




Modèles Linéaires

Projet Licence 3 Informatique (2023-2024)

Guillaume Metzler

Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Le projet sera à rendre par mail pour la date du 26 avril. Il est attendu un fichier  avec le code ainsi que des commentaires qui expliquent la démarche effectuée.

Pour ce travail, vous travaillerez avec le jeu de données pima qui est associé à ce travail. Ce jeu de données répertorie les cas de diabète dans une population indigène ainsi que différentes caractéristiques comme le nombre de grossesses, le taux de glucose, etc.

Plus précisément, les informations sont les suivantes :

- y : présence ou absence de diabète, c'est la variable que nous chercherons à prédire.
- $X1$: nombres de grossesses enregistrées
- $X2$: pression diastolique
- $X3$: age
- $X4$: insuline
- $X5$: épaisseur de la peau au niveau du triceps
- $X6$: indice de masse corporelle
- $X7$: fonction de risque de diabète
- $X8$: taux de glucose deux heures après un test
- $X9$: indice de mesure 1
- $X10$: indice de mesure 2

Travail attendu L'objectif est de construire un bon modèle prédictif permettant de prédire si la personne est atteinte de diabète ou non. Pour cela vous pourrez suivre les indications suivantes qui devront figurer dans votre étude.

1. Séparer votre jeu de données en deux groupes : un groupe pour entraîner, un groupe pour tester le modèle.
2. Identifier la nature du problème : régression ou classification ?
3. On cherchera à détecter d'éventuelles colinéarités pour écarter les redondances.
4. On pourra trouver le meilleur sous-ensemble de variables qui conduit à un bon modèle sur la base d'un indicateur statistique que vous préciserez.
5. On effectuera une analyse des résidus de ce que vous pensez être le meilleur modèle.
6. on étudiera les performances du modèle en apprentissage et en test.
7. On cherchera à améliorer les performances de ce modèle.