

**REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE**

**MINISTERE DE L’ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE**

**ECOLE NATIONALE SUPÉRIEURE D’INFORMATIQUE**

**Groupe : 2CS - SIQ1 Année : 2024/2025**

| **Module : Analyse et fouille de données** |
| --- |

**Rapport du TP1 :**

**Extraction et Interprétation des Informations Essentielles d'un Grand Tableau de Données**

* **DoctorVisits : 5190 Obs , 12 Var -**

**Réalisé par :**

* FLISSI Loubna.
* LARBAOUI Yasmine Badr El Houda.

**Encadré par :**

Madame BESSAH Naima.

### **Table des Matières**

1. Introduction ………………………………………………….………..…………2
2. Préparation des Données………………………………………………….….…..2
3. Analyse des résultats de l'ACP et justification du choix de trois dimensions .......3
4. Corrélations et contributions des variables ………….…….…….…….……..….4
5. Visualisation et Interprétation ……………………………,……,……,……,…...6
6. Conclusion ……………………………………………….…….……….………..6

# **Introduction :**

Dans ce travail, nous avons choisi d'analyser le tableau de données intitulé **"DoctorVisits"**. Nous avons commencé par analyser la nature et la structure des données pour comprendre les différentes variables, qu'elles soient quantitatives ou qualitatives. Cette première étape nous a permis d'identifier les principales caractéristiques du tableau.

Nous avons ensuite choisi d'appliquer l'Analyse en Composantes Principales (ACP) pour extraire les informations importantes et simplifier les données. Avant cela, certaines transformations étaient nécessaires, comme la conversion ou la suppression de variables non pertinentes. Chaque transformation a été justifiée pour améliorer l'analyse des relations entre les variables.

1. **Préparation des données :**

Avant de commencer l'analyse en composantes principales (ACP), il est important de préparer et nettoyer les données pour garantir des résultats fiables. Voici les étapes réalisées :

| 1. **Exploration des données :** Nous avons d'abord examiné la structure des données pour comprendre les types de variables et détecter d'éventuelles anomalies.   Figure 1.Analyse Préliminaire : Structure et Types des Variables dans le Jeu de Données 'DoctorVisits' |  |
| --- | --- |

1. **Suppression des variables inutiles :** Certaines variables, comme *rownames*, ne sont pas utiles pour l'analyse, nous les avons donc supprimées afin de simplifier les données.
2. **Gestion des valeurs manquantes et infinies :** Nous avons vérifié la présence de valeurs manquantes ou infinies. Aucune n’a été trouvée.
3. **Conversion des variables qualitatives :** Les variables catégorielles, comme *gender*, ont été transformées en variables numériques (0 ou 1) pour pouvoir être utilisées dans l'ACP.
4. **Standardisation des données ( centrer et réduire ) :** Pour que toutes les variables aient le même poids dans l'analyse, nous avons standardisé les données en leur donnant une moyenne de 0 et un écart-type de 1.
5. **Préparation finale :** Après ces étapes, les données sont prêtes pour l'ACP, assurant ainsi des analyses fiables et des résultats précis.

**3. Analyse des résultats de l'ACP et justification du choix de trois dimensions**

|  | Figure 2. Cercle des corrélations des variables dans l’espace des dimensions principales (ACP) |
| --- | --- |

1. **Présentation des pourcentages de variance expliquée par les composantes principales**

### **Présentation des pourcentages de variance :**

| Figure 3. Scree plot de l'ACP sur les données de [DoctorVisits in Australia ] | D'après le **graphique de scree plot** fourni, nous pouvons extraire les pourcentages de variance expliquée par chacune des dix dimensions principales. Ces pourcentages représentent la proportion de la variance totale des données qui est captée par chaque composante principale. |
| --- | --- |

* **Interprétation:**
* Les trois premières dimensions expliquent cumulativement environ 50% de la variance totale. Cela signifie que ces trois dimensions capturent la moitié de l'information contenue dans les données initiales.

1. **Justification du choix de dimensions**
2. **Critère du coude :** Le scree plot révèle une rupture notable entre la 3ᵉ et la 4ᵉ dimension, signalant une diminution marginale de l’inertie expliquée à partir de la 4ᵉ composante. Cela suggère que les trois premières dimensions capturent l’essentiel de la structure des données.
3. **Inertie cumulée :** Les trois premières composantes expliquent 49,8 % de l’inertie totale, ce qui est suffisant pour résumer une part significative de l’information contenue dans les données initiales.
4. **Critère d’interprétation :** Les trois premières dimensions offrent une interprétation claire et cohérente, tandis que les composantes suivantes ajoutent peu d'information interprétable.

**En retenant les trois premières composantes principales, nous optimisons la réduction de la dimensionnalité tout en préservant une part significative de l'inertie totale (49,8 %), garantissant ainsi une interprétation claire des données.**

| Figure 4.Cercle des corrélations avec la contribution des variables aux dimensions principales (ACP) | En plus , l'analyse des contributions des variables à chaque composante basée sur le **graphique des contributions** permet de mieux comprendre l'interprétation de chaque axe factoriel. Par exemple, on peut observer que les variables "visits", "illness" et "health" contribuent fortement à la première composante, ce qui suggère que cette composante est liée à l'état de santé et à l'utilisation des services de santé. |
| --- | --- |

**4. Corrélations et contributions des variables**

#### Variables les plus contributives sur chaque dimension :

Les graphiques fournis (cercle des corrélations et graphiques de contribution) permettent d'identifier les variables contributives pour chaque dimension.

* **Dimension 1 (23.2 % de variance expliquée) :**
  + D'après le cercle des corrélations et la carte des contributions :
* Variables fortement contributives (fortes flèches) :**Visits,Illness,Health**
* Ces variables influencent Dim 1, qui semble être liée à l'**état de santé et l'utilisation des services de santé**.
  + Contribution des variables (graphique des contributions) :
* Les variables comme **"visits"** et **"illness"** ont une contribution proche de 15 %, indiquant qu'elles expliquent une grande part de la variance capturée par Dim 1.
* **Dimension 2 (15.4 % de variance expliquée) :**
  + Variables contributives :**Income,Freepoor**

Ces variables influencent Dim 2, qui semble liée aux **conditions socio-économiques**.

* + Contribution des variables (graphique des contributions) :
* Les variables **"income"** et **"freepoor"** ont une contribution dominante, indiquant leur importance dans la construction de cet axe.

#### Interprétation de l’influence des variables supplémentaires :

* Les variables supplémentaires influencent les axes factoriels de deux manières :
* **Corrélation forte avec une dimension** : Une flèche longue et alignée sur l'axe.
* **Faible influence sur une dimension** : Une flèche courte ou un angle proche de 90° avec l'axe.
* À partir du cercle des corrélations :
  + Les variables supplémentaires influencent les deux premières dimensions comme suit :
    - **Dim 1** :
* Variables avec une corrélation positive forte : **Visits**, **Illness**, **Health**.
* Ces variables contribuent à interpréter Dim 1 comme une mesure de **l’état de santé et de l’utilisation des services de santé**.
  + - **Dim 2** :
* Variables avec une corrélation positive forte : **Income**, **Freepoor**.
* Ces variables renforcent l’interprétation de Dim 2 comme une dimension liée aux **conditions socio-économiques**.

### **En Résumé**

1. Les **contributions des variables** sont mathématiquement dérivées des coordonnées de chaque variable divisées par la valeur propre de la dimension, puis multipliées par 100. Ces données sont représentées sur le graphique de contribution.
2. Les **variables influentes** sont identifiées par des flèches longues et des contributions élevées dans les graphiques.
3. Les **dimensions principales** se résument comme suit :
   * **Dim 1 (23.2 %) :** Variables liées à la santé (**Visits**, **Illness**, **Health**).
   * **Dim 2 (15.4 %) :** Variables socio-économiques (**Income**, **Freepoor**).

#### la Qualité de représentation (cos²) dans l'ACP :

1. **Variables les mieux représentées** :
   * Les variables **freerepat** (**0.294**) et **reduced** (**0.266**) présentent la meilleure qualité de représentation sur les deux premiers axes principaux, indiquant leur forte contribution à la variance capturée.
   * D'autres variables bien représentées incluent **age** (**0.244**), **health** (**0.215**), et **visits** (**0.197**), soulignant leur rôle important dans l'explication des relations globales dans les données.
2. **Variables peu représentées** :
   * La variable **freepoor** a une qualité de représentation très faible (**0.004**), ce qui suggère qu'elle n'apporte pas d'information significative sur les deux premiers axes et pourrait être exclue des analyses ultérieures.
3. **Représentation des individus** :
   * Certains individus sont bien représentés sur les deux premiers axes, comme l'Individu 1 (**0.257**) et l'Individu 3 (**0.225**). Cela reflète leur forte contribution à la structuration des données dans ce plan.
   * À l’inverse, des individus comme l'Individu 4 (**0.072**) ont une représentation faible, montrant qu'ils contribuent moins à l'interprétation des axes principaux

**5. Visualisation et Interprétation :**

Les résultats de l’ACP révèlent des dimensions principales qui permettent une meilleure compréhension des données, mais surtout, des implications pratiques pour les parties concernées, notamment les décideurs dans le domaine de la santé publique, les responsables politiques et les chercheurs en socio-économie.

#### 1. Synthèse des résultats analytiques :

* Dimension 1 (23.2 % de la variance totale) : Capturant les états de santé et l’utilisation des services médicaux, cette dimension est dominée par les variables Visits, Illness, et Health.
  + Interprétation pratique : Cette dimension permet d’identifier les groupes d’individus ayant des besoins accrus en services de santé.
* Dimension 2 (15.4 % de la variance totale) : Liée aux conditions socio-économiques, avec des contributions importantes de Income et Freepoor.
  + Interprétation pratique : Cette dimension met en évidence les disparités socio-économiques influençant l'accès aux soins.

#### 2. Implications pratiques et utilité de l'analyse :

* Amélioration de la politique de santé :  
  Les résultats permettent d’identifier deux principaux leviers d’action : l’état de santé global et les disparités socio-économiques. En adaptant les politiques publiques à ces résultats, il est possible de réduire les inégalités en matière de santé.
* Identification des priorités régionales :  
  Les zones où la Dimension 1 est dominante peuvent être des régions nécessitant des infrastructures médicales supplémentaires. Celles où la Dimension 2 est dominante pourraient bénéficier de programmes socio-économiques ciblés.
* Éducation et sensibilisation :  
  Les groupes socio-économiquement défavorisés, identifiés via la Dimension 2, peuvent être la cible de campagnes d'éducation pour promouvoir des comportements de santé préventifs.
* Planification à long terme :  
  Les informations extraites peuvent être utilisées pour orienter la planification stratégique, comme la répartition des ressources médicales, le développement d'assurances santé accessibles ou encore des politiques de lutte contre la pauvreté.

**6. Conclusion :**

En rendant les résultats exploitables sur le plan pratique, cette analyse ACP contribue non seulement à une meilleure compréhension des données, mais aussi à la formulation de stratégies efficaces pour améliorer la santé et le bien-être des populations. Les décideurs peuvent utiliser ces résultats pour prioriser les interventions et maximiser leur impact en ciblant les principaux facteurs identifiés.