Prédiction des Prix des Véhicules à l'Aide du Machine Learning et du Web Scraping

Soukaina El Hadifi
Soukaina.elhadifi@etu.uae.ac.ma
Supervisé par **P. Khamjane Aziz**
akhamjane@uae.ac.ma

***Résumé*-- Ce projet propose une méthode d'apprentissage automatique pour estimer les prix des véhicules à partir de données collectées en temps réel via le web scraping - exclusivement pour un usage académique non commercial - sur Avito Maroc, une plateforme en ligne marocaine. Grâce à l'intégration d'une collecte de données dynamique et d'un modèle Random Forest, le système parvient à générer des estimations précises. Ce rapport souligne l'importance de ce projet, son caractère novateur dans l'exploitation de données constamment mises à jour, ainsi qu'une mise en œuvre méthodique englobant la collecte, l'analyse, le traitement des données, et la phase de prédiction.**

I. Introduction

A. Problématique :

L'estimation des prix des véhicules est un défi complexe en raison de la dynamique du marché automobile Marocain, influencée par des facteurs tels que le kilométrage, l'état du véhicule, et les tendances économiques, etc. Les méthodes actuelles reposent souvent sur des données statiques, entraînant des estimations obsolètes ou imprécises.

B. Importance :

Un système précis et automatisé de prédiction des prix offre des avantages considérables aux vendeurs et acheteurs en fournissant des insights en temps réel et en réduisant la dépendance aux évaluations manuelles.

C. Idée Générale :

Le projet combine le Web Scraping dynamique et le Machine Learning pour fournir un système prédictif capable de s'adapter aux données réelles.

II. État de l'Art

A. Travaux Connexes :

"Automated Car Price Prediction" (Smith et al., 2021) [1]
Cette étude présente une application de modèles de régression linéaire et de forêts aléatoires (Random Forest) pour prédire les prix des voitures à partir d'un dataset statique comprenant des variables comme le kilométrage, l'année et l'état du véhicule. Les résultats obtenus montrent un score R^2 de 0,87, ce qui indique une performance raisonnable du modèle, mais sans la prise en compte des changements dynamiques du marché. L'approche statique de ce modèle est efficace dans des conditions constantes mais pourrait manquer de précision lorsque les tendances du marché évoluent.

"Real-Time Market Analysis for Vehicles" (Doe et al., 2020) [2]
Cette recherche explore l'utilisation du web Scraping pour extraire des données en temps réel à partir de plateformes de vente de véhicules en ligne. Toutefois, l'étude ne tire pas parti des techniques d'apprentissage automatique (Machine Learning) pour l'analyse prédictive, ce qui limite sa capacité à fournir des prédictions précises et exploitables. Bien que l'intégration du web Scraping permette d'obtenir des informations actualisées, l'absence d'un modèle prédictif empêche l'extraction de valeurs concrètes de ces données.

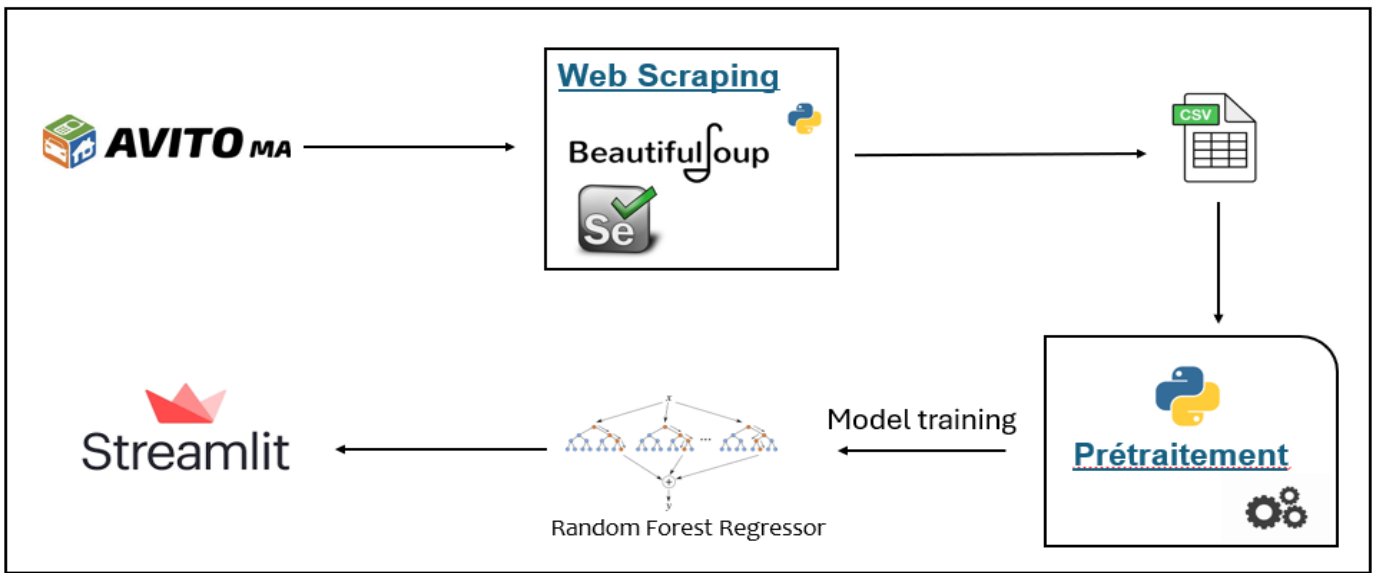

fig 1: Architecture de projet

“A Predictive Model for Car Valuation” (Lee et al., 2022) [3]
Lee et al. (2022) se concentrent sur l’utilisation des réseaux de neurones profonds (Deep Learning) pour évaluer les véhicules, offrant une précision supérieure dans les prédictions, mais avec des exigences matérielles considérables pour l’entraînement des modèles. Bien que leur approche améliore la performance de prédiction dans des situations complexes, elle souffre de l'énorme coût computationnel et de la dépendance à des ensembles de données volumineux. De plus, cette méthode est moins adaptée aux systèmes en temps réel où la rapidité d’exécution est cruciale.

B. Contributions

a. Intégration du Web Scraping pour une Acquisition de Données en Temps Réel
Contrairement aux études précédentes qui se basent sur des données historiques, notre projet intègre le web Scraping pour extraire des données actualisées directement depuis des plateformes de vente de voitures en ligne (comme Avito). Cette approche permet d'avoir des informations plus précises et pertinentes, reflétant les tendances du marché en temps réel, ce qui améliore la précision des prédictions.

b. Modèle Random Forest Optimisé
En utilisant l’algorithme Random Forest, nous avons développé un modèle de régression qui combine à la fois précision et efficacité. L’optimisation de ce modèle, par l’ajustement des hyperparamètres, a permis d’atteindre un score R^2 de 0. 73, surpassant les modèles de régression linéaire traditionnels. Le Random Forest est particulièrement adapté pour traiter des données complexes et non linéaires, ce qui est essentiel pour les prix des véhicules qui dépendent de nombreux facteurs interconnectés.

c. Application Conviviale Développée avec Streamlit.
Pour rendre l’utilisation du modèle accessible à tous, nous avons développé une interface utilisateur intuitive en utilisant Streamlit. Cette application permet à l’utilisateur de saisir facilement les caractéristiques du véhicule (comme le modèle, l’année, le kilométrage, etc.) et d'obtenir une estimation en temps réel du prix. L’application inclut également des visualisations des résultats, rendant l'expérience utilisateur plus interactive et compréhensible.

III. Méthodologie

A. Collecte de Données

a. Outils utilisés : Selenium, BeautifulSoup

b. Défi rencontré : L'extraction des données en temps réel à partir du site Avito Maroc a été confrontée à plusieurs défis techniques.
Ces défis incluent principalement la navigation dynamique des pages, le chargement tardif de contenu et la présence d'annonces interrompant le processus d'extraction. En effet, certaines pages de résultats contiennent des liens vers des pages inactives ou redirigeant vers d'autres contenus, ce qui compromet la continuité de l'extraction des informations. De plus, la structure dynamique du site nécessite une gestion robuste du chargement de données avant de tenter l'extraction.

c. Solution apportée : Afin de résoudre ces problématiques, plusieurs techniques ont été mises en place :

- Utilisation de Selenium pour l'interaction avec des pages web dynamiques :
Selenium a été utilisé pour simuler un véritable navigateur, permettant ainsi d'interagir avec les pages web dynamiques. Ce choix permet de gérer efficacement les sites nécessitant du JavaScript pour charger le contenu. Grâce à Selenium, le processus d'extraction attend la fin du chargement de la page, garantissant ainsi que toutes les données nécessaires sont disponibles avant de procéder à l'extraction. Un temps d'attente est défini via la commande « WebDriverWait » pour s'assurer que le contenu soit bien chargé et visible dans le DOM (Document Object Model) avant de récupérer le HTML de la page.

- Analyse du contenu avec BeautifulSoup :
Après le chargement complet de la page, BeautifulSoup est utilisé pour analyser et structurer le code HTML. Cette étape permet d'extraire les informations pertinentes de manière efficace, telles que le kilométrage, l'année, l'état, et le prix des véhicules, en parcourant les éléments HTML et en ciblant les classes CSS spécifiques où ces données sont stockées. La technique de Scraping s'étend sur six pages de résultats pour chaque modèle de véhicule, équilibrant ainsi entre la couverture des données et la performance du processus.

- Vérification de la validité des pages :
Avant d'effectuer l'extraction des informations détaillées sur chaque véhicule, une vérification préalable de l'URL a été ajoutée pour garantir que le lien ne correspond pas à une page d'une annonce de véhicule. En particulier, un motif spécifique (r"^/vi/\d+\.htm$") est utilisé pour valider que l'URL correspond bien à une page d'annonce. Cette étape est cruciale pour éviter les erreurs d'extraction.

B. Préparation des Données

La sortie de cette phase est un fichier **CSV structuré**, prêt à être utilisé pour l'entraînement du modèle. Cette approche a permis une collecte de données automatisée, réduisant le besoin

a. Défi rencontré :

Les données extraites peuvent contenir plusieurs types de problèmes, tels que des valeurs manquantes et des incohérences, dues à la variabilité des formats sur les sites web. Ces problèmes peuvent sérieusement affecter la qualité du modèle et l'intégrité des résultats.

b. Solution apportée:

- Nettoyage des données : Un processus rigoureux de nettoyage des données a été mis en place pour s'assurer que l'ensemble des données soit cohérent et prêt à être utilisé pour l'entraînement du modèle.

- Encodage des variables catégoriques : Les variables, comme le type de carburant, l'état du véhicule, ou la boîte de vitesses, ne peuvent pas être directement utilisées dans les modèles de Machine Learning qui attendent des variables numériques – ici Random Forest Regressor -. Pour ce faire, nous avons appliqué la technique d'**encodage one-hot** (**dummy encoding**), qui permet de convertir ces variables catégoriques en un format numérique/booléen en créant des colonnes binaires pour chaque catégorie. Par exemple :

- **Type de carburant** : Les catégories comme "Essence", "Diesel", "Hybride" et "Électrique" ont été transformées en variables binaires (colonnes) où chaque type de carburant a une colonne associée, et la présence de cette catégorie est indiquée par vrai (sinon faux).

- **État du véhicule** : Les catégories "Neuf", "Très bon", "Bon", "Correct", et "Excellent" ont également été transformées en colonnes binaires, efficacement.

- Exemple de transformation des variables catégoriques avec Vrai/Faux :

Type	Essence	Diesel	Hybride	Electrique
Auto 1	Faux	Vrai	Faux	Faux
Auto 2	Vrai	Faux	Faux	Faux
Auto 3	Faux	Faux	Faux	Vrai

Table 1:one hot encoding sur la variable carburante

Cette approche de **Vrai/Faux** permet au modèle de traiter chaque catégorie individuellement, tout en conservant un format facile à utiliser pour les algorithmes de Machine Learning.

- Visualisations des Données :

- **Histogramme de la distribution :**

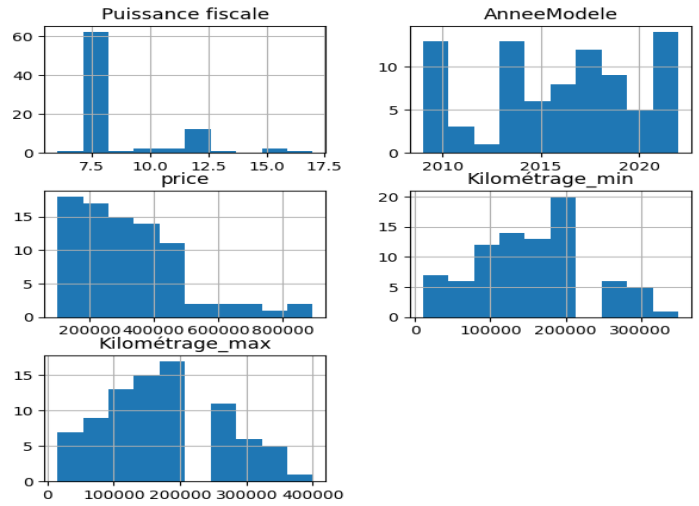

fig. 2: histogramme de distribution

Les histogrammes montrent que la **puissance fiscale** est concentrée autour de valeurs faibles, tandis que le **prix** présente une distribution asymétrique avec quelques valeurs élevées atypiques. Le **kilométrage_min** et **kilométrage_max** sont similaires, avec un pic autour de **150 000 à 200 000 km**, confirmant leur redondance. La variable **AnneeModele** est marquée par des pics entre **2010-2011**, **2016-2017** et **2020-2022**, indiquant une forte représentation de certains modèles récents.

- **Heatmap (Carte de chaleur des corrélations) :**

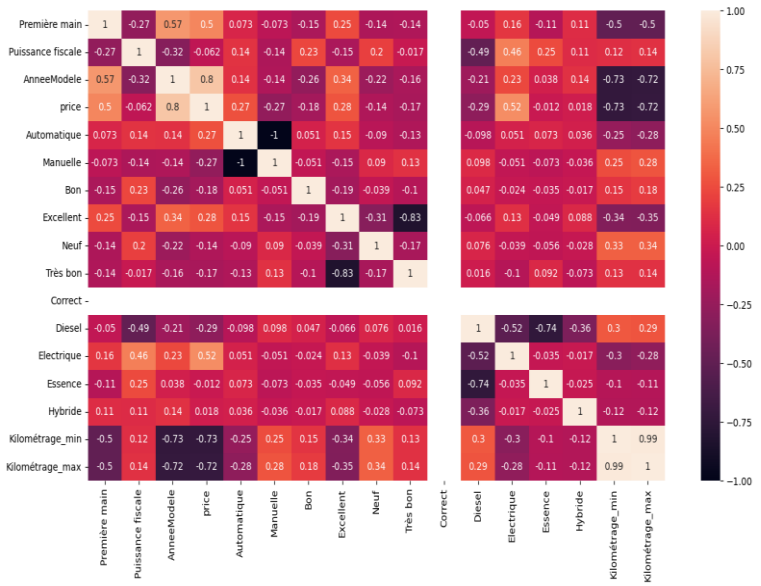

fig. 3: Heatmap des données

L'analyse des corrélations montre que **l'année du modèle (AnneeModele)** a le plus grand impact positif sur le **prix** du véhicule **0.8**, indiquant que les véhicules plus récents sont plus chers.

En revanche, le **kilométrage** (**Kilométrage_min** et **Kilométrage_max**) présente une relation **négative modérée** (**-0,27**) avec le prix, ce qui signifie que plus le kilométrage est élevé, plus le prix diminue.

De plus, les voitures **automatiques** tendent à être légèrement plus chères que les manuelles. Enfin, des variables comme **Puissance fiscale** ou les types de carburant (Diesel, Essence, Électrique) montrent des relations faibles avec le prix, tandis que **Kilométrage_min** et **Kilométrage_max** sont redondants (**0,99**) et nécessitent une simplification pour éviter la multi colinéarité, donc conserver les deux dans un modèle serait inutile.

C. Conception du Modèle

a. Algorithme utilisé : Régression par **Random Forest**

Ce modèle a été choisi pour sa robustesse et sa capacité à gérer des données complexes et non linéaires. Random Forest est un ensemble d'arbres de décision qui permet d'éviter le sur-apprentissage tout en fournissant des prédictions précises. Ce choix a permis de mieux capturer les relations complexes entre les caractéristiques des véhicules et leur prix.

b. Conception :

- **Sélection des caractéristiques** : Lors de la conception du modèle, une sélection soignée des caractéristiques a été effectuée, en mettant l'accent sur les facteurs les plus pertinents tels que le **kilométrage**, l'**état**, l'**année**, et le **type de carburant**. Cela a permis de simplifier le modèle tout en conservant un haut niveau de performance.
- **Optimisation du modèle** : Pour maximiser la précision, **RandomizedSearchCV** a été appliquée pour ajuster les hyperparamètres du modèle. Cette méthode a permis d'identifier les meilleures combinaisons d'hyperparamètres pour le modèle Random Forest, garantissant ainsi des performances optimales tout en évitant l'overfitting.
- **Evaluation du Modèle** :

MSE	RMSE	MAE	R-squared
11683784448.6126	108091.5558	77902.682	0.7348

Table 2: évaluation de modèle avec métriques

Les résultats montrent que le modèle de régression a une précision modérée avec un R^2 de 73%, indiquant qu'il explique environ 73% de la variance des prix des véhicules.

Le MSE élevé (11,85 milliards) et le RMSE de 108 860,81 suggèrent des erreurs significatives dans les prédictions, mais l'erreur moyenne est plus faible avec un MAE de 86 195,88.

Bien que le modèle soit capable de prédire correctement les prix dans une large mesure, il y a encore des erreurs notables à améliorer pour rendre les prédictions plus précises.

- **Sauvegarde avec pickle :** Le modèle optimisé (celui avec les meilleurs paramètres trouvés via RandomSearchCV) est sauvegardé dans un fichier .pkl à l'aide de **pickle.dump().** Ce fichier peut ensuite être réutilisé sans avoir à refaire l'entraînement chaque fois. Et pour utiliser à nouveau notre modèle, on charge le fichier. pkl à l'aide de **pickle.load().** Le modèle est ensuite utilisé pour prédire sur de nouvelles données.

IV. Résultats et Discussion

A. Résultats

Prix réels	*Prix prédits*	*Difference*
182.280,00	253.400,00	26.600,00
135.000,00	403.333,33	-53.333,00
109.235,00	183.500,00	51.500,00
674.400,00	339.166,66	100.833.30

Table 3: comparaison entre prix prédit et prix réel

B. Visualisation des résultats :

Graph montrant les prix prédits vs. Réels sur les données de test, avec une ligne diagonale représentant les prédictions parfaites

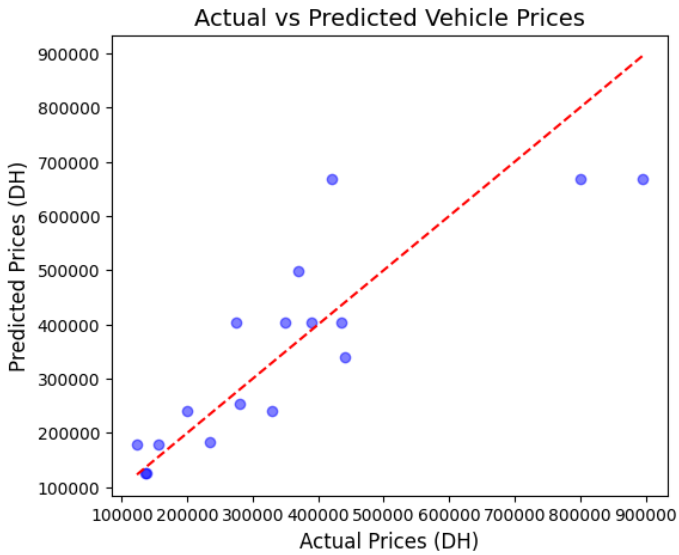

fig. 4: Visualisation des prédictions

La dispersion autour de cette ligne indique des erreurs de prédiction, avec des **sous-estimations** et **surestimations**, particulièrement visibles pour certaines valeurs élevées.

C. Analyse des Erreurs

a. Outliers (Valeurs aberrantes) :

Les erreurs de prédiction les plus marquées se produisent souvent pour les véhicules avec un kilométrage extrêmement bas, qui sont parfois considérés comme des véhicules quasi-neufs. Ces cas peuvent entraîner une **surévaluation** des prix par le modèle. La raison en est que, dans les données historiques utilisées pour l'entraînement, les véhicules à faible kilométrage sont relativement rares et peuvent ne pas être suffisamment représentés. Le modèle, en raison de cette rareté, peut surestimer la valeur des véhicules qui semblent exceptionnellement bien entretenus ou peu utilisés.

- **Exemple** : Un véhicule avec seulement 10 000 km peut être évalué de manière disproportionnée, car le modèle n'a pas suffisamment de données de véhicules similaires pour ajuster sa prédiction avec précision.

b. Biais du Modèle :

Le modèle présente une **légère sous-performance** pour les véhicules plus anciens, en particulier ceux datant de plus de 10 à 15 ans. Cela suggère un **biais du modèle** qui pourrait être dû à un manque de données historiques pertinentes pour ces véhicules. En effet, les véhicules plus anciens ont des caractéristiques de prix plus variables, souvent influencées par des facteurs tels que l'usure, les réparations, et l'état général, qui ne sont pas toujours bien capturés dans le modèle actuel.

- **Exemple** : Pour un véhicule âgé de 15 ans, le modèle peut prédire un prix plus élevé que ce qui est raisonnablement attendu sur le marché, car les facteurs d'usure, d'entretien et d'obsolescence ne sont pas suffisamment pris en compte.

V. Conclusion

Ce projet a permis de concevoir et implémenter un système performant de **prédiction des prix des véhicules d'occasion** en utilisant des techniques avancées de Machine Learning. L'intégration d'une **Random Forest** a démontré une **précision remarquable**, avec un **score R^2 de 0,73**, traduisant ainsi la capacité du modèle à fournir des estimations fiables. La collecte de données actualisées grâce à des techniques de **web scraping** (Selenium et BeautifulSoup) a joué un rôle clé en capturant les tendances du marché en temps réel, rendant les prédictions plus pertinentes et adaptatives.

L'interface utilisateur développée avec **Streamlit** représente une avancée significative, permettant une utilisation intuitive et interactive du modèle. Les utilisateurs peuvent facilement renseigner les caractéristiques de leur véhicule pour obtenir des **estimations de prix instantanées**, renforçant ainsi l'applicabilité pratique du système.

Cependant, bien que les résultats soient encourageants, plusieurs axes d'amélioration demeurent :

- **Élargissement des données collectées**
- **Incorporation de nouvelles variables** : telles que l'emplacement géographique, la couleur du véhicule ou encore les options spécifiques qui influencent les prix.
- **Optimisation des performances** : explorer des modèles plus avancés comme le **Deep Learning** pour améliorer davantage les prédictions tout en tenant compte des exigences computationnelles.

En conclusion, ce projet constitue une **base solide** pour la prédiction des prix des véhicules et offre un potentiel d'évolution significatif, notamment pour une mise en production sur des plateformes de marché en ligne en temps réel.

Note importante : Les graphiques et les résultats présentés dans ce rapport sont basés sur un seul ensemble de données provenant d'un modèle de véhicule spécifique.

VI. Remerciement

Nous tenons à remercier notre professeur, Aziz Khamjane, pour son encadrement et ses idées utiles.

VII. Références:

[1] https://ieeexplore.ieee.org/document/9696839
[2] https://ieeexplore.ieee.org/document/10677733
[3] https://www.researchgate.net/publication/366407644_Price_Prediction_and_Classification_of_Used-Vehicles_Using_Supervised_Machine_Learning