Institut National de Statistique et d'Economie Appliquée - Rabat

Projet de Fin d'Année

Prédiction du prix des logements

Élèves :
Soufiane MADHI
Imad LAKBABTI

Encadrant:
Badreddine BENYACOUB

juin 2024

Remerciements

Nous tenons à exprimer notre profonde gratitude au corps professoral de l'Institut National de Statistique et d'Économie Appliquée (INSEA), dont le dévouement et l'engagement dans l'enseignement et la recherche ont été des sources inestimables de savoir et de motivation tout au long de ce projet. Leur expertise et leur soutien constant ont été des piliers essentiels à la réalisation de ce travail.

Nous souhaitons remercier tout particulièrement M. Badreddine BENYACOUB, coordinateur de la filière Data Science, pour son encadrement exceptionnel, ses conseils avisés, et sa disponibilité sans faille. M. BENYACOUB a su nous guider avec perspicacité et nous inspirer une passion pour l'analyse des données et la modélisation statistique. Son encouragement à explorer de nouvelles idées et à adopter des approches innovantes a largement contribué à la qualité et au succès de ce projet.

Nous sommes également reconnaissants envers tous les enseignants et collègues de l'INSEA, qui ont partagé leurs connaissances et leur expérience, et qui ont favorisé un environnement d'apprentissage stimulant et enrichissant. Leur soutien moral et intellectuel a été d'une grande aide pour surmonter les défis rencontrés au cours de ce projet.

Enfin, nous remercions nos familles et amis pour leur soutien indéfectible et leur encouragement tout au long de cette aventure académique. Leur patience et leur compréhension ont été des moteurs précieux pour mener à bien ce travail.

À tous, merci pour votre soutien et votre confiance, qui ont été les clés de l'accomplissement de ce projet.

Imad LAKBABTI
Soufiane MADHI

Résumé

Ce projet vise à prédire la valeur des maisons dans les banlieues de Boston à l'aide de modèles de régression, en utilisant des données socio-économiques et géographiques. Après une analyse exploratoire des données pour comprendre les caractéristiques et les tendances, nous avons sélectionné les variables les plus pertinentes pour la prédiction.

Nous avons testé et comparé plusieurs modèles, dont des forêts aléatoires, des ExtraTrees et du Gradient Boosting, en optimisant leurs hyperparamètres pour améliorer leur précision. Les résultats montrent que les modèles basés sur des techniques d'ensemble, comme les forêts aléatoires et les ExtraTrees, offrent les meilleures performances pour prédire les valeurs des maisons. Ces modèles peuvent être utilisés pour des analyses immobilières et des prises de décision informées.

Introduction générale

Dans ce projet, nous visons à prédire les prix des logements dans les banlieues de Boston en utilisant un ensemble de données bien connu de l'étude de 1978 par Harrison et Rubinfeld sur les prix hédoniques. Cet ensemble de données, communément appelé le Boston Housing Dataset, fournit une collection complète de caractéristiques pouvant être utilisées pour modéliser et prédire la valeur médiane des maisons occupées par leurs propriétaires (medv) dans diverses villes. L'objectif principal est de développer un modèle prédictif qui estime avec précision les valeurs médianes des maisons dans ces banlieues, en utilisant 'medv' comme variable cible, représentant la valeur médiane des maisons occupées par leurs propriétaires en milliers de dollars. Le Boston Housing Dataset comprend 506 observations et 14 attributs couvrant divers facteurs socio-économiques et environnementaux susceptibles d'influencer les prix des logements. Les données proviennent des études de Harrison et Rubinfeld (1978) et de Belsley, Kuh, et Welsch (1980). Notre approche pour prédire les prix des logements impliquera plusieurs étapes : le prétraitement des données (nettoyage, gestion des valeurs manquantes, encodage des variables catégorielles et normalisation des caractéristiques numériques), l'analyse exploratoire des données (compréhension de la distribution, identification des motifs et visualisation des relations entre les variables), la sélection des caractéristiques les plus pertinentes, le développement et l'entraînement des modèles de régression (en utilisant divers algorithmes comme la régression linéaire, les arbres de décision, la forêt aléatoire, etc.), l'évaluation des modèles (à l'aide de métriques telles que l'erreur absolue moyenne, l'erreur quadratique moyenne et le coefficient de détermination), et enfin, l'optimisation et l'ajustement des paramètres du modèle pour améliorer précision et robustesse. La réussite de ce projet fournira des informations précieuses sur les facteurs influençant les prix des logements dans les banlieues de Boston et démontrera l'application de la modélisation prédictive dans l'évaluation immobilière. En utilisant des techniques statistiques et d'apprentissage automatique, nous visons à développer un modèle fiable qui peut aider les parties prenantes à prendre des décisions éclairées en matière d'investissements immobiliers et de planification urbaine.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Analyse exploratoire des données EDA

L'analyse exploratoire des données (Exploratory Data Analysis, EDA) constitue une étape cruciale dans tout projet d'analyse de données ou de modélisation prédictive. Dans le cadre du projet "Prédiction du prix des logements", l'EDA joue un rôle fondamental pour plusieurs raisons spécifiques et se concentre sur des objectifs clairs visant à maximiser la compréhension des données et à préparer efficacement les étapes suivantes de l'analyse.

1.1 Aperçu du Jeu de Données

Le Boston Housing Dataset comprend 506 observations et 14 attributs. Ces attributs encapsulent une gamme de facteurs socio-économiques et environnementaux qui peuvent influencer les prix des logements. Voici une brève description de chaque variable incluse dans l'ensemble de données :

- **crim** : Taux de criminalité par habitant par ville.
- **zn** : Proportion de terrains résidentiels zonés pour des lots de plus de 25 000 pieds carrés.
- **indus** : Proportion d'acres d'activités non commerciales par ville.
- **chas** : Variable fictive Charles River (1 si le secteur borde la rivière ; 0 sinon).
- **nox** : Concentration en oxydes d'azote (parties par 10 millions).
- **rm** : Nombre moyen de pièces par logement.
- **age** : Proportion de logements occupés par leurs propriétaires construits avant 1940.
- **dis** : Moyenne pondérée des distances aux cinq centres d'emploi de Boston.
- **rad** : Indice d'accessibilité aux autoroutes radiales.
- **tax** : Taux d'imposition foncière à pleine valeur par $10,000.
- **ptratio** : Ratio élèves-enseignant par ville.
- **black** : $1000(Bk - 0.63)^2$ où Bk est la proportion de noirs par ville.
- **lstat** : Pourcentage de la population à statut socio-économique faible.
- **medv** : Valeur médiane des maisons occupées par leurs propriétaires en milliers de dollars.

1.2 Analyse Statistique Descriptive

Les statistiques descriptives des variables du jeu de données d'entraînement sont présentées dans le tableau ci-dessous. Ces statistiques fournissent une vue d'ensemble des caractéristiques des données, incluant la moyenne, la médiane, l'écart-type, les valeurs minimales et maximales, ainsi que les quartiles.

TABLE 1.1 – Statistiques descriptives des variables du jeu de données d'entraînement

	ID	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
count	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000	333.000000
mean	250.951952	3.360341	10.689189	11.293483	0.060060	0.557144	6.265619	68.226426	3.709934	9.633634	409.279279	18.448048	359.466096	12.515435	22.768769
std	147.859438	7.352272	22.674762	6.998123	0.237956	0.114955	0.703952	28.133344	1.981123	8.742174	170.841988	2.151821	86.584567	7.067781	9.173468
min	1.000000	0.006320	0.000000	0.740000	0.000000	0.385000	3.561000	6.000000	1.129600	1.000000	188.000000	12.600000	3.500000	1.730000	5.000000
255075max	506.000000	73.534100	100.000000	27.740000	1.000000	0.871000	8.725000	100.000000	10.710300	24.000000	711.000000	21.200000	396.900000	37.970000	50.000000

TABLE 1.2 – Valeurs nulles dans le dataset d'entraînement

	0
ID	False
crim	False
zn	False
indus	False
chas	False
nox	False
rm	False
age	False
dis	False
rad	False
tax	False
ptratio	False
black	False
lstat	False
medv	False

Les statistiques descriptives révèlent plusieurs points importants :

— **Taux de Criminalité (crim)** : La moyenne est de 3.36, mais avec une large gamme allant de 0.006 à 73.534, indiquant une grande variabilité dans les taux de criminalité parmi les localités.
— **Proportion de Terrains Résidentiels (zn)** : La médiane est à 0, ce qui signifie que dans au moins la moitié des localités, il n'y a pas de terrains résidentiels de grande taille (plus de 25,000 sq.ft).
— **Concentration en Oxydes d'Azote (nox)** : Avec une moyenne de 0.56 et une faible écart-type, la concentration en polluants est relativement uniforme à travers les localités.
— **Nombre Moyen de Pièces (rm)** : La moyenne est de 6.27 pièces par logement, avec une gamme allant de 3.561 à 8.725, suggérant une diversité significative dans la taille des logements.
— **Valeur Médiane des Maisons (medv)** : La médiane est de 21,600 $ avec une dispersion significative, allant de 5,000 $ à 50,000 $, indiquant des différences importantes dans les valeurs immobilières parmi les banlieues de Boston.

Cette section offre une vue d'ensemble essentielle qui servira de base pour les analyses plus détaillées à venir, telles que l'évaluation des relations entre les variables et la construction de modèles prédictifs pour estimer la valeur des maisons.

1.3 Distribution de la variable cible

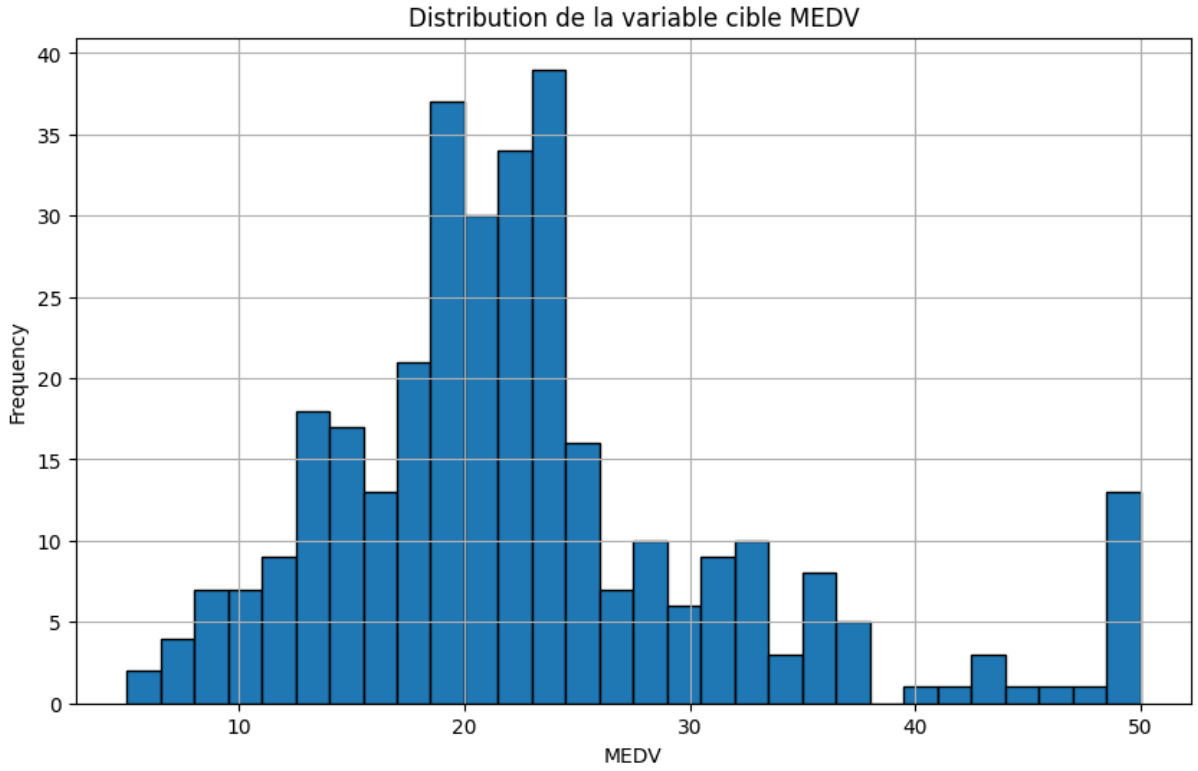

FIGURE 1.1 – Distribution de la variable cible MEDV

L'histogramme de la Figure 1.1 montre la distribution de la variable cible `MEDV`, qui représente la valeur médiane des maisons (en milliers de dollars) dans les banlieues de Boston. Cette analyse de la distribution révèle plusieurs caractéristiques clés :

— **Concentration des Valeurs** : La majorité des valeurs de `MEDV` se situent entre 15 000 et 30 000 dollars, avec une concentration particulièrement élevée autour de 20 000 à 25 000 dollars. Cela indique que la plupart des maisons dans les banlieues de Boston ont une valeur médiane dans cette fourchette de prix, ce qui suggère une prépondérance de quartiers de classe moyenne.

— **Distribution Bimodale** : La distribution montre deux pics principaux. Le premier, le plus significatif, est centré autour de 20 000 à 25 000 dollars, tandis que le second, plus petit mais distinct, se trouve à 50 000 dollars. Cette bimodalité pourrait indiquer la présence de deux groupes distincts de propriétés : des maisons abordables dans des quartiers standards et des maisons de valeur maximale dans des quartiers plus exclusifs ou soumis à une censure des données.

— **Pic à la Valeur Maximale** : Un nombre notable de valeurs se situent à la limite supérieure de 50 000 dollars. Ce pic suggère une limite artificielle ou une censure dans les données, indiquant que plusieurs maisons ont été évaluées au plafond de la valeur maximale enregistrée. Cela peut refléter une évaluation plafonnée des propriétés les plus chères ou des restrictions dans la collecte des données.

— **Asymétrie de la Distribution** : La distribution présente une légère asymétrie positive, avec une queue plus longue à droite. Cela indique qu'il y a moins de maisons avec des valeurs très élevées comparées à la concentration de valeurs modérées. Cette asymétrie peut être attribuée à la présence de quelques propriétés de luxe ayant des valeurs exceptionnellement élevées par rapport à la majorité des maisons.

— **Implications des Valeurs Élevées et Basses** : La présence de valeurs élevées (proche de 50 000 dollars) peut refléter des maisons situées dans des quartiers très prisés ou dotées d'aménagements de haute qualité. Inversement, les valeurs plus basses suggèrent des maisons dans des quartiers moins chers ou avec moins d'attractivité économique.

— **Variabilité des Prix** : L'écart important des valeurs de `MEDV` et la large gamme de prix indiquent une variabilité significative dans le marché immobilier de Boston. Cette diversité peut être influencée par divers facteurs tels que la qualité de l'environnement, l'accès aux services, et les caractéristiques socio-économiques des différents quartiers.

Cette analyse fournit une vue d'ensemble précieuse des valeurs des maisons dans les banlieues de Boston, mettant en lumière des aspects importants de la distribution de `MEDV` et des tendances sous-jacentes du marché immobilier local.

1.3.1 comparison à une distrubution normal

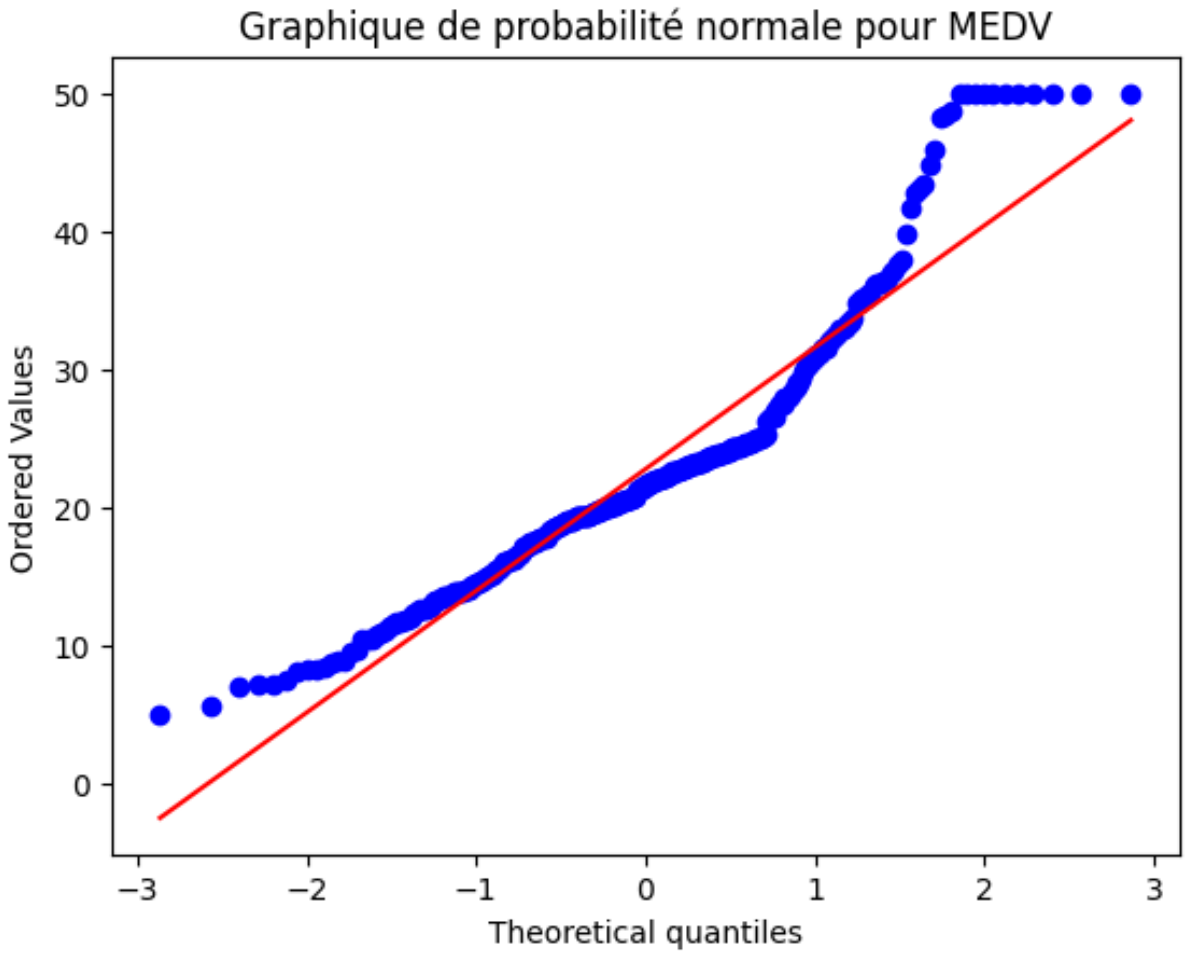

FIGURE 1.2 – Graphique de probabilité normale pour la variable `MEDV`

Le graphique de la Figure 1.2 compare la distribution des valeurs de `MEDV` avec une distribution normale théorique. Ce graphique, aussi appelé Q-Q plot, permet de visualiser les écarts par rapport à la normalité de la variable étudiée.

- **Alignement avec la Ligne Théorique** : La ligne rouge représente une distribution normale théorique. Les points qui suivent cette ligne indiquent une correspondance avec la distribution normale pour la partie centrale des données. Les valeurs médianes des maisons semblent suivre approximativement une distribution normale dans cette gamme.
- **Comportement aux Extrémités** : Aux extrémités, les points dévient de la ligne droite, indiquant une divergence par rapport à la normalité. À gauche, les valeurs sont plus faibles que prévu, tandis qu'à droite, les valeurs sont plus élevées que prévu. Cela suggère la présence de valeurs extrêmes ou des queues longues dans la distribution.
- **Présence de Censure ou de Limite** : La concentration de points à la valeur maximale de `MEDV` (50 000 dollars) indique une possible censure dans les données,

où les valeurs au-delà de cette limite ne sont pas enregistrées ou sont très fréquentes.

— **Asymétrie et Skewness** : La courbure des points vers le bas à gauche et vers le haut à droite indique une asymétrie positive de la distribution. Cela signifie que les valeurs élevées sont plus fréquentes ou plus élevées que celles attendues dans une distribution normale, ce qui est typique d'une queue longue à droite.

— **Implications pour l'Analyse Statistique** : Cette déviation par rapport à la normalité suggère que des méthodes statistiques basées sur l'hypothèse de normalité pourraient nécessiter des ajustements, comme une transformation des données pour améliorer leur normalité.

En conclusion, le graphique de probabilité normale pour `MEDV` révèle une distribution qui diffère de la normalité, avec des asymétries et des valeurs extrêmes fréquentes, ce qui peut influencer les analyses statistiques ultérieures et la modélisation des données.

1.4 Analyse de corrélation

1.4.1 Matrice de Corrélation

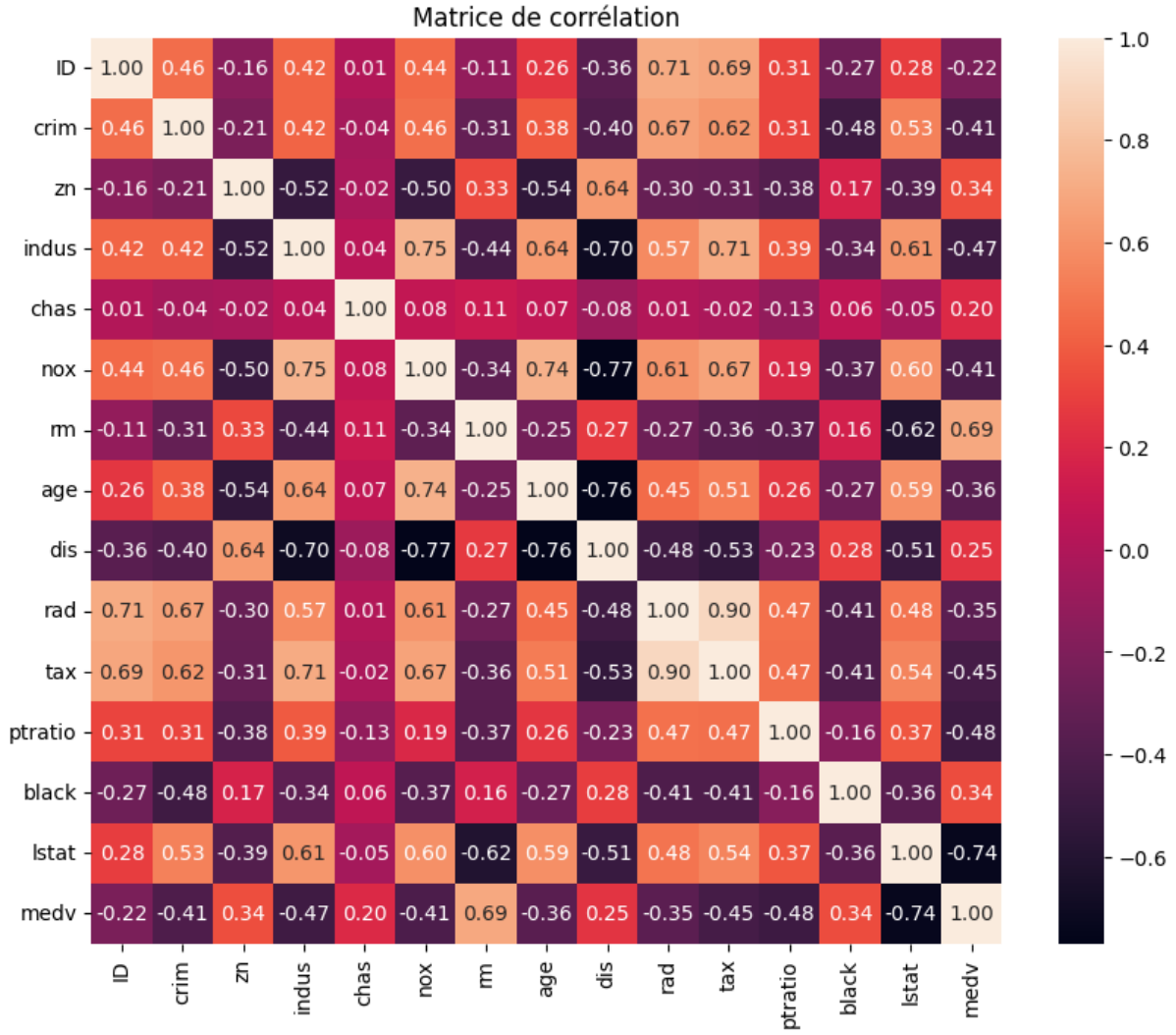

FIGURE 1.3 – Matrice de corrélation des variables du jeu de données Boston Housing

La Figure 1.3 montre la matrice de corrélation pour les différentes variables du jeu de données Boston Housing. Cette matrice permet de visualiser les relations linéaires entre les variables, avec des coefficients de corrélation allant de -1 à 1.

— **Corrélations avec la Variable Cible** `MEDV` :
  — **Variables Positivement Corrélées avec** `MEDV` : Les variables `rm` (0.69) et `zn` (0.34) montrent des corrélations positives significatives avec `MEDV`, suggérant que des maisons avec plus de pièces et situées dans des zones résidentielles sont généralement plus chères.
  — **Variables Négativement Corrélées avec** `MEDV` : Les variables `lstat` (-0.74), `ptratio` (-0.51), `tax` (-0.45), `nox` (-0.41), et `age` (-0.36) montrent des corréla-

tions négatives, indiquant que des valeurs plus élevées de ces variables sont associées à des valeurs des maisons plus faibles.

- **Corrélations Entre Variables Indépendantes** :
  - **Corrélations Fortes Positives** : `tax` et `rad` (0.91) montrent une forte corrélation positive, indiquant que les zones avec un meilleur accès aux autoroutes ont des taux d'imposition plus élevés.
  - **Corrélations Fortes Négatives** : `dis` et `nox` (-0.77) montrent une forte corrélation négative, suggérant que les maisons plus éloignées des centres d'emploi sont situées dans des zones moins polluées.
- **Implications Importantes** :
  - Les indicateurs socio-économiques, tels que `lstat` et `ptratio`, montrent des relations négatives significatives avec `MEDV`, soulignant l'impact de la qualité de vie et de l'éducation sur la valeur des maisons.
  - La pollution (`nox`) et l'industrialisation (`indus`) sont corrélées négativement avec la valeur des maisons, ce qui indique une préférence pour des zones moins polluées et moins industrialisées.
  - La variable `age` montre que les maisons dans des quartiers avec une proportion plus élevée de logements anciens tendent à avoir des valeurs plus basses.
  - L'accessibilité (`dis`, `rad`) montre des corrélations complexes avec la valeur des maisons et d'autres variables, reflétant l'importance de l'emplacement dans la détermination des valeurs immobilières.

En conclusion, la matrice de corrélation met en évidence des relations significatives entre certaines variables et la valeur médiane des maisons, ce qui peut guider l'analyse future pour comprendre les facteurs influençant la valeur des propriétés dans cette région.

1.4.2 Graphique de paires des variables clés

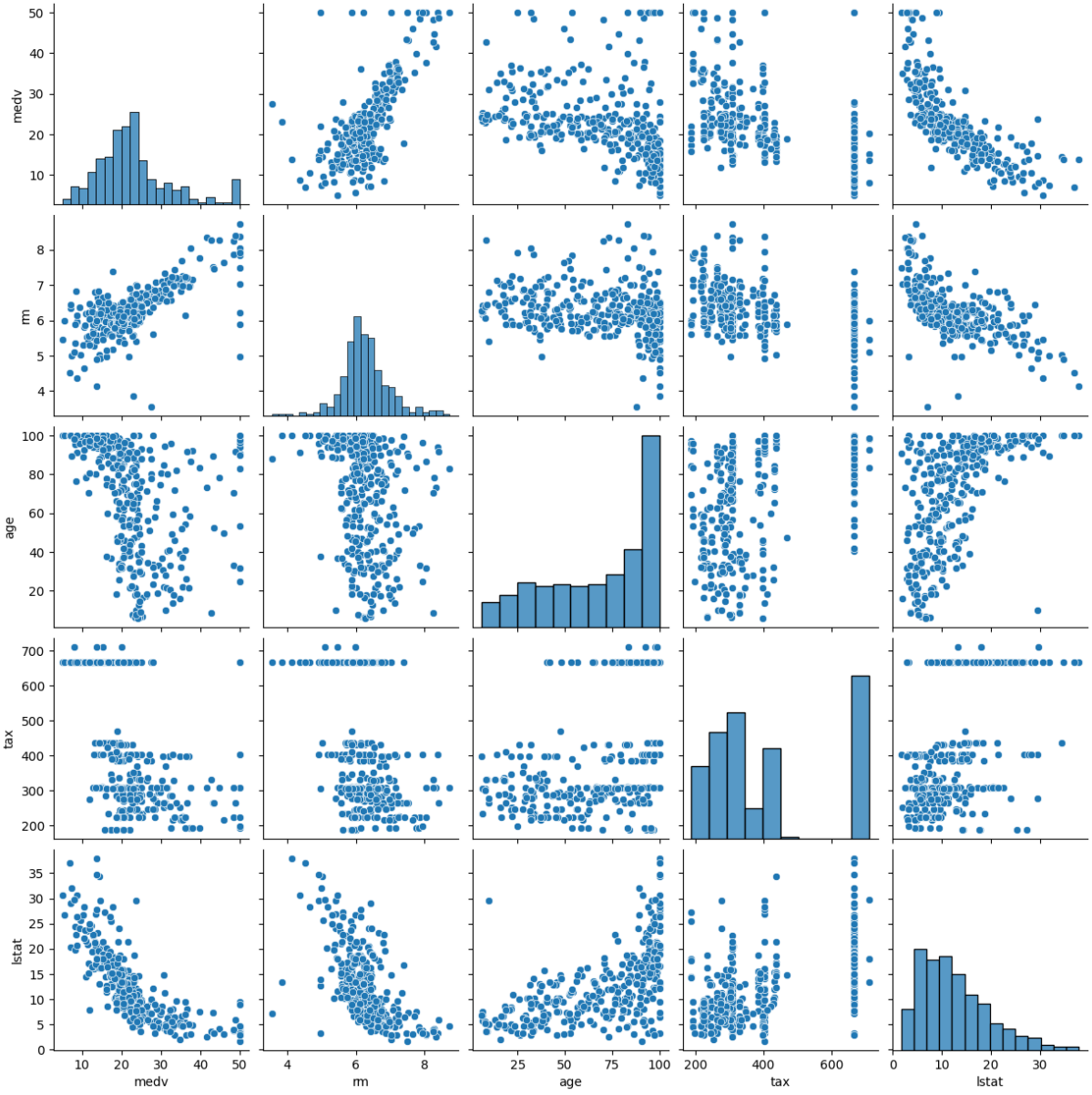

FIGURE 1.4 – Graphique de paires des variables clés du jeu de données Boston Housing

La Figure 1.4 montre un graphique de paires pour les principales variables du jeu de données Boston Housing, permettant de visualiser les relations bivariées entre les variables. Voici une analyse des principales observations :

— **Relation entre `medv` et Autres Variables** :

  — `medv` vs `rm` : Une relation positive et linéaire est observée, suggérant que des maisons avec plus de pièces tendent à avoir une valeur médiane plus élevée.

  — `medv` vs `age` : Une relation négative non linéaire montre que des quartiers avec une proportion plus élevée de logements anciens tendent à avoir des valeurs des maisons plus faibles.

  - `medv` vs `tax` : La relation semble légèrement négative, indiquant que des taux d'imposition plus élevés sont associés à des valeurs des maisons plus faibles.
  - `medv` vs `lstat` : Une relation négative claire, où une proportion plus élevée de population à faible statut socio-économique est associée à des valeurs des maisons plus basses.

- **Distribution des Variables** :
  - `medv` : Distribution légèrement asymétrique avec un pic autour de 20 000 à 25 000 dollars et une queue longue vers les valeurs plus élevées.
  - `rm` : Distribution presque symétrique, centrée autour de 6 à 7 pièces par logement.
  - `age` : Distribution montrant une concentration de maisons anciennes, avec un grand nombre autour de 80 à 100
  - `tax` : Distribution bimodale avec des pics autour de 200-300 et de 600-700.
  - `lstat` : Distribution asymétrique avec la majorité des valeurs entre 0 et 20

- **Relations entre Autres Variables** :
  - `rm` vs `age` : Relation négative non linéaire, suggérant que les maisons avec plus de pièces tendent à être dans des quartiers avec des logements plus récents.
  - `rm` vs `tax` : Relation diffuse sans tendance claire, indiquant que le nombre de pièces par logement n'est pas directement lié au taux d'imposition foncière.
  - `rm` vs `lstat` : Relation négative, montrant que les logements avec plus de pièces sont souvent dans des quartiers avec une proportion plus faible de population à faible statut socio-économique.
  - `age` vs `tax` : Relation légèrement positive, suggérant que les quartiers avec des logements plus anciens peuvent avoir des taux d'imposition plus élevés.
  - `age` vs `lstat` : Relation positive, indiquant que les quartiers avec des logements plus anciens ont tendance à avoir une proportion plus élevée de population à faible statut.
  - `tax` vs `lstat` : Relation positive, montrant que des taux d'imposition plus élevés sont associés à des quartiers avec une proportion plus élevée de population à faible statut.

En conclusion, ce graphique de paires révèle des relations significatives entre la valeur médiane des maisons (`medv`) et plusieurs caractéristiques socio-économiques et démographiques des quartiers, ce qui est essentiel pour comprendre les dynamiques du marché immobilier de Boston.

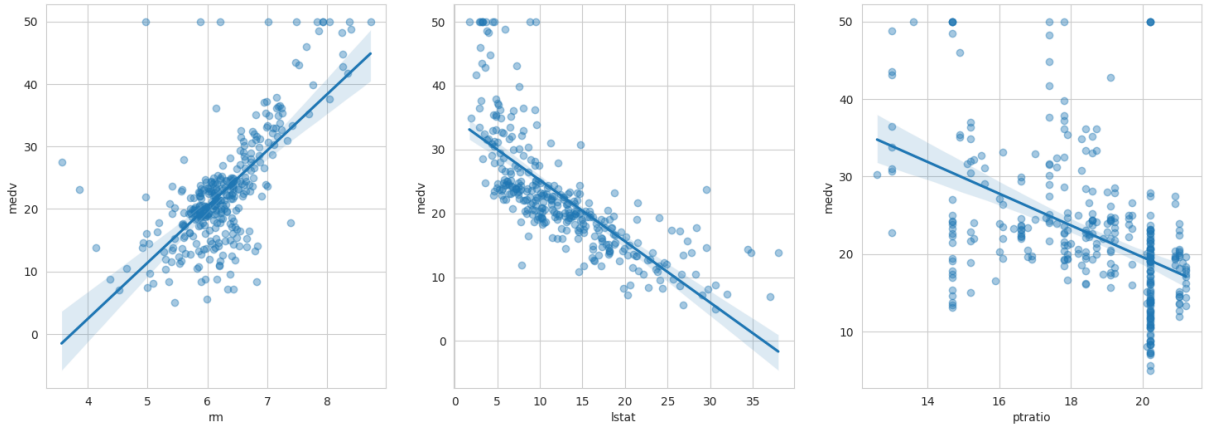

FIGURE 1.5 – Graphiques de régression entre `medv` et les variables `rm`, `lstat`, et `ptratio`

La Figure 1.5 présente des graphiques de régression qui montrent les relations entre la valeur médiane des maisons (`medv`) et trois variables explicatives clés : le nombre moyen de pièces par logement (`rm`), le pourcentage de population à faible statut socio-économique (`lstat`), et le ratio élèves/enseignant (`ptratio`). Voici une analyse de ces relations :

- **Relation entre `rm` et `medv` :**
  - La relation est clairement positive, indiquant que les maisons avec un plus grand nombre de pièces tendent à avoir des valeurs médianes plus élevées.
  - La bande de confiance est étroite, ce qui suggère que les prédictions de la valeur des maisons en fonction du nombre de pièces sont relativement fiables, malgré la présence de quelques valeurs aberrantes.
- **Relation entre `lstat` et `medv` :**
  - La relation est négative, montrant qu'une augmentation du pourcentage de population à faible statut socio-économique est associée à une diminution de la valeur médiane des maisons.
  - La bande de confiance est plus large pour les valeurs élevées de `lstat`, ce qui indique une incertitude accrue dans ces prédictions.
- **Relation entre `ptratio` et `medv` :**
  - La relation est également négative, suggérant qu'un ratio élèves/enseignant plus élevé est associé à des valeurs médianes des maisons plus basses.
  - La bande de confiance large indique une variabilité importante, ce qui suggère que `ptratio` seul peut ne pas être suffisant pour prédire avec précision la valeur des maisons.

Ces graphiques de régression montrent des relations significatives entre la valeur médiane des maisons (`medv`) et des facteurs socio-économiques et démographiques, suggérant que ces variables sont des indicateurs importants pour prédire la valeur des maisons dans cette région.

1.4.3 Graphiques de Résidus pour `medv`

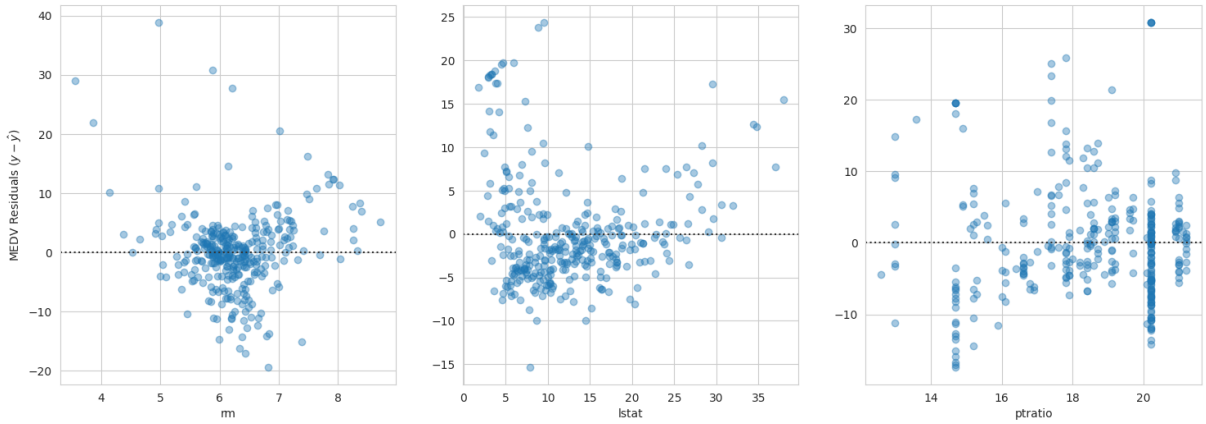

FIGURE 1.6 – Graphiques de résidus pour `medv` par rapport à `rm`, `lstat`, et `ptratio`

La Figure 1.6 présente des graphiques de résidus pour la valeur médiane des maisons (`medv`) par rapport à trois variables explicatives clés : le nombre moyen de pièces par logement (`rm`), le pourcentage de population à faible statut socio-économique (`lstat`), et le ratio élèves/enseignant (`ptratio`). Ces graphiques permettent de visualiser les erreurs de prédiction du modèle de régression et d'évaluer la qualité de ce modèle.

- **Résidus pour `rm` vs. `medv` :**
  - Les résidus sont principalement concentrés autour de zéro, suggérant une bonne capacité prédictive du modèle pour cette variable. Il existe cependant quelques valeurs aberrantes qui pourraient indiquer des erreurs de modélisation.
  - La dispersion des résidus autour de la ligne zéro est uniforme, ce qui indique une bonne homogénéité des variances sans signes évidents d'hétéroscédasticité.
- **Résidus pour `lstat` vs. `medv` :**
  - Les résidus montrent une dispersion accrue pour les valeurs élevées de `lstat`, ce qui suggère que le modèle ne capture pas parfaitement la relation entre `lstat` et `medv` pour ces observations.
  - La tendance à une hétéroscédasticité est visible, indiquant que la variance des erreurs augmente avec la valeur de `lstat`, ce qui peut affecter la fiabilité des prédictions pour cette variable.
- **Résidus pour `ptratio` vs. `medv` :**
  - La majorité des résidus sont concentrés autour de zéro, mais une légère tendance à des résidus plus dispersés pour des valeurs élevées de `ptratio` est observée.
  - La dispersion des résidus est plus uniforme que pour `lstat`, indiquant une moindre hétéroscédasticité, bien qu'une certaine variabilité reste présente.

Ces graphiques de résidus montrent que, bien que les modèles de régression capturent en grande partie les relations avec `medv`, il existe des indications d'hétéroscédasticité, particulièrement pour `lstat`. Cela suggère qu'il pourrait être bénéfique d'explorer des transformations de données ou des modèles plus complexes pour améliorer la qualité des prédictions.

1.5 Analyse des Valeurs Aberrantes

1.5.1 Détection des Valeurs Aberrantes

La détection des valeurs aberrantes est une étape cruciale dans l'analyse des données car elle permet d'identifier des observations qui s'écartent significativement des autres valeurs de l'ensemble de données. Ces valeurs peuvent fausser les résultats des analyses statistiques et des modèles prédictifs. Pour ce projet, nous avons utilisé la méthode de l'intervalle interquartile (IQR) pour détecter les valeurs aberrantes dans chaque variable du jeu de données Boston Housing.

L'intervalle interquartile est défini comme la différence entre le troisième quartile (Q3) et le premier quartile (Q1). Une valeur est considérée comme une valeur aberrante si elle est inférieure à Q1 - 1.5 * IQR ou supérieure à Q3 + 1.5 * IQR.

Table 1.3 – Pourcentage de Valeurs Aberrantes par Colonne

Colonne	Valeurs Aberrantes (%)
ID	0.00
crim	12.61
zn	12.31
indus	0.00
chas	100.00
nox	0.00
rm	6.31
age	0.00
dis	0.30
rad	0.00
tax	0.00
ptratio	2.70
black	14.11
lstat	2.10
medv	8.41

1.5.2 Interprétation des Résultats de la Détection des Valeurs Aberrantes

L'analyse des valeurs aberrantes révèle plusieurs points clés :

— **Colonne** `chas` : Toutes les valeurs de cette colonne sont considérées comme aberrantes (100%). Cela est dû à la nature binaire de cette variable, qui prend la valeur 0 ou 1 pour indiquer si la zone est proche de la rivière Charles. En raison de sa nature discrète, cette variable est intrinsèquement en dehors des critères de l'IQR utilisés pour la détection des valeurs aberrantes.

— **Colonne** `crim` : 12.61% des valeurs sont considérées comme aberrantes, ce qui reflète la grande variabilité des taux de criminalité par habitant dans les différentes localités.

— **Colonne** `zn` : 12.31% des valeurs sont aberrantes, indiquant que certaines zones ont des proportions extrêmement élevées ou faibles de terrains résidentiels.

— **Colonne** `rm` : 6.31% des valeurs sont aberrantes, ce qui indique que certaines maisons ont un nombre de pièces nettement supérieur ou inférieur à la moyenne.
— **Colonne** `black` : 14.11% des valeurs sont aberrantes, suggérant une distribution large dans la proportion de la population noire par localité.
— **Colonne** `medv` : 8.41% des valeurs sont aberrantes. Cela est principalement dû à la présence d'un grand nombre de maisons avec une valeur médiane de 50 000 dollars, ce qui pourrait indiquer une censure dans les données.

1.5.3 Élimination des Valeurs Aberrantes pour `medv`

Pour améliorer la qualité de l'analyse des données et la robustesse des modèles prédictifs, nous avons décidé d'éliminer les valeurs aberrantes de la variable cible `medv` qui étaient égales ou supérieures à 50 000 dollars. Cette valeur élevée peut indiquer une censure ou une limite artificielle dans les données.

Après l'élimination des valeurs aberrantes, la taille du jeu de données a été réduite de 333 observations à 322 observations, ce qui représente une élimination de 3.3% des données. La Table 1.4 montre la forme du jeu de données avant et après l'élimination des valeurs aberrantes.

TABLE 1.4 – Forme du Jeu de Données Avant et Après l'Élimination des Valeurs Aberrantes

État	**Nombre d'Observations**
Avant Élimination	333
Après Élimination	322

Chapitre 2

Modélisation des Données

2.1 Sélection des Caractéristiques

2.1.1 Hypothèses de Sélection des Caractéristiques

La sélection des caractéristiques est une étape cruciale dans la construction d'un modèle prédictif efficace. En se basant sur les connaissances préalables et l'analyse exploratoire des données, les hypothèses suivantes ont été formulées :

- **Nombre de Pièces (RM)** : Nous supposons que les maisons avec plus de pièces (`RM`) ont une valeur plus élevée. Cela signifie qu'une augmentation de la valeur de `RM` entraîne une augmentation de la valeur de `MEDV`, la valeur médiane des maisons. Ces variables sont directement proportionnelles.
- **Statut Socio-économique (LSTAT)** : Les maisons situées dans des quartiers avec une proportion plus élevée de résidents à faible statut socio-économique (`LSTAT`) auront une valeur plus faible. En d'autres termes, plus la valeur de `LSTAT` est basse, plus la valeur de `MEDV` est élevée. Ces variables sont inversement proportionnelles.
- **Ratio Élèves/Enseignant (PTRATIO)** : Les maisons dans les quartiers avec un ratio élèves/enseignant plus élevé (`PTRATIO`) ont tendance à avoir une valeur plus faible. Une diminution de la valeur de `PTRATIO` est donc associée à une augmentation de la valeur de `MEDV`. Ces variables sont également inversement proportionnelles.

Notre objectif est de développer un modèle capable de prédire la valeur des maisons. Par conséquent, nous avons sélectionné les caractéristiques pertinentes (`RM`, `LSTAT`, et `PTRATIO`) et la variable cible `MEDV`. Les autres caractéristiques jugées non pertinentes ont été exclues de l'analyse.

2.1.2 Préparation des Données

Pour la sélection des caractéristiques, nous avons extrait les variables pertinentes (`RM`, `LSTAT`, et `PTRATIO`) du jeu de données, ainsi que la variable cible `MEDV`. Les autres caractéristiques ont été exclues.

Les premières lignes des données filtrées sont présentées dans le Tableau 2.1.

TABLE 2.1 – Données Filtrées avec les Caractéristiques Sélectionnées

ID	RM	PTRATIO	LSTAT	MEDV
1	6.575	15.3	4.98	24.0
2	6.421	17.8	9.14	21.6
4	6.998	18.7	2.94	33.4
5	7.147	18.7	5.33	36.2
7	6.012	15.2	12.43	22.9

2.1.3 Forme des Données Sélectionnées

Après la sélection des caractéristiques, nous avons examiné les formes des données pour nous assurer de leur alignement correct. La forme des jeux de données est la suivante :

— **Forme des Caractéristiques (`features`)** : (333, 4)
— **Forme de la Variable Cible (`prices`)** : (333,)

Ces formes indiquent que nous avons 333 observations et 4 caractéristiques (y compris l'ID, qui sera utilisé pour le suivi) pour les prédictions de la variable cible `MEDV`. Cette vérification est essentielle pour garantir que les données sont prêtes pour l'analyse et la modélisation ultérieures.

2.2 Standardisation des Caractéristiques

2.2.1 Introduction à la Standardisation

La standardisation est une étape clé dans le prétraitement des données, surtout lorsque les caractéristiques des données ont des échelles différentes. Elle permet de centrer les données autour de zéro et de les mettre à la même échelle en utilisant leur moyenne et leur écart-type. Cette technique est particulièrement utile pour les algorithmes de machine learning qui sont sensibles à l'échelle des données, tels que la régression linéaire et les méthodes basées sur la distance.

Pour ce projet, nous avons standardisé les caractéristiques sélectionnées pour s'assurer qu'elles contribuent de manière égale à l'analyse et à la modélisation. La standardisation est réalisée en soustrayant la moyenne de chaque caractéristique et en divisant par son écart-type, ce qui donne des caractéristiques ayant une moyenne de zéro et un écart-type de un.

2.2.2 Processus de Standardisation

Nous avons utilisé la méthode `StandardScaler` de la bibliothèque `scikit-learn` pour standardiser les caractéristiques. Cette méthode ajuste et transforme les données de manière à ce que chaque caractéristique ait une moyenne de zéro et un écart-type de un. Les étapes suivies sont les suivantes :

1. Création d'un objet `StandardScaler`.
2. Ajustement de l'objet aux caractéristiques (`features`) et transformation de celles-ci.

3. Vérification de la moyenne et de l'écart-type des caractéristiques standardisées pour s'assurer de leur centrage et mise à l'échelle corrects.

Le code utilisé pour la standardisation est présenté ci-dessous :

```
# Création d'un objet StandardScaler
scaler = StandardScaler()

# Ajustement et transformation des caractéristiques
scaled_X = scaler.fit_transform(X)

# Vérification de la moyenne et de l'écart-type
print("Mean of scaled features:", scaled_X.mean(axis=0))
print("Standard deviation of scaled features:", scaled_X.std(axis=0))
```

2.2.3 Résultats de la Standardisation

Les résultats de la standardisation montrent que les caractéristiques ont été correctement centrées autour de zéro et mises à l'échelle pour avoir un écart-type de un. Les résultats sont résumés dans le Tableau 2.2.

TABLE 2.2 – Moyenne et Écart-Type des Caractéristiques Standardisées

Caractéristique	**Moyenne**	**Écart-Type**
RM	8.54e-17	1.00
PTRATIO	1.92e-16	1.00
LSTAT	1.92e-16	1.00
ID	-2.35e-16	1.00

2.2.4 Interprétation des Résultats

Les résultats de la standardisation des caractéristiques indiquent que :

— **Moyenne proche de zéro** : Les moyennes des caractéristiques standardisées sont toutes très proches de zéro (de l'ordre de 10^{-17} à 10^{-16}). Cela signifie que la méthode de standardisation a centré correctement les données autour de zéro, ce qui est essentiel pour éviter les biais dans les algorithmes de machine learning.

— **Écart-Type égal à un** : Les écarts-types des caractéristiques standardisées sont tous égaux à un, ce qui confirme que les données ont été mises à l'échelle de manière uniforme. Cela permet de garantir que toutes les caractéristiques auront un poids équivalent dans l'analyse et la modélisation, empêchant les caractéristiques ayant des échelles plus grandes de dominer le modèle.

La standardisation des données est une étape fondamentale pour s'assurer que les caractéristiques sont sur une échelle commune, facilitant ainsi l'interprétation des coefficients de régression et améliorant la performance des modèles prédictifs.

2.3 Division des Données en Ensembles d'Entraînement et de Test

2.3.1 Introduction à la Division des Données

La division des données en ensembles d'entraînement et de test est une étape cruciale dans le développement et la validation de modèles prédictifs. Elle permet de tester la performance du modèle sur des données non vues pendant l'entraînement, ce qui est essentiel pour évaluer sa capacité de généralisation. Pour ce projet, nous avons divisé les données en un ensemble d'entraînement, utilisé pour ajuster le modèle, et un ensemble de test, utilisé pour évaluer sa performance.

La division des données garantit que le modèle n'est pas surajusté (overfitting) aux données d'entraînement et qu'il peut fournir des prédictions précises sur de nouvelles données.

2.3.2 Processus de Division des Données

Nous avons utilisé la fonction `train_test_split` de la bibliothèque `scikit-learn` pour diviser les données en ensembles d'entraînement et de test. Un paramètre de `test_size` de 20% a été choisi pour que 20% des données soient réservées pour l'évaluation.

Les résultats de la division montrent que 266 échantillons ont été affectés à l'ensemble d'entraînement et 67 échantillons à l'ensemble de test, comme le montre le Tableau 2.3.

TABLE 2.3 – Division des Données en Ensembles d'Entraînement et de Test

Ensemble	**Nombre d'Échantillons**	**Nombre de Caractéristiques**
Entraînement (`X_train`)	266	4
Test (`X_test`)	67	4
Entraînement (`y_train`)	266	-
Test (`y_test`)	67	-

La division des données en ensembles d'entraînement et de test est équilibrée, avec une proportion de 80% pour l'entraînement et 20% pour le test, ce qui est une pratique courante pour s'assurer que l'évaluation du modèle est fiable.

Chapitre 3

Construction des Modèles

3.1 Comparaison des Modèles de Régression

Pour identifier le modèle de régression le plus performant pour prédire la valeur des maisons (`MEDV`), nous avons comparé plusieurs modèles de régression. Cette comparaison nous permet d'évaluer la performance des modèles sur l'ensemble d'entraînement et de sélectionner le modèle qui minimise l'erreur de prédiction.

L'objectif est de déterminer quel modèle offre la meilleure précision et la meilleure capacité de généralisation pour notre jeu de données.

3.1.1 Modèles de Régression Testés et Méthodologie

Nous avons testé plusieurs modèles de régression couramment utilisés dans l'analyse des données. Les modèles testés incluent :

- `LinearRegression`
- `Lasso`
- `Ridge`
- `ElasticNet`
- `KNeighborsRegressor`
- `DecisionTreeRegressor`
- `SVR`
- `AdaBoostRegressor`
- `GradientBoostingRegressor`
- `RandomForestRegressor`
- `ExtraTreesRegressor`

Nous avons utilisé la validation croisée avec 10 plis (`KFold(n_splits=10)`) pour évaluer chaque modèle. La métrique utilisée pour la comparaison est l'erreur quadratique moyenne négative (- MSE), qui mesure l'opposé de l'erreur moyenne au carré entre les valeurs prédites et les valeurs réelles. Une MSE plus faible indique une meilleure performance du modèle.

Le code utilisé pour effectuer la comparaison est présenté ci-dessous :

```
models = {}
models["Linear"] = LinearRegression()
models["Lasso"] = Lasso()
```

```
models["Ridge"] = Ridge()
models["ElasticNet"] = ElasticNet()
models["KNN"] = KNeighborsRegressor()
models["DecisionTree"] = DecisionTreeRegressor()
models["SVR"] = SVR(gamma='scale')
models["AdaBoost"] = AdaBoostRegressor()
models["GradientBoost"] = GradientBoostingRegressor()
models["RandomForest"] = RandomForestRegressor()
models["ExtraTrees"] = ExtraTreesRegressor()

model_results = []
folds = 10
metric = 'neg_mean_squared_error'
model_names   = []
for model_name in models:
    model   = models[model_name]
    k_fold  = KFold(n_splits=folds, shuffle=True, random_state=42)
    results = cross_val_score(model, X_train, y_train, cv=k_fold, scoring=metric)

    model_results.append(results)
    model_names.append(model_name)
    print("{}: {}, {}".format(model_name, round(results.mean(), 3), round(results.std
```

3.1.2 Résultats de la Comparaison des Modèles

Les résultats de la comparaison des modèles de régression sont résumés dans le Tableau 3.1. La métrique utilisée pour l'évaluation est la moyenne de l'erreur quadratique moyenne (MSE) obtenue par validation croisée.

TABLE 3.1 – Comparaison des Modèles de Régression

Modèle	**- Moyenne MSE**	**Écart-Type**
Linear	-34.168	12.499
Lasso	-35.560	14.019
Ridge	-34.162	12.525
ElasticNet	-38.120	16.560
KNN	-21.341	9.184
DecisionTree	-37.868	19.248
SVR	-32.768	17.575
AdaBoost	-17.486	8.528
GradientBoost	-15.850	6.501
RandomForest	-19.188	7.345
ExtraTrees	-15.773	5.455

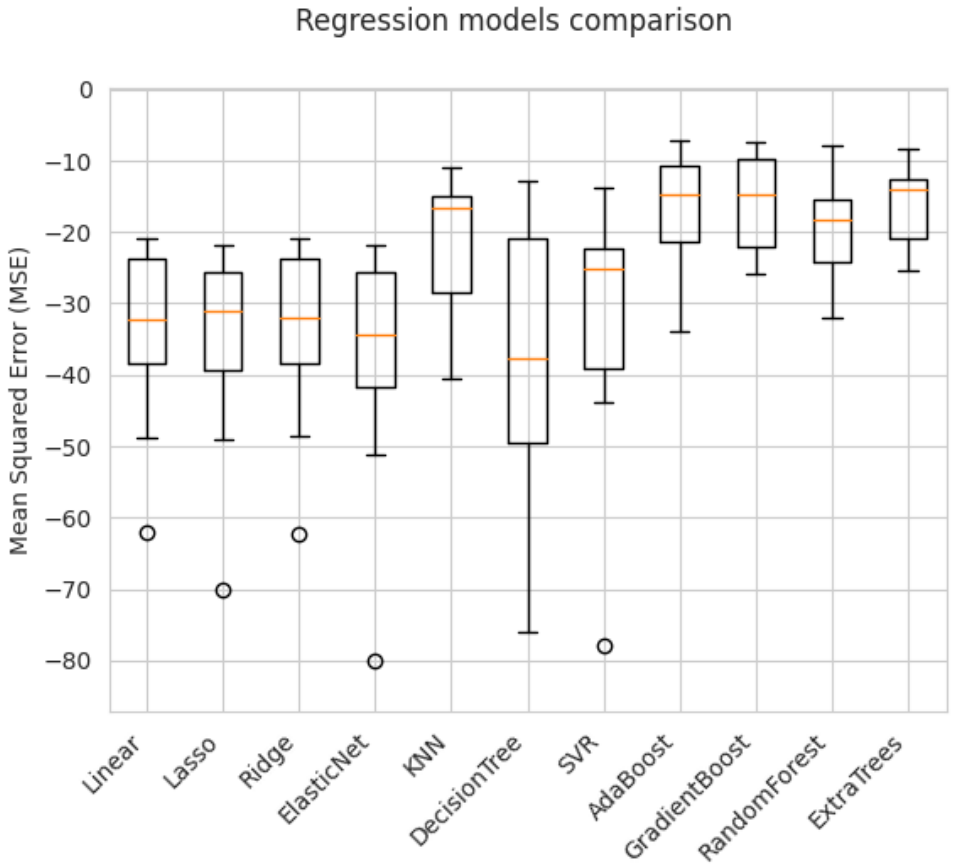

FIGURE 3.1 – Comparaison des Modèles de Régression Basée sur l'Erreur Quadratique Moyenne (MSE)

3.1.3 Interprétation des Résultats

Les résultats de la comparaison des modèles de régression montrent les performances suivantes :

- **Modèle Linéaire (`Linear`)** : Ce modèle présente une erreur moyenne de -34.168 avec un écart-type de 12.499, ce qui indique une performance modérée.
- **Lasso et Ridge** : Les performances de ces modèles sont similaires à celles du modèle linéaire, avec des erreurs moyennes proches de -34 et -35 respectivement. Cela suggère que la régularisation n'améliore pas significativement la performance pour ce jeu de données.
- **KNN** : Le modèle de régression par les k plus proches voisins (`KNN`) affiche une erreur moyenne de -21.341, ce qui est une amélioration notable par rapport aux modèles linéaires. Cela indique que les relations non linéaires entre les caractéristiques peuvent être mieux capturées par ce modèle.
- **AdaBoost et GradientBoosting** : Ces modèles basés sur des arbres de décision présentent les meilleures performances, avec des erreurs moyennes de -17.486 et -15.850 respectivement. L'utilisation de techniques d'ensemble permet de réduire l'erreur de prédiction de manière significative.
- **ExtraTrees** : Ce modèle montre également une très bonne performance avec une erreur moyenne de -15.773 et un écart-type relativement faible de 5.455, indiquant

une grande stabilité dans les prédictions.
— **DecisionTree** : Le modèle de régression par arbre de décision montre une performance inférieure avec une erreur moyenne de -37.868 et un écart-type élevé de 19.248, suggérant une variabilité élevée dans les prédictions.

En conclusion, les modèles basés sur des techniques d'ensemble, tels que `RandomForest`, `GradientBoosting` et `ExtraTrees`, offrent les meilleures performances en termes de précision et de stabilité pour la prédiction de la valeur des maisons dans notre jeu de données.

3.2 Optimisation des Modèles Basés sur des Techniques d'Ensemble

Les modèles basés sur des techniques d'ensemble, tels que `RandomForestRegressor`, `ExtraTreesRegressor`, et `GradientBoostingRegressor`, sont connus pour leur capacité à améliorer la précision et la robustesse des prédictions en combinant les prédictions de plusieurs modèles plus simples. L'optimisation de ces modèles permet d'améliorer encore leur performance en ajustant leurs hyperparamètres pour minimiser l'erreur de prédiction.

3.2.1 Formation Initiale des Modèles

Nous avons formé trois modèles de régression basés sur des techniques d'ensemble en utilisant les données d'entraînement. Voici les paramètres initiaux utilisés pour chaque modèle :

— **RandomForestRegressor** : Nombre d'estimateurs (`n_estimators`) = 100, état aléatoire (`random_state`) = 42.
— **ExtraTreesRegressor** : Nombre d'estimateurs (`n_estimators`) = 100, état aléatoire (`random_state`) = 42.
— **GradientBoostingRegressor** : État aléatoire (`random_state`) = 9.

Le code utilisé pour ajuster ces modèles est le suivant :

```
# Initialisation du RandomForestRegressor
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Initialisation du ExtraTreesRegressor
et_model = ExtraTreesRegressor(n_estimators=100, random_state=42)
et_model.fit(X_train, y_train)

# Initialisation du GradientBoostingRegressor
gbr = GradientBoostingRegressor(random_state=9)
gbr.fit(X_train, y_train)
```

3.2.2 Prédictions et Évaluation des Modèles

Pour évaluer la performance des modèles, nous avons utilisé l'erreur quadratique moyenne (MSE) et le score de variance expliquée. Une MSE plus faible et une variance expliquée plus élevée indiquent une meilleure performance du modèle.

```
# Prédictions et évaluation pour RandomForest
rf_predictions = rf_model.predict(X_test)
rf_mse = mean_squared_error(y_test, rf_predictions)
print("RandomForest MSE:", rf_mse)

# Prédictions et évaluation pour ExtraTrees
et_predictions = et_model.predict(X_test)
et_mse = mean_squared_error(y_test, et_predictions)
print("ExtraTrees MSE:", et_mse)

# Prédictions et évaluation pour GradientBoosting
predictions = gbr.predict(X_test)
print("MSE : {}".format(round(mean_squared_error(y_test, predictions), 3)))
print('Variance score: {}'.format(round(r2_score(y_test, predictions), 2)))
```

Les résultats obtenus sont présentés dans le Tableau 3.2.

TABLE 3.2 – Performance Initiale des Modèles de Régression

Modèle	**MSE**	**Score de Variance**
`RandomForest`	13.306	-
`ExtraTrees`	6.370	-
`GradientBoosting`	12.176	0.80

3.3 Optimisation des Hyperparamètres avec GridSearchCV

Pour améliorer la performance du modèle `RandomForestRegressor`, nous avons utilisé la recherche par grille (`GridSearchCV`) pour trouver les meilleures combinaisons d'hyperparamètres.

3.3.1 Procédure d'Optimisation

Nous avons défini une grille de paramètres incluant le nombre d'estimateurs (`n_estimators`), le nombre maximum de caractéristiques (`max_features`), et la profondeur maximale de l'arbre (`max_depth`). La recherche par grille a évalué chaque combinaison de paramètres en utilisant la validation croisée avec 5 plis.

```
from sklearn.model_selection import GridSearchCV

param_grid = {
    'n_estimators': [100, 200, 300],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth': [None, 10, 20, 30]
}

grid_search = GridSearchCV(estimator=RandomForestRegressor(random_state=42), param_gri
```

```
grid_search.fit(X_train, y_train)
print("Best parameters for RandomForest:", grid_search.best_params_)
```

Les meilleurs paramètres trouvés pour le modèle `RandomForestRegressor` sont :
— **max_depth** : 10
— **max_features** : sqrt
— **n_estimators** : 300

3.3.2 Évaluation du Modèle Optimisé

Nous avons formé le modèle `RandomForestRegressor` optimisé avec les meilleurs paramètres trouvés et évalué sa performance en termes d'erreur quadratique moyenne (MSE).

```
# Initialisation du RandomForestRegressor avec les meilleurs paramètres
rf_model1 = RandomForestRegressor(n_estimators=300, random_state=42, max_features='sq
rf_model1.fit(X_train, y_train)

# Prédictions et évaluation pour RandomForest
rf_predictions1 = rf_model1.predict(X_test)
rf_mse = mean_squared_error(y_test, rf_predictions1)
print("RandomForest MSE:", rf_mse)
```

Le modèle optimisé a montré une MSE de 7.216, indiquant une amélioration significative par rapport aux modèles précédents.
— **RandomForestRegressor Optimisé** : MSE = 7.216
Les résultats détaillés de l'optimisation et de l'évaluation sont résumés dans le Tableau 3.3.

TABLE 3.3 – Performance Optimisée du `RandomForestRegressor`

Modèle	**MSE Initial**	**MSE Optimisé**
`RandomForestRegressor`	13.306	7.216

Chapitre 4

Résultats et Conclusions

4.1 Synthèse des Résultats des Modèles

4.1.1 Comparaison des Modèles de Régression

Dans ce projet, plusieurs modèles de régression ont été comparés pour prédire les valeurs des maisons à Boston. Les modèles basés sur des techniques d'ensemble, comme `RandomForestRegressor`, `ExtraTreesRegressor`, et `GradientBoostingRegressor`, ont été optimisés et ont montré des performances supérieures en termes de précision et de stabilité.

TABLE 4.1 – Performance des Modèles de Régression

Modèle	**Erreur Quadratique Moyenne (MSE)**	**Score de Variance**
`RandomForestRegressor`	7.216	-
`ExtraTreesRegressor`	6.370	-
`GradientBoostingRegressor`	12.176	0.80

Les résultats montrent que le modèle `ExtraTreesRegressor` a la meilleure performance avec une MSE de 6.370, suivi du `RandomForestRegressor` optimisé avec une MSE de 7.216. Le `GradientBoostingRegressor` a également montré une bonne performance avec une MSE de 12.176 et un score de variance de 0.80, indiquant une bonne capacité à expliquer la variance des données.

4.1.2 Analyse des Caractéristiques Pertinentes

L'analyse exploratoire a permis de sélectionner les caractéristiques les plus pertinentes pour la prédiction, notamment le nombre moyen de pièces par logement (`RM`), le pourcentage de la population à faible statut socio-économique (`LSTAT`), et le ratio élèves/enseignant (`PTRATIO`). Ces variables ont montré des relations significatives avec la valeur médiane des maisons, ce qui a été confirmé par les résultats des modèles de régression.

4.2 Conclusions Principales

4.2.1 Efficacité des Modèles d'Ensemble

Les modèles basés sur des techniques d'ensemble, tels que `RandomForestRegressor` et `ExtraTreesRegressor`, se sont révélés être les plus efficaces pour prédire les valeurs des maisons dans ce contexte. Leur capacité à gérer des données complexes et à réduire les erreurs de prédiction a été démontrée par les faibles valeurs de MSE obtenues après optimisation.

4.2.2 Importance de l'Optimisation des Hyperparamètres

L'optimisation des hyperparamètres, en particulier pour le `RandomForestRegressor`, a montré une amélioration significative des performances du modèle. L'utilisation de `GridSearchCV` pour ajuster des paramètres tels que le nombre d'estimateurs, la profondeur des arbres, et le nombre de caractéristiques a permis d'obtenir des prédictions plus précises.

4.2.3 Contribution des Caractéristiques Sélectionnées

Les caractéristiques sélectionnées, telles que `RM`, `LSTAT`, et `PTRATIO`, ont prouvé leur pertinence pour la prédiction des valeurs des maisons. Ces variables doivent être prises en compte pour toute analyse future ou pour la construction de modèles prédictifs dans un contexte immobilier similaire.

4.3 Perspectives et Travaux Futurs

4.3.1 Amélioration des Modèles

Bien que les modèles d'ensemble aient montré une grande efficacité, il serait intéressant d'explorer d'autres techniques de modélisation telles que les réseaux de neurones ou les modèles de boosting plus avancés comme XGBoost, pour voir si des performances encore meilleures peuvent être obtenues.

4.3.2 Études sur des Données Plus Riches

L'utilisation de données supplémentaires, telles que des informations sur l'économie locale, des données climatiques, ou des données sur les infrastructures, pourrait enrichir les modèles et améliorer encore la précision des prédictions.

4.3.3 Applications Pratiques

Les modèles développés dans ce projet peuvent être utilisés pour des analyses immobilières approfondies, pour évaluer des politiques publiques, ou pour des conseils en investissement immobilier. L'intégration de ces modèles dans des outils décisionnels pourrait fournir une aide précieuse aux professionnels de l'immobilier.

Conclusion Générale

Ce projet de prédiction des valeurs des maisons dans les banlieues de Boston a démontré l'efficacité des techniques d'apprentissage automatique pour l'analyse prédictive dans le domaine immobilier. En utilisant des modèles de régression, nous avons pu analyser les relations complexes entre diverses caractéristiques des maisons, telles que le nombre moyen de pièces, le statut socio-économique de la population, et le ratio élèves/enseignant, et leurs valeurs. Nos résultats ont montré que les modèles basés sur des techniques d'ensemble, comme `RandomForestRegressor`, `ExtraTreesRegressor`, et `GradientBoosting Regressor`, sont particulièrement performants pour cette tâche, offrant une précision et une robustesse accrues après optimisation des hyperparamètres. Ces modèles ont non seulement démontré leur capacité à gérer des relations non linéaires et des interactions complexes, mais ils se sont également avérés précieux pour des applications pratiques telles que l'évaluation immobilière, la prise de décision en matière d'investissement, et l'analyse de marché. Toutefois, notre étude a également révélé certaines limitations, notamment la dépendance aux données disponibles et la nécessité d'intégrer davantage de variables pour mieux capturer les dynamiques du marché immobilier. À l'avenir, l'exploration de techniques de modélisation avancées, telles que les réseaux de neurones et l'intégration de données en temps réel, pourrait améliorer encore plus la précision des prédictions. En conclusion, ce projet a mis en lumière le potentiel significatif des approches d'apprentissage automatique dans l'analyse prédictive des valeurs immobilières, ouvrant la voie à des applications innovantes et des améliorations futures dans ce domaine.

Références

Boston Housing (2024). *Housing Values in Suburbs of Boston*. Disponible à l'adresse : `https://www.kaggle.com/c/boston-housing/overview`

Boston Housing (2024). *Le dataset utilisé*. Disponible à l'adresse : `https://www.kaggle.com/c/boston-housing/data`

scikit-learn *Machine Learning in Python*. Disponible à l'adresse : `https://scikit-learn.org/stable/index.html`