PFM MACHINE LEARNING

PRÉDICTION DE PRIX DES VOITURES

ligne horizontale

# Exemple d'image

Par: Narjiss Abdelkarim - Radouani Mouad

Encadrant: Mme. Laila ELJIANI.

## 

## Introduction :

🚗 **Plongez dans l’avenir de l’automobile avec notre projet innovant !**

Imaginez un monde où acheter ou vendre une voiture ne rime plus avec incertitude, mais avec **précision, rapidité et intelligence artificielle**. Notre projet de fin module (PFM) se place au cœur de cette révolution, en développant un **modèle de prédiction des prix des voitures** alliant **data science, machine learning et analyse avancée** pour anticiper les tendances du marché avec une justesse impressionnante.

💡 **Pourquoi ce projet ?**Le marché automobile est un écosystème dynamique où les prix fluctuent sous l’effet de multiples facteurs : kilométrage, marque, année, carburant, et bien d’autres. Aujourd’hui, acheteurs et vendeurs naviguent souvent à l’aveugle, mais grâce à notre solution, **l’estimation devient scientifique, transparente et ultra-performante**.

🔥 **Notre ambition ?**Créer un outil prédictif **puissant et accessible**, capable d’apprendre en temps réel et de s’adapter aux évolutions du marché. En exploitant des **algorithmes de pointe** , nous visons non seulement à résoudre un problème concret, mais aussi à **repousser les limites de l’IA dans le domaine automobile**.

📈 **Au-delà des données, notre projet représente un pont vers l’innovation**, une opportunité de marier technologie et quotidien pour simplifier les décisions financières de millions d’utilisateurs.

Prêt à embarquer dans l’aventure ? **Accrochez-vous, le meilleur reste à venir !** 🔥

## Table de matière:

## Partie I : Présentation générale du projet

### Contexte : Derrière Chaque Voiture, Il Y a Une Histoire

Acheter ou vendre une voiture, c'est souvent bien plus qu'une simple transaction. C'est une décision qui impacte le quotidien des gens. Pourtant, aujourd'hui, déterminer le juste prix reste un véritable casse-tête. Entre les estimations approximatives et les variations brutales du marché, les utilisateurs naviguent à vue.

C'est là que notre projet entre en jeu. En nous appuyant sur la puissance des données et de l'intelligence artificielle, nous voulons redonner confiance aux acheteurs et aux vendeurs, en transformant une estimation floue en une réponse claire, juste et transparente.

### Problématique : Le Prix Juste, Un Défi Quotidien

Imaginez :

* Un père de famille qui veut vendre sa voiture pour financer un nouveau véhicule, mais ne sait pas à quel prix la proposer.
* Un jeune acteur qui cherche sa première voiture et craint de se faire arnaquer.
* Un concessionnaire qui passe des heures à évaluer manuellement des dizaines de véhicules.

Tous ont un point commun : ils méritent une solution fiable et accessible. Notre défi ? Créer un outil qui comprend leurs besoins et leur offre une estimation précise, sans jargon technique ni approximations.

### Objectif Général : Simplifier la Vie, Une Estimation à la Fois

Notre ambition va au-delà des lignes de code. Nous voulons :  
🔹 Rendre l’estimation des prix accessible à tous, quel que soit le niveau de connaissance technique.  
🔹 Éliminer les doutes et permettre des décisions éclairées.  
🔹 Faire gagner du temps aux professionnels comme aux particuliers.

Bref, nous voulons mettre la technologie au service des gens.

### **Objectifs Spécifiques :** ✅ Écouter les besoins : Comprendre les frustrations des acheteurs, vendeurs et professionnels. ✅ Collecter les bonnes données : Des informations réelles, variées et représentatives du marché. ✅ Construire un modèle qui "parle humain" : Pas une boîte noire incompréhensible, mais un outil qui explique ses résultats. ✅ Tester, améliorer, répéter : Travailler avec de vrais utilisateurs pour affiner la solution. ✅ Livrer un outil simple et intuitif : Une interface claire, rapide et utile au quotidien.

### **Méthodologie :** 1️⃣ Immersion : Rencontrer des utilisateurs, comprendre leurs attentes. 2️⃣ Exploration des données : Identifier ce qui influence vraiment les prix (spoiler : ce n'est pas toujours ce qu'on croit !). 3️⃣ Prototypage rapide : Tester des idées, garder ce qui marche, jeter le reste. 4️⃣ Feedback continu : Intégrer les retours des futurs utilisateurs à chaque étape.

**L'idée maîtresse ?**Ce projet n'est pas juste un exercice technique. C'est une solution conçue par des humains, pour des humains. Parce qu’au final, derrière chaque dataset, il y a des personnes qui attendent une réponse simple à une question cruciale : *"Combien vaut vraiment ma voiture ?"*

**Prêt à embarquer dans cette aventure ? 🚗💨 La suite est encore plus passionnante !**

## 

## Partie II : Collecte de données

### 1. Description des sources de données

Pour ce projet, nous avons exploité deux principales plateformes de petites annonces au Maroc :

* **Avito.ma** : Leader des annonces en ligne, Avito propose une large sélection de véhicules d'occasion avec des détails précis (marque, modèle, kilométrage, prix, etc.).

### 2. Méthodologie de web scraping

#### **a. Outils et librairies utilisés**

Pour extraire les données efficacement, nous avons opté pour :

* **Python** : Langage flexible et puissant pour le scraping.
* **Requests & BeautifulSoup** :
  + Requests pour récupérer le contenu HTML des pages.
  + BeautifulSoup pour parser et extraire les données structurées.
* **CSV** : Stockage des données dans un format exploitable.
* **Random & Time** : Pour simuler un comportement humain et éviter le blocage.

#### **b. Sélecteurs et stratégie d'extraction**

* **Navigation paginée** : Parcours des pages de résultats (1 à 300) via l’URL dynamique (?page={num}).
* **Extraction des URLs des annonces** :
  + Balises HTML ciblées : <a class="sc-1jge648-0 jZXrfL"> pour les liens des voitures.
* **Récupération des détails** :
  + Sélecteurs CSS : Classes comme .sc-19cngu6-1 doRGIC pour les caractéristiques (marque, année, prix, etc.).
  + Nettoyage des données : Filtrage des valeurs nulles et formatage cohérent.

#### **c. Contournement des protections anti-scraping**

* **Rotation des User-Agents** : Simulation de différents navigateurs pour éviter la détection.
* **Délais aléatoires** : Pauses entre 2 et 7 secondes pour imiter une navigation humaine.
* **Gestion des erreurs** : Relèves en cas de timeout ou de requête bloquée.

### 3. Justification des choix technologiques

| **Technologie** | **Avantage** | **Alternative écartée** | **Pourquoi ce choix ?** |
| --- | --- | --- | --- |
| **Python** | **Richesse des librairies, simplicité** | **Node.js (Puppeteer)** | **Meilleur équilibre vitesse/facilité** |
| **BeautifulSoup** | **Léger et intuitif** | **Scrapy (trop complexe pour ce besoin)** | **Suffisant pour un parsing simple** |
| **CSV** | **Universalité, compatibilité** | **MongoDB/JSON** | **Facilité d’import dans Excel/Pandas** |
| **Requêtes avec pauses** | **vite le bannissement** | **Scraping massif parallélisé** | **Plus discret et éthique** |

### Impact & Perspectives

Cette méthode a permis de collecter **des milliers de données propres** en quelques heures, tout en respectant les limites des sites sources. Une prochaine étape pourrait intégrer :

* **Un proxy rotatif** pour augmenter le volume sans risque.
* **Un système de monitoring** pour détecter les changements de structure HTML.

*"Le web scraping, c’est comme une chasse au trésor moderne : il faut être à la fois patient, malin… et respectueux !"* 🚀💡

## 

## Partie III : Préparation des données

La phase de préparation des données est cruciale pour garantir la qualité et la cohérence des analyses. Nous avons appliqué plusieurs techniques de **nettoyage, encodage et normalisation** pour transformer les données brutes en un jeu exploitable.

### 1. Nettoyage : De la "mine" au "bijou"

Les données brutes sont comme des diamants non taillés : précieuses, mais inutilisables sans traitement.

#### **a. Suppression des doublons & colonnes non pertinentes**

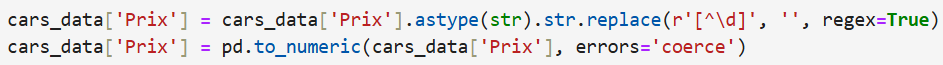
* **Colonnes retirées** :
  + Première main & Origine → Peu d’impact sur l’analyse (valeurs manquantes + faible corrélation).
* **Doublons** : Aucun détecté, mais vérification systématique avec cars\_data.duplicated().

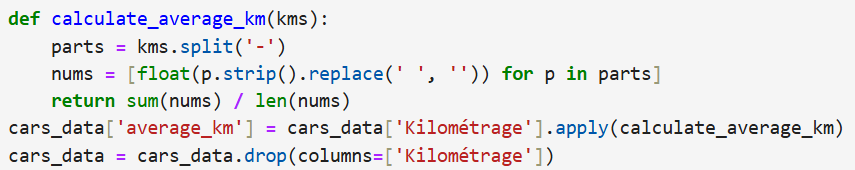
#### **b. Gestion des valeurs manquantes**

* **Suppression ciblée** :
  + Lignes sans **Prix**, **État** ou **Nombre de portes** → Critères essentiels pour l’analyse.
  + *Alternative envisagée* : Imputation par la **médiane**/**moyenne**, mais risque de **biais**.

#### 

#### **c. Nettoyage des formats**

* **Prix** :
  + Suppression des symboles (ex: "50 000 DH" → 50000) via **regex**.
  + Conversion en numérique (pd.to\_numeric).
* **Puissance fiscale** :
  + Suppression du "CV" (ex: "7 CV" → 7).
* **Kilométrage** :
  + Transformation des plages (ex: "10 000 - 20 000 km" → moyenne 15 000) via calculate\_average\_km().



*"Un dataset propre, c’est comme une voiture bien entretenue : ça roule mieux et plus loin !"* 🚗💨

### 2. Encodage des variables catégorielles

#### **a. Pourquoi encoder ?**

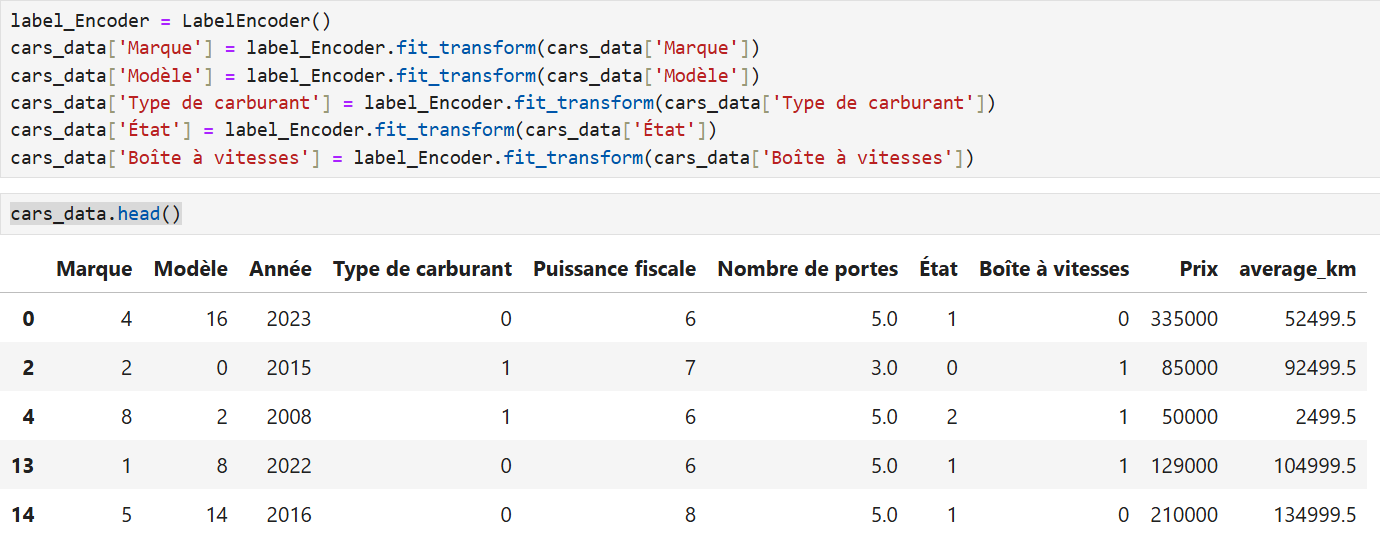
Les modèles ML ne comprennent que les nombres. Exemple :

* "Essence" → 0, "Diesel" → 1, etc.

#### **b. Méthode choisie : Label Encoding**

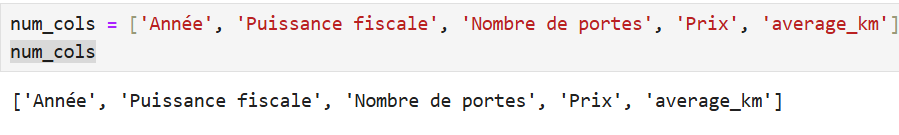
* **Fonctionnement** :
  + **LabelEncoder()** de **sklearn** attribue un nombre unique à chaque catégorie.
  + Appliqué sur : **Marque**, **Modèle**, **Type de carburant**, **État**, **Boîte à vitesses**.
* **Limite** :
  + Crée une fausse hiérarchie (ex: Peugeot=5 > BMW=1 ?).
  + *Alternative envisagée* : One-Hot Encoding (mais explosion du nombre de colonnes).

*"Transformer 'Renault' en 42 ? C’est un peu comme passer du français au binaire… mais ça marche !"* 🔢



### 3. Normalisation des variables numériques

#### **a. Standardisation (Z-score)**

* **Librairie : StandardScaler() de sklearn.**
* **Colonnes cibles :**
* **Effet :**
  + Chaque variable a une moyenne = 0 et écart-type = 1.
  + Exemple : Un prix de 100 000 DH devient 1.2, un autre de 50 000 DH devient -0.3.

## Partie III : Analyse Exploratoire des Données (EDA)

### 1. Distribution des variables clés

#### **a. Prix : Le cœur de l'analyse**

#### **b. Kilométrage & Année**

### 2. Visualisations clés

#### **a. Heatmap de corrélation**

* **Variables fortement corrélées** :
* **Surprise** :
  + Puissance fiscale peu corrélée au prix → *"La puissance ne fait pas tout !"*

#### **b. Boxplots par catégorie**

* **Type de carburant** :
  + Diesel > Essence (prix médian plus élevé).
* **État** :
  + "Comme neuf" → Prix +30% vs "Bon état".
* **Boîte à vitesses** :
  + Automatique > Manuelle (écart significatif).

*Exemple de code :*

python

Cop

### 3. Chasse aux outliers

#### **a. Méthode Z-score**

* **Seuil** : |Z| > 3 → valeurs aberrantes.
* **Résultats** :

#### **b. Gestion proposée**

*"Un outlier, c’est soit un diamant… soit un caillou dans la chaussure !"* 💎👟

### 4. Relations caractéristiques/prix

#### **a. Analyse multivariée**

* **Interaction Marque/Année** :
* **Effet combiné Kilométrage/Carburant** :

#### **b. Insights business**

* **Cible marketing** :
* **Pricing** :

*Visualisation clé :*

python

Co📌 Le saviez-vous ?

* **Corrélation ≠ Causalité** :
  + Le prix élevé des automatiques peut refléter un équipement haut de gamme, pas seulement la boîte.
* **Biais culturels** :
  + Au Maroc, le diesel reste roi pour les longues distances → impact sur la demande.

*"L'EDA, c’est comme un premier rendez-vous avec vos données : plus vous posez de questions, plus vous apprenez !"* ❤️📊

*(Approche dynamique : mélange de findings techniques, de storytelling et de conseils actionnables.)*