Projet d'analyse de données

On présente ici les différentes étapes d'un projet d'analyse de données.

1 Projet d'analyse données

Un projet d'analyse de données peut se découper en cinq grandes étapes :

1. Définition des objectifs
2. Données
3. Élaboration et validation des modèles
4. Mise en oeuvre
5. Suivi de la performance et amélioration

Lors de la planification d'un projet, il faut prendre en compte que chaque étape à une importance différente, mais aussi que chacune ne prend pas le même temps d'exécution. Pyle (1999) donne une estimation du temps de chaque étape, ainsi que de leur importance dans la réussite du projet (donné en pourcentage du total, cf. Table 1).

Table 1 – Découpage d'un projet d'analyse des données.

Étape	Temps	Importance
Comprendre le problème	10%	15%
Explorer la solution	9%	14%
Implementer la solution	1%	51%
Préparer les données	60%	15%
Analyser les données	15%	3%
Modéliser les données	5%	2%

On remarque deux faits importants : ce n'est pas parce qu'une étape est très importante qu'elle va prendre beaucoup de temps. L'implémentation de la solution est très importante (sinon il

n'y a pas de résultat), mais ne sera généralement pas très longue à faire (possiblement en quelques lignes de code). À l'inverse, la préparation des données est un étape d'importance moyenne (encore que c'est discutable), mais elle prend la majeur partie du temps du projet. En effet, il faut, par exemple, gérer les données manquantes, les données aberrantes, les éventuels accents pour des données en français, etc.

2 Définition des objectifs

Est-ce que l'on veut : visualiser les données ? explorer et émettre des hypothèses ? tester ? regrouper ? comprendre ? prédire ?

Comment fait-on en pratique ? On pose des questions ! Tout d'abord, il faut clarifier les termes. Qui va utiliser le modèle et comment ? Quelle est la population cible ?

Pourquoi est-ce important d'avoir des objectifs clairs lors d'un projet d'analyse de données ? Cela permet de guider la collecte des données et leur mise en forme. Cela permet de définir un modèle adéquat (e.g. classification vs prédiction). Cela permet d'analyser les résultats à la lumière de l'objectif et donc de permettre à d'autres personnes de juger de la pertinence de ceux-ci. Il est important de définir les objectifs avant de s'intéresser aux données pour ne pas être biaisé par celles-ci.

Exemple

La Banque National du Canada voudrait lancer un nouveau produit d'épargne et vous donne accès à sa base de données clients.
Mauvais objectif : Analysez les données de la base clients.
Meilleur objectif : Pouvez-vous prédire quels clients vont acheter le nouveau produit d'épargne ?

Exemple

L'équipe du hockey des Canadiens de Montréal souhaite mieux connaître ses adversaires pour développer des nouvelles tactiques de jeu.
Mauvais objectif : Analysez les données des adversaires.
Meilleur objectif : Pouvez-vous caractériser le style de jeu des adversaires dans l'optique d'y détecter des points faibles ?

Exemple

Pharmascience souhaite savoir si son nouveau médicament est efficace.
Mauvais objectif : Analysez les données du médicament.
Meilleur objectif : Pouvez-vous déterminer un protocole de tests (statistiques) permettant

de déterminer si le médicament est efficace ?

3 Données

Les données sont le coeur du sujet. Pour être utile, les données doivent être disponibles et de bonnes qualités. Une fois les objectifs définis, on effectue une traitement préliminaire et une exploration basique des données pour ensuite aller vers des modèles plus développés.

3.1 Où trouver des données ?

Réponse simple : Internet ! Voici une liste de sites (non-exhaustives) qui regroupent des jeux de données :

- Google datasets ;
- Kaggle ;
- UC Irvine Machine Learning Repository ;
- Time Series Machine Learning website ;
- Physionet Database.

On peut aussi regarder les sites officiels de sources de données que l'on peut trouver pour une grande partie des pays du monde :

- Canada : StatCan ;
- France : data.gouv.fr ;
- USA : data.gov ;
- Angleterre : data.gouv.uk ;
- etc.

Pour des données sur des sujets plus spécifiques, vous pouvez regarder les différentes branches des gouvernements. Par exemple, le Centre Canadien de cartographie et d'observation de la terre fournit les données géospatiales du Canada (ici).

Lorsque que l'on travaille pour une entreprise, on a généralement accès aux sources de données internes, e.g. base de données sur la production, les clients et les employés, les listes de transactions et de clients potentiels, des informations sur les visites web.

3.2 Qualité

Il y a un dicton populaire en informatique, s'appliquant aussi en analyse de données : "Garbage in, garbage out". En gros, cela veut dire que peut importe à quel point le modèle est sophistiqué, si les données en entrée sont de mauvaise qualité, biaisé, incomplète, … alors les résultats en

sortie seront de mauvaise qualité. Ainsi, cela assure une certaine crédibilité, reproductibilité et utisabilité à nos conclusions.

Mais qu'est-ce que l'on veut dire par qualité des données ? Pour nous assurer de la qualité des données, on pourra se poser les questions suivantes :

— Est-ce que les données sont représentatives de la population cible ?
— Est-ce que les données sont correctes et pertinentes ?
— Est-ce qu'il y a des données manquantes, redondantes, … ?

3.3 Constitution de la base de données

Une fois nos données récupérées, on doit les charger en mémoire pour ensuite pouvoir faire des analyses. En Python, la librairie `pandas` permet de lire la plupart des formats de fichiers auxquels on aura affaire. En ce qui concerne R, il faut utiliser différentes libraries pour charger différents types de données (cf. Table 2).

TABLE 2 – Différentes libraries pour différents formats de fichiers.

Format	Extension	Librarie
Texte	.txt ; .csv	`readr`
Excel	.xlsx	`readxl`
SAS	.sas7bdat	`haven`
SPSS	.sav ; .zsav	`haven`
JSON	.json	`jsonlite`

Depuis une dizaine d'année, le concept de "tidy data" a emergé (cf. Wickham (2014)). Chaque jeu de données "tidy" respecte trois principes :

1. Chaque variable est une colonne du tableau.
2. Chaque observation est une ligne du tableau.
3. Chaque cellule du tableau correspond à une unique mesure.

Cela permet d'avoir une approche unifiée pour l'analyse de données. De manière général, on essaiera toujours de mettre son jeu de données sous format "tidy". Le package `tidyr` en R et la librairie `pandas` en Python permettent de mettre en forme les données en format "tidy".

3.4 Exploration et traitement préliminaire

Une fois les données chargées et sous le format “tidy”, on fait une première exploration des données avant de passer à l’étape d’élaboration du modèle en elle-même. Cette étape, bien que **très importante**, n’est pas le coeur de ce cours. Voici quelques trucs à faire concernant cette première exploration :

— Nettoyage de données : retirer les doublons, uniformiser les modalités, vérifier le format des valeurs spéciales, etc.
— Exploration des données : modalités rares, modalités trop nombreuses, asymétrie, débalancement des classes, valeurs extrêmes ou aberrantes, variables fortement corrélées, valeurs manquantes.

4 Élaboration et validation des modèles

Ce cours concerne l’élaboration et la validation de modèles. Pour l’instant, on peut retenir quatre composantes principales pour ceci :

1. un **espace** (mathématique) dans lequel travailler ;
2. une **distance** permettant de comparer des observations ;
3. un **modèle** (ou algorithme) ;
4. une **fonction** permettant de mesurer la qualité du modèle.

On verra en détail chacune de ses composantes dans les parties suivantes.

5 Mise en oeuvre

Une fois que le modèle est choisi et valider, on peut vouloir le **mettre en production**. La mise en production est le fait de rendre le modèle disponible pour le plus grand nombre. Généralement, cela consiste à automatiser la collecte et le nettoyage des données, pour ensuite “nourir” le modèle créé et sortir des rapports d’analyse des résultats. Cette partie est ce qu’on appelle le **data engineering**. Un **data engineer** s’occupe donc de toute mise en place de la *pipeline* de traitement des données, de sa collecte à la sortie du modèle.

6 Suivi de la performance et amélioration

Finalement, une fois que le modèle est mis en production, il faut assurer un suivi de la performance. En effet, lors de l'arrivée de nouvelles données, il peut arriver que le modèle que l'on a considéré ne soit plus très adapté (e.g. les hypothèses faites ne sont plus correctes). Pour prévenir ce phénomène, on fait un **monitoring** du modèle en vérifiant régulièrement ces performances. On peut aussi vouloir améliorer notre modèle ; par exemple parce qu'on a des hypothèses plus précise sur celles-ci ou bien des données de meilleures qualité.

Références

Pyle, Dorian. 1999. *Data Preparation for Data Mining*. Morgan Kaufmann.

Wickham, Hadley. 2014. « Tidy Data ». *Journal of Statistical Software* 59 (septembre) : 1-23. https://doi.org/10.18637/jss.v059.i10.