1 Présentation de l'activité

Dans cette activité nous allons mettre en œuvre l'algorithme des k plus proches voisins.

La base de travail est une banque de données sur les rugbymen du Top14 (saison 2019-2020). Cette base de données fournit diverses informations sur chaque joueur et notamment son poste et sa corpulence (information sur sa taille et son poids).

Nous allons devoir essayer de prévoir le poste d'un rugbyman à partir de sa corpulence.

L'activité se déroulera en plusieurs étapes :

1. Préparation des données :
   (a) Charger les données issues du `fichierJoueursTop14.csv`.
   (b) Séparer cette banque de données en un partie « données de travail » et une partie « données de tests ».
   (c) Extraire seulement les données d'intérêt pour notre travail : catégorie de poste et corpulence (taille, poids) des joueurs.
2. Présentation visuelle dans un graphique chaque joueur.
   Un joueur est identifié par un point dont l'abscisse est sa taille et l'ordonnée est son poids. Par ailleurs, l'aspect du point représentatif du joueur dépend de la catégorie de son poste.
3. Implémenter l'algorithme des k plus proches voisin. Ceci nécessite de :
   (a) Définir une distance entre 2 joueurs (« proximité de corpulence »).
   (b) Pour un joueur-test identifié seulement par sa corpulence, calculer les distances le séparant de tous les joueurs de la banque de données.
   (c) Parmi ses k plus proches voisins, chercher quelle est la catégorie de poste la plus représentée.
4. Tester la pertinence de l'algorithme.
   Vérifier les prédictions de l'algorithme en s'appuyant sur le jeu de « données de tests ».

2 Préliminaires : visualisation de la banque de données dans un tableur

Avant de se lancer l'activité Jupyter `P7-3-knn-top14`, on peut commencer par visualiser le contenu du fichier `JoueursTop14.csv` dans un éditeur de texte ou un tableur.

1. Ouvrir le fichier `JoueursTop14.csv` dans un éditeur de texte. Observer rapidement la présentation de ce fichier.
2. Ouvrir le fichier avec un tableur. Choisir le séparateur approprié.
3. Supprimer les colonnes B (Nom), C (Poste) et E (Date de Naissance). On y voit plus clair en ne conservant que les données qui serviront dans l'activité.
4. Remarquer que la banque de données classe tous ces joueurs par équipe. Dans l'activité, nous conserverons toutes les équipes sauf Toulouse (la dernière alphabétique) pour les données de travail, et les joueurs toulousains seront les données de test.