Rapport de Projet :

Analyse et Prédiction des Prix des Maisons

Objectif :

L'objectif de ce projet est de prédire les prix des logements à partir d'un dataset de caractéristiques telles que la surface, la localisation, le nombre de chambres, etc. Pour cela, nous avons utilisé un pipeline complet de prétraitement et de modélisation automatique.

Data utilisée :

https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Technologies et Bibliothèques utilisées :

- Python 3.x

- Pandas, NumPy pour la manipulation de données

- Scikit-learn pour le pipeline de machine learning

- linear regression pour un modèle performant

- Matplotlib / Seaborn pour la visualisation. Pipeline de Machine Learning

Étapes du Projet : Prétraitement, Modélisation et Évaluation

1. Prétraitement des Données :

- Séparation automatique des variables numériques et catégorielles
- Encodage des variables de type objet via OneHotEncoder
- Normalisation des variables numériques via StandardScaler
- Intégration dans un pipeline avec ColumnTransformer

2. Modélisation :

- Modèle principal : Régression Linéaire (Linear Regression)
- Optimisation basique des hyperparamètres

3. Évaluation du Modèle :

- Séparation des données avec train_test_split (80% entraînement / 20% test)
- Métriques utilisées :
    - MSE (Erreur Quadratique Moyenne) : 1 754 318 687 330.67
    - R^2 (Score de détermination) : 0.65

Perspectives d'Amélioration :

- Application du Feature Engineering avancé
- Utilisation de GridSearchCV ou Optuna pour optimiser les hyperparamètres
- Implémentation de méthodes d'ensemble (Stacking)
- Analyse plus poussée des erreurs (ex : prédictions aberrantes)