

Cahier thématique

Analyse de dataset et validation de modeles, sélection de prédicteurs

**Préparé par :** Nesrine Zemirli  
  
© Hafed Benteftifa et Nesrine Zemirli 2015-2018

Ce document ne peut être utilisé dans le cadre d’une formation, publication papier, site internet ou tout support sans mon accord express.

Aucune reproduction, même partielle, ne peut être faite de ce document et de l'ensemble de son contenu : textes, images, etc. sans mon autorisation express. Pour toutes informations, communiquer avec moi sur [info@degenio.com](mailto:info@degenio.com).

|  |  |  |
| --- | --- | --- |
| **Date** | **Version** | **Changement** |
| 20 décembre 2016 | 1.0 | Version initiale |
|  |  |  |
|  |  |  |

**Approche**

Dans le cas de l’apprentissage supervisé, un ensemble de prédicteurs doit être choisi. Dans ce cadre, on va procéder à l’exploration du dataset pour établir la pertinence d’utiliser certaines des variables disponibles dans le dataset.

Un pré-traitement pourra aussi être appliqué sur le dataset afin de corriger des valeurs manquantes ou aberrantes.

**Les étapes sont les suivantes**

* Analyse de données
  + Summarize data
    - Structure des données
  + Évaluation de modèles sur la base de l’erreur RSS
    - Recherche de prédicteurs pertinents avec recherche exhaustive de combinaisons adéquates de prédicteurs
    - Recherche de prédicteur avec approche forward stepwise
    - Recherche de prédicteurs avec approche backward stepwise

**Références**

1. “Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshiran