# Groupe de TPT 1 – Raspcar

# 25 février 2020

**Membres :**

* Bah Mamadou Saliou
* Chatti Nader
* Diaz Gabriel
* Iben Salah Roukaya
* Ramoul Inès



# I. Descriptif du projet

Pour ce projet, nous disposons de 5 fichiers CSV :

* 2 Fichiers Client regroupant les clients ayant acheté une voiture au cours de l’année
* 1 fichier Immatriculations contient les véhicules vendus au cours de l’année
* 1 fichier Catalogue contenant les véhicules disponibles à la concession
* 1 fichier Marketing contenant les clients susceptibles d’acheter une voiture

Le but est de prédire quel type de véhicule les clients de la base marketing sont susceptibles d’acheter. La prédiction s’opère à l’aide d’un classifieur d’apprentissage supervisé.

Joint au dossier, vous trouverez le script R sous le nom « Projet.R » ainsi que le CSV Marketing contenant les prédictions.

Le source R est scrupuleusement commentée vous permettant de suivre notre raisonnement à chaque étape.

# II. Pré-traitement des données

Lors du chargement des fichiers dans la base R, nous nous sommes aperçus que les fichiers Client, Marketing et Immatriculations nécessitaient d’être traités afin de ne pas fausser les algorithmes de prédictions de nos classifieurs.

## II.I Clients

Dans un premier temps, nous avons binder nos deux fichiers client en un seul, puis, nous avons renommer la colonne « 2eme voiture » en « sec\_voiture » afin de la rendre manipulable plus facilement.

Le fichier Client comportait des valeurs manquantes sous différentes formes (« -1 », « N/D », « », « ? »). Seule une petite partie du fichier (environ 2000 lignes soit 1% des records) était concerné. De ce fait, nous avons pu supprimer ces lignes sans que cela impacte notre échantillon.

La colonne sexe contenait plusieurs types de valeurs (« M », « Homme », « Masculin », « F », « Femme », « Féminin »). Nous avons homogénéisé ces valeurs en « M » et « F ».

Les types des colonnes n’étaient pas corrects (quasiment tous en factor). Nous les avons modifiés afin de les rendre traitables plus efficacement par les classifieurs.

## II.2 Immatriculation

Nous sommes partis du principe qu’une immatriculation est, par définition, unique. Or, le fichier contenait des doublons. Nous avons supprimé ces lignes.

## II.3 Marketing

Comme pour le fichier Client, nous avons renommé la colonne « 2eme voiture » en « sec\_voiture » et nous avons transformé les types des différentes zones.

# III. Construction des catégories de voitures

A partir du fichier Catalogue, nous avons pu établir 5 types de véhicules :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Puissance | Longueur | Nb Place | Prix |
| Familial | <= 130 | Longue  Très longue | >= 5 |  |
| Citadine | <= 150 | Courte |  |  |
| Luxe | >= 180 |  |  | >= 90 000 |
| Sport | >= 180 |  |  | >= 30 000 |
| Berline | <= 180 |  |  |  |

En sortie de l’algorithme, voilà les prédictions définies :

| **Classification** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
| **1** | S80 T6 | Volvo | 272 | très longue | 5 | 5 | false | 50500 | Sport |
| **2** | Touran 2.0 FSI | Volkswagen | 150 | longue | 7 | 5 | false | 27340 | Familiale |
| **3** | Polo 1.2 6V | Volkswagen | 55 | courte | 5 | 3 | false | 12200 | Citadine |
| **4** | New Beatle 1.8 | Volkswagen | 110 | moyenne | 5 | 5 | false | 26630 | Berline |
| **5** | Golf 2.0 FSI | Volkswagen | 150 | moyenne | 5 | 5 | false | 22900 | Berline |
| **6** | Superb 2.8 V6 | Skoda | 193 | très longue | 5 | 5 | false | 31790 | Sport |
| **7** | Toledo 1.6 | Seat | 102 | longue | 5 | 5 | false | 18880 | Familiale |
| **8** | 9.3 1.8T | Saab | 150 | longue | 5 | 5 | false | 38600 | Berline |
| **9** | Vel Satis 3.5 V6 | Renault | 245 | très longue | 5 | 5 | false | 49200 | Sport |
| **10** | Megane 2.0 16V | Renault | 135 | moyenne | 5 | 5 | false | 22350 | Berline |
| **11** | Laguna 2.0T | Renault | 170 | longue | 5 | 5 | false | 27300 | Berline |
| **12** | Espace 2.0T | Renault | 165 | longue | 7 | 5 | false | 30350 | Familiale |
| **13** | 1007 1.4 | Peugeot | 75 | courte | 5 | 5 | false | 13750 | Citadine |
| **14** | Primera 1.6 | Nissan | 109 | longue | 5 | 5 | false | 18650 | Familiale |
| **15** | Maxima 3.0 V6 | Nissan | 200 | très longue | 5 | 5 | false | 30000 | Sport |
| **16** | Almera 1.8 | Nissan | 115 | moyenne | 5 | 5 | false | 16450 | Berline |
| **17** | Copper 1.6 16V | Mini | 115 | courte | 5 | 5 | false | 18200 | Citadine |
| **18** | S500 | Mercedes | 306 | très longue | 5 | 5 | false | 101300 | Luxe |
| **19** | A200 | Mercedes | 136 | moyenne | 5 | 5 | false | 25900 | Berline |
| **20** | Ypsilon 1.4 16V | Lancia | 90 | courte | 5 | 3 | false | 13500 | Citadine |
| **21** | Picanto 1.1 | Kia | 65 | courte | 5 | 5 | false | 8990 | Citadine |
| **22** | X-Type 2.5 V6 | Jaguar | 197 | longue | 5 | 5 | false | 37100 | Sport |
| **23** | Matrix 1.6 | Hyundaï | 103 | longue | 7 | 5 | false | 15960 | Familiale |
| **24** | FR-V 1.7 | Honda | 125 | longue | 7 | 5 | false | 19550 | Familiale |
| **25** | Mondeo 1.8 | Ford | 125 | longue | 5 | 5 | false | 23900 | Familiale |
| **26** | Croma 2.2 | Fiat | 147 | longue | 5 | 5 | false | 24780 | Berline |
| **27** | Cuore 1.0 | Daihatsu | 58 | courte | 5 | 3 | false | 8850 | Citadine |
| **28** | Logan 1.6 MPI | Dacia | 90 | moyenne | 5 | 5 | false | 7500 | Berline |
| **29** | M5 | BMW | 507 | très longue | 5 | 5 | false | 94800 | Luxe |
| **30** | 120i | BMW | 150 | moyenne | 5 | 5 | false | 35800 | Berline |
| **31** | A3 2.0 FSI | Audi | 150 | moyenne | 5 | 5 | false | 28500 | Berline |
| **32** | A2 1.4 | Audi | 75 | courte | 5 | 5 | false | 18310 | Citadine |

# IV. Construction des ensembles d’apprentissage et de test

A l’aide d’SQL, nous avons effectué une double jointure entre Clients, Immatriculation et Catalogue afin de rapatrier dans le fichier Client la catégorie de véhicule relative à chaque immatriculation.

A partir du fichier Client, nous avons pu construire un ensemble d’apprentissage égal à 70% des lignes et un ensemble de test avec les 30% restants.

# V. Phase d’apprentissage

Nous avons cherché à déterminer le meilleur classifieur en nous appuyant sur 7 méthodes d’apprentissage :

* Rpart
* C5.0
* Tree
* Random Forest
* Naive Bayes
* NNET
* KKNN

Lors de cette phase, nous avons entrainé nos classifieurs à l’aide de l’ensemble d’apprentissage créé précédemment.

# VI. Phase de test

Cette phase nous a permis de déterminer quel est le classifieur le plus précis.

Dans un premier temps, nous générons les prédictions sur l’ensemble de test. En voici un échantillon :

| **Classification** | | **Rpart** | | **c50** | | **tree** | | **RandomForest** | | **NaiveBayes** | | **NNET** | | **KKNN** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  | |  | |  | |  | |  | |  | |  |
| **122176** | Sport | | Sport | | Familiale | | Sport | | Familiale | | Berline | | Sport | | Familiale |
| **122177** | Berline | | Citadine | | Berline | | Citadine | | Citadine | | Berline | | Berline | | Citadine |
| **122178** | Luxe | | Sport | | Sport | | Sport | | Sport | | Sport | | Sport | | Familiale |
| **122179** | Citadine | | Citadine | | Citadine | | Citadine | | Citadine | | Citadine | | Citadine | | Citadine |
| **122180** | Sport | | Sport | | Sport | | Luxe | | Sport | | Sport | | Sport | | Luxe |
| **122181** | Citadine | | Citadine | | Citadine | | Citadine | | Citadine | | Citadine | | Citadine | | Citadine |

A première vue, nous pouvons d’ores et déjà relever que nos prédictions ne sont pas tout à fait conformes à la réalité (1er colonne)

Afin de déterminer lequel appliquer, nous sommes partis du constat qu’il est plus dommageable pour un commerce de mal conseiller leurs clients que de les conseiller correctement à tous les coups. Nous avons donc choisi de calculer les pourcentages d’erreurs de nos classifieurs à l’aide de leurs matrices de confusion et de conserver celui qui possède le plus petit.

| **Pourcentage\_FP** | |
| --- | --- |
|  |  |
| **rpart** | 36.51503 |
| **C5.0** | 34.31878 |
| **tree** | 38.24720 |
| **RandomForest** | 34.56323 |
| **NaiveBayes** | 46.69226 |
| **NNET** | 46.33131 |
| **KKNN** | 41.3716 |

Au vu des résultats, nous pouvons en déduire que la meilleure prédiction est obtenue grâce au classifieur C5.0.

# VII. Application au fichier Marketing

Voici les prédictions obtenues

| **age** | | **sexe** | | **taux** | | **situationFamiliale** | | **nbEnfantsAcharge** | | **sec\_voiture** | | **Prediction** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | |  | |  | |  | |  | |  | |  |
| **1** | 21 | | F | | 1396 | | Célibataire | | 0 | | FALSE | | Citadine |
| **2** | 35 | | M | | 223 | | Célibataire | | 0 | | FALSE | | Citadine |
| **3** | 48 | | M | | 401 | | Célibataire | | 0 | | FALSE | | Citadine |
| **4** | 26 | | F | | 420 | | En Couple | | 3 | | TRUE | | Citadine |
| **5** | 80 | | M | | 530 | | En Couple | | 3 | | FALSE | | Berline |
| **6** | 27 | | F | | 153 | | En Couple | | 2 | | FALSE | | Sport |
| **7** | 59 | | F | | 572 | | En Couple | | 2 | | FALSE | | Berline |
| **8** | 43 | | F | | 431 | | Célibataire | | 0 | | FALSE | | Berline |
| **9** | 64 | | M | | 559 | | Célibataire | | 0 | | FALSE | | Berline |
| **10** | 22 | | M | | 154 | | En Couple | | 1 | | FALSE | | Sport |
| **11** | 79 | | F | | 981 | | En Couple | | 2 | | FALSE | | Berline |
| **12** | 55 | | M | | 588 | | Célibataire | | 0 | | FALSE | | Berline |
| **13** | 19 | | F | | 212 | | Célibataire | | 0 | | FALSE | | Citadine |
| **14** | 34 | | F | | 1112 | | En Couple | | 0 | | FALSE | | Sport |
| **15** | 60 | | M | | 524 | | En Couple | | 0 | | TRUE | | Citadine |
| **16** | 22 | | M | | 411 | | En Couple | | 3 | | TRUE | | Citadine |
| **17** | 58 | | M | | 1192 | | En Couple | | 0 | | FALSE | | Berline |
| **18** | 54 | | F | | 452 | | En Couple | | 3 | | TRUE | | Citadine |
| **19** | 35 | | M | | 589 | | Célibataire | | 0 | | FALSE | | Berline |
| **20** | 59 | | M | | 748 | | En Couple | | 0 | | TRUE | | Citadine |