**FC INTERNAZIONALE DATA SCIENCE CHALLENGE:**

**Chi sono i giocatori più affini a Francesco Acerbi sul mercato? Un approccio attraverso dati e modelli di similarità**

**Davide Prati, Università degli studi di Milano Bicocca, CDLM in Data Science**

Il progetto parte dall'obiettivo di soddisfare una richiesta: fornire i 5 giocatori che siano più simili a Francesco Acerbi. Per farlo, una volta raccolti e preparati i dati adeguati, le caratteristiche di Acerbi saranno valutate in relazione a quelle di più di altri 800 giocatori, dei quali verrà studiata tramite opportuni algoritmi la similarità col centrale dell'Inter.

In particolare io ho voluto interpretare questo progetto in un'ottica anche il più possibile realistica per quanto riguarda il calciomercato, quindi buona parte delle scelte e strategie sviluppate nel corso dello svolgimento terranno ben presente questo aspetto.

Nella cartella Google Drive sono presenti i seguenti files:

* **Report.pdf** : il report scritto in LaTEX con un formato su due colonne, che riassume il lavoro svolto, contenente una sezione per ogni notebook.
* I seguenti notebooks Google Colab, in cui si trova il codice:
* **1- Data Retrieval.ipynb**: ottenimento dei dataset di partenza, o con un semplice download e caricamento su notebook oppure con tecniche di scraping;
* **2- Data Preparation.ipynb**: insieme di operazioni di concatenazione dei dataset, pulizia, restringimento, feature selection, merging;
* **3- Data Modelling.ipynb:** dopo una fase di normalizzazione e pesatura, in questo notebook ottengo i risultati, con tecniche specifiche e bilanciando opportunamente i dati;
* **4- Data Visualization.ipnyb**: presentazione di grafici - in particolare scatterplots - per mettere in risalto le caratteristiche del giocatore di riferimento e di quelli scelti rispetto alla distribuzione della popolazione obiettivo completa.
* file csv scaricati, creati e usati nel corso del progetto:
* **Stats22-23.csv**: dataset ottenuto da Kaggle sulla stagione 2022/2023
* **Stats21-22.csv**: dataset ottenuto da Kaggle sulla stagione 2021/2022
* **df\_values.csv**: dataset ottenuto dallo scraping
* **df\_finale.csv**: dataset risultante della fase di preparation.
* **df\_per\_viz.csv**: dataset con gli stessi dati di df\_finale, ma con l’aggiunta di score creati durante la fase di modelling e da riportare in fase di visualization
* **df\_simili.csv**: dataset che è semplicemente un restringimento di df\_per\_viz, ad Acerbi e ai 5 giocatori a lui più simili.

Tutti i pacchetti e le librerie vengono caricati all'inizio dei file di codice. Nel caso non siano già installati, possono essere installati utilizzando la procedura standard.