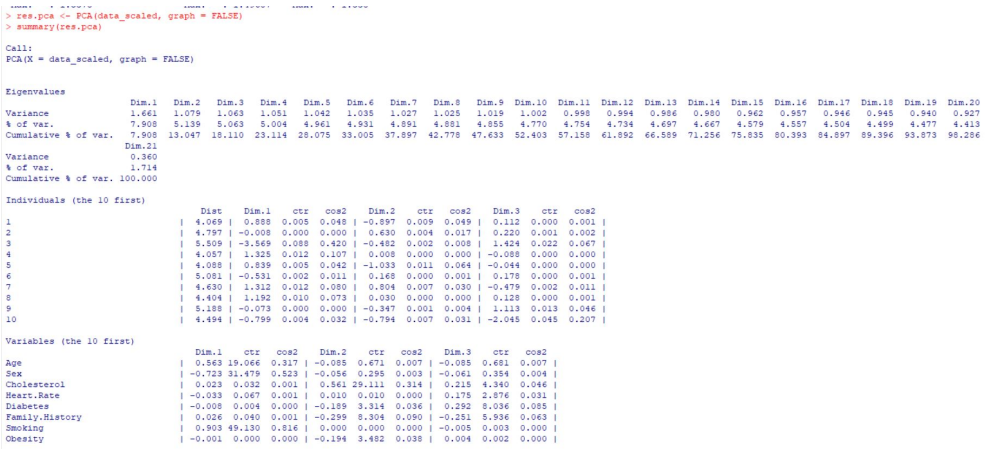
10.4.3 Application de l'ACP sur R

L'Analyse en Composantes Principales (ACP) a été réalisée à l'aide du langage R avec la fonction PCA() du package FactoMineR. Cette méthode permet de réduire la dimensionnalité du dataset tout en conservant l'essentiel de l'information.

Afin de déterminer le nombre optimal de composantes principales à retenir, nous avons appliqué le **C**ritère **d**e **K**aiser, qui recommande de ne conserver que

les composantes dont la valeur propre est supérieure à 1. La figure ci-dessous présente la distribution des valeurs propres des composantes principales.

Les résultats montrent que seules les X premières composantes vérifient cette condition, ce qui indique que ces dimensions capturent l'essentiel de la variance présente dans les données. Cette approche permet ainsi d'améliorer l'interprétabilité du modèle tout en réduisant le nombre de variables.



```
> res.pca <- PCA(data_scaled, graph = FALSE)
> summary(res.pca)

Call:
PCA(X = data_scaled, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14  Dim.15  Dim.16  Dim.17  Dim.18  Dim.19  Dim.20
Variance               1.661   1.079   1.063   1.051   1.042   1.035   1.027   1.025   1.019   1.002   0.998   0.994   0.986   0.980   0.962   0.957   0.946   0.945   0.940   0.927
% of var.              7.908   5.139   5.063   5.004   4.961   4.931   4.891   4.881   4.855   4.770   4.754   4.734   4.697   4.667   4.579   4.557   4.504   4.499   4.477   4.413
Cumulative % of var.   7.908  13.047  18.110  23.114  28.075  33.005  37.897  42.778  47.633  52.403  57.158  61.892  66.589  71.256  75.835  80.393  84.897  89.396  93.873  98.286
                       Dim.21
Variance                0.360
% of var.               1.714
Cumulative % of var. 100.000

Individuals (the 10 first)
                          Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
1                      |  4.069 |  0.888  0.005  0.048 | -0.897  0.009  0.049 |  0.112  0.000  0.001 |
2                      |  4.797 | -0.008  0.000  0.000 |  0.630  0.004  0.017 |  0.220  0.001  0.002 |
3                      |  5.509 | -3.569  0.088  0.420 | -0.482  0.002  0.008 |  1.424  0.022  0.067 |
4                      |  4.057 |  1.325  0.012  0.107 |  0.008  0.000  0.000 | -0.088  0.000  0.000 |
5                      |  4.088 |  0.839  0.005  0.042 | -1.033  0.011  0.064 | -0.044  0.000  0.000 |
6                      |  5.081 | -0.531  0.002  0.011 |  0.168  0.000  0.001 |  0.178  0.000  0.001 |
7                      |  4.630 |  1.312  0.012  0.080 |  0.804  0.007  0.030 | -0.479  0.002  0.011 |
8                      |  4.404 |  1.192  0.010  0.073 |  0.030  0.000  0.000 |  0.128  0.000  0.001 |
9                      |  5.188 | -0.073  0.000  0.000 | -0.347  0.001  0.004 |  1.113  0.013  0.046 |
10                     |  4.494 | -0.799  0.004  0.032 | -0.794  0.007  0.031 | -2.045  0.045  0.207 |

Variables (the 10 first)
                          Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
Age                    |  0.563 19.066  0.317 | -0.085  0.671  0.007 | -0.085  0.681  0.007 |
Sex                    | -0.723 31.479  0.523 | -0.056  0.295  0.003 | -0.061  0.354  0.004 |
Cholesterol            |  0.023  0.032  0.001 |  0.561 29.111  0.314 |  0.215  4.340  0.046 |
Heart.Rate             | -0.033  0.067  0.001 |  0.010  0.010  0.000 |  0.175  2.876  0.031 |
Diabetes               | -0.008  0.004  0.000 | -0.189  3.314  0.036 |  0.292  8.036  0.085 |
Family.History         |  0.026  0.040  0.001 | -0.299  8.304  0.090 | -0.251  5.936  0.063 |
Smoking                |  0.903 49.130  0.816 |  0.000  0.000  0.000 | -0.005  0.003  0.000 |
Obesity                | -0.001  0.000  0.000 | -0.194  3.482  0.038 |  0.004  0.002  0.000 |
```

Figure 10.1: Résultats de la réduction de dimensionnalité par ACP.

10.5 Résultats de l'Analyse en Composantes Principales (ACP)

Ce résultat provient d'une Analyse en Composantes Principales (ACP) réalisée avec la fonction `PCA()` de `FactoMineR` dans R. Voici une description détaillée des différentes parties affichées :

10.5.1 Valeurs propres (Eigenvalues)

Cette section montre la variance expliquée par chaque dimension principale (Dim.1, Dim.2, ..., Dim.21).

- La ligne **Variance** donne la valeur absolue de la variance pour chaque dimension.
- La ligne **% of var.** indique le pourcentage de variance expliquée par chaque dimension.
- La ligne **Cumulative % of var.** montre le cumul de la variance expliquée par les dimensions successives.

Interprétation

- La première dimension (Dim.1) explique 7.9% de la variance totale.
- La deuxième dimension (Dim.2) explique 5.1% supplémentaires, ce qui donne un cumul de 13.0%.
- Le graphique **Scree Plot** associé permet d'identifier un "coude", indiquant combien de dimensions conserver.

10.5.2 Analyse du Scree Plot

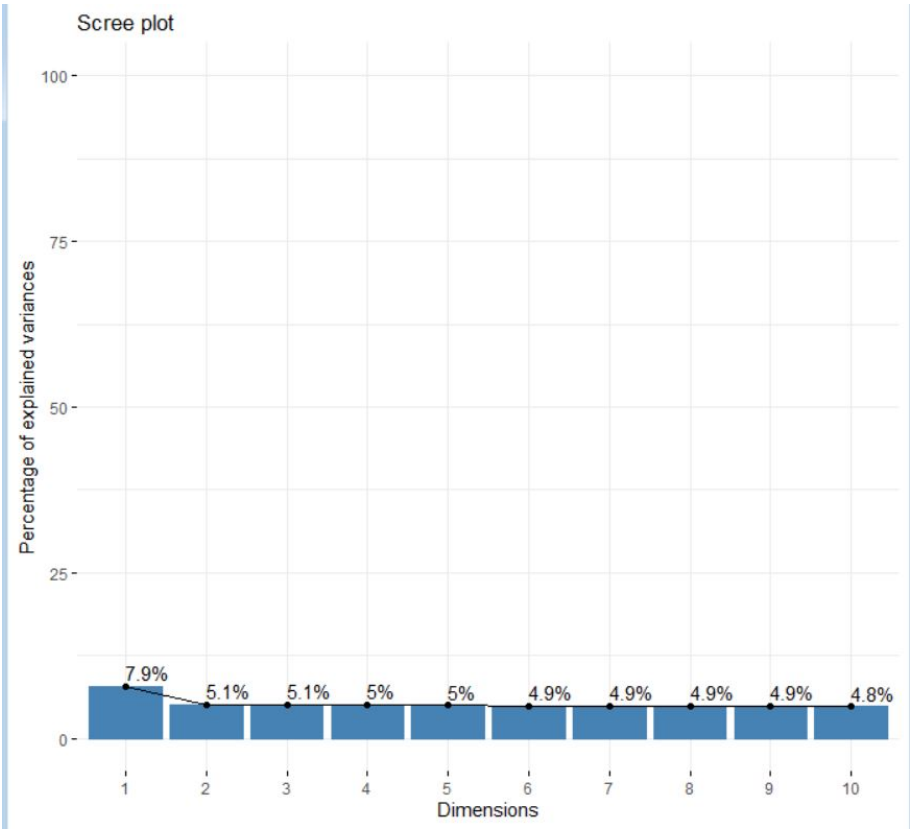


Figure 10.2: Scree plot montrant le pourcentage de variance expliquée par chaque dimension.

Le Scree Plot (voir Figure 10.2) représente le pourcentage de variance expliquée par chaque dimension principale. Voici une analyse détaillée de cette courbe :

- **Coude (Elbow)** : Le point où la courbe commence à se stabiliser (appelé "coude") est observé autour de la $2^{\text{ème}}$ ou $3^{\text{ème}}$ dimension. Cela suggère que les dimensions au-delà de ce point expliquent une proportion relativement faible de la variance totale.

– **Dimensions significatives** : Les premières dimensions (Dim.1 et Dim.2) expliquent une part importante de la variance, avec Dim.1 expliquant 7.9% et Dim.2 expliquant 5.1%. Ces dimensions sont donc les plus importantes pour l'analyse.

– **Dimensions supplémentaires** : Les dimensions suivantes (Dim.3 à Dim.10) expliquent entre 5.1% et 4.8% de la variance, ce qui montre une contribution relativement faible et homogène, sans réelle rupture marquée après la $2^{\text{ème}}$ ou $3^{\text{ème}}$ dimension.

10.6 Analyse du Graphique de Corrélation des Variables (ACP)

Ce graphique est un cercle de corrélation obtenu à partir de l'Analyse en Composantes Principales (ACP). Il permet d'analyser la contribution des variables initiales aux deux premières composantes principales Dim.1 et Dim.2.

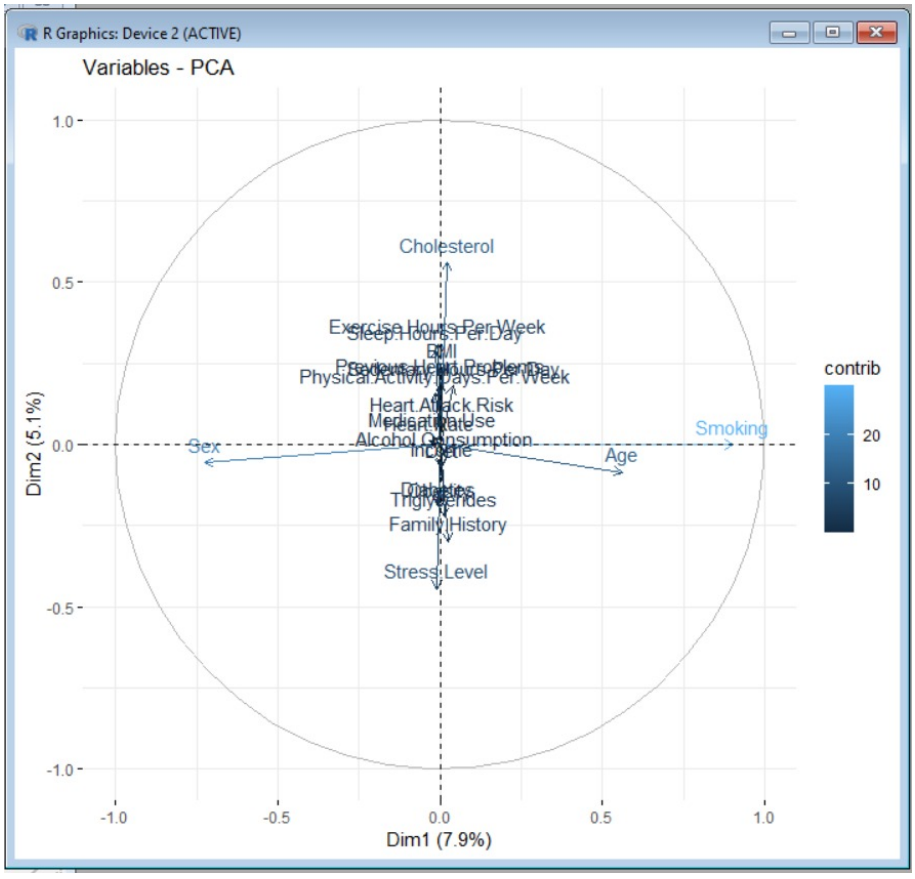


Figure 10.3: Cercle de corrélation des variables dans l'ACP.

10.6.1 Lecture et Interprétation du Graphique

Axes des Composantes Principales

- **Dim.1 (7.9%)** : Cette première composante explique 7.9% de la variance totale des données. Elle est principalement influencée par l'âge (`Age`) et le tabagisme (`Smoking`), qui sont bien représentés sur cet axe.
- **Dim.2 (5.1%)** : Cette seconde composante explique 5.1% de la variance et est fortement corrélée avec le niveau de stress (`Stress Level`), les antécédents familiaux (`Family History`) et les triglycérides (`Triglycerides`).

10.6.2 Explication des Éléments du Graphique

Position et Direction des Flèches (Variables)

- Les variables situées loin du centre sont fortement corrélées avec au moins une des deux dimensions et influencent fortement l'ACP.
- Les flèches pointant dans la même direction indiquent des variables corrélées positivement.
- Les flèches en directions opposées révèlent des corrélations négatives entre les variables.

Intensité de la Couleur (Échelle "contrib")

- Les couleurs varient en fonction de la contribution des variables à la formation des composantes principales.
- Plus une variable est foncée (bleu foncé), plus elle contribue fortement à la variance de l'axe.

Analyse des Variables Clés

- **Dim.1 (Axe horizontal)** :
  - Les variables les plus influentes sont `Smoking` (Tabagisme) et `Age`, qui sont bien projetées sur cet axe.
  - Cela signifie que l'âge et le tabagisme sont des facteurs importants dans la distinction des données selon Dim.1.

- **Dim.2 (Axe vertical)** :
  - Des variables comme `Stress Level`, `Triglycerides` et `Family History` sont les plus influentes.
  - Cela indique que le stress et les antécédents familiaux sont déterminants pour différencier les individus selon Dim.2.

Variables Centrales (Faible Corrélation)

- Certaines variables sont proches du centre du cercle, ce qui signifie qu'elles ont peu d'influence sur Dim.1 et Dim.2.
- Exemple : `Heart Attack Risk`, `Income`, `Medication Usage`, ce qui peut indiquer que ces facteurs sont moins discriminants selon ces deux dimensions.

10.7 Applications de l'ACP

10.7.1 Partie Théorique : Pourquoi Créer une Nouvelle Dataset ?

Après avoir réalisé une Analyse en Composantes Principales (ACP), il est souvent utile de créer une nouvelle dataset contenant les scores des composantes principales. Cette nouvelle dataset présente plusieurs avantages :

- **Réduction de la dimensionnalité** : Les composantes principales capturent l'essentiel de la variance des données originales, tout en réduisant le nombre de variables. Cela simplifie les analyses ultérieures.
- **Élimination de la multicollinéarité** : Les composantes principales sont orthogonales (non corrélées), ce qui élimine les problèmes de multicollinéarité dans les modèles statistiques.
- **Visualisation simplifiée** : Avec seulement deux ou trois composantes principales, il est plus facile de visualiser les données dans un espace réduit (par exemple, un graphique 2D ou 3D).

10.7.2 Partie Pratique : Création de la Nouvelle Dataset

Pour créer cette nouvelle dataset, nous utilisons les scores des deux premières composantes principales (Dim.1 et Dim.2), qui expliquent la plus grande partie

de la variance des données. Ces scores sont ensuite combinés avec les données transformées (par exemple, normalisées) pour former une nouvelle dataset.

Voici le code R utilisé pour réaliser cette étape :

[caption=Code R pour créer une nouvelle dataset avec les dimensions de l'ACP., label=code:r] Charger les bibliothèques nécessaires library(FactoMineR) library(factoextra)

Exemple de données (remplacez par vos propres données) data ¡- read.csv("votre$_fichier.csv$")

Transformation des données (si nécessaire) data$_transformed <- -scale(data)Normalisationdesdonn$

Réaliser l'ACP res.pca ¡- PCA(data$_transformed, scale.unit = FALSE, ncp = 5, graph = FALSE$)

Extraire les deux premières composantes principales data$_reduced <- -res.pca$ind$coord[, 1 : 2]Garderles2premièrescomposantes$

Combiner avec les données transformées data$_final <- -cbind(data_transformed, data_reduced)$

Afficher les premières lignes de la nouvelle dataset head(data$_final$)

Sauvegarder la nouvelle dataset dans un fichier CSV write.csv(data$_final, "data_final.csv", row.n$ $FALSE$)

Résultats de la Réduction de Dimensionnalité

Après l'application de l'Analyse en Composantes Principales (ACP), les deux premières composantes principales (Dim.1 et Dim.2) ont été extraites, capturant **13% de la variance totale**.

Bien que ce pourcentage soit relativement faible, ces dimensions permettent de structurer les données en mettant en évidence les principales tendances. Elles ont été intégrées au dataset d'origine afin de faciliter l'interprétation des résultats et d'améliorer les analyses ultérieures.

Description des Nouvelles Dimensions

Les nouvelles composantes principales obtenues, Dim.1 et Dim.2, représentent des combinaisons linéaires des variables initiales. Voici leurs caractéristiques principales :

- **Dim.1** Cette dimension semble être fortement influencée par les variables `Smoking` et `Age`, ce qui pourrait refléter un facteur lié au mode de vie et

aux risques de santé associés au tabagisme et au vieillissement. Elle pourrait aussi capturer une opposition entre certaines habitudes de vie (comme la consommation d'alcool) et des facteurs de santé.

– **Dim.2** :Cette dimension est davantage corrélée avec des variables comme "Stress level" et "family history". Elle pourrait ainsi représenter une distinction entre des facteurs biologiques (sexe, cholestérol) et certains comportements liés à la santé, comme l'activité physique et le stress.

Extrait du Dataset Réduit

Un extrait des données finales après la réduction dimensionnelle est présenté dans le tableau suivant :

ID	Dim.1	Dim.2
1	2.31	-1.12
2	-0.87	0.95
3	1.56	-0.42

Table 10.1: Extrait du dataset après réduction dimensionnelle

10.7.3 Applications de la Nouvelle Dataset

La nouvelle dataset (`data_final.csv`) contient :

– Les données transformées (par exemple, normalisées) dans leurs colonnes d'origine.

– Deux nouvelles colonnes correspondant aux scores des deux premières composantes principales (`Dim.1` et `Dim.2`).

Cette dataset peut maintenant être utilisée pour des analyses ultérieures, telles que :

– **Visualisation** : Les individus peuvent être visualisés dans un espace 2D en utilisant `Dim.1` et `Dim.2` comme axes.

– **Classification** : Les composantes principales peuvent servir de variables explicatives pour des algorithmes de classification (par exemple, k-means ou SVM).

– **Régression** : Les composantes principales peuvent être utilisées comme prédicteurs dans des modèles de régression.

10.7.4 Pourquoi les Premières Lignes Diffèrent-elles ?

Lors de la création de la nouvelle dataset, il est possible que les premières lignes affichées avec `head(data_final)` ne correspondent pas exactement aux premières lignes du dataset original. Cela peut être dû à plusieurs raisons :

- **Transformation des données** : Si les données ont été normalisées ou centrées-réduites avant l'ACP, les valeurs affichées dans `data_final` peuvent différer de celles du dataset original.
- **Ajout des composantes principales** : Les colonnes `Dim.1` et `Dim.2` sont ajoutées à la fin du dataset original, ce qui peut donner l'impression que les premières lignes ont changé.
- **Affichage partiel** : La fonction `head()` n'affiche que les premières lignes du dataset, ce qui peut ne pas refléter l'ensemble des données.

10.7.5 Interprétation

La nouvelle dataset (`data_final`) contient toutes les colonnes du dataset original, ainsi que les deux premières composantes principales (`Dim.1` et `Dim.2`). Les différences observées dans les premières lignes sont dues aux transformations appliquées avant l'ACP et à l'ajout des nouvelles colonnes.

10.7.6 Conclusion Finale

En conclusion, ce travail a démontré l'efficacité de l'ACP pour explorer et structurer des données complexes, tout en ouvrant des perspectives pour des analyses futures. Les résultats obtenus fournissent une base solide pour des recherches ultérieures et pourraient contribuer à une meilleure compréhension des facteurs de risque cardiovasculaires. Les hypothèses formulées dans cette étude pourront guider des travaux futurs visant à valider ces prédictions et à améliorer les stratégies de prévention en santé publique.

11 Ressources

Voici quelques ressources qui peuvent être utiles pour approfondir vos connaissances sur ce sujet :

- **Documentation de l'Analyse en Composantes Principales (ACP)**: https://fr.wikipedia.org/wiki/Analyse$_e n_c omposantes_p rincipales Wikipedia - ACP$**Livre** $https : //www.math.univ - toulouse.fr/ besse/Wikistat/StatisticalMethodsfortheAn$
- **Tuto sur LaTeX pour la création de documents**: https://www.latex-project.org/LaTeX Project