

UNIVERSITÉ SIDI MOHAMED BEN ABDLLAH

FACULTÉ DES SCIENCES DHAR EL MAHRAZ

machine learning

LES ALGORITHMES D'APPRENTISSAGE

SUPERVISÉ

Encadré par:

Pr. Ismail EL BATTEOUI

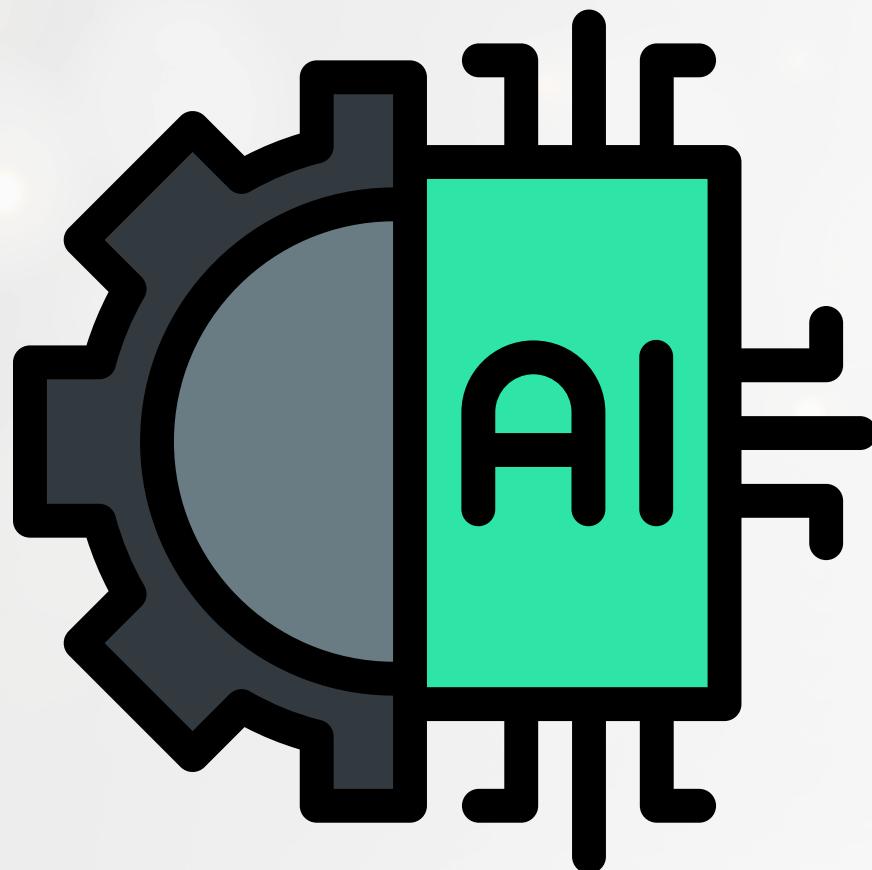
Préparé par:

**AZZOUI wassima
EL IDRISI LAOUKILI Mohammed**

PLAN

Dataset

Régression Linéaire



Régression Logistique

K-Nearest Neighbors

Arbre de Décision

Comparaison des Algorithmes

Conclusion

dataset

Breast Cancer Wisconsin

Maisons

- Nombre d'échantillons : 569
- Nombre de caractéristiques (features) : 30
- Noms des features (extraits) : ['mean radius', 'mean texture', 'mean perimeter', 'mean area', 'mean smoothness', ...]
- Classes (target) : ['malignant', 'benign']

dataset

Breast Cancer Wisconsin

Maisons

- Nombre d'échantillons : 1000
- Nombre de caractéristiques (features) : 4
- Caractéristiques : ['surface', 'nb_pieces', 'age', 'quartier']
- Nom de la variable cible (target) : 'prix'

Différences entre les Algorithmes Classiques en Supervised Learning

1. K-Nearest Neighbors (KNN) – K plus proches voisins

- Type : Apprentissage paresseux (non paramétrique)
- Utilisation : Classification ou régression
- Principe : Prédit une valeur ou une classe en se basant sur les K voisins les plus proches
- Avantages : Simple, efficace pour petits jeux de données
- Inconvénients : Lent sur grands jeux, sensible au bruit et aux variables inutiles

Différences entre les Algorithmes Classiques en Supervised Learning

2. Régression Linéaire

- Type : Modèle paramétrique
- Utilisation : Régression (valeurs continues)
- Principe : Modélise la relation entre les variables d'entrée et une variable de sortie continue à l'aide d'une droite
- Avantages : Facile à comprendre, rapide
- Inconvénients : Suppose une relation linéaire, sensible aux valeurs extrêmes (outliers)

Différences entre les Algorithmes Classiques en Supervised Learning

3. Régression Logistique

- Type : Modèle paramétrique
- Utilisation : Classification binaire ou multi-classes
- Principe : Utilise la fonction sigmoïde pour estimer la probabilité d'une classe
- Avantages : Probabilités interprétables, rapide
- Inconvénients : Moins performant si les données sont complexes ou non linéaires

Différences entre les Algorithmes Classiques en Supervised Learning

4. Arbre de Décision

- Type : Non paramétrique
- Utilisation : Classification ou régression
- Principe : Divise les données en branches selon les valeurs des caractéristiques (tests de type si... alors...)
- Avantages : Interprétable, gère bien les relations non linéaires
- Inconvénients : Risque fort de surapprentissage (overfitting)

Différences entre les Algorithmes Classiques en Supervised Learning

5. Forêt Aléatoire (Random Forest)

- Type : Ensemble d'arbres de décision
- Utilisation : Classification ou régression
- Principe : Crée plusieurs arbres de décision et fait une moyenne (régression) ou un vote (classification)
- Avantages : Plus robuste, réduit le surapprentissage
- Inconvénients : Moins interprétable, plus lent qu'un seul arbre

Régression Linéaire

les avantages

- Bonne performance quand la relation est linéaire
- Simplicité et facilité d'interprétation
- Faibles exigences en données
- Modèle de base robuste et fiable
- Transparence et explicabilité
- Facile à intégrer dans des systèmes existants
- Rapidité d'entraînement et d'exécution

Régression Linéaire

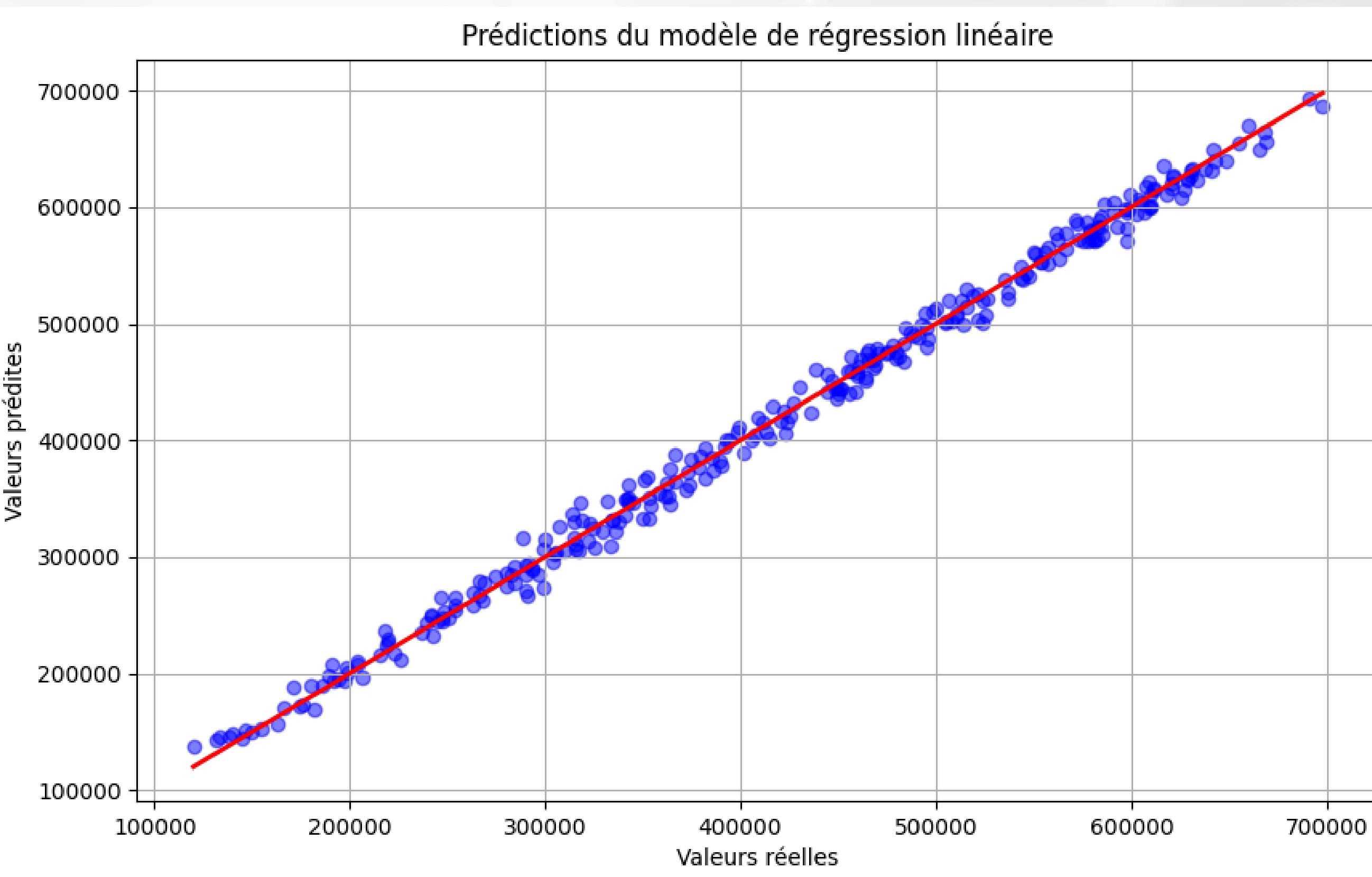
les inconvénients

- Relation linéaire obligatoire
- Sensibilité aux valeurs aberrantes
- Multicolinéarité
- les erreurs sont homoscédastiques et indépendantes
- Ne capture pas les interactions complexes ou non linéaires
- Pas adaptée aux variables qualitatives sans transformation

Régression Linéaire

resultat obtenu

- MAE: 8091.25
- MAPE: 2.23%
- ERMSE: 10032.44
- MSE: 100649760.57
- R²: 1.00



Régression Logistique

les avantages

- Interprétabilité facile
- Adaptée à la classification binaire
- Modèle probabiliste
- Moins sensible aux hypothèses strictes
- Rapide et efficace
- Facile à régulariser
- Extensible à la classification multiclasse

Régression Logistique

les inconvénients

- Extensible à la classification multiclasse
- Sensibilité aux variables corrélées
- Limité aux problèmes de classification
- Peu efficace avec des classes non linéairement séparables
- Difficile à utiliser avec des données déséquilibrées

Régression Logistique

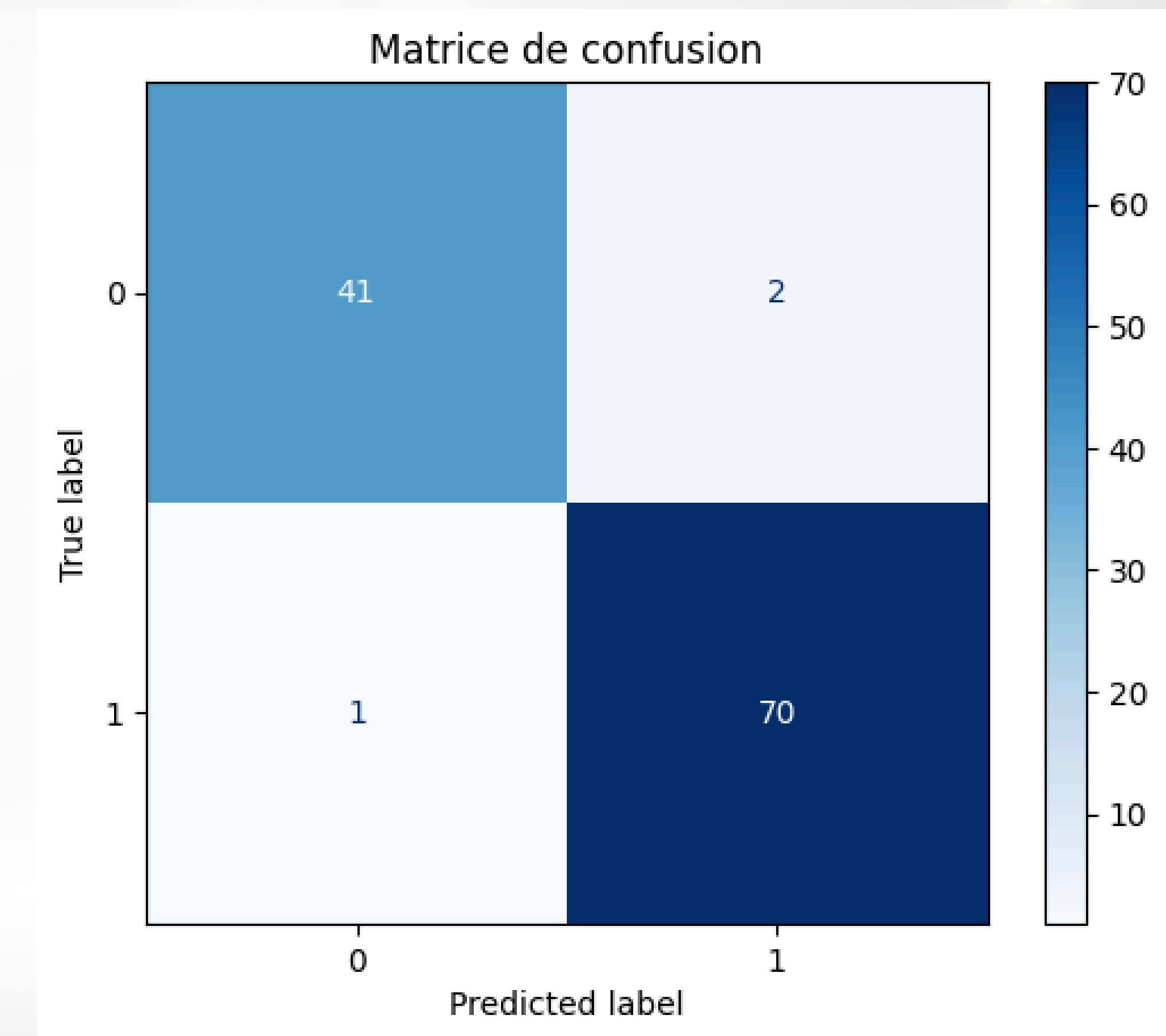
resultat obtenu

Accuracy : 0.97

Precision : 0.97

Recall : 0.97

F1-score : 0.97



K-NN

les avantages

- Simplicité et intuitivité
- Adapté à la classification et à la régression
- Pas de phase d'apprentissage
- Non paramétrique
- Robuste aux données bruitées si k est bien choisi

K-NN

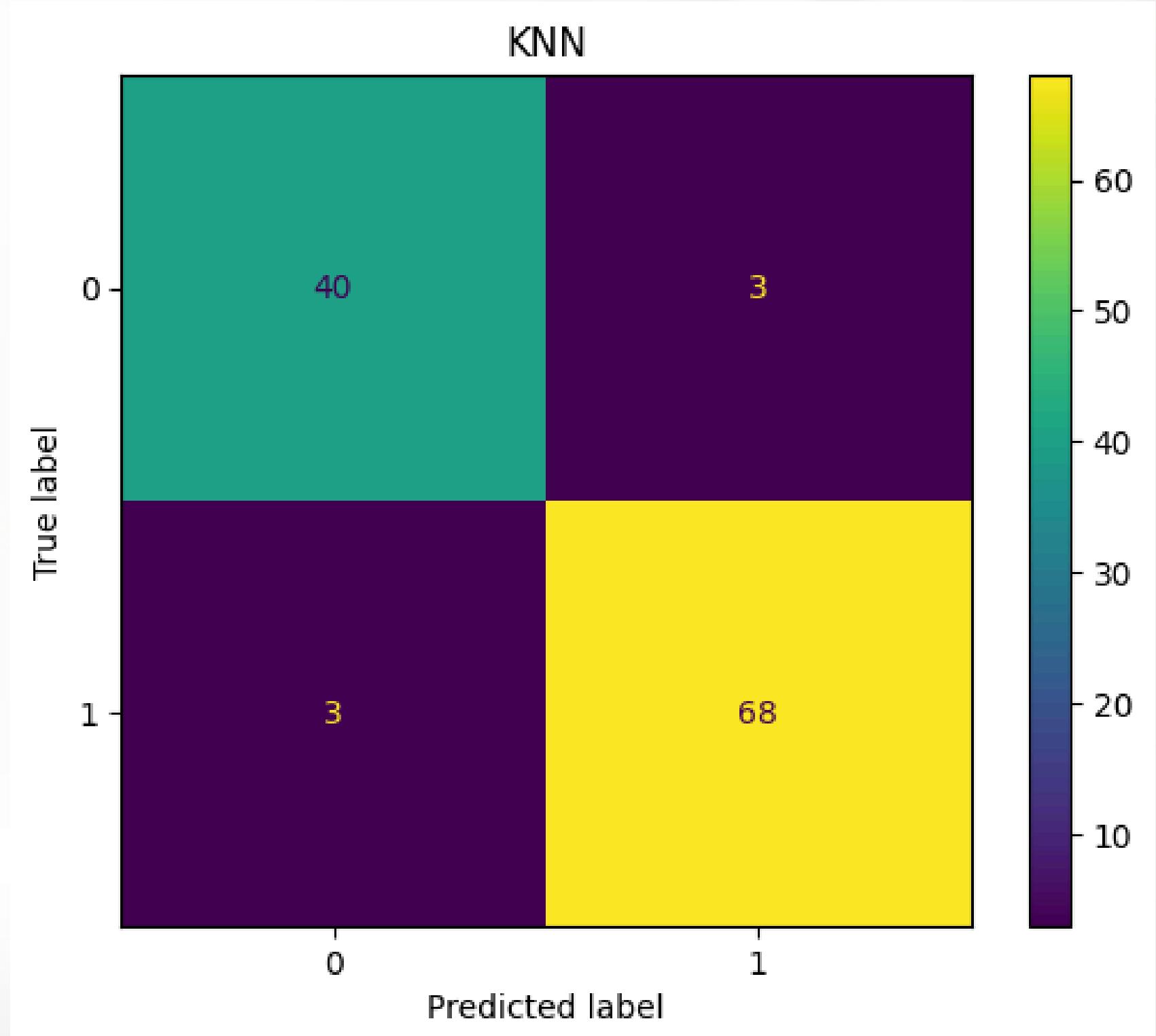
les inconvénients

- Lenteur en phase de prédiction
- Sensibilité à la dimensionnalité élevée
- Choix du paramètre k délicat
- Besoin d'un bon choix de métrique de distance
- Nécessite souvent une normalisation ou standardisation des données

K-NN

resultat obtenu pour classification

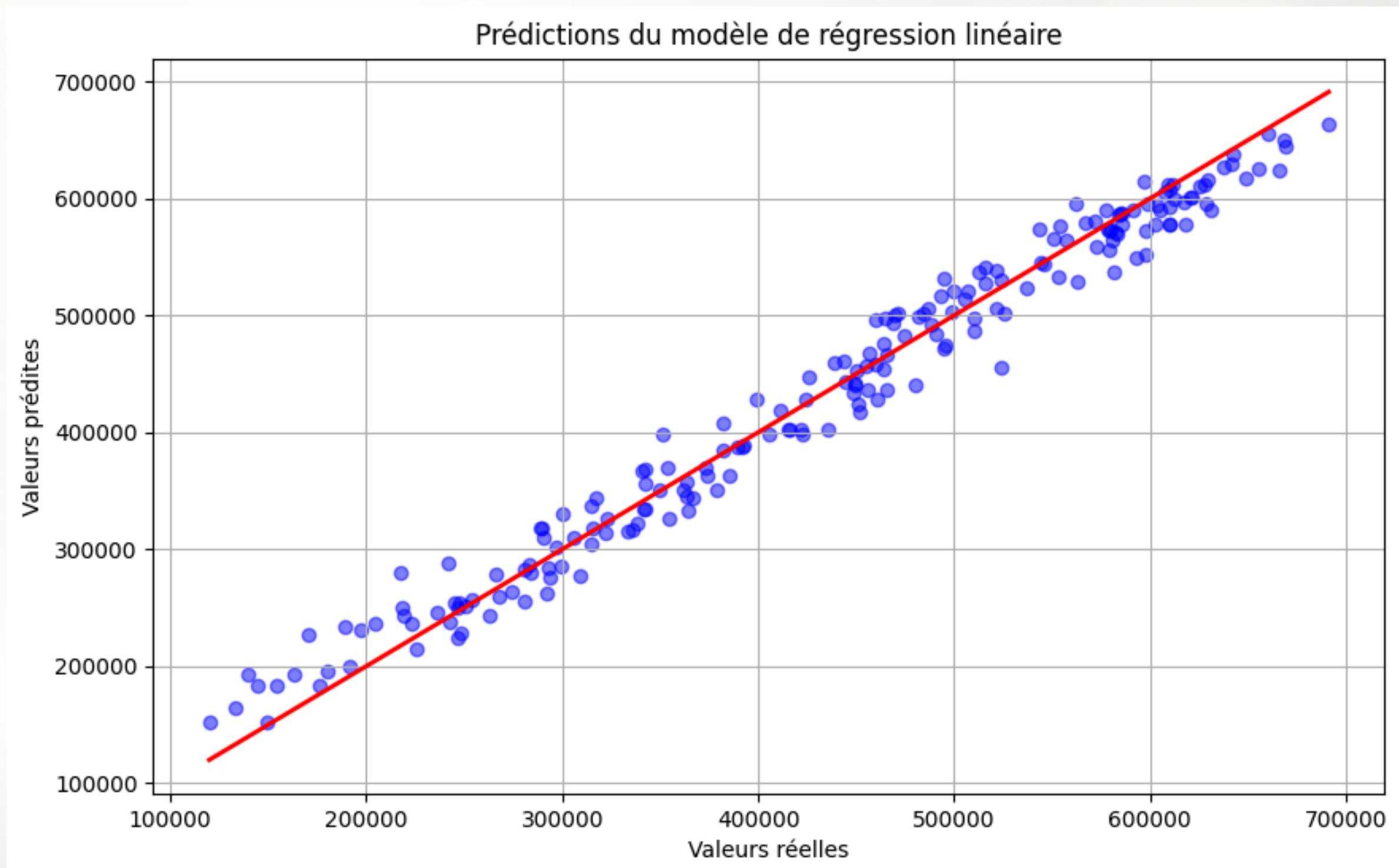
- Accuracy : 0.94
- Precision : 0.94
- Recall : 0.94
- F1-score : 0.94



K-NN

resultat obtenu pour regression

- MSE : 501613641.82
- MAE : 18161.10
- R^2 : 0.98
- MAPE: 5.16%
- RMSE: 22396.73



arbre de décision

les avantages

- Interprétabilité élevée
- Peu de prétraitement nécessaire
- Gère les variables numériques et catégoriques
- Capture les relations non linéaires et interactions
- Rapide à entraîner et à prédire

arbre de décision

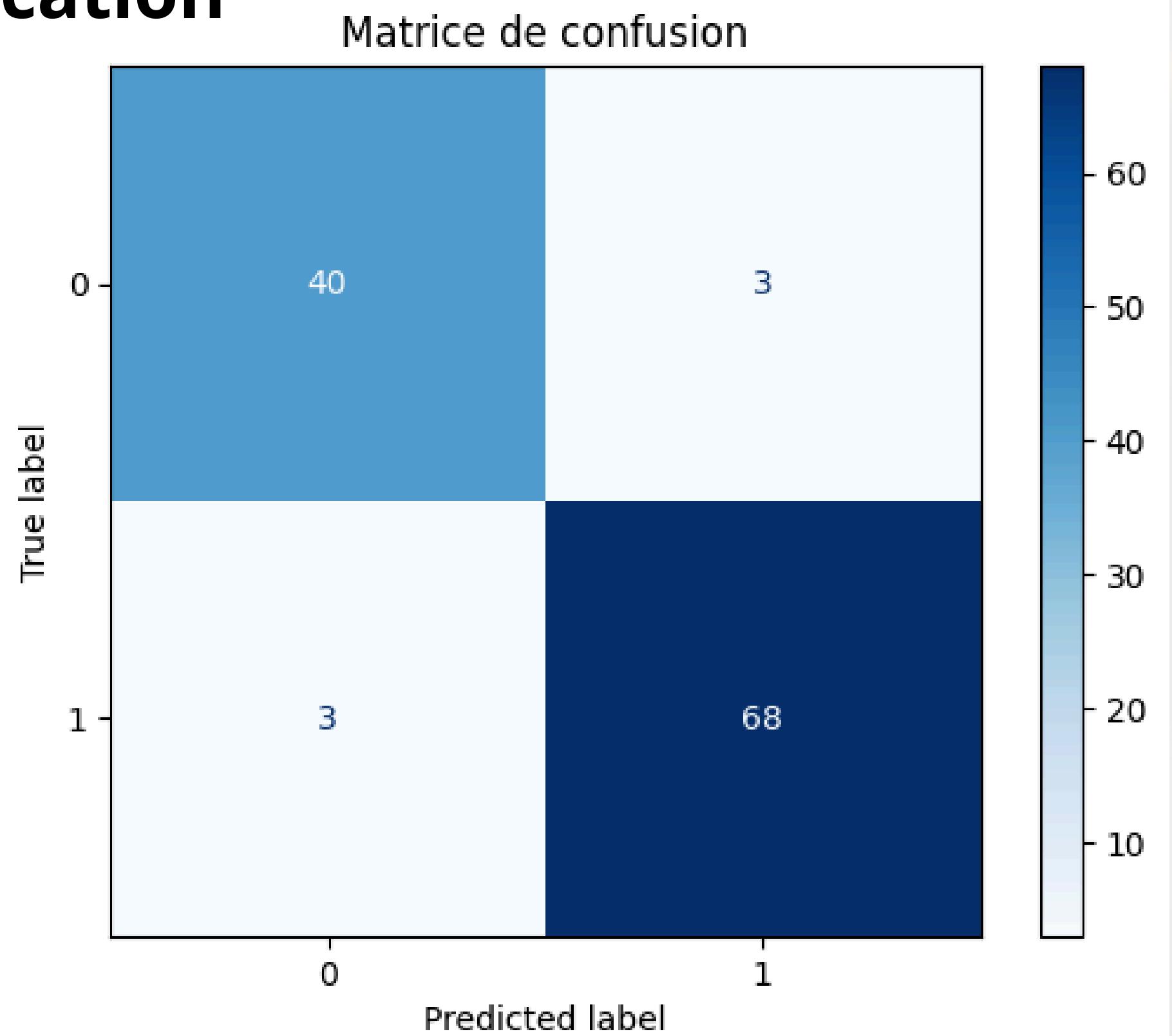
les inconvénients

- Surapprentissage
- Instabilité
- Moins performant seul
- Biais vers les variables avec beaucoup de catégories
- Peut être volumineux et complexe

arbre de décision

resultat obtenu pour classification

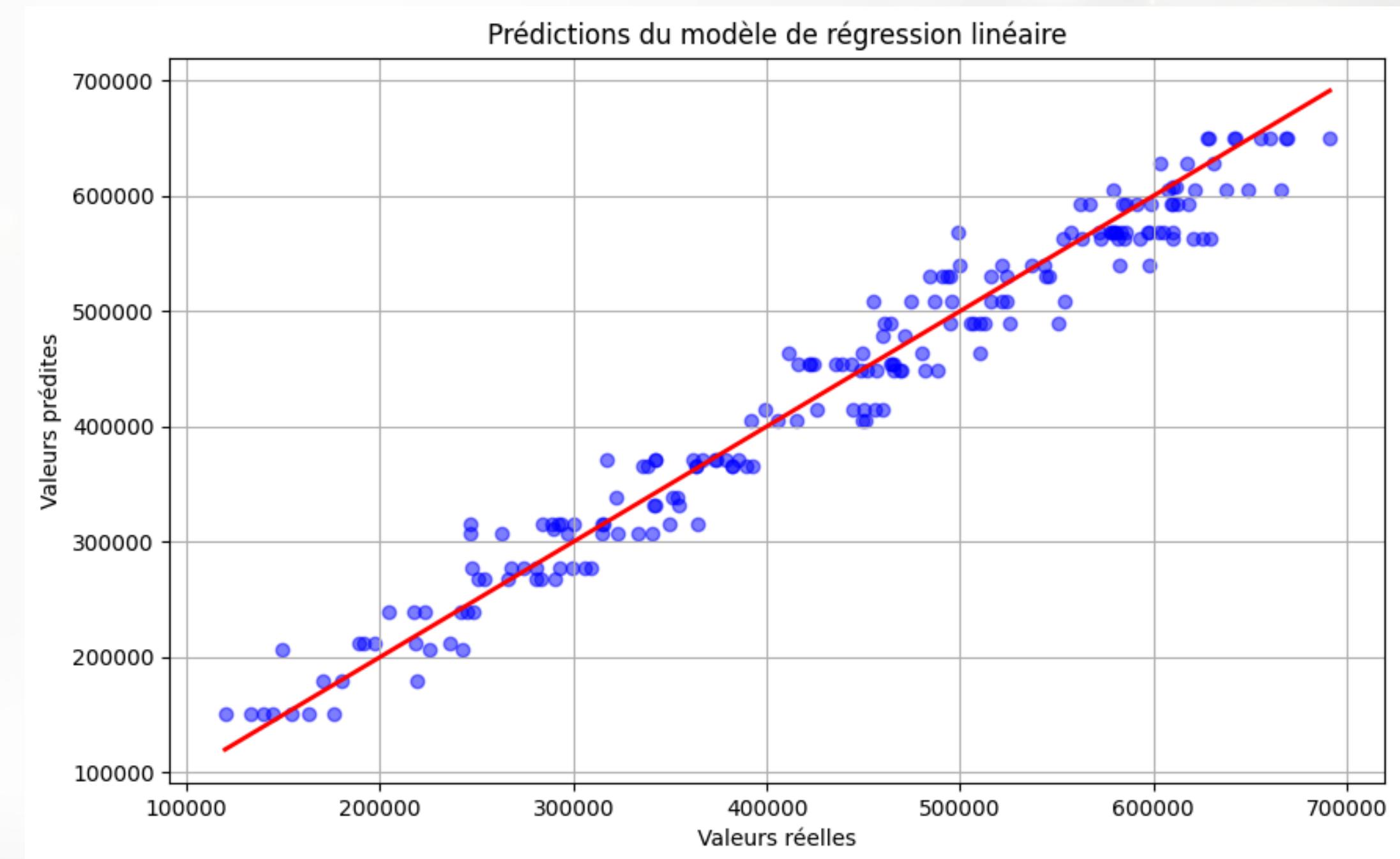
- Accuracy : 0.94
- Precision : 0.94
- Recall : 0.94
- F1-score : 0.94



arbre de décision

resultat obtenu pour regression

- MSE : 727552897.11
- MAE : 21824.65
- R^2 : 0.97
- MAPE: 5.71%
- RMSE: 26973.19



Comparaison des Algorithmes

Algorithme	Type	Performant	Remarque
Régression linéaire	Régression	Oui	Bon si relation linéaire simple
KNN	Classification / Régression	Oui	Performant sur petits datasets
Régression logistique	Classification	Oui	Efficace en classification binaire
Arbre de décision	Classification / Régression	Oui	sujet à l'overfitting sans élagage

Conclusion

La performance d'un algorithme dépend fortement du type de données, du problème à résoudre, et du volume de données. Aucun algorithme n'est "le meilleur" en toutes circonstances, mais certains se démarquent par leur robustesse ou leur simplicité.