# Dédicaces

# Remerciement

# Table des matières

Table des matières

[Dédicaces 2](#_Toc197622303)

[Remerciement 3](#_Toc197622304)

[Table des matières 4](#_Toc197622305)

[Liste des abréviations 6](#_Toc197622306)

[Liste des figures 7](#_Toc197622307)

[Introduction générale 8](#_Toc197622308)

[Chapitre 1 : Contexte du projet 9](#_Toc197622309)

[I. Introduction 10](#_Toc197622310)

[II. Marché automobile marocain 10](#_Toc197622311)

[1. La prédiction des prix des voitures d’occasion 10](#_Toc197622312)

[2. Problématique et défis de la prédiction 11](#_Toc197622313)

[3. Techniques de prédiction utilisées : Régression 11](#_Toc197622314)

[III. Objectifs du modèle de prédiction. 12](#_Toc197622315)

[IV. Problématique et enjeux : 12](#_Toc197622316)

[1. Problématique 12](#_Toc197622317)

[2. Enjeux 13](#_Toc197622318)

[V. Conclusion : 14](#_Toc197622319)

[Chapitre 2 : Collecte et Préparation des Données 15](#_Toc197622320)

[I. Sources de données 16](#_Toc197622321)

[1. Définition de web scraping : 16](#_Toc197622322)

[2. Problématiques et comparaison des méthodes 16](#_Toc197622323)

[3. Explication détaillée d’approche académique de collecte des données automobiles : 17](#_Toc197622324)

[II. Exploration et Prétraitement des Données 19](#_Toc197622325)

[1. Chargement & Exploration Initiale des Données 19](#_Toc197622326)

[2. Nettoyage et Transformation Initiale 20](#_Toc197622327)

[3. Gestion des Données Catégorielles 22](#_Toc197622328)

[4. Gestion des Valeurs Manquantes 23](#_Toc197622329)

[5. Analyse de Corrélation 26](#_Toc197622330)

[6. Détection et traitement des outliers 27](#_Toc197622331)

[Chapitre 3 : Construction et Évaluation des Modèles 30](#_Toc197622332)

[I. Préparation des Données pour la Modélisation 31](#_Toc197622333)

[II. Ingénierie des Features 31](#_Toc197622334)

[III. Sélection et Optimisation des Modèles 32](#_Toc197622335)

[1. Forêts Aléatoires 32](#_Toc197622336)

[2. Boosted Tree 33](#_Toc197622337)

[3. Arbre de décision 33](#_Toc197622338)

[IV. Évaluation et Interprétabilité 34](#_Toc197622339)

[V. Résultats et Sélection Finale 34](#_Toc197622340)

# Liste des abréviations

**SOMACA** : Société Marocaine de Construction Automobile

# Liste des figures

# Introduction générale

Le marché automobile marocain connaît une croissance notable, portée par une demande soutenue en véhicules neufs et d’occasion. Cette dynamique s’accompagne d’une diversification des offres et d’une transformation progressive des comportements d’achat. Cependant, d’estimer le prix d’un véhicule demeure un défi majeur. Cette difficulté résulte de la multitude de facteurs influençant le prix (âge, kilométrage, marque, etc.) et des disparités entre les annonces.

Face à ce constat, ce projet vise à concevoir un modèle prédictif reposant sur des techniques de Machine Learning, capable d’estimer le prix d’un véhicule au Maroc. En s’appuyant sur des données réelles issues de plateformes locales telles qu’Avito.ma, notre objectif est de proposer un outil accessible qui facilite les décisions des acheteurs et des vendeurs, tout en contribuant à la transparence du marché.

Le projet repose sur plusieurs enjeux fondamentaux. D’un point de vue technique, il s’agit de collecter, nettoyer et traiter des données hétérogènes provenant de sources locales, tout en sélectionnant des algorithmes de Machine Learning pertinents et adaptés au contexte marocain. Ce travail nécessite une compréhension fine des spécificités du marché pour garantir des prédictions fiables.

Sur le plan économique, le modèle vise à réduire les asymétries d’information entre acheteurs et vendeurs, en leur fournissant une estimation objective de la valeur des véhicules. Cela permet de renforcer la transparence du marché et de favoriser des transactions plus équitables.

Enfin, sur le plan pratique, le projet comprend la conception d’une interface conviviale, destinée à un large public, permettant d’obtenir instantanément une estimation du prix d’un véhicule. Cet outil a pour vocation de rendre la technologie accessible et utile au quotidien.

Le rapport est structuré en cinq chapitres complémentaires :

Le Chapitre 1, intitulé « Contexte général », introduit le marché automobile au Maroc, en mettant en lumière ses caractéristiques, les défis liés à la fixation des prix et les motivations ayant conduit à la mise en place de ce projet.

Le Chapitre 2, « Techniques de base », présente les fondements du Machine Learning. Il introduit notamment les notions essentielles de régression, de traitement de données, ainsi que les algorithmes retenus pour leur pertinence dans le contexte.

Le Chapitre 3, « État de l’art », propose une revue des recherches existantes sur la prédiction de prix des véhicules. Il explore les méthodes déjà utilisées, les modèles pertinents ainsi que les outils technologiques mobilisés dans ce domaine.

Le Chapitre 4, « Collecte et Préparation des Données », décrit en détail le processus de récupération des données — notamment via la plateforme Avito.ma — ainsi que les étapes de nettoyage, de transformation et de sélection des variables qui serviront à la modélisation.

Enfin, le Chapitre 5, « Construction et Évaluation des Modèles », expose les différents modèles construits, leurs performances comparées à l’aide de métriques comme le MSE et le R², ainsi que le choix du modèle final validé selon des critères de robustesse et de précision.

# Chapitre 1 : Contexte du projet

## Introduction

Le marché automobile marocain connaît une forte croissance, notamment dans le secteur des véhicules d’occasion. Cependant, évaluer correctement le prix d’un véhicule reste complexe à cause de nombreux facteurs. Ce chapitre présente le contexte économique et technique du projet, ainsi que les problématiques liées à la prédiction de prix.

## Marché automobile marocain

Le secteur automobile marocain, lancé dans les années 1950 avec la création de la **SOMACA**, s’est progressivement imposé comme un pilier stratégique de l’économie nationale. Aujourd’hui, le pays abrite plus de 50 entreprises spécialisées dans la fabrication de composants automobiles — notamment des faisceaux électriques et des pièces de rechange — attirant des investisseurs internationaux tels que Renault, PSA et Lear Corporation. Ces acteurs alimentent principalement les chaînes d’approvisionnement européennes et africaines, avec environ 90 % de la production destinée à l’exportation.

En 2025, le marché automobile marocain connaît une croissance soutenue et une diversification remarquable. Les ventes ont augmenté de plus de 35 % au premier trimestre par rapport à 2024, plaçant le pays en tête du continent africain en termes de volume et de croissance. Le segment des véhicules particuliers domine toujours le marché, tandis que les véhicules électriques (VE) enregistrent une progression spectaculaire, avec une hausse de plus de 265 % sur la même période, stimulée par des incitations gouvernementales et l’expansion des infrastructures de recharge. Le marché combine une offre de véhicules assemblés localement et importés, répondant ainsi à une demande croissante et variée. Par ailleurs, le développement technologique, les investissements dans les technologies vertes, notamment l’électrification, ainsi que les partenariats avec des constructeurs mondiaux, s’inscrivent dans la stratégie nationale de transition énergétique. Enfin, la modernisation des infrastructures et le soutien gouvernemental contribuent à positionner le Maroc comme un acteur compétitif, innovant et tourné vers l’avenir.

### La prédiction des prix des voitures d’occasion

Le marché des véhicules d’occasion au Maroc est en pleine expansion, notamment grâce à la digitalisation des plateformes de vente. Néanmoins, la détermination du juste prix d’un véhicule reste une tâche complexe en raison de nombreux facteurs influents : marque, modèle, année, kilométrage, type de carburant, état général, etc.

Cette complexité génère une forte hétérogénéité des prix pour des véhicules parfois très similaires, ce qui peut entraîner des déséquilibres dans les transactions. C’est dans ce contexte qu’intervient la nécessité de développer des modèles de prédiction automatisés et fiables pour estimer le prix réel d’un véhicule d’occasion.

### Problématique et défis de la prédiction

La détermination du prix des véhicules d’occasion représente un défi complexe en raison du grand nombre de variables en jeu, souvent redondantes ou fortement corrélées, ainsi que de la nature non linéaire des relations entre certaines caractéristiques et le prix final. Ce défi est accentué par la présence de valeurs aberrantes, la variabilité conjoncturelle du marché et l’absence de normalisation des données — notamment des descriptions non structurées et des formats hétérogènes selon les sources. Ces contraintes réduisent la fiabilité des modèles prédictifs classiques. L’analyse devient ainsi à la fois technique et contextuelle, exigeant des approches robustes capables de gérer l’hétérogénéité des données et l’évolution constante des dynamiques économiques.

### Techniques de prédiction utilisées : Régression

Pour répondre à cette problématique, le Machine Learning offre une panoplie d’approches adaptées à la prédiction de valeurs numériques. Parmi les plus utilisées dans ce contexte, on retrouve :

* **Régression linéaire multiple** : Modèle de base permettant de quantifier l’impact de chaque variable sur le prix.
* **Régression Ridge et Lasso** : Versions régularisées qui réduisent le sur-apprentissage et sélectionnent les variables les plus pertinentes.
* **Arbres de décision et forêts aléatoires (Random Forest)** : Méthodes puissantes capturant les interactions complexes entre variables.
* **Gradient Boosting Machines (XGBoost, LightGBM)** : Modèles très performants pour des données hétérogènes.
* **Réseaux de neurones (Deep Learning)** : Utilisés pour modéliser des relations très complexes, surtout avec de grandes quantités de données.

L’utilisation de ces techniques permet non seulement de prédire les prix, mais aussi d’analyser les facteurs déterminants, de détecter des anomalies, et de fournir un outil d’aide à la décision fiable et transparent.

## Objectifs du modèle de prédiction.

Ce modèle de prédiction des prix des véhicules d’occasion a pour objectif de répondre à plusieurs besoins complémentaires. Tout d’abord, il vise à fournir une estimation précise du prix de vente, en prenant en compte ses caractéristiques techniques et historiques. Cela permet de proposer une fourchette de prix réaliste, offrant ainsi une orientation tant pour les acheteurs que pour les vendeurs. Ensuite, le modèle facilite une analyse détaillée des déterminants de prix, en identifiant les variables les plus influentes et en mesurant l’impact de chaque facteur. Par exemple, il peut quantifier l’effet d’une augmentation du kilométrage ou de l’âge du véhicule sur sa valeur.

Parallèlement, le modèle joue un rôle crucial en détectant les anomalies, en repérant les annonces dont le prix est soit nettement inférieur, soit supérieur à la valeur estimée. Enfin, ce modèle constitue un véritable outil d’aide à la décision, permettant aux utilisateurs de simuler différents scénarios (comme une réduction du kilométrage) et de disposer d’une base objective pour faciliter les négociations.

## Problématique et enjeux :

### Problématique

Le marché marocain des véhicules d’occasion est caractérisé par plusieurs défis majeurs. Tout d’abord, une **asymétrie d’information** persiste entre acheteurs et vendeurs, ces derniers ne disposant pas toujours des mêmes données, ce qui engendre des déséquilibres dans les transactions. Ensuite, on observe une **forte hétérogénéité des prix**, avec des écarts significatifs entre des véhicules pourtant similaires, principalement à cause de l’absence de standardisation dans l’évaluation. Enfin, le marché souffre de **pratiques opaques**, où certains vendeurs profitent du manque de transparence pour surévaluer leurs véhicules, tandis que des annonces frauduleuses, notamment concernant le kilométrage, induisent les acheteurs en erreur.

### Enjeux

#### Économiques :

Pour les acheteurs, l’objectif est d’éviter les surcoûts liés à des prix injustifiés, en ayant accès à une estimation fiable du prix d’un véhicule. Pour les vendeurs, il s’agit d’optimiser leurs marges sans risquer de décourager les acheteurs, en trouvant un équilibre entre compétitivité et rentabilité. En ce qui concerne les plateformes, l’enjeu est d’augmenter la confiance des utilisateurs, ce qui peut entraîner une plus grande fidélité des clients et attirer davantage de trafic sur leurs sites.

#### Sociaux :

L’un des enjeux sociaux majeurs est de réduire les risques de conflits lors des négociations en fournissant une estimation neutre du prix, ce qui permet de faciliter les échanges entre acheteurs et vendeurs. De plus, rendre le marché plus accessible aux primo-accédants, souvent désavantagés par leur manque d’expertise technique dans l’évaluation des prix. Cela permet de niveler les inégalités et de favoriser une participation plus large à ce marché.

#### Techniques :

L’un des enjeux techniques majeurs réside dans la gestion de la complexité des données, qui incluent des variables hétérogènes telles que des données numériques et catégorielles, ainsi que des corrélations non linéaires. De plus, il est essentiel d’assurer la scalabilité du modèle afin qu’il puisse s’adapter à l’évolution constante du marché, en prenant en compte de nouveaux modèles de véhicules ainsi que les effets de l’inflation.

#### Éthiques :

Les enjeux éthiques du modèle incluent la nécessité de garantir sa neutralité, en évitant tout biais, comme la survalorisation de certaines marques. Il est également crucial de protéger la vie privée des utilisateurs, en veillant à anonymiser les données utilisées lors de l’entraînement du modèle afin de respecter les principes de confidentialité et de sécurité des informations personnelles.

## Conclusion :

La prédiction des prix de voitures d’occasion est un défi justifié par la diversité des données et l’absence de références claires. Les techniques de régression du Machine Learning offrent des solutions adaptées, ouvrant la voie à une modélisation efficace que les chapitres suivants vont explorer.

# Chapitre 2 : Techniques de base

## Introduction

Ce chapitre présente les fondements théoriques et les outils essentiels mobilisés dans le cadre de ce projet. Il aborde les concepts clés du Machine Learning, notamment les types de modèles supervisés, les techniques de régression, ainsi que les notions de surapprentissage, de validation croisée et d’évaluation des performances. Ces éléments constituent la base méthodologique sur laquelle repose la construction du modèle prédictif du prix des véhicules.

## Apprentissage Automatique

L’apprentissage automatique, sous-domaine de l’intelligence artificielle, trouve ses origines dans les travaux pionniers du milieu du XXe siècle. Le terme a été conceptualisé par Arthur Samuel, notamment lors de ses recherches sur les jeux de dames, marquant le début d’algorithmes capables d’apprendre à partir de l’expérience, c’est-à-dire des données d’entrée.

Il permet à une machine d’apprendre à partir de données, sans être explicitement programmée pour chaque tâche. L’objectif est de construire des modèles capables de faire des prédictions ou de prendre des décisions basées sur des exemples passés. Dans le cadre de ce projet, le Machine Learning est utilisé pour estimer le prix d’un véhicule d’occasion à partir de ses caractéristiques (kilométrage, marque, année, carburant, etc.).

Les modèles d’apprentissage automatique peuvent être classés selon différents critères. Selon le paradigme d’apprentissage, on distingue l’apprentissage supervisé, l’apprentissage non supervisé et l’apprentissage par renforcement. Selon l’objectif d’apprentissage, les modèles se divisent en modèles discriminants et génératifs.

### Apprentissage supervisé

L’apprentissage supervisé utilise des données étiquetées pour apprendre la correspondance entre des entrées et des sorties connues. Parmi les exemples courants figurent les machines à vecteurs de support (SVM), les arbres de décision, et les perceptrons multicouches, réseaux de neurones capables de modéliser des relations complexes et non linéaires

### Apprentissage Non Supervisé

# Le Chapitre 3 : État de l’art

## Introduction

# Chapitre 2 : Collecte et Préparation des Données

**Collecte des données** : Processus système de rassemblement d'informations brutes à partir de sources variées (bases de données, web scraping, partenariats institutions) pour alimenter des modèles d'IA. Elle implique l'extraction structurée de données quantitatives (prix, kilométrage) et qualitatives (descriptions, images), tout en garantissant leur représentativité et leur conformité légale (RGPD) [[1]](#endnote-1).

**Préparation des données** : Phase de transformation des données collectées en un format exploitable, incluant le nettoyage (suppression des doublons, imputation des vailleurs manquantes), la normalisation (harmonisation des unités et formats) et l'enrichissement via le *ingénierie des caractéristiques* (création de variables composites). Cette étape réprésente jusqu'à 80 % du temps des projets IA et vue à optimiser la précision des préférences [[2]](#endnote-2).

## Sources de données

Pour ce projet, nous avons exploité des données exclusivement locales, centrées sur le marché automobile marocain et nous avons utilisé le web scraping pour le collecte de données.

### Définition de web scraping :

Le **web scraping** est une méthode de collecte de données en ligne qui consiste à extraire automatiquement le contenu de pages web à partir de leurs URL, généralement après la visite des utilisateurs (ex-situ). Il s'oppose aux collectes **in-situ**, qui capturent le contenu directement dans l'environnement de l'utilisateur (navigateur, cookies, sessions) en temps réel [[3]](#endnote-3).

### Problématiques et comparaison des méthodes

Le web scraping ex-situ présente plusieurs limites majeures. D'une part, il introduit des biais de mesure, car les contenus récupérés diffèrent souvent de ceux effectivement consultés par les utilisateurs, notamment en raison de paywalls, de personnalisations ou du caractère dynamique des pages. D'autre part, les délais temporels (par exemple, un scraping effectué 30 à 90 jours après la visite) aggravent cnes erreurs, avec une divergence pouvant atteindre +6,5 %. Ces biais ne sont pas uniformes : certains types de contenus, comme les articles politiques protégés par des paywalls, sont plus affectés que d'autres (par exemple, les informations trompeuses).

Contrairement à l'approche ex-situ, qui repose sur des serveurs externes et ignore souvent le contexte utilisateur (cookies, abonnements), la collecte in-situ capture fidèlement l'expérience réelle des internautes. Cependant, cette méthode est techniquement plus complexe à mettre en œuvre, ce qui limite son accessibilité pour les chercheurs [[4]](#endnote-4).

### Explication détaillée d’approche académique de collecte des données automobiles :

#### Extraction JSON optimisée via l'analyse du store Redux

Le script exploite la structure de données interne du framework Next.js en extrayant le JSON intégré dans la balise <script id="\_\_NEXT\_DATA\_\_">. Cette approche contourne l'instabilité du DOM en accédant directement au store Redux (data.props.pageProps.initialReduxState), où les annonces sont prétraitées côté serveur avant le rendu client. L'utilisation de params.secondary[] permet un mapping hiérarchique des métadonnées techniques (marque, modèle, année), évitant les erreurs de parsing liées aux sélecteurs CSS. Cette méthode garantit une cohérence structurelle même lors des mises à jour front-end mineures.

#### Normalisation et encodage des données hétérogènes

Le traitement des données intègre une **normalisation monétaire** via la dissociation explicite des valeurs numériques (price.value) et des devises associées (price.currency), permettant des analyses comparatives inter-marchés. La **géolocalisation standardisée** est obtenue par découpage sémantique des entrées utilisateur (location.split(",")), isolant systématiquement la ville primaire malgré les formats variables. Parallèlement, un **encodage sémantique** transforme les indicateurs qualitatifs (is\_premium, is\_urgent) en variables binaires (0/1), facilitant leur intégration dans les pipelines d'apprentissage automatique tout en préservant leur signification opérationnelle. Cette triple approche assure une cohérence sémantique et structurelle des données collectées, répondant aux exigences des modèles prédictifs tout en atténuant les biais liés aux formats hétérogènes.

#### Persistance incrémentielle et intégrée des données

Le système de sauvetage implémente une **stratégie de versionnage implicite** pour assureur la cohérence longitudinale des jeux de données. La vérification précalable de l'existence du fichier via os.path.exists prévient les conflits d'écrasement en isolant les jeux historiques des nouvelles entrées, La fusion sécurisée via pd.concat avec ignore\_index=True garantit l'unité des index tout en conservant les relations sémantiques entre les droits (ex. véhicule ↔ options), éliminant les doublons structures liés aux demandes répétées, L'export Excel intègre système les URLs sources (urljoin), fourni un traçabilité complète pour audit manuel ou rejejeu des demandes, conforme aux principes FAIR de reproduction scientifique, Cette approche hybride (fichier + métadonnées embarquées) équilibre simplicité d'implémentation et rigueur académique, s'alignant sur les bonnes pratiques de persistance incrémentielle observées dans les systèmes de gestion de dons critiques, Mécanisme de prévention d'écrasement inspiré des stratégies de sauvegarde incrémentielle, Alignement avec les méthodes de clustering incrémental pour données persistantes[,](http://www.fastercapital.com/fr/contenu/Strategies-de-persistance---regroupement-de-donnees---une-methode-d-organisation-des-donnees-persistantes.html) Respect des exigences de la traite des *documents de données* académiques, Approche similaire à l'actualisation incrémentielle des modèles Power BI, Méthodologie compatible avec les architectures de bases de données convergées.

#### Développement complémentaire :

La logique de versionnage, bien qu'implicite, s'apparente aux systèmes de *instantané* des bases de données transactionnelles, ù chaque export Excel représente un état figé du marché automobile à un instant *t*. L'absence de mécanisme de *diff* expliquer (ex. *git*) est compensée par la conservation des URLs sources, permanente une reproduction à l'identique des demandes futures. Ce choix méthodologique répond aux exigences des revues scientifiques en sciences des données, ù la traite des sources prime sur l'optimisation du stockage , Recommandations numéros des politiques de science ouverte pour la gestion des dons de recherche.

#### Stratégies anti-détection et conformité

Le script intègre des **mesures de masquage passif** pour minimiser les risques de blocage tout en conservant une architecture simplifiée. L'en-tête User-Agent simulé (Chrome 120) réduit l'empreinte numérique du grattoir en imitant un navigateur standard, évitant les filtrages basés sur les agents non conventionnels. Le **fixette étrangleur** (time.sleep(2)) respect empiriquement les limites de demandes d'Avito.ma, bien qu'un délai dynamique (ex. ajusté au temps de réponse du serveur) optimiserait le débit tout en évitant les surtaxes. L’**absence de procurations rotatiques** simplifie le déploitement mais expose au risque de blocage IP lors des collectes massives, nécessitant une future intégration de middlewares de rotation (résidentiels ou datacenter) pour distributeur les demandes. Cette approche équilibre opérationnel et discrétion, bien qu'elle néglige des vecteurs de détection avancés comme l'analyse complémentaire ou les pièges *pot de miel*.

#### Approfondissement méthodologique :

La structure actuelle privée la maintenabilité sur l'optimisation des performances, avec une séparation claire entre la logique d'extraction (scrape\_cars), de transformation (car\_info), et de persistance (pd.to\_excel). L'absence de parallélisation (ex. via ThreadPoolExecutor) est compensée par la fiabilité du flux séquentiel, cruciale pour les jeux de dons de référence.

## Exploration et Prétraitement des Données

### Chargement & Exploration Initiale des Données

Le jeu de données a été chargé à partir d'un fichier CSV concurrent 20 747 annonces de voiture d'occasion, avec 42 variables descriptives pour chaque véhicule. Les premières lignes du jeu de données ont été affichées afin de vérifier la structure des données et s'assureur que le chargement s'est déroulé correction. Ces variables incluent des informations clés telles que le prix, le kilomètre (mileage), la marque (brand), le modèle (model), l'année de fabrication (year), ainsi que diverses caractéristiques techniques et options (climatisation, toit ouvrant, airbags, caméra de recul, etc.).

Une exploration initiale a été réalisée pour mieux comprendre la nature et la qualité des données. Nous avons examiné le nombre total d'observations et de variables, ainsi que les types de données présentes : numériques (entiers et flottants), catégorielles (chaînes de caractères) et booléennes. Par exemple, la variable cible « prix » est de type numérique flottante, tandis que des colonnes comme « marque » ou « fuel\_type » sont catégorielles.

L'analyse des informations générales via la méthode df.info() a révélé la présence de valeurs manquantes dans plusieurs colonnes importantes, notamment dans la variable cible « prix », qui ne contient que 16 079 valeurs non nulles sur 20 747, ainsi que dans des variables explicatives dit que « kilométrage » ou « modèle ». Certaines colonnes, comme « old\_price », sont environnement vides et seront exclusives des étapes ultérieures.

Nous avons également effectué une analyse statistique descriptive des variables numériques à l'aide de la méthode df.describe(). Cette analyser un permis d'observateur la distribution des données, notamment les valeurs minimales, maximales, moyennes, médianes et l'écart-type, ce qui est essentiel pour faire des événements valeurs aberrantes ou anomalies. Par exemple, le prix moyen des véhicules est d'environ 217 000 MAD, avec une forte variation reflétant la diversité des annonces.

Cette exploration initiale a permis d'identificateur les anomalies et les points clés à traitre dans les états suivants, notamment la gestion des valeurs manquantes, le nettoyage des données, et la transformation des variables catégorielles en formats exploitables pour la modélisation.

### Nettoyage et Transformation Initiale

Après l’exploration initiale des données, une étape essentielle a consisté à nettoyer et transformer le jeu de données afin d’améliorer sa qualité et sa pertinence pour la modélisation.

#### Gestion des Valeurs Manquantes

Nous avons commencé par identifier les colonnes contenant des valeurs manquantes à l’aide de la méthode isnull().sum(). Pour optimiser la qualité du dataset, nous avons supprimé les colonnes où plus de 70 % des valeurs étaient manquantes, en utilisant un seuil calculé à partir de la taille totale du jeu de données. Cette approche permet de conserver uniquement les variables les plus informatives et évite d’introduire du bruit dans les modèles.

#### Suppression des Colonnes Non Pertinentes

Certaines colonnes telles que id, list\_id, phone, description, seller\_name, seller\_verified et title, ne contribuaient pas directement à l’analyse ou contenaient des informations redondantes ou personnelles. Elles ont donc été supprimées pour simplifier le dataset, réduire sa dimension et améliorer son efficacité.

#### Élimination des Doublons

Nous avons détecté et supprimé les lignes dupliquées afin d’éviter tout biais dans les analyses et l’entraînement des modèles. Cette étape a permis de garantir l’unicité des observations et d’améliorer la qualité globale des données.

#### Filtrage des Annonces Pertinentes

Le dataset initial comprenait des annonces de différents types, notamment « à vendre » et « à louer ». Nous avons filtré les données pour ne conserver que les annonces de voitures « à vendre », puis supprimé la colonne ad\_type devenue inutile. De plus, nous avons vérifié les catégories et types de véhicules présents, excluant les motos et autres catégories non pertinentes telles que « Autre Immobilier ». Après cette sélection, les colonnes category et type ont été supprimées car elles ne présentaient plus d’intérêt.

#### Nettoyage et Transformation des Attributs Numériques

Certaines variables nécessitaient une transformation spécifique pour être exploitables :

* **Kilométrage (mileage)** : Certaines valeurs étaient exprimées sous forme de plages (par exemple, « 50 000 - 70 000 »). Nous avons développé une fonction pour convertir ces plages en valeurs numériques moyennes, facilitant ainsi leur utilisation par les modèles.
* **Puissance fiscale (fiscal\_power)** : Cette colonne contient parfois des valeurs mélangées à du texte. Nous avons extrait uniquement les nombres entiers pertinents à l’aide d’expressions régulières, assurant une cohérence et une uniformité des données.

#### Standardisation des Données Catégorielles

Enfin, nous avons standardisé les valeurs des colonnes catégorielles importantes telles que fuel\_type, gearbox, condition, seller\_type et location. Cette standardisation a consisté à convertir les chaînes de caractères en minuscules et à supprimer les espaces superflus. Cette étape est essentielle pour éviter les doublons dus à des variations de casse ou de format (par exemple, « Essence » vs « essence »), et pour facilitateur le traitement ultérieur des variables.

Ces opérations de nettoyage et de transformation ont permis d'obtenir un jeu de données plus cohérent, homogène et prêt à être utilisé pour les phases suicidées de modélisation et d'analyse, garantissant ainsi la robustesse et la fiabilité de notre approche préventive.

### Gestion des Données Catégorielles

La gestion des variables catégorielles est une étape clé dans la préparation des dons, car la plupart des algorithmes de machine learning nécessitent des entrées numériques. Dans cette phase, nous avons procédé à un encodage adapté des différentes variables catégorielles présentes dans le jeu de données.

#### Identification des valeurs manquants

Nous avons d'abord identifié les valeurs manquantes dans les colonnes catégorielles à l'aide de la méthode isnull(). sum(). Cette étape permet de cibler les variables nécessitant un traitement spécifique avant l'encodage.

#### Encodage des variables binaires et ordinales

Pour la variable binaire seller\_type, nous sommes appliqués un encodage simple par mapping, en attribuant la valeur 0 aux vendeurs particuliers (privé) et 1 aux vendeurs professionnels(shop).  
L'attribut condition, qui explique l'état du véhicule, un été encodé de manière ordinale en respectant un ordre logique allant de « pour pièces » (0) à « neuf » (6). Cette représentation numérique reflète la qualité croissante du véhicule et permet au modèle de capturer l'importance relative de cette variable.

#### Encodage des variables catégorielles nominales

Variable  gearbox un été encodée par mapping binaire, avec 0 pour « manuel » et 1 pour «automatique».  
Pour la variable fuel\_type, qui comporte plus modalités, nous avons utilisés un encodage (variables indicatrices) à chaud via la fonction get\_dummies(). Cette méthode crée une colonne binaire pour chaque type de carburant, évitant ainsi toute hiérarchie artificielle entre les catégories.

#### Gestion des variables géographiques

Variable location contient de nombres villes au Maroc, ce qui peut traîner une forte cardinalité. Pour répertoire cette complexité, nous avons regroupés les villes en régions administratives selon une correspondance précise. Ce regroupement a permis de passer d'une granularité trop fine à un nouveau régional plus pertinent pour la compétence des prix.  
Suite, la variable location une variable d'été supprimée et remplacée par la nouvelle colonne qui nommé region. Celle-ci a été encodée également par one-hot-encoding, ce qui facilite son intégration dans les modèles.

#### Encodage différé pour certaines variables

Enfin, pour les variables à très forte cardinalité telles que brand, model et origin, nous avons décidé de différer leur encodage à la phase d'entrainement. Cette stratégie vise à éviter les problèmes de fuite de données (data leakage) et de surapprentissage (overfitting), notamment lorsque certaines modalités sont peu représentées dans le jeu d'entrainement.

Cette gestion rigoureuse des données catégorielles permet d'assureur une meilleure qualité des entrées pour les modèles prédictifs, tout en conservant la richesse et la pertinence des informations contenues dans les variables initiales.

### Gestion des Valeurs Manquantes

La gestion des valeurs manquantes est une étape cruciale pour garantir la qualité et la fiabilité des modèles préventifs. Dans cette phase, nous avons d'abord identifiés les colonnes compostant des données manquantes à l'aide de la méthode isnull().sum(). Nous avons porté une attention particulière aux variables clés tell que gearbox, year, condition, fiscal\_power, origin, doors, model, brand, mileage et la variable cible price.

Concernant la colonne year, qui contenait une valeur particulière « 1980 ou plus ancien », nous l'avons remplacé par la valeur numérique 1980 afin de standardiser le format. Ensuite, nous avons converti cette colonne en type numérique, en gérant les erreurs événements grâce à la fonction pd.to\_numeric().

La variable cible price étant essentiel pour la compétence, nous avant supprimé toutes les lignes où cette valeur était manquante, car il n'est pas possible d'imputer cette variable sans introduire de biais.

Pour mieux comprendre la distribution de la variable cible, nous avons visualisé son histogramme en limitante l'axe des prix à 700 000 MAD. Cette représentation graphique a permis de détecter la présence de valeurs extrêmes et de mieux appréhender la répartition des prix dans le jeu de données.

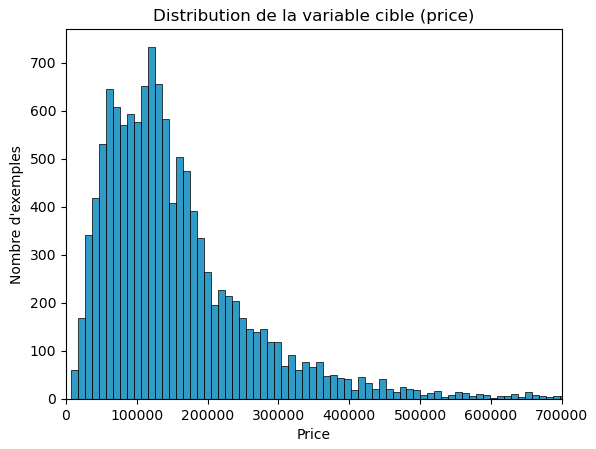


Figure 1:diagramme pour voir la distribution de prix

Pour les variables explicatives, nous adoptons différentes stratégies d'imputation en fonction de leur nature. Par exemple, pour la variable catégorielle gearbox, les valeurs manquantes ont été remplacées par la modalité la plus fréquente (mode), assurant ainsi une imputation simple et cohérente.

Pour la variable year, nous avons accord les valeurs manquantes par la valeur la plus fréquente, ce qui correspond à l'année la plus représente dans le jeu de données.

Pour les variables numériques continue ou ordinales telles que mileage, fiscal\_power, doors et condition, nous avons d'abord analysé leur distribution à l'aide d'histogrammes et de boxplots. Cette analyser visuelle a permis de détecter la présence d'éventuelles valeurs aberrantes et de compréhension la dispersion des données. Ensuite, les valeurs manquantes ont été imputées par la médiane, une méthode robuste face aux valeurs extrêmes.

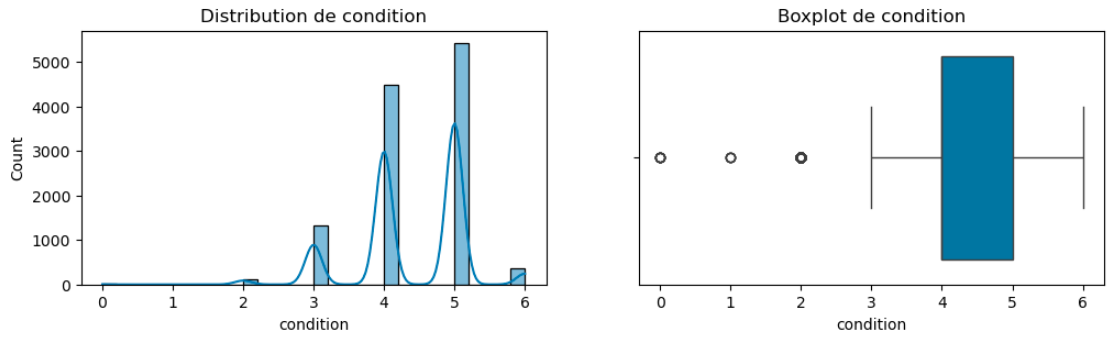


Figure 2:Boxplot et histogramme pour visualiser les valeurs numériques

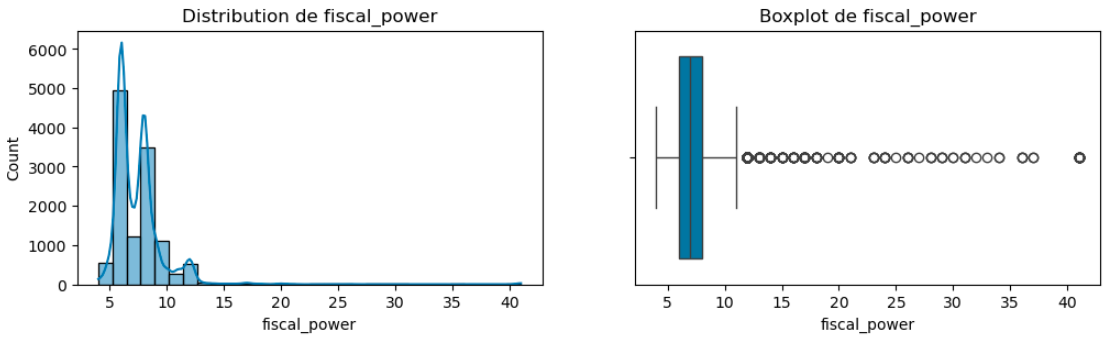


Figure 3:Boxplot et histogramme pour visualiser les valeurs numériques

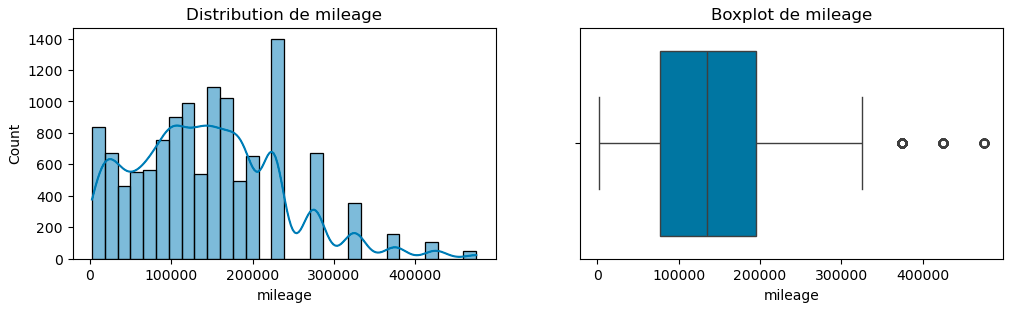


Figure 4Boxplot et histogramme pour visualiser les valeurs numériques

Au terme de ces opérations, nous avant vérifié que le jeu de données ne contenait plus de valeurs manquantes dans les colonnes essentielles, garantissant ainsi la qualité des données pour les phases suicidantes d'entrainement et d'évaluation des modèles.

### Analyse de Corrélation

L'analyse de corrélation vue à étudiant les relations linéaires entre les différentes variables numériques du jeu de données, afin d'identifier cellules qui sont fortes entre elles ou avec la variable cible. Cette étape est essentielle pour comprendre la structure des données, détecter la multicoléarité et orienteur la sélection des variables pour la modélisation.

Nous avons extrait les colonnes numériques du jeu de données et calculé la matrice de corrélation à l'aide de la méthode corr(). Cette matrice quantifie la force et la direction des relations linéaires entre chaque paire de variables par un coefficient compris entre -1 et 1.

Pour facilitateur l'interprétation, nous avons visualisé cette matrice sous forme de carte thermique (heatmap) avec la bibliothèque Seaborn. La heatmap utilise une échelle de couleurs allant du bleu (corrélation négative forte) au rouge (corrélation positive forte), avec des annotations numériques précises pour chaque coefficient. La taille de la figure a été adaptée (30x30 pouces) afin de rendre lisibles toutes les corrélations dans ce jeu de données riches en variables.

Cette analyser un permis d'identifiant les variables les plus corrélées avec la variable cible « prix », ainsi que les relations fortes entre certaines variables explicatives. Ces informations sont précieuses pour la sélection des caractéristiques pertinentes, la réduction de la redondance et l'amélioration de la performance des modèles prédictifs.

### Détection et traitement des outliers

Un outlier (ou valeur aberrante) est une observation qui s'écarte significativement des autres observations dans un ensemble de données. Ces valeurs peuvent être le résultat d'erreurs de mesure, d'erreurs de saisie ou de phénomènes rares mais réels.

#### Analyse Exploratoire Initiale

Notre analyse a débuté par une exploration visuelle des distributions à l'aide de boxplots et d'histogrammes. Cette approche nous a permis d'obtenir une première intuition sur la présence potentielle de valeurs aberrantes dans les variables numériques clés de notre jeu de données, notamment 'year', 'fiscal\_power', 'mileage' et 'price'. Les boxplots ont particulièrement révélé la présence d'observations extrêmes dans plusieurs variables, tandis que les histogrammes ont mis en évidence des distributions asymétriques, confirmées par le calcul du coefficient d'asymétrie (skewness).

#### Méthodologie de Détection des Outliers

**Nous avons** adopté une approche systématique pour identifier les valeurs aberrantes en utilisant la méthode de l'intervalle interquartile (IQR). Pour chaque variable numérique, nousavons calculé le premier quartile (Q1), le troisième quartile (Q3) et l'IQR. Les observations en dehors de l'intervalle [Q1 - 1.5×IQR; Q3 + 1.5×IQR] ont été considérées comme des outliers. Cette méthode robuste, moins sensible à la distribution des données que les méthodes basées sur les scores Z, s'est avérée particulièrement adaptée à notre jeu de données présentant des distributions non normales.

#### Stratégie de Traitement des Outliers

* **Pour les variables explicatives**

Pour les variables 'year', 'condition', 'fiscal\_power' et 'mileage', nous avons choisi une stratégie d'imputation par la médiane. Cette approche permet de conserver toutes les observations tout en réduisant l'impact des valeurs extrêmes. La médiane, étant moins sensible aux outliers que la moyenne, préserve mieux la distribution centrale des données. Ce traitement a été appliqué de manière sélective uniquement aux valeurs identifiées comme aberrantes.

* **Pour la variable cible 'price'**

La variable 'price' a fait l'objet d'un traitement différent. Après avoir identifié 691 outliers (soit environ **5,8%** des observations), nous avons pris la décision de les supprimer complètement du jeu de données. Cette approche plus radicale se justifie par le fait que ces valeurs extrêmes pouvaient biaiser significativement les résultats de modélisation ultérieurs. Une analyse complémentaire a révélé la présence de 459 observations avec des prix supérieurs à 300,000 unités, que nous avons également exclues après avoir vérifié qu'elles représentaient des cas exceptionnels non représentatifs de la population générale.

* **Validation des Résultats**

Après traitement, nous avons réexaminé les distributions via de nouveaux boxplots pour vérifier l'efficacité de nos approches. Le jeu de données final, réduit de 691 observations (de 11831 initiales à 11140 après traitement), présente désormais des distributions plus équilibrées tout en conservant la variabilité naturelle des données. Cette étape de validation visuelle confirme que les traitements appliqués ont atteint leur objectif de réduire l'impact des valeurs aberrantes sans pour autant dénaturer les caractéristiques principales du jeu de données.

* **Justification des Choix Méthodologiques**

La stratégie différentiée entre variables explicatives et variable cible repose sur des considérations statistiques pratiques. Pour les prédicteurs, la conservation des observations via l'imputation préserve la taille de l'échantillon tout en limitant l'effet des outliers. Pour la variable cible, la suppression s'imposait car ces valeurs extrêmes auraient pu fausser gravement les performances des modèles prédictifs. Le choix du coefficient 1.5 (et parfois 1.2) pour l'IQR a été ajusté après plusieurs itérations pour trouver un équilibre entre rigueur statistique et conservation d'un échantillon suffisamment large.

Après avoir effectué les étapes de **nettoyage, traitement des outliers et préparation des données**, nous avons sauvegardé le dataset prétraité dans un fichier Excel pour une utilisation ultérieure dans l'analyse ou la modélisation.

# Chapitre 3 : Construction et Évaluation des Modèles

## Préparation des Données pour la Modélisation

Notre processus de modélisation a débuté par la lecture du dataset prétraité (df\_pretraiter2.xlsx), ce qui nous a permis de garantir une parfaite cohérence avec les étapes de nettoyage et de prétraitement réalisées précédemment. Dans un premier temps, nous avons procédé à la séparation des variables en deux groupes distincts : les variables explicatives (X) comprenant toutes les colonnes à l'exception de 'price', et la variable cible (y) correspondant uniquement à la colonne 'price'. Cette étape fondamentale nous a permis de définir clairement les éléments utilisés pour la prédiction et celui à prédire.

Ensuite, nous avons divisé notre jeu de données en ensembles d'entraînement et de test selon une répartition 70%/30%, en utilisant la fonction train\_test\_split avec un random\_state fixé à 42. Ce choix nous a assuré une reproductibilité parfaite des résultats tout en conservant un échantillon de test suffisamment représentatif (30% des données) pour une évaluation finale rigoureuse des performances de nos modèles.

Par ailleurs, nous avons mis en place une stratégie sophistiquée pour gérer les éventuelles valeurs manquantes dans nos données. Pour les variables catégorielles (origin, brand et model), nous avons opté pour une imputation intelligente basée sur le mode conditionnel. Cette approche hiérarchique consistait à prioriser d'abord le mode par modèle (model), puis par marque (brand), et enfin à recourir au mode global lorsque les informations précédentes n'étaient pas disponibles. L'implémentation vectorisée de cette méthode nous a permis d'optimiser significativement les performances du traitement tout en maintenant une grande précision dans l'imputation des valeurs manquantes.

## Ingénierie des Features

Afin d'optimiser le traitement des variables catégorielles tout en évitant l'explosion dimensionnelle (problème connu sous le nom de "fléau de la dimension"), nous avons adopté une approche sophistiquée d'encodage. Plutôt que d'utiliser un simple **one-hot encoding** qui aurait considérablement augmenté la dimensionnalité de nos données, nous avons mis en œuvre un target encoding avec lissage (paramétré par α=10) pour les variables 'brand', 'model' et 'origin'. Cette technique combine astucieusement la moyenne globale du prix avec les statistiques spécifiques à chaque catégorie, permettant ainsi de préserver l'information sémantique tout en maintenant une représentation compacte. Une attention particulière a été portée à éviter tout **data leakage** en calculant ces encodages exclusivement sur l'ensemble d'entraînement avant de les appliquer à l'ensemble de test.

Parallèlement, pour les variables numériques, nous avons appliqué une standardisation via **MinMaxScaler** afin de normaliser l'échelle des différentes features. Cette transformation, essentielle pour de nombreux algorithmes de machine learning, permet d'améliorer la convergence et les performances en ramenant toutes les valeurs dans une plage commune [0, 1]. Dans une perspective de déploiement futur, le scaler a été soigneusement sauvegardé dans un fichier .pkl, garantissant ainsi qu'une normalisation cohérente pourra être appliquée aux nouvelles données lors de la phase de mise en production. Ces deux traitements complémentaires - target encoding pour les catégorielles et normalisation pour les numériques - ont constitué la base d'une représentation optimale de nos données pour l'apprentissage automatique.

## Sélection et Optimisation des Modèles

Notre étude a systématiquement comparé quatre familles d'algorithmes de machine learning modernes afin d'identifier la solution optimale pour notre problème de prédiction :

### Forêts Aléatoires

Nous avons implémenté une recherche aléatoire d'hyperparamètres (RandomizedSearchCV) explorant notamment :

* + Le nombre d'arbres (n\_estimators)
  + La profondeur maximale (max\_depth)
  + Le nombre de features considères (max\_features)
  + Le meilleur modèle a été sélectionné grâce à une validation croisée rigoureuse (5 folds), garantissant ainsi une robustesse optimale.

### Boosted Tree

Cette famille d'algorithmes a fait l'objet d'une investigation approfondie :

* + **HistGradientBoosting** : Privilégié pour sa vitesse d'exécution grâce à son binning automatique des données
  + **XGBoost** : Optimisé par réglage manuel du taux d'apprentissage (learning rate) et de la profondeur des arbres
  + **LightGBM** : Exploitant le paramètre num\_leaves pour un contrôle précis de la complexité du modèle
  + **CatBoost** : Apprécié pour sa gestion native des variables catégorielles et son grid search systématique

### Arbre de décision

Nous avons maintenu ce modèle comme baseline, avec une profondeur volontairement limitée à 8 pour prévenir tout risque de surapprentissage (overfitting). Cette approche conservatrice a servi de point de référence pour évaluer l'apport des méthodes plus sophistiquées.

Cette analyse comparative exhaustive nous a permis non seulement d'identifier l'algorithme le plus performant pour notre cas d'usage spécifique, mais aussi de comprendre les avantages respectifs de chaque approche en termes de précision, vitesse d'exécution et robustesse. La combinaison de techniques de validation croisée et de recherche systématique d'hyperparamètres a assuré l'objectivité de notre sélection finale.

## Évaluation et Interprétabilité

Pour évaluer de manière complète les performances de nos modèles prédictifs, nous avons mis en place une méthodologie rigoureuse combinant des mesures quantitatives et des analyses qualitatives. Les principales métriques utilisées incluent la RMSE (racine carrée de l'erreur quadratique moyenne) comme indicateur principal de performance, car elle pénalise davantage les grandes erreurs de prédiction tout en s'exprimant dans l'unité originale des prix, ce qui facilite son interprétation. Nous avons également calculé la MAE (erreur absolue moyenne) qui représente l'écart moyen en dirhams entre les valeurs prédites et réelles, offrant ainsi une mesure plus intuitive de l'erreur de prédiction. Le coefficient de détermination R² nous a permis de quantifier la proportion de variance expliquée par le modèle, avec une valeur se rapprochant de 1 indiquant une meilleure adéquation aux données.

Au-delà de ces métriques traditionnelles, nous avons approfondi notre analyse grâce à des techniques avancées d'interprétabilité des modèles. L'approche SHAP (SHapley Additive exPlanations), basée sur la théorie des jeux, nous a permis de quantifier précisément la contribution de chaque variable aux prédictions, à la fois globalement sur l'ensemble des données et localement pour des prédictions individuelles. Cette analyse a révélé l'importance relative des différentes caractéristiques dans notre modèle final. En complément, nous avons utilisé la méthode LIME (Local Interpretable Model-agnostic Explanations) qui fournit des explications locales en approximant le comportement du modèle autour de prédictions spécifiques. Ces deux approches combinées nous ont offert une compréhension approfondie du fonctionnement interne de nos modèles, garantissant non seulement de bonnes performances prédictives mais aussi la transparence nécessaire pour des applications pratiques. Cette analyse détaillée a notamment mis en évidence le rôle prépondérant de certaines variables clés comme l'année de production, la marque du véhicule et son kilométrage dans la détermination du prix.

## Résultats et Sélection Finale

Le modèle CatBoost final a démontré des résultats particulièrement convaincants sur nos différents jeux de données. Sur l'ensemble d'entraînement, il atteint une erreur absolue moyenne (MAE) de 11 488,73DH et une erreur quadratique moyenne (RMSE) de 16 485,30DH, avec un coefficient de détermination R² exceptionnel de 0,937, indiquant qu'il explique 93,7% de la variabilité des prix. Ces performances remarquables sur les données d'apprentissage se maintiennent de manière très satisfaisante sur l'ensemble de test, avec une MAE de 14 635,42DH et une RMSE de 22 633,46DH, tout en conservant un R² élevé à 0,883.

Cette légère différence entre les résultats d'entraînement et de test confirme que notre modèle parvient à éviter le surapprentissage tout en maintenant une grande précision prédictive. La performance sur l'ensemble de test, avec une erreur moyenne absolue inférieure à 15 000DH, démontre l'adéquation de CatBoost à notre problème de prédiction de prix. Ces résultats sont d'autant plus remarquables qu'ils ont été obtenus après une optimisation rigoureuse des hyperparamètres et une validation croisée stricte, garantissant ainsi la robustesse de nos conclusions.

1. ‘Collecte de données pour l’apprentissage automatique et l’IA : un guide complet - Shaip’, 23 November 2021 <https://fr.shaip.com/blog/ai-data-collection-buyers-guide/>; ‘Extraction de Données : Fondation Des Modèles IA’ <https://www.innovatiana.com/post/data-extraction-for-ai> [accessed 28 April 2025]; ‘Comment Optimiser l’utilisation d’une Data Source ? - IA School’ <https://www.intelligence-artificielle-school.com/ecole/technologies/optimiser-utilisation-data-source/> [accessed 28 April 2025]. [↑](#endnote-ref-1)
2. ‘L’IA Pour l’analyse de Données | Google Cloud’ <https://cloud.google.com/use-cases/ai-data-analytics> [accessed 28 April 2025]; ‘Services de Collecte de Données d’IA | Collecte de Données de Formation ML - Shaip’ <https://fr.shaip.com/offerings/data-collection/> [accessed 28 April 2025]; Aviv Regev and others, ‘The Human Cell Atlas’, ed. by Thomas R Gingeras, *eLife*, 6 (2017), p. e27041, doi:10.7554/eLife.27041. [↑](#endnote-ref-2)
3. Roberto Ulloa and others, ‘Beyond Time Delays: How Web Scraping Distorts Measures of Online News Consumption’, *Communication Methods and Measures*, 2025, pp. 1–22, doi:10.1080/19312458.2025.2482538. [↑](#endnote-ref-3)
4. Ulloa and others, ‘Beyond Time Delays’. [↑](#endnote-ref-4)