Université Moulay Ismail

Faculté des sciences de Meknès

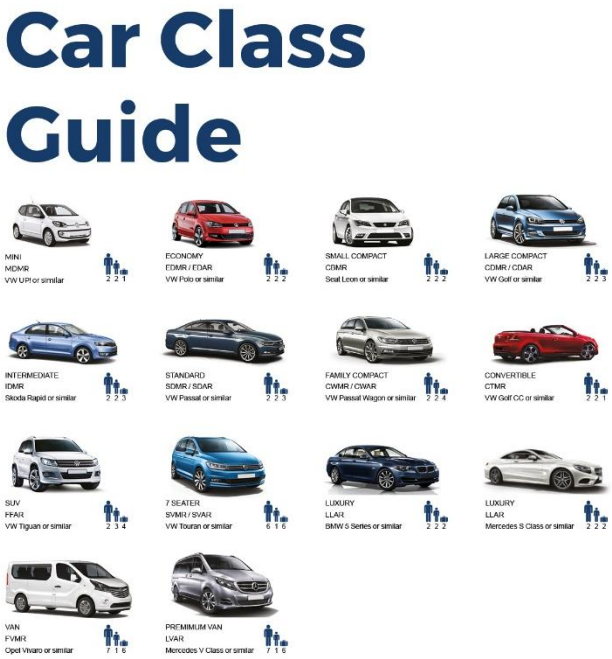
MASTER Intelligence Artificielle et Analyse des Données

Car Dealer – Data Mining

Prepared by: *Abdessamad Alhaouil*

Introduction

In today's highly competitive automotive market, leveraging data-driven insights is crucial for understanding customer preferences, optimizing inventory, and boosting sales. This project aims to build a predictive model that helps classify and recommend vehicles to clients based on their demographic and behavioral characteristics. By analyzing and modeling dealership data, we can uncover valuable patterns and trends that align with customer needs.



Objective

The primary objective of this project is to develop a robust and efficient machine learning pipeline that classifies vehicles into predefined categories such as **City Car**, **Family**, **Luxury**, **VAN**, **Sports Car**, and more. This classification is based on a combination of client attributes (e.g., age, family situation, financial profile) and vehicle specifications (e.g., price, power, and brand).

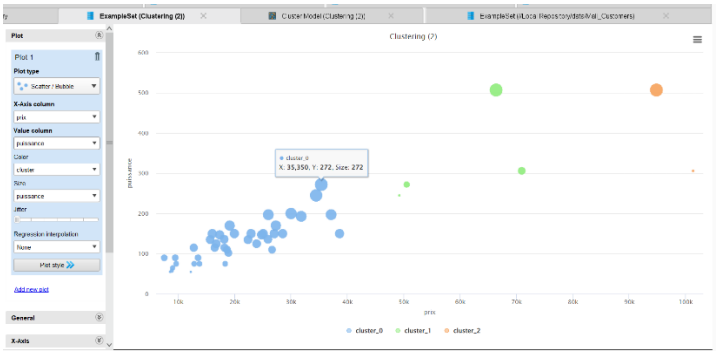
The goal is twofold:

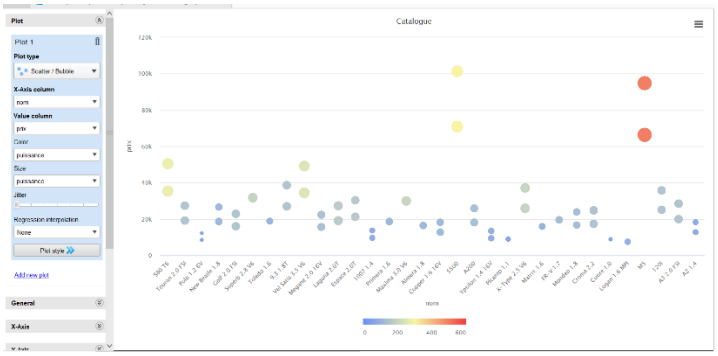
1. **For Clients**: Recommend the most suitable vehicle categories based on individual profiles, ensuring an enhanced customer experience.
2. **For Dealerships**: Enable better inventory management by clustering vehicles into categories and providing actionable insights into customer preferences.

Data Preprocessing and Cleaning

1. Overview of Data

To kick off the project, I visualized the datasets using RapidMiner to identify abnormalities, missing values, and potential issues in the data. This preliminary analysis guided my cleaning strategy.





age *integer*	sexe *polynominal*	taux *polynominal*	situationFamili... *polynominal*	nbEnfantsAch... *integer*	2eme voiture *binominal*	immatriculation *polynominal*
22	M	538	En Couple	1	false	3489 DA 72
18	M	585	En Couple	2	false	9611 IL 78
80	M	739	En Couple	3	false	1610 WA 51
83	F	772	En Couple	0	false	2350 JD 35
47	M	549	En Couple	2	false	7812 MP 13
82	F	599	En Couple	3	true	7744 HE 84
25	M	156	En Couple	0	false	5546 WA 24
39	F	998	Célibataire	0	false	9010 PG 45
22	M	402	En Couple	?	false	5681 TB 11
24	M	436	En Couple	4	false	3054 KO 12
30	M	228	En Couple	0	false	4665 BV 94
63	M	553	En Couple	2	false	4983 RH 81
67	M	596	Célibataire	0	false	8651 JJ 24
39	M	418	En Couple	1	false	2509 BI 77
38	M	221	En Couple	3	true	6458 DQ 84
20	M	597	En Couple	2	false	7752 QC 65
42	Homme	430	?	0	false	2148 XW 49
28	M	218	En Couple	0	true	8282 NO 59

Célibataire

2. Loading Data

The datasets were loaded into a Google Colab environment using Python to ensure flexibility and scalability in processing.

3. Cleaning the Datasets

A systematic approach was applied to handle missing and abnormal values for each dataset:

Clients Dataset:

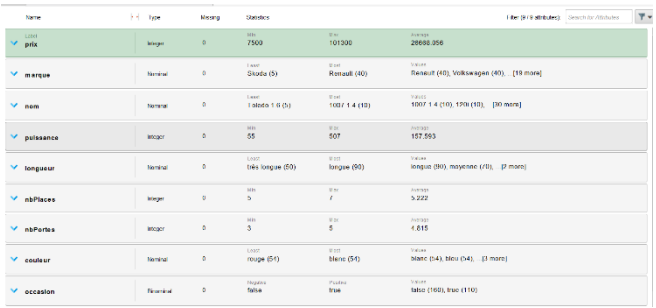
- **Age**: Replaced ? and NaN values with the mean, then transitioned to KNNImputer for a more accurate imputation based on neighboring values. Applied a lambda function to restrict ages to the range [18, 84].
- **Taux**: Handled similarly to age and restricted values to the range [544, 74185].
- **Sexe**: Standardized values to Male (True) and Female (False), dropping 189 null values.
- **Situation Familiale**: Mapped values to En Couple (True) and Célibataire (False).
- **2ème Voiture**: Replaced '?', empty values, and 'false' (string) with False (boolean) and 'true' (string) with True.
- **Immatriculation**: Removed duplicate rows based on the immatriculation column.
- **NbEnfantsAcharge**: Replaced '?', empty values, and -1 with 0. Restricted values to the range [0, 4].

Immatriculations Dataset:

- Corrected inconsistent values such as 'trÃ¨s longue' to 'tres longue'.

Catalogue Dataset:

- Handled inconsistencies like 'trÃ¨s longue', ensuring consistent feature formatting.



Feature Encoding

Encoding Categorical Features

To prepare for clustering and modeling:

- Numerical features remained untouched.
- Boolean features were already encoded during cleaning.
- **Nominal features**: Encoded using LabelEncoder() from sklearn.
- **Ordinal features**: Encoded using OrdinalEncoder() with predefined orderings.

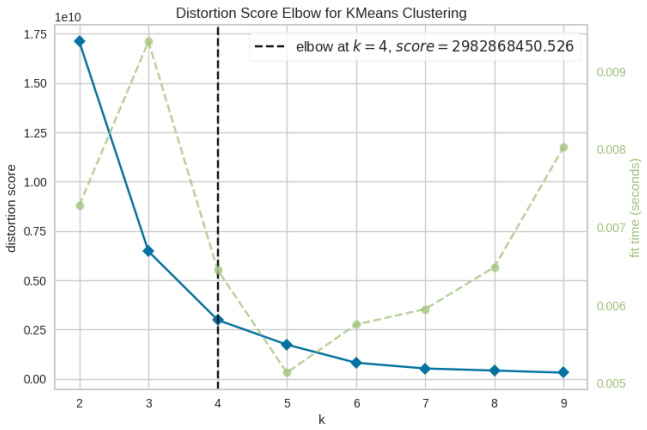
Clustering

1. Dividing the Catalogue Dataset

The catalogue dataset was split into **new cars** and **old cars** due to disparities in pricing and features, allowing more meaningful clustering.

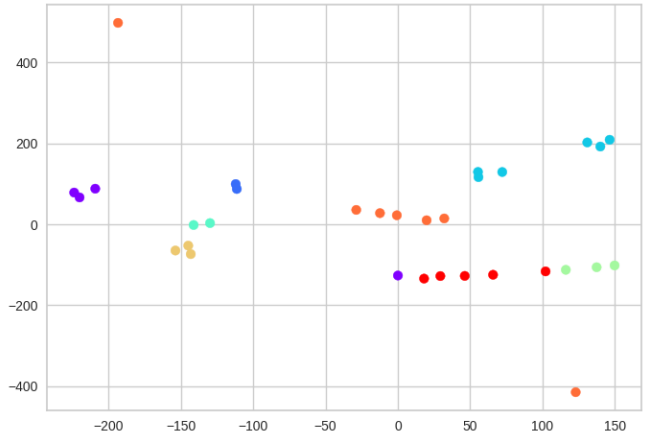
2. Identifying Optimal Clusters

- **Elbow Method**: Used Yellowbrick's ElbowVisualizer to identify potential cluster counts between 4 and 10.

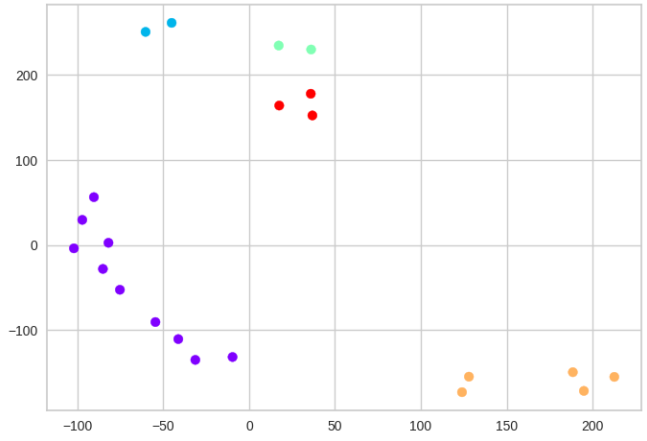


- **t-SNE Plotting**: Visualized clustering tendencies based on reduced dimensions.

New cars



Old cars



- **Silhouette Scores**: Calculated for different cluster counts:

New Cars: Optimal at 8 clusters (Silhouette Score: 0.683).

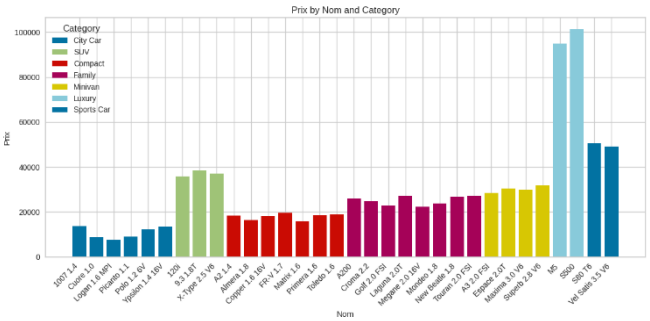
Old Cars: Optimal at 5 clusters (Silhouette Score: 0.789).

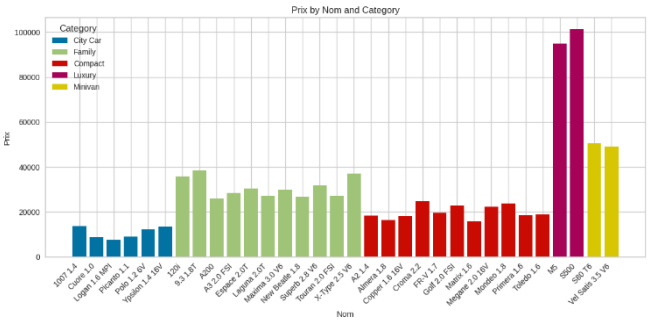
3. KMeans Clustering

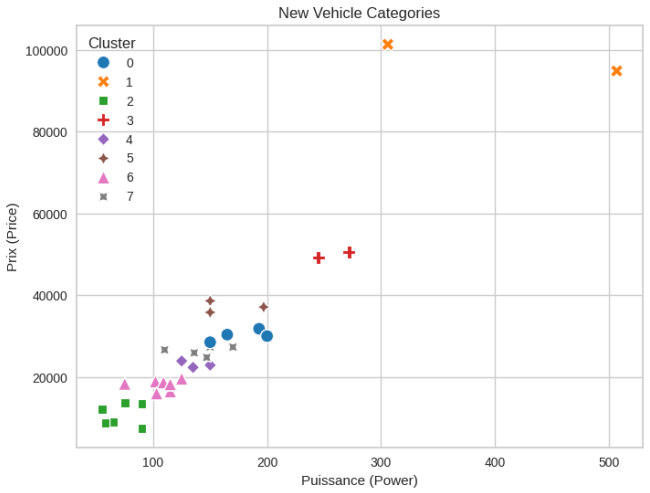
Performed KMeans clustering on both new and old cars:

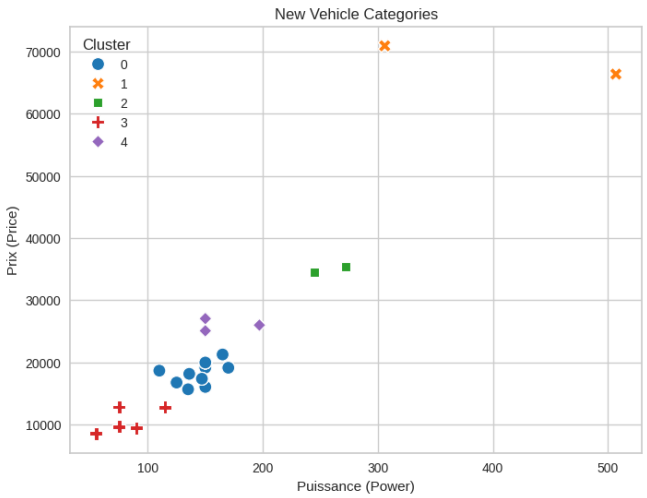
- Mapped clusters to 8 car categories inspired by car classification research: **Mini**, **Economy**, **Compact**, **Luxury**, **Standard**, **Sports Car**, and **Luxury v1.0**.

New cars









Data Merging

1. Dividing Immatriculations

Split the immatriculations dataset into old and new cars, aligning with the catalogue split.

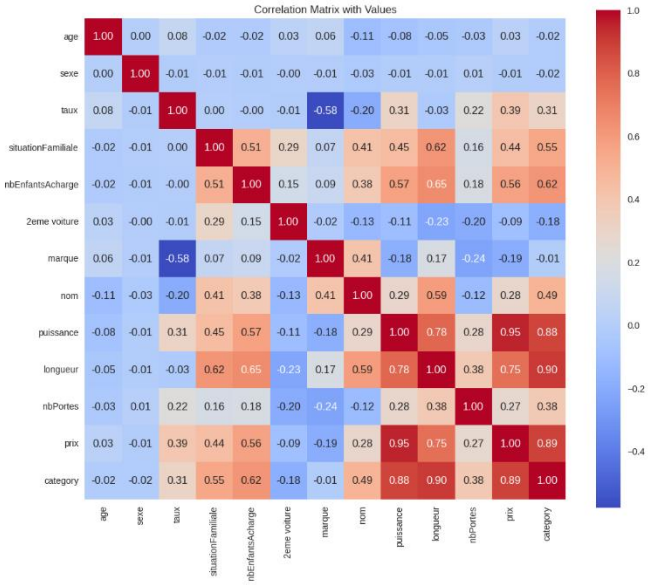
2. Merging Datasets

Merged the catalogue (with clusters) and immatriculations, then combined the result with the clients dataset on the immatriculation column.

Feature Selection

1. Correlation Analysis

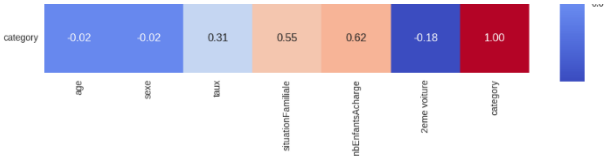
Plotted a correlation matrix to identify features most correlated with the target (vehicle category). Low-correlated features were removed to simplify the model.



Model Building and Training

1. Data Preparation

Prepared the cleaned and merged dataset, ensuring categorical features were encoded and numerical features scaled.

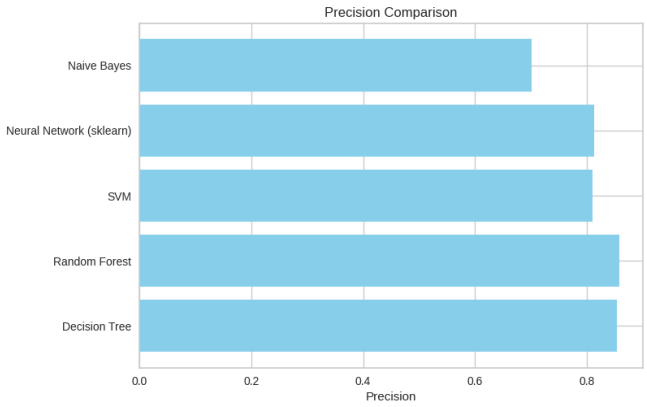


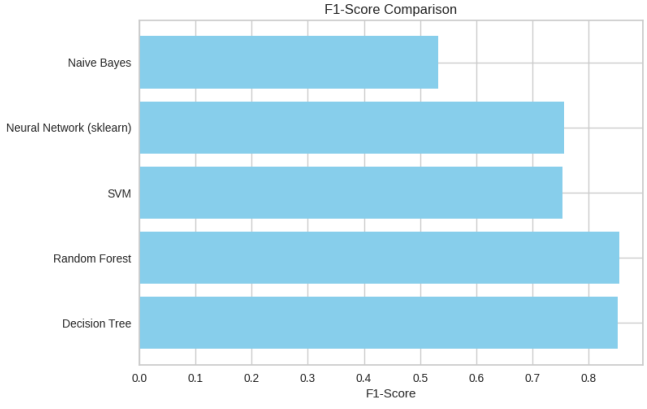
2. Model Training

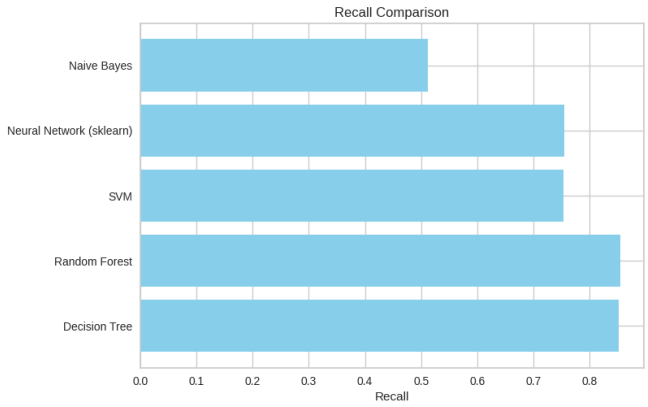
- Tested multiple models, including Random Forest, Decision Tree, Support Vector Machine, Neural Networks, and Naive Bayes.
- Hyperparameter tuning was performed to optimize performance.

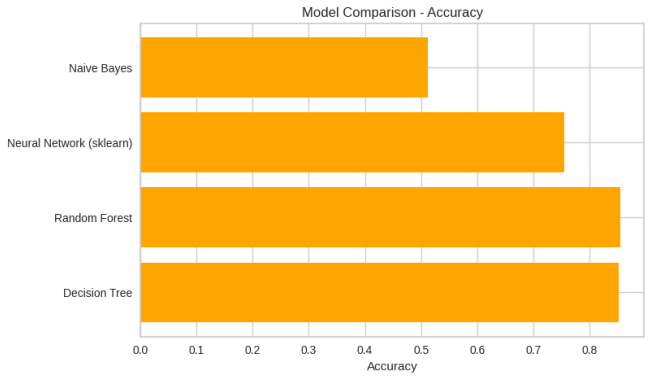
3. Evaluation

Each model was evaluated using metrics such as accuracy, precision, recall, F1-score, and cross-validation scores.









Prediction

The marketing dataset was used as input to predict the most suitable vehicle category for each client based on their demographic and financial profile.

```
raw = {
        'age': 53,
        'sexe': True,
        'taux': 594,
        'situationFamiliale': True,
        'nbEnfantsAcharge': 2,
        '2eme voiture': False
        }
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 42ms/step
probability of the class: 1.0
best class for this client: 0    Luxury v1.0
dtype: object
```

Car Prediction

Age:

45

Sex:

Female

Taux:

541

Situation Familiale:

Single

Number of Children:

2

2nd Car:

No

Predict

Cars in Predicted Category: Large Compact

Nom	Prix	Puissance	Couleur
Touran 2.0 FSI	27340	150	rouge
New Beatle 1.8	26630	110	rouge
Laguna 2.0T	27300	170	gris
A200	25900	136	blanc
Croma 2.2	24780	147	rouge

Back to Form

Select Category to View Cars

Category:

Sports Car

View Cars

Cars in Category: Sports Car

Nom	Prix	Puissance	Couleur
S500	101300	306	bleu
M5	94800	507	gris