prédictives afin de comparer sa performance avec d'autres algorithmes de la littérature. Ce chapitre est divisé en deux grandes parties. La première partie est consacrée au premier type du clustering prédictif (voir Section 6.2). Pour ce type d'algorithmes, l'axe de prédiction est privilégié. Dans ce cadre d'étude, afin d'atteindre notre objectif, nous allons comparer les performances prédictives de l'algorithme des K-moyennes prédictives avec celles obtenues par les algorithmes les plus répandus dans la littérature. La deuxième partie de ce chapitre est consacrée au deuxième type du clustering prédictif (voir Section 6.3). Pour ce type d'algorithmes, aucun axe n'est privilégié par rapport à l'autre. Il s'agit ici de réaliser un bon compromis entre la description et la prédiction sous la contrainte d'interprétation des résultats. Dans cette partie expérimentale, on cherche à connaître, pour un jeu de données illustratif, la capacité de notre algorithme des K-moyennes prédictives à découvrir la structure interne de la variable cible et donc à découvrir les différentes raisons qui peuvent mener à une même prédiction.

Note : L'ensemble des approches présentées dans les sections précédentes ont été codées sur le logiciel R. Des spécifications de codes ont également été fournies à un prestataire afin de faire intégrer les approches proposées dans le logiciel interne **Khiops Ennéade**. Ce dernier est disponible sur le site suivant : www.khiops.predicsis.com. Il est à signaler donc que l'ensemble des résultats obtenus dans cette thèse sont reproductibles.

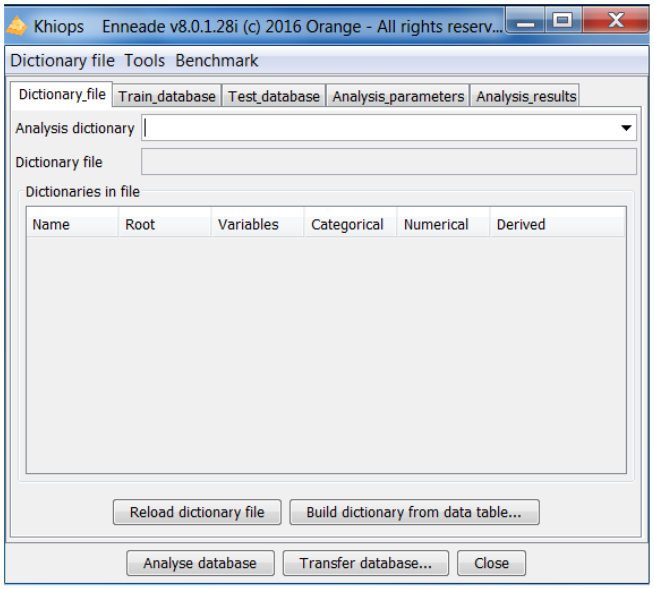

FIGURE 6.3

6.2 Clustering prédictif du premier type

Le clustering prédictif du premier type englobe l'ensemble des algorithmes du clustering modifiés permettant de prédire correctement la classe des nouvelles instances sous la contrainte d'avoir un nombre minimal de clusters. Dans ce cadre d'étude, l'axe de prédiction est principalement privilégié. L'algorithme des K-moyennes prédictives du premier type proposé dans cette thèse est donc l'algorithme incorporant les méthodes de prétraitement et d'initialisation des centres les plus performants en termes de prédictions.

Entrée
- Un ensemble de données D, où chaque instance X_i est décrite par un vecteur de d dimensions et par une classe $Y_i \in \{1, \ldots, J\}$.
- Le nombre de clusters souhaité, noté K.

Début
1) Prétraitement des données : *Conditional Info (CI-CI)*
2) Initialisation des centres : *Rocchio-And-Split*

Répéter
3) *Affectation :* générer une nouvelle partition en assignant chaque instance X_i au groupe dont le centre est le plus proche.

$$X_i \in C_k \forall j \in 1, \ldots, K \qquad k = argmin_j \parallel X_i - \mu_j \parallel$$

avec μ_k est le centre de gravité du cluster C_k.
4) *Représentation :* calculer les centres associés à la nouvelle partition

$$\mu_k = \frac{1}{N_k} \sum_{X_i \in C_k} X_i$$

jusqu'à ce que (convergence de l'algorithme)

5) Attribution des classes aux clusters formés : *vote majoritaire*
6) Prédiction de la classe des nouvelles instances : *un plus proche voisin*

Fin

Algorithme 8 – Algorithme des K-moyennes prédictives du premier type pour le cas CI-RS

En s'appuyant sur les résultats présentés dans la figure 6.1, l'algorithme des K-moyennes prédictives du premier type proposé est l'algorithme intégrant la méthode supervisée du prétraitement des données Conditional Info (CI) et la méthode supervisée d'initialisation des centres Rocchio-And-Split (RS). Pour un nombre fixe de clusters (K), l'algorithme 8 présente sous forme des lignes de code l'algorithme des K-moyennes prédictives du premier type.

L'un des principaux avantages de cet algorithme est qu'il n'est exécuté qu'une seule fois en raison de l'utilisation d'une méthode d'initialisation déterministe (*i.e.,* la méthode Rocchio-And-Split).

Cette section est consacrée à la comparaison des performances prédictives de cet algorithme des K-moyennes prédictives avec celles d'autres algorithmes du clustering prédictif les plus répandus dans la littérature. Cette section expérimentale est divisée en deux grandes parties. Dans la première partie (Section 6.2.1), on considère le nombre de clusters (K) comme une entrée de l'algorithme. Pour chaque jeu de données, on considère que le nombre de clusters (K) est égal au nombre de classes (J). Dans ce cas, le problème du départ devient un problème de classification supervisée. L'objectif de cette première partie est de tester la capacité de l'algorithme des K-moyennes prédictives présenté ci-dessus à atteindre l'objectif des algorithmes de la classification supervisée (*i.e.,* prédire correctement la classe des nouvelles instances).

La deuxième partie (Section 6.2.2) considère le nombre de clusters comme une sortie de l'algorithme (K≥J). L'objectif de cette partie est de tester la capacité de l'algorithme des K-moyennes prédictives à atteindre l'objectif des algorithmes de clustering prédictif du premier

type (*i.e.*, prédire correctement la classe des nouvelles instances sous contrainte d'obtenir un nombre minimal de clusters).

Les jeux de données utilisés dans cette partie expérimentale sont des jeux de données de l'UCI. Le tableau 6.1 présente les caractéristiques de ces jeux de données. Ces derniers ont été choisis afin d'avoir des bases de données illustratives diverses dans ce chapitre de synthèse en termes de nombre de classes J, de variables (continues M_n et/ou catégorielles M_c) et d'instances N. Pour chacun de ces jeux de données, on effectue un 2×5 folds cross validation.

ID	**Nom**	M_n	M_c	**N**	**J**	J_{maj}
1	Glass	10	0	214	6	36
2	Soybean	0	35	376	19	14
2	Breast	9	0	683	2	65
2	LED	7	0	1000	10	11
5	German	24	0	1000	2	70
6	Mushroom	0	22	8416	2	53

TABLE 6.1 – Liste des jeux de données utilisés- (J_{maj} représente $\approx$ pourcentage classe majoritaire).

6.2.1 Le nombre de clusters (K) est une entrée

Dans cette partie expérimentale, on cherche à tester la capacité de l'algorithme des K-moyennes prédictives présenté dans l'algorithme 8 à atteindre l'objectif des algorithmes de la classification supervisée. Les performances prédictives de l'algorithme des K-moyennes prédictives seront d'une part comparées à celles de l'algorithme des K-moyennes standard. Cette comparaison nous permet de savoir à quel point la version modifiée parvient à dépasser la version originale dans le contexte de la classification supervisée. D'autre part, l'algorithme des K-moyennes prédictives sera comparé à un des algorithmes de la classification supervisée le plus interprétable et le plus répandu dans la littérature, à savoir l'arbre de décision. Ce dernier est considéré comme une hiérarchie de clusters où chaque feuille représente un cluster. Pour une comparaison cohérente, le nombre de feuilles généré par l'arbre de décision est contrôlé de telle sorte d'avoir un nombre égal au nombre de classes du jeu de données utilisé (la taille du modèle est fixé K = J). Pour évaluer la performance prédictive de ces trois algorithmes, le critère "Variation d'Information" (VI) est utilisé. Plus la valeur de VI est proche de 0, meilleure est la performance prédictive du modèle.

Les deux figures 6.4 et 6.5 présentent les performances prédictives (en termes de VI) des trois algorithmes d'apprentissage lorsque le nombre de clusters (K) est égal au nombre de classes (J). Les résultats des deux figures montrent que l'algorithme des K-moyennes prédictives parvient à atteindre soit de meilleures performances prédictives par rapport à l'arbre de décision (résultats de la figure 6.4) ou des performances compétitives avec celles de l'arbre de décision (résultats de la figure 6.5). De plus, l'algorithme des K-moyennes prédictive arrive à atteindre des performances prédictives significativement meilleures que celles obtenues par l'algorithme des K-moyennes standard sachant que ce dernier est exécuté 100 fois avec différentes initialisations (en utilisant la même méthode K++) pour choisir la meilleure partition.

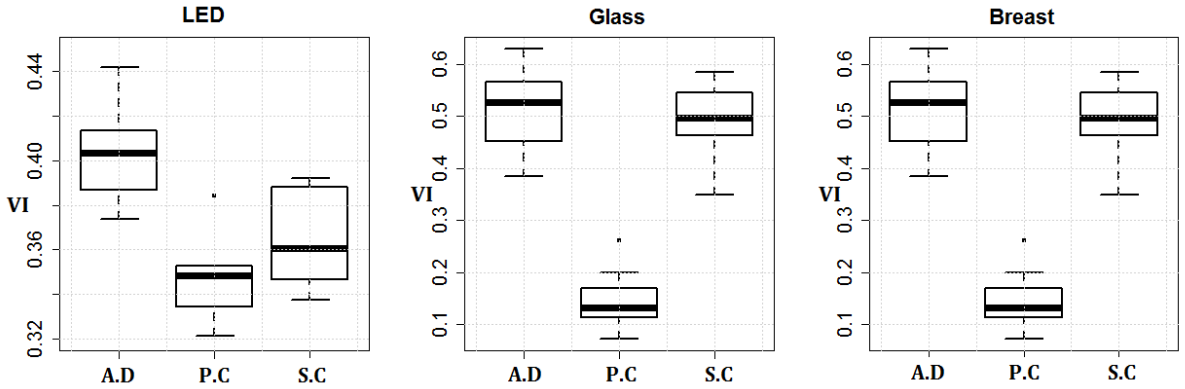

FIGURE 6.4 – Comparaison des performances prédictives (en termes de VI) pour les trois méthodes d'apprentissage (A.D = Arbre de décision, P.C= Prédictif clustering (K-moyennes prédictives) et S.C = K-moyennes standard)

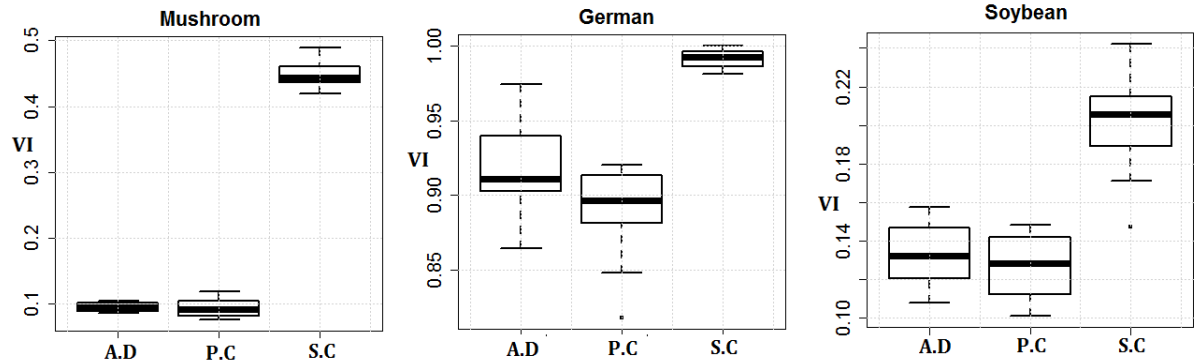

FIGURE 6.5 – Comparaison des performances prédictives (en termes de VI) pour les trois méthodes d'apprentissage (A.D = Arbre de décision, P.C= Prédictif clustering (K-moyennes prédictives) et S.C = K-moyennes standard)

6.2.2 Le nombre de clusters (K) est une sortie

Dans cette partie expérimentale, on cherche à tester la capacité de l'algorithme des K-moyennes prédictives présenté par l'algorithme 8 à atteindre l'objectif du clustering prédictif du premier type. Il s'agit ici de prédire correctement la classe des nouvelles instances sous la contrainte d'obtenir un nombre minimal de clusters (K).

Les algorithmes utilisés dans cette comparaison sont :

1. *L'arbre de décision* (A.D) : l'algorithme utilisé est l'arbre de décision de type CART. Dans cette étude expérimentale, nous utilisons l'algorithme élagué existant dans le logiciel R dans la librairie *rpart* [98]. Le nombre de clusters dans ce cas correspond au nombre de feuilles généré dans la phase d'apprentissage.
2. *Algorithme de Eick* (Eick) : cet algorithme nécessite un paramètre utilisateur β qui permet d'équilibrer le critère permettant d'évaluer la pureté des clusters en termes de classes et la contrainte sur le nombre de clusters générés. Les résultats présentés dans cette étude expérimentale sont obtenus lorsque β est égal à 0.1. Dans [46], Eick et al. montrent qu'avec cette valeur, ils parviennent à obtenir de meilleures performances. Pour plus de détails sur cet algorithme, voir la section 2.5.2 du chapitre 2.

3. *L'arbre du clustering prédictif* (PCT) : l'algorithme utilisé dans cette partie expérimentale est l'algorithme présenté dans "http ://clus.sourceforge.net/doku.php". Les résultats présentés dans cette section sont obtenus en utilisant l'arbre élagué présenté par cet algorithme. Dans ce cas, le nombre de clusters (K) correspond au nombre de feuilles générées par celui-ci dans la phase d'apprentissage. Pour plus de détails sur cet algorithme, voir la section 2.5.2 du chapitre 2.
4. *Les K-moyennes prédictives* (P.C) : pour avoir le nombre de clusters comme une sortie, l'algorithme des K-moyennes prédictives présenté par l'algorithme 8, est exécuté plusieurs fois avec différents nombres de clusters K dans le but de sélectionner la partition optimale au sens du clustering prédictif du premier type. Cette sélection est effectuée à l'aide de l'indice de rand ajusté "ARI".
5. *Les K-moyennes standard* (S.C) : pour avoir le nombre de clusters comme une sortie, l'algorithme des K-moyennes standard est exécuté plusieurs fois avec différent nombres de clusters K dans le but de sélectionner la partition optimale au sens du clustering prédictif du premier type. Cette sélection est effectuée à l'aide de l'indice de rand ajusté "ARI".

Les deux figures 6.6 et 6.7 présentent les performances prédictives (en termes de VI) des cinq modèles d'apprentissage. Les résultats de ces figures montrent qu'avec la supervision des deux étapes de prétraitement des données et d'initialisation des centres, l'algorithme des K-moyennes prédictives parvient à être très compétitif avec l'arbre de clustering prédictif (PCT). Cependant, l'algorithme des K-moyennes prédictives parvient à obtenir un nombre plus faible de clusters par rapport aux autres algorithmes. À titre d'exemple, pour le jeu de données German présenté dans la partie milieu de la figure 6.6, l'algorithme des K-moyennes prédictives obtient deux à trois clusters. Par contre, les arbres de décision, PCT, l'algorithme de Eick et les K-moyennes standard obtiennent respectivement 6, $\{63, 75\}$, $\{5-9\}$, 9 clusters.

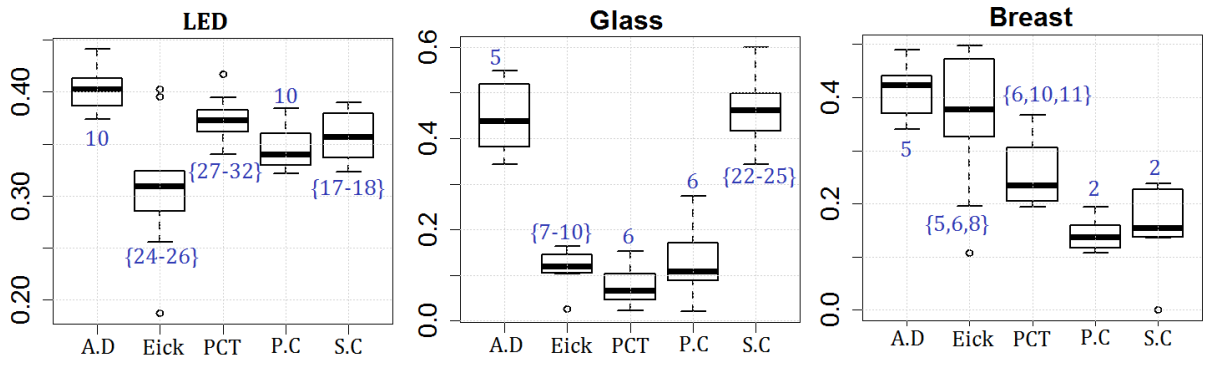

FIGURE 6.6 – Comparaison des performances prédictives (en termes de VI) des modèles lorsque le nombre de clusters est une sortie

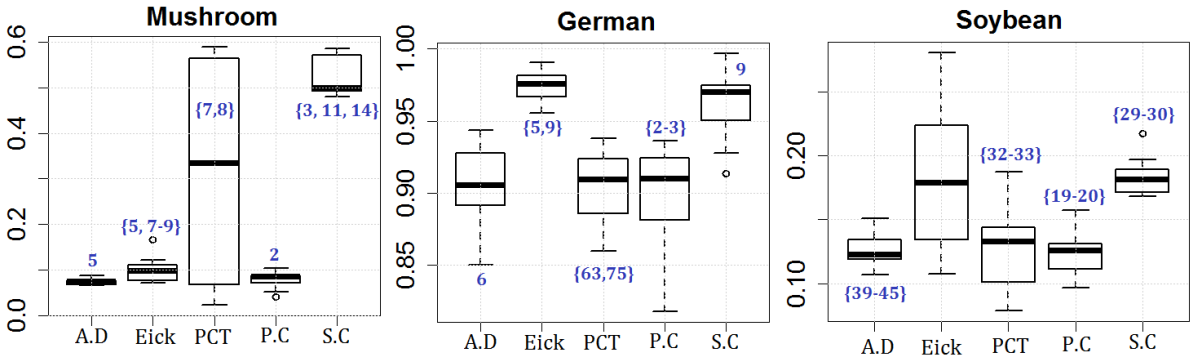

FIGURE 6.7 – Comparaison des performances prédictives (en termes de VI) des modèles lorsque le nombre de clusters est une sortie (suite)

6.2.3 Discussion

Dans cette première partie expérimentale, nous avons pu montrer qu'avec la supervision des deux étapes (prétraitement des données et initialisation des centres) de l'algorithme des K-moyennes standard, cet algorithme parvient à atteindre des performances (en termes de prédiction) meilleures ou très compétitives de celles obtenues par les algorithmes les plus répandus dans la littérature (*e.g.*, les arbres de décision et les arbres du clustering prédictif). Cet algorithme parvient également à avoir un nombre restreint de clusters par rapport à ces deux algorithmes. De plus, dans le contexte du clustering prédictif du premier type, on a pu montrer que l'algorithme des K-moyennes prédictives proposé est significativement meilleur que l'algorithme des K-moyennes standard.

6.3 Clustering prédictif du deuxième type

Le clustering prédictif du deuxième type englobe l'ensemble des algorithmes permettant de décrire et de prédire d'une manière simultanée. Il s'agit ici d'établir un compromis entre la description et la prédiction sous la contrainte d'interprétation des résultats. En s'appuyant sur la synthèse présentée dans 6.1 (voir la figure 6.2) et la discussion dans le chapitre d'initialisation (chapitre 4), l'algorithme des K-moyennes prédictives du deuxième type proposé dans cette thèse est l'algorithme des K-moyennes précédé par le prétraitement supervisé Conditional Info (CI) et la méthode d'initialisation KMeans++R (K++R). Pour le choix de la meilleure partition au sens du clustering prédictif du deuxième type, la version supervisée du critère Davies-Bouldin (SDB) est utilisée. Pour plus de détails sur ce critère, voir la section 5.3 du chapitre 5. L'algorithme 9 présente sous forme des lignes de code l'algorithme des K-moyennes prédictives du deuxième type.

Entrée
- Un ensemble de données D, où chaque instance X_i est décrite par un vecteur de d dimensions et par une classe $Y_i \in \{1, \ldots, J\}$.
- Le nombre de clusters souhaité, noté K.

Début
1) Prétraitement des données :*Conditional Info (CI-CI)*
2) Initialisation des centres : *Kmeans++R*
Pour K allant de J jusqu'à K_{max} **faire**
Répéter
3) *Affectation :* générer une nouvelle partition en assignant chaque instance X_i au groupe dont le centre est le plus proche.

$$X_i \in C_k \forall j \in 1, \ldots, K \qquad k = argmin_j \parallel X_i - \mu_j \parallel$$

avec μ_k est le centre de gravité du cluster C_k.
4) *Représentation :* calculer les centres associés à la nouvelle partition

$$\mu_k = \frac{1}{N_k} \sum_{X_i \in C_k} X_i$$

jusqu'à ce que (convergence de l'algorithme)
Fin Pour
5) Sélection du nombre de clusters optimal, noté K_{opti} en utilisant SDB.
6) Attribution des classes aux clusters formés : *vote majoritaire*
7) Prédiction de la classe des nouvelles instances. *un plus proche voisin.*
Fin

Algorithme 9 – Algorithme des K-moyennes prédictives du deuxième type pour le cas CI et K++R

L'objectif de cette section est donc de tester la capacité de cet algorithme à découvrir la structure interne *complète* de la variable cible. Il s'agit ici de découvrir les différentes raisons qui peuvent mener à une même prédiction. Cette partie expérimentale sert également à tester la capacité de cet algorithme à fournir des résultats facilement interprétables par l'utilisateur.

6.3.1 Description du jeu de données

Pour atteindre l'objectif cité ci-dessus, nous allons utiliser le jeu de données German de l'UCI. Ce jeu de données estime le risque crédit pour un demandeur donné. German contient 1000 demandeurs de crédit. Chaque demandeur est décrit par 20 variables descriptives et un score décrivant son risque crédit. 700 demandeurs ont été qualifiés à risque faible (classe 1) et 300 à risque élevé (classe 2). Une description détaillée de ce jeu de données est présentée dans "http ://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)".

La description des 20 variables descriptives est donnée dans ce qui suit :

1. Attribute 1 (V1) : Status of existing checking account (qualitative)
   — A11 : ... < 0 DM
   — A12 : 0 <= ... < 200 DM

- A13 : ... >= 200 DM
- A14 : no checking account

2. Attribute 2 (V2) : Duration in month (numerical)
3. Attribute 3 (V3) : Credit history (qualitative)
   - A30 : no credits taken/all credits paid back duly
   - A31 : all credits at this bank paid back duly
   - A32 : existing credits paid back duly till now
   - A33 : delay in paying off in the past
   - A34 : critical account/other credits existing (not at this bank)
4. Attribute 4 (V4) : Purpose (qualitative)
   - A40 : car (new)
   - A41 : car (used)
   - A42 : furniture/equipment
   - A43 : radio/television
   - A44 : domestic appliances
   - A45 : repairs
   - A46 : education
   - A47 : (vacation - does not exist ?)
   - A48 : retraining
   - A49 : business
   - A410 : others
5. Attribute 5 (V5) : Credit amount (numerical)
6. Attibute 6 (V6) : Savings account/bonds (qualitative)
   - A61 : ... < 100 DM
   - A62 : 100 <= ... < 500 DM
   - A63 : 500 <= ... < 1000 DM
   - A64 : .. >= 1000 DM
   - A65 : unknown/ no savings account
7. Attribute 7 (V7) : Present employment since (qualitative)
   - A71 : unemployed
   - A72 : ... < 1 year
   - A73 : 1 <= ... < 4 years
   - A74 : 4 <= ... < 7 years
   - A75 : .. >= 7 years
8. Attribute 8 (V8) : Installment rate in percentage of disposable income (numerical)
9. Attribute 9 (V9) : Personal status and sex (qualitative)
   - A91 : male : divorced/separated
   - A92 : female : divorced/separated/married
   - A93 : male : single
   - A94 : male : married/widowed
   - A95 : female : single
10. Attribute 10 (V10) : Other debtors / guarantors (qualitative)
    - A101 : none
    - A102 : co-applicant
    - A103 : guarantor

11. Attribute 11 (V11) : Present residence since (numerical)
12. Attribute 12 (V12) : Property (qualitative)
    — A121 : real estate
    — A122 : if not A121 : building society savings agreement/life insurance
    — A123 : if not A121/A122 : car or other, not in attribute 6
    — A124 : unknown / no property
13. Attribute 13 (V13) : Age in years (numerical)
14. Attribute 14 (V14) : Other installment plans (qualitative)
    — A141 : bank
    — A142 : stores
    — A143 : none
15. Attribute 15 (V15) : Housing (qualitative)
    — A151 : rent
    — A152 : own
    — A153 : for free
16. Attribute 16 (V16) : Number of existing credits at this bank (numerical)
17. Attribute 17 (V17) : Job (qualitative)
    — A171 : unemployed/ unskilled - non-resident
    — A172 : unskilled - resident
    — A173 : skilled employee / official
    — A174 : management/ self-employed/highly qualified employee/ officer
18. Attribute 18 (V18) : Number of people being liable to provide maintenance for (numerical)
19. Attribute 19 (V19) : Telephone (qualitative)
    — A191 : none
    — A192 : yes, registered under the customers name
20. Attribute 20 (V20) : foreign worker (qualitative)
    — A201 : yes
    — A202 : no

6.3.2 Résultats

Pour le jeu de données présenté ci-dessus, on cherche à connaître les différentes raisons qui peuvent mener à un faible ou à un fort risque crédit. Pour ce faire, l'algorithme des K-moyennes présenté dans l'algorithme 9, est exécuté avec différent nombres de clusters (voir la boucle **Pour** de l'algorithme 9) dans le but de sélectionner la partition optimale en utilisant le critère SDB. Il est à rappeler que la partition optimale au sens du clustering prédictif du deuxième type est celle qui réalise un bon compromis entre la description et la prédiction.

La figure 6.8 présente la valeur du critère SDB pour différentes partitions générées par l'algorithme des K-moyennes prédictives ayant différents nombres de clusters ($K \in [2, 14]$). Cette figure montre que la partition optimale au sens du clustering prédictif du deuxième type est la partition ayant 7 clusters.

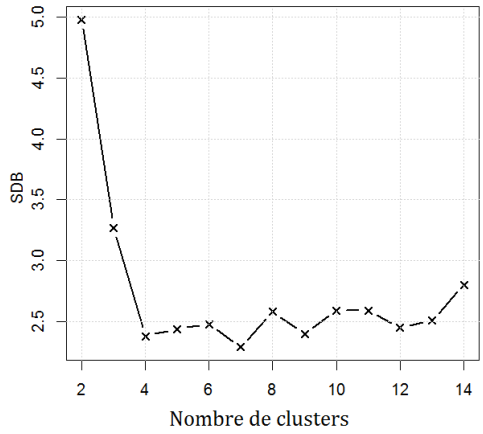

FIGURE 6.8 – Nombre de clusters K versus SDB

Cette partition optimale contient deux différents clusters formés par des demandeurs à fort risque crédit (clusters 2 et 4) et 5 différents clusters (clusters 1, 3, 5, 6 et 7) formés par des demandeurs à faible risque crédit. Le tableau 6.2 présente la probabilité que les demandeurs d'un cluster appartiennent à une des classes. Pour la première classe, c'est-à-dire, la classe à faible risque, les résultats du tableau 6.2 montrent que les trois clusters 1, 5 et 6 sont les clusters les plus purs en termes de la classe 1.

	Classe 1	Classe 2	# instances
Cluster 1	0.84	0.16	273
Cluster 5	0.82	0.18	91
Cluster 6	0.80	0.21	200
Cluster 3	0.65	0.35	193
Cluster 7	0.54	0.46	46
Cluster 2	0.48	0.52	154
Cluster 4	0.26	0.74	43

TABLE 6.2 – La probabilité d'appartenir à une classe j ($j \in \{1, 2\}$) conditionnellement au cluster i ($i \in \{1, \ldots, 7\}$)

Bien que les demandeurs qui forment ces 3 clusters ont la même étiquette, ils ont des profils différents. En effet, la figure 6.9 présente la description des différents clusters en utilisant les variables les plus discriminantes (V1, V2, V3, V4, V6, V12 et V15). Cette figure présente, pour chaque cluster et pour chacune des variables, les quantités d'informations (obtenues grâce au prétraitement Conditional Info) existant dans chaque intervalle de discrétisation ou groupe de modalités. Chaque bâton de cette figure présente la probabilité d'appartenir à un intervalle de discrétisation ou à un groupe de modalités (selon la nature de la variable en question) conditionnellement aux clusters. Cette figure montre que :

— Les gens du **cluster 1** sont spécifiquement des gens ayant une grand somme dans leur compte épargne (voir la variable V6).
— Les gens du **cluster 5** sont spécifiquement des gens ayant une grand somme dans leur compte courant (voir la variable V1) et ils possèdent des biens tels que des voitures, une assurance de vie, *etc.* (voir la variable V12).

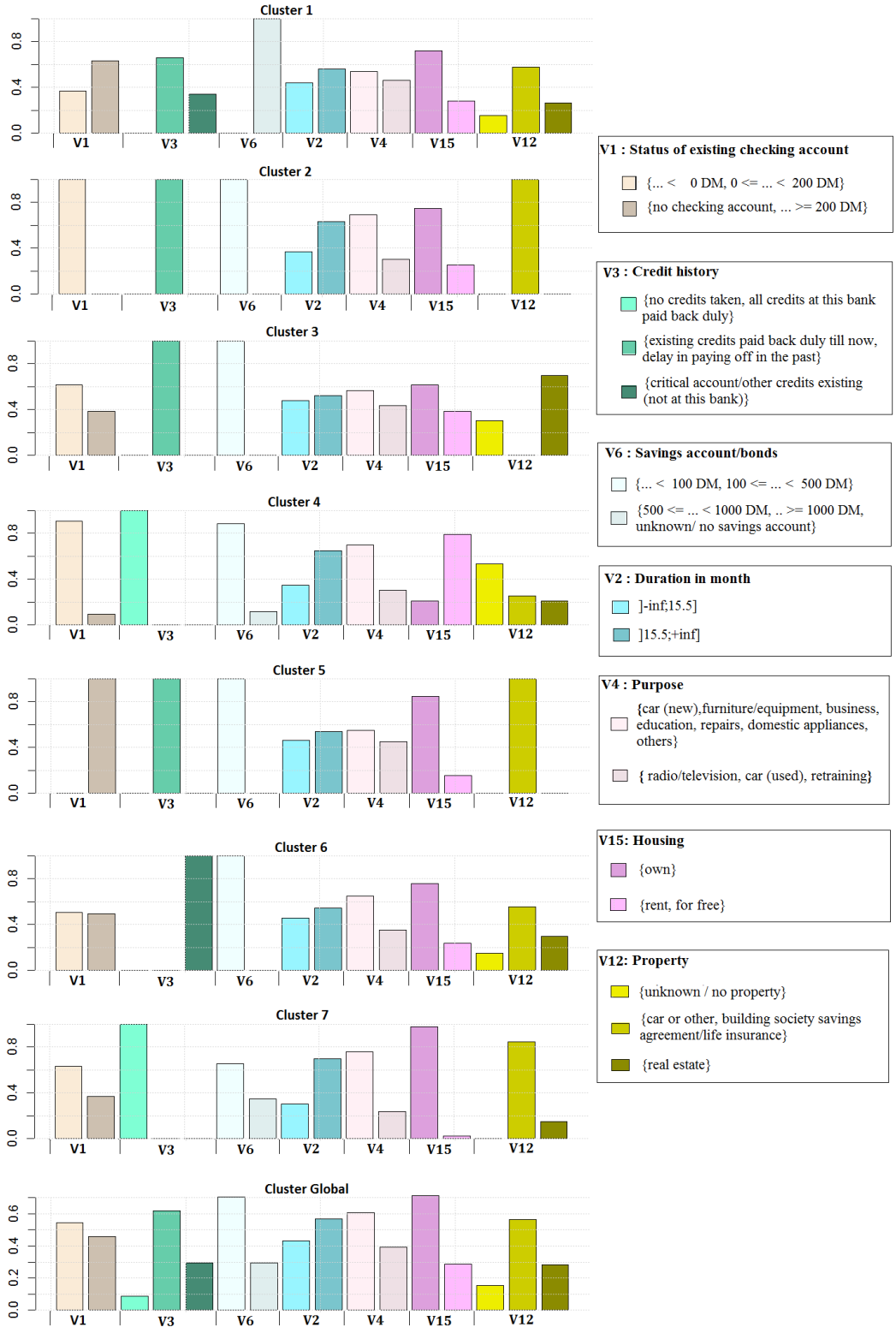

FIGURE 6.9 – Description des clusters de la partition optimale

— Les gens du **cluster 6** sont de gens ayant un autre crédit dans une autre banque (voir la variable V3). Cependant, la majorité de ces gens possèdent des biens immobiliers ou une assurance de vie, des voitures, etc (voir la variable V12). Ils ont également leur propre maison (voir la variable V15).

Bien que les deux clusters 3 et 7 sont les moins purs en termes de la classe 1, mais ils fournissent des informations différentes des trois premiers clusters et qui semblent être pertinentes :

— **Le cluster** 3 contient des gens qui ont déjà un crédit dans cette banque. Les dates du remboursement des parts de ce crédit ont été respectées jusqu'à présent. La majorité de ces gens possèdent des biens immobiliers.

— **Le cluster 7** contient spécifiquement des gens qui n'ont pas eu de crédits auparavant ou bien tous les crédits déjà pris par ces gens ont été remboursés. Ces gens possèdent soit des biens immobiliers, soit des assurances de vie ou des voitures, etc.

En ce qui concerne la deuxième classe, c'est-à-dire, la classe ayant des gens à fort risque crédit, on trouve que celle-ci peut être partitionnée en deux différents sous-groupes (cluster 2 et cluster 4).

— **Le cluster 2** contient spécifiquement des gens qui ont une petite somme dans leur compte courant et dans leur compte d'épargne. Ils ont déjà un crédit en cours dans cette banque ou les crédits qu'ils ont eu dans le passé n'ont pas été remboursés dans leurs délais. Ces gens possèdent des voitures, des assurances de vie, etc.

— **Le cluster 4** contient des gens qui ont une somme entre 0 et 200 DM dans leur compte courant. Ces gens n'ont pas eu de crédits auparavant ou bien tous les crédits déjà pris par ces gens dans cette banque ont été remboursés. Cependant, la majorité de ces gens n'ont pas de propriétés. Ils louent une maison ou ils sont hébergés gratuitement chez quelqu'un de proche.

6.3.3 Discussion

À l'aide de cet exemple illustratif, nous avons pu montrer la capacité de l'algorithme des K-moyennes prédictives proposé dans cette thèse à atteindre l'objectif du clustering prédictif du deuxième type et donc découvrir les différentes raisons qui peuvent mener à une même prédiction. Pour le jeu de données German, nous avons pu montrer les différents profils des demandeurs de crédits existant dans la base de données. De plus, nous avons montré la capacité de cet algorithme à fournir des résultats facilement interprétables par l'utilisateur.

6.4 Conclusion & Perspectives

6.4.1 Conclusion générale

Le clustering prédictif est un cadre dans lequel décrire et prédire d'une manière simultanée est un nouvel aspect d'apprentissage supervisé qui englobe à la fois les caractéristiques de la classification supervisée et du clustering. Afin de mettre en évidence les difficultés liés à ce type d'apprentissage, nous avons commencé dans cette thèse par introduire les concepts clés du clustering et de la classification supervisée. Nous avons ensuite présenté une liste non exhaustive des différentes approches potentielles permettant d'atteindre l'objectif souhaité.

Pour atteindre l'objectif de la thèse qui est la recherche d'un modèle "interprétable" permettant de décrire et de prédire d'une manière simultanée, nous avons choisi de modifier l'algorithme des K-moyennes standard. Cette version modifiée contient 7 étapes dont chacune peut être supervisée indépendamment des autres. Dans cette thèse, nous nous sommes intéressés à la supervision

de quatre étapes de cet algorithme, à savoir : 1) le prétraitement des données, 2) l'initialisation des centres, 3) le choix de la meilleure partition et 4) la mesure d'importance des variables.

Pour l'étape du prétraitement supervisé, nous avons pu montrer dans le chapitre 3 que les deux prétraitements proposés "Conditional Info" (CI) et Binarization (BIN) ont la capacité d'aider l'algorithme des K-moyennes standard à atteindre l'objectif du clustering prédictif. En effet, CI et BIN permettent d'écrire une distance dépendante de la classe permettant d'établir une relation entre la proximité des instances en termes de distance et leur classe d'appartenance. Ces méthodes de prétraitement sont donc capables de modifier indirectement la fonction du coût de l'algorithme des K-moyennes standard dans le but de l'adapter au problème du clustering prédictif. Les expérimentations menées dans le chapitre 3 ont montré que : *i*) lorsque l'axe du prédiction est privilégié, l'algorithme des K-moyennes standard précédé par Conditional Info fournit des résultats significativement meilleurs que l'algorithme des K-moyennes standard (précédé par les prétraitements non supervisés), *ii*) lorsque l'axe de description est privilégié, les prétraitements supervisés (Conditional Info et Binarization) parviennent à construire de bonnes matrices de Gram relativement à la variable cible. Nous avons pu montrer également que ces méthodes de prétraitement supervisées aident l'algorithme des K-moyennes standard à fournir des résultats facilement interprétables par l'utilisateur.

Pour l'étape d'initialisation des centres, nous avons présenté dans le chapitre 4 l'influence de l'utilisation d'une méthode d'initialisation supervisée ou non supervisée sur la qualité (au sens du clustering prédictif) des résultats générés par l'algorithme des K-moyennes standard. Dans ce cadre d'étude, nous avons proposé trois méthodes supervisées d'initialisation des centres (une déterministe et deux basées sur l'aléatoire) : Rocchio-And-Split (RS), KMeans++R et S-Bisecting. À l'aide de ces méthodes, nous avons pu montrer qu'une bonne méthode supervisée d'initialisation a la capacité d'aider l'algorithme des K-moyennes standard à atteindre l'objectif du clustering prédictif. Lorsque l'axe de prédiction est privilégié (le cas du clustering prédictif du premier type), nous avons pu montrer que quel que soit le prétraitement utilisé (supervisé ou non supervisé), l'algorithme des K-moyennes standard précédé par la méthode Rocchio-And-Split (RS) fournit de meilleures performances prédictives par rapport aux autres méthodes. Il est important de signaler qu'en raison de l'utilisation de RS (méthode déterministe), l'algorithme des K-moyennes n'est exécuté qu'une seule fois. Lorsqu'on cherche à réaliser le compromis entre la description et la prédiction (cas du clustering prédictif du deuxième type), nous avons pu montrer que quel que soit le prétraitement utilisé, la méthode S-Bisecting (SB) et la méthode K++R sont celles qui aident l'algorithme des K-moyennes standard à obtenir un bon compromis entre la description et la prédiction par rapport aux autres méthodes.

Pour l'étape du choix de la meilleure partition, nous avons commencé par présenter dans le chapitre 5 l'influence de l'utilisation d'un critère supervisé ou non supervisé sur la qualité des résultats obtenus par l'algorithme des K-moyennes au sens du clustering prédictif du premier type. Les résultats expérimentaux ont montré que l'utilisation de l'indice de rand ajusté (ARI) pour choisir la meilleure partition permet à l'algorithme des K-moyennes standard de gagner sur l'axe de prédiction. En raison d'absence d'un critère analytique permettant de mesurer la qualité des résultats issus des algorithmes du clustering prédictif du deuxième type, nous avons proposé dans le chapitre 5 une version supervisée de l'indice de Davies-Bouldin, nommée SDB. Cet indice est basé sur une nouvelle mesure de dissimilarité capable d'établir une relation entre la proximité des instances en termes de distance et leur classe d'appartenance : deux instances sont considérées comme similaires suivant cette nouvelle mesure si et seulement si elles sont proches en termes de distance et ont également la même étiquette. Grâce à cette version supervisée de l'indice de Davies-Bouldin, nous avons pu surmonter le problème de la non corrélation entre les clusters et les classes. Dans ce cadre, les résultats expérimentaux ont montré que l'indice SDB

parvient bien à mesurer le compromis entre la description et la prédiction.

Pour l'étape de la mesure d'importance des variables, nous avons proposé dans l'annexe E une méthode supervisée permettant de mesurer l'importance des variables après la convergence du modèle. L'importance d'une variable est mesurée dans cette étude par son pouvoir à prédire l'ID-clusters. Le problème dans ce cas devient un problème "uni-varié" de la classification supervisée. Il s'agit ici de remplacer la variable cible par l'ID-clusters obtenu par l'algorithme du clustering prédictif et ensuite de mesurer la capacité de chaque variable à prédire correctement l'ID-cluster en utilisant un des algorithmes de la classification supervisée. Dans cette étude, nous n'avons pas considéré les interactions qui peuvent exister entre les variables (e.g., une variable n'est importante qu'en présence des autres). Ceci fait l'objet des futurs travaux (parmi d'autre, voir la section 6.4.2).

Dans les chapitres précédents, chaque étape a été traitée individuellement. Dans le début de ce chapitre de synthèse, nous avons regroupé l'ensemble des méthodes supervisées proposées au cours de cette thèse (prétraitement, initialisation et le critère d'évaluation pour choisir la meilleure partition) dans un seul algorithme nommé, K-moyennes prédictives. Dans le cadre du clustering prédictif du premier type, les résultats expérimentaux ont montré qu'avec la supervision des deux étapes de prétraitement et d'initialisation, l'algorithme des K-moyennes parvient à être meilleur ou très compétitif avec les algorithmes de la littérature tel que l'arbre de décision et l'arbre de clustering prédictif. Dans le cadre du clustering prédictif du deuxième type, nous avons pu montrer que l'algorithme proposé dans cette thèse arrive à découvrir les différentes raisons qui peuvent mener à une même prédiction et donc découvrir la structure interne de la variable cible. Les résultats fournis par notre algorithme sont présenté sous forme d'histogrammes facilement interprétables par l'utilisateur.

6.4.2 Perspectives

Parmi les différents travaux futurs qui pourraient être envisagés suite à cette thèse, nous proposons les pistes suivantes :

1. **Modification de la fonction du coût de l'algorithme des K-moyennes** : l'algorithme des K-moyennes standard est basé sur une distance (souvent la distance Euclidienne) pour mesurer la proximité entre les instances. Cette distance n'accorde aucune importance à l'étiquetage des instances : deux instances de différentes étiquettes peuvent être considérées comme similaires si elles sont proches l'une de l'autre. Pour surmonter ce problème l'utilisation d'une mesure qui permet d'établir une relation entre la proximité des instances et leur classe d'appartenance s'avère nécessaire. Dans ce contexte, l'utilisation de la nouvelle mesure de dissimilarité proposée dans cette thèse dans le chapitre 5 pourrait résoudre ce problème.

2. **La prédiction de la classe des nouvelles instances** : durant cette thèse, la prédiction de la classe des nouvelles instances est effectuée à l'aide de l'approche un plus proche voisin : chaque nouvelle instance prend la classe du cluster qui lui est le plus proche. Dans certains cas (*e.g.,* le cas de déséquilibre des classes), l'utilisation du modèle un plus proche voisin s'avère insuffisant. En effet, il se peut que le cluster le plus proche de la nouvelle instance ait une étiquette différente de celle-ci. La solution la plus intuitive qui pourrait résoudre ce problème et d'améliorer d'avantage la qualité prédictive de l'algorithme des K-moyennes prédictives est l'utilisation du modèle k-plus proches voisins ($k > 2$).

3. **Amélioration des performances prédictives** : dans le cadre du clustering prédictif du premier type, l'axe de prédiction est privilégié. Pour améliorer davantage la performance prédictive de l'algorithme des K-moyennes prédictives du premier type, le modèle LVQ pourrait être utilisé. Ce dernier prendrait en entrée les centres générés par l'algorithme des K-moyennes prédictives après convergence. Ce modèle pourrait également être appliqué après l'étape d'initialisation des centres.

4. **Mesure d'importance des variables** : dans les travaux réalisé à ce sujet (voir l'annexe E), nous avons suivi une méthode uni-variée où chaque variable est traitée indépendamment des autres. Dans ce cas, nous n'avons pas étudié les interactions qui peuvent exister entre les variables. En effet, dans certains cas, une variable descriptives n'est importante qu'en présence d'une ou d'autres variables. Pour les travaux futurs, il est important de considérer donc ces interactions.

Liste des publications

Revue internationale

[10] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols. **Supervised pre-processings are useful for supervised clustering**. In *Springer Series Studies in Classification, Data Analysis, and Knowledge Organization*, Bremen, 2015.

Conférences internationales

[13] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols. **Evaluation of predictive clustering quality**, in MBC2, on Model Based clustering and classification (MBC2,2016).

[70] Vincent Lemaire, Oumaima Alaoui Ismaili, and Antoine Cornuéjols. **An initialization scheme for supervized k-means**. In *2015 International Joint Conference on Neural Networks, IJCNN 2015, Killarney, Ireland*, July 12-17, 2015, pages 1–8, 2015.

[59] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols. **A Supervised Methodology to Measure the Variables Contribution to a Clustering**, in Neural Information Processing - 21st International Conference, (ICONIP 2014), Kuching, Malaysia. pp. 159–166, 2014.

Revue nationale

[12] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols : **Clustering prédictif : décrire, prédire et interpréter simultanément** à venir sur invitation suite à la conférence RJCIA, in Revue d'intelligence Artificielle (RIA).

Conférences nationales

[60] Oumaima Alaoui Ismaili, Vincent Lemaire, and Antoine Cornuéjols. **Une méthode supervisée pour initialiser les centres des k-moyennes**. *Extraction et Gestion des Connaissances (EGC), Reims*, 2016.

[11] Oumaima Alaoui Ismaili, Vincent Lemaire, and Antoine Cornuéjols. **Une initialisation des K-moyennes à l'aide d'une décomposition supervisée des classes**. *Congrès de la Société Française de Classification (SFC)*, Nantes, 2015.

[9] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols. **Classification à base de clustering ou comment décrire et prédire simultanément ?**. Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA),Rennes, pages 7-12, 2015.

[8] Oumaima Alaoui Ismaili, Vincent Lemaire, Antoine Cornuéjols. **Une méthode basée sur des effectifs pour calculer la contribution des variables à un clustering**, In Atelier CluCo de la conférence Extraction et Gestion des Connaissances (EGC 2014), Rennes.

Démonstration

[69] Vincent Lemaire, Oumaima Alaoui Ismaili. **Un outil pour la classification à base de clustering pour décrire et prédire simultanément.** In Atelier Clustering and Co-clustering (CluCo), EGC 2016, Reims.