AVIGNON UNIVERSITÉ

Démissions d'un organisme bancaire

Groupe IA-CLA
Damien Dallon
Nathanaël Lefèvre

15 janvier 2023

Master 2 informatique
ILSEN/IA
UE Business intelligence & Systèmes décisionnels
ECUE Application Business Intelligence

Responsable
Vincent Labatut

UFR
SCIENCES
TECHNOLOGIES
SANTÉ

AVIGNON UNIVERSITÉ
CERI

CENTRE
D'**ENSEIGNEMENT**
ET DE **RECHERCHE**
EN **INFORMATIQUE**
ceri.univ-avignon.fr

Sommaire

1 Présentation

Si vous utilisez LaTeX pour écrire votre rapport, veuillez consulter le tutoriel fourni à l'adresse suivante : https://www.overleaf.com/latex/templates/modele-rapport-uapv/pdbgdpzsgwrt. Attention, il vous faut accéder au *code source* pour bénéficier des commentaires qu'il contient, et qui complètent le texte apparaissant dans le PDF produit.

Le reste du document présent décrit la structure imposée pour votre rapport. Vous devez obligatoirement la suivre, en respectant les titres et la numérotation indiquée. Les listes de points à l'intérieur des sections sont là pour décrire le contenu que vous devez produire. **Ne reprenez pas ces points *verbatim* : il s'agit simplement d'indications.**

> **Remarque :** Si vous faites une copie de ce document, configurez Overleaf pour qu'il le compile avec LuaLaTeX. De plus, vérifiez que le correcteur orthographique sélectionné est bien celui destiné au français.

1.1 Contexte

- Rappelez brièvement le contexte du projet et ses objectifs.

Un organisme bancaire a fait appel à nous pour mettre en place une solution de Machine Learning afin de détecter ses clients sociétaires qui sont sur le point de le quitter. Le but principal pour la banque est d'adapter sa relation client auprès de ceux identifiés comme démissionnaire afin de les convaincre de rester.

Par ailleurs, l'explicabilité du modèle est également un point important pour l'organisme bancaire qui souhaite savoir quelles caractéristiques influent le plus sur la classification.

1.2 Organisation

- Décrivez la composition du groupe, la répartition du travail.
- Indiquez comment votre travail a été organisé dans le temps.
- Indiquez aussi comment les tâches ont été distribuées entre les membres du groupe (qui a fait quoi ?). Il s'agit de décrire les tâches **individuelles**, donc vous devez les décomposer à un niveau suffisamment détaillé pour permettre de décrire ce que chaque membre du groupe a fait.
- Indiquez quelles bibliothèques vous avez utilisées, en expliquant leur rôle. S'il s'agit de bibliothèques différentes de celles utilisées en TP, **expliquez** la raison de votre choix.

2 Données

2.1 Caractéristiques

- Décrivez les données et l'exploration que vous en avez faite.
- Soyez exhaustifs : fichiers de données, liste des attributs, nature et codage des valeurs, interprétation, unité dans laquelle la variable est exprimée (pour les variables numériques), etc.

Bon commençons à travailler, car c'est important le travail :)

Nous avons accès à des données extraites en 2007 décrivant les sociétaires de l'organisme

2.2 Nettoyage

- Listes les types d'erreurs ou d'incohérences que vous avez rencontrées dans les données.

- Expliquez comment vous les avez corrigées, ou plus généralement, comment vous avez résolu ces problèmes. Vous pouvez envisager plusieurs méthodes pour traiter ces problèmes, afin de les comparer plus tard à travers les résultats obtenus.

2.3 Analyse descriptive

- Donnez les résultats de votre analyse descriptive des données nettoyées. Si vous envisagez plusieurs méthodes de nettoyage, concentrez-vous ici sur celle qui aboutit ensuite aux meilleurs résultats en Section 4.
- Considérez **chaque attribut** séparément : distribution, principales statistiques, et discussion. Si plusieurs attributs présentent les mêmes caractéristiques, vous pouvez les présenter de façon groupée.
- Étudiez également les associations entre **paires** d'attributs (y compris la classe à prédire), en procédant là encore visuellement (via des graphiques) et objectivement (via des statistiques). Discutez.

Remarque : en pratique, vous devez faire une première analyse descriptive *avant* le nettoyage, pour détecter les problèmes dans les données ; puis une seconde analyse descriptive *après* le nettoyage, pour étudier les propriétés des données propres. Pour éviter les redondances, on ne vous demande pas de décrire les deux dans ce rapport.

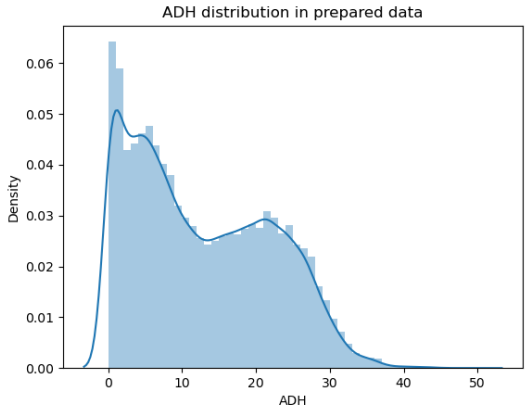
En Section 2.2, vous devez uniquement vous concentrer sur les problèmes détectés dans les données et comment vous les résolvez, sans donner l'analyse descriptive exhaustive.

En Section 2.3, vous devez décrire votre analyse descriptive complète des données nettoyées.

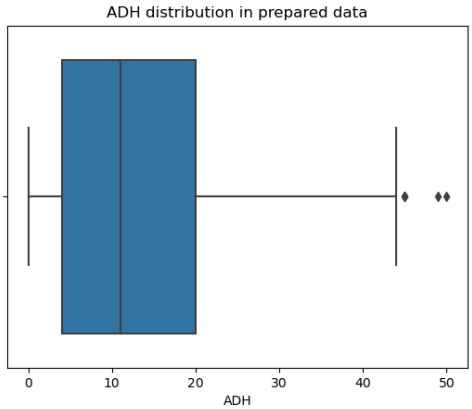
2.3.1 Analyse par attribut

Une fois nos données nettoyées, voici les profils pour chaque attribut :

ADH - Durée de la période d'adhésion, en années



(a) Densité [ADH]



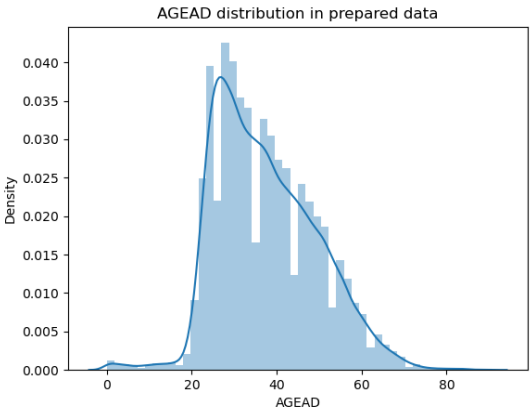
(b) Diagramme moustache [ADH]

	count	mean	std	min	25%	50%
ADH	45185.00	12.57	9.33	0.00	4.00	11.00

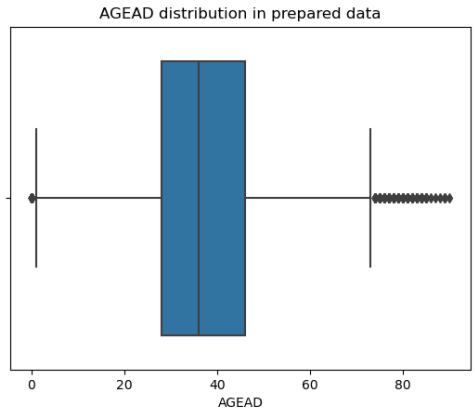
	75%	max
ADH	20.00	50.00

On peut observer deux profils de sociétaires : ceux qui restent entre 0 et 13 ans et ceux qui restent entre 10 et 30 ans voire plus. Notons que les deux groupes sont assez bien réparti puisque la moyenne est à 12.57, c'est à dire à la frontière des deux groupes.

AGEAD - Âge du client à l'adhésion, en années



(a) Densité [AGEAD]

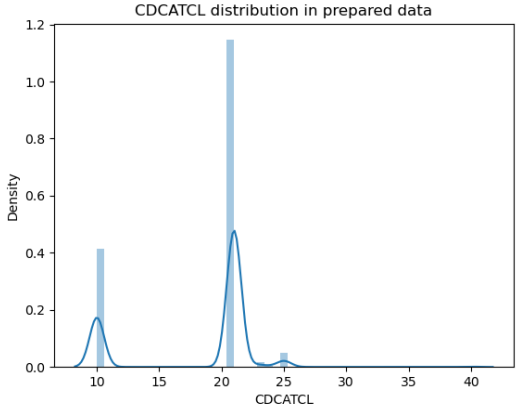


(b) Diagramme moustache [AGEAD]

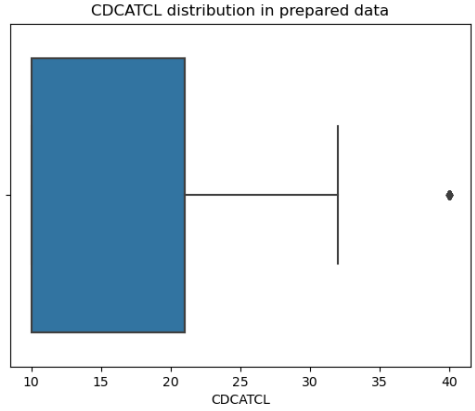
	count	mean	std	min	25%	50%	75%	max
AGEAD	45185.00	37.45	11.90	0.00	28.00	36.00	46.00	90.00

Nous observons qu'il y a très peu d'adhésion entre 0 et 20 ans. On peut imaginer qu'à de bas âges, ce sont surtout les parents qui ouvrent un compte à leurs enfants. Le nombre de sociétaire diminue pour un age d'adhésion allant d'environ 25 ans jusqu'à 80 ans voire plus, avec un maximum de 90 ans. notons que 75% des sociétaire ont 46 ans ou moins et que la moyenne d'age à l'adhésion est de 37.45 ans.

CDCATCL - Type de client (catégorie)

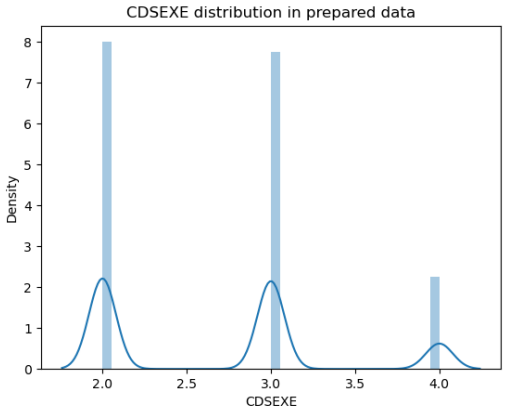


(a) Densité [CDCATCL]

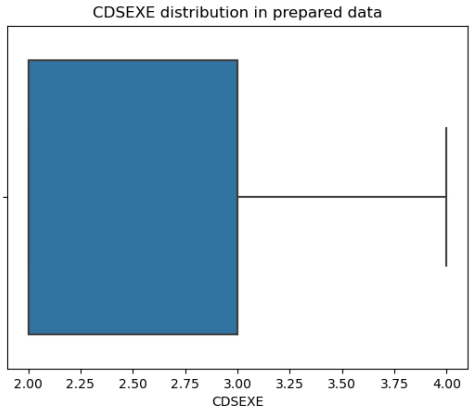


(b) Diagramme moustache [CDCATCL]

L'organisme bancaire classe ses clients en 3 catégories et on observe que la catégorie 21 est largement majoritaire Répond sur discord !

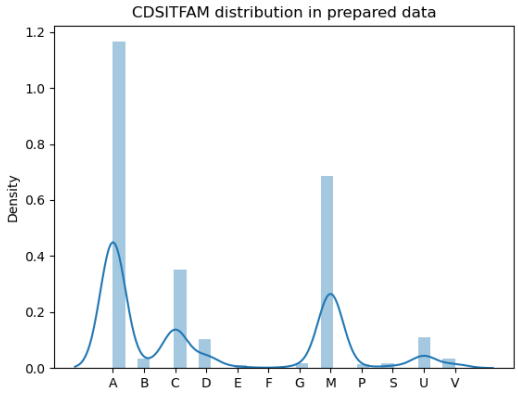


(a) Densité [CDSEXE]

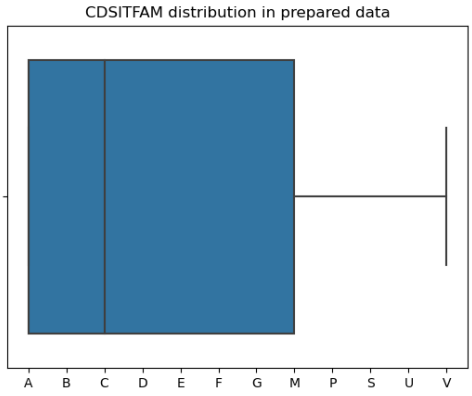


(b) Diagramme moustache [CDSEXE]

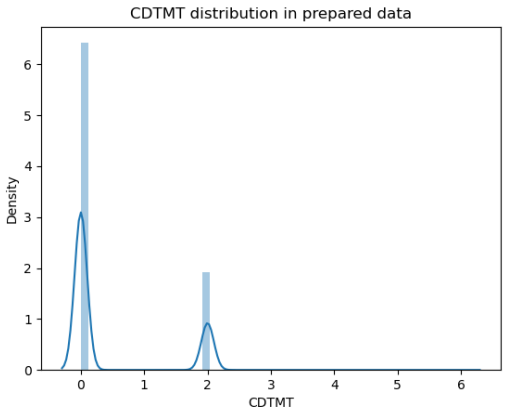
	count	mean	std	min	25%	50%	75%	max
CDSEXE	45185.00	2.68	0.68	2.00	2.00	3.00	3.00	4.00



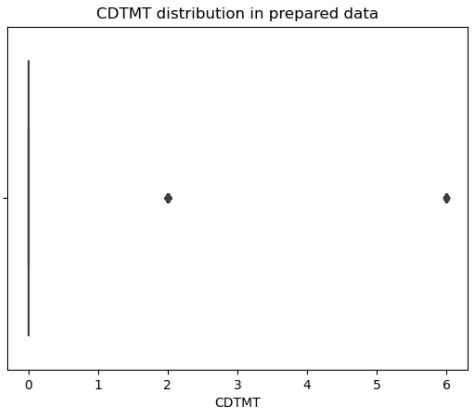
(a) Densité [CDSITFAM]



(b) Diagramme moustache [CDSITFAM]

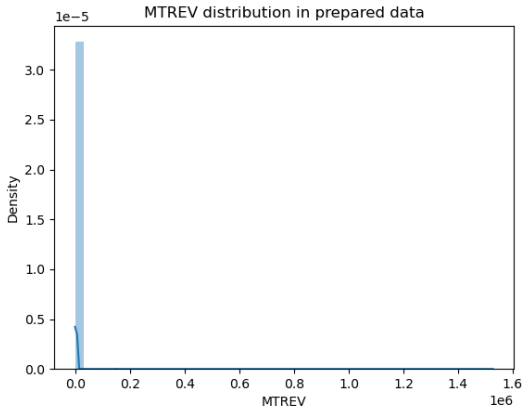


(a) Densité [CDTMT]

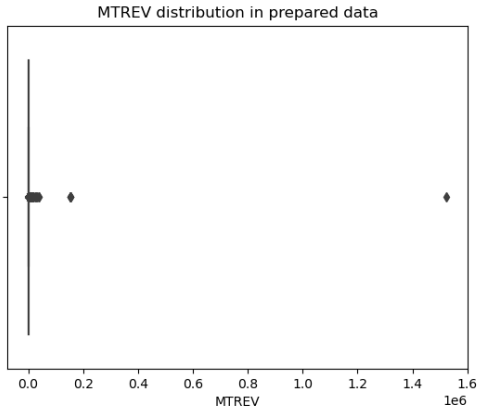


(b) Diagramme moustache [CDTMT]

	count	mean	std	min	25%	50%	75%	max
CDTMT	45185.00	0.46	0.85	0.00	0.00	0.00	0.00	6.00



(a) Densité [MTREV]



(b) Diagramme moustache [MTREV]

	count	mean	std	min	25%	50%	75%	max
MTREV	45185.00	420.65	7377.32	0.00	0.00	0.00	0.00	1524490.00

3 Méthodes

3.1 Outils de fouille

- Décrivez très brièvement les algorithmes que vous avez appliqués, en indiquant ceux qui ont été imposés (le cas échéant), ceux que vous avez sélectionnés, ceux qui ont été vus en cours mais écartés, et en **justifiant** ces choix.
- Pour les algorithmes retenus, indiquez quels sont les paramètres et options acceptés par les implémentations Python utilisées. Soyez **exhaustifs**, en listant tous les paramètres et options, et en expliquant pour chacun son rôle vis-à-vis de l'algorithme concerné.
- Indiquez sur quels paramètres vous avez joué pour tenter d'améliorer les résultats, en **justifiant** vos choix.

3.2 Recodage

- Certaines méthodes nécessitent un recodage des données pour pouvoir être appliquées : le cas échéant, expliquez comment vous avez procédé.
- Pour chaque décision que vous prenez, vous devez **expliquer** et **justifier** votre choix.

3.3 Évaluation

- Expliquez la méthode expérimentale utilisée pour évaluer la qualité des résultats, en **justifiant** vos choix (décomposition des données en apprentissage/validation/test, validation croisée, etc.).

- Décrivez la (ou les) mesure(s) utilisée(s) pour quantifier les performances, en **justifiant** là encore. Vous devez notamment donner une description **formelle** de la mesure (i.e. sa formule).
- Le cas échéant, indiquez la (ou les) méthode(s) statistiques utilisée(s) pour comparer ces mesures entre elles, en **justifiant** votre décision.

3.4 Implémentation

- Décrivez le script rendu, en expliquant quel traitement est réalisé, notamment quelles classes de quelles bibliothèques sont utilisées, et comment elles s'enchaînent.
- Incluez dans cette description les éventuels prétraitements (en plus des méthodes de classification proprement dites).
- Attention, vous devez **décrire** votre script, et non **pas** inclure du code source dans votre rapport.

4 Résultats

> **Attention :** de façon générale, dans cette section, ne vous contentez pas de donner des résultats bruts. Vous devez montrer que vous êtes allés plus loin que cela en expliquant comment vous interprétez vos résultats par rapport au contexte (données, objectifs, application…).

4.1 Performances individuelles

- Donnez les résultats obtenus pour les différents algorithmes appliqués sur le jeu d'apprentissage (du moins : pour ceux qui possèdent une étape d'apprentissage), en présentant ça sous forme compacte au moyen de tableaux.
- Commentez et interprétez ces résultats. Détectez-vous des cas de *sous*-apprentissage ?
- Définissez une sous-section par algorithme.

4.2 Comparaison

- Donnez les résultats individuels obtenus pour les différents algorithmes/paramétrages appliqués sur le jeu de **validation**. Discutez l'évolution par rapports aux résultats obtenus sur le jeu d'apprentissage.
- Là encore, vous devez donner votre interprétation des résultats, et ne pas vous arrêter à une succession de tableaux et de graphiques. Détectez-vous des cas de *sur*-apprentissage ?
- Comparez les résultats obtenus par les différents algorithmes/paramétrages, de manière à identifier celui qui semble le plus adapté à nos besoins.

4.3 Généralisation

- Donnez les résultats pour l'algorithme/paramétrage sélectionné sur le jeu de test. Pour rappel, il ne doit y en avoir qu'**un seul** : il ne s'agit plus de comparer les modèles entre eux, mais d'évaluer le pouvoir de généralisation du meilleur modèle obtenu à l'étape précédente.
- Discutez de sa faculté de généralisation : les résultats obtenus sur le jeu de test sont-ils du même niveau que ceux obtenus auparavant sur les autres jeux de données ? Statistiquement parlant, sont-ils **significativement** différents ou pas ?

4.4 Interprétation

- Décrivez les résultats de votre analyse destinée à identifier les attributs (et leurs valeurs) pertinents pour effectuer la prédiction demandée.
- Discutez ces résultats, notamment la nature des attributs et valeurs identifiés. Par exemple, la nature des attributs est-elle surprenante ou pas, relativement au problème posé ? Quels enseignements pouvez-vous en tirer du point de vue applicatif, toujours pour le problème posé dans le sujet ?

5 Conclusion

- Résumez très brièvement le travail accompli.
- Critiquez le projet : indiquez ce que vous avez apprécié, expliquez ce que le projet vous a apporté, précisez les aspects qui posent problème ou qui étaient ignorés mais que vous auriez voulu aborder. Ce point-là ne sera pas pris en compte pour l'évaluation du projet, mais permettra de l'améliorer le semestre prochain.
- Critiquez votre travail en indiquant les points positifs et les points négatifs (notamment les aspects que vous n'avez éventuellement pas traités).
- Proposez des solutions permettant de résoudre les limitations de votre travail.
- Proposez des perspectives sur ce projet, en indiquant comment le travail pourrait être étendu : analyses supplémentaires, problèmes connexes, etc.

Bibliographie. En ce qui concerne les références bibliographiques :

- Listez toutes les références bibliographiques citées dans le reste du document (en utilisant **BibTeX** si vous écrivez le rapport en LaTeX : par exemple [1], cf. le tutoriel fourni).
- Toute référence listée doit être citée **explicitement** et **à propos**, quelque part dans votre document.

Références

[1] Y.-C. Wei et C.-K. Cheng. « Towards efficient hierarchical designs by ratio cut partitioning ». In : *IEEE International Conference on Computer Aided Design*. 1989, p. 298-301. doi : 10.1109/ICCAD.1989.76957.