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 se déroule généralement en plusieurs phases distintes. On en décompte cinq :

1. Définition des objectifs
2. Collecte et préparation des données
3. Élaboration et validation des modèles
4. Implémentation
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, souvent sous-estimée, mobilise une grande partie du temps du projet, notamment pour gérer les données manquantes, les données aberrantes, ou encore les éventuels accents pour des données en français.

2 Définition des objectifs

Avant toute chose, il faut clarifier ce que l'on cherche à obtenir. Souhaite-t-on visualiser les données ? Tester une hypothèse ? Prédire un comportement ? Segmenter une population ? Une formulation précise des objectifs est donc nécessaire pour orienter le projet.

Pourquoi cette étape est-elle cruciale ? Cela permet de guider la collecte et la structuration des données. Cela permet de définir un modèle adéquat (e.g. classification, régression, …). Cela permet de faciliter l'interprétation et la communication des résultats. Cela permet d'éviter certains biais liés à une exploration aveugle des données.

Comment fait-on en pratique pour formuler un bon objectif ? On pose des questions ! Tout d'abord, il faut clarifier les termes. Qui va utiliser le modèle et comment ? Quelle est la population cible ? Quelle décision sera influencée par les résultats ?

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 : Analyser les données de la base clients.
Meilleur objectif : Peut-on prédire quels clients sont susceptibles d'acheter ce nouveau produit d'épargne ?

Exemple

L'équipe de hockey des Canadiens de Montréal souhaite mieux connaître ses adversaires pour développer de nouvelles tactiques de jeu.
Mauvais objectif : Analyser les données des adversaires.
Meilleur objectif : Peut-on caractériser le style de jeu des adversaires pour identifier leurs faiblesses ?

Exemple

Pharmascience souhaite évaluer l'efficacité d'un nouveau médicament.
Mauvais objectif : Analyser les données du médicament.
Meilleur objectif : Peut-on concevoir un protocole statistique permettant de tester l'efficacité du médicament ?

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, les agences gouvernementales sont souvent de bonnes ressources. 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". Même le meilleur modèle ne peut compenser des données biaisées, incomplètes ou erronées.

Pour nous assurer de la qualité des données, on pourra se poser les questions suivantes :

- Les données sont-elles représentatives de la population cible ?
- Sont-elles exactes, complètes, pertinentes ?

— Y a-t-il des valeurs manquantes, des doublons, des incohérences ?

3.3 Constitution de la base de données

Une fois nos données collectées, il faut 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, plusieurs packages sont utilisés selon le format (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 contient une valeur unique.

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 mise sous le format "tidy", une phase d'exploration préliminaire est nécessaire avant l'étape de modélisation. Cette étape, bien que souvent négligée, est **très importante**, mais elle n'est pas le coeur de ce cours. Cette étape permet de détecter les problèmes potentiels, de mieux comprendre la structure des données et d'orienter les choix méthodologiques. Voici quelques trucs à faire concernant cette première exploration :

— Nettoyage de données : supprimer les doublons, uniformiser les modalités, vérifier le format des valeurs spéciales, etc.

— Exploration des données : identification des modalités rares ou trop nombreuses, analyse des éventuelles asymétries, détection des classes déséquilibrées, identification des valeurs extrêmes ou aberrantes, recherche des corrélations fortes entre les variables, évaluation des 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 :

1. Un **espace** (mathématique) de représentation : il s'agit du cadre mathématique dans lequel on travaille.
2. Une **distance** (ou similarité) : elle permet de comparer les observations entre elles.
3. un **modèle** (ou algorithme) : c'est la méthode utilisée pour apprendre à partir des données.
4. une **fonction de coût** : elle meusre la qualité du modèle.

Ces éléments seront étudiés en détails dans les sections suivantes du cours.

5 Mise en oeuvre

Une fois le modèle choisi et validé, il peut être déployé en **production**. La mise en production signifie le rendre opérationnel dans un environnement réel, souvent en automatisant l'ensemble du processus de traitement des données. Généralement, cela consiste à automatiser la collecte, le nettoyage et la transformation des données, à intégrer le modèle créé dans une application ou un système décisionnel, et à générer des rapports ou des prédictions en temps réel ou à interalles réguliers. Cette partie est le domaine du **data engineering**. Un **data engineer** conçoit et maintient la *pipeline* de traitement depuis la source des données jusqu'à la sortir 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 sa performance dans le temps. En effet, les données évoluent, de même que les comportements qu'elles décrivent. Ainsi, les distributions des données peuvent changer (un phénomène appelé *data drift*), les hypothèses initiales peuvent ne plus être valides ou encore de nouvelles données ou de nouvelles variables peuvent améliorer la performance. Pour surveiller la performance du modèle, on peut faire un **monitoring** régulier des performances. On peut aussi réentraîner le modèle avec des données récentes ou l'améliorer en intégrant de nouvelles hypothèses.

Un bon modèle n'est donc pas seulement performant à un instant donné, il est aussi robuste et adaptable dans le temps.

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.