Mini-rapport de projet

Lien du dataset : https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats

L'objectif du projet

A partir des statistiques sur une saison (ici, 2021-2022) on cherche le futur gagnant du Ballon d'Or, en considérant que ce dernier est celui qui a le plus d'impact dans le jeu.

- On est dans un apprentissage non supervisé : on ne sait pas si le joueur a gagné auparavant ou va gagner le Ballon d'Or.
- On cherche à classer les joueurs en fonction du Top X (Top 100, Top 50, Top 20, Top 10, …). Le Top 1 correspond ainsi au gagnant.

Nettoyage, conversion ou remplacement de certaines entrées

- suppression des lignes avec le nombre de matchs joués inférieur ou égale à 5
- les lignes pour lesquels il manque au moins une donnée
- suppressions des colonnes qui ne concernent pas directement le problème : le club, la nation, le championnat et l'âge

Encodage des données

- position : entier
- age du joueur : entier
- autres informations numériques : entier ou flottant

Suppression de certaines colonnes

- les données brutes qui ont des pourcentages (ex : nombre de passe et pourcentage de passe réussie)
- les statistiques trop précises (ex : nombre de tirs cadrés, nombre de deuxième carton jaune, nombre d'actions défensives qui ont mené vers une tentative de tir, …), elles concernent une grande partie des features.

Pré-traitement en dehors de l'encodage des données

- normalisation entre [0;1] ou [-1;1] (ex : pourcentage ramené à 1)

Choix d'algorithme

- k-moyennes (pour des clusters de même taille)
- Propagation d'affinité (pour des clusters de tailles différentes)

L'équilibre des classes à prédire est à prendre en compte.