

Analyse de données

Project d'analyse données

Un projet d'analyse de données se découpe en cinq étapes :

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

Remarque

Dans ce cours, on s'intéressera principalement à l'élaboration et à la validation de modèles.

TODO : Planification d'un projet

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 ?

Exemples

1. La Banque National du Canada veut lancer un nouveau produit d'épargne et souhaite mieux connaître ses clients pour prédire s'ils veulent l'acheter.
2. L'équipe de hockey des Canadiens de Montréal souhaite mieux connaître ses adversaires pour développer des nouvelles tactiques de jeu.
3. Pharmascience souhaite savoir si son nouveau médicament est efficace.

Données

- Inventaire et qualité
- Constitution de la base de données
- Exploration et traitement préliminaire

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 permettront de tirer des conclusions de causalité ?
- Est-ce que les données sont fiables ?

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, déséquilibre des classes, valeurs extrêmes ou aberrantes, variables fortement corrélées, valeurs manquantes.

Statistiques descriptives