Introduction à l'IA et Robotique
Rapport – Projet d'apprentissage machine

BOUDEFLA Dounia

Ce rapport détaille un projet d'apprentissage machine qui a exploré l'utilisation de la régression pour prédire les prix des logements en Californie. À travers le jeu de données "California Housing", des méthodes statistiques ont été appliquées pour le traitement et l'analyse des données, et des modèles ont été construits pour estimer les valeurs immobilières. L'objectif était de déchiffrer les facteurs influençant les prix des maisons et de fournir des prédictions à la fois fiables et explicables.

1. Exploration statistique des données

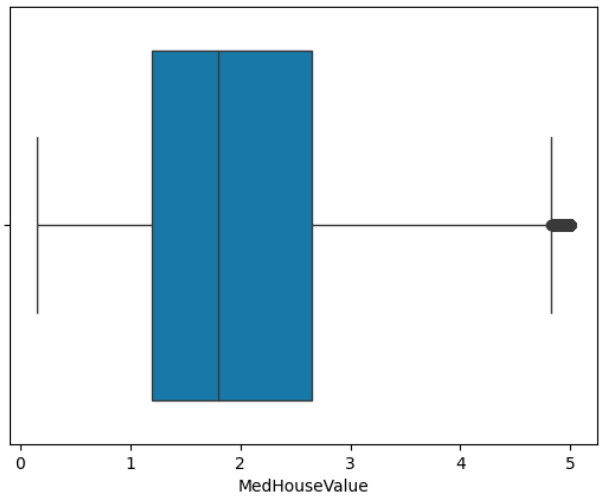
Le jeu de données "California Housing" a été chargé via la fonction fetch_california_housing() de la bibliothèque scikit-learn. Les caractéristiques et la variable cible ont été séparées pour créer un DataFrame.

Informations sur le DataFrame :

Le DataFrame contient 20640 entrées et 9 colonnes, toutes de type float64, ce qui indique que les caractéristiques sont numériques.

Visualisation des données :

Une boîte à moustaches a été utilisée pour visualiser la distribution des valeurs de 'MedHouseValue', permettant de comprendre la répartition des prix médians des maisons et d'identifier d'éventuelles valeurs absurdes.



Valeurs manquantes :

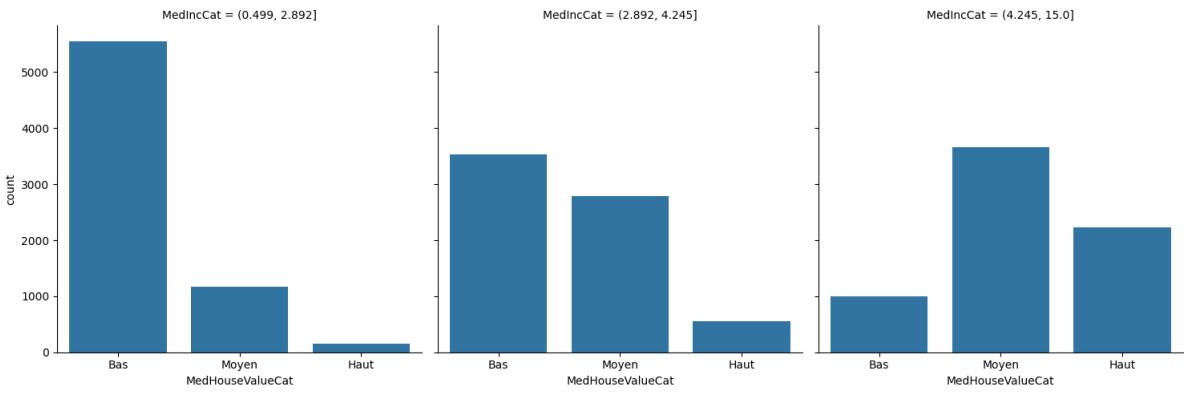
L'examen du DataFrame a confirmé l'absence de valeurs manquantes dans toutes les colonnes.

```
Poucentage des valeurs manquantes:

MedInc          0.0
HouseAge        0.0
AveRooms        0.0
AveBedrms       0.0
Population      0.0
AveOccup        0.0
Latitude        0.0
Longitude       0.0
MedHouseValue   0.0
```

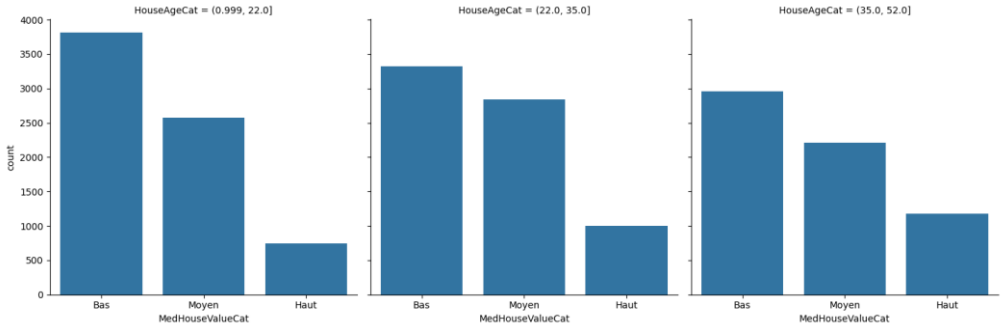
Corrélation entre les variables :

```
Corrélation entre les variables

                MedInc  HouseAge  ...  Longitude  MedHouseValue
MedInc        1.000000 -0.119034  ...  -0.015176       0.688075
HouseAge     -0.119034  1.000000  ...  -0.108197       0.105623
AveRooms      0.326895 -0.153277  ...  -0.027540       0.151948
AveBedrms    -0.062040 -0.077747  ...   0.013344      -0.046701
Population    0.004834 -0.296244  ...   0.099773      -0.024650
AveOccup      0.018766  0.013191  ...   0.002476      -0.023737
Latitude     -0.079809  0.011173  ...  -0.924664      -0.144160
Longitude    -0.015176 -0.108197  ...   1.000000      -0.045967
MedHouseValue 0.688075  0.105623  ...  -0.045967       1.000000
```

La caractéristique 'MedInc' montre une forte corrélation positive avec 'MedHouseValue', ce qui suggère que le revenu médian est un bon prédicteur du prix des maisons.



On remarque que la corrélation positive est faible entre la caractéristique 'HouseAge' et 'MedHouseValue', ce qui pourrait indiquer que l'âge de la maison n'est pas un prédicteur dominant du prix des maisons dans ce jeu de données.



2. Prétraitement des données

Le jeu de données "California Housing" ayant été vérifié et ne présentant aucune valeur manquante, il n'a pas été nécessaire d'appliquer des méthodes de traitement de valeurs manquantes.

Normalisation des données :

Afin de standardiser les données et de permettre une comparaison équitable entre les caractéristiques, un objet StandardScaler a été utilisé pour ajuster et transformer les données.

Sauvegarde des données :

Après la normalisation, le DataFrame a été enregistré dans un fichier CSV pour faciliter l'accès et la réutilisation lors des phases de modélisation.

3. Mise au point de modèles de régression et mesure des performances des modèles

Séparation des données :

Les données ont été divisées en un ensemble d'entraînement et un ensemble de test, avec 20% des données réservées pour le test.

Entraînement des modèles :

Trois modèles de régression ont été initialisés et entraînés sur l'ensemble d'entraînement :

- LinearRegression
- Lasso
- RandomForestRegressor

Évaluation des modèles :

Les modèles ont été évalués sur l'ensemble de test pour déterminer leur coefficient de détermination (R^2).

Voici les résultats obtenus :

- Coefficient de détermination pour LinearRegression : 57.67%
- Coefficient de détermination pour Lasso : 28.89%
- Coefficient de détermination pour RandomForestRegressor : 81.7%

Ces résultats indiquent que le modèle RandomForestRegressor a surpassé les autres modèles en termes de capacité de prédiction sur l'ensemble de test.

4. Amélioration de la performance d'un modèle

Après avoir évalué les performances initiales des modèles de régression, l'étape suivante a consisté à affiner le modèle le plus prometteur, le RandomForestRegressor, en utilisant la recherche par grille pour optimiser ses hyperparamètres.

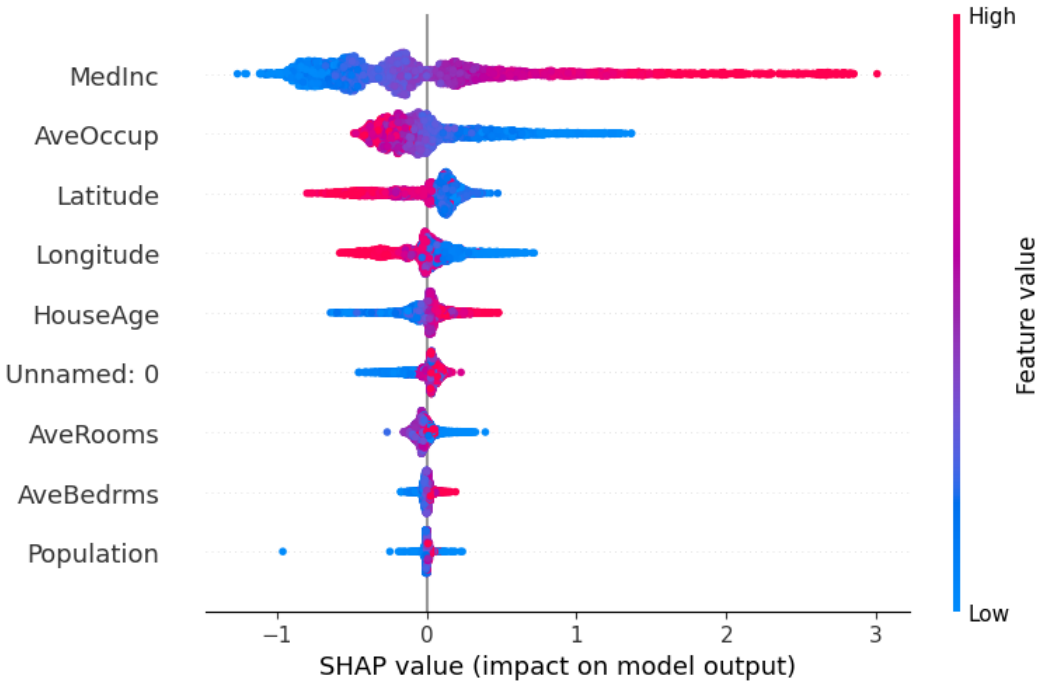
La fonction GridSearchCV de scikit-learn a été employée pour automatiser la recherche des meilleurs hyperparamètres pour le modèle RandomForestRegressor. La grille de recherche comprenait des variations du nombre d'estimateurs, de la profondeur maximale, du nombre minimum d'échantillons pour diviser un nœud, du nombre minimum d'échantillons requis à chaque feuille et de l'utilisation du bootstrap.

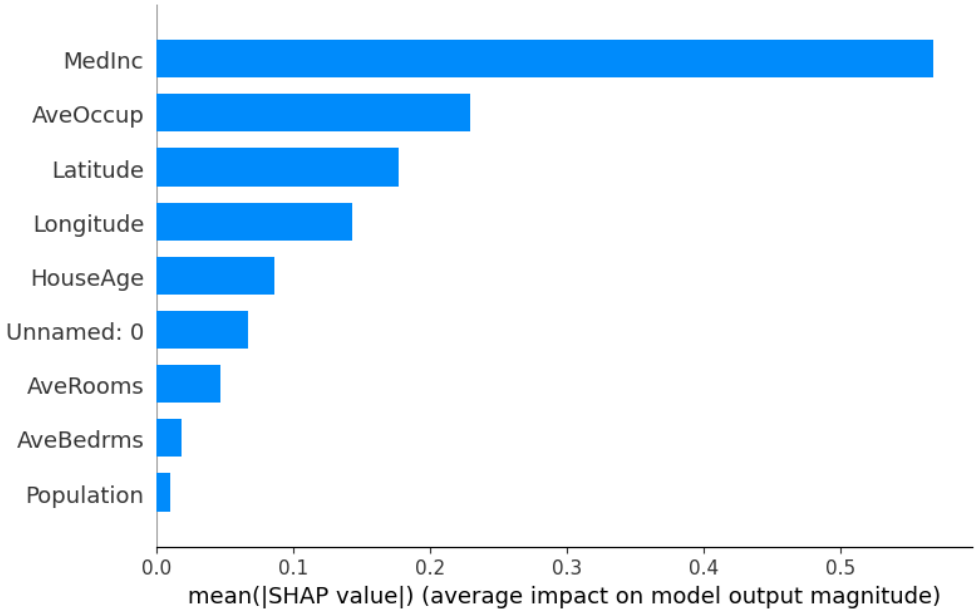
Avec les hyperparamètres optimisés, le modèle a été réévalué sur l'ensemble de test, résultant en un coefficient de détermination (R^2) de 81.83%

```
Coefficient de détermination: 0.8183058515177228
```

5. Analyse de l'importance des caractéristiques dans la prédiction des prix des maisons

L'analyse SHAP a révélé que certaines caractéristiques ont un impact plus significatif sur la prédiction des prix des maisons.





- Revenu Médian (MedInc) : Cette caractéristique a le plus grand impact sur les prédictions. Un revenu médian plus élevé est généralement associé à des valeurs immobilières plus élevées.
- Occupation Moyenne (AveOccup) : L'impact de l'occupation moyenne sur les prix des maisons est également notable, bien que moins important que le revenu médian.
- Latitude et Longitude : Ces caractéristiques géographiques influencent les prédictions, reflétant l'importance de l'emplacement dans la détermination des prix des maisons.
- Âge Moyen des Maisons (HouseAge) : L'âge des maisons a un impact modéré sur les prédictions, ce qui peut indiquer que les acheteurs valorisent certaines qualités des constructions plus anciennes ou plus récentes.
- Taille Moyenne des Pièces (AveRooms) : La taille des pièces affecte les prédictions dans une certaine mesure, suggérant que des pièces plus grandes peuvent augmenter la valeur des maisons.
- Taille Moyenne des Chambres (AveBedrms) et Population : Ces caractéristiques ont le moins d'impact sur les prédictions parmi celles analysées.

En résumé, le modèle indique que le revenu médian est le facteur le plus déterminant pour les prix des maisons, tandis que d'autres caractéristiques comme l'emplacement et la taille des maisons jouent également un rôle important, mais dans une moindre mesure.