Au cours de ce projet, nous avons mené une étude visant à développer et évaluer un modèle de prédiction des données anthropométriques. L'objectif était de créer un algorithme capable de prédire des mesures telles que la taille, le poids, le tour de poitrine (etc.) en utilisant des informations disponibles sur un sous-ensemble de la population. Le projet s'est déroulé en quatre grandes étapes : analyse de la qualité de la base de données, sélection des caractéristiques pertinentes et clustering, conception et entraînement du modèle puis validation et évaluation des performances.

Analyse de la Qualité de la Base de Données

La première étape a consisté à analyser la qualité de la base de données utilisée pour l'étude. Nous avons travaillé avec un jeu de données anthropométriques de KLEEP qui contenait des informations telles que la taille, le poids, l'âge, le sexe, etc pour un total de 17 données par individu. L'analyse de la qualité a porté sur l'examen des valeurs manquantes, des valeurs aberrantes, et de la cohérence des données. Nous avons trouvé que certaines variables présentaient un taux élevé de valeurs manquantes, ce qui a nécessité l'utilisation de techniques d'imputation comme la moyenne ou la régression linéaire. Les valeurs aberrantes ont été identifiées et traitées pour minimiser leur impact sur les résultats finaux. Une fois la base de données nettoyée et normalisée, elle était prête pour l'étape suivante.

Sélection des caractéristiques pertinentes et clustering

Dans cette phase, l'objectif était de comprendre les relations entre les différentes variables et de sélectionner celles qui seraient les plus pertinentes pour la prédiction. Nous avons utilisé des techniques de visualisation telles que les matrices de corrélation et les diagrammes de dispersion pour analyser les interdépendances entre les variables. En parallèle, des méthodes statistiques comme l'analyse de la variance et l'élimination des caractéristiques par sélection récursive ont été appliquées pour réduire la dimensionnalité du problème. Cette étape nous a permis de retenir un sous-ensemble optimal de variables explicatives, telles que l'âge, le sexe, et le tour de taille, qui montraient une forte corrélation avec les mesures anthropométriques à prédire.

Conception et Entraînement du Modèle

Une fois les caractéristiques sélectionnées, nous avons conçu et exploré différents modèles prédictifs. Les méthodes utilisées comprenaient des réseaux de neurones, des arbres de décision, et des techniques de clustering. Chaque modèle a été testé pour évaluer son efficacité à prédire les mesures anthropométriques à partir des données disponibles. Les réseaux de neurones, en particulier, ont été paramétrés avec plusieurs couches et neurones pour capturer les relations complexes entre les variables. Parallèlement, les arbres de décision ont été utilisés pour créer des modèles basés sur des règles de tri, permettant une interprétation plus facile des résultats. Les techniques de clustering ont été appliquées pour identifier des groupes naturels au sein des données, ce qui a permis d'améliorer la précision des prédictions en segmentant la population en sous-groupes homogènes.

Validation et Évaluation des Performances

La dernière étape a consisté à valider les modèles en utilisant un jeu de données de test indépendant. Nous avons comparé les performances des différents modèles à l'aide de métriques tel que le coefficient de détermination R^2. Parmi les modèles testés, le réseau de neurones s'est révélé être le plus performant, avec une marge d'erreur de seulement 1,9%. Les réseaux de neurones ont démontré une meilleure capacité à capturer les relations non linéaires complexes dans les données, offrant ainsi une prédiction plus précise et fiable.

Cette étude a permis de développer un modèle prédictif performant pour estimer des données anthropométriques à partir de variables partiellement observées. L'analyse rigoureuse de la qualité des données, combinée à une exploration approfondie des caractéristiques et à l'entraînement de différents modèles, a conduit à des résultats significatifs.