Plan de travail

Collecte de données

- Télécharger le dataset sur Kaggle
- Domaine de la sante
- 20 instances de training / 20 instances de test
- Problème de classification.

Exploration des données

- La typologie des données
- Le sommaire des principales caractéristiques du dataset
- Explorer les données en utilisant des méthodes visuelles

EDA : Analyse des données explorées

- Quelles variables utiliser ?
- Peut-on prédire à l'aide des infos sur un individus s'il est malade ?
- Transformation des données catégorielles
- Données manquantes
- Données aberrantes

Normalisation des données en utilisant l'une des méthodes

- Decimal scaling
- Min-max normalization
- Z-score

Nettoyage de bruits (en appliquant l'une des méthodes de Binning) :

- En utilisant Equal Width Binning
- Equal Frequency Binning

Exploration des données

- Nettoyer les données (conversion des types)
- Visualisons les données 😊 !

Feature engeneering and Feature selection

- Redondants
- Colinéaires

Choix de l'algorithme : dans notre cas KNN

Apprentissage

- Choix de la valeur de K
- En utilisant le training set

Test et évaluation du model

- En utilisant le testing set
- Ressortir la matrice de confusion
- Déterminons les performances du model !😊