Le fichier python du TP à compléter se trouve à l'adresse suivante :

https : *//github.com/vrunge/TP_Python*

Nous allons analyser les données du dataset Vowel avec différentes méthodes de Machine Learing vues en cours.

1 Statistique descriptive

0) Afficher la taille des tableaux train et test.

L'analyse de statistique descriptive se fera sur la partie d'entraînement des données (train).

1) Afficher le résumé numérique avec la fonction *describe()*.
2) Afficher les boxplots de chaque variable.
3) Afficher les histogrammes de chaque variable.
4) Donner la matrice de corrélation entre les variables.

2 Naive Bayes / LDA / QDA

5) Y-a-t-il indépendance entre les variables conditionnellement à Y?
6) Peut-on considérer les distribution $X^j|Y$ comme gaussienne?
7) Donner les matrices de corrélation conditionnelles à Y.
8) À la vue des résultats obtenus, peut-on dire quelle méthode entre Naive Bayes, LDA et QDA semble la mieux adaptée?

9) Calculer la performance de chacune des trois méthodes (Naive Bayes, LDA et QDA) sur les données de test.
10) Comparer les performances label par label. Que remarquez-vous?
11) BONUS. Comparer les résultats obtenus dans cette section avec des méthodes naïves (régression simples, multiples...)

3 k-NN

12) Avec la méthode des k plus proches voisins, trouver le k qui donne le meilleur résultat sur les données de test.
13) Répéter cette méthode après avoir centré et réduit les variables.

4 Decision Tree

14) Tracez différents arbres de décision. Que remarquez-vous?