Aix*Marseille université

Rapport du projet Consultance

M2 DS

Traitement des Réponses Incomplètes - Influence Démographique sur l'Attention

---

Author:

Issa Goukouni — issagoukouni96@gmail.com
Fatou Niass — fatou.niass@etu.univ-amu.fr
Adama BA — adama-abdoul.ba@etu.univ-amu.fr

Group ♣

2023/2024 – 2nd Semester

Contents

1 Abstrac

Attention is a crucial cognitive resource that governs our ability to perceive, process, and respond to environmental stimuli. Its importance is undeniable in various aspects of daily life, from succeeding in tasks at work to maintaining safety while driving. Demographic characteristics of individuals can significantly influence attention. Understanding these influences is crucial both for individuals and society. This study aims to systematically examine the relationships between demographic factors and attention to determine the primary determinants of this cognitive faculty. By identifying population groups most likely to be affected by attention problems, this research aims to inform the management and promotion of mental health.

Keys words : **Attention,Demographic characteristics,cognictive faculty.**

2 Introduction

L'attention est une ressource cognitive cruciale qui régit notre capacité à percevoir, à traiter et à répondre aux stimuli de notre environnement. Son importance est indéniable dans de nombreux aspects de la vie quotidienne, qu'il s'agisse de réussir une tâche au travail, de maintenir la sécurité lors de la conduite automobile, ou encore de se concentrer lors d'une conversation. Cependant, l'attention peut être influencée par une multitude de facteurs, parmi lesquels les **caractéristiques démographiques** des individus jouent un rôle significatif.

Comprendre quels **facteurs démographiques** ont le plus d'effets sur **l'attention** revêt donc une importance cruciale tant sur le plan individuel que sociétal. Cette question complexe nécessite une exploration approfondie, car les différences démographiques telles que **l'âge, le sexe, le niveau d'éducation** et d'autres variables peuvent avoir des répercussions variées sur la capacité d'une personne à maintenir son attention.

Dans cette étude, nous nous attacherons à examiner de manière systématique les relations entre les facteurs démographiques et l'attention, dans le but de déterminer quels sont les principaux déterminants de cette faculté cognitive essentielle. En analysant ces interactions, nous chercherons à identifier les groupes de population les plus susceptibles d'être affectés par des problèmes d'attention, ainsi que les implications potentielles de ces résultats pour la prise en charge et la promotion de la santé mentale.

À travers cette exploration, nous visons à apporter des éclaircissements significatifs sur la manière dont les caractéristiques démographiques façonnent notre capacité d'attention, ouvrant ainsi la voie à des interventions ciblées et des stratégies de prévention plus efficaces pour améliorer la santé cognitive et le bien-être des individus.

Nous allons essayer de repondre a notre problematique qui est : **quels facteurs demographiques ont le plus d'effets sur l'attention?**

3 Détermination des scores de questionnaire

Pour répondre à notre questionnement, nous avons d'abord entrepris de calculer les scores liés à l'attention. Voici comment cela a été réalisé :

Description
Nous avons calculé le score A en additionnant les réponses des six premiers items (ATT1 à ATT6).
Le score B a été obtenu en additionnant les réponses des items de 7 à 18 (ATT7 à ATT18).
Le score d'inattention a été calculé en sommant les réponses des items 1, 2, 3, 4, 7, 8, 9, 10 et 11.
De même, le score d'hyperactivité et d'impulsivité a été obtenu en additionnant les réponses des items 5, 6, 12, 13, 14, 15, 16, 17 et 18.
Enfin, le score total lié à l'attention a été calculé en additionnant les réponses des 18 items.

Table 1: Description des scores d'attention

En ce qui concerne les scores de dépression (HAD), nous avons procédé de la même manière :

Description
Le score d'anxiété a été obtenu en sommant les réponses des sept premiers items (HAD1 à HAD7).
Le score de dépression a été calculé en additionnant les réponses des items de 8 à 14 (HAD8 à HAD14).
Le score total du questionnaire HAD a été obtenu en additionnant les réponses de tous les items, de 1 à 14.

Table 2: Description des scores de depression

4 Préparation des données

Dans cette section, nous décrivons les principales étapes de traitement des données que nous avons effectuées pour préparer notre ensemble de données à l'analyse.

4.1 Traitement des données

Pour préparer nos données pour l'analyse, nous avons suivi plusieurs étapes de traitement et de nettoyage. Voici un aperçu des principales étapes que nous avons suivies :

1. **Sélection des colonnes d'attention** : Nous avons identifié les colonnes pertinentes liées à l'attention dans notre ensemble de données. Ces colonnes comprennent différentes

mesures de l'attention, telles que ATT1, ATT2, ..., ainsi que des indicateurs de difficulté d'attention et de concentration.

2. **Renommage des colonnes** : Pour rendre nos données plus compréhensibles et plus faciles à manipuler, nous avons renommé les colonnes en utilisant un dictionnaire de correspondance. Par exemple, les colonnes précédemment nommées sous la forme 'attention1[ATT1]' ont été renommées pour ne conserver que le nom de la mesure (par exemple, 'ATT1').

3. **Sélection des caractéristiques démographiques** : Nous avons également identifié les caractéristiques démographiques pertinentes pour notre analyse. Celles-ci incluent des informations telles que le sexe, l'âge, le niveau d'éducation, la situation professionnelle, le nombre d'enfants, etc.

4. **Création d'un sous-ensemble de données** : En combinant les colonnes d'attention sélectionnées avec les caractéristiques démographiques pertinentes, nous avons créé un nouveau sous-ensemble de données contenant les informations nécessaires à notre analyse.

En effectuant ces étapes de traitement des données, nous avons préparé notre ensemble de données pour une analyse plus approfondie de la relation entre les caractéristiques démographiques et les mesures d'attention chez notre population d'intérêt.

4.2 Traitement des valeurs manquantes

Pour évaluer la qualité de nos données et identifier les lacunes potentielles dans notre ensemble de données, nous avons utilisé un code Python pour visualiser le pourcentage de données manquantes pour chaque variable. Voici comment cela a été réalisé :

4.3 Visualisation

1. **Calcul du pourcentage de données manquantes** : Nous avons calculé le pourcentage de données manquantes pour chaque variable en divisant le nombre de valeurs manquantes par la taille totale de l'ensemble de données, puis en multipliant par 100. Cela nous a donné une mesure du degré de complétude de chaque variable.

2. **Création d'un diagramme en barres** : Pour visualiser ces pourcentages, nous avons utilisé la bibliothèque Matplotlib pour créer un diagramme en barres. Chaque barre représente une variable, et sa hauteur représente le pourcentage de données manquantes pour cette variable.

3. **Personnalisation du graphique** : Nous avons ajouté un titre à notre graphique pour le rendre informatif et compréhensible. De plus, nous avons étiqueté les axes x et y pour indiquer les variables et les pourcentages respectivement. Enfin, nous avons ajusté l'orientation des étiquettes de l'axe x pour une meilleure lisibilité.

4. **Affichage du graphique** : Enfin, nous avons affiché le graphique pour une visualisation immédiate des données manquantes dans notre ensemble de données.

En résumé, cette démarche nous a permis d'identifier rapidement les variables qui comportent des données manquantes et de visualiser leur impact sur notre ensemble de données.

Comme le montre ce graphique :

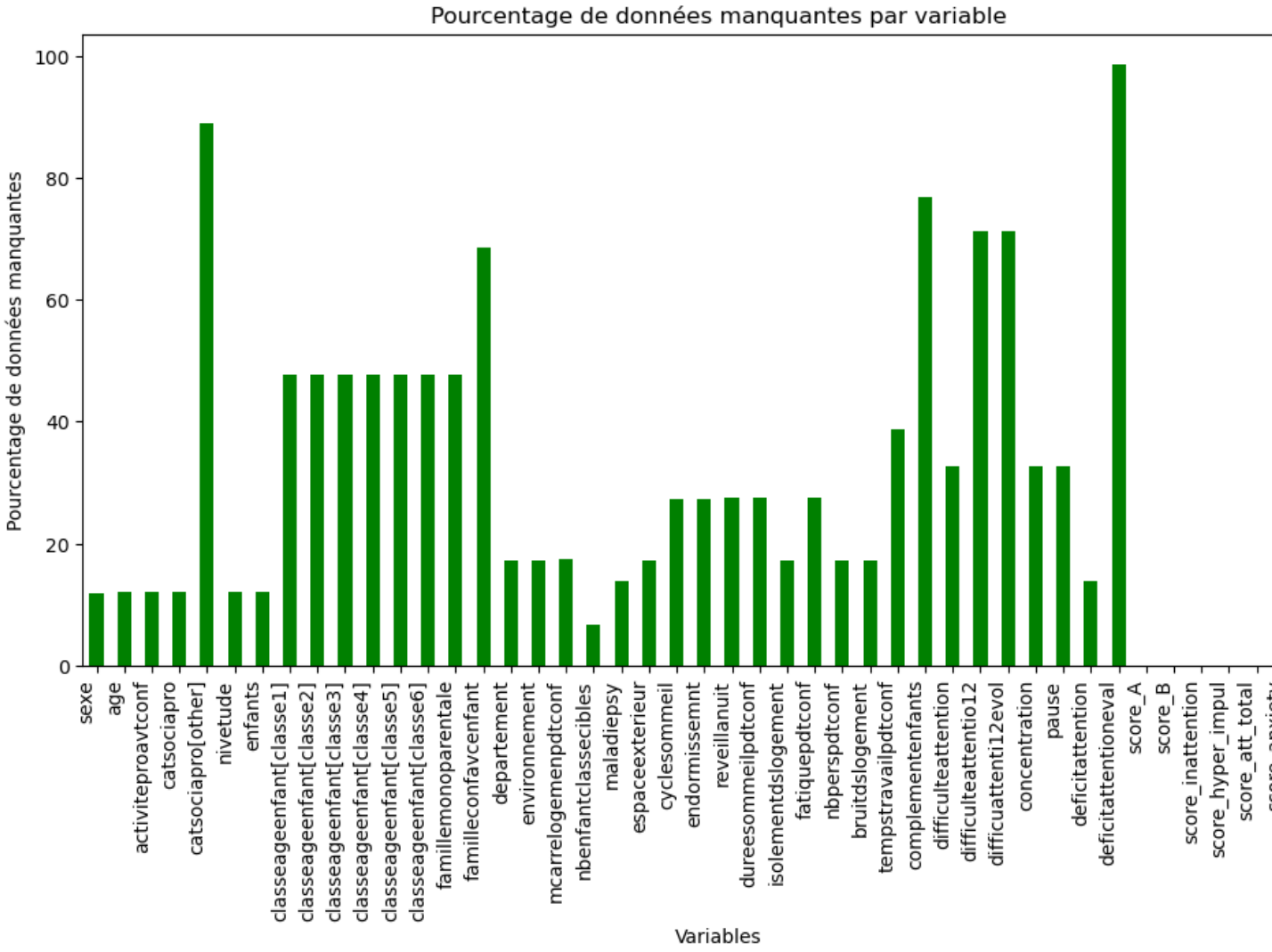

Figure 1: Pourcentage de donnees manquantes

Il est intéressant de noter que les variables *sexe*, *âge*, *activité professionnelle avant confinement* et *catégorie socioprofessionnelle* présentent des pourcentages de données manquantes similaires. En revanche, les variables *catégorie socioprofessionnelle [autre]* et *complementenfants* sont celles qui présentent le plus grand nombre de valeurs manquantes. Nous observons également un pourcentage identique pour le nombre d'enfants

4.4 Anlyse de correlation des données manquantes

Dans cette section de notre analyse, nous avons examiné les données manquantes dans notre ensemble de données à l'aide des bibliothèques Python Pandas, Seaborn et Matplotlib.

Tout d'abord, nous avons créé un DataFrame pour les données manquantes en extrayant un sous-ensemble de notre ensemble de données contenant uniquement les données manquantes des variables d'intérêt.

Ensuite, nous avons calculé la corrélation entre les données manquantes des différentes

variables. Cette analyse de corrélation nous a permis de déterminer s'il existe une tendance ou une relation entre les données manquantes de différentes variables.

Enfin, pour visualiser ces relations, nous avons utilisé une heatmap de corrélation. Cette heatmap nous a fourni une représentation graphique des corrélations entre les données manquantes. Les nuances de couleur dans la heatmap indiquent le degré de corrélation : les valeurs plus proches de 1 indiquent une corrélation positive, tandis que les valeurs plus proches de -1 indiquent une corrélation négative.

L'utilisation de cette heatmap nous a permis de visualiser rapidement les relations entre les données manquantes des différentes variables, ce qui nous a aidés à mieux comprendre la nature des données manquantes dans notre ensemble de données.

formule de correlation

$$\text{corrélation} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \cdot \sum(Y_i - \bar{Y})^2}}$$

où :

X_i : valeurs manquantes de la variable X
Y_i : valeurs manquantes de la variable Y
$\bar{X}$: moyenne des valeurs manquantes de la variable X
$\bar{Y}$: moyenne des valeurs manquantes de la variable Y

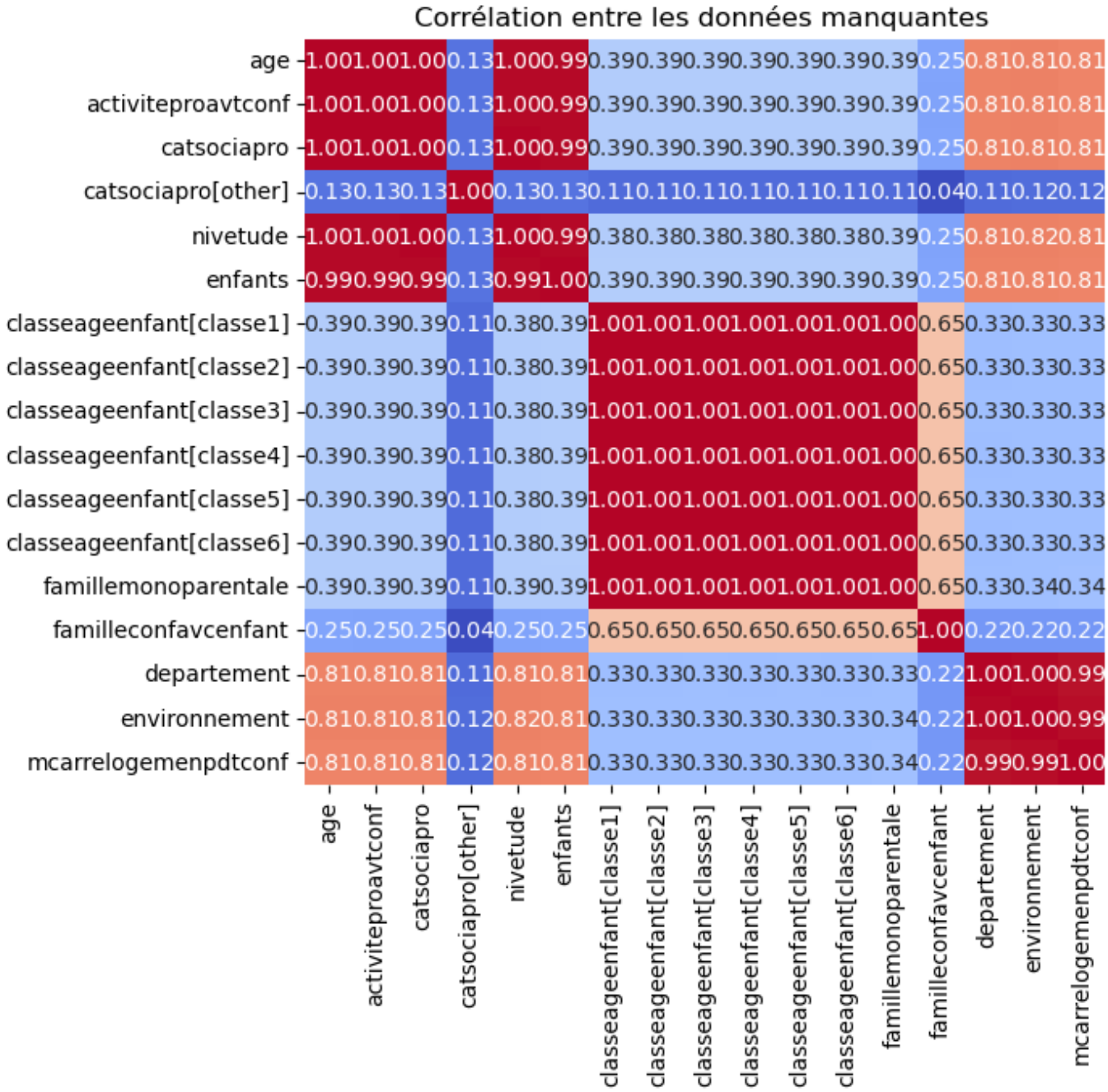

Figure 2: Heatmap

4.5 Gestion des valeurs manquantes dans les données catégorielles

Dans cette étape de prétraitement des données, nous avons abordé la gestion des valeurs manquantes dans les colonnes catégorielles de notre ensemble de données. Voici un résumé des étapes que nous avons suivies :

1. **Parcours des colonnes du DataFrame** : Nous avons utilisé une boucle pour parcourir chaque colonne du DataFrame, vérifiant si elle est de type objet, ce qui indique

généralement des données catégorielles.

2. **Remplacement des valeurs nulles** : Pour chaque colonne catégorielle contenant des valeurs manquantes, nous les avons remplacées par une chaîne de caractères indiquant le nom de la colonne suivi de "_non renseigné". Cela nous permet de conserver l'information sur les données manquantes tout en préservant la nature catégorielle de la variable.

En suivant ces étapes, nous avons pu traiter efficacement les valeurs manquantes dans les données catégorielles, ce qui est essentiel pour préparer nos données avant de les utiliser dans des analyses ultérieures ou des modèles d'apprentissage automatique.

4.6 Gestion des valeurs manquantes dans les autres colonnes

Dans cette étape supplémentaire de prétraitement des données, nous avons abordé la gestion des valeurs manquantes dans les autres colonnes de notre ensemble de données. Voici un aperçu des opérations effectuées :

- **Sélection des colonnes d'intérêt** : Nous avons sélectionné un ensemble de colonnes spécifiques à partir de notre DataFrame encodé, comprenant des variables telles que le sexe, l'âge, l'activité professionnelle, etc.
- **Traitement des valeurs manquantes pour la colonne "cyclesommeil"** : Nous avons rempli les valeurs manquantes dans la colonne "cyclesommeil" en utilisant la méthode de la dernière observation valide (ffill), puis en remplaçant les valeurs restantes par la moyenne des valeurs non manquantes de cette colonne.
- **Traitement des valeurs manquantes pour d'autres colonnes** : Nous avons également appliqué la méthode de la dernière observation valide (ffill) pour remplir les valeurs manquantes dans les colonnes "nbenfantclassecibles", "mcarrelogemenpdtconf" et "nbperspdtconf".

Malgré ces opérations d'imputation, il est important de noter qu'une grande différence dans les distributions des données n'a pas été observée après l'imputation des valeurs manquantes. Cela peut indiquer que l'imputation n'a pas introduit de biais significatif dans nos données.

En suivant ces étapes supplémentaires, nous avons pu gérer efficacement les valeurs manquantes dans les autres colonnes de notre ensemble de données, ce qui est essentiel pour assurer la qualité et la fiabilité de nos données pour les analyses ultérieures.

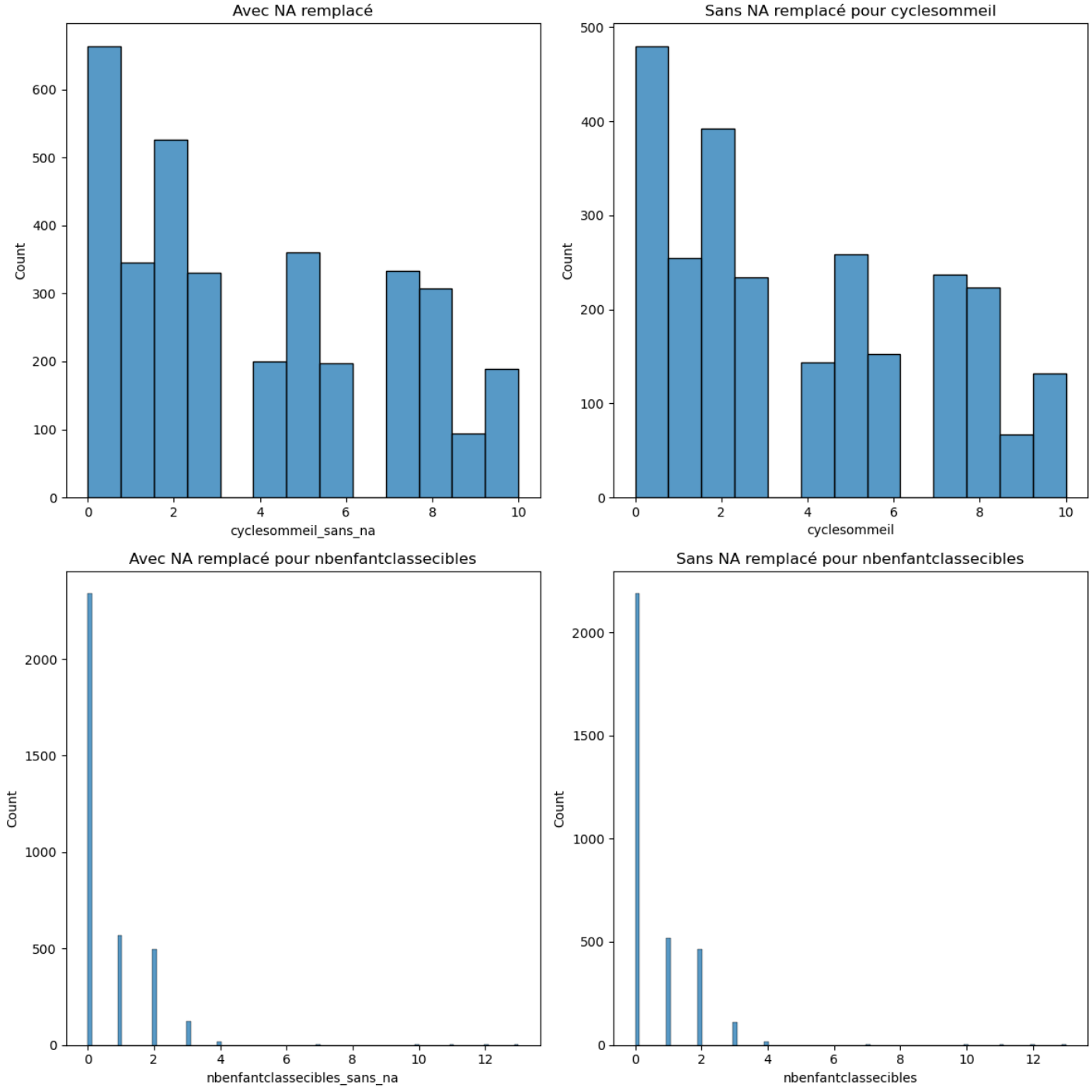

Figure 3: diagramme

5 Analyse exploratoire

5.1 Analyse univariée

5.2 Diagrammes circulaires de quelques variables

Dans cette partie, nous avons sélectionné quelques variables pour effectuer une analyse statistique classique afin d'observer la répartition des caractéristiques.

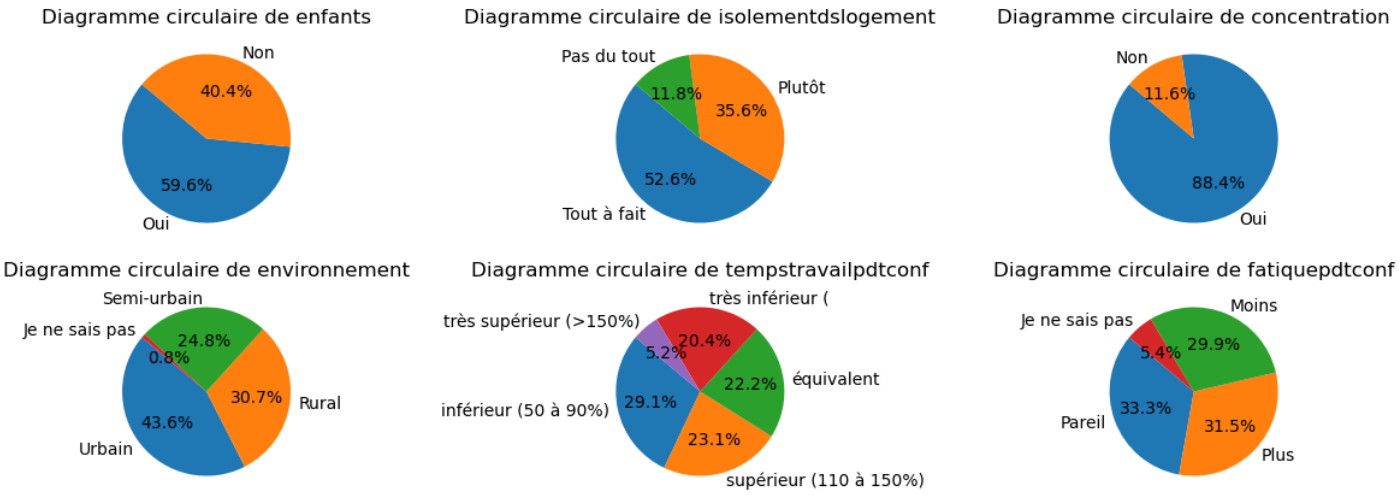

Figure 4: diagramme

Nous constatons que dans l'étude, 59,6% des participants ont des enfants, tandis que 40,4% n'en ont pas. En ce qui concerne le diagramme relatif à la variable 'isolementdsloement', nous remarquons que 52,6% des individus ont été en isolement, contre 11,8% qui ne l'ont pas du tout été, et 35,6% qui ont été en isolement à un certain moment.

5.3 Histogrammes des Variables Catégorielles

Dans cette section, nous avons exploré la répartition des données pour un ensemble de variables catégorielles sélectionnées. L'objectif était de visualiser graphiquement la distribution des valeurs pour chaque variable afin de mieux comprendre la répartition des données et d'identifier d'éventuels schémas ou tendances. Nous avons utilisé des histogrammes pour représenter cette répartition, ce qui nous a permis d'obtenir rapidement des informations sur la fréquence de chaque catégorie ou classe dans les données.

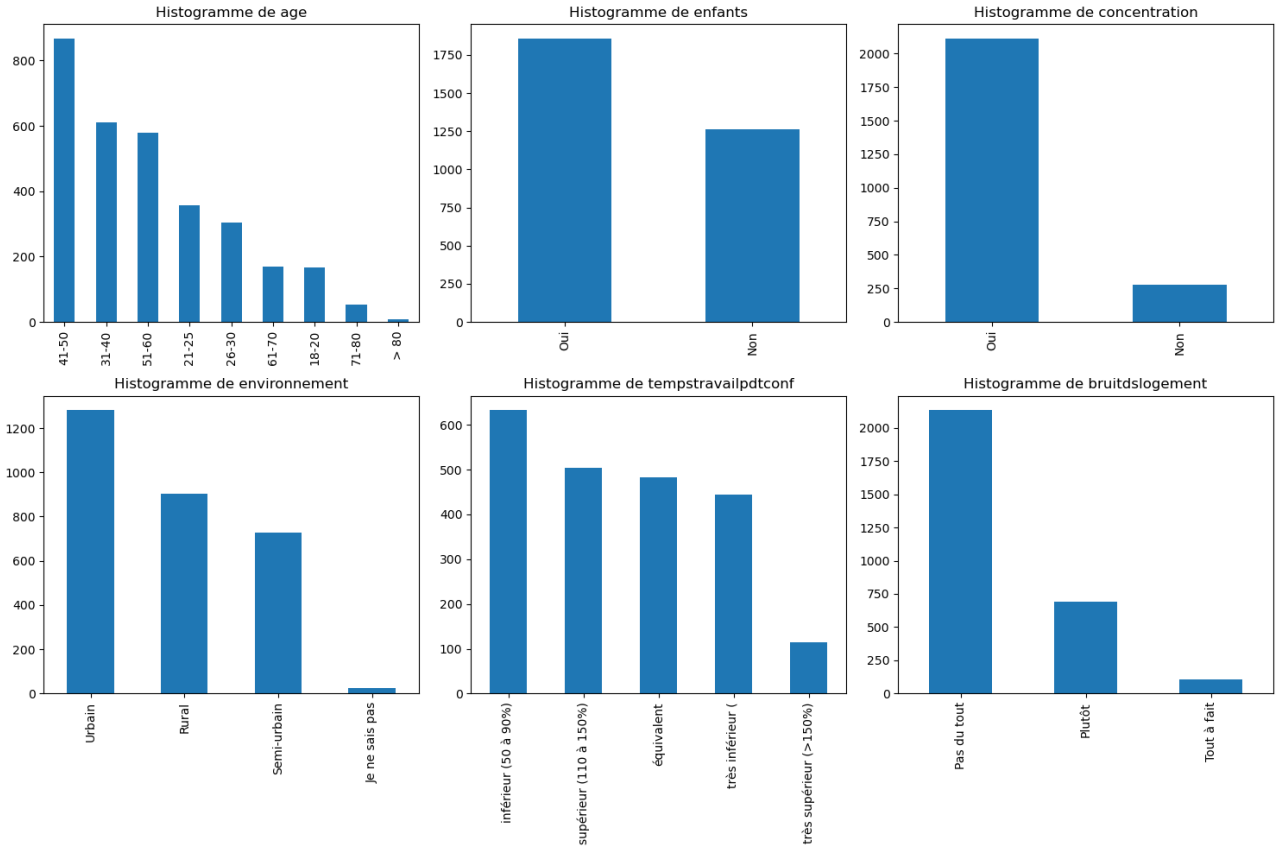

Figure 5: Histogramme

On remarque que la majorité des individus avaient un âge compris entre 41 et 50 ans, habitaient en milieu urbain et avaient un temps de travail inférieur à 50 à 90

5.4 Analyse Bivariée

5.5 Exploration de l'Impact des Autres Variables sur la Variable Cible

Dans cette section, nous avons exploré l'impact des autres variables sur notre variable cible, le "score_A". Pour ce faire, nous avons utilisé des graphiques de type boxplot pour chaque variable catégorielle sélectionnée.

Nous avons choisi un ensemble de variables catégorielles pertinentes telles que l'âge, l'activité professionnelle avant le confinement, la catégorie socioprofessionnelle, le niveau d'éducation, le niveau de bruit dans le logement et le niveau d'isolement dans le logement.

Pour chaque variable catégorielle, nous avons tracé un boxplot pour visualiser la distribution des valeurs de la variable cible ("score_A") en fonction des différentes catégories de la variable catégorielle. Les boxplots nous permettent de visualiser la médiane, les quartiles et les valeurs aberrantes de la variable cible pour chaque catégorie de la variable catégorielle.

L'utilisation de ces boxplots nous permet d'observer visuellement les variations de la variable cible en fonction des différentes catégories des autres variables, ce qui nous aide à mieux comprendre l'impact de ces variables sur notre variable cible.

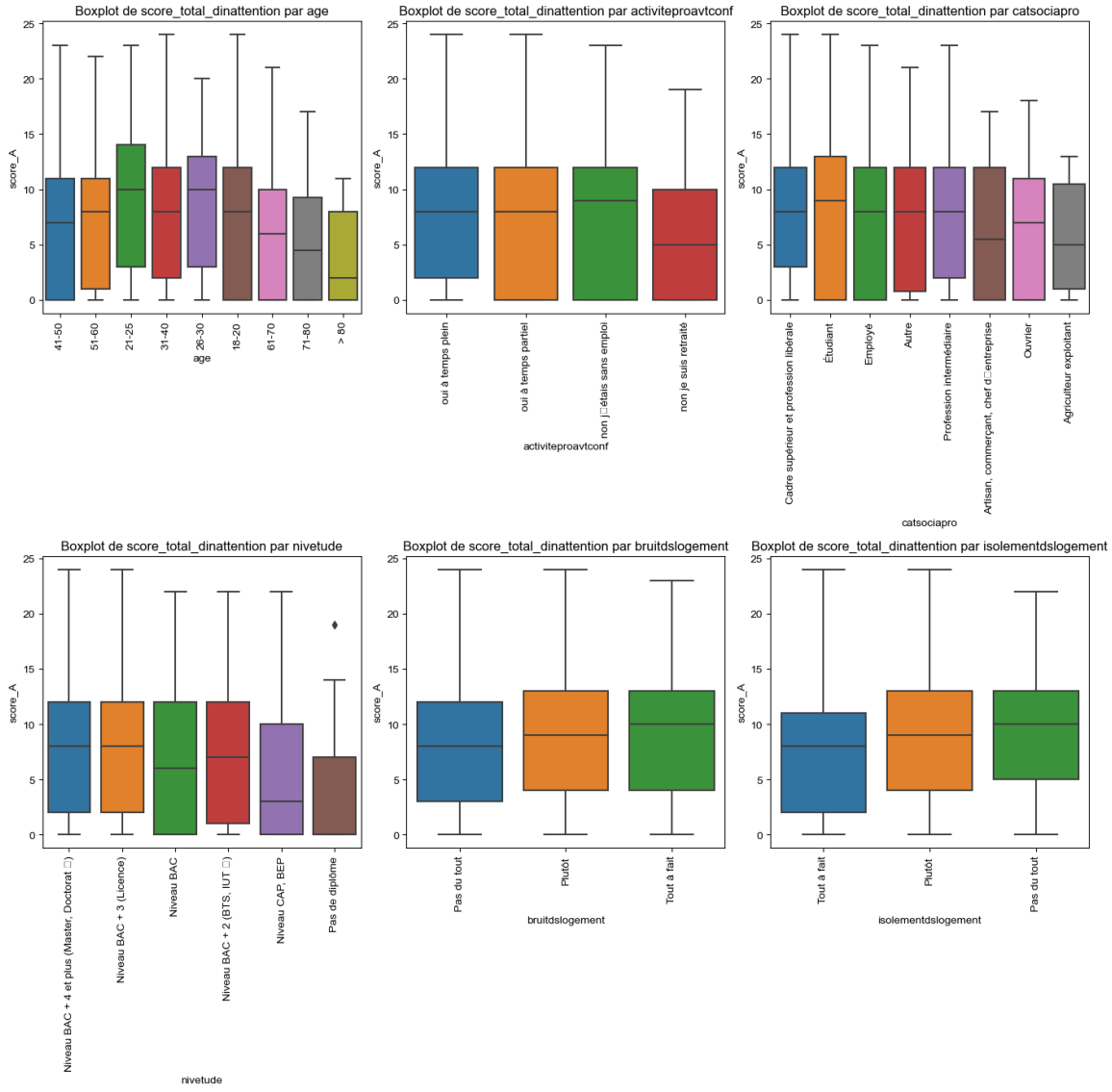

Figure 6: Boxplot

On constate que l'intervalle d'âge de $21 - 25$ semble avoir une influence significative sur le score A, car il a une médiane de 10, légèrement plus élevée que les autres intervalles d'âge. En revanche, les personnes âgées de plus de 80 semblent avoir une influence moins importante.

Pareil pour les gens qui ont fait des longues études (bac+3 ou plus) ont une influence très importante sur le scoreA que ceux qui n'ont pas fait des études du tout.

6 Modelisation

Nous allons utiliser des modèles explicatifs pour tenter d'expliquer notre phénomène, qui est l'attention, à partir des caractéristiques démographiques. Pour cela, nous allons recourir à deux modèles classiques : la régression linéaire et la forêt aléatoire.

Avant de faire une modélisation nous allons faire une brève introduction de ces méthodes,comment elles fonctionnent,une introduction théorique.

6.1 Regression Lineaire

Le modèle linéaire est utilisé dans un grand nombre de champs disciplinaires. Il en résulte une grande variété dans la terminologie. Soit le modèle suivant :

$$Y = X\beta + \varepsilon \tag{1}$$

La variable Y est appelée variable expliquée, variable dépendante, variable endogène ou encore réponse. Les variables X sont appelées variables explicatives, variable indépendante, variables exogènes ou encore prédicteurs. est appelé terme d'erreur ou perturbation.

On note généralement $\hat{\beta}$ le vecteur des paramètres estimés. On définit la valeur prédite ou ajustée $\hat{Y} = X\hat{\beta}$ et le résidu comme la différence entre la valeur observée et la valeur prédite : $\hat{\varepsilon} = Y - \hat{Y}$.

On définit aussi la somme des carrés des résidus (SCR, ou SSR en anglais) comme la somme sur toutes les observations des carrés des résidus :

$$\mathrm{SCR} = \mathrm{SSR} = \hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

6.2 Foret Aleatoire

Les forêts d'arbres décisionnels, également appelées forêts aléatoires, constituent une technique d'apprentissage utilisée pour notre problématique liée à l'attention. Cette méthode repose sur l'utilisation d'arbres de décision et fait partie des méthodes d'apprentissage ensembliste, qui exploitent la diversité des modèles pour améliorer les performances prédictives. Initialement proposées par Ho en 1995, elles ont été formellement présentées en 2001 par Leo Breiman et Adele Cutler. L'algorithme des forêts d'arbres décisionnels combine les concepts de bagging pour la sélection des données et d'utilisation de sous-espaces aléatoires. Il effectue un apprentissage sur plusieurs arbres de décision, chacun entraîné sur des sous-ensembles de données légèrement différents, afin de capturer au mieux la complexité de notre problème d'attention..

7 Applications et Resultats

7.1 Regression Lineaire

Ici, nous allons utiliser le modèle linéaire pour identifier les variables pertinentes susceptibles d'influencer le phénomène que nous souhaitons expliquer, à savoir l'attention.

la figure ci-dessous qui illustre les caractéristiques les plus influentes en termes de coefficients.

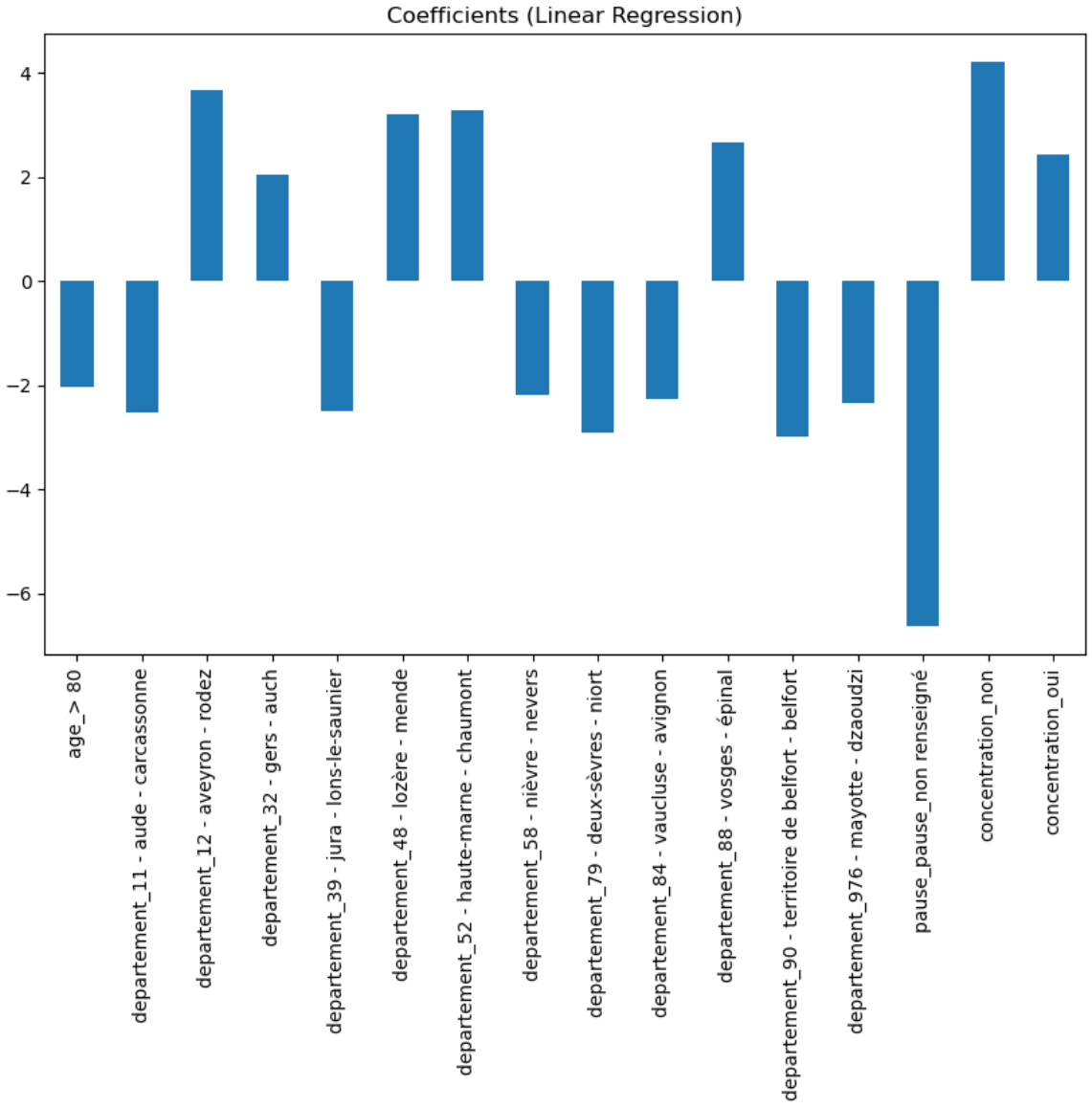

Figure 7: Contribution des Variables discrimantes

Nous constatons que la variable pause contribue de manière significative, ou bien en d'autres termes, elle explique bien notre phénomène (l'attention), de même que les variables département et âge.

Par contre la variable concentration contribue de manière négative,c'est à dire qu'elle est corrélé de manière négative à notre phénomène.C'est qui peut tout aussi se comprendre,car plus on sollicite notre cerveau plus on fatigue nos muscles de l'attention et au bout d'un moment on perd notre capacité d'attention.

7.2 Random forest regression

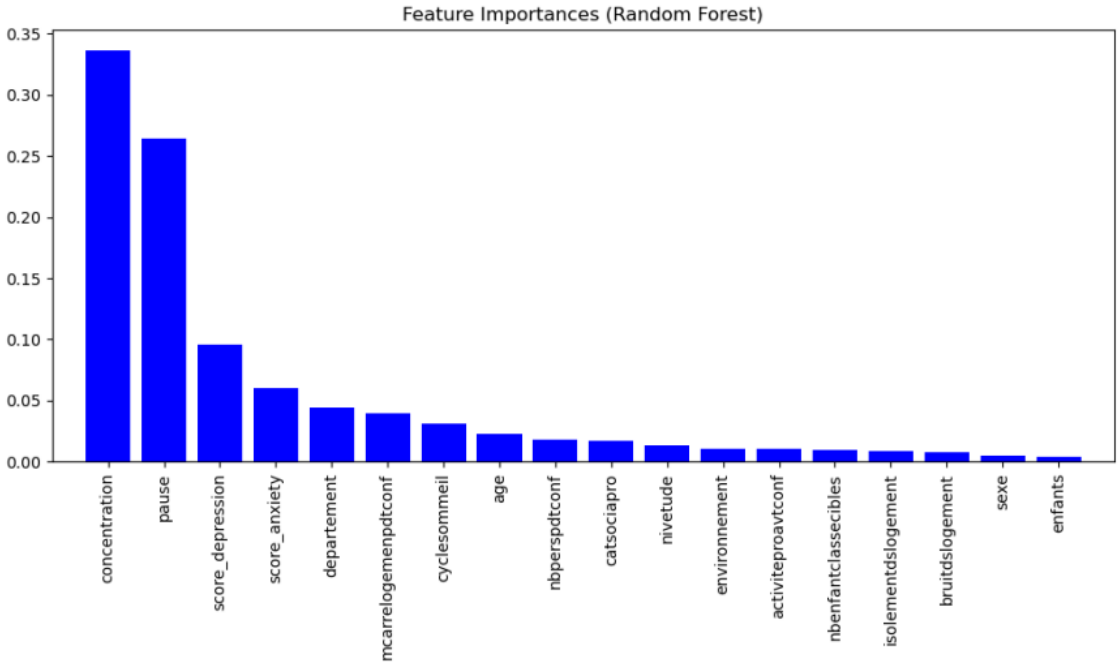

Figure 7: Variables importantes

Figure 8: Variables importantes

Ici nous constatons aussi que les variables pause ,nombre de mètre carré par logement nbr de personne pendant confinement ont une forte influence sur l'attention en terme d'importance.

8 Conclusion

En conclusion, cette étude a mis en lumière l'impact significatif des facteurs démographiques sur l'attention. À travers une analyse systématique, nous avons identifié des tendances importantes, révélant que des caractéristiques telles que l'âge, le sexe et le niveau d'éducation ont des répercussions variées sur la capacité d'une personne à maintenir son **attention**. Ces résultats soulignent l'importance de prendre en compte ces facteurs dans la conception d'interventions ciblées visant à améliorer la santé cognitive et le bien-être des individus. En comprenant mieux les influences démographiques sur l'attention, nous sommes mieux équipés pour développer des stratégies de prévention et d'intervention efficaces, ouvrant ainsi la voie à une amélioration significative de la santé mentale et du fonctionnement cognitif.