Projet I2E

Filière : Génie Énergétique et Environnement

Construction d'un prédicteur du confort thermique

Réalisé par :

- Hicham ESSABBAR
- NASSA Ayoub
- BABA Anas
- HATAFI Rami

Encadré par :

❖ Pr. Benjamin BERTIN

INSA Lyon
Année universitaire 2023/2024

Liste des figures

Table des matières

Introduction

Dans le domaine du confort des occupants de bâtiments, la sensation thermique est un élément crucial. Les approches conventionnelles, telles que le modèle Predicted Mean Vote (PMV), sont reconnues pour leur efficacité. Cependant, elles ne reflètent pas toujours avec exactitude la sensation thermique réellement perçue par les occupants. Pour adresser ce défi, notre projet se propose d'utiliser des techniques avancées de machine learning afin de réduire ces écarts.

L'objectif principal de notre travail est de développer un modèle de prédiction qui est capable de prédire avec une grande précision la sensation thermique ressentie par les occupants d'un bâtiment. Pour cela, nous exploitons la base de données ASHRAE, qui contient non seulement des mesures environnementales variées mais aussi des données sur les perceptions subjectives des occupants. Ce modèle se veut donc plus adaptatif et précis que les méthodes traditionnelles.

Notre approche méthodologique repose sur l'utilisation de techniques d'apprentissage automatique pour analyser les données fournies par ASHRAE. Nous nous concentrons particulièrement sur l'étude des relations entre les conditions environnementales et les réactions des occupants. L'ambition est de créer un modèle capable de surpasser les modèles actuels, comme le PMV, en termes de prédiction de la sensation thermique.

Nous envisageons que la mise en place d'un modèle plus précis et adaptatif aura un impact significatif sur deux fronts principaux. D'une part, il contribuera à améliorer le confort thermique des occupants, ce qui est un aspect non négligeable de la qualité de vie dans les espaces intérieurs. D'autre part, ce modèle permettra une gestion plus efficace de l'énergie dans les bâtiments, aboutissant à des économies d'énergie et une empreinte carbone réduite.

Chapitre I: Étude bibliographique et présentation des données

1. Étude bibliographique

1.1. Predicted Mean Vote PMV

Le modèle Predicted Mean Vote (PMV) est une composante essentielle dans l'étude du confort thermique des occupants de bâtiments [1]. En effet, Le PMV est un indicateur qui prédit la moyenne des réponses de confort thermique d'un grand groupe de personnes. Basé sur la théorie de l'équilibre thermique du corps humain, le modèle intègre divers facteurs environnementaux et personnels. Ces facteurs incluent la température ambiante, la vitesse de l'air, l'humidité relative, la température radiante, l'activité métabolique et l'isolation des vêtements.

Fanger a conçu le PMV pour quantifier de manière objective la sensation de confort thermique. Le modèle produit un score sur une échelle allant de -3 (froid) à +3 (chaud), où 0 représente un confort thermique optimal. Ce modèle est largement utilisé pour évaluer et concevoir des environnements intérieurs confortables, et il est devenu une référence dans les normes internationales de confort thermique, y compris la norme ASHRAE 55.

Toutefois, il est important de noter que le PMV a ses limites. Bien qu'il soit basé sur des principes physiologiques et psychologiques solides, le PMV ne prend pas toujours en compte la diversité des préférences individuelles et des adaptations comportementales. De plus, le modèle est principalement conçu pour des environnements modérés et peut ne pas être aussi précis dans des conditions extrêmes ou atypiques. C'est pourquoi il y a un intérêt croissant pour développer des modèles plus adaptatifs et personnalisés, en utilisant les techniques de machine learning, pour mieux répondre aux besoins individuels en matière de confort thermique.

1.2. Limites de la norme ASHRAE 55

La norme ASHRAE 55, largement utilisée pour évaluer et assurer le confort thermique dans les bâtiments, se concentre principalement sur les paramètres de l'environnement intérieur tels que la température, l'humidité, et la vitesse de l'air. Toutefois, cette norme a été initialement conçue en tenant compte des bâtiments avec des systèmes de climatisation et de chauffage mécaniques, ce qui pose des défis lorsqu'il s'agit de bâtiments ventilés naturellement.

De Dear et Brager soulignent que les bâtiments ventilés naturellement peuvent présenter des variations plus importantes et dynamiques des conditions environnementales internes, qui ne sont pas toujours adéquatement prises en compte par la norme ASHRAE 55 [2]. Par exemple, cette norme ne tient pas suffisamment compte des variations de température à court terme et de l'effet de l'adaptation personnelle des occupants à l'environnement thermique, qui sont des aspects cruciaux dans les bâtiments ventilés naturellement.

De plus, ASHRAE 55 s'appuie principalement sur des modèles statiques pour évaluer le confort, ce qui peut ne pas refléter fidèlement la réalité des occupants qui se déplacent et s'adaptent constamment dans des environnements dynamiques. Cette limitation souligne le besoin d'approches plus flexibles et adaptatives pour évaluer le confort thermique, en particulier dans les contextes où les conditions environnementales ne sont pas strictement contrôlées.

En conclusion, bien que la norme ASHRAE 55 soit un outil précieux pour garantir le confort thermique dans de nombreux contextes, ses limitations en ce qui concerne les bâtiments ventilés naturellement nécessitent une considération attentive. La recherche et le développement de nouvelles méthodes ou l'adaptation des normes existantes sont essentiels pour répondre aux défis spécifiques de ces environnements.

1.3. Modèles de prédiction existants

Dans l'article de Victor H [3], plusieurs algorithmes de machine learning sont discutés en termes de leur applicabilité et de leur efficacité dans la prédiction du confort thermique dans les bâtiments. Voici un résumé des algorithmes mentionnés et une comparaison de leurs caractéristiques et performances :

1.3.1. Random Forest :

- Description : Cet algorithme utilise un ensemble d'arbres de décision pour effectuer des prédictions. Chaque arbre est construit à partir d'un échantillon aléatoire des données et fait une prédiction indépendante. La prédiction finale est obtenue en moyennant les prédictions de tous les arbres.
- Avantages : Les Random Forests sont souvent privilégiés pour leur robustesse et leur capacité à gérer de grandes quantités de données avec des caractéristiques complexes.
- Limitations : Ils peuvent être sensibles aux données sur lesquelles ils sont entraînés, et il est crucial de choisir des données appropriées pour obtenir de bons résultats.

1.3.2. Deep Learning :

- Description : Les algorithmes de deep learning, tels que les réseaux de neurones profonds, apprennent des niveaux hiérarchiques de caractéristiques à partir des données, ce qui leur permet de capturer des relations complexes et subtiles.
- Avantages : Ils sont particulièrement efficaces avec de grands ensembles de données et peuvent capturer des relations non linéaires complexes.
- Limitations : Cependant, leur utilisation dans le contexte des logements locatifs est limitée en raison de la quantité restreinte de données disponibles. Le deep learning nécessite de grands volumes de données pour être efficace et peut être compliqué à mettre en œuvre.

1.3.3. Comparaison des Approches :

- Adéquation aux Données : Les Random Forests sont plus adaptés aux situations où la quantité de données est limitée, comme c'est souvent le cas dans les études sur le confort thermique dans les logements locatifs. En revanche, le deep learning nécessite des ensembles de données plus conséquents pour fonctionner efficacement.

- Complexité du Modèle : Les modèles de deep learning sont généralement plus complexes et peuvent capter des nuances plus subtiles dans les données. Cependant, cette complexité vient avec une augmentation des ressources computationnelles nécessaires et une plus grande difficulté à interpréter les modèles.
- Précision : Les Random Forests offrent un bon équilibre entre précision et complexité, en particulier lorsque les données disponibles sont limitées ou présentent des caractéristiques variées.

En résumé, l'article [3] met en évidence que le choix de l'algorithme de traitement dépend fortement des caractéristiques des données disponibles et des objectifs spécifiques de l'étude. Les Random Forests se distinguent comme une option fiable et robuste pour de nombreux cas d'usage en raison de leur capacité à gérer des ensembles de données hétérogènes et de taille modérée. Le deep learning, bien qu'offrant un potentiel de capture de relations plus complexes, est limité dans ce contexte par les exigences en termes de volume de données et de complexité computationnelle. En fin de compte, la sélection de l'algorithme approprié doit être guidée par une évaluation soignée des ressources disponibles, des caractéristiques des données et des objectifs spécifiques du projet.

2. Présentation des données

2.1. ASHRAE Global Thermal Comfort Database II

La base de données ASHRAE Global Thermal Comfort Database II se distingue par sa diversité de données, rassemblant des informations issues d'études de terrain menées dans une variété de pays. Cette collection offre un aperçu riche et varié des contextes environnementaux et culturels à travers le monde. Elle comprend des détails spécifiques sur le type de bâtiment, la stratégie de refroidissement utilisée, la région géographique, ainsi que le climat selon la classification de Köppen, offrant ainsi une vision complète des différentes variables qui influencent le confort thermique dans les bâtiments.

Un aspect clé de cette base de données est l'association des mesures objectives et subjectives. Les paramètres environnementaux objectifs comme la température de l'air, l'humidité relative, la vitesse de l'air et la température radiante sont méticuleusement enregistrés et couplés aux évaluations subjectives des occupants. Ces dernières incluent la sensation thermique, la préférence thermique et l'acceptabilité du mouvement de l'air. Cette approche intégrée permet une compréhension plus nuancée du confort thermique, tenant compte de la perception subjective des individus.

La base de données est également enrichie par des informations démographiques détaillées sur les occupants, telles que l'âge, le sexe, la taille, le poids et l'isolation des vêtements. Ces données sont essentielles pour saisir les variations individuelles dans la perception du confort thermique, permettant une analyse plus personnalisée et ciblée des besoins en matière de confort.

Enfin, la dynamique de mise à jour et d'évolution de la base de données est un atout considérable. Elle est régulièrement actualisée pour intégrer de nouvelles études, améliorant ainsi continuellement la qualité et la pertinence des données. Ouverte aux contributions de la

communauté scientifique, cette base de données encourage une collaboration et une expansion constantes, garantissant ainsi qu'elle reste un outil précieux et à jour pour la recherche dans le domaine du confort thermique.

En utilisant cette base de données, nous espérons développer un modèle qui reflète fidèlement la sensation thermique des occupants, en tenant compte des variables environnementales et personnelles. Notre approche vise à combiner les mesures objectives et les retours subjectifs pour obtenir un modèle qui soit non seulement scientifiquement solide, mais aussi pertinent et applicable dans divers contextes de bâtiments.

2.2. Présentation des paramètres

La base de données ASHRAE Global Thermal Comfort Database II est structurée en trois catégories principales de données, offrant une vue complète et détaillée des facteurs qui influencent le confort thermique dans les bâtiments.

2.2.1. Données Relatives aux Personnes

Cette catégorie englobe des informations démographiques et physiques sur les individus qui ont participé aux études. Ces données sont cruciales pour comprendre comment différentes caractéristiques personnelles peuvent influencer la perception du confort thermique. Les éléments clés comprennent :

- **Âge :** Indicateur important qui peut influencer la perception de la température et du confort.
- **Genre :** Permet d'analyser les différences de perception du confort thermique entre les sexes.
- **Poids et Taille :** Ces mesures physiques sont essentielles pour calculer le métabolisme basal des individus, un facteur déterminant dans la sensation de confort.
- **Isolation des Vêtements :** Informations sur l'habillement qui joue un rôle dans la régulation de la température corporelle et la sensation thermique.

2.3. Données Relatives à l'Environnement d'Étude

Les variables environnementales du lieu de l'étude sont un autre aspect crucial de la base de données. Elles permettent de comprendre comment l'environnement bâti et les conditions climatiques affectent le confort thermique. Parmi ces données, on trouve :

- **Paramètres du Bâtiment :** Type de bâtiment, méthode de refroidissement (naturelle, mécanique, mixte), etc.
- **Conditions Climatiques Extérieures** : Température extérieure, humidité relative, vitesse du vent, et d'autres données météorologiques qui peuvent influencer les conditions intérieures et le confort.

2.4. Données Mesurées sur le Confort

Cette catégorie comprend les mesures objectives et les perceptions subjectives relatives au confort thermique. Ces données sont essentielles pour évaluer l'efficacité des environnements bâtis en termes de confort des occupants. Les principaux éléments sont :

- **Predicted Mean Vote (PMV) :** Un indice prédictif du confort thermique basé sur les variables environnementales et personnelles.
- **Sensation Thermique** : Évaluations subjectives des occupants sur leur sensation de chaleur ou de froid, souvent mesurée sur une échelle standardisée.
- **Autres Indicateurs de Confort** : Comme le Predicted Percentage Dissatisfied (PPD), la température effective standardisée (SET), et d'autres mesures qui aident à comprendre la satisfaction ou l'inconfort des occupants.

En somme, la base de données ASHRAE Global Thermal Comfort Database II offre une perspective exhaustive et multidimensionnelle sur le confort thermique, combinant des éléments personnels, environnementaux et perceptuels. Cette approche intégrée est essentielle pour développer des modèles de confort thermique qui sont à la fois précis et représentatifs de la diversité des expériences humaines dans différents environnements bâtis. En tenant compte de ces variables variées et interconnectées, la base de données permet une analyse approfondie et nuancée du confort thermique, cruciale pour l'optimisation du design des bâtiments et la création d'espaces de vie et de travail plus confortables et énergétiquement efficaces.

3. Analyse des données

3.1. Comparaison entre PMV est la sensation thermique

Dans un premier temps, on a examiné la différence entre le PMV (Predicted Mean Vote) et la sensation thermique (Figure 1).

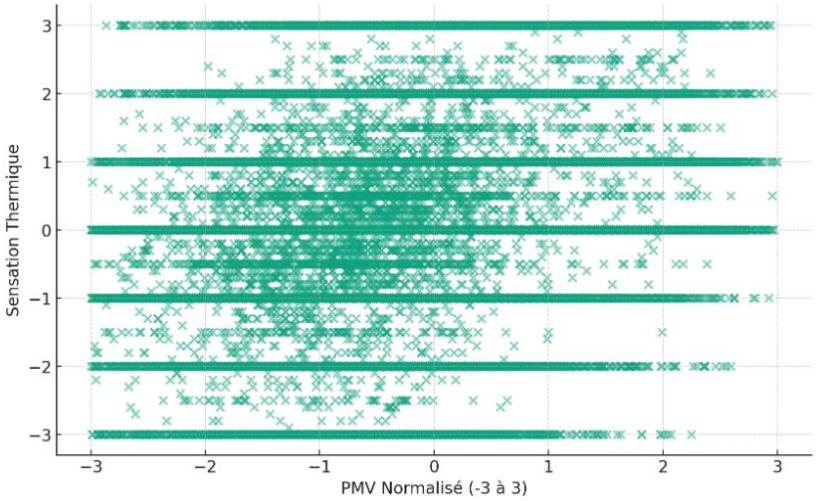

Figure 1 : Comparaison entre le PMV normalisé et la sensation thermique

La figure montre que bien que le PMV soit un indicateur reconnu du confort thermique, il existe des écarts notables entre le PMV et les sensations thermiques réelles rapportées par les individus. Cette observation suggère que le PMV, même normalisé, ne capture pas entièrement l'expérience subjective de la sensation thermique. Par conséquent, l'exploration de nouvelles variables ou la modification des modèles existants pourrait être nécessaire pour obtenir une représentation plus précise du confort thermique individuel.

3.2. Données manquantes

L'examen de notre dataset révèle un défi notable : la présence d'un grand nombre de valeurs manquantes dans plusieurs variables clés comme on peut voir sur la Figure 2:

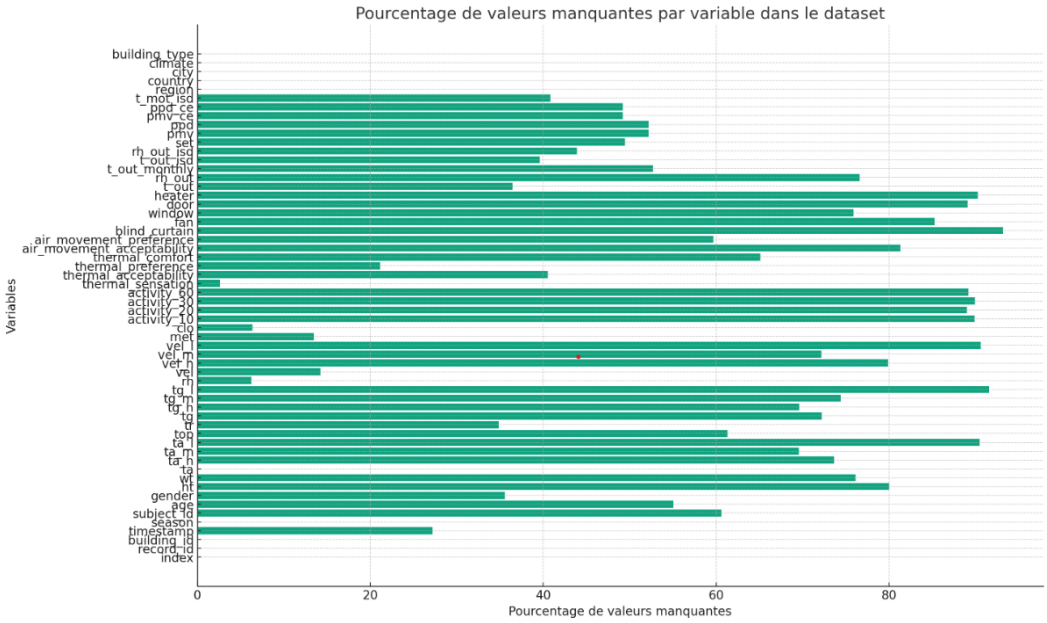

Figure 2 : Pourcentage de valeurs manquantes pour chaque paramètre

La présence de ces valeurs manquantes peut poser des problèmes significatifs pour l'analyse des données. Elles peuvent fausser les résultats et limiter l'efficacité des modèles statistiques et des analyses prédictives. Par conséquent, il est crucial de traiter ces valeurs manquantes de manière appropriée.

Il existe plusieurs techniques pour gérer les valeurs manquantes, chacune ayant ses propres avantages et inconvénients. Les approches utilisées pour résoudre ce problème sont :

Suppression des Données : Cette méthode implique la suppression des lignes ou des colonnes comportant des valeurs manquantes. Bien qu'elle soit simple à mettre en œuvre, cette technique peut entraîner la perte d'informations précieuses, surtout si le nombre de données manquantes est élevé.

Imputation des Valeurs Manquantes : L'imputation consiste à remplir les valeurs manquantes en utilisant différentes méthodes, telles que la moyenne, la médiane, ou des techniques plus complexes comme l'imputation multiple ou l'utilisation de modèles prédictifs. L'objectif est de conserver autant de données que possible en faisant des hypothèses raisonnables sur les valeurs manquantes.

Chapitre II: Analyse exploratoire de données

1. La répartition des données dans le monde

La Figure 3 illustre la répartition des données de notre dataset collectées dans le monde, la majorité des données provient de l'Angleterre, les êtas unis et de l'Inde. Plus une partie importante vient de l'Australie et de la Chine. Le reste des données (de 0.04% à 3.8%) provient des pays sud asiatiques, Canada, Mexique, Portugal, Italie, Tunisie, Japon...

Figure 3 : la répartition des données selon les pays

2. Matrice de corrélation

Dans la base de données disponible, les variables présentent une diversité de caractéristiques, incluant des données catégorielles et numériques. Pour faciliter l'analyse, une conversion des variables catégorielles en codes numériques a été effectuée, regroupant l'ensemble des données dans une matrice de corrélation, comme représenté dans la Figure 4. Cette visualisation permet d'explorer les différentes relations entre les variables.

Les observations issues de cette matrice de corrélation mettent en évidence plusieurs corrélations significatives. Les fortes corrélations qui sont de l'ordre de 0.9 entre les variables peuvent s'expliquer par le fait que ces variables partagent des liens intrinsèques ou sont influencées par des facteurs communs. Voici quelques explications possibles pour les fortes corrélations observées :

- **Températures à différents niveaux de la pièce (tg, ta, ta_h, ta_g, etc.)** : Ces variables sont toutes des mesures de température à divers endroits de l'environnement. Il est naturel de s'attendre à une forte corrélation entre ces mesures, car la variation de la température dans une pièce est généralement cohérente.

- **Vitesse de l'air à différents niveaux (vel, vel_h, vel_m, vel_l) :** La corrélation élevée entre ces variables peut être expliquée par le fait que la vitesse de l'air à différents emplacements dans la pièce est souvent liée. Par exemple, la vitesse de l'air à une hauteur donnée peut être influencée par la vitesse de l'air globale dans la pièce.
- **Métabolisme dans les différentes minutes passées (Activity 10, 20, 30, 60) :** Les différentes mesures de métabolisme au fil du temps peuvent être fortement corrélées, car l'activité métabolique d'une personne ne change généralement pas de manière significative d'une minute à l'autre.
- **Corrélations entre pmv, pmv_ce, ppd_ce et ppd :** Ces corrélations peuvent découler du fait que ces variables sont toutes liées à la perception thermique et au confort thermique. Même si les formules de calcul diffèrent selon les normes ISO ou ASHRAE, les facteurs qui influent sur le confort thermique restent les mêmes, d'où la similitude dans leurs variations.

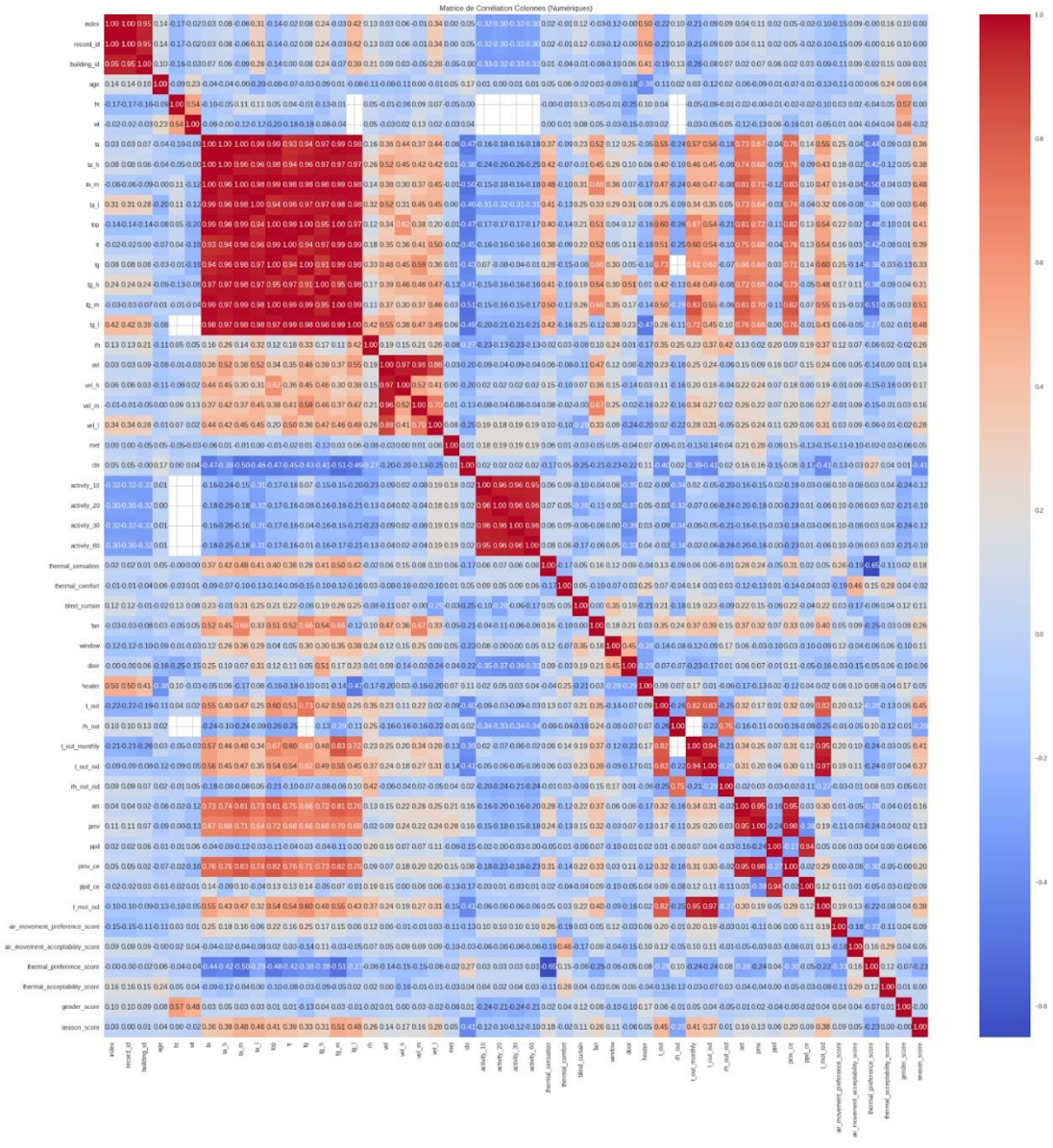

Figure 4 : Matrice des corrélations entre les différents variables numériques et catégorial de la dataset

3. Méthodes d'imputation

3.1. Acceptation thermique

Le remplissage des colonnes vides d'acceptation thermique sera effectué en fonction de l'acceptation de la vitesse de l'air. Les corrélations potentielles entre l'acceptation thermique et l'acceptation de la vitesse de l'air seront examinées en utilisant la vitesse de l'air. Un clustering des vitesses de l'air sera réalisé, classant les vitesses en "low velocity", "medium velocity", et "high velocity", comme illustré dans la Figure 5. Les clusters de vitesses sont définis comme suit :

- Low velocity : v < 0.4
- Medium velocity : 0.4 < v < 1.3
- High velocity : v > 1.3

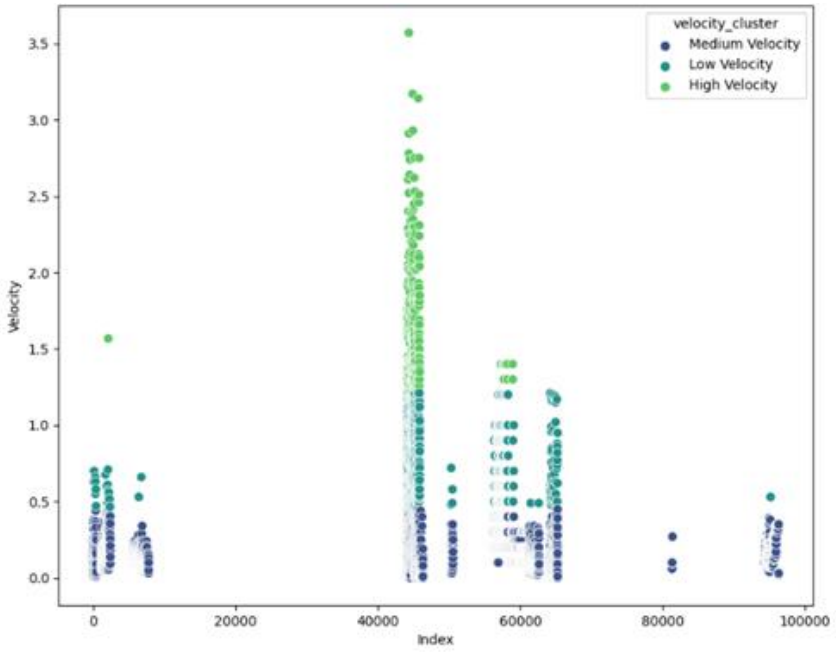

Figure 5 : Clusters des vitesses avec les frontières de décisions

Il est observable, d'après la Figure 6 venant de notre dataset et la Figure 7 prise de [4] et analysante une dataset d'une population au Brézil, que l'acceptabilité du mouvement d'air est liée à l'acceptabilité thermique. En revanche, la non-acceptabilité du mouvement d'air ne génère ni la non-acceptabilité thermique ni l'acceptabilité. Cependant, cela se manifeste partiellement, ce qui nous motive à compléter les cases vides de l'acceptabilité thermique associées à l'acceptabilité du mouvement d'air acceptable par acceptable.

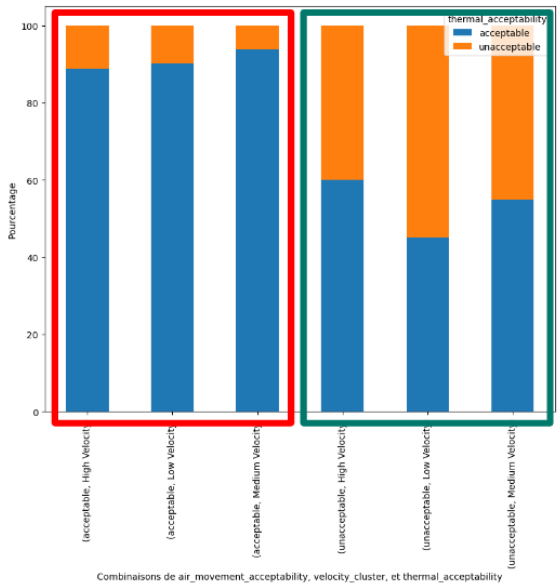

Figure 6 : Pourcentage de thermal acceptability en fonction de air mouvement acceptability et les clustrers des vitesses

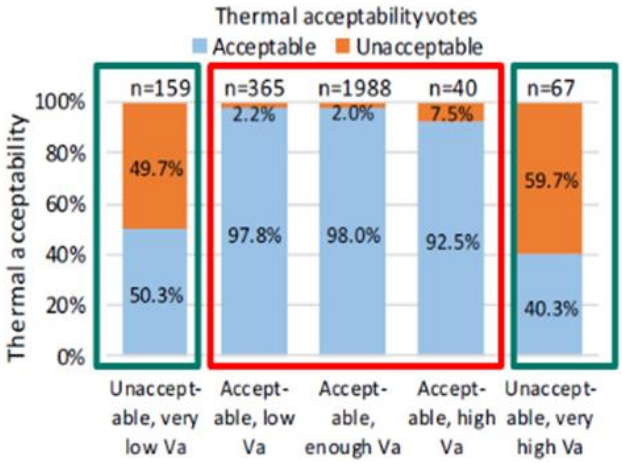

Figure 7 : Relation entre l'acceptabilité thermique et acceptation de la vitesse d'air d'après [4]

3.2. Utilisation de ventilateur

Selon la matrice de corrélation, une forte corrélation est observée entre le ventilateur (FAN) et les variables telles que ta_m, tg, vel_m, tg_m, avec des coefficients de corrélation égaux à 0.67. La division de l'ensemble des données en ensembles d'entraînement (80%) et de test (20%) par rapport aux données totales a été effectuée. Un modèle de régression logistique a été utilisé pour prédire la valeur binaire de FAN, atteignant une précision (accuracy) de 0.9, comme illustré dans la Figure 8 qui représente la matrice de confusion. La valeur 0 indique que le ventilateur est fermé, tandis que la valeur 1 indique que le ventilateur est ouvert.

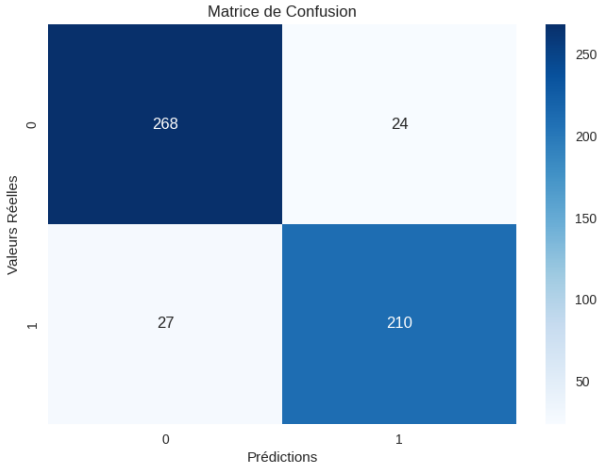

Figure 8 : Matrice de confusion montrant les valeurs réelles en fonction des valeurs prédites avec La valeur 0 indique que le ventilateur est fermé et 1 qu'il est ouvert

3.3. Acceptabilité du mouvement de l'air

En utilisant les mêmes clusters mentionnés dans la section 3.1, les résultats relatifs à l'acceptabilité du mouvement de l'air en fonction des vitesses sont obtenus. Pour les faibles vitesses (v < 0.4), l'acceptabilité s'établit à 80%. Pour les vitesses moyennes (0.4 < v < 1.3), l'acceptabilité atteint 90%, tandis que pour les hautes vitesses (v > 1.3), l'acceptabilité se situe à 85%. Ces résultats mettent en évidence la corrélation entre les différentes vitesses de l'air et les niveaux d'acceptabilité associés, comme illustré dans la Figure 9.

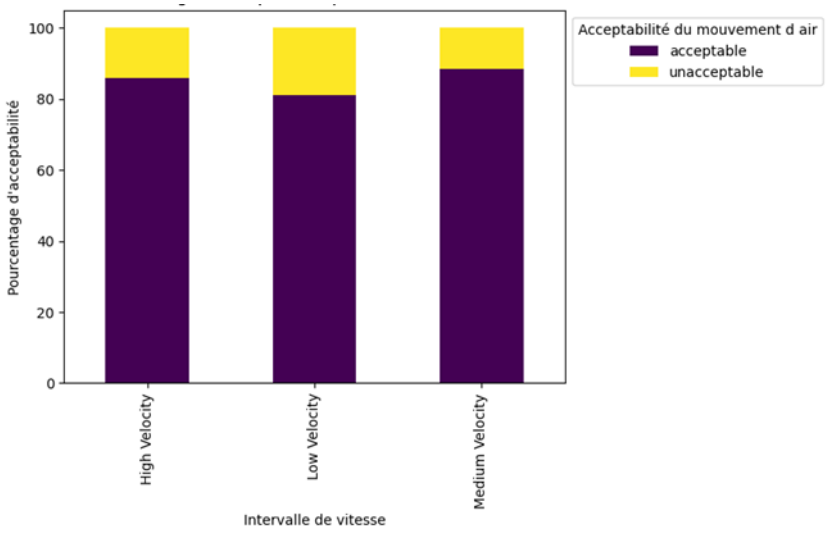

Figure 9 : Pourcentage d'acceptabilité du mouvement d'air ne fonction des clusters des vitesses

Une étude menée sur un pays ayant un climat tropical pour des valeurs très petites et moyenne valeurs de vitesses (v<1 m/s) [5] confirme ces observations, comme le démontre la Figure 10 et donc, un remplissage de toutes les cases vides par une acceptabilité positive.

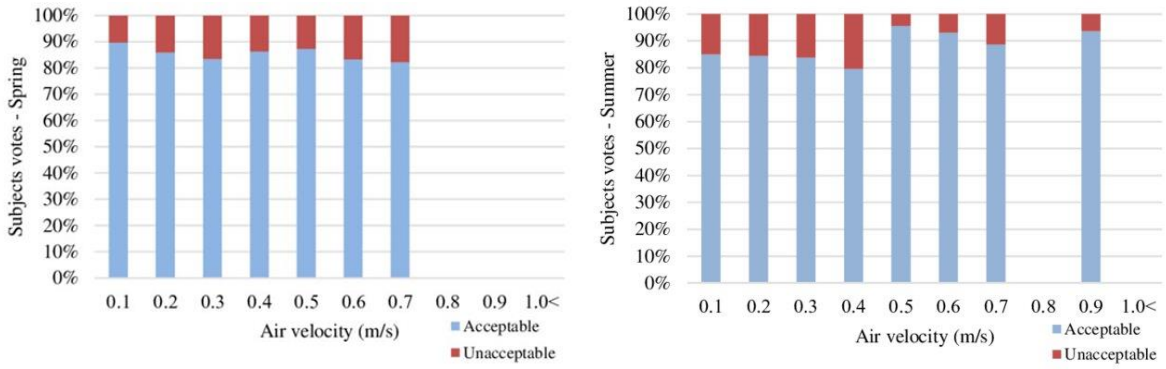

Figure 10 : Acceptabilité de mouvement d'air en fonction de la vitesse de l'air en été à droite et en hiver à gauche d'après [5]

3.4. Le métabolisme

Dans cette étape, nous allons essayer de remplir les valeurs manquantes (NAN) du métabolisme (l'attribut 'met') qui est un paramètre primordial pour mesurer le confort thermique, une analyse des valeurs des niveaux d'activité après des durées spécifiques de temps a montré une incohérence au niveau des unités des données de chaque étude. La figure 11 présente un ensemble des activités après 10 minutes qui ont des valeurs de l'ordre de 1 à 4 en unité met (1 met = 58 W). Les autres valeurs qui dépassent 60, ne peuvent être que des métabolismes en unité W. La première imputation était la correction des unités des activités.

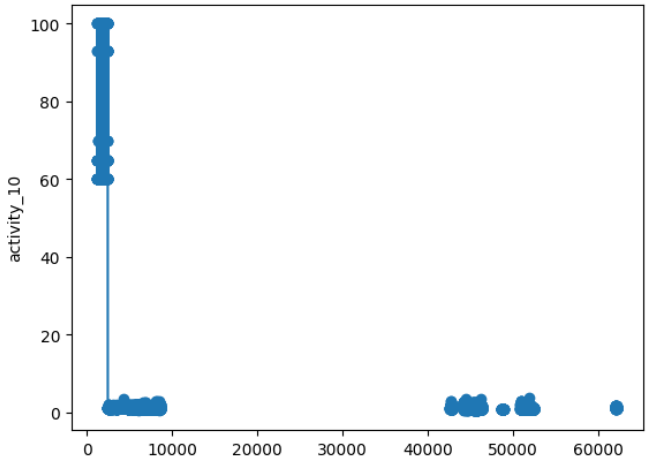

Figure 11 . Le tracé des valeurs de l'activité des sujets après 10 minutes

La régression linéaire entre le métabolisme et les quatre niveaux activités à des durés différentes a montré un coefficient de 0.25 pour chaque niveau d'activité. Donc la deuxième imputation était de remplir les valeurs manquantes de métabolisme par la valeur moyenne des activités à des durés différentes. 18 % des NAN ont été remplis.

Le métabolisme est une caractéristique intrinsèque des sujets, il est corrélé aux paramètres spécifiques de l'individus comme l'âge, le genre, le poids, la hauteur et le niveau d'habillement ("clo"). Il dépend aussi de l'activité qu'exerce l'individu, ce qui est traduit par le type de bâtiment (école, bureau...). Pour remplir les valeurs manquantes, nous avons séparé nos donnes selon le genre et le type du bâtiment, et nous avons testé des différentes corrélations pour les

autres attributs en cherchant la combinaison qui donne la plus petite erreur quadratique moyenne (MSE). Le MSE maximale qu'on a trouvé ne dépasse pas 0.03, alors nos corrélations sont importantes.

Le reste des NAN ont été remplis par la valeur de métabolisme la plus fréquente pour chaque type de bâtiment.

3.5. La température radiante

Le pourcentage des valeurs manquantes de la température radiante est de l'ordre de 38 % des données. Par définition de la température opérative (top), cette dernière est égale à la moyenne entre la température radiante (tr) et celle ambiante (ta), la première imputation de était de calculer "tr" manquantes à partir de "ta" et "top".

La corrélation entre "ta" et "tr" était de l'ordre de 95% avec un MSE de 0.85, la Figure 12 illustre la relation entre "tr" et "ta". Cette forte corrélation nous a permis de remplir tout le reste des valeurs manquantes.

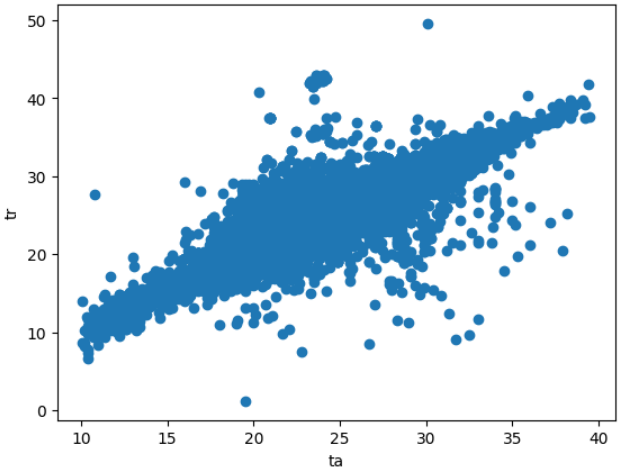

Figure 12 : La température radiante en fonction de la température ambiante

3.6. La vitesse de l'air

Le pourcentage des valeurs manquantes de la vitesse de l'air est de l'ordre de 14,2 % des données, le processus d'imputation des NAN de la vitesse de l'air ("vel") était similaire à celui du métabolisme. La première étape s'est agie de remplir les vitesses de l'air manquantes avec la moyenne des trois valeurs vitesses à différentes hauteurs (24 % des NAN).

La vitesse de l'air due au mouvement naturel (pas de ventilateur, portes et fenêtres fermées) dépend des paramètres thermodynamiques de l'air ambiant (Les températures ambiantes et radiantes et l'humidité relative). Le MSE de cette corrélation n'a pas dépassé 0.04 et a permis de remplir 39 % des NAN.

Pour le reste des valeurs ont été imputées par la valeur de la vitesse la plus fréquente selon chaque cas (ventilateur allumé, porte ouverte, fenêtre ouverte), exemple la Figure 13.

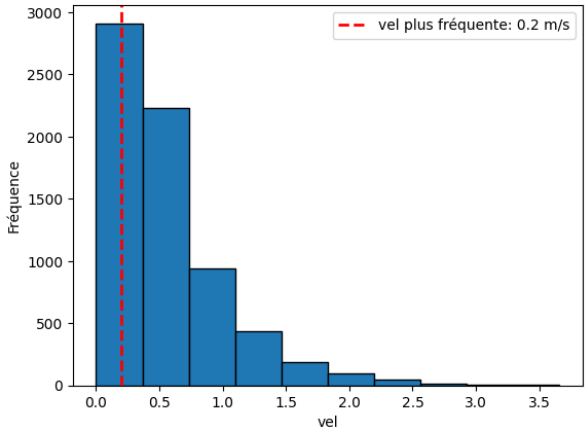

Figure 13 : la distribution des valeurs de la vitesse de l'air

3.7. Le PMV et le PPD

Les valeurs NAN des deux PMV (pmv : calculé par la méthode ISO, pmv_ce : celui par la méthode ASHRAE) représentent 57 % des lignes de notre dataframe. Grâce à la fonction PMV de la bibliothèque Thermalcomfort de Python, on avait la possibilité de remplir une grande partie des valeurs manquantes des PMV. Mais les paramètres d'entrée de la fonction qui sont les températures ambiante et radiante, la vitesse de l'air, l'humidité, le métabolisme et le niveau d'habillement, doivent respecter certaines plages de valeurs (Figure 14). Après cette imputation les NAN représentent 15 % pour PMV_ce et 20 % pour PMV.

Le paramètre	PMV	PMV_ce
Ta (°C)	10 - 30	10 - 40
Tr (°C)	10 – 40	10 - 40
vel (m/s)	0 – 0,88	0 - 1,88
met (met)	0,8 - 4	1 - 4
clo (clo)	0 - 2	0 – 1,5

Figure 14 : les plages des valeurs acceptables pour le calcul des PMV à l'aide de la bibliothèque thermalcomfort

3.8. L'âge et le genre

D'après une analyse des données, chaque type de bâtiment a une plage d'âge convenable, par exemple pour le centre pour personnes âgées, l'âge des individus ne peut pas descendre 60 ans. Donc nous avons rempli les valeurs manquantes de l'attribut 'age' par la moyenne pour chaque type de bâtiment (Figure 15).

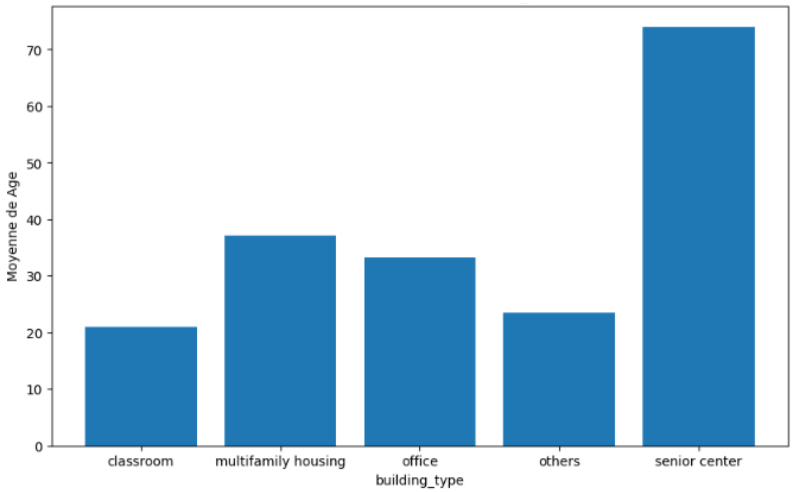

Figure 15 : la moyenne des valeurs d'âge pour chaque type de bâtiment

Pour les NAN de l'attribut genre, une petite corrélation entre le genre des sujets et leurs poids et taille a donné une erreur quadratique moyenne de 0,38. Cette corrélation a été adopté pour remplir toutes les valeurs manquantes du genre des sujets.

4. Matrice de remplissage

En procédant à la visualisation des matrices de données avant et après imputation, la Figure 16 révèle que 699 402 valeurs ont été imputées, représentant ainsi 12% de la matrice totale (53 colonnes * 109 033 lignes). Cette opération a significativement impacté la complétude de nos données, contribuant ainsi à l'amélioration globale de la qualité du jeu de données, comme démontré visuellement

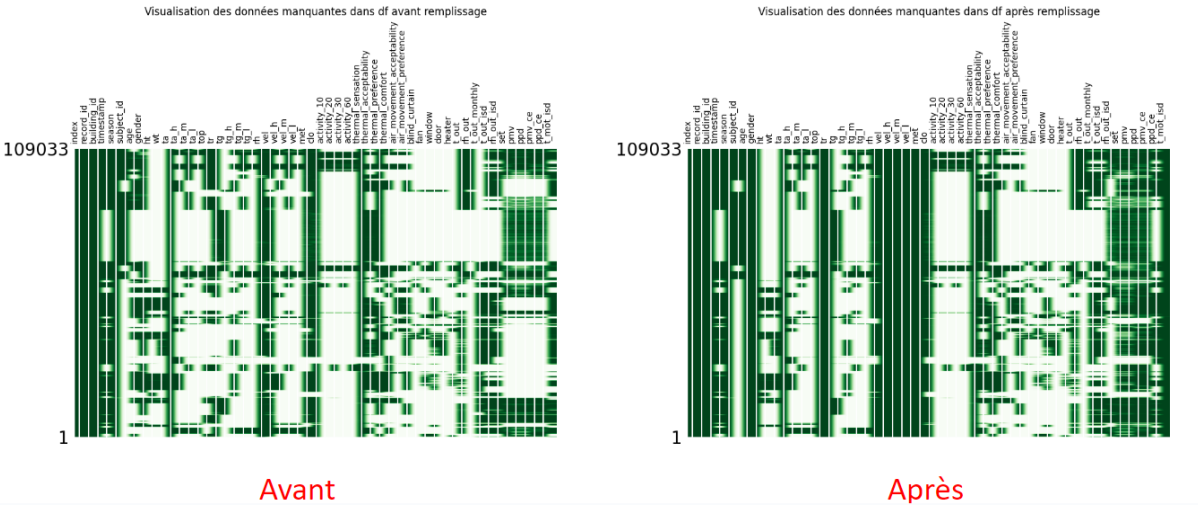

Figure 16 : Visualisation des donnée existants (en vert) avant et après imputation

Chapitre III:Construction du prédicteur de confort thermique

Les deux modèles de prédiction utilisés partagent un ensemble commun de caractéristiques d'entrée, sélectionnées pour leur lien avec l'objectif de l'étude, à savoir la sensation thermique ressentie par les occupants. Ces attributs d'entrée sont classés en trois catégories distinctes. La première catégorie concerne les paramètres environnementaux, incluant des éléments tels que la température ambiante, la température radiante, l'humidité relative, et la vitesse de l'air. La deuxième catégorie regroupe les paramètres individuels, qui comprennent des facteurs tels que le métabolisme, l'habillement, l'âge et le sexe de la personne. Enfin, la troisième catégorie englobe des paramètres démographiques et climatiques, tels que la région géographique, la saison de l'année, et le type de bâtiment. Ces divers éléments contribuent ensemble à fournir une vue d'ensemble complète nécessaire pour anticiper avec précision la sensation thermique des occupants.

1. Modèle prédictif N°1

Après avoir complété le processus d'imputation pour préparer les données requises, différents modèles d'apprentissage automatique ont été entraînés sur 80 % de ces données et évalués sur les 20 % restants. Lors de la comparaison de l'erreur quadratique moyenne (MSE) parmi divers algorithmes de régression, le modèle XGBoost a été sélectionné en raison de son MSE le plus bas, qui était de 1,108, comme le montre la Figure 17.

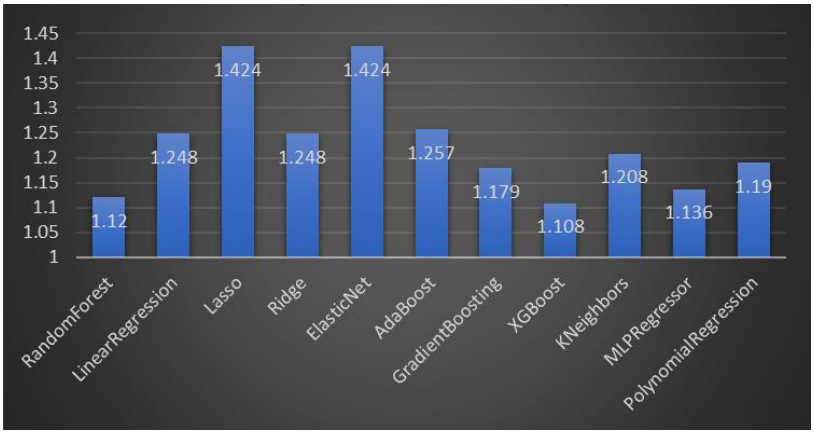

Figure 17 : Erreur quadratique moyenne entre les valeurs prédites et les valeurs réelles pour différentes méthodes de régression par le modèle (1)

La Figure 18 présentée est un nuage de points qui sert à comparer les valeurs prédites le modèle (1) avec les valeurs réelles observées. La dispersion des points est assez large, indiquant une variabilité significative dans les prédictions. Concernant la performance du modèle prédicteur, un MSE (Mean Squared Error) de 1,108 indique que les prédictions du modèle s'écartent en moyenne de la racine carrée de cette valeur par rapport aux valeurs réelles. Un coefficient de détermination de 0,22 révèle que seulement 22% de la variance des données

réelles est expliquée par le modèle, ce qui est relativement faible. Cela signifie que la capacité du modèle à prévoir avec précision les valeurs réelles est limitée, et que 78% de la variance est due à d'autres facteurs non pris en compte par le modèle ou à un bruit aléatoire. Ces indicateurs montrent que le modèle a une précision modeste et un pouvoir prédictif limité pour les données analysées.

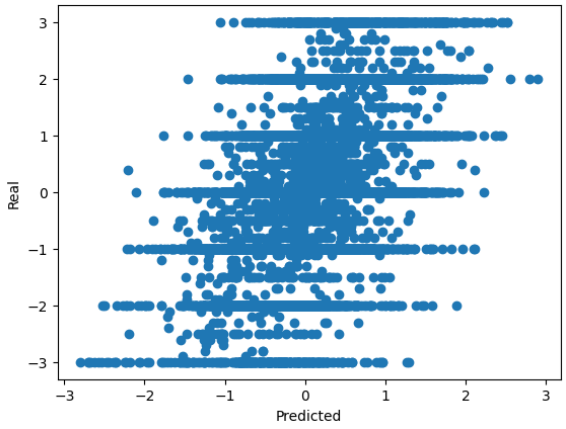

Figure 18 : Comparaison entre les valeurs réelles et les valeurs prédites par le modèle

2. Modèle prédictif N°2

La valeur de la sensation thermique est déterminée par le vote des répondants sur une échelle de -3 à 3, tandis que la valeur PMV est obtenue à partir d'une formule mathématique impliquant des variables telles que la température de l'air, le taux d'humidité, le métabolisme et la vitesse de l'air.

Des incohérences ont été identifiées dans les données actuelles. En effet, dans le tableau de la Figure 19, on observe que le PMV de la première personne indique une sensation thermique neutre, alors que selon les réponses au questionnaire, la personne déclare se sentir plus chaude, et une deuxième personne, bien que déclarant se sentir très froide dans le vote, le PMV indique que sa sensation thermique est neutre.

thermal_sensation	pmv
2	0.5
-3	-0.03

Figure 19 : Incohérence entre Thermal sensation et le PMV des personnes questionnées

Pour remédier à cette situation, une nouvelle variable sera créée en combinant les données de la sensation thermique et du PMV, en prenant en compte à la fois les votes subjectifs des individus, susceptibles d'être exagérés, et la composante scientifique et logique qui est représentée par le PMV.

La variable PCA combinée sera calculée à l'aide de la formule suivante :

$$\textbf{Combined PCA Variable} = w1 \times \textbf{Variable 1} + w2 \times \textbf{Variable 2}$$

Cette approche vise à intégrer de manière équilibrée les aspects subjectifs et scientifiques dans la création de la nouvelle variable.

Création de la colonne combined pca variable par la fonction de python sklearn.decomposition.PCA qui capture la variance maximale dans les données standardisées. Cette nouvelle variable représente une version compressée des informations Thermal sensation et PMV avec w1= 0.707 et w2= 0.707 obtenues par le vecteur propre w dans le contexte de l'Analyse en Composantes Principales (PCA) selon [6] est un concept essentiel lié à la décomposition spectrale de la matrice de covariance des variables d'origine. La matrice de covariance, notée C, est une mesure statistique qui quantifie les relations linéaires entre différentes paires de variables.

Le problème aux valeurs propres associé à la matrice de covariance, exprimé mathématiquement comme une équation caractéristique :

$$\mathbf{Cw = \lambda w}$$

Ici, w est le vecteur propre que l'on cherche à déterminer, λ est la valeur propre correspondante, et Cw représente le produit de la matrice de covariance C par le vecteur propre w.

La solution de cette équation caractéristique donne les vecteurs propres w et les valeurs propres λ. Les vecteurs propres indiquent la direction dans laquelle les données d'origine ont le plus de variabilité, et les valeurs propres quantifient l'importance de cette variabilité dans ces directions. En PCA, les vecteurs propres sont utilisés pour former les nouvelles variables (composantes principales) qui capturent le maximum de variance dans les données initiales.

Après avoir effectué une normalisation des valeurs de cette nouvelle variable entre -3 et 3, une étape de visualisation de la Combined PCA Variable a été entreprise. L'influence de la température sur cette variable est examinée dans la Figure 20. En effet, pendant l'hiver, avec des températures plus basses, les individus ont tendance à ressentir plus de froid, et inversement en été, où les températures plus élevées sont associées à une sensation de chaleur, comme illustré dans la Figure 20.

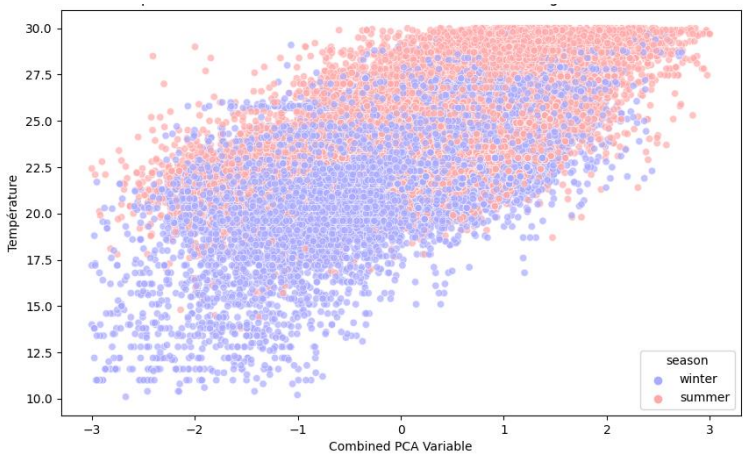

Figure 20 : Température en fonction de la nouvelle variable Combined PCA variable selon les saisons hiver ou été

Il est important de noter que la Combined PCA Variable peut différer considérablement des variables Thermal Sensation et PMV, comme le montre la Figure 21. Cette différence s'explique par le fait que la Combined PCA Variable combine ces deux variables, offrant ainsi une représentation intégrée qui peut diverger de manière significative du PMV et thermal sensation.

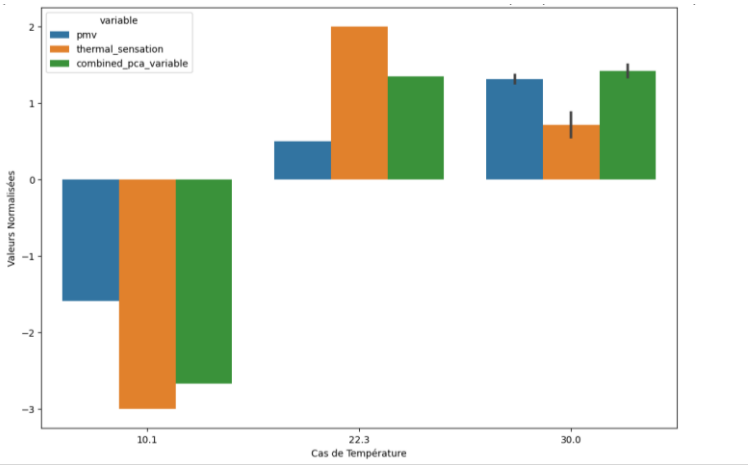

Figure 21 : Valeur de combined PCA variable en comparaison avec celle du PMV et Thermal sensation

L'ensemble de données a été stratifié en ensembles d'entraînement (80%) et de test (20%) par rapport aux données totales. Un modèle XGBoost Regressor a été développé en utilisant les variables ['region', 'season', 'building_type', 'age', 'gender', 'ta', 'tr', 'rh', 'vel', 'met', 'clo'] pour prédire la valeur de la Combined PCA Variable car il présente une erreur quadratique moyenne la plus faible par rapport à d'autres méthodes de régression comme le montre la Figure 22.

Après l'entraînement du modèle, l'évaluation de sa performance a été réalisée à l'aide du Mean Squared Error (MSE), qui s'est établi à 0,259. Cette métrique quantifie la moyenne des carrés des erreurs entre les valeurs prédites par le modèle et les valeurs réelles de la Combined PCA Variable, fournissant ainsi une indication de l'efficacité de la prédiction.

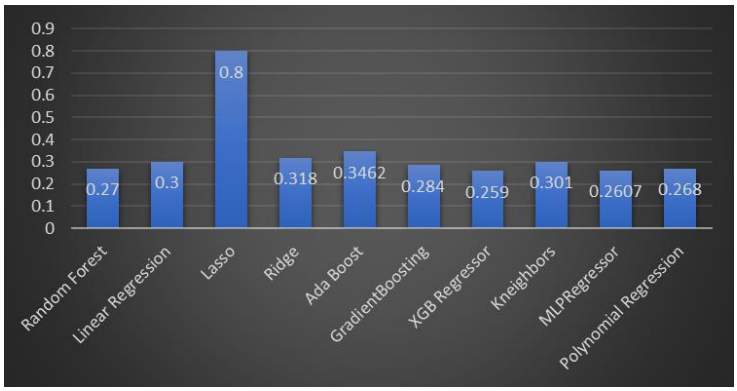

Figure 22 : Erreur quadratique moyenne entre les valeurs prédites et les valeurs réelles du Combined PCA variable pour différentes méthodes de régression par le modèle (2)

La performance de ce modèle peut être évaluée à l'aide de la Figure 23, où l'on observe que la relation entre les valeurs prédites et les valeurs réelles forme presque une droite linéaire.

Cette représentation graphique offre une visualisation de la cohérence entre les prédictions du modèle et les observations réelles, suggérant ainsi une adéquation linéaire dans la qualité des prédictions du modèle.

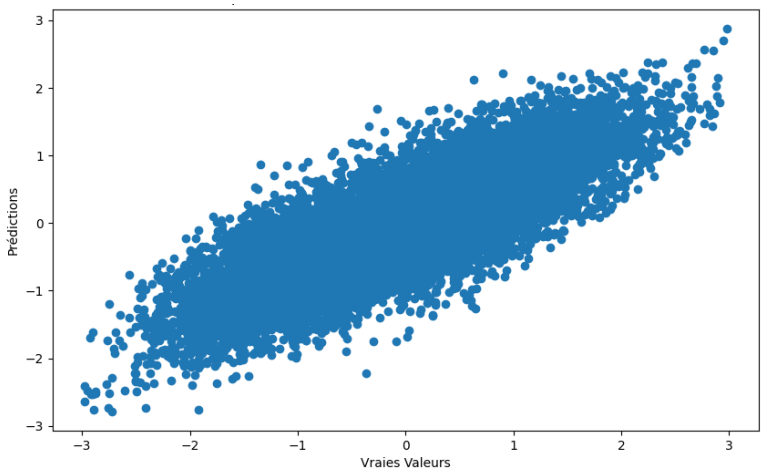

Figure 23 : Comparaison entre les valeurs prédites par XGBoost Regressor et les vraies valeurs du combined pca variable sur l'ensemble de test

3. Analyses de sensibilité des modèles

L'analyse de sensibilité est utilisée souvent dans l'évaluation de la fiabilité et de la robustesse des modèles de prédiction. En tant qu'outil méthodologique, elle permet d'examiner comment les différentes incertitudes présentes dans les variables d'entrée d'un modèle influencent ses résultats. Le principe de l'analyse de sensibilité est d'identifier les variables qui ont le plus grand impact sur les résultats du modèle. Cela implique de varier les entrées et d'observer comment ces variations modifient les sorties. Un modèle est considéré comme robuste si de petites variations dans les entrées n'entraînent pas de grands changements dans les sorties.

Nous procéderons à une analyse de sensibilité univariable. Nous allons se concentrer sur un seul paramètre tout en maintenant les autres constants. Pour les parties suivantes de notre analyse, les valeurs de référence choisies sont les suivantes : région (Europe), saison (hiver), type de bâtiment (salle de classe), sexe (masculin), âge (23 ans), température ambiante (ta : 16 °C), température radiante (tr : 17 °C), vitesse de l'air (vel : 0,2 m/s), métabolisme (met : 1,2) et vêtement (clo : 1).

3.1. Modèle prédictif N°1

L'analyse examine les attributs suivants dans leurs plages spécifiques : "ta" (température ambiante) entre 10 et 40 °C, "tr" (température radiante) entre 1 et 50 °C, "rh" (humidité relative) entre 2 et 100 %, "vel" (vitesse de l'air) entre 0 et 4 m/s, ainsi que "met" et "clo" qui varient selon des standards de référence [7].

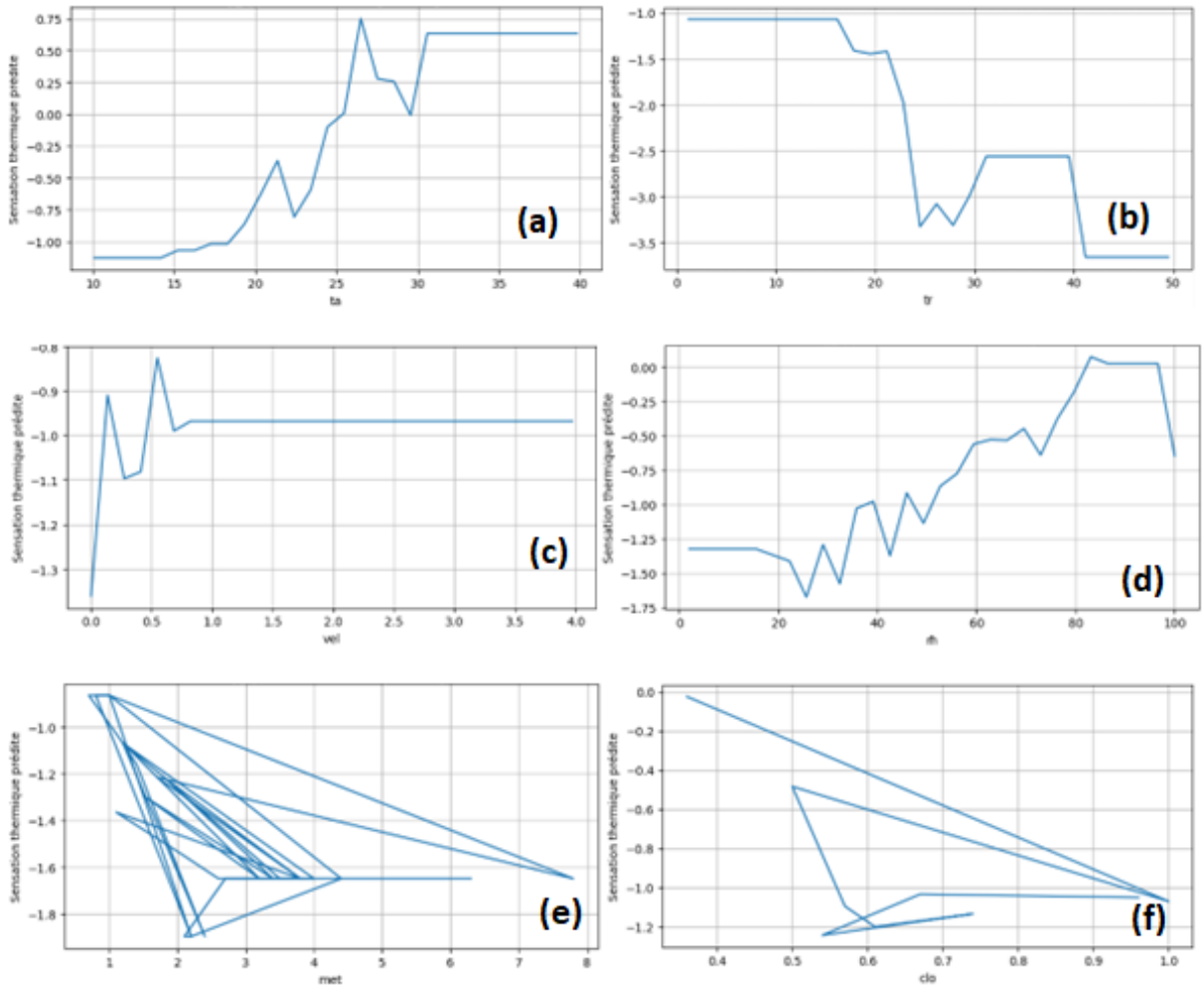

Figure 24 : Sensibilité de la sortie du modèle (1) en fonction des variables d'entrée (a) "ta" ; (b) "tr" (c) "vel" ; (d) "rh" ; (e) "met" ; (f) "clo"

Cependant, l'analyse est limitée dans certains scénarios, particulièrement dans les cas (a) et (b), où les variations de "ta" et "tr" semblent illogiques et imprévisibles. Cette irrégularité est attribuée à la forte corrélation entre ces deux températures, suggérant qu'une variation de l'une entraîne inévitablement une variation similaire de l'autre. L'évolution de la sortie en fonction des attributs "met" et "clo" illustrée dans la Figure 24 (e) et (b) montre la réponse du modèle face à des ensembles spécifiques de valeurs discrètes que ces deux attributs peuvent prendre.

Selon les observations faites à partir de la figure précédente, il est évident que les changements dans les paramètres d'entrée du modèle (1) influencent considérablement la perception de la sensation thermique prédite. Cela indique que le modèle est très sensible et manque de robustesse. Cette conclusion est cohérente avec les limitations précédemment mentionnées du modèle, qui ne parvient à expliquer que 22% de la variance observée.

3.2. Modèle prédictif N°2

Nous procédons de la même manière pour examiner la robustesse du deuxième modèle prédictif.

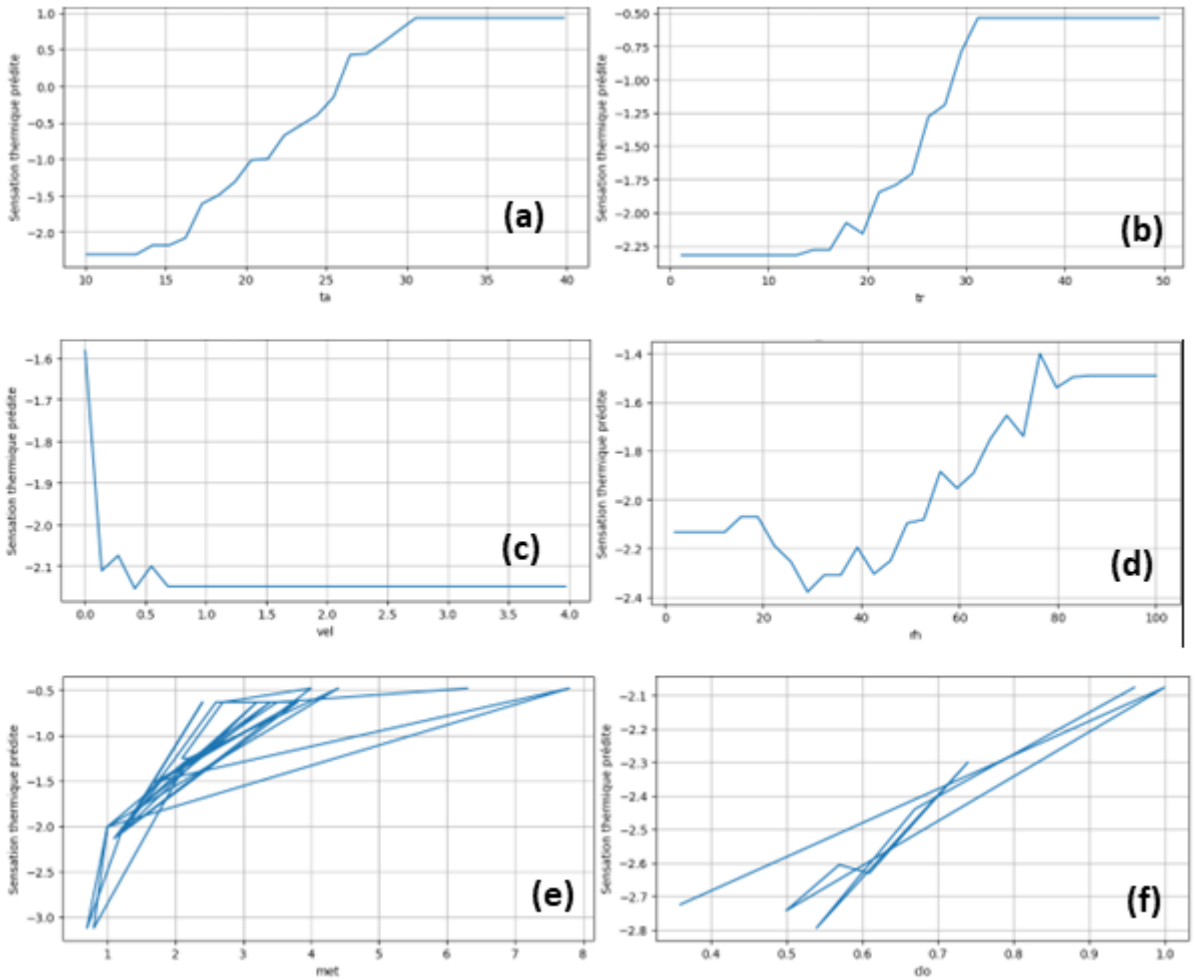

Figure 25 : Sensibilité de la sortie du modèle (2) en fonction des variables d'entrée (a) "ta" ; (b) "tr" (c) "vel" ; (d) "rh" ; (e) "met" ; (f) "clo"

L'observation de l'analyse de sensibilité montre que le modèle (2) réagit de manière significativement plus modérée aux changements des paramètres d'entrée par rapport au modèle (1). Cette observation, mise en évidence dans la Figure 25, indique que les variations des sorties du modèle (2) sont moins fluctuantes face aux variations des paramètres d'entrée. Cette moindre sensibilité du modèle (2) est cohérente avec ses performances supérieures, notamment sa capacité à expliquer jusqu'à 67% de la variance de la sensation thermique.

Conclusions et perspectives

Dans ce projet visant à aborder la question complexe de la prédiction du confort thermique, deux approches distinctes ont été développées pour aborder la problématique de la divergence entre la sensation thermique subjective (Thermal Sensation) et l'indice objectif PMV. La deuxième approche, est centrée sur la création d'une nouvelle variable, la Combined PCA Variable, en intégrant de manière équilibrée les aspects subjectifs des réponses individuelles et les mesures objectives du PMV. Cette méthode, alliant analyse en composantes principales et régression, a montré une capacité prometteuse à refléter de manière plus fidèle la complexité du confort thermique, comme le suggère le MSE de 0.259 et la corrélation linéaire dans les prédictions du modèle.

Le premier prédicteur, bien que plus simple dans son approche, a utilisé un modèle de régression non linéaire, en intégrant des variables telles que des facteurs environnementaux et personnels. Ses performances, avec un MSE de 1.05 et un R^2 de 22%, indiquent des limites dans sa capacité à anticiper entièrement la variabilité de la sensation thermique. Cette divergence dans les performances peut être en partie attribuée à la nature subjective de la sensation thermique, où les réponses des individus peuvent parfois être exagérées ou influencées par des facteurs non mesurables. Cette situation est illustrée par l'effet Hawthorne [8], lequel postule que la simple conscience des participants d'être impliqués dans une expérience peut influencer leur comportement naturel. Cela rend complexe la tâche de reproduire des conditions authentiquement réalistes dans le cadre de l'étude.

Pour l'avenir, l'intégration de variables externes telles que les conditions climatiques et environnementales (température, humidité) pourraient être une étape cruciale pour améliorer la précision des modèles. En tenant compte de l'influence de l'environnement extérieur, il est possible que les modèles puissent mieux anticiper la complexité de la sensation thermique et offrir des prédictions plus précises et personnalisées. De plus, la poursuite de la recherche sur la manière dont les perceptions individuelles du confort thermique sont formées et influencées pourrait également contribuer à affiner les modèles existants et à développer de nouvelles approches plus robustes et inclusives. Pour collecter des données subjectives avec une plus grande exactitude, l'emploi de technologies telles que les montres connectées pourraient être envisagé.

Références

[1] Fanger PO. Thermal comfort. Analysis and applications in environmental engineering. Therm Comf Anal Appl Environ Eng 1970.

[2] de Dear RJ, Brager GS. Thermal comfort in naturally ventilated buildings: revisions to ASHRAE Standard 55. Energy Build 2002;34:549–61. https://doi.org/10.1016/S0378-7788(02)00005-1.

[3] Hoyet V, Pannier M-L, Robillart M, Rousseau D. Vers le développement de modèles prédictifs individualisés du confort thermique pour les logements connectés. Conférence IBPSA Fr. 2022, 2022.

[4] De Oliveira CC, Rupp RF, Ghisi E. Influence of Air Movement and Air Humidity on Thermal Comfort in Office Buildings in Florianópolis, Brazil. 35th PLEA Conf.-Sustain. Archit. Urban Des. Plan. Post Carbon Cities Univ. Coruña Spain, 2020.

[5] de Freitas NVS, Mikuri LP, Andreasi WA. Influence of air movement on human thermal sensation in a tropical humid climate. Int J Sci Eng Investig 2018;7:70–6.

[6] Principal Component Analysis with NumPy – Wendy Navarrete 2020. https://wendynavarrete.com/principal-component-analysis-with-numpy/ (accessed January 15, 2024).

[7] Tartarini F, Schiavon S. pythermalcomfort: A Python package for thermal comfort research. SoftwareX 2020;12:100578. https://doi.org/10.1016/j.softx.2020.100578.

[8] Pastore L, Andersen M. Building energy certification versus user satisfaction with the indoor environment: Findings from a multi-site post-occupancy evaluation (POE) in Switzerland. Build Environ 2019;150:60–74. https://doi.org/10.1016/j.buildenv.2019.01.001.