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).

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 Inventaire et qualité

Qu'est-ce que l'on veut dire par qualité des données ?

— Est-ce que les données sont représentatives de la population cible ?
— Est-ce que les données permetternt de tirer des conclusions de causalité ?
— Est-ce que les données sont fiables ?

3.2 Constitution de la base de données

3.3 Exploration et traitement préliminaire

Source de données :

Quelques liens pour récupérer des données.

Nettoyage de données : cf R (importation, nettoyage, tidyverse, types de variables, retirer les doublons, uniformiser les modalités, vérifier le format des valeurs spéciales, pivot, opérateur pipe, jointure).

Exploration des données : modalités rares, modalités trop nombreuses, asymétrie, débalancement des classes, valeurs extrêmes ou aberrantes, variables fortement corrélées, valeurs manquantes.

Statistiques descriptives

4 Élaboration et validation des modèles

5 Mise en oeuvre

6 Suivi de la performance et amélioration

Références

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