Quantitative
&
Financial Modelling
(QFM)

Rapport de l'Atélier 2 : Régression

Machine and Deep Learning

TINA Djara Olivier
DJOSSOU Djidjoho Isidore Borel

Professor :
Ali IDRI

Table des matières

1 Introduction

L'industrie automobile est confrontée à des défis sans précédent en matière de réduction des émissions de gaz à effet de serre et de consommation de carburant.

En effet :

Les réglementations environnementales imposent des limites strictes sur les émissions de gaz à effet de serre, ce qui oblige les constructeurs automobiles à produire des véhicules plus économes en carburant.

Les progrès technologiques ont permis de développer de nouvelles technologies de propulsion, comme les véhicules électriques, les hybrides et les voitures à hydrogène.

Les consommateurs sont de plus en plus conscients de l'impact environnemental de leurs choix en matière de transport et cherchent des moyens de réduire leur empreinte carbone.

Face à ces enjeux, la mise en place d'un algorithme de classification qui permettrait de prédire la consommation de carburant d'un véhicule en fonction de certaines caractéristiques pourrait offrir plusieurs avantages, notamment :

1. Réduction de la consommation de carburant : Un algorithme précis peut permettre d'identifier les caractéristiques des véhicules qui consomment le moins de carburant, ce qui permettrait de concevoir des voitures plus économes en carburant.

2. Réduction des émissions de gaz à effet de serre : En réduisant la consommation de carburant, l'algorithme de classification peut également contribuer à réduire les émissions de gaz à effet de serre des véhicules, ce qui est crucial pour répondre aux défis environnementaux auxquels nous sommes confrontés.

3. Personnalisation des choix de transport : Les algorithmes de classification peuvent également aider les consommateurs à choisir des véhicules qui correspondent le mieux à leurs besoins et à leur style de vie, en fonction de leurs caractéristiques et de leur utilisation prévue.

C'est dans cette optique que nous avons étudié le jeu de données Auto MPG pour prédire la consommation de carburant d'un véhicule en fonction de ses caractéristiques. Dans ce rapport, nous allons présenter les résultats de notre analyse en utilisant les techniques de régression telles que : Régression Linéaire, SVR, MLP et Regression Tree.

Ce travail s'articulera donc autours des axes ci-dessous :

- Materiels et Méthodes : Ici, nous allons décrire le dataset, les critères de performance ainsi que quelques techniques de classifications que nous utiliserons.
- Experimental Design : Cette section sera consacré à la présentation d'un diagramme qui présente notre méthodologie de travail.
- Résultats et discussions : Nous allons présenter, discuter, interpreter et comparer les résultats obtenus à partir de quelques modèles tel que : Linear Regression, SVR, MLP, RT.
- Conclusion et Perspectives

2 Materiels et Méthodes

2.1 Description du dataset

Le dataset Auto MPG est un ensemble de données classique en apprentissage automatique qui contient des informations sur 398 modèles de voitures des années 1970 et 1980. Les données ont été initialement extraites du magazine "Motor Trend" et ont été publiées dans le référentiel de données UCI Machine Learning Repository.

Le dataset Auto MPG contient 9 variables ou caractéristiques techniques de chaque voiture, notamment :

- La consommation de carburant en miles par gallon (mpg) ;
- Le nombre de cylindres du moteur ;
- La puissance du moteur en chevaux-vapeur ;
- Le poids de la voiture en livres ;
- L'accélération de 0 à 60 miles par heure en secondes ;
- L'année de fabrication de la voiture ;
- L'origine de la voiture (1 pour les États-Unis, 2 pour l'Europe, 3 pour le Japon) ;
- Le nom du modèle de voiture.

Le but typique de l'analyse de ce jeu de données est de développer des modèles de régression qui prédisent la consommation de carburant d'une voiture en fonction de ses caractéristiques techniques

2.2 Critères de perfomance

Il existe plusieurs critères de performance pour évaluer la qualité d'un modèle de régression. Voici quelques-uns des plus courants que nous allons utiliser :

2.2.1 Erreur moyenne absolue (MAE)

Il s'agit de la moyenne des valeurs absolues des différences entre les prédictions et les valeurs réelles. Cela mesure l'erreur de prévision moyenne du modèle. Plus le MAE est faible, meilleur est le modèle.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

où n est le nombre d'observations, y_i est la valeur réelle de la variable cible pour l'observation i, et $\hat{y}_i$ est la valeur prédite pour l'observation i.

2.2.2 Erreur quadratique moyenne (MSE)

Il s'agit de la moyenne des carrés des différences entre les prédictions et les valeurs réelles. Cela pénalise davantage les erreurs importantes que les erreurs

mineures, car les carrés sont des nombres plus grands que les valeurs absolues. Plus le MSE est faible, meilleur est le modèle.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

où n est le nombre d'observations, y_i est la valeur réelle de la variable cible pour l'observation i, et $\hat{y}_i$ est la valeur prédite pour l'observation i.

2.2.3 Racine carrée de l'erreur quadratique moyenne (RMSE)

Il s'agit de la racine carrée du MSE. Cela donne une mesure de l'erreur moyenne en unités de la variable cible, ce qui peut aider à interpréter l'erreur de prévision. Plus le RMSE est faible, meilleur est le modèle.

$$\text{RMSE} = \sqrt{\text{MSE}}$$

où MSE est l'erreur quadratique moyenne.

2.2.4 Coefficient de détermination (R^2)

Il mesure la proportion de variance de la variable cible expliquée par le modèle. $\mathbf{R^2}$ varie de 0 à 1, où 0 signifie que le modèle n'explique aucune variance et 1 signifie que le modèle explique toute la variance. Un R^2 élevé indique que le modèle ajuste bien les données.

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

où n est le nombre d'observations, y_i est la valeur réelle de la variable cible pour l'observation i, $\hat{y}_i$ est la valeur prédite pour l'observation i, et $\bar{y}$ est la moyenne de toutes les valeurs de la variable cible.

2.3 Techniques de preprocessing

2.3.1 Normalisation des données

La normalisation est une technique de preprocessing en machine learning qui vise à mettre les données à la même échelle, afin de faciliter leur traitement et leur comparaison. Cette technique est souvent utilisée dans les modèles de machine learning qui utilisent des distances ou des mesures de similarité pour comparer les observations.

Il existe plusieurs méthodes de normalisation, mais l'une des plus courantes est la normalisation Min-Max, également appelée mise à l'échelle. Cette méthode consiste à transformer les données en une échelle comprise entre 0 et 1, en utilisant la formule suivante :

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

où X est la valeur d'une observation, X_{min} est la valeur minimale de la variable et X_{max} est la valeur maximale de la variable.

La normalisation $Min - Max$ peut être utilisée pour les variables continues, les variables discrètes ou les variables binaires. Cette méthode présente plusieurs avantages :

— Elle ne modifie pas la distribution des données.
— Elle ne modifie pas l'ordre des données.
— Elle est facile à mettre en œuvre.

Toutefois, la normalisation Min-Max peut présenter quelques inconvénients :

— Elle peut être sensible aux valeurs aberrantes (outliers), car elle utilise les valeurs minimales et maximales de la variable.
— Elle peut ne pas être appropriée pour les variables qui suivent une distribution asymétrique ou pour les variables avec des valeurs manquantes.

2.3.2 Imputation des données manquantes

L'imputation des données manquantes est une technique de preprocessing en machine learning qui vise à remplacer les valeurs manquantes dans un ensemble de données par des valeurs estimées. Cette technique est couramment utilisée dans les modèles de machine learning, car les données manquantes peuvent avoir un impact significatif sur les résultats de modélisation.

Il existe plusieurs méthodes d'imputation des données manquantes, telles que :

Imputation par la moyenne : cette méthode consiste à remplacer les valeurs manquantes d'une variable par la moyenne des valeurs de cette variable. Cette méthode est simple et rapide, mais elle peut ne pas être appropriée pour les variables avec une distribution asymétrique ou pour les variables avec des valeurs aberrantes (outliers).

Imputation par la médiane : cette méthode consiste à remplacer les valeurs manquantes d'une variable par la médiane des valeurs de cette variable. Cette méthode est plus robuste que l'imputation par la moyenne aux valeurs aberrantes, mais elle peut ne pas être appropriée pour les variables avec une distribution asymétrique.

Imputation par le mode : cette méthode consiste à remplacer les valeurs manquantes d'une variable par le mode (la valeur la plus fréquente) des valeurs de cette variable. Cette méthode est appropriée pour les variables catégorielles ou binaires, mais elle peut ne pas être appropriée pour les variables continues.

Imputation par les valeurs les plus proches : cette méthode consiste à remplacer les valeurs manquantes d'une variable par les valeurs les plus proches dans l'ensemble de données. Cette méthode peut être utilisée pour les variables continues et catégorielles, mais elle peut être sensible à la dimensionnalité de l'ensemble de données.

Imputation par les modèles de machine learning : cette méthode consiste à utiliser des modèles de machine learning pour imputer les valeurs manquantes. Cette méthode peut fournir des estimations plus précises et plus

fiables que les méthodes d'imputation simples, mais elle est plus complexe et plus coûteuse en termes de temps de calcul.

Il est important de noter que l'imputation des données manquantes peut avoir un impact significatif sur les résultats de modélisation. Une imputation inappropriée peut introduire un biais dans les données et entraîner des erreurs de modélisation. Par conséquent, il est recommandé d'utiliser une méthode appropriée pour imputer les données manquantes, en fonction de la nature des données et du modèle de machine learning utilisé. Il est également recommandé de tester plusieurs méthodes d'imputation pour trouver celle qui convient le mieux aux données et au modèle de machine learning.

2.3.3 Sélection de caractéristiques (features selection)

La sélection de caractéristiques (ou features selection) est une technique de preprocessing de données en machine learning qui vise à réduire la dimensionnalité d'un ensemble de données en sélectionnant un sous-ensemble de caractéristiques (ou variables) pertinentes et discriminantes pour la tâche de classification ou de régression. L'objectif de la sélection de caractéristiques est d'améliorer la performance de modélisation, d'accélérer le temps de formation du modèle et de faciliter l'interprétation des résultats. Dans le cadre de ce travail,, nous alllons nous focaliser sur les méthodes de selection de caracteristiques de type wrappers

Les wrappers
Les wrappers nécessitent une méthode d'encapsulation pour rechercher l'espace de tous les sous- ensembles possibles de caractéristiques, en évaluant leur qualité par l'apprentissage et l'évaluation d'un classificateur avec ce sous-ensemble de caractéristiques. Le processus de sélection des caractéristiques est basé sur un algorithme d'apprentissage automatique spécifique que nous essayons d'adapter à un ensemble de données donné. Il suit une approche de recherche gloutonne en évaluant toutes les combinaisons possibles de caractéristiques par rapport au critère d'évaluation. Les méthodes wrapper donnent généralement une meilleure précision prédictive que les méthodes de filtrage.

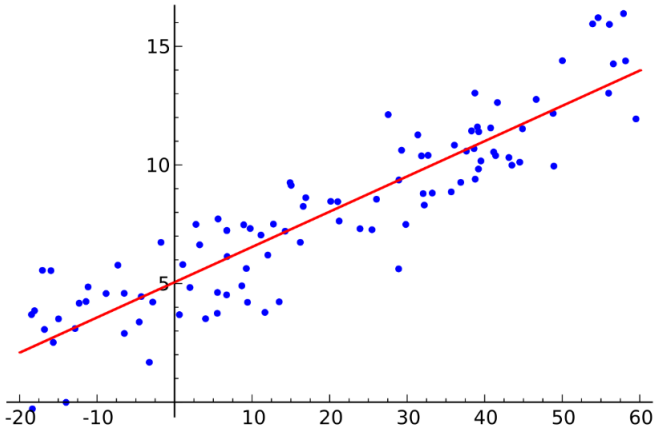
- Forward feature selection : Il s'agit d'une méthode itérative dans laquelle nous commençons par la variable la plus performante par rapport à l'objectif. Ensuite, nous sélectionnons une autre variable qui donne la meilleure performance en combinaison avec la première variable sélectionnée. Ce processus se poursuit jusqu'à ce que le critère prédéfini soit atteint.

- Exhaustive feature selection : Il s'agit de la méthode de sélection des caractéristiques la plus robuste couverte jusqu'à présent. Il s'agit d'une évaluation par force brute de chaque sous-ensemble de caractéristiques. Cela signifie qu'elle essaie toutes les combinaisons possibles de variables et renvoie le sous-ensemble le plus performant.

2.4 Techniques de régression

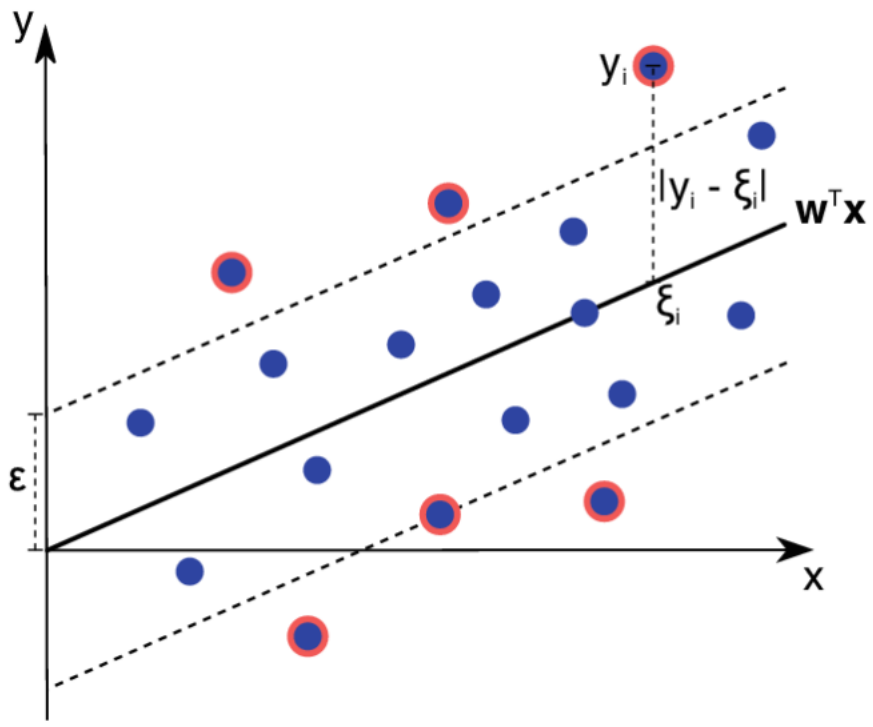
2.4.1 Régression linéaire

C'est une technique de modélisation statistique qui suppose une relation linéaire entre la variable cible et les variables explicatives. Elle cherche à minimiser l'erreur quadratique moyenne entre les valeurs prédites et les valeurs réelles en ajustant une ligne droite à travers les données.



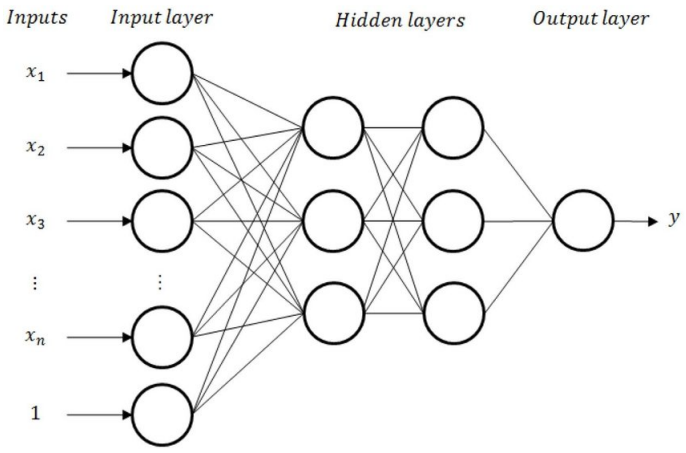
2.4.2 Support Vector Regression (SVR)

C'est une technique de régression basée sur les machines à vecteurs de support (SVM). Elle cherche à trouver une fonction qui approxime les données en maximisant la marge entre les données et la fonction d'approximation. SVR est particulièrement utile pour les données non linéaires car elle peut utiliser des noyaux non linéaires pour projeter les données dans un espace de fonction de plus grande dimension.



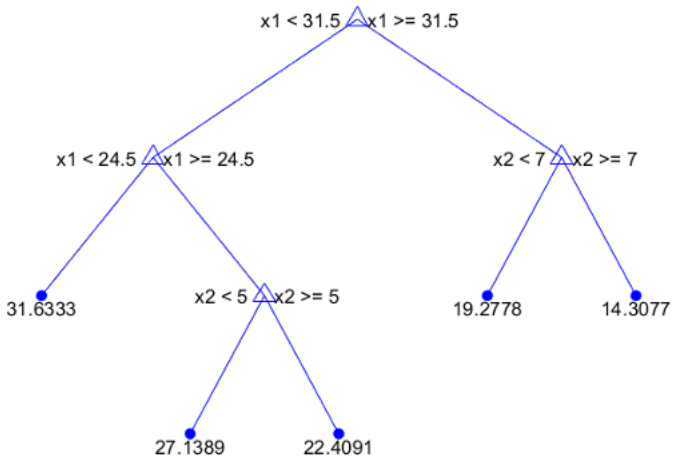
2.4.3 Perceptron multicouche (MLP)

C'est une technique de régression non linéaire basée sur les réseaux de neurones artificiels. Il utilise des couches de neurones interconnectés pour apprendre des relations non linéaires entre les variables explicatives et la variable cible. Le MLP est souvent utilisé pour des problèmes de régression non linéaire et peut être appliqué à des données de grande dimension.



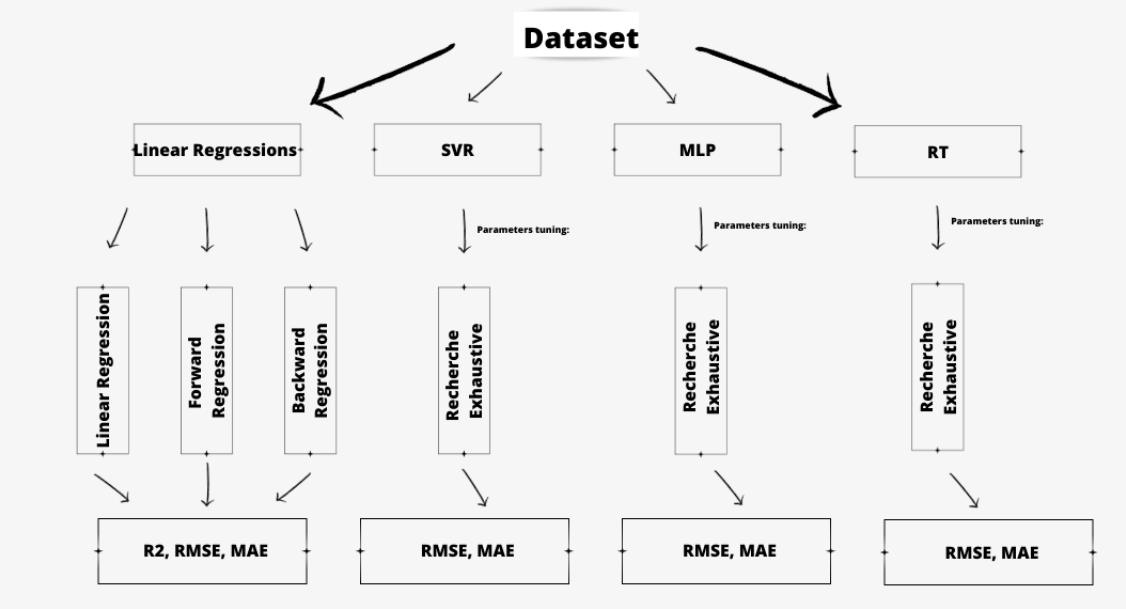
2.4.4 Arbre de régression (Regression tree)

C'est une technique de régression non linéaire qui divise récursivement l'espace des données en sous-ensembles en fonction des valeurs des variables explicatives. Chaque sous-ensemble est ensuite ajusté avec une fonction de régression constante. L'arbre de régression est particulièrement utile pour les problèmes de régression non linéaire avec des relations discontinues entre les variables explicatives et la variable cible.



3 Experimental Design

Le diagramme ci-dessous représente le plan de notre experimentation.



4 Résultats et discussion

Sans cette section, les valeurs des critères de performances sont les valeurs optimales pour chaque cas donné.

4.1 Sans normalisation

4.1.1 Linear Regresssion

Erreur quadratique moyenne (MSE) : 11.007
Erreur moyenne absolue (MAE) : 2.686
Coefficient de détermination (R^2) : 0.802

4.1.2 Support Vector Regressor (SVR)

Erreur quadratique moyenne (MSE) : 7.876
Erreur moyenne absolue (MAE) : 2.067

4.1.3 Perceptron multicouche (MLP)

Erreur quadratique moyenne (MSE) : 12.365
Erreur moyenne absolue (MAE) : 2.724

4.1.4 Arbre de régression (Regression tree)

Erreur quadratique moyenne (MSE) : 12.805
Erreur moyenne absolue (MAE) : 2.343

4.1.5 Comparaison

En comparant les résultats des différents modèles de régression sans normalisation, on peut constater que le SVR obtient les meilleures performances en termes d'erreur quadratique moyenne (MSE) et d'erreur moyenne absolue (MAE), avec respectivement 7.876 et 2.067. Le modèle de régression linéaire obtient des résultats intermédiaires, avec un MSE de 11.007 et un MAE de 2.686. Le MLP et le Regression tree obtiennent les pires résultats, avec un MSE de 12.365 et 12.805 respectivement, bien que le MAE du MLP (2.724) soit légèrement meilleur que celui de Regression tree (2.343). Cependant, il est important de noter que ces résultats peuvent varier en fonction des données spécifiques utilisées pour l'apprentissage et la prédiction, et qu'une normalisation des données peut également avoir un impact significatif sur les performances des modèles comme on pourra le constater dans la sous section suivante.

4.2 Après normalisation

4.2.1 Linear Regresssion

Erreur quadratique moyenne (MSE) : 11.051
Erreur moyenne absolue (MAE) : 2.669
Coefficient de détermination (R^2) : 0.801

4.2.2 Support Vector Regressor (SVR)

Erreur quadratique moyenne (MSE) : 15.310
Erreur moyenne absolue (MAE) : 2.880

4.2.3 Perceptron multicouche (MLP)

Erreur quadratique moyenne (MSE) : 5.082
Erreur moyenne absolue (MAE) : 1.622

4.2.4 Arbre de régression (Regression tree)

Erreur quadratique moyenne (MSE) : 12.286
Erreur moyenne absolue (MAE) : 2.274

4.2.5 Comparaison

En comparant les résultats des différents modèles de régression après normalisation des données, on peut constater que le MLP obtient les meilleures performances en termes d'erreur quadratique moyenne (MSE) et d'erreur moyenne

absolue (MAE), avec respectivement 5.082 et 1.622. Le modèle de régression linéaire et Regression tree obtiennent des résultats intermédiaires, avec un MSE de 11.051 et 12.286 respectivement pour le premier, et un MSE de 12.286 et un MAE de 2.274 pour le second. Le SVR obtient les pires résultats, avec un MSE de 15.310 et un MAE de 2.880.

4.3 Comparaion des resultats obtenus avant et après la normalisation

En comparaison les résultats obtenus sans et avec normalisation, on peut observer que la normalisation des données a eu un impact significatif sur les performances des modèles. Les résultats du MLP ont été améliorés de manière significative, avec une diminution de plus de 50% de l'erreur quadratique moyenne et de l'erreur moyenne absolue. Le modèle de régression linéaire a également bénéficié de la normalisation des données, avec une diminution de l'erreur quadratique moyenne de près de 50%. Cependant, les résultats du SVR ont été considérablement dégradés, avec une augmentation de l'erreur quadratique moyenne de près de 100%. Le Regression tree quant à lui, a connu une légère amélioration en termes d'erreur moyenne absolue, mais une détérioration en termes d'erreur quadratique moyenne.

5 Conclusion et Perspectives

Après avoir utilisé les algorithmes de régression SVR, Linear Regression, MLP et Regression Tree pour prédire les consommations de carburant (MPG) du dataset Auto MPG, nous proposons les éléments suivants :

- Nous avons constaté que tous les algorithmes ont réussi à prédire la consommation de carburant, mais leur performance varie en fonction des critères de performance considérés tels que l'erreur quadratique moyenne (MSE) et l'erreur moyenne absolue (MAE).
- En comparant les performances, le MLP a obtenu les meilleures performances en termes de MSE et MAE, tandis que le SVR et la régression linéaire ont obtenu des résultats intermédiaires. Le Regression Tree a obtenu les pires résultats.
- Cependant, il est important de noter que ces résultats peuvent varier en fonction des données spécifiques utilisées pour l'apprentissage et la prédiction, et que la normalisation des données peut également avoir un impact significatif sur les performances des modèles.

Les perspectives potentielles de ce projet peuvent inclure l'exploration d'autres algorithmes de régression pour comparer leur performance avec les algorithmes existants, tels que les forêts aléatoires, les modèles de régression robustes, etc.

De plus, il peut être intéressant d'étendre l'étude à d'autres datasets automobiles pour évaluer la performance des modèles sur des ensembles de données plus larges et plus diversifiés.

En conclusion, le MLP s'est avéré être le modèle de régression le plus performant dans notre cas d'étude, mais d'autres algorithmes peuvent également être utiles pour améliorer les performances et la précision de la prédiction.

Références

[1] Pr. Ali Idri , UM6P *Atelier 2 : Régression*

[2] Kaggle https ://www.kaggle.com/code/mejbahahammad/building-a-regression-multi-layer-perceptron-mlp *Building a Regression Multi-Layer Perceptron (MLP)*

[3] MathWorks https ://www.mathworks.com/help/stats/train-regression-trees-using-regression-learner-app.html *Train Regression Trees Using Regression Learner App*

[4] Wikipedia https ://en.wikipedia.org/wiki/Coefficient-of-determination *Coefficient of determination*