

Fouille de données

Travaux pratiques sous le logiciel R

Nous allons travailler sur une base de données décrivant les clients d'une banque allemande :

```
http://eric.univ-lyon2.fr/~jjjacques/Download/DataSet/GermanCredit.data
```

qui comporte des renseignements sur 1000 clients d'une banque allemande, chaque client étant décrit par 21 variables. La table contient la description des 21 variables.

1. Identifier les variables quantitatives et les variables catégorielles. Réaliser une ACP sur les variables quantitatives.
2. Réaliser un clustering sur les données quantitatives.
3. Réaliser une classification pour prédire la qualité du client. Évaluer votre modèle à l'aide d'un échantillon test que vous aurez au préalable extrait de l'échantillon total.

numero	nom de la variable	valeur
1	état du compte chèque (en DM)	A11 : < 0 A12 : ∈ [0, 200[A13 : ≥ 200 ou versement des salaires pendant au moins un an A14 : pas de compte chèque
2	durée en mois du crédit	∈ ℕ
3	historique des crédits	A30 : pas de crédit / tous remboursés A31 : tous les crédits dans la banque remboursés A32 : crédits en cours A33 : retard de paiement dans le passé A34 : compte critique / crédit existant dans d'autre banque
4	but du crédit	A40 : voiture neuve A41 : voiture occasion A42 : équipement / fourniture A43 : radio / télévision A44 : appareils ménagers A45 : réparation A46 : éducation A47 : vacances A48 : recyclage A49 : professionnel A410 : autre
5	montant du crédit (en DM)	∈ ℝ
6	montant de l'épargne (en DM)	A61 : < 100 A62 : ∈ [100, 500[A63 : ∈ [500, 100[A64 : ≥ 1000 A65 : inconnu
7	ancienneté dans le travail actuel (an)	A71 : sans emploi A72 : < 1 A73 : ∈ [1, 4[A74 : ∈ [4, 7[A75 : ≥ 7
8	taux d'apport	∈ ℝ
9	état marital	A91 : homme divorcé / séparé A92 : femme divorcé / séparé / mariée A93 : homme célibataire A94 : homme marié / veuf A95 : femme célibataire
10	autres demandeurs / garants	A101 : aucun A102 : co-demandeur A103 : garant
11	durée d'habitation dans la résidence actuelle (an)	∈ ℕ
12	biens	A121 : immobilier A122 : si pas A121 : placement (assurance vie ou part dans la banque) A123 : si pas A121 et A122 : voiture ou autre, non compris dans la variable 6 A124 : inconnu
13	âge (an)	∈ ℕ
14	autres demandes de crédits	A141 : banque A142 : magasins A143 : aucun
15	situation dans la résidence actuelle	A151 : locataire A152 : propriétaire A153 : occupant à titre gratuit
16	nombre de crédits dans la banque	∈ ℕ
17	emploi	A171 : sans emploi / non qualifié - étranger A172 : non qualifié - non étranger A173 : emploi qualifié / fonctionnaire A174 : gestion / indépendant / emploi hautement qualifié / haut fonctionnaire
18	nombre de personnes pouvant rembourser le crédit	∈ ℕ
19	téléphone	A191 : aucun A192 : oui, enregistré au nom du client
20	travailleur étranger	A201 : oui A202 : non
21	qualité client	1 : bon 2 : mauvais

TABLE 1 – Variables du jeu de données GermanCredit.data