Projet Clustering – Sujet B – Demande de crédit

MASTER 1 INFORMATIQUE

Objectif de l'étude

(1) Construire une typologie à partir d'un jeu de données. Le processus peut intégrer des étapes de préparation de données. Le choix du nombre de groupes est à votre discrétion, il doit être justifié.

(2) Fournir une appréciation quantitative de la qualité de votre partition.

(3) Fournir une interprétation des groupes c.-à-d. une lecture éclairée de la nature des groupes. Appuyer les calculs numériques (moyennes ou proportions conditionnelles vs. marginales, valeurs tests, rapports de corrélation, v de Cramer, …) et le discours par des représentations graphiques sera apprécié.

(4) Fournir un programme de déploiement Python permettant d'appliquer votre modèle d'affectation sur des individus supplémentaires. Elle prend en entrée a minima un fichier Excel avec exactement la même structure que le fichier des variables actives.

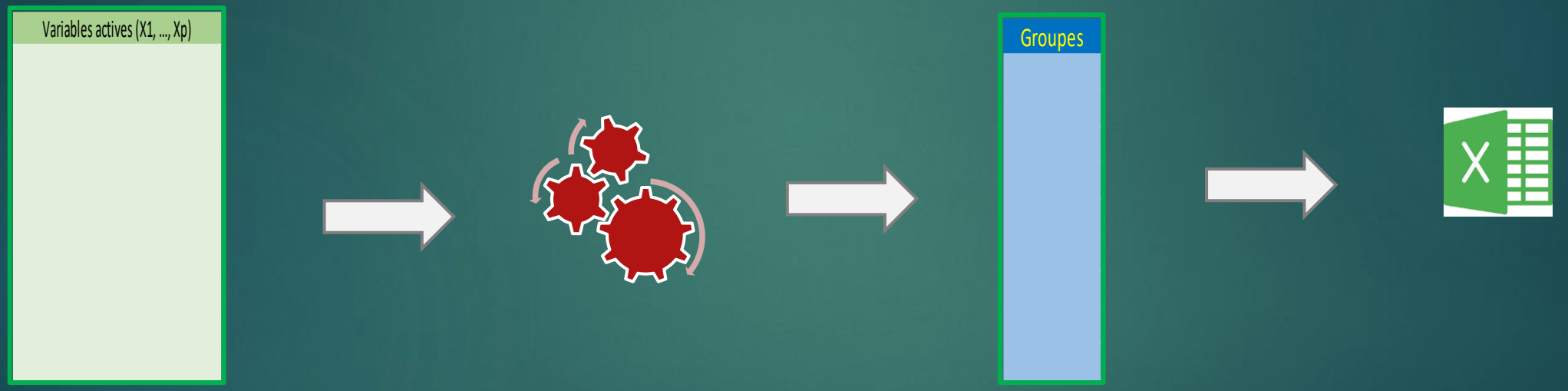
Notebook (.IPYNB)

Vous retracez votre étude dans un notebook, avec les caractéristiques suivantes :

1. Subdivisé en sections apparaissant clairement (voir Markdown : https://www.markdownguide.org/basic-syntax/)
2. Pour chaque traitement, vous devez annoncez ce que vous souhaitez faire ou montrer
3. Il doit comporter du code Python
4. Les sorties des traitements doivent être lisibles et parfaitement séquencées (ne faites pas de successions de print dans la même cellule de traitements par ex.)
5. Les résultats doivent commentés. Vous devez indiquer ce que l'on doit lire et comprendre.
6. Des graphiques sont souvent plus parlants que de longs tableaux illisibles et de grands discours.
7. Pour un plan type votre notebook, regardez du côté de la méthodologie CRISP-DM (doc envoyée en fichier attaché dans un message préparatoire, voir en particulier la section « CRISP-DM Outputs »).

Programme de déploiement (.PY)

Vous devez fournir un programme PYTHON permettant de déployer votre modèle d'affectation sur un fichier de données (Excel) avec des individus supplémentaires décrits à l'aide des variables actives.



Fichier XLSX avec les individus supplémentaires décrits par les variables actives (son nom est « deploiement.xlsx »)

Votre programme PYTHON (.PY), qui prend en entrée le fichier et qui calcul le groupe d'affectation pour chaque individu

Groupe attribué pour chaque individu supplémentaire

Cette colonne « groupes » doit être automatiquement sauvegardée dans un fichier « output.xlsx »

Jeux de données à traiter (1)

Décrire le jeu de données avec ses caractéristiques et ses particularités.

Décrire les objectifs supplémentaires en lien avec les spécificités du jeu de données.

Il vous appartient d'introduire tous traitements additionnels qui permettent de mieux asseoir votre analyse exploratoire des données et approfondir notamment l'interprétation des résultats.

Jeux de données à traiter (2)

Le fichier « sujet_b_demande_credit.xlsx » décrit les caractéristiques de n = 1500 demandeurs de crédit. Ils peuvent être de sexe masculin ou féminin (Title), effectuent une demande individuellement ou en couple (conjoint : Spouse_Title). Les variables actives sont quantitatives ou qualitatives (Reason : motif de la demande, Marital Status : statut marital, …, Refunding : vitesse de remboursement) ; les significations des variables sont relativement faciles à comprendre).

Les principaux enjeux ici sont (sans être exhaustif) : réaliser une typologie avec des données mixtes ; l'interprétation mêlant des variables quantitatives et qualitatives ; coder au mieux le programme de déploiement (.py) sur le fichier des individus supplémentaires tenant des transformations (j'imagine) des variables (on n'a pas un pipeline basique, attention).

A rendre

Par e-mail, un fichier archive (ZIP, RAR, 7z, …) contenant :

1. Votre Notebook au format IPYNB
2. Votre Notebook exporté au format HTML

Lorsque je relance le fichier « .ipynb » à partir du fichier de données que je vous ai fourni, je dois obtenir exactement ce que je vois dans le rapport HTML.

3. Votre programme « .py » de déploiement

Lorsque je présente le fichier « deploiement.xlsx » à votre programme Python, il doit produire automatiquement le fichier « output.xlsx ».

Adresse mail

Envoyer par e-mail votre travail (fichier archive en attaché) à l'adresse :

ricco.rakotomalala@yahoo.fr

Avec le sujet : [M1 Info – Clustering – N° de groupe] Nom_1 + Nom_2

Mettez-vous en CC de l'e-mail pour conserver une trace de votre travail.