

Cahier thématique

Projet intégrateur  
Classificateur pour prêt bancaire

**Préparé par :** Hafed Benteftifa  
  
© Hafed Benteftifa et Nesrine Zemirli 2015-2016

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 juin 2016 | 1.0 | Version initiale |
|  |  |  |
|  |  |  |

**Approche**

Dans le cas de l’apprentissage supervisé, on développe un projet complet en se basant sur trois algorithmes de type d'apprentissage supervisé.

**Les étapes sont les suivantes**

* + Chargement des données/dataset
  + Sommaire des données
    - Structure des données
    - Distribution des données
      * Correlations avec la variable cible
      * Correlations entre attributs (redondance)
  + Visualisation des données
    - Histogrammes des attributs
    - Utilisation du boxplot pour détecter les valeurs extremes
    - Analyse des variables numériques
    - Analyse des variables catégorielles
* Nettoyage des données
  + Détermination des valeurs manquantes
  + Problème des outliers ou valeurs extremes/réduction d'impact
* Développement de modèle
  + Encodage des attributs catégoriels
  + Fonction d'apprentissage
  + Détermination des métriques
* Utilisation et comparaison de différents modèles
  + Modèle logistic regression
  + Decision Tree
  + KNN
* Sauvegarde de modele pour utilisation ultérieure
* Utilisation de modele sauvegardé pour la prédiction
* Utilisation d'une interface graphique pour l'utilisation

**Références**

<http://www.neural.cz/dataset-exploration-boston-house-pricing.html>

<http://www.analyticsvidhya.com/blog/2016/01/12-pandas-techniques-python-data-manipulation/>