## 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 | $R^2$ Test | $R^2$ CV (moyen ± écart-type) |
|---|---|---|---|---|
| Régression Linéaire | ≈ 280 000 | ≈ 370 000 | 0.16 | 0.16 ± 0.06 |
| Random Forest | 272 677.57 | 359 874.18 | 0.18 | 0.11 ± 0.08 |
| Gradient Boosting | 273 818.86 | 366 679.85 | 0.15 | 0.16 ± 0.06 |
| SVR | 300 373.99 | 400 680.74 | -0.02 | -0.02 ± 0.02 |

## 6. Conclusion

**Meilleur modèle selon $R^2$ test :** *Random Forest* ($R^2 = 0.18$)

**Meilleur modèle selon validation croisée :** *Régression Linéaire* & *Gradient Boosting* ($R^2 = 0.16$)

**SVR** ne performe pas bien sur ces données probablement à cause de la distribution des prix très dispersée.

**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.