<https://www.kaggle.com/hugomathien/soccer>

The ultimate Soccer database for data analysis and machine learning

What you get:

* +25,000 matches
* +10,000 players
* 11 European Countries with their lead championship
* Seasons 2008 to 2016
* Players and Teams' attributes\* sourced from EA Sports' FIFA video game series, including the weekly updates
* Team line up with squad formation (X, Y coordinates)
* Betting odds from up to 10 providers
* Detailed match events (goal types, possession, corner, cross, fouls, cards etc...) for +10,000 matches

\**16th Oct 2016: New table containing teams' attributes from FIFA !*

Original Data Source:

You can easily find data about soccer matches but they are usually scattered across different websites. A thorough data collection and processing has been done to make your life easier. I must insist that you do not make any commercial use of the data. The data was sourced from:

* <http://football-data.mx-api.enetscores.com/> : scores, lineup, team formation and events
* <http://www.football-data.co.uk/> : betting odds. [Click here to understand the column naming system for betting odds:](http://www.football-data.co.uk/notes.txt)
* <http://sofifa.com/> : players and teams attributes from EA Sports FIFA games. *FIFA series and all FIFA assets property of EA Sports.*

When you have a look at the database, you will notice foreign keys for players and matches are the same as the original data sources. I have called those foreign keys "api\_id".

Improving the dataset:

You will notice that some players are missing from the lineup (NULL values). This is because I have not been able to source their attributes from FIFA. This will be fixed overtime as the crawling algorithm is being improved. The dataset will also be expanded to include international games, national cups, Champion's League and Europa League. Please ask me if you're after a specific tournament.

Please get in touch with me if you want to help improve this dataset.

[CLICK HERE TO ACCESS THE PROJECT GITHUB](https://github.com/hugomathien/football-data-collection/tree/master/footballData)

*Important note for people interested in using the crawlers:* since I first wrote the crawling scripts (in python), it appears sofifa.com has changed its design and with it comes new requirements for the scripts. The existing script to crawl players ('Player Spider') will not work until i've updated it.

Exploring the data:

Now that's the fun part, there is a lot you can do with this dataset. I will be adding visuals and insights to this overview page but please have a look at the kernels and give it a try yourself ! Here are some ideas for you:

The Holy Grail... ... is obviously to predict the outcome of the game. The bookies use 3 classes (Home Win, Draw, Away Win). They get it right about 53% of the time. This is also what I've achieved so far using my own SVM. Though it may sound high for such a random sport game, you've got to know that the home team wins about 46% of the time. So the base case (constantly predicting Home Win) has indeed 46% precision.

Probabilities vs Odds

When running a multi-class classifier like SVM you could also output a probability estimate and compare it to the betting odds. Have a look at your variance vs odds and see for what games you had very different predictions.

Explore and visualize features

With access to players and teams attributes, team formations and in-game events you should be able to produce some interesting insights into [The Beautiful Game](https://en.wikipedia.org/wiki/The_Beautiful_Game) . Who knows, Guardiola himself may hire one of you some day!