ÉCOLE NATIONALE SUPÉRIEURE DES ARTS ET MÉTIERS – MEKNÈS

PROJET DATA MINING

Analyse prédictive des préférences clients

pour un concessionnaire automobile

Réalisé par :
Ayman Naaimi

Encadré par :
M. E. Zemmouri

Filière :
(DS2M)

Année universitaire :
2024 – 2025

جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

ENSAM
المدرسة الوطنية العليا للفنون و المهن
ÉCOLE NATIONALE SUPÉRIEURE D'ARTS ET MÉTIERS

Table des matières

Table des figures

1 Introduction

Cette analyse descriptive vise à mieux comprendre les caractéristiques des véhicules du catalogue ainsi que celles des clients. Elle permet d'identifier des tendances générales, de détecter d'éventuelles anomalies, et d'orienter les étapes de prétraitement et de modélisation, notamment dans le cadre d'une segmentation ou d'une prédiction.

2 Analyse Descriptive

2.1 Analyse du catalogue des véhicules

2.1.1 Distribution du prix des véhicules

La distribution du prix est centrée autour de 20 000 €, avec une concentration nette entre 15 000 € et 30 000 €. Quelques véhicules haut de gamme atteignent jusqu'à 70 000 €, mais ils restent marginaux.

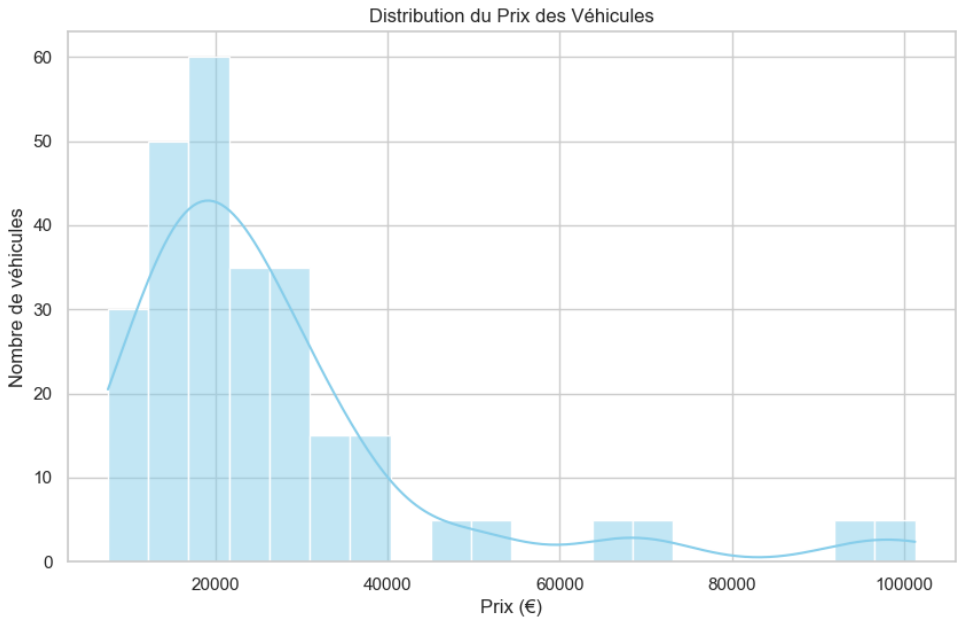

FIGURE 1 – Distribution du prix des véhicules

2.1.2 Distribution de la puissance

La puissance des véhicules est majoritairement comprise entre 75 et 130 chevaux DIN, indiquant une prédominance de modèles standards. Les pics sont visibles autour de 90 et 110 chevaux.

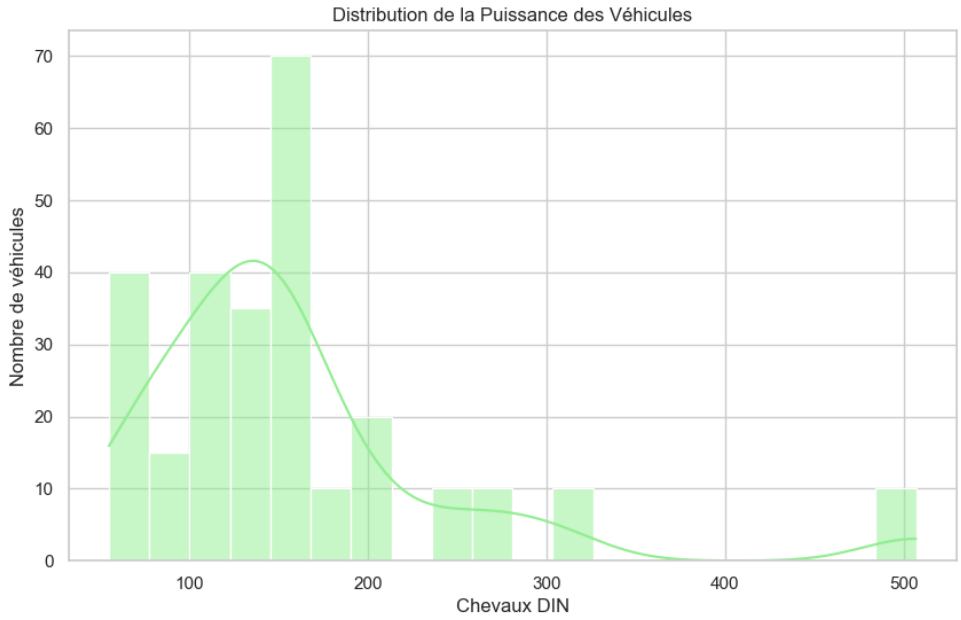

FIGURE 2 – Distribution de la puissance des véhicules

2.1.3 Répartition par marque

Les marques les plus représentées sont Renault, Peugeot et Citroën, suivies de Volkswagen. Cela reflète une forte présence de constructeurs français dans le catalogue.

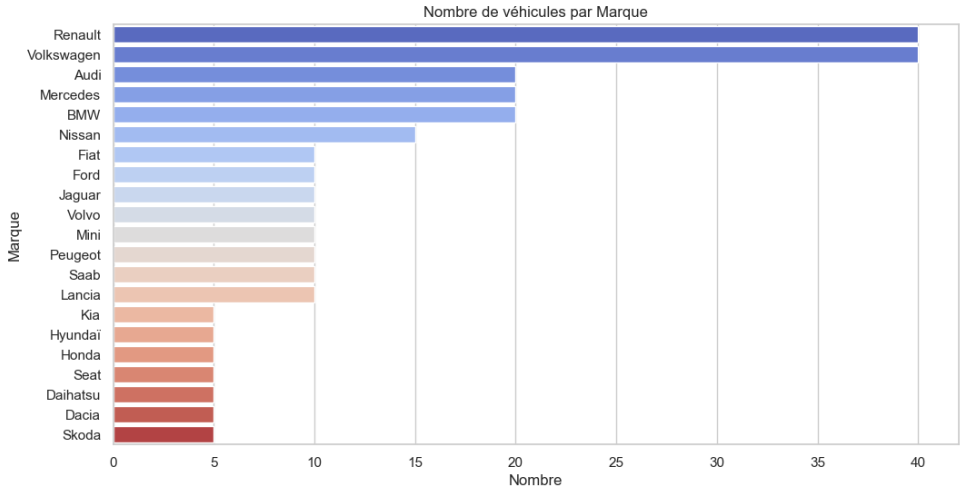

FIGURE 3 – Nombre de véhicules par marque

2.1.4 Prix en fonction de la longueur du véhicule

Les véhicules de type « Long » et « Très long » présentent des médianes de prix plus élevées, avec une forte variabilité. Les modèles « Courts » ont des prix plus homogènes et généralement inférieurs.

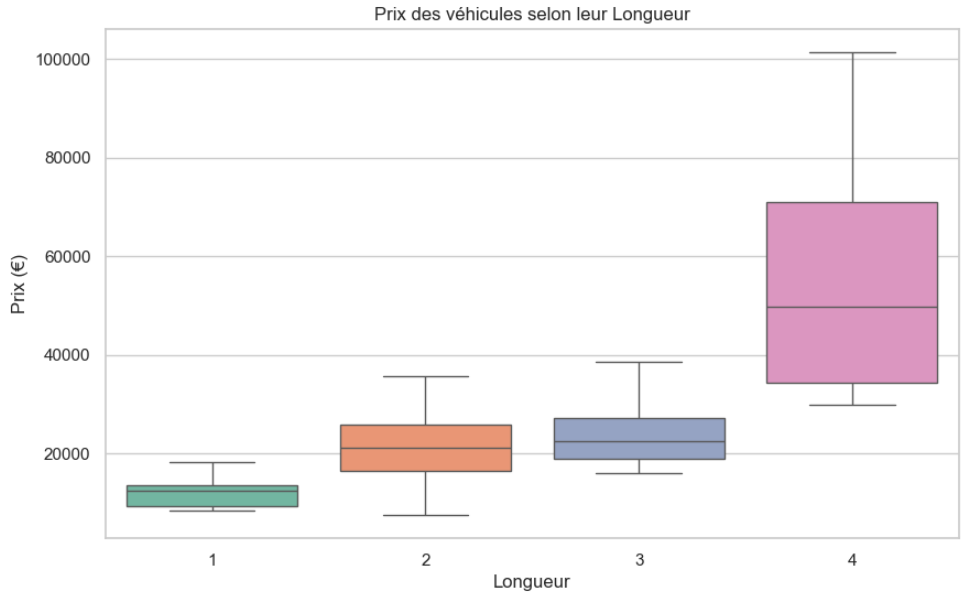

FIGURE 4 – Prix selon la longueur du véhicule

2.1.5 Couleur des véhicules

Les couleurs les plus fréquentes sont le gris, le noir, et le blanc, qui représentent à eux seuls la majorité du parc. Les couleurs vives comme le rouge ou le bleu sont plus rares.

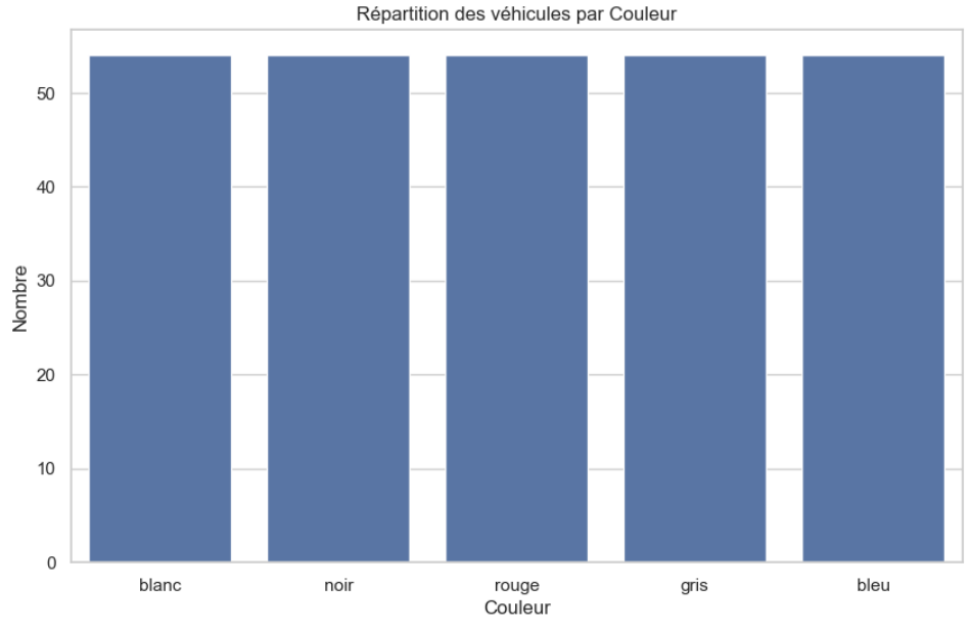

FIGURE 5 – Répartition des véhicules par couleur

2.2 Analyse des clients

2.2.1 Répartition par sexe

Les femmes représentent environ 30 % des clients, contre 70 % pour les hommes. La différence est significative.

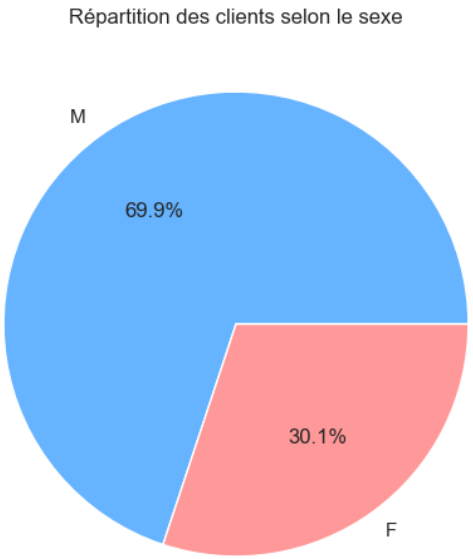

FIGURE 6 – Répartition des clients selon le sexe

2.2.2 Distribution de l'âge

La majorité des clients sont âgés de 20 à 50 ans. La distribution montre une baisse progressive après 50 ans.

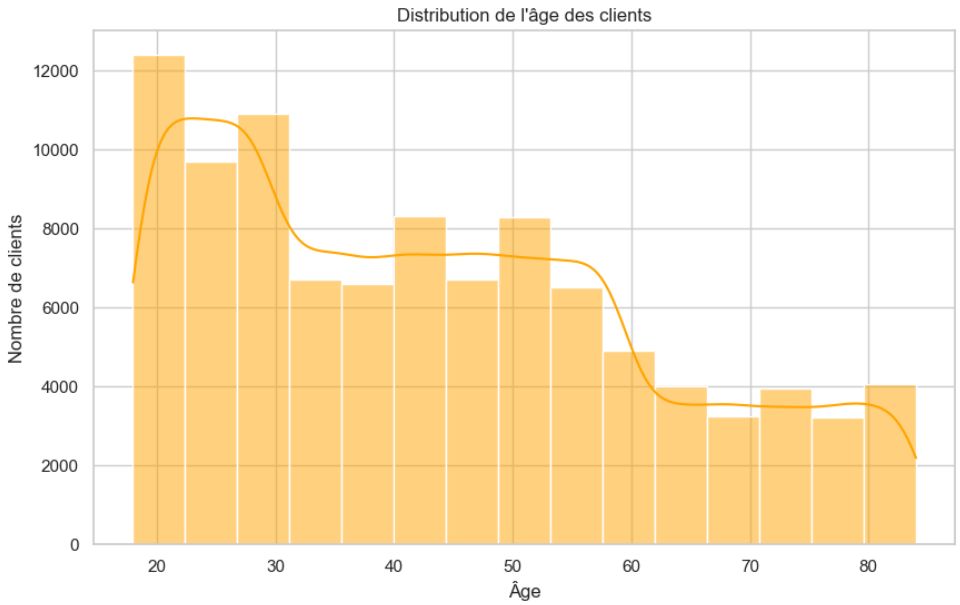

FIGURE 7 – Distribution de l'âge des clients

2.2.3 Situation familiale

Les clients en couple sont les plus représentés, suivis des célibataires.

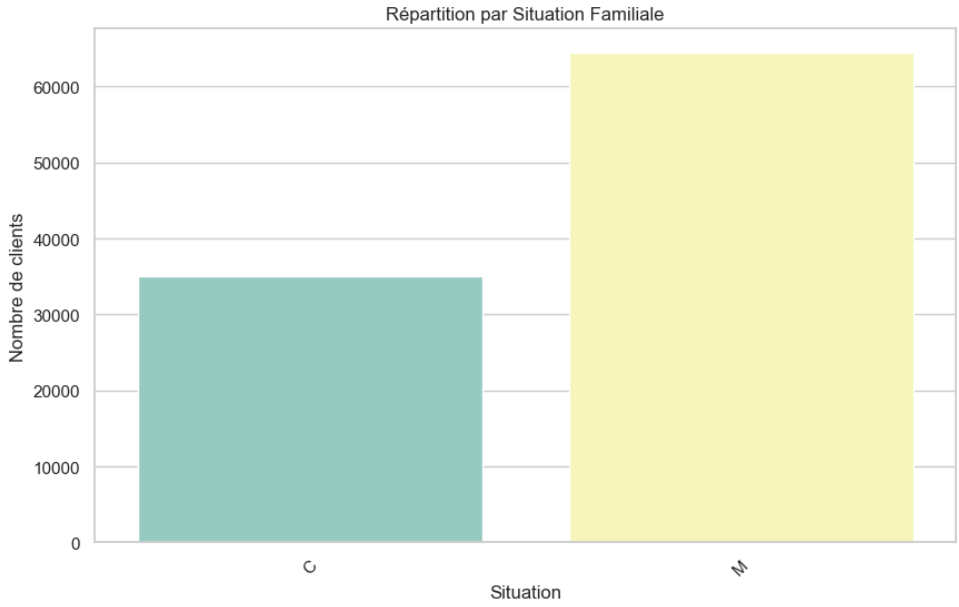

FIGURE 8 – Répartition par situation familiale

2.3 Synthèse de l'analyse descriptive

Cette analyse descriptive révèle plusieurs éléments clés :

— Le catalogue est dominé par des véhicules de milieu de gamme, de marques françaises, avec des puissances modérées et des couleurs neutres.

— Les clients sont principalement âgés de 20 à 50 ans, majoritairement en couple, avec une prédominance masculine.

— Ces caractéristiques offrent une base solide pour définir des segments clients et adapter les offres marketing lors de la phase de clustering ou de modélisation prédictive.

3 Clustering

3.1 Prétraitement des données initiales

Avant d'appliquer un algorithme de clustering sur les véhicules du catalogue, un prétraitement a été effectué dans Power BI afin d'améliorer la qualité et la pertinence des données utilisées :

— **Suppression des variables non pertinentes** : les champs `couleur`, `prix` et `occasion` ont été supprimés, car ils n'apportaient pas de valeur discriminante pour la catégorisation structurelle des véhicules.

— **Déduplication par nom de modèle** : pour éviter les doublons liés à la présence de plusieurs variantes du même véhicule, un seul enregistrement par modèle (champ `Nom`) a été conservé.

— **Normalisation implicite** : seules les variables structurelles pertinentes telles que `puissance`, `longueur`, `nombre de portes` et `nombre de places` ont été conservées pour décrire objectivement les véhicules.

Traitement de la variable `situationFamiliale` : Afin de simplifier l'information tout en conservant une logique métier pertinente, la variable `situationFamiliale` a été regroupée en deux catégories :

— **C** : regroupant les situations de type « Célibataire », « Divorcé(e) », « Seul(e) » ;

— **M** : regroupant « Marié(e) » et « En Couple ».

Ce regroupement permet de distinguer deux grands types de profils familiaux : individuels et en couple, ce qui peut influencer les besoins en véhicule (taille, nombre de places, etc.).

Traitement des variables catégorielles :

— **Sexe :** la variable `sexe` a été conservée telle quelle avec ses deux modalités : `M` (homme) et `F` (femme), considérées comme suffisamment représentatives et directement exploitables après encodage.

Les valeurs manquantes ont été supprimées afin de garantir la cohérence des données utilisées pour l'entraînement des modèles.

Ces étapes ont permis de préparer un jeu de données propre, épuré et représentatif pour appliquer efficacement un algorithme de clustering visant à regrouper les véhicules selon des critères techniques et dimensionnels.

3.2 Clustering des véhicules

L'objectif de cette étape était de regrouper les véhicules en catégories homogènes selon leurs caractéristiques techniques afin d'identifier des segments types : citadines, familiales, etc.

3.2.1 Prétraitement des données pour le clustering

Les opérations suivantes ont été réalisées :

— **Suppression des variables non pertinentes** : la couleur, l'occasion et le prix ont été exclus.

— **Encodage numérique** : la longueur a été convertie en variable numérique (1 = courte, 2 = moyenne, 3 = longue, 4 = très longue).

— **Normalisation** : les variables quantitatives ont été standardisées avec StandardScaler (sklearn).

— **variables sélectionnées** : la longueur, la puissance.

3.2.2 Algorithme utilisé

Nous avons utilisé l'algorithme **K-Means**, en testant plusieurs valeurs de k via la méthode du coude. Le meilleur compromis a été trouvé pour $k = 4$.

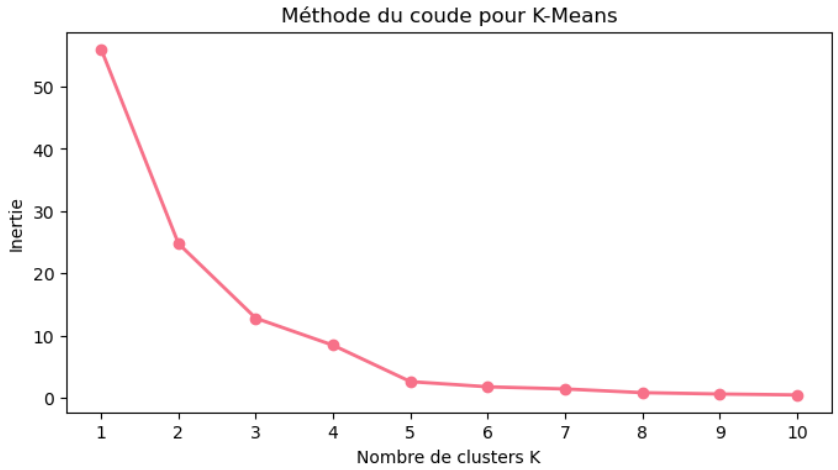

FIGURE 9 – Courbe du coude pour le choix du nombre de clusters

3.2.3 Interprétation des clusters

Après analyse des statistiques de chaque groupe, les étiquettes suivantes ont été attribuées :

— **Cluster 0** : *Familiales Puissantes* → puissance moyenne ∼287 ch, longueur très longue
— **Cluster 1** : *Compactes Polyvalentes* → puissance moyenne ∼129 ch, longueur moyenne
— **Cluster 2** : *Petites Citadines* → puissance ∼76 ch, longueur courte
— **Cluster 3** : *Compactes Dynamiques* → puissance moyenne ∼143 ch, longueur longue

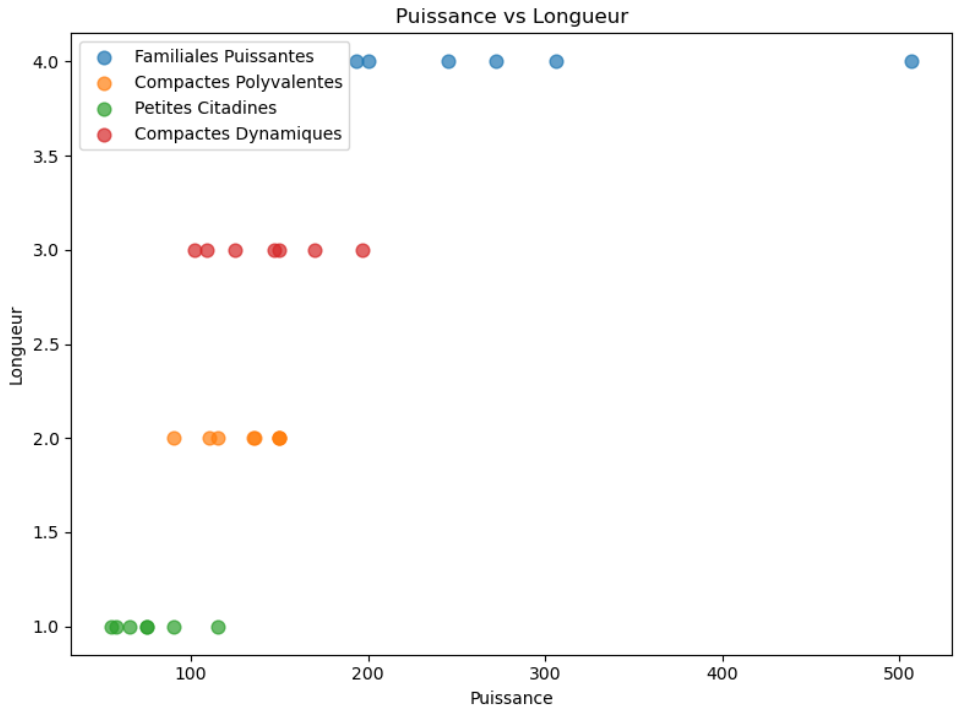

FIGURE 10 – Projection des clusters identifiés

3.3 Attribution des catégories aux données

3.3.1 Attribution aux immatriculations

Une jointure a été effectuée entre le fichier `Immatriculations` et le catalogue enrichi de la variable `catégorie`. La correspondance s'est faite via le champ `nom` du véhicule. Cela a permis d'attribuer une catégorie à chaque véhicule vendu cette année.

3.3.2 Fusion des données clients et immatriculations

Une seconde jointure a été réalisée entre le fichier `Clients` et le résultat précédent, sur le champ `immatriculation`. Le jeu de données final associe chaque client à la catégorie de véhicule achetée. Ce jeu constituera l'ensemble d'apprentissage pour la classification.

4 Classification

4.1 Modélisation : classification supervisée

4.1.1 Prétraitement des données

— Encodage des variables qualitatives : `sexe`, `situationFamiliale`, `2eme voiture`
— Imputation des valeurs manquantes par moyenne
— Sélection des variables prédictives : `age`, `sexe`, `taux`, `situationFamiliale`, `nbEnfantsAcharge`, `2eme voiture`
— Cible : `catégorie` (obtenue du clustering)

4.1.2 Algorithmes testés

Deux modèles ont été comparés via cross-validation :
— **Decision Tree** (max_depth=7)
— **Random Forest** (n_estimators=100, max_depth=10)

4.1.3 Résultats obtenus

— **Modèle choisi** : arbre de décision
— **Accuracy (test)** : 74.3%
— **F1-score (test)** : 73.9%

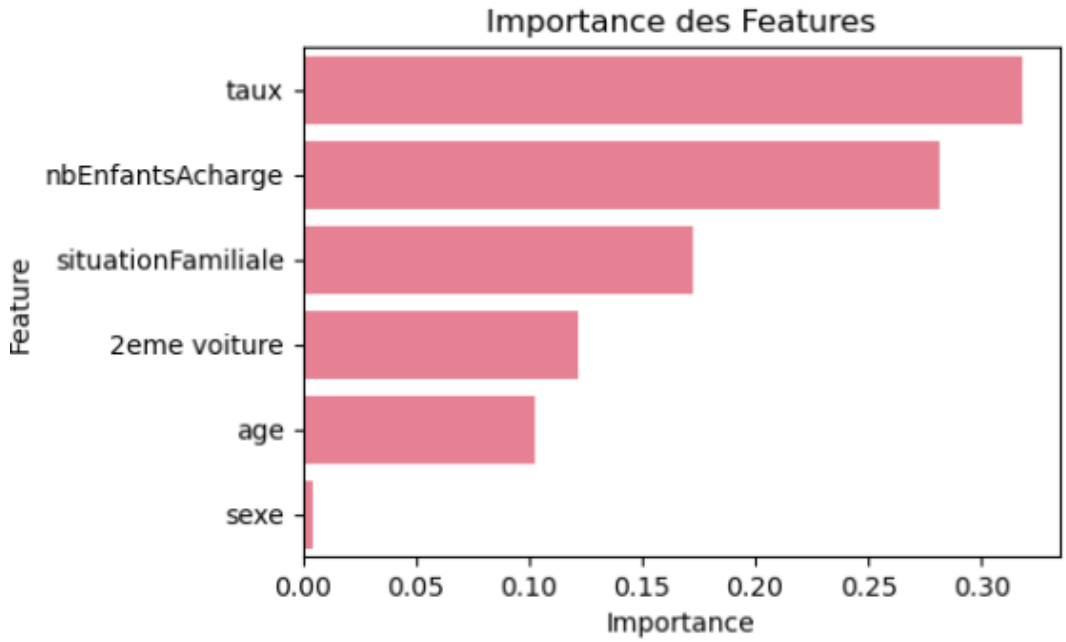

FIGURE 11 – Importance des variables selon Random Forest

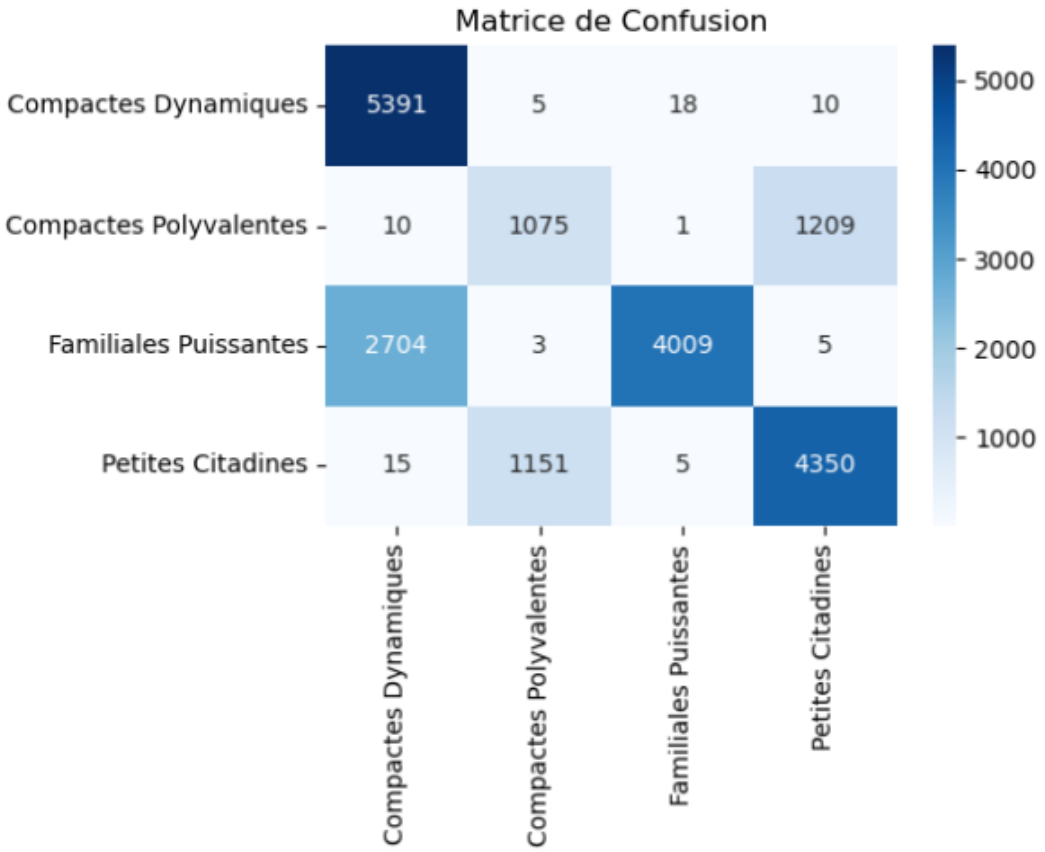

FIGURE 12 – Matrice de confusion du modèle final

4.2 Analyse de la performance du modèle de classification

4.2.1 Interprétation de la matrice de confusion

La matrice de confusion permet d'évaluer la performance du classifieur par arbre de décision en analysant les prédictions pour les 4 classes de véhicules. Voici les observations détaillées :

TABLE 1 – Métriques par classe (Accuracy globale : 74,3%, F1-score pondéré : 73,9%)

Classe	Precision	Recall	F1-Score	Observations
Compactes Dynamiques	0.85	0.99	0.91	Excellente détection mais risque de sur-prédict
Petites Citadines	0.79	0.78	0.78	Bon équilibre precision/recall
Familiales Puissantes	0.99	0.60	0.75	Prédictions fiables mais nombreux faux négati
Compactes Polyvalentes	0.47	0.48	0.48	Classe la moins bien prédite

4.2.2 Analyse détaillée par catégorie

- **Compactes Dynamiques** :
  - Recall exceptionnel (99%) indiquant une excellente détection
  - Precision légèrement inférieure (85%) révélant une tendance à sur-prédire cette classe

  - F1-score élevé (0.91) confirmant la robustesse sur cette catégorie
- **Petites Citadines** :
  - Métriques équilibrées (precision 79%, recall 78%)
  - Bonne fiabilité sans biais particulier
- **Familiales Puissantes** :
  - Precision excellente (99%) mais recall modéré (60%)
  - Modèle très conservateur pour cette catégorie (manque de faux positifs mais nombreux faux négatifs)
- **Compactes Polyvalentes** :
  - Performance limitée (F1-score 0.48)
  - Importantes confusions avec d'autres catégories
  - Nécessite potentiellement des features supplémentaires pour mieux la distinguer

4.3 Synthèse de la performance

La performance globale du modèle est satisfaisante avec une accuracy de 74,3%. Les résultats suggèrent que :

- Le modèle excelle sur 2 des 4 catégories
- La catégorie Compactes Polyvalentes nécessiterait des investigations complémentaires
- L'équilibre global precision/recall est acceptable pour une première itération

4.4 Prédiction sur les données marketing

Le modèle final a été appliqué aux clients du fichier `Marketing` afin de prédire la catégorie de véhicule la plus adaptée à chacun. Les prédictions ont été intégrées dans un fichier livré au service marketing pour ciblage précis.

Table 2 – Résultats avec prédiction

Âge	Sexe	Taux	Situation Familiale	Nb Enfants à charge	2ème voiture	Catégorie Prédite
21	0	1396	0	0	0	Petites Citadines
35	1	223	0	0	0	Compactes Polyvalentes
48	1	401	0	0	0	Compactes Polyvalentes
26	0	420	1	3	1	Familiales Puissantes
80	1	530	1	3	0	Familiales Puissantes
27	0	153	1	2	0	Compactes Dynamiques
59	0	572	1	2	0	Compactes Dynamiques
43	0	431	0	0	0	Compactes Polyvalentes
64	1	559	0	0	0	Petites Citadines
22	1	154	1	1	0	Compactes Dynamiques
79	0	981	1	2	0	Compactes Dynamiques
55	1	588	0	0	0	Petites Citadines
19	0	212	0	0	0	Compactes Polyvalentes

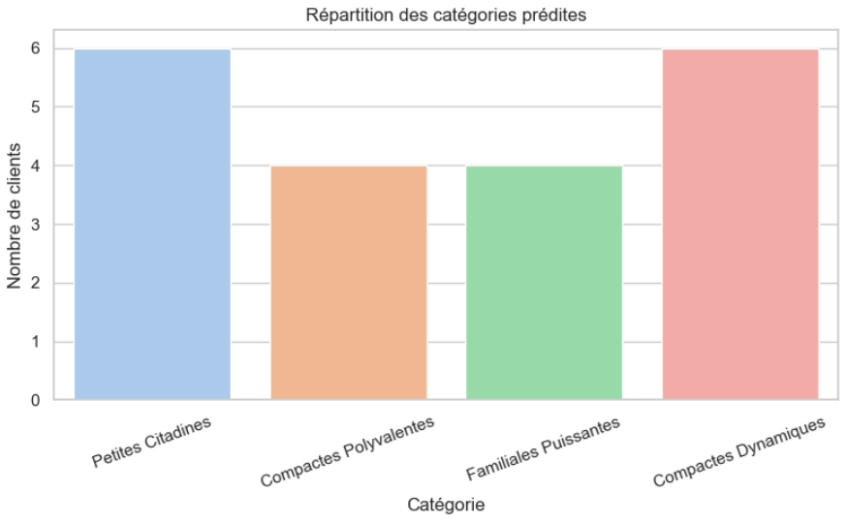

FIGURE 13 – Distribution prédite des catégories pour les clients marketing

5 Application

5.1 Application interactive de prédiction

Afin de permettre une utilisation simple et intuitive de notre modèle prédictif, une application interactive a été développée à l'aide de la bibliothèque `Tkinter` en Python.

5.1.1 Fonctionnalité

Cette application permet à un vendeur en concession de saisir en quelques secondes les informations clés d'un client potentiel :

— Âge
— Sexe
— Taux d'endettement
— Situation familiale
— Nombre d'enfants à charge
— Possession d'un second véhicule

Une fois les données saisies, l'application :

1. Prétraite les informations (encodage, typage)
2. Prédit la **catégorie de véhicule la plus adaptée** à l'aide du modèle d'arbre de décision entraîné
3. Affiche immédiatement un message contenant la catégorie
4. Liste tous les véhicules du catalogue appartenant à cette catégorie

5.1.2 Interface utilisateur

L'interface utilisateur est ergonomique et ne nécessite aucune compétence technique. Un exemple d'écran est présenté ci-dessous :

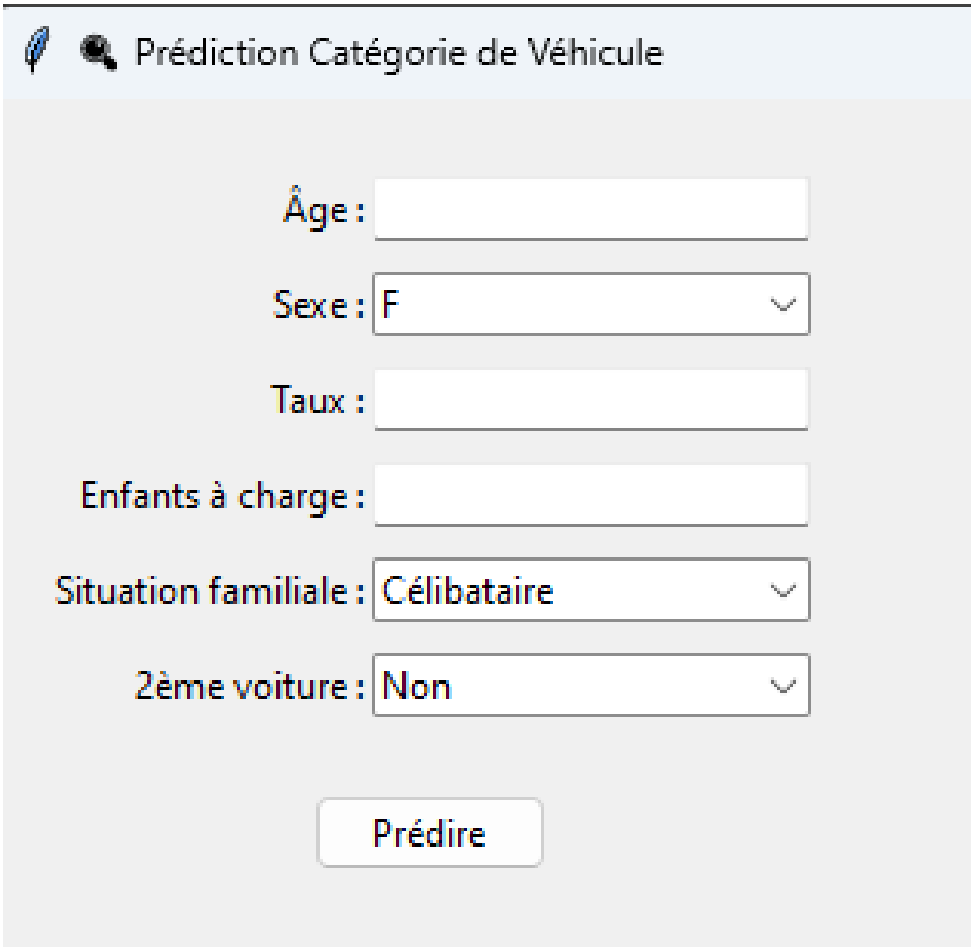

FIGURE 14 – Interface de l'application Tkinter pour la prédiction de catégorie

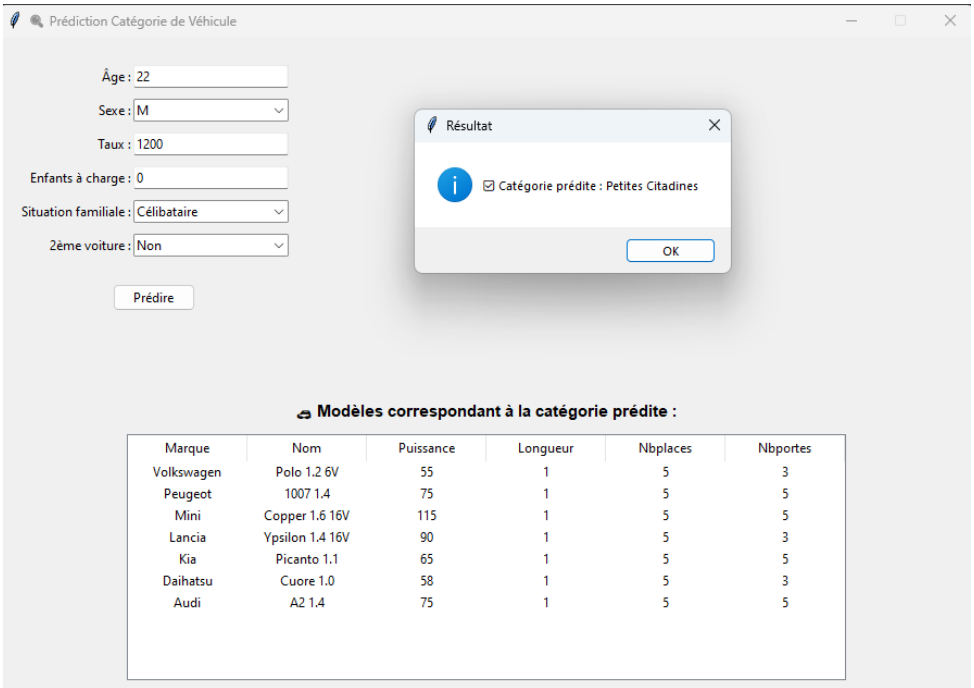

FIGURE 15 – Interface de l'application Tkinter pour la prédiction de catégorie

6 Conclusion

Le projet a permis de structurer les véhicules en catégories pertinentes via clustering, et de créer un modèle performant pour prédire la catégorie de véhicule correspondant aux profils clients. Ce modèle peut être déployé en concession ou auprès du service marketing.

Les quatre phases du projet ont démontré leur complémentarité :

— L'**analyse descriptive** a révélé les caractéristiques fondamentales des données
— Le **clustering** a permis d'identifier des categories
— La **classification** a abouti à un modèle prédictif performant (74,3% d'accuracy)
— L'**application** concrétise l'utilisation pratique en concession

Cette approche méthodologique offre au concessionnaire un outil complet d'aide à la décision pour optimiser ses ventes et personnaliser ses offres clients.