Projet Groupe

PYTHON
ANALYSE DE DONNEES POUR UNE IA

IPSSI Paris - [PRS] MIA4 26.2

COODIEN Govindarajen (Arbre de Décision/Rédaction)
BAKAYOKO Moussa (Réseau de Neurones)
TCHINDA D. Stevie C. (Clustering)

Table de Matière

1. Description Dataset

Notre dataset contient les données sur les voitures Allemandes de 2011 à 2021, collectées sur AutoScout24. Est est l'un des plus grands marchés automobiles européens pour les voitures neuves et d'occasion.

2. Arbre de Décision

(TP_DecisionTree.py)

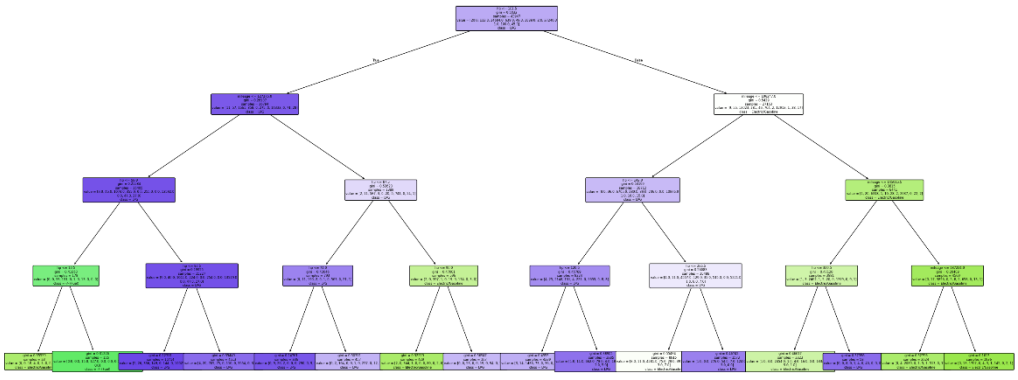
2.1 Chargement et Nettoyage des Données :

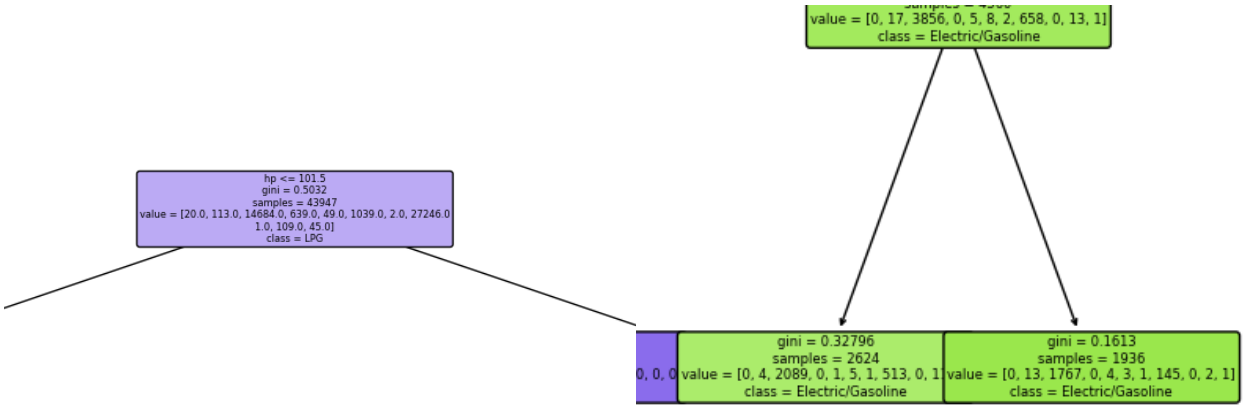
Le fichier cars.csv est chargé en utilisant pandas.
Le code retire les valeurs manquantes (NaN) et les doublons afin de garantir que les données utilisées sont propres et complètes.
L'index de la table est réinitialisé pour une meilleure organisation.

2.2 Préparation des Données pour le Modèle :

Deux ensembles de caractéristiques (features) sont créés pour entraîner deux arbres de décision différents :

Arbre de Décision (1) :

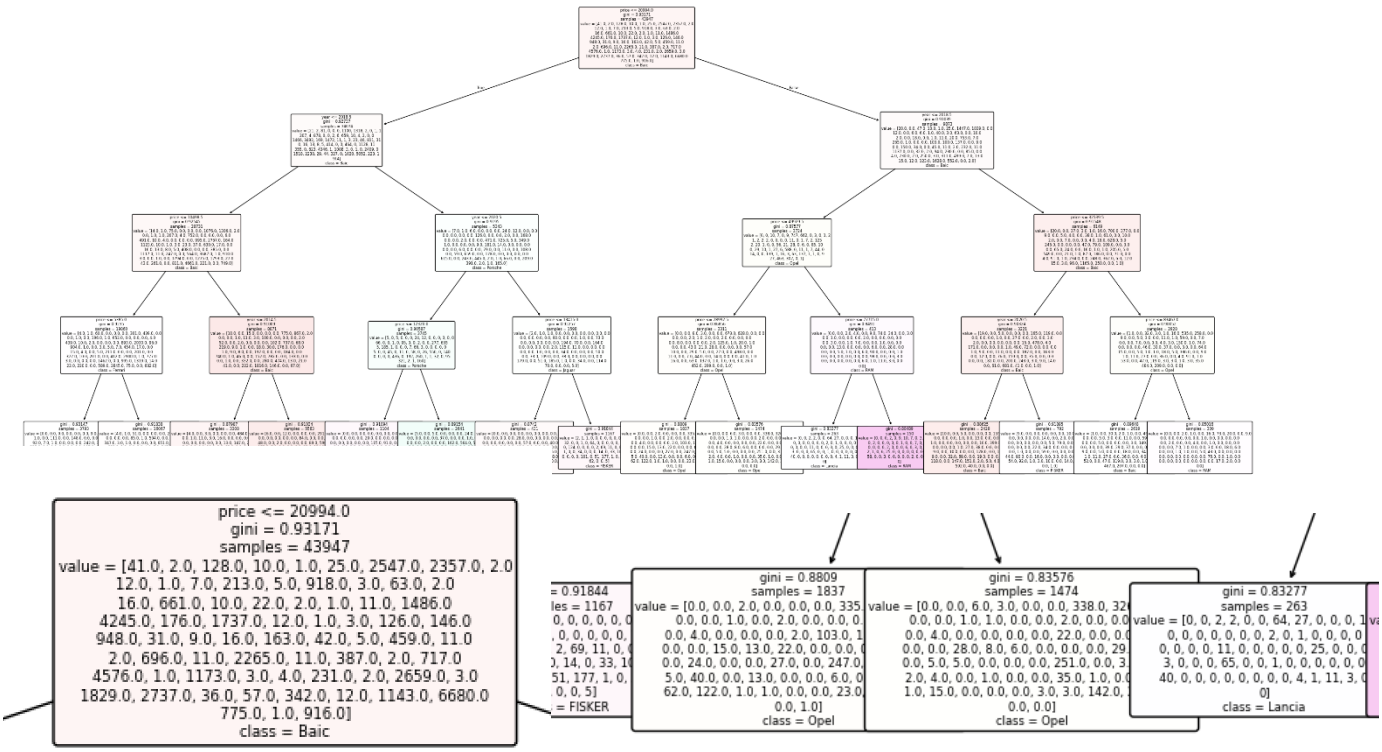




Premier Arbre de Décision : Utilise les colonnes mileage (kilométrage) et hp (chevaux) pour prédire le type de carburant (fuel).

Les colonnes mileage et hp sont les variables explicatives (X), et la colonne fuel est la variable cible (y).
Un classificateur DecisionTreeClassifier est créé avec une profondeur maximale de 4.
L'arbre est entraîné sur les données, puis visualisé. L'arbre montrent comment les décisions sont prises pour prédire le type de carburant.

Arbre de Décision (2):



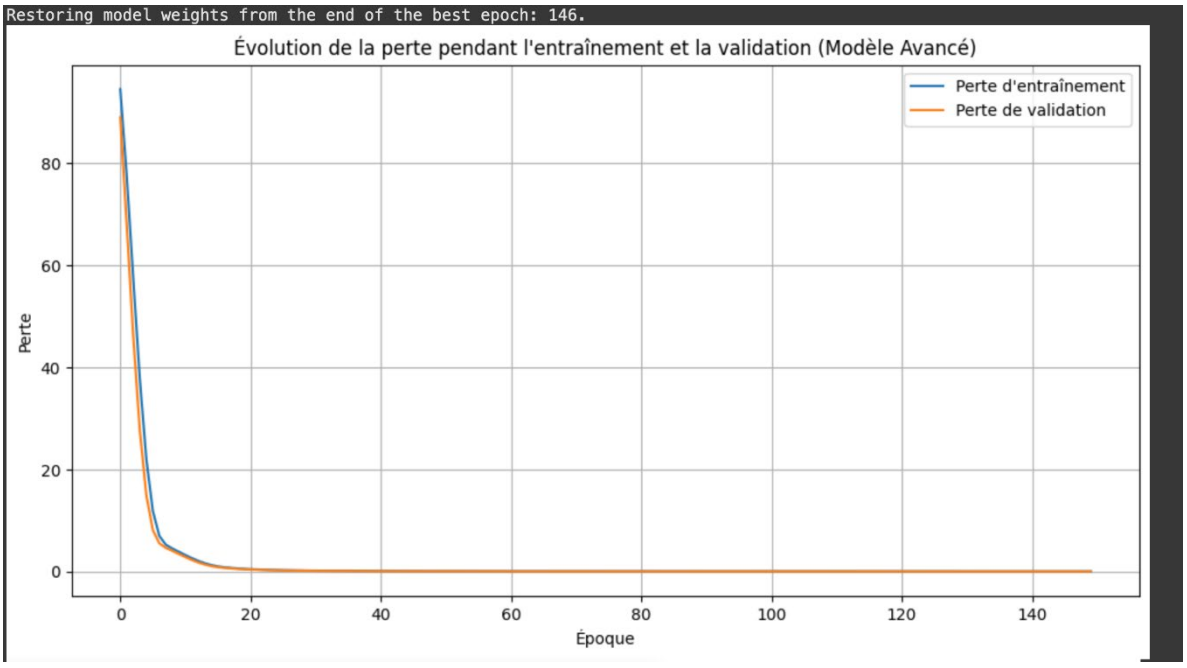
Deuxième Arbre de Décision : Utilise les colonnes price (prix) et year (année) pour prédire la marque de la voiture (make).

Les colonnes price et year sont utilisées comme variables explicatives (X_2) pour prédire la marque de la voiture (y_2).
Un deuxième classificateur DecisionTreeClassifier est entraîné et visualisé de la même manière.
Cela permet de comprendre comment les prix et les années influencent les prédictions de la marque de voiture.

3. Réseau de Neurones

Notre modèle de réseau de neurones est codé afin de prédire les prix des voitures par rapport la Puissance (HP), le carburant (FUEL), la marque (MAKE) et le model (models).

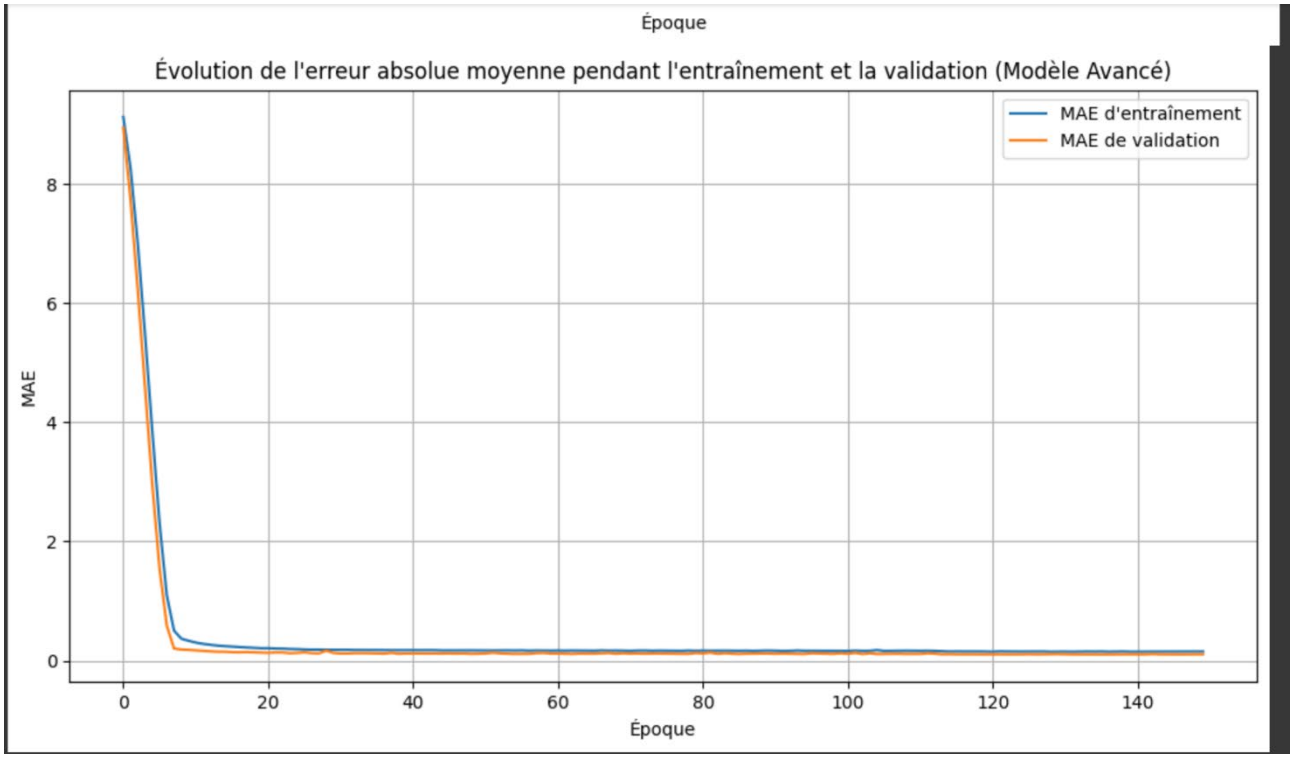
2.1. Évolution de la Perte Pendant l'Entraînement et la Validation



La courbe de perte (loss) pour l'entraînement et la validation diminue très rapidement, puis se stabilise à un niveau proche de zéro.
Cela indique que le modèle s'est très bien ajusté sur les données d'entraînement ainsi que sur les données de validation.

Conclusion sur la Perte :
Le fait que les deux courbes (entraînement et validation) se stabilisent à des valeurs proches et très faibles montres que le modèle a réussi à apprendre les relations présentes dans les données de manière très efficace. Cependant, une perte proche de zéro peut parfois indiquer que le modèle s'est surajusté (overfit) aux données d'entraînement. Heureusement, les courbes d'entraînement et de validation sont très proches, ce qui signifie que l'overfitting est probablement minimal.

2.2. Évolution de l'Erreur Absolue Moyenne (MAE)



Observation Générale :
Le MAE pour l'entraînement et la validation diminue rapidement au début, atteignant presque zéro avant de se stabiliser.
Cela montre que le modèle fait de très bonnes prédictions, avec une erreur très faible, sur les deux ensembles (entraînement et validation).
Conclusion sur le MAE :
Les valeurs de MAE proches de zéro indiquent que le modèle a une très bonne précision et parvient à prédire des valeurs très proches de la réalité, que ce soit sur les données d'entraînement ou de validation.

La similarité entre les courbes d'entraînement et de validation confirme que le modèle est bien généralisé.

Analyse Générale et Conclusion :
Bonne Généralisation : Les courbes pour l'entraînement et la validation sont très proches, ce qui signifie que le modèle généralise bien aux nouvelles données.
Absence Apparente d'Overfitting : Le modèle n'affiche pas de signes évidents de surajustement, ce qui est un très bon point. Cela est probablement dû à l'utilisation efficace de la régularisation L2, de Dropout, et de Batch Normalization.

Modèle Très Performant : Le modèle a une perte et un MAE proches de zéro, ce qui suggère une très grande précision sur les prédictions.
Recommandations Supplémentaires
Validation Croisée : Pour vous assurer que les performances sont robustes, je vous recommande d'effectuer une validation croisée avec K-Fold. Cela permettrait de vérifier la robustesse des résultats sur différentes parties des données.
Test sur Données Réelles : Vous pourriez également tester le modèle sur un jeu de données complètement différent (un ensemble de test jamais vu auparavant) pour vérifier qu'il fonctionne aussi bien que sur les données de validation.

4. Clustering

Étape 1 : Chargement et nettoyage des données

Le code charge les données depuis un fichier CSV, supprime les doublons et les lignes contenant des valeurs manquantes.
Cela garantit un ensemble de données propre et prêt pour l'analyse.

Étape 2 : Filtrage des données sur le carburant

Seules les voitures ayant comme carburant Diesel, Essence (Gasoline) ou Électrique (Electric) sont conservées.
Cela permet de restreindre l'analyse à ces trois types de carburant.

Étape 3 : Sélection et encodage des colonnes

Le code conserve les colonnes pertinentes (prix, modèle, carburant, transmission, puissance, année).
Les colonnes catégoriques sont converties en valeurs numériques avec LabelEncoder ou un mapping manuel.

Étape 4 : Clustering avec K-Means

L'algorithme K-Means est utilisé pour regrouper les voitures en 3 clusters. Chaque voiture est assignée à un cluster, ajouté sous forme d'une nouvelle colonne 'Cluster'.

Étape 5 : Analyse des clusters

Les Clusters regroupé les voitures par des caractéristiques similaire

```
Cluster 0:
size: 11456
Prix moyen: 18281.31
Puissance moyenne: 135.46
Voiture la plus  (ancienne/ressente): 2011/2021
Répartition carburant: Diesel=3915, Gasoline=412, Electric=7129

Cluster 1:
size: 15078
Prix moyen: 15314.28
Puissance moyenne: 132.45
Voiture la plus  (ancienne/ressente): 2011/2021
Répartition carburant: Diesel=5144, Gasoline=39, Electric=9895

Cluster 2:
size: 16035
Prix moyen: 15561.49
Puissance moyenne: 130.78
Voiture la plus  (ancienne/ressente): 2011/2021
Répartition carburant: Diesel=5625, Gasoline=188, Electric=10222
```

Différentes analyses sont effectuées sur les clusters, comme :
- Taille de chaque cluster
- Statistiques moyennes : prix, puissance, année
- Répartition des carburants (Diesel, Gasoline, Electric)

1. Analyse par répartition par Année :

```
Cluster 0 :
   Diesel: Année moyenne = 2016
   Gasoline: Année moyenne = 2017
   Electric: Année moyenne = 2016
Cluster 1 :
   Diesel: Année moyenne = 2015
   Gasoline: Année moyenne = 2020
   Electric: Année moyenne = 2016
Cluster 2 :
   Diesel: Année moyenne = 2015
   Gasoline: Année moyenne = 2019
   Electric: Année moyenne = 2016
```

2. Analyse par prix moyen et puissance moyenne catégorisé par type d'essence :

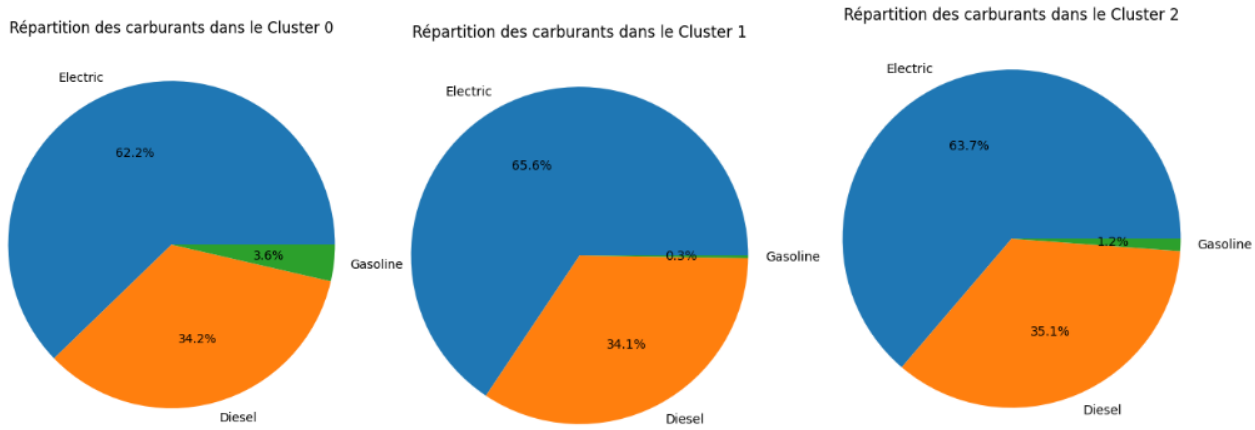
```
Cluster 0 :
   Electric: Prix moyen = 23990.08, Puissance moyenne = 166.68
   Diesel: Prix moyen = 20230.55, Puissance moyenne = 110.45
   Gasoline: Prix moyen = 15033.60, Puissance moyenne = 119.76
Cluster 1 :
   Electric: Prix moyen = 16579.35, Puissance moyenne = 155.36
   Diesel: Prix moyen = 28854.67, Puissance moyenne = 148.95
   Gasoline: Prix moyen = 14603.26, Puissance moyenne = 120.48
Cluster 2 :
   Electric: Prix moyen = 15326.17, Puissance moyenne = 141.07
   Diesel: Prix moyen = 28979.54, Puissance moyenne = 165.75
   Gasoline: Prix moyen = 15444.20, Puissance moyenne = 124.48
```

3. Analyse du Ratio entre la puissance et le prix :

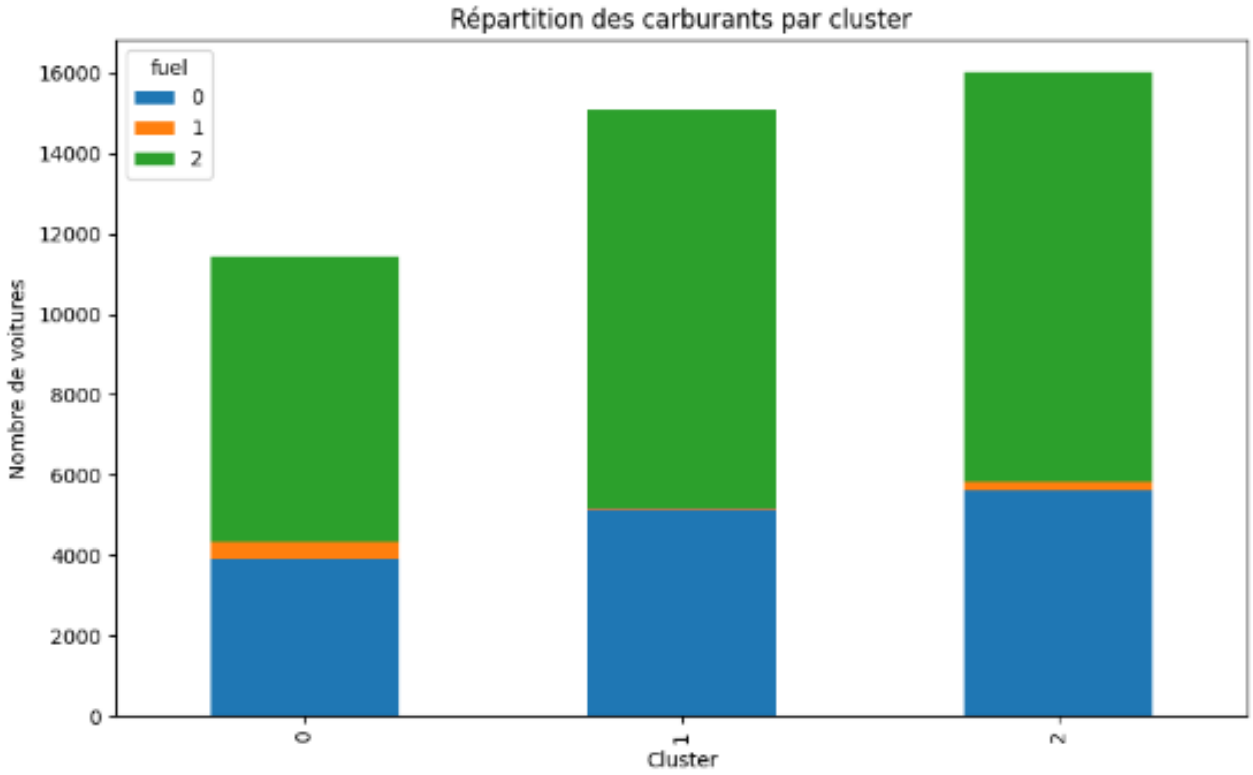
```
Cluster 0:
Ratio moyen Puissance/Prix: 0.0102
Cluster 1:
Ratio moyen Puissance/Prix: 0.0113
Cluster 2:
Ratio moyen Puissance/Prix: 0.0111
```

Étape 6 : Visualisation

Des graphiques sont générés pour visualiser la composition des clusters :
- Diagrammes circulaires pour la répartition des carburants de chaque cluster



- Graphiques en barres empilées pour comparer les clusters



Conclusion Cluster :

Domination des voitures électriques (Electric) :

Dans les trois clusters, les voitures électriques représentent la majorité. Cela peut refléter une montée en puissance des modèles modernes ou écologiques.

Part stable de Diesel :

Chaque cluster contient une part non négligeable de voitures Diesel, bien que celles-ci soient toujours inférieures aux voitures électriques.

Marginalité de Gasoline :

Les voitures à essence sont très rares, représentant une faible proportion dans chaque cluster.