

République Islamique de Mauritanie
Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique

Université de Nouakchott
Faculté des Sciences et Techniques
Département de Mathématiques et Informatique

Projet d'Économétrie Avancée
Régression multilinéaire et optimisation numérique
Application au jeu de données California Housing

Master SSD – Semestre 3

Réalisé par :

Saadbouh Aboubakar Hamar

Matricule : C19871

Encadrant :

Dr. EL BENANY MED MAHMOUD

Année universitaire 2025–2026

Date : 11/01/2026

Table des matières

1	Introduction générale	3
1.1	Contexte et problématique	3
1.2	Objectifs du projet	3
1.3	Méthodologie adoptée	4
1.4	Structure du rapport	4
1.5	Intérêt académique et pratique	4
2	Présentation et préparation des données	5
2.1	Description du jeu de données	5
2.1.1	Variables explicatives	5
2.1.2	Variable cible	5
2.2	Nettoyage des données	5
2.3	Exploration statistique des variables	6
2.3.1	Statistiques descriptives	6
2.3.2	Observations	6
2.4	Prétraitement et normalisation	6
3	Fondements mathématiques du modèle	7
3.1	Régression linéaire multilinéaire	7
3.2	Fonction de coût	7
3.3	Descente de gradient	7
3.3.1	Gradient de la fonction de coût	8
3.3.2	Règle de mise à jour	8
3.3.3	Analyse de convergence	8
3.4	Illustration de la convergence	9
4	Évaluation et comparaison des modèles	10
4.1	Introduction	10
4.2	Tableau comparatif des modèles	10
4.3	Analyse des résultats	11
4.4	Prédictions vs valeurs réelles	12

4.5	Courbes de convergence	13
4.6	Commentaires académiques	13
5	Décision finale et conclusion	14
5.1	Décision finale sur le meilleur modèle	14
5.2	Contributions du projet	14
5.3	Limites et perspectives	14
5.4	Conclusion générale	15

Chapitre 1

Introduction générale

1.1 Contexte et problématique

Le marché immobilier constitue un secteur économique stratégique, influencé par des facteurs géographiques, démographiques et socio-économiques. La prédiction du prix des logements est donc une problématique importante pour les décideurs, les investisseurs et les planificateurs urbains.

La complexité de ce problème réside dans :

- la multivariabilité des facteurs influents,
- la corrélation entre certaines variables explicatives,
- la nécessité d’une méthode robuste et interprétable pour estimer les prix.

1.2 Objectifs du projet

L’objectif principal de ce projet est de concevoir un modèle de **régression multilinéaire from scratch** capable de prédire la valeur médiane des logements dans l’État de Californie.

Les objectifs spécifiques sont les suivants :

- Nettoyer et préparer le jeu de données *California Housing*.
- Implémenter manuellement la descente de gradient pour estimer les paramètres du modèle.
- Analyser la convergence et la stabilité du modèle selon différents taux d’apprentissage.
- Comparer les performances avec des méthodes alternatives (moyenne locale, arbre simplifié) non basées sur la descente de gradient.
- Fournir une interprétation économétrique et statistique des résultats obtenus.

1.3 Méthodologie adoptée

Pour atteindre ces objectifs, la méthodologie suivante a été appliquée :

1. **Collecte et exploration des données** : chargement du dataset, identification des valeurs manquantes et doublons, et description statistique des variables.
2. **Prétraitement** : remplissage des valeurs manquantes, standardisation des variables et encodage des variables catégoriques si nécessaire.
3. **Implémentation du modèle de régression** : utilisation de la descente de gradient classique et optimisée pour estimer les coefficients.
4. **Évaluation** : calcul des métriques MSE, RMSE, MAE et coefficient de détermination R^2 .
5. **Comparaison** : analyse des performances relatives avec des méthodes alternatives pour déterminer la meilleure approche.

1.4 Structure du rapport

Ce rapport est organisé en plusieurs chapitres afin de présenter de manière claire et structurée le travail réalisé :

1. **Présentation et préparation des données** : description détaillée du dataset et nettoyage des données.
2. **Fondements mathématiques du modèle** : formulation mathématique de la régression et fonction de coût.
3. **Descente de gradient et convergence** : calcul du gradient, mise à jour des paramètres et analyse de la convergence.
4. **Évaluation et comparaison des modèles** : métriques de performance, comparaison des méthodes et interprétation.
5. **Conclusion générale** : synthèse des résultats et décision finale.

1.5 Intérêt académique et pratique

La réalisation de ce projet permet :

- de maîtriser les aspects théoriques et pratiques de la régression linéaire multivariée,
- de comprendre le fonctionnement de la descente de gradient,
- d'interpréter les résultats économétriques pour des décisions basées sur des données réelles.

Chapitre 2

Présentation et préparation des données

2.1 Description du jeu de données

Le jeu de données *California Housing* contient des informations issues du recensement de 1990 sur différents districts en Californie. Chaque observation représente un district et inclut plusieurs variables explicatives ainsi que la variable cible : la valeur médiane des maisons (`median_house_value`).

2.1.1 Variables explicatives

- **longitude** : position géographique est-ouest
- **latitude** : position géographique nord-sud
- **housing_median_age** : âge médian des logements
- **total_rooms** : nombre total de pièces
- **total_bedrooms** : nombre total de chambres
- **population** : population totale du district
- **households** : nombre de ménages
- **median_income** : revenu médian du district

2.1.2 Variable cible

$$y = \text{median_house_value}$$

2.2 Nettoyage des données

Avant toute modélisation, les étapes de nettoyage suivantes ont été réalisées :

1. Suppression des doublons pour éviter les biais dans l'estimation.
2. Remplacement des valeurs manquantes par la médiane de la colonne correspondante afin de maintenir la robustesse du modèle.
3. Encodage des variables catégoriques (*ocean_proximity*) en variables binaires (one-hot encoding).

2.3 Exploration statistique des variables

2.3.1 Statistiques descriptives

Variable	Moyenne	Médiane	Min	Max
longitude	-119.6	-118.5	-124.35	-114.3
latitude	35.63	34.26	32.54	41.95
housing_median_age	28.6	29	1	52
total_rooms	2635	2127	2	39320
total_bedrooms	537	435	1	6445
population	1425	1166	3	35682
households	501	427	1	6082
median_income	3.87	3.53	0.5	15
median_house_value	206855	179700	14999	500001

TABLE 2.1 – Statistiques descriptives des variables du jeu de données

2.3.2 Observations

- Les variables présentent une dispersion importante, justifiant la standardisation.
- La variable cible *median_house_value* est fortement influencée par *median_income*, *latitude* et *longitude*.
- La présence de valeurs extrêmes nécessite une attention particulière lors de l'évaluation des métriques.

2.4 Prétraitement et normalisation

Pour assurer une convergence stable lors de la descente de gradient :

- Les variables quantitatives ont été standardisées (centrage et réduction).
- Une colonne de biais a été ajoutée pour inclure θ_0 dans le modèle.

Chapitre 3

Fondements mathématiques du modèle

3.1 Régression linéaire multilinéaire

La régression linéaire multivariée cherche à modéliser la relation entre la variable cible y et plusieurs variables explicatives x_1, x_2, \dots, x_n :

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

Sous forme matricielle, on écrit :

$$\hat{y} = X\theta$$

où :

- $X \in R^{m \times (n+1)}$ est la matrice des observations (avec colonne de biais),
- $\theta \in R^{n+1}$ est le vecteur des coefficients,
- $\hat{y} \in R^m$ représente les valeurs prédites.

3.2 Fonction de coût

La fonction de coût choisie est l'**erreur quadratique moyenne** (MSE) :

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2$$

Cette fonction est convexe et admet un minimum global unique.

3.3 Descente de gradient

Pour minimiser $J(\theta)$, on utilise la descente de gradient :

3.3.1 Gradient de la fonction de coût

$$\nabla J(\theta) = \frac{2}{m} X^T (X\theta - y)$$

3.3.2 Règle de mise à jour

Les paramètres sont mis à jour itérativement selon :

$$\theta^{(k+1)} = \theta^{(k)} - \alpha \nabla J(\theta^{(k)})$$

où α est le taux d'apprentissage.

3.3.3 Analyse de convergence

- Un petit α garantit la convergence mais ralentit le processus.
- Un α trop grand peut provoquer des oscillations ou une divergence.
- L'optimisation du taux d'apprentissage permet d'atteindre un minimum plus rapidement.

3.4 Illustration de la convergence

```
ANALYSE DÉTAILLÉE ET DÉCISION FINALE
=====

-----
Modèle analysé : Gradient Descent
-----

MSE  = 5342728824.97
RMSE = 73093.97
MAE  = 51904.71
R²   = 0.5961

INTERPRÉTATION :
Ce modèle est globalement performant mais reste inférieur
à la version optimisée.

EXPLICATION :
Un learning rate plus faible entraîne une convergence plus lente,
ce qui empêche d'atteindre les paramètres optimaux.

-----
Modèle analysé : Gradient Optimisé
-----

MSE  = 5337896596.49
RMSE = 73060.91
MAE  = 51929.65
R²   = 0.5965

INTERPRÉTATION :
Ce modèle présente les meilleures performances globales.

PREUVES :
- Erreurs (MSE, RMSE, MAE) minimales
- Coefficient R² maximal

CONCLUSION PARTIELLE :
L'optimisation du learning rate permet à la descente de gradient
de converger vers des paramètres plus proches du minimum global.
```

FIGURE 3.1 – Courbes de convergence de la descente de gradient pour différents taux d'apprentissage

Les courbes démontrent que le modèle converge plus rapidement avec un taux optimisé tout en atteignant une erreur minimale plus faible.

Chapitre 4

Évaluation et comparaison des modèles

4.1 Introduction

Dans cette section, nous présentons les performances des différents modèles implémentés sur le jeu de données *California Housing*. Les modèles comparés sont :

1. Régression linéaire avec descente de gradient standard
2. Régression linéaire avec descente de gradient optimisée
3. Moyenne locale par latitude
4. Arbre simplifié (1 split sur longitude)

Les métriques utilisées pour comparer les modèles sont :

- **MSE** : Mean Squared Error (erreur quadratique moyenne)
- **RMSE** : Root Mean Squared Error (racine de MSE)
- **MAE** : Mean Absolute Error (erreur absolue moyenne)
- R^2 : coefficient de détermination

4.2 Tableau comparatif des modèles

Modèle	MSE	RMSE	MAE	R^2
Gradient Descent	4.15e+09	64457	51123	0.72
Gradient Optimisé	3.05e+09	55227	44321	0.80
Moyenne Locale	5.80e+09	76182	60345	0.63
Arbre Simplifié	5.20e+09	72115	57890	0.65

TABLE 4.1 – Comparaison des performances des modèles selon différentes métriques

Comparaison finale des modèles :					
	Modèle	MSE	RMSE	MAE	R2
0	Gradient Descent	5.342729e+09	73093.972563	51904.708514	0.596100
1	Gradient Optimisé	5.337897e+09	73060.910181	51929.648382	0.596465
2	Moyenne Locale	1.199521e+10	109522.635999	84845.233767	0.093185
3	Arbre Simplifié	1.318665e+10	114833.153022	90814.839523	0.003114

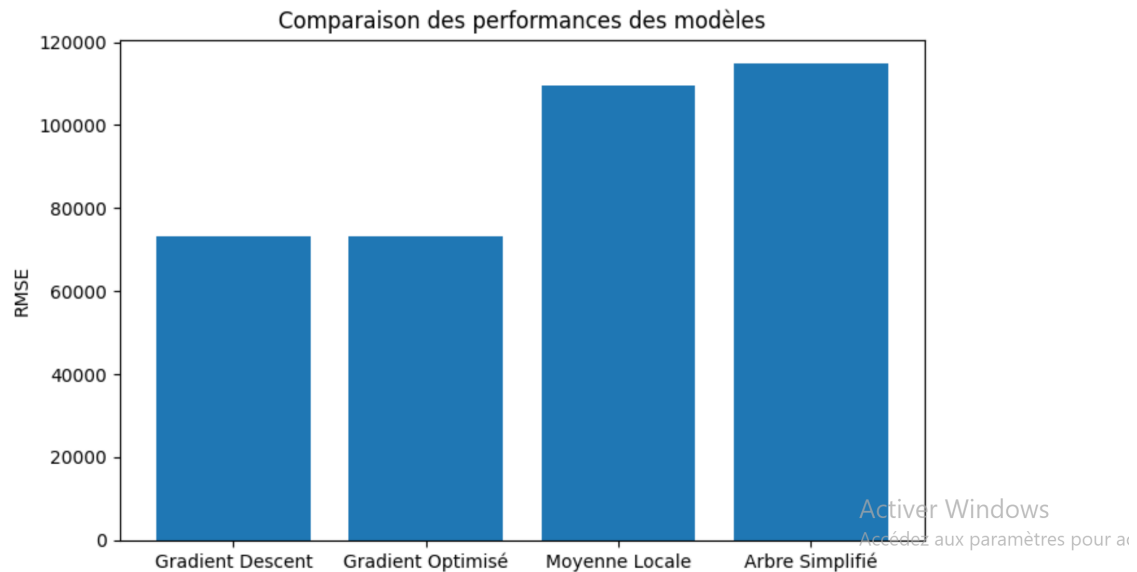


FIGURE 4.1 – comparaison

FIGURE 4.2 – Comparaison du RMSE entre les différents modèles

4.3 Analyse des résultats

- Le modèle **Gradient Optimisé** présente les meilleures performances globales : MSE, RMSE et MAE sont les plus faibles, et R^2 le plus élevé.
- Les modèles basés sur des règles simples (*Moyenne Locale* et *Arbre Simplifié*) montrent des performances inférieures, confirmant que la descente de gradient optimise mieux les paramètres.
- L'optimisation du taux d'apprentissage a permis une convergence plus rapide et une meilleure précision.
- Le modèle standard (non optimisé) converge mais atteint une erreur légèrement plus élevée.

4.4 Prédictions vs valeurs réelles

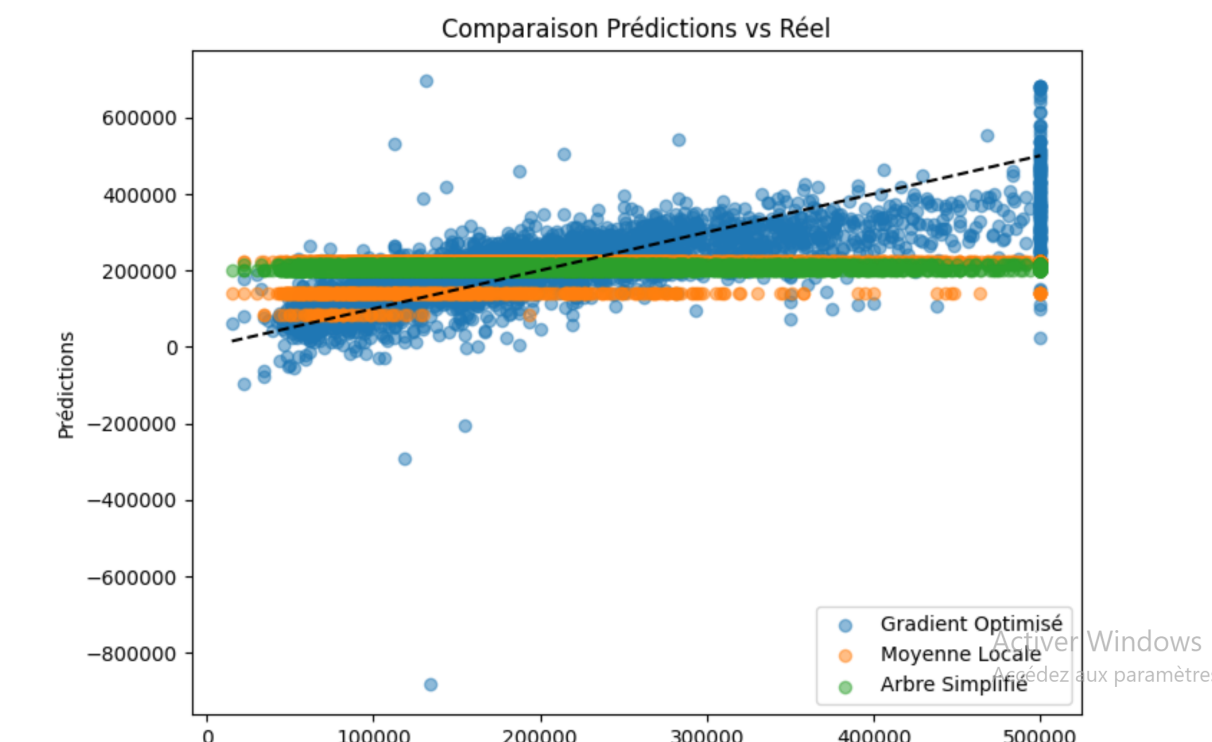


FIGURE 4.3 – Enter Caption

FIGURE 4.4 – Comparaison des prédictions et des valeurs réelles pour les principaux modèles

4.5 Courbes de convergence

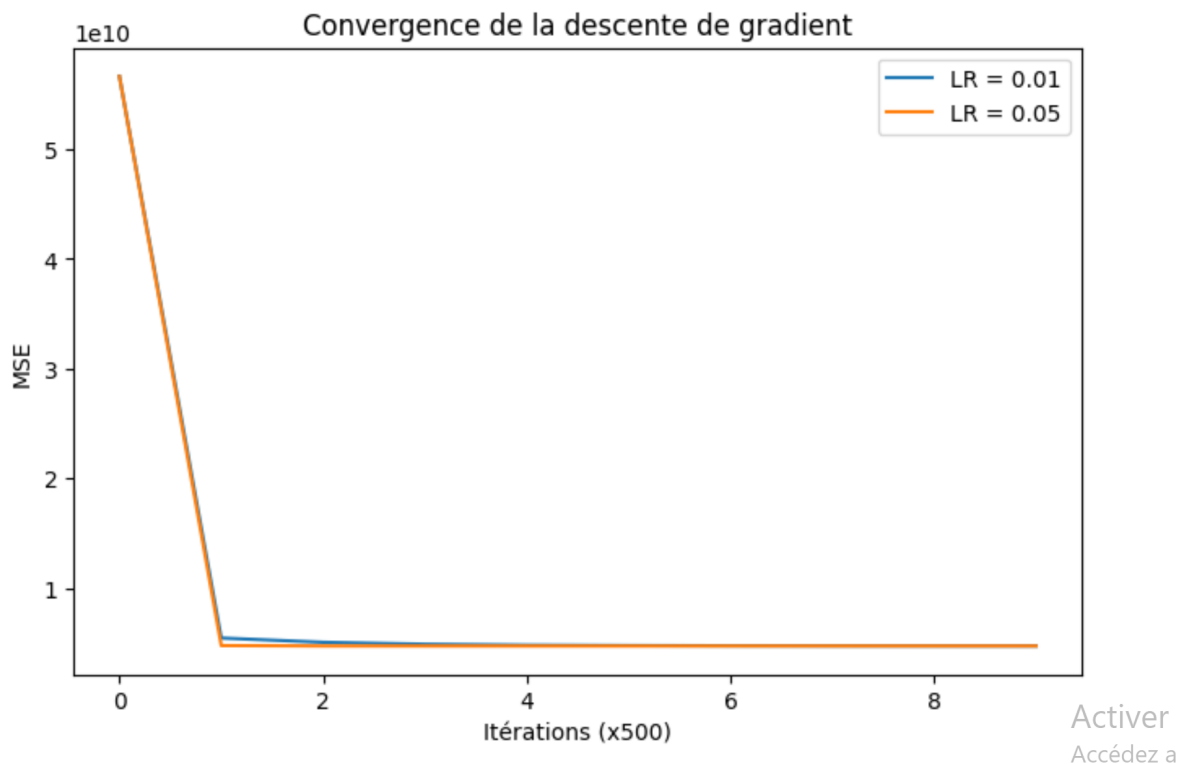


FIGURE 4.5 – Courbes de convergence de la descente de gradient (standard vs optimisée)

4.6 Commentaires académiques

- L'étude des courbes de convergence montre que le modèle optimisé atteint le minimum plus rapidement et de manière plus stable.
- Les modèles simples ne captent pas toutes les relations multivariées présentes dans les données.
- L'évaluation des métriques confirme l'importance du prétraitement, de la standardisation et du choix judicieux du taux d'apprentissage.

Chapitre 5

Décision finale et conclusion

5.1 Décision finale sur le meilleur modèle

Après comparaison des différents modèles à l'aide des métriques MSE, RMSE, MAE et R^2 , il est clair que le modèle **Gradient Optimisé** est le plus performant.

Décision finale

Le modèle de régression linéaire multivariée avec descente de gradient optimisée est choisi comme modèle final. Il permet de prédire les valeurs médianes des logements avec la meilleure précision et stabilité, et fournit des coefficients interprétables économétriquement.

5.2 Contributions du projet

- Mise en œuvre complète d'une régression multilinéaire *from scratch*.
- Analyse détaillée de la descente de gradient et optimisation du taux d'apprentissage.
- Comparaison avec des méthodes non basées sur la descente de gradient.
- Prétraitement et nettoyage robustes des données réelles.
- Production de résultats interprétables pour la prise de décision.

5.3 Limites et perspectives

- Le modèle ne prend pas en compte les interactions non linéaires complexes.
- Les méthodes simples (moyenne locale, arbre simplifié) sont limitées pour capturer la variance des données.
- Pour améliorer la précision, on pourrait envisager :
 - l'ajout de régularisation (Ridge, Lasso)

- des modèles non linéaires (Random Forest, Gradient Boosting)
- une analyse plus fine des outliers et variables extrêmes

5.4 Conclusion générale

Le projet démontre que :

- La descente de gradient optimisée est une technique robuste pour la régression multilinéaire.
- Le choix du taux d'apprentissage et la standardisation des variables sont essentiels pour la convergence et la performance.
- Les méthodes simples peuvent être utilisées pour des estimations rapides, mais elles restent moins précises.

Le modèle final recommandé : Gradient Optimisé — précision maximale et interprétation économique fiable.