✓ **Benlahcen Souad**
✓ **Supervised learning : Rgression Project**
✓ **18/07/2025**

1. Méthodologie

L'objectif de ce projet est de développer un modèle prédictif fiable pour estimer les prix des biens immobiliers au Maroc à partir d'un ensemble de données collectées en ligne. Plusieurs étapes ont été suivies : collecte des données, exploration, nettoyage, transformation, sélection de variables, modélisation et évaluation.

2. Analyse Exploratoire des Données (EDA)

- **Dimensions** du jeu de données : 9595 lignes × 9 colonnes.
- **Colonnes principales** : surface, ville, nombre_chambres, prix, type_bien, meublé, garage, balcon, ascenseur.
- **Valeurs manquantes** : Quelques valeurs nulles ont été détectées dans surface et nombre_chambres.
- **Valeurs aberrantes** : Des prix supérieurs à 10 millions MAD ou des surfaces supérieures à 1000 m² ont été considérés comme extrêmes.
- **Corrélation** : surface, nombre_chambres, et la ville ont montré une certaine corrélation avec prix.

3. Prétraitement des Données

✓ **Nettoyage** :

Suppression des doublons.

Remplissage ou suppression des valeurs manquantes.

✓ **Conversion de texte** :

Un dictionnaire de traduction a été utilisé pour convertir les villes de l'arabe au français.

✓ **Encodage** :

Variables catégorielles (ville, type_bien) encodées en One-Hot.

✓ **Normalisation** :

Les variables numériques ont été standardisées pour les modèles sensibles à l'échelle comme SVR.

✓ **Division des données** :

Train-test split : 80% entraînement / 20% test.

4. Modèles Testés

Les modèles suivants ont été testés avec validation croisée et optimisation d'hyperparamètres :

- **Régression Linéaire**
- **Random Forest**
- **Gradient Boosting**
- **SVR (Support Vector Regression)**

Chaque modèle a été évalué avec :

- MAE (Mean Absolute Error)
- RMSE (Root Mean Squared Error)
- R^2 (coefficient de détermination)
- Validation croisée (R^2 moyen et écart-type)

5. Résultats

Modèle	MAE (↓)	RMSE (↓)	R^2 (↑)
Linear Regression	296 631.49	406 709.09	0.22
Random Forest	310 784.40	442 306.76	0.07
SVR	335 500.54	462 741.35	-0.01
Gradient Boosting	**291 608.24**	416 827.42	0.18

Meilleur modèle : Linear Regression

6. Conclusion

SVR est le **pire** ici ($R^2 < 0$ → le modèle fait pire qu'une moyenne constante).

Random Forest ne donne pas de bons résultats non plus (faible R^2).

Gradient Boosting est **prometteur** en MAE, mais pas supérieur à la régression linéaire en R^2.

Linear Regression reste **le meilleur compromis** ici.

Améliorations possibles :

- Ajouter plus de variables explicatives (quartier, année de construction…).
- Appliquer un logarithme sur prix pour réduire la variance.
- Entraîner des modèles plus robustes ou des réseaux de neurones.