

### **Justification du choix :**

Pour le projet n°8, j'ai décidé de reprendre un projet personnel, celui qui m'a amené à la data science. Il s'agit d'un travail concernant les paris sportifs, plus particulièrement les paris sur les matchs de football européens. La technique utilisée pourra être très facilement élargie plus tard à d'autres continents ou d'autres sports.

### **Plan prévisionnel du projet n°8.**

Le concept de base reprend un utilisateur qui souhaite améliorer des pronostics sur les matchs européens. Pour cela, il va s'appuyer sur les matchs déjà passés, et plus particulièrement sur les matchs « similaires » déjà déroulés. En essayant de regarder les résultats des matchs précédents (et autres statistiques précédant ces matchs), le programme que je souhaite créer doit améliorer une pari « de base » par une personne lambda.

Pour pouvoir comparer l'efficacité du programme, 4 baselines seront utilisées pour comparaison :

- Une personne qui parie uniquement sur les favoris des matchs.
- Une personne qui parie uniquement sur les underdogs (non-favori) des matchs.
- Une personne qui parie uniquement sur les matchs nuls.
- Une personne qui parie au hasard.

Ces 4 baselines, bien que très basiques, représentent un nombre important de pronostiqueurs « amateurs » dans le monde.

### **Algorithmes envisagés :**

Un certain nombre d'algorithmes de classification seront utilisés et confrontés :

- KNeighborsClassifier
- RandomForestClassifier
- GradientBoostingClassifier
- DecisionTreeClassifier
- RidgeClassifier
- LogisticRegression
- MultinomialNB

En fonction de la confrontation des résultats, d'autres pourront être rajoutés.

Les deux métriques importants seront le gain et l'accuracy des pronostics des résultats. Les deux très importants car on pourrait facilement augmenter l'accuracy au détriment du gain (en ne pariant que sur les favoris, qui rapportent peu). De même, on peut augmenter le gain au détriment de l'accuracy (en ne pariant que sur les underdogs, qui rapportent beaucoup mais ne gagnent que peu souvent).

### **Dataset utilisé :**

Un dataset sera construit à partir des résultats stockés sur un site de référence : <http://www.football-data.co.uk>.

Les données fournies ici le sont de manière brute. Une grande partie de data engineering sera nécessaire afin de pouvoir créer un dataset avec les données exploitables.

### **Références bibliographiques :**

Cette liste est non-exhaustive.

- Pattern Detection Applied to Soccer Results Forecast, Diogo Reis, Janvier 2018
- "Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data par Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr and Iain Matthews, 2015
- A Study on Soccer Prediction using Goals and Shots on Target par Snorre Gebhardt Stenerud, 2016