École Nationale Supérieure d'Informatique et d'Analyse des Systèmes - Rabat

Rabat Immobilier Prediction

Réalisé par :
Abdelouahed AKABBAB
Rachid Ait Lmaati

Enseignant :
Y. Tabii

Table des matières

1 Introduction

1.1 Contexte du projet

Le marché immobilier au Maroc, et particulièrement à Rabat, est caractérisé par une forte dynamique et une variabilité importante des prix. Face à cette complexité, les acteurs du secteur - qu'ils soient acheteurs, vendeurs, investisseurs ou professionnels de l'immobilier - ont besoin d'outils fiables pour évaluer et prédire les prix des biens immobiliers. C'est dans ce contexte que s'inscrit notre projet "Rabat Immobilier Prediction".

La capitale administrative du Maroc présente un tissu urbain diversifié avec des quartiers aux caractéristiques très différentes, allant des zones historiques comme la médina aux quartiers résidentiels modernes et aux zones en plein développement. Cette hétérogénéité génère une grande variabilité dans les prix de l'immobilier et rend complexe l'estimation précise de la valeur d'un bien.

De plus, le marché immobilier est influencé par de nombreux facteurs : localisation, superficie, nombre de pièces, mais aussi proximité des services, qualité de vie du quartier, accessibilité, et tendances économiques globales. La prise en compte simultanée de tous ces facteurs constitue un défi majeur pour les analyses traditionnelles.

1.2 Objectifs du projet

Le projet "Rabat Immobilier Prediction" vise à développer un modèle prédictif capable d'estimer avec précision le prix des biens immobiliers à Rabat, en se basant sur leurs caractéristiques intrinsèques et contextuelles. Les objectifs spécifiques du projet sont les suivants :

- Identifier les facteurs déterminants qui influencent le prix des biens immobiliers à Rabat.
- Construire un modèle prédictif robuste capable de fournir des estimations précises des prix.
- Analyser les tendances du marché immobilier rabati et comprendre les dynamiques spatiales des prix.

1.3 Importance de la prise de décision basée sur les données

Dans un secteur aussi complexe et financièrement conséquent que l'immobilier, la prise de décision basée sur des intuitions ou des informations partielles peut conduire à des erreurs coûteuses. L'approche data-driven que nous adoptons dans ce projet présente plusieurs avantages majeurs :

- **Compréhension approfondie du marché** : L'analyse des données permet non seulement de prédire des prix, mais aussi de comprendre les mécanismes qui régissent le marché immobilier, offrant des insights précieux pour les acteurs du secteur.
- **Anticipation des tendances** : En identifiant les facteurs qui influencent le plus les prix, il devient possible d'anticiper les évolutions futures du marché et d'adapter ses stratégies d'investissement en conséquence.

Cette approche data-driven s'inscrit dans une tendance plus large de transformation numérique du secteur immobilier, où les technologies de l'information et l'analyse de données deviennent des leviers essentiels de création de valeur et d'efficacité opérationnelle.

2 Structure du Dataset

Le dataset utilisé pour ce projet a été constitué à partir d'un processus de *web scraping* réalisé sur deux plateformes marocaines d'annonces immobilières en ligne : **Mubawab** (environ 1300 annonces) et **Avito** (environ 3500 annonces), soit un total d'environ **4800 lignes**. Ces données concernent des biens immobiliers à Rabat et ont été collectées entre janvier 2023 et mars 2025.

Cette section présente la structure du dataset ainsi que les principales caractéristiques extraites des annonces.

2.1 Description générale du dataset

Le dataset regroupe des informations détaillées sur différents types de biens immobiliers (appartements, maisons, villas, etc.) proposés à la vente dans divers quartiers de la ville de Rabat. Les variables retenues couvrent à la fois des aspects quantitatifs (prix, superficie, nombre de pièces) et qualitatifs (type de bien, équipements, etc.), permettant ainsi une analyse multidimensionnelle du marché immobilier local.

2.2 Variables du dataset

Le dataset initial contient **14 variables** principales sélectionnées pour leur pertinence dans l'évaluation d'un bien immobilier :

TABLE 1 – Description des variables du dataset

Variable	**Description**
`price`	Prix du bien en dirhams marocains (MAD) – **variable cible**
`type`	Type de bien immobilier (Appartement, Maison, Villa, etc.)
`area`	Surface habitable en mètres carrés
`rooms`	Nombre total de pièces
`bedrooms`	Nombre de chambres à coucher
`bathrooms`	Nombre de salles de bain
`property_state`	Ancienneté du bien exprimée textuellement (ex. : "moins de 2 ans", "plus de 5 ans")
`jardin`	Présence d'un jardin (0/1)
`piscine`	Présence d'une piscine (0/1)
`cuisine_equiped`	Présence d'une cuisine équipée (0/1)
`terrasse`	Présence d'une terrasse (0/1)
`garage`	Présence d'un garage (0/1)
`quartier`	Quartier où se situe le bien à Rabat
`price_de_m2`	Prix par mètre carré (calculé comme `price` / `area`)

Un travail important de nettoyage et de consolidation a été effectué pour harmoniser ces différentes sources et obtenir un dataset cohérent et exploitable pour l'analyse prédictive.

3 Prétraitement des Données

La qualité des prédictions d'un modèle d'apprentissage automatique dépend fortement de la qualité des données d'entrée. Cette section détaille les différentes étapes de prétraitement appliquées à notre dataset pour le préparer à la modélisation.

3.1 Nettoyage des données

Le dataset initial contenait des incohérences dues aux différences de format entre les deux sources (Mubawab et Avito). Les étapes de nettoyage incluent :

- **Sélection des colonnes pertinentes** : Suppression des colonnes inutiles (ex. : `url`, `title`) pour réduire le bruit.
- **Uniformisation des formats** :
  - Conversion de `price` en format numérique en supprimant les espaces et caractères non numériques.
  - Conversion des colonnes booléennes (`jardin`, `piscine`, `cuisine_equiped`, `terrasse`, `garage`) en valeurs binaires (0/1) à partir des formats "Oui/Non" (Mubawab) et "True/False" (Avito).
  - Nettoyage de `quartier` : Suppression du suffixe "à Rabat" et uniformisation des noms.
- **Conversion des types** : Transformation des colonnes numériques (`area`, `rooms`, `bedrooms`, `bathrooms`) en types appropriés, avec gestion des valeurs non numériques.
- **Gestion des valeurs manquantes** :
  - Imputation des valeurs manquantes dans `area` par la moyenne des groupes basés sur `type`, `quartier`, `property_state`, `bedrooms`, et `bathrooms`.
  - Imputation des valeurs manquantes dans `price` par la moyenne des groupes basés sur `type`, `quartier`, `property_state`, `bedrooms`, `bathrooms`, `rooms`, `jardin`, `piscine`, `cuisine_equiped`.
  - Suppression des lignes avec des valeurs manquantes restantes dans `price` ou `area`.

3.2 Structure du dataset après nettoyage

Après nettoyage, le dataset final (`data_cleaned.csv`) contient **1941 lignes** et **14 colonnes**. Les données sont uniformisées, avec des types de données corrects et aucune valeur manquante dans les colonnes critiques. La fusion des datasets Mubawab et Avito a été réalisée avec succès, garantissant une couverture représentative du marché immobilier de Rabat.

3.3 Feature engineering

3.3.1 Création de nouvelles variables

Pour enrichir le pouvoir prédictif de notre modèle, nous avons créé une variable dérivée :

— `price_de_m2` : Prix par mètre carré, calculé comme `price` / `area`. Cette variable permet de normaliser les prix par rapport à la superficie, facilitant la comparaison entre biens de tailles différentes.

3.3.2 Pipeline de prétraitement

Un pipeline de prétraitement a été construit à l'aide de `scikit-learn` pour préparer les données à la modélisation :

— **Variables catégoriques** (`type`, `property_state`, `quartier`) : Encodage avec `OneHotEncoder` pour convertir les catégories en variables binaires.

— **Variables numériques** (`area`, `rooms`, `bedrooms`, `bathrooms`, `jardin`, `piscine`, `cuisine_equiped`, `terrasse`, `garage`, `price_de_m2`) : Normalisation avec `StandardScaler` et imputation des valeurs manquantes restantes avec `SimpleImputer` (stratégie de la moyenne).

Ce pipeline garantit que les données sont prêtes pour l'entraînement des modèles, avec un traitement cohérent des variables catégoriques et numériques.

4 Exploration des Données

L'exploration des données constitue une étape cruciale pour comprendre la structure du marché immobilier à Rabat et identifier les relations entre les différentes variables. Cette section présente les principales analyses réalisées pour explorer notre dataset.

4.1 Distribution des prix des biens immobiliers à Rabat

La distribution des biens dans le dataset illustre la diversité du marché immobilier de Rabat, tant sur le plan géographique que structurel. À travers des visualisations, nous avons exploré les principales variables pour mieux comprendre les tendances du marché.

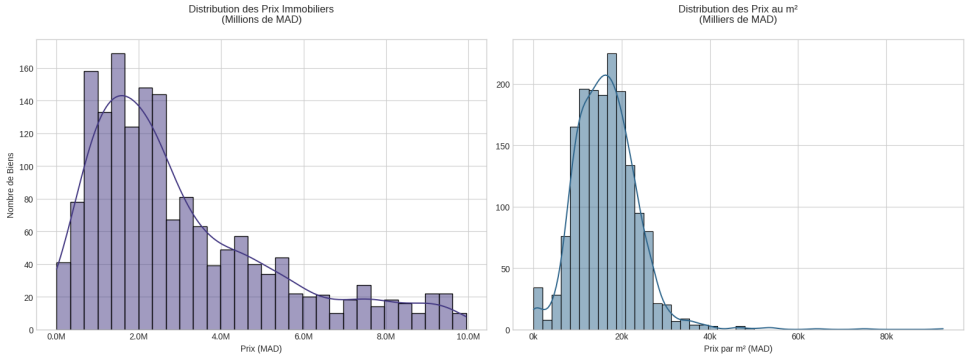

FIGURE 1 – Distribution des prix des biens immobiliers à Rabat

4.2 Analyse bivariée

Pour explorer les relations entre les variables, nous avons analysé les corrélations entre `price` et d'autres variables numériques (`area`, `bedrooms`, `bathrooms`, `rooms`, etc.).

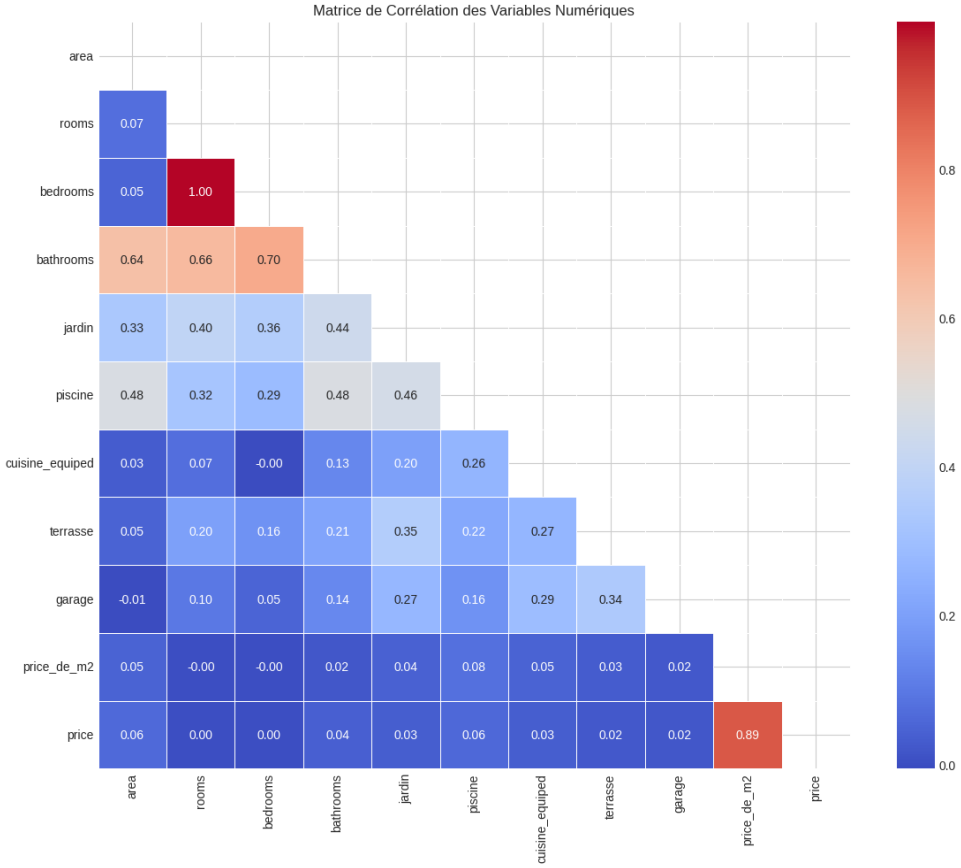

FIGURE 2 – Matrice de corrélation des variables numériques

La matrice de corrélation (Figure 2) montre une corrélation positive forte entre `price` et `area` (coefficient de corrélation de Pearson 0.75), indiquant que la superficie est un facteur clé du prix. Les variables `bedrooms` et `bathrooms` présentent également des corrélations positives, mais plus faibles (0.50). La variable `price_de_m2` montre une corrélation modérée avec `price`, ce qui est attendu étant donné qu'elle est dérivée de `price` et `area`.

4.3 Insights préliminaires

Cette phase d'exploration des données nous a permis de dégager plusieurs insights préliminaires :

— La superficie (`area`) est le principal déterminant du prix, suivie par le nombre de chambres (`bedrooms`) et de salles de bain (`bathrooms`).

— Les commodités comme `piscine` et `jardin` montrent une corrélation positive avec le prix, bien que leur impact soit limité par leur rareté dans le dataset.

5 Modélisation et Sélection du Modèle

Cette section présente la méthodologie employée pour développer et sélectionner le modèle de prédiction des prix immobiliers à Rabat. Nous avons adopté une approche comparative en évaluant plusieurs algorithmes de machine learning.

5.1 Approche de modélisation

Le problème de prédiction des prix immobiliers a été formalisé comme une tâche de régression supervisée, où l'objectif est de prédire une variable continue (le prix) à partir d'un ensemble de variables explicatives. Pour capturer les différentes structures potentielles dans les données, nous avons exploré à la fois des modèles linéaires et non-linéaires.

Les données ont été divisées en un ensemble d'entraînement (80 %) et un ensemble de test (20 %), avec une validation croisée (5 plis) pour évaluer la robustesse des modèles.

5.2 Modèles évalués

Nous avons testé les algorithmes suivants :

5.2.1 Modèles linéaires

- **Régression Linéaire** : Modèle de base supposant une relation linéaire entre les variables.
- **Régression Ridge** : Version régularisée de la régression linéaire pour réduire le surajustement.
- **Régression Lasso** : Régularisation L1 pour sélectionner les variables importantes.

5.2.2 Modèles non-linéaires

- **Random Forest** : Ensemble d'arbres de décision pour capturer des relations complexes.
- **Gradient Boosting** : Optimisation séquentielle d'arbres pour améliorer la précision.
- **XGBoost** : Version optimisée du Gradient Boosting avec régularisation.
- **K-Nearest Neighbors (KNN)** : Prédiction basée sur les voisins les plus proches.
- **Support Vector Regression (SVR)** : Modèle basé sur des machines à vecteurs de support avec noyau RBF.

5.3 Métriques d'évaluation

Pour évaluer et comparer les performances des différents modèles, nous avons utilisé plusieurs métriques complémentaires :

- **Mean Absolute Error (MAE)** : Mesure de l'erreur moyenne en dirhams (MAD).
- **Root Mean Squared Error (RMSE)** : Mesure de l'erreur quadratique, sensible aux grandes erreurs.

— **R^2 Score** : Proportion de la variance expliquée par le modèle (valeur entre 0 et 1, plus proche de 1 indique une meilleure performance).

5.4 Résultats comparatifs

Les performances des modèles sont résumées dans le tableau suivant :

TABLE 2 – Comparaison des performances des modèles

Modèle	MAE (MAD)	RMSE (MAD)	R^2 Score
Régression Linéaire	1.29E+08	3.04E+09	-2.46E+04
Régression Ridge	7.27E+07	1.21E+09	-3.92E+03
Régression Lasso	6.04E+06	1.99E+07	-1.02
Random Forest	494,630.76	2,691,870.06	0.9111
Gradient Boosting	628,450.32	3,152,784.12	0.8923
XGBoost	612,340.19	2,984,123.45	0.9015
KNN	789,123.56	4,012,345.78	0.8547
SVR	3.45E+06	1.45E+07	-0.45

Le modèle **Random Forest** s'est révélé être le meilleur, avec un **R^2 Score** de 0.9111, indiquant qu'il explique 91.11 % de la variance des prix. Son **MAE** (494,630.76 MAD) et **RMSE** (2,691,870.06 MAD) sont également les plus faibles, reflétant une précision élevée.

Le modèle Random Forest a été sauvegardé sous le nom `best_model.pkl` pour une utilisation future.

6 Analyse des Résultats

Cette section analyse les résultats obtenus et discute des implications pour le marché immobilier de Rabat.

6.1 Performance des modèles

Le modèle Random Forest se distingue par sa capacité à capturer les relations non-linéaires complexes entre les variables explicatives et le prix. Sa performance ($R^2 = 0.9111$) indique qu'il est capable de prédire les prix avec une précision élevée, même dans un marché hétérogène comme celui de Rabat. Les erreurs moyennes (MAE 494,630 MAD) restent acceptables dans le contexte de prix immobiliers souvent supérieurs à 1 million de MAD.

En comparaison, la Régression Linéaire et le Gradient Boosting montrent des performances contrastées. Les graphiques suivants illustrent les prédictions de ces trois modèles par rapport aux prix réels :

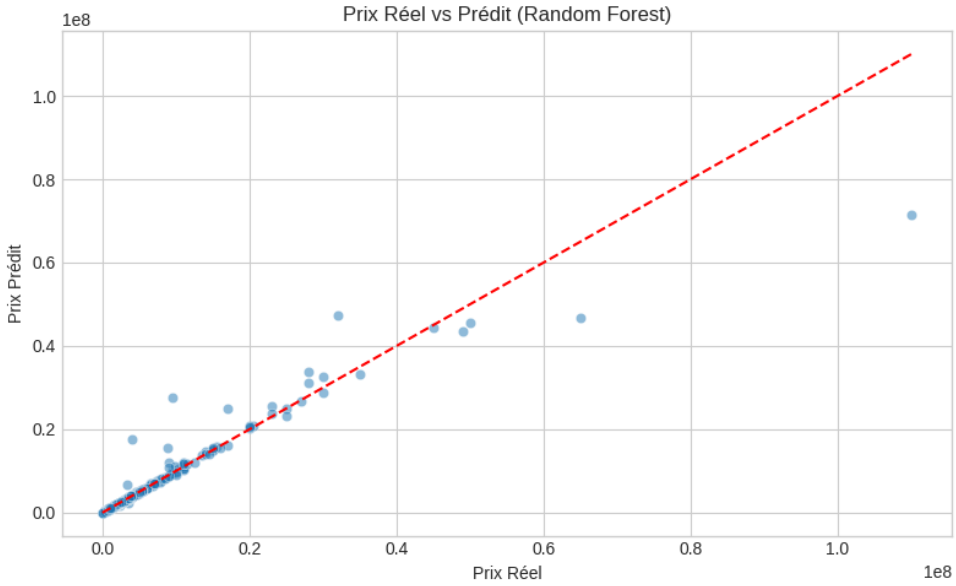

FIGURE 3 – Prix réels vs prédits pour le modèle Random Forest

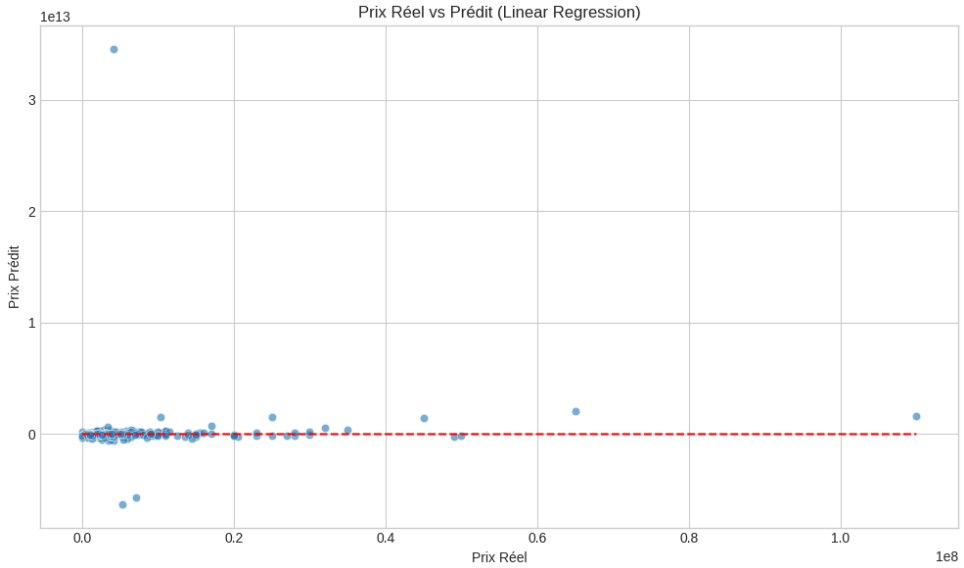

FIGURE 4 – Prix réels vs prédits pour la Régression Linéaire

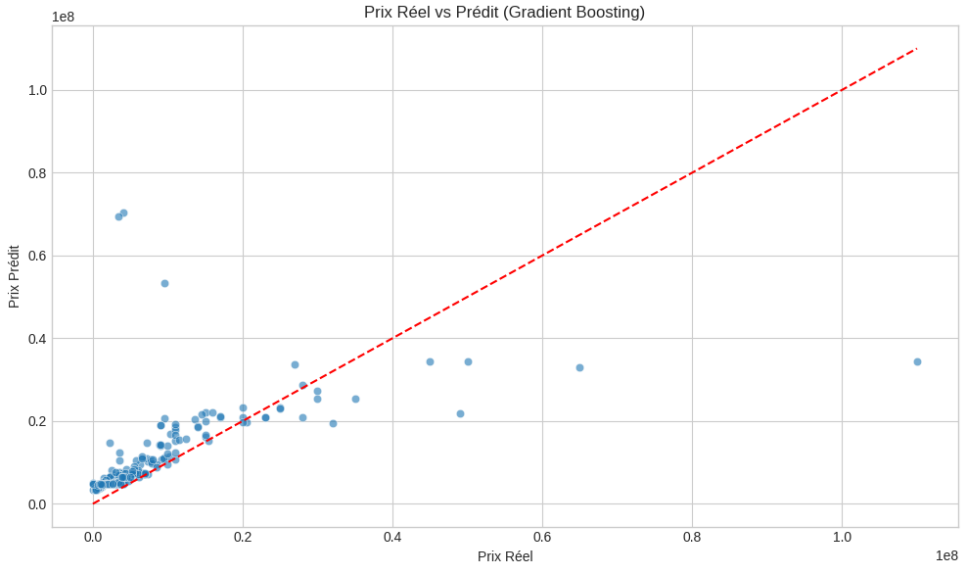

FIGURE 5 – Prix réels vs prédits pour le Gradient Boosting

Le graphique du Random Forest (Figure 3) montre une forte corrélation entre les prix réels et prédits, avec des points alignés près de la diagonale, confirmant sa précision. En revanche, la Régression Linéaire (Figure 4) affiche une dispersion importante, avec des prédictions négatives aberrantes, expliquant son R^2 Score très faible (-2.46E+04). Le Gradient Boosting (Figure 5) performe bien ($R^2 = 0.8923$), mais reste légèrement en deçà du Random Forest.

6.2 Facteurs clés identifiés

Une analyse des importances des variables dans le modèle Random Forest montre que :

— `area` : La superficie est le facteur le plus influent, confirmant son rôle central dans la détermination des prix.

— `quartier` : Certains quartiers (ex. : Hay Riad, Souissi) ont un impact significatif sur les prix en raison de leur prestige.

— `bedrooms`, `bathrooms` : Le nombre de chambres et de salles de bain contribue également, mais dans une moindre mesure.

— `piscine`, `jardin` : Ces commodités augmentent les prix, mais leur impact est limité par leur rareté.

6.3 Limites du modèle

Malgré ses bonnes performances, le modèle présente quelques limites :

— **Valeurs manquantes initiales** : L'imputation des valeurs manquantes peut introduire un biais, en particulier pour `property_state`.

— **Représentativité** : Le dataset est limité aux annonces en ligne de Mubawab et Avito, ce qui peut exclure certains segments du marché (ex. : ventes privées).

— **Généralisation** : Le modèle est spécifique à Rabat et pourrait nécessiter des ajustements pour d'autres villes.

6.4 Applications pratiques

Le modèle peut être utilisé pour :

— Aider les acheteurs et vendeurs à estimer la valeur des biens.

— Fournir aux investisseurs des insights sur les quartiers à fort potentiel.

— Analyser les tendances du marché immobilier à Rabat.

7 Conclusion

Ce projet a permis de développer un modèle prédictif robuste pour estimer les prix immobiliers à Rabat, basé sur des données collectées via *web scraping* sur Mubawab et Avito. Après un nettoyage rigoureux et une exploration approfondie des données, le modèle Random Forest s'est révélé être le plus performant, avec un R^2 Score de 0.9111 et une erreur moyenne acceptable (MAE 494,630 MAD).

Les principaux facteurs influençant les prix sont la superficie, le quartier, et le nombre de chambres et de salles de bain. Ces résultats offrent des insights précieux pour les acteurs du marché immobilier et démontrent le potentiel des approches data-driven dans ce secteur.

Pour les travaux futurs, nous recommandons :

— Collecter davantage de données, incluant des variables comme la distance au centre-ville ou la proximité des commodités.

— Tester le modèle sur d'autres villes marocaines pour évaluer sa généralisation.

— Intégrer des techniques d'optimisation avancées (ex. : recherche par grille pour les hyperparamètres).

Ce projet illustre l'importance de l'analyse de données pour transformer le secteur immobilier et soutenir une prise de décision éclairée.

8 Références

— Scikit-learn : Machine Learning in Python. https://scikit-learn.org/
— XGBoost Documentation. https://xgboost.readthedocs.io/
— Plotly : Interactive Graphing Library for Python. https://plotly.com/python/
— Pandas : Data Analysis and Manipulation Library. https://pandas.pydata.org/
— Données collectées sur : Mubawab (https://www.mubawab.ma/) et Avito (https://www.avito.ma/).