# Analyse des tendances immobilières

Analyse, Prédiction de prix et Cartographie


Réalisé et présenté par :

- BENELMAHI Hicham
- EL HAFIDI Mohamed
- EL KARCHAL Soufiane

# Sommaire

# 01 Introduction & Objectifs

# Introduction

Le marché immobilier marocain
se caractérise par une grande diversité
et une forte fragmentation,
tant au niveau des types de biens
que des zones géographiques,
ce qui rend son analyse globale
complexe.

Les acheteurs et les locataires
rencontrent des difficultés majeures
pour comparer objectivement
les prix pratiqués,
en raison de la dispersion
de l'information
et de l'absence d'indicateurs
de référence fiables.

Il devient alors nécessaire
de disposer d'un outil centralisé
d'analyse des données immobilières,
capable d'exploiter et de croiser
les informations issues
des plateformes Avito et Mubawab
afin d'identifier les tendances
réelles du marché.

# Objectifs du projet

## Automatiser la veille

*Scraping des données*

Automatiser

la récupération des annonces

immobilières

## Fiabiliser la donnée

*Nettoyage & normalisation*

Nettoyer, corriger et

standardiser les données

collectées

## Analyse exploratoire

*EDA & visualisation*

Analyser et visualiser

les tendances du marché

immobilier

## Prédiction des prix

Modèles de Machine Learning

Prédire les prix

de l'immobilier à l'aide

du Machine Learning

# 02 Collecte des Données-Scraping

# Choix des Sources de Données

## Avito.ma


- Volume massif d’annonces
- Données hétérogènes et variées
- Représentation du marché global

VS

## Mubawab.ma

- Données plus structurées
- Annonces détaillées et fiables
- Segment immobilier premium

# Méthodologie de Scraping

## Analyse du site

*Analyse du site web*

- Structure des pages
- Contenu dynamique / statique
- Chargement JavaScript
- Comportement utilisateur

## Plateformes cibles

*Sites immobiliers dynamiques*

- Plateformes immobilières dynamiques **(HTML & JavaScript)**

## Selenium

*Selenium WebDriver*

- Simulation des clics
- Scroll & pagination
- Rendu JavaScript

- ***Sélection par TEXTE et non par classes CSS (classes instables sur Avito)***

## Parsing HTML

*BeautifulSoup*

- Parsing du DOM
- Extraction rapide
- Données structurées

- Après chargement complet de la page

## Données brutes

- Prix
- Localisation
- Surface

# Défis et éthique du scraping web

Cette phase vise à garantir une collecte responsable des plateformes web ciblées.

- Respect du fichier robots.txt
- Gestion des délais (time.sleep) pour éviter le bannissement IP
- Rotation des User-Agents

# 03 Nettoyage des Données

# Pourquoi nettoyer les données ?

- Présence de valeurs aberrantes (ex : surface = 0 m²)
- Doublons entre plateformes
- Formats de prix incohérents

# Nettoyage des données et logique métier

**Traitement des valeurs manquantes :**

| Problème | Solution appliquée |
|---|---|
| Valeurs manquantes (salle_de_bain) | Imputation intelligente basée sur le nombre de chambres |
| Absence du prix | Suppression des lignes concernées |
| Absence de la ville | Suppression des lignes concernées |

Identification automatique du type de bien immobilier lorsqu'il est manquant.

Règle métier appliquée :

- Si nombre de chambres > 6
- Et présence d'un jardin

→ Bien classé comme Villa

Sinon → Appartement

# Prétraitement des données pour le Machine Learning

## Qu'est-ce que le preprocessing ?

Le preprocessing regroupe l'ensemble des transformations appliquées aux données afin de les rendre exploitables par les modèles de Machine Learning.

## Comment le preprocessing est appliqué ?

- Création de la variable « Prix au m² »
- Encodage des villes par Target Encoding
- Normalisation des surfaces (Scaling)

## Pourquoi le preprocessing est nécessaire ?

Les données brutes ne sont pas directement compatibles avec les algorithmes de.

Un bon preprocessing permet :

- d'améliorer la performance des modèles
- de réduiAre les biais liés aux échelles
- de garantir la stabilité de l'apprentissage

# 04 ANALYSE EXPLORATOIRE (EDA)

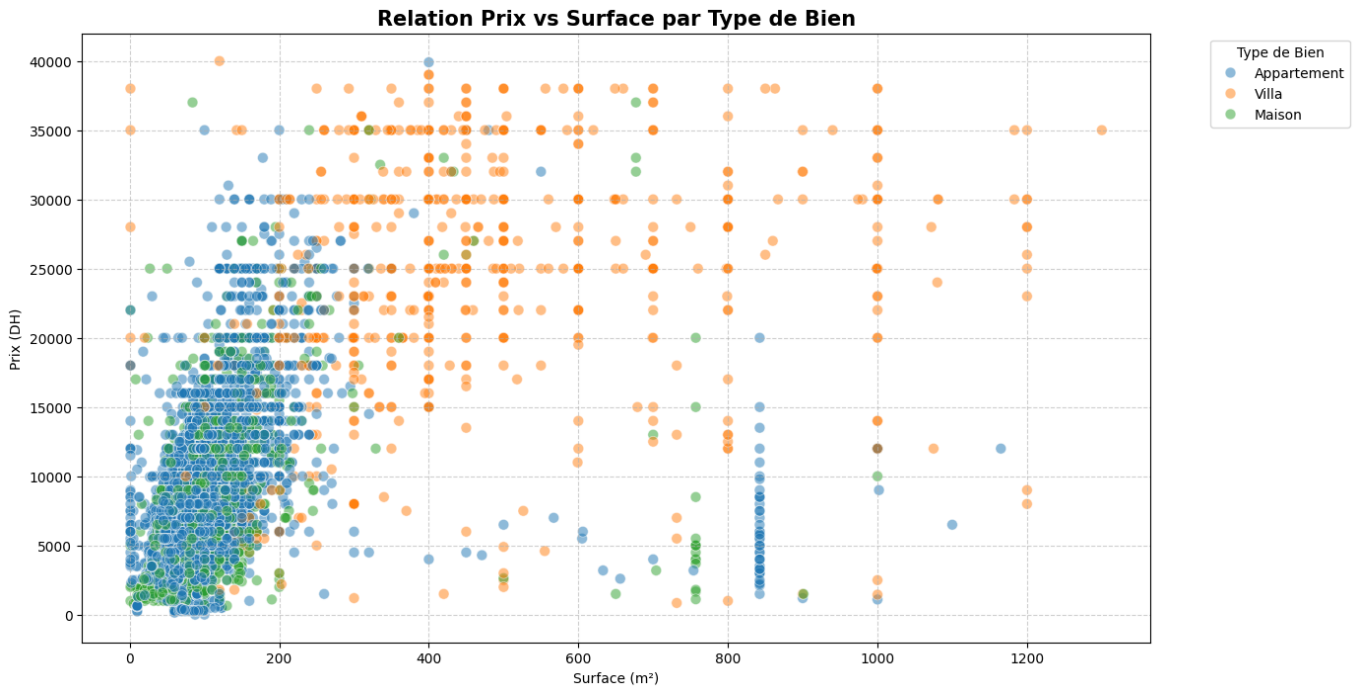
# Objectifs de l’Analyse Exploratoire

Comprendre la répartition géographique des annonces

Détecter les corrélations entre variables

Valider les hypothèses avant le Machine Learning

# Dynamique du Marché : Corrélation Prix et Surface




**1.Segmentation:** Appartements < 200 m² vs Villas dominantes > 600 m².

**2.Tendance :** Hausse constante des prix avec un plafond "luxe" à 40 000 DH.

**3.Volume :** Offre massive concentrée sous 200 m² et 15 000 DH.

**En synthèse :** Si la surface dicte la tendance générale du prix, le type de bien reste le facteur déterminant qui verrouille les paliers de valeur, particulièrement sur le segment haut de gamme.

# 05 Modélisation machine learning

# Modélisation Prédictive

**Analyse des Corrélations:** Identification des variables clés (Surface, Ville, Quartier) ayant le plus d'impact sur le prix final.

**Préparation & Robustesse:** Définition d'un protocole de test rigoureux (Split Train/Test) pour garantir que le modèle fonctionne sur de nouvelles données.

**Optimisation Algorithmique:** Comparaison de modèles (Linéaire vs Ensemble) pour atteindre le meilleur compromis entre précision et généralisation.

# Stratégie de Modélisation & Choix des Algorithmes

### Modèle de base (Baseline)

- Simple et rapide
- Facile à interpréter
- Sert de référence (baseline)

### Modèles avancés (Ensemble)

- Agrégation de plusieurs arbres
- Réduction de la Variance
- Robuste au bruit

### Objectif de la modélisation

- Éviter le sur-apprentissage
- Maximiser la généralisation

### Pipelines

- Encodage des variables
- Normalisation
- Entraînement du modèle

### Gradient Boosting / XGBoost

- Apprentissage séquentiel
- Correction progressive des erreurs
- Excellentes performances prédictives

### Outils & Méthodes

- Scikit-Learn
- XGBoost
- Pipelines de prétraitement

**Avantage :**

Réduction des fuites de données et reproductibilité des résultats.

# Choix Final : Architecture par Stacking

## Concept du Stacking

Fusion des prédictions de plusieurs modèles pour améliorer la précision

## Avantage clé

- Erreur résiduelle réduite
- Prédictions plus stables
- Robustesse accrue

## Base-Learners

- Random Forest
- XGBoost
- ExtraTrees

Chaque modèle capture des patterns différents du marché immobilier

## Meta-Modèle

Régression Ridge

- Agrégation des prédictions
- Régularisation
- Réduction du sur-apprentissage

# Protocole d’Entraînement & Robustesse

**Découpage des données (Vente):**

| 70 % → Entraînement
| 15 % → Validation
| 15 % → Test

→ Évaluer le modèle sur des données jamais vues

**Robustesse du modèle:**

**Validation croisée**
| 5-Fold Cross-Validation
| Réduction du hasard
| Meilleure généralisation

**Export production**
| model_location_final_stacking.pkl

**Pipeline complet :**
| Encodage
| Scaling
| Modèle

→ Modèle directement déployable en production

# Résultats du Modèle Stacking (Location)

**Résultats quantitatifs:**

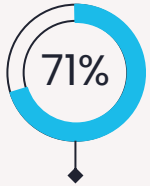
|$R^2$ = 0.73
|MAE = 2 400 DH
| RMSE = 4 844 DH

→ 73 % de la variance du prix expliquée

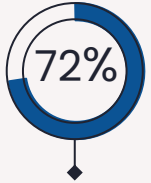
**Interprétation simple**

|Erreur moyenne raisonnable
|Stabilité des prédictions
| Bon compromis précision / robustesse

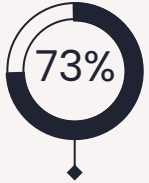
**Comparaison modèles**




**XGBoost seul**

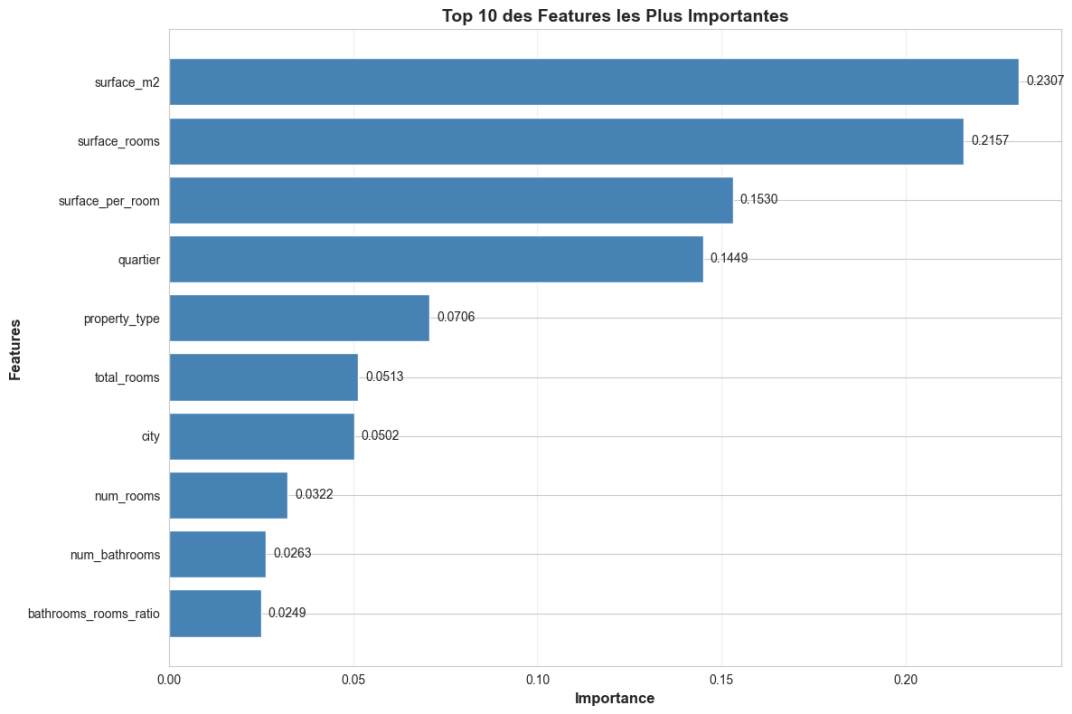



**Voting**




**Stacking**

# Facteurs Déterminants du Prix de location




**Surface (m²)**

| Premier levier de valorisation du bien.

| Plus la surface est grande, plus le prix augmente.

→ Ces variables capturent la même information sous des formes complémentaires.

**Quartier**

| Impact spatial majeur.

| Un même bien peut voir son prix fortement varier selon la localisation.

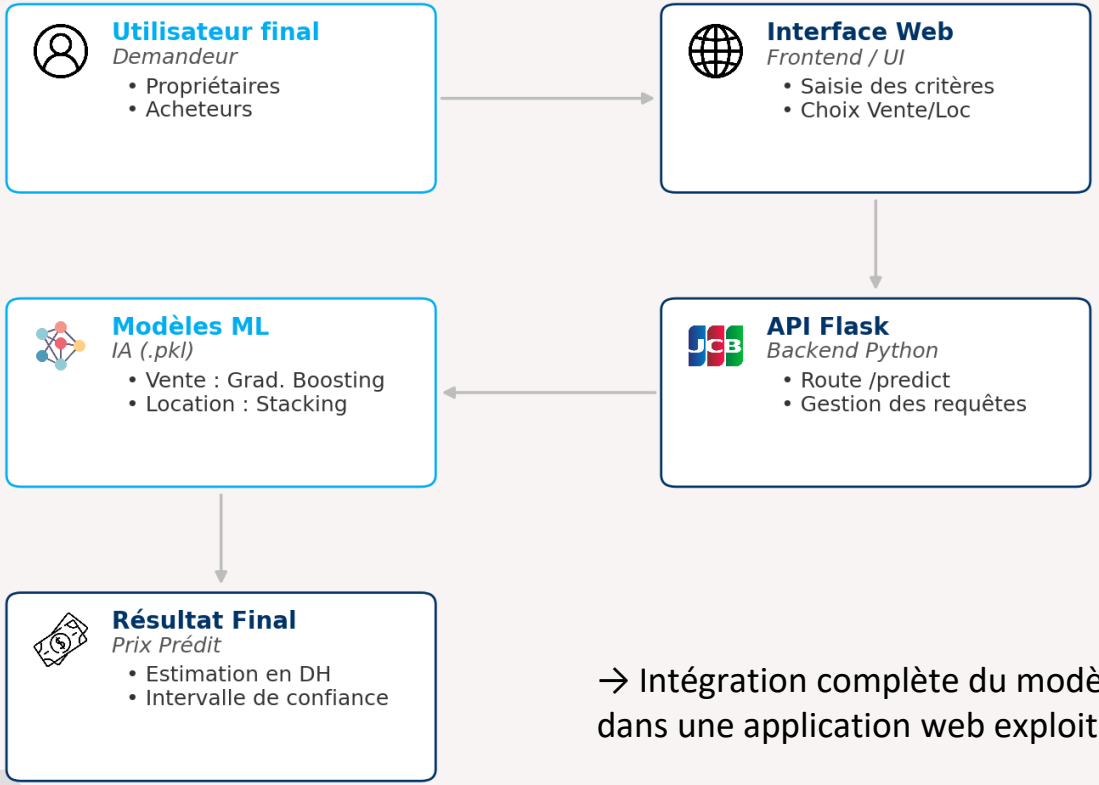
**Type de bien**

| Différenciation claire entre Appartement et Villa.

**Observation avancée:** Les quartiers haut de gamme agissent comme des multiplicateurs du prix basé sur la surface.

# 06 Déploiement & Conclusion

# Architecture de l’Application Web de Prédiction




→ Intégration complète du modèle ML dans une application web exploitable

# Démonstration de l’Outil


**Simulation réelle:** L’utilisateur peut saisir librement les caractéristiques du bien afin d’obtenir une estimation instantanée.

# Conclusion

Une démarche complète en science des données est appliquée, allant de la collecte automatisée des données jusqu'au déploiement de modèles prédictifs fiables. Cette approche repose sur la qualité rigoureuse des données, le choix judicieux des algorithmes et une évaluation méthodique des performances. Ces éléments essentiels garantissent la robustesse, la généralisation et la fiabilité des prédictions obtenues, permettant ainsi d'exploiter efficacement les données pour la prise de décision.

# Merci
# de votre attention

N’hésitez pas si vous avez des questions.